Analysis II for INFK

E. Kowalski

ETH Zürich – Fall 2019
Version of February 19, 2023
kowalski@math.ethz.ch

Contents

CHAPTER 1

Preliminaries

We assume knowledge of the contents of Analysis I (for instance, properties of real numbers, sequences and series, continuous functions on intervals, differentiable functions on intervals, Riemann integral, improper integrals). These can be found in the script [1] of Prof. M. Burger.

We denote by $\mathbf{N} = \{0, 1, 2, \ldots\}$ the set of all natural numbers, by $\mathbf{Q}$ the rational numbers, by $\mathbf{R}$ the real numbers and by $\mathbf{C}$ the complex numbers.

In addition to continuous and differentiable functions defined on intervals with values in $\mathbf{R}$, as in [1, Kap. III, IV], we will also consider functions $f\colon I \to \mathbf{R}^d$, where $d \geqslant 2$ and I is an interval. This means that

$$f(x) = (f_1(x), \ldots, f_d(x))$$

for some functions $f_i\colon I \to \mathbf{R}$. We then say:

- That f is continuous (on I, or at a point $x_0 \in I$) if each coordinate function f_i is continuous (on I or at x_0);
- That f is differentiable (on I, or at a point $x_0 \in I$) if each coordinate function f_i is differentiable (on I or at x_0), in which case we write

$$f'(x_0) = (f_1'(x_0), \ldots, f_d'(x_0)).$$

A primitive F of a continuous function $f\colon I \to \mathbf{R}^d$ is a differentiable function $F\colon I \to \mathbf{R}^d$ such that $F' = f$. A primitive always exists, for instance

$$F(x) = \Big(\int_{x_0}^{x} f_1(t)dt, \ldots, \int_{x_0}^{x} f_d(t)dt\Big)$$

when writing $f = (f_1, \ldots, f_d)$ as before.

Acknowledgments. Thanks to T. Rieder for sending a number of corrections.

CHAPTER 2

Ordinary differential equations

2.1. Introduction

A *differential equation* is an equation where the unknown (or unknowns) is a function f, and the equation relates values of f at a point x with values of derivatives of the function at the *same point* x. If the function has one variable only (as is the case in this chapter), one speaks of *ordinary differential equations.*[1]

EXAMPLE 2.1.1. (1) The exponential function $f(x) = e^x$, defined for $x \in \mathbf{R}$, satisfies $f'(x) = f(x)$ for all $x \in \mathbf{R}$. One says that this function is a solution (on $\mathbf{R}$) of the differential equation $y' = y$. This is not the only solution: in fact, for any constant $a \in \mathbf{R}$, the function $f_a(x) = ae^x$ also satisfies $f'_a(x) = f_a(x)$ for all $x \in \mathbf{R}$. Later, we will see that there are no other solutions.

(2) In the mechanics of Newton, the movement of a particle P with mass $m > 0$, given by its position $f(t) = (x(t), y(t), z(t)) \in \mathbf{R}^3$ for all times t is determined by the equation

$$mf''(t) = \text{sum of forces acting on } P \text{ at time } t,$$

and by the "initial condition", which means the specification of the position $f(0)$ and speed $f'(0)$ at some starting time t_0. Note that the forces acting on the particle at time t are expressions involving $f(t)$ (position) and $f'(t)$ (speed), at the *same* time t. Also, the solution is unique because of the initial conditions (otherwise, as in Example (1), there would be infinitely many solutions).

Since f is a function with one variable t but with values in $\mathbf{R}^3$, we recall again that the derivatives of f are simply taken for each coordinate separately

$$f'(t) = (x'(t), y'(t), z'(t)), \qquad f''(t) = (x''(t), y''(t), z''(t)).$$

For example, a particle subject to no external force satisfies the equation $mf''(t) = 0$, so that $f''(t) = 0$ for all t, which means that the motion is a straight line (each of the coordinates is of the form $x(t) = a_0t + b_0$, $y(t) = a_1t + b_1$, $z(t) = a_2t + b_2$).

Classical newtonian mechanics (and its solutions) forms a basic tool to simulate physical behavior of objects in applications (such as computer games, computer generated videos, etc).

(3) For any given continuous function a, the differential equation $f' - a = 0$ has a solution, namely any primitive of the function a. The existence of the solutions follows from the Fundamental Theorem of Calculus ([1, §5.4]): we may define

$$f(x) = \int_{x_0}^{x} a(t)dt.$$

In general, differential equations are closely related with integration theory.

(4) An equation like $f'(x + 1) - f(x) = 0$ is *not* an ordinary differential equation, because it relates the value of f at the point x with the derivative at *another* point.

[1] When there is more than one variable, one speaks of *partial differential equations*, referring to partial derivatives in multi-variable calculus (see Chapter 3).

Remark 2.1.2. In computer science, besides simulations of physical systems, differential equations arise frequently in the analysis of algorithms, for instance the running time of certain algorithms might be a solution of a differential equation, or might be approximated by such a solution.

It is customary to write down a differential equation without writing the evaluation $f(t)$ or $f(x)$ but only the function's name (or its derivatives), and to use the letter x or t for the variable when it appears elsewhere in the equation. In physics, the time variable t plays an important role, and derivation with respect to time is often denoted by a dot: $\dot{y}$ instead of y', and $\ddot{y}$ instead of y''. When one needs to specify initial conditions (values of the unknown function at some fixed value that specify the solution uniquely), one writes for instance $y(0) = a$, $y'(0) = b$ to say that the function f should satisfy $f(0) = a$ and $f'(0) = b$.

Example 2.1.3. The important function $f(x) = e^{-x^2}$, for $x \in \mathbf{R}$, satisfies the differential equation $y' = -2xy$, i.e., for every $x \in \mathbf{R}$, we have $f'(x) = -2xe^{-x^2} = -2xf(x)$. In a physics context, where the variable is understood as time, this might be written $\dot{y} = -2ty$.

Ordinary differential equations are classified according to their *order*, which is the highest derivative that appears in the equation. So Newton's equations are of order 2 (because forces are expressed in terms of y and y', and acceleration involves y'').

There is a trick to reduce any ordinary differential equation of order $k \geqslant 2$ to an equation of order 1, but for a function that takes values in a higher-dimensional space (keeping however a single variable). We illustrate it with an example.

Example 2.1.4. The differential equation

$$y'' = x(x+1)y' - 3y, \tag{2.1}$$

with unknown a differentiable function $f\colon \mathbf{R} \to \mathbf{R}$, can be transformed into the equation

$$Y' = \begin{pmatrix} 0 & 1 \\ -3 & x(x+1) \end{pmatrix} Y \tag{2.2}$$

with unknown a differentiable function $F\colon \mathbf{R} \to \mathbf{R}^2$, where the right-hand side is a matrix product. Indeed, if F is a solution of this equation and we write $F(x) = \begin{pmatrix} f_0(x) \\ f_1(x) \end{pmatrix}$, then the equation for F means that

$$\begin{pmatrix} f_0'(x) \\ f_1'(x) \end{pmatrix} = \begin{pmatrix} 0 & 1 \\ -3 & x(x+1) \end{pmatrix} \begin{pmatrix} f_0(x) \\ f_1(x) \end{pmatrix} = \begin{pmatrix} f_1(x) \\ -3f_0(x) + x(x+1)f_1(x) \end{pmatrix}$$

for all x. So we have $f_1 = f_0'$, and therefore $F(x) = \begin{pmatrix} f_0(x) \\ f_0'(x) \end{pmatrix}$, where the function $f_0\colon \mathbf{R} \to \mathbf{R}$ satisfies (second row of the equation) the ordinary differential equation

$$f_0''(x) = f_1'(x) = -3f_0(x) + x(x+1)f_0'(x)$$

for all x. Conversely, given a solution f of (2.1), putting $F(x) = \begin{pmatrix} f(x) \\ f'(x) \end{pmatrix}$ gives a solution of (2.2).

This trick explains why many general results are stated for equations of order 1. On the other hand, the solution of specific equations of order 2 or higher might be easier without using this trick.

Although it is “physically” clear that Newton’s equation have solutions, it is not at all obvious that ordinary differential equations should have solutions in general. In fact, it is possible that a solution only exists “locally” around an initial point.

Example 2.1.5. Consider the equation $2yy' = 1$ on $\mathbf{R}$ with the initial condition $y(0) = 1$. Writing the left-hand side as $(y^2)'$, we see that y^2 satisfies $y^2 = x + a$ for some constant $a \in \mathbf{R}$, and we have $a = 1$ because of the initial condition. Hence the solution is $f(x) = \sqrt{x+1}$. But although the equation can be asked for all $x \in \mathbf{R}$, here the solution only makes sense for $x > -1$.

At least one can prove that this local existence always holds, for nice enough equations of the form $y' = F(x, y)$ (and many others that can be brought to this form).

Theorem 2.1.6. *Suppose $F\colon \mathbf{R}^2 \to \mathbf{R}$ is a differentiable function of two variables* (see Chapter 3). *Let $x_0 \in \mathbf{R}$ and $y_0 \in \mathbf{R}^2$. Then the ordinary differential equation*

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

has a unique solution f defined on a “largest” open interval I containing x_0 such that $f(x_0) = y_0$. In other words, there exists I and a function $f\colon I \to \mathbf{R}$ such that for all $x \in I$, we have $f'(x) = F(x, f(x))$, and one cannot find a larger interval containing I with such a solution.

As is the case for polynomial equations, it is in general impossible to write down “explicitly” the solution to such an equation.

Example 2.1.7. The function F can be arbitrary, for instance

$$F(t, u) = u^3 \exp(\cos(tu^2 - 1)) + 3\sin(t)$$

for the complicated differential equation

$$y' = y^3 \exp(\cos(xy^2 - 1)) + 3\sin(x)$$

whose solutions $f\colon I \to \mathbf{R}$ (for some interval I) satisfy

$$f'(x) = f(x)^3 \exp(\cos(xf(x)^2 - 1)) + 3\sin(x) = 0,$$

for all $x \in I$.

2.2. Linear differential equations

The simplest differential equations are the *linear differential equations*.

Definition 2.2.1. Let $I \subset \mathbf{R}$ be an open interval and $k \geqslant 1$ an integer. An *homogeneous linear ordinary differential equation* of order k on I is an equation of the form

$$y^{(k)} + a_{k-1}y^{(k-1)} + \cdots + a_1 y' + a_0 y = 0$$

where the coefficients $a_0, \ldots, a_{k-1}$ are complex-valued *functions* on I, and the unknown is a complex-valued function from I to $\mathbf{C}$ that is k-times differentiable on I.

An equation of the form

$$y^{(k)} + a_{k-1}y^{(k-1)} + \cdots + a_1 y' + a_0 y = b, \tag{2.3}$$

where $b\colon I \to \mathbf{C}$ is another function, is called an *inhomogeneous linear ordinary differential equation*, with associated homogeneous equation the one with $b = 0$.

Note that if the coefficients are real-valued, it is often of interest to find only the *real-valued* solutions.

Example 2.2.2. The equations

$$y' = y, \qquad y'' = -y, \qquad y' + 2xy = 0$$

(which admit as particular solutions the functions $\exp(x)$, $\cos(x)$ and $\exp(-x^2)$, respectively) are linear and homogeneous. The equation

$$y' - y = \cos(x)$$

is linear and inhomogeneous. The equations

$$2yy' = 1, \qquad y' = y^2, \qquad \cos(y'') = \exp(x+y), \qquad (y+y')^3 = 1, \qquad y' - y = xe^y$$

are not linear.

The main property of linear differential equations (explaining the adjective) is that if we write $D(f)$ for the left-hand side of the equation, so that

$$D(f) = f^{(k)} + a_{k-1}f^{(k-1)} + \cdots + a_1 f' + a_0 f,$$

then the operation D is linear: for any numbers z_1 and z_2 and (k-times differentiable) functions f_1 and f_2, we have $D(z_1 f_1 + z_2 f_2) = z_1 D(f_1) + z_2 D(f_2)$. Indeed, let $f = z_1 f_1 + z_2 f_2$, then

$$\begin{aligned} D(f) &= f^{(k)} + \cdots + a_1 f' + a_0 f \\ &= z_1(f_1^{(k)} + a_{k-1}f^{(k-1)} + \cdots + a_0 f_1) + z_2(f_2^{(k)} + a_{k-1}f^{(k-1)} \cdots + a_0 f_2) \\ &= z_1 D(f_1) + z_2 D(f_2). \end{aligned}$$

The main theoretical results concerning linear differential equations are summarized in the following result:

Theorem 2.2.3. *Let $I \subset \mathbf{R}$ be an open interval and $k \geqslant 1$ an integer, and let*

$$y^{(k)} + a_{k-1}y^{(k-1)} + \cdots + a_1 y' + a_0 y = 0$$

be a linear differential equation over I with continuous coefficients.

(1) *The set $\mathcal{S}$ of k-times differentiable solutions $f\colon I \to \mathbf{C}$ of the equation is a complex vector space which is a subspace of the space of complex-valued functions on I.*

(1bis) *If the functions a_i are real-valued, the set $\mathcal{S}$ of real-valued solutions is a real vector space which is a subspace of the space of real-valued functions on I.*

(2) *The dimension of $\mathcal{S}$ is k, and for any choice of $x_0 \in I$ and any $(y_0, \ldots, y_{k-1}) \in \mathbf{C}^k$, there exists a unique $f \in \mathcal{S}$ such that*

$$f(x_0) = y_0, \quad f'(x_0) = y_1, \quad \ldots, \quad f^{(k-1)}(x_0) = y_{k-1}.$$

(2bis) *If the functions a_i are real-valued, the dimension of the space of real-valued solutions, as a real vector space, is k, and for any choice of $x_0 \in I$ and any $(y_0, \ldots, y_{k-1}) \in \mathbf{R}^k$, there exists a unique real-valued solution f such that*

$$f(x_0) = y_0, \quad f'(x_0) = y_1, \quad \ldots, \quad f^{(k-1)}(x_0) = y_{k-1}.$$

If b and the coefficients a_i are real-valued, there exists a real-valued solution.

(3) *Let b be a continuous function on I. There exists a solution f_0 to the inhomogeneous equation*

$$y^{(k)} + a_{k-1}y^{(k-1)} + \cdots + a_1 y' + a_0 y = b,$$

and the set $\mathcal{S}_b$ is the set of functions $f + f_0$ where $f \in \mathcal{S}$.

(4) *For any $x_0 \in I$ and any $(y_0, \ldots, y_{k-1}) \in \mathbf{C}^k$, there exists a unique $f \in \mathcal{S}_b$ such that*

$$f(x_0) = y_0, \quad f'(x_0) = y_1, \quad \ldots, \quad f^{(k-1)}(x_0) = y_{k-1}.$$

Remark 2.2.4. (1) If $b \neq 0$, the set $\mathcal{S}_b$ of solutions is *not* a vector space.

(2) Statement (1) of this theorem is elementary: the set $\mathcal{S}$ is just the kernel of the linear map that sends a function f to $D(f)$. In other words, if f_1 and f_2 are elements of $\mathcal{S}$ and z_1, z_2 are complex numbers and $f = z_1 f_1 + z_2 f_2$, then

$$D(f) = z_1 D(f_1) + z_2 D(f_2) = 0.$$

Also, if we can find any element f_0 of the set $\mathcal{S}_b$, then it is elementary that all other elements are of the form $f + f_0$ where $\mathcal{S}_b$, since for $f_1 \in \mathcal{S}_b$, we get

$$(f_1 - f_0)^{(k)} + a_{k-1}(f_1 - f_0)^{(k-1)} + \cdots + a_0(f_1 - f_0) \\ = D(f_1 - f_0) = D(f_1) - D(f_0) = b - b = 0,$$

so that $f_1 = f + f_0$ where $f = f_1 - f_0 \in \mathcal{S}$.

We will illustrate this result in the next sections by explaining how to solve, in practice, two important types of linear differential equations.

Remark 2.2.5. The linearity of the equation has also consequences when trying to solve the inhomogeneous equation. Indeed, for instance, if we know a function f_1 solving (2.3) with the right-hand side b_1, and one function f_2 solving (2.3) with the right-hand side b_2, then $f_1 + f_2$ solves (2.3) with right-hand side $b_1 + b_2$, since $D(f_1 + f_2) = D(f_1) + D(f_2) = b_1 + b_2$.

2.3. Linear differential equations of order 1

Let $I \subset \mathbf{R}$ be an open interval. We consider here the linear differential equation

$$y' + ay = b,$$

when a and b are general continuous functions defined on I.

The solution has two steps: first solving the homogeneous equation $y' + ay = 0$ (say that $\mathcal{S}$ is the space of solutions, which is a one-dimensional vector space according to Theorem 2.2.3), and then finding a solution f_0 of the inhomogeneous equation, so that the set $\mathcal{S}_b$ contains exactly the functions $f_0 + f$ where $f \in \mathcal{S}$. If f_1 is a basis of $\mathcal{S}$ (which only means that f_1 is in $\mathcal{S}$ and is not the zero function), this means that the solutions are given by $f_0 + zf_1$, where $z \in \mathbf{C}$ is arbitrary.

If a is real-valued, then there exists a real-valued non-zero element f_1 of $\mathcal{S}$, and a real-valued solution f_0, so that the real-valued solutions of the equation are the functions of the form $f_0 + xf_1$, where $x \in \mathbf{R}$ is arbitrary.

If one wishes the solve the problem with initial value $f(x_0) = y_0$, then it suffices to solve the equation

$$f_0(x_0) + zf_1(x_0) = y_0$$

to determine the value of z. (It might be thought that there is a problem if $f_1(x_0) = 0$, but we will see that this never happens for a non-zero function $f_1 \in \mathcal{S}$).

Step 1 (solving the homogeneous equation). Formally, the idea is to transform $y' + ay = 0$ into $y'/y = -a$, so that $(\log |y|)' = -a$, which implies by integration that

$$y = z \exp(-A),$$

where $z \in \mathbf{C}$ and A is a primitive of the function a. There is a potential problem with this argument, since we divided by y, which is a function so that y might vanish at some point in the interval. However, it is easy to check that the conclusion is correct. First, if we define a function $f(x) = z\exp(-A(x))$, then we get by the chain rule the relation

$$f'(x) = -zA'(x)\exp(-A(x)) = -a(x)f(x),$$

so that f is a solution of the equation $y' + ay = 0$.

Conversely, suppose that f is a solution of the equation $y' + ay = 0$, and define $g(x) = f(x)\exp(A(x))$. Then we obtain, by the Leibniz rule and the chain rule, the formula

$$\begin{aligned} g'(x) &= f'(x)\exp(A(x)) + A'(x)f(x)\exp(A(x)) \\ &= -a(x)f(x)\exp(A(x)) + a(x)f(x)\exp(A(x)) = 0, \end{aligned}$$

which means that g is a constant, say z, in which case $f(x) = z\exp(-A(x))$, as we guessed.

We conclude:

Proposition 2.3.1. *Any solution of $y' + ay = 0$ is of the form $f(x) = z\exp(-A(x))$ where A is a primitive of a. The unique solution with $f(x_0) = y_0$ is*

$$f(x) = y_0\exp(A(x_0) - A(x)).$$

Step 2 (solving the inhomogeneous equation). Now consider the equation

$$y' + ay = b.$$

We know that it suffices to find a single solution f_0 to obtain all of them by adding one of the solutions of $y' + ay = b$ found in the previous step.

Example 2.3.2. Sometimes, we can make a clever guess that finds a suitable function f_0. Consider for instance the equation $y' = y + x^2$. We might guess that a polynomial can be a solution; it should be of degree 2, so we can try $f(x) = ax^2 + bx + c$, for some constants a, b, c. In that case we get

$$f'(x) - f(x) = 2ax + b - (ax^2 + bx + c) = -ax^2 + (2a - b)x + b - c$$

so this function is a solution provided

$$\begin{cases} a & = -1 \\ 2a - b & = 0, \text{ hence } b = -2 \\ b - c & = 0, \text{ hence } c = -2. \end{cases}$$

So we can take $f_0(x) = -x^2 - 2x - 2$.

If there is no obvious guess of the form of a special solution f_0, there is a general method that works (but might lead to complicated formulas). It is called "variation of the constant", because it starts with the formula for a solution of the homogeneous equation, namely

$$f(x) = z\exp(-A(x)), \qquad z \in \mathbf{C},$$

and looks for a solution f_0 of this form, but where now z is considered to be itself a *function* of x. If we assume this, and compute the derivative f', then we see that $f_0(x) = z(x)\exp(-A(x))$ is a solution of $y' + ay = b$ if and only if

$$z'(x)\exp(-A(x)) - A'(x)z(x)\exp(-A(x)) + a(x)z(x)\exp(-A(x)) = b(x),$$

which (since $A'(x) = a(x)$) translates into

$$z'(x) = b(x)\exp(A(x)).$$

In other words, we can take z to be a primitive $C(x)$ of the continuous function $b(x)\exp(A(x))$, and the special solution is

$$f_0(x) = C(x)\exp(-A(x)).$$

If we use the fundamental theorem of calculus to write primitives (taking the value 0 at x_0) of a and $b\exp(A)$, this becomes the rather complicated expression

$$f_0(x) = \exp\Bigl(-\int_{x_0}^{x} a(t)dt\Bigr)\int_{x_0}^{x} b(t)\exp\Bigl(\int_{x_0}^{t} a(u)du\Bigr)dt,$$

which is a special solution such that $f_0(x_0) = 0$.

Remark 2.3.3. When solving concrete equations, *do not forget* the last step of multiplying the "constant" $z(x)$ by $\exp(-A(x))$ at the end!

2.4. Linear differential equations with constant coefficients

Let $k \geqslant 1$ be an integer, and let $a_0, \ldots, a_{k-1}$ be complex *constant* coefficients. We consider the linear differential equation

$$y^{(k)} + a_{k-1}y^{(k-1)} + \cdots + a_1y' + a_0y = b.$$

Note that the coefficients a_i are fixed numbers, but the right-hand side b is still assumed to be a general continuous function.

Example 2.4.1. The equation $y'' - xy = 0$ does *not* belong to this class, but the equations $y' - y = 0$ and $y'' + y = 0$ (satisfied by the exponential and by trigonometric functions) have constant coefficients.

The solution of the *homogeneous equation* is very simple in principle. One looks for solutions of the special form $f(x) = e^{\alpha x}$ for some complex number $\alpha \in \mathbf{C}$. Then we have $f^{(j)}(x) = \alpha^j e^{\alpha x}$ for all $j \geqslant 0$ and for all x, which means that

$$f^{(k)}(x) + a_{k-1}f^{(k-1)}(x) + \cdots + a_1f'(x) + a_0f(x) = e^{\alpha x}(\alpha^k + a_{k-1}\alpha^{k-1} + \cdots + a_1\alpha + a_0).$$

We conclude that f is a solution of the homogeneous equation if and only if $P(\alpha) = 0$, where P is the polynomial with coefficients $a_0, \ldots, a_{k-1}$:

$$P(X) = X^k + a_{k-1}X^{k-1} + \cdots + a_1X + a_0.$$

According to the Fundamental Theorem of Algebra, this polynomial of degree k has k complex roots, counted with multiplicity: there exist complex numbers $\alpha_1, \ldots, \alpha_k$ such that

$$P(X) = (X - \alpha_1)\cdots(X - \alpha_k).$$

This polynomial is called the *companion* or *characteristic polynomial* of the homogeneous differential equation.

Remark 2.4.2. We repeat that this is only defined when the coefficients of the equation are constant.

Remark 2.4.3. Although it is natural to look for complex-valued solutions, one is often interested in situations where the coefficients a_i are real and we know that the solution should take real values, or we want such solutions.

Suppose that a root $\alpha = \beta + i\gamma$ is not real, so the imaginary part γ is non-zero. Then the solution $f(x) = e^{\alpha x}$ does not take real values. However, in that case, the conjugate

$\beta - i\gamma = \bar{\alpha} \neq \alpha$ is also a root of the companion polynomial (which has real coefficients, so that $P(\bar{z}) = \overline{P(z)}$ for any $z \in \mathbf{C}$) and one can replace the two solutions

$$f_1(x) = e^{\alpha x}, \qquad f_2(x) = e^{\bar{\alpha} x}$$

by the real-valued functions

$$\widetilde{f}_1(x) = e^{\beta x}\cos(\gamma x), \qquad \widetilde{f}_2(x) = e^{\beta x}\sin(\gamma x)$$

(note for instance that $f_1 = \widetilde{f}_1 + i\widetilde{f}_2$ and that $\widetilde{f}_1 = f_1 + f_2$ since $e^{i\theta} = \cos(\theta) + i\sin(\theta)$).

The possible existence of multiple roots requires some care in the next step, so we begin by discussing the simple case where this does not happen.

Case 1: no multiple roots. Assume that $\alpha_i \neq \alpha_j$ for $i \neq j$. Then we have found k distinct solutions $f_j(x) = e^{\alpha_j x}$ of the homogeneous equation. It is not very difficult to check that these functions are linearly independent, so that the space of linear combinations of these functions has dimension k. According to Theorem 2.2.3, (2), this must be the full vector space $\mathcal{S}$ of solutions of the homogeneous linear differential equation. In other words, any solution of

$$y^{(k)} + a_{k-1}y^{(k-1)} + \cdots + a_1 y' + a_0 y = 0$$

is of the form

$$f(x) = z_1 e^{\alpha_1 x} + \cdots + z_k e^{\alpha_k x},$$

for some complex numbers $(z_1, \ldots, z_k)$ that can be chosen arbitrarily.

If one wishes to find the unique solution with

$$f(x_0) = y_0, \ldots, f^{(k-1)}(x_0) = y_{k-1}$$

for given $(y_0, \ldots, y_{k-1})$, one may simply view z_1, …, z_k as unknowns. Substituting $x = x_0$ in the formula for $f(x)$ and solving for these initial conditions becomes a linear system with unknowns z_1, …, z_k. It is a fact that the system has a unique solution (the determinant is always non-zero), which provides the required function.

Remark 2.4.4. If the constants a_i are real, the space of real-valued solutions of the equation is obtained as follows: order the roots α_j so that α_1, …, α_m are the real solutions of the polynomial P, and α_{m+1}, …, α_k are the solutions which are not real. Write $\alpha_j = a_j + ib_j$ for $j \geqslant m + 1$. (Note that we may have $m = 0$, if there is no real solution, or $m = k$, if all solutions are real). Then the space of real-valued solutions of the homogeneous differential equation is the space of functions of the form

$$\begin{aligned} f(x) = x_1 e^{\alpha_1 x} + \cdots + x_m e^{\alpha_m x} + \\ x_{m+1}e^{a_{m+1}x}\cos(b_{m+1}x) + y_{m+1}e^{a_{m+1}x}\sin(b_{m+1}x) + \\ \cdots + x_k e^{a_k x}\cos(b_k x) + y_k e^{a_k x}\sin(b_k x). \end{aligned}$$

Because such expressions are more complicated to handle, it is often better to work with complex-valued solutions as long as possible.

Example 2.4.5. (1) Consider the equation $y' + ay = 0$, with a constant. The companion polynomial is $X + a$, so the only solution is $\alpha_1 = -a$, and we get the solutions $f(x) = ze^{-ax}$. This coincides with the solution in Section 2.3, since a primitive of a is $A(x) = ax$.

(2) Consider the equation $y'' - ay = 0$. The companion polynonial is $P = X^2 - a$. There are then three cases.

- (Case 1). If $a > 0$, then $P = (X - \sqrt{a})(X + \sqrt{a})$ has two real roots, and the solutions take the form
$$f(x) = z_1 e^{\sqrt{a}\,x} + z_2 e^{-\sqrt{a}\,x}.$$
- (Case 2). If $a < 0$, then $P = (X - i\sqrt{|a|})(X + i\sqrt{|a|})$, and the solutions take the form
$$\begin{aligned} f(x) &= z_1 e^{i\sqrt{|a|}\,x} + z_2 e^{-i\sqrt{|a|}\,x} \\ &= (z_1 + z_2)\cos(\sqrt{|a|}\,x) + i(z_1 - z_2)\sin(\sqrt{|a|}\,x). \end{aligned}$$
- (Case 3). If $a = 0$, then we have only found one solution (namely $f(x) = 1$). However, the equation is easily solved in that case: f is a solution of $y'' = 0$ means that $f(x) = z_1 x + z_2$ for some complex numbers z_1 and z_2. So the function $f_1(x) = x$ is a second solution linearly independent of the first.

(3) What is the solution f to $y'' + y' + y = 0$ such that $f(0) = 1$ and $f'(0) = 0$?

The companion polynomial is $P(X) = X^2 + X + 1 = (X - \alpha)(X - \bar{\alpha})$ with $\alpha = (-1 + i\sqrt{3})/2$. Since we are interested in real solutions, it is easier to work with the two basic solutions
$$f_1(x) = e^{-x/2}\cos\Bigl(\frac{\sqrt{3}}{2}x\Bigr), \qquad f_2(x) = e^{-x/2}\sin\Bigl(\frac{\sqrt{3}}{2}x\Bigr).$$
We know that there exist numbers z_1 and z_2 such that
$$f(x) = z_1 f_1(x) + z_2 f_2(x),$$
and the initial conditions transform into the linear equations
$$\begin{cases} z_1 & = 1 \\ -\frac{1}{2}z_1 + \frac{\sqrt{3}}{2}z_2 & = 0 \end{cases}$$
for z_1 and z_2 (since, for instance, we have
$$f_1'(x) = -\frac{1}{2}e^{-x/2}\cos\Bigl(\frac{\sqrt{3}}{2}x\Bigr) - \frac{\sqrt{3}}{2}e^{-x/2}\sin\Bigl(\frac{\sqrt{3}}{2}x\Bigr),$$
so $f_1'(0) = -1/2$, and similarly for $f_2'(0) = \sqrt{3}/2$). It follows that $z_1 = 1$ and $z_2 = 1/\sqrt{3}$.

Case 2 (multiple roots). Suppose that α is a multiple root of order j of the polynomial P, with $2 \leqslant j \leqslant k$. Then the j functions
$$f_{\alpha,0}(x) = e^{\alpha x}, \quad f_{\alpha,1}(x) = xe^{\alpha x}, \quad \cdots, \quad f_{\alpha,j-1}(x) = x^{j-1}e^{\alpha x}$$
are linearly independent, and are solutions of the homogeneous linear differential equation. Taking the union of the functions $f_{\alpha,j}$ for all roots of P, each with its multiplicity, gives a basis of the space of solutions.

Remark 2.4.6. To say that α is a root of P with multiplicity $j \geqslant 1$ means either of the following two equivalent conditions:

(1) We have $P(\alpha) = \cdots = P^{(j-1)}(\alpha) = 0$.

(2) We have a factorization $P(X) = (X - \alpha)^j Q(X)$, where Q is a polynomial and $Q(\alpha) \neq 0$.

We now check the assertion about $f_{\alpha,j}$ being a solution in the case of a double root ($j = 2$). Note that
$$f_1'(x) = \alpha x e^{\alpha x} + e^{\alpha x}, \qquad f_1''(x) = \alpha^2 x e^{\alpha x} + 2\alpha e^{\alpha x}, \qquad \ldots$$

so that we find the formula

$$f_1^{(k)}(x) + a_{k-1}f_1^{(k-1)}(x) + \cdots + a_1 f_1'(x) + a_0 f_1(x) = xe^{\alpha x}P(\alpha) + e^{\alpha x}P'(\alpha).$$

Since $P(\alpha) = P'(\alpha) = 0$ for a double root, the function f_1 is a solution of the homogeneous differential equation. The general case is similar.

Once a basis of $\mathcal{S}$ is found (using this kind of functions for each root), one can find the unique solution with given initial conditions by again substituting x_0 in a linear combination, and solving a system of linear equations.

Example 2.4.7. Suppose that the companion polynomial factors as

$$P(X) = X(X-4)^3(X-(1+i))(X-(1-i)).$$

Then a basis of the solution space $\mathcal{S}$ are the functions

$$f_0(x) = 1 \text{ (for the solution 0 of } P)$$

$$f_1(x) = e^{4x}, \quad f_2(x) = xe^{4x} \quad f_3(x) = x^2e^{4x} \text{ (for the solution 4, which is a triple root)}$$

$$f_4(x) = e^{(1+i)x} = (\cos(x) + i\sin(x))e^x, \quad f_5(x) = e^{(1-i)x} = (\cos(x) - i\sin(x))e^x.$$

If one is interested in real-valued solutions, it might be easier to use the alternate basis where f_4 and f_5 are replaced by

$$\widetilde{f}_4(x) = e^x\cos(x), \quad \widetilde{f}_5(x) = e^x\sin(x).$$

We now go back to the general case. If we need to solve an inhomogeneous equation, there remains to find a special solution for

$$y^{(k)} + a_{k-1}y^{(k-1)} + \cdots + a_1y' + a_0y = b. \tag{2.4}$$

There are some useful tricks that can be used to avoid the analogue of the method of variation of constants, which is often rather complicated to implement (as we will see below). The first is Remark 2.2.5 (following from linearity). The second is that there are special cases of right-hand sides b where one can search explicitly for solutions of a special form. The most important are the following:

(1) If $b(x) = x^de^{\beta x}$ for some integer $d \geqslant 0$ and some number β which is *not* a root of the companion polynomial P, then one looks for a solution of the form $f(x) = Q(x)e^{\beta x}$, where Q is a polynomial of degree d.

(2) If $b(x) = x^d\cos(\beta x)$ or $b(x) = x^d\sin(\beta x)$ for some integer $d \geqslant 0$ and some number β which is not a root of the companion polynomial P, then one can either transform it to a combination of complex exponentials (and apply linearity and (1)), or one may look for a solution of the form

$$f(x) = Q_1(x)\cos(\beta x) + Q_2(x)\sin(\beta x),$$

where Q_1 and Q_2 are polynomials of degree d.

(3) If b is of the form of the previous two examples but where β is a root of multiplicity j of the companion polynomial, then one looks for $f(x) = Q(x)e^{\beta x}$ (or the analogue with cosine and sine), but where Q has degree $d + j$.

(4) The special case $\beta = 0$ of (1), (2), (3) corresponds to the situation when b is a polynomial of degree $d \geqslant 0$. So one should search for a polynomial solution f of the same degree d, unless 0 is a root of the companion polynomial, in which case one should look for a polynomial of degree $d + j$, where j is the multiplicity of 0 as a root of P.

Example 2.4.8. (1) We can illustrate Example (3) (and in fact remember the way it works) by considering the case where $P(X) = X^j$, so that $\alpha = 0$ is a root of order j. The equation is $y^{(j)} = b$, and if $b(x) = x^d e^{\alpha x} = x^d$, then a solution is

$$f(x) = \frac{1}{(d+1)\cdots(d+j)} x^{d+j},$$

which is indeed a polynomial of degree $d + j$.

(2) Consider the equation $y'' + 3y' + y = 3x^2 + \cos(x)$. Here we use linearity to find a special solution: a solution is $f = 3f_1 + f_2$, where f_1 is a solution of $y'' + 3y' + y = x^2$ and f_2 is a solution of $y'' + 3y' + y = \cos(x)$.

The companion polynomial is $P = X^2 + 3X + 1$ with roots $\alpha_1 = (-3 + \sqrt{5})/2$ and $\alpha_2 = (-3 - \sqrt{5})/2$.

To find f_1, we note that $\beta = 0$ is not a root of P, so we look for $f_1(x) = ax^2 + bx + c$. Then

$$f_1'' + 3f_1' + f_1 = ax^2 + (b + 6a)x + (c + 3b + 2a),$$

and the linear system to solve is

$$\begin{cases} a & = 1 \\ b + 6a & = 0 \text{ hence } b = -6 \\ c + 3b + 2a & = 0 \text{ hence } c = 18 - 2 = 16. \end{cases}$$

This means that $f_1(x) = x^2 - 6x + 16$.

To find f_2, since $\beta = 1$ is not a root of P, we consider $f_2(x) = a\cos(x) + b\sin(x)$. Then

$$f_2'' + 3f_2' + f_2 = (a + 3b - a)\cos(x) + (b + 3a - b)\sin(x) = 3b\cos(x) + 3a\sin(x),$$

and that means that we can take $b = 1/3$, $a = 0$, and $f_2(x) = \frac{1}{3}\sin(x)$.

We conclude that a special solution of the inhomogeneous equation is

$$f(x) = 3f_1(x) + f_2(x) = 3x^2 - 18x + 48 + \frac{1}{3}\sin(x).$$

Finally we discuss the method of variation of constants for linear differential equations of order $\geqslant 2$; it does not, in fact, necessarily require that the coefficients are constant, although the computations are often very difficult in general situations.

