Mathematics, Information Technologies and Applied Sciences 2022

post-conference proceedings of extended versions of selected papers

Editors:

Michal Novák and Miroslav Hrubý

Brno, Czech Republic, 2022

ISBN 978-80-7582-185-0

Aims and target group of the conference:

The conference **MITAV 2022** is the ninth annual MITAV conference. It should attract in particular teachers of all types of schools and is devoted to the most recent discoveries in mathematics, informatics, and other sciences as well as to the teaching of these branches at all kinds of schools for any age groups, including e-learning and other applications of information technologies in education. The organizers wish to pay attention especially to the education in the areas that are indispensable and highly demanded in contemporary society. The goal of the conference is to create space for the presentation of results achieved in various branches of science and at the same time provide the possibility for meeting and mutual discussions of teachers from different kinds of schools and orientation. We also welcome presentations by (diploma and doctoral) students and teachers who are just beginning their careers, as their novel views and approaches are often interesting and stimulating for other participants.

Organizers:

Union of Czech Mathematicians and Physicists, Brno branch (JČMF),
in co-operation with
Faculty of Military Technology, University of Defence, Brno,
Faculty of Education and Faculty of Economics and Administration, Masaryk University in Brno,
Faculty of Electrical Engineering and Communication, Brno University of Technology.

Venue:

Club of the University of Defence, Šumavská 4, Brno, Czech Republic
June 16 and 17, 2022.

Conference languages:

English, Czech, Slovak

International Scientific committee:

Prof. Leonid BEREZANSKY
Ben Gurion University of the Negev, Beer Sheva, Israel

Prof. Zuzana DOŠLÁ
Masaryk University in Brno, Faculty of Science, Czech Republic

Prof. Irada DZHALLADOVA
Kyiv National Economic Vadym Getman University, Ukraine

Prof. Mihály PITUK
University of Pannonia, Faculty of Information Technology, Veszprém, Hungary

Prof. Ľubica STUCHLÍKOVÁ
Slovak University of Technology in Bratislava, Faculty of Electrical Engineering and Information Technology, Slovakia

International Programme committee:

Chair: **Jaromír BAŠTINEC**
Brno University of Technology, Faculty of Electrical Engineering and Communication, Brno, Czech Republic

Members:
Jan HODICKÝ
M&S Technical SME at NATO HQ SACT, Norfolk, Virginia, USA

Miroslav HRUBÝ
University of Defence, Faculty of Military Technology, Brno, Czech Republic

Edita KOLÁŘOVÁ
Brno University of Technology, Faculty of Electrical Engineering and Communication, Brno, Czech Republic

Piet KOMMERS
Professor of UNESCO Learning Technologies, Emeritus University of Twente, Enschede, the Netherlands

Nataliia MORZE
Borys Grinchenko Kiev University, Kiev, Ukraine

Václav PŘENOSIL
Masaryk University, Faculty of Informatics, Brno, Czech Republic

Magdalena ROSZAK
Poznan University of Medical Sciences, Department of Computer Science and Statistics, Poznań, Poland

Eugenia SMYRNOVA-TRYBULSKA
University of Silesia, Katowice – Cieszyn, Poland

Olga YAKOVLEVA
Herzen State Pedagogical University of Russia, St. Petersburg, Russia

National steering committee:

Chair: **Karel Lepka**
Masaryk University in Brno, Faculty of Education, Department of Mathematics

Members:
Luboš Bauer
Masaryk University in Brno, Faculty of Economics and Administration, Department of Applied Mathematics and Informatics

Jaroslav Beránek
Masaryk University in Brno, Faculty of Education, Department of Mathematics

Milan Jirsa
University of Defence in Brno, Faculty of Military Technology, Department of Informatics and Cyber Operations

Tomáš Ráčil
University of Defence, Faculty of Military Technology, Department of Informatics and Cyber Operations

Each MITAV 2022 participant received printed collection of abstracts **MITAV 2022** with ISBN 978-80-7582-456-1. CD supplement of this printed volume contains all the accepted contributions of the conference.

Now, in autumn 2022, this **post-conference proceedings** were published, containing extended versions of selected MITAV 2022 contributions. The proceedings are published in English and contain extended versions of 10 selected conference papers. Published articles have been chosen from 17 conference papers and every article was once more reviewed.

Webpage of the MITAV conference:

https://mitav.unob.cz

Content:

List of reviewers:

Leonid Berezansky

Hashem Bordbar

Irina Cristea

Alessandro Linzi

Marijan Marković

Božidar Popović

Zdeněk Šmarda

REMARKS ON $M_{p,g}$ SUMMABILITY AND $\mathcal{I}_c^g$-CONVERGENCE OF SEQUENCES OF REAL NUMBERS

Vladimír Baláž, Alexander Maťašovský and Tomáš Visnyai
Faculty of Chemical and Food Technology, Slovak University of Technology in Bratislava
Radlinského 9, 812 37 Bratislava, Slovakia, vladimir.balaz@stuba.sk,
alexander.matasovsky@stuba.sk, tomas.visnyai@stuba.sk

Abstract: *The aim of the article is to show a new type of summability of sequences of real numbers. Its properties and connections with the $\mathcal{I}_c^g$-convergence are shown, where $\mathcal{I}_c^g$ is a special type of ideal that is generated by a real function g.*

Keywords: sequences of real numbers, convergence, ideal, summability.

INTRODUCTION

We recall the basic definitions and notations that will be used throughout the paper. Let $\mathbb{N}$ be the set of all positive integers, $\mathbb{N}_0 = \mathbb{N} \cup \{0\}$, and $\mathbb{R}^+$ be the set of all positive real numbers. A system $\mathcal{I}$, $\emptyset \neq \mathcal{I} \subseteq 2^{\mathbb{N}}$ is called an ideal, provided that $\mathcal{I}$ is additive ($A, B \in \mathcal{I}$ implies $A \cup B \in \mathcal{I}$) and hereditary ($A \in \mathcal{I}$, $B \subset A$ implies $B \in \mathcal{I}$). The ideal is called nontrivial if $\mathcal{I} \neq 2^{\mathbb{N}}$. If $\mathcal{I}$ is a nontrivial ideal, then $\mathcal{I}$ is called admissible if it contains the singletons ($\{n\} \in \mathcal{I}$ for every $n \in \mathbb{N}$). The fundamental notation which we shall use is $\mathcal{I}$-convergence introduced in the paper [9] (see also [3] where $\mathcal{I}$-convergence is defined by means of the dual notion of an ideal so-called filter). The notion $\mathcal{I}$-convergence corresponds to the natural generalization of the notion of statistical convergence (see [1], [2], [5], [6], [13]).

Definition 1. *Let $x = (x_n)$ be a sequence of real (complex) numbers. We say that the sequence $\mathcal{I}$-converges to a number L, and write $\mathcal{I} - \lim x_n = L$, if for each $\varepsilon > 0$ the set $A_\varepsilon = \{n : |x_n - L| \geq \varepsilon\}$ belongs to the ideal $\mathcal{I}$.*

In the following, we suppose that $\mathcal{I}$ is an admissible ideal. Then for every sequence (x_n) we immediately have that $\lim_{n\to\infty} x_n = L$ (classic limit) implies that (x_n) also $\mathcal{I}$-converges to the same number L but the opposite is not true. In other words, for an admissible ideal $\mathcal{I}$ we have $\mathcal{I}_{fin} \subseteq \mathcal{I}$, where $\mathcal{I}_{fin}$ is the ideal of all finite subsets of $\mathbb{N}$ and $\mathcal{I}_{fin}$ convergence coincides with the usual convergence.
Let $\mathcal{I}_d = \{A \subseteq \mathbb{N} : d(A) = 0\}$, where $d(A)$ is the asymptotic density of $A \subseteq \mathbb{N}$. The numbers $\underline{d}(A) = \liminf_{n\to\infty} \frac{\#\{a \leq n : a \in A\}}{n}$ and $\overline{d}(A) = \limsup_{n\to\infty} \frac{\#\{a \leq n : a \in A\}}{n}$ are called the lower and upper asymptotic density of the set A, respectively, where $\#M$ denotes the cardinality of the set M. If $\underline{d}(A) = \overline{d}(A) = d(A)$ then $d(A)$ is said to be the asymptotic density of A. The usual $\mathcal{I}_d$-convergence is called statistical convergence (see [5] and [13]). For $0 < q \leq 1$ the ideal $\mathcal{I}_c^{(q)} = \left\{A \subset \mathbb{N} : \sum_{a\in A} a^{-q} < \infty\right\}$ is an admissible ideal. The ideal $\mathcal{I}_c^{(1)} = \left\{A \subset \mathbb{N} : \sum_{a\in A} \frac{1}{a} < \infty\right\}$ is usually denote by $\mathcal{I}_c$ (see [2] and [8]).

$\mathcal{I}$-convergence satisfies usual axioms of convergence i.e. the uniqueness of limit, arithmetical properties etc. The class of all $\mathcal{I}$-convergent sequences is a linear space (see [9]). In [8] was proved the necessary and sufficient condition for the equivalence between the $\mathcal{I}_c^{(q)}$-convergence and the matrix method of summability however, a special class of regular matrices is needed (see [7]).

1 DEFINITIONS AND NOTIONS

Definition 2. *Let $g\colon \mathbb{R}^+ \to \mathbb{R}^+$ be a real function such that $\sum_{n=1}^{\infty} \frac{1}{g(n)} = +\infty$. Then we can define an ideal $\mathcal{I}_c^g = \left\{A \subset \mathbb{N} \,:\, \sum_{n\in A} \frac{1}{g(n)} < +\infty\right\}$. The ideal $\mathcal{I}_c^g$ is an admissible ideal.*

If $g(n) = c$, where $c \in \mathbb{R}$ then the ideal $\mathcal{I}_c^g$ contains only finite sets, hence $\mathcal{I}_c^g = \mathcal{I}_{fin}$. Next if $g(n) = n$ (we can write $g(x) = x$), then $\mathcal{I}_c^g = \mathcal{I}_c$ and finally if we take $g(x) = x^q,\ q \in (0, 1\rangle$ then the ideal $\mathcal{I}_c^g = \mathcal{I}_c^{(q)}$.

Let us denote s the set of all sequences of real numbers and let $s_1 \subset s$. The map $T\colon s_1 \to s$ is called a linear transformation if for all $x, y \in s_1$ and $a \in \mathbb{R}$ such that $x + y \in s_1$ and $ax \in s_1$ we have that the map T satisfies the following two conditions:

(i) $T(x + y) = Tx + Ty$ (additivity),

(ii) $T(ax) = a.Tx$ (homogeneity),

where Tx is also a sequence of real numbers.

The linear transformation T is regular if its convergence field $\mathcal{F}(T) = \{x = (x_n) \,:\, T - \lim x_n \in \mathbb{R}\}$ contains all convergent sequences and, moreover, the T-limit of a convergent sequence and its limit (in the usual sense) are the same, i.e. $T - \lim x_n = \lim_{n\to\infty} x_n \in \mathbb{R}$ (see [4] and [11]).

Recall the axioms of convergence (see [10]).

(S) Every constant sequence $(\xi, \xi, \dots, \xi, \dots)$ converges to ξ.

(H) The limit of any convergent sequence is uniquely determined.

(F) If a sequence $x = (x_n)$ has the limit ξ, then each of its subsequence has the same limit.

(U) If each subsequence of the sequence $x = (x_n)$ has a subsequence which converges to ξ, then $x = (x_n)$ converges to ξ.

It is well known that $\mathcal{I}$-convergence satisfies all axioms except for axiom (F) (see [9]). In the next part, we will focus on a certain summable method which can be defined as follows.

Definition 3. *Let $p > 0$ and $g\colon \mathbb{R}^+ \to \mathbb{R}^+$. We say that the sequence $x = (x_k) \in \ell_\infty$ is $M_{p,g}$-summable to the real number L (and write $M_{p,g} - \lim x_k = L$) if*

$$K = \sum_{k=1}^{\infty} \frac{|x_k - L|^p}{g(k)} < +\infty.$$

There is a natural question, is the convergence field of the method $M_{p,g}$ equal to the set of all bounded sequences? The next example shows that all bounded real sequences are not $M_{p,g}$-summable.

Example 4. Consider the function $g(x) = x^q$ where $q \in (0,1\rangle$ and $p > 0$. Define a sequence $x = (x_k)$ as follows

$$x_k = \begin{cases} 1 & \text{if } k \text{ is even,} \\ 0 & \text{if } k \text{ is odd.} \end{cases}$$

From the definition of $M_{p,g}$-summability follows that

$$\begin{aligned} K = \sum_{k=1}^{\infty} \frac{|x_k - L|^p}{k^q} &= \sum_{n=1}^{\infty} \frac{|L|^p}{(2n)^q} + \sum_{n=1}^{\infty} \frac{|1-L|^p}{(2n-1)^q} \\ &= L^p \sum_{n=1}^{\infty} \frac{1}{(2n)^q} + |1-L|^p \sum_{n=1}^{\infty} \frac{1}{(2n-1)^q} \\ &= +\infty \end{aligned}$$

because $q \in (0,1\rangle$. This means that the sequence $x = (x_k)$ of zeroes and ones is not $M_{p,g}$-summable.

It is easy to show that the summable method $M_{p,g}$ for $p > 0$ and a real function $g\colon \mathbb{R}^+ \to \mathbb{R}^+$ such that $\sum_{n\in\mathbb{N}} \frac{1}{g(n)} = +\infty$ is a linear transformation. We will show that $M_{p,g}$ summability of sequences implies $\mathcal{I}_c^g$-convergence to the same real number.

Theorem 5. *Let $p > 0$ and $g\colon \mathbb{R}^+ \to \mathbb{R}^+$ such that $\sum_{n=1}^{\infty} \frac{1}{g(n)} = +\infty$. If the sequence $x = (x_k)$ is $M_{p,g}$-summable to $L \in \mathbb{R}$, then $\mathcal{I}_c^g - \lim x_k = L$.*

Proof. Let $\varepsilon > 0$ and $g\colon \mathbb{R}^+ \to \mathbb{R}^+$ be a positive real function. Denote $A_{(\varepsilon)} = \{k \in \mathbb{N} : |x_k - L| \geq \varepsilon\}$. Then we have

$$K = \sum_{k=1}^{\infty} \frac{|x_k - L|^p}{g(k)} \geq \varepsilon^p \sum_{k\in A_{(\varepsilon)}} \frac{1}{g(k)}.$$

From this we get the inequality

$$\sum_{k\in A_{(\varepsilon)}} \frac{1}{g(k)} \leq \frac{K}{\varepsilon^p} < +\infty,$$

which means that $A_{(\varepsilon)} \in \mathcal{I}_c^g$, hence $\mathcal{I}_c^g - \lim x_k = L$. □

Moreover, we can ask whether the opposite of the previous theorem is true? Again, we give a negative answer.

Example 6. Let $p > 0$ and $g(x) = x^q,\ q \in (0,1\rangle$. Define a sequence $x = (x_k)$ as follows

$$x = (x_k) = \left(\frac{1}{(\log(k+1))^{\frac{1}{p}}} \right),\ k = 1, 2, \ldots .$$

It is clear that $\lim_{k\to\infty} x_k = 0$. Therefore $\mathcal{I}_c^g - \lim x_k = 0$. Calculate

$$K = \sum_{k=1}^{\infty} \frac{|x_k - 0|^p}{k^q} = \sum_{k=1}^{\infty} \frac{\left(\frac{1}{(\log(k+1))^{\frac{1}{p}}}\right)^p}{k^q} \geq \sum_{k=1}^{\infty} \frac{1}{k\log(k+1)} = +\infty$$

for $q \in (0,1\rangle$, therefore $M_{p,g} - \lim x_k \neq 0$.

This example shows much more. It is easy to show that the sequence $x = (x_n)$ is not $M_{p,g}$-summable to any real number L.

The previous example also showed that the $M_{p,g}$-summability is not a regular method i.e. it does not preserve a classical limit. So, we have the method of summability, which is not regular, but implies $\mathcal{I}_c^g$-convergence of sequences of real numbers to the same limit.

In the end, we give a result on how $M_{p,g}$-summability is related to the axioms of convergence.

Theorem 7. *Let $p > 0$, $g\colon \mathbb{R}^+ \to \mathbb{R}^+$ be a positive function such that $\sum_{n=1}^{\infty} \frac{1}{g(n)} = +\infty$. Then*

(i) $M_{p,g}$-summability has the properties (S) and (H),

(ii) $M_{p,g}$-summability does not have the properties (F) and (U).

Proof. (i) It is clear that the axiom (S) holds. Let $g(x) = x^q$. Suppose that $M_{p,g} - \lim x_k = \xi$ and simultaneously $M_{p,g} - \lim x_k = \mu$ where $\xi \neq \mu$. Choose $\varepsilon \in \left(0, \frac{|\xi-\mu|}{2}\right)$ and denote $A_{(\varepsilon)} = \{k \in \mathbb{N} \,:\, |x_k - \xi| \geq \varepsilon\}$ and $B_{(\varepsilon)} = \{k \in \mathbb{N} \,:\, |x_k - \mu| \geq \varepsilon\}$. Then $\overline{d}\left(A_{(\varepsilon)}\right) > 0$ or $\overline{d}\left(B_{(\varepsilon)}\right) > 0$, where $\overline{d}$ is the upper asymptotic density. Let $\overline{d}\left(A_{(\varepsilon)}\right) > 0$, then on the basis of article [12] we have $\sum_{k\in A_{(\varepsilon)}} \frac{1}{k} = +\infty$. Next

$$K = \sum_{k=1}^{\infty} \frac{|x_k - \xi|^p}{k^q} \geq \varepsilon^p \sum_{k\in A_{(\varepsilon)}} \frac{1}{k^q} \geq \varepsilon^p \sum_{k\in A_{(\varepsilon)}} \frac{1}{k} = +\infty.$$

This is in contradiction with the assumption of $M_{p,g}$-convergence of the sequence $x = (x_k)$.

(ii) We show that the $M_{p,g}$-summability does not satisfy the axiom (F). Let $A = \{n! \,:\, n \in \mathbb{N}\}$ and $g(x) = x^q,\ q \in (0,1\rangle$. Define a sequence $x = (x_k)$ as it follows:

$$x_k = \begin{cases} 1 & \text{if } k \in A, \\ 0 & \text{if } k \in \mathbb{N} \setminus A. \end{cases}$$

Then

$$\sum_{k=1}^{\infty} \frac{|x_k - 0|^p}{k^q} = \sum_{k\in A} \frac{1}{k^q} = \sum_{n=1}^{\infty} \frac{1}{(n!)^q} < +\infty.$$

So the sequence $x = (x_k)$ is $M_{p,g}$-summable to zero. But the sequence $y = (y_n)$, where $y_n = x_k$, $k = n!$ is a constant sequence $y_n = 1$, $n = 1, 2, \ldots$, hence $M_{p,g} - \lim y_n = 1$.

Now we are going to show that $M_{p,g}$-summability does not also satisfy the axiom (U). Take into consideration Example 6. Let $M = \{k_1 < k_2 < \cdots < k_n < \cdots\}$ be an arbitrary subsequence of $\mathbb{N}$. Put $y_n = x_{k_n}$, $n = 1, 2, \ldots$. In this way, we get a subsequence of the sequence $x = (x_k)$. Since $\lim_{k\to\infty} x_k = 0$ we can choose from every subsequence of x another subsequence that has the same limit. Take $y_n = x_{k_n}$. For $r = 1$ there exists $n_1 \in \mathbb{N}$ such that $y_{n_1} = x_{k_{n_1}} < 1$ then $z_1 = y_{n_1}$. For $r = 2$ there exists $n_2 \in \mathbb{N}$ $(n_2 > n_1)$ such that $y_{n_2} = x_{k_{n_2}} < \frac{1}{2}$ then $z_2 = y_{n_2}$ etc., for $r \in \mathbb{N}$ there exists $n_r \in \mathbb{N}$ $(n_r > n_{r-1})$ such that $y_{n_r} = x_{k_{n_r}} < \frac{1}{2^{r-1}}$ then $z_r = y_{n_r}$. By this construction we get the subsequence $z = (z_r)$, $r = 1, 2, \ldots$. Simple calculation gives $M_{p,g}$-summability of the sequence z to zero:

$$\sum_{r=1}^{\infty} \frac{|z_r - 0|^p}{r^q} = \sum_{r=1}^{\infty} \frac{\left|x_{k_{n_r}} - 0\right|^p}{k_{n_r}^q} \leq \sum_{r=1}^{\infty} \frac{\left(\frac{1}{2^{r-1}}\right)^p}{k_{n_r}^q} = \sum_{r=1}^{\infty} \frac{1}{2^{(r-1)p} k_{n_r}^q} < +\infty,$$

because $p > 0$ and $q \in (0, 1\rangle$. Therefore $M_{p,g} - \lim z_r = 0$, and the sequence $x = (x_k)$ is not $M_{p,g}$-summable (as it was shown in Example 6). We are finished the proof. □

CONCLUSION

We showed, that for the ideal $\mathcal{I}_c^g$, where g is a positive real function such that $\sum_{n=1}^{\infty} \frac{1}{g(n)} = +\infty$, from $M_{p,g}$-summability follows the $\mathcal{I}_c^g$-convergence of bounded sequences of real numbers to the same limit. We also showed some properties of $M_{p,g}$-summability e.g. axioms of convergence and a description of the convergence field of the given method.

References

[1] Vladimír Baláž. On generalized notion of convergence by means of ideal and its applications. *Mathematics, Information Technologies and Applied Sciences*, pages 9–20, 2017.

[2] Vladimír Baláž and Tomáš Visnyai. *$\mathcal{I}$-convergence of Arithmetical Functions*, chapter 8, pages 125–145. Number Theory and Its Applications. IntechOpen, London, 2020.

[3] Nicolas Bourbaki. *Eléments de Mathématique, Topologie Générale Livre III*. Nauka, Moscow, 1968. Translated to Russian under the title Obščaja topologija Osnovnye struktury.

[4] Richard G. Cooke. *Infinite matrices and sequence spaces*. Moscow, 1960.

[5] H. Fast. Sur la convergence statistique. *Colloquium Mathematicae*, 2(3-4):241–244, 1951.

[6] Rafał Filipów and Jacek Tryba. Ideal convergence versus matrix summability. *Studia Math*, 245(2):101–127, 2019.

[7] J. A. Fridy and H. I. Miller. A matrix characterization of statistical convergence. *Analysis*, 11(1):59–66, 1991.

[8] J. Gogola, M. Mačaj, and T. Visnyai. On $\mathcal{I}_c^{(q)}$-convergence. *Annales Mathematicae et Informaticae*, 38:27–36, 2011.

[9] Pavel Kostyrko, Tibor Šalát, and Władysław Wilczyński. $\mathcal{I}$-convergence. *Real Analysis Exchange*, 26(2):669–686, 2000.

[10] P. Mikusiński. Axiomatic theory of convergence. *Pr. Nauk. Uniw. ŚI Katow*, 12:13–21, 1982.

[11] Gordon Marshall Petersen. *Regular matrix transformations*. McGraw-Hill London, 1966.

[12] Barry J. Powell and Tibor Šalát. Convergence of subseries of the harmonic series and asymptotic densities of sets of positive integers. *Publications de l'Institut Mathématique. Nouvelle Série*, 50(64):60–70, 1991.

[13] I. J. Schoenberg. The integrability of certain functions and related summability methods. *The American Mathematical Monthly*, 66(5):361–775, 1959.

Acknowledgement

The author Vladimír Baláž wishes to thank The Slovak Research and Development Agency (research project VEGA No. 2/0119/23) for financial support.

The author Alexander Maťašovský wishes to thank The Slovak Research and Development Agency (research project VEGA No. 1/0386/21) for financial support.

BOUNDED SOLUTIONS OF A SYSTEM OF TWO DISCRETE EQUATIONS WITH COMPLEX EIGENVALUES OF THE MATRIX OF LINEAR TERMS

Jaromír Baštinec, Josef Diblík, Zuzana Piskořová
Brno University of Technology,
FEEC, Technická 8, 616 00 Brno, Czech Republic
bastinec@vutbr.cz, diblik@vut.cz, 155597@vut.cz

Abstract: *In the paper we consider a two-dimensional linear non-homogeneous system of discrete equations*

$$\begin{aligned} y_1(k+1) &= \quad py_1(k) + qy_2(k) + g_1(k), \\ y_2(k+1) &= -qy_1(k) + py_2(k) + g_2(k), \end{aligned}$$

where $k = a, a+1, \dots$ with a fixed integer $a \in \mathbb{N}$, p, q are real constants, $g_i \colon \{a, a+1, \dots\} \to \mathbb{R}$, $i = 1, 2$ are given functions. Sufficient conditions are derived guaranteeing the existence of a solution $y(k) = (y_1(k), y_2(k))$, $k = a, a+1, \dots$ satisfying $y_1^2(k) + y_2^2(k) < M$, where M is a given positive constant.

Keywords: bounded solution, linear discrete system, retract principle, mapping.

1 INTRODUCTION

This paper is concerned with a system of linear non-homogeneous discrete equations

$$y_1(k+1) = \quad py_1(k) + qy_2(k) + g_1(k), \tag{1}$$
$$y_2(k+1) = -qy_1(k) + py_2(k) + g_2(k), \tag{2}$$

where $k \in \mathbb{N}(a) := \{a, a+1, a+2, \dots\}$, $a \in \mathbb{N} := \{1, 2, \dots\}$ is a fixed integer, p and q are real constants, $q \neq 0$ and $g_i \colon \mathbb{N}(a) \to \mathbb{R}$, are given functions. Let M be a positive constant. In the paper we indicate sufficient conditions for the existence of at least one solution $y(k) = (y_1(k), y_2(k))$, $k \in \mathbb{N}(a)$ satisfying the inequality

$$y_1^2(k) + y_2^2(k) < M, \ \ \forall k \in \mathbb{N}(a). \tag{3}$$

The existence of bounded (within the meaning of various definitions) solutions for scalar, planar and n-dimensional linear and non-linear discrete systems has been considered, e.g., in [2, 3, 4, 5, 6, 8, 9, 10, 11, 12, 13, 16]. System (1), (2) is a particular case of a linear non-homoneneous system

$$y_1(k+1) = a_{11}y_1(k) + a_{12}y_2(k) + g_1(k), \tag{4}$$
$$y_2(k+1) = a_{21}y_1(k) + a_{22}y_2(k) + g_2(k), \tag{5}$$

where a_{ij}, $i, j = 1, 2$ are real numbers, $k \in \mathbb{N}(a)$. Set

$$A = \begin{pmatrix} a_{11} & a_{12} \\ a_{21} & a_{22} \end{pmatrix}, \ \ g(k) = \begin{pmatrix} g_1(k) \\ g_2(k) \end{pmatrix}, \ \ y(k) = \begin{pmatrix} y_1(k) \\ y_2(k) \end{pmatrix}.$$

Then, the system $(4), (5)$ can be rewritten in a vector-matrix form

$$y(k+1) = Ay(k) + g(k). \tag{6}$$

Let us transform the system (6) by a change of variables $y(k) = Sz(k)$, where $z(k)$ is a new dependent variable and the transformation matrix S is a 2 by 2 nonsingular matrix. Then, the system (6) is transformed into a system

$$z(k+1) = Jy(k) + f(k) \tag{7}$$

where $J = S^{-1}AS$ and $f(k) = S^{-1}g(k)$. If S is properly chosen, then the matrix A can be transformed into its Jordan form. For 2 by 2 matrices, Jordan forms $J = J_i$, $i = 1, 2, 3, 4$ are possible where

$$J_1 = \begin{pmatrix} \lambda_1 & 0 \\ 0 & \lambda_2 \end{pmatrix} \tag{8}$$

if the matrix A has two different real eigenvalues λ_1, λ_2,

$$J_2 = \begin{pmatrix} \lambda & 0 \\ 0 & \lambda \end{pmatrix}, \quad J_3 = \begin{pmatrix} \lambda & 1 \\ 0 & \lambda \end{pmatrix} \tag{9}$$

if the matrix A has one real eigenvalue λ with geometrical multiplicity equaling 2 or 1, and

$$J_4 = \begin{pmatrix} p & q \\ -q & p \end{pmatrix} \tag{10}$$

if the matrix A has two complex conjugate eigenvalues $\lambda = p \pm iq$ where i is the complex unit.

Although the above system (6) is solvable, to establish the existence of a bounded solution may be a daunting task. We will demonstrate this on the example of a scalar equation

$$y(k+1) = 3y(k) + \frac{1}{k+1} - \frac{3}{k} \tag{11}$$

having a bounded solution $y(k) = 1/k$. By the well-known formula for the general solution to equation (11), we have

$$y(k) = 3^{k-a}y(a) + \sum_{i=a}^{k-a} 3^{k-a}\left(\frac{1}{k+1} - \frac{3}{k}\right)$$

and it is not clear how this formula could be used to deduce the existence of a bounded solution.

Also, the above mentioned known results are either not applicable in principle or are not capable of solving the problem for system (1), (2). Let us mention at least the recent paper [5] where so-called triangular systems are considered. The results of this paper are applicable to systems (7) with $J = J_1$, $J = J_2$ or $J = J_3$ but not so if $J = J_4$.

2 PRELIMINARIES

In the paper we will apply a result published in [10]. Below we give its short description. Consider a system of nonlinear discrete equations

$$u(k+1) = F(k, u(k)) \tag{12}$$

where $u = (u_1, \ldots, u_n)$, $F\colon \mathcal{N}(a) \times \mathbb{R}^n \to \mathbb{R}^n$, and $k \in \mathcal{N}(a)$ is the independent variable. Consider an initial problem

$$u(s) = u^s, \tag{13}$$

where $s \in N(a)$ and $u^s \in \mathbb{R}^n$ is fixed. A solution $u = u(k)$, $k \in N(s)$ of the initial problem (12), (13) is defined as an infinite sequence

$$u(s) = u^s, u(s+1), u(s+2), \ldots$$

such that, for any $k \in N(s)$, equality (12) holds. The solution of initial problem (12), (13) exists and is unique. Below we assume that the vector $F(k, u)$ is continuous with respect to argument u. Then, the initial problem (12), (13) depends continuously on the initial data.

Let a set $\Omega(k)$, $k \in N(a)$, be an n-dimensional open bounded and simply connected subset of the set

$$S(k) := \{(k, u)\colon u \in \mathbb{R}^n\}.$$

Due to the above-formulated properties, every set $\Omega(k)$, $k \in N(a)$ is topologically equivalent to an n-dimensional open ball in $\mathbb{R}^n$. The boundary $\partial\Omega(k)$ of $\Omega(k)$ is defined in the space $S(k)$ in the usual way, as well as the closure $\overline{\Omega}(k) = \Omega(k) \cup \partial\Omega(k)$.

Definition 2.1 *Let a point $M = (k, u^0) \in S(k)$ with a fixed $k \in N(a)$ be given. The point $M^c = (k+1, F(k, u^0))$ is called the first consequent point to the point M and denoted by $M^c = \mathcal{C}[M]$.*

Definition 2.2 *Let a set $\mathcal{S} \subset S(k)$ with a fixed $k \in N(a)$ be given. We say that a set $\mathcal{S}^c$ is the first consequent set to the set $\mathcal{S}$ if $\mathcal{S}^c := \{M^c, M \in \mathcal{S}\}$ and write $\mathcal{S}^c = \mathcal{C}[\mathcal{S}]$.*

Define a mapping $\mathcal{F}\colon N(a) \times \mathbb{R}^n \to N(a) \times \mathbb{R}^n$ by the formula $\mathcal{F}(k, u) = (k+1, F(k, u))$.

Theorem 2.3 *Let, for every fixed $s \in N(a)$, the mapping $\mathcal{F}\colon \partial\Omega(s) \to \mathcal{C}[\partial\Omega(s)]$ be bijective. Suppose that, for every fixed $s \in N(a)$, the set $\mathcal{C}[\partial\Omega(s)]$ is the boundary of an n-dimensional closed domain $\mathcal{D}(s+1)$, homeomorphic with n-dimensional closed ball such that*

$$\overline{\Omega}(s+1) \subset \mathcal{D}(s+1) \quad \textit{and} \quad \overline{\Omega}(s+1) \cap \partial\mathcal{D}(s+1) = \emptyset. \tag{14}$$

Then, there exists at least one initial point $u_*(a) = u_*^a$ with $(a, u_*^a) \in \Omega(a)$ such that the solution $u = u_*(k)$, $k \in N(a)$ of system (12) *satisfies*

$$(k, u_*(k)) \in \Omega(k) \tag{15}$$

for every $k \in N(a)$.

3 BOUNDED SOLUTIONS TO SYSTEM (1), (2)

In this part we consider system (1), (2) and give sufficient conditions for the existence of at least one bounded solution $y(k) = (y_1(k), y_2(k))^T$, $k \in \mathcal{N}(a)$ such that

$$y_1^2(k) + y_2^2(k) < M, \ \ \forall k \in \mathcal{N}(a) \tag{16}$$

where M is a fixed positive number. We use the below notation

$$w(k) := y_1^2(k) + y_2^2(k) - M, \ \ k \in \mathcal{N}(a), \tag{17}$$

$$a(k) := qg_1(k) + pg_2(k), \quad b(k) := pg_1(k) - qg_2(k), \quad k \in \mathcal{N}(a),$$
$$c := p^2 + q^2,$$

and

$$D(k) := \left(\frac{a(k)}{c} + \frac{a^3(k)}{c}\frac{1}{b^2(k)} + M\frac{a(k)c}{b^2(k)} - M\frac{a(k)}{b^2(k)}\right)^2 - 2\left(\left(\frac{a(k)}{b(k)}\right)^2 + 1\right)$$
$$\cdot\left(\frac{1}{4}\left(\frac{b(k)}{c}\right)^2 + \frac{1}{4}\left(\frac{a^2(k)}{b(k)c}\right)^2 + \left(\frac{1}{2}M\frac{c}{b(k)}\right)^2 + \left(\frac{1}{2}\frac{M}{b(k)}\right)^2\right.$$
$$\left. + \frac{a^2(k)}{c^2} + M - \frac{M}{c} + M\frac{a^2(k)}{b^2(k)} - \frac{a^2(k)}{c}\frac{M}{b^2(k)} - M^2\frac{c}{b^2(k)} - 2M\right).$$

Theorem 3.1 *Let $M > 0$ be a fixed number and, for every $k \in \mathcal{N}(a)$, $b(k) \neq 0$, $D(k) < 0$,*

$$(p^2 + q^2)M + g_1^2(k) + g_2^2(k) > 2\sqrt{M}(|a(k)| + |b(k)|) + M, \tag{18}$$

and

$$a^2(k) + b^2(k) < Mc. \tag{19}$$

Then, the system (1), (2) *has at least one solution $y(k) = (y_1(k), y_2(k))^T$, $k \in \mathcal{N}(a)$ satisfying inequality* (16).

