

# MATHEMATICAL EXERCISES

with solutions

DENYS DUTYKH  
CNRS INSMI – LAMA UMR 5127  
Université Savoie Mont Blanc



August 2015 – v0.0.2



*To my wife.*

## ABSTRACT

This document contains some carefully selected problems and exercises from various fields of Mathematics, Physics and Computer Science. The author usually uses these exercises during the Lectures to warm up the students and enable their neural activity. There is another positive *side effect* of these exercises — they allow to recall some knowledge of classical fields of Mathematics and Physics, which is already quietly stored in the deepest parts of the brain by the times of Graduate studies.

## ACKNOWLEDGMENTS

First of all, I am infinitely grateful to the authors of exercises quoted throughout this text with whom this collection would not be possible.

The main part of this manuscript was written at my home laboratory — LAMA UMR #5127 at the University of Savoie Mont Blanc, France. During this period I was supported by my employer — Centre National de la Recherche Scientifique (CNRS). The constant support of my friends and colleagues Dr. Marx CHHAY, Dr. Marguerite GISCLON, Dr. Michel RAIBAUT and many others is also greatly acknowledged.

# CONTENTS

i	ALGEBRA	3
1	GENERAL PROBLEMS	5
ii	ANALYSIS	6
2	FUNCTIONS AND THEIR PROPERTIES	8
2.1	Continuity . . . . .	8
2.2	Differentiation . . . . .	9
3	SERIES SUMMATION	10
3.1	Divergent series . . . . .	10
4	EXTREMA OF A FUNCTION	11
iii	DIFFERENTIAL EQUATIONS	15
5	ORDINARY DIFFERENTIAL EQUATIONS	17
6	PARTIAL DIFFERENTIAL EQUATIONS	18
iv	PHYSICS	19
7	MECHANICS	21
v	COMPUTER SCIENCE	23
8	SCIENTIFIC COMPUTING	25
	BIBLIOGRAPHY	30
	INDEX	31

# LIST OF FIGURES

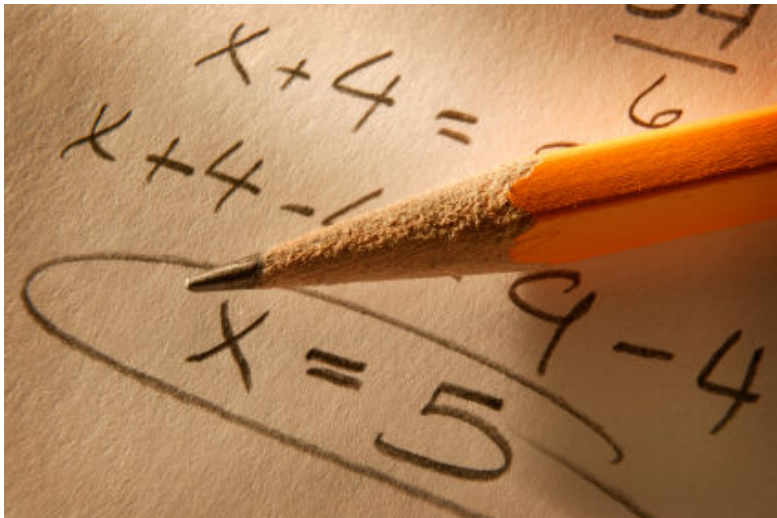
Figure 8.1	Function $G(x)$ from Problem 8.2 computed in MATLAB. . . .	28
Figure 8.2	Graph of the function $y(x)$ from Problem 8.3 plotted in MATLAB with two slightly different resolutions ( $N = 1000$ and $1001$ correspondingly, $N$ being the number of discretization points). . . . .	28

## INTRODUCTION



## Part I

## ALGEBRA



The idea to denote unknowns with the variable  $x$  belongs to © François VIÈTE (1540 – 1603)n who was a decypher in the court of ...

# 1

## GENERAL PROBLEMS

*Mathematicians are like lovers. Grant a mathematician the least principle, and he will draw from it a consequence which you must also grant him, and from this consequence another.*

— Bernard Le Bovier de Fontenelle (1657 – 1757)

*Algebra and money are essentially levelers; the first intellectually, the second effectively.*

— Simone Weil (1909 – 1943)

**Problem 1.1** (V.I. ARNOLD, Interview, 2001). *One has a bag with 100 kg of cucumbers which are composed of 99% of water. Then, the cucumbers were dried and now the water constitutes only 98% of their mass. What is the total mass of cucumbers after drying?*

*It seems that this problem has been posed during the hiring interviews at the Boeing Company.*

*Solution.* Initially the mass of the solid rest constitutes 1% of the mass, i.e. 1 kg and it is not going to change during the drying process. Let us denote<sup>1</sup> the mass of water after drying by  $x$ . Thus, the total mass of cucumbers will be equal to  $x + 1$  kg. By problem statement, the water constitutes now 98% of the total mass. Thus,

$$\frac{x}{x + 1} = 0.98 \implies x = 49.$$


Hence, the mass of cucumbers will be equal to 50 kg. This result might appear *counterintuitive* at the first sight.  $\square$

---

<sup>1</sup> The tradition to denote unknowns with letters  $x$ ,  $y$ ,  $z$  comes from François Viète (1540 – 1603) who was a cryptographer of Henri IV. During his work on a Spanish cipher he denoted the undeciphered symbols with  $x$ ,  $y$  and  $z$ .