We consider the inhomogeneous equation

$$y^{(k)} + a_{k-1}y^{(k-1)} + \cdots + a_1 y' + a_0 y = b, \tag{2.5}$$

and we assume that a basis $(f_1, \ldots, f_k)$ of the space $\mathcal{S}$ of solutions of the homogeneous equation

$$y^{(k)} + a_{k-1}y^{(k-1)} + \cdots + a_1 y' + a_0 y = 0$$

has been found (it may be any basis). We then search for a solution to (2.5) of the form

$$f(x) = z_1(x)f_1(x) + \cdots + z_k(x)f_k(x),$$

where $z_1, \ldots, z_k$ are functions such that, moreover, we have

$$\begin{cases} z_1'(x)f_1(x) + \cdots + z_k'(x)f_k(x) = 0 \\ z_1'(x)f_1'(x) + \cdots + z_k'(x)f_k'(x) = 0 \\ \cdots \\ z_1'(x)f_1^{(k-2)}(x) + \cdots + z_k'(x)f_k^{(k-2)}(x) = 0 \end{cases}$$

for all x. The justification for requiring these $k-1$ extra constraints is that we need to find k different functions, and we may hope to succeed if they satisfy k different equations; one of these will be the original one (2.5), in combination with the $k-1$ extra conditions. Indeed, one can prove that this method works.

The most important example is $k=2$. Write again $f = z_1 f_1 + z_2 f_2$, and the constraint

$$z_1' f_1 + z_2' f_2 = 0.$$

The reason this condition is useful is that we get by differentiation the formulas

$$\begin{aligned} f' &= z_1' f_1 + z_2' f_2 + z_1 f_1' + z_2 f_2' = z_1 f_1' + z_2 f_2' \\ f'' &= z_1' f_1' + z_2' f_2' + z_1 f_1'' + z_2 f_2'', \end{aligned}$$

and therefore

$$y'' + a_1 y' + a_0 y = z_1(f_1'' + a_1 f_1' + a_0 f_1) + z_2(f_2'' + a_1 f_2' + a_0 f_2) + z_1' f_1' + z_2' f_2'.$$

But f_1 and f_2 solve the homogeneous equation, and hence

$$y'' + a_1 y' + a_0 y = z_1' f_1' + z_2' f_2'.$$

We conclude that z_1, z_2 lead to a solution of the inhomogeneous equation provided they satisfy the equations

$$\begin{cases} z_1' f_1 + z_2' f_2 = 0 \\ z_1' f_1' + z_2' f_2' = b. \end{cases}$$

For any given value of x, this is a linear system of equations with unknowns $(z_1'(x), z_2'(x))$. Once it is solved, we can obtain (in principle) the required functions z_1 and z_2 by computing primitives of (z_1', z_2'). It is a fact that the determinant $f_1 f_2' - f_1' f_2$ of the system will not vanish when solving this linear system of equations, corresponding to the fact that (f_1, f_2) is a basis of the space $\mathcal{S}$ of solutions of the homogeneous equation.

Example 2.4.9. We wish to solve the inhomogeneous equation

$$y'' + y' - 6y = \frac{1}{1+x^2}.$$

The roots of the companion polynomial $X^2 + X - 6$ are $\alpha_1 = 2$ and $\alpha_2 = -3$, so we search for a solution of the type

$$f(x) = z_1(x)e^{2x} + z_2(x)e^{-3x}$$

satisfying

$$z_1'(x)e^{2x} + z_2'(x)e^{-3x} = 0$$

for all x. Substituting into the equation, we obtain the system

$$\begin{cases} z_1'(x)e^{2x} + z_2'(x)e^{-3x} = 0 \\ 2z_1'(x)e^{2x} - 3z_2'(x)e^{-3x} = \frac{1}{1+x^2}. \end{cases}$$

The determinant is $-5e^{-x}$ so is indeed never zero, and we find the solutions for z_1' and z_2' given by

$$\begin{cases} z_1'(x) = \frac{e^{-2x}}{5(1+x^2)} \\ z_2'(x) = -\frac{e^{3x}}{5(1+x^2)}. \end{cases}$$

This means that a solution is

$$f(x) = \frac{1}{5}e^{2x}\int_0^x \frac{e^{-2t}}{1+t^2}dt - \frac{1}{5}e^{-3x}\int_0^x \frac{e^{3t}}{1+t^2}dt.$$

2.5. An example: the harmonic oscillator

One of the most basic example of linear differential equation with constant coefficients is given by the harmonic oscillator.

Case 1 (harmonic oscillator without friction). Here we have a particle with mass $m > 0$ attached at the end of a vertical spring, moving without the effect of gravity or of any friction. We measure its position along the axis of movement by a single function $y(t)$, where t is time, and where the origin of the y-axis refers to the equilibrium position. Then the only force acting on the particle is the restoring force from the spring, which is of the form $F = -ky$ for some coefficient $k > 0$ that depends on the "strength" of the spring.

The Newtonian equations of motion takes the form of the differential equation

$$m\ddot{y} = -ky,$$

or in other words x is solution of the homogeneous linear differential equation of order 2 given by

$$\ddot{y} + \frac{k}{m}y = 0.$$

Since $k/m > 0$, the real-valued solutions are of the forme

$$y(t) = a\cos(\omega t) + b\sin(\omega t)$$

where $\omega = \sqrt{k/m}$, for some real numbers a and b. It is customary to rephrase this in the form

$$y(t) = A\cos(\omega t + \varphi)$$

where $A = (a^2 + b^2)^{1/2}$ and φ is some real number. The advantage of this formula is that it clearly shows not only that the movement of the particle is periodic, with period $2\pi/\omega$, but also that its maximal amplitude (around the equilibrium position corresponding to $y = 0$) is A.

To see why this formula holds, note that

$$\Big(\frac{a}{A}\Big)^2 + \Big(\frac{b}{A}\Big)^2 = 1,$$

so that there exists a real number φ such that $\cos(\varphi) = a/A$ and $\sin(\varphi) = -b/A$; we get

$$a\cos(\omega t) + b\sin(\omega t) = A(\cos(\varphi)\cos(\omega t) - \sin(\varphi)\sin(\omega t)) = A\cos(\omega t + \varphi).$$

Case 2 (damped harmonic oscillator). Suppose now that the particle also encounters resistance, and that this other force is proportional to velocity (this is an assumption true in many cases, at least approximately). Then the Newton equation for $y(t)$ becomes

$$m\ddot{y} = -b\dot{y} - ky,$$

where $b > 0$ is another parameter measuring the strength of the friction force. We write this as

$$\ddot{y} + \frac{b}{m}\dot{y} + \frac{k}{m}y = 0,$$

which has companion polynomial $X^2 + \frac{b}{m}X + \frac{k}{m}$. There are correspondingly three cases, depending on the sign of

$$\Delta = \frac{b^2 - 4km}{m^2}.$$

If $\Delta > 0$, a basis of the space $\mathcal{S}$ of solutions is

$$y_1(t) = \exp\Bigl(\Bigl(-\frac{b}{2m} + \frac{1}{2}\sqrt{\Delta}\Bigr)t\Bigr), \qquad y_2(t) = \exp\Bigl(\Bigl(-\frac{b}{2m} - \frac{1}{2}\sqrt{\Delta}\Bigr)t\Bigr).$$

Observe that the sum of the two solutions of the quadratic equation is $-b/m < 0$ and the product is $k/m > 0$, so that both solutions of the quadratic equations are negative. This means that, as $t \to +\infty$, we have $y_1(t) \to 0$ and $y_2(t) \to 0$. Since the condition $\Delta > 0$ corresponds to b "large", the physical behavior is that the friction force is strong enough to essentially bring the motion to a stop, without oscillations.

If $\Delta = 0$, there is a double root, and a basis of the space $\mathcal{S}$ of solutions is

$$y_1(t) = \exp\Bigl(-\frac{bt}{2m}\Bigr), \qquad y_2(t) = t\exp\Bigl(-\frac{bt}{2m}\Bigr).$$

We have then also an exponentially fast "return to equilibrium" without oscillations.

If $\Delta < 0$, we get oscillatory functions as basis for the real solutions of the equation, namely

$$y_1(t) = \exp\Bigl(-\frac{bt}{2m}\Bigr)\cos(\tfrac{1}{2}\sqrt{|\Delta|}\ t) \qquad y_2(t) = \exp\Bigl(-\frac{bt}{2m}\Bigr)\sin(\tfrac{1}{2}\sqrt{|\Delta|}\ t).$$

The solution can now, as above, be expressed in the form

$$y(t) = Ae^{-bt/(2m)}\cos(\tfrac{1}{2}\sqrt{|\Delta|}\ t + \varphi),$$

(with $A > 0$ and $\varphi \in \mathbf{R}$). Since $b > 0$, the physical behavior is again return to equilibrium due to friction, but in an oscillatory manner around the equilibrium position. Note that the period $2\pi/(\frac{1}{2}\sqrt{|\Delta|})$ is larger than the period $2\pi/(k/m)$ of the oscillator with the same parameters but without friction.

2.6. Other methods

Besides the techniques described in the previous sections, it is useful to know two other commong methods that can be helpful to solve certain differential equations that are not of the type previously considered.

Change of variable. If a function $f(x)$ is replaced by $h(y) = f(g(y))$, where g is a "new variable", then any equation satisfied by f corresponds to an equation satisfied by h, and this equation may be simpler to solve, leading to a solution of the original one.

Example 2.6.1. If we make the change of variable $h(t) = f(e^t)$, then we have relations

$$h'(t) = e^t f'(e^t), \qquad h''(t) = e^t f'(e^t) + e^{2t} f''(e^t).$$

If, for instance, we try to solve $x^2y'' + xy' = y$, for $x > 0$, then we see that his is equivalent to

$$h''(t) = h(t)$$

for $h(t) = f(e^t)$. So the solutions are given by

$$h(t) = ae^t + be^{-t}$$

which means that

$$f(x) = ax + \frac{b}{x}.$$

Separation of variable. Suppose that a differential equation of order 1 can be written in the form $(g(y))' = b$ for some functions g and b (in other words, $g'(y)y' = b$). Then this can be solved by writing $g(f(x)) = B(x)$, where B is a primitive of b, and then "inverting" g.

Example 2.6.2. Consider the equation $e^{2y}y' = x$ with $x > 0$. To say that f is a solution means that the derivative of $\frac{1}{2}e^{2f(x)}$ is x, hence

$$e^{2f(x)} = x^2 + a$$

for some constant a, or in other words

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

CHAPTER 3

Differential calculus in $\mathbf{R}^n$

In this chapter, n and m are always integers $\geqslant 1$.

3.1. Introduction

We are interested in functions defined on subsets of $\mathbf{R}^n$ which take values in $\mathbf{R}$, or $\mathbf{C}$, or even in another space $\mathbf{R}^m$, where $m \geqslant 1$ is an integer.

Here are some basic examples of such functions that should be kept in mind.

(1) *Linear maps* $f\colon \mathbf{R}^n \to \mathbf{R}^m$, or in other words, functions defined by $f(x) = Ax$, where A is a matrix with n columns and m rows, and x is interpreted as a column vector. For instance, for $n = 2$ and $m = 1$, one can consider $f(x, y) = x + y$ for $(x, y) \in \mathbf{R}^2$. Slightly more generally, if in addition we fix $y_0 \in \mathbf{R}^m$, we can define the *affine-linear map* $f(x) = y_0 + Ax$.

(2) *Quadratic forms* $Q\colon \mathbf{R}^n \to \mathbf{R}$, or in other words, functions of the type

$$Q(x) = \sum_{i=1}^{n}\sum_{j=1}^{n} a_{i,j}x_ix_j$$

for all $x = (x_1, \ldots, x_n)$, where $(a_{i,j})$ are real numbers. For instance, for $n = 2$, one can consider $Q(x, y) = xy$; for arbitrary n, one has the quadratic form

$$Q(x_1, \ldots, x_n) = x_1^2 + \cdots + x_n^2.$$

(3) Polynomials in n variables: these generalize the previous two examples. Given an integer $d \geqslant 0$, a polynomial in n variables of degree $\leqslant d$ is a finite sum of *monomials* of degree $e \leqslant d$, namely a finite sum of functions $\mathbf{R}^n \to \mathbf{R}$ of the type

$$f(x_1, \ldots, x_n) = \alpha x_1^{d_1} \cdots x_n^{d_n} \tag{3.1}$$

where the degree of the monomial, that is the integer

$$e = d_1 + \cdots + d_n,$$

satisfies $e \leqslant d$. For instance, the function

$$f(x, y, z) = x^3 - 12xy^5z + xyz$$

is a polynomial of degree 7. Example (1) (affine-linear maps) corresponds to polynomials of degree $\leqslant 1$, and Example (2) to certain polynomials of degree 2.

(4) "Cartesian product" functions: two functions $f_1\colon \mathbf{R}^n \to \mathbf{R}^{m_1}$ and $f_2\colon \mathbf{R}^n \to \mathbf{R}^{m_2}$ combine to produce a function $f = (f_1, f_2)\colon \mathbf{R}^n \to \mathbf{R}^{m_1+m_2}$, defined by

$$f(x) = (f_1(x), f_2(x)).$$

An important point is that any function $f\colon \mathbf{R}^n \to \mathbf{R}^m$ is a cartesian product $f = (f_1, \ldots, f_m)$ of functions $f_i\colon \mathbf{R}^n \to \mathbf{R}$, where $f_i(x)$ is just the i-th coordinate of $f(x)$ as a vector in $\mathbf{R}^m$. This means that many definitions and results for

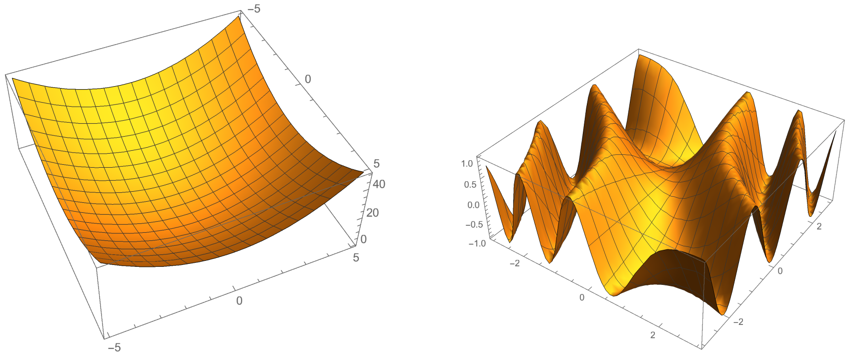

Figure 3.1. Graphs of $f(x,y) = x^2 + y^2$ and $f(x,y) = \sin(\frac{1}{2}x^2 + xy)$

functions $\mathbf{R}^n \to \mathbf{R}^m$ may be reduced easily to the case $m = 1$ by considering each coordinate separately.

(5) Functions with separated variables: if $f_1, \ldots, f_n$ are functions on $\mathbf{R}$ (or on a subset of $\mathbf{R}$, the same for each of them), we can define a function $f\colon \mathbf{R}^n \to \mathbf{R}$ by

$$f(x_1, \ldots, x_n) = f_1(x_1) \cdots f_n(x_n),$$

where the variables are "separated". For instance, any monomial (3.1) is a function with separated variables.

(6) Composition of functions: given any function $f\colon \mathbf{R}^n \to \mathbf{R}$, and a function $g\colon \mathbf{R} \to \mathbf{R}$, we can consider the composition $g \circ f$. For instance, composing the quadratic form

$$Q(x_1, \ldots, x_n) = x_1^2 + \cdots + x_n^2$$

with the square root, one obtains

$$\sqrt{x_1^2 + \cdots + x_n^2},$$

which is the euclidean norm (length from the origin to the point $x \in \mathbf{R}^n$). Composing with $\exp(-y)$, one gets

$$\exp(-(x_1^2 + \cdots x_n^2)).$$

Note that this last function is a function with separated variables (but the euclidean norm is not).

For functions $f\colon \mathbf{R}^2 \to \mathbf{R}$, one can visualize the graph of f, which is

$$\{(x,y,z) \in \mathbf{R}^3 \ : \ z = f(x,y)\}$$

as a surface in $\mathbf{R}^3$ (see Figure 3.1 for two examples; using an interactive software is better to understand such pictures, as one can manipulate the graph easily). This visualization is not possible anymore when there are 3 variables or more. This is one reason why multivariable calculus is often more difficult to understand intuitively than the one-variable case.

Remark 3.1.1. Another interesting visualization possibility concerns the case of $f\colon \mathbf{R}^2 \to \mathbf{R}^2$, where one can show $f(x)$ as a vector based at x, at least for a subset of values of x. Figure 3.2 illustrates this for the function

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

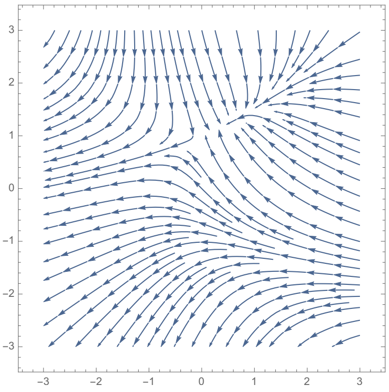

FIGURE 3.2. Vector plot

3.2. Continuity in $\mathbf{R}^n$

The first notion that we want to generalize is that of a continuous function. To follow the example of functions of one variable, we need first to recall the definition of convergence and limit of a sequence (or of a function) in $\mathbf{R}^n$.

We define

$$\|x\| = \sqrt{x_1^2 + \cdots + x_n^2}$$

for $x \in \mathbf{R}^n$ (the norm of x in the euclidean space $\mathbf{R}^n$; see Section 1.2 in [1])). This function satisfies the following properties:

$$\|x\| \geqslant 0, \text{ and } \|x\| = 0 \text{ if and only if } x = 0$$
$$\|tx\| = |t|\|x\| \text{ for all } t \in \mathbf{R}$$
$$\|x + y\| \leqslant \|x\| + \|y\| \qquad \text{(triangle inequality)}.$$

The definition of convergence on $\mathbf{R}^n$ is given in [1, Def. 2.6.1].

DEFINITION 3.2.1. Let $(x_k)_{k \in \mathbf{N}}$ where $x_k \in \mathbf{R}^n$. Write

$$x_k = (x_{k,1}, \ldots, x_{k,n}).$$

Let $y = (y_1, \ldots, y_n) \in \mathbf{R}^n$. We say that the sequence (x_k) converges to y as $k \to +\infty$ if for all $\varepsilon > 0$, there exists $N \geqslant 1$ such that for all $n \geqslant N$, we have

$$\|x_k - y\| < \varepsilon.$$

LEMMA 3.2.2. *The sequence (x_k) converges to y as $k \to +\infty$ if and only if one of the following equivalent conditions holds:*

(1) For each i, $1 \leqslant i \leqslant n$, the sequence $(x_{k,i})$ of real numbers converges to y_i.

(2) The sequence of real numbers $\|x_k - y\|$ converges to 0 as $k \to +\infty$.

PROOF. The equivalence of the two conditions is elementary: first, since

$$|x_{k,i} - y_i|^2 \leqslant \sum_{j=1}^{n} |x_{k,j} - y_j|^2 = \|x_k - y\|^2,$$

the second condition implies that $x_{k,i} \to y_i$ for each i; conversely, if the first condition holds, then

$$\|x_k - y\|^2 = \sum_{j=1}^{n} |x_{k,j} - y_j|^2$$

is the sum of n sequences, each converging to 0, hence converges to 0. The fact that these are equivalent to the convergence of (x_k) to y is proved in [1, Satz 2.6.3]. □

We next extend the definition of continuity given in [1, Def. 3.2.1, 3.2.2].

Definition 3.2.3. Let $X \subset \mathbf{R}^n$ and $f\colon X \to \mathbf{R}^m$.

(1) Let $x_0 \in X$. We say that f is *continuous* at x_0 if for all $\varepsilon > 0$, there exists $\delta > 0$ such that, if $x \in X$ satisfies $\|x - x_0\| < \delta$, then

$$\|f(x) - f(x_0)\| < \varepsilon.$$

(2) We say that f is *continuous on* X if it is continuous at x_0 for all $x_0 \in X$.

Similarly to [1, Satz 3.2.4], we can test if a function is continuous using sequences.

Proposition 3.2.4. *Let $X \subset \mathbf{R}^n$ and $f\colon X \to \mathbf{R}^m$. Let $x_0 \in X$. The function f is continuous at x_0 if and only if, for every sequence $(x_k)_{k\geqslant 1}$ in X such that $x_k \to x_0$ as $k \to +\infty$, the sequence $(f(x_k))_{k\geqslant 1}$ in $\mathbf{R}^m$ converges to $f(x)$.*

From this proposition, we can immediately see that most functions that we encounter are continuous.

In a similar way, we can define the limit of a function at a point (see [1, §3.10]).

Definition 3.2.5. Let $X \subset \mathbf{R}^n$ and $f\colon X \to \mathbf{R}^m$. Let $x_0 \in X$ and $y \in \mathbf{R}^m$. We say that f *has the limit y as $x \to x_0$ with $x \neq x_0$* if for every $\varepsilon > 0$, there exists $\delta > 0$, such that for all $x \in X$, $x \neq x_0$, such that $\|x - x_0\| < \delta$, we have $\|f(x) - y\| < \varepsilon$. We then write

$$\lim_{\substack{x\to x_0\\ x\neq x_0}} f(x) = y.$$

Remark 3.2.6. In this definition, we could also remove the assumption that $x_0 \in X$, because if $x_0 \notin X$, we could always extend f to $X \cup \{x_0\}$ by, for instance, defining $f(x_0) = 0$.

The "sequence" test for this condition is:

Proposition 3.2.7. *Let $X \subset \mathbf{R}^n$ and $f\colon X \to \mathbf{R}^m$. Let $x_0 \in X$ and $y \in \mathbf{R}^m$. We have*

$$\lim_{\substack{x\to x_0\\ x\neq x_0}} f(x) = y.$$

if and only if, for every sequence (x_k) in X such that $x_k \to x$ as $k \to +\infty$, and $x_k \neq x_0$, the sequence $(f(x_k))$ in $\mathbf{R}^m$ converges to y.

Example 3.2.8. Let $X \subset \mathbf{R}^n$ and $f\colon X \to \mathbf{R}^m$. Let $x_0 \in X$. Then f is continuous at x_0 if and only if

$$\lim_{\substack{x\to x_0\\ x\neq x_0}} f(x) = f(x_0).$$

The easiest way to prove continuity is in general to use composition:

Proposition 3.2.9. *Let $X \subset \mathbf{R}^n$, $Y \subset \mathbf{R}^m$ and $p \geqslant 1$ an integer. Let $f\colon X \to Y$ and $g\colon Y \to \mathbf{R}^p$ be continuous functions. Then the composite $g \circ f$ is continuous.*

Proof. We apply Proposition 3.2.4. If (x_k) is a sequence in X converging to $x \in X$ in $\mathbf{R}^n$, then by continuity of f, the sequence $(f(x_k))$ is a sequence in Y converging to $y = f(x)$. Then by continuity of g, the sequence $(g(f(x_k))$ converges to $g(f(x))$. By definition of $g \circ f$, this implies that $g \circ f$ is continuous. □

EXAMPLE 3.2.10. (1) Cartesian products of continuous functions are continuous: if $f_1\colon \mathbf{R}^n \to \mathbf{R}^{m_1}$ and $f_2\colon \mathbf{R}^n \to \mathbf{R}^{m_2}$ are continuous, then $f = (f_1, f_2)\colon \mathbf{R}^n \to \mathbf{R}^{m_1+m_2}$ is also continuous. In particular, a function $f\colon \mathbf{R}^n \to \mathbf{R}^m$ is continuous if and only if its coordinates $f_1, \ldots, f_m$ are continuous. This follows from the definition and is left as an exercise.

(2) Any linear map $f\colon \mathbf{R}^n \to \mathbf{R}^m$ is continuous. In particular, the identity map is continuous. To see this, note that according to (1) it is enough to assume that $m = 1$. Then there exist numbers $a_1, \ldots, a_n$ such that

$$f(x_1, \ldots, x_n) = a_1x_1 + \cdots a_nx_n.$$

Let $y = (y_1, \ldots, y_n) \in \mathbf{R}^n$ and let (x_k) be a sequence converging to y. Then, writing $x_k = (x_{k,1}, \ldots, x_{k,n})$, we have $x_{k,i} \to y_i$ for all i, and therefore it follows that $a_ix_{k,i} \to a_iy_i$, and then that

$$f(x_k) = a_1x_{k,1} + \cdots + a_nx_{k,n} \to a_1y_1 + \cdots + a_ny_n = f(y)$$

(see [1, Satz 2.1.8]: for convergent sequences, $\lim(a_k + b_k) = \lim a_k + \lim b_k$).

A similar argument shows that if $f_1\colon X \to \mathbf{R}^m$ and $f_2\colon X \to \mathbf{R}^m$ are continuous on X, then $f_1 + f_2$ is also continous. Alternatively, one can write $f_1 + f_2 = a \circ (f_1, f_2)$, where $a\colon \mathbf{R}^m \times \mathbf{R}^m \to \mathbf{R}^m$ is the addition map. Since a is linear and (f_1, f_2) is continuous, the sum $f_1 + f_2$ is continuous by composition (Proposition 3.2.9).

(3) Functions with separated variables are continuous if the factors are continuous: if $f_1, \ldots, f_n$ are continuous functions on $\mathbf{R}$ (or on a subset of $\mathbf{R}$, the same for each of them), then f defined by

$$f(x_1, \ldots, x_n) = f_1(x_1) \cdots f_n(x_n)$$

is continuous on $\mathbf{R}^n$. This follows easily from the rule

$$\lim_{k\to+\infty} a_kb_k = ab$$

for convergent sequences ([1, Satz 2.1.8]).

(4) Combining addition and functions with separated variables, one deduces that polynomials in $x_1, \ldots, x_n$ are continuous.

(5) Similarly, using the rules

$$\lim_{k\to+\infty} a_kb_k = ab, \qquad \lim_{k\to+\infty} a_k/b_k = a/b$$

when real numbers $a_k \to a$ and $b_k \to b$, with $b \neq 0$ in the second case (again [1, Satz 2.1.8]), one checks that if f_1 and f_2 are continuous functions from $X \subset \mathbf{R}^n$ to $\mathbf{R}$, then f_1f_2 is continuous, and if moreover $f_2(x) \neq 0$ for all $x \in X$, then f_1/f_2 is continuous.

(6) Analogues of the previous results exist for limits of functions, for instance

$$\lim_{x\to x_0} (f(x) + g(x)) = \lim_{x\to x_0} f(x) + \lim_{x\to x_0} g(x), \qquad \lim_{x\to x_0} f(x)g(x) = \lim_{x\to x_0} f(x) \lim_{x\to x_0} g(x)$$

if both f and g have limits as $x \to x_0$.

(7) Suppose that $f\colon \mathbf{R}^2 \to \mathbf{R}$ is continuous. Then, if we fix a value $y_0 \in \mathbf{R}$, then the function g defined on $\mathbf{R}$ by $g(x) = f(x, y_0)$ is continuous (for instance, it is the composition of f and of the function $x \mapsto (x, y_0)$, which is continuous). However the converse is *not* true. For instance, define

$$f(x, y) = \begin{cases} x & \text{if } y \geqslant 0 \\ -x & \text{if } y < 0. \end{cases}$$

Then each function $g(x) = f(x, y_0)$ is continuous, but f itself is not continuous. For instance, we have $f(1, 0) = 1$. However, the sequence $(x_k, y_k) = (1, -1/k)$ converges to $(1, 0)$ but we have $f(1, -1/k) = -1$ for any $k \geqslant 1$, which does not converge to 1.

One of the first difficulties in extending the definitions of Analysis I to functions of $n \geqslant 2$ variables is that the sets on which they are defined can be much more complicated than those used with one variable, which are often just intervals. For $n = 2$, one can draw many different possible "two-dimensional" shapes, each of which is a possible definition set for a function.

We need analogues of *closed* and *compact* intervals (Definitions 2.5.1 and 3.4.2 in [1]).

Definition 3.2.11. (1) A subset $X \subset \mathbf{R}^n$ is *bounded* if the set of $\|x\|$ for $x \in X$ is bounded in $\mathbf{R}$.

(2) A subset $X \subset \mathbf{R}^n$ is *closed* if for every sequence (x_k) in X that converges in $\mathbf{R}^n$ to some vector $y \in \mathbf{R}^n$, we have $y \in X$.

(3) A subset $X \subset \mathbf{R}^n$ is *compact* if it is bounded and closed.

Example 3.2.12. (1) The empty set and $\mathbf{R}^n$ are both closed.

(2) Let $r > 0$ and $x_0 \in \mathbf{R}^n$. The open disc $D = \{x \in \mathbf{R}^n : \|x - x_0\| < r\}$ is bounded (since, by the triangle inequality, we have

$$\|x\| \leqslant \|x - x_0\| + \|x_0\| \leqslant r + \|x_0\|$$

for all $x \in D$). It is not closed, since for instance the sequence

$$x_k = x_0 + (r - 1/k, 0, \ldots, 0) \to x_0 + r$$

where $x_0 + r \notin D$.

(3) The *closed* disc $\Delta = \{x \in \mathbf{R}^n : \|x - x_0\| \leqslant r\}$ is closed and bounded. Indeed, let $x_k \in \Delta$ be a sequence that converges to $y \in \mathbf{R}^n$. We have

$$\sqrt{(x_{k,1} - x_{0,1})^2 + \cdots + (x_{k,n} - x_{0,n})^2} \leqslant r$$

for all k and $x_{k,i} \to y_i$. Taking $k \to +\infty$, and using the property

$$(a_k \leqslant a \text{ for all } k) \Rightarrow \lim_k a_k \leqslant a$$

for converging sequences of real numbers, we deduce that

$$\sqrt{(y_1 - x_{0,1})^2 + \cdots + (y_n - x_{0,n})^2} \leqslant r$$

so that $y \in \Delta$.

In particular, for $n = 1$, this means that a closed interval is a closed set. An interval is compact if, furthermore, it is bounded.

(4) If $X_1 \subset \mathbf{R}^n$ and $X_2 \subset \mathbf{R}^m$ are bounded (resp. closed, resp. compact), then so is $X_1 \times X_2 \subset \mathbf{R}^{n+m}$. In particular, the set

$$B = I_1 \times \cdots \times I_n = \{(x_1, \ldots, x_n) \in \mathbf{R}^n : x_i \in I_i\}$$

is closed (resp. compact) if each interval I_i is closed (resp. compact).

Using basic examples of closed sets as above, one can construct many more using the following fundamental property:

Proposition 3.2.13. *Let $f\colon \mathbf{R}^n \to \mathbf{R}^m$ be a continuous map. For any closed set $Y \subset \mathbf{R}^m$, the set*

$$f^{-1}(Y) = \{x \in \mathbf{R}^n : f(x) \in Y\} \subset \mathbf{R}^n$$

is closed.

PROOF. Indeed, let $X = f^{-1}(Y)$. If (x_k) is a sequence in X that converges to $y \in \mathbf{R}^n$, then by continuity we get $f(x_k) \to f(y)$. But then $f(y) \in Y$ because it is the limit of $f(x_k) \in Y$ and Y is closed. This means that $y \in f^{-1}(Y)$. □

EXAMPLE 3.2.14. Let $f \colon \mathbf{R}^n \to \mathbf{R}$ be a continuous function. The *zero set* $Z = \{x \in \mathbf{R}^n \,:\, f(x) = 0\}$ is closed in $\mathbf{R}^n$ because $\{0\} \subset \mathbf{R}$ is closed.

For instance for $r \geqslant 0$, a circle or a sphere of radius r, defined by

$$\{x \in \mathbf{R}^2 \,:\, \|x - x_0\| = r\}, \qquad \{x \in \mathbf{R}^3 \,:\, \|x - x_0\| = r\},$$

is closed.

Similarly, for any $r \geqslant 0$, the set

$$\{x \in \mathbf{R}^n \,:\, |f(x)| \leqslant r\}$$

is $f^{-1}([-r, r])$, hence is closed since the interval $[-r, r]$ is closed.

In practice, the closed sets that we will use will very often be of one of these forms.

The following Theorem generalizes Theorem 3.4.5 of [1] to more than one variable.

THEOREM 3.2.15. *Let $X \subset \mathbf{R}^n$ be a non-empty compact set and $f \colon X \to \mathbf{R}$ a continuous function. Then f is bounded and achieves its maximum and minimum, or in other words, there exist x_+ and x_- in X such that*

$$f(x_+) = \sup_{x \in X} f(x), \qquad f(x_-) = \inf_{x \in X} f(x).$$

Another difficulty in working with functions of $n \geqslant 2$ variables is that for $n \geqslant 2$, the notion of continuity (or of limit) is much stronger than in dimension 1. One intuitive reason is that there are "many more ways" for a sequence to converge to $x \in \mathbf{R}^n$ than in $\mathbf{R}$.

For instance, all of the following sequences converge to $(0, 0)$ in $\mathbf{R}^2$, but the way they do it is quite different:

1. (Limit along a ray) Take $(\cos(\theta)/k, \sin(\theta)/k)$, where $\theta \in \mathbf{R}$ is fixed. All these points are on the line with angle θ from the x-axis.
2. (Spiraling limit) Take $(\cos(k)/k, \sin(k)/k)$; here the angle from the x-axis is k, and there is no special direction of convergence.

A priori, the limit of $f(x_k, y_k)$ could exist but be different for each of these sequences, or there could be limits in some directions but not others, the "spiraling" limit may or may not exist even if the "ray" limits exist, etc.

EXAMPLE 3.2.16. Define $f(0, 0) = 0$ and

$$f(x, y) = \frac{xy}{x^2 + y^2}.$$

Note that f is continuous when defined on $\mathbf{R}^2 \backslash \{(0, 0)\}$, since the denominator is continuous and is never zero there.

Then

$$f(\cos(\theta)/k, \sin(\theta)/k) = \frac{\sin(\theta)\cos(\theta)/k^2}{\cos^2(\theta)/k^2 + \sin^2(\theta)/k^2} = \cos(\theta)\sin(\theta)$$

so the limit exists for every θ, but its value depends on θ. This implies in particular that the function f is not continuous at $(0, 0)$, hence is not continuous on $\mathbf{R}^n$.

On the other hand, for the spiral, we get

$$f(\cos(k)/k, \sin(k)/k) = \cos(k)\sin(k)$$

which has no limit as $k \to +\infty$.

3.3. Partial derivatives

We now consider the generalization of derivability in $\mathbf{R}^n$. In one variable, we restricted to open intervals to define the derivative. The analogue in $\mathbf{R}^n$ is the following:

Definition 3.3.1. A subset $X \subset \mathbf{R}^n$ is *open* if, for any $x = (x_1, \ldots, x_n) \in X$, there exists $\delta > 0$ such that the set

$$\{y = (y_1, \ldots, y_n) \in \mathbf{R}^n \ : \ |x_i - y_i| < \delta \text{ for all } i\}$$

is contained in X.

In other words: any point of $\mathbf{R}^n$ obtained by changing any coordinate of x by at most δ is still in X.

The basic example to keep in mind is just $X = \mathbf{R}^n$ (and one may assume at first that this is the case for the definitions of partial derivatives and of the differential below).

The following proposition often leads to an easy way to show that a set is open:

Proposition 3.3.2. *A set $X \subset \mathbf{R}^n$ is open if and only if the complement*

$$Y = \{x \in \mathbf{R}^n \ : \ x \notin X\}$$

is closed.

Corollary 3.3.3. *If $f\colon \mathbf{R}^n \to \mathbf{R}^m$ is continuous and $Y \subset \mathbf{R}^m$ is open, then $f^{-1}(Y)$ is open in $\mathbf{R}^n$.*

Proof. This is because the complement of $f^{-1}(Y)$ is the set of points $x \in X$ such that $f(x)$ belongs to the complement of Y, which is closed according to the proposition, so this follows from Proposition 3.2.13. □

The following examples are the most important open sets for us.

Example 3.3.4. (1) The empty set and $\mathbf{R}^n$ are open. In fact, they are the only two sets in $\mathbf{R}^n$ that are both open and closed (this is intuitively reasonable, although a rigorous proof requires some care).

(2) The open ball of center x_0 and radius r

$$D = \{x \in \mathbf{R}^n \ : \ \|x - x_0\| < r\}$$

is open in $\mathbf{R}^n$. We can check this both using the definition and the corollary.