PROOF. The proof is divided into several parts below. Define, for every $k \in \mathcal{N}(a)$, a circle

$$\mathbb{S}_1(k) := \{(k, y_1, y_2) \colon y_1 \in \mathbb{R}, y_2 \in \mathbb{R}, y_1^2 + y_2^2 = M\}$$

as the boundary of a disc

$$\mathbb{D}_1(k) := \{(k, y_1, y_2) \colon y_1 \in \mathbb{R}, y_2 \in \mathbb{R}, y_1^2 + y_2^2 \leq M\}.$$

$i)$ Mapping of the circle $\mathbb{S}_1(k)$ by system (1), (2). To map the circle $\mathbb{S}_1(k)$, let us imagine that the points are initial generating the solutions to system (1), (2). In other words, we will discuss the properties of solutions $(y_1(k), y_2(k))^T$ to system (1), (2) such that their starting points satisfy the restriction $w(k) = 0$. For it, consider the value $w(k+1)$ and show that $w(k+1) > 0$. By our computation, we have

$$w(k+1) = y_1^2(k+1) + y_2^2(k+1) - M$$

$$
\begin{aligned}
&= (py_1(k) + qy_2(k) + g_1(k))^2 + (-qy_1(k) + py_2(k) + g_2(k))^2 - M \\
&= p^2y_1^2(k) + q^2y_2^2(k) + g_1^2(k) + 2pqy_1(k)y_2(k) + 2py_1(k)g_1(k) + 2qy_2(k)g_1(k) \\
&\quad + q^2y_1^2(k) + p^2y_2^2(k) + g_2^2(k) - 2pqy_1(k)y_2(k) - 2qy_1(k)g_2(k) + 2py_2(k)g_2(k) - M \\
&= p^2y_1^2(k) + q^2y_2^2(k) + g_1^2(k) + 2py_1(k)g_1(k) + 2qy_2(k)g_1(k) \\
&\quad + q^2y_1^2(k) + p^2y_2^2(k) + g_2^2(k) - 2qy_1(k)g_2(k) + 2py_2(k)g_2(k) - M \\
&= p^2(y_1^2(k) + y_2^2(k)) + q^2(y_2^2(k) + y_1^2(k)) + g_1^2(k) + g_2^2(k) \\
&\quad + 2py_1(k)g_1(k) + 2qy_2(k)g_1(k) - 2qy_1(k)g_2(k) + 2py_2(k)g_2(k) - M \\
&= (p^2 + q^2)(y_2^1(k) + y_2^2(k)) + g_1^2(k) + g_2^2(k) + 2y_1(k)(pg_1(k) - qg_2(k)) \\
&\quad + 2y_2(k)(qg_1(k) + pg_2(k)) - M = (*)
\end{aligned}
$$

Because $y_1^2(k) + y_2^2(k) = M$, we have $|y_1(k)| \le \sqrt{M}$ and $|y_2(k)| \le \sqrt{M}$. The inequality $w(k+1) > 0$ will hold if

$$
(*) \ge (p^2 + q^2)M + g_1^2(k) + g_2^2(k) - 2\sqrt{M}(|a(k)| + |b(k)|) - M > 0.
$$

The last inequality holds being equivalent with the assumption (18). The relation $w(k+1) = 0$, as it follows from the analysis of the expression $(*)$, represents a circle in the plane

$$
\mathbb{P}(k+1) := \{(k+1, y_1, y_2) \colon y_1 \in \mathbb{R}, y_2 \in \mathbb{R}\}.
$$

Denote this circle by $\mathbb{S}_2(k+1)$. Below we find its canonical form, the centre and radius. Equation $w(k+1) = 0$ is equivalent with

$$
y_1^2(k) + y_2^2(k) + \frac{g_1^2(k) + g_2^2(k)}{p^2 + q^2} + 2y_1(k)\frac{pg_1(k) - qg_2(k)}{p^2 + q^2} + 2y_2(k)\frac{qg_1(k) + pg_2(k)}{p^2 + q^2} - \frac{M}{p^2 + q^2} = 0.
$$

Modifying the left-hand side of this equality we derive

$$
\begin{aligned}
&y_1^2(k) + 2y_1(k)\frac{pg_1(k) - qg_2(k)}{p^2 + q^2} + y_2^2(k) + 2y_2(k)\frac{qg_1(k) + pg_2(k)}{p^2 + q^2} + \frac{g_1^2(k) + g_2^2(k)}{p^2 + q^2} \\
&\quad - \frac{M}{p^2 + q^2} \\
&= \left(y_1(k) + \frac{pg_1(k) - qg_2(k)}{p^2 + q^2}\right)^2 - \left(\frac{pg_1(k) - qg_2(k)}{p^2 + q^2}\right)^2 + \left(y_2(k) + \frac{qg_1(k) + pg_2(k)}{p^2 + q^2}\right)^2 \\
&\quad - \left(\frac{qg_1(k) + pg_2(k)}{p^2 + q^2}\right)^2 + \frac{g_1^2(k) + g_2^2(k)}{p^2 + q^2} - \frac{M}{p^2 + q^2} \\
&= \left(y_1(k) + \frac{pg_1(k) - qg_2(k)}{p^2 + q^2}\right)^2 + \left(y_2(k) + \frac{qg_1(k) + pg_2(k)}{p^2 + q^2}\right)^2 + \frac{g_1^2(k) + g_2^2(k)}{p^2 + q^2} \\
&\quad - \left(\frac{pg_1(k) - qg_2(k)}{p^2 + q^2}\right)^2 - \left(\frac{qg_1(k) + pg_2(k)}{p^2 + q^2}\right)^2 - \frac{M}{p^2 + q^2} \\
&= \left(y_1(k) + \frac{pg_1(k) - qg_2(k)}{p^2 + q^2}\right)^2 + \left(y_2(k) + \frac{qg_1(k) + pg_2(k)}{p^2 + q^2}\right)^2 + \frac{g_1^2(k) + g_2^2(k)}{p^2 + q^2}
\end{aligned}
$$

$$
\begin{aligned}
&- \left(\frac{p^2 g_1^2(k) - 2pg_1(k)qg_2(k) + q^2 g_2^2(k)}{(p^2+q^2)^2}\right) - \left(\frac{q^2 g_1^2(k) + 2qg_1(k)pg_2(k)) + p^2 g_2^2(k)}{(p^2+q^2)^2}\right) \\
&- \frac{M}{p^2+q^2} \\
&= \left(y_1(k) + \frac{pg_1(k) - qg_2(k)}{p^2+q^2}\right)^2 + \left(y_2(k) + \frac{qg_1(k) + pg_2(k)}{p^2+q^2}\right)^2 + \frac{g_1^2(k) + g_2^2(k)}{p^2+q^2} \\
&- \frac{p^2 g_1^2(k) - 2pg_1(k)qg_2(k) + q^2 g_2^2(k) + q^2 g_1^2(k) + 2qg_1(k)pg_2(k)) + p^2 g_2^2(k)}{(p^2+q^2)^2} - \frac{M}{p^2+q^2} \\
&= \left(y_1(k) + \frac{pg_1(k) - qg_2(k)}{p^2+q^2}\right)^2 + \left(y_2(k) + \frac{qg_1(k) + pg_2(k)}{p^2+q^2}\right)^2 + \frac{g_1^2(k) + g_2^2(k)}{p^2+q^2} \\
&- \frac{(p^2+q^2)(g_1^2(k) + g_2^2(k))}{(p^2+q^2)^2} - \frac{M}{p^2+q^2}\,.
\end{aligned}
$$

Therefore

$$
w(k+1) = \left(y_1(k) + \frac{pg_1(k) - qg_2(k)}{p^2+q^2}\right)^2 + \left(y_2(k) + \frac{qg_1(k) + pg_2(k)}{p^2+q^2}\right)^2 - \frac{M}{p^2+q^2}
$$

and the equation $w(k+1) = 0$ defines the circle $\mathbb{S}_2(k+1)$,

$$
\begin{aligned}
\mathbb{S}_2(k+1) := \Bigg\{ &(k+1, y_1, y_2) \colon y_1 \in \mathbb{R}, y_2 \in \mathbb{R}, \\
&\left(y_1 + \frac{pg_1(k) - qg_2(k)}{p^2+q^2}\right)^2 + \left(y_2 + \frac{qg_1(k) + pg_2(k)}{p^2+q^2}\right)^2 - \frac{M}{p^2+q^2} = 0 \Bigg\}. \quad (20)
\end{aligned}
$$

The circle (20) has the centre $\mathbb{C}_2(k+1)$ at the point

$$
\mathbb{C}_2(k+1) = \left(k+1, -\frac{pg_1(k) - qg_2(k)}{p^2+q^2}\,, -\frac{qg_1(k) + pg_2(k)}{p^2+q^2}\right)
$$

and its radius r_2 equals

$$
r_2 = \sqrt{\frac{M}{p^2+q^2}}\,.
$$

Using the previously defined symbols $a(k)$, $b(k)$ and c we can abbreviate

$$
w(k+1) = \left(y_1(k) + \frac{b(k)}{c}\right)^2 + \left(y_2(k) + \frac{a(k)}{c}\right)^2 - \frac{M}{c}\,,
$$

with equation (20) of the circle $\mathbb{S}_2(k+1)$ being transformed into

$$
\mathbb{S}_2(k+1) = \left\{ (k+1, y_1, y_2) \colon y_1 \in \mathbb{R}, y_2 \in \mathbb{R}, \left(y_1 + \frac{b(k)}{c}\right)^2 + \left(y_2 + \frac{a(k)}{c}\right)^2 - \frac{M}{c} = 0 \right\}.
$$

and

$$\mathbb{C}_2(k+1) = \left(k+1, -\frac{b(k)}{c}\,, -\frac{a(k)}{c}\right), \;\; r_2 = \sqrt{\frac{M}{c}}\,.$$

$ii)$ On a relationship of two circles. Consider in the plane $\mathbb{P}(k+1)$ a relationship between the circle $\mathbb{S}_2(k+1)$ being the boundary of a disc

$$\mathbb{D}_2(k+1) := \left\{(k+1, y_1, y_2)\colon y_1 \in \mathbb{R}, y_2 \in \mathbb{R}, \left(y_1 + \frac{b(k)}{c}\right)^2 + \left(y_2 + \frac{a(k)}{c}\right)^2 - \frac{M}{c} \leq 0\right\}$$

and the circle $\mathbb{S}_1(k+1)$. Let us search for the points of their intersection provided they exist. Analyzing the equation

$$y_1^2 + y_2^2 - M = \left(y_1 + \frac{b(k)}{c}\right)^2 + \left(y_2 + \frac{a(k)}{c}\right)^2 - \frac{M}{c}, \;\; k \in N(a+1),$$

we derive

$$y_1^2 + y_2^2 - M = y_1^2 + 2y_1\frac{b(k)}{c} + \left(\frac{b(k)}{c}\right)^2 + y_2^2 + 2y_2\frac{a(k)}{c} + \left(\frac{a(k)}{c}\right)^2 - \frac{M}{c}\,,$$

and, finally, coordinates y_1, y_2 satisfy the equation

$$y_1 = -y_2\frac{a(k)}{b(k)} - \frac{1}{2}\frac{b(k)}{c} - \frac{1}{2}\frac{a^2(k)}{b(k)c} - \frac{1}{2}M\frac{c}{b(k)} + \frac{1}{2}\frac{M}{b(k)}\,. \tag{21}$$

From equation

$$y_1^2 + y_2^2 - M = 0,$$

where y_1 is expressed by (21), we obtain

$$\left(y_2\frac{a(k)}{b(k)} + \frac{1}{2}\frac{b(k)}{c} + \frac{1}{2}\frac{a^2(k)}{b(k)c} + \frac{1}{2}M\frac{c}{b(k)} - \frac{1}{2}\frac{M}{b(k)}\right)^2 + y_2^2(k) - M = 0,$$

and, after some computation,

$$\begin{aligned} y_2^2\left(\left(\frac{a(k)}{b(k)}\right)^2 + 1\right) + y_2\left(\frac{a(k)}{c} + \frac{a^3(k)}{cb^2(k)} + M\frac{a(k)c}{b^2(k)} - M\frac{a(k)}{b^2(k)}\right) \\ + \frac{1}{4}\left(\frac{b(k)}{c}\right)^2 + \frac{1}{4}\left(\frac{a^2(k)}{b(k)c}\right)^2 + \left(\frac{1}{2}M\frac{c}{b(k)}\right)^2 + \left(\frac{1}{2}\frac{M}{b(k)}\right)^2 + \frac{1}{2}\frac{a^2(k)}{c^2} \\ + \frac{1}{2}M - \frac{1}{2}\left(\frac{M}{c}\right) + \frac{1}{2}M\frac{a^2(k)}{b^2(k)} - \frac{1}{2}\frac{a^2(k)}{b^2(k)}\frac{M}{c} - \frac{1}{2}M^2\frac{c}{b^2(k)} - M = 0, \end{aligned} \tag{22}$$

Equation (22) is quadratic with respect to y_2 in the form

$$\alpha(k)y_2^2 + \beta(k)y_2 + \gamma(k) = 0 \tag{23}$$

with coefficients

$$\alpha(k) = \left(\frac{a(k)}{b(k)}\right)^2 + 1,$$

$$\beta(k) = \frac{a(k)}{c} + \frac{a^3(k)}{b^2(k)c} + M\frac{a(k)c}{b^2(k)} - M\frac{a(k)}{b^2(k)},$$

$$\begin{aligned}\gamma(k) =& \frac{1}{4}\left(\frac{b(k)}{c}\right)^2 + \frac{1}{4}\left(\frac{a^2(k)}{b(k)c}\right)^2 + \left(\frac{1}{2}M\frac{c}{b(k)}\right)^2 + \left(\frac{1}{2}\frac{M}{b(k)}\right)^2 \\ &+ \frac{1}{2}\frac{a^2(k)}{c^2} + \frac{1}{2}M - \frac{1}{2}\frac{M}{c} + \frac{1}{2}M\frac{a^2(k)}{b^2(k)} - \frac{1}{2}\frac{a^2(k)}{c}\frac{M}{b^2(k)} - \frac{1}{2}M^2\frac{c}{b^2(k)} - M.\end{aligned}$$

For the discriminant $D^*(k)$ of equation (23) we derive

$$\begin{aligned}D^*(k) =& \beta^2(k) - 4\alpha(k)\gamma(k) = \left(\frac{a(k)}{c} + \frac{a^3(k)}{b^2(k)c} + M\frac{a(k)c}{b^2(k)} - M\frac{a(k)}{b^2(k)}\right)^2 - 4\left(\left(\frac{a(k)}{b(k)}\right)^2 + 1\right) \\ &\cdot \left(\frac{1}{4}\left(\frac{b(k)}{c}\right)^2 + \frac{1}{4}\left(\frac{a^2(k)}{b(k)c}\right)^2 + \left(\frac{1}{2}M\frac{c}{b(k)}\right)^2 + \left(\frac{1}{2}\frac{M}{b(k)}\right)^2\right. \\ &\left. + \frac{1}{2}\frac{a^2(k)}{c^2} + \frac{1}{2}M - \frac{1}{2}\frac{M}{c} + \frac{1}{2}M\frac{a^2(k)}{b^2(k)} - \frac{1}{2}\frac{a^2(k)}{c}\frac{M}{b^2(k)} - \frac{1}{2}M^2\frac{c}{b^2(k)} - M\right).\end{aligned}$$

Obviously, $D^*(k) = D(k)$. Since, by the hypotheses of Theorem 3.1, we have $D(k) < 0$, the equation (23) has no real roots and the circles $\mathbb{S}_1(k+1)$, $\mathbb{S}_2(k+1)$ have no point of intersection. Let us clarify the mutual position of discs $\mathbb{D}_1(k+1)$ and $\mathbb{D}_2(k+1)$. We show that the disc $\mathbb{D}_1(k+1)$ is embedded into the disc $\mathbb{D}_2(k+1)$. As circles $\mathbb{S}_1(k+1)$ and $\mathbb{S}_2(k+1)$ have no points of intersection it is sufficient to prove that a point lies both in disc $\mathbb{D}_1(k+1)$ and disc $\mathbb{D}_2(k+1)$. For this test choose the point

$$O = (k+1, 0, 0).$$

Obviously, $O \in \mathbb{D}_1(k+1)$. If also $O \in \mathbb{D}_2(k+1)$, then, as it follows from the definition of $\mathbb{D}_2(k+1)$, formula (20), the inequality

$$\left(\frac{pg_1(k) - qg_2(k)}{p^2+q^2}\right)^2 + \left(\frac{qg_1(k) + pg_2(k)}{p^2+q^2}\right)^2 - \frac{M}{p^2+q^2} < 0$$

must hold. This inequality can be rewritten as

$$\frac{b^2(k)}{c^2} + \frac{a^2(k)}{c^2} - \frac{M}{c} < 0$$

and holds obviously due to hypothesis (19).

iii) Application of Theorem 2.3. To apply Theorem 2.3, set $n = 2$,

$$F(k, y) = (F_1(k, y_1, y_2), F_2(k, y_1, y_2)) = (py_1(k) + qy_2(k) + g_1(k), -qy_1(k) + py_2(k) + g_2(k)),$$

$$\mathcal{F}(k,y) = (k+1, F(k,y)) = (k+1, py_1(k) + qy_2(k) + g_1(k), -qy_1(k) + py_2(k) + g_2(k)),$$
$$\Omega(k) := \operatorname{int} \mathbb{D}_1(k), \quad \partial\Omega(k) = \mathbb{S}_1(k)$$

Then, $\mathcal{C}[\partial\Omega(k] = \mathbb{S}_2(k+1)$ and the mapping

$$\mathcal{F}\colon \partial\Omega(s) = \mathbb{S}_1(s) \to \mathcal{C}[\partial\Omega(s)] = \mathbb{S}_2(s+1)$$

is bijective because $\det J_4 = c \neq 0$. The set $\mathcal{C}[\partial\Omega(s)]$ is the boundary of a 2-dimensional closed domain $\mathcal{D}_2(s+1)$, homeomorphic with the 2-dimensional closed ball,

$$\left(\overline{\Omega}(s+1) = \mathbb{D}_1(s+1)\right) \subset (\mathcal{D}(s+1) = \mathbb{D}_2(s+1))$$

and

$$\left(\overline{\Omega}(s+1) = \mathbb{D}_1(s+1)\right) \cap (\partial\mathcal{D}(s+1) = \mathbb{S}_2(s+1)) = \emptyset.$$

All hypotheses of Theorem 2.3 hold. Therefore, there exists at least one initial point $y_*(a) = y_*^a$ with $(a, y_*^a) \in \mathbb{D}_1(a)$ such that the solution $y = y_*(k)$, $k \in N(a)$ of system (1), (2) satisfies

$$(k, u_*(k)) \in \Omega(k) = \operatorname{int} \mathbb{D}_1(k)$$

for every $k \in N(a)$, that is

$$(k, u_*(k)) \in \Omega(k) = \operatorname{int} \mathbb{D}_1(k) = \{(k, y_1, y_2)\colon y_1 \in \mathbb{R}, y_2 \in \mathbb{R}, y_1^2 + y_2^2 - M < 0\}.$$

The inequality

$$y_1^2 + y_2^2 - M < 0 \quad \forall k \in N(a)$$

is equivalent with inequality (16). □

4 EXAMPLE

Let the system (1), (2) be reduced to the following one:

$$y_1(k+1) = \quad 2y_1(k) + 2y_2(k) + 1, \tag{24}$$
$$y_2(k+1) = -2y_1(k) + 2y_2(k) - 1, \tag{25}$$

where $k \in \mathbb{N}(a)$ and $a \in \mathbb{N}$ is arbitrary fixed. We have $p = q = 2$, $g_1(k) \equiv 1$ and $g_2(k) \equiv -1$. In the considered case we have

$$a(k) := qg_1(k) + pg_2(k) = 0, \quad b(k) := pg_1(k) - qg_2(k) = 4, \quad k \in \mathcal{N}(a),$$

$$c := p^2 + q^2 = 8.$$

Set $M = 4$. Then

$$D(k) := \left(\frac{a(k)}{c} + \frac{a^3(k)}{c}\frac{1}{b^2(k)} + M\frac{a(k)c}{b^2(k)} - M\frac{a(k)}{b^2(k)}\right)^2 - 2\left(\left(\frac{a(k)}{b(k)}\right)^2 + 1\right)$$
$$\cdot \left(\frac{1}{4}\left(\frac{b(k)}{c}\right)^2 + \frac{1}{4}\left(\frac{a^2(k)}{b(k)c}\right)^2 + \left(\frac{1}{2}M\frac{c}{b(k)}\right)^2 + \left(\frac{1}{2}\frac{M}{b(k)}\right)^2\right.$$

$$+\frac{a^2(k)}{c^2}+M-\frac{M}{c}+M\frac{a^2(k)}{b^2(k)}-\frac{a^2(k)}{c}\frac{M}{b^2(k)}-M^2\frac{c}{b^2(k)}-2M\Bigg)$$
$$=-2\left(\frac{1}{4}\left(\frac{16}{64}\right)+\left(\frac{1}{4}\cdot\frac{16\cdot 64}{16}\right)+\left(\frac{1}{4}\cdot\frac{16}{16}\right)\right.$$
$$\left.+4-\frac{4}{8}-16\frac{8}{16}-8\right)=-\frac{61}{8}<0.$$

Moreover

$$(p^2+q^2)M+g_1^2(k)+g_2^2(k)=8\cdot 4+2=34,$$
$$2\sqrt{M}(|a(k)|+|b(k)|)+M=2\cdot 2\cdot 4+4=20,$$

and inequality (18) holds. Inequality (19) holds as well because

$$a^2(k)+b^2(k)=16<Mc=32.$$

All hypotheses of Theorem 3.1 are fulfilled. Therefore, the system (24), (25) has at least one solution $y(k)=(y_1(k),y_2(k))^T$, $k\in\mathcal{N}(a)$ satisfying inequality (16), i.e.

$$y_1^2(k)+y_2^2(k)<4,\ \ \forall k\in\mathcal{N}(a). \tag{26}$$

System (24), (25) is an autonomous one and it is easy to see that there exist a bounded constant solution

$$y_1(k)=-\frac{3}{5}\,,\quad y_2(k)=-\frac{1}{5},\ \ \forall k\in\mathcal{N}(a)$$

satisfying inequality (26).

5 CONCLUDING REMARKS

The paper proves the existence of a bounded solution to system of two difference equations (1), (2). This system is a particular case of system (4), (5) provided that the matrix A has two complex conjugate eigenvalues $\lambda=p\pm iq$ where p, q are real constants, $q\neq 0$ and i is the complex unit. Previous results (e.g. [4, 5]) are, in general, applicable to systems where the matrix A has two different real eigenvalues, or has one real eigenvalue with geometrical multiplicity equaling 2 or 1 while the case of A having two complex conjugate eigenvalues has not yet been considered. For other asymptotic investigations of the behavior of solutions to systems of discrete equations, we refer, e.g., to [1, 7, 14, 15, 17] and to the references therein.

References

[1] Agarwal, R. P., *Difference Equations and Inequalities. Theory, Methods and Applications*, 2nd edition, Monographs and Textbooks in Pure and Applied Mathematics, Marcel Dekker, New York, 2000.

[2] Baštinec, J., Diblík J., Determination of initial data generating solutions of Bernoulli's type difference equations with prescribed asymptotic behavior, *Proceedings of the Eighth International Conference on Difference Equations and Applications* (Chapman & Hall/CRC, Boca Raton, FL, 2005), pp. 39–49.

[3] Baštinec, J., Diblík, J., Korobko, E., Bounded solutions of a triangular system of two nonlinear discrete equations, *AIP Conference Proceedings* **2425**, 270008-1–270008-4 (2022); https://doi.org/10.1063/5.0081826, International Conference of Numerical Analysis and Applied Mathematics ICNAAM 2020. Published by AIP Publishing.
[4] Baštinec, J., Diblík, J., Pinelas, S., Initial data generating bounded solutions of a system of two linear discrete equations, *AIP Conference Proceedings 2293, 340010-1–340010-4 (2020); https://doi.org/10.1063/5.0026616, International Conference of Numerical Analysis and Applied Mathematics ICNAAM 2019.* Published by AIP Publishing. ISBN: 978-0-7354-4025-8.
[5] Baštinec, J., Diblík, J., Pinelas, S., Vala, J., Determining the initial data generating solutions with prescribed behaviour of a triangular system of linear discrete equations, *Appl. Math. Comput.* **425**, 126533, 1–18.
[6] Baštinec, J., Diblík, J., Růžičková, M., Initial data generating bounded solutions of linear discrete equations, *Opuscula Math.* **26** (2006), No. 3, 395–406.
[7] Bodine, S., Lutz, D. A., *Asymptotic Integration of Differential and Difference Equations*, Lecture Notes in Mathematics **2129**, Springer, Cham, 2015.
[8] Diblík, J., Discrete retract principle for systems of discrete equations, *Comput. Math. Appl.* **42** (2001), 515–528.
[9] Diblík, J., Asymptotic behaviour of solutions of discrete equations, *Funct. Differ. Equ.* **11** (2004), 37–48.
[10] Diblík, J., Migda, M., Schmeidel, E., Bounded solutions of nonlinear discrete equations, *Nonlinear Analysis* **65** (2006), 845–853.
[11] Diblík, J., Růžičková, I., Růžičková, M., A general version of the retract method for discrete equations, *Acta Math. Sin. (Engl. Ser.)* **23** (2007), No 2, 341–348.
[12] Diblík, J., Růžičková, M., Václavíková,B., A retract principle on discrete time scales, *Opuscula Math.* **26**, No 3, (2006), 445–455.
[13] Diblík, J., Václavíková, B., Bounded solutions of discrete equations on discrete real time scales, *Funct. Differ. Equ.* **14** (2007), no. 1, 67–82.
[14] Elaydi, S. N., *An Introduction to Difference Equations*, 3rd edition, Undergraduate Texts in Mathematics, Springer, New York, 2005.
[15] Goldberg, S., *Introduction to difference equations with illustrative examples from economics, psychology, and sociology*, Dover Publications, New York, 1986.
[16] Hlavičková, I., How to find initial data generating bounded solutions of discrete equations, *Tatra Mt. Math. Publ.* **48** (2011), 83–90.
[17] Radin, M. A., *Difference Equations for Scientists and Engineering: Interdisciplinary Difference Equations,* World Scientific Publishing, Singapore, 2019.

Acknowledgement

This work has been supported by the grant of Faculty of Electrical Engineering and Communication, Brno University of Technology (research project No. FEKT-S-20-6225).

ITERATIVE THEORY OF FUNCTIONS AND SOLVING OF FUNCTIONAL EQUATIONS OF A SINGLE VARIABLE

Jaroslav Beránek
Faculty of Education, Masaryk University
Poříčí 31, 603 00 Brno
beranek@ped.muni.cz

Abstract: *The aim of the article is to show how it is possible to introduce to students the problem of functional equations and the study of orbital structures of real functions connected with them. These orbital structures enable solving some types of functional equations of a single variable. Some of the problems shown in the article can be solved with the help of the computer programs, which can be used as the motivation for introducing programs MAPLE, DERIVE etc.*

Keywords: Functional equation; vertex graph; iterative cycle; conjugacy

INTRODUCTION

Functional equations of a single variable are a significant section of a general theory of functional equations (e.g. [2], [8], [9], [10]). Frequently, they serve for mathematic modelling of real situations and appear as tasks in the highest levels of mathematic Olympiad. Therefore, it is important to introduce their solving to students. Solving functional equations of more variables is generally quite simple as students can use various alternatives while substituting specific values of variables or they can use Cauchy method of solving (See e.g. [2], [5], [9], [10]). However, solving functional equations of a single variable is more challenging for them. The use of iterative theory of functions and their interpretation with the help of vertex graphs seems beneficial. With respect to the didactic orientation of the article, let us present a brief outline of the theory which is necessary for solving functional equations of a single variable. The detailed survey of the theory, including proofs of following theorems, could be found in e.g. [8], [9], [10], [11].

1. ITERATIVE THEORY OF FUNCTIONS

1.1. Orbits and vertex graphs of functions

The mapping $f: X \to X$ of the set X into itself will be called transformation of the set X. For $n \in \mathbf{N_0}$ let us define n-th iteration f of the set X as follows: $f^0(x) = x$, $f^1(x) = f(x)$, $f^n(x) = (f \circ f^{n-1})(x)$ for every $x \in X$; in the shortened form we can note $f^n = f \circ f^{n-1}$. If the transformation f is a bijective mapping of the set X onto itself, the definition of this given set iteration can be broadened also for a non-negative integer n in the following way: let f^{-1} be an inverse function to the function f on the set X, then $f^{-2} = f^{-1} \circ f^{-1}$, $f^{-n} = (f^{-1})^n$. It is necessary to carefully distinguish between the notation of the n-th iteration of the function f, which is f^n (the value of this iteration for the element x is $f^n(x)$), and the formula $[f(x)]^n$, which equals $f(x) \cdot f(x) \cdot \ldots \cdot f(x)$.

Every transformation f of the set X determines the equivalence $\sim_f$ on X as follows: $x \sim_f y$, if and only if there exists such pair of positive integers m, n that $f^m(x) = f^n(y)$. The blocks of the

decomposition of the set X determined by the equivalence $\sim_f$ are called orbits of the transformation f, in short f-orbits. The set containing elements $x, f(x), f^2(x), f^3(x), \ldots$ is called the iterative sequence starting in x or the f-splinter of the element x.

Let $k \in \mathbf{N}$, then the cycle of the order k (k-cycle) of the mapping $f: X \to X$ is the set $\{x_0, x_1, \ldots, x_{k-1}\}$ consisting of the set X elements for which there applies $f(x_m) = x_{m+1}$ pro $0 \leq m < k-1$ and $f(x_{k-1}) = x_0$. The orbit containing a cycle is called a cyclic one, otherwise an acyclic one. For $k = 1$, the element $x \in X$ with the property $f(x) = x$ is called the fixed point of the transformation f.

Now let us give some orbit properties which will be further used:

- Every orbit contains one cycle at most.
- The orbit is acyclic if and only if for its every element there applies that the corresponding iterative sequence contains infinitely many elements.
- Every finite orbit is cyclic (the chain ending in in the cycle is not infinite, although it contains infinitely many elements).

In the case of injective transformation f, the orbits are either isolated cycles, two-sidedly infinite chains or infinite chains bounded from below by the least elements; if f is a bijection, its orbits are either cycles or two-sidedly infinite chains. The set of orbits of the function f is also called the orbit structure. The graphic representation of orbits is a vertex graph. In the most general sense, we can encounter a vertex graph as the graphic representation of binary relations. If X is a non-empty set and R is a binary relation on the set X, then a pair (X, R) is called the oriented graph. The elements of the set X are called vertices or nodes (represented as points in the plane), the pairs $(a, b) \in R$ are called oriented edges, the vertex a is called the initial one, and b is called the end one. While representing we draw an arrow leading from the point a to the point b; in the case $a = b$, we draw a loop around the point a). The oriented graph (X, R) is called a functional one (the vertex graph of a mapping), if the relation $R \subset X \times X$ is the mapping of the set X into itself (i.e. every vertex is the initial vertex of just one edge). It is evident that a vertex graph could be plotted as the whole only for transformations of finite sets with not very many elements. In other cases we will only outline the orbits. The comparison of vertex and Cartesian graphs is quite instructive. Here follow two illustrative examples.

a) Let $X = \mathbf{R}$, for every $x \in X$ there applies $f(x) = -x$. The only fixed point is number zero, for other elements of the set X there holds $f^2(x) = x$. The orbits of the function f are then cycles of order 2 (they are infinitely many) and one loop. Here is also a Cartesian graph for comparison.

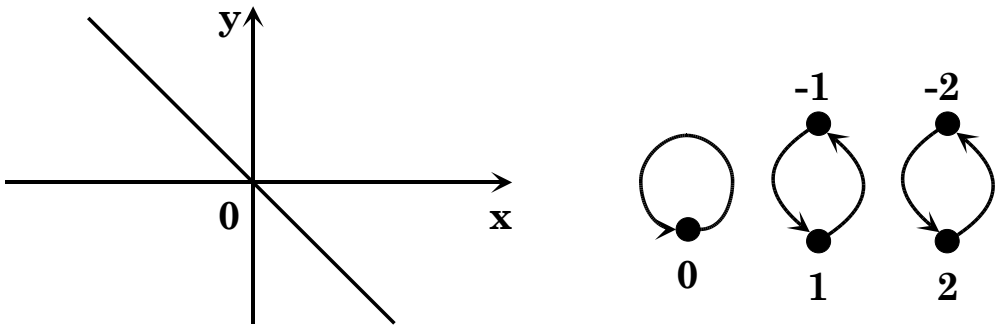


Fig. 1. Cartesian and vertex graph of function $f(x) = -x$.
Source: own

b) Let $X = \boldsymbol{R} - \{0\}$, for every $x \in X$ there applies $f(x) = x^{-1}$. For $x \in \{-1, 1\}$ there holds $f(x) = x$, for other elements of the set X there holds $f^2(x) = x$. The orbits of the function f are then two loops (fixed points) and uncountably many cycles of order 2.

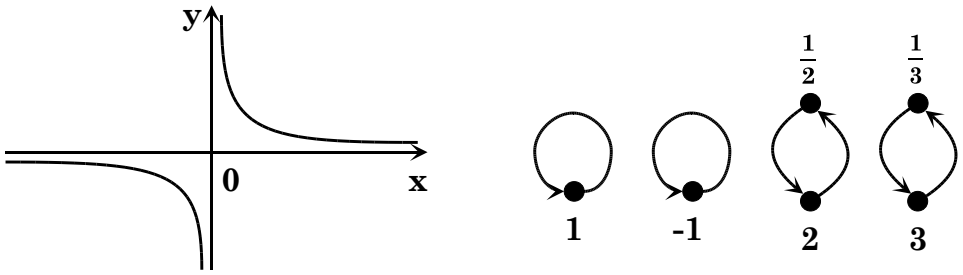


Fig. 2. Cartesian and vertex graph of function $f(x) = x^{-1}$.
Source: own

Let us notice the comparison of the vertex and Cartesian graphs on both examples. Although the Cartesian graphs are completely different, the vertex graphs are almost identical (only save for the number of loops and the marking of vertices).

In the last part of the theoretical outline, let us introduce the notion *conjugacy of functions*. Let X, Y be sets. The function g: $Y \to Y$ is called orbitally homomorphic to function f: $X \to X$, if there exists a function h: $X \to Y$ with the property $h \circ f = g \circ h$. The mapping h is called an orbital homomorphism. In addition, if it is bijective, then it is called an orbital isomorphism. In the case of $X = Y$ we speak about an orbital endomorphism or an orbital automorphism. Orbitally isomorphic functions are called conjugated. Functions f: $X \to X$, g:$Y \to Y$ are conjugated if and only if there exists a bijection h: $X \to Y$ such that for each pair of elements $x, y \in X$ and each pair of non-negative integers m, n there holds: $f^m(x) = f^n(y) \Leftrightarrow g^m(h(x)) = g^n(h(y))$. A more detailed description of conjugated functions is in [4]. It is possible to expect that conjugated functions have "the same" vertex graphs. If they are not the same, during solving the problem of conjugacy it is necessary to restrict the domains of functions so that the orbits with different vertex graphs are "removed". Let us give an example. Here are two real functions f: $\boldsymbol{R} \to \boldsymbol{R}$, g: $\boldsymbol{R} \to \boldsymbol{R}$, $f(x) = ax + b$, $g(x) = ax + c$, $a,b \neq 0$, $c \in \boldsymbol{R}$. Functions f, g are conjugated because there exists a bijection h: $\boldsymbol{R} \to \boldsymbol{R}$ which is defined for $a = 1$ by a formula $h(x) = \frac{c}{b}x + n$, $n \in \boldsymbol{R}$ arbitrary, for $a \neq 1$ by a formula $h(x) = mx + \frac{mb - c}{a - 1}$, $m \in \boldsymbol{R} - \{0\}$ arbitrary. The conjugacy can be easily verified from the definition

1.2. Iterative roots

In this part we will devote our attention to the problem of iterative roots and their determining. At first sight, it seems that this topic does not belong to the functional equations theory, but we will further show that it is not the case. For detailed information see [2], [9], [11].

Let $X \neq \phi$, let f be the mapping of the set X into itself, the number $m \in \boldsymbol{N}$, $m > 1$. The main problem of the iterative theory is to find such an arbitrary mapping g of the set X into itself that for every element x of the set X there applies:

$$g^m = f$$

The mapping g is called the iterative root of the order m of the function f or the m-th iterative root of the function f. We will now briefly outline the general theory of the existence and construction of iterative roots. For the didactic purposes, it is essential that in special cases (real elementary functions, bijective functions, ...) while solving functional equations of one variable it is not necessary to apply complicated theorems from the general theory, but a more efficient solution is possible. The following theorems 1 to 13 are taken from the publication [11], where you can find their proofs.

Theorem 1: Let $X \neq \phi$, let f, g be such mappings of the set X that $g^m = f$, $m \in \mathbf{N}$. Then the mapping is surjective if and only if f is surjective.

Theorem 2: Let $X \neq \phi$, let f, g be such mappings of the set X that $g^m = f$, $m \in \mathbf{N}$. Then the mapping g is injective if and only if f is injective.

Theorem 3: Let $X \neq \phi$, let f, g be such mappings of the set X that $g^m = f$, $m \in \mathbf{N}$. Then the mapping g is bijective if and only if f is bijective.

Now we will show the solution of a simple functional equation of one variable $g^m(x) = x$ on a non-empty set X, m is a natural number bigger than one. A trivial solution is the identical equation $g(x) = x$ itself, therefore we search for the non-trivial solution. The orbital structure of the identical function consists of isolated fixed points. The identical function is bejective, therefore according to Theorem 3 function g is bijective. The orbital structure of the bihection can contain only cycles and two-sidedly infinite chains. However, as $g^m(x) = x$, such chains are excluded at the mapping g. Therefore the desired function g must containonly cycles, while the order of these cycles must be the divisor of number m. Now the discussion of the solution depends on number m and the cardinality of the set X. If e.g. m is a prime number, the orbital structure of the desired function g contains either m-cycles or fixed points. The general solution can be described as follows: Let number m has $r+1$ divisors $m_0, ..., m_r$. Without detriment to generality, these divisors can be denoted so that it holds $1 = m_0 < m_1 < ... < m_{r-1} < m_r = m$. We will decompose the set X to $r+1$ blocks so that the elements of the set X in each block will form only m_i-cycles for $i = 0,..., r$ (some of these blocks can be empty). The mapping g, whose orbital structure contains blocks with m_i-cycles, is the solution of the equation $g^m(x) = x$.

Theorem 4: Let g be the m-th iterative root ($m \in \mathbf{N}$, $m \geq 2$) of the mapping f of the non-empty set X. Then every g-orbit is the union of p f-orbits, where p/m. If $p < m$, then all g-orbits are n-cyclic, while p/n. In addition, all f-orbits are $\frac{n}{p}$ - cyclic and at the same time the greatest common divisor (GCD) of numbers m, n equals p.

Theorem 4 describes properties of iterative roots provided that they exist. Now let us state the necessary and sufficient conditions for the existence of iterative roots.

Let f be the mapping of the set X into itself, let r, m be natural numbers with the property r/m. Let the mapping f contain at least r orbits and let there be given arbitrary r f-orbits. These orbits will be denoted m-mateable (by any mapping g), if g is the m-th iterative root of the function f, has one orbit and maps the union of the given r f-orbits into themselves. For $r = 1$ this only f-orbit is called m-self-mateable.

Theorem 5: If in the previous text there applies $r < m$, then the necessary condition for the m-mateability of r f-orbits is the fact that each of them is k-cyclic (with the same k) and tere applies

that GCD $(k, \frac{m}{r}) = 1$. The corollary of this Therem is, among others, the fact that an acyclic *f*-orbit cannot be *m*-self-mateable for any *m*.

Theorem 6: An arbitrary mapping of a non-emptyset has the *m*-th iterative root ($m \in \mathbf{N}$) if and only if the set of orbits of this mapping can be decomposedto disjoint blocks with following properties:

1° The number of orbits in each block is finite and is the divisor of the number *m*.

2° Orbits in each block are *m*-mateable.

Theorem 7: For the existence of the *m*-th iterative root ($m \in \mathbf{N}, m \geq 2$) of the mapping $f: X \to X$ it is sufficient if in the orbit structure of the function *f* there exist for each occurring orbit type either infinitely many orbits of such type or their number is divisible by number *m.*

Theorem 8: Let *f* be the bijection of any set into itself. Let us denote l_0 the number of the two-sidedly infinite chains, l_k be the number of *k*-cycles of the mapping *f*, $k \in \mathbf{N}$. Then there exists the *m*-th iterative root ($m \geq 2, m \in \mathbf{N}$) of the mapping *f* if and only if for every non-negative number *k* there applies either $l_k = \infty$ or $d_k | l_k$, where $d_0 = m$, $d_k = \frac{m}{m_k}$ $(k \in \mathbf{N})$, while m_k denotes the greatest common divisor of the number *m*, which is coprime to the number *k*.

Theorem 9: Let $f: X \to X$ be such bijection that in its orbit structure for every $k \in \mathbf{N}_0$ there applies either $l_k = 0$ or $l_k = \infty$ (according to the notation in Theorem 8). Then *f* has the *m*-th iterative root for every natural number *m.* For the orbits of this iterative root there also applies either $l_k = 0$ or $l_k = \infty$ for all $k \in \mathbf{N}_0$.

Theorem 10: Every strictly increasing and continuous bijection $\boldsymbol{R}$ on $\boldsymbol{R}$ has iterative roots of all orders.

Theorem 11: The strictly decreasing and continuous bijection of the set $\boldsymbol{R}$ has iterative roots of all orders if and only if it has either infinitely many 2-cycles or none.

Theorem 12: Every strictly decreasing and continuous bijection $\boldsymbol{R}$ has iterative roots of all odd orders.

Theorem 13: Let *f* be an arbitrary transformation of the set $\boldsymbol{R}$ which contains for a fixed natural number *m*, $m \geq 2$, at least *m f*-orbits. Let be given *m* such orbits. Then there applies: If these *f*-orbits are orbitally isomorphic to each other (i.e. of the same type), then they are *m*-mateable.

2. FUNCTIONAL EQUATIONS OF A SINGLE VARIABLE - EXERCISES

Exercise 1. At the 28th year of the International Mathematics Olypiad in 1987 appeared the following exercise ([13]):

Prove that there does not exist the function f mapping the set $\mathbf{N}_0 = \{0, 1, 2,...\}$ of all non-negative numbers to $\mathbf{N}_0$ such that $f(f(n)) = n + 1987$ for every $n \in \mathbf{N}_0$.

In the authorial solution of this exercise (according to [13]) there was used the proof by contradiction. Let us show this solution for information.

Let us assume that such function f exists. Obviously, for function f there applies from the task:

$$f\,(x + 1987) = f(f(f(x))) = f(x) + 1987,$$

from this using induction we will prove the validity of the equality

$$f\,(x + k.1987) = f(x) + k.1987 \text{ pro } x \in \mathbf{N}_0, k \in \mathbf{N}_0.$$

The function f is therefore given unambiguously by its values on the set $M = \{0, 1, 2, ..., 1986\}$. We will define on M a new function g as follows: For $x \in M$ we will express $f(x)$ in the form

$$f(x) = y + p.1987, \text{ where } y \in M, p \in \mathbf{N}_0$$

and state $g(x) = y$. As from the previous form there applies

$$x + 1987 = f(f(x)) = f\,(y + p.1987) = f(y) + p.1987,$$

then necessarily there must apply

$$f(y) = x + (1 - p)1987 \in \mathbf{N}_0,$$

so $p = 1$ or $p = 0$ or $0 < p < 1$. From the last expression of $f(y)$ there results $g(y) = x$, so there holds $g^2(x) = x$. As the number of the elements of the set M is odd, (*card* $M = 1987$), there must exist $x_0 \in M$ with the property $g(x_0) = x_0$. However, from the definition of g it means that there will arise one of two possibilities: $f(x_0) = x_0$ or $f(x_0) = x_0 + 1987$. In the first case then there applies $x_0 + 1987 = f(f(x_0)) = f(x_0) = x_0$, in the second case we will obtain $x_0 + 1987 = f(f(x_0)) = f(x_0 + 1987) = x_0 + 3974$. In both cases we reached the contradiction, so the function f cannot exist.