Part II

ANALYSIS



$$e(s) = \frac{y_1 - y_0}{x_1 - x_0} = \frac{g(x+h) - g(x)}{(x+h) - x} = \frac{g(x+h) - g(x)}{h}$$

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

$$f(x) = \lim_{h \rightarrow 0} \frac{(x+h)^2 - x^2}{h}$$

$$= \lim_{h \rightarrow 0} \frac{x^2 + 2xh + h^2 - x^2}{h}$$

$$= \lim_{h \rightarrow 0} \frac{2xh + h^2}{h}$$

$$= \lim_{h \rightarrow 0} (2x + h)$$

$$= 2x$$

$$\text{Slope}(T) = \lim_{h \rightarrow 0} \frac{g(x+h) - g(x)}{h}$$

$$\frac{df}{dx} \left[ \frac{d}{dx} (x^n) = nx^{n-1} \right]$$

$$= \lim_{h \rightarrow 0} \frac{h(x+h)}{h}$$

$$= \lim_{h \rightarrow 0} (x+h)$$

$$= x$$

$$f'(x) = \lim_{\Delta x \rightarrow 0} \frac{f(x+\Delta x) - f(x)}{\Delta x}$$

$$f(a) = \lim_{h \rightarrow 0} \frac{f(a+h) - f(a)}{h}$$

$$f'(a) = \lim_{h \rightarrow 0} \frac{f(a+h) - f(a)}{h}$$

# 2

## FUNCTIONS AND THEIR PROPERTIES

*Who has not be amazed to learn that the function  $y = e^x$ , like a phoenix rising again from its own ashes, is its own derivative?*

— François le Lionnais (1901 – 1984)

### 2.1 CONTINUITY

**Problem 2.1.** Give an example of a real everywhere discontinuous function  $f : \mathbb{R} \rightarrow \mathbb{R}$  satisfying the following functional relation

$$f(x + y) = f(x) + f(y). \quad (2.1)$$

*Solution.* Consider the infinite dimensional vector space of real numbers  $\mathbb{R}$  over the field of rationals  $\mathbb{Q}$ . Let  $\{r_\alpha\}$  is a Hamel basis of this linear vector space. Then, for every  $x \in \mathbb{R}$  there exists the representation

$$x = k_{\alpha 1} r_{\alpha 1} + k_{\alpha 2} r_{\alpha 2} + \dots k_{\alpha n} r_{\alpha n}, \quad k_{\alpha i} \in \mathbb{Q}, \quad n \in \mathbb{Z}^+.$$

We can define the function  $f$  as

$$f(x) = k_{\alpha 1} + k_{\alpha 2} + \dots k_{\alpha n}.$$

This function is linear by construction. Thus, it satisfies the functional relation (2.1). However, this function  $f$  cannot be continuous since it takes only rational values.  $\square$

## 2.2 DIFFERENTIATION

**Problem 2.2** (Hadamard's lemma). Let  $f : \mathcal{U}_0 \rightarrow \mathbb{R}$  be a smooth function defined on a neighbourhood  $\mathcal{U}_0 \subset \mathbb{R}^d$  of point  $\mathbf{x}_0 = (0, 0, \dots, 0) \in \mathcal{U}_0$  and such that  $f(\mathbf{x}_0) = 0$ . Show that there exist smooth functions<sup>1</sup>  $g_i : \mathcal{U}_0 \rightarrow \mathbb{R}$ ,  $i = 1, 2, \dots, d$  such that

$$f(x_1, x_2, \dots, x_d) = \sum_{i=1}^d x_i g_i(x_1, x_2, \dots, x_d), \quad (2.2)$$

and moreover  $g_i(0, 0, \dots, 0) = \frac{\partial f}{\partial x_i}(0, 0, \dots, 0)$ .