For the definition: let $x \in D$ and define $s = \|x - x_0\| < r$. Put $\delta_0 = \frac{1}{2}(r - s) > 0$. Then any $z \in \mathbf{R}^n$ such that $\|z\| < \delta_0$ satisfies

$$\|x + z - x_0\| \leqslant \|x - x_0\| + \|z\| \leqslant s + \delta_0 < r.$$

Define $\delta = \delta_0/\sqrt{n}$. If $|x_i - y_i| < \delta$ for all i, then putting $z = y - x$, we get

$$\|y - x\| = \sqrt{(y_1 - x_1)^2 + \cdots + (y_n - x_n)^2} < \delta\sqrt{n} = \delta_0$$

so $\|y - x_0\| = \|x + z - x_0\| < \delta$.

Using the corollary, let $f(x) = \|x - x_0\|$, which is a continuous function; then $D = f^{-1}(] - r, r[)$, so it is open.

On the other hand, the closed ball Δ is not open: for instance, if we take $x = x_0 + (r, 0, \ldots, 0)$, then for any $\delta > 0$, the point

$$x_0 + (r + \delta, 0, \ldots, 0)$$

is not in Δ.

(3) Let I_1, $\ldots$, I_n be open intervals in $\mathbf{R}$. Then $I_1 \times \cdots \times I_n$ is open in $\mathbf{R}^n$.

(4) Arguing as in Example (2), we see more generally that $X \subset \mathbf{R}^n$ is open if and only if, for any $x \in X$, there exists $\delta > 0$ such that the open ball of center x and radius δ is contained in X.

Now we can define partial derivatives.

DEFINITION 3.3.5. Let $X \subset \mathbf{R}^n$ be an open set. Let $f\colon X \to \mathbf{R}^m$ be a function. Let $1 \leqslant i \leqslant n$. We say that f has a partial derivative on X with respect to the i-th variable, or coordinate, if for all $x_0 = (x_{0,1}, \ldots, x_{0,n}) \in X$, the function defined by

$$g(t) = f(x_{0,1}, \ldots, x_{0,i-1}, t, x_{0,i+1}, \ldots, x_{0,n})$$

on the set

$$I = \{t \in \mathbf{R} \,:\, (x_{0,1}, \ldots, x_{0,i-1}, t, x_{0,i+1}, \ldots, x_{0,n}) \in X\}$$

is differentiable at $t = x_{0,i}$. Its derivative $g'(x_{0,i})$ at $x_{0,i}$ is denoted

$$\frac{\partial f}{\partial x_i}(x_0), \qquad \partial_{x_i} f(x_0), \qquad \partial_i f(x_0).$$

Intuitively, this definition means that we "freeze" all variables except the i-th one, and consider the derivative of the corresponding function of one variable. We recall once more that if $m \geqslant 2$, so that $g(t) = (g_1(t), \ldots, g_m(t))$ for some real-valued functions $g_j\colon I \to \mathbf{R}$, then g is differentiable if and only if all g_j are differentiable, and that

$$g'(t) = (g_1'(t), \ldots, g_m'(t)).$$

REMARK 3.3.6. (1) Note that by definition of an open set, the set I always contains an open interval containing $x_{0,i}$, so that it makes sense to ask that g be differentiable at $x_{0,i}$.

(2) The notation $\partial_{x_i} f$ can sometimes be confusing. It is important to remember that here x_i refers to a *variable*, and not to a specific real value. This is especially a problem when one writes a value of the partial derivative at a point: in

$$\partial_{x_1} f(x_1, \ldots, x_n),$$

we think of x_1 in the partial derivative as a variable (indicating for which variable we compute the derivative), but we think of $(x_1, \ldots, x_n)$ as a point in $\mathbf{R}^n$ where we evaluate the partial derivative. Writing $\partial_i f(x)$ is sometimes clearer for this reason.

It follows immediately from the definition that partial derivatives have all the properties of the usual derivative of a function of one variable.

PROPOSITION 3.3.7. *Consider $X \subset \mathbf{R}^n$ open and f, g functions from X to $\mathbf{R}^m$. Let $1 \leqslant i \leqslant n$.*

(1) *If f and g have partial derivatives with respect to the i-th coordinate on X, then $f + g$ also does, and*

$$\partial_{x_i}(f + g) = \partial_{x_i}(f) + \partial_{x_i}(g).$$

(2) *If $m = 1$, and if f and g have partial derivatives with respect to the i-th coordinate on X, then fg also does and*

$$\partial_{x_i}(fg) = \partial_{x_i}(f)\ g + f\partial_{x_i}(g).$$

Furthermore, if $g(x) \neq 0$ for all $x \in X$, then f/g has a partial derivative with respect to the i-th coordinate on X, with

$$\partial_{x_i}(f/g) = (\partial_{x_i}(f)\ g - f\partial_{x_i}(g))/g^2.$$

Moreover, computing partial derivatives is as easy as computing ordinary derivatives.

EXAMPLE 3.3.8. (1) Let f be linear from $\mathbf{R}^n$ to $\mathbf{R}$. Then if we write

$$f(x_1, \ldots, x_n) = a_1 x_1 + \cdots + a_n x_n,$$

then we see that

$$\partial_i f(x) = a_i$$

for all $x \in \mathbf{R}^n$ and $1 \leqslant i \leqslant n$.

(2) Let f be a function with separated variables, say

$$f(x) = f_1(x_1) \cdots f_n(x_n).$$

If each f_i is differentiable on $\mathbf{R}$, then f has partial derivatives, which are

$$\partial_i f(x) = f_1(x_1) \cdots f_{i-1}(x_{i-1}) f_i'(x_i) f_{i+1}(x_{i+1}) \cdots f_n(x_n)$$

(so all partial derivatives also have separated variables).

(3) Let $f(x, y, z) = \cos(xy^2z^3) - 12x^2$. Then we have

$$\begin{aligned} \partial_x f &= -y^2z^3 \sin(xy^2z^3) - 24x \\ \partial_y f &= -2xyz^3 \sin(xy^2z^3) \\ \partial_z f &= -3xy^2z^2 \sin(xy^2z^3). \end{aligned}$$

(4) Let $f(x, y)$ be the function of Example 3.2.16. Since $f(0, y) = f(x, 0) = 0$, we obtain the partial derivatives $\partial_x f(0, 0) = \partial_y f(0, 0) = 0$.

DEFINITION 3.3.9. Let $X \subset \mathbf{R}^n$ open and $f : X \to \mathbf{R}^m$ a function with partial derivatives on X. Write

$$f(x) = (f_1(x), \ldots, f_m(x)).$$

For any $x \in X$, the matrix

$$J_f(x) = (\partial_{x_j} f_i(x))_{\substack{1 \leqslant i \leqslant m \\ 1 \leqslant j \leqslant n}}$$

with m rows and n columns is called the *Jacobi matrix* of f at x.

EXAMPLE 3.3.10. Let $f : \mathbf{R}^2 \to \mathbf{R}^3$ be defined by

$$f(x, y) = \begin{pmatrix} \cos(x^2 + y) \\ e^{\sin(\pi xy)} - 1 \\ y + \frac{1}{x^2+1} \end{pmatrix}$$

(the variables (x, y) should be thought of as a column vector). Then the function has partial derivatives, and for any $(x, y) \in \mathbf{R}^2$, the Jacobi matrix is

$$J_f(x, y) = \begin{pmatrix} -2x \sin(x^2 + y) & -\sin(x^2 + y) \\ \pi y \cos(\pi xy) e^{\sin(\pi xy)} & \pi x \cos(\pi xy) e^{\sin(\pi xy)} \\ \frac{-2x}{(1+x^2)^2} & 1 \end{pmatrix}$$

(the first column has the partial derivatives with respect to x, and the second with respect to y).

If we want to evaluate this at some point, say $(1, 0)$, we obtain

$$J_f(1, 0) = \begin{pmatrix} -2\sin(1) & -\sin(1) \\ 0 & -\pi \\ \frac{-1}{2} & 1 \end{pmatrix}$$

As is clear from examples, often the partial derivatives $\partial_{x_i} f$ of a function themselves admit partial derivatives $\partial_{x_j}(\partial_{x_i} f)$, and so on. Some of thee notation that are used for multiple partial derivatives are:

$$\partial_{x_i}(\partial_{x_i} f) = \partial_{x_i^2} f = \frac{\partial^2 f}{\partial x_i^2}, \qquad \partial_{x_i}(\partial_{x_j} f) = \partial_{x_i, x_j} f = \frac{\partial^2 f}{\partial x_i \partial x_j}.$$

Definition 3.3.11 (Gradient, Divergence). Let $X \subset \mathbf{R}^n$ be open.

(1) Let $f\colon X \to \mathbf{R}$ be a function. If all partial derivatives of f exist at $x_0 \in X$, then the column vector

$$\begin{pmatrix} \partial_{x_1} f(x_0) \\ \cdots \\ \partial_{x_n} f(x_0) \end{pmatrix}$$

is called the *gradient* of f at x_0, and is denoted $\nabla f(x_0)$.

(2) Let $f = (f_1, \ldots, f_n)\colon X \to \mathbf{R}^n$ be a function with values in $\mathbf{R}^n$ such that all partial derivatives of all coordinates f_i of f exist at $x_0 \in X$. Then the real number

$$\mathrm{Tr}(J_f(x_0)) = \sum_{i=1}^{n} \partial_{x_i} f_i(x_0),$$

the trace of the Jacobi matrix, is called the *divergence* of f at x_0, and is denoted $\mathrm{div}(f)(x_0)$.

3.4. The differential

Although partial derivatives are very easy to define and compute, their existence is not the correct analogue of differentiability. To be more precise, we want this analogue to provide a way to approximate a function by a linear map, just as the fact that a function $f\colon \mathbf{R} \to \mathbf{R}$ is differentiable with derivative a at 0 means that

$$f(x) = f(0) + ax + E(x)$$

where the "error" $E(x)$ has the property that $\lim_{x \to 0} E(x)/x = 0$, so that the affine-linear map $g(x) = f(0) + ax = f(0) + f'(0)x$ is a good approximation to $f(x)$ when x is close to 0.

If we consider a function $f\colon \mathbf{R}^n \to \mathbf{R}$ with $n \geqslant 2$, the problem is that $\partial_{x_1} f(0)$, for instance, only gives some information on how f behaves when the first variable tends to 0, the others being fixed. It is quite believable that, for certain functions, we will not be able to deduce an approximation for $f(x)$ when x is close to 0 from the approximations along the coordinate axes!

Example 3.4.1. (1) Let $f(x, y)$ be the function of Examples 3.2.16 and 3.3.8. We have seen that $\partial_x f(0,0) = \partial_y f(0,0) = 0$, but from Example 3.2.16, the function f is not continuous at $(0,0)$! So the partial derivatives can not be combined in any reasonable manner to give a good approximation of f for (x, y) close to $(0,0)$.

(2) Let $g\colon \mathbf{R}^2 \to \mathbf{R}$ be defined by $g(0,0) = 0$ and

$$g(x, y) = \frac{xy}{\sqrt{x^2 + y^2}}$$

(see Figure 3.3 for its graph). This function is now continuous at $(0,0)$ because, for $(x, y) \neq (0,0)$, we have

$$|g(x, y)| \leqslant \frac{\frac{1}{2}(x^2 + y^2)}{\sqrt{x^2 + y^2}} = \frac{1}{2}\sqrt{x^2 + y^2} \to 0$$

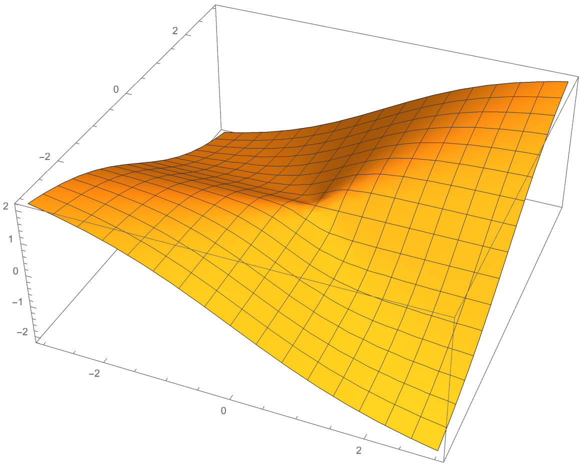

FIGURE 3.3. Graph of $g(x,y) = xy/\sqrt{x^2+y^2}$

as $(x,y) \to (0,0)$. Since $g(x,0) = g(0,y) = 0$, the partial derivatives exist and are both 0 again. But if we compute $g(r\cos(\theta), r\sin(\theta))$ as $r \to 0$, corresponding to approximating g along a line with angle θ with respect to the x axis, then we get for $r > 0$ the formula

$$g(r\cos(\theta), r\sin(\theta)) = \frac{r^2\cos(\theta)\sin(\theta)}{r} = r\cos(\theta)\sin(\theta),$$

which is a linear approximation, in terms of r, but one that cannot be constructed reasonably from the values of the partial derivatives.

It turns out that the correct definition of the generalization of differentiability is to take the approximation property as the defining condition.

DEFINITION 3.4.2. Let $X \subset \mathbf{R}^n$ be open and $f\colon X \to \mathbf{R}^m$ be a function. Let u be a linear map $\mathbf{R}^n \to \mathbf{R}^m$ and $x_0 \in X$. We say that f is *differentiable at x_0 with differential u* if

$$\lim_{\substack{x\to x_0\\ x\neq x_0}} \frac{1}{\|x-x_0\|}(f(x) - f(x_0) - u(x-x_0)) = 0$$

where the limit is in $\mathbf{R}^m$. We then denote $df(x_0) = u$.

If f is differentiable at every $x_0 \in X$, then we say that f is differentiable on X.

This definition means that, close to x_0, we can approximate $f(x)$ by the affine-linear function $g\colon \mathbf{R}^n \to \mathbf{R}^m$ defined by

$$g(x) = f(x_0) + u(x - x_0),$$

with an error that becomes much smaller than $\|x - x_0\|$ as x gets close to x_0.

REMARK 3.4.3. (1) If we write

$$f(x) = (f_1(x), \ldots, f_m(x))$$

and similarly write

$$u(x) = (u_1(x), \ldots, u_m(x))$$

where f_1, ..., f_m are functions $X \to \mathbf{R}$ and u_1, ..., u_m are linear maps $\mathbf{R}^n \to \mathbf{R}$, then the definition of limit shows that f is differentiable with differential u if and only if, for each i, the function f_i is differentiable with differential u_i.

Furthermore, a linear map $u\colon \mathbf{R}^n \to \mathbf{R}$ (or *linear form*) has the simple form

$$u(x_1, \ldots, x_n) = a_1x_1 + \cdots + a_nx_n$$

for some coefficients a_1, ..., a_n in $\mathbf{R}$. So, in the case $m = 1$, the approximation for $f(x)$ is

$$f(x_0) + a_1x_1 + \cdots + a_nx_n,$$

and depends only on the n numbers $(a_1, \ldots, a_n)$. These are the analogues of the single derivative $f'(x_0)$ when $n = 1$, and we will see that these coefficients are just the values of the partial derivatives of f at x_0.

(2) Suppose $n = 1$ and $m = 1$. Then the definition is equivalent to

$$0 = \lim_{\substack{x \to x_0 \\ x \neq 0}} \frac{f(x) - f(x_0) - a(x - x_{)}}{x - x_0} = \lim_{\substack{x \to x_0 \\ x \neq 0}} \frac{f(x) - f(x_0)}{x - x_0} - a$$

where a is the (unique) coefficient representing the linear map $u\colon \mathbf{R} \to \mathbf{R}$ (because $\|x - x_0\| = |x - x_0|$ and a function tends to 0 if and only if its absolute value does). In other words, f is differentiable according to the definition above if and only if f is differentiable at x_0 in the sense of Analysis I, with derivative $f'(x_0) = a$.

The following proposition shows that differentiable functions have some good properties: they are continuous, and have partial derivatives, which can be computed easily in terms of the differential.

Proposition 3.4.4. *Let $X \subset \mathbf{R}^n$ be open and $f\colon X \to \mathbf{R}^m$ be a function that is differentiable on X.*

(1) *The function f is continuous on X.*

(2) *The function f admits partial derivatives on X with respect to each variable.*

(3) *Assume that $m = 1$. Let $x_0 \in X$, and let*

$$u(x_1, \ldots, x_n) = a_1x_1 + \cdots + a_nx_n$$

be the differential of f at x_0. We then have

$$\partial_{x_i} f(x_0) = a_i$$

for $1 \leqslant i \leqslant n$.

Proof. (1) Let $x_0 \in X$. For $x \neq x_0$, write

$$f(x) = f(x_0) + u(x - x_0) + E(x)$$

for some $E(x) \in \mathbf{R}$. According to the definition, we have

$$\lim_{x \to x_0} \frac{E(x)}{\|x - x_0\|} = 0,$$

which implies that $E(x) \to 0$ as $x \to x_0$. Since u is continuous and $u(0) = 0$, we deduce that

$$\lim_{x \to x_0} f(x) = f(x_0),$$

which means that f is continuous on X.

(2) and (3): we consider only the case $n = 2$, $m = 1$ and $i = 1$ for simplicity, using (x, y) for the coordinates. Let $(x_0, y_0) \in X$. We define $E(x, y)$ by

$$f(x, y) = f(x_0, y_0) + a_1(x - x_0) + a_2(y - y_0) + E(x, y).$$

It follows that if we put $y = y_0$ and vary x only, we have

$$\frac{f(x, y_0) - f(x_0, y_0)}{x - x_0} = a_1 + 0 + \frac{E(x, y_0)}{x - x_0}.$$

Since $|x - x_0| = \|(x, y) - (x_0, y_0)\|$, the definition implies that

$$\lim_{x \to x_0} \frac{E(x, y_0)}{x - x_0} = 0,$$

and therefore

$$\lim_{x \to x_0} \frac{f(x, y_0) - f(x_0, y_0)}{x - x_0} = a_1,$$

which means that the partial derivative $\partial_x f$ exists at (x_0, y_0) and is equal to a_1. □

Example 3.4.5. (1) The simplest example of a differentiable function is an affine linear function

$$f(x) = y_0 + u(x)$$

where $y_0 \in \mathbf{R}^m$ and $u \colon \mathbf{R}^n \to \mathbf{R}^m$ is linear. Indeed, since $f(x_0) = y_0 + u(x_0)$, we get

$$f(x) = y_0 + u(x) = f(x_0) + u(x - x_0)$$

which means that f is differentiable at all x_0, with differential $df(x_0) = u$, independent of x_0.

(2) Consider the function $g \colon \mathbf{R}^2 \to \mathbf{R}$ of Example 3.4.1 (2). This is not differentiable at $(0, 0)$. Indeed, if it were, then since the two partial derivatives at $(0, 0)$ are equal to 0 (as we saw earlier), the proposition shows that the differential $u = df(0, 0)$ would be the zero linear map. But then we find that

$$\frac{1}{\|(x, y)\|}(g(x, y) - g(0, 0) - u(x, y)) = \frac{g(x, y)}{\|(x, y)\|} = \frac{xy}{x^2 + y^2},$$

and from Example 3.2.16, this quantity does not have a limit as $(x, y) \to (0, 0)$.

(3) Consider the case $m = 1$ in general. If a function $f \colon X \to \mathbf{R}$ is differentiable, then according to Proposition 3.4.4 (3), its differential at x_0 is the linear map $u \colon \mathbf{R}^m \to \mathbf{R}$ such that

$$u(t_1, \ldots, t_n) = \sum_{i=1}^{n} \frac{\partial f}{\partial x_i}(x_0) t_i$$

for all $t = (t_i) \in \mathbf{R}^n$. A convenient way to represent this is to write

$$u(t) = \nabla f(x_0) \cdot t,$$

where $\nabla f(x_0)$ is the gradient of f at x_0, and $x \cdot y$ denotes the scalar product of two vectors:

$$x \cdot y = x_1 y_1 + \cdots + x_n y_n.$$

The affine linear map that approximates f is then

$$g(x) = f(x_0) + \nabla f(x_0) \cdot (x - x_0).$$

The next issue is to know when a function is differentiable and to construct more differentiable functions (they would not be useful if they didn't exist). For this purpose, there are two basic results: (1) showing that various operations preserve differentiability; (2) giving a supply of functions for which it is easy to know that they are differentiable.

Proposition 3.4.6. *Let $X \subset \mathbf{R}^n$ be open, $f \colon X \to \mathbf{R}^m$ and $g \colon X \to \mathbf{R}^m$ differentiable functions on X.*

(1) The function $f + g$ is differentiable with differential $d(f + g) = df + dg$, and if $m = 1$, then fg is differentiable.

(2) If $m = 1$ and if $g(x) \neq 0$ for all $x \in X$, then f/g is differentiable.

The next proposition immediately implies that most elementary functions are differentiable:

Proposition 3.4.7. *Let $X \subset \mathbf{R}^n$ be open, $f\colon X \to \mathbf{R}^m$ a function on X. If f has all partial derivatives on X,* and if the partial derivatives of f are continuous *on X, then f is differentiable on X, with differential determined by its partial derivatives, in the sense that the matrix of the differential $df(x_0)$, with respect to the canonical basis of $\mathbf{R}^n$ and $\mathbf{R}^m$, is the Jacobi matrix of f at x_0.*

Example 3.4.8. (1) Let $n = 2$, $m = 1$ and consider $f(x,y) = \cos(x+y^2) - xe^y$. Pick $(x_0,y_0) = (\pi/4, 0)$. The function f is differentiable on $\mathbf{R}^2$ because its has partial derivatives and Jacobi matrix

$$J_f(x,y) = (-\sin(x+y^2) - e^y, -2y\sin(x+y^2) - xe^y)$$

where the components are continuous functions on $\mathbf{R}^2$.

At (x_0,y_0), the Jacobi matrix becomes

$$J_f(\pi/4,0) = (-\sin(\pi/4) - 1, 0 - \pi/4) = -(1 + 1/\sqrt{2}, \pi/4).$$

so that the differential $u = df(x_0,y_0)$ is the linear form

$$u(x,y) = -(1+1/\sqrt{2})x + \pi y/4,$$

and the affine-linear approximation $g(x,y)$ to $f(x,y)$ close to (x_0,y_0) is given by

$$g(x,y) = f(\pi/4,0) + u(x-x_0, y-y_0) = \frac{\sqrt{2}}{2} - \frac{\pi}{4} - \Big(1 + \frac{1}{\sqrt{2}}\Big)\Big(x - \frac{\pi}{4}\Big) + \frac{\pi y}{4}.$$

(2) Any polynomial in n variables is differentiable on $\mathbf{R}^n$. Its partial derivatives are also polynomials in n variables.

(3) If f_1, ..., f_n are functions of class C^1 on $\mathbf{R}$ (so that their derivatives are defined and continuous), then the function

$$f(x) = f_1(x_1)\cdots f_n(x_n)$$

is differentiable on $\mathbf{R}^n$.

The other important rule about differentiable functions is the *chain rule.*

Proposition 3.4.9 (Chain rule). *Let $X \subset \mathbf{R}^n$ be open, $Y \subset \mathbf{R}^m$ be open, and let $f\colon X \to Y$ and $g\colon Y \to \mathbf{R}^p$ be differentiable functions. Then $g \circ f\colon X \to \mathbf{R}^p$ is differentiable on X, and for any $x \in X$, its differential is given by the composition*

$$d(g \circ f)(x_0) = dg(f(x_0)) \circ df(x_0).$$

In particular, the Jacobi matrix satisfies

$$J_{g\circ f}(x_0) = J_g(f(x_0))J_f(x_0)$$

where the right-hand side is a matrix product.

Example 3.4.10. (1) To see this formula concretely, assume $n = m = p = 2$, and write

$$f(x,y) = \begin{pmatrix} f_1(x,y) \\ f_2(x,y) \end{pmatrix}, \quad g(u,v) = \begin{pmatrix} g_1(u,v) \\ g_2(u,v). \end{pmatrix}$$

Then the Jacobi matrices are

$$J_f(x,y) = \begin{pmatrix} \partial_x f_1 & \partial_y f_1 \\ \partial_x f_2 & \partial_y f_2 \end{pmatrix}, \qquad J_g(u,v) = \begin{pmatrix} \partial_u g_1 & \partial_v g_1 \\ \partial_u g_2 & \partial_v g_2 \end{pmatrix}.$$

The matrix product $J_g J_f$ gives us the Jacobi matrix of $g \circ f$, namely

$$J_{g\circ f}(x,y) = \begin{pmatrix} \partial_u g_1 \partial_x f_1 + \partial_v g_1 \partial_x f_2 & \partial_u g_1 \partial_y f_1 + \partial_v g_1 \partial_y f_2 \\ \partial_u g_2 \partial_x f_1 + \partial_v g_2 \partial_x f_2 & \partial_u g_2 \partial_y f_1 + \partial_v g_2 \partial_y f_2 \end{pmatrix}.$$

When evaluating such a Jacobi matrix at a given point x_0, it must be remembered that all partial derivatives of f are evaluated at x_0, and all partial derivatives of g are evaluated at $y_0 = f(x_0)$.

(2) Suppose $p = 1$, so that $g \circ f$ is real-valued. For the partial derivative of $g \circ f$ with respect to x_1, for instance, we get

$$\frac{\partial(g \circ f)}{\partial x_1}(x_0) = \frac{\partial g}{\partial y_1}(y_0)\frac{\partial f_1}{\partial x_1}(x_0) + \frac{\partial g}{\partial y_2}(y_0)\frac{\partial f_2}{\partial x_1}(x_0) + \cdots + \frac{\partial g}{\partial y_m}(y_0)\frac{\partial f_m}{\partial x_1}(x_0),$$

or in other words

$$\frac{\partial(g \circ f)}{\partial x_1}(x_0) = \sum_{j=1}^{m} \frac{\partial g}{\partial y_j}(y_0)\frac{\partial f_j}{\partial x_1}(x_0),$$

where $y_0 = f(x_0)$ and the variables in $\mathbf{R}^m$ are $(y_1, \ldots, y_m)$.

(A way to remember the formula is to think that the j-th coordinate variable y_j in the "denominator" of the partial derivative for g corresponds to the "numerator" f_j, which is the j-th coordinate of f).

(3) Let f, $g\colon \mathbf{R}^n \to \mathbf{R}$ be two functions. Define $h(x, y) = (f(x, y), g(x, y))$ and $m(u, v) = uv$, so that $m \circ h(x, y) = f(x, y)g(x, y)$. The Jacobi matrices of h and m are

$$J_h(x, y) = \begin{pmatrix} \partial_x f & \partial_y f \\ \partial_x g & \partial_y g \end{pmatrix}, \qquad J_m(u, v) = (v, u)$$

(the Jacobi matrix for m is just a row vector). It follows therefore that

$$\frac{\partial(fg)}{\partial x} = v\partial_x f + u\partial_x g,$$

evaluated at (x, y), which (since we must replace u and v by the coordinates of $h(x, y)$) means that

$$\frac{\partial(fg)}{\partial x} = g(x, y)\partial_x f(x, y) + f(x, y)\partial_x g(x, y),$$

a formula that we can recognize as the Leibniz rule.

(4) Let $I \subset \mathbf{R}$ be an open interval. Consider $f\colon I \to \mathbf{R}^m$ and $g\colon \mathbf{R}^m \to \mathbf{R}$, so that the composite is a function $g \circ f\colon I \to \mathbf{R}$. If f is differentiable on I (which means that each component is a differentiable function of one variable) and if g is differentiable on $\mathbf{R}^n$, then $g \circ f$ is differentiable on I, and its derivative, which is just the partial derivative with respect to the only variable is determined by

$$(g \circ f)'(t) = dg(f(t))\ f'(t),$$

i.e., the linear map $dg(f(t))\colon \mathbf{R}^m \to \mathbf{R}$ (whose coefficients are the partial derivatives of g), applied to the vector $f'(t) \in \mathbf{R}^m$. If we write $f(t) = (f_1(t), \ldots, f_m(t))$, this is just

$$(g \circ f)'(t) = \frac{\partial g}{\partial y_1}(f(t))f_1'(t) + \cdots + \frac{\partial g}{\partial y_m}(f(t))f_m'(t).$$

Another convenient expression as a scalar product is just

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

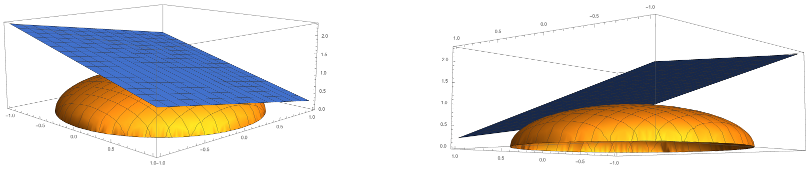

Figure 3.4. Graph of $f(x,y) = \sqrt{1-x^2-y^2}$ and tangent space at $(1/2, 1/3)$

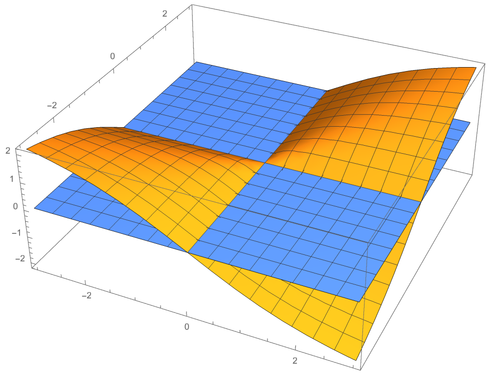

Figure 3.5. Graph of $g(x,y) = xy/\sqrt{1-x^2-y^2}$ and horizontal plane

Definition 3.4.11. Let $X \subset \mathbf{R}^n$ be open and $f\colon X \to \mathbf{R}^m$ a function that is differentiable. Let $x_0 \in X$ and $u = df(x_0)$ be the differential of f at x_0. The graph of the affine linear approximation

$$g(x) = f(x_0) + u(x - x_0)$$

from $\mathbf{R}^n$ to $\mathbf{R}^m$, or in other words the set

$$\{(x,y) \in \mathbf{R}^n \times \mathbf{R}^m \ : \ y = f(x_0) + u(x - x_0)$$

is called the *tangent space at* x_0 to the graph of f.

The tangent space at a point generalizes the tangent line for the graph of a function of one variable. It is the affine subspace in $\mathbf{R}^m$ that is “the best” fit to the graph of the function f around x_0. It is an affine space of dimension n, since it can be parameterized by $x \in \mathbf{R}^n$, which determines uniquely the corresponding point $y = f(x_0) + u(x - x_0)$ such that (x,y) belongs to the tangent space.

We can also write the points of the tangent space in the form

$$(x,y) = (x_0, f(x_0)) + (x - x_0, u(x - x_0))$$

which shows that it is the set of points $(x_0, y_0) + w$, where w belongs to the graph of u, which is a linear subspace of dimension n in $\mathbf{R}^{n+m}$. We say that this linear subspace is the *linear subspace parallel* to the tangent space at x_0.

Example 3.4.12. (1) Figure 3.4 illustrates (from two different angles) the graph of the function

$$f(x,y) = \sqrt{1-x^2-y^2}$$

(which is demi-sphere of radius 1 centered at $(0,0)$) and the tangent space at the point $(x,y) = (1/2, 1/3)$.

(2) Consider again the function $g(x,y) = xy/\sqrt{x^2+y^2}$ of Example 3.4.5. Figure 3.5 shows the graph of g and the horizontal plane $z = 0$ in $\mathbf{R}^3$ that “would be” the tangent plane if the function was differentiable at $(0,0)$.

(3) Define

$$f(x, y) = \sqrt{x^2 + y^2}.$$

Let $(x_0, y_0) = (3, 4)$. The tangent plane to the graph of f at the point (x_0, y_0) is the set of all (x, y, z) in $\mathbf{R}^3$ such that

$$z = f(3, 4) + \nabla f(3, 4) \cdot (x - 3, y - 4).$$

We have $f(3, 4) = \sqrt{9 + 16} = 5$, and the gradient at an arbitrary point is given by

$$\nabla f(x, y) = \begin{pmatrix} \frac{x}{\sqrt{x^2+y^2}} \\ \frac{y}{\sqrt{x^2+y^2}} \end{pmatrix}$$

so that $\nabla f(x_0, y_0) = (3/5, 4/5)$. The equation of the tangent plane becomes

$$z = 5 + 3(x - 3)/5 + 4(y - 4)/5.$$

If a function is differentiable at a point $x_0 \in \mathbf{R}^n$, one meaning of the linear map $u = df(x_0)$ is that the value $u(v)$, for a vector $v \in \mathbf{R}^n$, gives the "directional derivative" in the direction v, in the sense of the following definition:

Definition 3.4.13. Let $X \subset \mathbf{R}^n$ be an open set and let $f\colon X \to \mathbf{R}^m$ be a function. Let $v \in \mathbf{R}^n$ be a non-zero vector and $x_0 \in X$. We say that f has *directional derivative* $w \in \mathbf{R}^m$ *in the direction* v, if the function g defined on the set

$$I = \{t \in \mathbf{R} \,:\, x_0 + tv \in X\}$$

by

$$g(t) = f(x_0 + tv)$$

has a derivative at $t = 0$, and this is equal to w.

In other words, this means that the limit

$$\lim_{\substack{t \to 0 \\ t \neq 0}} \frac{f(x_0 + tv) - f(x_0)}{t}$$

exists and is equal to w.

Remark 3.4.14. It is easy to see that because X is open, the set I contains an open interval $] - \delta, \delta[$ for some $\delta > 0$, so that the derivability of g at $t = 0$ makes sense.

Proposition 3.4.15. *Let $X \subset \mathbf{R}^n$ be an open set and let $f\colon X \to \mathbf{R}^m$ be a differentiable function. Then for any $x \in X$ and non-zero $v \in \mathbf{R}^n$, the function f has a directional derivative at x_0 in the direction v, equal to $df(x_0)(v)$.*

Remark 3.4.16. (1) What is important to notice in this proposition, is that the values of the directional derivatives are linear with respect to the vector v. So if we know the directional derivatives w_1 and w_2 in directions v_1 and v_2, then it follows that the directional derivative in direction $v_1 + v_2$ is $w_1 + w_2$.

(2) If we take v to be the vector e_i of the canonical basis of $\mathbf{R}^n$, then the directional derivative in direction e_i is simply the partial derivative with respect to the i-th variable.

Example 3.4.17. (1) Consider the function $g(x, y) = xy/\sqrt{x^2 + y^2}$ of Example 3.4.5, (2). Although it is not differentiable at $(0, 0)$, it has directional derivatives in all directions $(u, v) \neq (0, 0)$, since $g(0, 0) = 0$ and

$$\frac{g(tu, tv) - g(0, 0)}{t} = \frac{uv}{\sqrt{u^2 + v^2}}.$$

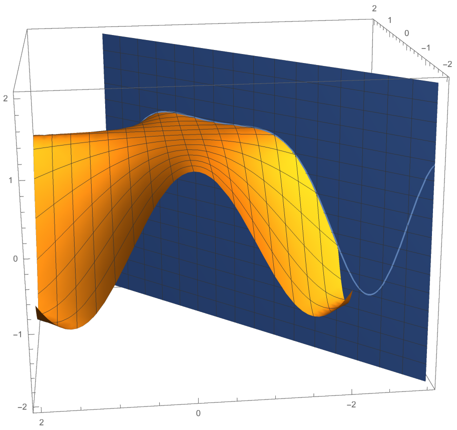

Figure 3.6. Directional derivative

(In fact, this is just $g(u, v)$). But this expression is not linear with respect to (u, v).

(2) Suppose that $m = 1$. Then the directional derivative in direction u is a real number which has the following geometric meaning: intersect the graph of f in $\mathbf{R}^{n+1}$ with the plane perpendicular to the hyperplane $\mathbf{R}^n = \mathbf{R}^n \times \{0\}$, which passes through $(x_0, 0)$ and $(x_0 + v, 0)$. This gives a set Γ which is the graph of the function $g(t) = f(x_0 + tv)$. Now, if v has length 1, then the slope of the tangent line to Γ at $(x_0, f(x_0))$ is equal to the directional derivative at that point.

For instance, define $f(x, y) = \cos(xy)$ and consider the point $(0, -1)$ and the direction $(1, 1)$. Figure 3.6 displays the graph and the corresponding perpendicular plane.

We now suppose $m = 1$. Let $f\colon X \to \mathbf{R}$ be differentiable, and let $x_0 \in X$. The tangent space at x_0 to the graph of f is the set of $(x, y) \in \mathbf{R}^n \times \mathbf{R}$ such that

$$y = f(x_0) + \nabla f(x_0) \cdot (x - x_0).$$

This is an affine space of dimension n, and the corresponding linear subspace in $\mathbf{R}^n$ is the graph of the linear map

$$x \mapsto \nabla f(x_0) \cdot x,$$

in other words the set of all $(x, y) \in \mathbf{R}^n \times \mathbf{R}$ such that $y = \nabla f(x_0) \cdot x$. A good way to visualize or interpret this linear space is to observe that it is the set of vectors orthogonal to the vector

$$n_0 = (-\nabla f(x_0), 1) \in \mathbf{R}^n \times \mathbf{R}.$$

Indeed, we have

$$y - \nabla f(x_0) \cdot x = (x, y) \cdot n_0$$

where the right-hand side is now a scalar product in $\mathbf{R}^{n+1}$.

