ANNALES MATHEMATICAE ET INFORMATICAE

VOLUME 60. (2024)

INSTITUTE OF MATHEMATICS AND INFORMATICS
ESZTERHÁZY KÁROLY CATHOLIC UNIVERSITY
HUNGARY, EGER

HU ISSN 1787-6117 (Online)

A kiadásért felelős az
Eszterházy Károly Katolikus Egyetem rektora
Megjelent a Líceum Kiadó gondozásában
Kiadóvezető: Dr. Nagy Andor
Műszaki szerkesztő: Dr. Tómács Tibor
Megjelent: 2024. december

Contents

Research papers

Methodological papers

Annales Mathematicae et Informaticae
60 (2024) pp. 1–9
DOI: 10.33039/ami.2024.11.001
URL: https://ami.uni-eszterhazy.hu

On the group of units of the semisimple group algebras of dimension 200

N. Abhilash[a], **E. Nandakumar**[b*]

[a]Department of Mathematics, Faculty of Engineering and Technology,
SRM Institute of Science & Technology,
Ramapuram-600089, Tamil Nadu, INDIA
abhilash18101994@gmail.com

[b]Department of Mathematics, College of Engineering and Technology,
SRM Institute of Science & Technology,
Kattankulathur-603203, Tamil Nadu, INDIA
nanda1611@gmail.com

Abstract. In this study, we examine the non-metabelian groups of order 200 and analyze the structure of the unit group within their corresponding group algebra. Among the fifty-two non-isomorphic groups of order 200, only two are non-metabelian. The paper focuses on characterizing the unit group of the semi-simple group algebras associated with these two groups over any finite field.

Keywords: finite field, group algebra, unit group, non-metabelian groups

AMS Subject Classification: Primary: 16U60, Secondary: 20C05

1. Introduction

The linear combination of elements from G with coefficients from the finite field $\mathcal{F}_q$, where G is a finite group, is the group algebra, denoted by $\mathcal{F}_qG$ over the field $\mathcal{F}_q$ with $q = p^k$ elements for a prime p and $k \in \mathbb{Z}^+$. Maschke's theorem [19] implies that the group algebra $\mathcal{F}_qG$ is semisimple if and only if $\mathrm{char}(\mathcal{F}_q) \nmid n$. Thus, via Wedderburn decomposition theorem [19], $\mathcal{F}_qG$ is isomorphic to the direct sum of matrix algebras over division rings, that is,

$$\mathcal{F}_qG \cong M_{n_1}(D_1) \oplus \cdots \oplus M_{n_l}(D_l),\ n_i, l \in \mathbb{Z}^+.$$

*Corresponding author

Submitted: February 23, 2024
Accepted: November 20, 2024
Published online: November 23, 2024

One may easily determine the structure of the unit group of $\mathcal{F}_qG$ from the aforementioned isomorphism. Recall that the unit group, $\mathbb{U}(\mathcal{F}_qG)$, is set of all of the invertible elements in $\mathcal{F}_qG$. See [1, 6, 22, 24, 26, 30–32] for a few noteworthy recent studies that have been conducted to determine the structure of the unit group of the semi-simple group algebra. The applications of units in number theory [14], coding theory [16], cryptography [20], etc. make the research in this area crucial. In addition to this, the recent counterexample to the renowned Kaplansky's unit conjecture further emphasizes the need of research in this area (see [15]). Furthermore, there have been significant developments in the exploration of the unit group of modular group algebras, in addition to integral and semisimple group algebras (see [9–11, 28] and the references therein for a comprehensive and recent literature in this direction).

In view of studying the unit group of all group algebras, the groups can be categorized into two divisions: metabelian and non-metabelian. Bakshi et al. conducted a thorough analysis of the first instance in [8]. As a result, we just need to consider the non-metabelian groups. In [27], Pazderski determined possible non-metabelian group orders. The smallest non-metabelian groups have order 24, and Khan et al. [17] and Maheshwari et al. [18] have examined the unit group of the corresponding group algebras. These findings may be readily ascertained with the use of [27]. Similarly, in [7, 21, 23, 24, 26, 29], Mittal et al. and Arvind et al. categorised the unit group of group algebras of non-metabelian groups up to order 120. Mittal et al. [25] and Abhilash et al. [2–5] recently finished the work for all groups up to order 180.

It is known that there exist non-metabelian groups of order 200 as a result of [27]. There exist 52 non-isomorphic groups of order 200, two of which are non-metabelian. The objective of this paper is to take these 2 groups into consideration and use the Wedderburn decomposition to derive the unit groups of their group algebras (see [19]).

The flow of this paper is as follows. The important definitions, results and the 2 non-metablian groups to be studied in this paper are introduced in Section 2 and Section 3, respectively. Moreover, Section 3 has the main results on the unit groups of the semisimple group algebras. The final section concludes the paper in nature.

2. Preliminaries

The results and prerequisites definitions needed to support the main result are provided in this section. The following notations apply to this entire paper.

$\mathcal{F}_q$	finite field of order $q = p^k$ with characteristic p and $k \geq 1$
G	finite group of order n with $p \nmid n$
e	exponent of the group G i.e., the l.c.m of the order of elements in G
ω	primitive e-th root of unity over $\mathcal{F}_q$
$\mathbb{G}$	Galois group of $\mathcal{F}_q(\omega)$ over $\mathcal{F}_q$, where $\mathcal{F}_q(\omega)$ is the splitting field of $\mathcal{F}_q$

$\mathcal{T}_{G,\mathcal{F}_q}$	collection of all s such that $\sigma(\omega) = \omega^s$, where $\sigma \in \mathbb{G}$
C_x	conjugacy class of x
$[x, y]$	denote the commutator $x^{-1}y^{-1}xy$ of $x, y \in G$
1	identity element of G

Definition 2.1 ([13]). *(i) For any prime p, an element $x \in G$ is said to be p'-element if order of x is not divisible by p.*
(ii) For any p'-element $x \in G$, the cyclotomic $\mathcal{F}_q$-class of $\gamma_x = \sum_{h \in C_x} h$ is the set $S_{\mathcal{F}_q}(\gamma_x) = \{\gamma_{x^s} \mid s \in \mathcal{T}_{G,\mathcal{F}_q}\}$.

Lemma 2.3 deals with the number of elements in a specific cyclotomic class, while the proposition following concerns the total count of cyclotomic $\mathcal{F}_q$-classes. Ferraz provides these two results in [13].

Proposition 2.2. *The set of simple components of $\mathcal{F}_qG/J(\mathcal{F}_qG)$ and the set of cyclotomic $\mathcal{F}_q$-classes in G, where $J(\mathcal{F}_qG)$ is the Jacobson radical of $\mathcal{F}_qG$, are in* 1-1 *correspondence.*

Lemma 2.3. *Let l be the number of cyclotomic $\mathcal{F}_q$-classes in G. If $\mathcal{F}_{q^{m_1}}, \mathcal{F}_{q^{m_2}}, \cdots, \mathcal{F}_{q^{m_l}}$ are the simple components of $Z\big(\mathcal{F}_qG/J(\mathcal{F}_qG)\big)$ and $S_1, S_2, \cdots, S_l$ are the cyclotomic $\mathcal{F}_q$-classes of G, then $|S_i| = [\mathcal{F}_{q^{m_i}} : \mathcal{F}_q]$ with a suitable ordering of the indices, assuming that $\mathbb{G}$ is cyclic.*

In order to uniquely characterize the Wedderburn decompositions of the semisimple group algebras, we need the following important result. See [19, Chapter 3] for its proof.

Lemma 2.4. *(i) Let $\mathcal{F}_qG$ be a semi-simple group algebra and let $N \trianglelefteq G$. Then*

$$\mathcal{F}_qG \cong \mathcal{F}_q(G/N) \oplus \Delta(G, N),$$

where $\Delta(G, N)$ is an ideal of $\mathcal{F}_qG$ generated by the set $\{n - 1 : n \in N\}$.
(ii) If $N = G'$ in part (i), then $\mathcal{F}_q(G/G')$ is the sum of all commutative simple components of $\mathcal{F}_qG$ and $\Delta(G, G')$ is the sum of all others.

Further, we discuss a necessary condition for the dimension of the matrix algebra in the Wedderburn decomposition (see [12]).

Lemma 2.5. *Suppose $\oplus_{i=1}^{t} M_{n_i}(\mathcal{F}_{q^{m_i}})$ is the summand of the semisimple group algebra $\mathcal{F}_qG$, where p is the characteristics of $\mathcal{F}_q$. Then p does not divide any of the n_i.*

Again, we discuss two very important results which helps us to find the unique Wedderburn decomposition. For proof of Lemma 2.6, we refer to [34].

Lemma 2.6. *Let p_1 and p_2 be two primes. Let $\mathcal{F}_{q_1}$ be a field with $q_1 = p_1^{k_1}$ elements and let $\mathcal{F}_{q_2}$ be a field with $q_2 = p_2^{k_2}$ elements, where $k_1, k_2 \geq 1$. Let both the group algebras $\mathcal{F}_{q_1}G, \mathcal{F}_{q_2}G$ be semisimple. Suppose that*

$$\mathcal{F}_{q_1}G \cong \oplus_{i=1}^{t} M(n_i, \mathcal{F}_{q_1}),\ n_i \geq 1$$

and $M\left(n, \mathcal{F}_{q_2^r}\right)$ *is a Wedderburn component of the group algebra* $\mathcal{F}_{q_2}G$ *for some* $r \geq 2$ *and any positive integer* n*, i.e.,*

$$\mathcal{F}_{q_2}G \cong \oplus_{i=1}^{s-1} M(m_i, \mathcal{F}_{q_{2,i}}) \oplus M(n, \mathcal{F}_{q_2^r}),\ m_i \geq 1.$$

Here $\mathcal{F}_{q_{2,i}}$ *is a field extension of* $\mathcal{F}_{q_2}$*. Then* $M(n, \mathcal{F}_{q_1})$ *must be a Wedderburn component of the group algebra* $\mathcal{F}_{q_1}G$ *and it appears atleast* r *times in the Wedderburn decomposition of* $\mathcal{F}_{q_1}G$*.*

Lemma 2.7 ([7, corollary **3.8**])**.** *Let* $\mathcal{F}_qG$ *be a finite semisimple group algebra. Then if there exists an irreducible representations of degree* n *over* $\mathcal{F}_q$*, then one of the Wedderburn components of* $\mathcal{F}_qG$ *is* $M_n(\mathcal{F}_q)$*. Also, if there exists* k *irreducible representations of degree* n *over* $\mathcal{F}_q$*, then* $M_n(\mathcal{F}_q)^k$ *is a summand of the group algebra* $\mathcal{F}_qG$*.*

Throughout this paper, if G has j conjugacy classes, then the representatives of the conjugacy classes are denoted by $g_1(=1), g_2, g_3, \ldots, g_j$.

3. Unit groups

The structure of the unit group of the group algebras of non-metabelian groups of order 200 is covered in this section. Two of the 52 non-isomorphic groups of order 200 are not metabelian. They are as follows:

1. $G_1 := (C_5 \times C_5) \rtimes D_8$
2. $G_2 := (C_5 \times C_5) \rtimes Q_8$

The conjugacy classes and the commutator subgroups of both the groups are taken from the GAP software (see [33]).

3.1. $G_1 := (C_5 \times C_5) \rtimes D_8$

The group G_1 has the following presentation:

$$G_1 = \langle x_1, x_2, x_3, x_4, x_5 \mid x_1^2, [x_2, x_1]x_3^{-1}, [x_3, x_1], [x_4, x_1]x_5^{-3}x_4^{-4}, [x_5, x_1]x_5^{-4}x_4^{-2}, x_2^2,$$
$$[x_3, x_2], [x_4, x_2]x_5^{-1}x_4^{-4}, [x_5, x_2]x_5^{-4}x_4^{-1}, x_3^2, [x_4, x_3]x_4^{-3}, [x_5, x_3]x_5^{-3}, x_4^5, [x_5, x_4], x_5^5\rangle.$$

The sizes (S), orders (O) and the representatives (R) of the 14 conjugacy classes of G_1 are:

R	1	x_1	x_2	x_3	x_4	x_1x_2	x_1x_4	x_1x_5	x_2x_4	x_4x_5	$x_2x_3x_4$	$x_4^2x_5$	$x_4^3x_5$	$x_4^2x_5^2$
S	1	10	10	25	8	50	20	20	20	4	20	4	4	4
O	1	2	2	2	5	4	10	10	10	5	10	5	5	5

It is clear that the exponent of G_1 is 20.

Theorem 3.1. *The unit group of the group algebra $\mathcal{F}_qG_1$ is as follows:*

1) *for $q \equiv \{1,9,11,19\}$ mod 20, $\mathbb{U}(\mathcal{F}_qG_1) \cong \mathcal{F}_q^{*4} \oplus GL_2(\mathcal{F}_q) \oplus GL_4(\mathcal{F}_q)^8 \oplus GL_8(\mathcal{F}_q)$.*

2) *for $q \equiv \{3,7,13,17\}$ mod 20, $\mathbb{U}(\mathcal{F}_qG_1) \cong \mathcal{F}_q^{*4} \oplus GL_2(\mathcal{F}_q) \oplus GL_8(\mathcal{F}_q) \oplus GL_4(\mathcal{F}_{q^2})^4$.*

Proof. The group algebra $\mathcal{F}_qG_1$ is Artinian and semisimple. We observe that the commutator subgroup $G_1' \cong (C_5 \times C_5) \rtimes C_2$ and $\frac{G_1}{G_1'} \cong C_2 \times C_2$. Since $q = p^k$ and $p \neq 2,5$, we split the proof in the following 2 cases.
Case 1: $q \equiv \{1,9,11,19\}$ mod 20. In this case, the cardinality of every cyclotomic $\mathcal{F}_q$-class is 1. Therefore, the decomposition of the group algebra by using Proposition 2.2 and Lemma 2.3 is $\mathcal{F}_qG_1 \cong \mathcal{F}_q \oplus_{i=1}^{13} M_{n_i}(\mathcal{F}_q), n_i \geq 1$. By applying (ii) of Lemma 2.4, we further deduce that

$$\mathcal{F}_qG_1 \cong \mathcal{F}_q^4 \oplus_{i=1}^{10} M_{n_i}(\mathcal{F}_q),\ n_i \geq 2.$$

Since the dimensions of both the sides are the same, we end up with $196 = \sum_{i=1}^{10} n_i^2$. This equation has 34 different solutions. By incorporating Lemma 2.5, we conclude that p can not be 3 and 7. This means that we are remaining with 5 choices of n_i's given as follows:

$$(2^8, 8, 10), (2^5, 4^3, 8^2), (2^4, 4, 5^4, 8), (2, 4^8, 8), (4^6, 5^3).$$

We observe that the subgroup $N := C_5 \times C_5$ is normal in G_1 and $F = G_1/N \cong D_8$. Using [8], we note that $\mathcal{F}_qF \cong \mathcal{F}_q^4 \oplus M_2(\mathcal{F}_q)$. Due to (i) of Lemma 2.4, we are remaining with $(2^8, 8, 10), (2^5, 4^3, 8^2), (2^4, 4, 5^4, 8), (2, 4^8, 8)$ choices of n_i's. Next, we define four group homomorphism f_1, f_2, f_3, f_4 from $G_1 \to GL_4(\mathcal{F}_{11})$, where $\mathcal{F}_{11}$ is the finite field having 11 elements, as follows:

$$x_1 \overset{f_1}{\longmapsto} \begin{pmatrix} 0 & 0 & 1 & 0 \\ 0 & 0 & 0 & 1 \\ 1 & 0 & 0 & 0 \\ 0 & 1 & 0 & 0 \end{pmatrix}, \quad x_2 \overset{f_1}{\longmapsto} \begin{pmatrix} 1 & 0 & 0 & 0 \\ 0 & 1 & 0 & 0 \\ 0 & 0 & 0 & 1 \\ 0 & 0 & 1 & 0 \end{pmatrix}, \quad x_3 \overset{f_1}{\longmapsto} \begin{pmatrix} 0 & 1 & 0 & 0 \\ 1 & 0 & 0 & 0 \\ 0 & 0 & 0 & 1 \\ 0 & 0 & 1 & 0 \end{pmatrix},$$

$$x_4 \overset{f_1}{\longmapsto} \begin{pmatrix} 4 & 0 & 0 & 0 \\ 0 & 3 & 0 & 0 \\ 0 & 0 & 9 & 0 \\ 0 & 0 & 0 & 5 \end{pmatrix}, \quad x_5 \overset{f_1}{\longmapsto} \begin{pmatrix} 4 & 0 & 0 & 0 \\ 0 & 3 & 0 & 0 \\ 0 & 0 & 5 & 0 \\ 0 & 0 & 0 & 9 \end{pmatrix},$$

$$x_1 \overset{f_2}{\longmapsto} \begin{pmatrix} 0 & 0 & 1 & 0 \\ 0 & 0 & 0 & 1 \\ 1 & 0 & 0 & 0 \\ 0 & 1 & 0 & 0 \end{pmatrix}, \quad x_2 \overset{f_2}{\longmapsto} \begin{pmatrix} 10 & 0 & 0 & 0 \\ 0 & 10 & 0 & 0 \\ 0 & 0 & 0 & 10 \\ 0 & 0 & 10 & 0 \end{pmatrix}, \quad x_3 \overset{f_2}{\longmapsto} \begin{pmatrix} 0 & 1 & 0 & 0 \\ 1 & 0 & 0 & 0 \\ 0 & 0 & 0 & 1 \\ 0 & 0 & 1 & 0 \end{pmatrix},$$

$$x_4 \overset{f_2}{\longmapsto} \begin{pmatrix} 4 & 0 & 0 & 0 \\ 0 & 3 & 0 & 0 \\ 0 & 0 & 9 & 0 \\ 0 & 0 & 0 & 5 \end{pmatrix}, \quad x_5 \overset{f_2}{\longmapsto} \begin{pmatrix} 4 & 0 & 0 & 0 \\ 0 & 3 & 0 & 0 \\ 0 & 0 & 5 & 0 \\ 0 & 0 & 0 & 9 \end{pmatrix},$$

$$x_1 \overset{f_3}{\longmapsto} \begin{pmatrix} 0 & 1 & 0 & 0 \\ 1 & 0 & 0 & 0 \\ 0 & 0 & 1 & 0 \\ 0 & 0 & 0 & 1 \end{pmatrix}, \quad x_2 \overset{f_3}{\longmapsto} \begin{pmatrix} 0 & 0 & 1 & 0 \\ 0 & 0 & 0 & 1 \\ 1 & 0 & 0 & 0 \\ 0 & 1 & 0 & 0 \end{pmatrix}, \quad x_3 \overset{f_3}{\longmapsto} \begin{pmatrix} 0 & 1 & 0 & 0 \\ 1 & 0 & 0 & 0 \\ 0 & 0 & 0 & 1 \\ 0 & 0 & 1 & 0 \end{pmatrix},$$

$$x_4 \overset{f_3}{\longmapsto} \begin{pmatrix} 5 & 0 & 0 & 0 \\ 0 & 9 & 0 & 0 \\ 0 & 0 & 4 & 0 \\ 0 & 0 & 0 & 3 \end{pmatrix}, \quad x_5 \overset{f_3}{\longmapsto} \begin{pmatrix} 4 & 0 & 0 & 0 \\ 0 & 3 & 0 & 0 \\ 0 & 0 & 5 & 0 \\ 0 & 0 & 0 & 9 \end{pmatrix},$$

$$x_1 \overset{f_4}{\longmapsto} \begin{pmatrix} 0 & 10 & 0 & 0 \\ 10 & 0 & 0 & 0 \\ 0 & 0 & 10 & 0 \\ 0 & 0 & 0 & 10 \end{pmatrix}, \quad x_2 \overset{f_4}{\longmapsto} \begin{pmatrix} 0 & 0 & 1 & 0 \\ 0 & 0 & 0 & 1 \\ 1 & 0 & 0 & 0 \\ 0 & 1 & 0 & 0 \end{pmatrix}, \quad x_3 \overset{f_4}{\longmapsto} \begin{pmatrix} 0 & 1 & 0 & 0 \\ 1 & 0 & 0 & 0 \\ 0 & 0 & 0 & 1 \\ 0 & 0 & 1 & 0 \end{pmatrix},$$

$$x_4 \overset{f_4}{\longmapsto} \begin{pmatrix} 5 & 0 & 0 & 0 \\ 0 & 9 & 0 & 0 \\ 0 & 0 & 4 & 0 \\ 0 & 0 & 0 & 3 \end{pmatrix}, \quad x_5 \overset{f_4}{\longmapsto} \begin{pmatrix} 4 & 0 & 0 & 0 \\ 0 & 3 & 0 & 0 \\ 0 & 0 & 5 & 0 \\ 0 & 0 & 0 & 9 \end{pmatrix}.$$

Clearly, the maps f_i's are irreducible representations of G_1 of degree 4 over $\mathcal{F}_{11}$, i.e., f_i's are group homomorphisms from G_1 to $GL_4(\mathcal{F}_{11})$ and f_i's are irreducible, that is there is no matrix $U \in GL_4(\mathcal{F}_{11})$ such that

$$U^{-1} f_i(g) U = \begin{pmatrix} A(g) & B(g) \\ 0 & C(g) \end{pmatrix} \quad \text{for all } g \in G_1 \text{ and } i = 1, 2, 3, 4$$

where $A(g)$, $B(g)$, and $C(g)$ are square matirces with entries from $\mathcal{F}_{11}$ depending on g. Therefore, Lemma 2.7 implies that $M_4(\mathcal{F}_{11})^4$ must be a summand of $\mathcal{F}_{11}G_1$. This confirms that $(2, 4^8, 8)$ is the final choice of the values of n_i's. Therefore, we have

$$\mathcal{F}_q G_1 \cong \mathcal{F}_q^4 \oplus M_2(\mathcal{F}_q) \oplus M_4(\mathcal{F}_q)^8 \oplus M_8(\mathcal{F}_q).$$

Case 2: $q \equiv \{3, 7, 13, 17\} \bmod 20$. In this case, the cyclotomic $\mathcal{F}_q$-class corresponding to g_7 includes g_8, g_9 includes g_{11}, g_{10} includes g_{14} and g_{12} includes g_{13}, while rest of the g_i's forms individual classes. Therefore, as per Proposition 2.2, Lemma 2.3 and (ii) of Lemma 2.4, the decomposition for this scenario is given by

$$\mathcal{F}_q G_1 \cong \mathcal{F}_q^4 \oplus_{i=1}^{2} M_{n_i}(\mathcal{F}_q) \oplus_{i=3}^{6} M_{n_i}(\mathcal{F}_{q^2}), \quad n_i > 1$$

which means $196 = n_1^2 + n_2^2 + 2\sum_{i=3}^{6} n_i^2$. There are 24 possibilities and by Lemma 2.6, $M_8(\mathcal{F}_q)$ must be a summand. Also, [8] implies that $\mathcal{F}_q F \cong \mathcal{F}_q^4 \oplus M_2(\mathcal{F}_q)$. Thus, (i) of Lemma 2.4 derives that

$$\mathcal{F}_q G_1 \cong \mathcal{F}_q^4 \oplus M_2(\mathcal{F}_q) \oplus M_8(\mathcal{F}_q) \oplus M_4(\mathcal{F}_{q^2})^4.$$

This completes the proof. □

3.2. $G_2 := (C_5 \times C_5) \rtimes Q_8$

The group G_2 has the following presentation:

$$G_2 = \langle x_1, x_2, x_3, x_4, x_5 \mid x_1^2x_3^{-1}, [x_2, x_1]x_3^{-1}, [x_3, x_1], [x_4, x_1]x_4^{-1}, [x_5, x_1]x_5^{-2}, x_2^2x_2^{-1}, [x_3, x_2], [x_4, x_2]x_5^{-1}x_4^{-4}, [x_5, x_2]x_5^{-4}x_4^{-4}, x_3^2, [x_4, x_3]x_4^{-3}, [x_5, x_3]x_5^{-3}, x_4^5, [x_5, x_4], x_5^5\rangle.$$

The sizes, orders and the representatives of the 8 conjugacy classes of G_2 are given below:

Representatives	1	x_1	x_2	x_3	x_4	x_1x_2	x_4x_5	$x_4^2x_5$
Size	1	50	50	25	8	50	8	8
Order	1	4	4	2	5	4	5	5

It is clear that the exponent of G_2 is 20.

Theorem 3.2. *The unit group of the group algebra $\mathcal{F}_qG_2$ for any field with characteristic not equal to 2 and 5 is given by:*

$$\mathbb{U}(\mathcal{F}_qG_2) \cong \mathcal{F}_q^{*4} \oplus GL_2(\mathcal{F}_q) \oplus GL_8(\mathcal{F}_q)^3.$$

Proof. The group algebra $\mathcal{F}_qG_2$ is Artinian and semisimple. We observe that the commutator subgroup $G_2' \cong (C_5 \times C_5) \rtimes C_2$ and $\frac{G_2}{G_2'} \cong C_2 \times C_2$. $T_{G,\mathcal{F}_q} = \{1, 3, 7, 9, 11, 13, 17, 19\}$. So, for any $g \in G_1$, we have

$$S_{\mathcal{F}_q}(\gamma_g) = \{\gamma_{g^t} \mid t \in T_{G,\mathcal{F}_q}\} \Rightarrow S_{\mathcal{F}_q}(\gamma_g) = \{\gamma_g, \gamma_{g^3}, \gamma_{g^7}, \gamma_{g^9}, \gamma_{g^{11}}, \gamma_{g^{13}}, \gamma_{g^{17}}, \gamma_{g^{19}}\}.$$

Further, it can be verified that g^t, where $t \in T_{G,\mathcal{F}_q}$ belong to the conjugacy class of g. Consequently, $S_{\mathcal{F}_q}(\gamma_g) = \{\gamma_g\}$ for any $g \in G_1$. Therefore, we conclude that the cardinality of every cyclotomic $\mathcal{F}_q$-class is 1. Therefore, as per Proposition 2.2, Lemma 2.3 and (ii) of Lemma 2.4, the decomposition is

$$\mathcal{F}_qG_2 \cong \mathcal{F}_q^4 \oplus_{i=1}^4 M_{n_i}(\mathcal{F}_q),$$

and $196 = \sum_{i=1}^4 n_i^2$, where $n_i > 1$. The possible choices of n_i's fulfilling above equation are $(2, 8, 8, 8), (3, 3, 3, 13), (3, 5, 9, 9), (4, 4, 8, 10), (5, 5, 5, 11), (7, 7, 7, 7)$. We observe that the subgroup $N := \langle x_4, x_5\rangle$ is normal in G_2 and $F = G_2/N \cong Q_8$. Using [8], we recall that $\mathcal{F}_qF \cong \mathcal{F}_q^4 \oplus M_2(\mathcal{F}_q)$. Therefore, $M_2(\mathcal{F}_q)$ must be the Wedderburn components of $\mathcal{F}_qG_2$ as per (i) of Lemma 2.4. So, $(2, 8, 8, 8)$ is the required choice, which means that $\mathcal{F}_qG_2 \cong \mathcal{F}_q^4 \oplus M_2(\mathcal{F}_q) \oplus M_8(\mathcal{F}_q)^3$. This completes the proof. □

4. Conclusion

This work concludes the study of the unit groups of semisimple group algebras of all groups up to order 200, with the exception of the groups of order 192. We have

previously investigated the unit group of the semisimple group algebras of non-metabelian groups of order 200. It is evident that in order to characterise the unit groups in a way that is distinct from one another, different strategies are needed as the group increases in size. Researchers are encouraged by this study to develop an algorithm that can calculate the semisimple group algebra of any finite group's Wedderburn decomposition.

References

[1] N. Abhilash, E. Nandakumar: *A brief about the units of Heisenberg group algebra of higher dimensions*, Asia Pac. J. Math. 10 (2023), p. 23, doi: 10.28924/APJM/10-23.

[2] N. Abhilash, E. Nandakumar, G. Mittal, R. K. Sharma: *Unit groups of semisimple group algebras of groups up to order 180*, Journal of Interdisciplinary Mathematics 27.6 (2024), pp. 1383–1404, doi: 10.47974/JIM-1858.

[3] N. Abhilash, E. Nandakumar, G. Mittal, R. K. Sharma: *Units of the semisimple group algebras of groups of order 162*, Math Vensik Online first.August 20 (2024), doi: 10.57016/MV-0dhsTT7h.

[4] N. Abhilash, E. Nandakumar, R. K. Sharma, G. Mittal: *On the unit group of the semisimple group algebras of groups up to order* 144, Palestine Journal of Mathematics 13.2 (2024), pp. 160–171.

[5] N. Abhilash, E. Nandakumar, R. K. Sharma, G. Mittal: *Structure of the unit group of the group algebras of non-metabelian groups of order* 128, Math Bohemica Online first.May 6 (2024), doi: 10.21136/MB.2024.0017-23.

[6] S. Ansari, M. Sahai: *Units in $F(C_n \times Q_{12})$ and $F(C_n \times D_{12})$*, Int. Elect. J. Algebra 34 (2023), pp. 182–196.

[7] N. Arvind, S. Panja: *Unit group of some finite semisimple group algebras*, Journal of Egyptian Mathematical Society 3 (2022), p. 17, url: https://joems.springeropen.com/articles/10.1186/s42787-022-00151-0.

[8] G. K. Bakshi, S. Gupta, I. B. S. Passi: *The algebraic structure of finite metabelian group algebras*, Communications in Algebra 43.1 (2015), pp. 2240–2257, doi: 10.1080/00927872.2014.888566.

[9] Z. A. Balogh: *The structure of the unit group of some group algebras*, Miskolc Math. Notes 21.2 (2020), pp. 615–620.

[10] Z. A. Balogh: *Unitary units of the group algebra of modular groups*, Journal of Algebra and Its Applications 21.02, 2250027 (2022), doi: 10.1142/S021949882250027X.

[11] A. Bovdi: *The group of units of a group algebra of characteristic p*, Publ. Math. Debrecen 52.1-2 (1998), pp. 193–244.

[12] C. Dietzel, G. Mittal: *Summands of finite group algebras*, Czechoslovak Mathematical Journal 71.4 (2021), pp. 1011–1014, doi: 10.21136/CMJ.2020.0171-20.

[13] R. A. Ferraz: *Simple components of the center of FG/J(FG)*, Communications in Algebra 36.9 (2008), pp. 3191–3199, doi: 10.1080/00927870802103503.

[14] W. Gao, A. Geroldinger, F. H.-Koch: *Group algebras of finite abelian groups and their applications to combinatorial problems*, Rocky Mountain J. Math. 39.3 (2009), pp. 805–823, doi: 10.1216/RMJ-2009-39-3-805.

[15] G. Gardam: *A counterexample to the unit conjecture for group rings*, Ann. of Math. 194.3 (2021), pp. 967–979.

[16] T. Hurley: *Convolutional codes from units in matrix and group rings*, International Journal of Information and Coding Theory 50.3 (2009), pp. 431–463, URL: https://www.ijpam.eu/contents/2009-50-3/9/index.html.

[17] M. Khan, R. K. Sharma, J. B. Srivastava: *The unit group of FS_4*, Acta Mathematica Hungarica 118.1-2 (2008), pp. 105–113, DOI: 10.1007/s10474-007-6169-4.

[18] S. Maheshwari, R. K. Sharma: *The unit group of the group algebra $\mathcal{F}_q SL(2,\mathbb{Z}_3)$*, Journal of Algebra Combinatorics Discrete Structures and Applications 3 (2016), pp. 1–6, URL: https://jacodesmath.com/index.php/jacodesmath/article/view/24/0.

[19] C. P. Milies, S. K. Sehgal: *An Introduction to Group Rings*, Netherlands: Springer Dordrecht, 2002, URL: https://link.springer.com/book/9781402002380.

[20] G. Mittal, S. Kumar, S. Narain, S. Kumar: *Group ring based public key cryptosystems*, J. Disc. Math. Sci. Crypt. 25.6 (2022), pp. 1683–1704.

[21] G. Mittal, R. K. Sharma: *Computation of the Wedderburn decomposition of semisimple group algebras of groups up to order 120*, Annales Mathematicae et Informaticae Preprint (2023), DOI: 10.33039/ami.2023.07.001.

[22] G. Mittal, R. K. Sharma: *Computation of Wedderburn decomposition of groups algebras from their subalgebra*, Bull. Korean Math. Soc. 59.3 (2022), pp. 781–787, DOI: 10.4134/BKMS.b210478.

[23] G. Mittal, R. K. Sharma: *On unit group of finite group algebras of Non-Metabelian groups of Order 108*, Journal of Algebra Combinatorics Discrete Structures and Applications 8.2 (2021), pp. 59–71, URL: https://jacodesmath.com/index.php/jacodesmath/article/view/158.

[24] G. Mittal, R. K. Sharma: *On unit group of finite group algebras of non-metabelian groups upto order 72*, Mathematica Bohemica 146.4 (2021), pp. 429–455, URL: https://articles.math.cas.cz/10.21136/MB.2021.0116-19.

[25] G. Mittal, R. K. Sharma: *The unit groups of semisimple group algebras of some non-metabelian groups of order* 144, Math. Bohemica 148.4 (2023), pp. 631–646, DOI: 10.21136/MB.2022.0067-22.

[26] G. Mittal, R. K. Sharma: *Unit group of semisimple group algebras of some non-metabelian groups of order* 120, Asian-European J. Math. 15.3 (2022), p. 2250059, DOI: 10.1142/S1793557122500590.

[27] G. Pazderski: *The orders to which only belong metabelian groups*, Netherlands: Math. Nachr., 1980.

[28] M. Sahai, S. Ansari: *Group of Units of Finite Group Algebras for Groups of Order 24*, Ukr. Math. J. 75 (2023), pp. 244–261.

[29] R. K. Sharma, G. Mittal: *Unit group of semisimple group algebra $\mathbb{F}_q SL(2,\mathbb{Z}_5)$*, Mathematica Bohemica 147.1 (2022), pp. 1–10, URL: https://articles.math.cas.cz/10.21136/MB.2021.0104-20.

[30] R. K. Sharma, J. B. Srivastava, M. Khan: *The unit group of $\mathbb{F}A_4$*, Publ. Math. Debrecen 71.1-2 (2007), pp. 21–26.

[31] R. K. Sharma, J. B. Srivastava, M. Khan: *The unit group of $\mathbb{F}S_3$*, Acta Math. Acad. Paedagog. Nyházi. 23.2 (2007), pp. 129–142.

[32] G. Tang, Y. Wei, Y. Li: *Units group of group algebras of some small groups*, Czechoslovak Math. J. 64.1 (2014), pp. 149–157.

[33] The GAP software – Groups: *Algorithms and Programming*, version 4.12.2, 2022.

[34] R. Waxman: *On the Wedderburn component of a group algebra* (2019), URL: https://math.stackexchange.com/questions/4677421/on-the-wedderburn-component-of-a-group-algebra.

Annales Mathematicae et Informaticae
60 (2024) pp. 10–18
DOI: 10.33039/ami.2024.02.001
URL: https://ami.uni-eszterhazy.hu

New methods for maximizing the smallest eigenvalue of the grounded Laplacian matrix

Ahmad T. Anaqreh, Boglárka G.-Tóth, Tamás Vinkó

Department of Computational Optimization, Institute of Informatics,
University of Szeged, Hungary
{ahmad,boglarka,tvinko}@inf.u-szeged.hu

Abstract. Maximizing the smallest eigenvalue of the grounded Laplacian matrix is an NP-hard problem that involves identifying the Laplacian matrix's $(n-k) \times (n-k)$ principal submatrix obtained after removing k rows and corresponding columns. The challenge is to determine optimally the rows and columns to be deleted. Our proposed approach, motivated by the Gershgorin circle theorem, is used together with the degree centrality of the corresponding graph. Moreover, integer linear programming for the vertex cover problem has been employed as an additional method of solving the problem. The efficiency of the methods is demonstrated on real-world graphs.

Keywords: Laplacian matrix, grounded Laplacian matrix, eigenvalues

1. Introduction

The simplest way to represent graphs is their topological representation, where the graph is a set of nodes and edges. However, the spectral representation, such as Adjacency matrix or Laplacian matrix, can significantly help in describing the structural and functional behavior of the graph. Let $G = (V, E)$ be a simple, undirected graph with nodes set V and edges set E, where $|V| = n$ and $|E| = m$. For an edge (i, j) we consider that $(j, i) \in E$ for symmetry, but they count only one edge in total. The Adjacency matrix A is defined as

$$A_{ij} = \begin{cases} 1 & \text{if } (i,j) \in E, \\ 0 & \text{otherwise.} \end{cases}$$

Submitted: July 31, 2023
Accepted: February 6, 2024
Published online: February 9, 2024

Let d_i denote the degree of node $i \in V$, i.e., $d_i = \sum_j A_{ij}$ for all $i \in V$. The Laplacian matrix L of graph G is defined as follows:

$$L_{ij} = \begin{cases} d_i & \text{if } i = j, \\ -1 & \text{if } (i,j) \in E, \\ 0 & \text{otherwise,} \end{cases}$$

Note that the Laplacian matrix is a symmetric positive semidefinite matrix. It is well known that its n eigenvalues are non-negative real numbers, bounded by the double of the maximum vertex degree [1]. As a consequence, we have $0 = \lambda_1 \leq \lambda_2 \leq \ldots \leq \lambda_n \leq 2 \max_{i \in V} d_i$, where λ_i stands for the i-th eigenvalue of L.

Some applications of the Laplacian. According to Mohar *et al.* [12], the eigenvalues of the Laplacian matrix have their applications in diverse fields. One of the main applications is in graph theory, where the number of spanning trees of a graph G is determined by the multiplication of all non-zero eigenvalues of L [9]. Moreover, the sum of resistance distances over all node pairs can also be determined using the eigenvalues of the Laplacian matrix [6]. Another important implication of the Laplacian matrix is the Fiedler value [4], which corresponds to the second smallest eigenvalue (λ_2) and plays a crucial role in determining the connectivity of a graph. A graph is considered connected if its Fiedler value is greater than zero. Finally, the number of components in G is equal to the multiplicity of the 0 eigenvalue of L.

Grounded Laplacian. Let $G = (V, E)$, a simple undirected and connected graph, be given together with its Laplacian matrix L. The grounded Laplacian matrix $L(S)$, which was introduced in [11], is an $(n-k) \times (n-k)$ submatrix obtained by deleting k rows and their corresponding columns from the Laplacian matrix L, where $S \subset V$, $|S| = k$, $0 < k \ll n$. The smallest eigenvalue of $L(S)$ is denoted by $\lambda(S)$. Note that $L(S)$ is a symmetric positive definite matrix, thus all its eigenvalues are strictly positive real numbers. Hence, $\lambda(S) > 0$ holds.

Applications, complexity and algorithms. Without completeness, we mention some applications of $L(S)$. The value of the smallest eigenvalue $\lambda(S)$ of matrix $L(S)$ determines the convergence rate of a leader-follower networked dynamical system [13], as well as the effectiveness of pinning scheme of pinning control of complex dynamical networks [10], with large $\lambda(S)$ corresponding to fast convergence speed and good pinning control performance.

Finding $L(S)$ with the maximum possible $\lambda(S)$ has been shown to be an NP-hard problem [14]. Thus, solution methods based on heuristics are desired. The authors in [14] introduced two greedy-type algorithms. The first one, referred as the NAÏVE algorithm, involves k iterations. In each iteration, a candidate is chosen if adding it to set S maximizes $\lambda(S)$. The second algorithm, referred to as the FAST algorithm, evaluates a candidate node based on the sum of the eigenvalues of

its adjacent nodes. The optimal candidate is chosen based on the maximum sum value. The eigenvalues for this computation are obtained from the eigenvector that corresponds to the smallest eigenvalue of a grounded Laplacian matrix, which is approximated using the SDDM solver [3] that determines the eigenvector without having to calculate the entire eigensystem.

2. Methodology

We propose two algorithms that differ from the approaches mentioned above. The first algorithm relies on the centrality of the nodes to select elements of set S, while the second algorithm is based on the vertex cover problem.

First approach. Our first method, denoted by DEGREE-G, is motivated by the well-known Gershgorin circle theorem [5]. The Gershgorin circle theorem provides bounds on the eigenvalues of a square matrix. Let B be a square matrix, with entries b_{ij}. For $i \in \{1, \ldots, n\}$, let $R_i = \sum_{i \neq j} |b_{ij}|$. Let $D(b_{ii}, R_i) \subseteq \mathbb{C}$ be a closed circle centered at b_{ii} with radius R_i.

Theorem. *Every eigenvalue of B lies within at least one of the Gershgorin circles $D(b_{ii}, R_i)$.*

Figure 1 represents the Gershgorin circles of a Laplacian matrix. Note that we obtain the sharp, yet trivial, lower bound 0 on the eigenvalues of L.

$$L = \begin{pmatrix} 3 & -1 & -1 & -1 & 0 \\ -1 & 2 & 0 & -1 & 0 \\ -1 & 0 & 3 & -1 & -1 \\ -1 & -1 & -1 & 4 & -1 \\ 0 & 0 & -1 & -1 & 2 \end{pmatrix}$$

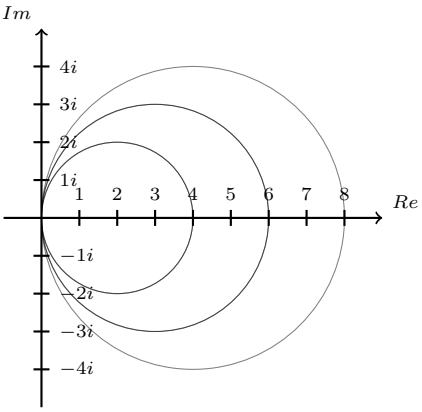


Figure 1. A Laplacian matrix (left) and its Gershgorin circles (right).

Before we give the details of our first approach, the concept of graph centralities needs to be briefly introduced. Given graph $G = (V, E)$, centrality is a function that assigns a non-negative real number to the nodes of G. Thus, upon calculating centrality values for G, it is possible to rank the nodes, which can be thought of as

Algorithm 1: Centrality-based Algorithm

```
node_cen = sort(centrality(V))
for i ∈ 1 → k do
    remove(L, node_cen[i])
compute(min_eigen(L))
```

assigning importance to the nodes. Traditionally, a larger centrality value indicates a higher importance of the node.

It is obvious that maximizing the lower bound on the smallest eigenvalue requires moving the circles further from the origin. Thus, the idea is to rank the nodes according to specific centrality, and then remove the corresponding row and column from the Laplacian matrix. The method is described in Algorithm 1. In this work, the degree centrality has been used.

As a simple demonstration, by applying the idea on the Laplacian matrix in Figure 1 with $k = 2$ we obtained $\lambda(S) = 1.27$, the output represented in Figure 2. In contrast, the greedy-type Naïve algorithm of [14] gives the set $S = \{2, 4\}$ for which we obtain $\lambda(S) = 1.2$. Note that none of these methods were able to obtain the optimal solution, that is $\lambda(S) = 1.47$ with $S = \{1, 5\}$, for this simple problem. Note that we ranked the nodes in ascending order for this method. We tried using the descending ranking as well, but overall it did not yield satisfactory results.

$$L = \begin{pmatrix} 3 & -1 & -1 \\ -1 & 3 & -1 \\ -1 & -1 & 4 \end{pmatrix}$$

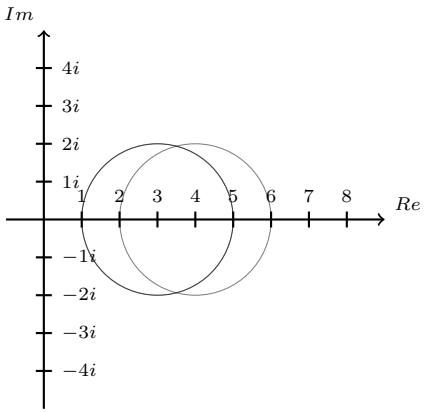


Figure 2. Illustration of the result of Algorithm 1 using degree centrality.

Second approach. The second method called Cover, also uses the Gershgorin circles, but it utilizes the so-called maximum k vertex cover problem as well. The maximum k vertex cover is based on the vertex cover problem [2], with the difference, that in the k vertex cover the search is for a set of k nodes that incident to the maximum number of edges of the graph rather than the minimum number of

nodes that each edge in the graph is incident to. The integer linear programming model of the vertex cover is as follows:

$$\begin{gathered}\min \sum_{i \in V} x_i, \\ x_i + x_j \geq 1 \quad \forall (i,j) \in E, \\ x_i \in \{0,1\} \quad \forall i \in V.\end{gathered}$$

Note that the objective function represents the minimum number of nodes that incident to all the edges in the graph, where $x_i = 1$ are the covering nodes. The constraint requires that for each edge at least one of its endpoints should be a covering node.

The integer linear program of the maximum k vertex cover is defined as follows:

$$\begin{gathered}\max \sum_{j \in V} y_j, \\ \sum_{i \in V} x_i = k, \\ y_j \leq \sum_{\forall i \in V : (j,i) \in E} x_i \quad \forall j \in V, \\ k \in \mathbb{N}, \quad x_i, y_i \in \{0,1\}, \ i = 1, \ldots, n.\end{gathered}$$

Again, the variables x_i stand for the nodes that cover, by the edges, the maximum number of vertices in the graph, while y_i represents the vertices that are covered. The constraints ensure that only k vertices can be selected and that the value of y_i is 1 iff at least one of its adjacent nodes, x_i, is selected. Once we solve the above IP, we delete the rows and columns that correspond to the solution from the Laplacian matrix and then determine its smallest eigenvalue.

The maximum k vertex cover problem can have multiple solutions, so we thought that combining vertex cover and degree centrality could enhance our results. As a result, we modified the objective function in our linear program to the following:

$$\max \sum_{j \in V} y_j - \delta \sum_{j \in V} d_j x_j,$$

where δ is a small number so as to not change the main objective. For instance $\delta = 1/\sum_{j \in V} d_j$ can be chosen. This modification aims to select k nodes with the lowest degree to maximize the objective. The approach is denoted as Cover1.

Moreover, as the maximum k vertex cover problem can have multiple solutions, we utilized a method to explore the possibility of obtaining a better value for $\lambda(S)$ by the alternate solutions. The idea is to solve the linear program iteratively while including a constraint that prevents the solution from resembling previous ones. By checking various solutions, we can obtain different values for $\lambda(S)$ and select the

one that gives the optimal result. This additional constraint is defined as follows for a given solution $S \subset V$ obtained before:

$$\sum_{i \in S} x_i \leq k - 1.$$

At each iteration, we need to ensure that the solution covers the maximum number of nodes, nc, that is, $\sum_{i \in V} y_i = nc$ must hold, otherwise, we need to stop the iteration. This approach is denoted as COVER2. The maximum number of iterations – and so the maximum number of alternative solutions – is fixed to be 100.

3. Numerical results

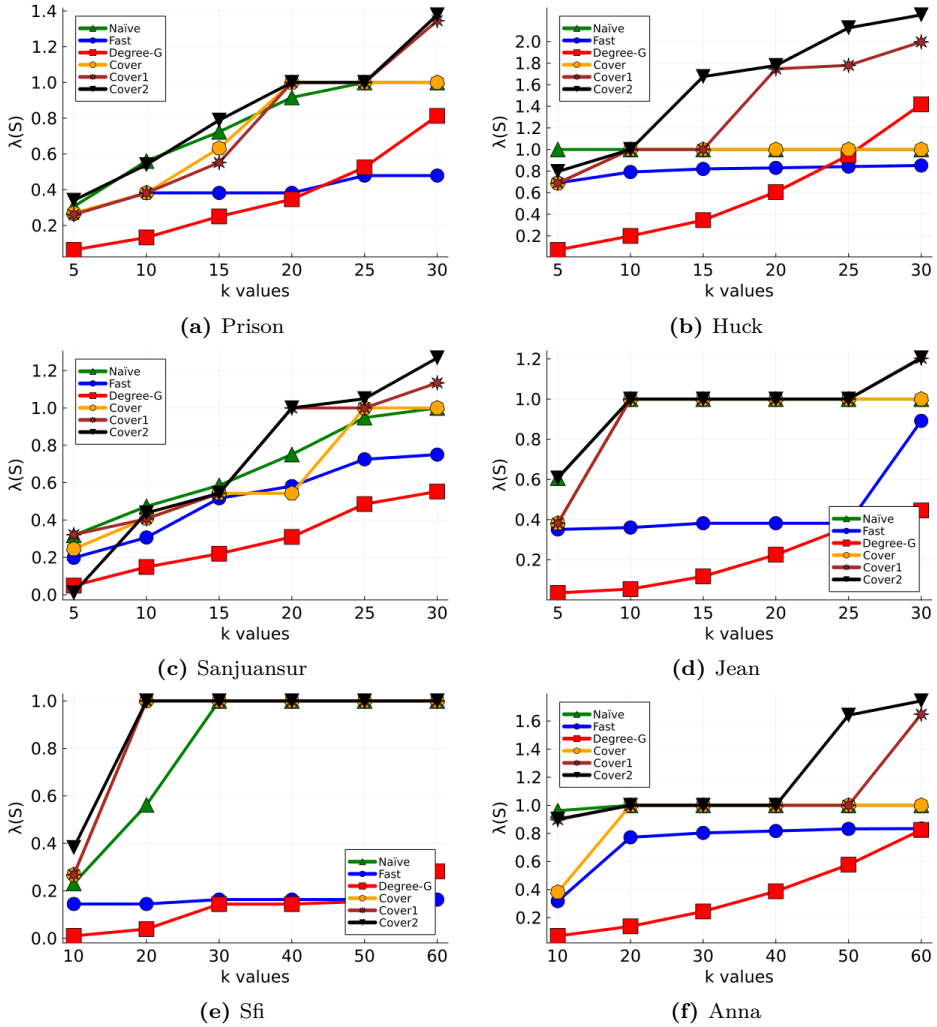
To demonstrate and compare the efficiency of the proposed methods we compare them with the two algorithms proposed by Wang *et al.* [14]. We implemented all the algorithms in Julia 1.7.0 using the package JuMP 0.22.1. As a solver, we used Gurobi 9.5.0., all on a computer with Intel Core i7-4600U CPU and 8GB RAM running Windows 10. Experiments were conducted on real-world networks, all of which are publicly accessible in KONECT [7], SNAP [8], and RWC at `http://tcs.uos.de/research/lip`.

Results for different k values. Figure 3 shows the results achieved by applying the tested methods to various real-world graphs. The figure shows the smallest eigenvalue, $\lambda(S)$, for six different values of k using the discussed methods on selected graphs. The evaluated cases are connected with a solid line, representing an approximate value, since $\lambda(S)$ increases monotonically with k for each method. The higher the lines, the better the results are. The presented results show that method efficiencies vary from graph to graph, but there are some common features. It is obvious that the DEGREE-G and FAST methods are inferior to the NAÏVE and COVER methods. Interestingly, the NAÏVE method outperforms the other methods, although the COVER methods perform equally well or even better than NAÏVE at specific values of k.

Results for specific k values. Instead of using ad-hoc k values, we utilized the vertex cover integer program to obtain the lowest value of k that can provide a sufficient value for certainly increasing the lower bound of $\lambda(S)$. Table 1, displays the corresponding values of k and $\lambda(S)$ for various real-world graphs using the different methods we have discussed. The NAÏVE and COVER methods are clearly more effective than DEGREE-G and FAST methods. However, the COVER methods perform equally or even better than NAÏVE in all cases.

Additionally, we performed a time comparison between the different methods. Table 2 presents the runtime for $k = 5$ of each method with a time limit of one hour. The results demonstrate that the DEGREE-G and FAST methods are notably

faster than the NAÏVE and COVER methods. It is clear that the NAÏVE and COVER algorithms exceeded the time limit for graphs containing thousands of nodes and edges, and it is evident that the COVER methods are generally faster than NAÏVE in completing the task.



(a) Prison **(b)** Huck

(c) Sanjuansur **(d)** Jean

(e) Sfi **(f)** Anna

Figure 3. Values of $\lambda(S)$ obtained by the algorithms for different k values.

Table 1. Values of $\lambda(S)$ for k value obtained from vertex cover.

Graph	N	M	k	Naïve	Fast	Degree-G	Cover	Cover1	Cover2
Prison	67	142	41	1	0.48	1.97	1.59	1.63	2.38
Huck	69	297	44	1	1	1.7	1	2.48	3.5
Sanjuansur	75	144	40	1.59	0.84	0.95	1.29	1.27	1.59
Jean	77	254	42	1	0.37	0.67	1	1.24	1.24
David	87	406	51	1	1	2.45	3.44	2.65	2.77
ieeebus	118	179	61	1	0.59	0.73	1	1.06	1.2
Sfi	118	200	53	1	0.27	0.24	1	1	1
Anna	138	493	58	1	0.83	0.87	1	1.65	1.65
Usair	332	2126	149	1	0.74	1	1	1.59	1.59
494bus	494	586	216	0.38	0.07	0.14	1	1	1
average				0.99	0.62	1.07	1.33	1.56	1.79

Table 2. The time (in seconds) for $k = 5$ to compute $\lambda(S)$.

Graph	N	M	Naïve	Fast	Degree-G	Cover	Cover1	Cover2
Prison	67	142	0.63	0.005	0.002	0.031	0.028	1.139
Huck	69	297	0.785	0.008	0.003	0.036	0.038	5.686
Sanjuansur	75	144	1	0.008	0.003	0.031	0.036	7.712
Jean	77	254	1.02	0.008	0.004	0.035	0.034	13.615
David	87	406	1.418	0.014	0.008	0.041	0.041	2.213
ieeebus	118	179	2.868	0.011	0.003	0.043	0.042	9.127
Sfi	118	200	2.708	0.011	0.008	0.041	0.041	7.724
Anna	138	493	3.88	0.016	0.014	0.067	0.066	14.543
Usair	332	2126	29.773	0.049	0.034	0.651	0.513	21.856
494bus	494	586	102.975	0.057	0.051	0.382	0.323	15.712
Email-Univ	1133	5451	1589.445	0.55	0.531	13.632	17.497	104.342
Routers-RF	2113	6632	>3600	1.448	1.515	55.288	57.114	272.99
US-Grid	4941	6594	>3600	14.255	15.551	312.706	315.643	1947.852
WHOIS	7476	56943	>3600	50.343	50.425	>3600	>3600	>3600
PGP	10680	24340	>3600	140.943	143.747	>3600	>3600	>3600
average			1075.789	13.852	14.126	505.532	506.094	641.634

4. Conclusion

Given the fact that maximizing the smallest eigenvalue of the grounded Laplacian matrix is NP-hard, it is desirable to establish efficient algorithms that can provide solutions of acceptable quality and reasonable running time. We have proposed two approaches and experimentally shown that, compared to the algorithms in the literature, these algorithms are competitive. The covering methods provide a better solution in a shorter time as the Naïve method, while Degree-G is almost as fast as the Fast method giving competitive results, especially for higher k values.

Acknowledgements. The research leading to these results has received funding from the national project TKP2021-NVA-09. Project no. TKP2021-NVA-09 has been implemented with the support provided by the Ministry of Innovation and Technology of Hungary from the National Research, Development, and Innovation Fund, financed under the TKP2021-NVA funding scheme. The work was also supported by the grant SNN-135643 of the National Research, Development, and Innovation Office, Hungary.

References

[1] W. Anderson, T. Morley: *Eigenvalues of the Laplacian of a Graph*, Linear and Multilinear Algebra 18.2 (1985), pp. 141–145, doi: 10.1080/03081088508817681.

[2] J. Chen, I. A. Kanj, W. Jia: *Vertex cover: further observations and further improvements*, Journal of Algorithms 41.2 (2001), pp. 280–301, doi: 10.1006/jagm.2001.1186.

[3] M. B. Cohen, R. Kyng, G. L. Miller, J. W. Pachocki, R. Peng, A. B. Rao, S. C. Xu: *Solving SDD linear systems in nearly m log1/2 n time*, in: Proceedings of the forty-sixth annual ACM symposium on Theory of computing, 2014, pp. 343–352, doi: 10.1145/2591796.2591833.

[4] M. Fiedler: *Algebraic connectivity of graphs*, Czechoslovak mathematical journal 23.2 (1973), pp. 298–305, doi: 10.21136/CMJ.1973.101168.

[5] G. H. Golub, C. F. Van Loan: *Matrix computations*, Johns Hopkins University Press, 2013.

[6] D. J. Klein, M. Randić: *Resistance distance*, Journal of mathematical chemistry 12 (1993), pp. 81–95, doi: 10.1007/BF01164627.

[7] J. Kunegis: *Konect: the koblenz network collection*, in: Proceedings of the 22nd international conference on world wide web, 2013, pp. 1343–1350, doi: 10.1145/2487788.2488173.

[8] J. Leskovec, R. Sosič: *Snap: A general-purpose network analysis and graph-mining library*, ACM Transactions on Intelligent Systems and Technology (TIST) 8.1 (2016), pp. 1–20, doi: 10.1145/2898361.

[9] H. Li, S. Patterson, Y. Yi, Z. Zhang: *Maximizing the number of spanning trees in a connected graph*, IEEE Transactions on Information Theory 66.2 (2019), pp. 1248–1260, doi: 10.1109/TIT.2019.2940263.

[10] H. Liu, X. Xu, J.-A. Lu, G. Chen, Z. Zeng: *Optimizing pinning control of complex dynamical networks based on spectral properties of grounded Laplacian matrices*, IEEE Transactions on Systems, Man, and Cybernetics: Systems 51.2 (2018), pp. 786–796, doi: 10.1109/TSMC.2018.2882620.

[11] U. Miekkala: *Graph properties for splitting with grounded Laplacian matrices*, BIT Numerical Mathematics 33.3 (1993), pp. 485–495, doi: 10.1007/BF01990530.

[12] B. Mohar: *Some applications of Laplace eigenvalues of graphs*, Springer, 1997, doi: 10.1007/978-94-015-8937-6_6.

[13] A. Rahmani, M. Ji, M. Mesbahi, M. Egerstedt: *Controllability of multi-agent systems from a graph-theoretic perspective*, SIAM Journal on Control and Optimization 48.1 (2009), pp. 162–186, doi: 10.1137/060674909.

[14] R. Wang, X. Zhou, W. Li, Z. Zhang: *Maximizing the Smallest Eigenvalue of Grounded Laplacian Matrix*, arXiv preprint arXiv:2110.12576 (2021).

Annales Mathematicae et Informaticae
60 (2024) pp. 19–26
DOI: 10.33039/ami.2024.02.002
URL: https://ami.uni-eszterhazy.hu

Interval based verification of adversarial example free zones for neural networks – Dependency problem

Tibor Csendes

University of Szeged, Institute of Informatics
csendes@inf.szte.hu

Abstract. Recent machine learning models are sensitive to adversarial input perturbation. That is, an attacker may easily mislead an otherwise well-performing image classification system by altering some pixels. It is quite challenging to prove that a network will have correct output when changing slightly some regions of the images. This is why only a few works targeted this problem. Although there are an increasing number of studies on this field, really reliable robustness evaluation is still an open issue. We will present some theoretical results on the dependency problem of interval arithmetic what is critical in interval based verification.

Keywords: verification, artificial neural network, interval arithmetic

AMS Subject Classification: 65G40, 68T07

1. Introduction

Szegedy et al. [7] showed first the phenomenon of adversarial examples. Since then the efforts for verification algorithms were concentrated around optimized models of the trained neural networks [9]. We were able to prove that these verification algorithms are unfortunately not reliable [11]. In the last ICAI conference we reported our first results with an interval based verification algorithm [4]. This approach applied simple natural interval extension for the calculation of inclusion functions, and we were able to produce realistic size adversarial example free zones for simple networks, that had good accuracy for the MNIST picture database distinguishing the hand written figures for the digits 3 and 7.