*Solution.* Indeed, let us construct a rectilinear path connecting the points  $\mathbf{x}_0 = (0, 0, \dots, 0)$  and  $\mathbf{x} = (x_1, x_2, \dots, x_d)$ :

$$p : [0, 1] \rightarrow \mathbb{R}^d \quad p(t) = t\mathbf{x}.$$

Along this path our function  $f(\mathbf{x})$  becomes a function of one variable  $t$ . Now we can perform an integration along this path:

$$\begin{aligned} f(x_1, x_2, \dots, x_d) &= \int_0^1 \frac{d f(tx_1, tx_2, \dots, tx_d)}{dt} dt = \\ &= \sum_{i=1}^d x_i \int_0^1 \frac{\partial f}{\partial x_i}(tx_1, tx_2, \dots, tx_d) dt. \end{aligned}$$

If we denote  $g_i(x_1, x_2, \dots, x_d) := \int_0^1 \frac{\partial f}{\partial x_i}(tx_1, tx_2, \dots, tx_d) dt$ , it yields the desired result (2.2). Finally, in order to show that  $g_i(0, 0, \dots, 0) = \frac{\partial f}{\partial x_i}(0, 0, \dots, 0)$  it is sufficient to set all  $x_i := 0$ ,  $\forall i = 1, \dots, d$  in the integral representation of  $g_i$ :

$$g_i(0, 0, \dots, 0) = \int_0^1 \frac{\partial f}{\partial x_i}(0, 0, \dots, 0) dt = \frac{\partial f}{\partial x_i}(0, 0, \dots, 0).$$

Equation (2.2) is basically a different form of the Taylor formula with integral reminder term.  $\square$

<sup>1</sup> It can be additionally shown that if  $f \in C^k$ , then  $g_i \in C^{k-1}$ ,  $i = 1, \dots, d$ .

# 3

## SERIES SUMMATION

*I recognize the lion by his paw.*

— Jacob Bernoulli (1654 – 1705)

[After reading an anonymous solution he realized was Newton's one.]

*Read Euler: he is our master in everything.*

— Pierre-Simon de Laplace (1749 – 1827)

### 3.1 DIVERGENT SERIES



# 4

## EXTREMA OF A FUNCTION

*The calculus is the greatest aid we have to the application of physical truth in the broadest sense of the word.*

— William Fogg Osgood (1864 – 1943)

Before tackling some problems about finding the extrema of functions of several variables, let us recall some key classical results from this area.

**Definition 4.1.** A function  $f : E \rightarrow \mathbb{R}$  defined on a set  $E \subset \mathbb{R}^m$  has a local maximum (local minimum) in an internal point  $\mathbf{x}_0 \in E$ , if there exists a neighbourhood  $U(\mathbf{x}_0) \subset E$  such that  $f(\mathbf{x}) \leq f(\mathbf{x}_0)$  ( $f(\mathbf{x}) \geq f(\mathbf{x}_0)$ ) respectively,  $\forall \mathbf{x} \in U(\mathbf{x}_0)$ .

Local minima and maxima are simply called local extrema. The following Theorem gives us necessary conditions of local extrema:

**Theorem 4.1.** Let a function  $f : U(\mathbf{x}_0) \rightarrow \mathbb{R}$  has all partial derivatives in point  $\mathbf{x}_0$ . Then, if the function  $f$  has a local extremum in  $\mathbf{x}_0$ , then necessarily we have

$$\frac{\partial f}{\partial x_1}(\mathbf{x}_0) = \frac{\partial f}{\partial x_2}(\mathbf{x}_0) = \dots = \frac{\partial f}{\partial x_m}(\mathbf{x}_0) \equiv 0.$$

The points  $\mathbf{x}_0$  where all partial derivatives vanish are called critical points of a function  $f$ . Finally the following Theorem provides us with sufficient conditions to have (or not) a local extremum:

**Theorem 4.2.** Let a function  $f : U(\mathbf{x}_0) \rightarrow \mathbb{R}$  is of class  $C^2(U(\mathbf{x}_0); \mathbb{R})$  and  $\mathbf{x}_0$  is a critical point of function  $f$ . Then, the point  $\mathbf{x}_0$  is a local minimum (maximum) of function  $f$  if the quadratic form  $Q(\mathbf{h}) = \sum_{i,j=1}^m \frac{\partial^2 f}{\partial x_i \partial x_j}(\mathbf{x}_0) h_i h_j$  is positive definite (negative definite respectively). If the quadratic form  $Q(\mathbf{h})$  can take the values of both signs, then the function  $f$  does not have an extremum in point  $\mathbf{x}_0$ .

**Problem 4.1 ([Zoro8]).** Find extrema of a function  $f : \mathbb{R}^2 \rightarrow \mathbb{R}$  given by the formula

$$f(x, y) = x^4 + y^4 - 2x^3.$$

*Solution.* From necessary conditions of a local extremum, we find

$$\begin{aligned}\frac{\partial f}{\partial x} &= 4x^3 - 4x = 0, \\ \frac{\partial f}{\partial y} &= 4y^3 = 0.\end{aligned}$$

The last system of equations admits three solutions  $\mathbf{x}_1 = (-1, 0)$ ,  $\mathbf{x}_2 = (0, 0)$ ,  $\mathbf{x}_3 = (1, 0)$ . Let us compute the coefficients of the quadratic form  $Q(\mathbf{h})$ :

$$\frac{\partial^2 f}{\partial x^2} = 12x^2 - 4, \quad \frac{\partial^2 f}{\partial x \partial y} = 0, \quad \frac{\partial^2 f}{\partial y^2} = 12y^2.$$