The gradient has another important interpretation, which generalizes the fact that for a function of one variable, the sign of the derivative indicates whether the function is (locally) increasing or decreasing. Precisely, suppose that the gradient vector $\nabla f(x_0)$ is non-zero. Then the vector $w_0 = \nabla f(x_0)$ points in the "direction of greatest increase" of the function f. In other words, it points in the direction where the directional derivative is the largest. This follows from the fact that

$$f(x) - f(x_0) = \nabla f(x_0) \cdot (x - x_0) + (\text{small error})$$

and that we know (Cauchy-Schwarz inequality) that

$$|\nabla f(x_0) \cdot (x - x_0)| \leqslant \|\nabla f(x_0)\| \, \|x - x_0\| = \|w_0\| \, \|x - x_0\|$$

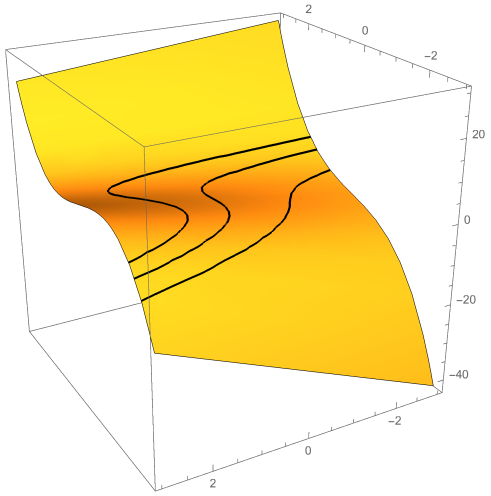

FIGURE 3.7. Some level curves of $f(x, y) = x^3 - xy + 2y$

with equality if $x - x_0$ is proportional to w_0, which corresponds to varying x in the direction of w_0.

Another way to see this is in terms of directional derivatives. Let $v \in \mathbf{R}^n$ be a vector of length one. If we remember that the scalar product of two vectors in $\mathbf{R}^n$ is the product of their lengths with the cosine of the angle, the directional derivative of f in the direction v at x_0 is

$$\nabla f(x_0) \cdot v = \|\nabla f(x_0)\| \cos(\theta)$$

where $\theta \in [0, \pi]$ is the angle between the gradient and the direction v. This is maximal when $\theta = 0$, which means that v is proportional to $\nabla f(x_0)$.

EXAMPLE 3.4.18. Think of the graph of $f\colon \mathbf{R}^2 \to \mathbf{R}$ as giving the height of a mountain above the point with coordinates (x, y) of the map of a region of the earth. Then the gradient $\nabla f(x_0)$ is a vector in $\mathbf{R}^2$, and it points in the direction in which the height grows faster: if one wants to climb the slope as quickly as possible, one should walk always in the direction of the gradient.

Yet another related geometric property of the gradient is that it is perpendicular to the "level sets" determined by an equation of the form $f(x) = c$, where $c \in \mathbf{R}$ is a fixed real number.

To be more precise, fix c, and denote by L_c the set of all $x \in X$ where $f(x) = c$. Let $x_0 \in L_c$ be any point in this set. Then, for any differentiable function of one variable $\gamma\colon]-1, 1[\to \mathbf{R}^n$ such that $f(\gamma(t)) = c$ for all $t \in I$ and $\gamma(0) = x_0$, the gradient $\nabla f(x_0)$ is orthogonal in $\mathbf{R}^n$ to the vector $\gamma'(0)$, which is "tangent" to the level set. This is simply because, by the Chain Rule, we have the relation

$$0 = (f \circ \gamma)'(0) = \nabla f(x_0) \cdot \gamma'(0).$$

EXAMPLE 3.4.19. The simplest example is $f(x, y) = x^2 + y^2$. Then the level sets L_c are empty if $c < 0$, a single point if $c = 0$, and a circle of radius $\sqrt{c}$ if $c > 0$. In this last case, the gradient vector at any point of L_c is $(2x, 2y)$, and therefore points in the direction orthogonal to the circle.

3.5. Higher derivatives

We can often straightforwardly compute partial derivatives of a function $f\colon \mathbf{R}^n \to \mathbf{R}^m$, and check that not only they exist, and are continuous, but also themselves admit further continuous partial derivatives, etc. This leads naturally to the notion of function of class C^k.

Definition 3.5.1. Let $X \subset \mathbf{R}^n$ be open and $f\colon X \to \mathbf{R}^m$.

We say that f is of class C^1 if f is differentiable on X and all its partial derivatives are continuous. The set of functions of class C^1 from X to $\mathbf{R}^m$ is denoted $C^1(X;\mathbf{R}^m)$.

Let $k \geqslant 2$. We say, by induction, that f is of class C^k if it is differentiable and each partial derivative $\partial_{x_i} f\colon X \to \mathbf{R}^m$ is of class C^{k-1}. The set of functions of class C^k from X to $\mathbf{R}^m$ is denoted $C^k(X;\mathbf{R}^m)$.

If $f \in C^k(X;\mathbf{R}^m)$ for *all* $k \geqslant 1$, then we say that f is of class C^∞. The set of such functions is denoted $C^\infty(X;\mathbf{R}^m)$.

In practical terms, this means that one has to check all possible combinations of k derivatives, with respect to any combination of k variables, and always obtain continuous functions.

Example 3.5.2. (1) If $f(x) = (f_1(x), \ldots, f_m(x))$, then f is of class C^k if and only if each f_i is of class C^k.

(2) If f, g are of class C^k, then so is $f+g$; if $m=1$, then so is fg, and if $g(x) \neq 0$ for all $x \in X$, then so is f/g of class C^k.

(3) If $f(x) = f_1(x_1)\cdots f_n(x_n)$ has separated variables, and if f_i is of class C^k, then f is of class C^k.

(4) Any polynomial in n variables is of class C^∞.

(5) Any partial derivative is a linear operation on the functions.

(6) Suppose that f is of class C^k, and that $f(X) \subset Y$, where $Y \subset \mathbf{R}^m$ is open, and that $g\colon Y \to \mathbf{R}^p$ is also of class C^k. Then the composite $g \circ f$ is also of class C^k. This follows, by induction on k, from the chain rule that expresses partial derivatives of $g \circ f$ in terms of partial derivatives of f and g.

Suppose that $k=2$. Then, in order to show that a function f is of class C^2, we first check that f is differentiable with continuous partial derivatives. There are n such checks to make since f has n partial derivatives. Next there are apparently n^2 second order derivatives, namely

$$\partial_{x_1}(\partial_{x_1} f), \quad \partial_{x_1}(\partial_{x_2} f), \quad \cdots \quad \partial_{x_1}(\partial_{x_n} f),$$

until

$$\partial_{x_n}(\partial_{x_1} f), \quad \partial_{x_n}(\partial_{x_2} f), \quad \cdots \quad \partial_{x_n}(\partial_{x_n} f).$$

However, if we do it in practice, we see that these derivatives are not independent at all.

Example 3.5.3. Let $f(x,y) = e^{-x^2\sqrt{y}}$ for $x \in \mathbf{R}$, $y > 0$. Then

$$\nabla f(x,y) = (-2x\sqrt{y}\exp(-x^2\sqrt{y}), -\tfrac{x^2}{2\sqrt{y}}\exp(-x^2\sqrt{y})).$$

Now we compute the four partial derivatives of order 2:

$$\begin{aligned}
\partial_{x^2} f &= -2\sqrt{y}\exp(-x^2\sqrt{y}) + 4x^2\sqrt{y}\exp(-x^2\sqrt{y})\\
\partial_{xy} f &= -\frac{x}{\sqrt{y}}\exp(-x^2\sqrt{y}) + \frac{x^3}{\sqrt{y}}\exp(-x^2\sqrt{y})\\
\partial_{yx} f &= -\frac{x}{\sqrt{y}}\exp(-x^2\sqrt{y}) + \frac{x^3}{\sqrt{y}}\exp(-x^2\sqrt{y})\\
\partial_{y^2} f &= \frac{x^2}{4y^{3/2}}\exp(-x^2\sqrt{y}) + \frac{x^4}{4y}\exp(-x^2\sqrt{y}).
\end{aligned}$$

We see here that $\partial_{xy} f = \partial_{yx} f$. This is a general fact.

Proposition 3.5.4 (Mixed derivatives commute). *Let $k \geqslant 2$. Let $X \subset \mathbf{R}^n$ be open and let $f\colon X \to \mathbf{R}^m$ be a function of class C^k. Then the partial derivatives of order k are independent of the order in which the partial derivatives are taken: for any variables x and y, we have*

$$\partial_{x,y} f = \partial_{y,x} f,$$

and for any variables x, y, z, we have

$$\partial_{x,y,z} f = \partial_{x,z,y} f = \partial_{y,z,x} f = \partial_{z,x,y} f = \cdots$$

etc...

Example 3.5.5. (1) To convince oneself that this should be true, it is best to look at a monomial first. Say

$$f(x,y,z) = x^a y^b z^c.$$

Then

$$\partial_{x,y} f = abx^{a-1}y^{b-1}z^c = \partial_{y,x} f$$

and

$$\partial_{x,y,z} f = abcx^{a-1}y^{b-1}z^{c-1}$$

is the same however we order x, y and z when taking the derivatives.

(2) Let $k = 2$. In order to ensure that $\partial_{x,y} f = \partial_{y,x} f$, it is essential to know that f is of class C^2 (so that all partial derivatives of order 2 are continuous), and there are counterexamples otherwise. For instance, one can easily check that

$$f(x,y) = \frac{xy(x^2-y^2)}{x^2+y^2}, \quad (x,y) \neq 0, \qquad f(0,0) = 0$$

defines a function $\mathbf{R}^2 \to \mathbf{R}$ which is differentiable (with $\nabla f(0,0) = 0$) and admits partial derivatives of order $\leqslant 2$, but at $(0,0)$, we have

$$\partial_{x,y} f(0,0) = 1, \qquad \partial_{y,x} f(0,0) = -1.$$

In polar coordinates, we have $f(x,y) = r\sin(4\theta)/4$.

Because of the symmetry, we introduce a more compact notation for mixed derivatives of "large order". If we want to take a derivative of order k, we select the first variable (say x_{i_1}), compute the partial derivative, then select a second (say x_{i_2}), compute the second derivative $\partial_{x_{i_2},x_{i_1}}$, etc, up to the i_k-th variable. But, provided Proposition 3.5.4 applies (i.e., f is of class C^k), the resulting partial derivative

$$\partial_{x_{i_k}}\partial_{x_{i_{k-1}}}\cdots\partial_{x_{i_1}} f$$

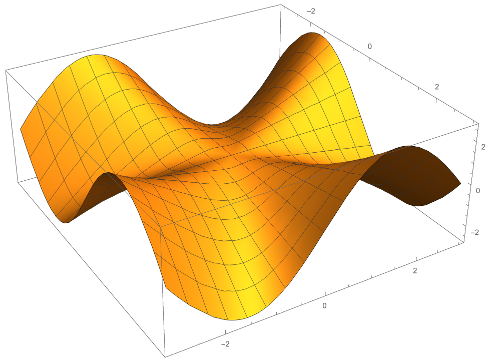

FIGURE 3.8. Function with non-symmetric mixed second derivatives

only depends on how many times we took the derivative with respect to each variable. In other words, let m_1 be the number of indices j such that $i_j = 1$, ..., m_n be the number of j such that $i_j = n$. Then

$$\partial_{x_{i_k}}\partial_{x_{i_{k-1}}}\cdots\partial_{x_{i_1}} f = \partial_{x_1^{m_1},x_2^{m_2},\dots,x_n^{m_n}} f.$$

Let $m = (m_1, \dots, m_n)$. This is a vector of non-negative integers, with

$$m_1 + \cdots + m_n = k$$

(since, in total, we have taken k derivatives). We may use any of the following notation for these expressions:

$$\partial_{x_1^{m_1},\dots,x_n^{m_n}} f = \frac{\partial^k f}{\partial x^m} = \partial_x^m f = D^m f = \partial^m f.$$

REMARK 3.5.6. The linearity of the partial derivatives means that

$$\partial_x^m(af_1 + bf_2) = a\partial_x^m f_1 + b\partial_x^m f_2$$

whenever both partial derivatives on the right-hand side exist.

EXAMPLE 3.5.7. Suppose $n = 3$ and $k = 4$. There are then 15 possible derivatives of order 4, corresponding to the tuples

$$\begin{array}{lllll} m = (4,0,0), & m = (3,1,0), & m = (3,0,1), & m = (2,2,0), & m = (2,1,1) \\ m = (2,0,2), & m = (1,3,0), & m = (1,2,1), & m = (1,1,2), & m = (1,0,3) \\ m = (0,4,0), & m = (0,3,1), & m = (0,2,2), & m = (0,1,3), & m = (0,0,4). \end{array}$$

For instance, $m = (1,1,2)$ corresponds to the derivative

$$\frac{\partial^4 f}{\partial x \partial y \partial^2 z}.$$

EXAMPLE 3.5.8. (Laplace operator) Let X be open in $\mathbf{R}^n$, and let $f \in C^2(X)$. The gradient of f belongs to $C^1(X; \mathbf{R}^n)$, so we can compute its divergence (Definition 3.3.11). We obtain

$$\operatorname{div}(\nabla(f)) = \sum_{i=1}^{n} \frac{\partial}{\partial x_i}\Big(\frac{\partial f}{\partial x_i}\Big) = \sum_{i=1}^{n} \frac{\partial^2 f}{\partial x_i^2}.$$

This differential expression is called the *Laplacian of f*, and is denoted Δf.

For the case $k = 2$, $m = 1$, we organize in a matrix the partial derivatives of order 2 of a function $X \to \mathbf{R}$, namely the derivatives

$$\frac{\partial^2 f}{\partial x_i \partial x_j},$$

where $1 \leqslant i, j \leqslant n$. For a function f of class C^2, this matrix will be symmetric.

Definition 3.5.9 (Hessian). Let $X \subset \mathbf{R}^n$ be open and $f\colon X \to \mathbf{R}$ a C^2 function. For $x \in X$, the *Hessian matrix* of f at x is the symmetric square matrix

$$\mathrm{Hess}_f(x) = \left(\partial_{x_i,x_j} f\right)_{1 \leqslant i,j \leqslant n}.$$

We also sometimes write simply $H_f(x)$.

Example 3.5.10. Let $n = 3$ and $f(x, y, z) = x^2 y - \cos(xz^3)$. Then we compute

$$\partial_x f = 2xy + z^3 \sin(xz^3), \quad \partial_y f = x^2, \quad \partial_z f = 3xz^2 \sin(xz^3)$$

and then we obtain the Hessian by further differentiation

$$\mathrm{Hess}_f(x, y, z) = \begin{pmatrix} 2y + z^6 \cos(xz^3) & 2x & 3z^2 \sin(xz^3) + xz^6 \cos(xz^3) \\ 2x & 0 & 0 \\ 3z^2 \sin(xz^3) + xz^6 \cos(xz^3) & 0 & 6xz \sin(xz^3) + 9x^2 z^6 \cos(xz^3) \end{pmatrix}.$$

3.6. Change of variable

An important application of the chain rule concerns the computation of partial derivatives after a *change of variable*. Here we have an open set $U \subset \mathbf{R}^n$ (with variables that we write $(y_1, \ldots, y_n)$, the "new" variables) and a change of variable $g\colon U \to X$ is a map that expresses the variables $(x_1, \ldots, x_n)$ in terms of $(y_1, \ldots, y_n)$, i.e., we consider

$$x_1 = g_1(y_1, \ldots, y_n), \qquad x_n = g_n(y_1, \ldots, y_n).$$

We should think of g as something "fixed" and very standard (such as going to polar coordinates, or to spherical coordinates, etc).

Whenever a function $f\colon X \to \mathbf{R}$ is given, the composite $h = f \circ g\colon U \to \mathbf{R}$ is the function f expressed in terms of the "new" variables y.

The chain rule then provides a way to express all partial derivatives of h in terms of those of f, and of the Jacobian matrix of the change of variable g. For instance

$$\partial_{y_1} h = \frac{\partial f}{\partial x_1}\frac{\partial g_1}{\partial y_1} + \cdots + \frac{\partial f}{\partial x_n}\frac{\partial g_n}{\partial y_1}.$$

Here, since we think that g is fixed, the corresponding partial derivatives are known quantities.

There are very common abuses of notation that may be very confusing at first, but that are extremely convenient:

(1) one thinks of f and h as being *the same function*, simply expressed in different coordinate systems, and one writes simply

$$\partial_{y_1} f = \frac{\partial f}{\partial x_1}\frac{\partial g_1}{\partial y_1} + \cdots + \frac{\partial f}{\partial x_n}\frac{\partial g_n}{\partial y_1}.$$

(2) one thinks of g_i as being the variable x_i, expressed in terms of the new variables $(y_1, \ldots, y_n)$, and replaces g_i by x_i, so the expression becomes

$$\partial_{y_1} f = \frac{\partial f}{\partial x_1}\frac{\partial x_1}{\partial y_1} + \cdots + \frac{\partial f}{\partial x_n}\frac{\partial x_n}{\partial y_1}.$$

Remark 3.6.1. The first of these two simplification is very natural if we think of a function like "the distance to the origin", which we can describe without referring to any particular choice of coordinate system.

The point of a change of variable is often to go back and forth, and one can solve for y in terms of x, and write down the corresponding relations

$$\partial_{x_1} f = \frac{\partial f}{\partial y_1}\frac{\partial y_1}{\partial x_1} + \cdots + \frac{\partial f}{\partial y_n}\frac{\partial y_n}{\partial x_1}.$$

In practice, this can be done by solving the linear system of equations represented by the chain rule.

Example 3.6.2. One of the most important example is the change of variable to polar coordinates in $\mathbf{R}^2$. The polar coordinates are $(r, \theta) \in U =]0, +\infty[\times\mathbf{R}$ (or sometimes $U =]0, +\infty[\times[0, 2\pi[$) and they parameterize the plane minus the origin $(0, 0)$ by

$$\begin{cases} x = r\cos\theta \\ y = r\sin\theta. \end{cases}$$

In other words, we consider the map

$$g \colon U \to \mathbf{R}^2$$

such that $g(r, \theta) = (r\cos\theta, r\sin\theta)$, and to express a function $f \colon \mathbf{R}^2 \to \mathbf{R}$ in polar coordinates means replacing f by $h = f \circ g \colon U \to \mathbf{R}$, so that

$$h(r, \theta) = f(r\cos\theta, r\sin\theta).$$

The Jacobian matrix of the change of variable is given by

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

(with determinant r). The chain rule leads to the formulas

$$\begin{aligned} \partial_r h &= \cos(\theta)\partial_x f + \sin(\theta)\partial_y f \\ \partial_\theta h &= -r\sin(\theta)\partial_x f + r\cos(\theta)\partial_y f \end{aligned}$$

(where all partial derivatives of f are evaluated implicitly at $(r\cos\theta, r\sin\theta)$.) This is also often expressed as

$$\begin{aligned} r\partial_r h &= x\partial_x f + y\partial_y f \\ \partial_\theta h &= -y\partial_x f + x\partial_y f. \end{aligned}$$

With the short-hand notation discussed earlier, this becomes

$$\begin{aligned} r\partial_r f &= x\partial_x f + y\partial_y f \\ \partial_\theta f &= -y\partial_x f + x\partial_y f. \end{aligned}$$

Solving for $\partial_x f$ and $\partial_y f$, we obtain the relations

$$\begin{aligned} \partial_x f &= \cos(\theta)\partial_r h - \frac{1}{r}\sin(\theta)\partial_\theta h \\ \partial_y f &= \sin(\theta)\partial_r h + \frac{1}{r}\cos(\theta)\partial_\theta h \end{aligned}$$

(where all partial derivatives of h are evaluated implicitly at (x, y) such that $x = r\cos\theta$, $y = r\sin\theta$), or

$$\partial_x f = \cos(\theta)\partial_r f - \frac{1}{r}\sin(\theta)\partial_\theta f \tag{3.2}$$

$$\partial_y f = \sin(\theta)\partial_r f + \frac{1}{r}\cos(\theta)\partial_\theta f \tag{3.3}$$

in abbreviated form.

One can iterate applying these partial derivatives to obtain expressions for higher derivatives. For instance, let us compute the Laplace operator

$$\Delta f = \partial_{x^2} f + \partial_{y^2} f$$

in polar coordinates (see Example 3.5.8). Using the formula (3.2) twice, we have

$$\begin{aligned}\partial_{x^2} f &= \cos(\theta)\partial_r(\partial_x f) - \frac{1}{r}\sin(\theta)\partial_\theta(\partial_x f)\\ &= \cos(\theta)\partial_r\Big(\cos(\theta)\partial_r f - \frac{1}{r}\sin(\theta)\partial_\theta f\Big) - \frac{1}{r}\sin(\theta)\partial_\theta\Big(\cos(\theta)\partial_r f - \frac{1}{r}\sin(\theta)\partial_\theta f\Big).\end{aligned}$$

Computing further these expressions, this gives

$$\begin{aligned}\partial_{x^2} f = {}&\cos(\theta)\Big\{\cos(\theta)\partial_{r^2} f + \frac{1}{r^2}\sin(\theta)\partial_\theta f - \frac{1}{r}\sin(\theta)\partial_{r\theta} f\Big\}\\ &- \frac{1}{r}\sin(\theta)\Big\{-\sin(\theta)\partial_r f + \cos(\theta)\partial_{r\theta} f - \frac{1}{r}\cos(\theta)\partial_\theta f - \frac{1}{r}\sin(\theta)\partial_{\theta^2} f\Big\}\\ &= \cos^2(\theta)\partial_{r^2} f + \frac{2}{r^2}\cos(\theta)\sin(\theta)\partial_\theta f - \frac{2}{r}\cos(\theta)\sin(\theta)\partial_{r\theta} f + \frac{1}{r}\sin^2(\theta)\partial_r f + \frac{1}{r^2}\sin^2(\theta)\partial_{\theta^2} f.\end{aligned}$$

A similar computation using instead (3.3) twice gives the formula

$$\partial_{y^2} f = \sin^2(\theta)\partial_{r^2} f - \frac{2}{r^2}\cos(\theta)\sin(\theta)\partial_\theta f + \frac{2}{r}\cos(\theta)\sin(\theta)\partial_{r\theta} f + \frac{1}{r}\cos^2(\theta)\partial_r f + \frac{1}{r^2}\cos^2(\theta)\partial_{\theta^2} f.$$

We conclude that (for a C^2 function f), we have

$$\partial_{x^2} f + \partial_{y^2} f = \partial_{r^2} f + \frac{1}{r}\partial_r f + \frac{1}{r^2}\partial_s\theta^2 f.$$

We look at a concrete example. Let $f(x, y) = \exp(x^2 + y^2)$. The corresponding expression in polar coordinates is $h(r, \theta) = \exp(r^2)$. We can compute the gradient of f and Δf using the polar coordinates by writing

$$\begin{aligned}\nabla f &= \begin{pmatrix}\partial_x f\\ \partial_y f\end{pmatrix} = \begin{pmatrix}\cos(\theta)\partial_r h - r^{-1}\sin(\theta)\partial_\theta h\\ \sin(\theta)\partial_r h + r^{-1}\cos(\theta)\partial_\theta h\end{pmatrix}\\ &= \begin{pmatrix}2r\cos\theta\exp(r^2)\\ 2r\sin\theta\exp(r^2)\end{pmatrix} = \begin{pmatrix}2x\exp(x^2 + y^2)\\ 2y\exp(x^2 + y^2)\end{pmatrix},\end{aligned}$$

and

$$\Delta f = \partial_{r^2} f + \frac{1}{r}\partial_r f + \frac{1}{r^2}\partial\theta^2 f = \partial_{r^2}(e^{r^2}) + 2e^{r^2} = (2 + 4r^2)e^{r^2} + 2e^{r^2} = 4(1 + x^2 + y^2)e^{x^2+y^2}.$$

3.7. Taylor polynomials

We consider in this section the case $m = 1$, and a function $f\colon X \to \mathbf{R}$. The affine-linear approximation for $f(x)$ when x is close to a point $x_0 \in X$ involves only the first derivatives of f, and is given by $T_1 f(x - x_0; x_0)$, where $T_1 f(y; x_0)$ is the function on $\mathbf{R}^n$ such that

$$T_1 f(y; x_0) = f(x_0) + \nabla f(x_0) \cdot y = f(x_0) + \sum_{i=1}^{n} \frac{\partial f}{\partial x_i}(x_0) y_i.$$

As a function of y, this is a polynomial of degree $\leqslant 1$ (it is of degree exactly 1 *unless* $\nabla f(x_0)$ is zero).

In the case $n = 1$, we know that we obtain better approximations to a function when considering higher derivatives, and building the Taylor polynomials of the function, defined by

$$T_k f(y; x_0) = f(x_0) + f'(x_0) y + \frac{f''(x_0)}{2} y^2 + \cdots + \frac{f^{(k)}(x_0)}{k!} y^k,$$

in the sense that

$$f(x) = T_k f(x - x_0; x_0) + (\text{remainder})$$

with, roughly speaking, a remainder that is much smaller than $|x - x_0|^k$ when $x \to x_0$.

The same is true in general, but the Taylor polynomials have now n variables.

Definition 3.7.1 (Taylor polynomials). Let $k \geqslant 1$ be an integer. Let $f\colon X \to \mathbf{R}$ be a function of class C^k on X, and fix $x_0 \in X$. The k-th Taylor polynomial of f at the point x_0 is the polynomial in n variables of degree $\leqslant k$ given by

$$T_k f(y; x_0) = f(x_0) + \sum_{i=1}^{n} \frac{\partial f}{\partial x_i}(x_0) y_i + \cdots$$
$$+ \sum_{m_1 + \cdots + m_n = k} \frac{1}{m_1! \cdots m_n!} \frac{\partial^k f}{\partial x_1^{m_1} \cdots \partial x_n^{m_n}}(x_0) y_1^{m_1} \cdots y_n^{m_n}$$

where the last sum ranges over the tuples of n non-negative integers such that the sum is k.

This seems a complicated formula, but comparing with the previous section, this means that the polynomial is a sum of monomials

$$\frac{1}{m_1! \cdots m_n!} \frac{\partial^j f}{\partial x_1^{m_1} \cdots \partial x_n^{m_n}}(x_0) y_1^{m_1} \cdots y_n^{m_n}$$

where j runs over all integers with $0 \leqslant j \leqslant k$, and for a given j, we consider all possible partial derivatives of order j (so that $m_1 + \cdots + m_n = j$) with the factorial coefficient.

Moreover, with clever notation, we can simplify this a lot. First, for any n-tuple $m = (m_1, \ldots, m_n)$ of non-negative integers, we define $|m| = m_1 + \cdots + m_n$ and we denote

$$m! = m_1! \cdots m_n!$$

and moreover, for variables y_1, ..., y_n, we denote by y^m the monomial

$$y^m = y_1^{m_1} \cdots y_n^{m_n}.$$

Then using the abbreviated notation for partial derivatives from the previous section, we can write

$$T_k f(y; x_0) = \sum_{|m| \leqslant k} \frac{1}{m!} \partial_x^m f(x_0) y^m$$

(by convention, the 0-th partial derivative is just the function f itself, and $(0,\ldots,0)! = 0! = 1$).

EXAMPLE 3.7.2. For $k = 1$, we recover the affine-linear map

$$T_1 f(y; x_0) = f(x_0) + \sum_{i=1}^{n} \partial_{x_i} f(x_0) y_i.$$

For $k = 2$, we obtain a polynomial of degree $\leqslant 2$ which is

$$T_1 f(y; x_0) = f(x_0) + \sum_{i=1}^{n} \partial_{x_i} f(x_0) y_i + \frac{1}{2}\sum_{i=1}^{n} \partial^2_{x_i^2} f(x_0) y_i^2 + \sum_{1\leqslant i<j\leqslant n} \partial^2_{x_i x_j} f(x_0) y_i y_j.$$

The term of order 2 corresponds to the partial derivatives of order 2, in other words to the tuples $(m_1,\ldots,m_n)$ with $m_1 + \cdots + m_n = 2$. Indeed, two cases can arise:

(1) either all except one m_i are zero, and $m_i = 2$, in which case we obtain the second derivative with respect to x_i taken twice, with coefficient $1/m! = 1/2! = 1/2$.

(2) or two of the m_i's are non-zero, equal to 1, and all others are zero; assume that $m_i = 1$ and $m_j = 1$ with $i < j$, then we get the partial derivative $\partial_{x_i x_j}$ with coefficient $1/m! = 1/(1!1!) = 1$.

Another way to express this second term (and to remember it) is to notice that

$$\frac{1}{2}\sum_{i=1}^{n} \partial^2_{x_i^2} f(x_0) y_i^2 + \sum_{1\leqslant i<j\leqslant n} \partial^2_{x_i x_j} f(x_0) y_i y_j = \frac{1}{2} y^t \operatorname{Hess}_f(x_0) y$$

where y^t is the transpose of the column vector y. Hence we can express the second Taylor polynomial concisely in the form

$$f(x_0) + \nabla f(x_0)\cdot y + \frac{1}{2} y^t \operatorname{Hess}_f(x_0) y$$

for $y \in \mathbf{R}^n$.

For instance, take $n = 2$, and suppose that

$$\operatorname{Hess}_f(x_0) = \begin{pmatrix} a & b \\ b & d \end{pmatrix}.$$

Then

$$\frac{1}{2} y^t \operatorname{Hess}_f(x_0) y = \frac{1}{2}(y_1\ y_2)\begin{pmatrix} a & b \\ b & d \end{pmatrix}\begin{pmatrix} y_1 \\ y_2 \end{pmatrix} = \frac{1}{2} a y_1^2 + b y_1 y_2 + \frac{1}{2} d y_2^2.$$

The following statement indicates one way that Taylor polynomials give a better and better approximation to a function of class C^k (there are more precise versions, but we will not need them).

PROPOSITION 3.7.3 (Taylor approximation). *Let $k \geqslant 1$ be an integer. Let $X \subset \mathbf{R}^n$ be open and $f\colon X \to \mathbf{R}$ be a function of class C^k. For x_0 in X, if we define $E_k f(x; x_0)$ by*

$$f(x) = T_k f(x - x_0; x_0) + E_k f(x; x_0)$$

then we have

$$\lim_{\substack{x\to x_0\\ x\neq x_0}} \frac{E_k f(x; x_0)}{\|x - x_0\|^k} = 0.$$

For $k = 2$, this means that for a function of class C^2, we have

$$\lim_{x\to x_0} \frac{1}{\|x - x_0\|^2}\Big(f(x) - \Big(f(x_0) + \nabla f(x_0)\cdot(x - x_0) + \frac{1}{2}(x - x_0)^t \operatorname{Hess}_f(x_0)(x - x_0)\Big)\Big) = 0.$$

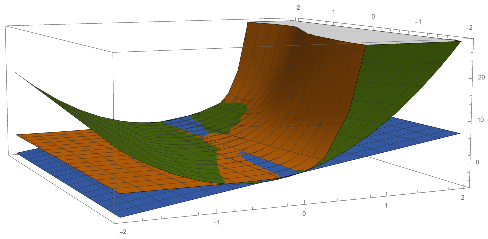

FIGURE 3.9. $f(x, y)$ and its approximations of order 1 and 2

EXAMPLE 3.7.4. Take $n = 2$ and $f(x, y) = e^{3x-\sin(xy)}$, around the point $(0, 0)$ where $f(0, 0) = e^0 = 1$.

The gradient is

$$\nabla f(x, y) = \begin{pmatrix} (3 - y\cos(xy))\exp(3x - \sin(xy)) \\ -x\cos(xy)\exp(3x - \sin(xy)) \end{pmatrix}$$

so that $\nabla f(0, 0) = \begin{pmatrix} 3 \\ 0 \end{pmatrix}$. The Hessian matrix is

$$\mathrm{Hess}_f(x, y) = e^{3x-\sin(xy)} \begin{pmatrix} a & b \\ b & d \end{pmatrix}$$

with

$$\begin{aligned} a &= y^2\sin(xy) + (3 - y\cos(xy))^2 \\ b &= -\cos(xy) + xy\sin(xy) - x\cos(xy)(3 - y\cos(xy)) \\ d &= x^2\sin(xy) + x^2\cos(xy)^2, \end{aligned}$$

so that

$$\mathrm{Hess}_f(0, 0) = \begin{pmatrix} 9 & -1 \\ -1 & 0 \end{pmatrix}$$

Hence the first order approximation at (x, y) close to $(0, 0)$ is

$$a(x, y) = 1 + 3x$$

and the second-order approximation at (x, y) is

$$g(x, y) = T_2 f(x, y; (0, 0)) = 1 + 3x + \frac{9x^2}{2} - xy.$$

As a numerical illustration, we find that

$$\begin{aligned} &f(-0.0015, 0.003) \approx 0.99551458963514434461139384694367021911 \\ &a(-0.0015, 0.003) = 0.9955 \\ &g(-0.0015, 0.003) = 0.995514625 \end{aligned}$$

so the precision has increased considerably (the difference goes from $\approx 1.46 \cdot 10^{-5}$ to $\approx 3.5 \cdot 10^{-8}$).

Figure 3.9 displays the graph of f, as well as that of a and g over $[-2, 2] \times [-2, 2]$.

3.8. Critical points

Recall that for a function of 1 variable, an important application of the derivative is its use for finding *extrema* of a function, using the necessary criterion that if a differentiable function f has a local maximum or minimum at a point x that is not a boundary of an interval, we have $f'(x) = 0$.

Proposition 3.8.1. *Let $X \subset \mathbf{R}^n$ be open and $f\colon X \to \mathbf{R}$ a differentiable function. If $x_0 \in X$ is such that*

$$f(y) \leqslant f(x_0) \text{ for all } y \text{ close enough to } x_0 \text{ (local maximum at } x_0)$$

or

$$f(y) \geqslant f(x_0) \text{ for all } y \text{ close enough to } x_0 \text{ (local minimum at } x_0).$$

Then we have $df(x_0) = 0$, or in other words $\nabla f(x_0) = 0$, or equivalently

$$\frac{\partial f}{\partial x_i}(x_0) = 0$$

for $1 \leqslant i \leqslant n$.

Proof. Let $1 \leqslant i \leqslant n$. Define

$$g(t) = f(x_0 + te_i)$$

for t such that $x_0 + te_i \in X$ (this contains an open interval around 0 since X is open). Then g has a local extremum at $t = 0$ by construction, and is differentiable, so $g'(t) = \partial_{x_i} f(x_0) = 0$. □

This proposition justifies the following definition:

Definition 3.8.2 (Critical point). Let $X \subset \mathbf{R}^n$ be open and $f\colon X \to \mathbf{R}$ a differentiable function. A point $x_0 \in X$ such that $\nabla f(x_0) = 0$ is called a *critical point* of the function f.

Proposition 3.8.1 is enough to determine the maximum and minimum of a function of more than one variable in many cases. One issue requires some care however: the existence of a point where a continuous function $f\colon X \to \mathbf{R}$ is maximal or minimal is not automatic if $X \subset \mathbf{R}^n$ is open.

Such points do exist, however, if f is defined on a set $\bar{X}$ that is compact (Definition 3.2.11), namely if the set $\bar{X}$ is bounded and closed. But the necessary condition of Proposition 3.8.1 does not apply in this case. The most common strategy is be in such a situation (a continuous function defined on a compact set $\bar{X}$), so that a maximum and a minimum are known to exist, and to have a decomposition

$$\bar{X} = X \cup B,$$

where X is open and B is a "boundary" part. Suppose then that the restriction of f to the open set X is differentiable. Then, *if* the maximum or minimum of f is reached in a point of X, this must be a critical point of the restriction of f to X. One can attempt to compute all these points, and evaluate f at these points to determine where the extremal points are. One must *in any case* also evaluate f on the boundary B in order to compare the values there, which might be larger (or smaller) than the values at the critical points in X.

Remark 3.8.3. This problem already occurs with one variable, where one must check the values $f(a)$ and $f(b)$ to find the maximum of a continuous function $f\colon [a, b] \to \mathbf{R}$, and not only the points $x \in]a, b[$ where $f'(x) = 0$.