In the present work we report on our new results. The algorithm was reim-

Submitted: July 19, 2023
Accepted: February 6, 2024
Published online: February 14, 2024

plemented in the Julia language [8]. Our computational test results on the full MNIST database of 10 different hand written digits with a modest but realistic size network (1 hidden layer of 128 neurons) could reach more modest verification results as those reported in [4]. We are also dealing with full-scale machine learning models: an artificial neural network based on the huBERT language model, trained to recognize Hungarian fake news in health-related texts, which contains approximately 110 million parameters. This is obviously out of reach for verification. On the other hand long standing open mathematical problems were solved by interval based computer aided methods, let us just mention our contribution to the solution of the Wright conjecture [2], and proving that the forced damped pendulum is chaotic [1].

The limitations of the naive interval arithmetic approach were felt already in our full MNIST test. It turned out that the critical question is how many activation functions obtain such an input interval that contains zero. If the inputs are real numbers then the probability of having a zero input for a ReLU activation function is in general zero for well trained networks. Table 1 gives the number of test cases from the picture database for which at least one of the ReLU activation functions had an input interval that contained zero. The moral of the results obtained is that the simple application of the algorithm introduced in [3] the application of which for neural network verification was reported in [4] is not sufficient, more sophisticated algorithms should be developed. As a preparatory step in that direction in the present paper a qualitative description is given on the possible amount of overestimation caused by the dependency problem of naive interval arithmetic.

Table 1. The number of critical cases, for which at least one of the ReLU activation functions had an input interval containing zero – as a function of the interval width.

	interval width							
	10^{-1}	10^{-2}	10^{-3}	10^{-4}	10^{-5}	10^{-6}	10^{-7}	10^{-8}
No. critical cases	9178	2389	310	32	4	1	1	0

2. Dependency problem for interval calculations

We consider the fully connected simple feedforward artificial neural network having an input layer, one or more hidden layers and an evaluation output layer. We use the ReLU activation function $\max(0, x)$. This is a fairly general framework, and our results can be transferred in a more or less straightforward way to other types of neural networks such as convolutional and recurrent ones. In our model each neuron number k is defined by the function

$$y_k = \max\left(0, \sum_{i=0}^{n} w_i x_i\right)$$

of its inputs x_i with weights w_i, and $x_0 = 1$. Here x_i stands for one of the n outputs of the earlier layer, or that of the inputs. Each neuron has as inputs all outputs of the previous layer, the first layers input is obviously the input of the network.

Our neural network can be seen as a multidimensional function of the inputs for each output. Without the ReLU activation functions these functions would be simple linear combinations of the input values, in which the product of respective weights give the coefficients of the linear combinations. The ReLU activation functions cause some parts of the computation tree disappear due to their zero branch. The whole neural network forms the surface of a polygon in the high dimensional space of input variables.

Verification of artificial neural networks means that for a given input we can prove that for a neighbourhood of that input, we obtain the same classification for each point inside the tested neighbourhood. In other words, the outputs of the neural network should not change that much which would result in a different class. Hence the range of the function should not change much inside the neighbourhood of the given input. Interval calculations provide a straightforward tool for bounding the range of functions reliably.

2.1. Interval calculation

Consider $\mathbb{I}$ the set of compact real intervals containing intervals $x = [a, b]$, where $a, b, \in \mathbb{R}$, $a \leq b$. The four basic operations can easily be calculated for intervals based on the real calculations on the real end points of the argument intervals:

$$\begin{aligned}
[a, b] + [c, d] &= [a + c, b + d], \\
[a, b] - [c, d] &= [a - d, b - c], \\
[a, b] \cdot [c, d] &= [\min(ac, ad, bc, bd), \max(ac, ad, bc, bd)], \\
[a, b] \,/\, [c, d] &= [a, b] \cdot [1/d, 1/c] \text{ if } 0 \notin [c, d].
\end{aligned}$$

Not only the basic operations, but also standard functions like sin, *log* etc. can easily be generalized for interval arguments. It is important that we have the inclusion property: $f(x) \in F(X)$, where the real number x is in the interval X. This property holds also for careful computer implementations, which use outside rounding to have rigorous bounds on the range $f(X)$ of the real function $f(x)$. The above arithmetic rules ensure sharp bounds on the resulting reals. The numeric effect of the outside rounding is usually negligible. Still, interval arithmetic keeps the inclusion property.

The main difficulty in applying interval arithmetic in the evaluation of trained neural network lies in the so-called dependency problem. In spite of the fact that addition and multiplication gives sharp bound for intervals, the hidden dependencies of input variables pose a substantial problem in terms of overestimation of the bounded ranges [6]. To illustrate it consider the inclusion function $F(X)$ of the function

$$f(x) = x^2 - x.$$

It gives $F([0,1]) = [-1,1]$, while the range of it, $f(X)$ is here just $[-0.25, 0.0]$. The width of the calculated inclusion $w(F(X)$ is eight time as large as the width of the range $w(f(X))$. It is not the last word, using more sophisticated techniques the problem of the too loose enclosure can be overcome – at the cost of higher computing times.

For example affine arithmetic and the interval propagation technique [10] can help. In the present work a description is given on the possible effect of the dependency problem on the overestimation of interval evaluation of trained artificial neural networks.

2.2. Results

In the following, we investigate the overestimation amounts we can face while evaluating trained fully connected feed forward networks with the ReLU activation function. We study the situation when the weights of such a network are given as real numbers, we fix an input (e.g. a picture), and we test how large intervals around the input values can be verified to result in the same classification we obtained for the real case. In the next theoretical investigations we assume that interval arithmetic is calculated in the precise way, i.e. we exclude the effect of outside rounding. Note that only the dependency problem of the addition and subtraction should be considered for artificial networks, because the multiplication or division of input values do not play a role in the calculation of the output values.

Assertion 2.1. *For a fully connected feed forward standard artificial neural network the overestimation size $w(F(X)) - w(f(X))$ of the inclusion function can be zero only if at least one of the following conditions are fulfilled:*

- *all input intervals are of zero width: $w(x_i) = b - a = 0$,*
- *for all input variables x_i in the computation of each of the outputs all weights of them are of the same sign: either all nonnegative, or all nonpositive, and*
- *all the final evaluation functions calculating the outputs of the network have negative arguments.*

These conditions are not only sufficient one by one, but a proper combination of them is also necessary.

Proof. If the input intervals are of zero width, then the possible harmful effect of dependency has no room, since the lower and the upper bounds of all input intervals are equal. In this way the weighted subtraction results in the same value, indifferent of which bounds we choose from the arguments.

The second condition is sufficient, since in this case it is the same which output we calculate, and also for all input variables, the weights have the same sign, and in this way the dependency problem cannot happen for their weighted summation.

If the third condition is met, then all ReLUs providing the output variables must have the value of interval zero ($[0,0]$). It is exactly the range of the network for the input studied.

The obvious effect of a ReLU activation function is, that all possible combinations of the subnetwork that produce its negative argument for the output disappear. In other words, we obtain an equivalent network for the given input, if we neglect the subnetwork producing the negative argument of the given ReLU activation function. If we evaluate our network on input intervals in stead of input real numbers, then the whole argument interval of the given ReLU activation function must be negative to have this feature. The remaining evaluation network will be sensitive for the first two earlier defined conditions, and if none holds, then the dependency problem will necessarily corrupt the result. Other combinations of the three conditions are possible that together ensure that the given neural network on the considered input will not produce positive overestimation size. □

The conditions in Assertion 2.1 are quite strong, and in full-size neural networks capable to solve real life problems, these can hardly be met. Note that the case when only a single hidden layer is in the network is implicitly covered by the second condition, since then each input interval will only be multiplied by a single weight. Single hidden layer fully connected feed forward neural networks seem to be too simple, but they have full recognition capacity with proper activation to get the output – at the cost of a large number of neurons in that single layer [5].

Consider now the question which are the major factors for the overestimation sizes in the same setting.

Assertion 2.2. *For a fully connected feed forward standard artificial neural network of k input intervals, m neurons in each of the even number of n hidden layers, and all weights w_i are bounded by $|w_i| \leq W$ the amount of overestimation $w(F(X)) - w(f(X))$ of the inclusion function of an output is not more than $2^{n/2} m^{n/2} W^n \sum_{i=1}^{k} w(X_i)$.*

Proof. As a consequence of Assertion 2.1 to have positive overestimation we must have positive width input intervals, two weights for their addition with opposing signs, and a path to output values through ReLU activation functions that have arguments which have positive values as well. Since our feed forward network is fully connected, in each neuron we have every interval from the previous layer multiplied by weights. No overestimation can occur in the first hidden layer, since here each input is just multiplied by a weight, but no subtraction of the same variable may be active.

One neuron on the second hidden layer can produce at most $2W^2 w(X_i)$ overestimation, where $w(X_i)$ is the width of the original input interval X_i, and W is the upper bound of the absolute value of the weights in the network. This is the case, when $w_l w_j X_i - w_l w_j X_i$ is in the actual sum of the neuron. The range of this sum is zero, and in this way the overestimation size is $2W^2 w(X_i)$.

Consider now the case when all weights of the network that will affect this overestimation are of the same sign. Then the calculated overestimation changes also by subsequent multiplications by weights as we calculate the output values. The isotonicity property of interval arithmetic implies that this overestimation cannot disappear in the evaluation of the network – with the exception of multiplication by

zero. For this kind of overestimation we obtain the upper bound of $2W^n w(X_i)$. For each input variable and neuron we can calculate with similar overestimation, that sums up to $2mW^n \sum_{i=1}^{k} w(X_i)$. According to this expression, the type of networks we discussed in the previous paragraph produce two times larger overestimation compared to fix sign weights considered in the present paragraph.

In all other scenarios the overestimation generated in the two first layers will be strengthened by the similar effects of the dependency problem, but this time the outputs of the earlier level neurons will play the role of the input intervals. The respective overestimation obtained on an output of a subsequent two level part of the network is $2mW^2 \sum_{i=1}^{m} w(Y_i)$, where Y_i are the output intervals of the preceding part of the network.

The total overestimation amount can be calculated for an output value of a deep neural network as

$$2^{n/2} m^{n/2} W^n \sum_{i=1}^{k} w(X_i),$$

for even integers n. For odd n-s this upper bound is accordingly smaller by the amount of the last overestimation caused by the effect of the dependency problem on the top layer. □

Example 2.3. To illustrate the overestimation caused by the dependency problem in interval based verification of artificial neural networks, consider a simple network having two neurons in the first layer and the ReLU activation functions after them, and one more layer having a single neuron. All neurons have zero bias for simplicity.

The function that describes it effect is

$$f(x) = w_3 \max(0, w_1 x) + w_4 \max(0, w_2 x)).$$

Fix the weights to $w_1 = w_2 = 1$ and $w_3 = 1$, $w_4 = -1$. Let's calculate first the inclusion function of $f(x)$ for the argument interval $[0, 1]$:

$$\begin{aligned} F([0,1]) &= \mathrm{ReLU}([0,1]) - \mathrm{ReLU}([0,1]) \\ &= \mathrm{ReLU}([0,1]) - \mathrm{ReLU}([0,1]) = [0,1] - [0,1] = [-1,1]. \end{aligned}$$

Here, for the sake of simplicity, the same notation was used for the interval version of the ReLU function as for the usual, real one. Otherwise, the interval version of ReLU can be e.g. $[\max(0, a), \max(0, b)]$ for an argument interval of $[a, b]$. The correct range of this network as a real function is $[0,\ 0]$. The amount of overestimation is now $2 - 0 = 2$. This is in accordance with Assertion 2.2, and shows that the upper bound given there is sharp (with $m = 2$, $n = 2$, $W = 1$, and $k = 1$).

Corollary 2.4. *A direct consequence of Assertion 2.2 is that we can have the same amount of overestimation due to the dependency problem with decreasing the number of hidden layers while increasing the number of neurons in a layer and vice versa.*

3. Conclusion

Interval arithmetic based methods are promising for the verification of artificial neural networks, but the dependency problem is a serious threat, and it should be handled carefully to obtain reasonable size adversarial example free zones in acceptable amount of computation time. We have characterized those few cases when the dependency problem does not corrupt the inclusion function of a network, and described how the parameters (input size, weight parameter bound, number of layers and neurons in the layers) affect the overestimation of interval inclusion functions.

The main consequence of our theoretical study is that we can control the amount of overestimation caused by the dependency effect of interval arithmetic by forcing advantageous parameters such as low absolute bound of weights, or minimizing the number of hidden layers – while keeping the expected level of precision and recall. Still, the obvious proven full solution for the dependency problem, a single hidden layer with proper activation to get the output, is probably computationally too complex to be applicable.

Acknowledgements. This research was supported by the project Extending the activities of the HU-MATHS-IN Hungarian Industrial and Innovation Mathematical Service Network EFOP3.6.2-16-2017-00015, 2018-1.3.1-VKE-2018-00033, and by the Subprogramme for Linguistic Identification of Fake News and Pseudoscientific Views, part of the Science for the Hungarian Language National Programme of the Hungarian Academy of Sciences (MTA). The author acknowledges the implementation and testing work done by Soma Timer.

References

[1] B. Bánhelyi, T. Csendes, B. Garay, L. Hatvani: *A computer-assisted proof for Sigma_3-chaos in the forced damped pendulum equation*, SIAM J. on Applied Dynamical Systems 7 (2008), pp. 843–867, doi: 10.1137/070695599.

[2] B. Bánhelyi, T. Csendes, T. Krisztin, A. Neumaier: *Global attractivity of the zero solution for Wright's equation*, SIAM J. on Applied Dynamical Systems 13 (2014), pp. 537–563, doi: 10.1137/120904226.

[3] T. Csendes: *An interval method for bounding level sets of parameter estimation problems*, Computing 41 (1989), pp. 75–86, doi: 10.1007/BF02238730.

[4] T. Csendes, N. Balogh, B. Bánhelyi, D. Zombori, R. Tóth, I. Megyeri: *Adversarial Example Free Zones for Specific Inputs and Neural Networks*, in: Proceedings of the 2020 ICAI, Eger, Hungary, 2020, url: https://ceur-ws.org/Vol-2650/paper9.pdf.

[5] G. Cybenko: *Approximation by superpositions of a sigmoidal function*, Mathematics of Control, Signals, and Systems 2 (1989), pp. 303–314, doi: 10.1007/BF02551274.

[6] H. Ratschek, J. Rokne: *Computer Methods for the Range of Functions*, Chichester: Ellis Horwood, 1984, doi: 10.2307/2008155.

[7] C. Szegedy, W. Zaremba, I. Sutskever, J. Bruna, D. Erhan, I. Goodfellow, R. Fergus: *Intriguing properties of neural networks*, in: Proceedings of the 2014 International Conference on Learning Representations, 2014, doi: 10.48550/arXiv.1312.6199.

[8] S. Timer: *Interval method for the verification of artificial neural networks in Julia language (in Hungarian)*, Szeged: University of Szeged, BSc Dissertation, 2023.

[9] V. Tjeng, K. Xiao, R. Tedrake: *Evaluating Robustness of Neural Networks with Mixed Integer Programming*, in: Proceedings of the 2019 International Conference on Learning Representations, 2019, doi: 10.48550/arXiv.1711.07356.

[10] S. Wang, K. Pei, J. Whitehouse, J. Yang, S. Jana: *Formal security analysis of neural networks using symbolic intervals*, in: Proceedings of the 27th USENIX Conference on Security Symposium, SEC'18, 2018, pp. 1599–1614, url: 10.48550/arXiv.1804.10829.

[11] D. Zombori, B. Bánhelyi, T. Csendes, I. Megyeri, M. Jelasity: *Fooling a complete neural network verifier*, in: Proceedings of the 2021 International Conference on Learning Representations, 2021, url: https://openreview.net/forum?id=4IwieFS44l.

Annales Mathematicae et Informaticae
60 (2024) pp. 27–36
DOI: 10.33039/ami.2024.03.002
URL: https://ami.uni-eszterhazy.hu

Isoptic curves of cycloids

Géza Csima

Department of Algebra and Geometry, Institute of Mathematics,
Budapest University of Technology and Economics,
Műegyetem rkp. 3., H-1111 Budapest, Hungary
csgeza@math.bme.hu

Abstract. The history of the isoptic curves goes back to the 19th century, but nowadays the topic is experiencing a renaissance, providing numerous new results and new applications. First, we define the notion of isoptic curve and outline some of the well-known results for strictly convex, closed curves. Overviewing the types of centered trochoids, we will be able to give the parametric equation of the isoptic curves of hypocycloids and epicycloids. Furthermore, we will determine the corresponding class of curves. Simultaneously, we show that a generalized support function can be given to these types of curves in order to apply and extend the results for strictly convex, closed curves. The calculation methods used during the procedure provide an excellent example of the application of univariate calculus, parametric curves, and vector calculus in geometry and can therefore be processed by either advanced high school students or university students.

Keywords: isoptic curves, trochoids

AMS Subject Classification: 51M04, 53A04, 97G70

1. Introduction

In this manuscript we work in the Euclidean plane $\mathbf{E}^2$. Let us introduce the following definition:

Definition 1.1 ([30])**.** The locus of the intersection of tangents to a curve (or curves) meeting at a constant angle α $(0 < \alpha < \pi)$ is the α-isoptic of the given curve (or curves). The isoptic curve with right angle called *orthoptic curve*.

Although the name "isoptic curve" was suggested by Taylor in 1884 ([26]), reference to former results can be found in [30]. In the obscure history of isoptic curves, we can find the names of la Hire (cycloids 1704) and Chasles (conics and

Submitted: April 11, 2023
Accepted: March 4, 2024
Published online: March 13, 2024

epitrochoids 1837) among the contributors of the subject, however, the details of the research results are not available in English. A very interesting table of isoptic and orthoptic curves is introduced in [30], unfortunately without any exact reference of its source. *Our goal in this paper is to independently reconstruct some of the missing computations for the isoptic curves of hypocycloids and epicycloids and to extend the results presented in [2] and [11].*

However, recent works are available on the topic, which shows its timeliness. In [2] the Euclidean isoptic curves of closed strictly convex curves are studied using their support function. Papers [15, 28, 29] deal with Euclidean curves having a circle or an ellipse for an isoptic curve. Further curves appearing as isoptic curves are well studied in Euclidean plane geometry $\mathbf{E}^2$, see e.g. [17, 23]. Isoptic curves of conic sections have been studied in [12] and [24]. There are results for Bezier curves by Kunkli et al. as well, see [14]. Many papers focus on the properties of isoptics, e.g. [18–20], and the references therein. There are some generalizations of the isoptics as well e.g. equioptic curves in [22] by Odehnal or secantopics in [21, 25] by Skrzypiec.

An algorithm for convex polyhedrons has been given by the authors in [8] in order to generalize the notion of isoptic curve into the space and it has been developed by Kunkli et al. for non convex cases in [13]. The spatial case encompasses many applications in both physical and architectural aspects, see [9].

There are some results in non-Euclidean geometries as well. The isoptic curves of the hyperbolic line segment and proper conic sections are determined in [4, 6, 7]. For generalized conic sections, and for their isoptics, see [5]. The isoptics of conic sections in elliptic geometry $\mathcal{E}^2$ are determined in [4].

There are some results in three dimensional Thurson geometries as well. The isoptic surface of segments has been determined in [10] in **Nil** geometry and in [3] for $\mathbf{S}^2\times\mathbf{R}$ and $\mathbf{H}^2\times\mathbf{R}$ geometries.

2. Preliminary results

In order to conduct further investigations on isoptics we need to summarize some preliminary results on the support function. We remind the dear reader that the complex number plane and $\mathbf{R}^2$ are isomorphic to each other, in such a way that real and imaginary parts give the first and second coordinates respectively. Furthermore, we will use the well-known Euler formula: $e^{it} = \cos(t) + i\sin(t) \Leftrightarrow (\cos(t), \sin(t))$.

Definition 2.1. Let $\mathcal{C}$ be a closed, strictly convex curve which surrounds the origin. Let $p(t)$ where $t \in [0, 2\pi[$ be the distance from 0 to the support line of $\mathcal{C}$ being perpendicular to the vector e^{it}. The function p is called a support function of $\mathcal{C}$.

It is well-known [1] that the support function of a planar, closed, strictly convex curve $\mathcal{C}$ is differentiable. For now we would like to express the isoptic of $\mathcal{C}$ using

the support function. We claim the following lemma omitting the proof which can be found for example in [27].

Lemma 2.2 ([27]). *If $f(x,y,t)=0$ is a family of straight lines, then the equation of the envelope of these lines can be obtained by eliminating the variable t from the two equations $f(x,y,t)=0$ and $\frac{\mathrm{d}}{\mathrm{d}t}f(x,y,t)=0$.*

This is used in [27] to prove the following theorem.

Theorem 2.3 ([27]). *Given a planar, closed, strictly convex curve $\mathcal{C}$ in polar coordinates with the radius z a function of angle t, where $t\in[0,2\pi)$. Then the following equation holds*

$$z(t)=p(t)e^{it}+\dot{p}(t)ie^{it}.$$

The corollary of this theorem is that we may use this parametrization to determine the isoptic curve of $\mathcal{C}$. The angle of $p(t)$ and $p(t+\pi-\alpha)$ is α, since the $p(t)$, $p(t+\pi-\alpha)$ and their support lines determine a cyclic quadrilateral (see Figure 1). Our goal is to determine the intersection of these tangent lines which is the fourth vertex opposite the origin. A proof can be found in [2].

Theorem 2.4 ([2]). *Let $\mathcal{C}$ be a plane, closed, strictly convex curve and suppose that the origin is in the interior of $\mathcal{C}$. Let $p(t)$, $t\in[0,2\pi]$ be the support function of $\mathcal{C}$. Then the α-isoptic curve of $\mathcal{C}$ has the form*

$$z_\alpha(t)=p(t)e^{it}+\left(-p(t)\cot(\pi-\alpha)+\frac{1}{\sin(\pi-\alpha)}p(t+\pi-\alpha)\right)ie^{it}. \tag{2.1}$$

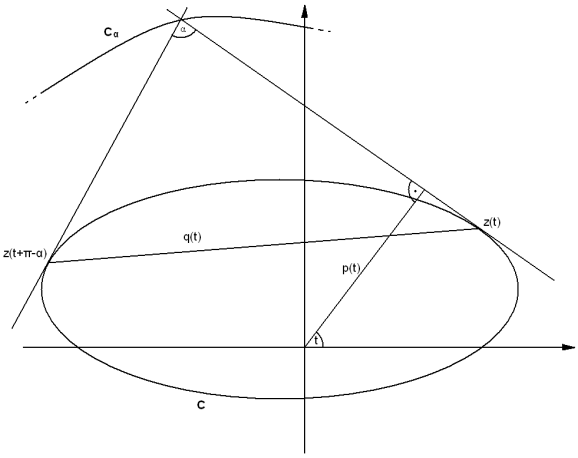


Figure 1

Definition 2.5 ([30]). A *hypocycloid* is generated by a point on a circle rolling internally upon a fixed circle. An *epicycloid* is generated by a point on a circle

rolling externally upon a fixed circle. A *hypotrochoid* is generated by a point rigidly attached to a circle rolling internally upon a fixed circle. An *epitrochoid* is generated by a point rigidly attached to a circle rolling externally upon a fixed circle.

We will use the following parametric equations of the hypo- and epicycloids, where we assumed that the radius of the fixed circle is 1, and the radius of the rolling circle is rational $\frac{1}{a} := \frac{p}{q} < 1$ in its lowest terms, otherwise the curve never closes, and fills the space between the circles. Then we have exactly p cusps and it is closed if and only if the length of parametric domain of t is greater than equal to $2q\pi$. In the case of hypocycloid, we also assume, that $2p \neq q$, which results in a segment. Since the parametrization of the hypo- and epicycloids are very similar, we consider them together. Hereinafter, in the notation $\pm$ or $\mp$, the upper sign refers to the hypocycloid, the lower one to the epicycloid.

$$\left\{\frac{(a\mp 1)\cos(t) \pm \cos((a\mp 1)t)}{a}, \frac{(a\mp 1)\sin(t) - \sin((a\mp 1)t)}{a}\right\}$$

Finally, we need the parametric equations of the hypo- and epitrochoids using the above sign convention:

$$\left\{(A \pm B)\cos(t) \pm H\cos\left(\frac{A\mp B}{B}t\right), (A\mp B)\sin(t) - H\sin\left(\frac{A\mp B}{B}t\right)\right\} \tag{2.2}$$

where the radius of the fixed and the rolling circles are A and B respectively, and H is the distance of the rigid point to the center of the rolling circle (see [16]).

3. Isoptic curves

Since the calculations of the isoptic curves of hypo- and epicycloids are also very similar, we consider them together. Our first step, to determine the isoptic curves, is always the tangent calculation. We need the derivative of the parametrization:

$$\begin{aligned} v_{H/E}(t) &= \left\{\mp\frac{2(a\mp 1)\sin\left(\frac{at}{2}\right)\cos\left(\frac{(a\mp 2)t}{2}\right)}{a}, \frac{2(a\mp 1)\sin\left(\frac{at}{2}\right)\sin\left(\frac{(a\mp 2)t}{2}\right)}{a}\right\} \\ &= \frac{2(a\mp 1)\sin\left(\frac{at}{2}\right)}{a}\left\{\mp\cos\left(\frac{(a\mp 2)t}{2}\right), \sin\left(\frac{(a\mp 2)t}{2}\right)\right\} \\ &= \frac{2(a\mp 1)\sin\left(\frac{at}{2}\right)}{a}\overline{v}_{H/E}(t) \end{aligned} \tag{3.1}$$

where we applied trigonometric product-to-sum and sum-to-product identities.

Remark 3.1. The tangent vector can be a null vector for discrete parameter values if $\sin\left(\frac{at}{2}\right) = 0$, but its direction may be determined in limit so that continuity remains.

Now, it is easy to see, that the angle of two tangents is equal to the the angle of the corresponding tangent vectors. Considering the $t+\phi$ and $t-\phi$ parametric values:

$$\begin{aligned}
&\frac{\langle v_{H/E}(t-\phi), v_{H/E}(t+\phi)\rangle}{\|v_{H/E}(t-\phi)\|\|v_{H/E}(t+\phi)\|}\\
&= \langle \overline{v}_{H/E}(t-\phi), \overline{v}_{H/E}(t+\phi)\rangle\\
&= \cos\left(\frac{(a\mp 2)(t-\phi)}{2}\right)\cos\left(\frac{(a\mp 2)(t+\phi)}{2}\right)\\
&\quad + \sin\left(\frac{(a\mp 2)(t-\phi)}{2}\right)\sin\left(\frac{(a\mp 2)(t+\phi)}{2}\right)\\
&= \cos\left(\frac{(a\mp 2)(t+\phi)}{2} - \frac{(a\mp 2)(t-\phi)}{2}\right) = \cos((a\mp 2)\phi)
\end{aligned} \tag{3.2}$$

that is *independent form the parameter value* of t. This is not necessarily true for other classes of curves. This uniformity gives us the possibility to determine the isoptic curve. Let $\phi := \frac{\alpha}{a\mp 2}$ be true, if we are interested in the α-isoptic curve. Then the angle of the oriented tangents that are drawn to points corresponding to the parameter values $t-\phi$ and $t+\phi$ is α.

Remark 3.2. In the case of the astroid ($a=4$), the value of ϕ is $\frac{\alpha}{2}$ so that the difference of considered two points in the parameter domain is exactly α.

From formula (3.1), we can derive the equation of the tangent respected to the parameter t:

$$x\sin\left(\frac{(a\mp 2)t}{2}\right) \pm y\cos\left(\frac{(a\mp 2)t}{2}\right) = \frac{(a\mp 2)\sin\left(\frac{at}{2}\right)}{a} \tag{3.3}$$

By replacing t with $t-\phi$ and $t+\phi$, we get an equation system. We are looking for the common point of the above tangents that will be a point of the isoptic curve related to the parameters t and α. When solving the system of equations, it is worth using Cramer's rule, since only 2×2 determinants need to be calculated. Due to the length of the solution process, we do not report it here, but after trigonometric simplification this is the result, which will be the parametrization of the isoptic curve, as well:

$$\begin{aligned}
x(t) &= \frac{(a\mp 2)\left(\sin\left(\frac{(a\mp 1)\alpha}{a\mp 2}\right)\cos(t) \pm \sin\left(\frac{\alpha}{a\mp 2}\right)\cos((a\mp 1)t)\right)}{a\sin(\alpha)}\\
y(t) &= \frac{(a\mp 2)\left(\sin\left(\frac{(a\mp 1)\alpha}{a\mp 2}\right)\sin(t) - \sin\left(\frac{\alpha}{a\mp 2}\right)\sin((a\mp 1)t)\right)}{a\sin(\alpha)}
\end{aligned}$$

We can propose the following theorem realizing the similarities to (2.2):

Theorem 3.3. *Let us be given a hypo- or epicycloid with its parametrization*

$$\left\{\frac{(a\mp 1)\cos(t)\pm\cos((a\mp 1)t)}{a},\frac{(a\mp 1)\sin(t)-\sin((a\mp 1)t)}{a}\right\}$$

where $a=\frac{q}{p}$ *and* $t\in[0,2q\pi]$ *such that* $p,q\in\mathbb{Z}^+\wedge p<q\wedge 2p\neq q$. *Then the* α*-isoptic curve of it is a hypo- or epitrochoid given by the parametrization*

$$\left\{(A\mp B)\cos(t)\pm H\cos\left(\frac{A\mp B}{B}t\right),(A\mp B)\sin(t)-H\sin\left(\frac{A\mp B}{B}t\right)\right\},$$

where

$$A=\frac{(a\mp 2)\sin\left(\frac{(a\mp 1)\alpha}{a\mp 2}\right)}{(a\mp 1)\sin(\alpha)},\quad B=\frac{(a\mp 2)\sin\left(\frac{(a\mp 1)\alpha}{a\mp 2}\right)}{a(a\mp 1)\sin(\alpha)},\quad H=\frac{(a\mp 2)\sin\left(\frac{\alpha}{a\mp 2}\right)}{a\sin(\alpha)}.$$

Remark 3.4. It is easy to see, that for the case of hypocycloid, if $\alpha=\frac{a-2}{a-1}\pi$, then $A=B=0$ in Theorem 3.3, so that the resulted parametric curve is a circle, centered at the origin with radius

$$\frac{(a-2)\sin\left(\frac{\pi}{a-1}\right)}{a\sin\left(\frac{a-2}{a-1}\right)}.$$

For the epicycloid, in Theorem 3.3 $A=B=0$ if $\alpha=\frac{a+2}{a+1}\pi$ but that angle is greater than π, therefore it is not a real isoptic curve.

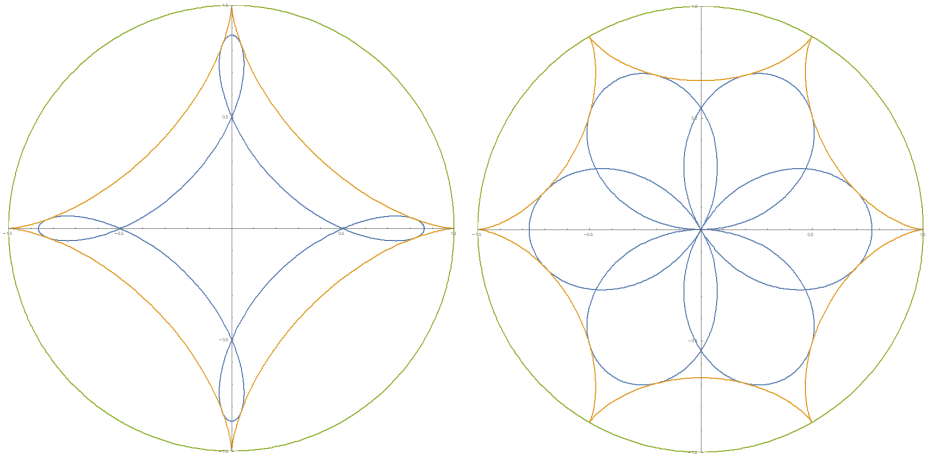


Figure 2. Isoptic curve for hypocycloid with $a=4$, $\alpha=\frac{\pi}{3}$ (left) and $a=6$, $\alpha=\frac{2\pi}{3}$ (right).

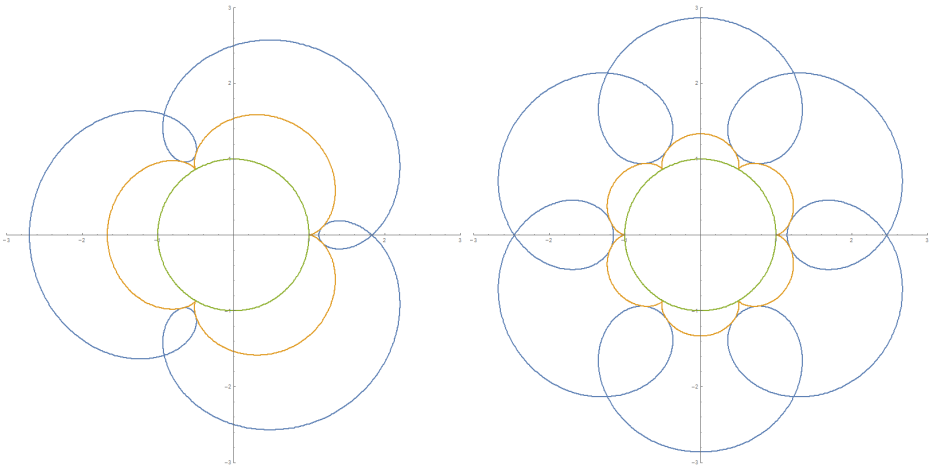


Figure 3. Isoptic curve for epicycloid with $a = 3$, $\alpha = \frac{\pi}{3}$ (left) and $a = 6$, $\alpha = \frac{\pi}{6}$ (right).

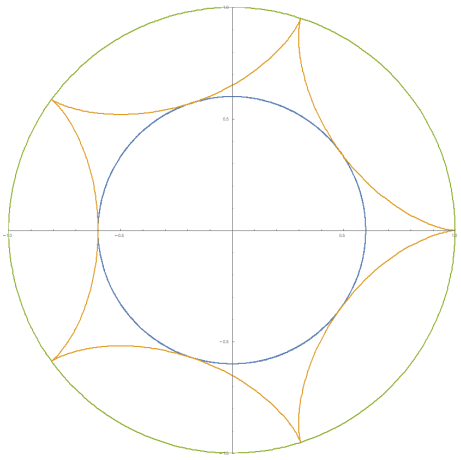


Figure 4. Isoptic curve as a circle for hypocycloid with $a = 5$, $\alpha = \frac{3\pi}{4}$.

4. Isoptic curves by support functions

One can realize that the tangents in formula (3.3) is in Hesse form, therefore it is easy to calculate the distance of the line to the origin. Despite hypocycloids and epicycloids are non-convex curves, we can define their support function nonetheless in order to give another approach of the isoptic curve by Theorem 2.4. We only

have to apply a substitution: $t = \frac{2}{a\mp 2}\left(\frac{\pi}{2} - u\right)$ in (3.1) to obtain:

$$x\cos(u) + y\sin(u) = \frac{(a\mp 2)}{a}\sin\left(\frac{a}{a\mp 2}\left(\frac{\pi}{2} - u\right)\right).$$

It is easy to see, that the transverse vector of the tangent is $e^{iu} = \{\cos(u), \sin(u)\}$ and its distance to the origin is $\frac{(a\mp 2)}{a}\sin\left(\frac{a}{a\mp 2}\left(\frac{\pi}{2} - u\right)\right)$. Then we can define the quasi-support functions:

$$p_{H/E}(u) = \frac{(a\mp 2)}{a}\sin\left(\frac{a}{a\mp 2}\left(\frac{\pi}{2} - u\right)\right). \tag{4.1}$$

From this point on, the procedure is fairly straightforward, especially with the help of some mathematical software. First, we apply (2.1) from Theorem 2.4 to (4.1), then we return to the original t variable. Finally, the resulting parametrization can be simplified with the help of the trigonometric product-to-sum and sum-to-product identities already used before, until we arrive at a form of the formulas that leads to the parametrization of the hypo- and epitrochoid by shifting the parameter range or, in the case of the epicycloid, by changing the direction as well.

Theorem 4.1. *Let us be given a $\mathcal{C}$ hypo- or epicycloid with its parametrization*

$$\mathcal{C}\colon \left\{\frac{(a\mp 1)\cos(t) \pm \cos((a\mp 1)t)}{a}, \frac{(a\mp 1)\sin(t) - \sin((a\mp 1)t)}{a}\right\}$$

where $a = \frac{q}{p}$ and $t \in [0, 2q\pi]$ such that $p, q \in \mathbb{Z}^+ \wedge p < q \wedge 2p \neq q$. Then the α-isoptic curve of $\mathcal{C}$ has the form

$$z_\alpha(t) = p(t)e^{it} + \left(-p(t)\cot(\pi - \alpha) + \frac{1}{\sin(\pi - \alpha)}p(t + \pi - \alpha)\right)ie^{it},$$

where $p(t) = \frac{(a\mp 2)}{a}\sin\left(\frac{a}{a\mp 2}\left(\frac{\pi}{2} - t\right)\right)$, $t \in [0, 2\pi]$ is the support function of $\mathcal{C}$.

References

[1] T. Bonnesen, W. Fenchel: *Ergebnisse der Mathematik und Ihrer Grenzgebiete. 1. Folge*, in: Theorie der Konvexen Körper, Berlin, Heidelberg: Springer, 1934, DOI: 10.1007/978-3-642-47404-0.

[2] W. Cieślak, A. Miernowski, W. Mozgawa: *Isoptics of a closed strictly convex curve*, in: Global Differential Geometry and Global Analysis, Berlin, Heidelberg: Springer, 1991, pp. 28–35, ISBN: 978-3-540-46445-7, DOI: 10.1007/BFb0083625.

[3] G. Csima: *Isoptic surfaces of segments in $\boldsymbol{S}^2\times\boldsymbol{R}$ and $\boldsymbol{H}^2\times\boldsymbol{R}$ geometries*, Journal of Geometry 115.1 (2023), DOI: 10.1007/s00022-023-00699-x.

[4] G. Csima, J. Szirmai: *Isoptic curves of conic sections in constant curvature geometries*, Mathematical Communications 19.2 (2014), pp. 277–290.

[5] G. Csima, J. Szirmai: *Isoptic curves of generalized conic sections in the hyperbolic plane*, Ukrainien Mathematical Journal 71 (12 2020), pp. 1929–1944, DOI: 10.1007/s11253-020-01756-3.

[6] G. Csima, J. Szirmai: *Isoptic curves of the conic sections in the hyperbolic and elliptic plane*, Stud. Univ. Žilina 24.1 (2010), pp. 15–22.

[7] G. Csima, J. Szirmai: *Isoptic curves to parabolas in the hyperbolic plane*, Pollac Periodica 1.1 (2012), pp. 55–64.

[8] G. Csima, J. Szirmai: *Isoptic surfaces of polyhedra*, Computer Aided Geom. Design 47 (2016), pp. 55–60, DOI: 10.1016/j.cagd.2016.03.001.

[9] G. Csima, J. Szirmai: *On the isoptic hypersurfaces in the n-dimensional Euclidean space*, KoG (Scientific and professional journal of Croatian Society for Geometry and Graphics) 17 (2013), pp. 53–57.

[10] G. Csima, J. Szirmai: *Translation-Like Isoptic Surfaces and Angle Sums of Translation Triangles in **Nil** Geometry*, Results in Mathematics 78 (5 2023), DOI: 10.1007/s00025-023-01961-z.

[11] T. Dana-Picard: *An automated study of isoptic curves of an astroid*, Special Issue on Dynamic Geometry and Automated Reasoning 97 (2020), pp. 56–68, DOI: 10.1016/j.jsc.2018.12.005.

[12] G. Holzmüller: *Einführung in die Theorie der isogonalen Verwandtschaft*, Leipzig-Berlin: B.G. Teuber, 1882.

[13] R. Kunkli, F. Nagy, M. Hoffmann: *New algorithm to find isoptic surfaces of polyhedral meshes*, Computer Aided Geometric Design 64 (1 2018), pp. 90–99, DOI: 10.1016/j.cagd.2018.04.001.

[14] R. Kunkli, I. Papp, M. Hoffmann: *Isoptics of Bézier curves*, Computer Aided Geometric Design 30 (1 2013), pp. 78–84, DOI: 10.1016/j.cagd.2012.05.002.

[15] Á. Kurusa: *Is a convex plane body determined by an isoptic?*, Beiträge zur Algebra und Geometrie 53 (2012), pp. 281–294, DOI: 10.1007/s13366-011-0074-2.

[16] J. D. Lawrence: *A Catalog of Special Plane Curves*, Mineola, New York: Dover Publications, Inc., 1972, pp. 165–168.

[17] G. Loria: *Spezielle algebraische und traszendente ebene Kurve*, Leipzig-Berlin: B.G. Teuber, 1911.

[18] M. Michalska: *A sufficient condition for the convexity of the area of an isoptic curve of an oval*, Rend. Semin. Mat. Univ. Padova 110 (2003), pp. 161–169.

[19] M. Michalska, W. Mozgawa: *α-isoptics of a triangle and their connection to α-isoptic of an oval*, Rend. Semin. Mat. Univ. Padova 133 (2015), pp. 159–172.

[20] M. Michalska, W. Mozgawa: *On some geometric condition for convexity of isoptics*, Rend. Semin. Mat. Univ. Pol. Torino 55 (2 1997), pp. 93–98.

[21] M. Michalska, M. Skrzypiec: *Crofton formulas and convexity condition for secantopics*, Bull. Belg. Math. Soc. - Simon Stevin 16 (3 2009), pp. 435–445, DOI: 10.36045/bbms/1251832370.

[22] B. Odehnal: *Equioptic curves of conic section*, J. Geom. Graphics 14 (1 2010), pp. 29–43, DOI: 10.36045/bbms/1251832370.

[23] T. Sch: *Spezielle ebene Kurven*, Monatshefte für Mathematik und Physik 21 (1 1910), A60–A61, DOI: 10.1007/BF01693325.

[24] P. Siebeck: *Ueber eine Gattung von Curven vierten Grades, welche mit den elliptischen Functionen zusammenhängen.* ger, Journal für die reine und angewandte Mathematik 57 (1860), pp. 359–370, URL: http://eudml.org/doc/147795.

[25] M. Skrzypiec: *A note on secantopics*, Beiträge zur Algebra und Geometrie 49 (1 2008), pp. 205–215.

[26] C. Taylor: *Note on a theory of orthoptic and isoptic loci*, Proc. R. Soc. London 37 (1884), pp. 138–141, doi: 10.1098/rspl.1884.0024.

[27] D. Widder: *Advanced Calculus, 2nd edition*, Englewood Cliffs, NJ: Prentice-Hall, 1962.

[28] W. Wunderlich: *Kurven mit isoptischem Kreis*, Aequationes Mathematicae 5 (2 1970), pp. 337–338, doi: 10.1007/BF01818468.

[29] W. Wunderlich: *Kurven mit isoptischer Ellipse*, Monatsh. Math. 75 (1971), pp. 346–362.

[30] R. C. Yates: *A handbook on curves and their properties*, Ann Arbor: J. W. Edwards, 1947.

Annales Mathematicae et Informaticae
60 (2024) pp. 37–43
DOI: 10.33039/ami.2024.03.006
URL: https://ami.uni-eszterhazy.hu

An identity for two sequences and its combinatorial interpretation

Yahia Djemmada[a], **Abdelghani Mehdaoui**[a], **László Németh**[b*†], **László Szalay**[c†‡]

[a]National Higher School of Mathematics, Sidi Abdellah, Algiers, Algeria
yahia.djem@gmail.com and mehabdelghani@gmail.com

[b]Institute of Basic Sciences, Departement of Mathematics
University of Sopron, Hungary
nemeth.laszlo@uni-sopron.hu

[c]Institute of Basic Sciences, Departement of Mathematics
University of Sopron, Hungary
and Department of Mathematics, J. Selye University, Slovakia
szalay.laszlo@uni-sopron.hu

Abstract. We recall a theorem on linear recurrences that we have already proved earlier, and we use it to provide new identities. The nature of the new result allows us to combine two linear recurrences of distinct order in the identity if they satisfy some prescribed conditions about their similarity. For example, we found a rule including consecutive k- and ℓ-generalized Fibonacci numbers. In addition, a combinatorial interpretation is explained if the coefficients of the recurrences are positive integers.

Keywords: linear recurrence, combinatorial interpretation, generalized Fibonacci number, generalized Pell number

AMS Subject Classification: 11B37, 11B39, 05A19, 05B45

*Corresponding author.

†For these authors the research was supported in part by National Research, Development and Innovation Office Grant 2019-2.1.11-TÉT-2020-00165.

‡For the author the research was also supported by Hungarian National Foundation for Scientific Research Grant No. 128088, and No. 130909.

Submitted: July 23, 2023
Accepted: March 4, 2024
Published online: April 19, 2024

1. Introduction

Suppose k is a positive integer, and $f_0, f_1, \ldots, f_{k-1}$ are complex numbers. Define

$$f_n = A_1 f_{n-1} + A_2 f_{n-2} + \cdots + A_k f_{n-k} \quad (n \geq k), \tag{1.1}$$

where the coefficients $A_1, \ldots, A_{k-1}$, and $A_k \neq 0$ are fixed complex numbers. Moreover, suppose that $(w_n)_{n\geq 0} \in \mathbb{C}^\infty$ is an arbitrary sequence. Based on the notation above, we construct the linear recurrence

$$G_n = A_1 G_{n-1} + A_2 G_{n-2} + \cdots + A_k G_{n-k} + w_{n-k} \quad (n \geq k), \tag{1.2}$$

assuming that the complex initial values $G_0, G_1, \ldots, G_{k-1}$ are also given. Note that formulae (1.1) and (1.2) differ essentially only in the term w_n.

Belbachir et al. [1] studied the connection between the sequences $(G_n)_{n\geq 0}$, $(f_n)_{n\geq 0}$ and $(w_n)_{n\geq 0}$, and proved the following general result.

Theorem 1.1. *For $n \geq k$, the terms of the sequences (f_n), (w_n) and (G_n) satisfy the identity*

$$\begin{aligned}\sum_{j=0}^{k-1} f_j G_{n+k-j} = &\sum_{j=0}^{k-1}\sum_{i=0}^{k-1-j} f_{n-j} A_{j+1+i} G_{k-1-i} \\ &+ \sum_{j=0}^{k-2}\sum_{i=1}^{k-1-j} f_j A_i G_{n+k-j-i} + \sum_{j=0}^{n} f_{n-j} w_j.\end{aligned} \tag{1.3}$$

The theorem is valid also for $k = 1$ (with an empty sum of the three on the right-hand side), but this case is not of much interest. Hence, we may suppose $k \geq 2$. Observe that the terms $A_{j+1+i}G_{k-1-i}$ and $f_j A_i$ on the right-hand side of (1.3) can take only finitely many values. Moreover, note that Theorem 1.1 is obviously true for arbitrary initial values of the sequence (f_n). Coefficients $A_1, \ldots, A_k$ in the definition of (f_n) are important in the sense that they, together with (w_n) also establish the sequence (G_n). But, generally, the initial values $f_0, \ldots, f_{k-1}$ can be chosen arbitrarily. Therefore, it is natural, if there is no other reason, to put $f_0 = \cdots = f_{k-1} = 0$, $f_{k-1} = 1$. The next corollary simplifies Theorem 1.1 as it describes this situation. (See [1] again.)

Corollary 1.2. *Assume that $f_0 = \cdots = f_{k-2} = 0$, $f_{k-1} = 1$. Then (1.3) simplifies to*

$$G_{n+1} = \sum_{j=0}^{k-1}\sum_{i=0}^{k-1-j} f_{n-j} A_{j+1+i} G_{k-1-i} + \sum_{j=0}^{n} f_{n-j} w_j \quad (n \geq k). \tag{1.4}$$

If we even specify the initial values $G_0 = G_1 = \cdots = G_{k-1} = 0$ (and keep the former conditions $f_0 = \cdots = f_{k-2} = 0$, $f_{k-1} = 1$), then we have

Corollary 1.3. *Under the condition above,* (1.4) *admits*

$$G_{n+1} = \sum_{j=0}^{n} f_{n-j} w_j \quad (n \geq k). \tag{1.5}$$

This corollary was not mentioned in [1], but it is obviously a direct consequence of Theorem 1.1. An illustration of Corollary 1.3 stands here.

Example 1.4. Recall [1] again. Let $\ell \geq 3$ be an integer. Moreover, let $f_n = F_n$, the n^{th} term of the Fibonacci sequence (see The On-Line Encyclopedia of Integer Sequences [5], sequence A000045). Put $w_n = F_{n-1}^{(\ell)} + \cdots + F_{n-(\ell-2)}^{(\ell)}$. Here $(F_n^{(\ell)})$ is the ℓ-generalized Fibonacci sequence (Fibonacci ℓ-step numbers or shortly ℓ-nacci sequence) defined by the initial values $F_0^{(\ell)} = F_1^{(\ell)} = \cdots = F_{\ell-2}^{(\ell)} = 0$, $F_{\ell-1}^{(\ell)} = 1$, and by the recurrence relation

$$F_n^{(\ell)} = F_{n-1}^{(\ell)} + F_{n-2}^{(\ell)} + \cdots + F_{n-\ell}^{(\ell)} \quad (n \geq \ell).$$

We may also need to extend $(F_n^{(\ell)})$ for some terms with negative subscripts follows from the recurrence rule above when it is applied backward. Hence, $F_{-1}^{(\ell)} = 1$, $F_{-2}^{(\ell)} = -1$, $F_{-3}^{(\ell)} = 0$, and so on. Finally, we fix $G_0 = F_0^{(\ell)} = 0$ and $G_1 = F_1^{(\ell)} = 0$. In this manner, we construct the sequence (G_n), which is obviously the ℓ-generalized Fibonacci sequence itself. Thus, (1.5) provides the identity

$$F_{n+1}^{(\ell)} = \sum_{j=0}^{n} F_{n-j} \left(\sum_{i=1}^{\ell-2} F_{j-i}^{(\ell)} \right). \tag{1.6}$$

Suppose $\ell = 3$ to obtain the terms of the so-called Tribonacci sequence $(T_n) = (F_n^{(3)})$ A000073. Then we have

$$T_{n+1} = \sum_{j=0}^{n} F_{n-j} T_{j-1}.$$

This property is also given in [4], see (2.5) therein; moreover, see Benjamin and Quinn's book [2, p. 47, Exercise 4(a)].

The main purpose of this paper is to extend (1.6), and to give a combinatorial interpretation if the coefficients are positive integers.

2. Results

2.1. New corollaries of Theorem 1.1

Let $k \geq 2$ and $\ell > k$ be positive integers. Assume that the sequence (f_n) is given by the initial values

$$f_0 = \cdots = f_{k-2} = 0, \ f_{k-1} = 1, \tag{2.1}$$

and by the recursive scheme (1.1). The coefficients $A_1, \ldots, A_k$ are fixed in (1.1). Define the recurrence (G_n) such that

$$G_n = A_1 G_{n-1} + \cdots + A_k G_{n-k} + A_{k+1} G_{n-k-1} + \cdots + A_\ell G_{n-\ell}. \tag{2.2}$$

Fix G_i for $i \in \{k-\ell, \ldots, 0, \ldots, k-1\}$, and put

$$w_n = A_{k+1} G_{n-1} + \cdots + A_\ell G_{n+k-\ell} \quad \text{for } n \geq 0.$$

Clearly,

$$G_n = A_1 G_{n-1} + \cdots + A_k G_{n-k} + w_{n-k} \quad \text{for } n \geq k,$$

and the situation fits the construction described in the Introduction. According to Corollary 1.2 we obtain

Theorem 2.1. *Using the notation above the identity*

$$G_{n+1} = \sum_{j=0}^{k-1} \sum_{i=0}^{k-1-j} f_{n-j} A_{j+1+i} G_{k-1-i} + \sum_{j=0}^{n} f_{n-j} \left(\sum_{i=1}^{\ell-k} A_{k+i} G_{j-i} \right)$$

follows.

In particular, we specify Corollary 1.3 in Theorem 2.2.

Theorem 2.2. *The condition* $G_0 = \cdots = G_{k-1} = 0$ *leads to*

$$G_{n+1} = \sum_{j=0}^{n} f_{n-j} \left(\sum_{i=1}^{\ell-k} A_{k+i} G_{j-i} \right). \tag{2.3}$$

2.2. Combinatorial explanation of Theorem 2.2

Consider the domino tilings of a $1 \times h$ chessboard with 1×1, 1×2, …, $1 \times \ell$ dominoes having A_1, A_2, …, A_ℓ different colors, respectively. Let C_h denote the total number of tilings, and K_h the number of tilings if the maximal length of the dominoes we can use is k, where $k < \ell$.

Since the length of the last domino in the tiling is one of $1, 2, \ldots, \ell$, the total number of ways to tile is

$$C_h = A_1 C_{h-1} + \cdots + A_\ell C_{h-\ell}.$$

The comparison of the initial values of sequences (C_n) and (G_n) provides the equality $C_h = G_{h+l-1}$. Indeed, we extend the initial conditions $G_0 = \cdots = G_{k-1} = 0$ with $G_k = \cdots = G_{\ell-2} = 0$, $G_{\ell-1} = 1$ in the recurrence (2.2). So $C_0 = G_{\ell-1} = 1$, $C_1 = G_\ell = A_1$, $C_2 = G_{\ell+1} = A_1^2 + A_2$, and so on.

Now we examine another approach to calculate the number of tilings. In order to do that, first, we notify that $K_j = f_{j+k-1}$ gives the number of tilings of a $1 \times j$

chessboard with 1×1, 1×2, …, $1\times k$ dominoes using the given colors. (It similarly follows from (1.1) and (2.1).)

Assume that the first j positions of the chessboard are tiled under the restriction of maximal length k, and then the next domino has size either $1\times(k+1)$ or $1\times(k+2)$ or, and so on, or $1\times\ell$. For the remaining part of the chessboard, we can use any dominoes with maximum length ℓ (for illustration, see Figure 1). Thus,

$$C_h = \sum_{j=0}^{h} K_j \left(\sum_{i=k+1}^{\ell} A_i C_{h-j-i} \right).$$

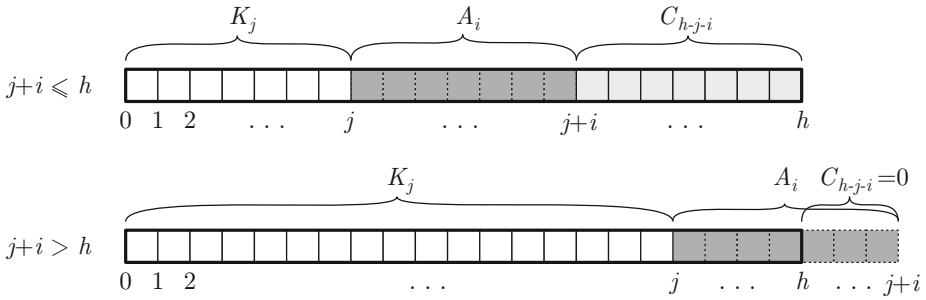


Figure 1. Chessboard and tilings.

Now we return to the sequences (G_n) and (f_n). Clearly,

$$G_{h+\ell-1} = \sum_{j=0}^{h} f_{j+k-1} \left(\sum_{i=k+1}^{\ell} A_i G_{h+\ell-1-j-i} \right).$$

If $h = n-\ell+2$, then we have

$$\begin{aligned} G_{n+1} &= \sum_{j=0}^{n-\ell+2} f_{j+k-1} \left(\sum_{i=k+1}^{\ell} A_i G_{n+1-j-i} \right) \\ &= \sum_{j=0}^{n-\ell+2} f_{n-(\ell-k-1)-j} \left(\sum_{i=k+1}^{\ell} A_i G_{j+(\ell-1)-i} \right) \\ &= \sum_{j=0}^{n-\ell+2} f_{n-(\ell-k-1)-j} \left(\sum_{i=1}^{\ell-k} A_{i+k} G_{j+(\ell-k-1)-i} \right) \\ &= \sum_{j=\ell-k-1}^{n-k+1} f_{n-j} \left(\sum_{i=1}^{\ell-k} A_{i+k} G_{j-i} \right). \end{aligned}$$

In the above equalities, first we used the swap $j \leftrightarrow h-j$, and then certain re-indexing.

Observe that the range of the first sum can be extended from $n-k+1$ to n. Indeed, the new coefficients $f_{k-2},\dots,f_0$ are all zero. Similarly, we can modify the sum by reducing the lower value from $\ell-k-1$ to 0 because for such j, the sum

$$\sum_{i=1}^{\ell-k} A_{i+k}G_{j-i}$$

vanishes. Finally, we have obtained

$$G_{n+1}=\sum_{j=0}^{n} f_{n-j}\left(\sum_{i=1}^{\ell-k} A_{i+k}G_{j-i}\right),$$

which is identical to formula (2.3) given in Theorem 2.2.

2.3. Formula for ℓ-generalized Fibonacci sequences

Let all the coefficients A_i be equal to 1. Writing the usual notation of k- and ℓ-generalized Fibonacci sequences (as in Example 1.4) formula (2.3) yields

$$F_{n+1}^{(\ell)}=\sum_{j=0}^{n} F_{n-j}^{(k)}\sum_{i=1}^{\ell-k} F_{j-i}^{(\ell)},\quad (2\le k<\ell).$$

This identity extends (1.6) of Example 1.4.

2.4. Formula for ℓ-generalized Pell sequences

Let $(P_n^{(\ell)})$ denote the ℓ-generalized Pell sequence (or shortly ℓ-Pell sequence), where the initial values are $P_0^{(\ell)}=P_1^{(\ell)}=\cdots=P_{\ell-2}^{(\ell)}=0$, $P_{\ell-1}^{(\ell)}=1$, and the recurrence is given by

$$P_n^{(\ell)}=2P_{n-1}^{(\ell)}+P_{n-2}^{(\ell)}+\cdots+P_{n-\ell}^{(\ell)}. \tag{2.4}$$

If $\ell=2$, then it gives the Pell sequence ($P_n=2P_{n-1}+P_{n-2}$, $P_0=0$, $P_1=1$, A000129 in OEIS [5]). When we apply the recurrence backward rule (2.4) we obtain the terms of the ℓ-Pell sequence with negative subscripts.

Recently, Bravo, Herrera, and Ramírez [3] presented some properties and combinatorial interpretations of the ℓ-generalized Pell sequences.

Lastly we provide a new identity involving the terms of k- and ℓ-generalized Pell sequences as a corollary of formula (2.3) given in Theorem 2.2. For this reason, we put $A_1=2$ and $A_2=A_3=\cdots=A_{\ell-1}=1$, and refer $(f_n)=(P_n^{(k)})$ and $(G_n)=(P_n^{(\ell)})$. Thus, (2.3) admits

$$P_{n+1}^{(\ell)}=\sum_{j=0}^{n} P_{n-j}^{(k)}\sum_{i=1}^{\ell-k} P_{j-i}^{(\ell)},\quad (2\le k<\ell).$$

Conflict of interest. The authors declare that they have no conflict of interest.

References

[1] H. Belbachir, F. Rami, L. Szalay: *A generalization of hyperbolic Pascal triangles*, J. Combin. Theory Ser. A. 188 (2022), p. 105574, ISSN: 0097-3165, DOI: 10.1016/j.jcta.2021.105574.

[2] A. Benjamin, J. Quinn, W. Watkins: *Proofs That Really Count: The Art of Combinatorial Proof*, Mathematical Association of America, 2003, ISBN: 9781614442080, DOI: 10.5948/9781614442080.

[3] J. J. Bravo, J. L. Herrera, J. L. Ramírez: *Combinatorial Interpretation of Generalized Pell Numbers*, J. Integer Seq. 23.2 (2020), Article 20.2.1, URL: https://cs.uwaterloo.ca/journals/JIS/VOL23/Bravo/bravo4.html.