Then, the quadratic form  $Q(\mathbf{h})$  in the vicinity of critical points takes the form

$$Q_1(\mathbf{h}) = 8h_1^2, \quad Q_2(\mathbf{h}) = -4h_1^2, \quad Q_3(\mathbf{h}) = 8h_1^2.$$

So, in all cases the quadratic form is semi-definite, which makes the application of Theorem 4 impossible. However, we can draw the conclusions by rewriting the function  $f$  as

$$f(x, y) = (x^2 - 1)^2 + y^4 - 1.$$

Now it becomes clear that points  $\mathbf{x}_1$  and  $\mathbf{x}_3$  are local minima. However, the point  $\mathbf{x}_2$  is not an extremum, since  $f(0, y) = y^4 > 0$  and for sufficiently small  $x \neq 0$  we have  $f(x, 0) = x^4 - 2x^2 < 0$ .  $\square$

**Problem 4.2** ([Zoro8]). (Huygens problem) *Imagine two absolutely rigid balls of masses  $m$  and  $M$  ( $0 \leq m \leq M$ ) having the speeds  $v$  and  $V$  correspondingly. Using the energy and momentum conservation laws one can show that after a central collision the speeds of the balls will be respectively*

$$\tilde{v} = \frac{(m - M)v + 2MV}{m + M}, \quad \tilde{V} = \frac{(M - m)V + 2mv}{m + M}.$$

*For the sake of simplicity we shall assume that the ball of mass  $m$  is initially at rest, i.e.  $v = 0$ . Then, the speed of this ball after the collision is*

$$\tilde{v} = \frac{2MV}{m + M}. \quad (4.1)$$

*From the last formula it is clear that  $V \leq \tilde{v} \leq 2V$ . How it would be possible to transfer even more kinetic energy to the ball at rest?*

*Solution.* In order to increase the energy transfer rate from the heavy ball, one can place the balls with intermediate masses  $m_1 < m_2 < \dots < m_n$  between  $m$  and  $M$ . Following Christian HUYGENS (1629 – 1695) we shall compute the masses  $m_i$ ,  $i = 1, \dots, n$  so that the ball at rest acquires the highest possible speed. According to formula (4.1), the new speed  $\tilde{v}$  after the interactions will be

$$\tilde{v} = \frac{2m_1}{m+m_1} \cdot \frac{2m_2}{m_1+m_2} \dots \frac{2m_n}{m_{n-1}+m_n} \cdot \frac{2M}{m_n+M} V.$$

Consequently, the problem is reduced to maximize the following function of  $n$  variables

$$f(m_1, m_2, \dots, m_n) = \frac{m_1}{m+m_1} \cdot \frac{m_2}{m_1+m_2} \dots \frac{m_n}{m_{n-1}+m_n} \cdot \frac{M}{m_n+M}.$$

The necessary condition of the extremum of function  $f(m_1, \dots, m_n)$  is reduced to the following nonlinear system

$$\begin{aligned} m m_2 - m_1^2 &= 0, \\ m_1 m_3 - m_2^2 &= 0, \\ &\dots\dots\dots \\ m_{n-1} M - m_n^2 &= 0. \end{aligned}$$

From this system it follows that numbers  $m, m_1, \dots, m_n, M$  form a geometrical progression with factor  $q = \sqrt[n+1]{\frac{M}{m}}$ . The resulting velocity becomes

$$\tilde{v} = \left( \frac{2q}{q+1} \right)^{n+1} V. \quad (4.2)$$

One can see that for  $n = 0$  we recover formula (4.1). The last step consists in checking that our solution truly corresponds to a maximum. By taking the limit  $m \rightarrow 0$  in (4.2), one obtains that  $\tilde{v} = 2^{n+1} V$ . Now it is clear that the system of intermediate masses increases considerably the speed of the ball comparing it to the result (4.1) of a simple pair-wise interaction.  $\square$

**Problem 4.3** ([Zoro8]). *Find extrema of a quadratic form*

$$Q(\mathbf{x}) = \sum_{i,j=1}^n a_{ij} x_i x_j, \quad (a_{ij} = a_{ji}, \quad \forall i, j = 1, \dots, n),$$

on the unit sphere

$$\sum_{i=1}^n x_i^2 = 1. \quad (4.3)$$

*Solution.* We write the Lagrange function for this problem

$$\mathcal{L} = \sum_{i,j=1}^n a_{ij} x_i x_j - \lambda \left( \sum_{i=1}^n x_i^2 - 1 \right).$$