Now let us use the iterative theory of functions. First, we will find the vertex graph of the function $\varphi(x) = x + 1987$ and then we will search for its second iterative roots. Function $\varphi(x) = x + 1987$ is not a bijection on the set $\mathbf{N}_0$, but it is an injection. It has no fixed points and its orbits are mutually isomorphic chains bounded from below. There are *1987* such chains, their least elements are *0, 1, ..., 1986*. The vertex graph is outlined in Fig. 3:

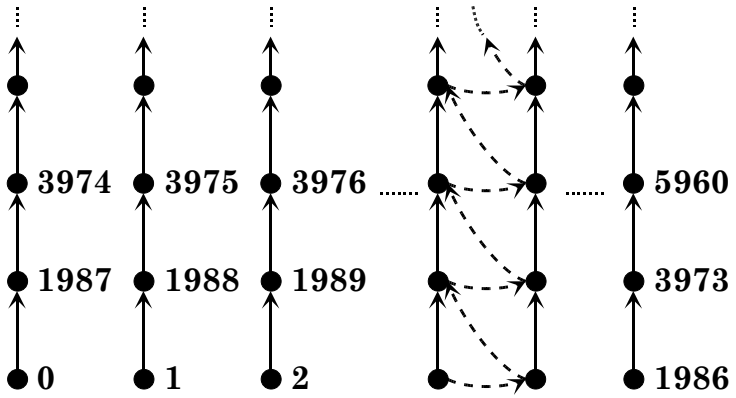


Fig. 3. Vertex graph of function $\varphi(x) = x + 1987$.
Source: Own

The difficulty of this task is to prove tht the function φ does not have the second iterative root, i.e. that φ-orbits are not nejsou *2*- mateable. The main idea of this proof is the fact that that there is an odd number of orbits. The orbits are not cyclic, therefore according to Theorem 5 they cannot be self-mateable. According to Theorems 5 and 6, for the existence of the second

iterative root the number of φ-orbits has to be even (orbits can be mated only in pairs). This does not hold, so the function φ des not have the iterative rootof order *2*.

Note: With the help of iterative theory it is possible to generalize this exercise. The first question is if the function $\varphi(n) = n + 1987$ has any own iterative roots (different from the trivial iterative root of order *1*, which is the functionkce $\varphi(n)$ itself). With respect to Theorems 5 and 6 it is obvious that we are searching the possibility of the mating of the existing *1987* orbits. As *1987* is a prime number, the only possible own iterative root is the root of order *1987*. Thus there exists the function $f: \mathbf{N}_0 \to \mathbf{N}_0$ with the property $f^{1987}(n) = n + 1987$ for $n \in \mathbf{N}_0$. This function f is the successor function v_0 on $\mathbf{N}_0$, defined by the formula $f(n) = n + 1$. The next difficulty is to find out when in general there exist own iterative roots of the function $\varphi(n) = n + c$, $n \in \mathbf{N}_0$, $c \in \mathbf{N}$. The vertex graph now contains just c isomorphic orbits (chains bounded from below with the least elements $0, 1, \ldots, c-1$). These chains have to be mated. Similarla as above, there always exists the iterative root of the order c (which is the function $f(n) = n + 1$). Furter, there always exist iterative roots of these orders which are the divisors of the number c. Therefore, the iterative root of the order *2* exists if the number c is an even number. If the problem was set for the function $f(n) = n + 1988$, the second iterative root would exist (further there would exist iterative roots of orders *4, 7, 14, 28, 71, 142, 284, 497, 994, 1988*).

Exercise 2. *There is given the quadratic function $f(x) = x^2 - 2$. Find the vertex graph of this function and with the help of it solve the functional equation $g(x^2) = [g(x)]^2 - 2$* (See [3])*.*

The main difficulty while designing the vertex graph will consist in searching the cycles of the given quadratic function $f(x) = x^2 - 2$. If we want to determine whether the function f has the cycle of the order n, we have to solve the equation $f^n(x) = x$, where f^n is the n-th iteration of the given function. For $n = 1$ while solving the quadratic equation we will easily discover that the function f has two fixed points -1 and 2. For $n = 2$ we will solve the equation $x^4 - 4x^2 - x + 2 = 0$. Although solving biquadratic equations is quite difficult algorithmically, in our case we will find the solution easily; we already know two roots $x_1 = -1$, $x_2 = 2$, therefore using e.g. Horner´s scheme we will find the decomposition of the given biquadratic equation $(x+1)(x-2)(x^2+x+1) = 0$, from which we will calculate $x_{3,4} = \frac{-1 \pm \sqrt{5}}{2}$. It is easy to verify that the two last mentioned irrational roots indeed form the 2-cycle of the function f. Naturally there arises a question whether the function $f(x) = x^2 - 2$ has further cycles of higher orders. For the search for the cycles there seems of great use the book by J. Smítala [10], which includes the Theorem of A. N. Šarkovskij. It says the following:

„Let f be a continuous function from the interval I to I. Let us introduce a new ordering on the set of all natural numbers as follows: $3 < 5 < 7 < \ldots < 2.3 < 2.5 < 2.7 < \ldots < 2^i.3 < 2^i.5 < 2^i.7 < \ldots < 2^{j+1} < 2^j < \ldots < 8 < 4 < 2 < 1$. Thus first there are odd numbers in the natural ordering, then their doubles, then their quadruples, etc. The ordering ends with the powers of number 2 in the descending order. Then there holds: If the function f has a cycle of order m and there holds $m < n$, then the function has also a cycle of order n".

According to Šarkovskij ordering, the next possible cycle will be the cycle of the order 4. We will determine it while solving the algebraic equation $f^4(x) = x$, which after substituting has the form

$$x^{16} - 16x^{14} + 104x^{12} - 352x^{10} + 660x^8 - 672x^6 + 336x^4 - 64x^2 + 2 = x.$$

For solving this equation of the order 16 it is necessary to use the computer technology. With the help of such tool we will find out thet such equation has 16 different real solutions, while

all of them lie in the interval (–2, 2). We already know four of them; these are all the elements which form cycles of orders 1 and 2 and have been given above. The next twelve solutions form three 4-cycles.

The next possible cycle is the cycle of the order 8. While its determining, it is necessary to solve the algebraic equation $f^{8}(x) = x$, which has the order 256. The use of a computer is inevitable. It results in 256 differentreal solutions which all lie in the interval $\langle -2, 2\rangle$. Among these roots certainly belong the 16 solutions which were found while searching for the cycle of the order 4. The conclusion is that the function $f(x) = x^2 - 2$ has also the cycle of the order 8, aanfd not only one, but thirty of them (256 solutions, among them two fixed points, one cycle of the order 2, three cycles of the order 4 and 30 cycles of the order 8).

It is evident that searching for other cycles of higher orders is practically impossible. According to Šarkovskij Theorem We will focus our attention to the cycle of the order 3. If we prove its existence, the function f will have cycles of all orders. Here we will show the connections with other mathematics areas as well.

The cycle of the order 3 will be determined while solving the equation $f^{3}(x) = x$ which after substitution has the form

$$x^8 - 8x^6 + 20x^4 - 16x^2 - x + 2 = 0.$$

With the help of a computer we will algebraically determine eight real solutions (as above, all of them lie in the interval $\langle -2, 2\rangle$). We already know two roots (fixed points –1 and 2). The next roots are:

$$x_1 = \frac{1}{2}\sqrt[3]{-4+4i\sqrt{3}} + \frac{2}{\sqrt[3]{-4+4i\sqrt{3}}},$$

$$x_2 = -\frac{1}{4}\sqrt[3]{-4+4i\sqrt{3}} - \frac{1}{\sqrt[3]{-4+4i\sqrt{3}}} + \frac{1}{2}i\sqrt{3}\left(\frac{1}{2}\sqrt[3]{-4+4i\sqrt{3}} - \frac{2}{\sqrt[3]{-4+4i\sqrt{3}}}\right),$$

$$x_3 = -\frac{1}{4}\sqrt[3]{-4+4i\sqrt{3}} - \frac{1}{\sqrt[3]{-4+4i\sqrt{3}}} - \frac{1}{2}i\sqrt{3}\left(\frac{1}{2}\sqrt[3]{-4+4i\sqrt{3}} - \frac{2}{\sqrt[3]{-4+4i\sqrt{3}}}\right),$$

$$x_4 = \frac{1}{6}\sqrt[3]{28+84i\sqrt{3}} + \frac{\frac{14}{3}}{\sqrt[3]{28+84i\sqrt{3}}} - \frac{1}{3},$$

$$x_5 = -\frac{1}{12}\sqrt[3]{28+84i\sqrt{3}} - \frac{\frac{7}{3}}{\sqrt[3]{28+84i\sqrt{3}}} + \frac{1}{2}i\sqrt{3}\left(\frac{1}{6}\sqrt[3]{28+84i\sqrt{3}} - \frac{\frac{14}{3}}{\sqrt[3]{28+84i\sqrt{3}}}\right),$$

$$x_6 = -\frac{1}{12}\sqrt[3]{28+84i\sqrt{3}} - \frac{\frac{7}{3}}{\sqrt[3]{28+84i\sqrt{3}}} - \frac{1}{2}i\sqrt{3}\left(\frac{1}{6}\sqrt[3]{28+84i\sqrt{3}} - \frac{\frac{14}{3}}{\sqrt[3]{28+84i\sqrt{3}}}\right).$$

Further we will determine the decomposition of the polynomial $x^8 - 8x^6 + 20x^4 - 16x^2 - x + 2$. We will get the product

$(x + 1)(x - 2)(x^3 - 3x + 1)(x^3 + x^2 - 2x - 1)$. Numbers x_1, x_2, x_3 are the roots of the polynomial $x^3 - 3x + 1$ and numbers x_4, x_5, x_6 are te roots of the polynomial $x^3 + x^2 - 2x - 1$. Using a precise calculation, we will verify the existence of the 3-cycle which is formed by numbers x_1, x_2, x_3.

First of all, we have to „rearrange“ these numbers. If we transform the complex number $-4+4i\sqrt{3}$ to the goniometric form $8(\cos\frac{2\pi}{3} + i \sin\frac{2\pi}{3})$, we can determine three values of the root $\sqrt[3]{-4+4i\sqrt{3}}$.

The first one is $2(\cos\frac{2\pi}{9} + i \sin\frac{2\pi}{9}$, further $2(\cos\frac{8\pi}{9} + i \sin\frac{8\pi}{9})$, $2(\cos\frac{14\pi}{9} + i \sin\frac{14\pi}{9})$. If we substitute the first of these values to the formula for the root x_1, we will get the real value $x_1 = 2\cos\frac{2\pi}{9}$. Now let us verify the existence of the 3-cycle:

$$x_1^2 - 2 = 2\cos\frac{4\pi}{9},\ (x_1^2-2)^2 - 2 = 2\cos\frac{8\pi}{9},\ ((x_1^2-2)^2-2)^2 - 2 = 2\cos\frac{16\pi}{9}.$$

As there applies $\cos\frac{2\pi}{9} = \cos\frac{16\pi}{9}$, real numbers $2\cos\frac{2\pi}{9}$, $2\cos\frac{4\pi}{9}$, $2\cos\frac{8\pi}{9}$ form the 3-cycle of the function $f(x) = x^2 - 2$. Analogically, we can verify (the calculation is a bit more laborious) that the triplet of roots x_4, x_5, x_6 also forms the 3-cycle.

In order to finish examining cycles of the function $f(x) = x^2 - 2$, there remains to verify the hypothesis that elements of all cycles lie in the interval $\langle -2, 2\rangle$ (thus hypothesis follows from the calculation while searching for the roots). For $x > 2$ we can determine through the calculation that $x^2 - 2 > x$ and thus trivially $x^2 - 2 > 2$. All natural iterations of the function f form an increasing sequence for $x > 2$, therefore there cannot hold $f^n(x) = x$. As the function f is even, there holds the corresponding assertion also for numbers $x < -2$. In the conclusion we can state that the vertex graph of the function $f(x) = x^2 - 2$ has cycles of all possible orders whose values lie in the interval $\langle -2, 2\rangle$.

Now we can already design the vertex graph of the function $f(x) = x^2 - 2$. As the range of this function is the interval $\langle -2, \infty)$, it is evident that numbers less than -2 cannot be the functional value of the function f for any real number x. Numbers from the interval $(-\infty, -2)$ are therefore the minimal elements, in the vertex graph they do not have any predecessor. Numbers from the interva $(2, \infty)$ have two predecessors; one of them is in the interval $(2, \infty)$, the secod one is in the interval $(-\infty, -2)$ because f is an even function. This follows from the following consideration: Let $x^2 - 2 = u$, where $u > 2$. Then $x = \sqrt{u+2} > \sqrt{4} = 2$. As the given function is even, we can claim that positive numbers higher than two form both-sidedly infinite chains on which at every point one opposite negative number less than -2 is “connected”. With the contraction of the domain of the function $f(x) = x^2 - 2$ to the union of intervals $(-\infty, -2) \cup (2, \infty)$ its orbits are therefore the same as infinite orbits of the quadratic function $q(x) = x^2$ (See Fig. 4). In the iteration theory these orbits are called both-sidedly infinite chains with short chains. The detailed description of the vertex graph of the function $q(x) = x^2$ is given in [6].

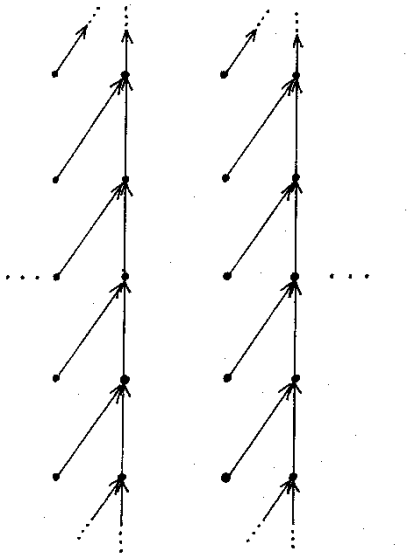


Fig. 4. Infinite orbits of the quadratic function $q(x) = x^2$
Source: Own

Therefore we can claim that there exist a function g on these intervals which is the solution of the functional equation $g(x^2) = [g(x)]^2 - 2$. Both functionscontain uncountably many orbits of the same form. We will create pairs of orbits (one f-orbit and one q-orbit) and the funvction g matches mutually corresponding elements of both orbits. The detailed information about this task can be found in the article [6]. For the chosen pair of orbits we will determine the functional formula for the function. We will denote O_f as the orbit of the function f, and O_q as the orbit of the function q. We will choose any element of the orbit O_q, which we will denote; in the orbir O_f we will also choose an element y_0. Let $x_0 > 2$, $y_0 > 2$. The function g: $O_q \to O_f$ is then defined e.g. as follows: For $k \in \boldsymbol{N_0}$ let there applies $g(x_0^{2^k}) = f^k(y_0)$, $g(-x_0^{2^k}) = -f^k(y_0)$, $g(\sqrt[2^k]{x_0}) = \left(f^{-1}\right)^k(y_0)$, $g(-\sqrt[2^k]{x_0}) = -\left(f^{-1}\right)^k(y_0)$, where f^{-1} is the inversive function to the function f in the interval $(2, \infty)$, defined by the formula $f^{-1}(y_0) = \sqrt{y_0 + 2}$. Let us further mention that for the zero iteration there holds $f^0(x) = x$ for every real number x and for every function f. In the function g defined by the above given way really satisfies the given functional equation.

For numbers from the interval $\langle -2, 2\rangle$ the situation is completely different. It is evident that in the interval $\langle -2, 2\rangle$ the sought for conjugating function g cannot exist, i.e. the equation $g(x^2) = [g(x)]^2 - 2$ does not have a solution. Vertex graphs of functions q and f are totally different – the orbits of the function q are acyclic (except for the fixed points 0 and 1) and the orbits of the function f contains cycles of all orders. The detailed justification of the non-existence of the conjugating function g could be found in the publications, e.g. [2], [7], [11].

Example 3. *There is given the quadratic function $f(x) = x^2 + 2$. Find the vertex graph of this function and with its help solve the functional equation $g(x^2) = [g(x)]^2 + 2$* (See [3]).

Let us consider the functional equation $g(x^2) = [g(x)]^2 + 2$. We will see that the modification which from the continunous aspect seems unimportant will substantially change the vertex graph of one of the functions. It deals with the question if the functions $q(x) = x^2$ and $\varphi(x) = x^2 + 2$ are conjugated. The vertex graph of the function q was described earlier, so we will deal with the vertex graph of the function φ. It was Descibed in detail in the author´s article [1], therefore we will mention it briefly. In this case neither the fixed points nor other cycles exist,

all orbits are infinite and acyclic. For their description we will use an auxiliary function ω, defined on the set of all natural number as follows:

$$\omega(n) = \begin{cases} n+2 \; for \; n \; odd \\ n+1 \; for \; n \; even \end{cases}.$$

The function $\varphi(x) = x^2 + 2$ contains the only orbit of the same form (i.e. isomorphic) to orbit ($\mathbf{N_0}$, ω) and further it contains uncountable many orbits same as ($\mathbf{N}$, ω). All orbits of the functions φ are elements bordered from below from the interval (–2, 2). As the orbits of the function q are both-sidedly infinite (i.e. not borbdered from below and above, if we do not consider two finite orbits for numbers 0, –1, 1) and the orbits of the function φ are bordered from below, it is impossible for the bijective function g to exist, with the help of which both functions would be conjugated for all real numbers.

Exercise 4. Prove that there exists an injective mapping **R** to **R** which does not have the n-th iterative root for any $n \geq 2$. (See [12]).

The proof of the existence of the sought for mapping will be prformed in two ways, although in the second one only theoretically without defining the given mapping by the functional formula. We want to prove that there exist the bijection on **R** which has no own iterative roots. Let us mention that the vertex grapg of the bijection contains only cycles and both-sidedly infinite chains. From the general theory there follows that the bijective mapping has no own iterative roots if and only if it contains only one both-sidedly infinite chain or if it contains infinitely many cycles, each of different orders. Now let us describe two possible cases:

a) $f(x) = x + 1$ for every $x \in \mathbf{Z}$, $f(x) = x$ for $x \in \mathbf{R} - \mathbf{Z}$. Such mapping f contains the onlyboth-sidedly infinite chain and infinitely many loops.

b) We will decompose the set of all natural numbers to infinitely many blocks of decomposition whose cardinalities are different to each other and thea are given by all natural numbers, so e.g. {1}, {2, 3}, {4, 5, 6}, {7, 8, 9, 10}, ... In every block we will define the mapping f as the cyclic permutation, for every $x \in \mathbf{R} - \mathbf{N}$ we will set $f(x) = x$. This mapping f contains infinitely many loops and infinitely many cycles of all possible orders.

In either of the two cases the described function f has no own iterative roots.

Note: As the last remark in conclusion let us mention some interesting assertions dealing with designing cycles of the function $f(x) = x^2 - 2$. In the previous text we established through a precise calculation the cycle of the order 3, whose values are $2\cos\frac{2\pi}{9}$, $2\cos\frac{4\pi}{9}$, $2\cos\frac{8\pi}{9}$. While calculation we used, among others, that for $k \in \mathbf{N}$ there applies generally the equation $(2\cos\frac{k\pi}{9})^2 - 2 = 2\cos\frac{2k\pi}{9}$. The just given relation can be generally formed for every real number x as follows: $(2\cos x)^2 - 2 = 2\cos 2x$. Therefore the sequence of iterations of the function $f(x) = x^2 - 2$ is for $x = 2\cos x$ as follows: $2\cos x, 2\cos 2x, 2\cos 4x, 2\cos 8x, ..., 2\cos 2^k x, ...$. Such sequence of iterations will contain the cycle of the order m, if and only if for some natural number k will hold $2\cos 2^k x = 2\cos 2^{k+m} x$. The last goniometric function can be solved generally using familiar formulas, to make it simpler we can set $k = 0$. The elements of the 2-cycle will be calculated with the help of the goniometric equation $2\cos x = 2\cos 4x$. As the result we have two possible 2-cycles:

$2\cos\frac{2\pi}{3}$, $2\cos\frac{4\pi}{3}$ a $2\cos\frac{2\pi}{5}$, $2\cos\frac{4\pi}{5}$. Both values of the elements of the first of the 2-cycles equals −1 (the fixed point of the function *f*, i.e. also the cycle of the order 2), values of the second of the 2-cycles have been already established as $\frac{-1\pm\sqrt{5}}{2}$. Then there must hold $2\cos\frac{2\pi}{5}=\frac{-1+\sqrt{5}}{2}$, $2\cos\frac{4\pi}{5}=\frac{-1-\sqrt{5}}{2}$, which is an interestingpossibility how to precisely express the values of goniometric functions with the help od roots. While searching for the 3-cycle we will use the equation $2\cos x = 2\cos 8x$. One of the solutions is already established 3-cycle $2\cos\frac{2\pi}{9}$, $2\cos\frac{4\pi}{9}$, $2\cos\frac{8\pi}{9}$; the next solution is one more 3-cycle $2\cos\frac{2\pi}{7}$, $2\cos\frac{4\pi}{7}$, $2\cos\frac{8\pi}{7}$. It is obvious that in the general case we will not get all cycles in this process – the given method using goniometric equation enables for every natural number *m* to establish precise values of at least one cycle of the order *m*. For example for $m = 4$ one of the 4-cycles is formed by elements $2\cos\frac{2\pi}{17}$, $2\cos\frac{4\pi}{17}$, $2\cos\frac{8\pi}{17}$, $2\cos\frac{16\pi}{17}$.

CONCLUSION

This part where we dealt with functional equations of one variable is relatively extensive and contains a lot of theoretical notions. However, they are essential for solving eqatinons. Solving functional equations of one variable of the type $f^3(x) = x$, $f^2(x) = x + 2$, $f^2(x) = x^2$, $f(x^2) = [f(x)]^2$ etc. without knowing vertex graphs, orbit theory and their mateability is rather complicated and in many cases it is impossible. Let us note that theoretical parts were designed so that after a certain simplification they can serve the secondary school students (e.g. while preparing for mathematics olympiads). On the other hand they could serve the university students as the initial information and motivation for further study of the theory of iterations and iterative roots. It is obvious that studying cycles and vertex graphs of real functions offers a wide range of possibilities and topics for students for their independent creative work, further it provides connections with other mathematics areas (here with solving algebraic equations, goniometric equations etc.), which can lead to liven up mathematics teaching. Of great importance is also the motivational role of these aspects of mathematics teaching because the above given theory can lead the readers to solving even such functional equations which whe approached for the first time could seem beyond their strength. For example, solving the functional equation $f^2(x) = \cosh(x) - 1$ is analogical to the equation $f^2(x) = x^2$ because the vertex graph of the quadratic function *q* and the function $\cosh(x) - 1$ is the same at first sight. The given types of functional equations of one variable are not the only ones; many of them can be solved e.g. with the help of the theory of recurrent sequences (e.g. the functional equation $f(n + 1) = 3 f(n) - 2 f(n - 1)$). Certainly, there exist much more complicated functional equations, including a lot of open problems in this area. Some of them could be found e.g. in mnographs by F. Neuman and M. Kuczma ([8], [9]).

REFERENCES

[1] Beránek, J. *O iterativních kořenech jisté polynomické funkce*. 1. vyd. Brno: Matematika a didaktika matematiky. Sborník prací Pedagogické fakulty MU. Masarykova Univerzita, 1993. s. 5-15.

[2] Beránek, J. *Funkcionální rovnice*. 1. vyd. Brno: Masarykova univerzita, 2004. 74 s. Matematika a didaktika matematiky, sv. 121. ISBN 80-210-3422-X.

[3] Beránek, J., Beránková, J. *Functional equations in problems of school mathematics.* In XXVIII International Colloquium on the Management of Educational Process. 1. vyd. Brno: Univerzity of Defence, 2010. s. 42 - 50. ISBN 978-80-7231-733-2.

[4] Coufalová, Y.,Francová, M., Chvalina, J. *Konjugace zobrazení a funkcí*. 1. vyd. Brno: Matematika a didaktika matematiky. Sborník prací Pedagogické fakulty MU, řada matematických věd č. 3, Masarykova Univerzita, 1993. s. 31-48.

[5] Davidov, L.: *Funkcionální rovnice*. ŠMM, Mladá fronta, Praha 1984.

[6] Chvalina. J., Beránek, J.: *O iteračních odmocninách kvadratické funkce*. In: Sborník prací pedagogické fakulty UJEP, řada matematických věd č. 1, Brno 1990, s. 7-19.

[7] Chvalina, J. *Funkcionální grafy, kvaziuspořádané množiny a komutativní hypergrupy*. 1. vyd. Brno: Masarykova univerzita, 1995, 205 s. ISBN 80-210-1148-3.

[8] Kuczma, M.: *Functional Equations in a Single Variable*. PWN, Warszawa 1968.

[9] Neuman, F.: *Funkcionální rovnice*. SNTL, Praha 1986.

[10] Smítal, J. *O funkciách a funkcionálnych rovniciach*. 1. vyd. Bratislava: Alfa, 1984. 143 s. ISBN 63-146-84.

[11] Targonski, Gy.: *Topics in Iteration Theory*. Vandenhoeck et Ruprecht, Göttingen and Zürich 1981.

[12] Zítek, F., a kol.: *33. ročník matematické olympiády*. SPN, Praha 1986.

[13] Zítek, F., a kol.: *36. ročník matematické olympiády na středních školách*. SPN, Praha 1989.

Optimal responding to cyber incidents:mathematical model and analitika

Dzhalladova Irada, prof., DrSc.,
Kyiv National Economic University, Department of Computer Mathematics and Information Security
Brno University of Technology, Faculty of Business and Management, Institute of Informatics
idzhalladova@gmail.com
Veronika Novotná, Ph.D.,doc., Mgr.
Brno University of Technology, Faculty of Business and Management, Institute of Informatics
veronika.novotna@vut.cz
Jan Luhan, Ph.D., Ing.
Brno University of Technology, Faculty of Business and Management, Institute of Informatics
Jan.Luhan@vut.cz

Abstract: *In this paper we offered an easy mathematical model to clarify how to define the time of a choice answered on the incident. It can depend upon the stakes involved within the present situation. The model deals with the question of when the resource should be used on the condition that its use today might prevent it from being available to be used later. The analysis provides concepts, theory, applications, and distinctions to market the understanding of strategy aspects of cyber conflict. Case of concept studies includes the same cyberattack, the persistent cyber espionage applied by some country's military. This paper focuses on one aspect of the problem: the timing of an answered cyber incident, either within of espionage or disruption. The goal of the paper is to push the understanding of this domain of cyber incidents to mitigate the harm of cyberattacks can do, and harness the capabilities they can provide.*

Keywords: optimal responding, a cyber incident, probability, non-Markovian process, Reliability and Stability, defense and attack.

INTRODUCTION

The paper takes a mathematical model offered to help analyze a choice for the optimal time for an attack. In other words when the own weapon should be used by the attacker, knowing that its use today may well prevent it from being effective later and find is the trade-off between waiting until the stakes of the present situation are high enough to warrant the use of the weapon, and not waiting so long that the vulnerability of the opponent might be discovered and patched. Studies [1,2] have clearly recognized that a cyber weapon has a strong tendency to depreciate once used. The implication has often explicitly drawn that it may pay to wait for an appropriate moment to deploy an attack. Namely, the longer one waits the more likely the opponent will have recognized and fixed the vulnerability one's resource is meant to exploit. Our model specifies how to comprehend the variables in this implicit logic of the timing of cyber conflict, to construct functional depend between these variables, and how to solve the decision problem inherent in the trade-offs among these variables.

In the present paper, we are going to be used here as a "cyber weapon". A cyber weapon needn't be a weapon in the sense of something which will cause damage by itself, it can be used for espionage within which case its use isn't necessarily an attack. Additionally, a cyber weapon to

take advantage of a target's vulnerability might include nontechnical means either on their own or in conjunction with technical means of intrusion. The present model is an adaption and extension of the model developed to review "the rational timing of surprise" [3,4] and therefore the mathematical instruments of this model [7,10]. Our model is presented from the angle of the attacker: when should a cyber weapon be wont to exploit a vulnerability in an exceeding target's network? The results, however, are equally relevant to a defender who wants to estimate how high the stakes must be so as for the offense to use an unknown vulnerability. Section 1 provides a model that expresses the worth of a "cyber weapon" for exploiting a vulnerability within the target's ADP system, then calculates when best to use that weapon. The event of our model provides some useful recommendations including the Attacker and Defender of a weapon for exploiting a cyber vulnerability. Section 2 concludes with a review of concepts and future avenues for research, and also the implications of our model for the analysis of the cyber incident.

1 Model of the optimal timing

For constructing our model we introduce some assumptions and characteristics of cyber weapons.

1.1 Assumptions

Assumption 1. Rate. You know the rate in current states. You do not know what the rates will be at any future state, although you do know the distribution of rates over time. The assumption about trades means you may know that the rates are low today, and you may be able to estimate the likelihood of various possible rates in the future, but you do not know when the rates associated with a particular event will occur. In other words rates changes according to the non-Markovian process [5,6].

Assumption 2. Characteristics and Value of probability a cyber weapon. For a cyber weapon, we can estimate two parameters that determine whether the weapon will be available next time t. These are the Reliability and the Stability of the resource. The Reliability of a resource is the probability that if you use it now it will still be usable in the next time period. The Stability of a resource is the probability that if you refrain from using it now, it will still be useable in the next time period.

$$p = Reliability = P(cyberweapon|ifuseit) \tag{1}$$

and

$$q = Stability = P(cyberweapon|ifnotuseit). \tag{2}$$

Both p and q depend not only on the resource itself but also on the capacity and vigilance of the intended target. The Reliability of a cyber weapon used against a well-protected target is likely to be less than the Reliability of the weapon against a target that is not particularly security conscientious. Likewise, a weapon will typically have less Stability against a target that keeps up-to-date on security patches than one that does not. In the case of a distributed denial of weapon, the effectiveness of the attack depends on the current capacity of the target to handle massive inputs, whereas the ability of the attacker to repeat the attack depends on the target's subsequent attainment of sufficient capacity to handle another such attack.

1.2 Main equations.

We will look at the implications of several distributions on the likelihood of various rates—such as how likely high routes state events are compared with routine low-states events. That is, we have a stochastic process — a set of random variables X_t, where the index $t < T$ is time, which can be discrete, but more often covers all real numbers in a certain interval. Just in time T (it s will has means as Thousholds), we will carry out optimization further. As you know, the stochastic properties X_t of are expressed by common distribution functions:

$$P_n\left(x_1, t_1; x_2, t_2; \ldots; x_n, t_n\right) dx_1 dx_2 \ldots dx_n =$$
$$x_1 < X_{t_1} < x_1 + dx_1,\ x_2 < X_{t_2} < x_2 + dx_2, \ldots\ ,\ x_n < X_{t_n} < x_n + dx_n$$

When $X_{t_1} = x_1,\ X_{t_2} = x_2, \ldots\ , X_{t_k} = x_k$ specified, other variables perform conditional probability distribution functions:

$$P\left(x_{k+1}, t_{k+1}; \ldots; x_n, t_n | x_1, t_1; \ldots; x_k, t_k\right) = \frac{P_n(x_1, t_1; \ldots; x_k, t_k; x_{k+1}, t_{k+1}; \ldots; x_n, t_n)}{P(x_1, t_1; \ldots; x_k, t_k)}) \quad (3)$$

This is the probability distribution $X_{t_{k+1}}, \ldots\ , X_{t_n}$, in which $x_1, \ldots\ , x_k$ act as parameters. Take the t_i in chronological order, then the process is Markov if this conditional probability depends only on the last value of the x_k in the t_k and does not depend on the previous values $x_{i<k}$. This should be done for all n, for any choice k, for any $t_1, \ldots\ , t_k$ and $x_1, \ldots\ , x_k$. If as well as P_1 and P_2 we can constructing all P_n. For example,

$$\begin{gathered} P_3\left(x_1, t_1; x_2, t_2; x_3, t_3\right) = P\left(x_3, t_3 | x_1, t_1; x_2, t_2\right) P_2\left(x_1, t_1; x_2, t_2\right) = \\ = P\left(x_3, t_3 | x_2, t_2\right) P(x_2, t_2 | x_1, t_1) P_1(x_1, t_1). \end{gathered} \quad (4)$$

The function for the Markov process $P(x_2, t_2|x_1, t_1)$ makes sense of the probability of moving from one state to another. For non-Markov processes, distribution functions (3) are defined by a completely different mathematical construct. In a one-dimensional space for responding to an opponent's dii, this is actually a symmetrical random wandering [9] with the probability of transition:

$$P\left(i, t+1 | i', t\right) = \frac{1}{2}\delta_{i,i'+1} + \frac{1}{2}\delta_{i,i'-1}$$

Here t and x takes entire values i . But suppose that the application of a cyber weapon tends to maintain the direction of movement: the probability of p in (1) to apply a cyber weapon, and q in (2) to avoid. Then the X_t is no longer Markov, since the probability of X_t depends not only on x_{t-1}, but also on x_{t-2}. This can be corrected by entering the two-component variable $\{X_t,\ X_{t-1}\}$. This common variable again becomes Markov with the probability of transition:

$$P\left(i_1, i_2, t+1 | {i_1}', {i_2}', t\right) = \delta_{i_2, {i_1}'}[p\delta_{i_1 - i_2, {i_1}' - {i_2}'} + q\delta_{i_1, {i_2}'}]$$

If we remember all the steps of our opponent's elms, that is, when we had the opportunity to apply cyber weapons (the process includes more previous steps), additional variables are needed. However, this no longer works if the memory extends to all previous steps.

Take the Markov process, in which t is the time when X_t takes discrete values $i\ =\ 0, 1,\ 2,\ \ldots$. In equation (4), take $t_3 = t_2 + \Delta t$:

$$\frac{P_3(i_1, t_1; i_2, t_2; i_3, t_2 + \Delta t)}{P_1(i_1, t_1)} = P(i_3, t_2 + \Delta t | i_2, t_2) P(i_2, t_2 | i_1, t_1)$$

Let's sum i_2 and take the threshold to get the basic equation:

$$\dot{P}\left(i,t|i_1,t_1\right)=\sum_{i'}W_{i,i'}P\left(i',t|i_1,t_1\right)-W_{i',i}P\left(i,t|i_1,t_1\right) \tag{5}$$

where $W_{i,i'}$ are probabilities of transition per i unit of time and are properties belonging to the physical system (for example, squares of matrix elements), while P refers to the state of the system.

Or the basic equation with memory [8],

$$\dot{P}\left(i,t|i_1,t_1\right)=\int_{t_1}^{t}dt\prime\sum_{i'}W_{i,i'}\left(t-t'\right)P\left(i',t'|i_1,t_1\right)-W_{i',i}\left(t-t'\right)P\left(i,t'|i_1,t_1\right), \tag{6}$$

with the statement that it determines the mathematical model of handling the probability of a strike and a possible contr-attack.

Let x also be continuous. Then the $W_{i,i'}$ takes the form of an integral core $W(x\mid x\prime)$. In our case, the process is such that during an infinitesimal Δt only infinitesimal jumps are possible, so the nucleus is reduced to a differential operator. We will use an analogy that perfectly describes our situation: the diffusion equation for the coordinate x of the Brownian particle [11]:

$$\frac{\partial P(x,t)}{\partial t}=D\frac{\partial^2 P(x,t)}{\partial x^2} \tag{7}$$

The solution of this equation given with the initial condition $P\left(x,t_1\right)=\delta(x-x_1))$ is will a probability of transition $P(x,\ t\ |x_1,\ t_1)$.

Consider one-dimensional diffusion in the potential field by the formula (7). Let it take place in a finite environment $x_a<x<x_c$ (Fig. 1). When we begin to apply cyber weapons at the inside point of the x_b, what are the chances that it will come out in x_a or x_c, respectively? That is, who will be the first to react to attacker or defender, or given the situation

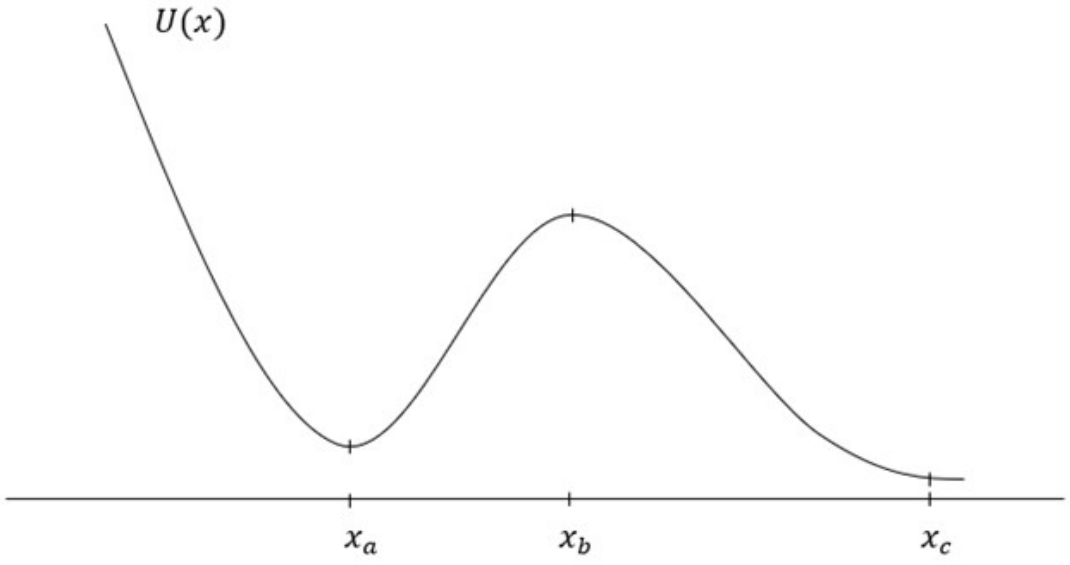


Figure 1: One-dimensional reaction in the attacker moment; Source: own

The answer is obtained by solving (7) with the marginal absorption conditions: $P(x_a,\ t)\ =\ 0$ and $P(x_c,\ t)\ =\ 0$ .The solution can be obtained explicitly due to the fact that in the equation for probabilities time is not included. It is clear that when the x_b is at the top of the highest maximum $U(x)$ the probability of exit will be equal to fifty to fifty.