Example 3.8.4. Let $\bar{X}$ be the square $[0,1] \times [0,1]$ in $\mathbf{R}^2$ and $f(x,y) = x^2 - 2y^2$. The set $\bar{X}$ is compact, and $\bar{X} = X \cup B$, where X is the open set $]0,1[\times]0,1[$, and B the boundary of the square, which is itself the union of four line segments.

The function f is differentiable on X, and its gradient is

$$\nabla f(x,y) = \begin{pmatrix} 2x \\ -4y \end{pmatrix}$$

so that the only critical point is $(0,0)$, where $f(0,0) = 0$. It is already clear that this is not the maximum, or the minimum, of f on $\bar{X}$.

On the boundary B, we compute

$$f(x,0) = x^2, \qquad f(x,1) = x^2 - 2, \qquad f(0,y) = -2y^2, \qquad f(1,y) = 1 - 2y^2.$$

The maximal values of f on these four segments are respectively

$$1, \qquad -1, \qquad -2, \qquad 1$$

and the minimal values are

$$0, \qquad -2, \qquad -2, \qquad -1.$$

We conclude that the maximum of f on $\bar{X}$ is equal to $1 = f(1,0)$, and that the minimum is $-2 = f(0,1)$.

In the case $n = 1$, the most convenient sufficient criterion for the existence of a local extremum at a point x where $f'(x) = 0$ is that the second derivative $f''(x)$ at this point should exist and be non-zero. Its sign then indicates whether x is a local maximum (if $f''(x) < 0$) or minimum ($f''(x) > 0$). The analogue question for $n \geqslant 2$ is more delicate. It is natural to think that the second partial derivatives (hence the Hessian matrix) should play the role of the second derivative, but the non-vanishing of $\mathrm{Hess}_f(x_0)$ is *not enough* to have a local extremum if $n \geqslant 2$, as the following important example shows.

Example 3.8.5. Let $n = 2$ and $f(x,y) = xy$. Then $\nabla f(x) = (y,x)$, so the only critical point is $(0,0)$, where $f(0,0) = 0$. We have

$$\mathrm{Hess}_f(0,0) = \begin{pmatrix} 0 & 1 \\ 1 & 0 \end{pmatrix}$$

which is non-zero, but nevertheless, the critical point $(0,0)$ is not a local maximum (since $f(x,x) = x^2 > f(0,0)$ for x arbitrarily small) and is not a local minimum (since $f(x,-x) = -x^2 < f(0,0)$ for x arbitrarily small).

This phenomenon reflects the fact that there is one line (namely, $y = x$) in which the restriction of the function has graph a downward parabola, and another (namely $y = -x$) in which it is an upward parabola. Such situations are called "saddle points".

Definition 3.8.6 (Non-degenerate critical point). Let $X \subset \mathbf{R}^n$ be open and $f\colon X \to \mathbf{R}$ a function of class C^2. A critical point $x_0 \in X$ of f is called *non-degenerate* if the Hessian matrix has non-zero determinant.

For a non-degenerate critical point x_0 of $f\colon X \to \mathbf{R}$, we can classify the behavior of the function f around x_0 in terms of the signs of the eigenvalues of the Hessian matrix. Recall, from linear algebra, that if a symmetric matrix H of size n is non-degenerate (has $\det(H) \neq 0$), then it is diagonalizable, with non-zero real eigenvalues, in an orthonormal basis of $\mathbf{R}^n$. Let p (resp. q) be the number of positive (resp. negative) eigenvalues of H. There exists an orthogonal basis $(v_1, \ldots, v_n)$ of $\mathbf{R}^n$ such that, for

$$y = t_1 v_1 + \cdots + t_n v_n \in \mathbf{R}^n,$$

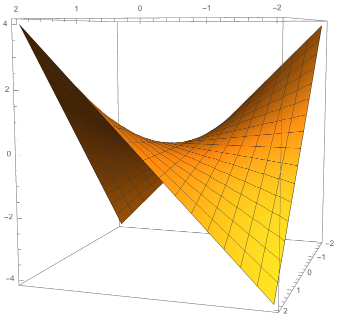

FIGURE 3.10. $f(x,y) = xy$

we have

$$y^t H y = t_1^2 + \cdots + t_p^2 - t_{p+1}^2 - \cdots - t_{p+q}^2,$$

where it is perfectly possible that $p = n$ (in which case, there are no terms with minus sign) or $q = n$ (in which case, there are no terms with plus sign).

The coefficients $t_1, \ldots, t_n$ are given by linear functions

$$t_i = \ell_i(y_1, \ldots, y_n)$$

if $y = (y_1, \ldots, y_n) \in \mathbf{R}^n$. Since $\nabla f(x_0) = 0$ (it is a critical point), the second Taylor polynomial of f at x_0 is then given by

$$f(x_0) + \frac{1}{2} y^t \operatorname{Hess}_f(x_0) y = f(x_0) + \frac{1}{2}\Big(\ell_1(y)^2 + \cdots + \ell_p(y)^2 - \ell_{p+1}(y)^2 - \cdots - \ell_{p+q}(y)^2\Big).$$

When x is very close to x_0, the function $f(x)$ is approximated very closely by

$$f(x_0) + \frac{1}{2}\Big(\ell_1(x-x_0)^2 + \cdots + \ell_p(x-x_0)^2 - \ell_{p+1}(x-x_0)^2 - \cdots - \ell_{p+q}(x-x_0)^2\Big),$$

and in particular the sign of $f(x) - f(x_0)$, which tells us whether x_0 is a local maximum, or minimum, or neither, is the same as the sign of

$$\ell_1(y)^2 + \cdots + \ell_p(y)^2 - \ell_{p+1}(y)^2 - \cdots - \ell_{p+q}(y)^2$$

This is very easy to determine, because when $x - x_0$ is in the direction of v_i, which means when only $\ell_i(x - x_0)$ is non-zero, and all other $\ell_j(x - x_0)$ are zero, we get the approximation

$$\ell_i(x-x_0)^2, \quad \text{if } 1 \leqslant i \leqslant p, \qquad -\ell_i(x-x_0)^2, \quad \text{if } p+1 \leqslant i \leqslant n.$$

If both of these cases occur for suitable choices of i, then there will be negative as well as positive values of $f(x) - f(x_0)$. So a local extremum is only possible if $p = n$ or $q = n$.

COROLLARY 3.8.7. *Let $X \subset \mathbf{R}^n$ be open and $f\colon X \to \mathbf{R}$ a function of class C^2. Let x_0 be a non-degenerate critical point of f. Let p and q be the number of positive and negative eigenvalues of* $\operatorname{Hess}_f(x_0)$.

(1) *If $p = n$, equivalently if $q = 0$, the function f has a local minimum at x_0.*

(2) *If $q = n$, equivalently if $p = 0$, the function f has a local maximum at x_0.*

(3) *Otherwise, equivalently if $pq \neq 0$, the function f does not have a local extremum at x_0.* One then says that f has a saddle point at x_0.

REMARK 3.8.8. (1) The condition $p = n$ means that the Hessian matrix H at x_0 is a positive definite symmetric matrix (and $q = n$ means that it is a negative definite matrix). This also means that $y^t H y > 0$ for any non-zero vector $y \in \mathbf{R}^n$. When $pq \neq 0$, the Hessian is also said to be *indefinite*.

(2) From linear algebra, we know an often convenient criterion for a symmetric matrix $A = (a_{i,j})_{1\leqslant i,j\leqslant n}$ to be positive definite: this is so if and only if the n submatrices

$$A_k = (a_{i,j})_{1\leqslant i,j\leqslant k},$$

for $1 \leqslant k \leqslant n$, have positive determinant. (For negative definite matrices, apply this to the opposite matrix; be careful that $\det(-A_k) \neq -\det(A_k)$ unless the size of the matrix is odd!) For instance, when $n = 2$, a matrix

$$\begin{pmatrix} a & b \\ b & d \end{pmatrix}$$

is positive definite if and only if

$$a > 0, \qquad ad - b^2 > 0.$$

It is negative definite if and only if

$$a < 0, \qquad ad - b^2 > 0,$$

and indefinite if and only if $ad - b^2 < 0$ (note that if $a = 0$, then the determinant is $-b^2 < 0$ since $a = b = 0$ is not possible for an invertible matrix).

For the Hessian matrix at a critical point x_0 of a C^2 function $f\colon \mathbf{R}^2 \to \mathbf{R}$, these conditions become

$$\frac{\partial^2 f}{\partial x^2}(x_0) > 0, \qquad \frac{\partial^2 f}{\partial x^2}(x_0)\frac{\partial^2 f}{\partial y^2}(x_0) - \Bigl(\frac{\partial^2 f}{\partial x \partial y}(x_0)\Bigr)^2 > 0$$

for a local minimum at x_0, or

$$\frac{\partial^2 f}{\partial x^2}(x_0) < 0, \qquad \frac{\partial^2 f}{\partial x^2}(x_0)\frac{\partial^2 f}{\partial y^2}(x_0) - \Bigl(\frac{\partial^2 f}{\partial x \partial y}(x_0)\Bigr)^2 > 0$$

for a local minimum at x_0, or

$$\frac{\partial^2 f}{\partial x^2}(x_0)\frac{\partial^2 f}{\partial y^2}(x_0) - \Bigl(\frac{\partial^2 f}{\partial x \partial y}(x_0)\Bigr)^2 < 0$$

for a saddle point.

If $n = 3$, the matrix

$$A = \begin{pmatrix} a & b & c \\ b & e & f \\ c & f & i \end{pmatrix}$$

is positive definite if and only if

$$a > 0, \qquad ae - b^2 > 0, \qquad \det(A) > 0.$$

(3) If pq is non-zero, the description with the Taylor polynomial is much more precise: it tells us that f behaves like a downward parabola in the directions corresponding to v_1, …, v_p, and like an upward parabola in the directions v_{p+1},…, v_n.

Example 3.8.9. (1) Consider again the function $f(x, y) = xy$ on $\mathbf{R}^2$ at the critical point $(0, 0)$, as in Example 3.8.5. Since it is a polynomial of degree 2, it is in fact equal to its second Taylor polynomial; the critical point is non-degenerate since $\det(\mathrm{Hess}_f(0, 0)) = -1$. An orthogonal basis of eigenvectors is (v_1, v_2) with $v_1 = (1, 1)$ (where $H(v_1) = v_1$) and $v_2 = (1, -1)$ (where $H(v_2) = -v_2$). The expression

$$f(x) = xy = \frac{1}{2}\Bigl(\frac{1}{2}(x+y)^2 - \frac{1}{2}(x-y)^2\Bigr)$$

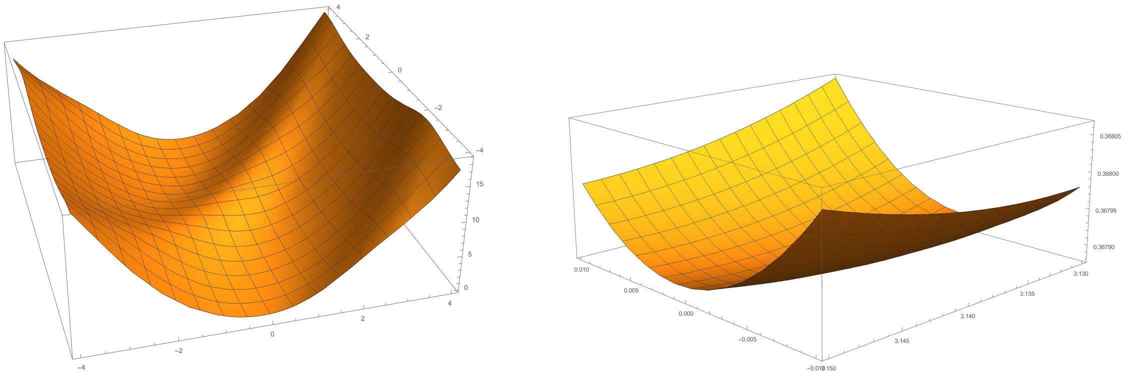

FIGURE 3.11. The graph of $e^{\cos(x-y)} + x^2$ and its behavior close to $(0, \pi)$

corresponds to our previous discussion, and we recover the directions $y = x$ and $y = -x$ where f has different behavior.

(2) Take $n = 2$ and

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

for $(x, y) \in X =]-4, 4[^2$. The gradient is

$$\nabla f(x, y) = \begin{pmatrix} -\sin(x-y)\exp(\cos(x-y)) + 2x \\ \sin(x-y)\exp(\cos(x-y)) \end{pmatrix}.$$

The critical points are determined by $\nabla f(x_0, y_0) = 0$. The second equation becomes $\sin(x_0 - y_0) = 0$, from which the first transforms to $x_0 = 0$, and hence $\sin(y_0) = 0$. We conclude that the critical points in the indicated region are $x_1 = (0, 0)$, $x_2 = (0, \pi)$ and $x_3 = (0, -\pi)$.

The Hessian is

$$\begin{aligned} \mathrm{Hess}_f(x, y) &= \begin{pmatrix} 2 & 0 \\ 0 & 0 \end{pmatrix} \\ &\quad + e^{\cos(x-y)} \begin{pmatrix} -\cos(x-y) + \sin^2(x-y) & \cos(x-y) - \sin^2(x-y) \\ \cos(x-y) - \sin^2(x-y) & -\cos(x-y) + \sin^2(x-y) \end{pmatrix}. \end{aligned}$$

The values H_1, H_2, H_3 of the Hessian of f at these three critical points are given respectively by

$$H_1 = \begin{pmatrix} 2-e & e \\ e & -e \end{pmatrix}, \qquad H_2 = H_3 = \begin{pmatrix} 2+e^{-1} & -e^{-1} \\ -e^{-1} & e^{-1} \end{pmatrix}.$$

The matrix H_1 is indefinite (the determinant being $-2e < 0$), but H_2 and H_3 are positive definite (since $2 + e^{-1} > 0$ and the determinant is $2/e > 0$). So $(0, \pi)$ and $(0, -\pi)$ are local minimum of f, while $(0, 0)$ is a saddle point.

It is interesting to note from the graphs that this is not so obvious!

REMARK 3.8.10. If x_0 is a *degenerate* critical point, the Hessian does not allow us to conclude anything concerning local extrema at x_0: there could be one (either a local maximum or local minimum) or not.

For instance, take $f_1(x, y) = x^4 + y^4$, $f_2(x, y) = x^4 - y^4$ and $f_3(x, y) = -x^4 - y^4$. The gradient of any of these functions vanishes if and only if $(x, y) = (0, 0)$, and $f_1(0, 0) = f_2(0, 0) = f_3(0, 0) = 0$. In all three cases, we also have $\mathrm{Hess}_{f_i}(0, 0) = 0$, so the information provided by the Hessian is the same. However, it is immediate that $(0, 0)$ is a local

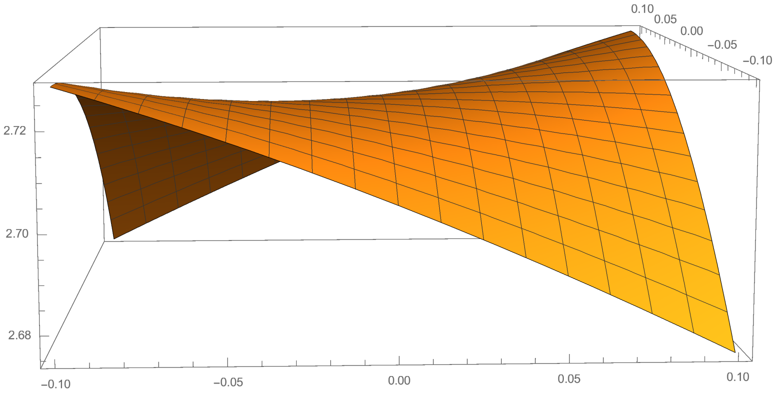

FIGURE 3.12. The behavior close to $(0,0)$

minimum of f_1 (even a global one), a local maximum of f_3, and that f_2 has a saddle point at $(0,0)$.

3.9. Lagrange multipliers

A common type of optimization problem does not simply asks for the maximum (or minimum) of a function, but adds *constraints* to the values of the variable. For instance, we might want to solve a problem like "what is the largest value of $f(x)$ if x is constrained to satisfy an equation $g(x) = 0$".

EXAMPLE 3.9.1. Let $(a,b,c) \in \mathbf{R}^3$ be non-zero, and let $(\alpha,\beta,\gamma) \in \mathbf{R}^3$, also non-zero. We want to find the maximum of the quadratic form

$$Q(x,y,z) = ax^2 + by^2 + cz^2$$

for (x,y,z) such that $\|(x,y,z)\| \leqslant 1$ and

$$\alpha x + \beta y + \gamma z = 0.$$

Geometrically, we intersect the sphere of radius 1 in $\mathbf{R}^3$ with a plane and we try to maximize $Q(x,y,z)$ on the intersection.

One idea to solve such a problem (which is often sufficient) is to parameterize the set of solutions of the constraint $g(x) = 0$ in terms of new variables (say u, so that $x = h(u)$ describes the set of solutions of $g(x) = 0$), and to maximize the function $f(h(u))$ for u in the set of parameters.

This method is often complicated because there is no simple parameterization of the solutions of $g(x) = 0$, or because the parameterization will destroy some natural symmetry of the problem, with the effect that the calculations become more complicated than they should.

The method of Lagrange multipliers can be used to solved this constrained maximization problems without involving a parameterization of the solution set.

Proposition 3.9.2. *Let $X \subset \mathbf{R}^n$ be open and let $f\colon X \to \mathbf{R}$ and $g\colon X \to \mathbf{R}$ be functions of class C^1. If $x_0 \in X$ is a local extremum of the function f restricted to the set*

$$Y = \{x \in X \ : \ g(x) = 0\}$$

then either $\nabla g(x_0) = 0$, or there exists $\lambda_0 \in \mathbf{R}$ such that

$$\begin{cases} \nabla f(x_0) = \lambda \nabla g(x_0) \\ g(x_0) = 0, \end{cases}$$

or in other words, there exists λ such that (x_0, λ) is a critical point of the differentiable function $h\colon X \times \mathbf{R} \to \mathbf{R}$ defined by

$$h(x, \lambda) = f(x) - \lambda g(x).$$

Such a value λ is called a Lagrange multiplier *at x_0.*

Intuitive explanation. Suppose there is a local extremum satisfying the constraint at x_0 and that $\nabla g(x_0) \neq 0$. If we "move" x around x_0, staying in the solution set of the equation $g(x) = 0$, which means moving perpendicularly to the gradient $\nabla g(x_0)$, the function f varies approximately by $(x - x_0) \cdot \nabla f(x_0)$. This will take values both positive and negative, unless all variations $x - x_0$ are orthogonal to $\nabla f(x_0)$. But all these possible variations represent the vectors orthogonal to $\nabla g(x_0)$, so the conclusion is that, for a local extremum, the gradients of f and g at x_0 are linearly dependent. And since $\nabla g(x_0) \neq 0$ by assumption, this means that there exists $\lambda \in \mathbf{R}$ such that

$$\nabla f(x_0) = \lambda \nabla g(x_0).$$

□

Compared to the problem of finding critical points of f (which has n equations $\partial x_i f(x_0) = 0$ and n unknowns), we have here $n + 1$ equations and $n + 1$ unknowns. Note that the values of the Lagrange multipliers λ is usually irrelevant to the final problem: they are just auxiliary quantities that are useful to find the local extrema.

As in the case of Proposition 3.8.1, it is important to remember that the solutions of the equations for Lagrange multipliers are only *candidates* for local extrema. As in the previous situation, we still need to check whether they are indeed extrema or not, and we may often need to handle a "boundary" component when f is defined on a set $\bar{X}$ that is compact, and is expressed as $X \cup B$ with X open, and B the boundary.

Remark 3.9.3. (1) Suppose that f is defined and continuous on a compact set $\bar{X} = X \cup B$. If the function g defining the constraint $g(x) = 0$ is also continuous, then the intersection

$$\bar{Y} = \bar{X} \cap \{x \in \bar{X} \ : \ g(x) = 0\}$$

is still a compact subset of $\mathbf{R}^n$ (indeed, it is bounded, as it is contained in $\bar{X}$, and it is the intersection of two closed sets – the second because g is continuous –, and it is elementary from Definition 3.2.11 that the intersection of two closed sets is closed). By restriction, f defines a continuous function $f|\bar{Y}\colon \bar{Y} \to \mathbf{R}$, and in particular Theorem 3.2.15 applies to $f|\bar{Y}$, which shows that f has a maximum and a minimum on $\bar{Y}$.

Suppose now that f is defined on $\mathbf{R}^n$, which is not compact. Then there is another important case in which the existence of a maximum and minimum for the constrained problem is ensured: this is so if the set Y defined by $g(x) = 0$ is itself compact, since then we are maximizing or minimizing the continuous function f on this compact set. And since g is continuous, the set Y is always closed, and therefore the question is whether it is bounded or not, which can often be determined very easily.

(2) Before deciding to use Lagrange multipliers, it is useful to check if some other method could apply, since the difficulty of the computations may vary a lot depending on the approach.

(3) The critical points of f on X are obvious candidates for local extrema of f restricted to Y, if they happen to be elements of Y. They occur in Proposition 3.9.2 precisely when the Lagrange multiplier λ is zero, since in that case the equation becomes $\nabla f(x_0) = 0$ (in addition to $g(x_0) = 0$).

Example 3.9.4. (1) Consider the problem of maximizing $f(x, y) = 2x^2 + 3xy - y^2$ on the circle of radius 1 in $\mathbf{R}^2$. The circle is compact, so we know that there exists a maximum. The circle is represented by the constraint $g(x, y) = 0$ with $g(x, y) = x^2+y^2-1$.

Since $\nabla g(x, y) = 0$ only if $(x, y) = 0$, for which $g(x, y) \neq 1$, only the case of a Lagrange multiplier can occur in Proposition 3.9.2. So we write down the equations

$$\begin{cases} x^2 + y^2 = 1 \\ 4x + 3y - 2x\lambda = 0 \\ 3x - 2y - 2y\lambda = 0. \end{cases}$$

The last two equations are linear with respect to x and y and have only the zero solution, which is incompatible with the first equation, unless the determinant is zero. This is

$$-(4 - 2\lambda)(2 + 2\lambda) - 9 = 4\lambda^2 - 4\lambda - 17.$$

The discriminant of this equation is $288 = 2^5 \cdot 3^2$, so the solutions are

$$\lambda_1 = \frac{4 + 12\sqrt{2}}{8} = \frac{1 + 3\sqrt{2}}{2}, \qquad \lambda_2 = \frac{4 - 12\sqrt{2}}{8} = \frac{1 - 3\sqrt{2}}{2}.$$

Writing $x = -3/(4 - 2\lambda)$, we obtain the possible values for y, namely

$$y = \pm\frac{4 - 2\lambda}{\sqrt{(4 - 2\lambda)^2 + 9}},$$

which gives, for the two values of λ, two values of y each, namely

$$\pm y_1 = \pm 0.382683432365089771\cdots, \qquad \pm y_2 = \pm 0.923879532511286756\cdots.$$

One can check that $y_1^2 + y_2^2 = 1$, so the corresponding values of x for a given y are ± the "other" value of y. Taking all possibilities of the sign into account, this shows that the maximum and minimum are taken at one of the values

$$\begin{matrix} (y_2, y_1), & (-y_2, y_1), & (y_2, -y_1), & (-y_2, -y_1) \\ (y_1, y_2), & (y_1, -y_2), & (-y_1, y_2), & (-y_1, -y_2). \end{matrix}$$

In fact, since $f(-x, -y) = f(x, y)$, we only need to check the first two values of each row, and for these we obtain

$$f(y_2, y_1) = \frac{1}{2}, \quad f(-y_2, y_1) = 2.621320343559642\cdots,$$
$$f(y_1, y_2) = -1.621320343559642\cdots, \quad f(y_1, -y_2) = \frac{1}{2}$$

(In fact, in this case, we have $f(-y_2, y_1) = \lambda_1$ and $f(y_1, y_2) = \lambda_2$, but this is a coincidence.)

In that case, it is however much simpler to represent the circle by the parameterization $(\cos\theta, \sin\theta)$, since this reduces the problem to maximizing or minimizing the function

$$f(\cos\theta, \sin\theta) = 2\cos^2\theta + 3\cos\theta\sin\theta - \sin^2\theta = 2 + 3\sin\theta(\cos\theta - \sin\theta).$$

Simply by differentiating, we find that the extreme values are achieved for $\theta = \pi/8$ (maximum) or $\theta = 5\pi/8$ (minimum).

(2) Consider the maximum and minimum of the function $f(x, y, z) = x^2 - y^2$ with the constraint $g(x, y, z) = 0$, where $g(x, y, z) = x^2 + 2y^2 + 3z^2 - 1$. Here the set of solutions of $g(x, y, z) = 0$ is closed, since g is continuous, and it is bounded since $x^2 \leqslant x^2 + 2y^2 + 3z^2 = 1$, and similarly $2y^2 \leqslant 1$ and $3z^2 \leqslant 1$ for any solution. Since f is also continuous, we know that there exist a maximum and a minimum.

The gradient of g is

$$\nabla g(x, y, z) = (2x, 4y, 6z)$$

and doesn't vanish when $g(x, y, z) = 0$. So we look for the Lagrange multipliers. The equations $\nabla f(x, y, z) = \lambda \nabla g(x, y, z)$ and $g(x, y, z) = 0$ are

$$\begin{cases} 2x = 2\lambda x \\ -2y = 4\lambda y \\ 0 = 6\lambda z \\ x^2 + 2y^2 + 3z^2 - 1 = 0. \end{cases}$$

The third equation shows that either $\lambda = 0$ or $z = 0$. In the first case, this implies that $x = y = 0$, and therefore $z = \pm 1/\sqrt{3}$, giving two possibilities $p_1 = (0, 0, 1/\sqrt{3})$ and $p_2 = (0, 0, -1/\sqrt{3})$. We have

$$f(p_1) = f(p_2) = 0.$$

If $z = 0$, then the equations for x and y become

$$\begin{cases} 2(1 - \lambda)x = 0 \\ -2(1 - 2\lambda)y = 0 \\ x^2 + 2y^2 = 1. \end{cases}$$

The first equation shows that either $x = 0$ or $\lambda = 1$. If $x = 0$, then we have the solutions with $\lambda = 1/2$, $y = \pm 1/\sqrt{2}$, in other words $p_3 = (0, 1/\sqrt{2}, 0)$, $p_4 = (0, -1/\sqrt{2}, 0)$. Then

$$f(p_3) = f(p_4) = -\frac{1}{2}.$$

Finally if $\lambda = 1$, then the second equation shows that $y = 0$, and the third gives the solutions $p_5 = (1, 0, 0)$ and $p_6 = (-1, 0, 0)$. Since

$$f(p_5) = f(p_6) = 1,$$

we conclude that the maximum of f with the constraint $g = 0$ is 1, and the minimum is $-1/2$.

(3) Here is an example with n arbitrarily large. Fix $(y_1, \ldots, y_n) \in \mathbf{R}^n$ non-zero. We want to maximize and minimize the function

$$f(x_1, \ldots, x_n) = x_1 y_1 + \cdots + x_n y_n$$

(which is in fact a linear function of x), subject to the constraint $x_1^2 + \cdots + x_n^2 = 1$. This constraint defines a compact set, so we know that the maximum exists.

Since $\nabla g(x) = (2x_1, \ldots, 2x_n)$ is non-zero for all x satisfying $g(x) = 0$, we solve the Lagrange multiplier equations. These are

$$\begin{cases} y_i = 2\lambda x_i & \text{for } 1 \leqslant i \leqslant n \\ x_1^2 + \cdots + x_n^2 = 1. \end{cases}$$

Since $y \neq 0$, we have $\lambda \neq 0$ from any of the first n equations. Then these equations state that $x_i = y_i/(2\lambda)$ for $1 \leqslant i \leqslant n$. It follows that

$$f(x) = \frac{1}{2\lambda}(y_1^2 + \cdots + y_n^2).$$

On the other hand, the last equation shows that

$$\frac{1}{4\lambda^2}(y_1^2 + \cdots + y_n^2) = 1,$$

and hence there are two solutions for λ, namely

$$\lambda = \pm\frac{1}{2\|y\|}.$$

We find the values $x = \pm y/(2\|y\|)$, and

$$f(x) = \pm\frac{1}{\sqrt{y_1^2 + \cdots + y_n^2}}(y_1^2 + \cdots + y_n^2) = \pm\sqrt{y_1^2 + \cdots + y_n^2} = \pm\|y\|.$$

Hence the constrained maximum of f is $\|y\|$ and the constrained minimum is $-\|y\|$.

If we now consider an arbitrary vector $x \neq 0$, and replace it with $\widetilde{x} = x/\|x\|$, which satisfies the constraint $g(\widetilde{x}) = 0$, the result implies by homogeneity that

$$-\|x\|\|y\| \leqslant x_1y_1 + \cdots + x_ny_n \leqslant \|x\|\|y\|.$$

This is the Cauchy-Schwarz inequality that we have recovered as a case of constrained optimization!

3.10. The inverse and implicit functions theorems

We finish this chapter by stating without proofs two important theoretical results that are often used in the study of functions of more than 1 variable, and of their level sets.

The first result is the analogue of the fact that a differentiable function $f\colon I \to \mathbf{R}$ defined on an interval is bijective from I to its image if its derivative is always > 0 (or always < 0). In other words, we want conditions that ensure that a function $f\colon X \to \mathbf{R}^n$ can be used as a change of variable, i.e., that we can recover x uniquely from the value $f(x)$.

Definition 3.10.1 (Change of variable). Let $X \subset \mathbf{R}^n$ be open and $f\colon X \to \mathbf{R}^n$ be differentiable. Let $x_0 \in X$. We say that f is a *change of variable* around x_0 if there is a radius $r > 0$ such that the restriction of f to the ball

$$B = \{x \in \mathbf{R}^n \ : \ \|x - x_0\| < r\}$$

of radius r around x_0 has the property that the image $Y = f(B)$ is open in $\mathbf{R}^n$, and if there is a differentiable map $g\colon Y \to B$ such that $f \circ g = \mathrm{Id}_Y$ and $g \circ f = \mathrm{Id}_B$.

Note that this definition is local: we do not require the existence of an inverse g for f defined everywhere. For a given $y \in Y$, there could well exist an element $x \in X$, not in B, such that $f(x) = y$.

Theorem 3.10.2 (Inverse function theorem). *Let $X \subset \mathbf{R}^n$ be open and $f\colon X \to \mathbf{R}^n$ be differentiable. If $x_0 \in X$ is such that $\det(J_f(x_0)) \neq 0$, i.e., such that the Jacobian matrix of f at x_0 is invertible, then f is a change of variable around x_0. Moreover, the Jacobian of g at x_0 is determined by*

$$J_g(f(x_0)) = J_f(x_0)^{-1}. \tag{3.4}$$

In addition, if f is of class C^k, then g is of class C^k.

In contrast to the case $n = 1$, there is no easy condition to ensure that f is a "global" change of variable: this must be investigated case by case.

It is easy to see that the requirement that $\det J_f(x_0) \neq 0$ is necessary for such a statement, and also to see that the formula (3.4) must be true if f is a change of variable. Indeed, if we assume that there exists g differentiable such that $g \circ f = \mathrm{Id}$, then by the chain rule, it follows that

$$J_g(f(x_0)) \cdot J_f(x_0) = J_{\mathrm{Id}}(x_0) = 1_n,$$

the identity matrix of size n (because the identity function is linear, so is its own differential). This formula implies that $J_f(x_0)$ is invertible with inverse $J_g(f(x_0))$.

Example 3.10.3. (1) Consider the function

$$f(x, y) = (\sin(xy), e^x + y).$$

Then

$$J_f(x, y) = \begin{pmatrix} y\cos(xy) & x\cos(xy) \\ e^x & 1 \end{pmatrix}$$

with determinant

$$\det J_f(x, y) = y\cos(xy) - xe^x\cos(xy) = \cos(xy)(y - xe^x).$$

This means that f is a change of variable around (x, y), unless either $xy = \pi/2 + k\pi$ for some $k \in \mathbf{Z}$, or $y = xe^x$.

(2) Consider the function

$$f(r, \theta, \varphi) = \begin{pmatrix} r\cos(\theta)\sin(\varphi) \\ r\sin(\theta)\sin(\varphi) \\ r\cos(\varphi) \end{pmatrix}$$

for $r \geqslant 0$, $0 \leqslant \theta \leqslant 2\pi$ and $0 \leqslant \varphi \leqslant \pi$ ("spherical coordinates").

The image of f is $\mathbf{R}^3$ and the function is differentiable and injective if the domain is the open set

$$X =]0, +\infty[\times]0, 2\pi[\times]0, \pi[.$$

In fact, r is the distance to the origin of (x, y, z), θ is the angle in the horizontal plane $z = 0$ from the x axis to the point (x, y), and φ is the angle between the vertical axis $x = y = 0$ and the line (x, y, z) (so it is between 0 and π).

The Jacobian of f is

$$J_f(r, \theta, \varphi) = \begin{pmatrix} \cos(\theta)\sin(\varphi) & -r\sin(\theta)\sin(\varphi) & r\cos(\theta)\cos(\varphi) \\ \sin(\theta)\sin(\varphi) & r\cos(\theta)\sin(\varphi) & r\sin(\theta)\cos(\varphi) \\ \cos(\varphi) & 0 & -r\sin(\varphi) \end{pmatrix} \tag{3.5}$$

with determinant

$$\begin{aligned} \det J_f(r, \theta, \varphi) = {} & -r^2\cos^2(\theta)\sin^3(\varphi) - r^2\sin^2(\theta)\sin^2(\varphi)\cos(\varphi) \\ & - r^2\cos^2(\theta)\cos^2(\varphi)\sin(\varphi) - r^2\sin^2(\theta)\sin^3(\varphi) = -r^2\sin(\varphi). \end{aligned} \tag{3.6}$$

This is non zero for all (r, θ, φ) in X, which confirms that the spherical coordinates give a change of variable around any point in X.

The last theorem of this chapter concerns the problem of transforming an equation $g(x, y) = 0$ into a functional relation $y = f(x)$ – in other words, of "parameterizing" the solutions of an equation.

We consider the case where y is a single value, whereas x runs over $\mathbf{R}^n$. As in the case of the Inverse Function Theorem, there is a general result that shows that such parameterizations exist, but a priori only for x close to a given x_0.

THEOREM 3.10.4 (Implicit Function Theorem). *Let $X \subset \mathbf{R}^{n+1}$ be open and let $g \colon X \to \mathbf{R}$ be of class C^k with $k \geqslant 1$. Let $(x_0, y_0) \in \mathbf{R}^n \times \mathbf{R}$ be such that $g(x_0, y_0) = 0$. Assume that*

$$\partial_y g(x_0, y_0) \neq 0.$$

Then there exists an open set $U \subset \mathbf{R}^n$ containing x_0, an open interval $I \subset \mathbf{R}$ containing y_0, and a function $f \colon U \to \mathbf{R}$ of class C^k such that the system of equations

$$\begin{cases} g(x, y) = 0 \\ x \in U, \quad y \in I \end{cases}$$

is equivalent *with $y = f(x)$. In particular, $f(x_0) = y_0$. Moreover, the gradient of f at x_0 is given by*

$$\nabla f(x_0) = -\frac{1}{(\partial_y g)(x_0, y_0)} \nabla_x g(x_0, y_0), \tag{3.7}$$

where $\nabla_x g = (\partial_{x_1} g, \ldots, \partial_{x_n} g)$.

IDEA OF THE PROOF. We will explain how to deduce at least the existence of the function from the Inverse Function Theorem. Consider the function

$$\varphi \colon X \to \mathbf{R}^{n+1}$$

defined by $\varphi(x, y) = (x, g(x, y))$. It is of class C^k. The Jacobian matrix is

$$J_\varphi(x, y) = \begin{pmatrix} 1_n & 0 \\ \nabla_x g & \partial_y g. \end{pmatrix}$$

Its determinant at (x_0, y_0) is

$$\det(J_\varphi(x_0, y_0)) = (\partial_y g)(x_0, y_0),$$

which is non-zero by assumption. This means that φ is a change of variable around (x_0, y_0), by the Inverse Function Theorem. Therefore, by Theorem 3.10.2, there exists an open set $V \subset \mathbf{R}^{n+1}$ containing $\varphi(x_0, y_0) = (x_0, 0)$ and a function $\psi \colon V \to U$ of class C^k such that

$$\varphi \circ \psi = \mathrm{Id}.$$

We use $(u, v) \in \mathbf{R}^n \times \mathbf{R}$ for the variables in V and write $\psi(u, v) = (\psi_1(u, v), \psi_2(u, v))$ where $\psi_1(u, v) \in \mathbf{R}^n$ and $\psi_2(u, v) \in \mathbf{R}$. Then the relation $\varphi \circ \psi = \mathrm{Id}$ means that

$$(u, v) = \varphi(\psi_1(u, v), \psi_2(u, v)) = (\psi_1(u, v), g(\psi_1(u, v), \psi_2(u, v))).$$

In particular, this means that $\psi_1(u, v) = u$, and taking $v = 0$, we get

$$0 = g(u, \psi_2(u, 0))$$

which shows that we can take $f(x) = \psi_2(x, 0)$ to solve the equation, namely that $g(x, f(x)) = 0$ for all x.