The necessary conditions of extrema read

$$\frac{\partial \mathcal{L}}{\partial x_i} = 2 \left( \sum_{j=1}^n a_{ij} x_j - \lambda x_i \right) = 0, \quad (4.4)$$

$$\frac{\partial \mathcal{L}}{\partial \lambda} = - \left( \sum_{i=1}^n x_i^2 - 1 \right) = 0. \quad (4.5)$$

where we took into account the symmetry of coefficients  $a_{ij}$ . Condition (4.3) can be rewritten as

$$\sum_{j=1}^n a_{ij} x_j = \lambda x_i, \quad i = 1, \dots, n.$$

Now it is clear that the Lagrange multiplier  $\lambda$  has to be an eigenvalue of the matrix  $\{a_{ij}\}$  and a critical point  $\mathbf{x} = (x_1, x_2, \dots, x_n)$  is the corresponding eigenvector.

Multiplying the first equation (4.4) by  $x_i$  and taking the sum for  $\forall i = 1, \dots, n$  yields

$$\sum_{i,j=1}^n a_{ij} x_i x_j = \lambda \underbrace{\sum_{i=1}^n x_i^2}_{=1 \text{ by (4.3)}} = \lambda.$$

The last identity has to be satisfied in any critical point. Thus, the Lagrange multiplier  $\lambda$  provides the extremal value of the quadratic form  $Q(\mathbf{x})$ . The existence of the solution comes from the fact that a continuous function on a compact set  $\{\mathbf{x} \in \mathbb{R}^n \mid \sum_{i=1}^n x_i^2 = 1\}$  has to attain its maximal and minimal values.  $\square$

Part III

DIFFERENTIAL EQUATIONS

$$\begin{aligned} \frac{\partial}{\partial a} \ln f_{a, \sigma^2}(\xi_1) &= \frac{(\xi_1 - a)}{\sigma^2} f_{a, \sigma^2}(\xi_1) = \frac{1}{\sqrt{2\pi\sigma}} \exp\left\{-\frac{(\xi_1 - a)^2}{2\sigma^2}\right\} \\ \int_{\mathbb{R}_n} T(x) \cdot \frac{\partial}{\partial \theta} f(x, \theta) dx &= M\left(T(\xi) \cdot \frac{\partial}{\partial \theta} \ln L(\xi, \theta)\right) \cdot \int_{\mathbb{R}_n} \frac{\partial}{\partial \theta} f(x, \theta) dx \\ \int_{\mathbb{R}_n} T(x) \cdot \left(\frac{\partial}{\partial \theta} \ln L(x, \theta)\right) \cdot f(x, \theta) dx &= \int_{\mathbb{R}_n} T(x) \cdot \left(\frac{\frac{\partial}{\partial \theta} f(x, \theta)}{f(x, \theta)}\right) \cdot f(x, \theta) dx \\ \frac{\partial}{\partial \theta} M(T(\xi)) &= \frac{\partial}{\partial \theta} \int_{\mathbb{R}_n} T(x) f(x, \theta) dx = \int_{\mathbb{R}_n} \frac{\partial}{\partial \theta} T(x) f(x, \theta) dx \end{aligned}$$