In our model, the coordinate is not Markov, and therefore it is not enough to know that $x(t_1) = x_b$: you also need to know its previous history. For example, if you want to calculate the autocorrelation function x:

$$< x\,(t_1)\,x\,(t_2) >= \int x_1 x_2 P\,(x_1, t_1; x_2, t_2)\,dx_1 dx_2 =$$

$$\int x_1 x_2 P\,(x_1, t_1) P(x_2, t_2 | x_1, t_1)\,dx_1 dx_2$$

You can also find the average time for any exit [9]. Obviously, to do this, you need to know the correct initial distribution of the $P_1(x, t_1)$). Remarque. For the non-Markov process, the initial value problem is not clearly defined unless additional information about the problem is provided. In setting our problem is the question of going beyond the threshold, such as x_a in (Fig. 2). How long does it take to break the barrier?

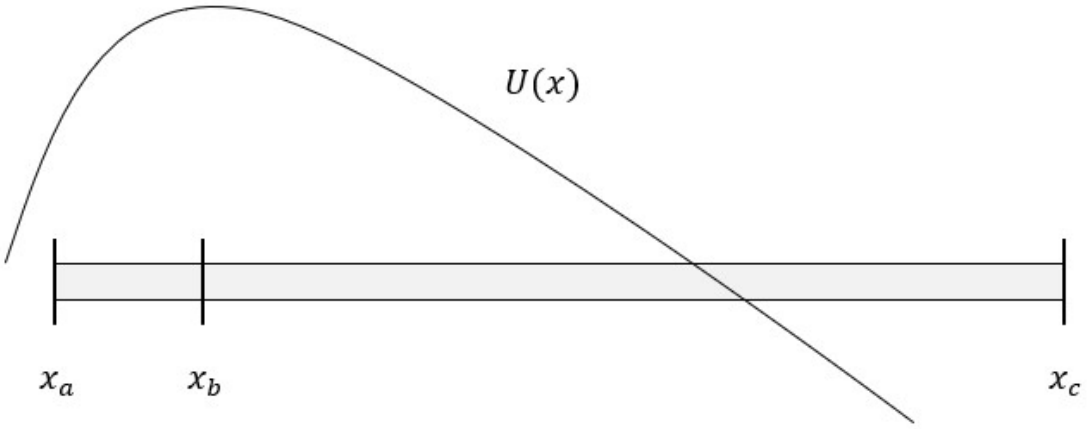


Figure 2: Out of the Threshold; Source: own

In the case of diffusion described in (7), you can take the average time of the first arrival in x_b– and multiply by 2, because in x_b the same probability of avoidance (the ability to hide) or return and respond. The average time of the first passage can again be calculated analytically. However, in larger dimensions, the question is: How long does it take to get out of the minimum x_a? This average time is determined by the lowest minimum on the curve.

The policy question is how to choose T to maximize the value of the cyber weapon. Because rates are not under our control, our best policy is to wait until the rates are high enough to risk losing the cyber weapon therefore of its limited reliability. This means our best policy can be expressed in terms of x(t)- the Thresholds of rates that will cause us to use the cyber weapon.

The value to the owner of a cyber weapon to exploit a target's vulnerability depends on its p and q, and the distribution of future rates as specified in equations (6) or (7).

We have an equation for the value of a cyber weapon for exploiting a target's vulnerability, we can evaluate what that weapon is worth. Even more useful is that we can calculate the best way to use the weapon in terms of the optimal Threshold, specifying how large the rates have to be to make it worthwhile to use the cyber weapon and take the added risk that it will no longer be available.

2 Application to Cyber incident

2.1 Cyber Espionage

Some country Army (called country R) has for years been deploying cyber weapons for espionage against the defense and industrial targets, in country A. Their cyber espionage often has only moderate Reliability against vigilant targets, so it is frequently discovered. It is able to continue because many of the targets have not maintained state-of-the-art defenses for known vulnerabilities.

A result of widely detected industrial espionage was hostility against this country's government. Country officials acknowledge that all countries spy on each other, but they say country R is unique in its theft of foreign technology. In terms of our model, one might well ask why the countries R are deploying their weapon for cyber exploitation now when the rates are not particularly high, rather than wait for a time when the rates are much higher? In other words, why might country R be operating with a low Threshold? One possibility is that it might have thought that the resource they were deploying had a low shelf life (low P). Another reason might be that country R expected high q against at least some targets because it has taken outliers several years to even detect that they have been compromised.

2.2 Country R Use of a High-Reliability, Low-Stability Cyber weapon.

In [13,14] and other sources to illustrate a situation in which the timing of the employment of a resource does not seem optimal, consider the case of the Countries R halt of its rare-earth exports pressure to provide strong economic pressure against J Country. The cyber incident started in, for example, September 3010, when a country R fishing trawler collided with a J Countries patrol ship near some disputed islands. On the next day and again on some days, the Country R demanded that the captain and crew be released. The next day, J country released the crew but continued to detain the captain. Tension continued to escalate, and after two weeks, Country R abruptly halted its exports of rare-earth materials. Country R controlled 97 percent of piles of earth, and J Country imported one-half of that supply, the effects on J country of the cutoff were immediate and drastic. J country complained that this was economic warfare. Country R waited a month to restore exports to most of the world, and before restoring exports to J country. After this demonstration of economic coercion, J Country, the A country, and others invested in the production of rare earth outside of Country R so as to never be subject to the same threat again. Clearly, Country R had the ability to stop the global supply of minerals essential for manufacturing electronics and automobiles

In terms of our model, this ability had very high Reliability because until Counties R actually stopped exports, other countries were happy to shut down their own production in favor of the cheaper Countries R supply. Country R's dominance could have persisted for years. When Country R did use its coercive power, it tried to achieve some Reliability by never acknowledging that the cut-off had any political purpose. However, the timing was so obviously connected to the J country detention of the trawler captain that there was little doubt that bringing it to a halt was quite deliberate. Additionally, once Countries R did deploy this ability to coerce by isolating exports of

rare piles of earth, they lost their ability to coerce again in the same way because importing nations awoke to the chance and took effective measures to finish their total dependence on Countries R exports.

In terms of our model, Countries R's ability to coerce others with a cut-off of rare-earth exports would have had very low Reliability. The resource had high Reliability because Countries R's dominance had persisted for years already and would probably have persisted for several more years had the cyber weapon for coercion not been used when it had been. A cyber weapon with both low Reliability and high Stability features a very high optimal Threshold to be used. Our model suggests that Countries R would have been happier had they'd the patience to attend to a situation with much higher rates before deploying this particular low-Reliability and high-Stability cyber weapon for economic coercion.

CONCLUSION

The cyber incident has already begun. The exploitation of vulnerabilities in computer systems has been used for both espionage and sabotage. The exploitation of vulnerabilities has also led to new ways of conducting crime and fighting crime; maintaining anonymity and destroying anonymity, resisting political authority, and reinforcing political authority. Within the near future, the cyber conflict will likely allow international sanctions to be more precisely targeted than economic sanctions alone and can provide powerful force multipliers for so-called cyber warfare. This paper clarified a number of the important considerations on optimal timing for such use. This type of study can help users make better choices and help defenders better understand what they're up against. In some situations, one might want to mitigate the potential harm from a cyber incident, and in other situations, one might want to harness the tools of cyber conflict. In some cases, one might want to try both. In any case, a vital step is to know the logic inherent in this new domain. The implications of our model are easy to summarize: Reliability and Stability are both desirable properties of a cyber weapon. However, they need opposite effects on the simplest time to use cyberweapons. Persistence ends up in more patience, meaning the rates have to meet a better Threshold before the resource is worth using. The rationale is that with high Reliability you are doing not have to worry considerably about the resource becoming obsolete before you employ it. High Reliability, however, promotes use even with relatively low stakes because the resource is probably going to be reusable. Moreover, in an exceedingly world of exponential rates, the prospect of occasional very high Gains increases the brink because those very high stakes are more worth expecting. Turning the attitude around, it might be a slip-up to judge one's own vulnerability by what one sees when the rates are low or moderate. The potential attacker may be expecting an occurrence of sufficiently high states to take advantage of the Cyber Weapon it already has.

References

[1] J Clapper Statement for the Record: Worldwide Threat Assessment of the US Intelligence Community (US Senate Select Committee on Intelligence, Washington, DC, 2013).

[2] L Milevski, Stuxnet and strategy: A special operation in cyberspace. Joint Force Quarterly 63, 64–69 (2011).

[3] Robert Axelrod and Rumen Iliev.Timing of cyber conflict. PNAS, January 13, 2014 — 111 (4) 1298-1303 — https://doi.org/10.1073/pnas.1322638111
[4] RA Clarke, R Knake Cyber War: The Next Threat to National Security (Harper Row, New York, 2010).
[5] Dzhalladova Irada and Ruzickova Miroslava. Dynamical system with random structure and their applications. Cambridge Sientific Publishers, 2020. ISBN 978-1-908106-66-7.
[6] Dzhalladova I., Ruzickova M., Dacjuk M. Non Markovian process: postulates and model problems. Modelling and information systems in economics, No. 101, 2022, p. 122-127. ISSN 1511-1100. — https://doi.org/10.33111/mise.101.5
[7] Bashtinec Ja. Stability of the Zero Solution of Stochastic Differential Systems with Four-Dimensional Brownian Motion. In: *Mathematics, Information Technologies and Applied Sciences 2016, post-conference proceedings of extended versions of selected papers*. Brno: University of Defence, 2016, p. 7-30. [Online]. [Cit. 2017-07-26]. Available at: <http://mitav.unob.cz/data/MITAV2016Proceedings.pdf>.ISBN 978-80-7231-400-3.
[8] Kampen N.G., Stochastic Processes in Physics and Chemistry (North-Holland, Amsterdam 1981, 1992).
[9] Gardiner C.W., Handbook of Stochastic Methods (Springer, Berlin 1983).
[10] Kramers H.A., Physica 7, 284 (1940)
[11] Feller W., An Introduction to Probability Theory and its Applications I (2nd ed., Wiley, New York 1966) p. 323.
[12] Thomas P. Rona. Weapon Systems and Information War, the Official Home of the Department of Defense. — http://www.dod.gov
[13] Martin C. Libicki. What is Information Warfare? United States Government Printing, Washington DC, 1995. — http://www.dodccrp.org/files/Libicki
[14] Irada Dzhalladova and Miroslava Růžičková. Mathematical tools for creating models of information and communication network security. p. 55-63. In: *Mathematics, Information Technologies and Applied Sciences 2016, post-conference proceedings of extended versions of selected papers*. Brno: University of DefenceBrno: University of Defence, Brno, 2018, Available at: <http://mitav.unob.cz/data/MITAV2016Proceedings.pdf>ISBN 978-80-7582-065-5

Acknowledgement

The work presented in this paper has been supported by the MeMovII.EV 902004007/21400 c.p. (research project No. 02.2.69/0.0./0.0/18/053/0016962).

MICROCONTROLLERS IN LABORATORY PRACTICE

Michal Kuba[1], Soňa Pavlíková[1], Dagmar Faktorová[1], Peter Fabo[1,2]
[1]Faculty of Special Technology, Alexander Dubcek University of Trencin, Ku kyselke 469, 911 06 Trencin, Slovak Republic, michal.kuba@tnuni.sk
[2] Research Centre, University of Žilina, Univerzitná 8215/1, 010 26 Žilina, Slovak Republic

Abstract: *Standard laboratory practice in technical fields of research and development is associated with the use of universal as well as specialized instrumentation. Current technologies allow some information and data acquisition procedures to be implemented on modern microcontrollers. These data collection technologies can be extended with other options such as digital processing of measurement data directly by a microcontroller, distributed data collection, in situ measurement or use in IoT technology. The contribution is devoted to an overview of the possibilities of implementing the basic methods of laboratory practice on a microcontroller chip by directly using standard measurement procedures as well as their specialized modifications using specific features of the technology.*

Keywords: measurement, microcontrollers, data processing, data acquisition

INTRODUCTION

A modern microcontroller (MCU) is a complex device that has a significant processing power of a processor core, which is usually based on ARM technology [1, 2], which in some models is supplemented by a mathematical coprocessor. A standard parts of the MCU are built-in programmable peripherals [3] for data transfer and communication with the superior system, typically an USART serial modem, on more powerful MCU models also LAN and recently modems supporting RF communication protocols are integrated too. For communication with external peripherals within the device, the microcontroller usually includes an elaborate GPIO pin management system as well as SPI, I^2C and CAN serial communication interfaces. Analog signal processing is enabled by ADC converters with an analog multiplexer for channel selection with an option of resolution and sampling rate setting. For the analog signals generation some models contain one or several DAC converters. An important part of practically every microcontroller are timers designed for processing the time parameters of signals as well as for generating the time sequences. The clock frequency of the processor core, the properties of the internal circuits and peripherals of the microcontroller can be modified, changed, activated and deactivated over time, which makes it possible to effectively manage the energy consumption of the entire device. A typical feature of programmable peripherals is the possibility of configuring them in such a way that they can perform a significant part of the activities autonomously or with only minimal intervention of the processor core, which can perform other activities.

The mentioned features of MCU, their internal structure and the implementation of various types of peripherals resulted from the primary application area of microcontrollers [4] which is the control of electronic devices of industrial as well as consumer electronics, automotive, robotics and communication systems. The peripherals of modern MCUs are designed as universal, so we can think about the possibilities of their use in laboratory practice for collecting and processing information and data. The terminology regarding MCU from the STM family of microcontrollers, including other designations and acronyms used in this contribution refer to the STM32L476 datasheet and manual [5, 6].The MCU core clock

frequency used has been 80 MHz. The presented circuit diagrams are only schematic, aimed to clarify the fundamentals of the presented procedures and do not replace their complete technical documentation.

1 IMPLEMENTATION OF STANDARD MEASUREMENT METHODS USING MCU

Many standard measurement methods and procedures can be implemented directly on microcontrollers by replicating known and proven procedures. The advantage of this approach is the possibility of simple verification the method functionality, it's parameters analysis and the results and their accuracy verification.

1.1 Measurement of the signals time parameters

As an example, the article shows the direct implementation of standard method of signals time parameters measurement (pulse repetition frequency and period), which is one of the basic tasks of laboratory practice.

1.1.1 Events counting

Counting various types of events, including e.g. the number of limit state crossings or the number of impulses, is encountered quite often in practice.The implementation of the event counter depends on several requirements, such as number of pulses per atime unit, the dead time (time during which the counter does not respond to the next pulse), the way of processing events (reaction to the beginning or the end of the event) and so on.
For a simple counter registering random events with a low frequency, such as counting the flips of a rain gauge, counting pulses of a Geiger-Müller counter in environmental applications, etc., a simple counter with MCU interrupt generation with firmware support is sufficient (Fig. 1). The implementation uses the possibility of triggering an interrupt when the state of the microcontroller pin changes (in Fig. 1 pin PC13). The conditions for triggering an interrupt are set in the external interrupt control block (EXTI), in which the triggering of an interrupt on the rising, falling or both edges of the signal change can be set.

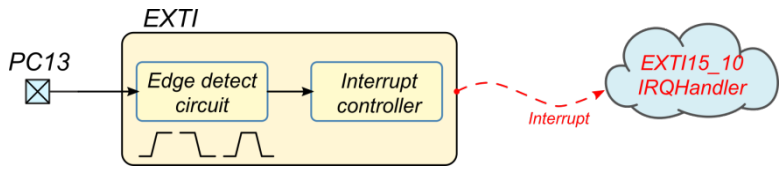


Fig. 1. Events counting by triggering an interrupt by event on a microcontroller pin.

The inherent dead time of an interrupt-implemented counter is approximately 500 ns. The signal on the input pin is assumed to have levels corresponding to binary logic levels (0 … 3.3V), but for some pins (in the documentation marked as FT) levels (0 … 5V) are also allowed.If different voltage levels from external signal sources occur, they need to be adjusted by means of suitable analog circuits. For example, in the case of long leads it may be necessary to protect the input from induced voltage with protective diodes. Alternatively, it may be required to detect pulses with a defined amplitude, in which case a comparator with level adjustment or a window detector has to be used.In the case of several counters collecting signals from several inputs, the priority of interrupts must be considered.

1.1.2 Frequency and period measurement

Integrated programmable universal timers of the microcontroller [6] allow to design of complicated circuits for signal processing and generation. The usual method of measuring the period of a pulse signal is counting the pulses with a known frequency by a counter that is gated by the measured signal.
There are usually several types of timers with different purposes and possibilities on a microcontroller chip. They are based on a GPT – General Purpose Timer, which provides a simple possibility of measuring the signal period without the intervention of the MCU core. When the signal edge arrives at the MCU input pin (TIMx_CH1), the timer generates signal to capture the counter value (this counter is marked as CNT counter).That edge is also used to generate the reset of the counter to zero value. The implementation of this measurement principle using GPT counter is shown in Fig. 2. The range of measured periods and the measurement resolution depends on the selection of the timer (CNT, 16/32 bit) and the frequency of the clock signal CK_CNT, which can be set in a wide range by pre-scalers. The accuracy of the measurement is determined by the stability of the oscillator from which the clock signal is derived.

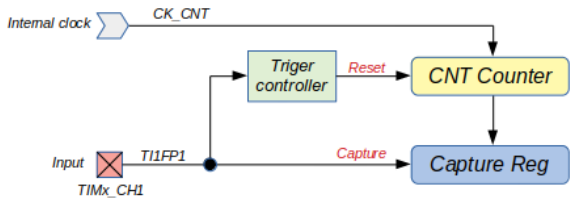


Fig. 2. Measuring the signal period using a universal timer counter.

For accurate frequency measurement using input signal gating method, the chained timers concept can be used. The principle of this method is shown in Fig. 3

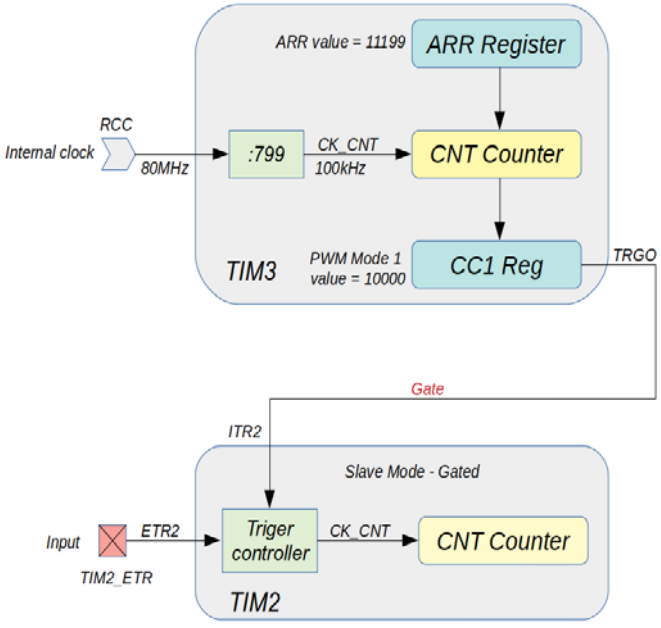


Fig. 3. Frequency measurement using chained timers.

Two chained timers are used for the implementation. The 16-bit timer creates a gating signal, the 32-bit timer is frequency counter. The gating signal is created in the PWM mode of the timer where it is possible to choose the gating time and the time interval for the next measurement. The upper limit of the measured frequencies is limited by the MCU production technology and usually is the same as the processor clock frequency, in the discussed case 80 MHz, with the stability of the used reference oscillator crystal usually in the order of 10^{-6}.
Since the configuration of the MCU peripherals is not static, but it is defined by program means, it is possible to change their configuration during the work of the MCU based on the measurement conditions. Therefore, it is possible, for example, to change the measurement of the period to the measurement of the frequency or vice versa, according to the required measurement accuracy.

1.2 Processing of analog signals

For analog signal processing, MCUs contain one or several analog to digital converters (ADC) [7], which are equipped with an input analog multiplexer.In principle, measurement of physical quantities using the MCU does not differ from standard procedures, when the input physical quantity is converted into a voltage. The voltage level represents an input that is subsequently amplified or weakened in a suitable way, and its bandwidth is adjusted by filtering in relation to the sampling frequency of the converter.Since the ADC converter is made up of a capacitive approximation register which also forms a sample & hold circuit, for measurement of faster signals it is necessary at the input of the ADC converter connect a low output impedance voltage follower. This follower also limits the input voltage range of the converter to the range (0 … 3.3V).
In some MCU types, the analog subsystem also includes digital-to-analog converters (DAC) [8] with relatively high output impedance, usually equipped with optional output amplifiers, but these do not cover the entire dynamic range of the converter for technological reasons. To use the full dynamic range of the DAC, it is suitable to use external rail-to-rail operational amplifiers.

2 OPTIMIZATION OF MEASUREMENT METHODS FOR MCU

The potential of MCU is not limited to replication of standard measurement procedures. With the help of specific MCU properties, it is also possible to implement measurement procedures optimized for use with the specific features of MCU peripherals.

2.1 Capacitance measurement using MCU

The capacitance measurement of the capacitor is among the standard procedures of laboratory practice when evaluating the dielectric properties of substances. The elementary procedure for measuring capacity is a principle based on measuring the time period after which the measured capacitor is charged or discharged to the selected predefined reference value from the defined initial value. The standard measurement procedure consists of measuring this duration and then charging or discharging the capacitor to the initial value. This time perioddepends on the capacity of the capacitor and inefficiently extends the measurement cycle.
To cut down on measurement time and enhance measurement evaluation, the charging and discharging times may be adjusted to restrict to active measurement times. This is enabled by

fundamental properties of the linear RC circuit, by which the charging and discharging times are the same if the measurement is starting from zero or maximum voltage values on the capacitor and the voltage on the capacitor is measured after reaching the reference value, which is half of the maximum voltage value [9].

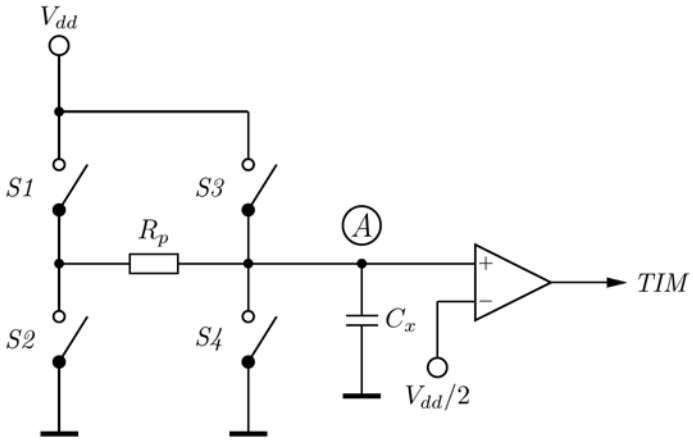


Fig. 4. Principle of capacity measurement.

The reference value is derived from the maximum voltage value to which the measured capacitor can be charged, thereby eliminating the influence of supply voltage fluctuations on the measurement accuracy. The principle of measurement is shown in Fig. 4.
At the beginning of the measurement cycle, the measured capacitor is assumed to be discharged and all switches are turned off. When the switch *S1* is turned on, the capacitor is charged through the resistor R_p, and after reaching the value of the voltage $V_{dd}/2$ on the capacitor, the measurement of the charging time ends. After turning the switch *S1* off and turning the switch *S3* on for a predefined time T_{S3}, the capacitor is charged to the maximum voltage value V_{dd}. After turning the switch *S3* off and turning the switch *S2* on, the capacitor begins to discharge through the resistance R_p to the final voltage value $V_{dd}/2$. When this voltage value is reached, the measurement of the discharge time ends. Next, when the switch *S2* is turned off and the switch *S4* is turned on for a predefined time T_{S4}, the capacitor is completely discharged. Next the switch *S4* is turned off and the measurement continues with the next measurement cycle.

The waveform of the voltage on the measured capacitor at point *A* in Fig. 4 with the described order of switching is shown in Fig. 5.

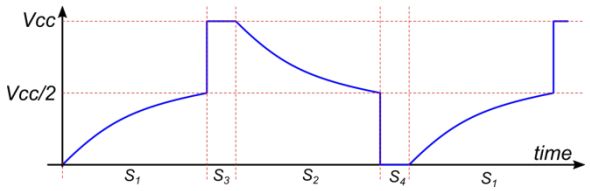


Fig. 5. Voltage waveform at point *A* in Fig. 4. (V_{cc} is V_{dd})

The total measurement time of one measurement cycle is determined using the following equation:

$$T = 2\,R_p\,C_x\ln(2) + T_{S3} + T_{S4}$$

The implementation of the aforementioned measurement principle is shown in Fig. 6. Switches *S1* – *S*4 are formed by driver output transistors of MCU ports, which are configured as output pins in digital mode. The output of the comparator, which is part of the MCU peripherals, controls the measurement of the measurement cycle period using the 32-bit timer TIM2. The RCC clock frequency of the timer counter is 80 MHz.
Controlling the measurement requires only minimal processor intervention. Within the interrupt handler routine, which is triggered by the comparator, the states of the individual switches are cyclically set, and the values of the timer counter are recorded. The configurable analog comparator is a part of the MCU peripherals, and the entire circuit requires only one external component – resistor R_p and uses 4 MCU pins.

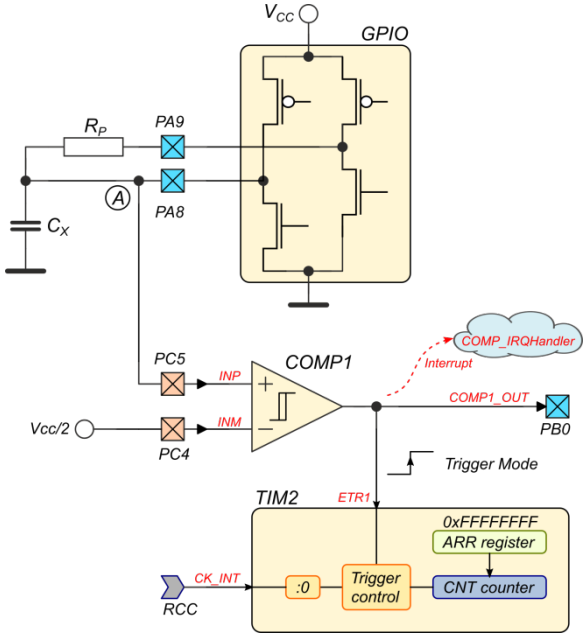


Fig. 6. Implementation of capacity measurement.

The maximum measured value of the capacity is limited by the current carrying capacity of the port pins when charging and discharging the measured capacitor, which can also be polarized. The calibration of the measurement is in one-shot, the stray and mounting capacitances C_s are determined by measuring without the connected calibration capacitor.This capacity is subtracted from the value of the measured capacity during corrections.

By using a 32-bit timer counter, the range of measured capacity values without the modification of parameters and circuit configuration is in the range of several orders of magnitude, in the realized version from 0.1 pF to 220 nF (Fig. 7).

2.2 Analog signals generation

The generation of signals with defined waveform, generation of harmonic signals with fixed or variable phase, the synchronization of parts of the measurement chain to the selected part of the generated waveform are tasks that frequently occur in the creation of complex measurement methods.

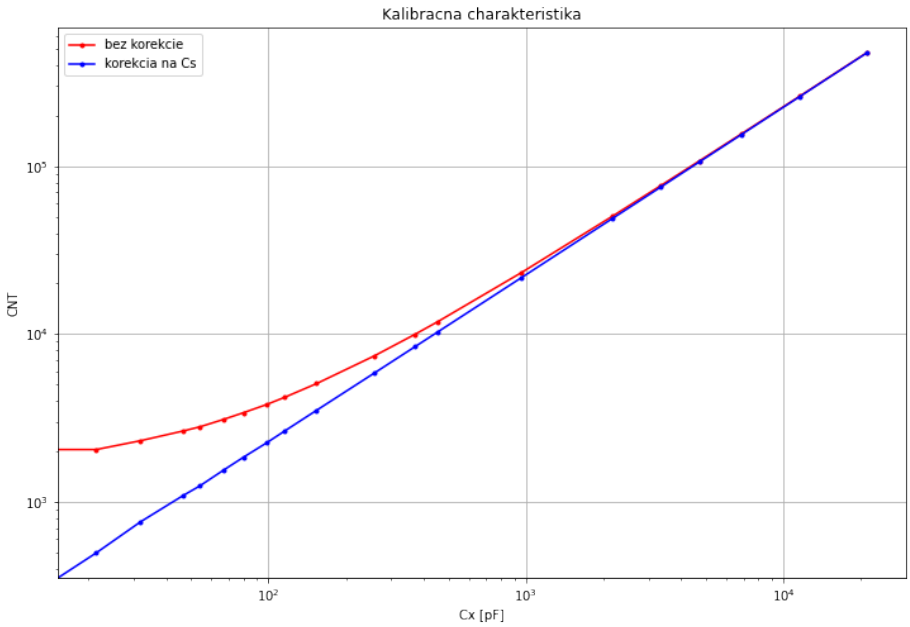


Fig. 7. Calibration curve with and without stray capacitance correction, the capacitance measurement is for R_p = 330 kΩ, the calibration capacitors values were measured using a GW Instek LCR 6020 reference bridge.

To generate the analog periodic signals with a constant frequency in the order of kHz, it is possible to use a pulse-width (PWM) modulator as a DAC converter, which is a standard part of microcontroller timers. The principle of operation of such DAC is based on the generation of pulses with variable width (PWM) [10], where the mean value of the generated sequence of pulses corresponds to the required voltage value of the generated waveform. An analog value is obtained from the generated sequence of pulses using an external RC low-pass filter, which is connected at the PWM output. An analog signal is obtained by changing the pulse width value at regular intervals according to the values stored in the microcontroller memory.

The mentioned concept has benefit in that the output voltage of the converter is in the range of the amplitude of the generated pulses, which is usually the range of the supply voltages of the microcontroller. On the other hand, the need to generate PWM pulses with a frequency in multiples of the frequency of the generated signal and the need to use an output filter limits the use of such a DAC to generate the slower periodic signals only. On the STM32 platform, it is possible to implement the signal generator concept with the use of PWM DAC through the timer and DMA transfer, so that it is not necessary to use any additional CPU intervention except for setting the basic parameters during initialization.

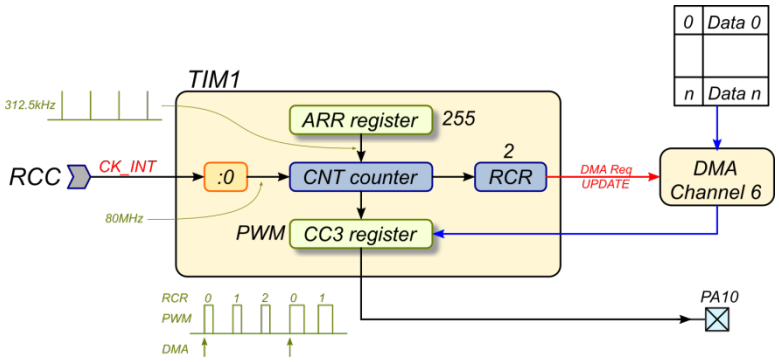


Fig. 8. Signal generator using a timer in PWM mode.

A simplified configuration of the DAC using PWM is shown in Fig. 8. The ARR register determines the resolution of the converter and together with the clock signal pre-scaler for the counter determine the repetition frequency of the PWM pulses. The pulse width at the timer output is determined by the value stored in the Compare register CC3. This value is modified by the DMA controller with the value from the table of values, which is stored in the MCU memory. The DMA transfer works in cyclic mode. After setting the last value from the table, the first value from the table is set in the next step.

If it is necessary to synchronize other parts of the system with the generated signal, the DMA controller enables the generation of an interruption after the transfer of half of the data block as well as at the end of its transfer. The signal waveform generated by the described method is shown in Fig. 9.

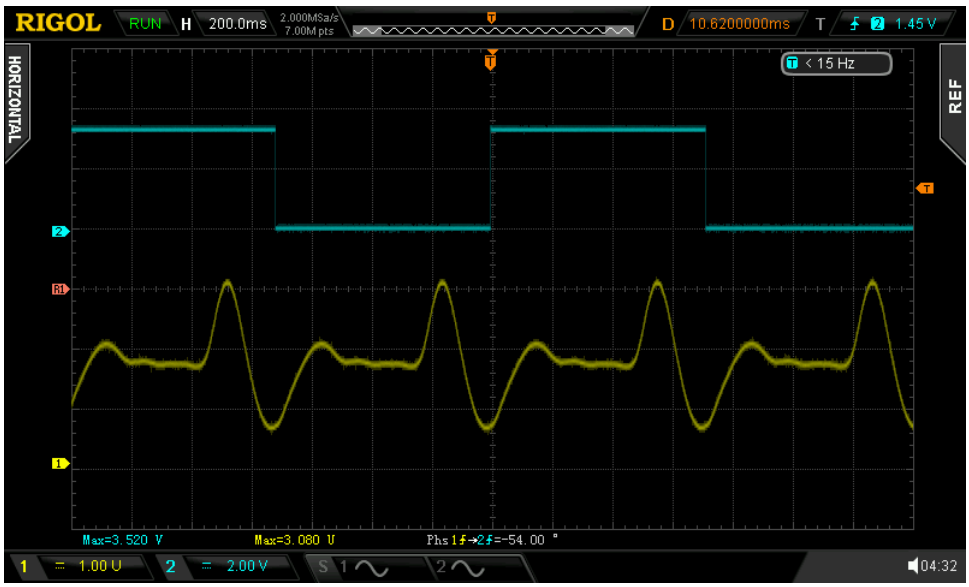


Fig. 9. Example of signal waveform generated by the timer in PWM mode.

CONCLUSION

The contribution in the overview describes the possibilities of applying modern MCU in laboratory practice in the field of research and development. Selected examples demonstrate the possibilities of simple implementation of standard measurement procedures as well as the

realization of specialized measurement procedures using features of the MCU. With the appropriate use of the software and the integrated MCU peripherals, it is possible to achieve results comparable to standard measuring instruments, while at the same time the entire device can be miniaturized.

References

[1] *Arm® Cortex®-M4 Processor, Technical Reference Manual*, Available at: <https://developer.arm.com/documentation/100166/0001/>
[2] Yiu J. *The Definitive Guide to the ARM Cortex-M3*, Elsevier Inc., 2007, ISBN: 978-0-7506-8534-4
[3] Brown, G. Discovering the STM32 Microcontroller, Available at: <https://www.st.com/content/st_com/ja/support/learning/stm32-education/text-books.html>
[4] *M68HC11 Reference Manual*, Motorola, 1996, Available at: <https://home.deec.uc.pt/~jlobo/tc/M68HC11_ref_man.pdf>
[5] *Datasheet - STM32L476xx - Ultra-low-power Arm® Cortex*, STMicroelectronics, 2020, Available at: <https://www.st.com>
[6] RM0351 - Reference manual, STMicroelectronics, 2020, Available at: <https://www.st.com>
[6] AN4013 Timer overview, STMicroelectronics, 2012, Available at: <https://www.st.com>
[7] AN3116 STM32s ADC modes and their applications, STMicroelectronics, 2010, Available at: <https://www.st.com>
[8] AN3216 Audio and waveform generation using the DAC in STM32 microcontroller families, STMicroelectronics, 2010, Available at: <https://www.st.com>
[9] Úžitkový vzor SK9356Y1, Senzor na priame meranie hodnoty kapacity kondenzátora, Vestník ÚPV SR č.: 20/2021
[10] AN4776 General-purpose timer cookbook for STM32 microcontrollers, STMicroelectronics, 2019, Available at: <https://www.st.com>

Acknowledgement

The authors are grateful to the VEGA agency (grant VEGA 2/0013/21).

UPPER DENSITY, QUASI-DENSITY OF SUBSET OF THE SETS OF NATURAL NUMBERS

Renáta Masárová
Faculty of Materials Science and Technology, Slovak University of Technology
Jána Bottu 25, 917 24 Trnava, Slovak Republic, renata.masarova@stuba.sk

Abstract: *In this paper, we show that the quasi-densities of subsets of the set of natural numbers $d_p(A)$ are not an upper density. We will show several properties of this density and the conditions that the sequence (p_n) must satisfy for the quasi-density $d_p(A)$ to exist and be finite. The relationship between asymptotic density and quasi-density of set A is described.*

Keywords: upper density, asymptotic density, statistical convergence, quasi-statistical convergence, quasi-density

INTRODUCTION

Let A be a subset of the set of natural numbers. The density of this set is a number that shows how densely are the elements of the set A distributed in the set of natural numbers [1,5,9]. Using densities, we can define the generalized convergence of sequences of real numbers. The most well-known of these densities is the asymptotic (natural) density, using which we define statistical convergence. In paper [8] the authors define generalization of statistical convergence – quasi-statistic convergence. This paper is focused on the properties of the quasi-density by which this convergence is defined. It also shows how quasi-density differs from the most well-known densities such as a asymptotic density and a logarithmic density.

1 THE UPPER DENSITY

In paper [4] the authors defined the term upper density.
Let $P(N)$ be a set of all subsets of N. A function $\mu: P(N) \to R$ is called the upper density on the set N, if for every set $A, B \subseteq N$ and positive integer constants k and h, the following holds true:
(V1) $\mu(N) = 1$
(V2) $\mu(A) \leq \mu(B)$ for $A \subseteq B$
(V3) $\mu(A \cup B) \leq \mu(A) + \mu(B)$
(V4) $\mu(k \cdot A) = \frac{1}{k}\mu(A)$, where $k \cdot A := \{ka: a \in A\}$
(V5) $\mu(A + h) = \mu(A)$, where $A + h := \{a + h: a \in A\}$.

From (V1) and (V2) we get
(V1*) $\mu(A) \leq 1$ for all $A \subseteq N$,
Combining (V4) and (V5) results in
(V4*) $\mu(k \cdot A + h) = \frac{1}{k}\mu(A)$ for every $A \subseteq N$ and $k, h \in Z^+$.

These criteria are satisfied by all commonly used densities. The most well-known of these are [2,4,7]:

1. Asymptotic density

Let $A \subseteq N$. We define $A(n) = |k \in A, k \leq n|$ as the number of elements of a set A smaller than n. The upper and lower asymptotic densities of the set $A \subseteq N$ are

$$\overline{d}(A) = \limsup_{n\to\infty} \frac{A(n)}{n} = \limsup_{n\to\infty} \frac{|A \cap \{1,2,\dots,n\}|}{n}$$

$$\underline{d}(A) = \liminf_{n\to\infty} \frac{A(n)}{n} = \liminf_{n\to\infty} \frac{|A \cap \{1,2,\dots,n\}|}{n}.$$

If $\overline{d}(A) = \underline{d}(A)$, then there exists an $\lim_{n\to\infty} \frac{A(n)}{n} = d(A)$ that is called the asymptotic density of the set A. It is evident that if for some set A there exists a $d(A)$, then $0 \leq d(A) \leq 1$.
While for every set there exists both an upper and a lower asymptotic density, the asymptotic density of the set A does not necessarily exist.