[4] R. Frontczak: *Relations for generalized Fibonacci and Tribonacci sequences*, Notes Number Theor. Disc. Math. 25.1 (2019), pp. 178–192, DOI: 10.7546/nntdm.2019.25.1.178-192.

[5] OEIS Foundation Inc.: *The On-Line Encyclopedia of Integer Sequences*, Published electronically at https://oeis.org, 2024, URL: http://oeis.org.

Annales Mathematicae et Informaticae
60 (2024) pp. 44–53
DOI: 10.33039/ami.2024.03.005
URL: https://ami.uni-eszterhazy.hu

A note on the exponential Diophantine equation $(a^x - 1)(b^y - 1) = az^2$

Yasutsugu Fujita[a], **Maohua Le**[b]

[a]Department of Mathematics, College of Industrial Technology, Nihon University, 2-11-1 Shin-ei, Narashino, Chiba, Japan
fujita.yasutsugu@nihon-u.ac.jp

[b]Institute of Mathematics, Lingnan Normal College, Zhanjiang, Guangdong, 524048 China
lemaohua2008@163.com

Abstract. Let a, b be fixed positive integers such that $(a \bmod 8, b \bmod 8) \in \{(0,3),(0,5),(2,3),(2,5),(4,3),(6,5)\}$. In this paper, using elementary methods with some classical results for Diophantine equations, we prove the following three results: (i) The equation $(*)$ $(a^x - 1)(b^y - 1) = az^2$ has no positive integer solutions (x, y, z) with $2 \nmid x$ and $x > 1$. (ii) If $a = 2$ and $b \equiv 5 \pmod 8$, then $(*)$ has no positive integer solutions (x, y, z) with $2 \nmid x$. (iii) If $a = 2$ and $b \equiv 3 \pmod 8$, then the positive integer solutions (x, y, z) of $(*)$ with $2 \nmid x$ are determined. These results improve the recent results of R.-Z. Tong: On the Diophantine equation $(2^x - 1)(p^y - 1) = 2z^2$, Czech. Math. J. 71 (2021), 689–696. Moreover, under the assumption that a is a square, we prove that $(*)$ has no positive integer solutions (x, y, z) even with $2 \mid x$ in some cases.

Keywords: polynomial-exponential Diophantine equation, Pell's equation, generalized Ramanujan-Nagell equation

AMS Subject Classification: 11D61

1. Introduction

Let $\mathbb{N}$ be the set of all positive integers. Let a, b be fixed positive integers with $\min\{a, b\} > 1$. In 2000, L. Szalay [7] completely solved the equation

$$(2^x - 1)(3^x - 1) = z^2, \quad x, z \in \mathbb{N}. \tag{1.1}$$

Submitted: November 25, 2021
Accepted: March 22, 2024
Published online: March 24, 2024

He proved that (1.1) has no solutions (x, z). Since then, this result has led to a series of related studies for the equation

$$(a^x - 1)(b^x - 1) = z^2, \quad x, z \in \mathbb{N} \tag{1.2}$$

(see [3]). Obviously, the solution of (1.2) involves a system of generalized Ramanujan-Nagell equations. Recently, R.-Z. Tong [8] discussed the equation

$$(2^x - 1)(p^y - 1) = 2z^2, \quad x, y, z \in \mathbb{N}, \tag{1.3}$$

where p is an odd prime with $p \equiv \pm 3 \pmod 8$. He proved the following two results: (i) (1.3) has no solutions (x, y, z) with $2 \nmid x$, $2 \mid y$ and $y > 4$. (ii) If $p \neq 2g^2 + 1$, where g is an odd positive integer, then (1.3) has no solutions (x, y, z) with $2 \nmid x$. In this paper, we will discuss the generalized form of (1.3) as follows:

$$(a^x - 1)(b^y - 1) = az^2, \quad x, y, z \in \mathbb{N}. \tag{1.4}$$

For any positive integer n, let r_n, s_n be the positive integers satisfying

$$r_n + s_n\sqrt{2} = \left(3 + 2\sqrt{2}\right)^n. \tag{1.5}$$

For any odd positive integer m, let R_m, S_m be the positive integers satisfying

$$R_m + S_m\sqrt{2} = \left(1 + \sqrt{2}\right)^m. \tag{1.6}$$

Using elementary methods with some classical results for Diophantine equations, we prove the following results:

Theorem 1.1. *If*

$$(a \mod 8, b \mod 8) \in \{(0,3), (0,5), (2,3), (2,5), (4,3), (6,5)\}, \tag{1.7}$$

then (1.4) *has no solutions* (x, y, z) *with* $2 \nmid x$ *and* $x > 1$.

Theorem 1.2. *If* $a = 2$ *and* $b \equiv 5 \pmod 8$, *then* (1.4) *has no solutions* (x, y, z) *with* $2 \nmid x$. *If* $a = 2$ *and* $b \equiv 3 \pmod 8$, *then* (1.4) *has only the following solutions* (x, y, z) *with* $2 \nmid x$:

(i) $b = 3$, $(x, y, z) = (1, 1, 1)$, $(1, 2, 2)$ *and* $(1, 5, 11)$.

(ii) $b = 2g^2 + 1$, $(x, y, z) = (1, 1, g)$, *where* g *is an odd positive integer with* $g > 1$.

(iii) $b = r_m$, $(x, y, z) = (1, 2, s_m)$, *where* m *is an odd positive integer with* $m > 1$.

Theorem 1.3. *Let* $N(a, b)$ *denote the number of solutions* (x, y, z) *of* (1.4) *with* $2 \nmid x$. *If* $a = 2$ *and* $b \equiv 3 \pmod 8$, *then*

$$N(2, b) = \begin{cases} 3, & \text{if } b = 3, \\ 2, & \text{if } b = 2g^2 + 1 \text{ and } g = R_m \text{ with } m > 1, \\ 1, & \text{if } b = 2g^2 + 1 \text{ and } g \neq R_m, \\ 0, & \text{otherwise.} \end{cases}$$

Obviously, the above theorems improve the result of [8].

The following results concern the solvability of (1.4) including even the case where $2 \mid x$.

Theorem 1.4. *If* $(a \bmod 8, b \bmod 8) \in \{(0,3),(0,5),(4,3)\}$ *and* a *is a square, then* (1.4) *has no solutions* (x,y,z) *with* $x > 1$.

Theorem 1.5. *Assume that one of the following conditions holds*:

(i) $a = 4$ *and either* $b = 3$ *or* b *has a prime divisor* p *with* $p \equiv 11 \pmod{24}$.

(ii) $a = 16$ *and either* $b \in \{3,5\}$ *or* b *has a prime divisor* p *with*

$$p \equiv 11, 13, 29, 37, 43, 59, 67 \text{ or } 101 \pmod{120}.$$

Then, (1.4) *has no solutions.*

2. Preliminaries

Let D be a nonsquare positive integer, and let D_1, D_2 be positive integers such that $D_1 > 1$, $D_1 D_2 = D$ and $\gcd(D_1, D_2) = 1$. By the basic properties of Pell's equation (see [5, 10] and [4, Lemma 1]), we obtain the following two lemmas immediately.

Lemma 2.1. *The equation*

$$u^2 - Dv^2 = 1, \quad u, v \in \mathbb{N} \tag{2.1}$$

has solutions (u,v)*, and it has a unique solution* (u_1, v_1) *such that* $u_1 + v_1\sqrt{D} \le u + v\sqrt{D}$*, where* (u,v) *runs through all solutions of* (2.1)*. The solution* (u_1, v_1) *is called the least solution of* (2.1)*. For any positive integer* n*, let* $u_n + v_n\sqrt{D} = \left(u_1 + v_1\sqrt{D}\right)^n$*. Then we have*

(i) $(u,v) = (u_n, v_n)$ $(n = 1, 2, \dots)$ *are all solutions of* (2.1).

(ii) *If* $2 \mid n$*, then each prime divisor* p *of* u_n *satisfies* $p \equiv \pm 1 \pmod{8}$.

(iii) *If* $2 \nmid n$*, then* $u_1 \mid u_n$.

Lemma 2.2. *If the equation*

$$D_1 U^2 - D_2 V^2 = 1, \quad U, V \in \mathbb{N} \tag{2.2}$$

has solutions (U,V)*, then it has a unique solution* (U_1, V_1) *such that* $U_1\sqrt{D_1} + V_1\sqrt{D_2} \le U\sqrt{D_1} + V\sqrt{D_2}$*, where* (U,V) *runs through all solutions of* (2.2)*. The solution* (U_1, V_1) *is called the least solution of* (2.2)*. For any odd positive integer* m*, let* $U_m\sqrt{D_1} + V_m\sqrt{D_2} = \left(U_1\sqrt{D_1} + V_1\sqrt{D_2}\right)^m$*. Then we have*

(i) $(U,V) = (U_m, V_m)$ $(m = 1, 3, \dots)$ *are all solutions of* (2.2).

(ii) $u_1 + v_1\sqrt{D} = \left(U_1\sqrt{D_1} + V_1\sqrt{D_2}\right)^2$, *where* (u_1, v_1) *is the least solution of* (2.1).

For any positive integer l, let $\mathrm{ord}_2(l)$ denote the order of 2 in the factorization of l.

Lemma 2.3. *If* (2.2) *has solutions* (U, V), *then every solution* (U, V) *of* (2.2) *satisfies* $\mathrm{ord}_2(D_1U^2) = \mathrm{ord}_2(D_1U_1^2)$, *where* (U_1, V_1) *is the least solution of* (2.2).

Proof. By (i) of Lemma 2.2, there exists an odd positive integer m which makes $U\sqrt{D_1} + V\sqrt{D_2} = \left(U_1\sqrt{D_1} + V_1\sqrt{D_2}\right)^m$, whence we get

$$U = U_1 \sum_{i=0}^{(m-1)/2} \binom{m}{2i} \left(D_1U_1^2\right)^{(m-1)/2-i} \left(D_2V_1^2\right)^i. \tag{2.3}$$

Since $D_1U_1^2 - D_2V_1^2 = 1$ implies that $D_1U_1^2$ and $D_2V_1^2$ have opposite parity, we have

$$2 \nmid \sum_{i=0}^{(m-1)/2} \binom{m}{2i} \left(D_1U_1^2\right)^{(m-1)/2-i} \left(D_2V_1^2\right)^i. \tag{2.4}$$

Hence, by (2.3) and (2.4), we get $\mathrm{ord}_2(U) = \mathrm{ord}_2(U_1)$. It implies that $\mathrm{ord}_2(D_1U^2) = \mathrm{ord}_2(D_1U_1^2)$. The lemma is proved. □

Lemma 2.4. *Let* r_n, s_n *be defined as in* (1.5). *Then* $(u, v) = (r_n, s_n)$ $(n = 1, 2, \dots)$ *are all solutions of the equation*

$$u^2 - 2v^2 = 1, \quad u, v \in \mathbb{N}, \tag{2.5}$$

and

$$r_n \equiv \begin{cases} 1 \pmod 8, & \text{if } 2 \mid n, \\ 3 \pmod 8, & \text{if } 2 \nmid n. \end{cases} \tag{2.6}$$

Proof. Since $(u_1, v_1) = (3, 2)$ is the least solution of (2.5), by (i) of Lemma 2.1, we see from (1.5) that $(u, v) = (r_n, s_n)$ $(n = 1, 2, \dots)$ are all solutions of (2.5). By (1.5) we have

$$r_n = \sum_{i=0}^{[n/2]} \binom{m}{2i} 3^{n-2i} \cdot 8^i,$$

where $[n/2]$ is the integer part of $n/2$. It follows that

$$r_n \equiv 3^n \pmod 8,$$

whence we obtain (2.6). The lemma is proved. □

Lemma 2.5. *For any odd positive integer* m, *we have* $r_m = 2R_m^2 + 1$, *where* r_m, R_m *are defined as in* (1.5) *and* (1.6) *respectively.*

Proof. Since $3+2\sqrt{2}=\left(1+\sqrt{2}\right)^2$ and $3-2\sqrt{2}=\left(1-\sqrt{2}\right)^2$, by (1.5) and (1.6), we have

$$\begin{aligned} r_m &= \frac{1}{2}\left(\left(3+2\sqrt{2}\right)^m+\left(3-2\sqrt{2}\right)^m\right)=\frac{1}{2}\left(\left(1+\sqrt{2}\right)^{2m}+\left(1-\sqrt{2}\right)^{2m}\right)\\ &= \frac{1}{2}\left(\left(\left(1+\sqrt{2}\right)^m+\left(1-\sqrt{2}\right)^m\right)^2-2\left(1+\sqrt{2}\right)^m\left(1-\sqrt{2}\right)^m\right)\\ &= \frac{1}{2}\left((2R_m)^2+2\right)=2R_m^2+1. \end{aligned}$$

The lemma is proved. □

Lemma 2.6 ([9])**.** *The equation*

$$2X^2+1=Y^3,\quad X,Y\in\mathbb{N}$$

has no solutions (X,Y)*.*

Lemma 2.7 ([6])**.** *The equation*

$$2X^2+1=Y^q,\quad X,Y\in\mathbb{N},\ q \text{ is an odd prime with } q>3$$

has only the solution $(X,Y,q)=(11,3,5)$*.*

Lemma 2.8 ([1, 2])**.** *The equation*

$$X^4-DY^2=1,\quad X,Y\in\mathbb{N}$$

has solutions (X,Y) *if and only if either* $X^2=u_1$ *or* $X^2=2u_1^2-1$*.*

Lemma 2.9. *The equation*

$$2X^2+1=Y^t,\quad X,Y,t\in\mathbb{N},\ t>2 \tag{2.7}$$

has only the solution $(X,Y,t)=(11,3,5)$*.*

Proof. Let (X,Y,t) be a solution of (2.7), and let q be the largest prime divisor of t. By Lemmas 2.6 and 2.7, (2.7) has only the solution $(X,Y,t)=(11,3,5)$ with $q\geq 3$. Since $t>2$, if $q=2$, then $4\mid t$ and the equation

$$(X')^4-2(Y')^2=1,\quad X',Y'\in\mathbb{N} \tag{2.8}$$

has a solution $(X',Y')=(Y^{t/4},X)$. However, since the least solution of (2.5) is $(u_1,v_1)=(3,2)$, neither $u_1=3$ nor $2u_1^2-1=17$ is a square. By Lemma 2.8, (2.8) has no solutions (X',Y'). Therefore, (2.7) has no solutions (X,Y,t) with $q=2$. The lemma is proved. □

3. Proof of Theorem 1.1

In this section, we assume that (1.7) holds and that (x, y, z) is a solution of (1.4) with $2 \nmid x$ and $x > 1$. Then we have

$$x \geq 3. \tag{3.1}$$

Since $\gcd(a, a^x - 1) = 1$, by (1.4), we get

$$a^x - 1 = df^2, \quad b^y - 1 = adg^2, \quad z = dfg, \quad d, f, g \in \mathbb{N}. \tag{3.2}$$

By the first equality of (3.2), we have

$$\gcd(a, d) = 1. \tag{3.3}$$

Since $2 \mid a$, by (3.1) and the first equality of (3.2), we get $2 \nmid f$ and

$$d \equiv df^2 \equiv a^x - 1 \equiv 0 - 1 \equiv 7 \pmod 8. \tag{3.4}$$

Hence, we see from (3.4) that

$$d \text{ is not a square.} \tag{3.5}$$

On the other hand, substituting (3.4) into the second equality of (3.2), we have

$$b^y \equiv 1 + 7ag^2 \equiv \begin{cases} 1 \pmod 8, & \text{if } a \equiv 0 \pmod 8 \text{ or } 2 \mid g, \\ 7 \pmod 8, & \text{if } a \equiv 2 \pmod 8 \text{ and } 2 \nmid g, \\ 5 \pmod 8, & \text{if } a \equiv 4 \pmod 8 \text{ and } 2 \nmid g, \\ 3 \pmod 8, & \text{if } a \equiv 6 \pmod 8 \text{ and } 2 \nmid g. \end{cases} \tag{3.6}$$

Further, since $b \equiv \pm 3 \pmod 8$, we get

$$b^y \equiv \begin{cases} 1 \pmod 8, & \text{if } 2 \mid y, \\ \pm 3 \pmod 8, & \text{if } 2 \nmid y. \end{cases} \tag{3.7}$$

Therefore, in view of (1.7), comparing (3.6) and (3.7), we obtain

$$2 \mid y. \tag{3.8}$$

We see from (3.8) and the second equality of (3.2) that the equation

$$u^2 - adv^2 = 1, \quad u, v \in \mathbb{N} \tag{3.9}$$

has a solution

$$(u, v) = (b^{y/2}, g). \tag{3.10}$$

By (3.3) and (3.5), ad is a nonsquare positive integer. Hence, applying (i) of Lemma 2.1 to (3.10), there exists a positive integer n' which makes

$$b^{y/2} + g\sqrt{ad} = \left(u_1 + v_1\sqrt{ad}\right)^{n'}, \tag{3.11}$$

where (u_1, v_1) is the least solution of (3.9).

For any positive integer n, let

$$u_n + v_n\sqrt{ad} = \left(u_1 + v_1\sqrt{ad}\right)^n. \tag{3.12}$$

If $2 \mid n'$, then from (3.11) and (3.12) we get $b^{y/2} = u_{n'}$ and, by (ii) of Lemma 2.1, $b \equiv \pm 1 \pmod 8$, which contradicts the assumption. So we get

$$2 \nmid n'. \tag{3.13}$$

Since $2 \nmid x$, we see from the first equality of (3.2) that the equation

$$aU^2 - dV^2 = 1, \quad U, V \in \mathbb{N} \tag{3.14}$$

has a solution

$$(U, V) = \left(a^{(x-1)/2}, f\right). \tag{3.15}$$

Let (U_1, V_1) be the least solution of (3.14). For any odd positive integer m, let

$$U_m\sqrt{a} + V_m\sqrt{d} = \left(U_1\sqrt{a} + V_1\sqrt{d}\right)^m. \tag{3.16}$$

Applying (i) of Lemma 2.2 to (3.15), by (3.16), there exists an odd positive integer m' which makes

$$\left(a^{(x-1)/2}, f\right) = (U_{m'}, V_{m'}). \tag{3.17}$$

Hence, by Lemma 2.3, we get from (3.1) and (3.17) that

$$\mathrm{ord}_2(aU_1^2) = \mathrm{ord}_2(aU_{m'}^2) = \mathrm{ord}_2(a^x) \geq x \geq 3. \tag{3.18}$$

By (ii) of Lemma 2.2, we find from (3.11), (3.13) and (3.16) that

$$\begin{aligned} b^{y/2} + g\sqrt{ad} &= \left(U_1\sqrt{a} + V_1\sqrt{d}\right)^{2n'} = \left(\left(U_1\sqrt{a} + V_1\sqrt{d}\right)^{n'}\right)^2 \\ &= \left(U_{n'}\sqrt{a} + V_{n'}\sqrt{d}\right)^2. \end{aligned} \tag{3.19}$$

Since $aU_{n'}^2 - dV_{n'}^2 = 1$, by (3.19), we have

$$b^{y/2} = aU_{n'}^2 + dV_{n'}^2 = 2aU_{n'}^2 - 1. \tag{3.20}$$

Further, by Lemma 2.3, we have $\mathrm{ord}_2(aU_{n'}^2) = \mathrm{ord}_2(aU_1^2)$. Hence, by (3.18), we get $\mathrm{ord}_2(aU_{n'}^2) \geq 3$ and $aU_{n'}^2 \equiv 0 \pmod 8$. Therefore, by (3.20), we obtain $b^{y/2} \equiv 7 \pmod 8$. But, since $b \equiv \pm 3 \pmod 8$, it is impossible. Thus, the theorem is proved.

4. Proof of Theorem 1.2

In this section, we assume that $a = 2$, $b \equiv \pm 3 \pmod 8$ and (x, y, z) is a solution of (1.4) with $2 \nmid x$. By Theorem 1.1, we have

$$x = 1. \tag{4.1}$$

Since $a = 2$, substituting (4.1) into (3.2), we get

$$d = f = 1 \tag{4.2}$$

and

$$b^y - 1 = 2g^2, \ z = g, \quad g \in \mathbb{N}. \tag{4.3}$$

If $b \equiv 5 \pmod 8$, then from the first equality of (4.3) we get $1 = (-2/b) = (2/b) = -1$, a contradiction, where $(*/b)$ is the Jacobi symbol. Therefore, if $a = 2$ and $b \equiv 5 \pmod 8$, then (1.4) has no solutions (x, y, z) with $2 \nmid x$.

We just need to consider the case $b \equiv 3 \pmod 8$. Applying Lemma 2.9 to the first equality of (4.3), by (4.1) and (4.3), equation (1.4) has only the solution

$$b = 3, \quad (x, y, z) = (1, 5, 11) \tag{4.4}$$

with $y > 2$.

When $y = 2$, by the first equality of (4.3), $(u, v) = (b, g)$ is a solution of (2.5) Since $(u_1, v_1) = (3, 2)$ is the least solution of (2.5), by (i) of Lemma 2.1, we get from (1.5) that

$$(b, g) = (r_{n'}, s_{n'}), \quad n' \in \mathbb{N}. \tag{4.5}$$

Further, since $b \equiv 3 \pmod 8$, by Lemma 2.4, we see from (4.5) that $2 \nmid n'$. Hence, by (4.1), (4.2), (4.3) and (4.5), we obtain

$$b = r_m, \ (x, y, z) = (1, 2, s_m), \quad m \in \mathbb{N}, \ 2 \nmid m. \tag{4.6}$$

When $y = 1$, by (4.1), (4.2) and (4.3), we have

$$b = 2g^2 + 1, \ (x, y, z) = (1, 1, g), \quad g \in \mathbb{N}, \ 2 \nmid g. \tag{4.7}$$

Thus, since $r_1 = 2 \cdot 1^2 + 1 = 3$, the combination of (4.4), (4.6) and (4.7) yields the solutions (i), (ii) and (iii). The theorem is proved.

5. Proof of Theorem 1.3

By Theorem 1.2, we get $N(2, 3) = 3$ immediately. By Lemma 2.5, if $b = 2g^2 + 1$ and $g = R_m$ with $m > 1$, then $b = r_m > 3$. Hence, by Theorem 1.2, we have $N(2, b) = 2$. In addition, if $b = 2g^2 + 1$ with $g \neq R_m$ or $b \neq 2g^2 + 1$, then $N(2, b) = 1$ or 0. The theorem is proved.

6. Proof of Theorems 1.4 and 1.5

Proof of Theorem 1.4. By Theorem 1.1, we may assume that $x = 2x_0$ for some $x_0 \in \mathbb{N}$. In addition, since a is a square, we may write $a = a_0^2$ for some $a_0 \in \mathbb{N}$. Then, by the first equality of (3.2), we get

$$(a_0^{x_0})^4 - df^2 = 1. \tag{6.1}$$

It is clear from (6.1) that

$$d \text{ is not a square.} \tag{6.2}$$

Applying Lemma 2.8 to (6.1), we see that either $a^{x_0} = u_1'$ or $a^{x_0} = 2(u_1')^2 - 1$, where (u_1', v_1') is the least solution of (2.1) with $D = d$. Since $2 \mid a$, we must have

$$a^{x_0} = u_1'. \tag{6.3}$$

On the other hand, we know by $4 \mid a$ and $2 \mid x$ that (3.4) holds, which together with (3.6) and (3.7) yields $2 \mid y$. Since $a = a_0^2$, we see from the second equality of (3.2) that (2.1) with $D = d$ has a solution $(u, v) = (b^{y/2}, a_0 g)$. By (i) of Lemma 2.1 and (6.2), we have

$$(u_n', v_n') = \left(b^{y/2}, a_0 g\right), \quad n \in \mathbb{N}, \tag{6.4}$$

where $u_n' + v_n'\sqrt{d} = \left(u_1' + v_1'\sqrt{d}\right)^n$. If $2 \mid n$, then, by (ii) of Lemma 2.1, $b \equiv \pm 1 \pmod 8$, which contradicts the assumption. If $2 \nmid n$, then, by (iii) of Lemma 2.1, $u_1' \mid u_n'$. However, by (6.3) and (6.4), we have $a \mid b^{y/2}$, which contradicts $2 \mid a$ and $b \equiv \pm 3 \pmod 8$. The theorem is proved. □

Proof of Theorem 1.5. By Theorem 1.4, we have

$$x = 1. \tag{6.5}$$

(i) Substituting $a = 4$ and (6.5) into (3.2), we get

$$d = 3, \ f = 1$$

and

$$b^y - 1 = 12g^2, \ z = 3g, \quad g \in \mathbb{N}. \tag{6.6}$$

Obviously, we have $b \neq 3$. If b has a prime divisor p with $p \equiv 11 \pmod{24}$, then by (6.6) we have

$$-1 = \left(\frac{-1}{p}\right) = \left(\frac{12g^2}{p}\right) = \left(\frac{3}{p}\right) = 1,$$

a contradiction. Thus, (i) is proved.

(ii) Substituting $a = 16$ and (6.5) into (3.2), we get

$$d = 15, \ f = 1$$

and

$$b^y - 1 = 15 \cdot 16g^2, \ z = 15g, \quad g \in \mathbb{N}. \tag{6.7}$$

Obviously, we have $b \notin \{3, 5\}$. If b has a prime divisor p with $p \equiv 11, 43, 59$ or $67 \pmod{120}$, then, by (6.7),

$$-1 = \left(\frac{-1}{p}\right) = \left(\frac{15}{p}\right) = 1,$$

a contradiction. If b has a prime divisor p with $p \equiv 13, 29, 37$ or $101 \pmod{120}$, then, by (6.7),

$$1 = \left(\frac{-1}{p}\right) = \left(\frac{15}{p}\right) = -1,$$

a contradiction. Thus, the theorem is proved. □

Acknowledgements. The authors thank the referee for careful reading and helpful comments.

References

[1] J. H. E. Cohn: *The Diophantine equation* $(a^n - 1)(b^n - 1) = x^2$, Period. Math. Hung. 44 (2002), pp. 169–175, doi: 10.1023/A:1019688312555.

[2] M.-H. Le: *A necessary and sufficient condition for the equation* $x^4 - Dy^2 = 1$ *to have positive integer solutions*, Chinese Sci. Bull. 30 (1984), p. 1698.

[3] M.-H. Le, G. Soydan: *A brief survey on the generalized Lebesgue-Ramanujan-Nagell equation*, Surv. Math. Appl. 15 (2020), pp. 473–523.

[4] L. Li, L. Szalay: *On the exponential Diophantine equation* $(a^n - 1)(b^n - 1) = x^2$, Publ. Math. Debrecen 77 (2010), pp. 465–470, doi: 10.5486/PMD.2010.4697.

[5] L. J. Mordell: *Diophantine equations*, London: Academic Press, 1969.

[6] T. Nagell: *Sur l'impossibilité de quelques equations á deux indéterminées*, Norsk Mat. Forenings Skr. 13 (1923), pp. 65–82.

[7] L. Szalay: *On the Diophantine equation* $(2^n - 1)(3^n - 1) = x^2$, Publ. Math. Debrecen 57 (2000), pp. 1–9, doi: 10.5486/PMD.2000.2069.

[8] R.-Z. Tong: *On the Diophantine equation* $(2^x - 1)(p^y - 1) = 2z^2$, Czech. Math. J. 71 (2021), pp. 689–696, doi: 10.21136/CMJ.2021.0057-20.

[9] R. W. van der Waall: *On the Diophantine equations* $x^2 + x + 1 = 3y^2$, $x^3 - 1 = 2y^2$, $x^3 + 1 = 2y^2$, Simon Stevin 46 (1972/1973), pp. 39–51.

[10] D. T. Walker: *On the Diophantine equation* $mx^2 - ny^2 = \pm 1$, Amer. Math. Monthly 74 (1967), pp. 504–513, doi: 10.1080/00029890.1967.11999992.

Annales Mathematicae et Informaticae
60 (2024) pp. 54–63
DOI: 10.33039/ami.2024.02.003
URL: https://ami.uni-eszterhazy.hu

Customer selection rules in competitive facility location

Boglárka G.-Tóth

University of Szeged
boglarka@inf.szte.hu

Abstract. In a competitive facility location setting, customers have a wide range of options to choose from when deciding which facility to patronize. This decision-making process is influenced by a variety of factors, including price, location, quality of service, reputation, and other amenities.

It is important for businesses to understand the different selection rules that customers use and to carefully select the best approximation to build models when decisions need to be made. It is also important for developing effective marketing and branding strategies. By addressing the needs and preferences of their target audience, companies can increase their chances of attracting and retaining customers, and ultimately gain a competitive advantage in the marketplace.

This study discusses the most important customer selection rules and introduces a general modelling scheme for them. A new hybrid rule is proposed to include many patronising behaviours.

Keywords: competitive location, customer choice rules

1. Introduction

When locating a new facility, one of the most important considerations is whether there are competitors in the market offering the same goods or services. If there are competitors in the area, then the locating firm will have to compete for the market, and the profit that the firm makes will be affected by the decisions of its competitors. Therefore, maximizing profit is a much more difficult problem to solve in the presence of competitors than in a monopolistic scenario.

Knowing how customers split their purchases between existing facilities helps to estimate the market share captured by each facility (see [3–5]). The existing customer selection rules assume that all customers follow the same selection rule,

Submitted: July 31, 2023
Accepted: February 6, 2024
Published online: February 14, 2024

however it is clear that in reality, it is not the case. There could be customers for each selection rule, and those should be taken into account in a competitive facility location problem. The aim of the present paper is thus to build a model where multiple customer selections rules are considered, and the new facility is sought accordingly.

If a chain is planning to locate a new facility, it is also important to know the type of competition it faces. If the characteristics of the competitors are known in advance and assumed to be fixed, static competition is assumed. However, the chain may anticipate that competitors will react by also locating a new facility, leading to Stackelberg-type models. This case is considered as competition with foresight, where the location of the leader is optimized, assuming that the follower also locates optimally. Considering dynamic competition assumes that there is an action-reaction cycle of the competing firms. In such settings, decisions are very difficult and can only be approached from a game theory perspective, thus strategies and equilibrium are sought.

This work deals with static competitive facility location problems, where demand is assumed to be inelastic and concentrated in a finite set of demand points. The attraction function considered is multiform, i.e. the facilities differ in location as well as in other aspects such as floor area, number of counters, parking, product mix, etc. The quality of the facility $j \in J$ as perceived by customer $i \in I$ follows the Multiplicative Competitive Interaction pattern [2] and is thus defined by

$$A_{ij} = \prod_k f_{ijk}^{\alpha_k},$$

where the kth accounted factor is measured in f_{ijk} and has importance α_k.

Now the attraction follows the Huff rule, depending on both the location and the quality of the facilities, being inversely proportional to a modified distance measure and proportional to the quality or other positive factors taken into account. In fact, the attraction (or utility) of facility j for customer i is expressed as

$$u_{ij} = \frac{A_{ij}}{g(d_{ij})},$$

where d_{ij} is the distance between facility j and customer i and $g(\cdot)$ is a non-negative, non-decreasing function that modifies the distance. Prices are not considered as decision variables, but they can be considered as part of the attraction factors that determine the qualities of the facilities. Most of the existing models of this type focus on market share maximisation [13–15], although profit maximisation has also been used in many recent works [7, 8].

A number of different customer selection rules have been presented in the literature and are now reviewed in the next section.

2. Patronizing behaviours

Patronizing behaviour is the way customers choose which facilities to favour based on their utility. First, we review the most relevant choice rules presented in litera-

ture, together with their models, using the notations introduced previously.

2.1. Binary or deterministic choice rule

An often used rule is that customers only travel to the nearest/cheapest facility to make their purchases, as occurs in Hotelling-like models [13]. It was the first rule introduced, and since equivalent products were assumed, only distance played a role. However, this role can also be based on the utility of the facilities, so that the most attractive facility gets all the demand. The market share of facility j can be calculated as

$$ms_j = \sum_{i \in I : u_{ij} > \max_{k \in J} u_{ik}} w_i$$

where w_i is the demand of customer i.

In the case of a tie, a tie-breaking rule is considered, which can be *New oriented*, where the new facility takes all the demand; *Conservative*, where the old facility takes all the demand; and a *Tie rule* can also be considered, where the demand is split between all the tied facilities according to the given rule.

2.2. Probabilistic rule

Another very frequently used rule in retailing is that each customer patronizes all available facilities offering the goods probabilistically, with a probability proportional to her/his attraction to each facility, as in Huff-like models [14, 15].

Using the probabilistic rule, the market share of facility j is

$$ms_j = \sum_{i \in I} w_i \frac{u_{ij}}{\sum_{k \in J} u_{ik}}.$$

The probabilistic rule is used for instance in [7, 11].

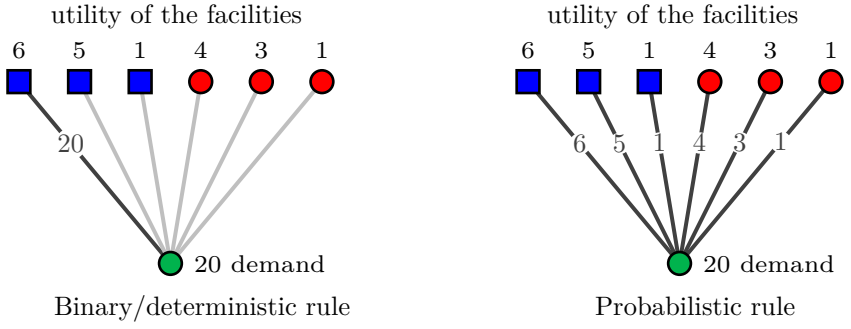


Figure 1. Example for the Binary and Probabilistic selection rules.

As an example of the two classic rules, see Figure 1, where there are two competing chains, denoted by squares and circles, with 3-3 facilities. Our customer's

demand is 20, and the utility of the facilities for this customer is written above each facility. The demand that each facility gains is shown on the edge between them (0 if nothing is written).

2.3. Multi-deterministic or Partially binary rule

This rule is important when there are multiple chains in the market [6]. For a general setting, we can assume that L is the set of chains, and each chain l has its set of facilities C_l. We assume that the set of all facilities $J = \bigcup_{l \in L} C_l$.

The customer splits his demand between all chains, but is served by only the most attractive facility from each chain. The demand is shared probabilistically among the most preferred facilities of each firm. For chain l, its total market share is given by

$$ms_l = \sum_{i \in I} w_i \frac{\max\limits_{j \in C_l} u_{ij}}{\sum\limits_{k \in L} \max\limits_{j \in C_k} u_{ij}}.$$

2.4. Partially probabilistic rule

Using the probabilistic rule, all facilities are patronized, even those with very low utility. This is not realistic, as customers tend to split their demand between facilities with high utility. Therefore, the partially probabilistic rule [9] aims to solve this issue: only the facilities with a minimum level of utility $\underline{u}$ will serve the customer, and the facilities that do not reach the minimum utility are left without demand. Among the facilities with higher utility, the demand is split probabilistically. Thus, the market share for facility j can be written as

$$ms_j = \sum_{i \in I} w_i \frac{U_{ij}}{\sum\limits_{k \in J} U_{ik}} \quad \text{where} \quad U_{ij} = \begin{cases} u_{ij}, & \text{if } u_{ij} > \underline{u}, \\ 0, & \text{otherwise.} \end{cases}$$

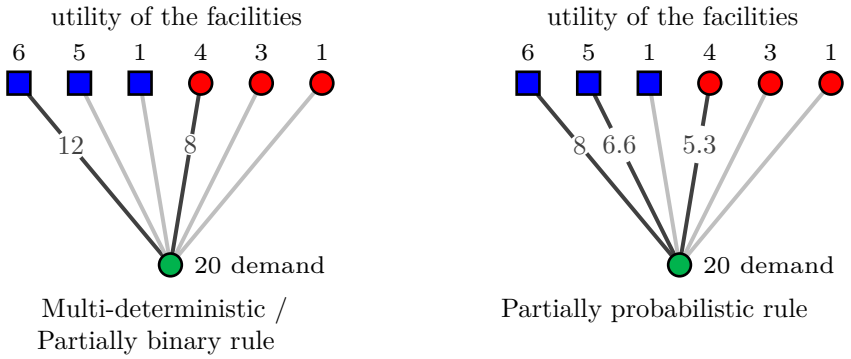


Figure 2. Example for the Multi-deterministic and Partially probabilistic selection rules.

An example of the Multi-deterministic and Partial probabilistic rule is shown in the graphs in Figure 2, where, as before, there are two competing chains, denoted by squares and circles, with 3-3 facilities. Our customer's demand is 20, and the utility of the facilities for this customer is written above each facility. The demand of each facility is written on the edge between them.

2.5. Pareto-Huff selection rule

This rule is based on the assumption that quality cannot be compared with distance. However, those facilities that are dominated by another facility, that is closer and of higher quality, should not be patronized by the customer. Thus, by selecting only the Pareto optimal facilities (minimizing distance and maximizing quality), the dominated facilities can be disregarded. Only the Pareto-optimal facilities, collected in the set P_i, can serve customer i. Among these facilities $j \in P_i \subseteq J$ for customer i, the demand is split probabilistically [10].

The market share of facility j is then

$$ms_j = \begin{cases} \sum_{i \in I} w_i \dfrac{u_{ij}}{\sum_{k \in P_i} u_{ik}} & \text{if } j \in P_i, \\ 0 & \text{otherwise.} \end{cases}$$

2.6. Brand preference

For some products, customers tend to choose by brand rather than by other factors. Therefore, a customer i splits his demand probabilistically among all facilities of his favourite brand $B(i)$. For a facility $j \in C_{B(i)}$, its market share is defined as

$$ms_j = \sum_{i \in I} w_i \frac{u_{ij}}{\sum_{k \in C_{B(i)}} u_{ik}}.$$

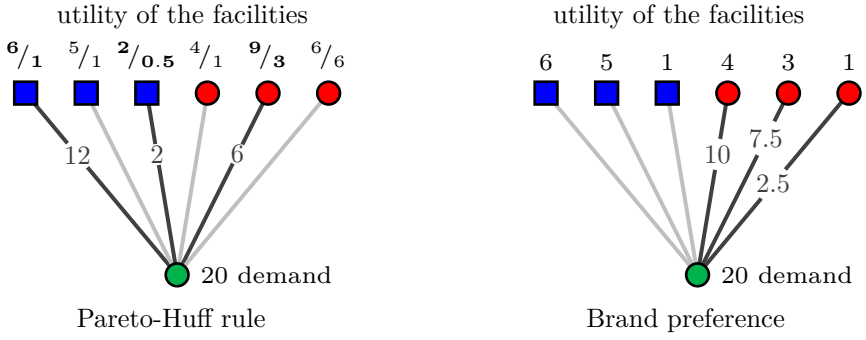


Figure 3. Example for the Pareto-Huff and Brand preference selection rules.

An example of the Pareto-Huff and Brand preference rules is shown in Figure 3, where the settings are the same as in Figures 1–2, except that for the Pareto-Huff case we have written the utilities as quality / distance, and highlighted in bold those in the Pareto-optimal front. For the Brand preference example, our client prefers the red circle chain to the blue square and divides his demand according to the utilities.

2.7. Covering-based choice rule

For some customers, or in some applications, distance is an important consideration and more distant facilities may not be accessible to customers. Thus, whenever a coverage-based customer selection rule is used, only those facilities within a given radius R are probabilistically patronized.

Therefore, the market share of facility j can be formalized as

$$ms_j = \sum_{i \in I} w_i \frac{U_{ij}}{\sum_{k \in J} U_{ik}} \quad \text{where} \quad U_{ij} = \begin{cases} u_{ij}, & \text{if } d_{ij} \leq R, \\ 0, & \text{otherwise,} \end{cases}$$

where it is assumed that all customers have at least one facility within the radius R, otherwise a dummy facility should take all their demand.

2.8. Multinomial Logit model (MNL)

It is a model based on random utility, where utility depends on some measurable characteristics, v_{ij}, but also on some random features, ε_{ij}, so $U_{ij} = v_{ij} + \varepsilon_{ij}$.

If we assume that ε_{ij} are identically independently distributed with the log-Weibull (also known as Gumbel) distribution, which allows us to express the probabilities of customer i to select a facility j as

$$\text{prob}_{ij} = \frac{e^{v_{ij}}}{\sum_{k \in J} e^{v_{ik}}}.$$

This means that in this case we can set the utility as $U_{ij} = e^{v_{ij}}$, and so again, the market share of a facility j looks almost equivalent to the probabilistic rule, i.e.

$$ms_j = \sum_{i \in I} w_i \frac{U_{ij}}{\sum_{k \in J} U_{ik}}.$$

The MNL rule is well studied in many papers, see [12] for some linearization approaches.

The example of the last rule can be the same as for the probabilistic rule in Figure 1, where we assume that the utilities are directly the $e^{v_{ij}}$ values.

Our first goal is the unification of the above rules, if possible, and the construction of a possible solution procedure on the basis of the unified model.

3. Unification of customer selection rules

Looking at the formulas for the different rules, we noticed that most of them are similar to the formula for the probabilistic rule. Therefore, we chose to generalize it to fit each rule. Let us denote by U_{ijr} the utility of a customer i for a facility j and a customer selection rule r, i.e.

$$U_{ijr} = \begin{cases} u_{ij}, & \text{if condition set by rule } r \text{ holds for } u_{ij}, \\ 0, & \text{otherwise.} \end{cases}$$

We can now write the market share of facility j for a given rule r

$$ms_{jr} = \sum_{i \in I} w_i \frac{U_{ijr}}{\sum_{k \in J} U_{ikr}}.$$

There is another way to do that, namely

$$ms_{jr} = \sum_{i \in I_{jr}} w_i \frac{u_{ij}}{\sum_{k \in S_{ir}} u_{ik}}, \tag{3.1}$$

where

$$S_{ir} = \{j \in J \mid \text{condition set by rule } r \text{ holds for } u_{ij}\},$$
$$I_{jr} = \{i \in I \mid \text{condition set by rule } r \text{ holds for } u_{ij}\}.$$

In this setting, the only rule that does not fit in is the MNL, because in that case the utility is changed by the exponential, $U_{ij} = e^{v_{ij}}$. We have therefore omitted this from further discussion.

To complete the unified description of the rules discussed, we need to define the set S_{ir} for each rule r. This can be done as follows:

Binary: $S_{ib} = \arg\max_{j \in J} u_{ij}$

Probabilistic: $S_{ip} = J$

Multi-deterministic: $S_{im} = \{j \in J \mid j = \arg\max_{k \in C_l} u_{ik},\ l \in L\}$

Partially probabilistic: $S_{iP} = \{j \in J \mid u_{ij} \geq \overline{u}\}$

Pareto-Huff: $S_{iH} = P_i$

Brand preference: $S_{iB} = C_{B(i)}$

Covering-based: $S_{ic} = \{j \in J \mid d_{ij} \leq R\}$

Placing the above defined sets S_{ir} in (3.1) gives the specific formula for the customer selection rule r.

Now we are ready to reach the second aim of the paper, to construct a hybrid customer selection rule that combines the ones we have discussed before.

4. Hybrid customer selection rules

Most studies assume that all demand follows a particular choice rule. However, we know that we are all different, and some customers may follow one rule while others apply another. Therefore, at a demand point i, there may be customers who belong to each of the rules mentioned. Suppose we can estimate the proportion of customers who follow each rule. Let p_{ir} the proportion of customers at demand point i following rule r.

Thus, the market share is

$$ms_j = \sum_{i \in I} w_i \sum_{r \in R_{ij}} p_{ir} \frac{u_{ij}}{\sum\limits_{k \in S_{ir}} u_{ik}}, \tag{4.1}$$

where R_{ij} is the set of rules for which $j \in S_{ir}$.

In general, (4.1) is highly nonlinear and non-convex as it is composed of such functions. Note that although it is not highlighted in the formula, S_{ir} and R_{ij} are sets that depend on the location variables, and in many cases the objective function is not even continuous. However, in different settings it is easier to solve.

For the planar case, where the location of the new facility is continuous, either a geometric or interval branch and bound algorithm can be used, or heuristics, as in [6, 7, 9]. In the network setting, where locating on edges is possible, a special branch and bound method can be designed to solve such problems, see [1, 11] for similar works. These works also show that although the problem become difficult, it is still solvable for medium-size problems.

If there is only a discrete set of choices, but more than one facility is to be located, the problem leads to a Mixed Integer Nonlinear Programming problem, which might be linearized.

The surely tractable case for larger instances is when only one new facility is sought among a discrete set of choices. The simplified model for this case is shown next.

Suppose there is a discrete set of choices for the new facility, $f \in F$, and one new facility is being sought. In such a setting, the model is built using binary variables $x_f, f \in F$, with the value 1 if the location f is chosen, or 0 otherwise. Of course, $\sum_{f \in F} x_f = 1$ have to be added as a constraint, since only one facility is to be located. What makes this case easy is that for a given location f, one can directly compute all utilities and S_{ir} and R_{ij} sets, so the market share at the new facility can be written as

$$\begin{aligned} ms_f &= \sum_{i \in I} w_i \sum_{f \in F} \sum_{r \in R_{if}} p_{ir} \frac{u_{if} x_f}{u_{if} x_f + \sum\limits_{k \in S_{ir}} u_{ik}} \\ &= \sum_{i \in I} w_i \sum_{f \in F} x_f \sum_{r \in R_{if}} p_{ir} \frac{u_{if}}{u_{if} + \sum\limits_{k \in S_{ir}} u_{ik}} \end{aligned}$$

$$= \sum_{i \in I} w_i \sum_{f \in F} \tilde{u}_{if} x_f, \tag{4.2}$$

where the parameters $\tilde{u}_{if}$ are calculated as

$$\tilde{u}_{if} = \sum_{r \in R_{if}} p_{ir} \frac{u_{if}}{u_{if} + \sum_{k \in S_{ir}} u_{ik}} \quad \forall i \in I, f \in F.$$

Note that we can omit x_f from the denominator in (4.2), since the whole fraction is directly zero if $x_f = 0$, and the fraction substituting x_f with 1 otherwise.

Now, with the calculated parameters, the problem becomes a rather easy to solve integer programming problem. The difficulty is rather to estimate the necessary data, that consist of estimating the different quality measures of the facilities, their importance, but also the proportion of customers belonging to each selection rule together with their corresponding details.

5. Summary

We have reviewed the most commonly used customer selection rules from the literature and found that most of them can be written in a similar form to the probabilistic selection rule, however they do not patronize all facilities.

After unifying the patronizing behaviours, we defined a hybrid selection rule, where it is assumed that at each demand point, customers may follow different selection rules. The hybrid rule is non-linear and non-convex, and as such is difficult to handle in general, although not worse than most of the individual rules.

Nevertheless, an integer programming model is given for the case where exactly one facility is to be located and there are discrete choices for the new facility. This model is easy to solve for even large data sets, however it is not trivial to collect all the needed data, as in all the customer selection rules described in the paper. Thus, as future work, it is planned to design a model which needs less amount of data to be estimated or simulated, but also to linearize the general model when more than one facility is to be located.

References

[1] R. Blanquero, E. Carrizosa, B. G.-Tóth: *Maximal Covering Location Problems on networks with regional demand*, Omega 64 (2016), pp. 77–85, ISSN: 0305-0483, DOI: 10.1016/j.omega.2015.11.004, URL: https://www.sciencedirect.com/science/article/pii/S0305048315002443.

[2] L. G. Cooper, M. Nakanishi: *Standardizing variables in multiplicative choice models*, Journal of Consumer Research 10.1 (1983), Publisher: The University of Chicago Press, pp. 96–108, DOI: 10.1086/208948.

[3] T. Drezner, H. Eiselt: *Facility location: applications and theory*, in: ed. by Z. Drezner, H. Hamacher, Section: Consumers in competitive location models, Berlin: Springer-Verlag, 2002, pp. 151–178.

[4] H. Eiselt, G. Laporte, J. Thisse: *Competitive location models: a framework and bibliography*, Transportation Science 27.1 (1993), pp. 44–54, doi: 10.1287/trsc.27.1.44.

[5] H. Eiselt, V. Marianov, T. Drezner: *Competitive location models*, in: Location science, ed. by G. Laporte, S. Nickel, F. Saldanha-da-Gama, Section: 14, Springer, 2015, pp. 365–398, doi: 10.1007/978-3-319-13111-5_14.

[6] J. Fernández, B. G.-Tóth, J. Redondo, P. Ortigosa, A. Arrondo: *A planar single-facility competitive location and design problem under the multi-deterministic choice rule*, Computers & Operations Research 78 (2017), pp. 305–315, doi: 10.1016/j.cor.2016.09.019.

[7] J. Fernández, B. Pelegrín, F. Plastria, B. Tóth: *Solving a Huff-like competitive location and design model for profit maximization in the plane*, European Journal of Operational Research 179.3 (2007), pp. 1274–1287, doi: 10.1016/j.ejor.2006.02.005.

[8] J. Fernández, B. Tóth, F. Plastria, B. Pelegrín: *Reconciling franchisor and franchisee: a planar biobjective competitive location and design model*, in: Recent advances in optimization, ed. by A. Seeger, vol. 563, Lectures Notes in Economics and Mathematical Systems, Berlin: Springer-Verlag, 2006, pp. 375–398, doi: 10.1007/3-540-28258-0_22.

[9] J. Fernández, B. G.-Tóth, J. L. Redondo, P. M. Ortigosa: *The probabilistic customer's choice rule with a threshold attraction value: Effect on the location of competitive facilities in the plane*, Computers & Operations Research 101 (2019), pp. 234–249, issn: 0305-0548, doi: 10.1016/j.cor.2018.08.001.

[10] P. Fernández, B. Pelegrín, A. Lančinskas, J. Žilinskas: *Exact and heuristic solutions of a discrete competitive location model with Pareto-Huff customer choice rule*, Journal of Computational and Applied Mathematics 385 (2021), Publisher: Elsevier, p. 113200, doi: 10.1016/j.cam.2020.113200.

[11] B. G.-Tóth, K. Kovács: *Solving a Huff-like Stackelberg location problem on networks*, Journal of Global Optimization 64.2 (2016), pp. 233–257, doi: 10.1007/s10898-015-0368-2.

[12] K. Haase, S. Müller: *A comparison of linear reformulations for multinomial logit choice probabilities in facility location models*, European Journal of Operational Research 232.3 (2014), pp. 689–691, issn: 0377-2217, doi: 10.1016/j.ejor.2013.08.009, url: https://www.sciencedirect.com/science/article/pii/S0377221713006747.

[13] H. Hotelling: *Stability in competition*, Economic Journal 39 (1929), pp. 41–57, doi: 10.2307/2224214.

[14] D. Huff: *Defining and estimating a trading area*, Journal of Marketing 28.3 (1964), pp. 34–38, doi: 10.1177/002224296402800307.

[15] D. Serra, H. Eiselt, G. Laporte, C. ReVelle: *Market capture models under various customer choice rules*, Environment and Planning B 26.5 (1999), pp. 141–150, doi: 10.1068/b260741.

Annales Mathematicae et Informaticae
60 (2024) pp. 64–74
DOI: 10.33039/ami.2024.02.006
URL: https://ami.uni-eszterhazy.hu

Testing the selection heuristic of the Accelerated Branch and Bound method

Emília Heinc[ab], **Balázs Bánhelyi**[ab]

[a]University of Szeged, Institute of Informatics
heincze,banhelyi@inf.u-szeged.hu

[b]University of Győr, Vehicle Industry Research Center

Abstract. This article examined the issue of selection heuristics for the ABB algorithm, a branch-and-bound method for determining the optimal solution structure in P-graphs. Previous studies have not investigated the possible effects of different heuristics on the running time of the ABB algorithm. In this study, we represent the results of applying three basic heuristics in randomly generated P-graphs. In particular, for P-graphs with matrix patterns, the LIFO heuristic is recommended because it performed the best, while the FIFO heuristic had the slowest running time.

Keywords: Branch and Bound, mixed integer programming, production models

AMS Subject Classification: 90C57

1. Introduction

Process network synthesis (or PNS) is a basic tool developed in the 1990s for chemical process problems [14]. Process synthesis aims to find the optimal sub-structure and configuration of an extensive system of functional units and materials [13]. The method is applied in many fields, such as supply chain optimization [12], energy optimization[1], and vehicle scheduling [3, 6]. The search for the optimal sub-network can be formulated as a MILP problem, where the number of binary variables equals the number of units, meaning the complexity of the problem is NP-hard.

The algorithm that finds the optimal sub-network is an accelerated branch-and-bound method (ABB). This method is very similar to the branch-and-bound method for MILP problems. Dakin [5] proposed depth-first search for MILP. This node selection rule always selects a node from the leaf queue with the maximum

Submitted: July 17, 2023
Accepted: February 6, 2024
Published online: February 21, 2024

depth in its search tree. Depth-first search is the preferred strategy for feasibility-only problems. Another technique is Breadth-first search algorithm. It starts at the root of the tree and examines all nodes at the current depth before moving to nodes at the next depth level. This technique is often used when the algorithm needs to parallelize [2]. Besides these two basic algorithms, there are of course other algorithms when additional information is available, such as a heuristic [11]. There are several summary studies of these algorithms in the literature [4], but nowadays more and more algorithms are appearing that use AI for decision making [15].

The ABB algorithm uses the axioms of synthesis to speed up the running of the algorithm. In previous studies, no heuristic for processing order of the non-closed branches to reduce running time was considered. In this work, we tried the simplest heuristics to create an order between the non-closed branches. We tried the following heuristics to determine the order of the non-closed branches: LIFO, FIFO, and random order. We ran the algorithm over randomly generated examples and compared the results based on the number of times the LP solver was called and the runtime.

2. Definitions

2.1. P-Graphs and feasible solutions

P-Graphs. P-Graph is a basic tool for determining the optimal sub-solution structure and configuration of partial solutions for large systems. The main strength of the method is that it combines combinatorial and graph-theoretic techniques.

P-Graphs consist of pairs (M, O), where M is a finite set of materials, and $O(\in \wp(M) \times \wp(M))$ is a finite set of operating units. The graphs can be described as directed bipartite graphs. The set of vertices are O and M sets, and the directed edges represent the connection between operating units and materials.

Process Network Synthesis (PNS) problems are defined by (P, R, O) triplets, where O is similar to set of operating units previously defined in Subsection 2.1. The following will be satisfied in Process In process network synthesis problems, all materials in M can be classified into three categories: P, products, R, raw materials, and I intermediate materials.

Combinatorially feasible solution. A combinatorially feasible solution or solution structure is a special P-Graph, that is made by the materials and operating units from a process network synthesis problem. A P-Graph(M, O) is a combinatorially feasible solution if and only if it satisfies the following five axioms:

1. All materials from P are in the graph.

2. None of the operating units in the graph produces any materials from R.

3. All operating units represented in the graph are from the set O.

4. Every operating unit in the graph has at least one path leading to a material that belongs to the set P.

5. Every material from set M is either produced or consumed by at least one operating unit belonging to O.

2.2. Basic notations linked to operating units and materials

An operating unit is defined by (α, β), where $\alpha, \beta \in \wp(M)$. Set of consumed materials is defined by $X \subseteq O$ denoted by $\mathrm{mat}^{\mathrm{in}}(X) = \bigcup_{(\alpha,\beta)\in o} \alpha$. Set of produced materials is $\mathrm{mat}^{\mathrm{out}}(X) = \bigcup_{(\alpha,\beta)\in o} \beta$. In both cases, the $X \subseteq O$. An operating units in the configuration execute operations and during one operation, a predefined constant amount of materials is consumed and produced. The operating unit $o \in O$ has a cost value, which is defined by the formula, $\mathrm{cost}(o) = \mathrm{fix\_cost} * y + \mathrm{operational\_cost} * x$, i.e. the weighted sum of $y \in \{0, 1\}$, which is set to 1 if the operating unit o is installed in the P-Graph, and $x \in \mathbb{R}_{\geq 0}$, which defines how many operations are executed. Besides the cost of operating units, the raw materials may also have a cost. It is the cost to buy them if they are not available by default.

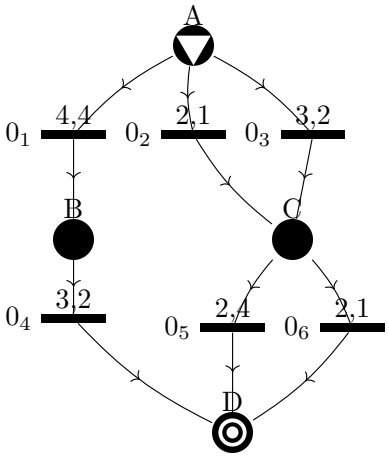
2.3. Decision mappings

It was shown that P-Graphs are basic tools to define the sub-structure of a PNS problem. Another tool, so-called decision mapping[9] equivalent to the P-Graph, can describe sub-substructures. Let $\Delta(m) = \{(\alpha, \beta) \mid (\alpha, \beta) \in O \text{ and } m \in \beta\}$. Let $\delta(m)$ be a subset of $\Delta(m)$, where $m \in M$. δ is also extended to sets, $\delta[m] = \{(X, \delta(X)) \mid X \in m\}$. The complement of decision mapping $\delta(m)$ is $\overline{\delta}(m) = \Delta(m) \setminus \delta(m)$, if $m \in M$. Included operating unit set in the decision mapping $\delta[m]$ will be $\mathrm{op}(\delta[m]) = \bigcup_{X\in m} \delta(X)$. The other notation of the included operating units of decision mapping $\delta[m]$ is O_I. The set of excluded operating units is denoted by $O_E = \mathrm{op}(\overline{\delta}[m])$.

3. Mathematical model of solution structure

Among possible solution structures of a PNS problem, the solution structure that contains all other possible solution structures of the problem is called the maximal solution structure. The algorithm that finds the maximal solution structure is called MSG (Maximal Solution structure Generation) algorithm [7]. Finding a solution structure with minimal total cost for the PNS problem is equivalent to finding an optimal sub-solution structure of the maximal solution structure of the problem. The cost function is composed of the cost of operating units and materials. As mentioned earlier, finding a sub-solution structure forms a MILP problem[10]. The variables come from the operating units. An example of the MILP transformation is shown in Figure 1. The graph in the figure on the left is the maximal structure of the problem. In the object function, the summed cost is

minimized. The first six inequalities define that x is 0 if an operating unit is not installed. B and C are intermediate materials (shown as black circles). Material balance conditions should be defined for intermediate products. These conditions state that at least as much of each intermediate product must be produced as is necessary for the operation of the operating units that use it, otherwise the manufacturing process would come to a standstill. A is the raw material. The aim is to produce the products from the raw materials.



Minimize obj:
$4y_1 + 2y_2 + 3y_3 + 3y_4 + 2y_5 + 2y_6 +$
$4x_1 + 1x_2 + 2x_3 + 3x_4 + 4x_5 + 1x_6$
Subject to:
O_1: $x_1 \leq 1000y_1$
...
O_6: $x_6 \leq 1000y_6$
B: $x_1 - x_4 \geq 0$
C: $x_2 + x_3 - x_5 - x_6 \geq 0$
$x_1, x_2, ..., x_6 \geq 0$
General
$x_1, x_2, ..., x_6$
Binary
$y_1, y_2, ..., y_6$

Figure 1. Transform the maximal structure into a MILP mathematical model.

4. Accelerated Branch and Bound (ABB) method

As it was mentioned in [8], the problem of finding the optimal sub-solution structure is NP-hard. To find the optimal solution, a branch-and-bound technique is used. The algorithm is given in Algorithm 1. The input of the algorithm is the set of materials M and the *PNS* problem whose optimal partial solution structure is to be found, as well as the *datastructure* used in the branching method to keep an ordering between the unsolved branches.