# 5

## ORDINARY DIFFERENTIAL EQUATIONS

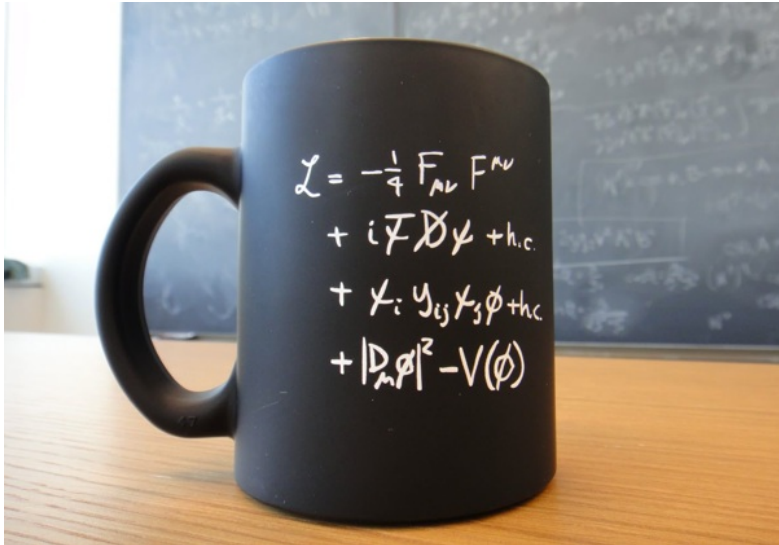
# 6

## PARTIAL DIFFERENTIAL EQUATIONS



Part IV

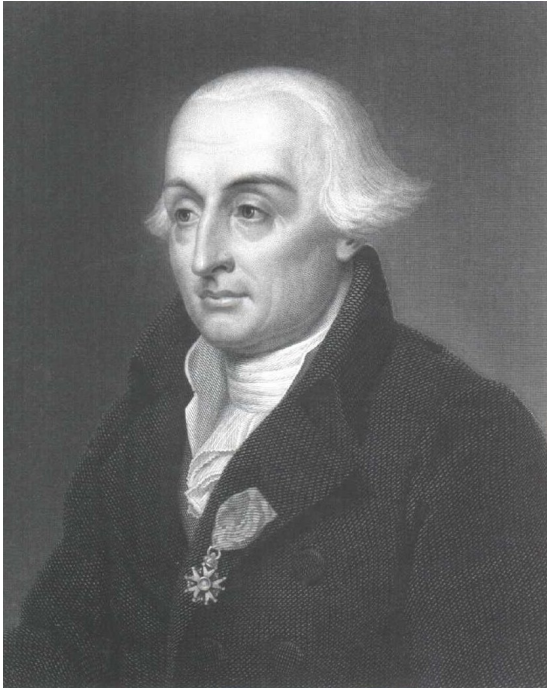
PHYSICS



CERN Mug. Lagrangian of the Standard Model in Particle Physics.

# 7

## MECHANICS



*La mécanique analytique* by © Joseph-Louis LAGRANGE (1736 – 1813)

**Problem 7.1** (Proper time, [LL75]). We observe from an inertial coordinate system a clock moving with the speed  $v$  from the moment of time  $t = t_1$  to  $t = t_2 > t_1$ . What will be the corresponding interval of time measured by the moving clock?

*Solution.* Denote by  $C = \{Otxyz\}$  and  $C' = \{O't'x'y'z'\}$  our (fixed) and moving coordinate frames of reference. During an infinitesimal interval of time  $dt$  the clock will travel the distance  $\sqrt{dx^2 + dy^2 + dz^2}$ . In the frame of reference  $C'$  the clock is steady, i.e.  $dx' = dy' = dz' = 0$ . We have to find the elapsed interval of time  $dt'$ .

Since the interval  $ds$  is invariant (see [LL75, Chapter 1, §2]), one has

$$ds^2 = c^2 dt^2 - dx^2 - dy^2 - dz^2 \equiv c^2 dt'^2,$$

where  $c$  is the speed of light in the vacuum. From where we find:

$$dt' = \sqrt{1 - \frac{dx^2 + dy^2 + dz^2}{c^2 dt^2}} dt, \quad (7.1)$$

but we know that  $\frac{dx^2 + dy^2 + dz^2}{dt^2} = v^2$ , where  $v$  is the speed of moving watch. Consequently, equation (7.1) can be rewritten as

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

and after integration we obtain the requested time interval

$$t'_2 - t'_1 = \int_{t_1}^{t_2} \sqrt{1 - \frac{v^2}{c^2}} dt.$$

□

*The invariance of interval is a consequence of the fact that the speed of light is a constant independent of the reference frame.*

**Part V**

**COMPUTER SCIENCE**

```
int    yourDick;  
long int myDick;
```

**Problem 8.1** (J.-M. Muller's sequence). The sequence  $\{x_n\}_{n=0}^{+\infty}$  is defined as

$$x_n = M(x_{n-1}, x_{n-2}), \quad n \geq 2, \quad (8.1)$$

where the iteration function  $M(\cdot, \cdot)$  is given by

$$M(y, z) = 108 - \frac{815 - \frac{1500}{z}}{y}.$$

The recurrence relation (8.1) is supplied with the following starting conditions:

$$x_0 = 4, \quad x_1 = 4.25 = \frac{5}{4}.$$

Compute numerically  $x_{30}$ .

*Explanations.* A typical numerically generated sequence  $\{x_n\}_{n=0}^{+\infty}$  looks as given in Table 8.1 (see [Kaho6] for more details). One can see that the sequence  $\{x_n\}_{n=0}^{+\infty}$  seems to converge numerically to 100.0. Let us see what the theory says about recurrence (8.1).

Now it can be easily seen<sup>1</sup> that for given starting conditions the sequence  $\{x_n\}_{n=0}^{+\infty}$  is

$$x_n = \frac{3^{n+1} + 5^{n+1}}{3^n + 5^n}, \quad n \geq 0.$$

<sup>1</sup> Perhaps, it is not that obvious and some further explanations are needed here (see [Kaho6] for even more details). Let us substitute  $x_n = \frac{z_{n+1}}{z_n}$  into the recurrence (8.1). One obtains

$$z_{n+2} = 108z_{n+1} - 815z_n + 1500z_{n-1}.$$

The last linear recurrence relation can be solved using the method of characteristic polynomials, i.e.  $z_n = \lambda^n$ :

$$\lambda^3 - 108\lambda^2 + 815\lambda - 1500 \equiv (\lambda - 3)(\lambda - 5)(\lambda - 100) = 0.$$

Consequently, the general solution to (8.1) is given by the following formula:

$$x_n = \frac{C_1 3^{n+1} + C_2 5^{n+1} + C_3 100^{n+1}}{C_1 3^n + C_2 5^n + C_3 100^n},$$

where  $C_i$ ,  $i = 1, 2, 3$  are some arbitrary constants. Our particular sequence is obtained for  $C_1 = C_2 = 1$  and  $C_3 = 0$ .