1. Logarithmic density

The upper and lower logarithmic densities are

$$\overline{\ell}(A) = \limsup_{n\to\infty} \frac{\sum_{a_i \in A, a_i \leq n} \frac{1}{a_i}}{\ln n}$$

$$\underline{\ell}(A) = \liminf_{n\to\infty} \frac{\sum_{a_i \in A, a_i \leq n} \frac{1}{a_i}}{\ln n}.$$

If $\overline{\ell}(A) = \underline{\ell}(A)$, then there exists an $\ell(A) = \lim_{n\to\infty} \frac{\sum_{a_i \in A, a_i \leq n} \frac{1}{a_i}}{\ln n}$ that is called the logarithmic density of the set A.
Between the lower and upper asymptotic and logarithmic densities $A \subseteq N$ there are the following relations:

$$0 \leq \underline{d}(A) \leq \underline{\ell}(A) \leq \overline{\ell}(A) \leq \overline{d}(A) \leq 1.$$

3. Schnirelmann density

It is defined (in contrast to the preceding two densities) for every set $A \subseteq N$. The Schnirelmann density for the set A is

$$\delta(A) = \inf_{n\geq 1} \frac{A(n)}{n} = \inf_{n\geq 1} \frac{|A \cap \{1,2,\dots,n\}|}{n}.$$

This density is „sensetive“ to changes that we make at the beginning of a set.
E.g. if $1 \notin A$, then $\delta(A) = 0$, if $2 \notin A$, then $\delta(A) = \frac{1}{2}$ [2]. For every set $A \subseteq N$ the following applies

$$0 \leq \delta(A) \leq \underline{d}(A) \leq 1.$$

In literature, we can find other densities (Dirichlet, Banach, exponential, …)
Using densities we can define different types of convergences of the sequences.

We say that $x = (x_n)$ statistically converges to the number $L \in R$, if $\forall \varepsilon > 0: d(N_\varepsilon) = 0$, where $d(N_\varepsilon) = \{k \epsilon N: |x_k - L| \geq \varepsilon\}$.
Many authors have generalized this convergence by replacing the statistical density with a different type of density [2,3,5]. In the next chapter, we will focus on a density by the use of which one of the generalizations was defined.

2 QUASI-DENSITIES OF SUBSETS OF THE SET OF NATURAL NUMBERS

In [8] the authors defined:
Let $p = (p_n)$ be a sequence of positive real numbers with the properties:
i) $\lim_{n\to\infty} p_n = +\infty$
ii) $\limsup_{n\to\infty} \frac{p_n}{n} < +\infty$

We say that the sequence $x = (x_n)$ quasi-statistically converges to the number $L \in R$ ($\mathrm{stq}_p - \lim x_k = L$), if $\forall \varepsilon > 0: d_p(E_\varepsilon) = \lim_{n\to\infty} \frac{1}{p_n} |\{k \epsilon E_\varepsilon, k \leq n\}| = 0$,
where $E_\varepsilon = \{k \epsilon N, |x_k - L| \geq \varepsilon\}$.

In case we choose the sequence $p = (p_n)$ to be a sequence of all natural numbers, we get a statistical convergence.

Definition: Let $p = (p_n)$ be a sequence of positive real numbers that satisfies the following properties:
i) $\lim_{n\to\infty} p_n = +\infty$
ii) $\limsup_{n\to\infty} \frac{p_n}{n} < +\infty$.

We will call such a sequence permissible.

Sequences that satisfy these properties are for example:
1. $(p_n) = (\log n)_{n=1}^{\infty}$,
2. $(p_n) = (n \cdot \alpha + d)_{n=1}^{\infty}, \alpha \in R^+, d \in R$,
3. $(p_n) = (n^\alpha)_{n=1}^{\infty}, \alpha \in (0,1)$.

We will use this sequence to define the density.

Definition: Let $p = (p_n)$ be a permissible sequence. The lower quasi-density of the set $A \subseteq N$ is

$$\underline{d_p}(A) = \limsup_{n\to\infty} \frac{A(n)}{p_n}.$$

The upper quasi-density of the set $A \subseteq N$ is

$$\overline{d}_p(A) = \liminf_{n\to\infty} \frac{A(n)}{p_n}.$$

In case the upper and lower quasi-densities of the set A are equal, there exists a quasi-density of the set A and we denote it as $d_p(A) = \underline{d_p}(A) = \overline{d_p}(A)$ and

$$d_p(A) = \lim_{n\to\infty} \frac{A(n)}{p_n}.$$

Note: If we choose $(p_n) = (n)_{n=1}^{\infty}$, we get an asymptotic density.

The basic properties of density $d_p(A)$ are described in paper [6]. We will show that this density does not satisfy all the criteria for an upper density.

Theorem 1. Let $A \subseteq N$ be a finite set. Then $d_p(A) = 0$ for every permissible sequence $p = (p_n)$.

Proof: If A is a finite set, then $d_p(A) = \lim_{n\to\infty} \frac{1}{p_n} |\{k\epsilon A, k \le n\}| \le \lim_{n\to\infty} \frac{|A|}{p_n} = 0.$

Theorem 2. Let $A, B \subseteq N$ be non-empty sets which their quasi-densities $d_p(A)$ and $d_p(B)$. Let $d_p: \mathcal{P}(N) \to \langle 0, \infty)$ be a function. Then

i) If $A \subseteq B$, then $d_p(A) \le d_p(B)$

ii) $\underline{d_p}(A \cup B) \le d_p(A) + d_p(B)$

iii) If $\mathrm{A} \cap \mathrm{B} = \emptyset$, then $\mathrm{d_p}(\mathrm{A} \cup \mathrm{B}) = \mathrm{d_p}(\mathrm{A}) + \mathrm{d_p}(\mathrm{B})$.

Proof: i) Let $A \subseteq B$. Then for every $n \in N$ the following holds true $|\{k\epsilon A, k \le n\}| \le |\{k\epsilon B, k \le n\}|$.

Thus $\frac{1}{p_n} |\{k\epsilon A, k \le n\}| \le \frac{1}{p_n} |\{k\epsilon B, k \le n\}|$.

We get $\lim_{n\to\infty} \frac{1}{p_n} |\{k\epsilon A, k \le n\}| \le \lim_{n\to\infty} \frac{1}{p_n} |\{k\epsilon B, k \le n\}|$, i.e. $d_p(A) \le d_p(B)$.

ii) It is evident that $|\{k\epsilon A \cup B, k \le n\}| \le |\{k\epsilon A, k \le n\}| + |\{k\epsilon B, k \le n\}|$.

From that we get

$$\underline{d_p}(A \cup B) = \liminf_{n\to\infty} \frac{1}{p_n} |\{k\epsilon(A \cup B), k \le n\}| \le \limsup_{n\to\infty} \frac{1}{p_n} |\{k\epsilon(A \cup B), k \le n\}| \le$$

$$\le \limsup_{n\to\infty} \frac{1}{p_n} (|\{k\epsilon A, k \le n\}| + |\{k\epsilon B, k \le n\}|) \le$$

$$\le \limsup_{n\to\infty} \frac{1}{p_n} |\{k\epsilon A, k \le n\}| + \limsup_{n\to\infty} \frac{1}{p_n} |\{k\epsilon B, k \le n\}| =$$

$$\overline{d_p}(A) + \overline{d_p}(B) = d_p(A) + d_p(B).$$

iii) If $A \cap B = \emptyset$ then $|\{k\epsilon A \cup B, k \le n\}| = |\{k\epsilon A, k \le n\}| + |\{k\epsilon B, k \le n\}|$.

$$\overline{d_p}(A \cup B) \le d_p(A) + d_p(B) = \liminf_{n\to\infty} \frac{1}{p_n} |\{k\epsilon A, k \le n\}| + \liminf_{n\to\infty} \frac{1}{p_n} |\{k\epsilon B, k \le n\}| \le$$

$$\le \liminf_{n\to\infty} \frac{1}{p_n} |\{k\epsilon(A \cup B), k \le n\}| = \underline{d_p}(A \cup B) \le \overline{d_p}(A \cup B).$$

.

It follows from that

$$d_p(A \cup B) = d_p(A) + d_p(B).$$

Quasi-density does not satisfy the criterion (V1). The quasi-density of the set of all natural numbers N is dependent on the sequence (p_n), and does not necessarily exist for some permissible sequences.

Example 1. Let sequence $p = (p_n) = \left(\sqrt{n}\right)_{n=1}^{\infty}$. This sequence is permissible.

In regards to this sequence the quasi-density of set $N = \{1,2, \dots\}$ is

$$d_p(N) = \lim_{n\to\infty} \frac{|N \cap \{1,2,\dots,n\}|}{p_n} = \lim_{n\to\infty} \frac{n}{\sqrt{n}} = \infty.$$

Example 2. We define the sequence $p = (p_n)$:

$$p_n = \begin{cases} 2n, & n = 2k \\ \dfrac{n}{2}, & n = 2k+1 \end{cases} \quad k = 1,2,\dots$$

This sequence is sequence permissible. The quasi-density of set $N = \{1,\ 2, \dots\}$ given the sequence p is

$$d_p(N) = \lim_{n\to\infty} \frac{|N \cap \{1,2,\dots,n\}|}{p_n} = \lim_{n\to\infty} \frac{n}{p_n}.$$

Quasi-density of set N given by the sequence p does not exist, because

$$\lim_{k\to\infty} \frac{2k}{p_{2k}} = \lim_{k\to\infty} \frac{2k}{4k} = \frac{1}{2},$$

$$\lim_{k\to\infty} \frac{2k+1}{p_{2k+1}} = \lim_{k\to\infty} \frac{2k+1}{\frac{2k+1}{2}} = 2.$$

For the quasi-density of set N the following theorem holds true:

Theorem 3. Let (p_n) be a permissible sequence. Then

i) If $\limsup_{n\to\infty} \frac{p_n}{n} = T \neq 0$, then $\overline{d_p}(N) = \frac{1}{T}$ (if $\limsup_{n\to\infty} \frac{p_n}{n} = 1$, then $\overline{d_p}(N) = 1$).

ii) If $\limsup_{n\to\infty} \frac{p_n}{n} = 0$, then $\overline{d_p}(N) = \infty$.

Proof:

i) $\overline{d_p}(N) = \limsup_{n\to\infty} \frac{|N \cap \{1,2,\dots,n\}|}{p_n} = \limsup_{n\to\infty} \frac{n}{p_n} = \frac{1}{\limsup_{n\to\infty} \frac{p_n}{n}} = \frac{1}{T}$.

ii) $\overline{d_p}(N) = \limsup_{n\to\infty} \frac{|N \cap \{1,2,\dots,n\}|}{p_n} = \limsup_{n\to\infty} \frac{n}{p_n} = \frac{1}{\limsup_{n\to\infty} \frac{p_n}{n}} = \infty$.

Note: Let $p = (p_n)$ be a permissible sequence and exists a finite $\lim_{n\to\infty} \frac{p_n}{n}$.

i) In the case of $\lim_{n\to\infty} \frac{p_n}{n} = l \neq 0$, then $d_p(N) = \frac{1}{l}$.

ii) In the case of $\lim_{n\to\infty} \frac{p_n}{n} = 0$, then $d_p(N) = \infty$.

iii) In the case of $\lim_{n\to\infty} \frac{p_n}{n} = 1$, then $d_p(N) = 1$.

The quasi-density does not satisfy the criterion (V1*). For any $A \subseteq N$ the following holds:
$0 \leq \underline{d_p}(A) \leq +\infty$,
$0 \leq \overline{d_p}(A) \leq +\infty$, i.e. the quasi-density can be a value greater than 1.

Theorem 4. Let the following hold true for sequences $p = (p_n)$

$$0 < \liminf_{n\to\infty} \frac{p_n}{n} \leq \limsup_{n\to\infty} \frac{p_n}{n} = T < \infty.$$

Then for any sequence $A \subseteq N$

$$0 \leq \underline{d_p}(A) \leq \overline{d_p}(A) \leq \frac{1}{T}$$

holds true.

Proof: $0 \leq \underline{d_p}(A) = \lim\limits_{n\to\infty} \inf \frac{|k\epsilon A, k\leq n|}{p_n} = \lim\limits_{n\to\infty} \inf \frac{n}{p_n} \cdot \frac{|k\epsilon A, k\leq n|}{n} \leq \lim\limits_{n\to\infty} \sup \frac{n}{p_n} \cdot \frac{|k\epsilon A, k\leq n|}{n} =$
$= \lim\limits_{n\to\infty} \sup \frac{|k\epsilon A, k\leq n|}{p_n} = \overline{d_p}(A).$
In addition to that $\overline{d_p}(A) = \lim\limits_{n\to\infty} \sup \frac{n}{p_n} \cdot \frac{|k\epsilon A, k\leq n|}{n} \leq \frac{1}{T} \cdot \overline{d}(A) \leq \frac{1}{T}.$
It is sufficient to realize that for every set $A \subseteq N$ there exists a $\underline{d}(A)$ and $\overline{d}(A)$ (an asymptotic density $d(A)$ does not have to exist).

Theorem 5. Let $A \subseteq N$ be such a set, for which its asymptotic density is $d(A) = m$, where $m\epsilon\langle 0,1\rangle$. Let there exists a non-zero $\lim\limits_{n\to\infty} \frac{p_n}{n} = l$. Then there also exists a quasi-density of set A and $d_p(A) = \frac{1}{l} \cdot m$ holds true.
Proof: When we use the definition of quasi-density we get the following
$d_p(A) = \lim\limits_{n\to\infty} \frac{1}{p_n} |\{k\epsilon A, k \leq n\}| = \lim\limits_{n\to\infty} \frac{n}{p_n} \cdot \frac{1}{n} |\{k\epsilon A, k \leq n\}| = \frac{1}{l} \cdot m.$

Corollary. Let $p = (p_n)$ be any arithmetic sequence of the type $p_n = n \cdot \alpha + d$, $n = 1,2, \ldots, \alpha\epsilon R^+, d\epsilon R$.
Let $A \subseteq N$ be such a set, that its asymptotic density $d(A) = m$.
Then $d_p(A) = \frac{m}{\alpha}$.

If the condition in the theorem is not satisfied, then the asymptotic density and quasi-density of a given set, even if they both exist, do not have to be the same.

Example 3. Let $p_n = \log n, n = 2,3, \ldots$. It is evident that a sequence (p_n) defined as such is permissible, as $\lim\limits_{n\to\infty} \log n = \infty$ and $\lim\limits_{n\to\infty} \frac{\log n}{n} = 0$.
Let us consider the sets $A = \{1^2, 2^2, \ldots\}$ and $B = N = \{1,2, \ldots\}$.
The asymptotic densities of these sets are $d(A) = 0$ and $d(B) = 1$.
The quasi-density of these sets in regards to the previously defined sequence (p_n) exists and is the same for both: $d_p(A) = \infty$ a $d_p(B) = \infty$.

Theorem 6. For every non-negative real number t there exists such a set $A \subseteq N$ and a permissible sequence $p = (p_n)$, that $d_p(A) = t$.
Proof: If $t = 0$, then we can choose A to be any finite set (theorem1).).
Let $t \in (0, \infty)$, and let us choose any $m \in \langle 0,1\rangle$.
For these chosen immutable numbers, we define a sequence $p = (p_n) = \left(\frac{m}{t} \cdot n\right)_{n=1}^{\infty}$. This sequence is permissible, because
$\lim\limits_{n\to\infty} p_n = \lim\limits_{n\to\infty} \frac{m}{t} \cdot n = +\infty$ and $\lim\limits_{n\to\infty} \sup \frac{p_n}{n} = \lim\limits_{n\to\infty} \frac{\frac{m}{t} \cdot n}{n} = \frac{m}{t} < +\infty.$
For every $m \in \langle 0,1\rangle$ the exists such a set $A \subseteq N$, for which its asymptotic density is $d(A) = m$.
Let A be such a subset of natural numbers, such that its asymptotic density is m and $(p_n) = \left(\frac{m}{t} \cdot n\right)_{n=1}^{\infty}$. Then
$d_p(A) = \lim\limits_{n\to\infty} \frac{1}{p_n} |\{k\epsilon A, k \leq n\}| = \lim\limits_{n\to\infty} \frac{t}{m \cdot n} |\{k\epsilon A, k \leq n\}| = \frac{t}{m} \lim\limits_{n\to\infty} \frac{|\{k\epsilon A, k\leq n\}|}{n} = \frac{t}{m} m = t.$

CONCLUSION

In this paper, we have shown that the quasi-density of subsets of the set of natural numbers has only some of the properties of an upper density. The density of the set of all natural numbers is always equal to 1. In this paper there were examples showing that the quasi-density of the set of all natural numbers $d_p(N)$ does not necessarily have to exist. If it does exist, it can have a value greater than 1(e.g., it can be $d_p(N) = \infty$). Under certain conditions (see Theorem 5) a quasi-density can be described using an asymptotic density. However, this does not hold true in general. Even if a quasi-density defines a certain type of generalized convergence (a quasi-statistical convergence), the paper has shown that its properties are different than other densities that were used so far.

References

[1] Baláž, V. On Generalized Notion of Convergence by Means of Ideal and Its Applications. In: *Mathematics, Information Technologies and Applied Sciences 2017, post-conference proceedings of extended versions of selected papers*. Brno: University of Defence, 2017, p. 9-20. [Online]. Available at: <http://mitav.unob.cz/data/MITAV 2017 Proceedings.pdf>. ISBN 978-80-7582-026-6.

[2] Baláž, V., Šalát, T. Uniform density u and corresponding I_u- convergence. *Mathematical Communications*, No. 11, 2006, p. 1-7.

[3] Baláž, V., Visnyai, T. I- convergence of aritmetical functions *Number Theory and Its Applications, Intech Open Limited,* London, 2020, p. 399-4211.

[4] Filipów, R., Tryba, J. Densities for sets of natural numbers vanishing on a given family. *Journal of Number Theory*, No. 211, 2020, p. 371-382.

[5] Kostyrko, P., Šalát, T., .; Wilczynski, W. I- convergence. *Real Anal. Exchange*, No. 26, 2000/2001, p. 669-686.

[6] Masárová, R., Visnyai, T., Vrábeľ, R. Quasi-Density of Sets, Quasi-Statistical Convergence and the Matrix Summability Method. *Axioms,* No. 11(3), 2022, p. 1-11.

[7] Niven, I. The asymptotic density of sequences. *Bull. Amer. Math. Soc.*, No. 57(6), 1951, p. 420-434.

[8] Ozguc, I. S., Yurdakadim, T. On quasi-statistical convergence. *Commun. Fac. Sci. Univ. Ank. Series A1*, No. 61(1), 2012, p. 11-17.

[9] Visnyai, T. Remarks on Compact Submeasures. In: *Mathematics, Information Technologies and Applied Sciences 2015, post-conference proceedings of extended versions of selected papers*. Brno: University of Defence, 2015, p. 156-161. [Online]. Available at: <http://mitav.unob.cz/data/MITAV 2015 Proceedings.pdf>. ISBN 978-80-7231-436-2.

Acknowledgement

This publication is the result of the project implementation: Strategic research in the field of SMART monitoring, treatment and preventive protection against coronavirus (SARS-CoV-2), supported by the Operational Programme Integrated Infrastructure funded by the European Regional Development Fund.

MINIMIZATION OF PARALLEL PHASES PERFORMED IN THE NUMERICAL COMPUTATION OF A CERTAIN TYPE OF PDE

Martin Nehéz
Institute of Computer Science and Mathematics,
Faculty of Electrical Engineering and Information Technology,
Slovak University of Technology in Bratislava
Ilkovičova 3, 812 19 Bratislava,
Slovak Republic
martin.nehez@stuba.sk

Abstract: *In this paper, we derive a finite-difference formula which solves a certain type of a second order linear partial differential equation. It enable us to accelerate the parallel algorithm which numerically computes the corresponding finite-difference equation.*

Keywords: finite difference, stencil, graph coloring, parallel processing.

INTRODUCTION

Finite difference algorithms are one of the traditional numerical methods commonly used to solve differential equations. Such algorithms are exploited both for ordinary and partial differential equations (abbreviated to PDEs). The main idea behind the finite difference method is based on the approximation of values at certain grid points which are derived from a spatial domain of a given continuous function. Such a principle is referred to as *discretization* [6].

All numerical methods are usually tailored with respect to computer-aided feasibility and the finite difference method is no exception. Requirements on computer resources needed for numerical methods are often very high, however. Due to this fact, the usage of parallel (or high-performance) computers for such tasks is highly desirable. Roughly, parallelization schemes usually rely on partitioning the computational problem in question into minor subproblems. In such a way, each processor is responsible for processing a corresponding subproblem. Comparing to sequential algorithms, exploitation of the above idea often leads to speedup of computation. Indeed, the design of suitable parallel algorithms requires a nontrivial knowledge and skills [7].

In this paper, we describe a method for speed up of traditional parallel schemes for certain type of partial differential equations solved by finite-difference method. Such a method is based on the combination of linear algebra and graph coloring. It is explained on an example of a certain type of second order partial differential equation and its discretization. In our example, the described method enables to reduce the number of phases in corresponding parallel algorithm from 4 to 2. It may lead to considerable acceleration of a computation compared to the original algorithm.

The organization of the paper is as follows. Section 1 contains definitions of necessary notions, basic notations and a description of useful principles and methods. The bulk of this paper, namely Section 2, is devoted to the proof of our main result. Our contribution is summarized in Conclusion.

1 PRELIMINARIES AND THE PROBLEM STATEMENT

1.1 Graphs, Coloring and Stencils

Standard graph-theoretic notions are mentioned here without definitions. We address [2, 3] for further definitions and results regarding the graphs and coloring, respectively. Nevertheless, we will mention some useful abbreviations here. Namely, for a positive integers m, n, let K_n denote an n-vertex complete graph and let $K_{m,n}$ denote a complete bipartite graph with two partitions of cardinality m and n, respectively.

A vertex coloring of graphs is introduced more precisely. Let $G = (V, E)$ be a simple undirected graph, consisting of a set of vertices V and a set of edges E. Let t be a positive integer and let $C = \{c_1, c_2, \ldots, c_t\}$ be a set of colors. A function $f : V(G) \to C$ is said to be a *legal vertex coloring of* G if $f(u) \neq f(v)$ for each edge $\{u, v\} \in E(G)$. Clearly, each legal vertex coloring of G is assigning different colors to each adjacent vertices. Instead of the "legal vertex coloring" we will refer only "legal coloring", as we deal with no other kinds of colorings. A graph is said to be t-colorable if it can be legally colored with at most t different colors. The *chromatic number* of G, abbreviated to as $\chi(G)$, is the minimum t such that G is t-colorable. If a clique K_i ($i \geq 2$) is a subgraph of G, then $\chi(G) \geq i$.

Let us explain the notion of "dependency graph" in the following part. Given a set of objects S and a transitive relation $R \subseteq S \times S$, *the dependency graph* is a directed graph $G = (S, T)$ with $T \subseteq R$ the transitive reduction of R [1]. More precisely, let $\mathcal{P}$ be a standard procedural programming language with an assignment operator denoted by "=" and with variables $x_1, x_2, \ldots$; let S be a set of such variables. If φ denotes an expression, syntactically correct in $\mathcal{P}$, with d variables (d is an positive integer) such that:

$$x_k \ = \ \varphi(x_1, \ldots, x_d);$$

then $(x_k, x_i) \in R$ iff x_i occurs in φ for $i = 1, \ldots, d$. It means that relation the (x_k, x_i) represents a dependency "x_k *dependns on* x_j" for every expression φ occurring in a given program written in $\mathcal{P}$. Each relation (x_k, x_i) is represented by the directed edge (or *arc*) $x_k \leftarrow x_i$ in a corresponding dependency graph G.

To avoid conflicts, dependency graphs are used as standard tools in concurrency theory [1]. Commonly used alternatives to dependency graphs in numerical analysis are "stencils" [4], [7]. However, unlike dependency graphs, stencils are undirected. More precisely, the *stencil* is a set of points (i.e. vertices, nodes) used for computing a finite difference by an assignment statement "$y = \varphi(\ldots);$" around a point y. A construction of stencils is similar to the dependency graphs, but all arcs in a given dependency graph must be undirected and, if necessary, all multiple occurrences of the same edge will be deleted.

1.2 A Discretiation Principle and Its Parallelization

Let us consider the partial differential equation (shortly PDE) in the form

$$\frac{\partial^2 u(x,y)}{\partial x^2} + \frac{\partial^2 u(x,y)}{\partial x \partial y} + \frac{\partial^2 u(x,y)}{\partial y^2} = g(x,y)\ , \tag{1}$$

where g is a function of independent variables x, y and u is a function (or dependent variable). Eq. (1) is frequently written in the equivalent form

$$u_{xx} + u_{xy} + u_{yy} = g(x, y). \tag{2}$$

Note that it is an elliptic PDE, since $A = B = C = 1$ and $B^2 - 4AC = -3 < 0$. In order to get a numerical solution of eq. (1), we will use a finite difference method. By Taylor polynomial for functions of two variables, we get [6]:

$$u_{xx} + u_{xy} + u_{yy} \;=\; \frac{1}{2h^2}\,[\,u(x+h, y+h) + u(x-h, y-h) +$$

$$+\, u(x+h, y) + u(x-h, y) + u(x, y+h) + u(x, y-h) - 6u(x, y)\,] + O(h^2)$$

This approximation leads to the discretization within 2-dimensional grid $n \times n$ (see Fig. 1), where $u_{i,j}$ stands for $u(x, y)$, $u_{i+1,j}$ stands for $u(x+h, y)$, $u_{i,j-1}$ stands for $u(x-h, y)$, etc. We address [4], for advanced issues of elliptic PDEs discretization.

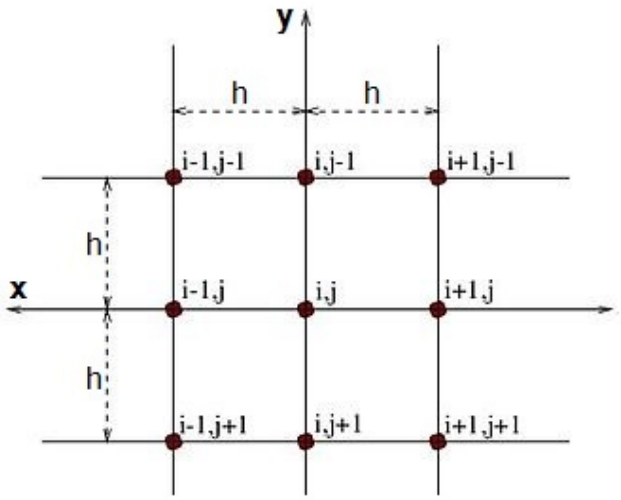


Fig. 1. Arrangement and labeling of a grid for numerical computation of the PDE (1).
Source: own

By a straightforward calculation, it follows:

$$u_{i,j}(t+1) \;=\; \frac{1}{6}\left[u_{i+1,j+1}(t) + u_{i-1,j-1}(t) + u_{i+1,j}(t) + u_{i-1,j}(t) + u_{i,j+1}(t) + u_{i,j-1}(t) - \frac{2g_{i,j}}{n^2}\right],$$

where $h = 1/n$ and t is an order of iteration during the algorithm computation. The above assignment statement leads to the 7-point stencil, see Fig. 2. The same stencil is also used for another factorizations described in [4], p. 1830. Alternatively, more accurate formula is suggested in [7] which leads to the 9-point stencil (called also *2D compact stencil*), see Fig. 2.

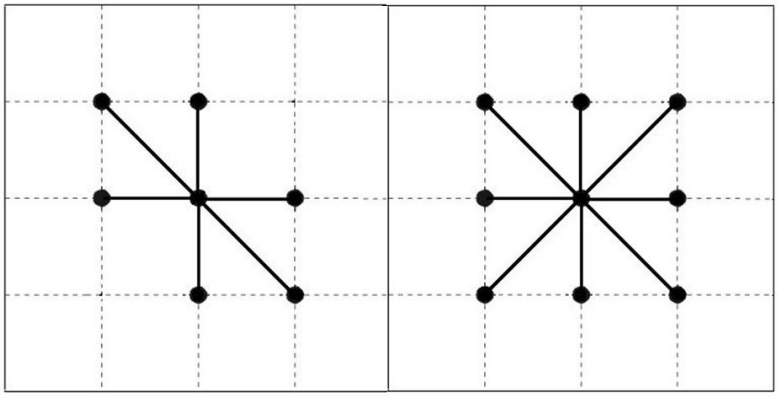


Fig. 2. The 7-point stencil (left) and the 9-point stencil (right). The stencils are isomorphic to two complete bipartite graphs, i.e. $K_{1,6}$ (left) and $K_{1,8}$ (right). [Source: own]

Stencils form patterns which repeatedly occur in the entire grid. Due to correct computation of each value $u_{i,j}$, the computation should proceed in phases. The values that can be computed simultaneously (within the same phase) must not be connected by an edge (in the same stencil). All such values must be "far enough" apart. That can be ensured by a legal vertex coloring of all points in the entire grid. An example of a legal vertex coloring with 4 colors for a grid with occurrence of 9-point stencils is shown in Fig. 3. Such a coloring is optimal (i.e. the chromatic number of the grid is $\chi = 4$) because the grid contains cliques K_4. (On the other hand, if a grid is formed by 7-point stencils, then its chromatic number is 3; corresponding figure is omitted.)

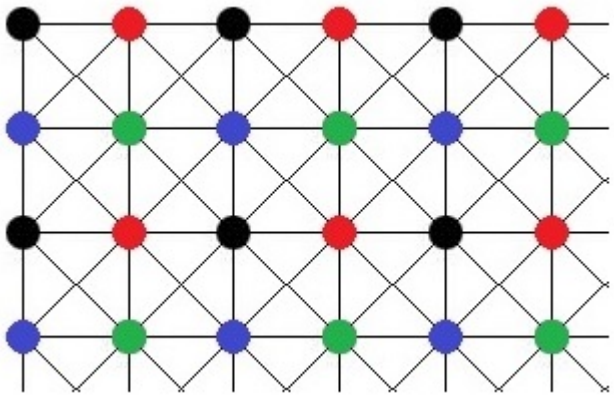


Fig. 3. The grid formed by 9-point compact stencil with chromatic number 4.
Source: own

The parallelization of the described numerical algorithm heavily rely on the legal coloring. Each phase corresponds to a single color. It means that the computation for each point with the same color can be realized in parallel. On the other hand, the use of different colors ensures to avoid collisions during computation. Such an idea leads to the Algorithm 1 with k phases. Clearly, the number of phases corresponds to the chromatic number of the corresponding grid. Of course, the number of phases affects the time complexity of the described algorithm. Moreover, in each phase only approximately $(1/k)$th of nodes is active while others are idle. (The details on how processors can be mapped into nodes are not discussed in this paper.) It is naturally advantageous to achieve

the smaller number of phases. In this regard, it is desirable to minimize the number of points (or simplify their layout) in a corresponding stencil. This is exactly the principal goal of our paper. Our solution is presented in the next section.

Algorithm 1 The parallel algorithm with k phases

repeat

Phase 1: **for all** grid points with color 1 **do in parallel** compute update;

...

Phase k: **for all** grid points with color k **do in parallel** compute update;

until the accuracy of the solution is not sufficient

2 THE MAIN RESULT

The main result of this paper is formulated in the following statement.

Theorem 1 *There exists a finite-difference equation which solves the PDE (1) numerically. Such an equation leads to the* 5*-point stencil and the corresponding parallel algorithm uses* 2 *phases.*

Proof. Let us use the general form of the Taylor series of a two-variable function u at a point (x, y):

$$u(x+\epsilon, y+\delta) \;=\; \sum_{k=0}^{\infty} \left\{ \frac{1}{k!} \sum_{i=0}^{k} \binom{k}{i} \frac{\partial^k u(x,y)}{\partial x^{k-i}\, \partial y^i} \cdot \epsilon^{k-i} \delta^i \right\} . \tag{3}$$

In the rest of this paper, all partial derivatives are written in compact forms, namely, the first-order partial derivatives are denoted by u_x and u_y, respectively, the second-order partial derivatives are denoted by u_{xx}, u_{yy}, u_{xy}, etc. By setting $\pm h$ for both ϵ and δ, 4 instances are derived from the Taylor series (3). The first one (the setting $\epsilon = \delta = h$ is chosen) is written in the form:

$$u(x+h, y+h) = u(x,y) + (u_x + u_y)h + \left(\frac{1}{2}u_{xx} + u_{xy} + \frac{1}{2}u_{yy}\right) h^2 +$$

$$+ \left(\frac{1}{6}u_{xxx} + \frac{1}{2}u_{xxy} + \frac{1}{2}u_{xyy} + \frac{1}{6}u_{yyy}\right) h^3 + O(h^4)$$

All four instances of eq. (3) differ only in signs of their right-hand side summands. The details are listed schematically in Tab. 1. In order to get a finite-difference formula solving eq. (1), we will use only the 5-point stencil with quantities $u_{i,j}, u_{i+1,j+1}, u_{i-1,j+1}, u_{i+1,j-1}, u_{i-1,j-1}$, respectively. Such a formula will be derived by approximation of functions $u(x+h, y+h), u(x-h, y+h), u(x+h, y-h), u(x-h, y-h)$, respectively. To do so, it is desirable to find such integer weights w_i (for $i = 1, \ldots, 4$), that the linear combination of $u(x+h, y+h), \ldots, u(x-h, y-h)$ should be proportional to the left-hand side of eq. (1).
More precisely,

$$w_1 \cdot u(x+h, y+h) + \cdots + w_4 \cdot u(x-h, y-h) \;=\; \tau \cdot h^2 \left(u_{xx} + u_{xy} + u_{yy}\right) , \tag{4}$$

for a suitable integer τ. Finding of suitable weights $w_1, \ldots, w_4$ can be handled by Tab. 1. Note that weighted sum of each column (except the second one) should yield either 0 or τ. More precisely,

Table 1: Signs of particular summands in 4 instances which were derived from eq. (3). Due to shortening, symbol u stands for $u(x, y)$ on the top of the 2^{nd} column.

$u(\cdot,\cdot)$	u	u_x	u_y	$\frac{1}{2}u_{xx}$	u_{xy}	$\frac{1}{2}u_{yy}$	$\frac{1}{6}u_{xxx}$	$\frac{1}{2}u_{xxy}$	$\frac{1}{2}u_{xyy}$	$\frac{1}{6}u_{yyy}$	weight
$u(x+h, y+h)$	+	+	+	+	+	+	+	+	+	+	w_1
$u(x-h, y+h)$	+	−	+	+	−	+	−	+	−	+	w_2
$u(x+h, y-h)$	+	+	−	+	−	+	+	−	+	−	w_3
$u(x-h, y-h)$	+	−	−	+	+	+	−	−	−	−	w_4

sums of all columns for $\frac{1}{2}u_{xx}, u_{xy}, \frac{1}{2}u_{yy}$ have to be τ. All other sums (except the column regarding u) should be 0. Columns in which the the occurrence of signs is the same are omitted. It follows that columns which regards u_x, $\frac{1}{6}u_{xxx}$ and $\frac{1}{2}u_{xyy}$, respectively, lead to the equation

$$w_1 - w_2 + w_3 - w_4 = 0 \, .$$

By the similar argument, the following equation is derived with respect to the columns regarding u_y, $\frac{1}{2}u_{xxy}$ and $\frac{1}{6}u_{yyy}$, respectively,

$$w_1 + w_2 - w_3 - w_4 = 0 \, .$$

Clearly, both u_{xx} and u_{yy} occur τ times in eq. (4). Therefore both corresponding columns lead to the equation

$$\frac{1}{2}(w_1 + w_2 + w_3 + w_4) = \tau \, .$$

Finally, the following equation is derived by taking into account the column regarding u_{xy}

$$w_1 - w_2 - w_3 + w_4 = \tau \, .$$

All four equations above are listed in a matrix form. Construction of the corresponding matrix is shown in Fig. 4.

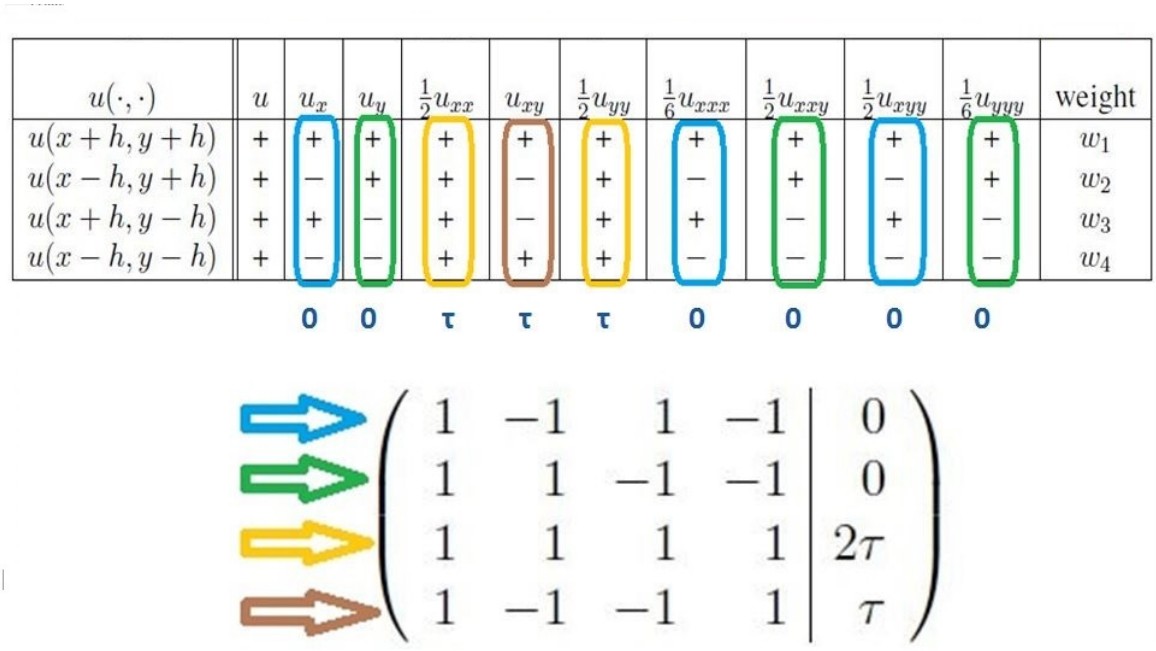


Fig. 4. Construction of the 4-row matrix from Tab. 1.
Source: own

By straightforward matrix row operations, we get

$$\left(\begin{array}{cccc|c} 1 & -1 & 1 & -1 & 0 \\ 1 & 1 & -1 & -1 & 0 \\ 1 & 1 & 1 & 1 & 2\tau \\ 1 & -1 & -1 & 1 & \tau \end{array}\right) \sim \left(\begin{array}{cccc|c} 1 & -1 & 1 & -1 & 0 \\ 0 & 1 & -1 & 0 & 0 \\ 0 & 0 & 1 & 1 & \tau \\ 0 & 0 & 0 & 4 & 3\tau \end{array}\right) .$$

Moreover, it holds that $w_1 = w_4$ and $w_2 = w_3$ and thus, the solution is

$$w_1 = w_4 = \frac{3\tau}{4} ,$$

$$w_2 = w_3 = \frac{\tau}{4} .$$

In order to obtain the smallest positive integer solution, we choose $\tau = 4$ and therefore $w_1 = w_4 = 3$ and $w_2 = w_3 = 1$. By knowing all weights, the resulting approximation is written in the following form:

$$3u(x+h, y+h) + u(x-h, y+h) + u(x+h, y-h) + 3u(x-h, y-h) =$$

$$= 4u(x,y) + 4h^2 \left(u_{xx} + u_{xy} + u_{yy}\right) + O(h^4) . \tag{5}$$

By transformation into finite differences and by substitution $h = n^{-1}$, we get

$$u_{i,j}(t+1) \;=\; \frac{1}{4}\left[\, 3u_{i+1,j+1}(t) + u_{i-1,j+1}(t) + u_{i+1,j-1}(t) + 3u_{i-1,j-1}(t) \,\right] - \frac{4g_{i,j}}{n^2} .$$

Such a formula leads to the 5-point stencil $K_{1,4}$ with only diagonal edges. The corresponding grid is 2-colorable (see Fig. 5) and thus the parallel algorithm uses 2 phases. □

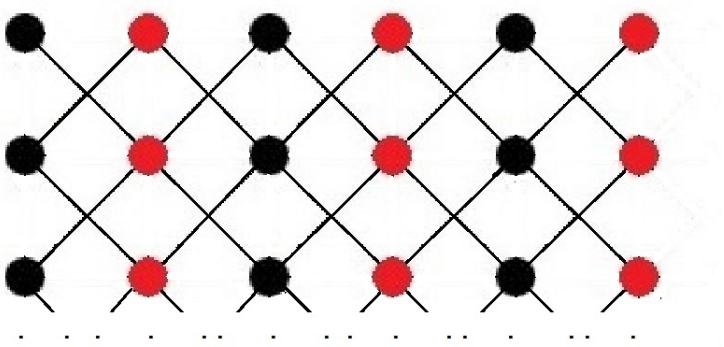


Fig. 5. The grid formed by 5-point stencils with chromatic number 2. It is a result of above proof.
Source: own

CONCLUSION

A commonly used approach to parallelization discretization-based numerical algorithms that solve PDEs consists of adapting a parallel scheme to a selected stencil-based numerical model [6], [7]. As discussed in Sec. 1.2, this one the parallel computation may be inefficient with respect to a fraction of the number of active and idle nodes in each phase.
We chose a different approach in this paper. Namely, we design a new discretization scheme for the partial differential equation of the second order formulated in Sect. 1.2. Our scheme is aimed at minimization of idle nodes in each parallel phase and it leads to a 5-point stencil with exclusively diagonal connections. Resulting finite-difference formula enables to reduce number o phases of a corresponding parallel algorithm to 2. Comparing to the standard parallel algorithms, which uses either 4 or 3 phases, such a result represents the significant improvement. The argument in our proof is based on the combination of standard ideas, such as exploitation of Taylor series and linear algebra, with graph coloring of stencils.