The *ABBD* sub-method is called for each node of the branch-and-bound tree. The parameters of the function are as follows. The first parameter is the set of materials to be produced. Initially, it is set to the set of products. The second parameter is the materials that were already produced. The third parameter is the current decision mapping. The materials that currently need to be produced are in the set p.

The halting condition of the algorithm is when there are no more materials to produce. A lower bound for the optimal value is calculated for each branch since the

value should be minimized. In each branch, a decision is made as to which operating unit set will produce material from p, i.e., a new entry is added to the current decision mapping. When the branching strategy runs, different strategies can be applied as to which node should be extracted next. The strategy is determined by the *datastructure*. The *datastructure* stores elements of p in a predefined order. The *GetElement*() method returns a material, and the sub-branches of the current branch are all possible decision mappings where the material x is produced by a set of operating units. The set of operating units is denoted by c in the algorithm, and it is checked in line 16 that c is consistent with current decision mapping.

The material set p' is the previous state of the materials to be produced, and p will be the state of the new materials to be produced in the sub-branch. It is input materials of the included operating units in the decision mapping, as to operate a unit the input materials have to be produced (except raw materials). The raw materials and currently produced materials (m') do not need to be produced. New materials to be produced ($p \setminus p'$) are added to the data structure, and materials newly left out from p are erased.

The ABBD sub-method is recursively called then for the sub-branch with the modified parameters.

Algorithm 1 ABB algorithm

Input $M, PNS(P, R, O), datastructure$
Global variables $R, \Delta(x), (x \in M), U, currentbest$

1: $U := \infty; currentbest := \infty$
2: $O := MSG(PNS(P, R, O))$
3: ABBD(P, $\emptyset$, $\delta[\emptyset]$)
4: **return**
5: **end procedure**

1: **procedure** ABBD(p, m, $\delta[m]$)
2: $bound = Lower\_Bound(PNS(P, R, 0), O_I, O_E)$
3: **if** $p = \emptyset$ **then** *Halting condition.*
4: **if** $U \geq bound$ **then**
5: $U = bound$;
6: update currentbest;
7: **end if**
8: **return**
9: **end if**
10: **if** $bound \geq U$ **then** *Cutting the branch.*
11: **return**
12: **end if**
13: $x := datastructure.GetElement()$;
14: $C := \wp(\Delta(x)) \setminus \{\emptyset\}$;
15: **for** $\forall c \in C$ **do**
16: **if** $\forall y \in m,\ c \cap \overline{\delta}(y) = \emptyset \& (\Delta(x) \setminus c) \cap \delta(y) = \emptyset$ **then**

```
17:       m' := m ∪ {x};
18:       if S(δ[m']) = ∅ then
19:         Continue;
20:       end if
21:       δ[m'] := δ[m] ∪ {(x, c)};
22:       p' := p;
23:       p := (mat^in(op(δ[m'])) ∪ P) \ (m' ∪ R);
24:       datastructure.Insert(p \ p');
25:       datastructure.Erase(p' \ p);
26:       O_I := op(δ[m']);
27:       O_E := op(δ̄[m']);
28:       ABBD(PNS(P, R, O), p, m', δ[m'])
29:     end if
30:   end for
31:   return
32: end procedure
```

4.1. Heuristics

When selecting a new material to process, various strategies can be used to expand the current decision mapping. In previous studies, nodes were selected according to the alphabetical order of materials. It was not examined, how the selection methods affect the total running time of the ABB algorithm. The following strategies were tried: FIFO, LIFO, and Random Pick (RND). During the LIFO, the algorithm works as a DFS (Depth-first search) algorithm, and for the FIFO case, it is a BFS (Breadth-first search) algorithm.

Consider the P-Graph shown in Figure 2 and assume that no heuristics is specified for the traversal of the graph. The default values are applied when there is no specified amount of materials produced and consumed.

The simplest way is to order the nodes in alphabetical order of the names. In this case, the G, H will be the first materials to produce the p. If G is first produced then there are two ways to produce G by the path of O_3, O_4 or by O_1, O_4. According to the ABB algorithm, in this scenario, the path O_3, O_4 is examined first, and then all possible steps are executed to produce material H. It took extra five steps to find out. The same is repeated for the other two additional branches of producing G, the cases O_4, O_1 and O_4, O_1, O_3. The reason is that branch in both cases cannot be cut by the relaxed lower bound. This is because the upper bound is 12 since all fixed and default costs of operating units are 1. If O_3 and also O_1 produce D, then the relaxed bound is calculated from the costs of the included operating units added to the operating costs of the free operating units. The number of steps evaluated to 15. If in the root node, when deciding whether to produce G or H first, it is decided that H should be produced first, then the algorithm would stop running after 11 steps. If the decision in the first step were completely random

in which orders the products, then the expected value of the run in the case of alphabetical order would be $\frac{1}{2} \cdot 17 + \frac{1}{2} \cdot 11 = 14$.

If the graph is evaluated according to the heuristic of LIFO, the last added material is always processed. In the original case, this is product H. In this heuristic, the algorithm finds the shortest chain of materials and operating units that leads to the raw material A. In the example, this takes five steps. After that, G is considered. If D is produced from raw material B, the process takes two additional steps, and the same is true if the source material is A, and if D is produced by both O_3 and O_4, one additional step is needed to examine this branch. This results in a total of 9 steps to find the optimal solution. This is the best-case scenario. If H and G are swapped, then the same worst-case scenario is executed when the algorithm is run in the order G, H using the alphabetic heuristic. The result is the same running time 17. Overall, the expected value of the LIFO heuristic for this example is $\frac{9+17}{2} = 13$.

In the case FIFO, if the order of processing is G, H first, then after G is produced by O_4, H is examined. Then H is produced by O_7, and after that D will be the selected material. The production of D has three branches, and these branches have the same additional four steps to reach the raw material A. So, in total, there are 15 steps. In the case of H, G, first H is produced by O_7, then G is examined and produced by O_4. After that, the F is produced by O_6, and then the branching is done how to produce the material D. A total of four steps are used to examine how D can be produced. After the three branches were examined, it took additional three steps to be processed. This means that in this case, the algorithm would stop running after 13 steps. The expectation value of the FIFO case heuristic will be $\frac{15+13}{2} = 14$.

This simple example shows that the LIFO strategy works the fastest on average (13 steps), compared to the random and FIFO algorithms (14 steps). Of course, this strategy also depends heavily on which product of the same step in the ABB algorithm is processed first. There are two cases, one with 9 and the other with 17 steps.

5. Results

The algorithm was also tried on randomly generated examples. The operating units and materials are ordered in a matrix pattern. The graph has a *height* number of layers and in each layer, there is a *width* number of operating units or neighbour materials. The first layer consists only of raw materials, and the last layer contains only product materials. The type of neighbor layers alternates between material and operating unit. Connections are made only between two consecutive layers. Every material in the i-th layer except raw materials is produced by at least one operating unit from the $(i-1)$-th layer. Accordingly, they are also consumed by at least one operating unit from the $(i+1)$-th layer. In addition to these connections, other connections with random p-value $(0 \geq p \geq 1)$ are randomly produced between two layers. All connections are generated by a uniform random distribution. The

graph generation pattern can be seen in Figure 4. In our experiment, the width and height were set to 5, and p was set to 0.2.

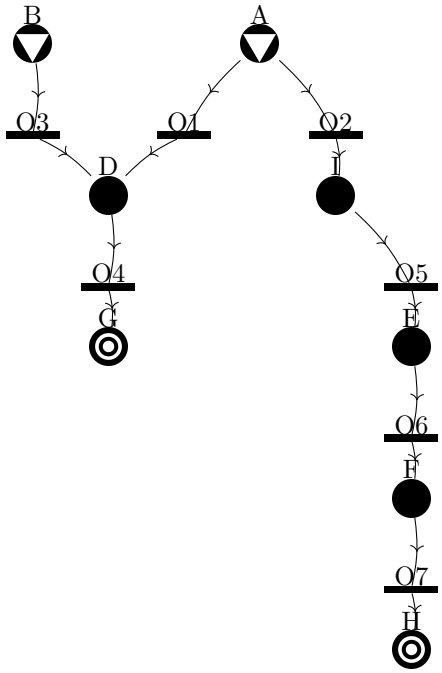


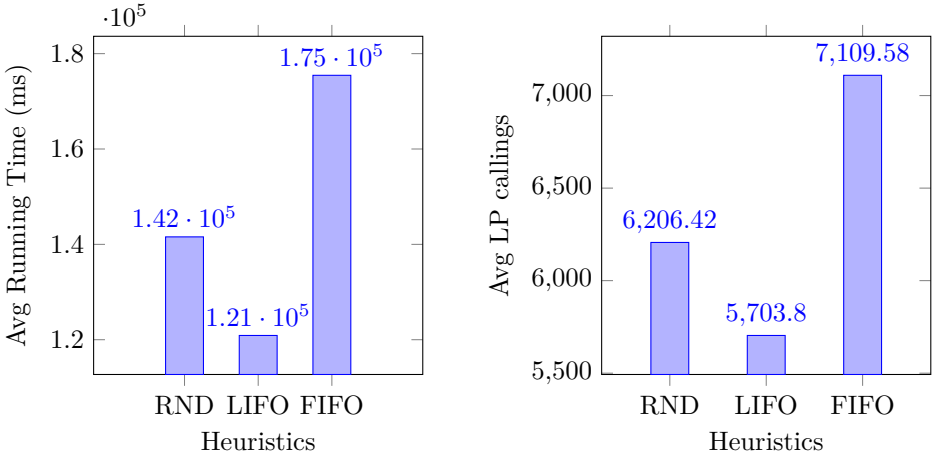
Figure 2. Example P-Graph for different heuristics.

100 examples were generated uniformly at random and the ABB algorithm was run with three different heuristic settings. The random selection heuristic was solved so that the nodes were indexed in random order. In the node selection part, the elements of the materials to be produced were ordered by increasing indexes. The results are grouped by heuristics and the average running time values are collected. Two types of running times were measured. The first is the CPU per clock per second multiplied by 1000. The second is the number of LP solver callings. The lower bound sub-method is used to calculate the relaxation of the MILP transform associated with the current sub-problem. It is particularly relevant to consider the number of solver callings, as well, since the operating units can differ from the simple linear transition between the produced and consumed materials, i.e. they may also be nonlinear or stochastic.

Running time was calculated as CPU clock per second multiplied by 1000. For these randomly generated graphs, the LIFO algorithm performed the best. The FIFO algorithm was the slowest, and the random pick over-performed it as well.

The random pick heuristic itself can be interpreted as choosing the nodes in the graph by alphabetical order where randomly assigning the name of the materials. In the previous cases, generally alphabetical order was used for every case. The

present paper shows that it is worth reconciling the node selection heuristic.



(a) Average running time of heuristics.

(b) Average number of LP callings of heuristics.

Figure 3. Comparison of different heuristics.

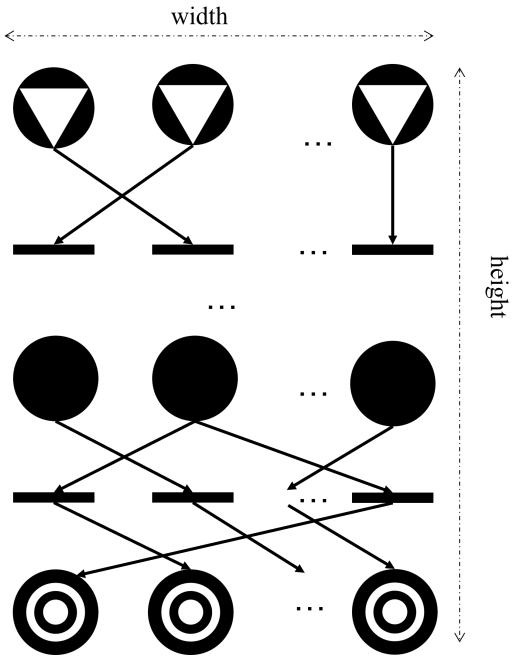


Figure 4. The basic structure of the examined p-graphs.

6. Conclusion and future work

In this article, we have presented the algorithm ABB, which is used in the context of the P-graph. This is very similar to a branch-and-bound method used for standard binary LP problems. We have studied the effect of different traversals of the branches on the running time and the number of LP problems to be solved. In the paper we studied only the 3 best known and simplest versions, but this also shows how differences arise on general graphs. From these results, it can be concluded that it is worth trying other algorithms that are even more computationally demanding. These selection heuristics can be used to achieve even further optimizations that also improve runtime. Such algorithms can be especially interesting when the evaluation of the optimization models is even more time-consuming, for example in nonlinear or stochastic cases.

In the future, we plan to investigate more sophisticated heuristics that use the cost functions and the structural nature of the possible solution structures or other factors.

Acknowledgements. The research presented in this paper was funded by the National Laboratories 2020 Program – Artificial Intelligence Subprogram – Establishment of the "National Artificial Intelligence Laboratory (MILAB) at Széchenyi István University (NKFIH-870-21/2020)" project.

References

[1] K. B. Aviso, J.-Y. Lee, J. C. Dulatre, V. R. Madria, J. Okusa, R. R. Tan: *A P-graph model for multi-period optimization of sustainable energy systems*, Journal of Cleaner Production 161 (2017), pp. 1338–1351, ISSN: 0959-6526, DOI: 10.1016/j.jclepro.2017.06.044.

[2] Z. Bagóczki, B. Bánhelyi: *A parallel interval arithmetic-based reliable computing method on a GPU*, Acta Cybernetica 23.2 (Jan. 2017), pp. 491–501, DOI: 10.14232/actacyb.23.2.2017.4.

[3] M. Barany, B. Bertók, Z. Kovács, F. Friedler, L. T. Fan: *Solving vehicle assignment problems by process-network synthesis to minimize cost and environmental impact of transportation*, Clean Technologies and Environmental Policy 13 (2011), pp. 637–642, DOI: 10.1007/s10098-011-0348-2.

[4] T. Berthold: *Primal Heuristics for Mixed Integer Programs*, in: 2006.

[5] R. J. Dakin: *A tree-search algorithm for mixed integer programming problems*, Comput. J. 8 (1965), pp. 250–255, DOI: 10.1093/comjnl/8.3.250.

[6] Z. Ercsey, A. Nagy, J. Tick, Z. Kovács: *Bus Transport Process Networks with Arbitrary Launching Times*, Acta Polytechnica Hungarica 18 (2021), pp. 125–141, DOI: 10.12700/APH.18.4.2021.4.7.

[7] F. Friedler, K. Tarjan, Y. Huang, L. Fan: *Combinatorial algorithms for process synthesis*, Computers & Chemical Engineering 16 (1992), European Symposium on Computer Aided Process Engineering—1, S313–S320, ISSN: 0098-1354, DOI: 10.1016/S0098-1354(09)80037-9.

[8] F. Friedler, J. B. Varga, E. Fehér, L. T. Fan: *Combinatorially Accelerated Branch-and-Bound Method for Solving the MIP Model of Process Network Synthesis*, in: State of the Art in Global Optimization: Computational Methods and Applications, Boston, MA: Springer US, 1996, pp. 609–626, ISBN: 978-1-4613-3437-8, DOI: 10.1007/978-1-4613-3437-8_35.

[9] F. Friedler, J. Varga, L. Fan: *Decision-mapping: A tool for consistent and complete decisions in process synthesis*, Chemical Engineering Science 50.11 (1995), pp. 1755–1768, ISSN: 0009-2509, DOI: 10.1016/0009-2509(95)00034-3.

[10] F. Friedler, Á. Orosz, J. P. Losada: *P-Graphs for Process Systems Engineering, Mathematical Models and Algorithms*, London: Springer, 2022, DOI: 10.1007/978-3-030-92216-0.

[11] P. E. Hart, N. J. Nilsson, B. Raphael: *A Formal Basis for the Heuristic Determination of Minimum Cost Paths*, IEEE Transactions on Systems Science and Cybernetics 4.2 (1968), pp. 100–107, DOI: 10.1109/TSSC.1968.300136.

[12] H. L. Lam, P. S. Varbanov, J. J. Klemeš: *Optimisation of regional energy supply chains utilising renewables: P-graph approach*, Computers & Chemical Engineering 34.5 (2010), Selected Paper of Symposium ESCAPE 19, June 14-17, 2009, Krakow, Poland, pp. 782–792, ISSN: 0098-1354, DOI: 10.1016/j.compchemeng.2009.11.020.

[13] N. Nishida, G. Stephanopoulos, A. W. Westerberg: *A review of process synthesis*, AIChE Journal 27.3 (1981), pp. 321–351, DOI: 10.1002/aic.690270302.

[14] J. J. Siirola: *Industrial Applications of Chemical Process Synthesis*, Advances in Chemical Engineering 23 (1996), pp. 1–62, ISSN: 0065-2377, DOI: 10.1016/S0065-2377(08)60201-X.

[15] J. Zhang, C. Liu, X. Li, H.-L. Zhen, M. Yuan, Y. Li, J. Yan: *A survey for solving mixed integer programming via machine learning*, Neurocomputing 519 (2023), pp. 205–217, ISSN: 0925-2312, DOI: 10.1016/j.neucom.2022.11.024.

Annales Mathematicae et Informaticae
60 (2024) pp. 75–92
DOI: 10.33039/ami.2024.03.004
URL: https://ami.uni-eszterhazy.hu

A secure key authentication scheme for cryptosystems based on DLP in group ring

Sandeep Kumar, Gaurav Mittal, Sunil Kumar

Defence Research and Development Organization, Near Metcalfe House, New Delhi, 110054, India
sandeepkumar.hqr@gov.in
gaurav.mittaltwins@gmail.com
sunilkumar.hqr@gov.in

Abstract. The public keys in a public key cryptosystem need not to be protected for confidentiality, however, it is important to confirm their legality. In this paper, motivated by Meshram et al. (2017), we develop a simple novel key authentication scheme for public key cryptosystems whose security rely on discrete logarithm problem in group ring. The advantage of our novel scheme is that it requires no authority unlike regular certificate based techniques. In our scheme, we consider a pair of secret key and password as the certificate of public key. We show that the security of our scheme relies on discrete logarithm problem in group ring (DLPGR). The DLPGR is an NP problem for which no known quantum algorithm exists that solves it in polynomial time.

Keywords: Authentication Scheme, Discrete Logarithm Problem, Public-key Cryptosystem, Group Ring, Certificate based scheme

AMS Subject Classification: 94A60, 20C05, 20C07

1. Introduction

The public key cryptography successfully tackled the serious issue of distribution of secret keys in symmetric key cryptography (cf. [4, 41, 42]). Moreover, various advanced digital signature schemes and cryptographic primitives have been created through the aid of public key cryptography (see, for example, [2, 3, 11–14, 17, 18, 28, 29, 31]). Typically, in a public key cryptographic scheme, there are two keys

Submitted: November 5, 2023
Accepted: March 20, 2024
Published online: March 23, 2024

related to an entity, namely a public key which is available in open domain and a private key which is available only with the entity. The public keys are kept open in a repository open to all. But this public availability leads to vulnerability against certain active attacks, such as an adversary can replace the actual public key of an entity with a false public key [7]. Consequently, a secure key authentication scheme is required to verify the legality of public keys. In the available literature, various key authentication schemes have been proposed. But for almost all of these schemes, there is a requirement of atleast one authority known as trusted center (TC) or key authentication center (KAC). Over this authority, the whole trust lies, therefore, it must be strong and safe against any external and internal attacks.

It is worth to mention that in a wide range of key authentication schemes, the hold of KAC over secret keys can be categorized in the following situations: (i) KAC is in total control of secret key of an entity; (ii) KAC can create an undetected false certificate, however, it does not possess the secret key of an entity; (iii) if KAC has produced a false certificate without possessing secret key, then it can be shown that KAC can also produce the false certificate [7]. Accordingly, Girault [7] classified the key authentication schemes into the following levels of trust: (a) schemes depending on identity, i.e., ID-based; (b) schemes depending on certificates; (c) schemes depending on public keys that are self-certified. Moreover, Girault suggested a design of self-certified public keys. The improvement of the model proposed by Girault was discussed in [24] by Laih et al. and their scheme was the combination of ID and certificate based schemes.

For cryptosystems whose security depend on discrete logarithm problem, Horng et al. [15] proposed a key authentication scheme. Their scheme requires no KAC, however, the design is similar to certificate based schemes. Any entity can create a certificate of the public key by combining secret key and password through some known function. The password's hash value is calculated and deposited at the server. Zhan et al. [44] shown that the design of Horng et al. was susceptible to an attack based on guessing the password. In order to protect against the password guessing attack, Lee and Wu [26] proposed another key authentication scheme. Further, in 2003, it is shown by Lee et al. [25] that there is a problem of non-repudiation of public key of an entity in the design of Zhan et al. In addition to this, Lee et al. [25] discussed an upgraded key authentication scheme. However, the scheme of Lee et al. has serious security flaws (see, [37, 43, 45]). Meshram et al. [32] presented another key authentication scheme for cryptosystems whose security depend on the problems such as generalized discrete logarithm and integer factorization. We also refer to the references within Meshram et al. [32] for a nice survey on several other key authentication schemes available in the literature.

Due to the availability of various quantum algorithms (see [2]), the hard problems such as discrete logarithm in a finite field and integer factorization problem are breakable on a sufficiently large quantum computer. Therefore, there is an urgent need to incorporate various other hard problems (possibly NP-hard) in designing new cryptographic primitives, for example, shortest vector problem that arises in lattices, decoding a general linear code etc., (see [2]). In this paper, we utilize the

recently discovered hard problem by Hurley et al. [19] in the algebraic structure of group ring [33]. Precisely, our contribution in this paper is as follows: we incorporate discrete logarithm problem in group ring (DLPGR) to design a novel key authentication scheme for public key cryptosystems whose security relies on the hardness of DLPGR.

To this end, we mention some of the available literature in the direction of group ring based cryptography. Rososhek [38] discussed cryptosystems in automorphism groups of group rings of abelian groups. These cryptosystems implicitly depend on the structure of group ring. Hurley et al. [19] discovered several hard problems in group ring that are important from the perspective of cryptography. Inam et al. [20] designed an ElGamal-like cryptosystem that is based on the matrices over group ring. Goel et al. [8] presented an undeniable signature scheme by utilizing group ring. A key exchange protocol by employing matrices over group ring was proposed by Gupta et al. [9]. Mittal et al. [34–36] constructed few encryption schemes using group ring.

In this paper, our main aim is to present a novel key authentication scheme. This scheme is especially for all the public key cryptosystems whose security relies on solving DLPGR. We show that our scheme works in the absence of any authority and its security relies on deducing the solution of DLPGR. This paper is organized as follows. The Section 2 contains some background material upon which we built our new scheme. Our novel key authentication scheme whose security relies on DLPGR is discussed in Section 3. We discuss the security analysis of our scheme in Section 4. The Section 5 involves a comparison analysis of our scheme with the already available key authentication schemes. In Section 6, we discuss an example to show the practicality of our scheme. Finally, the last section draws some concluding remarks.

2. Preliminaries

2.1. Group ring and units

Definition 2.1. Let R be a ring having unity and let G be a group. Let RG be the set of all R-linear combinations of the form

$$u = \sum_{g \in G} r(g)g, \quad r(g) \in R,$$

where the summation runs over finitely many elements of G. In other words, the set RG contains all finite R-linear combinations of the elements of G. Let $\cdot$ be the operation defined on the group G and

$$u_1 = \sum_{g \in G} r(g)g \quad \text{and} \quad u_2 = \sum_{g \in G} r'(g)g,$$

where $r(g), r'(g) \in R$ and $g \in G$. In order to multiply two elements, we write $u_2 = \sum_{h \in G} r(h)h$ for notational convenience. Then we consider the addition $(+)$

and multiplication $(*)$ operations in RG as follows:

$$u_1 + u_2 = \big(\sum_{g\in G} r(g)g\big) + \big(\sum_{g\in G} r'(g)g\big) = \sum_{g\in G}(r(g) + r'(g))g,$$
$$u_1 * u_2 = \big(\sum_{g\in G} r(g)g\big) * \big(\sum_{g\in G} r'(g)g\big) = \sum_{g,h\in G} r(g)r'(h)(g\cdot h) = \sum_{g'\in G} r''(g')g',$$

where

$$r''(g') = \sum_{g\cdot h=g'} r(g)r(h) = \sum_{g\in G} r(g)r(g'\cdot g^{-1}) = \sum_{h\in G} r(g'\cdot h^{-1})r(h).$$

It is straight-forward to see that both the addition and multiplication operations defined above are well-defined. This is because a ring already has addition and multiplication operations. The set RG along with operations $+$ and $*$ is known as group ring.

Example 2.2. Let $R = \mathbb{Z}_5 = \{0, 1, 2, 3, 4\}$ be a ring of 5 elements and let $G = \mathbb{C}_4 = \{e, a, a^2, a^3\}$ be a cyclic group containing 4 elements. Let $u_1 = 2e + a$ and $u_2 = a + 2a^3$ be the elements of the group ring $\mathbb{Z}_5C_4$. Then we have

$$\begin{aligned} u_1 + u_2 &= (2e + a) + (a + 2a^3) = 2e + 2a + 2a^3, \\ u_1 * u_2 &= (2e + a) * (a + 2a^3) = 2e(a + 2a^3) + a(a + 2a^3) \\ &= 2a + 4a^3 + a^2 + 2a^4 = 2e + 2a + a^2 + 4a^3, \end{aligned}$$

where we have used the fact that $a^4 = e$ for $a \in G$.

Definition 2.3. Units: Let $u_1, u_2 \in RG$ be such that

$$u_1 * u_2 = e = u_2 * u_1.$$

Then the element u_2 is inverse of u_1 (or u_1 is inverse of u_2) and we denote $u_2 = u_1^{-1}$. The elements u_1 and u_2 are known as units of the group ring.

Definition 2.4. Order: The order of an element $u \in RG$ is the smallest positive integer k such that $u^k = e$.

Example 2.5. Let $R = \mathbb{Z}_2 = \{0, 1\}$ be a ring of 2 elements and let G be the quaternion group of 8 elements, i.e.,

$$Q_8 = \langle x, y : x^4 = y^4 = e, x^2 = y^2, yx = x^{-1}y\rangle.$$

Let $w = 1 + x + y$. Using Sharma et al. [39], we know that

$$w^{-1} = (1 + x + y)^3.$$

Thus, $w \in \mathbb{Z}_2Q_8$ is a unit of the group ring.

Definition 2.6. Augmentation map: Let RG be a group ring. Then the following map

$$\mathcal{J}\colon RG \to R \quad \text{defined by} \quad \mathcal{J}\big(\sum_{g\in G} r(g)g\big) = \sum_{g\in G} r(g)$$

is known as the augmentation map. Clearly, $\mathcal{J}$ maps any element $w \in RG$ to the sum of all the coefficients $r(g)$ of the elements g of G appearing in w.

Example 2.7. Let $R = \mathbb{Z}_5$ and $G = Q_8$, where Q_8 is the quaternion group of order 8 (same as discussed in Example 2.5). We take

$$w = \sum_{g\in G} r(g)g = 1 + 2x + y \in RG.$$

Then, we observe that $r(1) = r(y) = 1$, $r(x) = 2$ and rest of the coefficients $r(g)$ are zero for $g \in \{xy, x^2y, x^3y, x^2, x^3\}$. So, we have

$$\mathcal{J}(w) = \mathcal{J}(1 + 2x + y) = r(1) + r(x) + r(y) + 0 = 1 + 2 + 1 = 4.$$

Thus, $\mathcal{J}(w) = 4$ for $w = 1 + 2x + y$.

2.2. Discrete logarithm problem in group ring (DLPGR)

Definition 2.8. Let $u_1, u_2 \in RG$ be given by

$$u_1 = \sum_{g\in G} r(g)g, \quad u_2 = \sum_{g\in G} r'(g)g,$$

where $r(g), r'(g) \in R$. Let k be a positive integer such that $u_1 = u_2^k$, i.e.,

$$\sum_{g\in G} r(g)g = \left(\sum_{g\in G} r'(g)g\right)^k.$$

The DLPGR is the problem of deducing k from the known values of u_1 and u_2.

There are several other versions of DLP in groups, for example, generalized DLP, Elliptic curve DLP (see [31, 32]). DLPGR was discovered by Hurley et al. [19] and used by various researchers to produce secure cryptosystems [34, 36]. To this end, we briefly discuss a public key cryptosystem whose security relies on DLPGR.

2.3. Public key cryptosystem based on DLPGR

Let R be a ring and let G be a finite group. Let u be a unit of the group ring RG with inverse u^{-1}. Both u and u^{-1} are open in public domain. It is worth to mention that unlike groups, computation of inverse of an element in a group ring is also, in general, a hard problem (see [19]). But for some instances, it is easy

(see [35] for a nice overview). Let k be a secret integer and $v = u^k$ be the public key. To encrypt a message M, an ephemeral key r needs to be chosen, where r is a positive integer. The ciphertext is as follows

$$\mathcal{C}_1 = (u^{-1})^r, \quad \mathcal{C}_2 = M * (v)^r.$$

It is worth to mention that the encryption mentioned above is not similar to ElGamal as the ciphertext generated by ElGamal's scheme does not involve the computation of inverse of an element. The decryption is straight-forward by using the private key as follows:

$$M = \mathcal{C}_2 * (\mathcal{C}_1)^k.$$

It is easy to see that the security of above scheme depends on DLPGR. Next, we recall the password authentication scheme for multi-user computing systems.

2.4. Password authentication procedure

A password is a series of characters that is anticipated to distinguish between an entity and the system. In a password authentication scheme for multi-user, entity must (i) register with various systems; (ii) must save purposely created various passwords to attain security of high level. While login to the system with his/her identity (ID), an entity needs to enter his/her password (Pw) to help the system in his/her recognition. The system authorize the entity by verifying the pair (ID, Pw). Basically, the system checks whether or not the pair (ID, Pw) belongs to the list of authorized pairs available with the system. Suppose that the list of authorized pairs available with the system is not encrypted. This situation would be extremely insecure. Since any adversary may get access to the system and can easily forge. Consequently, Evans et al. [6] recommended a cryptographic solution of the same that keeps authorized passwords safe from snooping. Furthermore, it was suggested that passwords can be mapped through some cryptographic one-way function to pictures. Therefore, the list of authorized pairs available with the system can then be a list of mapping results (see [21, 22, 46]).

Based on DLPGR, we select a unit u of the group ring RG of large order. Let Pw be the password of the entity. For this password Pw, we use the capacity u^{Pw} as a picture. The benefit of such pictures is that they can be placed openly in the table of passwords since they can only leak the information about the password if DLPGR is solvable. But DLPGR is a hard problem.

3. The novel key authentication scheme

Let R be a ring and let G be a finite group. For an entity j, let his/her password be represented by Pw_j. Let sk_j be the private key of j and the corresponding public key pk_j be

$$pk_j = u^{sk_j},$$

where u is an element of the group ring RG (it may be taken as a unit of large order). We define the notations to be used in our key authentication scheme in Table 1.

Table 1. Notations incorporated in the novel scheme.

Notations	Descriptions
Pw_j	Password of an entity j
sk_j, pk_j	Private and public keys of entity j
$\oplus$	XOR operation
RG	Group ring
u	Unit of group ring
C_j	Certificate of public key pk_j
$\mathrm{E}(\cdot)$	Exponentiation (one-way) function
$\mathcal{J}$	Augmentation map
$\mathcal{H}$	Public hash function

3.1. Registration phase

Our scheme bank on the following assumptions:

(1) Let the one-way exponential function

$$E\colon \mathbb{Z} \to RG \quad \text{where} \quad z \mapsto u^z,$$

where $R = \mathbb{Z}$ or $\mathbb{F}_p$ for some prime p. The function $E(\cdot)$ is open in public domain. It is important to see that by using repeated square and multiply algorithm [31], one can easily compute $E(z)$ for any given z. This one-way exponential function is nothing but same as the function utilized in the cryptosystems that are based on DLPGR.

(2) In order to guard against the password guessing attacks, for the password Pw_j of an entity j, we apply the one-way exponential function E and obtain the encrypted password as $E(\mathrm{Pw}_j \oplus sk_j)$. This encrypted password is saved in the password table.

(3) For saving the storage space, augmentation map $\mathcal{J}$ (see Definition 2.6) and a public hash function $\mathcal{H}$ can be utilized. For this, first apply $\mathcal{J}$ on the multiplication of $E(\mathrm{Pw}_j \oplus sk_j)$, with $pk_j^{\mathcal{J}(pk_j)}$ and then apply $\mathcal{H}$ on the result, i.e., store the picture

$$\mathcal{H}\big(\mathcal{J}(E(\mathrm{Pw}_j \oplus sk_j)\big(pk_j\big)^{\mathcal{J}(pk_j)}))\big)$$

of password Pw_j. We note that $\mathcal{J}(E(\mathrm{Pw}_j \oplus sk_j)\big(pk_j\big)^{\mathcal{J}(pk_j)}))$ is an integer as $R = \mathbb{Z}$ or $\mathbb{Z}_p$. By writing its binary representation, we can compute the above-mentioned hash value by using any of the recently developed state-of-the-art hash functions

such as hash functions based on Gluškov product of automata [5, 10], quantum hash function [27], hash functions based on chaotic maps [40] etc., or conventional hash function such as SHA-3.

(4) The password table can now be openly placed in public domain, since we have stored the picture of passwords or encrypted passwords.

3.2. Certificate generation phase

Since we are working on the self-certified keys, the certificate is generated itself by the entity. There is no need of KAC or TC. The entity j can pair his/her secret key sk_j along with password Pw_j to obtain the certificate

$$\mathcal{C}_j = \mathrm{Pw}_j \oplus sk_j + sk_j \mathcal{J}(pk_j).$$

For verification purpose, both the public key pk_j and certificate $\mathcal{C}_j$ are placed in the network that is open to public.

3.3. Authentication and verification phase

Each entity needs to present the certificate, public key and encrypted password for authentication. The password image can be written as

$$\begin{aligned} E(\mathcal{C}_j) &= u^{\mathrm{Pw}_j \oplus sk_j + sk_j \mathcal{J}(pk_j)} \\ &= u^{\mathrm{Pw}_j \oplus sk_j} u^{sk_j \mathcal{J}(pk_j)} \\ &= E(\mathrm{Pw}_j \oplus sk_j)\big(pk_j\big)^{\mathcal{J}(pk_j)}. \end{aligned} \tag{3.1}$$

If hash function is used on the encrypted password, then we must have

$$\mathcal{H}(\mathcal{J}(E(\mathcal{C}_j))) = \mathcal{H}\big(\mathcal{J}(E(\mathrm{Pw}_j \oplus sk_j)\big(pk_j\big)^{\mathcal{J}(pk_j)}))\big). \tag{3.2}$$

So, whenever an entity t needs to utilize the public key of an entity j, he/she obtains j's certificate $\mathcal{C}_j$ and public key pk_j from the network. Also, t obtains $E(\mathrm{Pw}_j \oplus sk_j)$ or the hash value $\mathcal{H}\big(\mathcal{J}(E(\mathrm{Pw}_j \oplus sk_j)\big(pk_j\big)^{\mathcal{J}(pk_j)}))\big)$ from the password table. Following this, the entity t can ensure the validity of pk_j through equation (3.2) or the following equation:

$$E(\mathcal{C}_j) = E(\mathrm{Pw}_j \oplus sk_j)\big(pk_j\big)^{\mathcal{J}(pk_j)}.$$

If any of these holds, then the entity t gets assurance about the legality of public key pk_j and can use it for the encryption purpose.

4. Security analysis

In the password table, in place of the entity j's password Pw_j, we have stored $E(\mathrm{Pw}_j \oplus sk_j)$. This means that one cannot alter or modify it illegally. However,

an adversary may attempts to illegally create the public key or deduce the private key or speculate the password in our scheme. We show that this would not be possible in our scheme via the following theorem.

Theorem 4.1. *The novel key authentication scheme of this paper can resist the public key based forgery attack.*

Proof. Suppose that an attacker attempts to forge the public key pk_j of an entity j with a wrong key wk_j. In order to certify wk_j as an actual key, the attacker must produce a false certificate $\mathcal{C}'_j$ such that any of the following holds:

$$E(\mathcal{C}'_j) = E(\mathrm{Pw}_j \oplus sk_j)\big(wk_j\big)^{\mathcal{J}(wk_j)}, \tag{4.1}$$
$$\mathcal{H}(\mathcal{J}(E(\mathcal{C}'_j))) = \mathcal{H}\big(\mathcal{J}(E(\mathrm{Pw}_j \oplus sk_j)\big(wk_j\big)^{\mathcal{J}(wk_j)}))\big).$$

From these equations, the adversary can deduce $\mathcal{C}'_j$ by solving any of the following:

$$\mathcal{C}'_j = E^{-1}\Big(E(\mathrm{Pw}_j \oplus sk_j)\big(wk_j\big)^{\mathcal{J}(wk_j)}\Big), \tag{4.2}$$
$$\mathcal{C}'_j = E^{-1}\Big(\mathcal{J}^{-1}\Big(\mathcal{H}^{-1}\Big(\mathcal{H}\big(\mathcal{J}(E(\mathrm{Pw}_j \oplus sk_j)\big(wk_j\big)^{\mathcal{J}(wk_j)}))\big)\Big)\Big). \tag{4.3}$$

Since the attacker cannot modify $E(\mathrm{Pw}_j \oplus sk_j)$ in the table of passwords without possessing the knowledge of entity j's password, the attacker cannot deduce the certificate $\mathcal{C}'_j$ from equation (4.2) without solving DLPGR. Meanwhile, as the cryptographic hash functions, by definition, are pre-image resistant, the attacker cannot deduce the certificate $\mathcal{C}'_j$ from equation (4.3) without solving DLPGR, without inverting $\mathcal{J}$ (which is a many-to-one function) and without inverting the hash function $\mathcal{H}$. Consequently, for the attacker it is not feasible to forge the public-key illegally. □

Next, we talk about the public-key forgery attack model discussed in Lee et al. [25].

Theorem 4.2. *The novel key authentication scheme is secure against the ingenious attack model of Lee et al. [25].*

Proof. Let t be a malicious legal entity and let sk_t be his/her private key and let wk_j be his/her wrong public key. The entity t utilizes his/her private key to sign any record and the signature $\mathcal{C}_t$ must be certified by entity t by using his/her public key pk_t. However, at a later stage, in place of the actual certificate $\mathcal{C}_t$, t may provide a false certificate $\mathcal{C}'_t$ and deny signing the record to infer that wrong public key was utilized in the first place in the following manner:
(1) uses equation (4.1) to compute

$$wk_j^{\mathcal{J}(wk_j)} = E(\mathcal{C}'_j)\big(E(\mathrm{Pw}_j \oplus sk_j)\big)^{-1}.$$

(2) tries to obtain wk_j from $wk_j^{\mathcal{J}(wk_j)}$.
We claim that deducing wk_j from $wk_j^{\mathcal{J}(wk_j)}$ is harder than solving DLPGR. To see this, suppose that $wk_j \in \mathbb{F}_p$. This means $\mathcal{J}(wk_j) = wk_j$. That is to say that we need to deduce wk_j from $wk_j^{wk_j}$. Due to Agnew et al. [1], we know that deducing wk_j from $wk_j^{wk_j}$ is harder than DLP. Further, it is easy to see that DLPGR is hard problem than DLP as the group ring contains the group. Consequently, it follows that deducing wk_j from $wk_j^{\mathcal{J}(wk_j)}$ is harder than solving DLPGR. Thus, the result holds. □

Next, we show that our scheme is secure against the password guessing attack.

Theorem 4.3. *The presented key authentication scheme can withstand the password guessing attack launched by a malicious server.*

Proof. If there is a closed network environment, then it is believed that the servers are trusted. Therefore, it is highly unlikely that any server will initiate a password guessing attack. Consequently, one can assume that password table will not be illegally amended. An example of such an environment is any closed environment in which all the servers are controlled by a single admin. However, in order to strike at the server end, an attacker must have to guess the password Pw_j as well as sk_j in order to deduce $E(\mathrm{Pw}_j \oplus sk_j)$. But this is computationally infeasible as both are randomly chosen and are known only to the entity. Further, even if an attacker guesses the password by some means, e.g., derives it from the certificate

$$\mathcal{C}_j = \mathrm{Pw}_j \oplus sk_j + sk_j\mathcal{J}(pk_j),$$

it is still computationally infeasible to guess the secret key. Therefore, our scheme is safe against any such attack. □

Next, we show that our scheme is secure even if certificate gets intercepted.

Theorem 4.4. *Suppose that the password used in the certificate of public key gets intercepted in the presented scheme. It is still not possible for an adversary to find the private key from the certificate.*

Proof. We suppose that password Pw_j of entity j gets compromised by any means. Then, in order to deduce the secret key, an adversary may try to utilize (i) $E(\mathrm{Pw}_j \oplus sk_j)$, which is available on the server; (ii) certificate

$$\mathcal{C}_j = \mathrm{Pw}_j \oplus sk_j + sk_j\mathcal{J}(pk_j).$$

It is clear that if the adversary somehow deduces the secret key from the known value of $E(\mathrm{Pw}_j \oplus sk_j)$, then he/she has solved DLPGR, i.e., adversary deduced sk_j from the values of $u^{\mathrm{Pw}_j \oplus sk_j}$, where Pw_j is known. But we know that it is computationally infeasible to solve DLPGR.

For the possibility (ii), we assume that the adversary may try to obtain the secret key from the known value of certificate $\mathcal{C}_j$. More precisely, adversary tries to find the couple (r_1, r_2), where

$$r_1 = \mathrm{Pw}_j \oplus sk_j, \quad r_2 = sk_j \mathcal{J}(pk_j).$$

Suppose that the adversary succeeds in deducing the above-mentioned couple. Then, by XORing the compromised password Pw_j with r_1, the private key sk_j can be computed, i.e.,

$$sk_j = \mathrm{Pw}_j \oplus sk_j \oplus \mathrm{Pw}_j.$$

The same can be obtained from r_2 also as $\mathcal{J}(pk_j)$ is public knowledge. However, if Pw_j and sk_j are sufficiently large, then it is computationally infeasible to obtain the secret key from the certificate via brute force. Thus, result. □

Next, we briefly discuss the hardness of DLPGR.

4.1. Hardness of DLPGR

There is no known classical/quantum algorithm that can find the solution of DLP in group rings. However, one can always apply the brute force attack. So, in order to utilize DLPGR in cryptography, we study its brute force complexity.

4.1.1. Brute force attack

Let G be a finite group with order z and

$$u_1 = \sum_{j=1}^{z} r_{g_j} g_j.$$

Let order of u_1 be s, i.e., s is the least positive integer for which $u_1^s = 1$. It is straight-forward to see that on s multiplications of u_1 with itself, one can solve DLPGR discussed in Definition 2.8. We note that on multiplying u_1 with u_1, $O(2z^2)$ multiplications are required that includes z^2 group and z^2 ring multiplications. That is to say that DLPGR can be solved in $O(2z^2 s)$ multiplications. Let

$$R = \mathbb{Z}_p \quad \text{(finite field of order } p\text{)} \quad \text{and} \quad G = \langle g \rangle \quad \text{(cyclic group of order } z\text{)},$$

where p is a k_1-bit number and z is a k_2-bit number. Then one can solve DLPGR in

$$\mathcal{T} = O(z^2 2^\alpha (k_1^2 + k_2^2)) \quad \text{bit operations},$$

where the size of s is α bits. Clearly, if we consider the input size as the order of u_1, then $\mathcal{T}$ is an exponential time.

4.1.2. Collision algorithms

The general effect of any collision algorithm is that it reduces the number of bit operations by square root times the operations required for brute force. Furthermore, it is easy to see that the collision algorithms available to solve DLP in groups can be extended to solve DLPGR. Consequently, the number of bit operations needed to solve DLPGR through collision algorithms can be reduced to

$$\mathcal{T}' = O(z^2 2^{\alpha/2}(k_1^2 + k_2^2)) \quad \text{bit operations.}$$

To the best of authors' knowledge, currently, there is no algorithm that takes lesser than $\mathcal{T}'$ bit operations to solve DLPGR. Therefore, DLPGR is an extremely hard problem. The security provided by DLPGR with the different parameters is provided in the following Table 2.

Table 2. Size of parameters and the security provided by DLPGR.

Parameters Size	security (atleast)
$\|G\| = z \geq 2^7$, $s \geq 2^{225}$	128 bits
$\|G\| = z \geq 2^8$, $s \geq 2^{485}$	256 bits

5. Advantages of our scheme and comparison analysis

The various advantages of our scheme are discussed as follows:

(1) The size of parameters required for DLPGR is considerably smaller than the related hard problems such as integer factorization problem (IFP), DLP, DLP in a subgroup, Elliptic curve discrete logarithm problem (ECDLP), DLP with conjugacy search problem (DLCSP) (cf. [8]). This is shown in Table 3.

Table 3. Parameters sizes and security.

Hard Problem	Parameters sizes and security (bits)
IFP [31]	((3072, primes of size 1536 bits), 128)
DLP [31]	(3072, 128)
DLP in a subgroup (DSA) [31]	(256, 128)
DLCSP [8]	(48, 128)
ECDLP [2]	(256, 128)
DLPGR (Table 2)	$(10, 256)$

(2) The proposed scheme attains the third and top level of security mentioned by Girault [7]. This is because the regular key authentication schemes are insecure due

to the presence of authorities as they can work together in a bad way. However, in our scheme, there is no involvement of any authority. Consequently, there would not be any malicious collaboration among the certifying authorities.

(3) It is known that a public key in an identity-based (ID-based) scheme is the ID of entity. However, in our presented scheme, entity can effortlessly change his/her private key as well as password. Accordingly, on changing his/her private key and (or) password, the entity can also modify the associated data which involves password's picture, public key and the certificate. In addition, the authentication phase can also be performed by the entity himself/herself.

(4) Our scheme can be executed even if the image of password is not produced by using the Hash function $\mathcal{H}$. This is discussed in Subsection 3.1. However, the use of a hash function can considerably reduce the storage requirement.

(5) Suppose that it is possible to compute inverse of an element in a group ring through an oracle and $\mathcal{J}(pk_j) = 1$. Then equation (3.1) implies that

$$pk_j = E(\mathcal{C}_j)E(\mathrm{Pw}_j \oplus sk_j)^{-1}. \tag{5.1}$$

That is one can compute entity's public key via equation (5.1), where $E(\mathrm{Pw}_j \oplus sk_j)$ can be obtained from the table of passwords and $\mathcal{C}_j$ can be considered as self-certified public key. Therefore, the original public keys may be deleted from the public domain, since they are no longer required to store there. As a result, we only need to store the picture of the password and the self-certified public key. Girault [7] discussed that the public key file can be removed from the public domain in the self-certified schemes, provided the cryptographic scheme is non-interactive. So, our scheme is in line with the discussion of Girault. Thus, the storage required for this portion of the scheme is equal to that of ID-based scheme.

Next, we compare our scheme with the already available key schemes in the literature. Basically, we show that the computation cost of our scheme is very much comparable to various other schemes. We refer to Table 4 and Figure 1 to study the comparison analysis with various other schemes such as Hsieh et al. [16], Kumaraswamy et al. [23], Lee et al. [25], Liu et al. [30], Peinado [37], Wu et al. [43], Zhang et al. [45], Meshram et al. [32]. The notations used in Table 4 are as follows:
T_{inv}: Time required in a modular inverse computation
T_{mul}: Time required in a modular multiplication computation
T_{exp}: Time required in a modular exponentiation computation
T_{add}: Time required in a modular addition computation
$T_{\mathcal{H}}$: Time required in hash computation
T_{XOR}: Time required in a XOR function computation
T'_{add}: Time required in an addition computation through map $\mathcal{J}$
T'_{exp}: Time required in exponentiation computation in a group ring
T'_{mul}: Time required in a multiplication computation in a group ring.

It is worth to mention that for the parameters sizes mentioned in Table 3, we have that

$$T'_{\text{exp}} \approx T_{\text{exp}}, \quad T'_{\text{add}} \approx T_{\text{add}}, \quad T'_{\text{mul}} \approx T_{\text{mul}}.$$

6. Example

In this scheme, we study a toy example related to our scheme. All the results are calculated using the software GAP (Groups, Algorithm, Programming). We consider

$$R = \mathbb{Z}_5 = \{0, 1, 2, 3, 4\} \quad \text{and} \quad G = Q_8 = \langle x, y : x^4 = y^4 = e, x^2 = y^2, yx = x^{-1}y \rangle.$$

Registration phase: Let $u = 1 + xy$. Let $sk_j = 2$ be the private key. Then the public key is

$$pk_j = u^2 = 1 + 2xy + y^2.$$

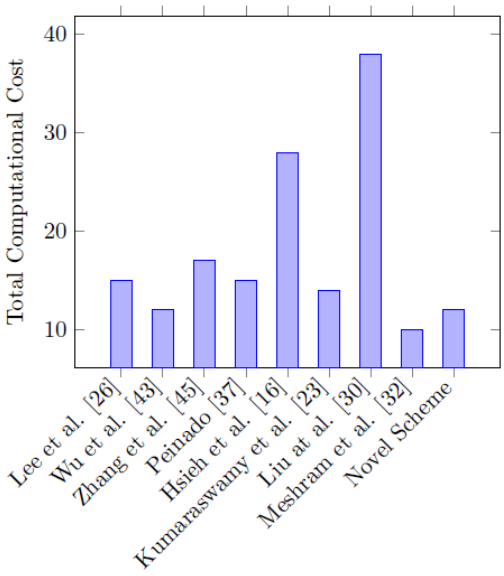


Figure 1. Key Authentication Schemes.

Let $\text{Pw}_j = 12$. Then

$$\text{Pw}_j \oplus sk_j = 14.$$

This means the encrypted password is

$$E(\text{Pw}_j \oplus sk_j) = u^{14}.$$

Also, we have

$$\mathcal{J}(pk_j) = 1 + 2 + 1 = 4 \quad \text{and} \quad (pk_j)^{\mathcal{J}(pk_j)} = (1 + 2xy + y^2)^4.$$

The picture to be stored is

$$\begin{aligned}&\mathcal{H}\big(\mathcal{J}(E(\mathrm{Pw}_j \oplus sk_j)\big(pk_j\big)^{\mathcal{J}(pk_j)}))\big)\\&= \mathcal{H}\big(\mathcal{J}((u^{14} * (1 + 2xy + y^2)^4)\big) = \mathcal{H}\big(\mathcal{J}(1 + y^2 + 2xy)\big) = \mathcal{H}(4). \qquad (6.1)\end{aligned}$$

Table 4. Computation cost in registration and authentication phases.

Authentication scheme	Registration phase	Authentication phase
Lee et al. (2003)	$T_{\text{inv}} + 4T_{\text{mul}} + 3T_{\text{exp}} + 2T_{\text{add}} + T_{\mathcal{H}}$	$2T_{\text{mul}} + 2T_{\text{exp}}$
Wu and Lin (2004)	$T_{\text{inv}} + 2T_{\text{mul}} + 4T_{\text{exp}} + 2T_{\text{add}} + T_{\mathcal{H}}$	$T_{\text{mul}} + T_{\text{exp}}$
Zhang and Kim (2005)	$2T_{\text{mul}} + T_{\text{exp}} + 3T_{\text{add}} + T_{\mathcal{H}}$	$3T_{\text{mul}} + T_{\text{exp}} + 4T_{\text{add}} + 2T_{\mathcal{H}}$
Peinado (2004)	$T_{\text{inv}} + 4T_{\text{mul}} + 3T_{\text{exp}} + 2T_{\text{add}} + T_{\mathcal{H}}$	$2T_{\text{mul}} + 2T_{\text{exp}}$
Hsieh and Leu (2012)	$9T_{\mathcal{H}} + 7T_{\text{XOR}}$	$7T_{\mathcal{H}} + 5T_{\text{XOR}}$
Kumaraswamy et al. (2015)	$3T_{\text{mul}} + 2T_{\text{exp}} + 3T_{\text{add}}$	$2T_{\text{mul}} + 3T_{\text{exp}} + T_{\text{add}}$
Liu et al. (2014b)	$4T_{\mathcal{H}} + 10T_{\text{XOR}} + 2T_{\text{mul}}$	$3T_{\mathcal{H}} + 6T_{\text{XOR}} + 13T_{\text{mul}}$
Meshram et al. (2017)	$2T_{\text{mul}} + T_{\text{exp}} + T_{\text{add}} + T_{\mathcal{H}} + 2T_{\text{XOR}}$	$2T_{\text{mul}} + T_h$
Our Scheme	$2T_{\text{mul}} + 2T_{\text{exp}} + 3T_{\text{add}} + T_{\mathcal{H}} + T_{\text{XOR}}$	$T_{\text{add}} + T_{\text{exp}} + T_h$

Certificate generation phase: The certificate corresponding to key pk_j is

$$\begin{aligned}\mathcal{C}_j &= \mathrm{Pw}_j \oplus sk_j + sk_j\mathcal{J}(pk_j)\\&= (12 + 2) + (2 \times 4) = 22.\end{aligned}$$

Verification phase: For verification, we must have

$$\mathcal{H}(\mathcal{J}(E(\mathcal{C}_j))) = \mathcal{H}\big(\mathcal{J}(E(\mathrm{Pw}_j \oplus sk_j)\big(pk_j\big)^{\mathcal{J}(pk_j)}))\big). \qquad (6.2)$$

We note that

$$\mathcal{H}(\mathcal{J}(E(\mathcal{C}_j))) = \mathcal{H}(\mathcal{J}(u^{22})) = \mathcal{H}(\mathcal{J}(1 + y^2 + 2xy)) = \mathcal{H}(4). \qquad (6.3)$$

Thus, verification is completed because of equations (6.1)–(6.3).

7. Conclusion

We have proposed a simple novel key authentication scheme for public key cryptosystems based on discrete logarithm problem in group ring. In our scheme, the

entity controls the certificate and authentication procedure is based on the table of passwords and there is no requirement of authorities in our scheme. We have carefully discussed the security of our scheme as well as the size of various parameters required in our scheme. Moreover, in order to show the worth of our scheme, we compared it with the several other related schemes. Finally, in order to show the practicality of our scheme, we have discussed a toy example.

Acknowledgements. The authors are thankful to the editor and anonymous reviewer of the manuscript for their valuable comments and suggestions that improved the paper to a great extent.

References

[1] G. Agnew, R. Mullin, S. Vanstone: *Improved digital signature scheme based on discrete exponentiation*, Electronics Letters 26.14 (1990), pp. 1024–1025, doi: 10.1049/el:19900663.

[2] D. Bernstein, J. Buchmann, E. Dahmen: *Post quantum cryptography*, Berlin, Heidelberg: Springer, 2009, doi: 10.1007/978-3-540-88702-7.

[3] C. Chang, Y. Chen, C. Lin: *A data embedding scheme for color images based on genetic algorithm and absolute moment block truncation coding*, Soft Computing 13.4 (2009), pp. 321–331, doi: 10.1007/s00500-008-0332-x.

[4] W. Diffie, M. Hellman: *New directions in cryptography*, IEEE Transactions on Information Theory 22.6 (1976), pp. 644–654, doi: 10.1109/TIT.1976.1055638.

[5] P. Dömösi, G. Horváth: *Hash functions based on Gluškov product of automata*, in: Eleventh Workshop on Non-Classical Models of Automata and Applications (NCMA 2019), Valencia, Spain, 2019, pp. 1–15.

[6] A. Evans, W. Kantrowitz, E. Weiss: *A user authentication system not requiring secrecy in the computer*, Communications of ACM 17.8 (1974), pp. 437–441, doi: 10.1145/361082.361087.

[7] M. Girault: *Self-certified public keys*, in: Proceddings of Eurocrypt'91, Valencia, Spain, 1991, pp. 490–497, doi: 10.1007/3-540-46416-6_42.

[8] N. Goel, I. Gupta, M. Dubey: *Undeniable signature scheme based over group ring*, Applicable Algebra in Engineering, Communication and Computing 27 (2016), pp. 523–535, doi: 10.1007/s00200-016-0293-8.

[9] I. Gupta, A. Pandey, M. Dubey: *A key exchange protocol using matrices over group ring*, Asian European Journal of Mathematics 12.5 (2019), p. 1950075, doi: 10.1142/S179355711950075X.

[10] C. Hannusch, G. Horváth: *Properties of Hash Functions based on Gluškov Product of Automata*, Journal of Automata, Languages and Combinatorics 26.1-2 (2021), pp. 55–65, url: jalc.de/issues/2021/issue_26_1-2/jalc-2021-055-065.php.

[11] D. He, N. Kumar, M. Khan, J. Lee: *Anonymous two-factor authentication for consumer roaming service in global mobility networks*, IEEE Transactions on Consumer Electronics 59.4 (2013), pp. 811–817, url: ieeexplore.ieee.org/document/6689693.

[12] D. He, N. Kumar, H. Shen, J. Lee: *One-to-many authentication for access control in mobile pay-TV systems*, Science China-Information Sciences 59.5 (2016), pp. 1–14, doi: 10.1007/s11432-015-5469-5.

[13] D. He, S. Zeadally, N. Kumar, J. Lee: *Anonymous authentication for wireless body area networks with provable security*, IEEE Systems Journal 11.4 (2016), pp. 2590–2601, url: ieeexplore.ieee.org/document/7458160.

[14] D. He, S. Zeadally, L. Wu: *Certificateless public auditing scheme for cloud-assisted wireless body area networks*, IEEE Systems Journal 12.1 (2015), pp. 64–73, URL: ieeexplore.ieee.org/document/7111218.

[15] G. Horng, C. Yang: *Key authentication scheme for cryptosystems based on discrete logarithms*, Computer Communications 19.9-10 (1996), pp. 848–850, DOI: 10.1016/S0140-3664(96)01112-7.

[16] W. Hsieh, J. Leu: *Exploiting hash functions to intensify the remote user authentication scheme*, Computers and Security 31.6 (2012), pp. 791–798, DOI: 10.1016/j.cose.2012.06.001.

[17] C. Hu, P. Liu, S. Guo: *Public key encryption secure against related-key attacks and key-leakage attacks from extractable hash proofs*, Journal of Ambient Intelligence and Humanized Computing 7.5 (2016), pp. 681–692, DOI: 10.1007/s12652-015-0329-0.

[18] C. Hu, P. Liu, Y. Zhou, S. Guo, Y. Wang, Q. Xu: *Public-key encryption for protecting data in cloud system with intelligent agents against side-channel attacks*, Soft Computing 20.12 (2016), pp. 4919–4932, DOI: 10.1007/s00500-015-1782-6.

[19] B. Hurley, T. Hurley: *Group ring cryptography*, International Journal of Pure and Applied Mathematics 69.1 (2011), pp. 67–86.

[20] S. Inam, R. Ali: *A new ElGamal-like cryptosystem based on matrices over groupring*, Neural Computing and Applications 29 (2018), pp. 1279–1283, DOI: 10.1007/s00521-016-2745-2.

[21] M. Khan, S. Kumari: *An authentication scheme for secure access to healthcare services*, Journal of medical systems 37.4 (2013), pp. 1–12, DOI: 10.1007/s10916-013-9954-3.

[22] M. Khan, S. Kumari: *Cryptanalysis and improvement of "an efficient and secure dynamic ID-based authentication scheme for telecare medical information systems"*, Security and Communication Networks 7.2 (2014), pp. 399–408, DOI: 10.1002/sec.791.

[23] P. Kumaraswamy, C. Rao, V. Janaki, K. Prashanth: *A new key authentication scheme for cryptosystems based on discrete logarithms*, Journal of Innovation in Computer Science and Engineering 5.1 (2015), pp. 42–47, URL: https://www.indianjournals.com/ijor.aspx?target=ijor:jicse&volume=5&issue=1&article=008.

[24] C. Laih, W. Chiou, C. Chang: *Authentication and protection of public keys*, Computers and Security 13.7 (1994), pp. 581–585, DOI: 10.1016/0167-4048(94)90009-4.