We can also quickly explain the formula (3.7): we start from the relation $g(x, f(x)) = 0$, which we write as $g \circ \widetilde{f} = 0$, where $\widetilde{f}(x) = (x, f(x))$. Since $\widetilde{f}(x_0) = (x_0, y_0)$, it follows by the chain rule that

$$0 = J_g(x_0, y_0) \cdot J_{\widetilde{f}}(x_0).$$

But we have

$$J_g(x, y) = ((\nabla_x g)^t, \partial_y g), \qquad J_{\widetilde{f}}(x) = \begin{pmatrix} 1_n \\ (\nabla f)^t \end{pmatrix}$$

(a matrix with $n+1$ rows and n columns), and writing down the coefficients of the matrix product, we obtain the relations

$$0 = \partial_{x_i} g(x_0, y_0) + (\partial_y g)(x_0, y_0) \partial_{x_i} f(x_0)$$

for $1 \leqslant i \leqslant n$. □

Example 3.10.5. (1) Let $g(x, y) = x^2 + y^2 - 1$ and (x_0, y_0) such that $g(x_0, y_0) = 0$. Then $(\partial_y g)(x_0, y_0) = 2y_0$. Therefore we can solve for y as a function of x, provided $y_0 \neq 0$. In fact the solution is simply

$$f(x) = \begin{cases} \sqrt{1 - x^2} & \text{if } y_0 > 0 \\ -\sqrt{1 - x^2} & \text{if } y_0 < 0. \end{cases} \tag{3.8}$$

Suppose that $y_0 > 0$ for instance. Then note that, for a given x, the point $(x, \sqrt{1 - x^2})$ is *not* the unique solution to $g(x, y) = 0$, since $(x, -\sqrt{1 - x^2})$ is also a solution. This explains the restriction to y belonging to some interval containing y_0 in the theorem, which is needed if we want to have an exact characterization of the solutions, and not just a sufficient condition that $g(x, f(x)) = 0$.

The formula (3.7) gives the derivative of f at x_0, namely

$$f'(x_0) = -\frac{1}{2y_0} \partial_x g(x_0) = -\frac{x_0}{y_0}.$$

If $y_0 > 0$, then this is equal to

$$f'(x_0) = -\frac{x_0}{\sqrt{1 - x_0^2}},$$

which is of course the same that one obtains from the formula (3.8).

If we consider $y_0 = 0$, then the picture of the circle shows indeed that the solution set is not the graph of a function of x when x is close to $x_0 = \pm 1$.

What can be done when $y_0 = 0$ (and this is a common occurence) is to use y as a variable to parameterize the solution, instead of x. Indeed, since $(\partial_x g)(\pm 1, 0) = \pm 2 \neq 0$, it follows from the Implicit Function Theorem applied to $\widetilde{g}(x, y) = g(y, x)$ that there is a parameterization as a function of y. In fact, it is simply $x = \sqrt{1 - y^2}$ or $x = -\sqrt{1 - y^2}$, depending on whether $x_0 > 0$ or $x_0 < 0$.

CHAPTER 4

Integration in $\mathbf{R}^n$

This chapter is devoted to integration in $\mathbf{R}^n$. There are in fact at least two different importants aspects: (1) integrating functions $f\colon X \to \mathbf{R}$, where $X \subset \mathbf{R}^n$; (2) relating integrals over different sets, of different dimensions.

In (1), besides defining integrals, one is led to analogues of the fundamental computational tools of the Riemann integral, such as the change of variable formula.

4.1. Line integrals

We begin with the simplest type of integrals in $\mathbf{R}^n$, namely integration of functions $I \to \mathbf{R}^n$, where I is an interval, and other integrals that involve a single variable, which is the integral of a function "along a curve".

We use again the scalar product in $\mathbf{R}^n$, which we denote

$$x \cdot y = \sum_{i=1}^{n} x_i y_i.$$

Definition 4.1.1. (1) Let $I = [a, b]$ be a closed and bounded interval in $\mathbf{R}$. Let

$$f(t) = (f_1(t), \ldots, f_n(t))$$

be a continuous function from I to $\mathbf{R}^n$, i.e., f_i is continuous for $1 \leqslant i \leqslant n$. Then we define

$$\int_a^b f(t)dt = \Big(\int_a^b f_1(t), \ldots, \int_a^b f_n(t)dt\Big) \in \mathbf{R}^n.$$

(2) A *parameterized curve* in $\mathbf{R}^n$ is a continuous map $\gamma\colon [a, b] \to \mathbf{R}^n$ that is piecewise C^1, i.e., there exists $k \geqslant 1$ and a partition

$$a = t_0 < t_1 < \cdots < t_{k-1} < t_k = b$$

such that the restriction of f to $]t_{j-1}, t_j[$ is C^1 for $1 \leqslant j \leqslant k$. We say that γ is a parameterized curve, or a *path*x, between $\gamma(a)$ and $\gamma(b)$.

(3) Let $\gamma\colon [a, b] \to \mathbf{R}^n$ be a parameterized curve. Let $X \subset \mathbf{R}^n$ be a subset containing the image of γ, and let $f\colon X \to \mathbf{R}^n$ be a continuous function. The integral

$$\int_a^b f(\gamma(t)) \cdot \gamma'(t)dt \in \mathbf{R}$$

is called the *line integral of f along γ*. It is denoted

$$\int_\gamma f(s) \cdot ds, \qquad \text{or} \qquad \int_\gamma f(s) \cdot d\vec{s}.$$

The integral of continuous functions $I \to \mathbf{R}^n$ satisfy much of the same rules as the Riemann integral of a function $I \to \mathbf{R}$, for instance

$$\int_a^b (f(t) + g(t))dt = \int_a^b f(t)dt + \int_a^b g(t)dt.$$

Also, as in the one-variable case, we define

$$\int_b^a f(t)dt = -\int_a^b f(t)dt,$$

if $a < b$.

In the line integral, $\gamma'(t)$ and $f(\gamma(t))$ are both vectors in $\mathbf{R}^n$ for all t (since γ takes values in $\mathbf{R}^n$), so that the final integral is a real number.

It is customary, when working with line integrals, to say that the function $f\colon X \to \mathbf{R}^n$ is a *vector field*: a function that sends each point x in $X \subset \mathbf{R}^n$ to a vector in $\mathbf{R}^n$, which we display as based at x.

Example 4.1.2. (1) Let $x = (x_1, \ldots, x_n)$ and $y = (y_1, \ldots, y_n)$ be elements of $\mathbf{R}^n$. The function

$$\gamma\colon [0,1] \to \mathbf{R}^n$$

defined by $\gamma(t) = (1-t)x + ty$ is a parameterized curve joining $\gamma(0) = x$ to $\gamma(1) = y$. Its image in $\mathbf{R}^n$ is exactly the line segment joining x to y. Note that $\gamma'(t) = y - x$ for all $t \in [0,1]$ (intuitively, this means that γ goes from x to y with constant speed).

Let f be a continuous function on X, expressed as $f(x) = (f_1(x), \ldots, f_n(x))$. Then we have

$$\int_\gamma f(s) \cdot d\vec{s} = \sum_{i=1}^n (y_i - x_i) \int_0^1 f_i((1-t)x + ty)dt.$$

In particular, suppose that $y_i = x_i$ for all i except a single value j (which means that the segment γ joins two points along one of the coordinate axes). Then we get

$$\int_\gamma f(s) \cdot d\vec{s} = (y_j - x_j) \int_0^1 f_j((1-t)x + ty)dt.$$

(2) Define $\gamma\colon [0, 2\pi] \to \mathbf{R}^2$ by $\gamma(t) = (\cos(t), \sin(t))$. This is a parameterized curve, whose image is the circle centered at $(0,0)$ with radius 1. For $f(x,y) = (f_1(x,y), f_2(x,y))$, we have

$$\int_\gamma f(s) \cdot d\vec{s} = \int_0^{2\pi} \Big(f_1(\cos(t), \sin(t))(-\sin(t)) + f_2(\cos(t), \sin(t)) \cos(t) \Big) dt.$$

Take for instance

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

Then we obtain

$$\int_\gamma f(s) \cdot d\vec{s} = \int_0^{2\pi} (\sin^2(t) + \cos^2(t))dt = 2\pi.$$

Take now $\gamma_1(t) = (\cos(t), \sin(t))$, but defined for $0 \leqslant t \leqslant 4\pi$. Then the parameterized curve γ_1 corresponds to “going twice over the circle”, so the image of γ_1 is the same as the image of γ. However, for the same vector field f as before, we have

$$\int_{\gamma_1} f(s) \cdot d\vec{s} = \int_0^{4\pi} dt = 4\pi.$$

(3) A parameterized curve $\gamma\colon [a,b] \to \mathbf{R}^n$ is not required to map different times t to different points: the trajectory described by γ may have points of self-intersection.

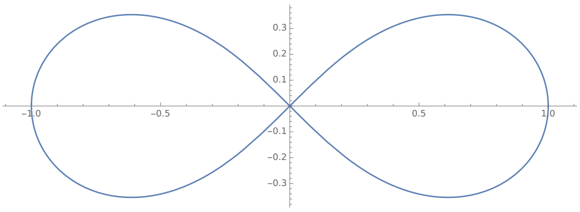

FIGURE 4.1. Lemniscate

An example is the circle taken twice over of the previous example, another one is the lemniscate of Bernoulli

$$\lambda(t) = \Big(\frac{\cos(t)}{1+\sin^2(t)}, \frac{\cos(t)\sin(t)}{1+\sin^2(t)}\Big) \tag{4.1}$$

for $0 \leqslant t \leqslant 2\pi$.

This is a closed curve, and we have also

$$\lambda(\pi/2) = \lambda(3\pi/2) = (0,0).$$

Note however that

$$\lambda'(t) = \frac{1}{(1+\sin^2(t))^2}\Big(-\sin(t) - \sin^3(t) - 2\sin(t)\cos^2(t), \cos^2(t) - 2\sin^2(t)\Big)$$

so that $(-1/2,-1/2) = \lambda'(\pi/2) \neq \lambda'(3\pi/2) = (1/2,1/2)$.

REMARK 4.1.3. (1) Let $f(x) = (f_1(x), \ldots, f_n(x))$. Another notation that is sometimes used, in relation with the notion of *differential form*, is

$$\int_\gamma f(s)\cdot d\vec{s} = \int_\gamma \omega$$

where one writes

$$\omega = f_1(x)dx_1 + \cdots + f_n(x)dx_n,$$

using linearly independent "formal symbols" dx_1, ..., dx_n to separate the components f_i of f.

(2) The line integral has a physical interpretation. Suppose we have a particle that moves from x_1 to x_2 along the path γ, where $\gamma(t)$ is the position and $\gamma'(t)$ is the speed of the particle at time t. Suppose further that a force f, represented by a vector giving its direction and intensity, is applied to the particle during the motion. Then the line integral

$$\int_\gamma f(s)\cdot d\vec{s}$$

is the "work" that is done by the force f along this trajectory. If there are no other forces, then the work is (in Newtonian mechanics) the difference in the kinetic energy of the particle between the starting and end points of the trajectory.

The most important property of the line integral is that it (essentially) only depends on the *image curve* $\gamma([a,b]) \subset \mathbf{R}^n$, and not on the chosen parameterization. More precisely:

DEFINITION 4.1.4. Let $\gamma\colon [a,b] \to \mathbf{R}^n$ be a parameterized curve. An *oriented reparameterization* of γ is a parameterized curve $\sigma\colon [c,d] \to \mathbf{R}^n$ such that $\sigma = \gamma\circ\varphi$, where $\varphi\colon [c,d] \to [a,b]$ is a continuous map, differentiable on $]a,b[$, that is strictly increasing and satisfies $\varphi(a) = c$ and $\varphi(b) = d$.

Note that the image $\sigma([c,d]) \subset \mathbf{R}^n$ of an oriented reparameterization σ of γ is the same as the image $\gamma([a,b])$. Also, γ is conversely an oriented reparameterization of σ, since $\gamma = \sigma \circ \varphi^{-1}$.

Proposition 4.1.5. *Let γ be a parameterized curve in $\mathbf{R}^n$ and σ an oriented reparameterization of γ. Let X be a set containing the image of γ, or equivalently the image of σ, and $f\colon X \to \mathbf{R}^n$ a continuous function. Then we have*

$$\int_\gamma f(s)\cdot d\vec{s} = \int_\sigma f(s)\cdot d\vec{s}.$$

Proof. This is a consequence of the change of variable formula for Riemann integrals (see [1, Satz 5.4.6]): since $\sigma = \gamma \circ \varphi$, we have $\sigma'(u) = \varphi'(u)\gamma'(\varphi(u))$ for $c \leqslant u \leqslant d$, and hence using the definition of line integrals, we get

$$\begin{aligned}\int_\sigma f(s)\cdot d\vec{s} &= \int_c^d f(\sigma(u))\cdot\sigma'(u)du\\ &= \int_c^d f(\gamma(\varphi(u)))\cdot\varphi'(u)\gamma'(\varphi(u))du\\ &= \int_c^d \Big(f(\gamma(\varphi(u)))\cdot\gamma'(\varphi(u))\Big)\varphi'(u)du\\ &= \int_a^b \Big(f(\gamma(t))\cdot\gamma'(t)\Big)dt = \int_\gamma f(s)d\vec{s},\end{aligned}$$

by applying the change of variable formula $t = \varphi(u)$, $dt = \varphi'(u)du$, since $c = \varphi(a)$ and $d = \varphi(b)$. □

Because of this proposition, one speaks, for instance, of the line integral of a vector field f along a circle, instead of fixing a parameterization of the circle. But one should keep in mind that this means "going over the circle only once, without repetition".

Example 4.1.6. (1) Let $n \geqslant 1$ and define

$$\gamma_n(t) = (\cos(2\pi t^n), \sin(2\pi t^n))$$

for $0 \leqslant t \leqslant 1$. Then $\gamma_n = \gamma_1 \circ \varphi_n$, where $\varphi_n(t) = t^n$. Hence all γ_n are common reparameterizations of γ_1. The curve described by γ_n is the circle of radius 1 centered at $(0,0)$. Note that

$$\gamma_n'(t) = (-2\pi n t^{n-1}\sin(2\pi t^n), 2\pi n t^{n-1}\cos(2\pi t^n)),$$

and in particular, if $n \geqslant 2$, we have $\gamma_n'(0) = 0$, which means that the trajectory described by γ_n starts from $(1,0)$ with very small speed, and then accelerates as t increases. Nevertheless, if $f(x,y) = (-y,x)$, we have always (for instance)

$$\int_{\gamma_n} f(s)\cdot d\vec{s} = \int_{\gamma_1} f(s)\cdot d\vec{s} = 2\pi.$$

(2) It is important that the reparameterizations that are used preserve the orientation, in other words that the endpoints are not "switched". For instance, suppose that $\gamma\colon [0,1] \to X$ is a parameterized curve. Let $\sigma(u) = \gamma(1-u)$; then σ is a parameterized curve, with the same image as γ, but it goes from $\sigma(0) = \gamma(1)$, the endpoint of γ, to $\sigma(1) = \gamma(0)$, the starting point of γ.

Let f be a continuous vector field on X. Then we compute

$$\int_\sigma f(s)\cdot d\vec{s} = \int_0^1 f(\gamma(1-u))\cdot(-\gamma'(1-u))du$$

and by substituting $t = 1-u$, this is

$$\int_1^0 f(\gamma(t))\cdot\gamma'(t)dt = -\int_\gamma f(s)\cdot d\vec{s}.$$

In other words: going along a parameterized curve "backwards" leads to the opposite value of the line integral.

The following example is extremely important, as it gives a very fast way to compute certain line integrals.

EXAMPLE 4.1.7. Let X be an open set in $\mathbf{R}^n$ and $g\colon X \to \mathbf{R}$ a function of class C^1. Define $f = \nabla g$, which is a vector field $X \to \mathbf{R}^n$. Let $\gamma\colon [a,b] \to X$ be a parameterized curve with image in X. We write $\gamma(t) = (\gamma_1(t), \ldots, \gamma_n(t))$.

We then have by definition

$$\int_\gamma f(s)\cdot d\vec{s} = \int_a^b \sum_{i=1}^n \frac{\partial g}{\partial x_i}(\gamma(t))\gamma_i'(t)dt.$$

But, by the Chain Rule, the function

$$\sum_{i=1}^n \frac{\partial g}{\partial x_i}(\gamma(t))\gamma_i'(t)$$

is the derivative of the C^1 function

$$h(t) = g(\gamma(t)).$$

Hence, by the fundamental theorem of calculus from Analysis I ([1, §5.4]), we have

$$\int_\gamma \nabla g(s)\cdot d\vec{s} = g(\gamma(b)) - g(\gamma(a)),$$

the difference between the value of g at the "end point" $\gamma(b)$ of the curve, and the value at the "start point" $\gamma(a)$.

What is striking in this example is that *the answer only depends on the extremities of the parameterized curve*! It is irrelevant how complicated the path joining $\gamma(a)$ to $\gamma(b)$ may be.

DEFINITION 4.1.8. Let $X \subset \mathbf{R}^n$ and $f\colon X \to \mathbf{R}^n$ a continuous vector field. If, for any x_1, x_2 in X, the line integral

$$\int_\gamma f(s)\cdot d\vec{s}$$

is independent of the choice of a parameterized curve γ in X from x_1 to x_2, then we say that the vector field is *conservative*.

REMARK 4.1.9. (1) Equivalently, f is conservative if and only if

$$\int_\gamma f(s)\cdot d\vec{s} = 0$$

for any *closed* parameterized curve in X (where a curve is said to be closed if $\gamma(a) = \gamma(b)$).

Indeed, if f is conservative, then the integral on a closed curve from x_1 to x_1 must be equal to the integral along the constant curve $\gamma(t) = x_1$, which is zero (the speed of γ being 0).

Conversely, suppose that this condition holds. Let γ_1, γ_2 be two paths in X from x_1 to x_2. Then the parameterized curve

$$\gamma(t) = \begin{cases} \gamma_1(2t) & \text{if } 0 \leqslant t \leqslant 1/2 \\ \gamma_2(2(1-t)) & \text{if } 1/2 \leqslant t \leqslant 1 \end{cases}$$

is a closed parameterized curve from x_1 to x_1, so that the integral of f along γ is zero by our assumption; but a simple computation shows that

$$0 = \int_\gamma f(s) \cdot d\vec{s} = \int_{\gamma_1} f(s) \cdot d\vec{s} - \int_{\gamma_2} f(s) \cdot d\vec{s}.$$

Hence the vector field is conservative.

(2) In physics, to say that a force is represented by a conservative vector field means that the work done by the force on a particle from one point to another is the same, whatever the trajectory between the two points.

(3) The equation $\nabla g = f$ is linear. It follows, for instance, that if f_1 and f_2 are both conservative, with respective potentials g_1 and g_2, then for any $(a, b) \in \mathbf{R}^2$, the vector field $af_1 + bf_2$ is conservative, with potential $ag_1 + bg_2$.

The previous example shows if X is open then any gradient vector field f on X is conservative, i.e., any vector field of the form $f = \nabla g$, where g is of class C^1 on X, is conservative. The converse is true:

Theorem 4.1.10. *Let X be an open set and f a conservative vector field. Then there exists a C^1 function g on X such that $f = \nabla g$.*

If any two points of X can be joined by a parameterized curve, then g is unique up to addition of a constant: if $\nabla g_1 = f$, then $g - g_1$ is constant on X.

Remark 4.1.11. (1) To say that any two points of X can be joined by a parameterized curve means that, for all x and y in X, there exists a parameterized curve $\gamma\colon [a, b] \to X$ such that $\gamma(a) = x$ and $\gamma(b) = y$. When this is true, we say that X is *path-connected*.

This is the case for instance when X is a disc in the plane, or a product of intervals. More generally, it is true whenever X is *convex*, which means that for any x and y in X, the line segment joining x to y is contained in X (this is because such a line segment is the image of a parameterized curve, as we saw in Example 4.1.2 (1)).

On the other hand, let X be the union of two discs that are disjoint, for instance, the discs of radius 1 around $(0, 0)$ and $(3, 0)$. Then X is *not* path-connected, since (by the intermediate value theorem) any curve $\gamma(t) = (\gamma_1(t), \gamma_2(t))$ joining $(0, 0)$ to $(3, 0)$ must be such that there exists t_0 with $\gamma_1(t_0) = 3/2$, which is impossible since the points of X have first coordinate in $[-1, 1] \cup [2, 4]$.

(2) If f is a conservative vector field on X, then a function g such that $\nabla g = f$ is called a *potential* for f. Note that it is not unique, since at least it is possible to add a constant to g without changing the gradient.

Idea of the proof. Write $f(x) = (f_1(x), \ldots, f_n(x))$ for $x \in X$. Assume that X is path-connected for simplicity. Then fix a point $x_0 \in X$. For any $x \in X$, select a parameterized curve γ_x from x_0 to x, and define

$$g(x) = \int_{\gamma_x} f(s) \cdot d\vec{s}.$$

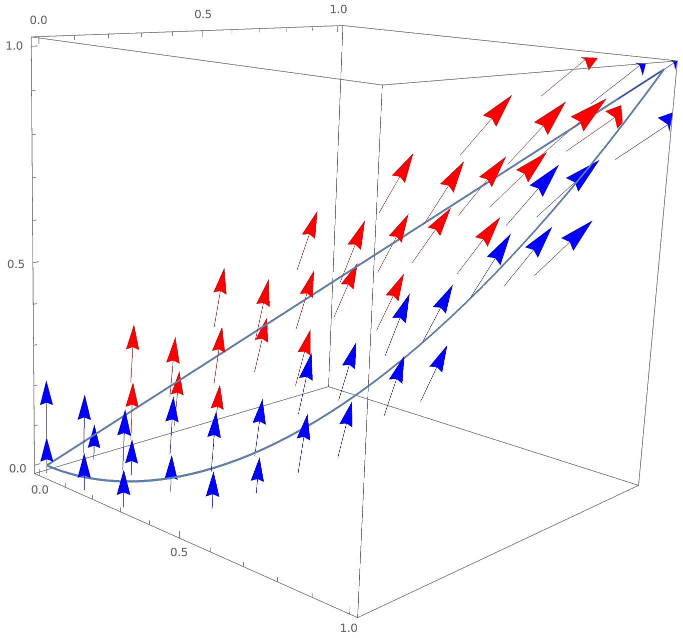

FIGURE 4.2. Vector field along two curves

This is a function of x, and $g(x)$ doesn't depend on the choice of γ_x, since the vector field f is conservative. In particular, to compute the partial derivative $\partial_{x_1} g$ of g at x, we can compute $g(x+te_1)$ for t small enough by selecting the curve γ_{x+te_1} to be the curve γ_x followed by the straight line segment $\ell_{x,t}$ from x to $x+te_1$ (which is contained in X, for t small enough, because X is open). Then we get easily

$$g(x+te_1)-g(x)=\int_{\ell_{x,t}} f(s)\cdot d\vec{s}=t\int_0^1 f_1((1-u)x+u(x+te_1))du$$

(since $\ell'_{x,t}(u)=te_1$ for all u; see again Example 4.1.2 (1))). By continuity of f, for t small enough, the function f_1 is almost constant on the segment $\ell_{x,t}$, equal to $f_1(x)$. So

$$g(x+te_1)-g(x)\approx tf_1(x)$$

which means that the partial derivative of g with respect to x_1 exists and is equal to $f_1(x)$.

Doing the same for all partial derivatives, we conclude that $\nabla g=f$. □

EXAMPLE 4.1.12. (1) Let $n=3$ and $f(x,y,z)=(y^2,xz,1)$. We will show that f is not conservative by computing the line integrals along two curves joining the same points, and showing that they are different.

We let $p_1=(0,0,0)$ and $p_2=(1,1,1)$. The parameterized curves

$$\gamma_1(t)=(t,t,t),\qquad \gamma_2(t)=(t,t^2,t^3)$$

for $0\leqslant t\leqslant 1$ both join p_1 to p_2. We have

$$\int_{\gamma_1} f(s)\cdot d\vec{s}=\int_0^1 (t^2,t^2,1)\cdot(1,1,1)dt=\int_0^1(2t^2+1)dt=\Big[\frac{2t^3}{3}+t\Big]_0^1=\frac{5}{3}.$$

On the other hand, we get

$$\int_{\gamma_2} f(s)\cdot d\vec{s} = \int_0^1 (t^4, t^4, 1)\cdot(1, 2t, 3t^2)dt = \int_0^1 (2t^5 + t^4 + 3t^2)dt =$$

$$\left[\frac{t^6}{3} + \frac{t^5}{5} + t^3\right]_0^1 = \frac{1}{3} + \frac{1}{5} + 1 = \frac{23}{15}.$$

Note that the second integral is smaller. In a physics interpretation, this would mean that less work (and energy) is needed to move the particle subject to the force f from p_1 to p_2 using the second path than the first path.

(2) Suppose that we know that a concrete vector field f is conservative. How does one find a potential g such that $\nabla g = f$?

One way to do that find g such that

$$\frac{\partial g}{\partial x_1} = f_1(x),$$

by integrating f_1 with respect to x_1, when other variables are fixed. This gives an answer up to a function g_1 that depends only on $(x_2, \ldots, x_n)$. We then solve for

$$\frac{\partial g}{\partial x_2} = f_2(x),$$

starting with the "partial" formula for g involving g_1, obtaining a new unknown function depending only on $(x_3, \ldots, x_n)$, and we repeat.

For instance, consider the vector field

$$f(x,y,z) = \begin{pmatrix} 9x^2\cos(yz) + z\sin(y) \\ -3x^3 z\sin(yz) + xz\cos(y) + 2y \\ -3x^3 y\sin(yz) + x\sin(y) + 2z \end{pmatrix}.$$

In order to have

$$\frac{\partial g}{\partial x} = 9x^2\cos(yz) + z\sin(y),$$

by the fundamental theorem of calculus, applied for each value of (y, z) separately, we must have

$$g(x,y,z) = 3x^3\cos(yz) + xz\sin(y) + h(y,z),$$

for some function $h\colon \mathbf{R}^2 \to \mathbf{R}$. Then, for g of this type to satisfy

$$\frac{\partial g}{\partial y} = -3x^3 z\sin(yz) + xz\cos(y) + 2y,$$

we must have

$$-3x^3 z\sin(yz) + xz\cos(y) + \frac{\partial h}{\partial y} = -3x^3 z\sin(yz) + xz\cos(y) + 2y,$$

which means that $\partial_y h = 2y$, or in other words that

$$h(y,z) = y^2 + k(z), \quad g(x,y,z) = 3x^3\cos(yz) + xz\sin(y) + y^2 + k(z),$$

for some function k. Finally, to have

$$\frac{\partial g}{\partial z} = -3x^3 y\sin(yz) + x\sin(y) + 2z,$$

we need to have

$$-3x^3 y\sin(yz) + x\sin(y) + k'(z) = -3x^3 y\sin(yz) + x\sin(y) + 2z,$$

which means that $k(z) = z^2 + c$ for some real number c. We can pick $c = 0$, which gives

$$g(x, y, z) = 3x^3 \cos(yz) + xz \sin(y) + y^2 + z^2.$$

The next general question is: how can one determine easily in practice if a concrete vector field f is conservative? The characterization in terms of line integrals is not really convenient, since many integrals are very hard to compute. There is however a very simple *necessary* condition.

Proposition 4.1.13. *Let $X \subset \mathbf{R}^n$ be an open set and $f\colon X \to \mathbf{R}^n$ a vector field of class C^1. Write*

$$f(x) = (f_1(x), \ldots, f_n(x)).$$

If f is conservative, then we have

$$\frac{\partial f_i}{\partial x_j} = \frac{\partial f_j}{\partial x_i}$$

for any integers with $1 \leqslant i \neq j \leqslant n$.

Proof. Indeed, let g be a potential for f, which is then of class C^2. Then $f_i = \partial_{x_i} g$ and hence

$$\frac{\partial f_i}{\partial x_j} = \frac{\partial^2 g}{\partial x_j \partial x_i} = \frac{\partial^2 g}{\partial x_i \partial x_j} = \frac{\partial f_j}{\partial x_i}$$

by Proposition 3.5.4. □

Example 4.1.14. (1) Consider again the example $f(x, y, z) = (y^2, xz, 1)$. Since

$$\partial_y(y^2) = 2x \neq z = \partial_x(xz),$$

we can see that f is not conservative without having to search for two curves where the line integrals are different.

(2) If $n = 2$, with $f = (f_1, f_2)$, then the condition is simply that

$$\frac{\partial f_2}{\partial x} = \frac{\partial f_1}{\partial y}.$$

It is natural to ask if this necessary criterion is also sufficient. This is not always true, and the answer depends on the set X where the vector field is defined.

Definition 4.1.15. A subset $X \subset \mathbf{R}^n$ is *star shaped* if there exists $x_0 \in X$ such that, for all $x \in X$, the line segment joining x_0 to x is contained in X. We then also say that X is *star-shaped around x_0*.

Example 4.1.16. (1) A ball

$$X = \{x \in \mathbf{R}^n \;:\; \|x - x_0\| < r\},$$

or a "rectangle"

$$X = [a_1, b_1] \times \cdots \times [a_n, b_n],$$

is a star-shaped subset. In fact, these are even *convex* sets, which means that for any x and y in X, the line segment between x and y is contained in X, or in other words, they are star-shaped around any point in X.

On the other hand, let X_1 and X_2 be two different discs in $\mathbf{R}^2$ that intersect in at least one point x_0. Then $X = X_1 \cup X_2$ is star-shaped (since, for any $x \in X$, the segment joining x_0 to x is either contained in X_1 or X_2, and hence is contained in X) but in general not convex.

(2) The complement $X = \{x \in \mathbf{R}^n \,:\, x \neq 0\}$ of the origin in $\mathbf{R}^n$ is not star-shaped: whatever value of $x_0 \neq 0$ we select in $\mathbf{R}^n$, the segment between x_0 and $-x_0$ is not contained in X, since it contains the origin $0 \notin X$.

(3) Let $0 < a < b$ be real numbers. The annulus

$$X = \{((x, y) \in \mathbf{R}^2 \,:\, a \leqslant x^2 + y^2 \leqslant b\} \subset \mathbf{R}^2$$

is not star-shaped, for the same reason as in (2): it does not contain $(0, 0)$, and the segment between (x, y) and $(-x, -y)$, which both belong to X if (x, y) does, passes through $(0, 0)$.

(4) If X is star-shaped, say around x_0, then it is path-connected: indeed, for any x and y in X, there is a parameterized curve from x to y obtained by following first the segment from x to x_0, and then the segment from x_0 to y, both of which are contained in X.

THEOREM 4.1.17. *Let X be a star-shaped open subset of $\mathbf{R}^n$. Let f be a C^1 vector field such that*

$$\frac{\partial f_i}{\partial x_j} = \frac{\partial f_j}{\partial x_i} \tag{4.2}$$

on X for all $i \neq j$ between 1 and n. Then the vector field f is conservative.

REMARK 4.1.18. The requirement that X is star-shaped is *not* necessary. Intuitively, the correct hypothesis on X should be that there is no “hole” in the middle around which a circle can go without it being possible to contract it within X.

EXAMPLE 4.1.19. (1) Let $X = \mathbf{R}^2 - \{0\}$. Define

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

on X. This is a C^1 vector field. We have

$$\partial_y\Big(\frac{y}{x^2 + y^2}\Big) = \frac{1}{x^2 + y^2} - \frac{2y^2}{(x^2 + y^2)^2} = \frac{x^2 - y^2}{(x^2 + y^2)^2}$$

and

$$\partial_x\Big(\frac{x}{x^2 + y^2}\Big) = \frac{1}{x^2 + y^2} - \frac{2x^2}{(x^2 + y^2)^2} = \frac{y^2 - x^2}{(x^2 + y^2)^2},$$

so that the condition (4.2) holds.

However, consider the closed parameterized curve $\gamma(t) = (\cos(t), \sin(t))$ for $0 \leqslant t \leqslant 2\pi$, which describes a circle of radius 1 around 0. Then we have

$$\int_\gamma f(s) \cdot d\vec{s} = \int_0^{2\pi} (\sin^2(t) + \cos^2(t))dt = 2\pi \neq 0,$$

which implies that the vector field f is *not* conservative.

For this particular choice of X, one can in fact prove that a C^1 vector field satisfying (4.2) is conservative if and only if

$$\int_\gamma f(s) \cdot d\vec{s} = 0,$$

for this particular curve γ.

(2) If we consider the same vector field as in (1), but defined on the open set $Y = \{(x, y) \in \mathbf{R}^2 \,:\, x > 0\}$ (a half-plane), then since this half-plane is convex, and therefore star-shaped, it follows that there exists a potential $g \colon Y \to \mathbf{R}$ such that $\nabla g = f$. In

fact, one can check that $g(x,y) = \arctan(y/x)$ has this property. Indeed, g is defined for $x > 0$, and since $\arctan'(x) = 1/(1+x^2)$, we get

$$\partial_x g = -\frac{y}{x^2}\frac{1}{1+(y/x)^2} = -\frac{y}{x^2+y^2}, \qquad \partial_y g = \frac{1}{x}\frac{1}{1+(y/x)^2} = \frac{x}{x^2+y^2}$$

on Y, which is the desired result.

(3) Let (a,b,c) be real parameters. For which values of (a,b,c) b is the vector field

$$f(x,y) = \begin{pmatrix} ax^3y + bxy^3 \\ bx^4 + cx^2y^2 \end{pmatrix}$$

conservative? Since f is defined on $\mathbf{R}^2$, we need to check if $\partial_y f_1 = \partial_x f_2$, which becomes the equation

$$ax^3 + 3bxy^2 = 4bx^3 + 2cxy^2.$$

Since x and y take arbitrary values, this is true if and only if

$$\begin{cases} a = 4b \\ 3b = 2c. \end{cases}$$

This means that we can use b as an arbitrary parameter and put

$$a = 4b, \qquad c = \frac{3b}{2}.$$

For instance, this is the case when $(a,b,c) = (4,1,3/2)$.

If $n = 3$, then there are three conditions (4.2). It is customary to view them as stating that an auxiliary vector field attached to f, its *curl*, is zero.

Definition 4.1.20. Let $X \subset \mathbf{R}^3$ be an open set and $f\colon X \to \mathbf{R}^3$ a C^1 vector field. Then the curl of f, denoted $\operatorname{curl}(f)$, is the continuous vector field on X defined by

$$\operatorname{curl}(f) = \begin{pmatrix} \partial_y f_3 - \partial_z f_2 \\ \partial_z f_1 - \partial_x f_3 \\ \partial_x f_2 - \partial_y f_1 \end{pmatrix},$$

where $f(x,y,z) = (f_1(x,y,z), f_2(x,y,z), f_3(x,y,z))$.

We see immediately that $\operatorname{curl}(f) = 0$ means precisely that the conditions (4.2) hold, for a 3-dimensional vector field.

Remark 4.1.21. To remember the definition, one can remember the (formal!) determinant

$$\operatorname{curl}(f) = \begin{vmatrix} e_1 & e_2 & e_3 \\ \partial_x & \partial_y & \partial_z \\ f_1 & f_2 & f_3 \end{vmatrix},$$

where (e_1, e_2, e_3) is the canonical basis of $\mathbf{R}^3$, expanding it "as usual", with the rule that $\partial_x \cdot f_i = f_i \cdot \partial_x = \partial_x f_i$, etc.

4.2. The Riemann integral in $\mathbf{R}^n$

We will now describe the Riemann integral in $\mathbf{R}^n$. The goal, for a closed bounded subset $X \subset \mathbf{R}^n$ and a continuous function $f\colon X \to \mathbf{R}$, is to define its *integral*

$$\int_X f(x_1, \ldots, x_n)dx$$

so that it has analogue properties to the Riemann integral for $n = 1$. The important difficulty, in comparison with the case $n = 1$, is that there possibilities for X have many more different shapes in higher dimension. Also, if X is a product of intervals

$$X = [a_1, b_1] \times \cdots \times [a_n, b_n] \subset \mathbf{R}^n,$$

(an n-dimensional "rectangle") then it is fairly natural to attempt to partition into smaller rectangles, by considering partitions of each interval $[a_i, b_i]$. However, if X is even as simple as a disc

$$X = \{(x, y) \in \mathbf{R}^2 \ : \ x^2 + y^2 \leqslant 1\} \subset \mathbf{R}^2,$$

then it cannot be decomposed in a finite union of rectangles, or even of smaller discs.