References

[1] Aalbersberg, I.J., Rozenberg, G. Theory of traces. *Theoretical Computer Science*, 60, 1988, p. 1-82. ISSN 0304-3975.

[2] Formanowicz, P., Tanaś, K. A survey of graph coloring - its types, methods and applications. *Foundations of Computing and Decision Sciences*, 37, No. 3, 2012, p. 223-238. ISSN 0867-6356.

[3] Gross, J.L., Yellen, J. *Graph Theory and Its Applications.* Boca Raton: Chapman & Hall/CRC, 2005, 800 pp. ISBN 978-1584885054.

[4] Jay Kuo, C.-C., Levy, B.C. Discretization and Solution of Elliptic PDEs-A Digital Signal Processing Approach. *Proceedings of the IEEE*, 78, No. 12, 1990, p. 1808 - 1842. ISSN 0018-9219, e-ISSN: 1558-2256.

[5] Mazumder, S. *Numerical Methods for Partial Differential Equations: Finite Difference and Finite Volume Methods.* Cambridge, Massachusetts: Academic Press, Elsevier, 2016, 462 pp. ISBN 978-0128498941.

[6] Tomov, S. Discretization of PDEs and Tools for the Parallel Solution of the Resulting Systems. The University of Tennessee 2015, 29 pp. [Online]. [Cit. 2022-09-22]. Available at: `https://netlib.org/utk/people/JackDongarra/WEB-PAGES/SPRING-2015/lect10-pdes.pdf`

[7] Tvrdík, P. *Parallel Algorithms and Computing.* Prague: Vydavatelství ČVUT, 2009, 200 pp.

FIXED PRINCIPAL PAYMENT AMORTIZATION SCHEDULE AND LINEAR DIFFERENCE EQUATIONS

Dana Říhová
Faculty of Military Leadership, University of Defence
Kounicova 65, 662 10 Brno, Czech Republic
dana.rihova@unob.cz

Abstract: *The contribution deals with an application of difference equations in the field of finance. It focuses on the loan schedule with fixed principal payments. The formulas for creating an amortization schedule are derived using the simple linear difference equations. In particular, we derive how to calculate the amount of the installment, the amount of interest and also the loan balance in each payment period. Unlike financial mathematics, where the sequence properties are used, all necessary formulas are obtained by solving difference equations. It is shown that recursive relations between two consecutive elements of the sequences appearing in the fixed principle amortization actually constitute the first order linear difference equations with constant coefficients. To find the rules for calculating an arbitrary element of such sequences means to solve these difference equations.*

Keywords: fixed principal payment, loan amortization, linear difference equation.

INTRODUCTION

The values of most financial products are given as a sequence of values observed at discrete time intervals and a number of models can be expressed by the recursion between two consecutive elements of the sequence where the initial value of the first element is given. However, the formulas used in finance represent the rule for calculating an arbitrary element of such a sequence and are usually derived from the recursion using properties of this sequence. The mentioned recursive rule actually constitutes a difference equation ([8] and [5]) and finding the appropriate formula leads to solving the difference equation. We encounter difference equations not only in finance but also in life insurance ([15], for example.

Loan repayment means a gradual regular repayment of the provided amount of money over a certain period of time. The each periodic payment consists of part of interest and part of principal, more detailed information can be found in [10], [16] or [18]. The loan can be repaid with constant annuities when installments are still the same, but the proportion of interest and principal changes during the each repayment. This case is discussed similarly in [14]. However in the following paper we deal with the case of the constant principal part when differently sized installments are used.

A fixed principal payment loan has a declining payment amount. The principal portion of the payment is the same and the interest portion is less each period due to the declining principal balance. Thus the constant principal plus the declining interest amount result in a declining periodic payment. The amortization schedule with fixed principal payments provides in the table the particular

amount of interest, the periodic payment and the outstanding debt in each payment during the entire loan repayment period.

1 LOAN REPAYMENT WITH FIXED PRINCIPAL

Suppose that the loan D is to be repaid also with interest in a total of n periodic payments, which are due at the end of each interest period. Let r stand for the annual interest rate and k denote the number of interest periods in a year. Thus the fraction $\frac{r}{k}$ means the interest rate per one interest period and will be denoted i. We will not consider any tax on interest income for simplicity. The loan repayment scheme described by the difference equations is shown in [3], [9], [2] and [4].

Let D_j represent the outstanding debt after the j-th payment. Each repayment consists of a constant principal and a variable interest, which depends on the loan balance. The principal M remains the same

$$M = \frac{D}{n} \tag{1}$$

and the interest decreases with increasing j. The new levels of debt form an arithmetic sequence and we gradually get them by subtracting the principal M. Thus the loan repayment process can be described by the following recursive relation

$$D_{j+1} = D_j - M, \qquad j = 0, 1, 2, \ldots, n-1$$

where the initial element is equal to the initial amount of debt

$$D_0 = D. \tag{2}$$

However the above recursion also represents the first order nonhomogeneous linear difference equation with constant coefficients (see [13] or [1])

$$D_{j+1} - D_j = -M, \qquad j = 0, 1, 2, \ldots, n-1. \tag{3}$$

1.1 Loan Balance in Amortization Schedule

To express an arbitrary element D_j of the loan balance sequence we can use the properties of the arithmetic sequence or as we show we need to solve the corresponding difference equation. Using the principle of superposition, we obtain the general solution of (3) by summing the general solution to the appropriate homogeneous equation and an arbitrary particular solution to the nonhomogeneous equation, for more information see [7] and [12], [17], [11].

The root of the corresponding characteristic equation of (3) is real number 1, thus the general solution of the appropriate homogeneous difference equation is given by a constant $C \in \mathbb{R}$. As the root 1 is a part of the right-hand side of (3), which is the constant $-M = -M \cdot 1^j$, then a particular solution can be estimated by the expression jb, where constant $b \in \mathbb{R}$, for details see [6].

By substituting jb into (3) and modifying it for b we get

$$b = -M.$$

Consequently the general solution of (3) takes form the sum

$$D_j = C - jM, \quad \text{where} \quad C \in \mathbb{R}.$$

Constant C can be determined from the initial condition (2) and choosing $j = 0$ into mentioned above equation

$$C = D.$$

Therefore the general solution of (3) can be expressed as

$$D_j = D - jM.$$

Combining with (1) we obtain the search formula for the outstanding debt after the j-th payment

$$D_j = \frac{D}{n}(n - j), \qquad j = 1, 2, \ldots, n, \tag{4}$$

which is used in the fixed principal amortization schedule to determine the loan balance. In the special case we have with index $j = 1$

$$D_1 = \frac{D}{n}(n - 1) \tag{5}$$

and with $j = n$

$$D_n = 0, \tag{6}$$

thus the total debt is actually repaid in a total of n installments.

1.2 Interest in Amortization Schedule

In this section we derive difference equation for calculation of interest. Let U_j denote the amount of interest on the j-the payment. As the interest is always paid from the remaining part of the debt at each period, we can write

$$U_{j+1} = iD_j, \qquad j = 0, 1, 2, \ldots n - 1 \tag{7}$$

and specially for $j = 0$ we have

$$U_1 = iD. \tag{8}$$

Multiplying the equation (3) by interest rate i we get

$$iD_{j+1} - iD_j = -iM.$$

Using (7) and the above relation we obtain by renumbering the following recursion

$$U_{j+1} - U_j = -iM, \qquad j = 1, 2, \ldots, n - 1. \tag{9}$$

This recurrence again represents the first order nonhomogeneous linear difference equation with constant coefficients.

With similar considerations as in the previous part according to the superposition principle, its general solution is the sum of the general solution to the appropriate homogeneous equation, which is constant $C \in \mathbb{R}$, and an arbitrary particular solution to the nonhomogeneous equation, which can be estimated by jb, where constant $b \in \mathbb{R}$. It is because the right side of the equation can be expressed in the form $-iM \cdot 1^j$ and the number 1 is the root of corresponding characteristic equation of (9). Further details concerning the solution of such linear difference equations you can find in [6].

Substituting jb into (9) we obtain

$$b = -iM.$$

Hence we can write general solution of (9) as follows

$$U_j = C - ijM, \quad \text{where} \quad C \in \mathbb{R}. \tag{10}$$

The constant C can be now specified considering the equality (8) and the above equation for $j = 1$

$$iD = C - iM,$$

which implies

$$C = i\,(D + M).$$

Combining the previous relation, (1) and (10) we get the following formula for calculating the amount of interest in the j-th payment

$$U_j = \frac{D}{n}\, i\,(n - j + 1), \qquad j = 1, 2, \ldots, n. \tag{11}$$

In the case of the last installment the amount of interest is expressed by

$$U_n = \frac{D}{n}\, i. \tag{12}$$

1.3 Periodic Payment in Amortization Schedule

Now we again use the difference equations to determine the amount of the regular installment, let us denote it as A_j. It arises from the sum of interest and principal in each payment, so it applies

$$A_j = U_j + M, \qquad j = 1, 2, \ldots, n. \tag{13}$$

Expressing U_j from the above relation

$$U_j = A_j - M$$

and substituting it into the equation (9) we obtain

$$A_{j+1} - M - (A_j - M) = -iM$$

which is again the first order nonhomogeneous linear difference equation with constant coefficients

$$A_{j+1} - A_j = -iM, \qquad j = 1, 2, \ldots, n - 1, \tag{14}$$

where the initial element according to (8) is

$$A_1 = iD + M. \tag{15}$$

From (1) further follows

$$A_1 = \frac{D}{n}(in + 1). \tag{16}$$

In the same way as in the previous section we get an estimate of the general solution of (14) as the sum

$$A_j = C - ijM, \quad \text{where} \quad C \in \mathbb{R}, \tag{17}$$

because real number 1 is the root of the characteristic equation of (14). From initial condition (15) and the above equation with index $j = 1$ we can specify the constant C

$$C = i(D + M) + M.$$

Therefore substituting this into (17) we have

$$A_j = i(D + M) + M - ijM,$$

and considering (1) we get

$$A_j = \frac{D}{n}\left[i(n - j + 1) + 1\right], \qquad j = 1, 2, \ldots, n \tag{18}$$

which is the formula for calculation of the amount of the j-th payment. This relation follows directly from (13), (11) and (1). Let us note that the final payment with index $j = n$ is equal to

$$A_n = \frac{D}{n}(i + 1). \tag{19}$$

CONCLUSION

We can include all the above derived formulas (16), (8), (5), (18), (11), (4) and (19), (12), (6) which were created by solving the difference equations in the following Table 1.

It should be noted that all the mentioned formulas were obtained by solving difference equations and not using the properties of sequences which is commonly used in financial mathematics. The table below is called the fixed principal loan amortization schedule and is commonly used in finance. The table shows that though the principal amount included in each payment remains the same, the interest amount and therefore also the total payment amount decreases over each payment period. Thus the loan balance is declining and equal to zero at the end.

j-th installment	Payment A_j	Interest U_j	Principal M_j	Loan Balance D_j
0	-	-	-	$n\,\dfrac{D}{n} = D$
1	$\dfrac{D}{n}\,(n\,i+1)$	$n\,\dfrac{D}{n}\,i$	$\dfrac{D}{n}$	$(n-1)\,\dfrac{D}{n}$
2	$\dfrac{D}{n}\,[(n-1)\,i+1]$	$(n-1)\,\dfrac{D}{n}\,i$	$\dfrac{D}{n}$	$(n-2)\,\dfrac{D}{n}$
3	$\dfrac{D}{n}\,[(n-2)\,i+1]$	$(n-2)\,\dfrac{D}{n}\,i$	$\dfrac{D}{n}$	$(n-3)\,\dfrac{D}{n}$
$\vdots$	$\vdots$	$\vdots$	$\vdots$	$\vdots$
j	$\dfrac{D}{n}\,[(n-j+1)\,i+1]$	$(n-j+1)\,\dfrac{D}{n}\,i$	$\dfrac{D}{n}$	$(n-j)\,\dfrac{D}{n}$
$\vdots$	$\vdots$	$\vdots$	$\vdots$	$\vdots$
$n-1$	$\dfrac{D}{n}\,(2\,i+1)$	$2\,\dfrac{D}{n}\,i$	$\dfrac{D}{n}$	$\dfrac{D}{n}$
n	$\dfrac{D}{n}\,(i+1)$	$\dfrac{D}{n}\,i$	$\dfrac{D}{n}$	0

Table 1: Amortization schedule

References

[1] Banasiak, J. *Mathematical Modelling in One Dimension: An Introduction via Difference and Differential Equations*. Cambridge University Press, 2013, ISBN 978-1-107-65468-6.

[2] Elaydi, S. *An Introduction to Difference Equations*. Springer, 2004, ISBN 978-0-387-27602-1.

[3] Fulford, G., Forrester, P., Jones, A. *Modeling with Differential and Difference Equations*. Cambridge University Press, 1997, ISBN 9781139172660.

[4] Goldberg, S. *Introduction to Difference Equations: With Illustrative Examples from Economics, Psychology, and Sociology*. Dover Publications Inc., United States, 2010, ISBN 978-0486650845.

[5] Kwapisz, M. On Difference Equations Arising in Mathematics of Finance. *Nonlinear Analysis: Theory, Methods and Applications*. No. 30, 1997, p. 1207–1218, ISSN 0362-546X.

[6] Moučka, J., Rádl, P. *Matematika pro studenty ekonomie*. Praha: Grada Publishing, Inc., 2015, ISBN 978-80-247-5406-2.

[7] Neusser, K. *Difference Equations for Economists*. Bern: University of Bern, 2012. [Online]. [Cit. 2022-03-21] Available at: <http://www.neusser.ch/downloads/DifferenceEquations.pdf>.

[8] Neusser, K. *Time Series Econometrics*. Springer, 2016, ISSN 2192-4333.

[9] Pražák, P. *Diferenční rovnice s aplikacemi v ekonomii*. Gaudeamus, 2013, ISBN 978-80-7435-268-3.

[10] Radová, J., Dvořák, P., Málek, J. *Finanční matematika pro každého*. Praha: Grada Publishing, a.s., 2005, ISBN 80-247-1230-X.

[11] Říhová, D., Viskotová, L. Loan Amortization and Linear Difference Equations. In: *Proceedings of 18th Conference of Enterprise and Competitive Environment*. Mendel University in Brno, 2015, p. 783–790. [Online]. [Cit. 2022-03-21] Available at: <https://ece.pefka.mendelu.cz/sites/default/files/imce/ece_2015_final.pdf>. ISBN 978-80-7509-342-4.

[12] Říhová, D., Viskotová, L. Some Applications of Linear Difference Equations in Finance with Wolfram—alpha and Maple. *Ratio Mathematica*, No. 27, 2014, p. 81–90, ISSN 1592-7415.

[13] Říhová, D. Several Products of Financial Mathematics as Linear Difference Equations. In: *Proceedings of Conference Aplimat 2019*. Bratislava: Slovak University of Technology, p. 992–1000, ISBN 978-80-227-4884-1.

[14] Říhová, D. Amortization Schedule via Linear Difference Equations. In: *Proceedings of Conference Mathematical Methods in Economics 2020*. Brno: Mendel University, p. 511–515. [Online]. [Cit. 2022-03-21] Available at: <https://mme2020.mendelu.cz/wcd/w-rek-mme/mme2020_conference_proceedings_final.pdf>. ISBN 978-80-7509-734-7.

[15] Sakalová, K. Difference Equations in Life Insurance Mathematics. In: *Proceedings of 7th International Scientific Conference on Managing and Modelling of Financial Risk*. Ostrava: Technical University of Ostrava, 2014, p. 684–690, ISSN 2464-6970.

[16] Šoba, O., Širůček, M., Ptáček, R. *Finanční matematika v praxi*. Praha: Grada Publishing, a.s., 2013, ISBN 978-80-247-4636-4.

[17] Viskotová, L., Říhová, D. Annuity Due and Ordinary Annuity Using Linear Difference Equations and CAS. In: *Proceedings of Conference Aplimat 2016*. Bratislava: Slovak University of Technology, 2016, p. 1093–1104, ISBN 978-80-227-4530-7.

[18] Investopedia. [Online]. [Cit. 2022-03-21] Available at: <http://www.investopedia.com>.

PROJECTIVE GEOMETRIC ALGEBRA - BARYCENTRIC AND PLÜCKER COORDINATES COMPUTATION

Vaclav Skala
Department of Computer Science and Engineering, Faculty of Applied Sciences,
University of West Bohemia, Plzen
Czech Republic
http://www.VaclavSkala.eu ORCID[0000-0001-8886-4281]

Abstract: *This paper presents the computation of the barycentric coordinates and Plücker coordinates using the projective extension of the Euclidean space and geometric algebra. Using the projective extension, it also presents a relationship between linear systems of equations* ***Ax=b*** *and* ***Ax=0*** *using the projective extension. An application of the principle of duality enables solving dual problems efficiently. The given approach uses vector notation leading to efficient implementation on GPU or efficient use of SSE instructions. As the presented approach is based on projective notation, the division operation is postponed and the proposed method leads to higher computational robustness.*

Keywords: barycentric coordinates, Plücker coordinates, principle of duality, outer product, geometric algebra

INTRODUCTION

Linear algebra and geometry are closely related research fields and many algorithms have been developed. Geometric calculus evolved from Euclid's geometry(300 BC), Descartes geometry(1637), Hamilton's Algebra of quaternions(1843), Grassmann's Extensive algebra(1844), Cayley's Matrix algebra (1854), Clifford's algebra(1878), Gibbs Vector algebra(1881), Ricci's Tensor calculus(1890) and Pauli&Dirac's Spin algebra to Geometric Algebra&Calculus, which was formulated by Hesteness[12] as Space-time algebra in 1996 [1].

Since then the Geometric Algebra (GA) has developed to the universal multi-dimensional calculus, see Calvet[5], Macdonald[21], Kanatani[17], Gunn[10]. The geometric algebra is used in many fields, e.g. physics Doran[6], computer graphics Dorst[7][8], Hildebrand[13], Vince[38][39], electrical engineering Joot[16], Esch[9], geometry Calvet[5], motion interpolation Halma[11] robotics Bayro-Corrochano[4][2], quantum computing Alves[1], applications Li[20], Perwass[24] etc.
The geometric algebra was extended to the Conformal Geometric Algebra(CGA), see Doran[6], Bayro-Corrochano[3], Li[19], Hildenbrand[14], etc. [2]

Today's linear algebra uses the Gibbs vector algebra and Cayley's matrix notation, which leads to problems if multi-dimensional formulation is to be used.

[1] http://geocalc.clas.asu.edu/html/Evolution.html
[2] A brief introduction to the CGA: https://en.wikipedia.org/wiki/Conformal_geometric_algebra

1 GEOMETRIC ALGEBRA

The vector algebra (Gibbs algebra) used nowadays uses two fundamental operations on two vectors $\mathbf{a},\mathbf{b}$ in E^n, i.e. the inner product (scalar product or dot product) $c = \mathbf{a}\cdot\mathbf{b}$, where c is a scalar value and outer product $\mathbf{c} = \mathbf{a}\wedge\mathbf{b}$ (the cross-product $\mathbf{c} = \mathbf{a}\times\mathbf{b}$ in E^3), [3] where $\mathbf{c}$ is a bivector and it has different properties than a vector as it represents an oriented area in n-dimensional space.

The Geometric Algebra (GA) uses a new product called Geometric product defined as:

$$\mathbf{ab} = \mathbf{a}\cdot\mathbf{b} + \mathbf{a}\wedge\mathbf{b} \tag{1}$$

where $\mathbf{ab}$ is a geometric product.

In the case of the n-dimensional space, vectors are defined as $\mathbf{a} = (a_1\mathbf{e}_1 + ... + a_n\mathbf{e}_n)$, $\mathbf{b} = (b_1\mathbf{e}_1 + ... + b_n\mathbf{e}_n)$ and the $\mathbf{e}_i$ vectors form orthonormal basis vectors in E^3 then we get:

1	0-vector (scalar)	$\mathbf{e}_{12}$, $\mathbf{e}_{23}$, $\mathbf{e}_{31}$	2-vectors (bivectors)
$\mathbf{e}_1,\mathbf{e}_2,\mathbf{e}_3$,	1-vector (vectors)	$\mathbf{e}_{123}$	3-vector (pseudoscalar)

It can be easily proved that the following operations are valid, including an inverse of a vector.

$$\mathbf{a}\cdot\mathbf{b} = \frac{1}{2}(\mathbf{ab}+\mathbf{ba}) \qquad \mathbf{a}\wedge\mathbf{b} = -\mathbf{b}\wedge\mathbf{a} \qquad \mathbf{a}^{-1} = \mathbf{a}/||\mathbf{a}||^2 \tag{2}$$

It can be seen, that geometric algebra is *anti-commutative* and the pseudoscalar I in E^3 has the basis $\mathbf{e}_1\mathbf{e}_2\mathbf{e}_3$, i.e.

$$\mathbf{e}_i\mathbf{e}_j = -\mathbf{e}_j\mathbf{e}_i \qquad \mathbf{e}_i\mathbf{e}_i = 1 \qquad \mathbf{e}_1\mathbf{e}_2\mathbf{e}_3 = I \qquad \mathbf{a}\wedge\mathbf{b}\wedge\mathbf{c} = q \tag{3}$$

where q is a scalar value and a short notation $\mathbf{e}_i\mathbf{e}_j = \mathbf{e}_{ij}$ can be used:

In general, the geometric product is represented as

$$\mathbf{ab} = \sum_{i,j=1}^{n} a_i\mathbf{e}_i b_j\mathbf{e}_j \qquad \mathbf{a}\cdot\mathbf{b} = \sum_{i=1}^{n} a_i\mathbf{e}_i b_i\mathbf{e}_i \tag{4}$$

$$\mathbf{a}\wedge\mathbf{b} = \sum_{i,j=1\&i\neq j}^{n} a_i\mathbf{e}_i b_j\mathbf{e}_j = \sum_{i,j=1,\&i>j}^{n} (a_i b_j - a_j b_i)\mathbf{e}_i\mathbf{e}_j \tag{5}$$

It is not a friendly user notation for a practical application and causes problems in practical implementations, primarily due to the anti-commutativity of the geometric product.

However, the geometric product can be easily represented by the tensor product, see Mochizuki[23], which can be represented by a matrix. As the homogeneous coordinates will be used in the following, the tensor product for the 4-dimensional case is presented [4] :

$$\mathbf{ab} \underset{\text{repr}}{\Longleftrightarrow} \mathbf{ab}^T = \mathbf{a}\otimes\mathbf{b} = \mathbf{Q} = \begin{bmatrix} a_1b_1\mathbf{e}_1\mathbf{e}_1 & a_1b_2\mathbf{e}_1\mathbf{e}_2 & a_1b_3\mathbf{e}_1\mathbf{e}_3 & a_1b_4\mathbf{e}_1\mathbf{e}_4 \\ a_1b_2\mathbf{e}_2\mathbf{e}_1 & a_2b_2\mathbf{e}_2\mathbf{e}_2 & a_2b_3\mathbf{e}_2\mathbf{e}_3 & a_2b_4\mathbf{e}_2\mathbf{e}_4 \\ a_1b_3\mathbf{e}_3\mathbf{e}_1 & a_3b_2\mathbf{e}_3\mathbf{e}_2 & a_3b_3\mathbf{e}_3\mathbf{e}_3 & a_3b_4\mathbf{e}_3\mathbf{e}_4 \\ a_1b_4\mathbf{e}_4\mathbf{e}_1 & a_4b_2\mathbf{e}_4\mathbf{e}_2 & a_4b_3\mathbf{e}_4\mathbf{e}_3 & a_4b_4\mathbf{e}_4\mathbf{e}_4 \end{bmatrix} = \mathbf{B}+\mathbf{U}+\mathbf{D} \tag{6}$$

where $\mathbf{B}+\mathbf{U}+\mathbf{D}$ are *B*ottom triangular, *U*pper triangular, *D*iagonal matrices, a_4, b_4 are the homogeneous coordinates, i.e. actually w_a, w_b (will be explained later), and the operator $\otimes$ means the anti-commutative tensor product.

[3] Massey[22] and Silagadze[25] use multi-dimensional cross-product term

[4] The vector basis $\mathbf{e}_i\mathbf{e}_j$, etc. will not be used explicitly

2 PROJECTIVE EXTENSION AND PRINCIPLE OF DUALITY

Let us consider the projective extension of the Euclidean space and the use of homogeneous coordinates. [5].

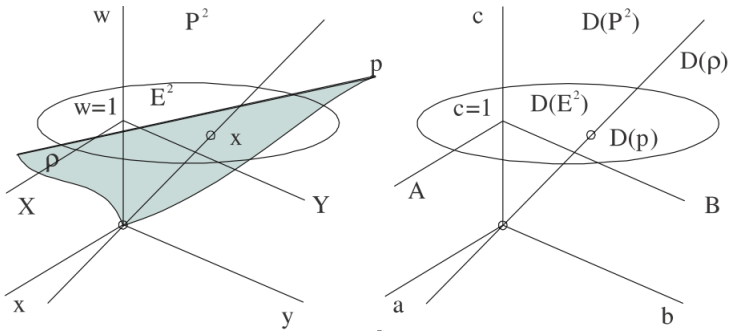


Figure 1: Projective extension and dual space

It uses homogeneous coordinates and two equivalent forms can be found:

- the form $[x_1,\ldots,x_n : x_w]$ is mostly used in computer graphics-related fields, namely $[x,y : w]$ in the case of P^2, resp. $[x,y,z : w]$ in the case of P^3, where w is the homogeneous coordinate.
- the form $[x_0 : x_1,\ldots,x_n]$ is used in the mathematical fields and the x_0 is the homogeneous coordinate. This form has the advantage that the homogeneous coordinate is on the first position.

It should be noted that ":" is used to emphasize that the x_w, resp x_0 has a different meaning as it is the "scaling factor", i.e. without a physical unit, while $x_1,\ldots,x_n$ has different physical units, e.g. meters[m] etc.

The mutual conversion between the Euclidean space and projective space is given as:

$$X_i = \frac{x_i}{x_0} \quad x_0 \neq 0 \quad \text{, resp.} \quad X_i = \frac{x_i}{x_w} \quad x_w \neq 0 \quad , \quad i = 1,\ldots,n \tag{7}$$

where X_i are coordinates in the Euclidean space.
In the case of the E^2 space

$$X = \frac{x}{x_0} \quad Y = \frac{Y}{x_0} \quad x_0 \neq 0 \quad \text{, resp.} \quad X = \frac{x}{w} \qquad Y = \frac{y}{w} \quad w \neq 0 \tag{8}$$

where (X,Y), resp.$[x,y : w]$ are coordinates in the Euclidean space E^2, resp.in the projective space P^2. The extension to the E^3, resp. E^n space is straightforward, see Vince[39], Yamaguchi[40].
The geometrical interpretation of the Euclidean ($x_w = 1$, resp. $x_0 = 1$) and the projective spaces is presented at Fig.1.
It should be noted, that a distance of a point $X = (X,Y)$, i.e. $x = [x,y : w]^T$ from a line in the E^2 is defined as

$$dist = \frac{aX + bY + c}{\sqrt{a^2 + b^2}} = \frac{ax + by + cw}{w\sqrt{a^2 + b^2}} \tag{9}$$

where (a,b) is the normal vector (actually it is a bivector) of the line.

[5]The concept of the projective extension for the CAD/CAM systems was deeply described in Yamaguchi[40]

2.1 Inner and outer products

The *inner product* and *outer product*, i.e. the *dot-product* and *cross-product* in the E^3, are known. However, if the projective extension of the Euclidean space is used, there are slightly different interpretations.

Let us consider vectors $\mathbf{a} = [a_1, a_2, a_3 : a_4]^T$ and $\mathbf{b} = [b_1, b_2, b_3 : b_4]^T$ in the projective space. They represents actually vectors $(a_1/a_4, a_2/a_4, a_3/a_4)$ and $(b_1/b_4, b_2/b_4, b_3/b_4)$ in the E^3 space. It can be seen, that the diagonal of the matrix $\mathbf{Q}$ actually represents the inner product in the projective representation:

$$\mathbf{a} \cdot \mathbf{b} = [(a_1 b_1 + a_2 b_2 + a_3 b_3) : a_4 b_4]^T \triangleq \frac{a_1 b_1 + a_2 b_2 + a_3 b_3}{a_4 b_4} \tag{10}$$

where $\triangleq$ means projective equivalence. The inner product represents the trace $tr(\mathbf{Q})$ of the matrix $\mathbf{Q}$ and $\mathbf{a} \cdot \mathbf{b}$ means a scalar value expressed using homogeneous coordinates.

The outer product in the E^3 *vector space* is represented respecting anti-commutativity as:

$$\mathbf{a} \wedge \mathbf{b} \underset{\text{repr}}{\Longleftrightarrow} \sum_{i,j=1 \& i>j}^{3,3} (a_i b_j \mathbf{e}_i \mathbf{e}_j - b_i a_j \mathbf{e}_i \mathbf{e}_j) = \sum_{i,j \& i>j}^{3,3} (a_i b_j - b_i a_j) \mathbf{e}_i \mathbf{e}_j \tag{11}$$

where $\mathbf{a}, \mathbf{b} \in E^3$ vector space.

However, if the projective extension is used,

$$\mathbf{a} \wedge \mathbf{b} = \underset{\text{repr}}{\Longleftrightarrow} \sum_{i,j=1 \& i>j}^{4,4} (a_i b_j \mathbf{e}_i \mathbf{e}_j - b_i a_j \mathbf{e}_i \mathbf{e}_j) \triangleq \frac{\sum_{i,j \& i>j}^{3,3} (a_i b_j - b_i a_j) \mathbf{e}_i \mathbf{e}_j}{a_4 b_4 \mathbf{e}_4 \mathbf{e}_4} \tag{12}$$

It means, that the result of the outer product $\mathbf{c} = \mathbf{a} \wedge \mathbf{b}$ is represented as $\mathbf{c} = [c_1, \ldots, c_3 : c_4]^T$, where $(c_1, \ldots, c_3)$, i.e. by a bivector (normal vector) of a plane in E^3, while $c_4 = a_4 b_4$ is actually a scaling factor.

It should be noted, that the outer product can be used for a solution of a linear system of equations $\mathbf{Ax} = \mathbf{b}$ or $\mathbf{Ax} = \mathbf{0}$, too.

2.2 Principle of duality

The principle of duality is essential principle, in general. Its application in geometry in connection with the implicit representation using projective geometry brings some new formulations or even new ones, see Johnson[15].

The duality principle for basic geometric entities and operators are presented by Tab.1 and Tab.2. It the E^2 case, a point is dual to a line and vice versa, the intersection of two lines is dual to a union of two points, i.e. line given by two points, similarly for the E^3 case.

Table 1: Duality of geometric entities

Duality of geometric entities						
Point in E^2	$\Longleftrightarrow$ DUAL	Line in E^2		Point in E^3	$\Longleftrightarrow$ DUAL	Plane in E^3

Table 2: Duality of operators

Duality of operators		
Union $\cup$	$\Longleftrightarrow$ DUAL	Intersection $\cap$

3 COMPUTATION WITH HOMOGENEOUS REPRESENTATION

The geometric algebra (GA) presented above has been formulated for vectors in the Euclidean space, as presented above. However, the concept can be extended using the projective extension of the Euclidean space. It enables handling geometric entities like points, lines and planes, efficiently.

3.1 SOLUTION OF LINEAR SYSTEM OF EQUATIONS

A solution of a linear system of equations is a part of linear algebra and is used in many computational systems. It should be noted, that linear equations $\mathbf{Ax} = \mathbf{b}$ can be transformed to an implicit the homogeneous system, i.e. to the form $\mathbf{B}\boldsymbol{\xi} = \mathbf{0}$, where $\mathbf{B} = [\mathbf{A}|-\mathbf{b}]$, $\boldsymbol{\xi} = [\xi_1, ..., \xi_n : \xi_w]^T$, $x_i = \xi_i \,/\, \xi_w$, $i = 1, ..., n$, see Skala[31, 33] [6].

As the solution of a linear system of equations is equivalent to the outer product (generalized cross-vector) of vectors formed by row vectors $\mathbf{a}_i$ of the matrix $\mathbf{B}$, the solution of the system is defined as:

$$\boldsymbol{\xi} = \mathbf{a}_1 \wedge \mathbf{a}_2 \wedge ... \wedge \mathbf{a}_n \qquad [\mathbf{A}|-\mathbf{b}]\boldsymbol{\xi} = 0 \qquad \mathbf{a}_i = [a_{i,1}, \ldots, a_{i,n}, -b_i] \tag{13}$$

which is equivalent to a solution of the linear system of equations:

$$\begin{bmatrix} a_{11} & \cdots & a_{1n} \\ \vdots & \ddots & \vdots \\ a_{n1} & \cdots & a_{nn} \end{bmatrix} \begin{bmatrix} x_1 \\ \vdots \\ x_n \end{bmatrix} = \begin{bmatrix} b_1 \\ \vdots \\ b_n \end{bmatrix} \quad \text{, i.e.} \quad \begin{bmatrix} a_{11} & \cdots & a_{1n} & -b_1 \\ \vdots & \ddots & \vdots & \vdots \\ a_{n1} & \cdots & a_{nn} & -b_n \end{bmatrix} \begin{bmatrix} \xi_1 \\ \vdots \\ \xi_n \\ \xi_w \end{bmatrix} = \begin{bmatrix} 0 \\ \vdots \\ 0 \end{bmatrix} \tag{14}$$

It is a significant result as a solution of a linear system of equations is formally the same for systems for the both cases, i.e. $\mathbf{Ax} = \mathbf{0}$ and $\mathbf{Ax} = \mathbf{b}$.
As the solution is formally determined, the formal linear operators can be used for further symbolic processing using formula manipulation, as the geometry algebra is multi-linear. Even more, it is capable to handle more complex objects generally in the n-dimensional space, i.e. oriented surfaces, volumes, etc.

However, more general rules can be derived for the n-dimensional space and the *outer product* application in Euclidean space. Let a matrix $\mathbf{M}$ is a $n \times n$ non-singular matrix representing a

[6] This can be also used in solution of ordinary differential equations using the Laplace transform, see Skala[34]

geometric transformation, see the Eq.15.

$$(\mathbf{Ma}) \wedge (\mathbf{Ma}_2) \wedge \ldots \wedge (\mathbf{Ma}_n) = det(\mathbf{M})^{n-1} (\mathbf{M}^{-1})^T (\mathbf{a}_1 \wedge \mathbf{a}_2 \wedge \ldots \wedge \mathbf{a}_n) \tag{15}$$

In the case pro of the projective extension of the Euclidean space, the Eg.15 is simplified to Eq.16 due to implicit representation, as the $det(\mathbf{M})^{n-1}$ is only a multiplicative constant.

$$\begin{aligned} (\mathbf{Ma}) \wedge (\mathbf{Ma}_2) \wedge \ldots \wedge (\mathbf{Ma}_n) &= det(\mathbf{M})^{n-1} (\mathbf{M}^{-1})^T (\mathbf{a}_1 \wedge \mathbf{a}_2 \wedge \ldots \wedge \mathbf{a}_n) \\ &\triangleq (\mathbf{M}^{-1})^T (\mathbf{a}_1 \wedge \mathbf{a}_2 \wedge \ldots \wedge \mathbf{a}_n) \end{aligned} \tag{16}$$

where $\triangleq$ means projective equivalence as we use the implicit formulation.

Now, it is possible to use the Functional analysis approach: Let L is a linear operator, then the following operation is valid..... As there are many linear operators like derivation, integration, Fourier and Laplace transforms etc., there is a wide variety of applications of those to the formal solution of the linear system of equations, i.e. $L(\xi)$. However, it is necessary to respect that in the case of the projective representation specific care is to be taken for deriving rules for derivation etc., as a fraction is to be processed; similarly to other operators.

3.2 Intersections and unions

The direct consequence of the principle of duality is that the intersection point $\mathbf{x}$ of two lines $\mathbf{p}_1, \mathbf{p}_2$, resp. a line $\mathbf{p}$ passing through two given points $\mathbf{x}_1, \mathbf{x}_2$, is given as:

$$\mathbf{x} = \mathbf{p}_1 \wedge \mathbf{p}_2 \underset{\text{DUAL}}{\Longleftrightarrow} \mathbf{p} = \mathbf{x}_1 \wedge \mathbf{x}_2 \tag{17}$$

where $\mathbf{p}_i = [a_i, b_i : c_i]^T$, $\mathbf{x} = [x, y : w]^T$ (w is the homogeneous coordinate), $i = 1, 2$; similarly in the dual case.

In the case of the E^3 space, a point is dual to a plane and vice versa. It means that the intersection point $\mathbf{x}$ of three planes $\boldsymbol{\rho}_1$,$\boldsymbol{\rho}_2$,$\boldsymbol{\rho}_3$, resp. a plane $\boldsymbol{\rho}$ passing through three given points $\mathbf{x}_1, \mathbf{x}_2, \mathbf{x}_3$ is given as:

$$\mathbf{x} = \boldsymbol{\rho}_1 \wedge \boldsymbol{\rho}_2 \wedge \boldsymbol{\rho}_3 \underset{\text{DUAL}}{\Longleftrightarrow} \boldsymbol{\rho} = \mathbf{x}_1 \wedge \mathbf{x}_2 \wedge \mathbf{x}_3 \tag{18}$$

where $\mathbf{x} = [x, y, z : w]^T$, $\boldsymbol{\rho}_i = [a_i, b_i, c_i : d_i]^T$, $i = 1, 2, 3$.