[25] C. Lee, M. Hwang, L. Li: *A new key authentication scheme based on discrete logarithms*, Applied Mathematics and Computation 139.2-3 (2003), pp. 343–349, DOI: 10.1016/S0096-3003(02)00192-3.

[26] W. Lee, Y. Wu: *A simple and efficient key authentication scheme*, in: Proceedings of The 18th workshop on combinational mathematics and computational theory, 2001, pp. 70–77.

[27] D. Li, J. Zhang, F. Guo, W. Huang, Q. Wen, H. Chen: *Discrete-time interacting quantum walks and quantum hash schemes*, Quantum information processing 12 (2013), pp. 1501–1513, DOI: 10.1007/s11128-012-0421-8.

[28] B. Liu, J. Bi, A. Vasilakos: *Toward incentivizing anti-spoofing deployment*, IEEE Transactions on Information Forensics and Security 9.3 (2014), pp. 436–450, DOI: 10.1109/TIFS.2013.2296437.

[29] C. Liu, K. Xie, Y. Miao, X. Zha, Z. Feng, J. Lee: *Study on the communication method for chaotic encryption in remote monitoring systems*, Soft Computing 10.3 (2006), pp. 224–229, DOI: 10.1007/s00500-005-0475-y.

[30] T. Liu, Q. Wang, H. Zhu: *A Multi-function Password Mutual Authentication Key Agreement Scheme with privacy preserving*, Journal of Information Hiding and Multimedia Signal Processing 5.2 (2014), pp. 165–178, URL: https://bit.nkust.edu.tw/~jihmsp/2014/vol5/JIH-MSP-2014-02-005.pdf.

[31] A. Menezes, P. Oorschot, S. Vanstone: *Handbook of applied cryptography*, CRC press, 2018, DOI: 10.1201/9780429466335.

[32] C. Meshram, C. Lee, C. Li, C. Chen: *A secure key authentication scheme for cryptosystems based on GDLP and IFP*, Soft Computing 21 (2017), pp. 7285–7291, doi: 10.1007/s00500-016-2440-3.

[33] C. Milies, S. Sehgal: *An introduction to group rings*, vol. 1, Springer Science and Business Media, 2002, url: https://link.springer.com/book/9781402002380.

[34] G. Mittal, S. Kumar, S. Kumar: *Novel public-key cryptosystems based on NTRU and algebraic structure of group rings*, Journal of Information and Optimization Sciences 42.7 (2021), pp. 1507–1521, doi: 10.1080/02522667.2021.1914811.

[35] G. Mittal, S. Kumar, S. Kumar: *A quantum secure ID-based cryptographic encryption based on group rings*, Sādhanā 47.1 (2022), p. 35, doi: 10.1007/s12046-022-01806-5.

[36] G. Mittal, S. Kumar, S. Narain, S. Kumar: *Group ring based public key cryptosystems*, Journal of discrete mathematical sciences and cryptography 25.6 (2022), pp. 1683–1704, doi: 10.1080/09720529.2020.1796868.

[37] A. Peinado: *Cryptanalysis of LHL-key authentication scheme*, Applied mathematics and computation 152.3 (2004), pp. 721–724, doi: 10.1016/S0096-3003(03)00590-3.

[38] S. Rososhek: *Cryptosystems in automorphism groups of group rings of Abelian groups*, Journal of Mathematical Sciences 154.3 (2008), pp. 386–391, doi: 10.1007/s10958-008-9168-2.

[39] S. Rososhek: *The unit group of Z_pQ_8*, Algebras Groups and Geometries 24 (2008), pp. 425–430.

[40] A. V. Tutueva, A. I. Karimov, L. Moysis, C. Volos, D. N. Butusov: *Construction of one-way hash functions with increased key space using adaptive chaotic maps*, Chaos, Solitons and Fractals 141 (2020), p. 110344, doi: 10.1016/j.chaos.2020.110344.

[41] T. Wang, Y. Liu, A. V. Vasilakos: *Survey on channel reciprocity based key establishment techniques for wireless systems*, Wireless Networks 21 (2015), pp. 1835–1846, doi: 10.1007/s11276-014-0841-8.

[42] L. Wei, H. Zhu, Z. Cao, X. Dong, W. Jia, Y. Chen, A. V. Vasilakos: *Security and privacy for storage and computation in cloud computing*, Information sciences 258 (2014), pp. 371–386, doi: 10.1016/j.ins.2013.04.028.

[43] T.-S. Wu, H.-Y. Lin: *Robust key authentication scheme resistant to public key substitution attacks*, Applied mathematics and computation 157.3 (2004), pp. 825–833, doi: 10.1016/j.amc.2003.08.074.

[44] B. Zhan, Z. Li, Y. Yang, Z. Hu: *On the security of HY-key authentication scheme*, Computer Communications 22.8 (1999), pp. 739–741, doi: 10.1016/S0140-3664(99)00032-8.

[45] F. Zhang, K. Kim: *Cryptanalysis of Lee–Hwang–Li's key authentication scheme*, Applied mathematics and computation 161.1 (2005), pp. 101–107, doi: 10.1016/j.amc.2003.12.012.

[46] J. Zhou, Z. Cao, X. Dong, N. Xiong, A. V. Vasilakos: *4S: A secure and privacy-preserving key management scheme for cloud-assisted wireless body area network in m-healthcare social networks*, Information Sciences 314 (2015), pp. 255–276, doi: 10.1016/j.ins.2014.09.003.

Annales Mathematicae et Informaticae
60 (2024) pp. 93–97
DOI: 10.33039/ami.2024.09.001
URL: https://ami.uni-eszterhazy.hu

Catalan numbers which are factoriangular numbers

Florian Luca[a], **Japhet Odjoumani**[b]*, **Alain Togbé**[c]

[a]Mathematics Division, Stellenbosch University, Stellenbosch, South Africa
fluca@sun.ac.za

[b]Institut de Mathématiques et de Sciences Physiques, Université d'Abomey-Calavi, Dangbo BENIN
japhet.odjoumani@imsp-uac.org

[c]Department of Mathematics and Statistics, Purdue University Northwest, 1401 S, U.S. 421, Westville IN 46391 USA
atogbe@pnw.edu

Abstract. In this paper, we prove that the only Catalan numbers or middle binomial coefficients which are factoriangular numbers are 1, 2 and 5.

Keywords: Diophantine equations, Catalan numbers, Factoriangular numbers

AMS Subject Classification: 11B65, 11D72, 11D61

1. Introduction

The Catalan numbers $\{C_n\}_{n\geq 0}$ are given by

$$C_n = \frac{(2n)!}{(n+1)!n!} = \frac{1}{n+1}\binom{2n}{n} \quad \text{for integer} \quad n \geq 0.$$

These numbers are named after the Belgian–French mathematician Eugène Charles Catalan (1814–1894). The Catalan numbers appear naturally when counting various structures. For more information on them, we refer interested readers to the books of Thomas Koshy [3] and Richard Stanley [8]. The first few Catalan numbers are:

$$1,\ 1,\ 2,\ 5,\ 14,\ 42,\ 132,\ 429,\ 1430,\ 4862,\ 16796,\ 58786,\ 208012,\ 742900, \ldots$$

*This work was made possible by a Grant from European Mathematical Society (EMS-Simons).

Submitted: August 15, 2022
Accepted: September 2, 2024
Published online: October 22, 2024

The middle binomial coefficients $\{B_n\}_{n\geq 0}$ are given by $B_n = (n+1)C_n = \binom{2n}{n}$. The first few middle binomial coefficients are

$$1,\ 2,\ 6,\ 20,\ 70,\ 252,\ 924,\ 3432,\ 12870,\ 48620,\ 184756,\ 705432,\ 2704156,\ldots$$

A factoriangular number is the number of the form:

$$Ft_m = m! + \frac{m(m+1)}{2} \quad \text{for integer} \quad m \geq 0.$$

The factoriangular numbers have been studied first by Castillo [1]. The first few factoriangular numbers are:

$$1,\ 2,\ 5,\ 12,\ 34,\ 135,\ 741,\ 5068,\ 40356,\ 362925,\ 3628855,\ 39916866,\ 479001678,\ldots$$

Diophantine equations with factoriangular numbers were studied before. For examples, all Fibonacci numbers, all Pell numbers, all Lucas numbers, all balancing and Lucas-balancing numbers, which are factoriangulars, were found in [2, 4, 5, 7]. In this manuscript, we prove the following result:

Theorem 1.1. *The only Catalan numbers or middle binomial coefficients which are factoriangular numbers are* $1,\ 2$ *and* 5.

2. The proof of the Theorem 1.1

We consider the Diophantine equation

$$Ft_m = B_n, C_n. \tag{2.1}$$

We generated $\{Ft_m : 0 \leq m < 10^5\}$ and $\{B_n, C_n : 0 \leq n < 2.5\cdot 10^5\}$ and intersected these two sets obtaining that their intersection is $\{1, 2, 5\}$. Assume now that there are other solutions to equation (2.1). Since $C_{2.5\cdot 10^5} < B_{2.5\cdot 10^5} < Ft_{10^5}$, it follows that $n \geq 2.5 \cdot 10^5$. Since $B_{2.5\cdot 10^5} > C_{2.5\cdot 10^5} > Ft_{3\cdot 10^4}$, it follows that $m \geq 3 \cdot 10^4$. Now

$$Ft_m = m! + m(m+1)/2 < m^m\left(\frac{2}{m^2} + \frac{1}{m^{m-2}}\right) < m^m,$$

where the above inequality is in fact true for $m \geq 3$, and also

$$Ft_m = m! + m(m+1)/2 > m! > (m/e)^m > m^{0.9m},$$

where the right-most inequality holds for all $m > e^{10} = 22026,46\ldots$. Further,

$$2^{2n} > B_n > C_n > \frac{2^{2n}}{(n+1)(2n+1)} > 2^{1.9n}, \tag{2.2}$$

where the last inequality is equivalent to $2^{0.1n} > (2n+1)(n+1)$ which holds for all $n > 200$. Thus, we have

$$m^{0.9m} < Ft_m = B_n, C_n < 2^{2n}, \quad \text{therefore} \quad \frac{0.9m\log m}{2\log 2} < n,$$

and

$$2^{1.9n} < B_n, C_n = Ft_m < m^m, \quad \text{therefore} \quad n < \left(\frac{2}{1.9}\right)\frac{m\log m}{2\log 2} < \frac{1.1m\log m}{2\log 2}.$$

We record this as a lemma.

Lemma 2.1. *If (n, m) satisfy (2.1) and $n \geq 2.5 \cdot 10^5$, then $m > 3 \cdot 10^4$ and*

$$\frac{0.9m\log m}{2\log 2} < n < \frac{1.1m\log m}{2\log 2}.$$

Lemma 2.2. *If p is a prime in the interval $(\sqrt{n}, \sqrt{2n})$, then $p \mid B_n$.*

Proof. By Kummer's theorem, $p \mid B_n$ if and only if there is at least a carry when adding n to itself in base p. Since $p < \sqrt{2n} < n < p^2$, it follows that $n = ap + b$, where $a, b \in \{0, \ldots, p-1\}$ with $a \neq 0$. If both a, b are at most $(p-1)/2$, then

$$2n = (2a)p + (2b) \leq (p-1)p + (p-1) = p^2 - 1 < p^2,$$

a contradiction. Thus, one of a, b must be in $[(p+1)/2, p-1]$, therefore $p \mid B_n$. □

Let $I = (\sqrt{n}, \sqrt{2n})$. In the equation

$$B_n, C_n = m! + m(m+1)/2,$$

let us consider primes $p \in I$. Such primes divide B_n by Lemma 2.2. At most one of them divides $n+1$. Indeed, if at least two of them say $p_1 < p_2$ divide $n+1$, we would then have that $n+1 \geq p_1 p_2 \geq \sqrt{n}(\sqrt{n}+2) = n + 2\sqrt{n}$, a contradiction. Thus, all primes $p \in I$ divide B_n and with at most one exception they divide C_n as well. If $p \in I$ divides C_n, then since

$$p < \sqrt{2n} < \left(\frac{2.2m\log m}{2\log 2}\right)^{1/2} < m,$$

it follows that p divides also $m!$; hence, $m(m+1)/2$. If there are at least four such primes say $p_1 < p_2 < p_3 < p_4$, then $m(m+1)/2$ is divisible by their product. Thus,

$$m^2 > \frac{m(m+1)}{2} \geq p_1 p_2 p_3 p_4 > \left(\frac{0.9m\log m}{2\log 2}\right)^{4/2} > (0.6m\log m)^2 > 0.3m^2(\log m)^2,$$

which is false. Thus, there can be at most three such prime factors of C_n showing that I contains at most four primes. Hence,

$$\pi(\sqrt{2n}) - \pi(\sqrt{n}) \leq 4.$$

By [6, Corollaries 1 and 2], we have that

$$\frac{x}{\log x} < \pi(x) < \frac{5x}{4\log x} \quad \text{for} \quad x > 114.$$

Applying this with $x \in \{\sqrt{n}, \sqrt{2n}\}$, we have that

$$\pi(\sqrt{2n}) \geq \frac{\sqrt{2n}}{\log(\sqrt{2n})} \quad \text{and} \quad \pi(\sqrt{n}) < \frac{5\sqrt{n}}{4\log(\sqrt{n})}.$$

We thus get that

$$4 \geq \pi(\sqrt{2n}) - \pi(\sqrt{n}) \geq \frac{\sqrt{2n}}{\log(\sqrt{2n})} - \frac{1.25\sqrt{n}}{\log(\sqrt{n})}$$

which gives $n < 80000$, a contradiction. This finishes the proof of Theorem 1.1.

3. Concluding remarks

A similar argument shows that Diophantine equations of the form

$$B_n, C_n = m! \pm P(m), \tag{3.1}$$

where $P(X) \in \mathbb{Q}[X]$ is an integer valued polynomial have only finitely many positive integer solutions n, m. Indeed, the estimates of Lemma 2.1 apply when n is sufficiently large with respect to the degree and height (maximum absolute value of the coefficients) of the polynomial $P(X)$. Lemma 2.2 shows that all primes $p \in I$ divide B_n and all such primes also divide C_n with at most one exception. Since they are also smaller than m, it follows that they divide $P(m)$. Since such primes are in fact larger than $c_1\sqrt{m \log m}$ with some suitable positive constant c_1, it follows that for large m there cannot be more than $2k$ such primes, where k is the degree of $P(X)$. This gives that $\pi(\sqrt{2n}) - \pi(\sqrt{n}) \leq 2k+1$, which implies that n; hence also m, is bounded. The left-hand side of equation (3.1) can be replaced by a binomial coefficient of the form $\binom{an+b}{cn+d}$ with integers $a > c \geq 1$, b, d and the resulting equations still have only finitely many solutions. We do not enter into further details.

Acknowledgements. J. Odjoumani worked on this paper during a visit to the School of Maths of Wits University in August 2022. This author thanks this institution for its hospitality and financial support. He also thanks EMS (European Mathematical Society), which granted him a collaboration grant for his visit to Wits.

References

[1] R. C. Castillo: *On the sum of corresponding factorials and triangular numbers: some preliminary results*, Asia Pacific Journal of Multidisciplinary Research 3.4 (2015), pp. 5–11.

[2] B. Kafle, F. Luca, A. Togbé: *Lucas factoriangular numbers*, Mathematica Bohemica 145.1 (2020), pp. 33–43, doi: 10.21136/MB.2018.0021-18.

[3] T. Koshy: *Catalan Numbers with Applications*, Oxford University Press, Nov. 2008, ISBN: 9780195334548, DOI: 10.1093/acprof:oso/9780195334548.001.0001.

[4] F. Luca, J. Odjoumani, A. Togbé: *Pell factoriangular numbers*, Publications de l'Institut Mathematique 105.119 (2019), pp. 93–100, DOI: 10.2298/PIM1919093L.

[5] S. G. Rayaguru, J. Odjoumani, G. K. Panda: *Factoriangular numbers in balancing and Lucas-balancing sequence*, Boletín de la Sociedad Matemática Mexicana 26 (2020), pp. 865–878, DOI: 10.1007/s40590-020-00303-1.

[6] J. B. Rosser, L. Schoenfeld: *Approximate formulas for some functions of prime numbers*, Illinois Journal of Mathematics 6.1 (1962), pp. 64–94, DOI: 10.1215/ijm/1255631807.

[7] C. A. G. Ruiz, F. Luca: *Fibonacci factoriangular numbers*, Indag. Math. 284 (2017), pp. 796–804, DOI: 10.1016/j.indag.2017.05.002.

[8] R. P. Stanley: *Catalan Numbers*, Cambridge University Press, 2015, DOI: 10.1017/CBO9781139871495.

Annales Mathematicae et Informaticae
60 (2024) pp. 98–107
DOI: 10.33039/ami.2024.08.001
URL: https://ami.uni-eszterhazy.hu

The structure of the unit group of the group algebras $\mathbb{F}_{3^k}D_{6n}$ and $\mathbb{F}_qD_{42}$

Sandeep Malik[a], **R. K. Sharma**[b], **Meena Sahai**[c]

[a]Department of Mathematics, Indian Institute of Technology Delhi, New Delhi, 110016, India
sandeeppmmalik@gmail.com

[b]Department of Mathematics, Indian Institute of Technology Delhi, New Delhi, 110016, India
rksharmaiitd@gmail.com

[c]Department of Mathematics and Astronomy, University of Lucknow, Lucknow, 226007, India
sahai_m@lkouniv.ac.in

Abstract. Let $\mathbb{F}_q$ be a finite field of order $q = p^k$ for some prime p and a positive integer k. In this article, we provide the structure of the unit group $\mathcal{U}(\mathbb{F}_{3^k}D_{6n})$ of the group algebra $\mathbb{F}_{3^k}D_{6n}$ when n is not divisible by 3. Also, a characterization of the unit group $\mathcal{U}(\mathbb{F}_qD_{42})$ of the group algebra $\mathbb{F}_qD_{42}$ has been provided for all the possible cases corresponding to different values of the characteristic p.

Keywords: group algebra, dihedral group, unit group, Wedderburn decomposition

AMS Subject Classification: 16U60, 16S34

1. Introduction

Let $\mathcal{U}(\mathbb{F}_qG)$ be the unit group of the group algebra $\mathbb{F}_qG$ of a group G over a finite field $\mathbb{F}_q$ of order $q = p^k$, for some prime p. For $H \lhd G$, one can extend the canonical homomorphism $\omega\colon G \to G/H$ to form an epimorphism $\omega'\colon \mathbb{F}_qG \to \mathbb{F}_q(G/H)$ which is defined by $\omega'(\sum_{g\in G}\alpha_g g) = \sum_{g\in G}\alpha_g\omega(g)$. Let $\Delta(G,H) = \mathrm{Ker}(\omega')$ and $J(\mathbb{F}_qG)$ be the Jacobson radical of $\mathbb{F}_qG$. The canonical involution $*\colon \mathbb{F}_qG \to \mathbb{F}_qG$ is defined by $(\sum_{g\in G}\alpha_g g)^* = \sum_{g\in G}\alpha_g g^{-1}$. The dihedral group of order $2n$ is represented by $D_{2n} = \langle r, s \mid r^n = s^2 = 1,\ rs = sr^{-1}\rangle$. For basic definitions and results, we

Submitted: December 19, 2023
Accepted: August 13, 2024
Published online: August 17, 2024

refer to [12].

The structure of $\mathcal{U}(\mathbb{F}_qG)$ has been presented for many different groups G in [6, 8, 9, 13–16]. In [7], Kaur and Khan studied $\mathcal{U}(\mathbb{F}_{2^k}D_{2p})$ for prime p. Furthermore, the structure of $\mathcal{U}(\mathbb{F}_{2^k}D_{2n})$ for odd integers n was described by Makhijani and Sharma [10]. In [11], authors have provided characterizations of $\mathcal{U}(\mathbb{Z}D_8)$ and $\mathcal{U}(\mathbb{Z}D_{12})$. Creedon and Gildea [3, 4] provided the structures of $\mathcal{U}(\mathbb{F}_{3^k}D_6)$ and $\mathcal{U}(\mathbb{F}_{2^k}D_8)$ in terms of explicit extensions of elementary cyclic groups. The unitary units of some group algebras have been studied in [1, 2]. The description of $\mathcal{U}(\mathbb{F}_qG)$ for a non semi-simple group algebra $\mathbb{F}_qG$ is quite challenging.

In this paper, we aim to establish the structures of the unit groups $\mathcal{U}(\mathbb{F}_{3^k}D_{6n})$ and $\mathcal{U}(\mathbb{F}_qD_{42})$. Some associated useful results are listed in Section 2. The result related to $\mathcal{U}(\mathbb{F}_{3^k}D_{6n})$ is discussed in Section 3. Section 4 of this article identifies the structure of $\mathcal{U}(\mathbb{F}_qD_{42})$ for characteristic 2 by employing the result in [10]. Additionally, the characterization of $\mathcal{U}(\mathbb{F}_qD_{42})$ is established for the other two non semi-simple cases for $p = 3,\ 7$. Finally, we discuss the semi-simple case for $\mathbb{F}_qD_{42}$ and consequently describe $\mathcal{U}(\mathbb{F}_qD_{42})$ by means of the Wedderburn decomposition.

2. Preliminaries

If $p = 2$, then from [10] we get a generalized result given as follows:

Lemma 2.1 ([10], Theorem 3.2)**.** *Let $q = 2^k$ and n be a positive odd integer. Then,*

$$\mathcal{U}(\mathbb{F}_qD_{2n}) \cong C_2^k \times C_{q-1} \times \prod_{d|n,d>1} GL(2, \mathbb{F}_{q^{c_d}})^{\frac{\phi(d)}{2c_d}}$$

where ϕ is the Euler totient function,

$$c_d = \begin{cases} \frac{b_d}{2}, & \text{if } b_d \text{ is even and } q^{\frac{b_d}{2}} \equiv -1 \bmod d; \\ b_d, & \text{otherwise} \end{cases}$$

and b_d is the multiplicative order of q under mod d.

We recall a useful result from [12, Proposition 3.6.11] to determine the Wedderburn decomposition of semi-simple group algebras which states that if $\mathbb{F}_qG$ is semi-simple, then

$$\mathbb{F}_qG \cong \mathbb{F}_q(G/G') \bigoplus \Delta(G, G')$$

where $\mathbb{F}_q(G/G')$ is the sum of all the commutative simple components of $\mathbb{F}_qG$ and $\Delta(G, G')$ is the sum of all others.

In order to describe the structure of $\mathbb{F}_qG/J(\mathbb{F}_qG)$, we utilize some results given by Ferraz [5]. Let G be a finite group. An element $g \in G$ is said to be a p'-element if the order of g is not divisible by p. Let A be the set of all p'-elements in G and e be the l.c.m. of the orders of all the elements in A. Let ξ be the primitive e-th root of unity over $\mathbb{F}_q$. Define the set

$$B = \{t \mid \xi \to \xi^t \text{ is an automorphism of } \mathbb{F}_q(\xi) \text{ over } \mathbb{F}_q\}.$$

Then $B = \{1, q, \ldots, q^{x-1}\} \bmod e$, where x is the multiplicative order of $q \bmod e$. Let $g \in G$ be a p'-element and β_g be the sum of all conjugates of g. The cyclotomic $\mathbb{F}_q$-class of β_g is defined by

$$S(\beta_g) = \{\beta_{g^t} \mid t \in B\}.$$

We use the above description and the following two results to characterize $\mathcal{U}(\mathbb{F}_q D_{42})$ when $\mathbb{F}_q D_{42}$ is semi-simple.

Lemma 2.2 ([5]). *The number of cyclotomic $\mathbb{F}_q$-classes in G is equal to the number of simple components of $\mathbb{F}_q G/J(\mathbb{F}_q G)$.*

Lemma 2.3 ([5]). *Let t be the number of cyclotmic $\mathbb{F}_q$-classes in G and ξ be the same as defined above. If $S_1, \ldots, S_t$ are the cyclotomic $\mathbb{F}_q$-classes in G and $P_1, \ldots, P_t$ are the simple components of the center of $\mathbb{F}_q G/J(\mathbb{F}_q G)$, then an appropriate ordering of the indices gives $|S_i| = [P_i : \mathbb{F}_q]$.*

3. The structure of $\mathcal{U}(\mathbb{F}_{3^k} D_{6n})$

Theorem 3.1. *Let $\mathbb{F}_q$ be a finite field of order $q = 3^k$ and n be a positive integer not divisible by 3. Then,*

$$\mathcal{U}(\mathbb{F}_q D_{6n}) \cong ((\cdots (C_3^{3nk} \rtimes \underbrace{C_3^k) \rtimes C_3^k) \rtimes \cdots \rtimes C_3^k}_{n \text{ times}}) \rtimes \mathcal{U}(\mathbb{F}_q D_{2n}).$$

Proof. Let $G = D_{6n}$ and $N = \langle r^n \rangle$. Then, $N \lhd G$ and $G/N \cong \langle r^3, s \rangle$. Let $K = \langle r^3, s \rangle$ and define a ring epimorphism $\phi \colon \mathbb{F}_q G \to \mathbb{F}_q K$ by

$$\phi\left(\sum_{j=0}^{n-1}\sum_{i=0}^{2} r^{ni+3j}(x_{i+3j} + x_{i+3j+3n}s)\right) = \sum_{j=0}^{n-1}\sum_{i=0}^{2} r^{3j}(x_{i+3j} + x_{i+3j+3n}s).$$

By restricting the map ϕ, we find a group epimorphism $\phi' \colon \mathcal{U}(\mathbb{F}_q G) \to \mathcal{U}(\mathbb{F}_q K)$. The inclusion map from $\mathbb{F}_q K \to \mathbb{F}_q G$ is a ring monomorphism. Restricting this map, we get a group monomorphism $\theta \colon \mathcal{U}(\mathbb{F}_q K) \to \mathcal{U}(\mathbb{F}_q G)$ given by

$$\theta\left(\sum_{i=0}^{n-1} r^{3i}(z_i + z_{i+n}s)\right) = \sum_{i=0}^{n-1} r^{3i}(z_i + z_{i+n}s).$$

Observe that $\phi' \circ \theta = 1_{\mathcal{U}(\mathbb{F}_q K)}$ and hence, $\mathcal{U}(\mathbb{F}_q G) \cong S \rtimes \mathcal{U}(\mathbb{F}_q K)$ where $S = \mathrm{Ker}(\phi')$.

Let $u = \sum_{j=0}^{n-1}\sum_{i=0}^{2} r^{ni+3j}(x_{i+3j} + x_{i+3j+3n}s) \in S$. Then, $\phi'(u) = 1$. Solving this, we obtain the following equations:

$$\begin{aligned} & x_0 + x_1 + x_2 = 1,\ x_{3m} + x_{3m+1} + x_{3m+2} = 0 && \text{for } m = 1, \ldots, 2n-1. \\ \implies\ & x_0 = 1 - x_1 - x_2,\ x_{3m} = -x_{3m+1} - x_{3m+2} && \text{for } m = 1, \ldots, 2n-1. \end{aligned}$$

In view of this, the set S can be equivalently written as

$$S = \left\{ 1 + \sum_{j=0}^{n-1} \sum_{i=1}^{2} (r^{ni} - 1) r^{3j} (y_{i+2j} + y_{i+2j+2n} s) \mid y_i \in \mathbb{F}_q \right\}.$$

It is trivial to check that S is a non-abelian group and that $S^3 = 1$. Since $q = 3^k$, therefore $|S| = 3^{4nk}$. Assume that $C(r^n)$ is the centralizer of r^n in S. Then,

$$C(r^n) = \{u \in S \mid ur^n = r^n u\}.$$

Let $u = 1 + \sum_{j=0}^{n-1} \sum_{i=1}^{2} (r^{ni} - 1) r^{3j} (y_{i+2j} + y_{i+2j+2n} s) \in C(r^n)$. Then,

$$ur^n - r^n u = \sum_{j=0}^{n-1} \sum_{i=1}^{2} (r^{ni} - 1) r^{3j+2n} y_{i+2j+2n} s - \sum_{j=0}^{n-1} \sum_{i=1}^{2} (r^{ni} - 1) r^{3j+n} y_{i+2j+2n} s.$$

We get,

$$ur^n - r^n u = \hat{r^n} \sum_{i=0}^{n-1} r^{3i} (y_{2i+2n+1} - y_{2i+2n+2}) s.$$

This results in the following condition

$$ur^n - r^n u = 0 \text{ if and only if } y_{2i+2n+1} = y_{2i+2n+2} \quad \text{for } i = 0, 1, \ldots, n-1.$$

In conclusion,

$$C(r^n) = \left\{ 1 + \sum_{j=0}^{n-1} \sum_{i=1}^{2} (r^{ni} - 1) r^{3j} h_{i+2j} + \hat{r^n} \sum_{i=0}^{n-1} r^{3i} h_{i+2n+1} s \mid h_i \in \mathbb{F}_q \right\}.$$

Let us consider some subgroups of S which are given by:

$$N_m = \{1 + a_1 \hat{r^n} + a_2 (r^n + 2r^{2n}) r^{3m} s \mid a_i \in \mathbb{F}_q\} \quad \text{for } m = 0, 1, \ldots, n-1,$$

and $W_0 = C(r^n)$, $W_n = S$,

$$W_m = \left\{ 1 + \sum_{i=1}^{2} (r^{ni} - 1) \left(\sum_{j=0}^{n-1} r^{3j} h_{i+2j} + \sum_{j=0}^{m-1} r^{3j} h_{i+2j+2n} s \right) + \hat{r^n} \sum_{i=m}^{n-1} r^{3i} h_{i+m+2n+1} s \mid h_i \in \mathbb{F}_q \right\} \quad \text{for } m = 1, \ldots, n-1.$$

Clearly N_m and W_m are subgroups of W_{m+1} and $I = N_m \cap W_m = \{1 + a_1 \hat{r^n} \mid a_1 \in \mathbb{F}_q\} \cong C_3^k$, for $m = 0, 1, \ldots, n-1$. Furthermore, N_m is an abelian group and therefore $N_m = I \times Q_m$ for some subgroup Q_m of N_m such that $Q_m \cong C_3^k$, for $m = 0, 1, \ldots, n-1$. We consider the following general elements

$$v_m = 1 + a_1 \hat{r^n} + a_2 (r^n + 2r^{2n}) r^{3m} s \in N_m \quad \text{for } m = 0, \ldots, n-1,$$

$$u_0 = 1 + \sum_{j=0}^{n-1}\sum_{i=1}^{2}(r^{ni}-1)r^{3j}h_{i+2j} + r^{\hat{n}}\sum_{i=0}^{n-1} r^{3i}h_{i+2n+1}s \in W_0,$$

$$u_m = 1 + \sum_{i=1}^{2}(r^{ni}-1)(\sum_{j=0}^{n-1} r^{3j}h_{i+2j} + \sum_{j=0}^{m-1} r^{3j}h_{i+2j+2n}s)$$
$$+ r^{\hat{n}}\sum_{i=m}^{n-1} r^{3i}h_{i+m+2n+1}s \in W_m \quad \text{for } m = 1,\ldots,n-1.$$

Let us define

$$H_1 = \sum_{j=0}^{n-1}\sum_{i=1}^{2}(r^{ni}-1)r^{3j}h_{i+2j},$$

$$H_{2,0} = 0,\ H_{2,m} = \sum_{j=0}^{m-1}\sum_{i=1}^{2}(r^{ni}-1)r^{3j}h_{i+2j+2n} \quad \text{for } m = 1,\ldots,n-1,$$

$$H_{3,m} = r^{\hat{n}}\sum_{i=m}^{n-1} r^{3i}h_{i+m+2n+1} \quad \text{for } m = 0,\ldots,n-1.$$

Then, we can write

$$u_m = 1 + H_1 + H_{2,m}s + H_{3,m}s \in W_m \quad \text{for } m = 0,\ldots,n-1.$$

Since $N_m \subseteq S$, therefore $N_m^3 = 1$. Hence, for $v_m \in N_m$, we have

$$v_m^{-1} = v_m^2 = 1 + 2(a_1 + a_2^2)r^{\hat{n}} + 2a_2(r^n + 2r^{2n})r^{3m}s \quad \text{for } m = 0,\ldots,n-1.$$

The aforementioned information combined with the following steps help to deduce the structure of S.

Step 1: Taking $u_0 \in W_0$ and $v_0 \in N_0$, we have

$$u_0^{v_0} = v_0^{-1}u_0v_0$$
$$= u_0 + a_2(H_1 - H_1^*)(r^n + 2r^{2n})s \in W_0.$$

In conclusion, N_0 normalizes W_0. It is trivial to show that W_0 is abelian and therefore, $W_0 \cong C_3^{3nk}$. Clearly, $W_0 \cap Q_0 = \{1\}$. Hence, $W_1 \cong W_0 \rtimes Q_0 \cong C_3^{3nk} \rtimes C_3^k$.

Step 2: Taking $u_1 \in W_1$ and $v_1 \in N_1$, we have

$$u_1^{v_1} = v_1^{-1}u_1v_1$$
$$= u_1 + a_2(H_1 - H_1^*)(r^n + 2r^{2n})r^3s$$
$$+ a_2(H_{2,1}(r^{2n} + 2r^n)r^{-3} - H_{2,1}^*(r^n + 2r^{2n})r^3) \in W_1.$$

It is concluded that N_1 normalizes W_1. Clearly, $W_1 \cap Q_1 = \{1\}$. Hence, $W_2 \cong W_1 \rtimes Q_1 \cong (C_3^{3nk} \rtimes C_3^k) \rtimes C_3^k$. Consequently, it can be shown that

$$u_m^{v_m} = u_m + a_2(H_1 - H_1^*)(r^n + 2r^{2n})r^{3m}s$$

$$+ a_2(H_{2,m}(r^{2n} + 2r^n)r^{-3m} - H^*_{2,m}(r^n + 2r^{2n})r^{3m}) \in W_m$$

for $m = 0, \ldots, n-1$.

The succeeding steps can be concluded by following a similar process to obtain that N_m normalizes W_m and therefore $W_{m+1} \cong W_m \rtimes Q_m$ for $m = 2, \ldots, n-1$. Finally, we get $W_n \cong W_{n-1} \rtimes Q_{n-1}$, that is

$$S \cong ((\cdots(C_3^{3nk} \rtimes \underbrace{C_3^k) \rtimes C_3^k) \rtimes \cdots \rtimes C_3^k}_{n \text{ times}}).$$

Moreover, since $K \cong D_{2n}$, we get

$$\mathcal{U}(\mathbb{F}_q D_{6n}) \cong ((\cdots(C_3^{3nk} \rtimes \underbrace{C_3^k) \rtimes C_3^k) \rtimes \cdots \rtimes C_3^k}_{n \text{ times}}) \rtimes \mathcal{U}(\mathbb{F}_q D_{2n}). \qquad \square$$

With the help of the above theorem, the characterization problem of unit groups of group algebras of dihedral groups is reduced to the unit groups of the group algebras of smaller dihedral groups.

4. The structure of $\mathcal{U}(\mathbb{F}_q D_{42})$

This section deals with the characterization of $\mathcal{U}(\mathbb{F}_q D_{42})$. The characterization is complete except in characteristic 7, for which we have partial results.

Theorem 4.1. *Let $\mathbb{F}_q$ be a finite field of order $q = p^k$ with characteristic p.*

1. *If* **Char $\mathbb{F}_q = 2$**, *then $\mathcal{U}(\mathbb{F}_q D_{42})$ is isomorphic to*
   (*i*) $C_2^k \times C_{q-1} \times GL(2, \mathbb{F}_q)^{10}$ *if* $k \equiv 0 \bmod 6$.
   (*ii*) $C_2^k \times C_{q-1} \times GL(2, \mathbb{F}_q) \times GL(2, \mathbb{F}_{q^3}) \times GL(2, \mathbb{F}_{q^6})$ *if* $k \equiv \pm 1 \bmod 6$.
   (*iii*) $C_2^k \times C_{q-1} \times GL(2, \mathbb{F}_q) \times GL(2, \mathbb{F}_{q^3})^3$ *if* $k \equiv \pm 2 \bmod 6$.
   (*iv*) $C_2^k \times C_{q-1} \times GL(2, \mathbb{F}_q)^4 \times GL(2, \mathbb{F}_{q^2})^3$ *if* $k \equiv 3 \bmod 6$.

2. *If* **Char $\mathbb{F}_q = 3$**, *then $\mathcal{U}(\mathbb{F}_q D_{42})$ is isomorphic to*
   (*i*) $S \rtimes (C_{q-1}^2 \times GL(2, \mathbb{F}_q)^3)$ *if* $q \equiv \pm 1 \bmod 7$,
   (*ii*) $S \rtimes (C_{q-1}^2 \times GL(2, \mathbb{F}_{q^3}))$ *if* $q \equiv \pm 2 \bmod 7$ *or* $q \equiv \pm 3 \bmod 7$
   where $S \cong (((((((C_3^{21k} \rtimes C_3^k) \rtimes C_3^k) \rtimes C_3^k) \rtimes C_3^k) \rtimes C_3^k) \rtimes C_3^k) \rtimes C_3^k)$.

3. *If* **Char $\mathbb{F}_q = 7$**, *then*

$$\mathcal{U}(\mathbb{F}_q D_{42}) \cong S \rtimes (\mathbb{F}_q^* \times \mathbb{F}_q^* \times GL(2, \mathbb{F}_q))$$

   where S is a non-abelian group such that $|S| = 7^{36k}$ and $S^7 = 1$.

4. *If* **Char** $\mathbb{F}_q \neq 2, 3, 7$, *then* $\mathcal{U}(\mathbb{F}_q D_{42})$ *is isomorphic to*
   (*i*) $C_{q-1}^2 \times GL(2, \mathbb{F}_q)^{10}$ *if* $q \equiv 1, 41 \bmod 42$.
   (*ii*) $C_{q-1}^2 \times GL(2, \mathbb{F}_q) \times GL(2, \mathbb{F}_{q^3})^3$ *if* $q \equiv 5, 17, 25, 37 \bmod 42$.
   (*iii*) $C_{q-1}^2 \times GL(2, \mathbb{F}_q) \times GL(2, \mathbb{F}_{q^3}) \times GL(2, \mathbb{F}_{q^6})$ *if* $q \equiv 11, 19, 23, 31 \bmod 42$.
   (*iv*) $C_{q-1}^2 \times GL(2, \mathbb{F}_q)^4 \times GL(2, \mathbb{F}_{q^2})^3$ *if* $q \equiv 13, 29 \bmod 42$.

Proof. The structure of the unit group $\mathcal{U}(\mathbb{F}_q D_{42})$ differs based on the values of the characterstic p.

1. Char $\mathbb{F}_q = 2$: The structure of $\mathcal{U}(\mathbb{F}_q D_{2n})$ for $q = 2^k$ and an odd integer n has been given by the formula in Lemma 2.1, which depends on the value of q as well. In this article, the structure of $\mathcal{U}(\mathbb{F}_q D_{42})$ is being categorized into four cases based on the values of k upto mod 6. The divisors of 21, which are greater than 1, are 3, 7 and 21. By using Lemma 2.1 for different values of k upto mod 6, we get the following results.
(**a**) If $k \equiv 0 \bmod 6$, then $c_3 = c_7 = c_{21} = 1$ and hence, $\mathcal{U}(\mathbb{F}_q D_{42})$ is isomorphic to

$$C_2^k \times C_{q-1} \times GL(2, \mathbb{F}_q)^{10}.$$

(**b**) If $k \equiv \pm 1 \bmod 6$, then $c_3 = 1,\ c_7 = 3,\ c_{21} = 6$ which gives that $\mathcal{U}(\mathbb{F}_q D_{42})$ is isomorphic to

$$C_2^k \times C_{q-1} \times GL(2, \mathbb{F}_q) \times GL(2, \mathbb{F}_{q^3}) \times GL(2, \mathbb{F}_{q^6}).$$

(**c**) If $k \equiv \pm 2 \bmod 6$, then $c_3 = 1,\ c_7 = c_{21} = 3$ and hence, $\mathcal{U}(\mathbb{F}_q D_{42})$ is isomorphic to

$$C_2^k \times C_{q-1} \times GL(2, \mathbb{F}_q) \times GL(2, \mathbb{F}_{q^3})^3.$$

(**d**) If $k \equiv 3 \bmod 6$, then $c_3 = c_7 = 1,\ c_{21} = 2$ and it can be concluded that $\mathcal{U}(\mathbb{F}_q D_{42})$ is isomorphic to

$$C_2^k \times C_{q-1} \times GL(2, \mathbb{F}_q)^4 \times GL(2, \mathbb{F}_{q^2})^3.$$

2. Char $\mathbb{F}_q = 3$: In particular, using Theorem 3.1 for $n = 7$, we obtain

$$\mathcal{U}(\mathbb{F}_q D_{42}) \cong S \rtimes \mathcal{U}(\mathbb{F}_q D_{14})$$

where

$$S \cong (((((((C_3^{21k} \rtimes C_3^k) \rtimes C_3^k) \rtimes C_3^k) \rtimes C_3^k) \rtimes C_3^k) \rtimes C_3^k) \rtimes C_3^k).$$

Moreover, on the lines of [14, Theorem 4.1], we get

$$\mathcal{U}(\mathbb{F}_q D_{14}) \cong \begin{cases} C_{q-1}^2 \times GL(2, \mathbb{F}_q)^3, & \text{if } q \equiv \pm 1 \bmod 7; \\ C_{q-1}^2 \times GL(2, \mathbb{F}_{q^3}), & \text{if } q \equiv \pm 2 \bmod 7 \text{ or } q \equiv \pm 3 \bmod 7. \end{cases}$$

Hence,

$$\mathcal{U}(\mathbb{F}_qD_{42}) \cong \begin{cases} S \rtimes (C_{q-1}^2 \times GL(2,\mathbb{F}_q)^3), & \text{if } q \equiv \pm 1 \bmod 7; \\ S \rtimes (C_{q-1}^2 \times GL(2,\mathbb{F}_{q^3})), & \text{if } q \equiv \pm 2 \bmod 7 \text{ or } q \equiv \pm 3 \bmod 7 \end{cases}$$

where $S \cong (((((((C_3^{21k} \rtimes C_3^k) \rtimes C_3^k) \rtimes C_3^k) \rtimes C_3^k) \rtimes C_3^k) \rtimes C_3^k) \rtimes C_3^k)$.

3. Char $\mathbb{F}_q = 7$: Let $G = D_{42}$ and $N = \langle r^3 \rangle$. Then, $N \lhd G$ and $G/N \cong \langle r^7, s \rangle \cong D_6$. Let $K = \langle r^7, s \rangle$ and $\phi\colon \mathbb{F}_qG \to \mathbb{F}_qK$ be the ring epimorphism defined by

$$\phi\left(\sum_{j=0}^{2}\sum_{i=0}^{6} r^{3i+7j}(x_{i+7j} + x_{i+7j+21}s)\right) = \sum_{j=0}^{2}\sum_{i=0}^{6} r^{7j}(x_{i+7j} + x_{i+7j+21}s).$$

By restricting the map ϕ, we find a group epimorphism $\phi'\colon \mathcal{U}(\mathbb{F}_qG) \to \mathcal{U}(\mathbb{F}_qK)$. The inclusion map from $\mathbb{F}_qK \to \mathbb{F}_qG$ is a ring monomorphism. A group monomorphism $\theta\colon \mathcal{U}(\mathbb{F}_qK) \to \mathcal{U}(\mathbb{F}_qG)$ is obtained by restricting this inclusion map which is defined by

$$\theta\left(\sum_{i=0}^{2} r^{7i}(z_i + z_{i+3}s)\right) = \sum_{i=0}^{2} r^{7i}(z_i + z_{i+3}s).$$

Observe that $\phi' \circ \theta = 1_{\mathcal{U}(\mathbb{F}_qK)}$ and hence, $\mathcal{U}(\mathbb{F}_qG) \cong S \rtimes \mathcal{U}(\mathbb{F}_qK) \cong S \rtimes \mathcal{U}(\mathbb{F}_qD_6)$ where $S = \mathrm{Ker}(\phi')$.

Let $u = \sum_{j=0}^{2}\sum_{i=0}^{6} r^{3i+7j}(x_{i+7j} + x_{i+7j+21}s) \in S$. Then, $\phi'(u) = 1$. This results in the following equations:

$$\begin{gathered} x_0 + x_1 + x_2 + x_3 + x_4 + x_5 + x_6 = 1, \\ x_{7m} + x_{7m+1} + x_{7m+2} + x_{7m+3} + x_{7m+4} + x_{7m+5} + x_{7m+6} = 0 \\ \text{for } m = 1, \ldots, 5. \end{gathered}$$

Hence, $S = \{1 + \sum_{j=0}^{2}\sum_{i=1}^{6}(r^{3i} - 1)r^{7j}(y_{i+6j} + y_{i+6j+18}s) \mid y_i \in \mathbb{F}_q\}$. It is clear that S is a non-abelian group and that $S^7 = 1$. Since $q = 7^k$, therefore $|S| = 7^{36k}$. From [16, Theorem 2.3] we get that $\mathcal{U}(\mathbb{F}_qD_6) \cong \mathbb{F}_q^* \times \mathbb{F}_q^* \times GL(2,\mathbb{F}_q)$ for $p > 3$. Hence,

$$\mathcal{U}(\mathbb{F}_qG) \cong S \rtimes (\mathbb{F}_q^* \times \mathbb{F}_q^* \times GL(2,\mathbb{F}_q)).$$

4. Char $\mathbb{F}_q \neq 2, 3, 7$: Pertaining to this case, $\mathbb{F}_qD_{42}$ is a semi-simple group algebra by Maschke's theorem and hence $J(\mathbb{F}_qD_{42}) = (0)$. Then,

$$\mathbb{F}_qD_{42} \cong \mathbb{F}_q(D_{42}/D'_{42}) \bigoplus \Delta(D_{42}, D'_{42}).$$

As $D_{42}/D'_{42} \cong C_2$, then $\mathbb{F}_q(D_{42}/D'_{42}) \cong \mathbb{F}_qC_2 \cong \mathbb{F}_q \bigoplus \mathbb{F}_q$. Therefore, the Wedderburn decomposition is

$$\mathbb{F}_qD_{42} \cong \mathbb{F}_q \bigoplus \mathbb{F}_q \bigoplus_{j=1}^{m} M(n_j, R_j)$$

where $n_j \geq 2$ and R_j's are division algebras over the finite field $\mathbb{F}_q$ for $j \in \{1, \ldots, m\}$.

The conjugacy classes of D_{42} are: $\{1\}$, $\{r^{\pm 1}\}$, $\ldots$, $\{r^{\pm 10}\}$, $\{s, rs, \ldots, r^{20}s\}$. Since the class sums form a basis for $Z(\mathbb{F}_q D_{42})$, therefore $\dim(Z(\mathbb{F}_q D_{42}))$ = number of conjugacy classes of $D_{42} = 12$. Hence, $m \leq 10$. Clearly, for the given characterstic p, we obtain $e =$ l.c.m. of the orders of all the p'-elements in $D_{42} = 42$.
(**a**) If $q \equiv 1, 41 \bmod 42$, then $B = \{1\} \bmod 42$ or $B = \{1, 41\} \bmod 42$. From this, we get $|S(\beta_g)| = 1$ for all $g \in G$. Then, by Lemma 2.2 and Lemma 2.3, we deduce that

$$\mathbb{F}_q D_{42} \cong \mathbb{F}_q \bigoplus \mathbb{F}_q \bigoplus_{j=1}^{10} M(n_j, \mathbb{F}_q).$$

After computing the dimension of both sides, we get the equation $\sum_{j=1}^{10} n_j^2 = 40$, which is only possible when $n_j = 2$ for all $j \in \{1, \ldots, 10\}$. Hence,

$$\mathbb{F}_q D_{42} \cong \mathbb{F}_q \bigoplus \mathbb{F}_q \bigoplus M(2, \mathbb{F}_q)^{10}.$$

(**b**) If $q \equiv 5, 17, 25, 37 \bmod 42$, then $B = \{1, 5, 17, 25, 37, 41\} \bmod 42$ or $B = \{1, 25, 37\} \bmod 42$. This gives $|S(\beta_g)| = 1$ for $g = 1, r^7, s$, and $|S(\beta_g)| = 3$ for $g = r, r^2, r^3$. Then, by Lemma 2.2 and Lemma 2.3, we can conclude that

$$\mathbb{F}_q D_{42} \cong \mathbb{F}_q \bigoplus \mathbb{F}_q \bigoplus M(n_1, \mathbb{F}_q) \bigoplus_{j=2}^{4} M(n_j, \mathbb{F}_{q^3}),$$

with the constraint $n_1^2 + 3n_2^2 + 3n_3^2 + 3n_4^2 = 40$. The only such possibility is $n_j = 2$ for all $j \in \{1, \ldots, 4\}$. Hence,

$$\mathbb{F}_q D_{42} \cong \mathbb{F}_q \bigoplus \mathbb{F}_q \bigoplus M(2, \mathbb{F}_q) \bigoplus M(2, \mathbb{F}_{q^3})^3.$$

(**c**) If $q \equiv 11, 19, 23, 31 \bmod 42$, then $B = \{1, 11, 23, 25, 29, 37\} \bmod 42$ or $B = \{1, 13, 19, 25, 31, 37\} \bmod 42$. Thus, $|S(\beta_g)| = 1$ for $g = 1, r^7, s$, $|S(\beta_g)| = 3$ for $g = r^3$, and $|S(\beta_g)| = 6$ for $g = r$. Then, following Lemma 2.2 and Lemma 2.3, the Wedderburn decomposition is

$$\mathbb{F}_q D_{42} \cong \mathbb{F}_q \bigoplus \mathbb{F}_q \bigoplus M(n_1, \mathbb{F}_q) \bigoplus M(n_2, \mathbb{F}_{q^3}) \bigoplus M(n_3, \mathbb{F}_{q^6}),$$

subject to the constraint $n_1^2 + 3n_2^2 + 6n_3^2 = 40$. The equation is satisfied only when $n_j = 2$ for all $j \in \{1, 2, 3\}$. Hence,

$$\mathbb{F}_q D_{42} \cong \mathbb{F}_q \bigoplus \mathbb{F}_q \bigoplus M(2, \mathbb{F}_q) \bigoplus M(2, \mathbb{F}_{q^3}) \bigoplus M(2, \mathbb{F}_{q^6}).$$

(**d**) If $q \equiv 13, 29 \bmod 42$, then $B = \{1, 13\} \bmod 42$ or $B = \{1, 29\} \bmod 42$. This gives $|S(\beta_g)| = 1$ for $g = 1, r^3, r^6, r^7, r^9, s$, and $|S(\beta_g)| = 2$ for $g = r, r^2, r^4$. Then Lemma 2.2 and Lemma 2.3 guarantees that

$$\mathbb{F}_q D_{42} \cong \mathbb{F}_q \bigoplus \mathbb{F}_q \bigoplus_{j=1}^{4} M(n_j, \mathbb{F}_q) \bigoplus_{j=5}^{7} M(n_j, \mathbb{F}_{q^2}),$$

with the constraint $\sum_{j=1}^{4} n_j^2 + \sum_{j=5}^{7} 2n_j^2 = 40$. The only such possibility is $n_j = 2$ for all $j \in \{1, \ldots, 7\}$. Hence,

$$\mathbb{F}_q D_{42} \cong \mathbb{F}_q \bigoplus \mathbb{F}_q \bigoplus M(2, \mathbb{F}_q)^4 \bigoplus M(2, \mathbb{F}_{q^2})^3.$$

For every case (**a**)–(**d**) discussed above, the structure of $\mathcal{U}(\mathbb{F}_q D_{42})$ is a direct implication of the obtained Wedderburn decomposition of $\mathbb{F}_q D_{42}$. □

References

[1] A. Bovdi, L. Erdei: *Unitary Units in Modular Group Algebras of 2-groups*, Communications in Algebra 28.2 (2000), pp. 625–630, doi: 10.1080/00927870008826848.

[2] V. A. Bovdi, A. N. Grishkov: *Unitary and Symmetric Units of a Commutative Group Algebra*, Proceedings of the Edinburgh Mathematical Society 62.3 (2019), pp. 641–654, doi: 10.1017/S0013091518000500.

[3] L. Creedon, J. Gildea: *The Structure of the Unit Group of the Group Algebra $F_{2^k} D_8$*, Canadian Mathematical Bulletin 54.2 (2011), pp. 237–243, doi: 10.4153/CMB-2010-098-5.

[4] L. Creedon, J. Gildea: *The Structure of the Unit Group of the Group Algebra $F_{3^k} D_6$*, International Journal of Pure and Applied Mathematics 45.2 (2008), pp. 315–320.

[5] R. A. Ferraz: *Simple Components of the Center of $FG/J(FG)$*, Communications in Algebra 36.9 (2008), pp. 3191–3199, doi: 10.1080/00927870802103503.

[6] J. Gildea, R. Taylor: *Units of the Group Algebra of the Group $C_n \times D_6$ over any Finite Field of Characteristic* 3, International Electronic Journal of Algebra 24 (2018), pp. 62–67, doi: 10.24330/ieja.440205.

[7] K. Kaur, M. Khan: *Units in $F_2 D_{2p}$*, Journal of Algebra and Its Applications 13.2 (2014), p. 1350090, doi: 10.1142/S0219498813500904.

[8] M. Khan: *Structure of the Unit Group of FD_{10}*, Serdica Mathematical Journal 35.1 (2009), pp. 15–24.

[9] N. Makhijani, R. K. Sharma, J. B. Srivastava: *The Unit Group of $F_q D_{30}$*, Serdica Mathematical Journal 41.2-3 (2015), pp. 185–198.

[10] N. Makhijani, R. K. Sharma, J. B. Srivastava: *Units in $F_{2^k} D_{2n}$*, International Journal of Group Theory 3.3 (2014), pp. 25–34.

[11] S. Malik, R. K. Sharma: *Describing the Group of Units of Integral Group Rings ZD_8 and ZD_{12}*, Asian-European Journal of Mathematics 17.1 (2024), p. 2350236, doi: 10.1142/S1793557123502364.

[12] C. P. Milies, S. K. Sehgal: *An Introduction to Group Rings*, Dordrecht: Kluwer Academic Publishers, 2002.

[13] G. Mittal, R. K. Sharma: *Wedderburn Decomposition of a Semisimple Group Algebra $F_q G$ from a Subalgebra of Factor Group of G*, International Electronic Journal of Algebra 32 (2022), pp. 91–100, doi: 10.24330/ieja.1077582.

[14] M. Sahai, S. F. Ansari: *Unit Groups of Group Algebras of Certain Dihedral Groups-II*, Asian-European Journal of Mathematics 12.4 (2019), p. 1950066, doi: 10.1142/S1793557119500669.

[15] M. Sahai, S. F. Ansari: *Unit Groups of Group Algebras of Groups of Order* 18, Communications in Algebra 49.8 (2021), pp. 3273–3282, doi: 10.1080/00927872.2021.1893740.

[16] R. K. Sharma, J. B. Srivastava, M. Khan: *The Unit Group of FS_3*, Acta Mathematica. Academiae Paedagogicae Nyíregyháziensis. New Series 23.2 (2007), pp. 129–142.

Annales Mathematicae et Informaticae
60 (2024) pp. 108–120
DOI: 10.33039/ami.2024.10.001
URL: https://ami.uni-eszterhazy.hu

On the Diophantine equation $(p^n)^x + (4^m + p)^y = z^2$ when p, $4^m + p$ are prime integers*

Pham Hong Nam

TNU-University of Sciences, Thai Nguyen, Vietnam
namph@tnus.edu.vn

Abstract. In this paper, we give methods to solve the Diophantine equation $(p^n)^x + (p + 4^m)^y = z^2$ where $p \geq 3$ and $p + 4^m$ are prime integers. Concretely, using the congruent method, one proves that this equation has no non-negative solutions if $p > 3$. For the case $p = 3$, using the elliptic curves, we will show that this equation has no solutions if $m \geq 3$. In this case, when $m = 1$ using the elliptic curves, we will show that this equation has only solution $(x, y, z) = (2, 1, 4)$ if $n = 1$ and $(x, y, z) = (1, 1, 4)$ if $n = 2$, and when $m = 2$ using the elliptic curves, we will show that this equation has only solutions is $(x, y, z) = (4, 1, 10)$ if $n = 1$ and $(x, y, z) = (2, 1, 10)$ if $n = 2$.

Keywords: Diophantine equation, elliptic curves, factor method

AMS Subject Classification: 11G07, 11D45, 11D61

1. Introduction

Equations with an exponential Diophantine nature, such as those encountered in the Fermat–Catalan and Beal's conjectures, take the form $a^m + b^n = c^k$ while imposing restrictions on the exponents. These equations arise when additional variables are introduced as exponents in a Diophantine equation. Despite efforts dedicated to addressing specific instances like Catalan's conjecture, a comprehensive theory for solving these equations is currently unavailable. Catalan's conjecture specifically states that the equation $x^p - y^q = 1$ possesses no other integer solutions besides $3^2 - 2^3 = 1$ (see [2]). Mihailescu successfully proved this conjecture

*The author is supported by TNU-University of Sciences.

Submitted: July 8, 2023
Accepted: October 24, 2024
Published online: November 6, 2024

(see [4]). Many authors have studied a generalized form of this equation, which is the Diophantine equation expressed as

$$b^x - c^y = z^2. \tag{1.1}$$

Consider the following examples: when $a^2+b^2 = c^2$ with $\gcd(a, b, c) = 1$ and a being an even number, Terai [7] proposed the conjecture that Equation (1.1) has a unique positive integer solution $(x, y, z) = (2, 2, a)$. In the case where $b = q \not\equiv 7 \pmod 8$ represents an odd prime and $x \equiv 1 \pmod 2$, the Diophantine equation

$$p^x - c^y = z^2 \tag{1.2}$$

has been solved by Arif and Muriefah [1] and Zhu [8]. In [6], Terai presented various results concerning Equation (1.2) (see [6, Theorem 1.2, 1.3, 1.4]). Additionally, Terai noted in [6] that there is currently no proof establishing the absence of solutions for the equations $12^x - 23^y = z^2$ and $24^x - 47^y = z^2$ in the scenario where both x and y are odd.

Turning our attention to a different aspect, the Diophantine equations $x^n+y^t = z^m$ have garnered significant attention from mathematicians. Generally speaking, this problem poses considerable challenges. Suppose m and n are positive integers. We define a solution $(x, y, z) \in \mathbb{Z}_3$ for the equation $x^n + y^n = z^m$ as primitive if $\gcd(x, y, z) = 1$. Conversely, a solution (x, y, z) is deemed trivial if xyz belongs to the set $\{-1, 0, 1\}$. In 1997, Darmon and Merel [3] succeeded in proving the following two theorems by employing the Shimura–Taniyama conjecture in conjunction with Frey curves.

Theorem 1.1. *The equation $x^n + y^n = z^2$ has no nontrivial primitive solutions for prime $n \geq 7$.*

Theorem 1.2. *Assume the Shimura–Taniyama conjecture. Then the equation $x^n + y^n = z^3$ has no nontrivial primitive solutions for prime $n \geq 7$.*

After that, Poonen [5] completed the proof of the above two theorems. Concretely, he proved Theorem 1.1 and Theorem 1.2 are true for all $n \geq 3$ and $n \geq 4$, respectively.

Now, we consider the Diophantine equations

$$p^x + q^y = z^2 \tag{1.3}$$

where p and q are distinct primes. This Diophantine equation has been widely studied for various fixed values of p and q. Note that if either $y = 0$ or $x = 0$ then Equation (1.3) becomes $z^2 - (p^n)^x = 1$ or $z^2 - q^y = 1$, a special form of the Catalan conjecture. However, there is no solution to the general equations.