Because of this, the construction of the Riemann integral is much more involved. Since we will not be able to give the details of the proofs that this construction succeeds anyway, we will present its *properties* first, and we will only discuss in a remark what is a precise limiting process that can be used as a definition (see Remark 4.2.7).

For any bounded closed subset $X \subset \mathbf{R}^n$ and any continuous function $f\colon X \to \mathbf{R}$, one can define the *integral of* f *over* X, denoted

$$\int_X f(x)dx,$$

which is a real number, depending of course on X and on f.

The integral satisfies the following properties:

(1) **(Compatibility)** If $n = 1$ and $X = [a, b]$ is an interval (with $a \leqslant b$), then the integral of f over X is the Riemann integral of f:

$$\int_{[a,b]} f(x)dx = \int_a^b f(x)dx.$$

(2) **(Linearity)** If f and g are continuous on X and a, b are real numbers, then

$$\int_X (af_1(x) + bf_2(x))dx = a\int_X f_1(x)dx + b\int_X f_2(x)dx.$$

(3) **(Positivity)** If $f \leqslant g$, then

$$\int_X f(x)dx \leqslant \int_X g(x)dx$$

and especially, if $f \geqslant 0$, then

$$\int_X f(x)dx \geqslant 0.$$

Moreover, if $Y \subset X$ is compact and $f \geqslant 0$, then

$$\int_Y f(x)dx \leqslant \int_X f(x)dx.$$

(4) **(Upper bound and triangle inequality)** In particular, since $-|f| \leqslant f \leqslant |f|$, we have

$$\Bigl|\int_X f(x)dx\Bigr| \leqslant \int_X |f(x)|dx,$$

and since $|f + g| \leqslant |f| + |g|$, we have

$$\Bigl|\int_X (f(x) + g(x))dx\Bigr| \leqslant \int_X |f(x)|dx + \int_X |g(x)|dx.$$

(5) **(Volume)** If $f = 1$, then the integral of f is the "volume" in $\mathbf{R}^n$ of the set X, and if $f \geqslant 0$ in general, the integral of f is the volume of the set

$$\{(x, y) \in X \times \mathbf{R} \,:\, 0 \leqslant y \leqslant f(x)\} \subset \mathbf{R}^{n+1}.$$

In particular, if X is a bounded "rectangle", say

$$X = [a_1, b_1] \times \cdots \times [a_n, b_n] \subset \mathbf{R}^n$$

and $f = 1$, then

$$\int_X dx = (b_n - a_n) \cdots (b_1 - a_1). \tag{4.3}$$

We write $\mathrm{Vol}(X)$ or $\mathrm{Vol}_n(X)$ for the volume of X.

(6) **(Multiple integral, or Fubini's Theorem)** If n_1 and n_2 are integers $\geqslant 1$ such that $n = n_1 + n_2$, then for $x_1 \in \mathbf{R}^{n_1}$, let

$$Y_{x_1} = \{x_2 \in \mathbf{R}^{n_2} \,:\, (x_1, x_2) \in X\} \subset \mathbf{R}^{n_2}. \tag{4.4}$$

Let X_1 be the set of $x_1 \in \mathbf{R}^n$ such that Y_{x_1} is not empty. Then X_1 is compact in $\mathbf{R}^{n_1}$ and Y_{x_1} is compact in $\mathbf{R}^{n_2}$ for all $x_1 \in X_1$. If the function

$$g(x_1) = \int_{Y_{x_1}} f(x_1, x_2) dx_2$$

on X_1 is continuous, then

$$\int_X f(x_1, x_2) dx = \int_{X_1} g(x_1) dx_1 = \int_{X_1} \Bigl(\int_{Y_{x_1}} f(x_1, x_2) dx_2\Bigr) dx_1.$$

Similarly, exchanging the role of x_1 and x_2, we have

$$\int_X f(x_1, x_2) dx = \int_{X_2} \Bigl(\int_{Z_{x_2}} f(x_1, x_2) dx_1\Bigr) dx_2,$$

where $Z_{x_2} = \{x_1 \,:\, (x_1, x_2) \in X\}$, if the integral over x_1 is a continuous function.

Remark 4.2.1. (1) The conditions we have stated are not independent, and are not the only properties of the integral that we will state. However, they are enough to get some intuition, and are sufficient to compute many concrete integrals. Moreover, they *characterize* the integral: there is at most one way to define an "integral" for all X and all f in order that all properties above are satisfied.

(2) Property (5) is somewhat ambiguous, and could be replaced by the special case (4.3) (which is itself a special case of the formula (4.5) below); the fact that $\int_X dx$ is the volume of X would then be the *definition* of the volume $\mathrm{Vol}(X)$.

(3) If the variables are $x_1, \ldots, x_n$, we also write

$$\int_X f(x_1, \ldots, x_n) dx_1 \cdots dx_n.$$

(4) There are at least two intuitive explanations of Fubini's Theorem. First, if we think of integrals as generalizations of sums, then a two-variable integral is like a sum of real numbers $a_{i,j}$ with two indices; then Fubini's formula amounts to

$$\sum_{i,j} a_{i,j} = \sum_i \Bigl(\sum_j a_{i,j}\Bigr) = \sum_j \Bigl(\sum_i a_{i,j}\Bigr)$$

which are just different ways of combining the sum of these numbers, and are equal because of the commutativity and associativity of addition.

Next, we think of volumes only, and take $n = 2$ for simplicity. Consider a compact subset $X \subset \mathbf{R}^2$ of the form

$$X = \{(x, y) \,:\, a \leqslant x \leqslant b, \quad f_1(x) \leqslant y \leqslant f_2(x)\},$$

where $f_1 \leqslant f_2$ are two continuous functions defined on $[a, b]$. Then Fubini's formula for the volume of X becomes

$$\mathrm{Vol}(X) = \int_a^b g(x)dx,$$

where

$$g(x) = \int_{f_1(x)}^{f_2(x)} dy = f_2(x) - f_1(x).$$

The function $g(x)$ is the length of the vertical interval in X over the coordinate x (which can be thought of as a vertical slice of X), and so we say that the area of X is the integral of the length of vertical slices, which is intuitively reasonable.

Note that for more complicated sets, the slices might not be just a single interval, but the same intuitive explanation applies. And similarly, the area is the integral of the length of horizontal slices of X.

Example 4.2.2. (1) The simplest case of Fubini's Theorem is when X is a "generalized rectangle", namely

$$X = X_1 \times X_2,$$

where $X_1 \subset \mathbf{R}^{n_1}$ and $X_2 \subset \mathbf{R}^{n_2}$. Then X_1 is the same set that was denoted X_1 in Property (6). Moreover, for any $x_1 \in X_1$, we have

$$Y_{x_1} = \{x_2 \in \mathbf{R}^{n_2} \,:\, (x_1, x_2) \in X_1 \times X_2\} = X_2 \subset \mathbf{R}^{n_2},$$

which is therefore independent of x_1. If f is continuous on X, one can then *prove* that the function

$$g(x_1) = \int_{Y_{x_1}} f(x_1, x_2)dx_2 = \int_{X_2} f(x_1, x_2)dx_2$$

is *always* continuous in that case. Hence Fubini's Theorem takes the simple form

$$\int_{X_1 \times X_2} f(x_1, x_2)dx_1dx_2 = \int_{X_1} \Big(\int_{X_2} f(x_1, x_2)dx_2\Big)dx_1 = \int_{X_2} \Big(\int_{X_1} f(x_1, x_2)dx_1\Big)dx_2$$

for any continuous function f on X.

(2) Suppose now that

$$X = [a_1, b_1] \times \cdots \times [a_n, b_n] \subset \mathbf{R}^n$$

and that f is a function with separated variables given by

$$f(x_1, \ldots, x_n) = f_1(x_1) \cdots f_n(x_n),$$

where each function f_i is continuous (so f is also continuous). Then we claim that the integral takes the easy form

$$\int_X f(x_1, \ldots, x_n)dx_1 \cdots dx_n = \Big(\int_{a_1}^{b_1} f_1(x)dx\Big) \cdots \Big(\int_{a_n}^{b_n} f_n(x)dx\Big). \tag{4.5}$$

Indeed, consider the case $n = 2$ (the general case follows by induction): we have by Fubini's Theorem

$$\int_X f(x, y)dxdx = \int_{a_1}^{b_1} g(x)dx$$

provided the function g, defined by

$$g(x) = \int_{a_2}^{b_2} f_1(x) f_2(y) dy = \Big(\int_{a_2}^{b_2} f_2(y) dy\Big) f_1(x)$$

is continuous – which is the case since f_1 and f_2 are continuous. Since g is a multiple of f_1, we get

$$\int_X f(x,y) dx dx = \Big(\int_{a_2}^{b_2} f_2(y) dy\Big) \int_{a_1}^{b_1} f_1(x) dx,$$

which gives (4.5).

(3) We want to compute the volume V of the ball of radius one in $\mathbf{R}^3$, namely

$$X = \{(x,y,z) \in \mathbf{R}^3 \ : \ x^2 + y^2 + z^2 \leqslant 1\}.$$

First approach. We use slices according to the z variable: since the z variable runs over $[-1,1]$, according to Fubini's Theorem, we have

$$V = \int_{-1}^{1} g(z) dz,$$

where $g(z)$ is the area of the subset $X_z = \{(x,y,z) \in X\}$ where the last coordinate is z. This is a disc (in the horizontal plane where this value of z is fixed) of radius $\sqrt{1-z^2}$. So

$$V = \int_{-1}^{1} \pi(1-z^2) dz = \pi\Big(2 - \frac{2}{3}\Big) = \frac{4\pi}{3}.$$

Second approach. According to geometric intuition, the volume V is twice the volume of the subset X_+ where $z \geqslant 0$, which is then

$$X_+ = \{(x,y,z) \in \mathbf{R}^3 \ : \ 0 \leqslant x^2 + y^2 \leqslant 1, \ 0 \leqslant z \leqslant \sqrt{1 - x^2 - y^2}\}.$$

By Property (5), this means that

$$V = 2 \int_D \sqrt{1 - x^2 - y^2} dx dy$$

where

$$D = \{(x,y) \in \mathbf{R}^2 \ : \ x^2 + y^2 \leqslant 1\}$$

is the disc of radius 1 in $\mathbf{R}^2$. We use Fubini's Theorem to compute this two-dimensional integral. Here the set X_1 corresponding to the disc D is $[-1,1]$ (the set of possible first coordinates of a point in D). For a given $x \in [-1,1]$, the possible set Y_x of values of y is

$$Y_x = [-\sqrt{1-x^2}, \sqrt{1-x^2}].$$

So, according to Property (6) and Property (1), we have

$$\int_D \sqrt{1-x^2-y^2} dx dy = \int_{-1}^{1} g(x) dx$$

where

$$g(x) = \int_{-\sqrt{1-x^2}}^{\sqrt{1-x^2}} \sqrt{1-x^2-y^2} dy,$$

if g is continuous. But this function $g(x)$ is half of the area of a disc of radius $1 - x^2$, so we know that $g(x) = \frac{1}{2}\pi(1-x^2)$. In particular, it is indeed continuous, and as a consequence, we get

$$\int_D \sqrt{1-x^2-y^2} dx dy = \frac{\pi}{2} \int_{-1}^{1} (1-x^2) dx = \frac{\pi}{2}\Big(2 - \frac{2}{3}\Big) = \frac{2\pi}{3},$$

and finally $V = 4\pi/3$.

(4) In applying Fubini's Theorem, it can indeed happen that the function $g(x)$ is not continuous, although this creates no difficulty in practice, because of the possibility of decomposing the domain of integration, as we will discuss next.

For instance, let

$$X = \{(x,y) \in \mathbf{R}^2 \ : \ 0 \leqslant x \leqslant 2 \text{ and } 0 \leqslant y \leqslant 1 + \lfloor x \rfloor\}$$

(in other words, we have $0 \leqslant y \leqslant 1$ if $0 \leqslant x < 1$ and $0 \leqslant y \leqslant 2$ if $1 \leqslant x \leqslant 2$). If we take $f = 1$ and therefore use the two-dimensional integral to compute the area of X using Fubini's Theorem, we get $X_1 = [0,2]$ and

$$Y_x = \begin{cases} [0,1] & \text{if } 0 \leqslant x < 1 \\ [0,2] & \text{if } 1 \leqslant x \leqslant 2 \end{cases}$$

for which

$$g(x) = \int_{Y_x} dy = \begin{cases} 1 & \text{if } 0 \leqslant x < 1 \\ 2 & \text{if } 1 \leqslant x \leqslant 2. \end{cases}$$

This function is not continuous at $x = 1$.

A useful tool to compute integrals in dimension $\geqslant 2$ is to partition the domain of integration X. For this, we have the property that integrals "add up" over disjoint pieces, and more generally:

(7) **(Domain addivitity)** If X_1 and X_2 are compact subsets of $\mathbf{R}^n$, and f is continuous on $X_1 \cup X_2$, then

$$\int_{X_1 \cup X_2} f(x)dx + \int_{X_1 \cap X_2} f(x)dx = \int_{X_1} f(x)dx + \int_{X_2} f(x)dx. \tag{4.6}$$

Note that $X_1 \cap X_2$ is also compact, so all integrals exist.

In particular, if $X_1 \cap X_2$ is empty, then

$$\int_{X_1 \cup X_2} f(x)dx = \int_{X_1} f(x)dx + \int_{X_2} f(x)dx,$$

which is often very convenient. This simple formula holds also if the intersection $X_1 \cap X_2$ is "negligible". For instance, in $\mathbf{R}^2$, the intersection might be a parameterized curve, and for such a set, the integral is 0 (intuitively, because it is a one-dimensional object and the integral in $\mathbf{R}^2$ measures area).

We make the following definitions to deal with more general situations:

Definition 4.2.3. (1) Let $1 \leqslant m \leqslant n$ be an integer. A *parameterized m-set* in $\mathbf{R}^n$ is a continuous map

$$f\colon [a_1, b_1] \times \cdots \times [a_m, b_m] \to \mathbf{R}^n$$

which is C^1 on

$$]a_1, b_1[\times \cdots \times]a_m, b_m[.$$

(2) A subset $B \subset \mathbf{R}^n$ is *negligible* if there exist an integer $k \geqslant 0$ and parameterized m_i-sets $f_i\colon X_i \to \mathbf{R}^n$, with $1 \leqslant i \leqslant k$ and $m_i < n$, such that

$$X \subset f_1(X_1) \cup \cdots \cup f_k(X_k).$$

For instance, note that a parameterized 1-set in $\mathbf{R}^n$ is just a parameterized curve. Intuitively, we think of a parameterized m-set in $\mathbf{R}^n$ as a way to describe an m-dimensional subset of $\mathbf{R}^n$, but one should be aware that the image of a parameterized m-set f might be of dimension smaller than m (for instance, it is possible that f is constant, in which case the image is a single point, which is an object of dimension 0).

Example 4.2.4. (1) Any subset of the real axis $\mathbf{R} \times \{0\} \subset \mathbf{R}^2$ is negligible in $\mathbf{R}^2$.

(2) More generally, if $H \subset \mathbf{R}^n$ is an affine subspace of dimension $m < n$, then any subset of $\mathbf{R}^n$ that is contained in H is negligible.

(3) The image of a parameterized curve $\gamma\colon [a,b] \to \mathbf{R}^n$ is negligible, since γ is a 1-set in $\mathbf{R}^n$,

Proposition 4.2.5. *Let $X \subset \mathbf{R}^n$ be a compact set. Assume that X is negligible. Then for any continuous function on X, we have*

$$\int_X f(x)dx = 0.$$

We do not prove this, but illustrate this (fairly natural) property with examples.

Example 4.2.6. (1) Consider the graph $X = \{(t, \gamma(t)) \,:\, a \leqslant t \leqslant b\}$ of a continuous function $g\colon [a,b] \to \mathbf{R}$. This is the image of the parameterized curve $t \mapsto (t, \gamma(t))$, so it is negligible. Indeed, we can check the proposition in that case using Fubini's Theorem: for any function f continuous on X, we have

$$\int_X g(x,y)dxdy = \int_a^b \Bigl(\int_{f(x)}^{f(x)} g(x,y)dy\Bigr)dx = 0,$$

since an integral over a one-point interval is zero, and the integral of the zero function is zero by linearity.

(2) The formula (4.6) also explains why the volume of the unit ball $X \subset \mathbf{R}^3$ in Example 4.2.2 is twice the volume of the hemisphere X_+ with $z \geqslant 0$. Indeed, let $X_1 = X_+$ and $X_2 = X_-$, the lower hemisphere. Since $X = X_+ \cup X_-$, by Property (7), we have

$$V = \int_X dxdydz = \int_{X_+} dxdydz + \int_{X_-} dxdydz - \int_{X_+ \cap X_-} dxdydz.$$

The intersection $X_+ \cap X_-$ is $D \times \{0\}$, where $D \subset \mathbf{R}^2$ is the disc of radius 1. So it is negligible by Example 4.2.4, (2) (one can also see that this is the image in $\mathbf{R}^3$ of the parameterized 2-set given by

$$(r, \theta) \mapsto (r\cos(\theta), r\sin(\theta))$$

on $[0,1] \times [0, 2\pi]$). It follows by the proposition that

$$\int_D dxdydz = 0,$$

and hence

$$V = \int_X dxdydz = \int_{X_+} dxdydz + \int_{X_-} dxdydz.$$

To show that the volume of X_- is the same as that of X_+, one can use the same method as in Example 4.2.2 (later, we will see the change of variable formula that allows us to do this more directly).

Remark 4.2.7. We will explain here one possible definition of the Riemann integral in $\mathbf{R}^n$. It goes in the following steps:

(1) Definition of integrable functions on a closed bounded rectangle

$$X = [a_1, b_1] \times \cdots \times [a_n, b_n].$$

Namely, consider finite partitions of each interval

$$a_i = t_{i,0} < t_{i,1} < \cdots < t_{i,k} = b_i$$

which induce a partition of X into smaller rectangles

$$X_{j_1,\ldots,j_n} = [t_{1,j_1}, t_{1,j_1+1}] \times \cdots \times [t_{n,j_n}, t_{n,j_n+1}].$$

Each such rectangle has n-dimensional volume

$$m(j_1, \ldots, j_n) = (t_{1,j_1+1} - t_{1,j_1}) \cdots (t_{n,j_n+1} - t_{n,j_n}).$$

Each such partition defines an *upper Riemann sum* and a *lower Riemann sum*:

$$S^+ = \sum_{j_1=0}^{k-1} \cdots \sum_{j_n=0}^{k-1} \Big(\sup_{x \in X_{j_1,\ldots,j_n}} f(x) \Big) m(j_1, \ldots, j_n)$$

$$S^- = \sum_{j_1=0}^{k-1} \cdots \sum_{j_n=0}^{k-1} \Big(\inf_{x \in X_{j_1,\ldots,j_n}} f(x) \Big) m(j_1, \ldots, j_n)$$

We say that f is Riemann-integrable over X if

$$\sup S_- = \inf S_+,$$

where we consider supremum and infimum over all upper and lower Riemann sums computed for every possible partition. We then define

$$\int_X f(x)dx = \sup S_- = \inf S_+.$$

Such functions are not necessarily continuous, but all continuous functions on X are Riemann-integrable.

(2) Definition of Jordan-measurable subsets $X \subset \mathbf{R}^n$, which are necessarily bounded in $\mathbf{R}^n$: we say that a bounded set X, contained in a closed rectangle $B = [-R, R]^n$ of "radius" $R > 0$ around 0 is *Jordan-measurable* if the function defined on B by

$$\varphi(x) = \begin{cases} 1 & \text{if } x \in X \\ 0 & \text{if } x \notin X \end{cases}$$

is integrable in the sense of (1). One then checks that this definition is independent of the choice of the radius R.

(3) For a Jordan-measurable subset $X \subset \mathbf{R}^n$, and a function $f \colon X \to \mathbf{R}^n$, consider a closed bounded rectangle X' such that $X \subset X'$. Then we say that f is integrable over X if the function

$$\tilde{f}(x_1, \ldots, x_n) = \begin{cases} f(x_1, \ldots, x_n) & \text{if } (x_1, \ldots, x_n) \in X \\ 0 & \text{otherwise,} \end{cases}$$

is integrable over the rectangle X', in the sense of the definition in Step (1), and we define

$$\int_X f(x_1, \ldots, x_n)dx = \int_{X'} \tilde{f}(x_1, \ldots, x_n)dx.$$

Note that $\tilde{f}$ is, in general, not continuous, even if f is. One can show that if X is Jordan-measurable, then every continuous function f on X is integrable in this sense.

To be precise, this definition leads to some restrictions on the compact sets X that are allowed, but all "usual" compact sets (such as rectangles, balls, etc) are Jordan-measurable, so this is not an issue in applications. The more general definition that leads to the integral over arbitrary compact subsets that we have discussed is that of the *Lebesgue integral.*

With this restriction concerning X, the Riemann integral whose definition is sketched above satisfies, for continuous functions, all Properties described above.

4.3. Improper integrals

As in the one-dimensional case, one is often interested in extending the integral to unbounded domains, or to open bounded regions with functions that are not bounded. This is done by taking appropriate limits of integrals over compact subsets of the region of interest. We indicate just some basic definitions in $\mathbf{R}^2$.

Let $I \subset \mathbf{R}$ be a bounded interval and let $J = [a, +\infty[$ for some $a \in \mathbf{R}$. Let f be a continuous function on $X = J \times I$. We say that it is Riemann-integrable on X if the limit

$$\lim_{x \to +\infty} \int_{[a,x]\times I} f(x,y)dxdy = \lim_{x\to+\infty} \int_a^x \Bigl(\int_I f(x,y)dy\Bigr)dx = \lim_{x\to+\infty} \int_I \Bigl(\int_a^x f(x,y)dx\Bigr)dy$$

exists (the equality being cases of Fubini's Theorem). We then denote this limit by

$$\int_{J\times I} f(x,y)dxdy.$$

If $f \geqslant 0$, or more generally if $|f|$ is Riemann integrable on X, one can prove the Fubini formula

$$\int_{J\times I} f(x,y)dxdy = \int_a^{\infty} \Bigl(\int_I f(x,y)dy\Bigr)dx = \int_I \Bigl(\int_a^{+\infty} f(x,y)dx\Bigr)dy,$$

where each improper integral is a one-variable integral (this formula is however not always true without some assumption).

Similarly, let f be continuous on $\mathbf{R}^2$. Assume that $f \geqslant 0$. We say that f is Riemann-integrable on $\mathbf{R}^2$, if the limit

$$\lim_{R\to+\infty} \int_{[-R,R]^2} f(x,y)dxdy$$

exists, which is then called the *integral of f over* $\mathbf{R}^2$ and denoted

$$\int_{\mathbf{R}^2} f(x,y)dxdy.$$

One can then show that this integral is also the limit of

$$\int_{D_R} f(x,y)dxdy$$

where D_R is the disc of radius R centered at 0 (or any increasing sequence of compact subsets of $\mathbf{R}^2$ whose union is $\mathbf{R}^2$). There is also the Fubini formula

$$\int_{\mathbf{R}^2} f(x,y)dxdy = \int_{-\infty}^{\infty} \Bigl(\int_{-\infty}^{+\infty} f(x,y)dy\Bigr)dx = \int_{-\infty}^{+\infty} \Bigl(\int_{-\infty}^{+\infty} f(x,y)dx\Bigr)dy,$$

again with "ordinary" improper integrals in the last two formulas.

Remark 4.3.1. In all these cases, we also often say that "the integral converges" to indicate that a function is Riemann-integrable on an unbounded set.

The following comparison principle is the easiest way to prove that a certain improper integral exists: if $|f| \leqslant g$ (resp. $0 \leqslant f \leqslant g$), and we know that

$$\int_{J\times I} g(x,y)dxdy \text{ or } \int_{\mathbf{R}^2} g(x,y)dxdy$$

exists, then so does

$$\int_{J\times I} f(x,y)dxdy \text{ or } \int_{\mathbf{R}^2} f(x,y)dxdy,$$

respectively.

Example 4.3.2. (1) Consider the improper Riemann integral

$$\int_{[0,+\infty[\times[1,2]} xe^{-xy}dxdy.$$

We have for any $R > 0$

$$\int_0^R \Bigl(\int_1^2 xe^{-xy}dy\Bigr)dx = \int_0^R x\Bigl[-\frac{1}{x}e^{-xy}\Bigr]_1^2 dx = \int_0^R \Bigl(e^{-x} - e^{-2x}\Bigr)dx.$$

This can be evaluated and is equal to

$$(1-e^{-R}) - \frac{1}{2}(1-e^{-2R}) = \frac{1}{2} - e^{-R} + e^{-2R} \to \frac{1}{2}.$$

Hence the integral converges and is equal to $1/2$.

(2) In Example 4.4.3 (3) below, we will see that the improper integral

$$\int_{\mathbf{R}^2} e^{-(x^2+y^2)}dxdy$$

exists and is equal to π.

4.4. The change of variable formula

We now consider the analogue for the integral in $\mathbf{R}^n$ of the change of variable formula

$$\int f(g(x))g'(x)dx = \int f(y)dy$$

of one-variable calculus.

Let $\bar{X} \subset \mathbf{R}^n$ and $\bar{Y} \subset \mathbf{R}^n$ be compact subsets. Let $\varphi\colon \bar{X} \to \bar{Y}$ be a continuous map. We assume that we can write

$$\bar{X} = X \cup B, \qquad \bar{Y} = Y \cup C$$

where

(1) the sets X and Y are open;
(2) the sets B and C are negligible, in the sense of Definition 4.2.3;
(3) the restriction of φ to the open set X is a C^1 bijective map from X to Y.

In this situation, the Jacobian matrix $J_\varphi(x)$ is invertible at all $x \in X$; we assume that we can find a continuous function on $\bar{X}$ that restricts to $\det(J_\varphi(x))$ on X (this is usually obvious because we have a formula for the Jacobian, which makes sense and is clearly continuous on X). We abuse notation and still write $\det(J_\varphi(x))$ for this function, even if $x \in B$.

Remark 4.4.1. (1) Note that there is no assumption concerning the image of B.

(2) It is very frequent that φ is the restriction of a C^1 map $\mathbf{R}^n \to \mathbf{R}^n$, in which case the determinant of the Jacobian matrix is continuous everywhere, so that the last issue doesn't require any argument.

Theorem 4.4.2 (Change of variable formula). *In the situation described above, for any continuous function f on $\bar{Y}$, we have*

$$\int_{\bar{X}} f(\varphi(x))|\det(J_\varphi(x))|dx = \int_{\bar{Y}} f(y)dy.$$

If one wants to remember this formula, the mnemonic is that when $y = \varphi(x)$, we have $dy = |\det(J_\varphi(x))|dx$.

Example 4.4.3. (1) The simplest (but very important) case of the formula is when $\varphi(x) = x + x_0$ is a translation. Intuitively, this shouldn't change the volume, or the integral. Indeed, since φ is affine-linear, we have $J_\varphi(x) = 1_n$, the identity matrix, for all x. The change of variable formula becomes

$$\int_{\bar{X}} f(x + x_0)dx = \int_{x_0+\bar{X}} f(x)dx$$

for any compact subset $\bar{X}$ and any continuous function f on $x_0 + \bar{X}$. With $f = 1$, we see that the volume of $\bar{X}$ and that of $x_0 + \bar{X}$ are the same.

(2) The next most important special case is when φ is the restriction of a bijective linear map, namely $\varphi(x) = Ax$, where A is an invertible matrix of size n. Then $J_\varphi(x) = A$ for all $x \in \mathbf{R}^n$, with constant determinant $\det(A)$.

Let $\bar{X} = X \cup B$ be a compact set as above and $\bar{Y} = \varphi(\bar{X})$. Then $\varphi(\bar{X}) = \varphi(X) \cup \varphi(B)$. The change of variable formula becomes

$$\int_{\bar{X}} f(\varphi(x))dx = \frac{1}{|\det(A)|}\int_{\bar{Y}} f(y)dy$$

for any continuous function f on $\bar{Y}$.

Take especially f to be the function equal to 1 on $\bar{Y}$, so that the integral of f over $\bar{Y}$ is the n-dimensional volume of Y. Note that $f(\varphi(x))$ is the characteristic function of the set $\{x \in \mathbf{R}^n \;:\; Ax \in Y\}$, in other words of $A^{-1}Y$. We get

$$\mathrm{Vol}(Y) = |\det(A)|\,\mathrm{Vol}(A^{-1}Y),$$

which shows how the volume is transformed (dilated or contracted) under a linear map. If we replace $A^{-1}Y$ by X, which means that $Y = AX$, then we get equivalently

$$\mathrm{Vol}(AX) = |\det(A)|\,\mathrm{Vol}(X)$$

for any compact subset $X \subset \mathbf{R}^n$.

For instance, let $X = [0,1]^n$ be the unit cube in $\mathbf{R}^n$. Its volume is 1, and therefore

$$\mathrm{Vol}(A[0,1]^n) = |\det(A)|,$$

which provides the geometric interpretation of the determinant of real matrices.

It is actually possible to prove directly this last formula. For instance, observe that if A is diagonal, with diagonal entries $a_1, \ldots, a_n$, then

$$A[0,1]^n = [0,a_1] \times \cdots \times [0,a_n],$$

which has volume $|a_1|\cdots|a_n| = |\det(A)|$. One can also argue directly when A is an "elementary" matrix, for instance

$$A = \begin{pmatrix} 1 & 0 & 1 \\ 0 & 1 & 0 \\ 0 & 0 & 1 \end{pmatrix}$$

for $n = 3$. Since $A(x, y, z) = (x + z, y, z)$, one can check that $A[0, 1]^3$ is the set

$$Y = \{(x, y, z) \, : \, 0 \leqslant y \leqslant 1, \quad 0 \leqslant z \leqslant 1, \quad z \leqslant x \leqslant 1 + z\}.$$

One can compute the volume of Y by applying Fubini's Theorem (using the variable x in the inner integral). This gives

$$\mathrm{Vol}(Y) = \int_{[0,1]^2} \Big(\int_z^{1+z} dx \Big) dydz = \int_{[0,1]^2} dydz = 1 = \det(A).$$

One can also intuitively observe that

$$Y = Y_1 \cup Y_2,$$

where

$$Y_1 = \{(x, y, z) \, : \, 0 \leqslant y \leqslant 1, \quad 0 \leqslant z \leqslant 1, \quad z \leqslant x \leqslant 1\},$$
$$Y_2 = \{(x, y, z) \, : \, 0 \leqslant y \leqslant 1, \quad 0 \leqslant z \leqslant 1, \quad 1 \leqslant x \leqslant 1 + z\},$$

and if translate Y_2 by the vector $(-1, 0, 0)$, we obtain

$$Y_3 = Y_2 - (1, 0, 0)2 = \{(x, y, z) \, : \, 0 \leqslant y \leqslant 1, \quad 0 \leqslant z \leqslant 1, \quad 0 \leqslant x \leqslant z\},$$

and then $Y_3 \cup Y_1 = [0, 1]^3$. Since $Y_3 \cap Y_1$ is negligible, and the volume of Y_2 is equal to that of Y_3 (by (1), since Y_3 is a translate of Y_2), we get $1 = \mathrm{Vol}([0, 1]^3) = \mathrm{Vol}(Y_1) + \mathrm{Vol}(Y_2) = \mathrm{Vol}(Y)$, again.

(3) We consider the function

$$f(x, y) = e^{-(x^2+y^2)}$$

and we want to compute its integral over the compact disc

$$\bar{Y}_R = \{(x, y) \in \mathbf{R}^2 \, : \, x^2 + y^2 \leqslant R^2\}$$

where $R > 0$ is a parameter. Note that $\bar{Y}_R = Y_R \cup C_R$ with

$$Y_R = \{(x, y) \in \mathbf{R}^2 \, : \, 0 < x^2 + y^2 < R^2, \quad y \neq 0 \text{ if } x < 0\},$$

which is open, and C_R is the union of the segment $[-R, 0] \times \{0\}$ and of the circle of radius R, each of which is a parameterized curve, so that C_R is negligible.

Consider the polar coordinate change of variable

$$\varphi \colon \bar{X}_R \to \bar{Y}_R,$$

where

$$\bar{X}_R = [0, R] \times [-\pi, \pi]$$

and

$$\varphi(r, \theta) = (r\cos(\theta), r\sin(\theta))$$

(see Example 3.6.2). Note that φ is continuous on $\bar{X}_R$, and that the restriction of φ to a map from X_R to Y_R, where

$$X_R =]0, R[\times] - \pi, \pi[,$$

is bijective and of class C^1 (see Figure 4.3).

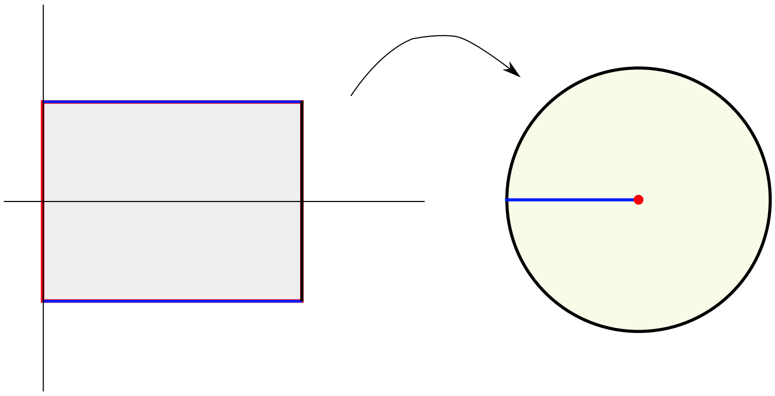

FIGURE 4.3. Polar coordinates and boundaries

The Jacobian matrix is

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

with determinant equal to

$$\det(J_\varphi(r,\theta)) = r.$$

We have $\bar{X}_R = X_R \cup B_R$ where

$$B_R = \{(r,\theta) \in X_R \;:\; r = 0 \text{ or } r = R \text{ or } |\theta| = \pi\}$$

is negligible (it is the union of four line segments). Note that the Jacobian matrix is a function that makes sense and is continuous on the whole of $\bar{X}_R$.

The change of variable formula is applicable, and it means that

$$\int_{\bar{X}_R} e^{-r^2} r dr d\theta = \int_{\bar{Y}_R} e^{-(x^2+y^2)} dx dy.$$

We can compute the integral in the left-hand side easily using Fubini's Theorem:

$$\int_{\bar{X}_R} e^{-r^2} r dr d\theta = \int_0^R re^{-r^2} \Bigl(\int_{-pi}^{\pi} d\theta\Bigr) dr = 2\pi \int_0^R e^{-r^2} r dr = 2\pi \Bigl[-\frac{1}{2} e^{-r^2}\Bigr]_0^R = \pi(1 - e^{-R^2}).$$

If we let $R \to +\infty$, we conclude that the improper Riemann integral of f over $\mathbf{R}^2$ converges and satisfies

$$\int_{\mathbf{R}^2} e^{-(x^2+y^2)} dx dy = \pi.$$

We can go further and derive an interesting consequence of this computation. Consider instead the integral of f over a square, namely

$$\int_{S_R} e^{-(x^2+y^2)} dx dy$$

where $S_R = [-R,R]^2$. Since f is a function with separated variables, we can reduce this integral to a one-variable integral by Fubini's Theorem (see (4.5)): we have

$$\int_{S_R} e^{-(x^2+y^2)} dx dy = \Bigl(\int_{-R}^R e^{-x^2} dx\Bigr)^2.$$

But now observe that $f \geqslant 0$ and that

$$Y_R \subset S_R \subset Y_{2R},$$

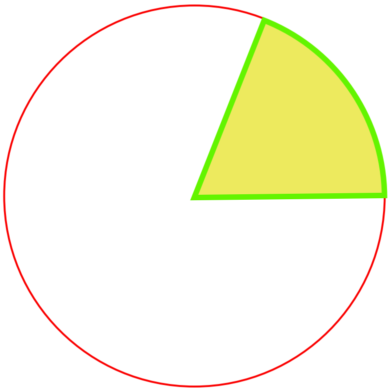

FIGURE 4.4. A sector

so that by positivity (Property (3) of the integral), we know that

$$\int_{Y_R} e^{-(x^2+y^2)}dxdy \leqslant \Big(\int_{-R}^{R} e^{-x^2}dx\Big)^2 \leqslant \int_{Y_{2R}} e^{-(x^2+y^2)}dxdy,$$

which means that

$$\pi(1-e^{-R^2}) \leqslant \Big(\int_{-R}^{R} e^{-x^2}dx\Big)^2 \leqslant \pi(1-e^{-4R^2}).$$

If we let $R \to +\infty$, both the first and the third quantities converge to π. We conclude that the improper Riemann integral of e^{-x^2} over $\mathbf{R}$ exists and satisfies

$$\int_{\mathbf{R}} e^{-x^2}dx = \sqrt{\pi}.$$

There are standard examples of change of variable (in the sense also of Section 3.6) that are often used to perform integrals over specific domains that have particularly nice parameterizations in the new variables.