It can be seen that the above formulae is equivalent to the extended cross-product, which in natively supported by GPU architecture. For an intersection computation, we get:

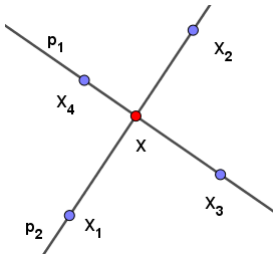
$$\mathbf{x} = \mathbf{p}_1 \wedge \mathbf{p}_2 = \begin{bmatrix} \mathbf{e}_1 & \mathbf{e}_2 & \mathbf{e}_w \\ a_1 & b_1 & c_1 \\ a_2 & b_2 & c_2 \end{bmatrix} \qquad \mathbf{x} = \boldsymbol{\rho}_1 \wedge \boldsymbol{\rho}_2 \wedge \boldsymbol{\rho}_3 = \begin{bmatrix} \mathbf{e}_1 & \mathbf{e}_2 & \mathbf{e}_3 & \mathbf{e}_w \\ a_1 & b_1 & c_1 & d_1 \\ a_2 & b_2 & c_2 & d_2 \\ a_3 & b_3 & c_3 & d_3 \end{bmatrix} \tag{19}$$

Due to the principle of duality, a dual problem solution is given as:

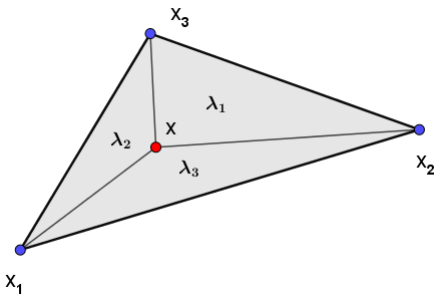
$$\mathbf{p} = \mathbf{x}_1 \wedge \mathbf{x}_2 = \begin{bmatrix} \mathbf{e}_1 & \mathbf{e}_2 & \mathbf{e}_w \\ x_1 & y_1 & w_1 \\ x_2 & y_2 & w_2 \end{bmatrix} \qquad \boldsymbol{\rho} = \mathbf{x}_1 \wedge \mathbf{x}_2 \wedge \mathbf{x}_3 = \begin{bmatrix} \mathbf{e}_1 & \mathbf{e}_2 & \mathbf{e}_3 & \mathbf{e}_w \\ x_1 & y_1 & z_1 & w_1 \\ x_1 & y_2 & z_2 & w_2 \\ x_3 & y_3 & z_3 & w_3 \end{bmatrix} \tag{20}$$

The above-presented formulae prove the strength of the geometric algebra approach [7] and also

[7] See Skala[29][30][32]



(a) Duality of a point and a line in E^2



(b) Barycentric coordinates in E^2

Figure 2: Duality in E^2: Lines and points, union and intersection, barycentric coordinates

simplifies geometric operations, e.g. line clipping Skala[35][36][37].

There is a natural question: What is the more convenient computation of the geometric product, as computation with the outer product, i.e. extended cross-product, using basis vector approach is not simple. Fortunately, the geometric product of $\boldsymbol{\rho}_1, \boldsymbol{\rho}_2$, resp. of $\mathbf{x}_1$ and $\mathbf{x}_2$ vectors using homogeneous coordinates given as anti-commutative tensor product is given as:

$\boldsymbol{\rho}_1\boldsymbol{\rho}_2$	a_2	b_2	c_2	d_2
a_1	a_1a_2	a_1b_2	a_1c_2	a_1d_2
b_1	b_1a_2	b_1b_2	b_1c_2	b_1d_2
c_1	c_1a_2	c_1b_2	c_1c_2	a_1d_2
d_1	d_1a_2	d_1b_2	d_1c_2	d_1d_2

$\mathbf{x}_1\mathbf{x}_2$	x_2	y_2	z_2	w_2
x_1	x_1x_2	x_1y_2	x_1z_2	x_1w_2
y_1	y_1x_2	y_1y_2	y_1z_2	y_1w_2
z_1	z_1x_2	z_1y_2	z_1z_2	x_1w_2
w_1	w_1x_2	w_1y_2	w_1z_2	w_1w_2

3.3 Plücker coordinates

A line in the E^3 space is given as an intersection of two planes or in a parametric form, see Eq.21:

$$\begin{aligned}\rho_1 &: a_1X + b_1Y + c_1Z + d_1 = 0\\ \rho_2 &: a_2X + b_2Y + c_2Z + d_2 = 0\end{aligned} \quad \text{, or} \quad \mathbf{X}(t) = \mathbf{X}_1 + (\mathbf{X}_2 - \mathbf{X}_1)\, t \tag{21}$$

where: $\rho_1 : \mathbf{n}_1^T\mathbf{X} + d_1 = 0$ and $\rho_2 : \mathbf{n}_2^T\mathbf{X} + d_2 = 0$.
The question is how to compute a line $\mathbf{p} \in E^3$ given as an intersection of two planes $\boldsymbol{\rho}_1$, $\boldsymbol{\rho}_2$, which is dual to a line determination given by two points $\mathbf{x}_1$, $\mathbf{x}_2$ as those problems are dual.

The parametric solution can be easily obtained using standard Plücker coordinates [8]. The above-given formula is difficult to derive [9] and not easy to understand and computation is complex.

$$q(t) = \frac{\omega \times \mathbf{v}}{||\omega||^2} + \omega t \qquad \mathbf{L} = \mathbf{x}_1\mathbf{x}_2^T - \mathbf{x}_2\mathbf{x}_1^T \quad \omega = [l_{41}, l_{42}, l_{43}]^T \quad \mathbf{v} = [l_{23}, l_{31}, l_{12}]^T \tag{22}$$

[8] The "reference" point of a line is the closest point to the origin of the coordinate system, which is a substantial property, e.g. in robotics and mechanical engineering

[9] https://en.wikipedia.org/wiki/Plücker_coordinates

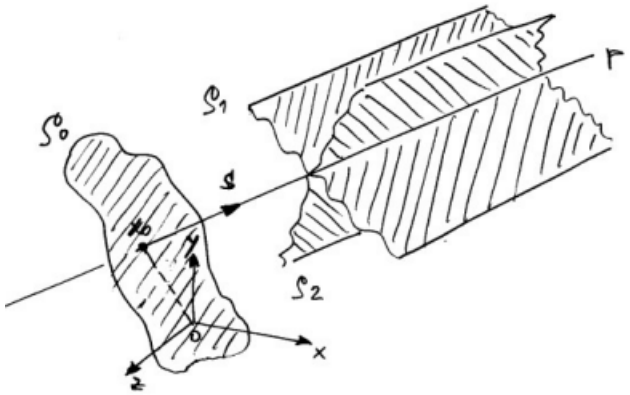


Figure 3: Plücker coordinates

In 1871, Klein[18] derived that $\omega\mathbf{v} = 0$, i.e. there is a dimension reduction, see Skala[26] for details.
However, using the outer product the formulation is easy and easy to understand, see Fig.3:

$$\mathbf{s} = \mathbf{n}_1 \wedge \mathbf{n}_2 \qquad \boldsymbol{\rho}_0 = [\mathbf{s}^T : 0]^T \qquad \mathbf{x}_0 = \boldsymbol{\rho}_1 \wedge \boldsymbol{\rho}_2 \wedge \boldsymbol{\rho}_0 \tag{23}$$

where $\mathbf{s}$ is the directional vector of and $\mathbf{x}_0$ is a "reference" point of a line, which is the closest point to the origin.

For the intersection of two planes, the principle of duality can be applied directly.

However, using geometric algebra, the principle of duality and projective representation, we can directly write:

$$\mathbf{p} = \rho_\mathbf{1} \wedge \rho_\mathbf{2} \underset{\text{DUAL}}{\Longleftrightarrow} \mathbf{p} = \mathbf{x}_1 \wedge \mathbf{x}_2 \tag{24}$$

It can be seen, that the formula given above keeps the duality in the final formulae, too.

From the formal point of view, the geometric product for the both cases is given as:

$$\boldsymbol{\rho}_1\boldsymbol{\rho}_2 \underset{\text{repr}}{\Longleftrightarrow} \boldsymbol{\rho}_1 \otimes \boldsymbol{\rho}_2 = \begin{bmatrix} a_1a_2 & a_1b_2 & a_1c_2 & a_1d_2 \\ b_1a_2 & b_1b_2 & b_1c_2 & b_1d_2 \\ c_1a_2 & c_1b_2 & c_1c_2 & c_1d_2 \\ d_1a_2 & d_1b_2 & d_1c_2 & d_1d_2 \end{bmatrix} \tag{25}$$

The dual problem formulation:

$$\mathbf{x}_1\mathbf{x}_2 \underset{\text{repr}}{\Longleftrightarrow} \mathbf{x}_1 \otimes \mathbf{x}_2 = \begin{bmatrix} x_1x_2 & x_1y_2 & x_1z_2 & x_1w_2 \\ y_1x_2 & y_1y_2 & y_1z_2 & y_1w_2 \\ z_1x_2 & z_1y_2 & z_1z_2 & z_1w_2 \\ w_1x_2 & w_1y_2 & w_1z_2 & w_1w_2 \end{bmatrix} \tag{26}$$

It means that we have computation of the Plücker coordinates for both cases, i.e. for the computation of a line $\mathbf{p} = \boldsymbol{\rho}_1 \wedge \boldsymbol{\rho}_2$ given as an intersection of two planes in E^3 and a line given by two points, i.e. as a union of two points, in E^3 as $\mathbf{p} = \mathbf{x}_1 \wedge \mathbf{x}_2$ using the projective representation and the principle of duality. It should be noted that the given approach offers: significant simplification of

computation of the Plücker coordinates as it is simple and easy to derive and explain, uses vector-vector operations, which is especially convenient for SSE and GPU application one code sequence for the both cases.

The Plücker coordinates are also in mechanical engineering applications, especially in robotics, due to their simple displacement and momentum specifications. In other fields simple explanation and derivation are important arguments for GA approach application.

3.4 Barycentric coordinates

The barycentric coordinates are often used in many applications, not only in geometry. The barycentric coordinates computation, see Fig.2b, leads to a solution of a system of linear equations.

$$X_1\lambda_1 + X_2\lambda_2 + \lambda_3 X_3 = X \qquad Y_1\lambda_1 + Y_2\lambda_2 + \lambda_3 Y_3 = Y \qquad \lambda_1 + \lambda_2 + \lambda_3 = 1 \tag{27}$$

In the matrix form:

$$\begin{bmatrix} X_1 & X_2 & X_3 \\ Y_1 & Y_2 & Y_3 \\ 1 & 1 & 1 \end{bmatrix} \begin{bmatrix} \lambda_1 \\ \lambda_2 \\ \lambda_3 \end{bmatrix} = \begin{bmatrix} X \\ Y \\ 1 \end{bmatrix} \quad \text{, resp.} \quad \begin{bmatrix} x_1 & x_2 & x_3 \\ y_1 & y_2 & y_3 \\ w_1 & w_2 & w_3 \end{bmatrix} \begin{bmatrix} \lambda_1 \\ \lambda_2 \\ \lambda_3 \end{bmatrix} = \begin{bmatrix} X \\ Y \\ w \end{bmatrix} \tag{28}$$

where $\mathbf{X} = (X,Y) \in E^2$ and $\mathbf{x} = [x,y:w]^T \in P^2$, i.e. in the projective space.
However, a solution of linear system equations is equivalent to the outer product application, as explained above; Skala[26][27]. Therefore, it is possible to compute the barycentric coordinates using the outer product, which is recommendable especially for the GPU oriented applications.

Let us consider the E^2 case and the barycentric interpolation between three points (a triangle vertices) given generally in the projective space as $\mathbf{x}_i = [x_i, y_i : w_i]^T$, $i = 1,\ldots,3$ & $w_i \neq 0$, of the given triangle, and vectors:

$$\boldsymbol{\xi} = [x_1, x_2, x_3, x] \qquad \boldsymbol{\eta} = [y_1, y_2, y_3, y] \qquad \boldsymbol{\omega} = [w_1, w_2, w_3, w] \tag{29}$$

Then the barycentric coordinates μ in the homogeneous coordinates of the point $\mathbf{x} = [x,y:w]^T$ are given as:

$$\begin{bmatrix} \boldsymbol{\xi} \\ \boldsymbol{\eta} \\ \boldsymbol{\omega} \end{bmatrix} \begin{bmatrix} \mu_1 \\ \mu_2 \\ \mu_3 \\ \mu_w \end{bmatrix} = \begin{bmatrix} 0 \\ 0 \\ 0 \end{bmatrix} \quad \text{, i.e.} \quad \boldsymbol{\mu} = \boldsymbol{\xi} \wedge \boldsymbol{\eta} \wedge \boldsymbol{\omega} \tag{30}$$

where $\boldsymbol{\mu} = [\mu_1, \mu_2, \mu_3 : \mu_w]^T$ and the barycentric coordinates in the Euclidean space λ are given as:

$$\boldsymbol{\lambda} = (\lambda_1, \lambda_2, \lambda_3) = (-\frac{\mu_1}{\mu_w}, -\frac{\mu_2}{\mu_w}, -\frac{\mu_3}{\mu_w}) \tag{31}$$

Similarly, for other dimensions, see Skala[28] for details. How simple and elegant solution!

It can be seen, that the presented computation of barycentric coordinates is simple and convenient for GPU or SSE applications. As we have assumed from the very beginning, there is no need to convert the coordinates of points from homogeneous coordinates to Euclidean coordinates. As a direct consequence of that, we save a lot of division operations and increase the robustness of the computation.

4 CONCLUSION

This contribution briefly presents geometry algebra, which is not generally known and used. However, it offers simple and efficient solutions to many computational problems if combined with the principle of duality and projective notation.
As the result of this contribution, a new formulation of the Plücker coordinates, often used in mechanical engineering and robotics, is given. As the operations are based on standard linear algebra formalism, it is simple to use. The presented approach supports direct GPU application with significant speed-up and parallelism potential. Also, the approach is applicable to d-dimensional problem solutions, as geometric algebra is multi-dimensional.
The presented approach efficiently computes the barycentric coordinates of a point in the given convex simplex, the Plücker coordinates of a line given by two points or two planes in the E^3 space. As the division operation is postponed, higher robustness of computation can be achieved.

References

[1] R. Alves, D. Hildenbrand, J. Hrdina, and C. Lavor. An online calculator for quantum computing operations based on geometric algebra. *Advances in Applied Clifford Algebras*, 32(1), 2022.

[2] E. Bayro-Corrochano. *Geometric algebra applications vol. II: Robot modelling and control*. Springer International Publishing, 2020.

[3] E. Bayro-Corrochano, L. Reyes-Lozano, and J. Zamora-Esquivel. Conformal geometric algebra for robotic vision. *Journal of Mathematical Imaging and Vision*, 24(1):55–81, 2006.

[4] E. Bayro-Corrochano and G. Scheuermann. *Geometric algebra computing: In engineering and computer science*. Springer London, 2010.

[5] R. G. Calvet. *Treatise of Plane through Geometric Algebra*. Cerdanyola del Valls, Spain, 1 edition, 2013.

[6] C. Doran, A. N. Lasenby, and J. Lasenby. Conformal geometry, euclidean space and geometric algebra. *CoRR*, cs.CG/0203026, 2002.

[7] L. Dorst, D. Fontijne, and S. Mann. *Geometric Algebra for Computer Science: An Object-Oriented Approach to Geometry*. Morgan Kaufmann Publishers Inc., San Francisco, CA, USA, 2009.

[8] L. Dorst and J. Lasenby, editors. *Guide to Geometric Algebra in Practice*. Springer, 2011.

[9] J. Esch. Geometric algebra for electrical and electronic engineers. *Proceedings of the IEEE*, 102(9):1338–1339, 2014.

[10] C. Gunn. Doing Euclidean plane geometry using projective geometric algebra. *Advances in Applied Clifford Algebras*, 27(2):1203–1232, 2017.

[11] A. Halma. *Interpolation in Conformal Geometric Algebra: Toward Unified Interpolation of Euclidean Motions in the Conformal Model of Geometric Algebra*. Lap Lampert Publ., Moldova, 1 edition, 2011.

[12] D. Hestenes. *Space-Time Algebra*. Springer/Birkhauser, Berlin, Germany, 2015.

[13] D. Hildebrand. *Foundations of Geometric Algebra Computing*. Springer-Verlag, Berlin, 1 edition, 2013.

[14] D. Hildenbrand. Geometric computing in computer graphics using conformal geometric algebra. *Computers and Graphics (Pergamon)*, 29(5):795–803, 2005.

[15] M. Johnson. Proof by duality: or the discovery of new theorems. *Mathematics Today*, December:138–153, 1996.
[16] P. Joot. *Geometric Algebra for Electrical Engineers: Multivector Electromagnetism*. CreateSpace Independent Publishing Platform, 2019.
[17] K. Kanatani. *Understanding geometric algebra: Hamilton, Grassmann, and Clifford for computer vision and graphics*. CRC Press, 2015.
[18] F. Klein. Notiz, betreffend den zusammenhang der liniengeometric mit der mechanik starrer körper. *Mathematische Annalen*, 4(3):403–415, 1871.
[19] H. Li. Conformal geometric algebra - a new framework for computational geometry. *Jisuanji Fuzhu Sheji Yu Tuxingxue Xuebao/Journal of Computer-Aided Design and Computer Graphics*, 17(11):2383–2393, 2005.
[20] H. Li, P. J. Olver, and G. Sommer, editors. *Computer Algebra and Geometric Algebra with Applications*, volume 3519 of *Lecture Notes in Computer Science*. Springer, 2005.
[21] A. Macdonald. A survey of geometric algebra and geometric calculus. *Advances in Applied Clifford Algebras*, 27(1):853–891, 2017.
[22] W. S. Massey. Cross products of vectors in higher dimensional Euclidean spaces. *The American Mathematical Monthly*, 90(10):697–701, 1983.
[23] N. Mochizuki. The tensor product of function algebras. *Tohoku Mathematical Journal*, 17(2):139–146, 1965.
[24] C. Perwass. *Geometric Algebra with Applications in Engineering*. Springer-Verlag, Berlin, 1 edition, 2009.
[25] Z. K. Silagadze. Multi-dimensional vector product. *Journal of Physics A: Mathematical and General*, 35(23):49494953, May 2002.
[26] V. Skala. A new approach to line and line segment clipping in homogeneous coordinates. *Visual Computer*, 21(11):905–914, 2005.
[27] V. Skala. Length, area and volume computation in homogeneous coordinates. *International Journal of Image and Graphics*, 6(4):625–639, 2006.
[28] V. Skala. Barycentric coordinates computation in homogeneous coordinates. *Computers and Graphics (Pergamon)*, 32(1):120–127, 2008.
[29] V. Skala. Intersection computation in projective space using homogeneous coordinates. *International Journal of Image and Graphics*, 8(4):615–628, 2008.
[30] V. Skala. Projective geometry and duality for graphics, games and visualization. *SIGGRAPH Asia 2012 Courses, SA 2012*, 2012.
[31] V. Skala. Modified Gaussian elimination without division operations. *AIP Conference Proceedings*, 1558:1936–1939, 2013.
[32] V. Skala. A new robust algorithm for computation of a triangle circumscribed sphere in E3 and a hypersphere simplex. *AIP Conference Proceedings*, 1738, 2016.
[33] V. Skala. extended cross-product and solution of a linear system of equations. *Lecture Notes in Computer Science*, 9786:18–35, 2016.
[34] V. Skala. Geometric algebra, extended cross-product and Laplace transform for multidimensional dynamical systems. *Advances in Intelligent Systems and Computing*, 661:62–75, 2018.
[35] V. Skala. Optimized line and line segment clipping in E2 and geometric algebra. *Annales Mathematicae et Informaticae*, 52:199–215, 2020.
[36] V. Skala. A new coding scheme for line segment clipping in E2. *Lecture Notes in Computer Science (including subseries Lecture Notes in Artificial Intelligence and Lecture Notes in Bioinformatics)*, 12953 LNCS:16–29, 2021.

[37] V. Skala. A novel line convex polygon clipping algorithm in e2 with parallel processing modification. *Lecture Notes in Computer Science (including subseries Lecture Notes in Artificial Intelligence and Lecture Notes in Bioinformatics)*, 12953 LNCS:3–15, 2021.
[38] J. Vince. *Geometric Algebra: An Algebraic System for Computer Games and Animation*. Springer Publishing Company, Incorporated, 1st edition, 2009.
[39] J. A. Vince. *Geometric Algebra for Computer Graphics*. Springer-Verlag TELOS, Santa Clara, CA, USA, 1 edition, 2008.
[40] F. Yamaguchi. *Computer-Aided Geometric Design*. Springer-Verlag Tokyo, Tokyo, Japan, 1 edition, 2002.

Acknowledgement

The author would like to thank to colleagues at the Shandong University in Jinan (China) and University of West Bohemia in Pilsen (Czech Rep.) for stimulating of this work and discussions made.

The research was supported by the University of West Bohemia - Institutional research support.

Appendix

The GPU implementation of the outer product for the E^3 case using the homogeneous coordinate is quite simple. It should be noted that only 4 clocks for the *outer product* and 4 clocks for the *inner product* are needed.

```
float4 a;
a.x = dot(x1.yzw, cross(x2.yzw, x3.yzw));
a.y = - dot(x1.xzw, cross(x2.xzw, x3.xzw));
a.z = dot(x1.xyw, cross(x2.xyw, x3.xyw));
a.w = - dot(x1.xyz, cross(x2.xyz, x3.xyz));
return a;
```

or more compactly as:

```
float4 cross_4D(float4 x1, float4 x2, float4 x3)
return(
    dot(x1.yzw, cross(x2.yzw, x3.yzw)),
    - dot(x1.xzw, cross(x2.xzw, x3.xzw)),
    dot(x1.xyw, cross(x2.xyw, x3.xyw)),
    - dot(x1.xyz, cross(x2.xyz, x3.xyz))
);
```

THE FINITE ELEMENT METHOD – 55 YEARS OF MATHEMATICAL THEORY

Jiří Vala
Brno University of Technology, Faculty of Civil Engineering,
Institute of Mathematics and Descriptive Geometry,
e-mail vala.j@fce.vutbr.cz

Abstract: *The series of pioneering papers from the years 1967 and 1968, including the substantial contribution of the research group at Brno University of Technology and further scientists in former Czechoslovakia, can be seen as the origin of the mathematical theory of the finite element method, whose industrial applications date back (at least) to 1952. This paper tries i) to present the motivation for the finite element method together with its classical theory of convergence of approximate solutions and ii) to sketch its later development, driven by the progress both in mathematical and numerical analysis and in computational hardware and software, iii) supplied by an example significant for the design of building structures.*

Keywords: finite element method, numerical mathematics, computational mechanics, quasi-brittle fracture.

INTRODUCTION

The years 1967 and 1968, well-known thanks to significant political changes in former Czechoslovakia, stopped by the Soviet military intervention, can be seen also as the early years of the mathematical theory of the finite element method with successful continuation, where the contribution of the Czechoslovak researchers cannot be omitted. The engineering formulations of such numerical approach, validated by important industrial applications, are (at least) 15 years elder: thus most engineering journals took part on the celebration of 70 years of the finite element methods in the several last months. However, unlike [1], devoted to the beginnings of the finite element method in engineering computations, and [2], trying to sketch its later development, in this paper we shall prefer the point of view of the mathematical theory, although it contains and generates still new open questions, even in some problems covered by numerical solvers of commercial software packages.

Most engineering formulations rely on the physical principles of classical thermomechanics (typically in some their simplified versions) and lead to boundary and initial value problems for partial differential equations (PDEs) and their systems. Their analytical solutions, unlike ordinary differential equations (ODEs), are known only for selected model problems, unknown or exotic in practice, but useful for the software development and testing. Also most semi-analytical solutions (using Fourier series, Laplace or Fourier transforms, Green functions, etc.) are restricted to rather special classes of problems; moreover some difficulties in their numerical approaches (as evaluation of infinite integrals) are not included in original problems. This motivated the development of fully discretized approaches, generating a finite (but sufficiently large) number of algebraic equations, linear if possible, with sparse system matrices: always for originally linear problems, for particular steps of iterative procedures for nonlinear ones. Such methods can be distinguished by their approach to the discretization: in their rough classification the (at least historically) natural first

choice is the finite difference method (FDM), the second (more advanced) one the finite element method (FEM), the third one the problem-oriented cooperation of FEM with further approaches, including FDM.

A first model (combined Dirichlet and Neumann) boundary value problem for an ODE reads

$$-(au')' + bu = f \ \text{ for all } x \in [0,\ell]\,, \qquad u(0) = 0\,, \qquad au'(\ell) = g\,, \tag{1}$$

the prime symbol denoting $\mathrm{d}/\mathrm{d}x$ for brevity; $f(x)$ and g (a constant) are prescribed, as well as both coefficients $a(x)$ and $b(x)$, $u(x)$ is an unknown function. Applying FDM, dividing the interval $[0,\ell]$ into n subintervals $[\ell_{i-1},\ell_i]$, taking $\ell_0 = 0$ and $\ell_i = ih$ for $h = \ell/n$ and any $i \in \{1,\dots,n\}$, introducing $f_i = f(ih)$, $a_i = a(ih)$, $b_i = b(ih)$ and similarly also $u_i \approx u(ih)$ (a priori unknown), setting $u_0 = 0$, we can write

$$-a_i(u_{i+1} - 2u_i + u_{i-1}) + h^2 b_i u_i = h^2 f_i\,, \qquad u_{n+1} - u_{n-1} = 2hg\,, \tag{2}$$

which generates a system of linear algebraic equations; let us notice the tricky value u_{n+1}, not contained in (1).

An alternative approach works with the integration of the first equation (1), multiplied by an appropriate test function v, satisfying $v(0) = 0$, by parts (at least in the distributive sense). The obvious result is

$$\int_0^\ell av'u'\,\mathrm{d}x + \int_0^\ell bvu\,\mathrm{d}x - [avu']_0^\ell = \int_0^\ell vf\,\mathrm{d}x\,. \tag{3}$$

Taking the last equation (1) into account, from (3) we obtain

$$\int_0^\ell au'v'\,\mathrm{d}x + \int_0^\ell bvu\,\mathrm{d}x = \int_0^\ell vf\,\mathrm{d}x + v(\ell)g\,. \tag{4}$$

Using the notation $(w,\tilde w)$ for the integrals of products of $w\tilde w$ on $[0,\ell]$ for any appropriate functions w and $\tilde w$ and, moreover, the notation $\langle v\,,g\rangle = v(\ell)g$ formally, we can rewrite (4) as

$$(v', au') + (v, bu) = (v, f) + \langle v, g\rangle\,. \tag{5}$$

Applying the standard definitions of Lebesgue and Sobolev spaces by the monograph of T. Roubíček on PDEs (2005) [3], Part 1, it is natural (cf. the second equation (1)) to introduce the space of test functions $V = \{w \in W^{1,2}(0,\ell) : w(0) = 0\}$, to assume $f \in L^2(0,\ell)$ and always positive $a, b \in L^\infty(0,\ell)$ and to seek for an unknown $u \in V$ by (5); $(..)$ can be than interpreted as the scalar product in $L^2(0,\ell)$. Such unique u can be also derived as minimum of the quadratic functional

$$G(v) = \frac{1}{2}(v', av') + \frac{1}{2}(v, bv) - (v, f) - \langle v, g\rangle \tag{6}$$

where $v \in V$ again, i. e. $G(u) \le G(v)$ for any $v \in V$, whereas $G(u) = G(v)$ just for $u = v$.

Usually (5) is referred as the weak (Galerkin) formulation of (1) and the requirement to minimization of $G(v)$ in (6) as the variational (Ritz) formulation of (1). Both formulations (5) and (6) admit various approximations of an unknown function u, exploiting some basis of test functions v, which can result in different classes of numerical algorithms. In particular, in the simplest choice of FEM,

taking u as linear Lagrange splines (continuous piecewise linear functions) using the same nodes $0, h, \ldots, nh = \ell$ as in FDM, which can be expressed as linear combinations of n piecewise linear basis functions $v_i(x) = (x - (i-1)h)/h$ for $x \in [(i-1)h, ih]$ and $v_i(x) = ((i+1)h - x)/h$ for $x \in [ih, (i+1)h]$ ($i = n$ is not allowed in this case), zero otherwise, with $i \in \{1, \ldots, n\}$, we are allowed to introduce V_h as a finite-dimensional subspace of V, supplied by the above sketched basis. Consequently the discretized form of (5) is

$$(v_h', a u_h') + (v_h, b u_h) = (v_h, f) + \langle v_h, g \rangle \,, \tag{7}$$

which for all $v_h \in V_h$ generates a system of linear algebraic equations again, very similar to (2), obtained from FDM: the unknowns are the approximations of u in particular nodes again, some differences can be observed due to numerical quadrature rules handling a, b and f, applied to particular additive terms of (7), moreover in FEM no tricky node is needed. All details, including simple examples, available in the MATLAB environment, and numerous illustrations, can be found e. g. in [4], Parts 8 (FDM), 9 (numerical quadrature) and 10 (FEM).

For most students of numerical mathematics without strong motivation for the study of FEM, coming from real physical, engineering, etc. applications areas, such conclusion is far from being satisfactory: the approach of FEM seems to be a strange and complicated way how to come to the same or very similar results as from the much more simple and transparent approach of FDM. A second model boundary value problem can be useful to overcome such opinion. Let us introduce a domain Ω in the N-dimensional Euclidean space $\mathcal{R}^N$ with a Lipschitz boundary $\partial\Omega$: frequently we have $N = 2$ or $N = 3$ in applications, for $N = 1$ we should come back to a first model boundary value problem. Let us assume that $\partial\Omega$ consists of two disjoint parts Θ and Γ where Θ has a non-zero Hausdorff measure on Ω. We shall work with some Cartesian coordinate system $x = (x_1, x_2, x_3)$ in $\mathcal{R}^N$, using the Hamilton operator $\nabla = (\partial/\partial x_1, \partial/\partial x_2, \partial/\partial x_3)$, the central dot for the scalar product in $\mathcal{R}^N$ and $\mathcal{N}$ for the outer unit normal to $\partial\Omega$ Thus an announced second model problem for a PDE reads

$$-\nabla \cdot (a \nabla u) + bu = f \ \text{ on } \Omega \,, \qquad u(0) = 0 \ \text{ on } \Theta \,, \qquad a \nabla u \cdot \mathcal{N} = g \ \text{ on } \Gamma \,; \tag{8}$$

f on Ω and g on Γ are prescribed, as well as both coefficients a and b on Ω, u is a unknown function on Ω. Analogously to (2), FDM works for a paralleliped $\Omega = [0, \ell_1] \times \ldots \times [0, \ell_N]$ efficiently (the details can be left the patient reader), in most other cases one comes to serious difficulties with the assertion of boundary conditions of both types on a (in general) curved $(N-1)$-dimensional boundary.

An alternative approach relies on the integration by parts again, assuming $v = 0$ on Θ and utilizing the Green - Ostrogradskiĭ theorem (whose all proofs are rather difficult, unlike the case $N = 1$), with the result

$$\int_\Omega a \nabla v \cdot \nabla u \, \mathrm{d}x + \int_\Omega bvu \, \mathrm{d}x - \int_{\partial\Omega} av \nabla u \cdot \mathcal{N} \, \mathrm{d}s(x) = \int_\Omega vf \, \mathrm{d}x \,. \tag{9}$$

Taking the last equation (8) into account, from (9) we obtain

$$\int_\Omega a \nabla v \cdot \nabla u \, \mathrm{d}x + \int_\Omega bvu \, \mathrm{d}x = \int_\Omega vf \, \mathrm{d}x + \int_\Gamma vg \, \mathrm{d}s(x) \,. \tag{10}$$

Using the notation $(w, \tilde{w})$ for the integrals of products of $w\,\tilde{w}$ on Ω, $(\!(\mathcal{W},$
$tilde\mathcal{W}$ for the integrals of $\mathcal{W} \cdot \tilde{\mathcal{W}}$ on Ω where W and $\mathcal{W}$ are vector functions with values in $\mathcal{R}^N$ and $\langle w, \tilde{w} \rangle$ for those on Γ, we can rewrite (10) as

$$(\!(\nabla v, a\nabla u)\!) + (v, bu) = (v, f) + \langle v, g \rangle\,. \tag{11}$$

Here it is natural (cf. the second equation (8)) to introduce the space of test functions $V = \{w \in W^{1,2}(\Omega) : w = 0 \text{ on } \Theta\}$, to assume $f \in L^2(\Omega)$, $g \in L^2(\Gamma)$ and always positive $a, b \in L^\infty(\Omega)$ and to seek for an unknown $u \in V$ by (11); $(.\,,.)$ $(\!(.\,,.)\!)$ and $\langle .\,,.\rangle$ can be than interpreted as the scalar products in $L^2(\Omega)$, $L^2(\Omega)^N$ and $L^2(\Gamma)$. Such unique u can be also derived as minimum of the quadratic functional

$$G(v) = \frac{1}{2}(\!(\nabla v, a\nabla v)\!) + \frac{1}{2}(v, bv) - (v, f) - \langle v, g \rangle \tag{12}$$

where $v \in V$ again.

Preferring (11) to (12) also in these considerations, as well as the basis of linear Lagrange splines on N-dimensional simplices (triangles, tetrahedra, ...), for Ω compound from such simplices and Θ and Γ just from their N-dimensional faces we can repeat the arguments from the case $N = 1$: V_h is certain n-dimensional subspace of V, whose basis consists of special Lagrange splines with unit values just in 1 node; finally we receive

$$(\!(\nabla v_h, a\nabla u_h)\!) + (v_h, bu_h) = (v_h, f) + \langle v_h, g \rangle \tag{13}$$

for each $v_h \in V_h$; h here can be understood as the longest applied edge in the above sketched decomposition of Ω to n-dimensional simplices. Let us remark that this approach becomes more delicate in the case of still more general Ω, Θ and Γ where the approximating n-dimensional spaces are not exactly subspaces of V; this brings an additional potential source of numerical errors. Nevertheless, FEM is much more robust than FDM to handle (at least) complicated geometrical configurations. Moreover, only the approximations of first derivatives (simple functions here) occur in (13), similarly to (7), unlike the second differences in (2) and its hypothetical N-dimensional generalization.

1 PREREQUISITES AND THE CLASSICAL THEORY BEFORE 1990

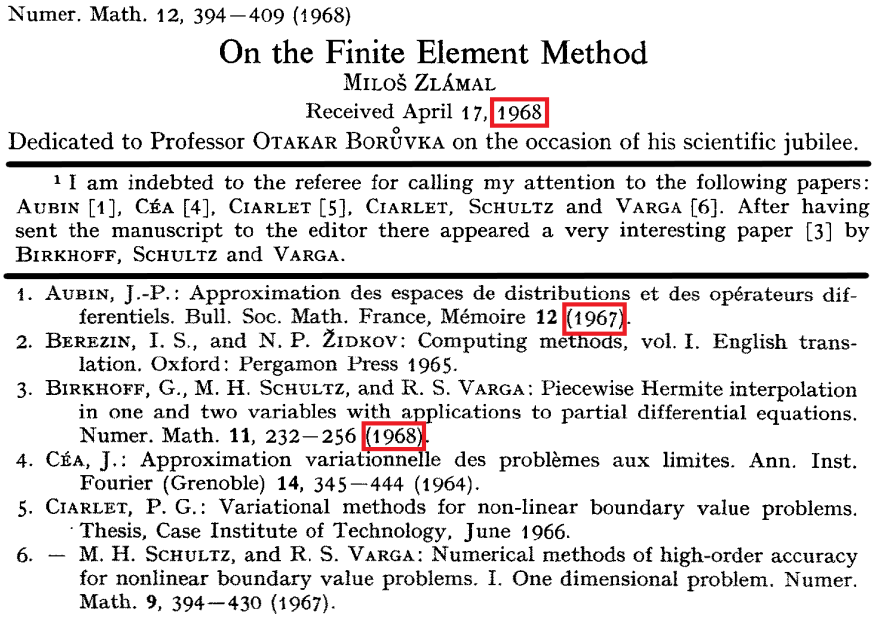
The evident drawback of FEM, in comparison to FDM, is that the proper study of its convergence cannot be performed using the standard mathematical approaches developed in the 18th and 19th century. However, the fundamentals of variational methods, as an important tool needed by FEM (and many other computational approaches), were invented by L. Euler (1744) in [5] yet. The approximate solution of the problem of minimal area was constructed using the piecewise linear functions on triangles by by K. H. Schellbach (1851) in [6]; this can be seen as the first ad hoc implementation of the finite element technique. Nevertheless, the non-existence of appropriate computational machines limited such approach strongly for the long time. The minimization of energy potential (functional) was used for the justification of the existence of solution of the Poisson equation $hu = f$ in certain space of integrable function with scalar product by D. Hilbert (1901), summarized (with numerous other results) in [7]; this is the origin of the modern nomenclature of Hilbert spaces where the existence of scalar products is important for the construction

of orthogonal projections, namely from original function spaces to approximating subspaces in numerical algorithms. A more general minimization principle was formulated by W. Ritz (1909) in [8]; after its publication J. W. Strutt, well-known as baron Rayleigh among physicists, criticized the formulation "new method" in the title: indeed, similar minimization approach using eigenfunctions can be traced in his much elder monograph (1877) [9]. The weak formulation of some engineering motivated problems without existence of potentials like (6) or (12) was presented by B. G. Galerkin (1915) in [10]. The still used approximation spaces for such formulation were designed by G. I. Petrov (1940) in [11]. Certain comparable numerical approach, referring to the linear problem of elasticity, was suggested by A. Hrennikoff (1941) in [12]. The progress in computer hardware and software brought the renaissance of approximations with piecewise linear functions of triangles: namely R. Courant applied such approach more times to for the analysis of the Laplace equation based on the energy minimization, typically (1922, 1943) in [13] and [14].

Fig. 1 Boeing YB52 prototype, designed using the intuitive finite element technique (historical photo from https://www.thisdayinaviation.com/tag/boeing-yb-52-stratofortress/, 15th April 1952).

The above mentioned new possibilities in scientific and technical computations supported also the design of real structures under mechanical loads. The engineering society refers namely to the successful FEM-based dynamical design of wings of Boeing YB52 – see Fig. 1. The plane stresses were approximated by clever intuitive combinations of simple functions, as announced by N. J. Turner (1952), followed (with certain delay, 1956) by the proper publication in [15]. As "the finite element method" this approach was presented later by R. W. Clough (1960) in [16] (or probably 1956 less officially yet). Many ideas built in modern FEM algorithms have their origin just in several following years: e. g. hybrid elements, Lagrange multipliers, and searching for saddle point of functionals are mentioned in the monograph of O. C. Zienkiewicz (1967) [17]. This significant applied research progress highlighted the absence of mathematical background of FEM. Thus in (at least) two following decades we can trace two quite different trends in the

theory of numerical methods: i) the effort to reduce FEM to a choice of special basis functions accompanied by some (not very transparent) mathematical manipulations, in the reasonable cases fully transferable to FDM, ii) the development of FEM as a general tool, applicable to nearly all physical and engineering problems, sending FDM, together with various semi-analytical methods, to the past of scientific computations. During these decades the trend ii) prevailed totally, although in a slightly less optimistic context: the careful problem-oriented application of FEM is needed, collaborating with further methods, including FDM.