In this paper, we consider the Diophantine equation

$$(p^n)^x + (p + 4^m)^y = z^2. \tag{1.4}$$

where n is a positive integer and $p \geq 3$ are prime integers. For the case $p > 3$, using the factor method, one proves that (1.4) has no solutions. However, there is still no answer for the case $p = 3$.

2. Preliminaries

In this section, we give some results that will be used in the sequel.

Lemma 2.1. *Any odd power of an integer of the form $4a+3, a>0$ is of the form $4B+3$, i.e., for every odd integer $k \geq 1$, we have $(4a+3)^k = 4B+3$, where B is a positive integer.*

Proof. We prove that by induction on k. For $k=1$, we have $(4a+3)^1 = 4a+3$. Hence the assertion is true when $k=1$.

Assume that the induction is true for all odd powers less than or equal to $k=2r+1$, where $r \geq 0$ is an integer. Then we have

$$(4a+3)^{2r+3} = (4a+3)^2(4a+3)^{2r+1}.$$

By inductive assumption, there exists a positive integer b such that

$$(4a+3)^{2r+1} = 4b+3.$$

Hence we have

$$\begin{aligned}(4a+3)^{2r+3} &= (4a+3)^2(4b+3)\\ &= 4[b(4a+3)^2 + 12a^2 + 18a + 6] + 3.\end{aligned}$$

We put $B = b(4a+3)^2 + 12a^2 + 18a + 6$. Then $(4a+3)^{2r+3} = 4B+3$, the claim is proved. □

For the case $y=0$, we have the following lemma.

Lemma 2.2. *The Diophantine equation*

$$(3^n)^x + 1 = z^2$$

has only solution is $(x,z) = (1,2)$ if $n=1$ and has no non-negative integer solutions if $n>1$.

Proof. Let x and z be non-negative integers such that $(3^n)^x + 1 = z^2$. We have

$$(3^n)^x = z^2 - 1 = (z+1)(z-1). \tag{2.1}$$

Since 3 is prime, we get by Equation (2.1) that there exists integers $a > b$ with $a+b = nx$ such that

$$z+1 = 3^a \quad \text{and} \quad z-1 = 3^b. \tag{2.2}$$

From Equation (2.2), we get that

$$3^b(3^{a-b} - 1) = 2. \tag{2.3}$$

Since 2 and 3 are primes, we get by Equation (2.3) that $b=0$, and therefore $3^{nx} = 3$. Hence $nx = 1$. Thus, if $n>1$ then the Diophantine equation $(3^n)^x + 1 = z^2$ has no solutions, and if $n=1$ then the Diophantine equation $(3^n)^x + 1 = z^2$ has only solution is $(x,z) = (1,2)$. □

For the case $x = 0$, we have the following lemma.

Lemma 2.3. *Let m be a positive integer such that $4^m + 3$ is prime. Then the Diophantine*

$$1 + (4^m + 3)^y = z^2$$

has no non-negative integer solutions.

Proof. Let y and z be non-negative integers such that $(3 + 4^m)^y + 1 = z^2$. We have

$$(3 + 4^m)^y = z^2 - 1 = (z + 1)(z - 1). \tag{2.4}$$

Since 3 is prime, we get by Equation (2.4) that there exists integers $a > b$ with $a + b = y$ such that

$$z + 1 = (3 + 4^m)^a \quad \text{and} \quad z - 1 = (3 + 4^m)^b. \tag{2.5}$$

From Equation (2.5), we get that

$$(3 + 4^m)^b[(3 + 4^m)^{a-b} - 1] = 2. \tag{2.6}$$

Since 2 and $3 + 4^m$ are primes, we get by Equation (2.6) that $b = 0$, and therefore $(3 + 4^m)^y = 3$. Since $3 + 4^m > 3$, Equation (2.6) has no solutions. □

The following lemma is the key to the proof of the main results of this paper.

Lemma 2.4. *Let A be a positive integer of the form $4^m + 3$. Then $A - 1$ has always a prime divisor q such that $q \neq 2$ and $q \neq 3$ for all $m \geq 3$.*

Proof. We have $A - 1 = 4^m + 2 = 2(2 \cdot 4^{m-1} + 1)$. It clear that 2 is not divisor of $2 \cdot 4^{m-1} + 1$. Suppose that $A - 1$ has only two prime divisors, 2 and 3. Then there exists an positive integer x such that $2 \cdot 4^{m-1} + 1 = 3^x$ that means $2 \cdot 4^{m-1} = 3^x - 1$. So, we have the following equation

$$2 \cdot (2^{m-1})^2 = 3^x - 1. \tag{2.7}$$

We divided it into three cases.

1. If $x = 3t$ for some positive integer t. Then Equation (2.7) becomes

$$4^2 \cdot (2^{m-1})^2 = 2^3 3^{3t} - 2^3. \tag{2.8}$$

We put $Y = 4 \cdot 2^{m-1}$ and $X = 2 \cdot 3^t$ then Equation (2.8) becomes

$$Y^2 = X^3 - 2^3$$

is an elliptic curve. Using the Magma Calculator this equation has non-negative solutions (X, Y) is $(2, 0)$. Hence Equation (2.8) has no non-negative integer solutions.

2. If $x = 3t + 1$ for some non-negative integer t. Then Equation (2.7) becomes

$$3^2 \cdot 4^2 \cdot (2^{m-1})^2 = 2^3 \cdot 3^{3t+3} - 2^3 \cdot 3^2. \tag{2.9}$$

We put $Y = 12 \cdot 2^{m-1}$ and $X = 2 \cdot 3^{t+1}$ then Equation (2.9) becomes

$$Y^2 = X^3 - 2^3 \cdot 3^2$$

is an elliptic curve. Using the Magma Calculator this equation has non-negative solutions (X, Y) is $(6, 12)$. Hence $t = 0$ and $m = 1$, a contradiction since $m \geq 3$. Hence Equation (2.9) has no non-negative integer solutions.

3. If $x = 3t + 2$ for some non-negative integer t. Then Equation (2.7) becomes

$$3^4 \cdot 4^2 \cdot (2^{m-1})^2 = 2^3 \cdot 3^{3t+6} - 2^3 \cdot 3^4. \tag{2.10}$$

We put $Y = 36 \cdot 2^{m-1}$ and $X = 2 \cdot 3^{t+2}$ then Equation (2.10) becomes

$$Y^2 = X^3 - 2^3 \cdot 3^4$$

is an elliptic curve. Using the Magma Calculator this equation has non-negative solutions (X, Y) are $(9, 9)$, $(18, 72)$, $(22, 100)$, $(54, 396)$, $(97, 955)$, $(1809, 76941)$. Hence $2.3^{t+2} = 18, 36.2^{m-1} = 72$, that means $t = 0, m = 2$, a contradiction. Therefore, Equation (2.10) has no non-negative integer solutions. □

3. Main results

The purpose of this section is to present some results on solutions to the Diophantine equation

$$(3^n)^x + (4^m + 3)^y = z^2$$

where m is an integer such that $q = 4^m + 3$ is a prime number, for example, $(m, q) \in \{(1, 7), (2, 19), (3, 67), \ldots\}$. For $x = 0$ or $y = 0$, according to Lemma 2.2 and Lemma 2.3. Therefore, from now on, we always consider $x, y \geq 1$. Firstly, we consider x an even integer, and $m \geq 3$ is an integer. Then we have the following theorem.

Theorem 3.1. *Let $m \geq 3$ be an integer such that $4^m + 3$ is prime and x is even. Then the Diophantine equation*

$$(3^n)^x + (4^m + 3)^y = z^2$$

has no non-negative integer solutions.

Proof. Since $x = 2t$ for some $t > 0$, we have

$$(4^m + 3)^y = z^2 - (3^n)^{2t} = (z + 3^{nt})(z - 3^{nt}). \tag{3.1}$$

Since $4^m + 3$ is prime and $y \geq 1$, we get by Equation (3.1) there exists integers $a > b$ such that

$$z + 3^{nt} = (4^m + 3)^a, \;\; z - p^{nt} = (4^m + 3)^b$$

where $a + b = y$. Therefore,

$$(4^m + 3)^b[(3 \cdot 4^m + 1)^{a-b} - 1] = 2 \cdot 3^{nt}. \tag{3.2}$$

Since $4^m + 3 > 3$ and 2, and 3 are distinct primes, we get by Equation (3.2) that $b = 0$. Combined with the condition $y \geq 1$, Equation (3.2) becomes

$$2 \cdot 3^{nt} = (4^m + 3)^y - 1 = (4^m + 2)[(4^m + 3)^{y-1} \ldots + 1]. \tag{3.3}$$

By Equation (3.3) that $(4^m + 2) \mid 2 \cdot 3^{nt}$. Since $m \geq 3$, we get by Lemma 2.4 that $4^m + 2$ has always a prime divisor q such that $q \neq 2$ and $q \neq 3$. Since $(4^m + 2) \mid 2 \cdot 3^{nt}$, we have $q \mid 3^{nt}$, a contradiction. □

For the case $m = 1$ then the Diophantine equation $(p^n)^x + (p + 4^m)^y = z^2$ becomes $(3^n)^x + 7^y = z^2$. Then we have the following theorem.

Theorem 3.2. *Let $x = 2t$ be even for some positive integer t. Then the Diophantine equation*

$$(3^n)^x + 7^y = z^2$$

has only solution is $(x, y, z) = (2, 1, 4)$ if $n = 1$, and has no solutions if $n \geq 2$.

Proof. Since $x = 2t$ are even integers, we have

$$7^y = z^2 - (3^n)^{2t} = (z + 3^{nt})(z - 3^{nt}). \tag{3.4}$$

Since 7 is prime and $y \geq 1$, we get by Equation (3.1) there exists integers $a > b$ such that

$$z + 3^{nt} = 7^a, \;\; z - 3^{nt} = 7^b$$

where $a + b = y$. Therefore,

$$7^b[7^{a-b} - 1] = 2 \cdot 3^{nt}. \tag{3.5}$$

Since 7 and 2, and 3 are distinct primes, we get by Equation (3.5) that $b = 0$. Hence Equation (3.5) becomes

$$2 \cdot 3^{nt} = 7^y - 1. \tag{3.6}$$

We put $nt = u$ then Equation (3.6) becomes

$$2 \cdot 3^u = 7^y - 1. \tag{3.7}$$

We divided it into two cases.

1. The case $u = 2r$. Then Equation (3.7) becomes

$$2 \cdot (3^r)^2 = 7^y - 1. \tag{3.8}$$

- If $y = 3l$, then Equation (3.8) becomes

$$(4 \cdot 3^r)^2 = 2^3 \cdot 7^{3l} - 2^3. \tag{3.9}$$

We put $Y = 4 \cdot 3^r$ and $X = 2 \cdot 7^l$ then Equation (3.9) becomes

$$Y^2 = X^3 - 2^3$$

is an elliptic curve. Using the Magma Calculator this equation has non-negative solutions (X, Y) is $(2, 0)$. Hence Equation (3.9) has no non-negative integer solutions.

- If $y = 3l + 1$, then Equation (3.8) becomes

$$2^3 \cdot 7^2 \cdot 2 \cdot (3^r)^2 = 2^3 \cdot 7^{3l+3} - 2^3 \cdot 7^2. \tag{3.10}$$

We put $Y = 4 \cdot 7 \cdot 3^r$ and $X = 2 \cdot 7^{l+1}$. Then Equation (3.10) becomes

$$Y^2 = X^3 - 2^3 \cdot 7^2$$

is an elliptic curve. Using the Magma Calculator this equation has no non-negative solutions (X, Y). So, in this case, Equation (3.8) has no solutions.

- If $y = 3l + 2$, then Equation (3.8) becomes

$$2^3 \cdot 7^4 \cdot 2 \cdot (3^r)^2 = 2^3 \cdot 7^{3l+6} - 2^3 \cdot 7^4. \tag{3.11}$$

We put $Y = 4 \cdot 7^2 \cdot 3^r$ and $X = 2 \cdot 7^{l+2}$. Then Equation (3.11) becomes

$$Y^2 = X^3 - 2^3 \cdot 7^4$$

is an elliptic curve. Using the Magma Calculator this equation has non-negative solutions (X, Y) is $(6402, 512240)$ that means $2 \cdot 7^{l+2} = 6402$, a contradiction. So, in this case, Equation (3.11) has no solutions.

2. The case $u = 2r + 1$ is odd. Then Equation (3.7) becomes

$$(6 \cdot 3^r)^2 = 6 \cdot 7^y - 6. \tag{3.12}$$

- If $y = 3l$, then Equation (3.12) becomes

$$6^2(6 \cdot 3^r)^2 = 6^3 \cdot 7^{3l} - 6^3. \tag{3.13}$$

We put $Y = 36 \cdot 3^r$ and $X = 6 \cdot 7^l$. Then Equation (3.13) becomes

$$Y^2 = X^3 - 6^3$$

is an elliptic curve. Using the Magma Calculator this equation has non-negative solutions (X, Y) are $(6, 0)$, $(10, 28)$, $(33, 189)$. So Equation (3.12) has no non-negative solutions.

- If $y = 3l + 1$, then Equation (3.12) becomes

$$6^2 \cdot 7^2(6 \cdot 3^r)^2 = 6^3 \cdot 7^{3l+3} - 6^3 \cdot 7^2. \tag{3.14}$$

We put $Y = 252 \cdot 3^r$ and $X = 6 \cdot 7^{l+1}$. Then Equation (3.14) becomes

$$Y^2 = X^3 - 6^3 \cdot 7^2$$

is an elliptic curve. Using the Magma Calculator this equation has non-negative solutions (X, Y) are $(22, 8)$, $(25, 71)$, $(42, 252)$, $(105, 1071)$, $(294, 5040)$, and $(394, 7820)$. Hence $6 \cdot 7^{l+1} = 42$ or $6 \cdot 7^{l+1} = 294$, that means $l = 0$ or $l = 1$. If $l = 0$ then $y = 1$ and $nt = 1$ that means $y = 1$ and $n = t = 1$, and so $x = 2, z = 4$. If $l = 1$ then $y = 4$. So, Equation (3.7) becomes $3^{nt} = 1200$, a contradiction.

- If $y = 3l + 2$, then Equation (3.12) becomes

$$6^2 \cdot 7^4(6 \cdot 3^r)^2 = 6^3 \cdot 7^{3l+6} - 6^3 \cdot 7^4. \tag{3.15}$$

We put $Y = 6^2 \cdot 7^2 \cdot 3^r$ and $X = 6 \cdot 7^{l+2}$. Then Equation (3.15) becomes

$$Y^2 = X^3 - 6^3 \cdot 7^4$$

is an elliptic curve. Using the Magma Calculator this equation has only non-negative solution (X, Y) is $(106, 820)$. Therefore, Equation (3.15) has no non-negative solutions.
Thus if x is even, then $(x, y, z) = (2, 1, 4)$ is only solution of the Diophantine equation $(3^n)^x + 7^y = z^2$ when $n = 1$. □

For the case $m = 2$ then the Diophantine equation $(p^n)^x + (p + 4^m)^y = z^2$ becomes $(3^n)^x + 19^y = z^2$. With the same of the proof of Theorem 3.2, we have the following theorem.

Theorem 3.3. *Let $x = 2t$ for some positive integer t. Then the Diophantine equation*

$$(3^n)^x + 19^y = z^2$$

has only solution is $(x, y, z) = (4, 1, 10)$ if $n = 1$ and $(x, y, z) = (2, 1, 10)$ if $n = 2$, and has no solutions if $n \geq 3$.

Proof. Similar to the proof of Theorem 3.2, we have the equation

$$2 \cdot 3^{nt} = 19^y - 1. \tag{3.16}$$

We put $nt = u$. Then Equation (3.16) becomes

$$2 \cdot 3^u = 19^y - 1. \tag{3.17}$$

We divided it into two cases.

1. The case $u = 2r$. Then Equation (3.17) becomes

$$2 \cdot (3^r)^2 = 19^y - 1. \tag{3.18}$$

- If $y = 3l$, then Equation (3.18) becomes

$$(4 \cdot 3^r)^2 = 2^3 \cdot 19^{3l} - 2^3. \tag{3.19}$$

We put $Y = 4 \cdot 3^r$ and $X = 2 \cdot 19^l$. Then Equation (3.19) becomes

$$Y^2 = X^3 - 2^3. \tag{3.20}$$

is an elliptic curve. Using the Magma Calculator, Equation (3.20) has non-negative solutions (X, Y) is $(2, 0)$. Hence $4 \cdot 3^r = 0$, a contradiction. Therefore, Equation (3.19) has no non-negative integer solutions.

- If $y = 3l + 1$, then Equation (3.18) becomes

$$2^3 \cdot 19^2 \cdot 2 \cdot (3^r)^2 = 2^3 \cdot 19^{3l+3} - 2^3 \cdot 19^2. \tag{3.21}$$

We put $Y = 4 \cdot 19 \cdot 3^r$ and $X = 2 \cdot 19^{l+1}$. Then Equation (3.21) becomes

$$Y^2 = X^3 - 2^3 \cdot 19^2$$

is an elliptic curve. Using the Magma Calculator, this equation has non-negative solutions (X, Y) are $(17, 45)$, $(38, 228)$, $(114, 1216)$ and $(209, 3021)$. Hence

$$2 \cdot 19^{l+1} = 38 \quad \text{and} \quad 4 \cdot 19 \cdot 3^r = 228$$

that means $l = 0, r = 1$. Therefore, we have $nt = 2, y = 1$. So, Equation (3.16) has no solutions if $n \geq 3$, and the non-negative integer solutions of Equation (3.16) are $(x, y, z) = (4, 1, 10)$ if $n = 1$ and $(x, y, z) = (2, 1, 10)$ if $n = 2$.

- If $y = 3l + 2$, then Equation (3.18) becomes

$$2^3 \cdot 19^4 \cdot 2 \cdot (3^r)^2 = 2^3 \cdot 19^{3l+6} - 2^3 \cdot 19^4. \tag{3.22}$$

We put $Y = 4 \cdot 19^2 \cdot 3^r$ and $X = 2 \cdot 19^{l+2}$. Then Equation (3.22) becomes

$$Y^2 = X^3 - 2^3 \cdot 19^4$$

is an elliptic curve. Using the Magma Calculator this equation has no non-negative solutions. So, in this case, Equation (3.18) has no solutions.

2. The case $u = 2r + 1$. Then Equation (3.17) becomes

$$(6 \cdot 3^r)^2 = 6 \cdot 19^y - 6. \tag{3.23}$$

- If $y = 3l$, then Equation (3.23) becomes

$$6^2(6 \cdot 3^r)^2 = 6^3 \cdot 19^{3l} - 6^3. \tag{3.24}$$

We put $Y = 36 \cdot 3^r$ and $X = 6 \cdot 19^l$. Then Equation (3.24) becomes

$$Y^2 = X^3 - 6^3$$

is an elliptic curve. Using the Magma Calculator, Equation (3.24) has non-negative solutions (X, Y) are $(6, 0), (10, 28), (33, 189)$. So, Equation (3.24) has no non-negative solutions.

- If $y = 3l + 1$, then Equation (3.23) becomes

$$6^2 \cdot 19^2(6 \cdot 3^r)^2 = 6^3 \cdot 19^{3l+3} - 6^3 \cdot 19^2. \tag{3.25}$$

We put $Y = 36 \cdot 19 \cdot 3^r$ and $X = 6 \cdot 19^{l+1}$. Then Equation (3.25) becomes

$$Y^2 = X^3 - 6^3 \cdot 19^2$$

is an elliptic curve. Using the Magma Calculator this equation has no non-negative solutions. So, in this case, Equation (3.23) has no solutions.

- If $y = 3l + 2$, then Equation (3.23) becomes

$$6^2 \cdot 19^4(6 \cdot 3^r)^2 = 6^3 \cdot 19^{3l+6} - 6^3 \cdot 19^4. \tag{3.26}$$

We put $Y = 6^2 \cdot 19^2 \cdot 3^r$ and $X = 6 \cdot 19^{l+2}$. Then Equation (3.26) becomes

$$Y^2 = X^3 - 6^3 \cdot 19^4$$

is an elliptic curve. Using the Magma Calculator this equation has no non-negative solutions. So, in this case, Equation (3.23) has no solutions. □

Next, we consider y to be an even number. Then we have the following theorem.

Theorem 3.4. *Let m be an integer such that $4^m + 3$ is prime and $y = 2s$ be even for some positive integer s. Then the Diophantine equation*

$$(3^n)^x + (4^m + 3)^y = z^2$$

has no non-negative integer solutions.

Proof. Since $y = 2s, s > 0$, the Diophantine equation

$$(3^n)^x + (4^m + 3)^y = z^2$$

becomes

$$3^{nx} = z^2 - (4^m + 3)^{2s} = [z + (4^m + 3)^s][z - (4^m + 3)^s]. \tag{3.27}$$

Since 3 is prime and $x \geq 1$, we get by Equation (3.27) there exists integers $a > b$ such that

$$z + (4^m + 3)^s = 3^a, \;\; z - (4^m + 3)^s = 3^b$$

where $a + b = nx$. Therefore,

$$3^b[3^{a-b} - 1] = 2 \cdot (4^m + 3)^s. \tag{3.28}$$

Since $4^m + 3 > 3$ and 2, and 3 are distinct primes, we get by Equation (3.28) that $b = 0$. Combined with the condition $x \geq 1$, Equation (3.28) becomes

$$2 \cdot (4^m + 3)^s = (3^n)^x - 1.$$

We set $nx = u$. Since $4^m + 3$ is a prime number, we have

$$3^{4^m+2} \equiv 1 \,(\mathrm{mod}\,(4^m + 3)).$$

Let i and j be integers such that $0 \leq i, j < 4^m + 2$ and $i \neq j$. Then we have

$$3^{i+(4^m+2)k} \not\equiv 3^{j+(4^m+2)k} \,(\mathrm{mod}(4^m + 3))$$

for all integers k. We divided it into two cases.

1. Let $u = i + (4^m + 2)k$ be a positive integer, where $0 < i < 4^m + 2$ and k is a non-negative integer. Then

   $$2 \cdot (4^m + 3)^s = 3^u - 1 \not\equiv 0 \,(\mathrm{mod}(4^m + 3))$$

   a contradiction.

2. Let $u = (4^m + 2)k$ be a positive integer, where k is a positive integer. Clearly, 3 is a divisor of $4^m + 2$ for all positive integers m. Hence, there exists a positive integer v such that $4^m + 2 = 3v$. Therefore,

   $$2 \cdot (4^m + 3)^s = 3^{3vk} - 1 = 27^{vk} - 1 = 2 \cdot 13 \cdot [27^{vk-1} + \ldots + 1]$$

   a contradiction, because 13 is not divisor of $2 \cdot (4^m + 3)^s$. Thus, in this case, the Diophantine equation

   $$(3^n)^x + (4^m + 3)^y = z^2$$

   has no non-negative integer solutions. □

Finally, we consider x and y to be both odd integers. Then we have the following theorem.

Theorem 3.5. *Let m be an integer such that $4^m + 3$ is prime. Suppose that $x = 2t + 1, y = 2s + 1$ for some positive integers s, t. Then the Diophantine equation*

$$(3^n)^x + (4^m + 3)^y = z^2$$

has no non-negative integer solutions for all $m \geq 3$ and has only solution is $(x, y, z) = (1, 1, 4)$ if $n = 2$ and $m = 1$, and has only solutions is $(x, y, z) = (1, 1, 10)$ if $n = 2$ and $m = 2$.

Proof. We divided it into two cases.

1. Let n is even. Then $n = 2l$ for some integer l. The equation $(3^n)^x + (4^m + 3)^y = z^2$ becomes
$$(3^l)^{2x} + (4^m + 3)^y = z^2 \tag{3.29}$$
   - Let $m \geq 3$. By Theorem 3.1, Equation (3.29) has no non-negative integer solutions.
   - Let $m = 1$. We get by Theorem 3.2 that Equation (3.29) has only solution is $(2x, y, z) = (2, 1, 4)$ if $l = 1$. Hence $(x, y, z) = (1, 1, 4)$ if $n = 2$.
   - Let $m = 2$. Since x is odd, we get by Theorem 3.3 that Equation (3.29) has only solution is $(2x, y, z) = (2, 1, 10)$ if $l = 2$. Hence $(x, y, z) = (1, 1, 10)$ if $n = 4$.

2. Let n be odd and $x = 2t + 1$, and $y = 2s + 1$ for some positive t and s. We get by Lemma 2.1 that there exists integers A, B such that $(3^n)^{2t+1} = 4A+3$ and $(4^m + 3)^{2s+1} = 4B + 3$. Since 3 and $4^m + 3$ are odd prime numbers, we have $z^2 = (3^n)^x + (4^m + 3)^y$ is even. Hence z is even, which means $z = 2c$ for some integer c. Therefore, $z^2 = 4c^2$, which means 4 is a divisor of z^2. On the other hand, we have
$$z^2 = (3^n)^{2t+1} + (4^m + p)^{2s+1} = 4(A + B) + 2,$$
   a contradiction. □

4. Conclusion

In this paper, we give methods to solve the Diophantine equation $(3^n)^x+(3+4^m)^y = z^2$. This method is perfectly applicable to similar Diophantine equations.

Conflict of interest

The authors declare that there are no potential conflicts of interest regarding the publication of this work. And there are no financial and personal relationships with other people or organizations that can inappropriately influence our work.

References

[1] S. Arif, F. A. Muriefah: *On the Diophantine equation* $x^2 + q^{2k+1} = y^n$, J. Number Theory 95.1 (2002), pp. 95–100, doi: 10.1006/jnth.2001.2750.

[2] E. Catalan: *A note on extraite dune lettre adressee a lediteur*, J. Reine Angew. Math. 27 (1884), pp. 192–192, doi: 10.1515/crll.1844.27.192.

[3] H. Darmon, L. Merel: *Winding quotients and some variants of Fermat's Last Theorem*, J. Reine Angew. Math. 490 (1997), pp. 81–100, doi: 10.1515/crll.1997.490.81.

[4] P. Mihailescu: *On primary Cyclotomic units and a proof of Catalan's conjecture*, J. Reine Angew. Math. 572 (2004), pp. 167–195, doi: 10.1515/crll.2004.048.

[5] B. Poonen: *Some diophantine equations of the form* $x^n + y^n = z^m$, Acta Arith. 86.3 (1998), pp. 193–205, doi: 10.4064/aa146-2-6.

[6] N. Terai: *A note on the Diophantine equation* $x^2 + q^m = c^n$, Bull. Aust. Math. Soc. 90.1 (2014), pp. 20–27, doi: 10.1017/S0004972713000981.

[7] N. Terai: *The Diophantine equation* $x^2 + q^m = p^n$, Acta Arith. 63.4 (1993), pp. 351–358, doi: 10.4064/AA-63-4-351-358.

[8] H. Zhu: *A note on the Diophantine equation* $x^2+q^m = y^3$, Acta Arith. 146.2 (2011), pp. 195–202, doi: 10.4064/aa146-2-6.

Annales Mathematicae et Informaticae
60 (2024) pp. 121–132
DOI: 10.33039/ami.2024.02.005
URL: https://ami.uni-eszterhazy.hu

Designing structured cabling systems documentation and model by using Building Information Modeling – Literature review

Sergey Pogorelskiy[a], **Imre Kocsis**[b]

[a]Doctoral School of Informatics, University of Debrecen, Debrecen, Hungary
pogorelski.serg@gmail.com

[b]Department of Basic Technical Studies, Faculty of Engineering, University of Debrecen, Debrecen, Hungary
kocsisi@eng.unideb.hu

Abstract. Offices, airports, factories and universities requires a local area network that combines computers, telephones, and peripheral equipment. A basis of computer network is a structured cabling system with the main elements as follows: telecommunication cabinets, copper and optical panels, cable lines and cable trays. The design of structured cabling systems as parts of complex building engineering systems—similarly to many engineering systems—is usually carried out by computer-aided design programs. This approach has a number of disadvantages. Therefore, more and more attention has recently been paid to the use of Building Information Modeling concept not only for the design of buildings and their engineering systems but also for the operation of them ([15, 20, 23, 26]). In a recent work, we touched upon the issue of designing telecommunication cabinets and the equipment inside them in the Building Information Modeling environment. We developed a novel 3D model of cabinets, which has a number of distinctive features: (1) the ability to select the equipment installed in a particular unit in the properties; (2) the ability to change and add equipment inside the cabinet; (3) automated creation of schemes for facades of cabinets and equipment inside; (4) automated creation of equipment specifications in cabinets [18]. Also, we analyzed the use of cable trays for modeling cable lines and found an optimal way to build tray routes and their elements. The goal of our further study is to build up the entire design cycle process using all the necessary elements in

Submitted: July 19, 2023
Accepted: February 10, 2024
Published online: February 17, 2024

Building Information Modeling and then use the model during the operation. Till now there is only a small part of the research covers topics related to use Building Information Modeling for engineering systems. The authors aim to fill this gap with a qualitative analysis of the existing literature and the application of Building Information Modeling in information technology. The methodology includes several steps: Traditional literature review on the use of Building Information Modeling in the design and management of facilities in the field of engineering systems and then qualitative analysis of researchers content related to the design of engineering systems. The qualitative investigation of the literature has identified five main areas of Information and Communications Technology where Building Information Modeling tools and methodologies are used, namely (1) analysis of cabling systems; (2) production of working drawings; (3) optimized data center design; (4) preparation of documentation and models for further facility management; (5) monitoring system parameters. Literary sources have different degrees of correlation with the main research questions: weak, medium, and strong. In our study, we used medium and strong correlated topics of the study. The ultimate goal is to find an optimal solution to designing structured cabling systems documentation and model by using Building Information Modeling with the improvement of techniques available in the relevant literature.

Keywords: Building Information Modeling, BIM, structured cabling systems

1. Introduction

Structured cabling systems (SCS) are the basis for smart buildings and data centers. It is difficult to imagine a modern office without cable systems. Now they include not only computers, telephones and, peripheral equipment. This SCS applies to intelligent buildings for several reasons. Literally, the networked link between a building's systems allows the company operate within to automatically control security, environmental conditions, lighting, communications, and other factors. It is now more important than ever for an enterprise's operations to be efficient, effective, and economical. Furthermore, one of the most challenging environments is a data center, where structured cabling can be extremely beneficial and even necessary. Given the numerous active equipment elements that require connectivity. SCS include main following elements:

- telecommunication cabinets and racks in cross rooms of buildings;
- copper and optical panels inside telecommunication cabinets;
- telecommunication outlets installed in the offices;
- copper and optical cable lines connecting panels in the cabinet and information sockets;
- cable trays and boxes for laying cable lines in them.

During the SCS design by using standard CAD systems, engineers faced with the following problems: (1) Drawings of cable trays are created in 2D; (2) Cabinet façade schemes are created manually; (3) The equipment for specifications are counted manually; (4) Cable lines specification counted manually; (5) Classic 2D schemes are not convenient to use for the further facility management; (6) Duplication of work between the engineers designing a system and engineers creating a model of the same system in Building Information Modeling (BIM). To avoid the above problems, it is necessary to use the concept of BIM design. BIM is a complicated idea that relies for management work, tools, and apps to facilitate information flow and, as a result, increase project productivity. The industry is evolving due to the use of BIM and the growing use of digital technology in building construction, building operations, and building maintenance. Instead of traditional application of BIM for buildings design, the technology could be applied for Engineering Design. Traditionally, the BIM concept in engineering systems is used to model and create projects for the following systems: heating, ventilation and air conditioning systems, pipelines, fire extinguishing systems. Less often we can find power supply systems. And the area of application of BIM for structured cabling systems remains unexplored. The aim of this research, therefore, is to start bridging that gap. In particular, we have prepared a traditional literature review on the application of BIM in information and communication technology and engineering systems.

2. Methodology

The methodology adopted to develop this qualitative review of the literature on the use of BIM in structured cabling design and data center research consisted of a traditional literature search on the use of (1) BIM in design and facility management; (2) Qualitative analysis of the content related to the design of structured cabling systems and data centers, disclosed in Step 1. Five main areas are identified in which BIM tools and methodologies are used; A breakdown was also made according to three criteria for the correspondence of literature sources to the main five questions:

- Weak: there is no BIM use with the same title proposed by the authors nor is there a BIM use that, in its description, focuses on the structured cabling design and data center engineering area that the authors identified.
- Medium: there is either a BIM use with the same title identified by the authors or there is a BIM use (or more than one) that focuses on the same topic proposed by the authors, even if the description in the guide is too general and never directly relates to the structured cabling design and data center engineering discipline.
- Strong: there is a BIM use with the same title identified by the authors and its description goes into detail about the structured cabling design and data center engineering area that the authors identified.

The qualitative investigation of the literature that the authors have conducted has highlighted 5 main BIM uses in ICT engineering: (1) cabling systems; (2) production of working drawings; (3) data center optimization; (4) preparation of documentation and models for further facility management; (5) monitoring system parameters.

3. Literature review

This article first reviews articles in the field of using BIM for the design and operation of engineering systems. The aim is to find the literature in recent years that has discussed these issues in general or for specific tasks. Based on the data obtained, there are a number of documents to date in which attention has been paid to various models and methods of operation. The most relevant documents are selected and discussed within the framework of this article, 27 articles-analyzed, and on the basis of them, a quantitative and qualitative analysis is carried out, as reflected in tables 1 and 2.

Table 1. Quantitative analysis.

Reference	Area of use				BIM Content		
	Plan	Design	Build	Operate	Is there a relationship with the BIM	Does the article correlate with 1 of the 5 main criteria	Level of correlation with main criteria's
[18]		+			+	+	Strong
[23]		+		+	+	+	Strong
[8]		+			+	+	Weak
[21]				+			Weak
[2]		+			+		Medium
[3]		+			+		Medium
[26]		+			+	+	Strong
[15]		+			+		Weak
[11]		+			+	+	Strong
[7]		+			+		Weak
[19]		+	+		+		Weak
[9]		+			+		Weak
[4]				+			Weak
[5]				+			Weak
[6]		+		+			Weak
[17]				+			Weak
[10]		+		+	+		Weak
[13]				+			Weak
[14]				+	+		Weak
[16]				+			Weak
[1]				+			Weak
[22]				+			Weak
[24]				+			Weak
[25]				+			Weak
[27]	+	+		+			Weak
[12]		+	+		+	+	Strong

Table 2. Qualitative analysis.

	Number of Reference Document
(1) analysis of cabling systems	6.9%
(2) production of working drawings	24.14%
(3) optimized data center design	20.69%
(4) preparation of documentation for further facility management	37.93%
(5) monitoring system parameters	37.93%

Table 1 shows a detailed analysis of all literature sources used in this article. Table 2 shows in percentage terms the number of sources related to the 5 main criteria for qualitative analysis. The related articles attempt to answer the following questions:

1. The purpose of this article [1] is to provide an overview of the different types of maintenance strategies for critical infrastructure facilities such as hospitals in Malaysia. To obtain data, interviews were conducted with institutional management and end-users in selected hospitals. Hospital management should have a strategic maintenance plan in place to monitor each facility and help it operate well with less chance of failure. Therefore, end-user facilities in hospitals must be maintained and controlled according to their function. The results show that there is a correlation between operating strategies and customer satisfaction levels.

2. The article [2] reveals the definition and essence of information modeling in construction. The content and effect of using information modeling of various objects of the object's life cycle is described. Analyzed short-term and long-term benefits. An exploratory review of Revit software was carried out in search of Autodesk according to the criteria: tools, cost characteristics and profitability. A predictive calculation of the effectiveness of information modeling technologies in construction is given, examples of the successful implementation of information modeling in construction abroad and in Russia are found.

3. In recent years, cloud computing has developed rapidly. In order to implement them, we have to have a physical infrastructure in the form of a data center. Accordingly[4], the data center must operate efficiently, with all the necessary monitoring systems for parameters, to ensure the best level of use of IT resources. Through the effective implementation of data center operations management for cloud computing, it is possible to reduce the workload of staff and improve the efficiency of operational staff, improve the current state of business systems, as a result improves the overall efficiency of enterprise management.

4. The growing development of data centers is causing problems with energy

consumption. More than 1.3 percent of global energy consumption comes from electricity used by data centers, and this rate is growing. Articles show that most of the energy consumed in a data center is mainly due to the electricity used to run servers and cool them (70 percent of the total cost of a data center). Therefore, the main factor in this power consumption is related to the number of running servers. The main goal of this article [5] is to manage the on/off of servers in the data center over time in order to adapt the system to changes in incoming traffic in order to ensure good performance and reasonable power consumption. The system begins to gradually turn on the servers at a high level of requests. And turn off the servers gradually when the rate of receiving requests becomes low.

5. Data centers are about 50 times more energy intensive than conventional office buildings. The main goal of this study [6] is determined by the process of energy analysis, numerical studies and simulation studies to evaluate the impact of each technical component in order to create energy-optimized data centers. The methodology and program developed in this paper for evaluating the energy consumption of data centers should be used by engineers and designers when building data centers to evaluate the efficiency and economic benefits of cooling systems.

6. This article [7] presents a comparison of the development of design in civil engineering: traditional design and information modeling (BIM). The advantages and disadvantages of traditional design and information modeling are described. An office complex in Warsaw, designed using BIM software, was analyzed. The shortcomings and problems in the implementation of BIM are analyzed.

7. The aim of this study [9] is to develop a 4D BIM research model in EPC projects. This study differs from previous work in the following ways: Firstly, previous studies have not considered the context of project contract types, and this study focuses on EPC projects where a quantitative research method is appropriate, as some redundant variables can be avoided. Secondly, most of the research on the use of information technology was carried out in developed countries. This study focuses on China, which is a typical developing country with a huge construction market.

8. The study [10] presents a new concept that shows how BIM can be used effectively during building maintenance. BIM is now widely used in construction projects for quality control, time management and financial control. After the construction is completed, the digital model is transferred to the client for subsequent use in operation. BIM can help the owner optimize facility maintenance by exporting relevant information about the building being built and the requirements to run the systems that will be used throughout the facility's life cycle. The critical factor is the availability of a method to combine BIM with active data. The most important outcome of this study is

the definition of a conditions data model solution that: combines active and passive big data with BIM; provides dynamic services based on the shape of the building, building services technologies, the Internet of things and information about the actions of residents in real time; solves IT problems with the processing of large BIM files through an Internet browser and mobile applications; and allows to provide the data needed for the building's digital twin.

9. There are currently no monitoring strategies utilized in the design of mechanical, electrical and plumbing (MEP) engineering systems due to the complex structure of the components. In order to address this issue, this work [11] generated a directed representative graph using BIM data and integrated the graph with Internet of Things (IoT) for the aim of monitoring MEP systems. In a directed representative graph, edges are the connections between two representative adjacency points, and vertices are the representative points. Six Revit-created BIM object models were utilized to test the suggested methodology. The developed simulated system shows how to use IoT for intelligent MEP system monitoring on a directed representative graph.

10. The paper [12] examines the benefits and drawbacks of adopting BIM technologies in the planning of the Shanghai Baoshan commercial center. The mechanical and electrical components of the building's engineering systems are given special consideration. In this project, three-dimensional mechanical and electrical models interact, and the pipeline conflict problems are found when building the electromechanical system. As a result, the construction period is extended and more materials are wasted, which lowers costs and improves construction efficiency.

11. High performance computing (HPC) is inextricably linked to effective data center infrastructure management (DCIM). The cost and complexity of DCIM quality assurance is constantly reviewed and evaluated by companies such as Google, Microsoft, and Facebook. This article [13] demonstrates a system that uses big data strategies and 3D game technology to successfully monitor and analyze multiple HPC systems and a modular data center on a single platform. Big data technology and a 3D gaming platform enable real-time monitoring of 5,000 environmental sensors, over 3,500 IT data points, and display visual analytics of the overall operational health of the data center.

12. This article [14] presents a scenario for integrating augmented reality (AR) and building information modeling (BIM) to create an intelligent environment (AmI) for facility managers, in which mobile user interfaces will have data to facilitate decision making. The technological requirements for creating such an intelligent environment are also discussed.

13. Efforts were made in the article [16] to reduce operational risk, increase responsiveness and improve monitoring in data center infrastructure using low

cost and low power wireless sensors to monitor power, temperature, humidity, air pressure drop and vibration in the data center. The purpose of the study was to collect and analyze information in order to ultimately reduce downtime and operating costs, improve energy efficiency, and properly plan the use of space in the data center. Finally, an approach to time monitoring and data center infrastructure management was proposed. As a prospective research in this area, the authors consider predicting the performance and operation of data centers in the future.

14. The aim of this paper [20] is to develop a new methodology based on BIM and integration of facility management systems supported by an information model. The process of implementing the information model is described, including the information technology involved, the data and process requirements, and the methods used to assess the performance of the facilities. A first pilot study has been carried out on the example of operating theatres in medical centers. The methodology can facilitate maintenance planning based on the current state of the facility and the achievement of organizational, environmental and technical requirements. The practical results are as follows: improved assessment of technical and environmental performance; better visualization of the state of the building; improved decision-making process; easier planning of maintenance tasks and management of facility parameters.

15. In this paper [17], a new statistical approach, based on Monte Carlo methodology, has been proposed to estimate the lifetime performance of modular data centers. The approach uses component failure probability distributions over time of use to calculate component-level failure penetration, called a snapshot. At the same time, a generalisation in the form of a cumulative Tanh-Log probability distribution has been proposed to better fit real system failure data. Using the proposed distribution and analysis approach, the performance of three well-known topologies in the context of modular data centers, i.e. FatTree, BCube and MDCube2D, was studied. In addition, in order to make these topologies more flexible and independent with respect to the hardware and to increase their resilience to failures, some extended versions of these topologies, designated as FatTreeE, BCubeE and MDCube2DE, were introduced. It was concluded that the extended BCube topology, BCubeE, could provide better fault tolerance in terms of various performance metrics (depending on the tolerance range of the distributions used).

16. The article [18] compares classical design methods and the use of information modeling methods, using the example of modeling cable trays and telecommunication cabinets. It provides an analysis of the available cable tray design solutions and the choice of the optimal solution for cable route modeling and automatic element specification. Also, this article presents a new dynamic family of telecommunication cabinets, which was developed by the authors to simplify the creation of schemes for cabinet facades in data centers and

further automatic specification of equipment inside racks. The labor costs of engineers in design and modeling have been shown to be lower as a result of this study.

17. In the article [23] the authors focus on data centers. This article provides a practical example showing the problem of designing and then preparing a model for use in Facility Management (FM). The importance of this study lies in the fact that the proposed method demonstrates a direct relationship between the following three components: temperature and humidity sensors, FM and Building Management System (BMS) software, and BIM. This method is implemented by a direct link between the BIM model and the program for FM, which in turn is linked to BMS.

18. The authors of the article [3] considered 3 stages of integrating BIM systems in the design of data centers: 1. Building an information model using the selected BIM tool. 2. Design of engineering systems depending on the level of complexity of the object. 3. A method for implementing the data center model with information modeling has been formulated, taking into account the requirements of the Uptime Institute classification. The main advantages of BIM in the design of data centers were also described.

19. BIM is mainly used for the design and construction of new buildings. However, one of the main challenges facing the implementation of BIM for existing or old buildings built without considering their modelling is to obtain accurate data about the existing building and convert it into a BIM model. Based on this challenge, this study [15] develops a framework that uses different data collection methods for existing buildings and then converts the data obtained into 3D BIM models, with which the facility management processes can be improved over the life of the building. In particular, this paper looks at: 3D laser scanning techniques for collecting data from the interior and exterior of buildings. The captured data was then converted into a high-resolution 3D model. This confirmed the accuracy of this engineering model by converting 2D engineering plans into 3D plans using well-known engineering design tools. Moreover, the 3D model was integrated with the web-based Building Management System (BMS) platform. This research base helps in the development of engineering facility management processes and modern digital transformation processes aimed at accelerating the management of facilities during operation using modern technology.

20. The study [25] of modern cyber-physical systems (CPS) has been an important area of research for Internet Data Centers (IDCs). IDCs - support the reliable operation of many important online services. Along with the expansion of Internet services and cloud computing, the energy consumption associated with IDC operations has increased significantly in recent years. This massive energy consumption has placed an extremely heavy burden on IDC

operators. While most previous work has only looked at IDC's dynamic optimisation within electricity markets, IDC's response to the electricity market has been overlooked. Due to the fact that IDCs are typically large users in the electricity market, they may have market power affecting the electricity price. This paper investigates how to address the interaction between the performance of IDCs and the market price of electricity. To this end, a function for modelling IDC market power is proposed and the problem of minimising the total electricity cost is formulated as non-linear programming. A CMC algorithm based on the economic concept is also presented. The CMC algorithm not only solves the optimization problem efficiently but also determines the dynamics of workload allocation. Extensive performance evaluations demonstrate that the proposed method can effectively minimize the overall power cost for IDC by adaptively managing the interaction between IDC and smart grid.

21. The article [26] discusses the use of wireless sensors for monitoring the temperature parameters of data centers and their integration with BIM. Integrating a Wide Area Network with an existing data center BMS has a number of advantages, including cheaper and faster installation, which allows more sensors to be deployed for more accurate measurements and control, and the associated flexibility to deploy the temporary infrastructure needed to perform measurements. In limited time. This technique for monitoring temperature parameters in the data center allows to increase energy efficiency.

As can be seen from the presented study, the number of papers aiming to answer questions connected to the design and operation of structured cabling systems in offices and data centers is quite minimal. It is important to go deeply into this area because it is understudied.

4. Research gap analysis and future research

Due to research topic related to SCS is unexplored, in our research we would like to concentrate on practical aspects and check the concept. Our next steps are as follows: creation of a model and design of structured cabling systems of a industrial project; developing script for automatic cable routing and adjusting for cable trays; linking elements of a structured cabling network with elements of other engineering systems in the model (clash detection); testing of cable lines and comparison of actually built results with lines obtained in the cable specification in BIM; comparison of the parameters of cable lines, actually received and calculated.

References

[1] N. A. A. Rani, M. R. Baharum, A. R. N. Akbar, A. H. Nawawi: *Perception of maintenance management strategy on healthcare facilities*, Procedia - Social and Behavioral Sciences 170 (2015), pp. 272–281.

[2] R. G. Abakumov, A. E. Naumov: *Building information model: advantages, tools and adoption efficiency*, in: IOP Conference Series: Materials Science and Engineering, vol. 327, 2018, p. 022001.

[3] R. Angelina, K. Pavel: *Application of Building Information Modeling in Data Center design*, in: IOP Conference Series: Materials Science and Engineering, vol. 869, 2020, p. 022006.

[4] W. Bai, W. Geng: *Research on operation management under the environment of cloud computing data center*, International Journal of Database Theory and Application 8.2 (2015), pp. 185–192.

[5] M. Bayati: *Managing energy consumption and quality of service in data centers*, in: 2016 5th International Conference on Smart Cities and Green ICT Systems (SMARTGREENS), IEEE, 2016, pp. 1–9.

[6] J. Cho, J. Yang, C. Lee, J. Lee: *Development of an energy evaluation and design tool for dedicated cooling systems of data centers: Sensing data center cooling energy efficiency*, Energy and Buildings 96 (2015), pp. 357–372.

[7] I. Czmoch, A. Pękala: *Traditional design versus BIM based design*, Procedia Engineering 91 (2014), pp. 210–215.

[8] K. Elena, K. Pavel, B. Tatyana: *Automation of the formation of organizational technological documentation*, in: Applied mechanics and materials, vol. 738, 2015, pp. 444–447.

[9] P. Gong, N. Zeng, K. Ye, M. König: *An empirical study on the acceptance of 4D BIM in EPC projects in China*, Sustainability 11.5 (2019), p. 1316.

[10] E. Halmetoja: *The conditions data model supporting building information models in facility management*, Facilities (2019).

[11] J. Han, X. Zhou, W. Zhang, Q. Guo, J. Wang, Y. Lu: *Directed representative graph modeling of MEP systems using BIM data*, Buildings 12.6 (2022), 834:1–834:21.

[12] C. Hong, X. Ming, K. Liu, B. Tu, J. Yu: *Application of BIM Technology in Building Mechanical and Electrical Engineering Modeling and Pipeline Inspection*, in: IOP Conference Series: Earth and Environmental Science, vol. 719, 2021, p. 022018.

[13] M. Hubbell, A. Moran, W. Arcand, D. Bestor, B. Bergeron, C. Byun, V. Gadepally, P. Michaleas, J. Mullen, A. Prout, et al.: *Big Data strategies for Data Center Infrastructure management using a 3D gaming platform*, in: 2015 IEEE High Performance Extreme Computing Conference, 2015, pp. 1–6.

[14] J. Irizarry, M. Gheisari, G. Williams, K. Roper: *Ambient intelligence environments for accessing building information*, Facilities 32.3/4 (2014), pp. 120–138.

[15] S. Karim, N. Khalid, G. Murat, T. Onur, F. Faisal, Z. Tarek: *BIM-based facility management models for existing buildings*, Journal of Engineering Research 10.1A (2022), pp. 21–37.

[16] M. Levy, J. O. Hallstrom: *A new approach to data center infrastructure monitoring and management (DCIMM)*, in: 2017 IEEE 7th Annual Computing and Communication Workshop and Conference (CCWC), 2017, pp. 1–6.

[17] R. F. Moghaddam, V. Asghari, F. F. Moghaddam, Y. Lemieux, M. Cheriet: *A Monte-Carlo approach to lifespan failure performance analysis of the network fabric in modular data centers*, Journal of Network and Computer Applications 87 (2017), pp. 131–146.

[18] S. Pogorelskiy, I. Kocsis: *Automation for structured cabling system in data centers using Building Information Modelling*, International Review of Applied Sciences and Engineering 13.3 (2022), pp. 335–345.

[19] H. Qin: *The advantages of BIM application in EPC mode*, in: MATEC Web of Conferences, vol. 100, EDP Sciences, 2017, p. 05058.

[20] M. Rossella, N. Maurizio, P. Francesco, T. Andrej: *A methodology for a performance information model to support facility management*, Sustainability 11.24 (2019), 7007:1–7007:25.

[21] F. SA: *Organization of a management system for the operation of a data processing center*, Electron. Sci. J. Age Quality in Russian language 2 (2018), pp. 35–59.

[22] S. Saha, J. Sarkar, A. Dwivedi, N. Dwivedi, A. M. Narasimhamurthy, R. Roy: *A novel revenue optimization model to address the operation and maintenance cost of a data center*, Journal of Cloud Computing 5.1 (2016), 1:1–1:23.

[23] P. Sergey, K. Imre: *Efficiency Improvement with Data Center Monitoring Based on Building Information Modeling on the Facility Management Stage*, Designs 7.1 (2023), 3:1–3:15, doi: 10.3390/designs7010003.

[24] H. Tang, J. Cao, Z. Shao: *Network simulation and vulnerability analysis on organization of facility management*, in: 2017 IEEE International Conference on Systems, Man, and Cybernetics, 2017, pp. 2533–2538.

[25] P. Wang, L. Rao, X. Liu, Y. Qi: *D-Pro: Dynamic data center operations with demand-responsive electricity prices in smart grid*, IEEE Transactions on Smart Grid 3.4 (2012), pp. 1743–1754.

[26] W. Wei, L. Wenjia, L. Deify, N. Woonki: *Improving data center energy efficiency using a cyber-physical systems approach: integration of building information modeling and wireless sensor networks*, Procedia engineering 118 (2015), pp. 1266–1273, doi: 10.1016/j.proeng.2015.08.481.

[27] M. Wiboonrat, U. Kaewsiri: *A chronological transformation of data center project management*, in: 2014 World Automation Congress (WAC), 2014, pp. 173–178.

Annales Mathematicae et Informaticae
60 (2024) pp. 133–140
DOI: 10.33039/ami.2024.12.003
URL: https://ami.uni-eszterhazy.hu

On a combinatorial identity associated with Pascal's triangle

Monika Sviteková[a], **László Szalay**[b*]

[a]Jan Selye University
129684@student.ujs.sk

[b]Jan Selye University, University of Sopron
szalay.laszlo@uni-sopron.hu

Abstract. Let $f(x) = \cos x$, and consider the sum $\tau_n^{(f)} = \sum_k \binom{n-k}{k} f(kt)$. Using a general method due to Ahmia and Szalay on weighted sums in generalized Pascal triangle an explicit formula is developed for $\tau_n^{(f)}$. An analogous result is provided if $f(x) = \sin x$, and a strong connection to Fibonacci polynomials is also discovered in both cases.

Keywords: combinatorial identity, Pascal's triangle, weighted sum, Fibonacci polynomial

AMS Subject Classification: 05A19, 11B39

1. Introduction

Pascal's triangle is one of the most studied objects in combinatorics, in particular there exists a huge number of identities on binomial coefficients. We refer here the two closed forms

$$\begin{aligned} \sum_{k=0}^{n} \binom{n}{k} \cos kx &= 2^n \cos \frac{nx}{2} \left(\cos \frac{x}{2}\right)^n \quad \text{and} \\ \sum_{k=0}^{n} \binom{n}{k} \sin kx &= 2^n \sin \frac{nx}{2} \left(\cos \frac{x}{2}\right)^n, \end{aligned} \tag{1.1}$$

see, for example (1.26) and (1.27) in [6]. On the other hand, theory of linear recurrences often plays an important role in solving combinatorial and number

*The second author was supported by the Hungarian National Foundation for Scientific Research Grant No. 130909.

Submitted: September 17, 2024
Accepted: December 13, 2024
Published online: December 20, 2024

theoretical problems. The present paper describes a method to investigate the explicit form of the sum

$$\tau_n^{(f)} = \sum_k \binom{n-k}{k} f(kt), \tag{1.2}$$

where the function $f(x)$ is equal to either $\cos x$ or $\sin x$. The precise result (i.e. explicit formula) for $\tau_n^{(\cos)}$ and $\tau_n^{(\sin)}$ is presented in Theorem 2 of Section Results. Closely related identities appear in [1] in particular, identity (2.3), furthermore on pages 75–76 of the excellent book of Riordan [7].

This work is motivated essentially by the two identities (1.1). We suggest two approaches, the starting point of the first one is the paper of Ahmia and Szalay [3]. This provides a recurrence relation for the sums of binomial coefficients weighted by the terms of a given linear recurrence such that the binomial coefficients lay along an arbitrary final direction. Paper [3] is an extension of [4] where the principal theorem is able to handle arbitrary diagonal sums without weights in a generalized Pascal triangle. We believe that, using our approach analogous questions like $\sum_k \binom{n-k}{c_1+ck} \cos kt$ and $\sum_k \binom{n-k}{c_1+ck} \sin kt$ can be solved for small values of c such as $c = 2, 3$ with $0 \le c_1 < c$, and the nature of the result has similar flavor to Theorem 2. There are other possible ways to handle such questions, for example Egorychev method, i.e. application of complex integral representation of the binomial coefficients (see Part 7 of the compendium [2]). The advantage of our method is the flexibility in binomial coefficients in the above sums. On the other hand, our solution is not very efficient, it requires probably more calculations than other ways do.

Let $C_n = C_n(t) = \cos nt$ and $S_n = S_n(t) = \sin nt$ be considered as two sequences of functions. It is known that

$$C_n = (2\cos t)C_{n-1} - C_{n-2} \quad \text{and} \quad S_n = (2\cos t)S_{n-1} - S_{n-2}, \tag{1.3}$$

in other words, both sequences (C_n) and (S_n) satisfy the same binary recurrence rule with the initial values $C_0 = 1$, $C_1 = \cos t$, and $S_n = 0$, $S_1 = \sin t$, respectively. This observation ensures that we can apply the main result of [3] to the problem above. It turned out that the explicit formula we gained has a strong connection to Fibonacci polynomials, which are the basement of the second approach. The sequence of Fibonacci polynomials is defined by the initial polynomials $F_0(t) = 0$, $F_1(t) = 1$, and by the recurrence $F_n(t) = tF_{n-1}(t) + F_{n-2}(t)$. Comparing the results (of Theorem 2.1 and 2.2) provided by the two approaches we gain identities related to (1.2). We note that our method is applicable even, at least in theory, for functions $\mathcal{F}_k = f(kt)$ $(k = 0, 1, \dots)$ if the sequence $(\mathcal{F}_k)$ satisfies a homogeneous linear recurrence relation. In particular, using our approach we could prove (1.1) although this proof is not the simplest one.

In this paragraph, we introduce certain necessary notation and recall the aforementioned result from [3], which plays a crucial role in the investigation. Let x and y denote two non-zero real numbers. (This criteria appears in [4], but the statements remain true even if we allow for x and y to be complex numbers.) Assume

that the element of the generalized Pascal triangle located in the kth position of row n is $\binom{n}{k}x^{n-k}y^k$ ($n \in \mathbb{N}$, $0 \le k \le n$), and consider the sum

$$T_n = T_n^{(r,q,p)} = \sum_{k=0}^{\omega} \binom{n-qk}{p+rk} x^{n-p-(r+q)k} y^{p+rk},$$

where the parameters r, q and p satisfy the conditions $r \in \mathbb{N}^+$, $q \in \mathbb{Z}$, $r+q > 0$ and $0 \le p < r$, further $\omega = \lfloor (n-p)/(q+r) \rfloor$. Recall the worldwide known example, when $r = q = 1$, $p = 0$, $x = y = 1$. This choice admits $T_n^{(1,1,0)} = \sum_{k=0}^{\lfloor n/2 \rfloor} \binom{n-k}{k} = F_{n+1}$, where $F_n = F_{n-1} + F_{n-2}$, $F_0 = 0$, $F_1 = 1$ is the Fibonacci sequence. Clearly, the vector (r, q) and the non-negative integer p determine uniquely a finite ray crossing the generalized Pascal triangle if n is given. The paper of Belbachir, Komatsu and Szalay [4] gives a precise description on the sum T_n by showing that it satisfies the linear recurrence relation

$$T_n = \binom{r}{1} x T_{n-1} - \binom{r}{2} x^2 T_{n-2} + \cdots + (-1)^{r+1} \binom{r}{r} x^r T_{n-r} + y^r T_{n-r-q}. \quad (1.4)$$

Obviously, the order of the recurrence is $r+q$ if q is non-negative, and r otherwise. Observe, that (1.4) does not depend on p. Now we slightly modify the problem by including a sequence (G_n) to have the weighted sum

$$T_n = T_n^{(r,q,p),(G)} = \sum_{k=0}^{\omega} \binom{n-qk}{p+rk} x^{n-p-(r+q)k} y^{p+rk} G_k.$$

In general, the problem to describe the behavior of sequence $(T_n^{(r,q,p),(G)})$ is rather difficult, but if (G_n) is a homogeneous linear recursive sequence, then we can conclude Theorem 1.1.

Suppose that (G_n) is a real (or complex) linear homogeneous recurrence of order $s \in \mathbb{N}^+$ with given initial values $G_0, \ldots, G_{s-1}$ and with the defining identity $G_n = \sum_{j=1}^{s} A_j G_{n-j}$, where we assume $A_s \neq 0$.

Theorem 1.1 (Theorem 3.1 in [3]). *The terms*

$$T_n = \sum_{k=0}^{\omega} \binom{n-qk}{p+rk} x^{n-p-(r+q)k} y^{p+rk} G_k$$

satisfy the recurrence relation

$$T_n = xT_{n-1} - \sum_{j=1}^{rs-1} (-1)^j \binom{rs-1}{j} x^j (T_{n-j} - xT_{n-j-1})$$
$$+ \sum_{t=1}^{s} A_t y^{rt} \sum_{j=0}^{r(s-t)} (-1)^j \binom{r(s-t)}{j} x^j T_{n-(r+q)t-j}$$

for all $n \ge \max\{rs, (r+q)s\}$.

Since our application is restricted to the case of binary recurrences $G_n = A_1G_{n-1} + A_2G_{n-2}$, further when $r = q = 1$, $p = 0$ hold we have the following consequence of Theorem 1.1.