*We assume that all computations are done in the standard floating point arithmetics with double precision.*

So, the limit of the sequence  $\{x_n\}_{n=0}^{+\infty}$  in the *exact* arithmetics can be readily computed:

$$\lim_{n \rightarrow +\infty} x_n = \lim_{n \rightarrow +\infty} \frac{3\left(\frac{3}{5}\right)^n + 5}{\left(\frac{3}{5}\right)^n + 1} = 5.$$

The peculiarity here is that the recurrence (8.1) has three fixed points: 3, 5, 100 and only the last one (100) is stable. Taking into account that rounding errors are unavoidable, the numerically generated sequence  $\{x_n\}_{n=0}^{+\infty}$  falls in the attraction basin of the stable fixed point.  $\square$

**Problem 8.2** (A smooth surprise, [Kaho6, § 6]). *Let*

$$T(x) = \begin{cases} \frac{\exp(x) - 1}{x}, & x \neq 0, \\ 1, & x = 0, \end{cases}$$

$$Q(x) = \left| x - \sqrt{1 + x^2} \right| - \frac{1}{x + \sqrt{1 + x^2}}.$$

*Compute numerically*  $G(x) = T(Q^2(x))$ .

*Indications.* One can notice that in the exact arithmetics  $Q(x) \equiv 0$  for  $\forall x \in \mathbb{R}$  (but not *complex*  $x$ !). Consequently,  $G(x) \equiv 1$  for  $\forall x \in \mathbb{R}$ . However, the numerical computation is necessarily performed in the presence of round-off errors. A graph of the function  $G(x)$  produced and computed with MATLAB in double accuracy is shown in Figure 8.1. The reader can see that round-off errors do matter! Function  $G(x) = 0$  for almost every  $x$  in contrast with its exact value. The reason for this discrepancy is that in floating point arithmetics  $Q(x)$  takes tiny values which are hardly ever zero.

This example may appear somehow artificial. Unfortunately it is not. The function  $T(z)$  appears routinely in the so-called exponential integrators of stiff ODEs [HO10].  $\square$

**Problem 8.3** (A spike, [Kaho6, § 7]). *Plot the graph of the function*  $y(x)$  *on*  $x \in [\frac{1}{2}, 2]$ , *where*

$$y(x) = 1 + x^2 + \frac{1}{80} \log(|1 + 3(1 - x)|).$$

*Indications.* Two sample graphs are shown in Figure 8.2. However, in both cases we fail to capture the correct behaviour of the function because of the discretization effects. In reality the function  $y(x)$  has to plunge down since  $y(4/3) = -\infty$  but it does not.  $\square$



$n$	$x_n$ (synthetic)	$x_n$ (exact)
0	4.00000000000000	4.00000000000000
1	4.25000000000000	4.25000000000000
2	4.4705882352941	4.4705882352941
3	4.6447368421052	4.6447368421052
4	4.7705382436260	4.7705382436260
5	4.8557007125890	4.8557007125890
6	4.9108474990828	4.9108474990827
7	4.9455374041250	4.9455374041239
8	4.9669625817851	4.9669625817627
9	4.9800457018084	4.9800457013556
10	4.9879794575704	4.9879794484783
11	4.9927704703332	4.9927702880620
12	4.9956595420973	4.9956558915066
13	4.9974643422978	4.9973912683813
14	4.9998961477637	4.9984339439448
15	5.0283045630311	4.9990600719708
16	5.5810310849684	4.9994359371468
17	15.420563287948	4.9996615241037
18	72.577658482982	4.9997969007134
19	98.110905976394	4.9998781354779
20	99.903728999705	4.9999268795046
21	99.995181883411	4.9999561270611
22	99.999759084721	4.9999736760057
23	99.999987954271	4.9999842055202
24	99.999999397715	4.9999905232822
25	99.999999969885	4.9999943139585
26	99.999999998494	4.9999965883712
27	99.99999999924	4.9999979530213
28	99.99999999996	4.9999987718123
29	99.99999999999	4.9999992630872
30	99.99999999999	4.9999995578522
31	100.00000000000	4.9999997347113
32	100.00000000000	4.9999998408267
33	100.00000000000	4.9999999044960

**Table 8.1:** Numerically generated J.-M. Muller's sequence  $\{x_n\}_{n=0}^{+\infty}$ . The synthetic values are taken from [Kaho6].