Numer. Math. 12, 394—409 (1968)

On the Finite Element Method

MILOŠ ZLÁMAL

Received April 17, 1968

Dedicated to Professor OTAKAR BORŮVKA on the occasion of his scientific jubilee.

[1] I am indebted to the referee for calling my attention to the following papers: AUBIN [1], CÉA [4], CIARLET [5], CIARLET, SCHULTZ and VARGA [6]. After having sent the manuscript to the editor there appeared a very interesting paper [3] by BIRKHOFF, SCHULTZ and VARGA.

1. AUBIN, J.-P.: Approximation des espaces de distributions et des opérateurs differentiels. Bull. Soc. Math. France, Mémoire **12** (1967).
2. BEREZIN, I. S., and N. P. ŽIDKOV: Computing methods, vol. I. English translation. Oxford: Pergamon Press 1965.
3. BIRKHOFF, G., M. H. SCHULTZ, and R. S. VARGA: Piecewise Hermite interpolation in one and two variables with applications to partial differential equations. Numer. Math. **11**, 232—256 (1968).
4. CÉA, J.: Approximation variationnelle des problèmes aux limites. Ann. Inst. Fourier (Grenoble) **14**, 345—444 (1964).
5. CIARLET, P. G.: Variational methods for non-linear boundary value problems. Thesis, Case Institute of Technology, June 1966.
6. — M. H. SCHULTZ, and R. S. VARGA: Numerical methods of high-order accuracy for nonlinear boundary value problems. I. One dimensional problem. Numer. Math. **9**, 394—430 (1967).

Fig. 2 The title of the celebrated article [25], supplied by the begin of *References*; the years of publishing of most relevant articles for its analysis are highlighted.

It is not easy to set some symbolic starting date for the mathematical theory of FEM. From 1960 the series of papers, working with functional analysis, Sobolev spaces etc., prepared all needed ingredients for the proper analysis of convergence properties of FEM for a sufficiently large class of problems. A system of piecewise linear functions on triangles, covering a rather general domain Ω in $\mathcal{R}^2$, was applied to the Laplace equation $hu = 0$, supplied by both Dirichlet and Neumann boundary conditions by K. O. Friedrichs (1962) in [18]; this report contains probably the first correct proof of convergence of FEM in the Hilbert spaces $L^2(\Omega)$ and $W^{1,2}(\Omega)$, although no convergence rate is guaranteed. The theory of approximation for variational problems was developed by J. Cèa (1964) in [19]. For sufficiently smooth solutions u and certain strategy of generation of a triangular mesh in $\mathcal{R}^2$, introducing selected boundary value problems for the Laplace equation and some its generalizations, the linear convergence rate of the type

$$\|u - u_h\|_{W^{1,2}(\Omega)} \leq \mathfrak{C}h\|u\|_{W^{2,2}(\Omega)} \tag{14}$$

where a positive $\mathfrak{C}$, independent of u, exists, was verified by L. A. Oganesyan (1964) in [20]. The estimate (14) can be still seen as a fundamental formula for the quality of convergence of the so-called h-version of FEM; later improvements replaced only the norm in $W^{2,2}(\Omega)$ by the seminorm, working with second derivatives only, specified some computable bounds for $\mathfrak{C}$, etc. For the linear

theory of elasticity in $\mathcal{R}^2$ the formal proof of convergence of FEM (without rate considerations) was done by F. Kang (1965) in [21], covering various element types: triangles, parallelipeds (beyond the piecewise linear test functions) hanging nodes (as the first step to discontinuous approximations), etc. The attempts to handle nonlinear problems can be traced from the contribution of P. G. Ciarlet (1967) in [22]: the solution of systems of (usually spares) systems of linear algebraic equations must be accompanied by additional iterative procedures.

For the further improvement of guaranteed convergence properties for FEM some deeper results from the theory of approximation and interpolation in general function spaces were needed. Such results for the approximation of operators in the spaces of distributions were derived by J.-P. Aubin (1967) in [23]. The Hermite interpolation for PDEs was elaborated by G. Birkhoff et al. (1968) in [24]. Consequently M. Zlámal (1968) in his famous article [25], whose title page is presented by Fig. 2, derived the convergence rate also for quadratic elements and a model equation of second order in $\mathcal{R}^2$, using the so-called minimum angle condition: the decomposition of Ω to triangles, introducing h as the maximal length over all applied triangle edges, passing $h \to \infty$, must be performed is such way that all triangular elements have still their minimal angles greater or equal to a fixed positive constant. This condition is equivalent to the requirement that each triangle must contain such circle with a radius ρ that ρ/h does not exceed a fixed positive constant; for $N > 2$ (not discussed in [25]) this requirement should be preferred because of its more transparent generalization from triangles to N-dimensional simplices and circles to N-dimensional spheres. In the theory of FEM, this is well-known as the regular family of decomposition of Ω to simplices.

Fig. 3 The hypothetical centre of universe in Brno, crossing of Demlova and Trávníky (Lawns) Streets (author's photo, 2012). The English translation of the Czech text: *The centre of universe is everywhere. This one is preferred by Dr. Jiří Grygar and Prof. Alexander Ženíšek.*

The results of M. Zlámal, who was the director of the Laboratory of Computing Machines at Brno University of Technology (BUT), with close connections to several faculties of BUT, were supported by the physical, mathematical, numerical and computational analysis of the team of his

collaborators, up to the software development and its application to selected engineering problems. Several these experts contributed to the theory of FEM significantly, as professors at BUT. Except M. Zlámal, they were physicists or engineers originally, but changed their research priorities to FEM. In particular, from 4 authors of the monograph [26] (1979) covering both engineering and mathematical approaches, only F. Leitner continued his career as engineer in the design of water structures, unlike V. Kolář (expert in traffic structures, Faculty of Civil Engineering BUT), A. Ženísek (theoretical physicist, Faculty of Mechanical Engineering BUT) and J. Kratochvíl (expert in water structures, Faculty of Civil Engineering BUT), as well as unlike F. Melkes, the author of [27] (1972) (expert in electrical machines, Faculty of Electrical Engineering and Communication BUT). This research was continued (even outside BUT) in some reasonable form also in the period of political repressions in Czechoslovakia 1970–1989. Moreover, all these persons were active in further, seemingly remote disciplines, as in astrophysics (with his friend J. Grygar), poetry, singing and advanced card games in the case of A. Ženíušek, in addition to his director's duties at the Institute of Mathematics of the Faculty of Mechanical Engineering of BUT, as documented only partially by Fig. 3. Let us remind also some further worldwide-known professors from former Czechoslovakia, contributed to the mathematical theory of FEM, mentioned in this paper (in the alphabetical order): I. Babuška (originally Prague, since 1968 College Park (Maryland) and Austin (Texas)), V. Dolejší (Prague), Z. Dostál (Ostrava), M. Feistauer, M. Křížek, K. Rektorys, T. Roubíček (all Prague), J. Sládek and V. Sládek (Bratislava).

Parallel to FEM, the development of variational methods for the numerical analysis of PDEs cannot be omitted, namely the books of K. Rektorys (1974, 1985) [28] (on elliptic equations) and [29] (on parabolic and hyperbolic equations). Another remarkable discussion was connected with the minimum angle condition by [25]: its partial improvement by P. Jamet (1976) in [30] had a negligible public acceptance, but its later modification by M. Křížek (1991) in [31] became a popular part in most studies of FEM convergence. Its principal idea is to replace the minimum angle condition by the maximum angle one: all triangular elements must still have their maximal angles lesser or equal to a fixed constant (lesser than π). This condition is equivalent to the requirement that each triangle must be contained in such circle with a radius $\mathfrak{R}$ that $h/\mathfrak{R}$ does not exceed a fixed positive constant; the generalization for $N > 2$ with N-dimensional spheres and simplices is obvious. In the theory of FEM, this is well-known as the semiregular family of decomposition of Ω to simplices. Unfortunately, unlike the case of piecewise linear test functions, this does not generate the estimates of the type (14) in more complicated cases automatically, thus the general strategy of choice of such families of decompositions, with the additional influence of numerical integration, cannot be seen as a quite closed problem yet. For problem-oriented algorithms, namely for those working with local mesh refinements in the classical h-version of FEM, the so-called p-version of FEM can be useful, working with the increasing degree of piecewise polynomial test functions; for the elastoplastic deformation such approach was elaborated by B. A. Szabó (1978) in [32]. The combined hp-version of FEM was introduced and studied by I. Babuška (1988) in [33].

2 EXPANDING RESEARCH DIRECTIONS AND APPLICATIONS

The proper critical overwiev of further development of FEM, its alternatives and their combinations could have (at least) the similar extent for any such particular method as all preceding sections of this paper, due to expanding research directions and applications since 1990, which is not accept-

able for the paper of this type. Thus we shall continue only with brief comments and references, without detailed explanation of principles. In general, the increasing abilities of hardware and software make it possible to come from 2-dimensional to 3-dimensional and even (in special applications) to more-dimensional problems. The utilization of hp-version of FEM supports the mesh adaptivity together with the choice (usually still polynomial) test functions, supported by various a priori and a posteriori error estimates, working with real discrete partial results, unlike (14) (where the right-hand side contains an unknown u). Moreoever, polynomials can be replaced by other quasi-orthogonal functions like wavelet, frames, etc., as explained by X. Chen et al. (2004) in [34]. Namely in damage mechanics special enrichment strategies for the bases of test functions have been elaborated: namely the partition of unity FEM (PU-FEM) by J. Melenk and I. Babuška (1996) in [35], the generalized FEM (G-FEM) by T. Stroubolis et al. (2001) in [36] and the (extrinsic) extended FEM (X-FEM) by N. Moës at al. (1999) in [37], whose intrinsic variant by T.-P. Fries (2006) in [38] needs no additional degrees of freedom thanks to an advanced adaptive strategy (i. e. the order n of a resulting system of linear algebraic is not increased). For large problems as more deformable bodies with contact conditions the distributed and parallel computations with successive updates of contact data can be useful, supported by the technique of finite element tearing and interconnecting (FETI) by Z. Dostál et al. (2010) in [39]. This could be coupled with the discontinous Galerkin formulation of FEM (DG-FEM) by M. Dolejší and M. Feistauer (2015) in [40], applicable to the Navier - Stokes equations of (both laminar and turbulent) fluid flow, although the existence, uniqueness and smoothness of solutions contain still a lot of open questions in such case.

Certain comeback of FDM, upgraded by L. Gavete at al. (2003) in [41], can be registered for PDEs of evolution of both parabolic and hyperbolic types, as heat transfer (energy conservation, a parabolic PDE), moisture or contaminant transfer in porous medium (mass conservation, a parabolic PDE), dynamics of deformable bodies under mechanical loads (energy conservation, a hyperbolic PDE), etc. Thanks to the properties of Rothe sequences in Bochner - Sobolev spaces, introduced by [3], Part 7, one can decompose, using the Cacuchy initial conditions, the solution od such PDE, step by step in time, to a series of solutions of elliptic PDEs, using FEM typically. Only for illustration: let us consider a parabolic model problem, slightly adopted from (11),

$$(v, c\dot{u}) + (\!(\nabla v, a\nabla u)\!) + (v, bu) = (v, f) + \langle v, g\rangle\,, \tag{15}$$

valid for each $t \in \mathcal{I}$, $\mathcal{I} = [0, \tau]$ being a finite time interval; the upper dot symbol means $\partial/\partial t$ for brevity. Unlike positive time-independent characteristics $a, b, c \in L^\infty(\Omega)$, taking $H = L^2(\Omega)$ and $Z = L^2(\Gamma)$, we suppose $f \in L^2(I, H) \cong L^2(\Omega \times \mathcal{I})$ and $g \in L^2(I, Z) \cong L^2(\Gamma \times \mathcal{I})$. The Cauchy initial condition $u(., 0) = u_0$ with $u_0 \in V$ is prescribed; the aim is to find an unknown $u \in L^2(I, V)$ such that its time derivative satisfies $u \in L^2(I, H)$, i. e. $u \in W^{1,2,2}(I, V, H)$. Thus the time discretization can be motivated by the classical Euler explicit method, taking ϕ_s instead of $\phi(s\delta)$, $s \in \{1, \ldots, m\}$, $\delta = \tau/m$, for appropriate time-dependent functions ϕ, similarly to functions of x iintroduced above (2), in particular $u_s \approx u(s\delta)$ (for an unkonwn u), with the result

$$(v, c_s(u_s - u_{s-1})) + \delta(\!(\nabla v, a_s\nabla u_s)\!) + \delta(v, b_s u_s) = \delta(v, f_s) + \delta\langle v, g_s\rangle\,, \tag{16}$$

valid for each $v \in V$. The announced Rothe sequences for any integer m can be then compound from $(u_s - u_{s-1})/\delta$ (linear Lagrange splines on $\mathcal{I}$) and u_s (simple functions on $\mathcal{I}$); the constructive proof of existence of a unique solution of (15) is than based on the limit passage from (16) to (15) assuming $m \to \infty$. For a model hyperbolic problem we could take $\ddot{u}$ instead of u in (15) and

incorporate the second Cauchy initial condition $\dot{u}(.,0) = \hat{u}_0$ where $\hat{u}_0 \in V$ is prescribed, too; thus it is natural to search for $u \in W^{2,2,2,2}(I, V, H, V^*)$, V^* being the adjoint space to V. An alternative approach for both parabolic and hyperbolic problems relies on the so-called method of lines: using the multiplicative Fourier decomposition, involving e, g. some FEM basis, we are able to come, instead of an original PDE, to a sparse system of time-variable ODEs; in most practical problems an additional time discretization is needed (because of the expensive eigenvalue analysis), thus such approach leads to a very similar algorithm to the original one.

A concurrent method to FEM is the (also integral) finite volume method (FVM), avoiding the integration by parts like (3) and (9), thus the better fulfilment of conservation principle at the discretized level (not only in the limit case $h \to 0$) can be expected, as derived by Cai (1990) in [42], paraphrasing the title of [25]. FVM is frequently used (for some variables) in combination with FEM: e. g. the primary triangular mesh for FEM in $\mathcal{R}^2$ is accompanied by the secondary polygonal mesh around the nodes of the primary one. In some cases also certain (typically incomplete) knowledge on general soultions can be exploited, even without FVM, e, g. using the theory of Green functions, which leads to the boundary element method (BEM), introduced by the brothers J. Sládek and V. Sládek (1990) in [43]: the dimension N for discretization is reduced to $N-1$, but on a curved boundary where Dirac distributions, Heaviside functions, etc. occur. Another idea could be the adaptation of the least square technique to FEM (LS-FEM), involving both a given PDE and all boundary conditions via appropriate weights; this rather rare approach was optimized by B.-N. Jiang and L. Povinelli (1993) in [44].

New research directions related to FEM were initiated by the development of advanced materials, structures and technologies, too, due to the necessity of incorporation of randomness of their structure and bridging available data from laboratory and in situ micro- and macrostructural experiments. The methodology of handling uncertain data in initial conditions, geometrical description and material properties, referred as stochastic FEM (S-FEM), was suggested by I. Babuška et al. (2005) in [45]. The more-scale computations (primary deterministic, mostly periodic) can be performed using the multiscale FEM (Ms-FEM), introduced by T. Hou and Y. Efendiev (2009) in [46]. Moreover, the recent modelling and simulation approaches to continuum mechanics, including direct, sensitivity and inverse problems, force the analysis of multi-physical, strongly nonlinear initial and boundary value problems for systems of PDEs of evolution frequently, thus the development of robust solvers of systems of nonlinear algebraic equations is needed (utilizing the inexact Newton method, the conjugate gradient method, the Nelder - Mead simplex method, some selected soft computing tricks, ...).

Even for nonlinear generalizations of problems like (11) and (15), the typical physical fundamentals are i) the principle of conservation of scalar quantities, usual in classical thermodynamics, as mass, components of (linear and angular) momentum and energy (the first thermodynamic law), ii) constitutive relations of selected types in form of inequalities, covering admissible irreversible processes, or direct incorporation of such special relations, as traditional in the engineering theories of plasticity, damage, contacts / impacts of deformable bodies, etc. (the second thermodynamic law), and iii) the positivity of Kelvin temperature (the third thermodynamic law). As typical examples of nonlinear problems, whose existence proofs (overcoming technical difficulties) can be performed similarly to that sketched above, in particular to (9) here, with the obvious application of FEM to

computational analysis, we can present

$$((\nabla v, a(u^{k-1})\nabla u^k)) + (v, bu^k) = (v, f) + \langle v, g\rangle\,, \tag{17}$$

(a generalized Poisson equation with a nonlinear coefficient), or

$$((\nabla v, a|\nabla u^{k-1}|^{p-2}\nabla u^k)) + (v, bu^k) = (v, f) + \langle v, g\rangle \tag{18}$$

with some finite real $p > 2$ (another generalization with the so-called p-Laplacian); the upper indices $k \in \{1, 2, \ldots\}$ refer to particular steps of the method of successive approximation. Unfortunately, such simple approach, as that suggested by (17) and (18), to the analysis of numerous strongly nonlinear problems of physical and engineering practice, is quite impossible, including the Navier - Stokes equations of fluid flow, or the Maxwell equation of an electromagnetic field: cf. the still unsolved 4th Millenium Prize Problem [47]. Moreover, all partial existence results, as that of E. Feireisl and M. Novotný (2022) in [48], work with dissipative varifold solutions, or other very abstract definitions, whose effective numerical characterization and reasonable physical interpretation form an additional serious problem. Other difficulties are contributed by physicists: still more complicated multiple-scale formulations include a lot of internal variables, whose reliable quantitative identification may be even more complicated than the analysis of the original direct simulation problem. Thus most engineering computational approaches, summarized by B. Szabó and I. Babuška (2021) in [49], can be seen as certain compromises between i) the requirement of practice, including the time and money for the complete analysis, ii) the formal existence and convergence theory iii) the robustness and effectiveness of computations and iv) the transparency and reliability of numerical results. Such compromise will be demonstrated on the following numerical example, referring to the recent research at the Faculty of Civil Engineering of BUT: cf. the research project mentioned in *Acknowledgement*.

3 AN ILLUSTRATIVE EXAMPLE

A significant problem of engineering practice is the reliable prediction of strain and stress development of cement-based composites under mechanical, thermal, etc. loads. Because of their poor behaviour in tension, causing the risk of irreversible damage, such composites contain some stiffening components in most applications, e. g. the metal fibres implemented in the following example. The crucial difficulty of all computational models is the incorporation of some scale bridging. The so-called quasi-brittle behaviour of such composites can be characterized roughly by 4 phases: i) elastic deformation, ii) creation of damaged zones with microscopic cracks, iii) initiation and development of macroscopic cracks, iv) total destruction of material structure. The model presented here comes from the both geometrical and physical linearization for i), implementing two different nonlinear modifications, taking ii) and iii) into account: the stiffness decrease due to certain damage factor, introduced by P. Havlásek et al. (2016) in [50], using the nonlocal stress evaluation by A. C. Eringen (1984), coming from [51], for ii), and the dynamics of cohesive interfaces, as analyzed by M. G. Pike and C. Oskay (2015) in [52]. However, in such simplified model the analysis of complicated systems of macrocropic cracks by iii), induced by extensive damaged zone by ii), tending to the final stage iv), is not realistic.

We shall demonstrate the problem formulation in $\mathcal{R}^N$ for $N = 3$; the details of its plane stress simplification with $N = 2$, as presented by Fig. 4, will be left to the curious reader. Let us consider

Ω as a union of a finite number of domains with Lipschitz boundaries in $\mathcal{R}^3$. The set of all such boundaries $\partial\Omega$ consists of 3 parts: Θ for Dirichlet boundary conditions, Γ for Neumann boundary conditions and Λ for interior interfaces (potential macroscopic cracks and matrix / fibre contacts). We shall study the development of a displacement u, related to the initial configuration of Ω, under the volume loads $f \in L^2(I, H)$ and the surface loads $g \in L^2(I, Z)$; here $H = L^2(\Omega)^3$ and $Z = L^2(\Gamma)^3$ (because all values of u, depending on x and t, have 3 components now); similarly we introduce $X = L^2(\Lambda)^3$. The conservation of (linear) momentum then reads

$$(v, \rho\ddot{u}) + (v, \alpha\rho\dot{u}) + (\!(\varepsilon(v), \sigma)\!) = (v, f) + \langle v, g\rangle + \langle \mathcal{D}v, \mathcal{T}\rangle_* \tag{19}$$

for any $v \in V$: here $V = \{w \in W^{1,2}(\Omega)^3 : w = o \text{ on } \Theta\}$ where o denotes the zero vector in $\mathcal{R}^3$. The initial Cauchy conditions are $u(.,0) = o$ and $\dot{u}(.,0) = \hat{u}_0$ for some prescribed initial velocity $\hat{u}_0 \in V$. In (19) the following scalar products occur now: $(.\,,.)$ in H, $\langle .\,,.\rangle$ in Z, $\langle .\,,.\rangle_*$ in X and $(\!(.\,,.)\!)$ in $H \times H$; moreover we utilize the standard small strain tensor $\varepsilon(v)$, whose components are $\varepsilon_{ij}(v) = (\partial v_i/\partial x_j + \partial v_j/\partial x_i)/2$ for all $i, j \in \{1, 2, 3\}$, and $\mathcal{D}v$ means the difference between the traces of v on Λ We consider also the material density $\rho \in L^\infty(\Omega)$ and the mass damping factor $\alpha \in L^\infty(\Omega)$; its implementation brings some energy dissipation during the strain and stress development into account even in the case i): we cannot have a closed physical system in practice. It is natural to expect $u \in W^{2,2,2,2}(I, V, H, V^\star)$ again.

Clearly (19) must be supplied by some constitutive equations for still undefined stresses σ on Ω and surface tractions $\mathcal{T}$ on Λ. Let us notice that for the case i) without any active cohesive interface Λ we can have still a linear problem, working with the empirical Hooke law $\sigma = C\varepsilon(u)$ where $C \in L^\infty(\Omega)^{(3\times3)\times(3\times3)}_{\text{sym}}$; thus C contains (in general) 21 independent parameters, reducible to the pair of the well-known Lamé coefficients (or to the Young modulus and the Poisson ratio) for any isotropic medium. Such problem could be handled using the standard arguments from [28] and [29]. To incorporate ii), let us take $(1 - \mathfrak{D})\, C$ instead of C in the strain - stress relation where $0 \le \mathfrak{D} \le \mathfrak{D}_*$; the prescribed upper bound $\mathfrak{D}_*$ must be lesser than 1 (to avoid iv)). The most delicate step is now the evaluation of $\mathfrak{D}$: working with the triple of principal stresses σ_1, σ_2 and σ_3 (to guarantee the objectivity), coming from the condition $\det(\sigma - \sigma_i I) = 0$ where $i \in \{1, 2, 3\}$, I being the unit matrix of order 3, we are able to evaluate their nonlocal values $\sigma_i^*(x) = \int_\Omega \mathcal{K}(x, \tilde{x})\sigma_i(\tilde{x})\,\mathrm{d}\tilde{x}$, applying an appropriate regularization kernel $\mathcal{K}$, taking σ_1^*, σ_2^* and σ_3^* as inputs for certain bounded continuous function $\omega(\sigma_1^*, \sigma_2^*, \sigma_3^*)$, we are able to set, in any time $t \in \mathcal{I}$, the factor $\mathfrak{D}$ locally as the maximum of values of ω over all times $\tilde{t} \in [0, t]$ (to respect the irreversibility of damage). To incorporate iii), we need, moreover, some continuous and bounded cohesive function $\varsigma(\mathcal{D}v)$, accepting all above mentioned traces, to be able to evaluate $\mathcal{T} = \varsigma(\mathcal{D}u)$. Thus, understanding $\mathfrak{D}$ here as $\mathfrak{D}(u)$, representing a very complicated function of u, from (19) we receive

$$(v, \rho\ddot{u}) + (v, \alpha\rho\dot{u}) + (\!(\varepsilon(v), (1 - \mathfrak{D})\, C\varepsilon(u))\!) = (v, f) + \langle v, g\rangle + \langle \mathcal{D}v, \varsigma(\mathcal{D}u)\rangle_*\,. \tag{20}$$

The time-discretized version of (20), analogous to (16), then reads

$$\begin{aligned}(v, \rho(u_s - 2u_{s-1} + u_{s-2})) + \delta(v, \alpha\rho(u_s - u_{s-1})) + \delta^2(\!(\varepsilon(v), (1 - \mathfrak{D}_s^*)\, C\varepsilon(u_s))\!)\\ = \delta^2(v, f_s) + \delta^2\langle v, g_s\rangle + \delta^2\langle \mathcal{D}v, \varsigma(\mathcal{D}u_s^*)\rangle_*\end{aligned} \tag{21}$$

for particular steps $s \in \{1, \ldots, n\}$; we have $u_0 = o$ and u_{-1} for the first step can be set using $u_{-1} = u_1 - 2\delta\hat{u}_0$. In the nonlinear terms by ii) and iii) u_s are replaced by u_s^*, which can be taken

as u_{s-1} for the first guess, which can start the iterative process for each fixed $s \in \{1, \ldots, n\}$. The reformulation of (21) in V_h instead of V seems to be straightforward, but the realistic analysis of crack development requires non-trivial X-FEM mesh refinement steps beyond the scope of this paper.

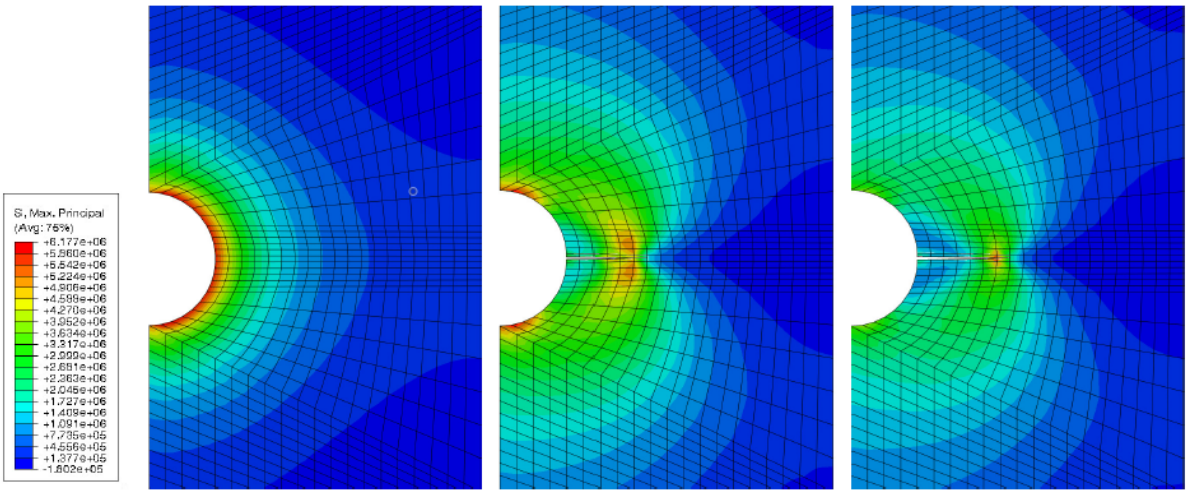


Fig. 4 The stress development in a cement-based composite structure with the metal fibre reinforcement, performed by the XFEM-based user-defined procedure in the Abaqus software, using the computational model, suggested by J. Mazars (2001).

Although the computational scheme (21) is not quite easy, it needs further improvements for practical implementations. Firstly, a more advanced dissipation scheme is required, combining the mass damping with the structural one, derived from the parallel Kelvin viscoelastic model, as presented by [53] (for the quasi-static case) and [54] (for the fully dynamic case). Secondly, the reasonable prediction of strain and stress redistributions in cement-based composites needs a more-parameter evaluation damage to distinguish between tension and compression; one rather simple model was suggested by J. Mazars (2001) in [55]. Fig. 4 documents such computational modelling for a simple plane structure. Nevertheless, further generalizations are needed for fast dynamical processes, as i) an advanced remeshing in particular time steps to suppress the effects of linearization, ii) replacement of (21) by a (at least conditionally stable) explicit computational scheme to avoid long iterations, together with the proper analysis of energy dissipation on contacts of multiple deformable bodies, as introduced by [56] in details, and iii) an effective contact detection, based on the results from the graph theory, forcing distributed and parallel computing platforms, suggested by [57].

CONCLUSION

The approximately 70 years of the FEM-based engineering computations and 55 years of the mathematical theory of FEM demonstrated its excellent properties and wide applicability, even beyond the scope of physical and engineering problems, sketched in this paper, and its ability to collaborate with numerous different numerical approaches in problem-oriented computations. Most commercial software packages of several last decades use FEM as their primary choice. The future

potential of the traditional FEM, or the dominance of some seemingly concurrent method, derived from FEM, or the assertion of a quite new method, as FEM instead of FDM at the beginning of the history discussed in this paper, depends on the success of its unified theoretical and numerical analysis for strongly nonlinear multiphysical problems, as the crucial research challenge for the near future.

References

[1] Gupta, K. K., Meek, J. L. A brief history of the beginning of the finite element method. *International Journal for Numerical Methods in Engineering* **39** (1996), p. 3761-3774.

[2] Sabat, L., Kundu, Ch. K. History of finite element method: a review. In: *Recent Developments in Sustainable Infrastructure* (Das, B. B., Barbhuiya, S., Gupta, R., Saha, P., eds.), p. 395-404.

[3] Roubíček, T. *Nonlinear Partial Differential Equations with Applications*. Basel: Birkhäuser, 2005.

[4] Vala, J. Numerická matematika. Brno: FCE BUT, 2021.

[5] Euler, L. *Methodus inveniendi lineas curvas maximi minimive proprietate gaudentes, sive Solutio problematis isoperimetrici latissimo sensu accepti*. Lausanne: M.-M. Bousquet & socios, 1744.

[6] Schellbach, K. H. Probleme der Variationsrechnung. *Journal fr reine und angewandte Mathematik* **41** (1851), p. 293-363.

[7] Courant, R., Hilbert, D. *Methoden der mathematischen Physik I, II*. Berlin: Springer, 1937.

[8] Ritz, W. Über eine neue Methode zur Lösung gewisser Variationsprobleme der mathematischen Physik. *Journal für reine und angewandte Mathematik* **135** (1909), p. 1-61.

[9] Strutt, baron Rayleigh, J. W. *The Theory of Sound*. London: Macmillan & Co., 1877.

[10] Galerkin, B. G. Rods and Plates, Series Occurring in Various Questions Concerning the Elastic Equilibrium of Rods and Plates. *Vestnik Inzhenerov i Tekhnikov* **19** (1915) (translated from Russian), p. 897-908.

[11] Petrov, G. I. Application of the method of Galerkin to a problem involving the stationary flow of a viscous fluid. *Prikladnaya Matematika i Mekhanika* **4** (1940) (translated from Russian), p. 3-12.

[12] Hrenikoff, A. Solution of problems of elasticity by the frame-work method. *ASME Journal of Applied Mechanics* **8** (1941), p. A619-A715.

[13] Hurwitz, A., Courant, R. *Gemeinsame Funktionentheorie*. In: Hurwitz, A., et al. *Vorlesungen über allgemeine Funktionentheorie und elliptische Funktionen*. Berlin: Springer, 1922.

[14] Courant, R. Variational methods for the solution of problems of equilibrium and vibrations. *Bulletin of the American Mathematical Society* **49** (1943), p. 1-23.

[15] Turner, M. J., Clough, R. W., Matrin H. C., Topp, L. J. Stifffness and deflection analysis of complex structures. *Journal of the Aeronautical Sciences* **23** (1956), p. 805-823.

[16] Clough, R. W. The finite element method in plane stress analysis. *Proceedings of the 2nd ASCE Conference on Electronic Computation* in Pittsburgh (1960), p. 345-378. Pittsburgh: ASCE, 1960,

[17] Zienkiewicz, O. C. *The Finite Element Method in Structural and Continuum Mechanics*. New York: McGraw-Hill, 1967.

[18] Friedrichs, K. O. *A finite difference scheme for the Neumann and the Dirichlet Problem*. Washington: US Depatrment of Energy (technical report), 1962.

[19] Cèa, J. Approximation variationnelle des problèmes aux limites. *Annales de l'Institut Fourier* **14** (1964), p. 345-444.
[20] Gusman, J. A., Oganesyan, L. A. Inequalities for the convergence of finite difference schemes for degenerate elliptic equations. *Zhurnal Vychislitelnoĭ Matematiki i Matematicheskoĭ Fiziki* **5** (1964) (translated from Russian), p. 351-357.
[21] Kang, F. Difference scheme based on variational principle. *Journal of Applied and Computational Mathematics* **2** (1965), p. 238-262.
[22] Ciarlet, P. G., Schultz, M. H., Varga, R. S. Numerical methods of high-order accuracy for nonlinear boundary value problems. *Numeriche Mathematik* **9** (1967), p. 394-430.
[23] Aubin, J.-P. Approximation des espaces de distributions et des opérateurs différentiels. *Bulletin de la Société Mathématique de France* **12** (1967), p. 1-139.
[24] Birkhoff, G., Schultz, M. H., Varga, R. S. Piecewise Hermite interpolation in one and two variables with applications to partial differential equations. *Numerische Mathematik* **11** (1968) p. 232-256.
[25] Zlámal, M. On the finite element method. *Numerische Mathematik* **12** (1968), p. 394-409.
[26] Kolář, V., Kratochvíl, J., Leitner, F., Ženíšek, A. *Výpočet plošných a prostorových konstrukcí metodou konečných prvků*. Prague: SNTL, 1979.
[27] Melkes, F. Reduced piecewise bivariate Hermite interpolations. *Numerische Mathematik* **19** (1972), p. 320-340.
[28] Rektorys, K. *Variační metody v inženýských problémech a v problémech matematické fyziky*. Prague: SNTL, 1974.
[29] Rektorys, K. *Metoda časové diskretizace a parciální diferenciální rovnice.* Prague: SNTL, 1985.
[30] Jamet, P. Error estimates for nearly degenerate finite elements. *RAIRO Analyse Numérique* **10** (1976), p. 43-61.
[31] Křížek, M. On semiregular families of triangulations and linear interpolation. *Applications of Mathematics* **36** (1991), p. 223-232.
[32] Szabó, B. A., Mehta, A. K. p-convergent finite element approximations in fracture mechanics. *International Journal for Numerical Methods in Engineering* **12** (1978), p. 551-560.
[33] Babuška, I., Suri, M. The hp-version of the finite element method with quasiuniform meshes. *RAIRO Modélisation mathématique et analyse numérique* **21** (1987), p. 199-238.
[34] Chen, X., Yang, S., Ma, J., He, Z. The construction of wavelet finite element and its application. *Finite Elements in Analysis and Design* **40** (2004), p. 541-554.
[35] Melenk, J. M., Babuka, I. The partition of unity finite element method: basic theory and applications. *Computer Methods in Applied Mechanics and Engineering* **139** (1996), p. 289-314.
[36] Strouboulis, T., Copps, K., Babuška, I. The generalized finite element method. *Computer Methods in Applied Mechanics and Engineering* **190** (2001), p. 4108-4113.
[37] Moës, N., Dolbow, J., Belytchko, T. A finite element method for crack growth without remeshing. *International Journal for Numerical Methods in Engineering* **46** (1999), p. 131-150.
[38] Fries, T.-P., Belytchko, T. The intrinsic XFEM: a method for arbitrary discontinuities without additional unknowns. *International Journal for Numerical Methods in Engineering* **68** (2006), p. 1358-1385.

[39] Dostál, Z., Kozubek, T, Vondrák, V., Brzobohatý, T., Markopoulos, A. Scalable TFETI algorithm for the solution of multibody contact problems of elasticity. *International Journal for Numerical Methods in Engineering* **82** (2010), p. 1384-1405.
[40] Dolejší, V., Feistauer, M. *Discontinuous Galerkin Method: Analysis and Applications to Compressible Flow*. Cham: Springer, 2015.
[41] Gavete, L., Gavete, M. L., Benito, J. Improvements of generalized finite difference method and comparison with other meshless method. *Applied Mathematical Modelling* **27** (2003), p. 831-847.
[42] Cai, Z. On the finite volume element method. *Numerische Mathematik* **58** (1990), p. 713-735.
[43] Balaš, J., Sládek, J., Sládek, V. *Stress Analysis by Boundary Element Methods*. Amsterdam: Elsevier, 1990.
[44] Jiang, B.-N., Povinelli, L. Optimal least-squares finite element methods for elliptic problems. *Computer Methods in Applied Mechanics and Engineering* **102** (1993), p. 199-212.
[45] Babuška, I., Tempone, R., Zouraris, G. E. Solving elliptic boundary value problems with uncertain coefficients by the finite element method: the stochastic formulation. *Computer Methods in Applied Mechanics and Engineering* **194** (2005), p. 1251-1294.
[46] Hou, T., Efendiev, Y. *Multiscale Finite Element Methods: Theory and Applications*. New York: Springer, 2009.
[47] Fefferman, Ch. Existence and smoothness of the Navier - Stokes equation. In: *Millenium Prize Problems*, Problem 4. Denver: Clay Mathematics Institute, 2000.
[48] Feireisl, E., Novotný, A. Two phase flows of compressible viscous fluids. *Discrete and Continuous Dynamical Systems* **15** (2022), p. 2215-2232.
[49] Szabó, B., Babuška, I. *Finite Element Analysis: Method, Verification and Validation*. Hoboken: J. Wiley & Sons, 2021.
[50] Havlásek, P., Grassl, P., Jirásek, M. Analysis of size effect of quasi-brittle materials using integral-type nonlocal models. *Engineering Fracture Mechanics* **157** (2016), p. 72-85.
[51] Eringen, A. C. *Theory of Nonlocal Elasticity and Some Applications*. Princeton: Princeton University (technical report), 1984.
[52] Pike, M. G. Oskay, C. XFEM modeling of short microfiber reinforced composites with cohesive interfaces. *Finite Element Analysis and Design* **106** (2015), p. 16-31.
[53] Vala, J., Kozák, V. Computational analysis of quasi-brittle fracture in fibre reinforced cementitious composites. *Theoretical and Applied Fracture Mechanics* **107** (2020), p. 102486/1-8.
[54] Vala, J., Kozák, V. Nonlocal damage modelling of quasi-brittle composites. *Applications of Mathematics* **66** (2021), p. 815-836.
[55] Pijaudier - Cabot, G., Mazars, J. Damage models for concrete. In: *Handbook of Materials Behavior Models* (Lemaitre, J., ed.). Cambridge (Massachusetts): Academic Press, 2001, p. 500-512.
[56] Štekbauer, H., Němec, I., Lang, R., Burkart, D., Vala, J. On a new computational algorithm for impacts of elastic bodies. *Applications of Mathematics* **67** (2022), in print, 28 p.
[57] Rek, V., Vala, J. On a distributed computing platform for contact - impact of elastic bodies. *Proceedings of SNA (Seminar on Numerical Analysis)* 2021 in Ostrava (virtual), p. 64-67. Ostrava: VŠB-TU, 2021.

Acknowledgement

The work presented in this paper has been supported by the project of specific university research at Brno University of Technology No. FAST-S-22-7867.