Corollary 1.2 (Corollary 3.3 in [3])**.** *The terms*

$$T_n = \sum_{k=0}^{\lfloor n/2 \rfloor} \binom{n-k}{k} x^{n-2k} y^k G_k$$

satisfy the recurrence relation

$$T_n = 2xT_{n-1} + (A_1y - x^2)T_{n-2} - A_1xyT_{n-3} + A_2y^2T_{n-4}. \tag{1.5}$$

Subsequently, equation (1.5) implies

$$T_n = 2T_{n-1} + (A_1 - 1)T_{n-2} - A_1T_{n-3} + A_2T_{n-4} \tag{1.6}$$

if $x = y = 1$.

2. Results

Let $z = \cos t + i \sin t \in \mathbb{C}$, where i is the imaginary unit. We also introduce the notation $\zeta = \sqrt{1+4z}$ with the condition $0 \le \arg(\zeta) \le \pi$. Clearly, the complex conjugate of ζ is

$$\overline{\zeta} = \overline{\sqrt{1+4z}} = -\sqrt{1+4\overline{z}}.$$

We study together the two cases $(G_n) = (C_n)$ and $(G_n) = (S_n)$ as far as it is possible. According to (1.3), the coefficients are $A_1 = 2\cos t$ and $A_2 = -1$ in the common binary recurrence rule. The characteristic polynomial of the recurrence (1.6) in this particular case is

$$p(X) = X^4 - 2X^3 - (2\cos t - 1)X^2 + (2\cos t)X + 1 = \prod_{j=1}^{4}(X - z_j), \tag{2.1}$$

where $z_1 = (1+\zeta)/2$, $z_2 = (1-\zeta)/2$, $z_3 = \overline{z_2}$, $z_4 = \overline{z_1}$. Note that all zeros are simple.

Using the notation above we can formalize the first result.

Theorem 2.1. *The identities*

$$\tau_n^{(\cos)} = \sum_{k=0}^{\lfloor n/2 \rfloor} \binom{n-k}{k} \cos kt = \frac{1}{2\zeta}(z_1^{n+1} - z_2^{n+1}) - \frac{1}{2\overline{\zeta}}(z_3^{n+1} - z_4^{n+1}),$$

$$\tau_n^{(\sin)} = \sum_{k=0}^{\lfloor n/2 \rfloor} \binom{n-k}{k} \sin kt = \frac{i}{2\zeta}(-z_1^{n+1} + z_2^{n+1}) - \frac{i}{2\overline{\zeta}}(z_3^{n+1} - z_4^{n+1})$$

hold.

Using Fibonacci polynomials and applying another method we can prove the following theorem.

Theorem 2.2. *We have the identities*

$$\tau_n^{(\cos)} = \sum_{k=0}^{\lfloor n/2\rfloor} \binom{n-k}{k} \cos kt = \frac{1}{2} z^{n/2} F_{n+1}\left(\frac{1}{\sqrt{z}}\right) + \frac{1}{2}\overline{z}^{n/2} F_{n+1}\left(\frac{1}{\sqrt{\overline{z}}}\right),$$

$$\tau_n^{(\sin)} = \sum_{k=0}^{\lfloor n/2\rfloor} \binom{n-k}{k} \sin kt = -\frac{i}{2} z^{n/2} F_{n+1}\left(\frac{1}{\sqrt{z}}\right) + \frac{i}{2}\overline{z}^{n/2} F_{n+1}\left(\frac{1}{\sqrt{\overline{z}}}\right).$$

Corollary 2.3. *The corresponding right-hand sides of the identities in Theorems 2.1 and 2.2 are equivalent.*

Remark 2.4. One may show directly, for instance, that

$$\frac{1}{2\zeta}(z_1^{n+1} - z_2^{n+1}) - \frac{1}{2\overline{\zeta}}(z_3^{n+1} - z_4^{n+1}) = \frac{1}{2} z^{n/2} F_{n+1}\left(\frac{1}{\sqrt{z}}\right) + \frac{1}{2}\overline{z}^{n/2} F_{n+1}\left(\frac{1}{\sqrt{\overline{z}}}\right),$$

but it would require some more calculation.

3. Proof of the theorems

Theorem 2.1. The definition of z_i $(i = 1, 2, 3, 4)$ implies $z_1 + z_2 = z_3 + z_4 = 1$. We also have $z_1 - z_2 = \zeta$ and $z_3 - z_4 = -\overline{\zeta}$, further $z_1 z_2 = -z$ and $z_3 z_4 = -\overline{z}$. Observe that $z + \overline{z} = 2\cos t$ and $z - \overline{z} = 2i \sin t$.

Since the zeros of the characteristic polynomial (2.1) are simple the fundamental theorem of linear recurrences (see [5, Theorem C.1.]) leads to the formulae

$$\tau_n^{(cos)} = \sum_{u=1}^{4} c_u z_u^n \quad \text{and} \quad \tau_n^{(sin)} = \sum_{u=1}^{4} d_u z_u^n, \tag{3.1}$$

where the coefficients c_u and d_u can be determined by verifying the above equations for $n = 0, 1, 2, 3$. This means that we must solve two systems of equations, each contains four linear equations with four unknowns.

First we find the powers of z_u in order to fix the system. Trivially, $z_1^0 = 1$, $z_1^1 = z_1$. Further

$$z_1^2 = \left(\frac{1+\zeta}{2}\right)^2 = \frac{1 + 2\zeta + \zeta^2}{4} = \frac{1 + 2\zeta + (1 + 4z)}{4} = \frac{1+\zeta}{2} + z = z_1 + z,$$

and $z_1^3 = (z_1 + z)z_1 = z_1^2 + z_1 z = z_1 + z + z_1 z$. Similar considerations admit the following more general result. For any $u = 1, 2$ we have

$$z_u^0 = 1, \quad z_u^1 = z_u, \quad z_u^2 = z_u + z, \quad z_u^3 = z_u + z + z_u z, \tag{3.2}$$

while

$$z_u^0 = 1, \quad z_u^1 = z_u, \quad z_u^2 = z_u + \overline{z}, \quad z_u^3 = z_u + \overline{z} + z_u\overline{z} \tag{3.3}$$

hold for $u = 3, 4$. Thus, according to (3.1) we have the system

$$\begin{aligned}
\tau_0^{(\cos)} &= 1 = c_1 + c_2 + c_3 + c_4, \\
\tau_1^{(\cos)} &= 1 = c_1 z_1 + c_2 z_2 + c_3 z_3 + c_4 z_4, \\
\tau_2^{(\cos)} &= 1 + \cos t = c_1 z_1^2 + c_2 z_2^2 + c_3 z_3^2 + c_4 z_4^2, \\
\tau_3^{(\cos)} &= 1 + 2\cos t = c_1 z_1^3 + c_2 z_2^3 + c_3 z_3^3 + c_4 z_4^3
\end{aligned}$$

of linear equations, where we replace the corresponding powers of z_u by (3.2) and (3.3) later. The Vandermonde determinant $V(z_1, z_2, z_3, z_4)$ of the system is non-zero, therefore we have a unique solution. The solution, after a few simplification steps, appears as

$$\begin{aligned}
c_1 &= -\frac{(\cos t - \overline{z})i}{2\zeta \sin t} z_1, & c_2 &= \frac{(\cos t - \overline{z})i}{2\zeta \sin t} z_2, \\
c_3 &= -\frac{(\cos t - z)i}{2\overline{\zeta} \sin t} z_3, & c_4 &= \frac{(\cos t - z)i}{2\overline{\zeta} \sin t} z_4.
\end{aligned}$$

Observe that $\cos t - \overline{z} = i \sin t$, and $\cos t - z = -i \sin t$. Hence $c_1 = z_1/(2\zeta)$, $c_2 = -z_2/(2\zeta)$, $c_3 = -z_3/(2\overline{\zeta})$, and $c_4 = z_4/(2\overline{\zeta})$. Combining these arguments with (3.1), finally we have

$$\sum_{k=0}^{\lfloor n/2 \rfloor} \binom{n-k}{k} \cos kt = \frac{1}{2\zeta}(z_1^{n+1} - z_2^{n+1}) - \frac{1}{2\overline{\zeta}}(z_3^{n+1} - z_4^{n+1}).$$

This proves the first part of Theorem 2.1.

Now turn our attention to the sum $\tau_n^{(sin)}$. In this case, the same machinery works, but now we have $\tau_0^{(sin)} = 0$, $\tau_1^{(sin)} = 0$, $\tau_2^{(sin)} = \sin t$, and $\tau_3^{(sin)} = 2 \sin t$ on the left-hand side of the current system of four equations. Of course, on the right-hand side we replace c_u by d_u. The solution is given by

$$d_1 = -\frac{i}{2\zeta} z_1, \; d_2 = \frac{i}{2\zeta} z_2, \; d_3 = -\frac{i}{2\overline{\zeta}} z_3, \; d_4 = \frac{i}{2\overline{\zeta}} z_4,$$

and we can conclude the second statement of the theorem immediately. □

Remark 3.1. Recall that $z = \cos t + i \sin t$. The characteristic polynomial $p(X)$ has the factorization

$$p(X) = X^4 - 2X^3 - (2\cos t - 1)X^2 + (2\cos t)X + 1 = (X^2 - X - z)(X^2 - X - \overline{z}).$$

This explains why we obtained two separated parts for both $\tau_n^{(cos)}$ and $\tau_n^{(sin)}$ in Theorem 2.1:

$$X^2 - X - z = (X - z_1)(X - z_2) \quad \text{and} \quad X^2 - X - \overline{z} = (X - z_3)(X - z_4).$$

Theorem 2.2. Let the polynomial sequence $(p(t))$ be defined by $p_0(t) = 0$, $p_1(t) = 1$, and $p_n(t) = p_{n-1}(t) + tp_{n-2}(t)$ for $n \geq 2$. The connection between these polynomials and Fibonacci polynomials can be given by $p_n(t) = t^{(n-1)/2}F_n(1/\sqrt{t})$. Indeed, it is true for $n = 0, 1$ and we see by induction that

$$\begin{aligned} t^{(n-1)/2}F_n(1/\sqrt{t}) &= t^{(n-1)/2}\frac{1}{\sqrt{t}}F_{n-1}(1/\sqrt{t}) + t^{(n-1)/2}F_{n-2}(1/\sqrt{t}) \\ &= t^{(n-2)/2}F_{n-1}(1/\sqrt{t}) + t \cdot t^{(n-3)/2}F_{n-2}(1/\sqrt{t}) \\ &= p_{n-1}(t) + tp_{n-2}(t) \\ &= p_n(t). \end{aligned}$$

Recall the explicit sum formula

$$F_{n+1}(t) = \sum_{k=0}^{\lfloor n/2 \rfloor} \binom{n-k}{k} t^{n-2k},$$

see, for example, [8]. Hence

$$p_{n+1}(t) = t^{n/2}F_{n+1}(1/\sqrt{t}) = t^{n/2}\sum_{k=0}^{\lfloor n/2 \rfloor} \binom{n-k}{k} t^{k-n/2} = \sum_{k=0}^{\lfloor n/2 \rfloor} \binom{n-k}{k} t^k,$$

i.e. the coefficients of polynomial $p_{n+1}(t)$ are the binomial coefficients $\binom{n-k}{k}$. Knowing that $z^k = \cos kt + i \sin kt$, we obtain immediately that

$$p_{n+1}(z) = \sum_{k=0}^{\lfloor n/2 \rfloor} \binom{n-k}{k} (\cos kt + i \sin kt).$$

Combining it with

$$p_{n+1}(\overline{z}) = \sum_{k=0}^{n/2} \binom{n-k}{k} (\cos kt - i \sin kt),$$

we have immediately

$$\begin{aligned} \sum_{k=0}^{\lfloor n/2 \rfloor} \binom{n-k}{k} \cos kt &= \frac{p_{n+1}(z) + p_{n+1}(\overline{z})}{2} \\ &= \frac{z^{n/2}F_{n+1}\left(\frac{1}{\sqrt{z}}\right) + \overline{z}^{n/2}F_{n+1}\left(\frac{1}{\sqrt{\overline{z}}}\right)}{2}, \end{aligned}$$

and

$$\begin{aligned} \sum_{k=0}^{\lfloor n/2 \rfloor} \binom{n-k}{k} \sin kt &= \frac{-p_{n+1}(z) + p_{n+1}(\overline{z})}{2} i \\ &= \frac{-z^{n/2}F_{n+1}\left(\frac{1}{\sqrt{z}}\right) + \overline{z}^{n/2}F_{n+1}\left(\frac{1}{\sqrt{\overline{z}}}\right)}{2} i. \end{aligned}$$

Now the proof is complete. □

References

[1] https://web.archive.org/web/20171118022028/http://www.math.wvu.edu/~gould/Vol.4.PDF, [Accessed 15-12-2024].

[2] https://web.archive.org/web/20171118022028/http://www.math.wvu.edu/~gould/Vol.7.PDF, [Accessed 15-12-2024].

[3] M. Ahmia, L. Szalay: *On the weighted sums associated to rays in generalized Pascal triangle*, Indian J. Discrete Math. 1 (2016), pp. 8–17.

[4] H. Belbachir, T. Komatsu, L. Szalay: *Linear recurrences associated to rays in Pascal's triangle and combinatorial identities*, Math. Slovaca 64 (2013), pp. 287–300.

[5] Z. Borevich, I. Shafarevich: *Number Theory*, New York, NY: Academic Press, 1966.

[6] H. W. Gould: *Combinatorial Identities, A Standardized Set of Tables Listing 500 Binomial Coefficient Summations. Revised Edition*, published by the author, 1972.

[7] J. Riordan: *Combinatorial identities. Reprint edition with corrections*, New York, NY: R. E. Krieger Publishing Company, 1979.

[8] M. N. S. Swamy: *Further properties of Morgan-Voyce polynomials*, Fibonacci Quart. 6 (1968), pp. 167–175.

Annales Mathematicae et Informaticae
60 (2024) pp. 141–150
DOI: 10.33039/ami.2024.02.007
URL: https://ami.uni-eszterhazy.hu

Effective inclusion methods for verification of ReLU neural networks

Attila Szász, Balázs Bánhelyi

University of Szeged, Institute of Informatics
{szasz,banhelyi}@inf.u-szeged.hu

Abstract. The latest machine learning models are sensitive to adversarial inputs, i.e., the neural network can give incorrect results even with small changes in the learning case. To avoid this, techniques are used during learning, or verification is also possible. In many cases, these methods use interval arithmetic, whose usefulness is severely limited by overestimation. In this paper, we present and compare such methods that can handle this problem.

Keywords: artificial neural network, verification, interval arithmetic, symbolic calculations

AMS Subject Classification: 68T07, 68N30, 65G30

1. Introduction

One of the most important topics in artificial intelligence research today is the verification of neural networks. The accuracy of the networks has continuously increased over the years, so more and more complex neural networks have been created and many tasks could be solved by them. Many modern teaching methods have been developed that have improved the quality of the networks. In certain fields, it is inevitable to be precise and to have fast networks.

In many works, it has been shown that these nets, which are considered to be safe, can also be wrong [10]. In many cases, noise on the input that is invisible to the human eye can lead to a wrong classification. There are many methods developed by researchers to solve these problems. The methods are mainly divided into 2 classes: robust learning and adversary example detection.

For this reason, neural network verification is an important topic in today's artificial intelligence research. The neural network technique focuses on speed and typically uses floating point arithmetic, while others prefer symbolic methods [15]

Submitted: July 17, 2023
Accepted: February 6, 2024
Published online: February 14, 2024

used for reliability. Other important family of deep neural networks, the Binarized Neural Networks (BNN) [4, 6], that are similar to regular feedforward neural networks. One difference is that the weights and activations in a BNN are constrained to be only two values: 1 and -1, which implied other verification technique.

The standard numerical systems often have significantly longer run times [2, 11, 13]. In this paper, other methods have been described using the numerical result. These methods have a correct evaluation and a manageable runtime. The system we wrote defines not only the inputs but also the interval of values for the outputs of the given network. When verifying the robustness of a neural network, these intervals must be as small as possible. During the evaluation, in addition to the nets with the ReLU activation function, the output widths and the running times were also compared.

2. Motivation

During our work, we have developed a system based on reliable network assessment. The system supports multiple evaluation methods, in both CPU and GPU environments. Many current contemporary systems use floating-point arithmetic with an emphasis on speed in evaluation. The big disadvantage of this is that some numerical errors in the various operations can accumulate during the evaluation [16].

A good way to get a handle on these errors is to use interval arithmetic [1, 5]. One solution is to compute an interval containing the given value instead of the floating-point number. The method is well suited for neural networks and also for their evaluation since the operations that can be performed on real numbers can be easily extended to intervals. In this case, the inputs of the network are intervals. One of the advantages of the method is that it increases the running time only minimally compared to floating point evaluation. Since it is reliable and robust in class, this method is also used in the evaluation phase.

At the last ICAI conference, the first results of an interval-based verification algorithm [3]. In this approach, a simple natural interval expansion was used to compute inclusion functions. This mod was able to include the values of the output neurons in the input interval. This proves the correct result for each point of the input.

The network shown in Figure 1 is evaluated according to the rules of naive interval arithmetic. The ReLU activation function is contained in some neurons of the network. The value of the output neuron x_5 lies in the interval $[0; 5]$. It can be observed that the upper limit of 5 would occur only if the neuron x_3 had the value 5 and the neuron x_4 had the value 6. It can be seen that under the given input conditions, these values can never occur simultaneously. As a result, the upper bound of x_5 was never sharp, leading to an overestimation in the evaluation. The main reason for the overestimation is the dependency problem, which can become quite strong and unmanageably wide in the results as the number of layers increases.

The goal of our work is to implement and compare numerically correct systems that handle the dependency problem but do not drastically increase the runtime.

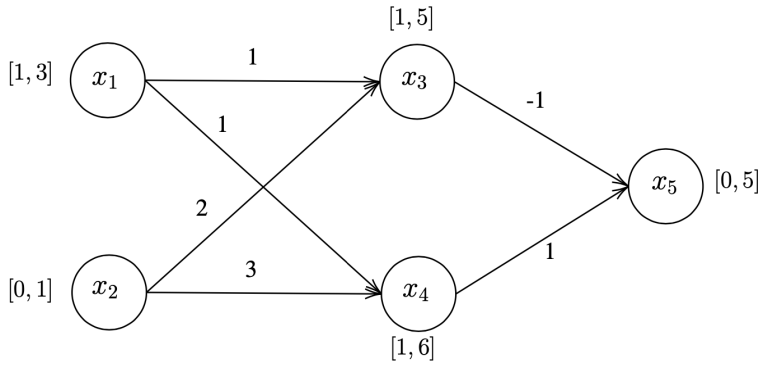


Figure 1. Naive interval arithmetic network evaluation.

3. Methods

In this work, three techniques were presented, each based on the use of a 1-1 function for the set of values of neurons on the given input set. The first technique is symbolic propagation [12], where we search for the best-fitting function and then handle the nonlinearity with new variables. In the second case, 2 separate functions were held for each neuron and calculated with them. In the third case, the well-known linear affine expression was used.

3.1. Symbolic function propagation

In our system, the overestimates resulting from the naive evaluation have been effectively handled. The method has a ReLU activation function that is fully feedforward and was used to evaluate networks with connected layers. The main feature of the system is that the evaluation captures dependency relations rather than the values of the current layer. In addition to a neuron of the input layer x_1 and x_2, the function $x_3 = x_1 + 2x_2$ is recorded for the neuron x_3 (Figure 1). When the network is evaluated, the corresponding function is determined for all neurons, and the result of its evaluation in intervals is the set of values of the neuron.

The transformations between layers can be easily extended to the functional representation of neurons. The result of the product of a given function x and a constant w is the function xw, which is obtained by multiplying all coefficients of the function x by the constant w. For the sum of the functions, we calculate a function formed by the sum of the coefficients. Since we examined ReLU networks in the evaluation, the activation function had to be extended to functions as well. To calculate the transformation, we used the method in the RefineZono article [8].

In the operation, we can distinguish 3 cases. If the lower bound of the interval is greater than 0, then ReLU is return with an original function, since there is no cut. If the upper bound is less than or equal to zero, then ReLU nullifies all coefficients of the input function, ensuring that the output is reduced to 0. On the other hand, if the input function contains the value 0, the symbolic function $\hat{y} = \hat{x} + x_{\text{new}}$ is calculated from the input interval and the calculated output (see Figure 2).

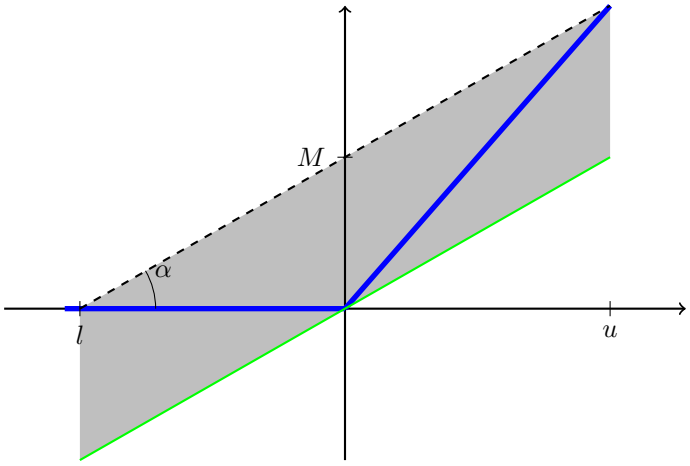


Figure 2. ReLU inclusion with Symbolic propagation.

The first step in determining y is the green line through 0 defined in Figure 2. To do this, we first calculate the slope of the line, which is arbitrarily $[l; u]$ in the case of an interval:

$$\lambda = \frac{u}{u - l}.$$

Then the coefficients of the input function x were multiplied by λ, and the result is a line passing through zero, which is the lower bound of inclusion. To determine the upper bound, the function $x + M$ must be calculated. For the result of the inclusion to be a function, a new variable is introduced. The output can be easily handled in the form of the function, only the coefficients need to be stored. Also, by introducing the new variables, we can handle a larger dependency. The value interval of the new variable can be calculated with the following formula:

$$x_{\text{new}} = [0, \lambda|l|].$$

The result of the transformation will be a new function $\hat{y} = \lambda\hat{x} + x_{\text{new}}$.

3.2. ReLU inclusion with 2 functions

The major advantage of the inclusion shown in Figure 2 is that it can be computed quickly, and the newly introduced variable makes it easier to handle dependency relationships in the computation. However, a disadvantage is that the inclusion

does not guarantee that the 0 lower bounds will be overestimated in the calculation of the lower bound. To deal with this, we also investigated an evaluation in which we maintain separate lower and upper inclusion functions for each neuron. The investigated to determine the value set of a neuron, both functions must be evaluated. The lower bound of the lower inclusion function gives the lower bound of the neuron. For the upper bound, we take the upper bound of the inclusion function.

In this case, we also had to handle the ReLU transformation. ReLU's input, in this case, was 2 functions that were evaluated to get the boundaries. If the lower bound is at least 0, the transformation leaves both functions unchanged. If the upper bound was less than 0, then the coefficients of the boundary function were set to zero by zero to obtain the interval $[0; 0]$. In the case where the interval encompasses zero, the inclusion shown in Figure 2 is calculated. By setting the coefficients of the lower bound function to zero, the lower bound 0 can be ensured so that no overestimation occurs here. Determining the upper bound for the method is shown in Figure 3 and proceeds similarly. First, we determine the value by which we multiply the above coefficients of the function λ, then we add the value of the shift M. Then we get the transformation:

$$\hat{y}_{\text{lower}} = 0,$$
$$\hat{y}_{\text{upper}} = \hat{x}\lambda + M.$$

The advantage of inclusion is that a zero lower bound can be guaranteed for output. However, the disadvantage is that a new variable was not introduced and use two separate limit functions, so the dependence information registered between neurons is reduced and the computing demand increases.

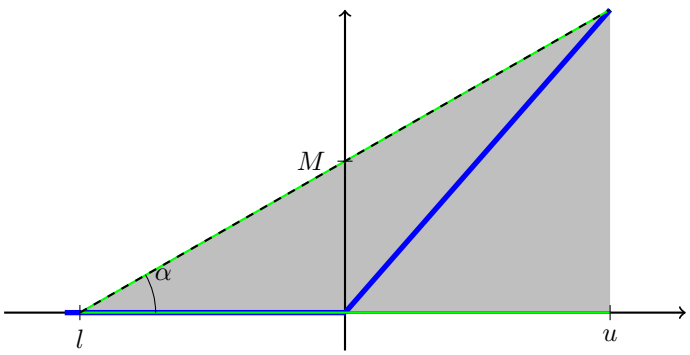


Figure 3. Layer transformation for separate boundary functions.

3.3. Affine representation of neurons

To further manage dependency information, we also explored a method in which input intervals were recorded in the affine form [7]. Then, any input x was treated

in the following form:

$$\hat{x} = x_0 + x_1\epsilon_1 \cdots x_n\epsilon_n.$$

The coefficients x_i are floating point values, and the ϵ_i are independent in the $[-1;1]$ interval is given. The x_0 value is used as the mean, and the ϵ_i are to be referred to as noise symbols.

The conversion between the affine and interval forms is easy to write. A x result interval in the case of an affine expression can be given in the following form: x_0 indicates the center of the interval and the radius can be calculated from the absolute sum of the corresponding coefficients.

$$x = [x_0 - r, x_0 + r], \text{ where } r = \sum_{i=1}^{n} |x_i|.$$

For an interval $x = [a;b]$ the corresponding affine form x can be calculated with the following relation: $\hat{x} = x_0 + x_k\epsilon_k$, where $x_0 = \frac{a-b}{2}$, $x_k = \frac{b-a}{2}$ and ϵ_k is a new variable.

The layer transformations can be easily extended in this case as well. The result of the product of an affine expression x and the constant α is an affine form where the coefficients of x have been multiplied by the value α. The sum of an affine expression and a constant value is also an affine form where the mean x_0 is increased by the specified number.

An extension of the ReLU transformation similar to the method shown in Figure 3. In the first step, we calculate the slope by which we multiply the coefficients of the input function. On the other hand, when introducing the new variables, so that the values remain in the range $[-1;1]$, we perform the following transformation:

$$[0, \lambda|l|] = \epsilon_k \left[0, \frac{\lambda|l|}{2}\right] + \frac{\lambda|l|}{2}, \text{ where } \epsilon_k = [-1;1].$$

Two methods have been used to handle numerical errors. One solution is the coefficients of the affine expression x are stored with interval inclusion and in the evaluation, and the transformations were calculated according to the rule of interval arithmetic. The other solution is to expand our expression with a new z_k k-error term during each operation. Here z_k is the upper bound of the absolute error in the coefficients of the affine expression and k is a completely new symbol.

4. Results

During the evaluation, our goal was to make as diverse as possible examples. Our methods were tested on different sizes, types, and quality networks. For this reason, the networks were examined on those published in the ERAN [9] system. 6 networks from the set, on which we also studied the runtime, output widths, and robustness. The parameters of each network are listed in Table 1. All trained networks have a ReLU activation function.

Table 1. Parameters of the ERAN networks.

Layer number	Layer width	Training method
3	50	regular
3	100	regular
4	1024	regular
6	500	regular
6	500	PGD (0.1) robust
6	500	PGD (0.3) robust

4.1. Results for 0.02 wide input intervals

The average total output width of the symbolic propagation (Table 3) was 0.41. For the symbolic system (Table 2), this value is 210, which is about a 510 times improvement over the naive method. The most significant reduction in width is the PGD (0.3) teaching network, where the improvement was nearly 3000 times. Overall, symbolic propagation produced the narrowest results in this case as well. Using the affine method (Table 4) produced an average improvement of about 108 times compared to the naive method. However, for the network with the most neurons (4X1024), we obtained outputs that were about 7 times wider than those of the symbolic propagation next to it. The results of the solution with separate constraint functions are visible in Table 5. Compared to the naive method, the average improvement was about 10 times.

Table 2. Naive interval arithmetic.

Network	Training	ϵ		Mean time(s)	
		$\epsilon = 0.02$	$\epsilon = 0.1$	CPU	GPU
3×50	regular	4.70	24.31	0.0006	0.0005
3×100	regular	9.62	53.41	0.001	0.0005
4×1024	regular	1088.51	5385.14	0.058	0.00089
6×500	regular	35.15	315.58	0.042	0.0012
6×500	PGD (0.1) robust	96.23	444.36	0.027	0.0012
6×500	PGD (0.3) robust	29.60	170.57	0.059	0.0012

4.2. Results for 0.1 wide input intervals

With wider input bounds, we obtain almost 578 times narrower intervals with symbolic propagation, than with the naive method. With the affine method, this quotient reduces to about 6. The average width with separate limit functions is 420, which turns out to be about 2.5 times narrower than with the naive method.

Table 3. Symbolic function propagation.

Network	Training	ϵ		Mean time(s)	
		$\epsilon = 0.02$	$\epsilon = 0.1$	CPU	GPU
3×50	regular	0.62	2.5	0.0038	0.0021
3×100	regular	0.56	2.41	0.006	0.0021
4×1024	regular	1.20	4.85	0.092	0.0046
6×500	regular	0.08	0.41	0.076	0.0046
6×500	PGD (0.1) robust	0.02	0.10	0.058	0.0047
6×500	PGD (0.3) robust	0.01	0.11	0.094	0.0046

Table 4. Affine propagation.

Network	Training	ϵ		Mean time(s)	
		$\epsilon = 0.02$	$\epsilon = 0.1$	CPU	GPU
3×50	regular	0.89	14.14	0.0045	0.0016
3×100	regular	1.06	28.48	0.0059	0.0016
4×1024	regular	9.35	843.41	0.093	0.0035
6×500	regular	0.19	131.77	0.061	0.0039
6×500	PGD (0.1) robust	0.06	15.67	0.056	0.0040
6×500	PGD (0.3) robust	0.05	13.39	0.062	0.0037

Table 5. Separated propagation.

Network	Training	ϵ		Mean time(s)	
		$\epsilon = 0.02$	$\epsilon = 0.1$	CPU	GPU
3×50	regular	1.15	12.83	0.179	0.0023
3×100	regular	1.85	26.91	0.325	0.0024
4×1024	regular	123.84	2260.3	11.10	0.0064
6×500	regular	2.83	101.89	25.21	0.0061
6×500	PGD (0.1) robust	3.67	82.27	5.99	0.0061
6×500	PGD (0.3) robust	1.36	36.38	50.91	0.006

4.3. Effect of robust training

For 6-layer networks with 500 neurons, we compared how robust training affects the average output width. During the test, we calculated the robust of the average output widths of the simple trained network and the robustly trained network, which represents the degree of improvement compared to the simple training method. Table 6 shows the symbolic propagation (S_ IA), the affine method (S_ AFF), and the results of the evaluation under separate boundary functions (S_ SEP). Robust teaching had a positive effect on the increase in initial width in almost all cases. The reason is that the networks trained in this way are prepared for a given

input change in their environment.

Table 6. Roboust teaching effect.

Methods	Training	ϵ	Mean improvement
S_IA	PGD(0.1)	0.02	3.46
		0.1	3.88
	PGD(0.3)	0.02	9.40
		0.1	10.95
S_AFF	PGD(0.1)	0.02	3.27
		0.1	10.80
	PGD(0.3)	0.02	16.10
		0.1	290.90
S_AFF	PGD(0.1)	0.02	0.74
		0.1	1.23
	PGD(0.3)	0.02	13.59
		0.1	3.49

5. Conclusion

In this work, we demonstrate the effectiveness of different techniques for multiple neural networks. The CPU/GPU time of the algorithms was shown on self-trained networks of different sizes and on ERAN networks. We will also separately explain how the computation time evolves in the case of networks trained with other robust techniques. We hope that these methods will be more effective in training than the interval method used robust training [14].

Acknowledgements. Support by the the European Union project RRF-2.3.1-21-2022-00004 within the framework of the Artificial Intelligence National.

References

[1] G. Alefeld, G. Mayer: *Interval analysis: theory and applications*, Journal of Computational and Applied Mathematics 121.1 (2000), pp. 421–464, ISSN: 0377-0427, DOI: 10.1016/S0377-0427(00)00342-3, URL: https://www.sciencedirect.com/science/article/pii/S0377042700003423.

[2] R. Bunel, J. Lu, I. Turkaslan, P. H. S. Torr, P. Kohli, M. P. Kumar: *Branch and Bound for Piecewise Linear Neural Network Verification*, J. Mach. Learn. Res. 21 (2019), 42:1–42:39.

[3] T. Csendes, N. Balogh, B. Bánhelyi, D. Zombori, R. Tóth, I. Megyeri: *Adversarial Example Free Zones for Specific Inputs and Neural Networks*, in: International Conference on Applied Informatics, 2020.

[4] G. Kovásznai, K. Gajdár, N. Narodytska: *Portfolio solver for verifying binarized neural networks*, Annales Mathematicae et Informaticae 53 (2021), Cited by: 0; All Open Access, Bronze Open Access, Green Open Access, pp. 183–200, doi: 10.33039/ami.2021.03.007.

[5] R. Moore, R. Kearfott, M. Cloud: *Introduction To Interval Analysis*, Cambridge Uni Press (CUP), 2009.

[6] N. Narodytska, S. Kasiviswanathan, L. Ryzhyk, M. Sagiv, T. Walsh: *Verifying Properties of Binarized Deep Neural Networks*, Proceedings of the AAAI Conference on Artificial Intelligence 32 (Sept. 2017), doi: 10.1609/aaai.v32i1.12206.

[7] E. Remm, M. Goze: *Affine structures on abelian Lie groups*, Linear Algebra and its Applications 360 (2003), pp. 215–230, issn: 0024-3795, doi: 10.1016/S0024-3795(02)00452-4, url: https://www.sciencedirect.com/science/article/pii/S0024379502004524.

[8] G. Singh, T. Gehr, M. Püschel, M. T. Vechev: *Boosting Robustness Certification of Neural Networks*, in: 7th International Conference on Learning Representations, ICLR 2019, New Orleans, LA, USA, May 6-9, 2019, OpenReview.net, 2019, url: https://openreview.net/forum?id=HJgeEh09KQ.

[9] G. Singh, T. Gehr, M. Püschel, M. T. Vechev: *Boosting Robustness Certification of Neural Networks*, in: 7th International Conference on Learning Representations, ICLR 2019, New Orleans, LA, USA, May 6-9, 2019, OpenReview.net, 2019, url: https://openreview.net/forum?id=HJgeEh09KQ.

[10] C. Szegedy, W. Zaremba, I. Sutskever, J. Bruna, D. Erhan, I. J. Goodfellow, R. Fergus: *Intriguing properties of neural networks*, in: 2nd International Conference on Learning Representations, ICLR 2014, Banff, AB, Canada, April 14-16, 2014, Conference Track Proceedings, ed. by Y. Bengio, Y. LeCun, 2014, url: http://arxiv.org/abs/1312.6199.

[11] V. Tjeng, K. Y. Xiao, R. Tedrake: *Evaluating Robustness of Neural Networks with Mixed Integer Programming*, in: International Conference on Learning Representations, 2017.

[12] S. Wang, K. Pei, J. Whitehouse, J. Yang, S. Jana: *Formal Security Analysis of Neural Networks Using Symbolic Intervals*, in: Proceedings of the 27th USENIX Conference on Security Symposium, SEC'18, Baltimore, MD, USA: USENIX Association, 2018, pp. 1599–1614, isbn: 9781931971461.

[13] S. Wang, K. Pei, J. Whitehouse, J. Yang, S. S. Jana: *Efficient Formal Safety Analysis of Neural Networks*, in: Neural Information Processing Systems, 2018.

[14] K. Xiao, V. Tjeng, N. Shafiullah, A. Mądry: *Training for faster adversarial robustness verification via inducing Relu stability*, in: International Conference on Learning Representations, May 2019.

[15] X. Xie, K. Kersting, D. Neider: *Neuro-Symbolic Verification of Deep Neural Networks*, in: Proceedings of the Thirty-First International Joint Conference on Artificial Intelligence, IJCAI-22, ed. by L. D. Raedt, Main Track, International Joint Conferences on Artificial Intelligence Organization, July 2022, pp. 3622–3628, doi: 10.24963/ijcai.2022/503.

[16] D. Zombori, B. Bánhelyi, T. Csendes, I. Megyeri, M. Jelasity: *Fooling a Complete Neural Network Verifier*, in: International Conference on Learning Representations, 2021.

Annales Mathematicae et Informaticae
60 (2024) pp. 151–158
DOI: 10.33039/ami.2024.05.001
URL: https://ami.uni-eszterhazy.hu

A note on the Bricard property of projective planes

Zoltán Szilasi

Institute of Mathematics, University of Debrecen
H-4010, Debrecen, Hungary
szilasi.zoltan@science.unideb.hu

Abstract. We show that the Bricard property does not hold in every Moufang plane.

Keywords: projective planes, Moufang planes, Bricard property, alternative division rings, octonions

AMS Subject Classification: 51A20; 51A25; 51A30; 51E15; 51A35; 51A05; 17D05

1. Preliminaries

Concerning preliminaries we refer to [1, 4, 5], however, for convenience of the reader we recall some basic definitions and theorems.

An incidence geometry $(\mathcal{P}, \mathcal{L}, \mathcal{I} \subset \mathcal{P} \times \mathcal{L})$ is a *projective plane* if

(P1) for every pair of distinct points A and B there is a unique line incident with A and B (we denote this line by $\overleftrightarrow{AB}$);

(P2) for every pair of distinct lines m and n there is a unique point incident with m and m (we denote this point by $m \cap n$);

(P3) there are four points no three of which are collinear.

In a projective plane an ordered triple of noncollinear points is a *triangle*. Then the points are called the *vertices*, and the lines joining the three possible distinct pairs of vertices are called *sides*.

We say that two triangles ABC and $A'B'C'$ are *centrally perspective* from a point O if the lines $\overleftrightarrow{AA'}$, $\overleftrightarrow{BB'}$ and $\overleftrightarrow{CC'}$ are incident with O. The triangles are

Submitted: February 5, 2024
Accepted: May 8, 2024
Published online: May 16, 2024

called *axially perspective* from a line l if the points $\overleftrightarrow{AB} \cap \overleftrightarrow{A'B'}$, $\overleftrightarrow{AC} \cap \overleftrightarrow{A'C'}$ and $\overleftrightarrow{BC} \cap \overleftrightarrow{B'C'}$ are incident with l. A projective plane is *Dersarguesian*, if *any two triangles that are perspective from a point are perspective from a line.* This holds of and only if it can be coordinatized by a skewfield.

In this paper we focus on the *Bricard property* of projective planes:

Let ABC and $A'B'C'$ be two triangles, and let $P := \overleftrightarrow{BC} \cap \overleftrightarrow{B'C'}$, $Q := \overleftrightarrow{AC} \cap \overleftrightarrow{A'C'}$ and $R := \overleftrightarrow{AB} \cap \overleftrightarrow{A'B'}$. If $\overleftrightarrow{A'P}$, $\overleftrightarrow{B'Q}$ and $\overleftrightarrow{C'R}$ are concurrent, then $D := \overleftrightarrow{BC} \cap \overleftrightarrow{AA'}$, $E := \overleftrightarrow{AC} \cap \overleftrightarrow{BB'}$ and $F := \overleftrightarrow{AB} \cap \overleftrightarrow{CC'}$ are collinear.

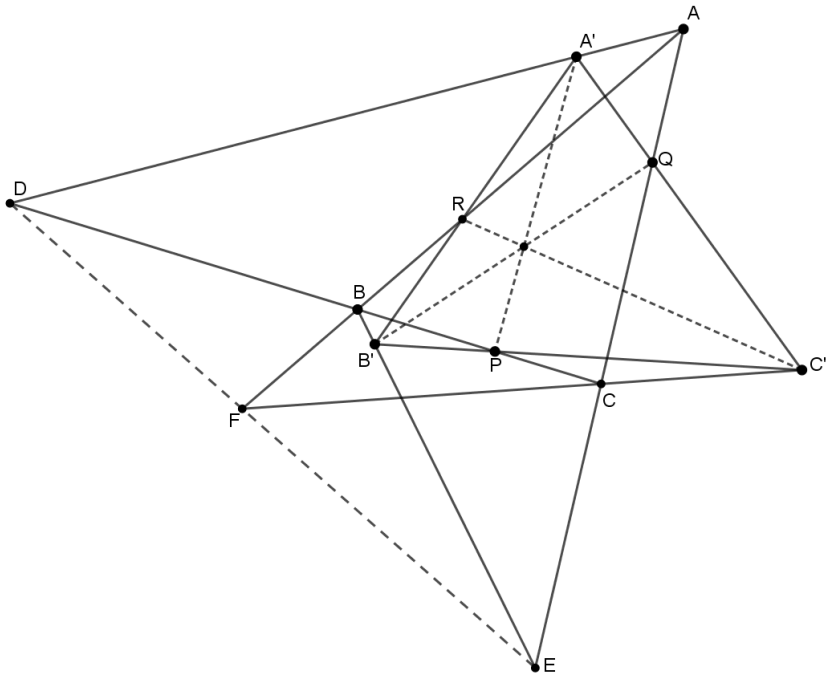


Figure 1. The Bricard property.

In [3] it is shown that the Bricard property follows from the Desargues property.

It is an open question if the Desargues property is necessary in a projective plane to satisfy the Bricard property. The author of [3] conjectures that the Bricard property follows from the following weaker version of the Desargues property:

(D9): *If the triangles $A_1B_1C_1$ and $A_2B_2C_2$ are perspective from a point O, and the triplets (A_1, B_2, C_1) and (A_2, B_1, C_2) are collinear, then the two triangles are perspective from a line.*

In [5] we proved that the converse of the Bricard property does not necessarily hold even under the following, somewhat stronger condition, which is valid in *Moufang planes*:

(D10): *If two triangles $A_1B_1C_1$ and $A_2B_2C_2$ are perspective from a point O, and O is incident to the line of $\overleftrightarrow{A_1B_1} \cap \overleftrightarrow{A_2B_2}$ and $\overleftrightarrow{A_1C_1} \cap \overleftrightarrow{A_2C_2}$, then the triangles are perspective from a line.*

However, it is unknown whether the Bricard property and its converse are equivalent, therefore it is still unknown if the Bricard property hold in every Moufang plane, or every projective plane satisfying (D10). In this paper we prove that neither (D9), nor (D10) implies the Bricard property, as we provide a counterexample for the Bricard property in a Moufang plane.

We recall that a projective plane is a Moufang plane if and only if it can be coordinatized by an alternative division ring, i.e., it is isomorphic to a projective plane over an alternative division ring. We recall that a triplet $(\mathcal{R}, +, \cdot)$ (briefly $\mathcal{R}$) is called an alternative division ring if

Let $\mathcal{R}$ be a set and $+$, $\cdot$ be binary operations on $\mathcal{R}$ such that

- $(\mathcal{R}, +)$ is a commutative group with zero element 0;
- $a \cdot 0 = 0 \cdot a = 0$ for all $a \in \mathcal{R}$;
- $(\mathcal{R}\backslash\{0\}, \cdot)$ is a loop (for a definition, see, e.g., [2]);
- $a \cdot (b + c) = a \cdot b + a \cdot c$,
- $(a + b) \cdot c = a \cdot c + b \cdot c$,
- $a \cdot (a \cdot b) = (a \cdot a) \cdot b$,
- $a \cdot (b \cdot b) = (a \cdot b) \cdot b$; $a, b, c \in \mathcal{R}$.

In the following we will write simply ab instead of $a \cdot b$. We denote the unit of $(\mathcal{R}\backslash\{0\}, \cdot)$ by 1. In an alternative division ring for all $a \in \mathcal{R}\backslash\{0\}$ there exists a unique element a^{-1} such that $aa^{-1} = a^{-1}a = 1$, canned the inverse of a. By a difficult theorem of Bruck-Kleinfield and Skornyakov, an alternative division ring either is associative or is a Cayley-Dickson algebra over some field. From this it follows that in every alternative division ring we have the *inverse property*

$$a(a^{-1}b) = (ba^{-1})a = b \quad \text{for all} \quad a \in \mathcal{R} \setminus \{0\}, b \in \mathcal{R},$$

since this holds in every Cayley-Dickson algebra.

Let $\mathcal{R}$ be an alternative division ring. The incidence structure $(\mathcal{P}, \mathcal{L}, \mathcal{I})$, where

- $\mathcal{P} := \{[x, y, 1], [1, x, 0], [0, 1, 0] \mid x, y \in \mathcal{R}\}$;
- $\mathcal{L} := \{\langle a, 1, b\rangle, \langle 1, 0, a\rangle, \langle 0, 0, 1\rangle \mid a, b \in \mathcal{R}\}$;
- $([x, y, z], \langle a, b, c\rangle) \in \mathcal{I}$ if and only if $xa + yb + zc = 0$

is a projective plane called *the projective plane over the alternative division ring* $\mathcal{R}$.

The most simple example of an alternative division ring that is not a skewfield is the alternative division ring of *octonions*. They can be constructed by the Cayley-Dickson procedure from the ring of quaternions. An octonion can be written in form

$$x = x_0 + x_1 i + x_2 j + x_3 k + x_4 l + x_5 I + x_6 J + x_7 K,$$

where x_i ($i \in \{0, 1, 2, 3, 4, 5, 6, 7\}$) are real numbers, and the rule of multiplication of the basic elements i, j, k, l, I, J, K is given by the the following table:

	i	j	k	l	I	J	K
i	-1	l	K	$-j$	J	$-I$	$-k$
j	$-l$	-1	I	i	$-k$	K	$-J$
k	$-K$	$-I$	-1	J	j	$-l$	i
l	j	$-i$	$-J$	-1	K	k	$-I$
I	$-J$	k	$-j$	$-K$	-1	i	l
J	I	$-K$	l	$-k$	$-i$	-1	j
K	k	J	$-i$	I	$-l$	$-j$	-1

The conjugate of $x = x_0 + x_1 i + x_2 j + x_3 k + x_4 l + x_5 I + x_6 J + x_7 K$ is

$$\overline{x} := x_0 - x_1 i - x_2 j - x_3 k - x_4 l - x_5 I - x_6 J - x_7 K,$$

and the norm of x is

$$\|x\| := \sqrt{x_0^2 + x_1^2 + x_2^2 + x_3^2 + x_4^2 + x_5^2 + x_6^2 + x_7^2}.$$

Then the inverse of x is

$$x^{-1} = \frac{\overline{x}}{\|x\|^2}.$$

The projective plane over the octonions is called the *octonion plane*.

2. A counterexample for the Bricard property in the octonion plane

Theorem 2.1. *The Bricard property does not hold in every Moufang plane.*

Proof. Consider the following triangles ABC and $A'B'C'$ in the octonion plane:

$$A'[1,0,0],\ B'[0,1,0],\ C'[0,0,1];$$
$$A\left[-\frac{1}{2}+\frac{1}{2}i-\frac{1}{2}k+\frac{1}{2}K, -\frac{1}{2}-\frac{1}{2}i-\frac{1}{2}k+\frac{1}{2}K, 1\right],$$
$$B\left[\frac{1}{2}+\frac{1}{2}i-j-l, \frac{1}{2}-\frac{1}{2}i-j-l, 1\right],$$

$$C\left[\frac{1}{2}-\frac{1}{2}j-\frac{1}{2}k+\frac{1}{2}I,\frac{1}{2}-\frac{1}{2}j-\frac{1}{2}k-\frac{1}{2}I,1\right].$$

It is easy to see that A is incident to $\langle -1,1,i\rangle$ and $\langle -k,1,k\rangle$; B is incident to $\langle -1,1,i\rangle$ and $\langle j,1,-1\rangle$; C is incident to $\langle j,1,-1\rangle$ and $\langle -k,1,k\rangle$, therefore

$$\overleftrightarrow{AB}=\langle -1,1,i\rangle,\ \overleftrightarrow{BC}=\langle j,1,-1\rangle,\ \overleftrightarrow{AC}=\langle -k,1,k\rangle.$$

Since

$$\overleftrightarrow{A'B'}=\langle 0,0,1\rangle,\ \overleftrightarrow{B'C'}=\langle 1,0,0\rangle,\ \overleftrightarrow{A'C'}=\langle 0,1,0\rangle,$$

we get

$$P=[0,1,1],\ Q=[1,0,1],\ R=[1,1,0].$$

Therefore $\overleftrightarrow{A'P}$, $\overleftrightarrow{B'Q}$ and $\overleftrightarrow{C'R}$ are concurrent at the point $O[1,1,1]$.

We are going to show that the points $D:=\overleftrightarrow{BC}\cap\overleftrightarrow{AA'}$, $E:=\overleftrightarrow{AC}\cap\overleftrightarrow{BB'}$ and $F:=\overleftrightarrow{AB}\cap\overleftrightarrow{CC'}$ are not collinear.

To obtain the coordinates of D, first we determine the line $\overleftrightarrow{AA'}$. Since $[1,0,0]$ is incident to it, it is of the form $\overleftrightarrow{AA'}=\langle 0,1,e\rangle$ for some octonion e. As the point A is incident to the line, we get

$$-\frac{1}{2}-\frac{1}{2}i-\frac{1}{2}k+\frac{1}{2}K+e=0,$$
$$e=\frac{1}{2}+\frac{1}{2}i+\frac{1}{2}k-\frac{1}{2}K.$$

So

$$\overleftrightarrow{AA'}=\left\langle 0,1,\frac{1}{2}+\frac{1}{2}i+\frac{1}{2}k-\frac{1}{2}K\right\rangle.$$

Next we calculate the intersection of $\overleftrightarrow{AA'}$ with the line $\overleftrightarrow{BC}=\langle j,1,-1\rangle$. If $D=[d_1,d_2,1]$, then

$$d_1j+d_2-1=0;$$
$$d_2+\frac{1}{2}+\frac{1}{2}i+\frac{1}{2}k-\frac{1}{2}K=0.$$

From the second equation we get d_2, and the first equation gives $d_1=(-d_2+1)j^{-1}=-(-d_2+1)j$; therefore

$$D=\left[-\frac{3}{2}j-\frac{1}{2}l+\frac{1}{2}I+\frac{1}{2}J,-\frac{1}{2}-\frac{1}{2}i-\frac{1}{2}k+\frac{1}{2}K,1\right].$$

We obtain the point $E=\overleftrightarrow{AC}\cap\overleftrightarrow{BB'}$ in a similar manner. Since $\overleftrightarrow{BB'}=\left\langle 1,0,-\frac{1}{2}-\frac{1}{2}i+j+l\right\rangle$ and $\overleftrightarrow{AC}=\langle -k,1,k\rangle$, the $[e_1,e_2,1]$ coordinates of E satisfy the following system of equations:

$$-e_1+e_2+k=0;$$

$$e_1 - \frac{1}{2} - \frac{1}{2}i + j + l = 0.$$

From the first equation $e_2 = (e_1 - 1)k$ and from the second equation e_1 can be expressed, therefore

$$E = \left[\frac{1}{2} + \frac{1}{2}i - j - l, -\frac{1}{2}k - I + J + \frac{1}{2}K, 1\right].$$

Finally, we determine the point $F = \overleftrightarrow{AB} \cap \overleftrightarrow{CC'}$. Since $C' = [0, 0, 1]$, the line $\overleftrightarrow{CC'}$ is of the form $\langle c, 1, 0\rangle$ for some octonion c. To obtain c we use the fact that $C \in \overleftrightarrow{CC'}$:

$$\left(\frac{1}{2} - \frac{1}{2}j - \frac{1}{2}k + \frac{1}{2}I\right)c + \left(\frac{1}{2} - \frac{1}{2}j - \frac{1}{2}k - \frac{1}{2}I\right) = 0.$$

From this equation,

$$\begin{aligned} c &= \left(\frac{1}{2} - \frac{1}{2}j - \frac{1}{2}k + \frac{1}{2}I\right)^{-1}\left(-\frac{1}{2} + \frac{1}{2}j + \frac{1}{2}k + \frac{1}{2}I\right) \\ &= \left(\frac{1}{2} + \frac{1}{2}j + \frac{1}{2}k - \frac{1}{2}I\right)\left(-\frac{1}{2} + \frac{1}{2}j + \frac{1}{2}k + \frac{1}{2}I\right) \\ &= -\frac{1}{2} + \frac{1}{2}I - \frac{1}{2}k + \frac{1}{2}j. \end{aligned}$$

Thus

$$\overleftrightarrow{CC'} = \left\langle -\frac{1}{2} + \frac{1}{2}j - \frac{1}{2}k + \frac{1}{2}I, 1, 0\right\rangle.$$

So for the coordinates $[f_1, f_2, 1]$ of F we have

$$\begin{gathered} -f_1 + f_2 + i = 0; \\ f_1\left(-\frac{1}{2} + \frac{1}{2}j - \frac{1}{2}k + \frac{1}{2}I\right) + f_2 = 0. \end{gathered}$$

From this we get

$$f_1\left(\frac{1}{2} + \frac{1}{2}j - \frac{1}{2}k + \frac{1}{2}I\right) = i,$$

whence

$$f_1 = i\left(\frac{1}{2} + \frac{1}{2}j - \frac{1}{2}k + \frac{1}{2}I\right)^{-1} = i\left(\frac{1}{2} - \frac{1}{2}j + \frac{1}{2}k - \frac{1}{2}I\right) = \frac{1}{2}i - \frac{1}{2}l + \frac{1}{2}K - \frac{1}{2}J.$$

Therefore

$$F = \left[\frac{1}{2}i - \frac{1}{2}l - \frac{1}{2}J + \frac{1}{2}K, -\frac{1}{2} - \frac{1}{2}l + \frac{1}{2}K - \frac{1}{2}J, 1\right].$$

It is well-known that if $A[a_1, a_2, 1]$ and $B[b_1, b_2, 1]$ are points in a projective plane over an alternative division ring $\mathcal{R}$, then the points of the line $\overleftrightarrow{AB}$ are of the form

$$[t(a_1, a_2, 1) + (1 - t)(b_1, b_2, 1)],\ t \in \mathcal{R} \quad \text{or} \quad [1, x, 0].$$

Therefore, if we want to check whether D, E and F are collinear, we need to check if the coordinates of F can be combined from the coordinates of D and E in this way. Suppose that such a t octonion exists. Then, from the first coordinates of D, E and F, we get

$$t\left(-\frac{3}{2}j-\frac{1}{2}l+\frac{1}{2}I+\frac{1}{2}J\right)+(1-t)\left(\frac{1}{2}+\frac{1}{2}i-j-l\right)=\frac{1}{2}i-\frac{1}{2}l-\frac{1}{2}J+\frac{1}{2}K.$$

This equation leads to

$$t\left(-\frac{1}{2}-\frac{1}{2}i-\frac{1}{2}j+\frac{1}{2}l+\frac{1}{2}I+\frac{1}{2}J\right)=-\frac{1}{2}+j+\frac{1}{2}l-\frac{1}{2}J+\frac{1}{2}K,$$

hence

$$\begin{aligned}t&=\left(-\frac{1}{2}+j+\frac{1}{2}l-\frac{1}{2}J+\frac{1}{2}K\right)\left(-\frac{1}{2}-\frac{1}{2}i-\frac{1}{2}j+\frac{1}{2}l+\frac{1}{2}I+\frac{1}{2}J\right)^{-1}\\&=\left(-\frac{1}{2}+j+\frac{1}{2}l-\frac{1}{2}J+\frac{1}{2}K\right)\left(-\frac{1}{3}+\frac{1}{3}i+\frac{1}{3}j-\frac{1}{3}l-\frac{1}{3}I-\frac{1}{3}J\right)\\&=-\frac{1}{6}-\frac{5}{6}i-\frac{1}{6}j-\frac{1}{6}k+\frac{1}{6}l-\frac{1}{6}I+\frac{1}{2}J-\frac{1}{2}K.\end{aligned}$$

We check if the second coordinates can be combinated using the same coefficient. In this case the following equation would hold:

$$t\left(-\frac{1}{2}-\frac{1}{2}i-\frac{1}{2}k+\frac{1}{2}K\right)+(1-t)\left(-\frac{1}{2}k-I+J+\frac{1}{2}K\right)=-\frac{1}{2}-\frac{1}{2}l+\frac{1}{2}K-\frac{1}{2}J.$$

Here the left side is

$$\begin{aligned}&\left(-\frac{1}{6}-\frac{5}{6}i-\frac{1}{6}j-\frac{1}{6}k+\frac{1}{6}l-\frac{1}{6}I+\frac{1}{2}J-\frac{1}{2}K\right)\left(-\frac{1}{2}-\frac{1}{2}i-\frac{1}{2}k+\frac{1}{2}K\right)\\&+\left(\frac{7}{6}+\frac{5}{6}i+\frac{1}{6}j+\frac{1}{6}k-\frac{1}{6}l+\frac{1}{6}I-\frac{1}{2}J+\frac{1}{2}K\right)\left(-\frac{1}{2}k-I+J+\frac{1}{2}K\right),\end{aligned}$$

whose real part is

$$\frac{1}{12}-\frac{5}{12}-\frac{1}{12}+\frac{1}{4}+\frac{1}{12}+\frac{1}{6}+\frac{1}{2}-\frac{1}{4}=\frac{1}{3}.$$

Otherwise, the real part of the right side is $-\frac{1}{2}$; therefore D, E and F are not collinear. □

References

[1] A. Heyting: *Axiomatic projective geometry*, Amsterdam: North-Holland, 1980, doi: 10.1016/C2013-0-11858-2.

[2] D. R. Hughes, F. C. Piper: *Projective Planes*, New York, Berlin, Heidelberg: Springer, 1973.

[3] A. Linehan: *Relationships between geometric propositions which characterise projective planes*, University of Western Australia: Thesis, 2021.

[4] F. W. Stevenson: *Projective Planes*, San Francisco: W. H. Freeman, 1972.

[5] Z. Szilasi: *A note on the converse Bricard property of projective planes*, arXiv (2023), doi: 10.48550/arXiv.2311.17209.

Annales Mathematicae et Informaticae
60 (2024) pp. 159–168
DOI: 10.33039/ami.2024.02.004
URL: https://ami.uni-eszterhazy.hu

Predicting somatic cell count in milk samples using machine learning*

Bence Tarr[a], **István Szabó**[a], **János Tőzsér**[b]

[a]Institute of Technical Sciences,
Hungarian University of Agriculture and Life Sciences,
Gödöllő, Hungary
tarr.bence.gyula@uni-mate.hu
szabo.istvan.prof@uni-mate.hu

[b]Department of Animal Science, Albert Kázmér Faculty,
Széchenyi István University,
Mosonmagyaróvár, Hungary
tozser.janos@ga.sze.hu

Abstract. Milk quality is an important factor both for the farmers to be able to sell their products and for the milk industry to be able to plan its production based on quantity and quality. Milk quality has a direct link with cow health, more specifically with utter health. One of the most common utter diseases is mastitis. It always captures a lot of interest based on its frequency and cost as a dairy disease which eventually leads to an involuntary and premature culling of milking cows and decreased milk yield. The genetic evaluation of mastitis is very difficult as it is a low heritable trait and categorical in nature [2]. That is why it is necessary to find markers that could predict the occurrence of mastitis. One of the widely used such markers is the somatic cell count (SCC) [9] which is considered to be the most suitable indicator trait for mastitis resistance given its medium to high genetic correlation with mastitis and its greater heritability than mastitis. The SCC is also easy to record in the practice. The selection for lower SCC in milk has a positive effect on the incidence of mastitis. The selection against high SCC also does not deteriorate the immune system of cattle and decreases the risk of infection at the same time. The genetic evaluation [1] of this trait is mostly based on somatic cell score (SCS), a logarithmic transformation of SCC to achieve normality of distribution. In our study, we used the milk database of Holstein cows from 3 different farms. From each farm, we had

*Special thanks to Laszló Dégen and Attila Monostori at ÁT Kft. for supporting this research.