(1) **Polar coordinates** (r, θ) are useful for integrating over a disc in $\mathbf{R}^2$ centered at 0, or more generally over a disc sector $\Delta = \Delta(a, b, R)$ defined by

$$0 \leqslant r \leqslant R, \qquad -\pi < a \leqslant \theta \leqslant b < \pi$$

for some parameters (a, b, R).

We computed the jacobian determinant in the previous example, and one gets the general formula

$$\int_{\Delta} f(x,y)dxdy = \int_0^R \int_a^b f(r\cos\theta, r\sin\theta)rdrd\theta. \tag{4.7}$$

Taking r to vary between $0 < r_0 \leqslant r \leqslant R$, we obtain an annulus.

(2) **Spherical coordinates** (r, θ, φ) in $\mathbf{R}^3$ (Example 3.10.3 (2)) are useful for integrating over balls centered at 0, or parts of them. We computed the jacobian and its determinant $-r^2\sin(\varphi)$ in (3.5) and (3.6). So, for integrating a function f over a ball B of radius R in $\mathbf{R}^3$, we have the formula

$$\int_B f(x,y,z)dxdydz = \int_0^R \int_0^{2\pi} \int_0^{\pi} f(r\cos\theta\sin\varphi, r\sin\theta\sin\varphi, r\cos\varphi)r^2\sin(\varphi)drd\theta d\varphi$$

(since it is easy to see that the boundary parts are neligible). Note that $\sin(\varphi) \geqslant 0$ for $0 \leqslant \varphi \leqslant \pi$, so that the absolute value of the jacobian determinant is indeed $r^2\sin(\varphi)$.

Example 4.4.4. (1) We compute the integral I of x^2y over the sector given by

$$\Delta = \{0 \leqslant r \leqslant 2, \quad \pi/6 \leqslant \theta \leqslant \pi/2\}.$$

In polar coordinates, this becomes

$$I = \int_0^2 \int_{\pi/6}^{\pi/2} r^4 \cos^2\theta \sin\theta dr d\theta = \Big[\frac{r^5}{5}\Big]_0^2 \int_{\pi/6}^{\pi/2} \cos^2(\theta)\sin(\theta)d\theta.$$

If we replace the trigonometric functions by their exponential versions, the function $\cos^2(\theta)\sin(\theta)$ becomes

$$\begin{aligned} \cos^2(\theta)\sin(\theta) &= \frac{1}{8i}(e^{i\theta}+e^{-i\theta})^2(e^{i\theta}-e^{-i\theta}) \\ &= \frac{1}{8i}(e^{2i\theta}+2+e^{-2i\theta})(e^{i\theta}-e^{-i\theta}) \\ &= \frac{1}{8i}(e^{3i\theta}-e^{-3i\theta}+e^{i\theta}-e^{-i\theta}) = \frac{1}{4}(\sin(3\theta)+\sin(\theta)). \end{aligned} \tag{4.8}$$

Therefore

$$I = \frac{32}{5} \times \frac{1}{4}\int_{\pi/6}^{\pi/2}(\sin(3\theta)+\sin(\theta))d\theta = \frac{8}{5}\Big[-\frac{1}{3}\cos(3\theta)-\cos(\theta)\Big]_{\pi/6}^{\pi/2} = \frac{8\cos(\pi/6)}{5} = \frac{4\sqrt{3}}{5}.$$

(2) We compute the integral I of z^2 over the spherical shell in $\mathbf{R}^3$ given by $1 \leqslant r \leqslant 2$ in spherical coordinates. Since

$$z = r\cos(\varphi),$$

we get

$$I = \int_1^2 \int_0^{2\pi} \int_0^{\pi} r^4 \cos^2(\varphi)\sin(\varphi) dr d\theta d\varphi$$

We use the formula (4.8) to write this finally as

$$= 2\pi \times \Big[\frac{r^5}{5}\Big]_1^2 \times \frac{1}{4}\int_0^{\pi}(\sin(3\varphi)+\sin(\varphi))d\varphi = 2\pi \times \Big(\frac{32}{5}-\frac{1}{5}\Big) \times \frac{2}{3} = \frac{124\pi}{15}.$$

4.5. Geometric applications of integrals

Besides the fact that the integral can be used to define and compute volumes of subsets of $\mathbf{R}^n$, there are quite a few other natural geometric quantities that can be expressed in terms of integrals. We present some of them in this section.

(1) **[Center of mass]** Let X be a compact subset of $\mathbf{R}^n$, such that the volume of X is positive. The *center of mass* (or *barycenter*) of X is the point $\bar{x} \in \mathbf{R}^n$ such that $\bar{x} = (\bar{x}_1, \ldots, \bar{x}_n)$ with

$$\bar{x}_i = \frac{1}{\mathrm{Vol}(X)} \int_X x_i dx.$$

Intuitively, $\bar{x}_i$ is the average over X of the i-th coordinate, and $\bar{x}$ is the point where X is "perfectly balanced".

Note that $\bar{x}$ is not necessarily in X (for instance, for an annulus

$$X = \{(x,y) \in \mathbf{R}^2 \,:\, 1 \leqslant x^2+y^2 \leqslant 2\}$$

in $\mathbf{R}^2$, the center of mass is $(0,0)$), but this is the case if X is convex.

(2) [**Surface area**] Consider a continuous function

$$f\colon [a,b]\times[c,d]\to\mathbf{R}$$

which is C^1 on $]a,b[\times]c,d[$. Let

$$\Gamma = \{(x,y,z)\in\mathbf{R}^3 \,:\, (x,y)\in[a,b]\times[c,d],\quad z=f(x,y)\}\subset\mathbf{R}^3$$

be the graph of f. Intuitively, this is a surface, and it should have an area. This is in fact given by

$$\int_a^b\int_c^d\sqrt{1+(\partial_x f(x,y))^2+(\partial_y f(x,y))^2}dxdy.$$

Such a result also holds for the graphs of functions defined on other sets, such as discs, provided they are C^1 in the "interior" of the domain.

There is an analogue formula for the length of the graph of a function $f\colon[a,b]\to\mathbf{R}$, namely it is equal to

$$\int_a^b\sqrt{1+f'(x)^2}dx.$$

Example 4.5.1. (1) What is the center of mass of a cone

$$X=\{(x,y,z)\in\mathbf{R}^3 \,:\, 0\leqslant z\leqslant 1,\quad x^2+y^2\leqslant(1-z)^2\}$$

in $\mathbf{R}^3$? (This is a cone because for a given z, the "slice" of X where z is fixed is a disc centered at 0 with radius $1-z$). For symmetry reasons, we have $\bar{x}=\bar{y}=0$ (you should check that), so the question is to compute $\bar{z}$. First we compute the volume, using Fubini's Theorem

$$\mathrm{Vol}(X)=\int_X dxdydz=\int_0^1\Big(\pi(1-z)^2\Big)dz=\frac{\pi}{3}.$$

Next we compute

$$\int_X zdxdydz=\int_0^1 z\Big(\pi(1-z)^2\Big)dz=\frac{\pi}{12},$$

so that the center of mass is $(0,0,1/4)$.

(2) What is the surface S of the sphere

$$X=\{(x,y,z) \,:\, x^2+y^2+z^2=1\}$$

of radius 1 in $\mathbf{R}^3$? Geometrically, this is twice the area of the graph of the function

$$f(x,y)=\sqrt{1-x^2-y^2}$$

defined for (x,y) such that $x^2+y^2\leqslant 1$. Although this is not defined over a rectangle, an analogue of the formula above holds, and we have

$$S=2\int_D\sqrt{1+(\partial_x f)^2+(\partial_y f)^2}dxdy$$

where D is the disc of radius 1 centered at $(0,0)$ in $\mathbf{R}^2$. We have

$$\partial_x f=-\frac{x}{\sqrt{1-x^2-y^2}},\qquad \partial_y f=-\frac{y}{\sqrt{1-x^2-y^2}},$$

hence the surface is

$$S=2\int_D\Big(1+\frac{x^2}{1-x^2-y^2}+\frac{y^2}{1-x^2-y^2}\Big)^{1/2}dxdy=2\int_D\frac{1}{\sqrt{1-x^2-y^2}}dxdy.$$

Using polar coordinates (4.7), this becomes

$$S = 4\pi \int_0^1 \frac{r}{\sqrt{1-r^2}} dr = 4\pi \Big[-\sqrt{1-r^2} \Big]_0^1 = 4\pi.$$

4.6. The Green formula

In the last sections, we will discuss two important formulas which are of the form

$$\int_{\partial X} f = \int_X Df,$$

where

(1) f is a C^1 vector field defined on $\mathbf{R}^n$;
(2) $X \subset \mathbf{R}^n$ is a compact m-dimensional subset, with $1 \leqslant m < n$;
(3) ∂X is the "boundary" of X, which has dimension $m-1$, and the integral on ∂X is a generalization of a line integral;
(4) Df is some expression computed using the partial derivatives of first order of f.

In fact, there exist versions of these results in all dimensions, but we focus here on the cases $n = m = 2$ (Green's formula) and, in the next section, on the case $n = m = 3$ (Gauss–Ostrogradski formula).[1]

In all cases, the prototype is the Fundamental Theorem of Calculus, in the form

$$\int_a^b f'(x)dx = f(b) - f(a), \tag{4.9}$$

where $X = [a, b]$ and the boundary is simply the set $\{a, b\}$ with two elements.

The Green formula concerns the case of relating an integral over a subset X of $\mathbf{R}^2$ with a line integral over its boundary. The typical example is an integral over a compact disc of radius $r > 0$ centered at x_0, which is related to a line integral over the circle of radius r centered at x_0.

The difficulty in a rigorous formulation of this formula is mostly in precisely understanding which subsets X are allowed, and what "boundary" means. Moreover there is an issue of *orientation* of the boundary (reflected in (4.9) in the fact that the sign of $f(b)$ and $f(a)$ is not the same on the right-hand side).

Definition 4.6.1. A simple closed parameterized curve $\gamma\colon [a, b] \to \mathbf{R}^2$ is a closed parameterized curve such that $\gamma(t) \neq \gamma(s)$ unless $t = s$ or $\{s, t\} = \{a, b\}$, and such that $\gamma'(t) \neq 0$ for $a < t < b$. (If γ is only piecewise C^1 inside $]a, b[$, this condition only applies where $\gamma'(t)$ exists).

Example 4.6.2. (1) A circle parameterized by

$$\gamma(t) = (x_0 + r\cos(t), y_0 + r\sin(t))$$

for $0 \leqslant t \leqslant 2\pi$ is a simple closed parameterized curve. But if we consider the circle twice over (i.e., for $0 \leqslant t \leqslant 4\pi$), then it is not.

(2) The lemniscate λ (Figure 4.1) defined by (4.1) is not a simple closed curve, since $\lambda(\pi/2) = \lambda(3\pi/2) = 0$.

[1] The most general statement is known as *the Stokes formula*.

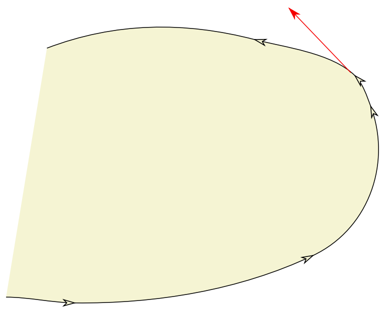

Figure 4.5. The set is on the left

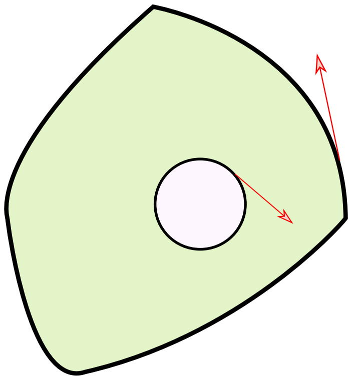

Figure 4.6. The set is on the left

Theorem 4.6.3 (Green's formula). *Let $X \subset \mathbf{R}^2$ be a compact set with a boundary ∂X that is the union of finitely many simple closed parameterized curves γ_1, ..., γ_k. Assume that*

$$\gamma_i \colon [a_i, b_i] \to \mathbf{R}^2$$

has the property that X lies always "to the left" of the tangent vector $\gamma_i'(t)$ based at $\gamma_i(t)$. Let $f = (f_1, f_2)$ be a vector field of class C^1 defined on some open set containing X. Then we have

$$\int_X \left(\frac{\partial f_2}{\partial x} - \frac{\partial f_1}{\partial y}\right) dxdy = \sum_{i=1}^{k} \int_{\gamma_i} f \cdot d\vec{s}.$$

The condition that X be on the left of the boundary is illustrated in Figure 4.5. We then say that the boundary is *positively oriented* by the corresponding parameterized curves γ_i.

Note that if this condition is not met, it simply means that one must "reverse" the corresponding curve, e.g., replace $\gamma \colon [0, 1] \to \mathbf{R}^2$ by $\tilde{\gamma}(t) = \gamma(1 - t)$ for $0 \leqslant t \leqslant 1$, which reverses the orientation of the tangent vector.

Another case, where there are two boundary curves, shows again the way the boundary must be oriented possibly in different directions depending on which part of the boundary is involved (see Figure 4.6).

Example 4.6.4. (1) Suppose that the set X has only one boundary curve γ, and that f is a conservative vector field. Then we see that the Green formula holds, since both sides are then zero (the right-hand side by Remark 4.1.9, and the left-hand side by Example 4.1.14 (2)).

(2) If X is a closed disc of radius $r > 0$ around $(x_0, y_0) \in \mathbf{R}^2$, then the boundary is the circle which is the image of the parameterized curve

$$\gamma(t) = (x_0 + r\cos(t), x_0 + r\sin(t))$$

for $0 \leqslant t \leqslant 2\pi$. Note that this is a simple closed curve; the tangent vector is

$$\gamma'(t) = (-r\sin(t), r\cos(t))$$

and one sees on a picture that the disc is to the left of $\gamma'(t)$ (e.g., $\gamma'(0) = (0, r)$ is a vertical vector based at $\gamma(0) = (r, 0)$, so the disc is located to the left).

Hence the Green formula becomes

$$\int_X \Big(\frac{\partial f_2}{\partial x} - \frac{\partial f_1}{\partial y}\Big) dxdy = \int_\gamma f \cdot d\vec{s}.$$

Let us specialize the vector field to $f(x, y) = (0, x)$. Then the formula becomes

$$\int_X dxdy = \int_\gamma f \cdot d\vec{s}.$$

Indeed, the left-hand side is the area πr^2 of the disc, and we can check that the right-hand side is

$$\int_0^{2\pi} (x_0 + r\cos(t))(r\cos(t))dt = \int_0^{2\pi} r^2\cos^2(t)dt = r^2\int_0^{2\pi} \frac{1}{2}\Big(1 + \cos(2t)\Big)dt = \pi r^2.$$

In this case, it is most likely the computation of the area of the disc that is the main interest. Many other vector fields have the property that

$$\frac{\partial f_2}{\partial x} - \frac{\partial f_1}{\partial y} = 1$$

(e.g. $f(x, y) = (g(x), x)$ where g is an arbitrary function) but it is of course best to choose a simple one to facilitate the computation of the line integral.

(3) More generally, we can always use the Green formula to compute an integral over X. Indeed, for any function g, we can find many vector fields $f = (f_1, f_2)$ such that

$$g = \partial_x f_2 - \partial_y f_1.$$

For instance, we can put $f_1 = 0$ and find f_2 by solving $\partial_x f_2 = g$ (computing a primitive with respect to the x variable).

As an example, let $g(x, y) = x^2y^2$ and let X be the interior of an ellipse centered at 0 with axes lengths $a > 0$ in the x-direction and $b > 0$ in the y-direction. We want to compute

$$\int_X g(x, y)dxdy.$$

We put $f(x, y) = (0, \frac{1}{3}x^3y^2)$ to have $\partial_x f_2 = g$, and we parameterize the boundary by

$$\gamma(t) = (a\cos(t), b\sin(t)), \qquad 0 \leqslant t \leqslant 2\pi,$$

which is a simple closed parameterized curve. So

$$\begin{aligned}\int_X g(x, y)dxdy &= \int_\gamma f \cdot d\vec{s} \\ &= \frac{1}{3}a^3b^2\int_0^{2\pi} \cos^3(t)\sin^2(t) \times b\cos(t)dt.\end{aligned}$$

Using trigonometric computations as in Example 4.4.4, we find that

$$\int_0^{2\pi} \cos^4(t)\sin^2(t)dt = \frac{\pi}{8},$$

so the integral is $\pi a^3 b^3/24$.

(4) Consider the square $X = [0,1]^2 \subset \mathbf{R}^2$ and the vector field $f(x,y) = (xy, x^2 - y^2)$. We want to compute the line integral over the boundary

$$\int_{\partial X} f \cdot d\vec{s},$$

where the boundary is taken counterclockwise (so that it satisfies the "set is on the left" condition). We do not even need to write a parameterization of the boundary square. By Green's Formula and Fubini's Formula, we get

$$\int_{\partial X} f \cdot d\vec{s} = \int_0^1 \int_0^1 \Big(2x - x\Big) dxdy = \frac{1}{2}.$$

(5) Green's formula is equivalent with a variant where we integrate the *divergence* of a vector field $f = (f_1, f_2)$, which we recall is defined by

$$\mathrm{div}(f) = \mathrm{Tr}(J_f) = \partial_x f_1 + \partial_y f_2$$

(see Definition 3.3.11). Indeed, note that

$$\mathrm{div}(f) = \partial_x \widetilde{f}_2 - \partial_y \widetilde{f}_1,$$

where $\widetilde{f}(x,y) = (-f_2, f_1)$. So we have, under the assumptions that Green's Formula is valid for X and its boundary, the relation

$$\int_X \mathrm{div}(f) dxdy = \sum_{i=1}^{k} \int_{\gamma_i} \widetilde{f} \cdot d\vec{s}.$$

It is customary to note that the line integral for the boundary component γ_i is the integral of

$$\widetilde{f}_1(\gamma_i(t))\gamma'_{i,1}(t) + \widetilde{f}_2(\gamma_i(t))\gamma'_{i,2}(t) = -f_2(\gamma_i(t))\gamma'_{i,1}(t) + f_1(\gamma_i(t))\gamma'_{2,i}(t) = f(\gamma_i(t)) \cdot \vec{n}(t)$$

where

$$\vec{n}(t) = (\gamma'_{i,2}(t), -\gamma'_{i,1}(t)).$$

For this reason, this variant of the Green formula is often written

$$\int_X \mathrm{div}(f) dxdy = \sum_{i=1}^{k} \int_{\gamma_i} f \cdot d\vec{n}.$$

For each parameterized curve, note that $\vec{n}(t) \cdot \gamma'(t) = 0$ for all t: in other words, $\vec{n}(t)$ is a vector perpendicular (or *normal*) to the tangent vector to the curve, and that it points "outwards" of X (i.e., it goes "to the right" since γ' has the property that X is "to the left"). In fact, this vector is characterized by the conditions that (1) the length of $\vec{n}(t)$ is the same as the length of $\gamma'(t)$; (2) it is perpendicular to $\gamma'(t)$; (3) $\vec{n}(t)$ is directed "outwards". One says that $\vec{n}$ is the "exterior normal vector".

As a further special case of the divergence form of the Green formula, when we apply it to the gradient field ∇g of a function g, then we obtain

$$\int_X \Delta(g) dxdy = \sum_{i=1}^{k} \int_{\gamma_i} \nabla(g) \cdot d\vec{n}$$

since $\operatorname{div}(\nabla g) = \Delta g$ is the Laplacian of g (Example 3.5.8). For instance, it follows that if ∇g is parallel to the boundary (i.e., orthogonal to $\vec{n}$), then the integral of Δg over X is zero, which is not at all obvious from the definition!

We state separately the general example of using the Green formula for a suitable vector field to compute the area of a region:

Corollary 4.6.5. *Let $X \subset \mathbf{R}^2$ be a compact set with a boundary ∂X that is the union of finitely many simple closed parameterized curves $\gamma_1, \ldots, \gamma_k$. Assume that*

$$\gamma_i = (\gamma_{i,1}, \gamma_{i,2})\colon [a_i, b_i] \to \mathbf{R}^2$$

has the property that X lies always "to the left" of the tangent vector $\gamma_i'(t)$ based at $\gamma_i(t)$. Then we have

$$\operatorname{Vol}(X) = \sum_{i=1}^{k} \int_{\gamma_i} x \cdot d\vec{s} = \sum_{i=1}^{k} \int_{a_i}^{b_i} \gamma_{i,1}(t)\gamma_{i,2}'(t)dt.$$

4.7. The Gauss–Ostrogradski formula

The Gauss–Ostrogradski formula is an analogue of the Green formula in $\mathbf{R}^3$. Thus it concerns a 3-dimensional compact set $X \subset \mathbf{R}^3$, with boundary $S = \partial X$ which is a surface (2-dimensional).

Definition 4.7.1. A parameterized surface $\Sigma\colon [a,b] \times [c,d] \to \mathbf{R}^3$ is a 2-set in $\mathbf{R}^3$ such that the rank of the Jacobian matrix is 2 at all $(s,t) \in]a,b[\times]c,d[$.

Note that since there are two variables, two is the maximal possible rank for the jacobian matrix.

Example 4.7.2. (1) Consider a function $g\colon [a,b]\times[c,d] \to \mathbf{R}$ that is C^1 in $]a,b[\times]c,d[$. Then the function

$$\Sigma(s,t) = (s, t, g(s,t))$$

defines a parameterized surface in $\mathbf{R}^3$, whose image is the graph of g. Indeed, the Jacobian matrix is

$$J_\Sigma(s,t) = \begin{pmatrix} 1 & 0 \\ 0 & 1 \\ \partial_s g & \partial_t g \end{pmatrix}$$

which has rank 2 for all (s,t), since the first two rows are linearly independent.

(2) The sphere of radius $r > 0$ centered at (x_0, y_0, z_0) is the image of the parameterized surface

$$\Sigma(s,t) = \begin{pmatrix} x_0 + r\cos(s)\sin(t) \\ y_0 + r\sin(s)\sin(t) \\ z_0 + r\cos(t) \end{pmatrix}$$

for $(s,t) \in [0, 2\pi] \times [0, \pi]$. The Jacobian matrix is

$$\begin{pmatrix} -r\sin(s)\sin(t) & r\cos(s)\cos(t) \\ r\cos(s)\sin(t) & r\sin(s)\cos(t) \\ 0 & -r\sin(t) \end{pmatrix}.$$

It has rank 2 if $(s,t) \in]0, 2\pi[\times]0, \pi[$ (in that case, the second and third rows define an invertible 2×2 matrix unless $\cos(s) = 0$; but when that is the case, namely $s = \pi/2$ or $3\pi/2$, the first and the third rows define an invertible 2×2 matrix).

(3) Another parameterization of the same sphere is given by

$$\Sigma(s,t) = \frac{1}{(1+s^2+t^2)} \begin{pmatrix} x_0 + 2rs \\ y_0 + 2rt \\ z_0 + r(1-s^2-t^2) \end{pmatrix}$$

for $(s,t) \in \mathbf{R}^2$ (although this is not a compact set in $\mathbf{R}^2$).

Indeed, first note that

$$\|\Sigma(s,t) - (x_0, y_0, z_0)\|^2 = \frac{1}{(1+s^2+t^2)^2}\Big(4r^2s^2 + 4r^2t^2 + r^2(1-s^2-t^2)^2\Big) = r^2$$

for all $(s,t) \in \mathbf{R}^2$, so that the image of Σ is contained in the sphere. It covers the whole sphere, except $(0,0,-1)$. Indeed, we may assume by translating that $(x_0, y_0, z_0) = 0$. Then consider (s,t) with $s^2 + t^2 = u^2$ fixed (in other words, a circle of radius u). The image of this subset of $\mathbf{R}^2$ is the circle centered at $(0,0,(1-u^2)/(1+u^2))$ that is contained in the unit sphere. The function $u \mapsto (1-u^2)/(1+u^2) = -1 + 2/(1+u^2)$ is strictly decreasing for $u \geqslant 0$, going from 1 to the limit -1 as $u \to +\infty$.

The Jacobian matrix is

$$\frac{1}{1+s^2+t^2} \begin{pmatrix} 2r - 4rs^2/(1+s^2+t^2) & -4rst/(1+r^2+s^2) \\ -4rst/(1+s^2+t^2) & 2r - 4rt^2/(1+s^2+t^2) \\ -4rs/(1+s^2+t^2) & -4rt/(1+s^2+t^2) \end{pmatrix}.$$

It is of rank 2 for all (s,t) (check that the second and third rows are independent unless $s = 0$, in which case the first and second rows are independent).

We next recall a definition from linear algebra.

Definition 4.7.3. Let x and y be two linearly independent vectors in $\mathbf{R}^3$. The *vector product*, or *cross product* $z = x \times y$ is the unique vector in $\mathbf{R}^3$ such that (x,y,z) is a basis of $\mathbf{R}^3$ with $\det(x,y,z) > 0$, and

$$\|z\| = \|x\|\,\|y\|\,\sin(\theta),$$

where θ is the angle between x and y.

If x and y are not linearly independent, we just define $x \times y = 0$, the zero vector. The formula for the length of the cross-product is still valid.

We recall that there is in fact an elementary formula: if $x = (x_1, x_2, x_3)$ and $y = (y_1, y_2, y_3)$, then

$$x \times y = \begin{pmatrix} x_2y_3 - x_3y_2 \\ x_3y_1 - x_1y_3 \\ x_1y_2 - x_2y_1 \end{pmatrix} = \det \begin{vmatrix} e_1 & e_2 & e_3 \\ x_1 & x_2 & x_3 \\ y_1 & y_2 & y_3 \end{vmatrix},$$

(with the same formal style of computation as in Remark 4.1.21, where (e_1, e_2, e_3) is the canonical basis of $\mathbf{R}^3$).

Remark 4.7.4. In particular, note the useful formulas

$$e_1 \times e_2 = e_3, \qquad e_2 \times e_3 = e_1, \qquad e_3 \times e_1 = e_2,$$

and $y \times x = -x \times y$.

If (f_1, f_2, f_2) is a basis basis in $\mathbf{R}^3$, there are two possibilities: either $\det(f_1, f_2, f_3) > 0$ or $\det(f_1, f_2, f_3) < 0$. The first type are called *positively oriented*. An example is the canonical basis (e_1, e_2, e_3), which has determinant 1.

If the basis (f_1, f_2, f_3) is orthogonal, it is possible to check that all positively oriented orthonormal bases, for instance $(f_1/\|f_1\|, f_2/\|f_2\|, f_3/\|f_3\|)$, are of the form (Ae_1, Ae_2, Ae_3)

where A is a rotation matrix (an element of $\mathrm{SO}_3(\mathbf{R})$). Intuitively, that means they can be obtained from the canonical basis by rotation.

Let $\Sigma\colon [a,b]\times[c,d]\to\mathbf{R}^3$ be a parameterized surface such that Σ is injective on $]a,b[\times]c,d[$. For all (s,t), the vector $\vec{n}=\partial_s\Sigma(s,t)\times\partial_t\Sigma(s,t)$ is orthogonal to the two vectors $\partial_s\Sigma(s,t)$ and $\partial_t\Sigma(s,t)$, which are linearly independent since the Jacobian matrix of Σ has rank 2. Intuitively, the two vectors span the tangent plane to the surface, hence this vector $\vec{n}$ is *perpendicular* to the surface.

Consider now a 3-dimensional compact subset X of $\mathbf{R}^3$ with boundary ∂X given by the image of the parameterized surface $\Sigma\colon [a,b]\times[c,d]\to\partial X$. (For instance, S could be a ball in $\mathbf{R}^3$ of some radius $r>0$, and the boundary ∂S would be the corresponding sphere sphere.)

For the boundary surface Σ, the orientation condition that is the correct analogue of that concerning the boundary curves in Theorem 4.6.3 is now that the normal vector $\vec{n}$ based at any point of the boundary should point *away* from X: it should be an "exterior normal vector".

Example 4.7.5. Consider the parameterized sphere of Example 4.7.2. Then

$$\partial_s\Sigma=\begin{pmatrix}-r\sin(s)\sin(t)\\ r\cos(s)\sin(t)\\ 0\end{pmatrix},\qquad \partial_t\Sigma=\begin{pmatrix}r\cos(s)\cos(t)\\ r\sin(s)\cos(t)\\ -r\sin(t)\end{pmatrix}.$$

We compute the cross product

$$\partial_s\Sigma\times\partial_t\Sigma=-r^2\sin(t)\begin{pmatrix}\cos(s)\sin(t)\\ \sin(s)\sin(t)\\ \cos(t)\end{pmatrix}.$$

One can check that this is an *interior* normal vector. For instance, let $s=\pi$ and $t=\pi/2$, so that $\Sigma(s,t)=(x_0-r,y_0,z_0)$; then $\partial_s\Sigma=-re_2$ and $\partial_t\Sigma=-re_3$, so that the cross product is

$$\partial_s\Sigma\times\partial_t\Sigma=r^2e_2\times e_3=r^2e_1,$$

which points inside the ball from the point (x_0-r,y_0,z_0).

Here is the formula:

Theorem 4.7.6 (Gauss–Ostrogradski formula). *Let $X\subset\mathbf{R}^3$ be a compact set with a boundary ∂X that is a parameterized surface $\Sigma\colon[a,b]\times[c,d]\to\mathbf{R}^3$. Assume that Σ is injective in $]a,b[\times]c,d[$, and that Σ has the property that the normal vector $\vec{n}$ points away from Σ at all points. Let $\vec{u}=\vec{n}/\|\vec{n}\|$ be the unit exterior normal vector.*

Let $f=(f_1,f_2,f_3)$ be a vector field of class C^1 defined on some open set containing X. Then we have

$$\int_X \operatorname{div}(f)dxdydz=\int_\Sigma (f\cdot\vec{u})d\sigma.$$

In this case, both the left and right-hand sides require come explanation:

(1) For a vector field $f=(f_1,f_2,f_3)$ on $X\subset\mathbf{R}^3$, we denote $\operatorname{div}(f)=\partial_x f+\partial_y f+\partial_z f$, which is called the *divergence* of the vector field f (similarly to the case $n=2$). Hence the left-hand side of the formula is

$$\int_X \operatorname{div}(f)dxdzdz=\int_X\Big(\frac{\partial f}{\partial x}+\frac{\partial f}{\partial y}+\frac{\partial f}{\partial z}\Big)dxdydz.$$

(2) For a parameterized surface $\Sigma\colon [a,b]\times[c,d]\to\mathbf{R}^3$ in $\mathbf{R}^3$ with exterior normal vector field $\vec{n}=(n_1,n_2,n_3)=\partial_s\Sigma\times\partial_t\Sigma$, and a function g defined on the image of Σ, we define the *surface integral*

$$\int_\Sigma g\ d\sigma=\int_a^b\int_c^d g(\Sigma(s,t))\sigma(s,t)dsdt$$

where

$$\sigma(s,t)=\|\partial_s\Sigma\times\partial_t\Sigma\|=\|\vec{n}(s,t)\|.$$

Like the line integral for a parameterized curve, the key property of the surface integral (and especially the explanation for the complicated-looking factor $\|\partial_s\Sigma\times\partial_t\Sigma\|$) is that it is *independent of the chosen parameterization* of the surface (see Proposition 4.1.5). This can be proved by applying the change of variable formula, as in the case of line-integrals.

Next, for a C^1 vector field $f=(f_1,f_2,f_3)$ on $\mathbf{R}^3$, we define

$$\int_\Sigma(f\cdot\vec{n})d\sigma=\int_\Sigma g\ d\sigma,$$

where

$$g(\Sigma(s,t))=f(\Sigma(s,t))\cdot\vec{u}(s,t)=\sum_{i=1}^{3}u_i(s,t)f_i(\Sigma(s,t)).$$

This particular surface integral is called the *flux* of the vector field f through the surface Σ.

Note that in the flux, the expression $\vec{u}(s,t)\sigma(s,t)$ simplifies always to $\vec{n}(s,t)$ since

$$\vec{u}(s,t)\sigma(s,t)=\frac{\vec{n}(s,t)}{\|\vec{n}(s,t)\|}\sigma(s,t)=\vec{n}(s,t).$$

Example 4.7.7. (1) We illustrate first the surface integral. Suppose Σ is a parameterized surface given by $\Sigma(s,t)=(s,t,f(s,t))$ for some function $f\colon[a,b]\times[c,d]\to\mathbf{R}$ (so the image is the graph of f). We take $g(x,y,z)=1$, and we claim that

$$\int_\Sigma d\sigma=\text{the surface area of the graph of } f,$$

which is a natural result. Indeed, we have

$$\partial_s\Sigma=\begin{pmatrix}1\\0\\\partial_s f\end{pmatrix},\qquad\partial_t\Sigma=\begin{pmatrix}0\\1\\\partial_t f\end{pmatrix},$$

hence

$$\partial_s\Sigma\times\partial_t\Sigma=\begin{pmatrix}-\partial_s f\\-\partial_t f\\1\end{pmatrix}$$

so that

$$\|\partial_s\Sigma\times\partial_t\Sigma\|=\Bigl((\partial_s f)^2+(\partial_t f)^2+1\Bigr)^{1/2},$$

hence

$$\int_\Sigma d\sigma=\int_a^b\int_c^d\Bigl((\partial_s f)^2+(\partial_t f)^2+1\Bigr)^{1/2}dsdt$$

is the surface area of the graph according to Section 4.5.

(2) We can use the Gauss–Ostrogradski formula to compute volumes, similarly to the computation of areas using the Green formula. Consider the vector field $f(x, y, z) = (x, 0, 0)$, so that $\mathrm{div}(f) = 1$. Then if $X \subset \mathbf{R}^3$ has boundary $\Sigma\colon [a, b] \times [c, d] \to \mathbf{R}^3$ (an injective parameterized surface) with positive orientation, we have

$$\mathrm{Vol}(X) = \int_\Sigma (f \cdot \vec{n}) d\sigma = \int_a^b \int_c^d n_1(s, t) x(x, t) \sigma(s, t) ds dt,$$

where $\Sigma(s, t) = (x(s, t), y(s, t), z(s, t))$.

Consider the example of the volume of a ball B centered at 0 with radius r in $\mathbf{R}^3$ again, where the boundary is parameterized as in Example 4.7.2. We computed $\partial_s\Sigma \times \partial_t\Sigma$ in Example 4.7.5. Since this normal vector is interior, and

$$\sigma(s, t) = \|\partial_s\Sigma \times \partial_t\Sigma\| = r^2 \sin(t)\Big(\cos^2(s)\sin^2(t) + \sin^2(s)\sin^2(t) + \cos^2(t)\Big)^{1/2} = r^2 \sin(t)$$

we get

$$\begin{aligned}\mathrm{Vol}(B) &= \int_0^{2\pi} \int_0^\pi r\cos(s)\sin(t) \times r^2 \cos(s)\sin^2(t) ds dt \\ &= r^3\Big(\int_0^{2\pi} \cos^2(s) ds\Big)\Big(\int_0^\pi \sin^3(t) dt\Big) = \frac{4\pi r^3}{3},\end{aligned}$$

using the formulas

$$\cos^2(s) = \frac{1}{2}(1 + \cos(2s))$$

$$\sin^3(t) = -\frac{1}{8i}(e^{3it} - 3e^{it} + 3e^{-it} - e^{-3it}) = \frac{1}{4}(3\sin(t) - \sin(3t)),$$

which imply that

$$\int_0^{2\pi} \cos^2(s) ds = \pi$$

and

$$\int_0^\pi \sin^3(t) dt = \frac{1}{4}\Big(3[-\cos(t)]_0^\pi - \frac{1}{3}[\cos(3t)]_0^\pi\Big) = \frac{1}{4}\Big(3 \cdot 2 - \frac{1}{3} \cdot 2\Big) = \frac{4}{3}.$$

Bibliography

[1] M. Burger, *Analysis II*, script for FS 2019 Lecture.