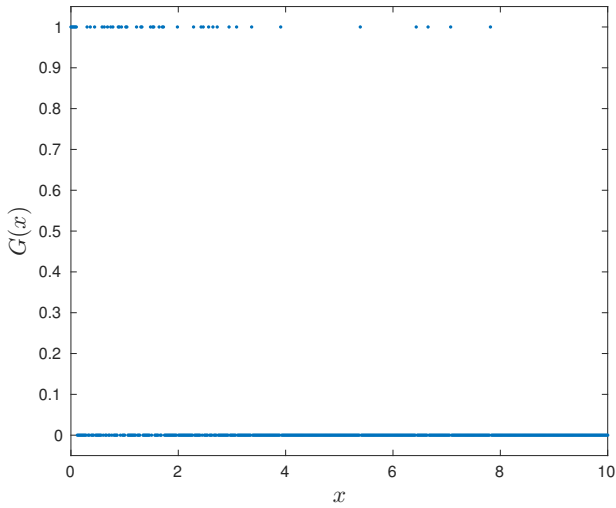


Figure 8.1: Function  $G(x)$  from Problem 8.2 computed in MATLAB.

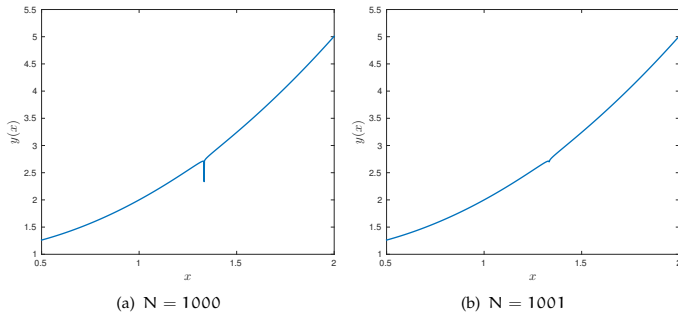


Figure 8.2: Graph of the function  $y(x)$  from Problem 8.3 plotted in MATLAB with two slightly different resolutions ( $N = 1000$  and  $1001$  correspondingly,  $N$  being the number of discretization points).

---

★ ★ ★

---

From the examples given above, one can see that in some situations the peculiarities of floating-point computations can lead to totally wrong and/or unpredictable results. As a general recommendation to diagnose such eventual floating-point arithmetics problems, Prof. W. KAHAN [Kaho6, § 4] suggests to follow this scheme:

- Repeat the computation in arithmetics of increasing precision. Significant digits in the numerical result must agree.
- Repeat the computation in arithmetics of the same precision but rounded differently (towards zero or towards infinity) and compare these results.
- Repeat the computation in arithmetics of the same precision but rounded randomly and treat the results statistically.
- Repeat the computation in arithmetics of the same precision but with slightly perturbed input data and see how the results spread.
- Perform the computations in *interval arithmetics* to obtain a bound on results.

However, one has to remember that the effects of round-off cannot be in general assessed without a mathematically rigorous and time-consuming error analysis [Kaho6].

## BIBLIOGRAPHY

- [HO10] M. Hochbruck and A. Ostermann. Exponential integrators. *Acta Numerica*, pages 209–286, 2010.
- [Kah06] W. Kahan. How futile are mindless assessments of roundoff in floating-point computation? Technical report, 2006.
- [LL75] L. D. Landau and E. M. Lifshitz. *The Classical Theory of Fields*. Butterworth-Heinenann, Oxford, 4 edition, 1975.
- [Zor08] V. A. Zorich. *Mathematical Analysis I*. Springer Verlag, Berlin Heidelberg New York, 2 edition, January 2008.

# INDEX

- Algebra, 4
- Analysis, 7
- Arnold Vladimir, 5
  
- Bernoulli Jacob, 10
  
- Computer science, 24
- Critical point, 11
  
- de Fontenelle Bernard Le
  - Bovier, 5
- de Laplace Pierre-Simon, 10
- Derivative, 7
- Differential equations, 16
  
- Error analysis, 29
- Euler Leonhard, 16
- Exponential integrator, 26
  
- Floating-point, 29
  
- Huygens Christian, 13
  
- Inertial coordinate system, 22
  
- Lagrange Joseph-Louis, 21
- Le Lonnais François, 8
- Leibniz Gottfried, 7
  
- Local extremum, 11
- Local maximum, 11
- Local minimum, 11
  
- Muller Jean-Michel, 25
  
- Newton Isaac, 7
  
- Ordinary Differential
  - Equations, 26
- Ordinary Differential
  - Equations (ODEs),  
17
- Osgood William Fogg, 11
  
- Partial Differential Equations
  - (PDEs), 18
- Physics, 20
  
- Round-off, 29
  
- Speed of light, 22
  
- Vacuum, 22
- Viète François, 4, 5
  
- Weil Simone, 5



## COLOPHON

This document was typeset using the  $\text{\LaTeX}$  `classicthesis` and `arsclassica` classes.