Submitted: July 27, 2023
Accepted: February 8, 2024
Published online: February 9, 2024

altogether 8000 samples tested. The samples were analyzed using chemical methods every month for a year. 11 different types of data were recorded from each sample. Our aim was to find the best mixture of recorded data that would predict the value of linearized somatic cell count. After the logarithmic linearization the SCC results were divided into 3 main groups (based on the probability of mastitis). Thus our prediction problem turned into a classification problem. We used machine learning to train our algorithm. We experimented with different types of classification methods and found good results for the prediction of SCC in milk samples. We changed the input variables as not all the 9 measured input variables will be necessary for good prediction results. Our preliminary results show that using machine learning it is possible to build a model that can be used to predict mastitis in dairy cows based on variables generally analyzed during milk quality checking tests.

Keywords: somatic cell count, machine learning, prediction using AI, milk quality

AMS Subject Classification: 92-02

1. Introduction

Precision agriculture is no more a question of the future but the reality of today. The ever-growing demand for more food, better animal husbandry and less environmental impact puts a high stress on agriculture experts to find better ways to optimize production. The importance of the quality of cattle breeding is growing as the demand for better food and better milk is emerging. Monitoring the quality of milk is important for several reasons: overall animal health, milk quantity, and manufacturing of dairy products. So monitoring milk quality and finding new tools to help farmers plan their production is of greater importance than ever. [12] There are already existing methods for milk quality monitoring. However the most accurate solution to this is laboratory testing which is done usually monthly. This method can not be than on farms and takes time. Since laboratory testing had been a standard monitoring method for many years, there are a lot of historical data available to work with. Milk consists of several important food constituents like fats, proteins, carbohydrates, several minerals, and vitamins. The quantity of these constituents will determine the quality of the milk as well as provide information about the animal's health. We collected the milk data from three different farms. There already exists mathematical methods for how laboratories analyze their data. The focus of milk data analysis is to find milk samples which reflect a possible health problem in the animal. When a possible infected milk sample is found, then the cow where the sample came from must be checked by an expert for further possible medication. Somatic cell count is the most frequently used indicator of subclinical mastitis in dairy cattle. The most important cause of increased SCC is a bacterial infection of the mammary gland [7]. Other factors influence SCC like age, stage of lactation, season, stress, food and many others. These factors are considered less important than bacterial factors. Milk samples

for SCC analysis as part of DHI programs are routinely collected at milking time. There are several factors which influence somatic cell counts and the high value not necessarily mean an animal with a real clinical problem. Usually, we consider infected milk to have an SCC value of 300,000 and 250,000 cells/mL [11]. Apart from an infection, several other factors to consider to evaluate high SCC values are lactation days, number of calving and also on-site weather circumstances. An easy-to-use, on-site prediction method for SCC value in milk yields great economic importance. Mastitis if realized in time can be cured faster, and cheaper with using much less medication.

2. Method

In this study our aim was to prove that with a carefully selected machine learning model, it is possible to build an algorithm that can predict SCC based on other milk constituent values that are easier to monitor. Input variables were treated before creating the training and testing databases. A few input parameters were transformed into categorical values for biological reasons and to help better model creation. Lactation days (0:n<100, 1:100<n<200, 2:n>200) were categorized into three categories. The number of calving was also divided into three categories: 1,2,3+. As a first step general data cleaning steps were made. We deleted all outlier and zero values from our database. We developed a software environment where all steps of data cleaning, transformation, modification and the selection teaching and testing of the model can be done automatically in a single workflow. Our solution was written in Python using Pandas and scikit-learn modules. Since SCC values have a huge variance we will certanly need to convert them to a better usable categorical output variable, so we will be able to use tree-based models. Our dependent variable (linear SCC score) was transformed to create 3 categories, thus we needed tree based algorithms that are suitable for multi-class classification. Using the logarithmic linearization method we created three SCC groups. We denoted the groups with 0, 1 and 2 values each of them corresponding to healthy, possible infected and infected categories. The distribution of the SCC categories can be seen in Table 1.

Table 1. Instances of our output variable.

Category	No. of instances
0	17 500
1	5 200
2	2 600

As we expected we were facing a very unbalanced output dataset. This is due to the fact that in the selected milk samples the majority of the animals were healthy with a relatively low number of possible infected milk samples. We had to find workarounds to balance categories to be able to build a better model. We

were considering two solutions to balance our dataset before teaching our model: The first one is Up-sampling of the dataset. In this case when we duplicated the underrepresented categories of the dependent variable and at the same time we altered the input variables. We used a random multiplication factor for each of the input variables (a factor of 0.98-1.02). To create a fully balanced dataset we also applied another solution to the same dataset. Additional farm data was used to make our dataset bigger, but in this case only the under-represented values were inserted into the final database. After all these treatments to the experimental dataset, we reached a very balanced dataset which was suitable for model creation using multi-class machine learning algorithms. To find the best prediction model we had to decide which ML algorithms would be used for the experiment. Based on our practice with other datasets 4 ML methods were selected for deeper testing: Random Forest, Support Vector Method, Decision Tree Regressor and Extra Trees Classifier [10]. We created two subsets from our data: one dataset for training and one dataset for testing. Our dataset was divided into two parts on an 80% (training) and 20% (testing) basis, the entries were selected randomly from the original dataset. The distribution of the output variable in both datasets was the same as in our original balanced dataset. To find the suitable input variables which will give us the best result for prediction we used the correlation matrix. Several trial runs were made to select the most suitable features.

3. Experiment

3.1. Dataset

The initial dataset covers 3 years (2019–2021) of laboratory data from 3 different farms. Farmers generally use monthly laboratory checking to monitor their milk quality so all together we used 36 months of data from each herd. The size of the dataset was 25000 measurements and each measurement resulted in 11 different parameters for each milk sample. Considering biological, chemical and statistical methods we focused on the following parameters: casein, lactoferrin, somatic cell count (SCC), proteins and milk fat. In this study our goal was to create a prediction model for somatic cell count, we have to examine the SCC variable at the first step. The statistical description of the SCC variable in our dataset can be seen in Table 2.

3.2. Somatic Cell count

Somatic cells (SCC) found in cow milk are a mixture of milk-producing cells and immune cells. SCC value can be used for estimating mammary health and milk quality [3] as cells are secreted into the milk generally. SC is related to mastitis as its main role is to fight infection and repair damaged tissues. Milk somatic cell counts (SCCs) are widely used as a marker to monitor the milk quality and animal health in dairy herds. The SCC is categorized based on the number of cells per ml of milk.

Table 2. Statistical evaluation of SCC values in our dataset.

No. samples	26 6686
Mean	266686
Std.	1245000
Min	2000
25%	50000
50%	150000
75%	401000
Max	900000

- If SCC $<$ 100,000 then it is considered an 'uninfected' cow
- If SCC $>$ 100,000 but $<$ 300,000 the cow will need special attention
- If SCC $>$ 300,000 it means Cows are highly likely to be infected

In real life the variance of SCC is large we need to convert these values into another scale for better prediction results. For this purpose the linearized somatic cell count number is used [6]. For analysis of the SCC variable on the test day, the transformation we used the following formula [5]:

$$\mathrm{SCC}t = [\log 2(\mathrm{SCC}/100{,}000)] + 3$$

The result of SCC conversion into categorical numbers can be seen in Table 3.

Table 3. Linear scores after SCC logaritmization.

Category Number	**SCC**	**Category Number**	**SCC**
1	25 000	6	800 000
2	50 000	7	1 600 000
3	100 000	8	3 200 000
4	200 000	9	6 400 000
5	400 000	–	–

Linear SCC scores were divided into 3 groups to create the final categorical variables to be predicted (not infected = 0, possible infection = 1 and infected = 2). These will be the values of the dependent variable for the model to predict.

3.3. Lactoferrin

Careful selection of input variables is the key to finding the best prediction model. Lactoferrin (Lfe), is an iron-binding glycoprotein. It plays a key role in the defence

mechanisms of the mammary gland, contributing to the prevention of microbiological infectious diseases. Lactoferrin also may limit the oxidative degeneration of cellular components during inflammation and involution of the mammary gland. Lfe concentration in milk was significantly associated with somatic cell count (SCC) in laboratory experiments [4]. However, in real life the correlation between Lfe and SCC was not so strong, animals with high Lactoferrin rarely had mastitis. So the problem to be solved was to find other parameters that together with Lfe can be used to predict SCC.

4. Verification

To benchmark our model's real-life accuracy we calculated the accuracy of the model separately for all categories and then calculated an average accuracy to validate our results. The accuracy score function we evaluated for our results computes the accuracy as the count of correct predictions. In the case of multi-label classification, this function returns the subset accuracy. In case the predicted categories for a sample match with the original measured set of categories, then the subset accuracy is 1; otherwise it is 0. For calculating the accuracy in multi label classification the following outcomes are defined:

- False positives (FP): This is when a classifier predicts a label that does not match the input data.
- False negatives (FN): This is when a classifier misses a label that exists in the input data.
- True positives (TP): This is when a classifier correctly predicts the existence of a label.
- True negatives (TN): This is when a classifier correctly predicts the inexistence of a label.

Accuracy is the proportion of examples that were correctly classified. It is the sum of the number of true positives and true negatives, divided by the number of examples in the dataset.

$$\texttt{accuracy} = \frac{\mathrm{TP} + \mathrm{TN}}{\mathrm{FP} + \mathrm{FN} + \mathrm{TP} + \mathrm{TN}}.$$

We checked the accuracy of the model independently for all three categories, we assumed that the average accuracy is calculated as the average of the three sub-results. A good model that would be suitable for real-life usage should give prediction success of over 80% and an over 90% success rate our model would be accurate enough to use instead of laboratory measurements.

5. Results

We run 10 different combinations of input variables on all 4 ML algorithms. To find the best combination of input variables, we used biological and statistical tools. Based on biological and chemical observations Lactoferrin (LF) and Protein were included in our input mix. For other candidates, we calculated the correlation between the variables and used these results as a hint for creating the input mixture. We have created a correlation heatmap of our variables. The correlation heatmap displays the correlation between multiple variables as a color-coded matrix. It's like a color chart that shows us how closely related different variables are. In the correlation heatmap, each variable is represented by a row and a column, and the cells show the correlation between them. The color of each cell represents the strength and direction of the correlation, with darker colors indicating stronger correlations. Our correlation heatmap in terms of our output variable (SCC) can be seen in Figure 1 .

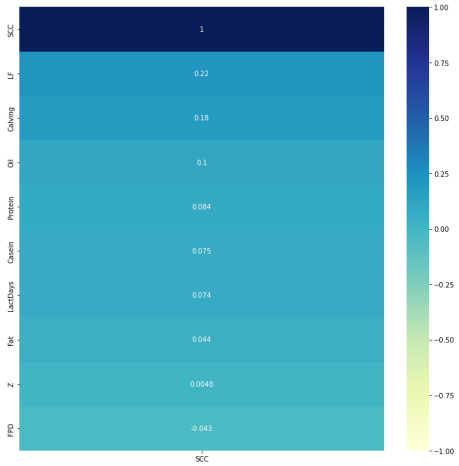


Figure 1. Heatmap of all possible variables. Showing the correlations.

The best model result used the following input variables: LNPC, Z, Urea, SUGAR, LF, FAT, PROTEIN. The calculations for average accuracy and the accuracy for each separate category (positive TRUE values) are shown in Table 4.

As we can see all selected algorithms gave us a true value of 80% or above. The best algorithm seems to be the Extra Trees Classifier. We used mainly tree-based models primarily due to their robust nature and because they can be used on any type of data (categorical/continuous), can be used on data that is not normally distributed, and require little data transformations. ExtraTrees is an ensemble machine learning model that trains several decision trees and aggregates the results from the group of decision trees to output a prediction. It is also

Table 4. Accuracy results of different models in this study.

ML algorithm	LSCC=0	LSCC=1	LSCC=2	Average
Random Forest	0.88	0.86	0.85	0.86
SVM	0.86	0.85	0.80	0.84
Decision Tree Regressor	0.83	0.80	0.82	0.82
Extra Trees Classifier	0.89	0.88	0.86	0.88

called an extremely randomized tree since to ensure sufficient differences between individual decision trees, it randomly selects the values where it splits a feature and creates child nodes [8]. It would be very interesting to see which input variable (which milk constituent) plays a greater role in the best algorithm to predict the SCC value. As from biological considerations, it is not yet clear which parameters will predict the future change in SCC. So we further investigated the algorithm in the best model. To find the importance of each input feature in our model we used scikit-learn's feature importance investigation tools. Our feature importance technique assigns a score to input features based on how important they are in the model that predicts the target variable. Investigating our best model, we calculated the feature score for each input variable. The result can be seen in Figure 2.

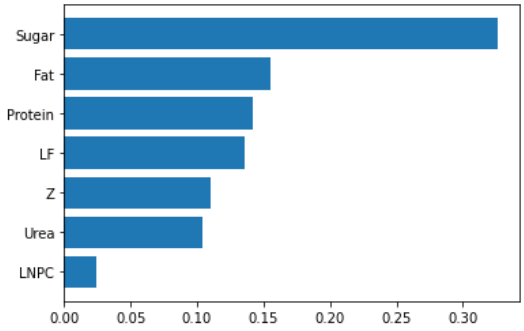


Figure 2. The result of the feature importance scoring for the best model's input parameters.

From the feature importance scoring results we can conclude that in our prediction algorithm sugar, fat, protein and lactoferrin variables play the greatest role in the prediction of SCC.

The accuracy checking for each output category separately was important to prove that our algorithm is not biased. As our model gave similarly good prediction values for less represented cases in our validation database it can be used in real life as well. This experiment shows that with enough data a good model can be built which can be used safely to predict the SCC category of a milk sample. Using this method the chemical analyses of milk samples can be cheaper or faster.

6. Conclusion

As a conclusion, we can see that regularly collected milk data can be used to predict somatic cell count in milk samples. Lactoferrin which seemed biologically the most significant impact factor on SCC values gives a much better result when used with other input variables. However, in the best-performing model lactoferrin was not the most significant input variable. The result of the feature importance scoring can serve as a basis to build another prediction model SCC scores well ahead of time, to give more time to farmers to watch the affected cows for possible mastitis. For this, we will need further data with more datestamp which can be obtained from farms using milking robots. We also proved that the logarithmic SCC value classified into 3 categories is well-suitable for prediction purposes and it also satisfies the milk health needs of the market.

Since most of the collected milk samples are from healthy cows the balancing of the input dataset must be done prior to teaching the algorithm. Regular balancing methods proved to be successful in balancing the dataset and for machine learning classification. Our results proved that multiclass machine learning can be used to predict high levels of somatic cell count in milk samples. This algorithm can serve as a basis for future prediction of somatic cell count it can help in predicting early mammary health of milking cows. If the farmer can start medication right before critical somatic cell count values are reached in milk samples which can lead to more effective milking of the farm.

References

[1] M. Alam, C. Cho, T. Choi, B. Park, J. Choi, Y. Choy, S. Lee, K. Cho: *Estimation of Genetic Parameters for Somatic Cell Scores of Holsteins Using Multi-trait Lactation Models in Korea*, Asian-Australas J Anim Sci 28.3 (2015), pp. 303–310, doi: 10.5713/ajas.13.0627.

[2] A. A. E. Amin: *Estimates of Heritability for Somatic Cell Count, Test-Day Milk Yield and Some Udder-Teat Characteristics in Saudi Dairy Goats using Random Regression Animal Model*, Adv. Anim. Vet. Sci 6.3 (2018), pp. 128–134, doi: 10.17582/journal.aavs/2018/6.3.128.134.

[3] C. Burvenich, et al.: *Physiological and Genetic Factors That Influence the Cows Resistance to Mastitis, Especially during Early Lactation*, Proceedings of the 5th IDF Mastitis Congress, Symposium on Immunology of Ruminant Mammary Gland (2000).

[4] J. B. Cheng, J. q. Wang, D. P. Bu, G. L. Liu, C. G. Zhang, H. Y. Wei, L. Y. Zhou, J. Z. Wang: *Factors Affecting the Lactoferrin Concentration in Bovine Milk*, Journal of Dairy Science 91 (2007), pp. 970–976, doi: 10.3168/jds.2007-0689.

[5] S. Dabdoub, G. Shook: *Phenotypic relations among milk yield, somatic cell count and clinical mastitis*, Journal of Dairy Science 67.1 (1984), pp. 163–164.

[6] L. Dégen, A. Monostori: *Az új tőgyegészségügyi rendszer bemutatása*, Állattenyésztési Teljesítményvizsgáló Kft.

[7] I. Dohoo, K. E. Leslie: *Evaluation of changes in somatic cell counts as indicators of new intramammary infections*, Preventive Veterinary Medicine 10.3 (1991), pp. 225–237, doi: 10.1016/0167-5877(91)90006-N.

[8] P. Geurts, D. Ernst, L. Wehenkel: *Extremely randomized trees*, Machine Learning 63 (2006), pp. 3–42, doi: 10.1007/s10994-006-6226-1.

[9] K. C. Halasa Tariq: *Differential Somatic Cell Count: Value for Udder Health Management*, Frontiers in Veterinary Science 7 (2020), doi: 10.3389/fvets.2020.609055.

[10] W. A. Muhammad, J. Reynolds, Y. Rezgui: *Predictive modelling for solar thermal energy systems: A comparison of support vector regression, random forest, extra trees and regression trees*, Journal of Cleaner Production 203 (2018), pp. 810–821, doi: 10.1016/j.jclepro.2018.08.207.

[11] N. Sharma, N. K. Singh, M. S. Bhadwal: *Relationship of Somatic Cell Count and Mastitis: An Overview*, Asian-Australasian Journal of Animal Sciences 24.3 (2011), pp. 429–438, doi: 10.5713/ajas.2011.10233.

[12] I. Tímár: *Versenyképesség a magyar tejágazatban*, PhD thesis, Budapesti Corvinus Egyetem, 2004.

Annales Mathematicae et Informaticae
60 (2024) pp. 169–177
DOI: 10.33039/ami.2024.02.008
URL: https://ami.uni-eszterhazy.hu

Online way to learn SQL

Tamás Balla, Sándor Király

Eszterházy Károly Catholic University,
Faculty of Informatics
balla.tamas@uni-eszterhazy.hu
kiraly.sandor@uni-eszterhazy.hu

Abstract. The new version of Hungarian National Base Curriculum (NAT 2020) makes many changes in public education, including many innovations in the field of Computer Science Education (newly known as Digital Culture). NAT 2020 makes knowledge of the Structured Query Language (SQL) mandatory for students in the advanced final examination. In the future, students must be able to independently formulate SQL statements based on the task without the help of graphical tools (query builders) in the advanced final exam. This poses new challenges for both students and their teachers.

Based on the changes in the Hungarian National Base Curriculum, we have already created an educational portal called kodolosuli.hu, which is a great success among both students and teachers related to teaching and learning programming languages (Java, C++, CSharp and Python)[2]. Based on the experience of the kodolosuli.hu portal, we thought it would be advisable to design and develop a new web portal related to relational database management. Our goal with the design and development of sqlsuli.hu was to give the teenage generation the opportunity to learn the basics of relational database management and SQL in a playful and effective way. In 2022, after continuous development, the site became available for free. In Hungary, this type of portal is not available, apart from our sqlsuli.hu portal, which was developed for people under the age of 18.

The framework has an extensive grader tool that helps students test their knowledge in some automatic way, so the portal provides a flexible learning way. The test databases are based on Harry Potter and Star Wars data, which is expected to increase students' engagement and motivation for learning. The course is available in the Hungarian language.

Keywords: SQL, Learning Management System, final examination, online platform

AMS Subject Classification: 97C90

Submitted: November 5, 2023
Accepted: February 22, 2024
Published online: March 18, 2024

1. Introduction

The structured query language (SQL) is the query and control language of relational database management systems. Relational database management systems provide SQL interface to perform each task, which means that SQL is a very rich language, providing several statements and structures from creating users to formulating very complex queries. We can divide the SQL into four different parts based on the functionality:

- Data Control Language (DCL): consists of the SQL commands that can be used to manage the users and their privileges belonging to the database and its objects.
- Data Definition Language (DDL): consists of the SQL commands that can be used to manage (create, modify, delete) database objects.
- Data Manipulation Language (DML): consists of the SQL commands that can be used to manage (add, update, delete) data content of the tables.
- Query Language (QL): consists of the SQL commands that can be used to query of the content of tables from different viewpoints.
- Transaction Control Language (TCL): consists of the SQL commands that can be used to manage transactions in the database.

If we wish to to be successful in database management in public education, we need to know the DDL, DML and QL elements of SQL. In most cases, DCL regulation is a database administrator task.

In Hungary, the educational content to be taught in public education and the main teaching methodology are regulated by the National Base Curriculum (NAT), which has been updated several times by the government in recent years. The new Hungarian National Base Curriculum (NAT 2020) makes many changes including many innovations in the field of Computer Science Education (newly known as Digital Culture). NAT 2020 makes knowledge of the Structured Query Language (SQL) mandatory for students in the advanced final examination. In the future, students must be able to independently formulate SQL statements based on the task without the help of graphical tools (query builders) in the advanced final exam. If we examine the content of the base curriculum, we can conclude that it will be difficult to meet the expectations in classroom conditions: in high school, only 25 lessons (1 lesson is 45 minutes) are spent on database management. The emerging new situation and new expectations present both students and teachers with new challenges.

Learning a programming language may be more difficult than learning SQL statements, meanwhile acquiring knowledge of SQL for students under the age of 18 is also time-consuming and not as easy as one might think. In the school, teachers try to teach the students to think in an algorithmic way, but the declarative characteristic of SQL requires the opposite thinking methods. It is quite interesting

how can we teach the students in public education two different thinking methods in one hour per week. This is clearly an extremely difficult task. On the one hand, it basically requires teacher supervision, similar to learning other content.

If we would like to learn SQL independently in some practical way without any teacher supervision, we can install some free database management system (like MySQL or PostgreSQL). There are a lot of vendors whose database management systems can be used to learn SQL by working with them. But for beginners these systems are not very helpful or in some way might be very harmful too. Database management systems are only limited to providing feedback on syntax errors [7]. It means that if we made a logical error, these systems do not show error messages. Accordingly, we can learn techniques in database management that we believe are correct, but in the future, they will produce an incorrect set of results. Realizing these problems, in the late 1990s and early 2000s, researchers developed various tools that provide a way to learn SQL [9].

Several researchers have been involved in teaching SQL: Al-Shuaily and Renaud proposed applying SQL patterns [10], Mitrovic developed a Knowledge-Based Teaching System for SQL [6], Quer et. al developed a software tool, LearnSQL (Learning Environment for Automatic Rating of Notions of SQL), that allows the automatic and efficient e-learning and assessment of relational database skills [8]. Garner and Mariani developed a graphical user interface centred around the textual translation of a query which has the potential to improve the way in which users gain an understanding of SQL [3]. Although books and course notes in Hungarian and online courses in English are available, it seems insufficient for most secondary school students. An early example SQLTutor in which the developers tried to examine the content of queries and make feedback belonging to the logical problems [6]. Another system, called eSQL concentrates on the sequences of steps that solve SQL problems. In this system during the implementation, we get sequences of images to describe what happened during the execution [4]. Learn-SQL is also an online SQL tool in which users' solution are checked by automatically in a logical way, so it can be very useful to learn SQL online [1].

Recently, books and lecture notes in Hungarian and online courses in English are available, which seem insufficient for most secondary school students. There are some portals that provide learning SQL playfully such as SQL Island [11] or SQL Murder Mystery [5]. Portals such as Udemy.com, CodeAcademy, tutorials on W3Schools or Tutorialspoint that teach SQL in an interactive way in English might be useful. In Hungary, this type of portal is not available, apart from our sqlsuli.hu portal, which was developed mainly for people under the age of 18.

2. Learning Management System

We started the development in 2021, and at the end of the year, we implemented our web portal which can help students to learn the main concept of the structured query language and relational data model. In 2021, the demo version of the Learning Managment System was published, and in 2022, after continuous development,

the site became available for free.

The webportal was created by using our own educational framework, which was developed using the PHP5 programming language and the MVC design pattern. We use the MySQL relational database management system to store the data. The portal provides the key features of Learning Management Systems.

Currently, there are three types of users who can access the system; students, teachers and the administrator. Students can reach the curriculum and they can check the progress and the solutions that they submitted earlier. The teachers can also use the system for learning and they access their students' progress as well, if the students enable this option. The administrators have full access to both the materials and the users' data. The administrator can create and manage new learning objects or tasks and they also can read all information about the students' progress and solutions.

The most interesting parts of the portal are the course editor and the grader subsystem.

2.1. Course editor

In the course editor subsystem, we can create new online courses with our own teaching content.

Courses can be divided into chapters in the system, which serve as logical containers. The series of lessons can be recorded in these chapters. A lesson is a teaching object that the student sees on one screen at the same time.

We currently distinguish between three lesson types: independent lesson material, independent task and lesson material with tasks.

During the preparation of the independent study material, we use the entire screen to present new knowledge. During the creation of the independent lesson material, we can use text minds, audio-visual content and attachments. This lesson type is used to assess our students' SQL knowledge. Applying the lesson material with task types, (Figure 1), we can also create tasks next to the curriculum, which appear on the right side of the screen. After reading the study material, students must complete the related tasks in order to move on to the next lesson. These tasks are typically short SQL statements that focus on a specific piece of text. In the case of the independent task (Figure 2), the entire screen is used to introduce the task that the student must solve in order to move on. These types of lessons typically occur at the end of each chapter with the aim of integrating the knowledge acquired in the chapter into more complex tasks.

2.2. Grader subsystem

In addition to the learning objects and materials, another important element of the web portal is the grader. With the help of the grader system, students have the opportunity to test their knowledge automatically.

Testing queries is particularly challenging, as there are several correct solutions to the same problem in SQL. A well-functioning learning tool must also ensure

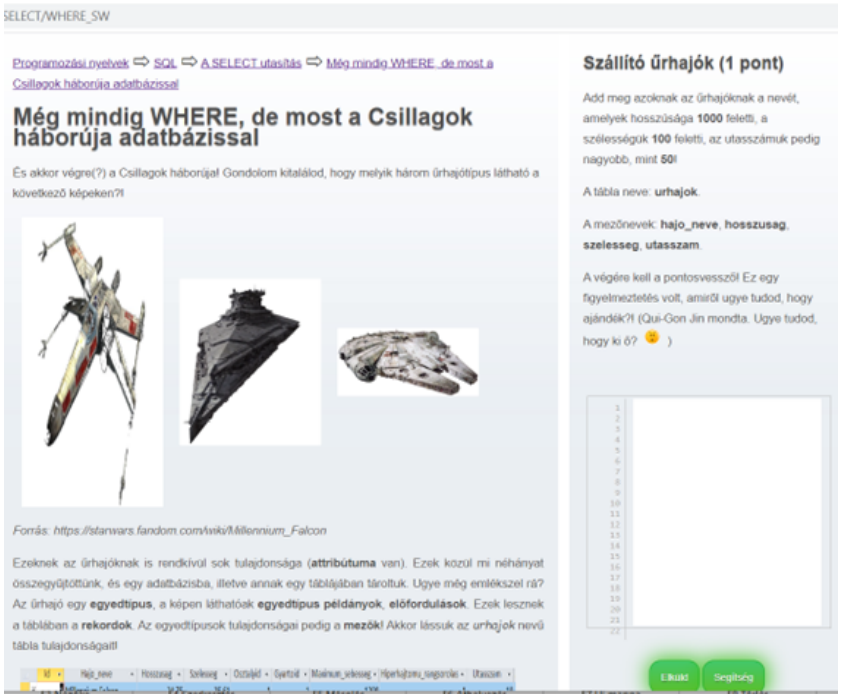


Figure 1. Lesson material with tasks.

the acceptance of several correct solutions since motivation is greatly reduced if a tool classifies a completely correct solution as incorrect. Accordingly, we discarded the use of only syntactic checking. A textual check of the specified query is also not expedient, because then we only accept the solution defined by us. In order to test the queries, the solution of the task must be specified when recording the tasks in the framework. The students' SQL solution is first run in the database by a database user with limited privileges. If the query execution is successful, we run the administrator solution. The result sets of the two solutions are compared, and if they completely match, the solution is accepted, otherwise, the solution is rejected and the error is indicated. Students do not see the contents of the data tables, so they cannot use tricks to generate the correct solution because they do not know the expected output. When creating the portal, we tried to focus on the fact that the error message helps the user to be able to correct their solution during each incorrect solution. Possible error messages that may appear in the course material are:

- *The query contains syntax errors.* We receive this error message if we create a query that does not comply with the SQL syntax rules, so the query cannot be executed.
- *Database user does not have enough sufficient privileges to execute it.* The

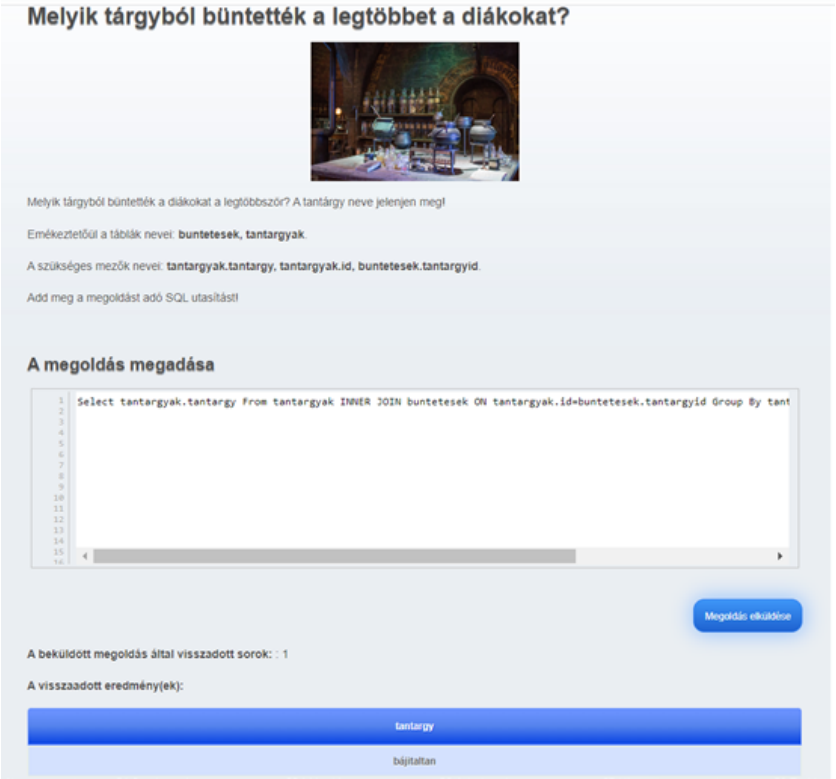


Figure 2. Independent task.

query does not contain a syntax error, but the database user does not have sufficient privileges to execute it.

- *The number of columns queried is incorrect.* The query is correct and gives a result, but the number of fields listed after SELECT does not correspond to the ones specified in the task description.

- *The queried columns are incorrect.* The query is correct, and the number of columns is correct, but their names (such as alias names) or their order are incorrect.

- *The number of queried rows is incorrect.* The query is correct, it gives a result, but the number of rows returned is not the same as expected. The WHERE condition or HAVING contains an error, or the wrong query type is used.

- *The queried rows are incorrect.* The query is correct, it gives a result, but the result set contains at least 1 incorrect record. In this case, the portal shows

an example of the faulty record, thereby facilitating a successful modification of command.

- *The order of the queried lines is incorrect.* The query is correct, it gives results, the number and content of the returned rows is correct, but their order is not.

These error descriptons or messages can help the students to develop a successful solution. Through modifications and errors, students also learn the subtleties of creating SQL, which would be very difficult to discuss in a handbook or on an educational portal.

The other type of the grader is when students have to work with DDL and DML language elements. To test the DML elements, when specifying the task, the creator provides a preliminary script, the solution script and the control script. When solving the tasks, we use temporary databases, the elements of the temporary database are deleted when the database connection open to the user is closed. At the beginning of submitting the solution, we run the preliminary script, thus creating the objects of the temporary database.

After that, the administrator solution and the control script will be executed, which is typically an SQL statement, so we get the special set of records that we should also get at the end of the student solution. Therefore, we rerun the preliminary script, then the student solution, and the verification script. If the two result sets are equal, the solution is correct. For successful testing, it may also happen that the preliminary script already contains a DML instruction (there is something to delete, modify and thus check success). The testing of the DDL instructions is similar, only in this case the control script checks the information of the temporary schema, i.e. whether the table structure was successfully created, deleted or modified.

3. Curriculum

Our goal with the design and development of sqlsuli.hu was to give the teenage generation the opportunity to learn the basics of relational database management and SQL in a playful and effective way.

Using our website, students can learn how to apply different SQL statements such as Select, Update, Delete, etc. The curriculum focuses on the Select statement since students primarily have to know how to apply this one in the final examination. Through 24 chapters, students can learn to use Select statements, beyond basic use, through Join statements to the nested Select statements. The statements of DDL (Data Definition Language) and DML (Data Manipulation Language) are also introduced to the users of the portal. DML statements (Update, Delete from, Insert into and Insert into with Select) are presented through four chapters with practice exercises. In a separate chapter, DDL statements (Create table and Alter table) are also presented with a practical exercise.

The text of our curriculum contains imaginary dialogues between the authors of the curriculum and the student, provocative questions and comments, and also interesting examples. The modified quotes from Star Wars and Harry Potter, such as "There are always two of the Sith, now the tasks too" or "Curiosity is not a sin but we should exercise caution with our curiosity but not in this portal" accompany the entire curriculum. We have inserted emoticons and pictures related to the movie in each unit. Videos have been created for a better understanding of inner join clause.

4. Results

The aim of this paper is to present the main functionalities of a new website, called sqlsuli.hu. The portal is an online interactive platform that offers a free SQL course. The goal is to reach secondary school students who wish to learn SQL and help them to be successful in the final exam in computer science. The website starts by presuming no prior knowledge at all and lets students work through short exercises with gradually increased complexity. This website offers a SQL curriculum that uses Harry Potter and Star Wars databases.

In the website we developed a special grade system which enables completely individual and successful learning.

The site is open for anyone but currently used by the authors' students. We count on feedback from the authors' students, which makes it possible to correct any errors that may still exist in the portal, or makes changes to lessons, wording and assignments. Then the portal will be promoted.

References

[1] A. Abelló, M. E. Rodríguez, T. Urpí, X. Burgués, M. J. Casany, C. Martín, C. Quer: *LEARN-SQL: Automatic assessment of SQL based on IMS QTI specification*, in: 2008 Eighth IEEE International Conference on Advanced Learning Technologies, IEEE, 2008, pp. 592–593.

[2] T. Balla, S. Király: *A Discussion of Developing a Programming Education Portal*, Central-European Journal of New Technologies in Research, Education and Practice 2.2 (2020), pp. 1–14, doi: 10.36427/CEJNTREP.2.2.

[3] P. Garner, J. A. Mariani: *Learning SQL in steps*, Journal of Systemics, Cybernetics and Informatics 13.4 (2015), pp. 19–24.

[4] R. Kearns, S. Shead, A. Fekete: *A teaching system for SQL*, in: Proceedings of the 2nd Australasian conference on Computer science education, 1997, pp. 224–231.

[5] K. Lab: *SQL Murder Mystery*, 2019, url: https://mystery.knightlab.com/ (visited on 04/08/2023).

[6] A. Mitrovic: *A knowledge-based teaching system for SQL*, in: Proceedings of ED-MEDIA, vol. 98, 1998, pp. 1027–1032.

[7] S. Prabhu, S. Jaidka: *SQL and PL-SQL: Analysing teaching methods*, in: CITRENZ Conference, 2019.

[8] C. Quer, A. Abelló Gamazo, X. Burgués Illa, M. J. Casany Guerrero, C. Martín Escofet, M. E. Rodríguez González, Ó. Romero Moral, A. Urpí Tubella: *E-assessment of relational database skills by means of LearnSQL*, in: EDULEARN17 proceedings: 9th International Conference on Education and New Learning Technologies: Barcelona, Spain, 3-5 July, 2017, International Association of Technology, Education and Development (IATED), 2017, pp. 9443–9448.

[9] S. Sadiq, M. Orlowska, W. Sadiq, J. Lin: *SQLator: an online SQL learning workbench*, in: Proceedings of the 9th annual SIGCSE conference on Innovation and technology in computer science education, 2004, pp. 223–227.

[10] H. Al-Shuaily, K. Renaud: *SQL patterns-a new approach for teaching SQL*, in: 8th HEA Workshop on Teaching, Learning and Assessment of Databases, Abertay-Dundee, 2010, pp. 29–40.

[11] S. Xinogalos, M. Satratzemi: *The Use of Educational Games in Programming Assignments: SQL Island as a Case Study*, Applied Sciences 12.13 (2022), p. 6563.

Annales Mathematicae et Informaticae
60 (2024) pp. 178–189
DOI: 10.33039/ami.2024.04.002
URL: https://ami.uni-eszterhazy.hu

The effect of problem-based learning on students' learning outcomes*

Emőke Báró

University of Debrecen, University of Nyíregyháza,
MTA-ELKH-ELTE Research Group in Mathematics Education
baro.emoke@nye.hu

Abstract. Given the lack of consensus in the literature regarding the impact of problem-based learning on students' learning outcomes, we aimed to identify and understand the possible underlying factors that may contribute to the effectiveness of problem-based learning. To this end, we designed a series of lessons using a problem-based approach supplemented by heuristic strategies to investigate these factors. Two-cycle action research was implemented to explore lower secondary students' learning outcomes affected by problem-based learning and the purposeful use of heuristic strategies. We found that the combination of problem-based learning and the purposeful use of heuristic strategies positively impacts students' learning outcomes, and we explored this effect from both algebraic and geometric perspectives.

Keywords: problem-based learning, secondary school, learning outcomes

AMS Subject Classification: 97D40

1. Introduction

Problem-based learning (PBL) has beneficial effects on students' motivation, attitude [17], and critical thinking [2, 3]. However, the literature has no consensus on its impact on learning outcomes. While some studies claim that PBL increases student's achievement, other studies and analyses have either found an adverse effect or no significant increase in achievement [1, 8]. Further research reports the effectiveness of the purposeful use of heuristic strategies on learning outcomes [18].

*This research was supported by the ÚNKP-23-3 New National Excellence Program of the Ministry for Culture and Innovation from the source of the National Research, Development and Innovation Fund.

Submitted: January 31, 2024
Accepted: April 9, 2024
Published online: April 20, 2024

Based on these studies, we wanted to analyze the effect of PBL on students' learning outcomes, paying particular attention to the purposeful use of heuristic strategies.

Therefore, we (the author and two university experts) designed a two-cycle action research transforming two chapters (one algebraic, one geometric) from the curriculum. These chapters consisted of six, respectively, seven lessons, of which four were designed based on the PBL principles, including heuristic elements. We expected the combination of PBL and the purposeful use of heuristic strategies to impact students' learning outcomes positively. We aimed to explore this effect from both algebraic and geometric perspectives.

2. Theoretical background

Mathematics is about problems and solutions [11]. In mathematics education, a problem implies an obstacle that hinders achieving the goal. The way to overcome the obstacle is problem-solving and purposeful reasoning [16]. Heuristics have been generally recognized as a crucial component for problem-solving [14] because they are general suggestions on a strategy that is designed to help when we solve problems. A method that builds on problems and problem-solving is problem-based learning.

PBL dates to the 1950s and 1960s. Dewey [7] was perhaps the first to formulate the idea that knowledge should be imparted to learners in an active, exploratory way. In his work, Dewey advocates the introduction of active learning, whereby the teacher's task is not simply to make the students learn specific theories. Instead, the teacher's task is to create learning situations (problem situations) where students can acquire knowledge independently and help them manage them [6]. In this way, problem-based learning represented a paradigm shift from previous teaching-learning strategies.

Csíkos [5] defines PBL in mathematics as requiring students to analyze mathematical problem situations, to critically approach their own and their peers' minds, and they must learn to explain and justify their reasoning (see also [13]).

Note that hereafter, by the analysis of problem situations we mean not only problem-solving but problem-posing as well [15] (Figure 1).

A meta-analysis [8], which synthesized the results of 43 studies, sought to answer the following question: Do learners who learn using a problem-based approach achieve learning goals more effectively than learners who do not receive problem-based instruction? The research found an instant and lasting positive effect on learners' skills and abilities, while a negative effect was found in the area of knowledge. This analysis indicates that learners using the problem-based method have slightly less knowledge but, at the same time, remember more of the knowledge they have acquired. This mixed picture is confirmed by Hattie's study [12], which found that problem-based learning had no significant effect on student achievement. However, some researchers support the conclusion that problem-based methods improve the emotional domain of learners' learning, increase performance on complex tasks, and promote long-term retention of knowledge [1]. These studies and the

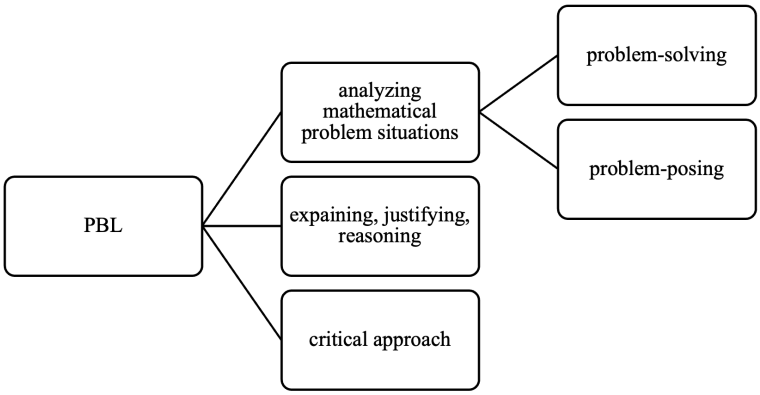


Figure 1. Definition of PBL.

contradicting results do not confirm the applicability of using problem-based learning to increase learning outcomes. However, the studies that have examined the impact of heuristic strategies on learning outcomes have all shown positive results.

Schoenfeld [18] used a control group study to show that all heuristics students improved from pre-test to post-test, while only one non-heuristics student made similar progress. In addition, heuristic students also had better persistence in problem-solving. A study by Singh et al. [19] also shows that the use of heuristics has a positive effect on the development of mathematical thinking of high school and university students, helping them to find ways to solve different problems through exploration.

Learning outcomes are closely related to the level of understanding. In the following subsection, we explore the different levels of understanding mathematics from two perspectives: algebra and geometry.

2.1. Levels of understanding mathematics

The Van Hiele levels [21–23] are regarded as a well-known model that suggest a possible way of structuring and describing people's understanding of geometry. They distinguish five levels of geometrical understanding (Table 1). According to this, a student advances sequentially from the initial to the highest level. Similar models were suggested for learning algebra as well; we considered the following six levels (Table 2) of algebraic thinking in primary and secondary education [10].

3. Research questions

Our research questions were formulated based on the previously presented research studies.

Table 1. Van Hiele levels [21–23].

Level	Description
0. vizualization	learners 'say what they see'
1. analysis	learners are aware of properties but do not reason based on them
2. abstraction	learners are able to recognize more formal properties and definitions
3. deduction	learners are able to use more formal reasoning, based on axioms, definitions and theorems
4. rigor	learners can argue precisely, comparing systems operating under different axioms and not being bound by the particularities of diagrams

Table 2. Levels of algebraic thinking [10].

Level	Description
Level 0:	learners can carry out operations with objects using natural, numerical, iconic, gestural languages
Level 1:	learners can use intensive objects (generic entities), the algebraic structure properties of N and the algebraic equality (equivalence)
Level 2:	learners can use symbolic–alphanumeric representations, although linked to the spatial, temporal, and contextual information; solving equations of the form $Ax \pm B = C$
Level 3:	learners use symbols analytically, without referring to contextual information
Level 4:	studying families of equations and functions using parameters and coefficients
Level 5:	analytical (syntactic) calculations are carried out involving one or more parameters

RQ1. How does the conscious use of heuristic strategies implemented in a problem-based approach affect students' learning outcomes?

RQ2. How do students reflect on their PBL process?

4. The setting of the study

The study was implemented with 61 students in two cycles. These students were 7th graders in the first cycle of the study and 8th graders in the second round, being

part of the lower secondary school system in Transylvania, Romania. Among them, there were two students with special needs. The instruction language was Hungarian since Hungarian was the maternal language of the students. Action research has been implemented. Action research happens when people are involved in their research-practice to improve it and better understand their practice situations [9]. The action research reported here involves one mathematics teacher–researcher (the author) from Romania teaching 7th and 8th graders and two university experts in mathematics education. We selected one-one chapter from the 7th and 8th grade curriculum and aimed to design these chapters according to the curriculum, considering previous teaching– and research experiences. The title of the chapter from the first cycle was: Equations and problems that can be solved by equations of the form $ax + b = 0,\ a \neq 0$. The second cycle's chapter topic was 3-D shapes: cube, cuboid, cone, cylinder, prism, pyramid, etc. The structure of the two chapters is shown in the following table (Table 3).

Table 3. The structure of the chapters.

	7th grade	**8th grade**
Pretest	May 2021	November 2021
Lessons	**Topic of the lesson**	
	PBL: Equations	PBL: Cube and cuboid
	Practice (2)	PBL: Prisms
	PBL: Solving $ax + b = 0, a \neq 0$ type equations (3)	PBL: Pyramids
	Practice (4)	Practice (4)
	PBL: Solving word-problems by equations (5)	PBL: Cylinders and Cones
	PBL: Practice (6)	Practice (6)
		Practice (6)
Posttest	May 2021	December 2021
	Interview – Reflection	

As the table shows four problem-based lessons were implemented in both cycles. The main heuristic strategy we used in the process of designing problem-based lessons was pattern recognition. After the intervention, as described in Table 3, a semi-structured interview was conducted with 12 randomly selected students. The random selection was done after creating three categories of students: above-average, average, and below-average learning outcomes. The average was the mean of the mathematics grades of the two classes. Definition of the groups:

Group 1: students with below-average learning outcomes (mathematics grade 6 or below 6 in the previous semester[1])

[1]In Romania, grades are on a scale of 1-10, the lowest passing grade being a 5.

Group 2: students with average learning outcomes (mathematics grade 7 or 8 grades in the previous semester)

Group 3: students with above-average learning outcomes (mathematics grade 9 or 10 in the previous semester).

Two students per class were drawn from each category for the semi-structured interview. The categorization of the students is justified by Csapó's [4] study on the relationship between grades in mathematics and attitude towards mathematics.

5. Data collection

In this paper, we examine the results of the pre-and post-tests of the two cycles (4 tests in total), as well as the students' opinions gained from the interview after the intervention. The maximum score for the first cycle's test was 20 points, and the tasks from the tests were scored with the following in mind:

- 4 points – correct solution
- 3 points – sign error or minor calculation error (slight error caused by inattention)
- 2 points – the student gave an incorrect result, but the solution was initially correct (here we do not mean slight errors caused by inattention)
- 1 point – the student tried to solve the problem but made a mistake in the first half of the solution
- 0 points – the student did not write a solution to the problem.

The maximum score for the second cycle's test was 16 points, using the same point-giving system. An exception was made for two tasks that did not need explanation (one drawing, one multiple choice type question). In these cases, the maximum point was 2 points. The students' results in both cycles were analyzed based on the three groups described above (group 1: below average, group 2: average, group 3: above average). The analysis of these results is presented in the next chapter.

6. Quantitative results of the tests

First, we analyze the results of the tests from the first cycle. The Wilcoxon test shows a significant improvement for all three groups in (Table 4).

Although the p-significance value for the 3rd group is higher than for the other two groups, this result is also significant at the 0.05 level. This is due to the fact that 10 out of 17 students in this group (above-average students) already scored very high on the pretest (18/20-20/20). The students' averages are also plotted on a bar chart (Figure 2).

Table 4. The results of the tests in the 1st cycle.

Group	Wilcoxon test (W)	Wilcoxon test (z)	Wilcoxon test (p)	Effect size
1st group	8.000	-3.243	< 0.001	-0.895
2nd group	1.000	-3.351	< 0.001	-0.983
3rd group	0.000	-2.666	0.004	-1.000

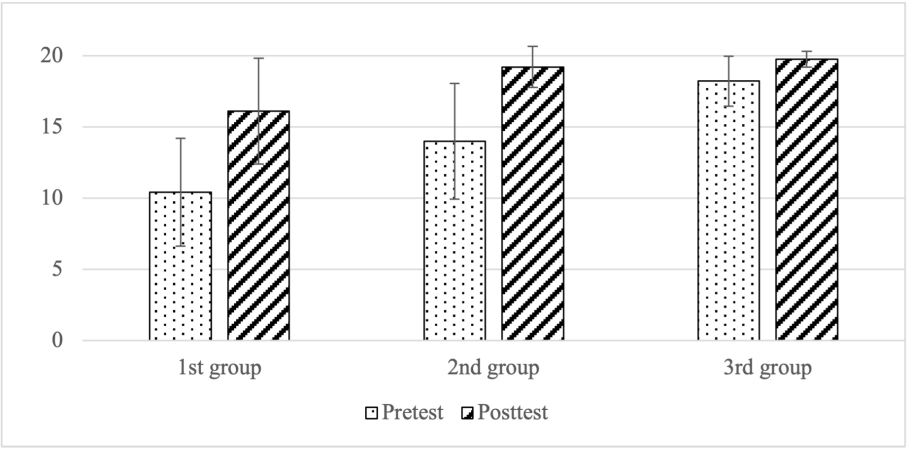


Figure 2. The results of tests from the first cycle.

The tests of the second cycle were analyzed similarly to the first one, with the following results (Table 5).

Table 5. The results of the tests in the 2nd cycle.

Group	Wilcoxon test (W)	Wilcoxon test (z)	Wilcoxon test (p)	Effect size
1st group	52.500	-0.801	0.217	-0.228
2nd group	10.000	-2.481	0.007	-0.780
3rd group	1.000	-3.110	0.001	-0.978

The results from the second cycle (Figure 3) show that although there was an improvement in the 1st group of students, it was not significant. However, in the 2nd and 3rd groups, learning outcomes improved significantly. The validity of this is made more significant in practical terms by the effect size.

The effect size for the Wilcoxon test is a measure of the difference's magnitude between two paired or matched samples. The value can range from -1 to 1, with values near 0 indicating that there is no effect and values near -1 or 1 indicating a

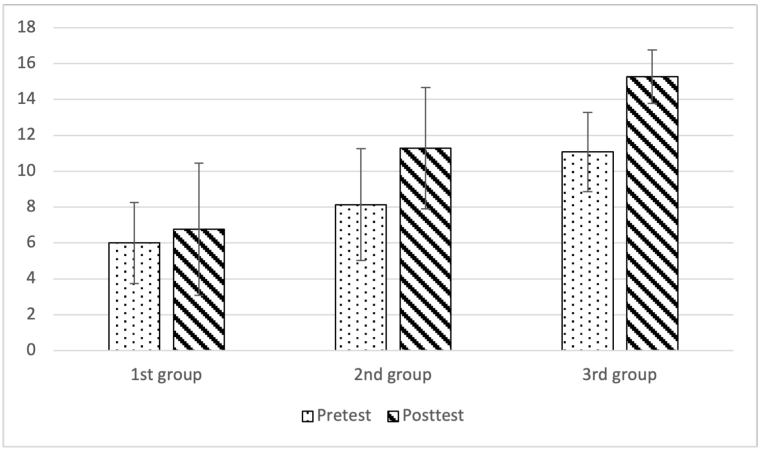


Figure 3. The results of tests from the second cycle.

strong effect. Although the analyses and effect size indicate a significant improvement overall, it cannot be said with certainty that this was due to the applied method. That is why we asked the students to share their experiences connected to these classes. The following sections consist of a qualitative analysis presenting the students' answers.

7. Qualitative analysis of students' answers

After having finished the teaching unit using a problem-based approach, we wanted to obtain the students' opinions, so we conducted a semi-structured interview with 12 students, as described earlier. In this section, we present why we are convinced that the significant increase in students' learning outcomes can be attributed to the intervention. We organized these reasons into six items by the motifs in the students' explanations:

1. the applied problem-solving activities can be used to develop logical thinking:

   S29: I like [. . .] that you have to think if I add another one, how it changes, and then you observe in a series how it changes and according to what equations, and this way, **you develop logical thinking.**

   S37: I **understood and had the logic in my head** how the results could come out.

2. the applied problem-solving activities allow exploring and expressing one's ideas:

   S53: [What I like is] that **you can discover something** that maybe **no one else has discovered**, and then **you're the first to notice** it, so to speak.

S13: What [I like] is that you can figure out the logic of how to do it, and then we usually **get the chance to figure out how to do it by ourselves**...

S32: I get a formula, or not necessarily a formula, just a rule that fits any of these numbers, I can apply it to any of these cases, **discover the relationship between them** and it feels good.

3. these activities can boost self-confidence and make the student feel good:

   S13: if I find out, then I'll **have a little confidence** so that I'm not such a lost cause after all. [Smiles] And **that makes me happier**...

   S32: I can [...] discover the relationship between them and **it feels good**.

   S17: In this class [...] **I felt confident** [...] **because**, if it's presented like that, I can **understand** it more easily than, [...] the exercises that are explained [...] if they say "expression", I'm like, "What?!" But if they tell me that it's the task with the toothpicks, then **I remember and understand how to solve it.**

4. the problem-posing activities require and develop creativity:

   S2: very, very good activities and I think **they also develop creativity**

   S54: I like it, it's good. **It's imaginative**, or how shall I say?

   S13: Well, **my creativity is not so unlimited**. [...] but if I have a starting point and I'm given what it's about, **I can figure it out**.

5. through problem-posing it is possible to have better insight into the structure of tasks:

   S28: [problem-posing] **allows me to see behind things** [...] how to solve them, in which case I'm the inventor, so **I set the boundaries** and do what I want with the task...

   S2: **you see the structure of the task, and how it is built**.

   S56: Well, I like to pose problems, because at least it's mine, and **I know that I put it together** and I know what it's about...

6. these kinds of activities help mathematical understanding:

   S2: [problem-posing] **contributes even more to our understanding**. And [...] we aren't solving only boring tasks, [...] they have **contributed to a better understanding.**

   S16: for example, if I have to solve a problem involving an equation, I may not understand it. However, if I have to formulate the equation or the text, **the whole thing becomes clearer and easier to understand.**

   S2: **helps you to understand** because you see the structure of the task, and how it is built.

   S29: I think it's a good idea to try to think a little bit backward [...] if you have a positive attitude, **it's helpful**...

> S28 Well, I think it's good if we can pose the problems, because **it will be much easier to understand** when we get such a task in the exam, and **it will be much clearer what to do**.
>
> S 58: [these activities] were good, because [...] **if you don't understand** it, and you **come up with an exercise** that's like the one you did in class, [...] **you might understand it better**.

Overall, we can say that students see problem-solving and -posing activities as logic developer activities, which also give them confidence. They find them challenging, but also creativity-boosting, which can promote better understanding. The perceived competence identified in students' opinions also supports the impact of the development.

8. Discussion

Considering the previously presented students' opinions we can claim that the significant increase in the learning outcomes must have a connection with the applied methods. This is supported by the perceived competence in students' answers. However, the results presented in the quantitative analysis show a symmetric structure in terms of the rate of development over the two cycles. Kruskal-Wallis test shows that the rate of development is significantly group dependent ($U(2) = 11.233, p = 0.004$). The reason why the 3rd group increased the least their learning results in the first cycle is apparent: their pretest scores were too high for that. The interesting question is how the 1st group managed to make a considerable improvement in the first cycle and barely improved in the second one. We assume that, although they were on the correspondent level in terms of content in the algebraic levels, they were not at the proper input level in the geometry chapter. The teaching process must start at the proper Van Hiele level to move from one level to the next. Moreover, if someone does not reach the expected entry-level, they will not be able to develop their understanding during the course [20]. This suggests that there could be several reasons why this group lacked in significant progress, for example: (1) the teacher did not design the chapter in such a way as to ensure development for the learners at the lower levels; (2) the material to be taught is too demanding for students of this ability, i.e., the curriculum is not based on age-appropriate Van Hiele levels for them. In any case, this means we need to explore the issue in more detail, which implies further research.

9. Conclusion

We designed a two-cycle action research to explore students' learning outcomes affected by PBL and the purposeful use of heuristic strategies. We applied a problem-based approach for two chapters (one algebraic, one geometric) from the curriculum. We found that the combination of PBL and conscious use of heuristic

strategies positively impact students' learning outcomes, and we managed to explore this effect from both algebraic and geometric perspectives. The students who took part in the study were separated into three groups, taking into account their previous learning outcomes. We report a significant increase in students' learning outcomes, shown in both cycles of the experiment. However, while the 1st group of students (below-average) produced a significant increase in the first cycle (algebra topic), their development in the second cycle was not significant. We concluded that factors affecting the rate of development in terms of learning outcomes need further research. Although based on the quantitative analysis, we cannot conclude with absolute certainty that the significant increase is due to the method we used, the students' responses after the intervention indicate this assumption. We have summarized in six items the effects of the activities we used on students' perceived competence: problem-solving that involves pattern recognition can develop logical thinking (1), allows exploring (2) and positively impacts self-confidence (3). On the other hand, problem-posing activities develop creativity (4), give a better insight into the structure of the tasks (5) thus help mathematical understanding (6). Considering the PBL's impact on learning outcomes, we can claim that the purposeful use of heuristic strategies during PBL activities contributes to successful development. Also, the literature argues that PBL has many more beneficial effects, for example, on student's motivation, attitude, and critical thinking.

Acknowledgements. This study was funded by the Research Program for Public Education Development of the Hungarian Academy of Sciences (KOZOKT2021-16).

References

[1] D. E. Allen, R. S. Donham, S. A. Bernhardt: *Problem-based learning*, New Directions for Teaching and Learning 2011.128 (2011), pp. 21–29, doi: 10.1002/tl.465.

[2] E. Báró: *Observing critical thinking during online pair work*, in: Critical Thinking Practices in Mathematics Education and Beyond, Rzeszów: Wydawnictwo Uniwersytetu Rzeszowskiego, 2022, pp. 128–136.

[3] E. Báró: *Teaching strategies for developing critical thinking skills*, in: Critical Thinking in Mathematics: Perspectives and Challenges, Rzeszów: Wydawnictwo Uniwersytetu Rzeszowskiego, 2022, pp. 128–136.

[4] B. Csapó: *A tantárgyakkal kapcsolatos attitűdök összefüggései*, MAGYAR PEDAGÓGIA 100.3 (2000), pp. 343–366.

[5] C. Csíkos: *Problémaalapú tanulás és matematikai nevelés*, Iskolakultúra 20.12 (2010), pp. 52–60, doi: 10.25656/01:7122.

[6] J. Dewey: *Democracy and education: An introduction to the philosophy of education*, Macmillan Publishing, 1916.

[7] J. Dewey: *How we think*, D C Heath, 1910, doi: 10.1037/10903-000.

[8] F. Dochy, M. Segers, P. Van den Bossche, D. Gijbels: *Effects of problem-based learning: a meta-analysis*, Learning and Instruction 13.5 (2003), pp. 533–568, issn: 0959-4752, doi: 10.1016/S0959-4752(02)00025-7.

[9] A. Feldman: *Editorial*, Educational Action Research 25.5 (2017), pp. 669–672, doi: 10.1080/09650792.2017.1388342.

[10] J. D. Godino, T. Neto, M. R. Wilhelmi, L. Aké, S. Etchegaray, A. Lasa: *Algebraic reasoning levels in primary and secondary education*, in: ed. by K. Krainer, N. Vondrová, Proceedings of the Ninth Congress of the European Society for Research in Mathematics Education, 2015, pp. 426–432.

[11] P. R. Halmos: *The Heart of Mathematics*, The American Mathematical Monthly 87.7 (1980), pp. 519–524, doi: 10.1080/00029890.1980.11995081.

[12] J. Hattie: *The paradox of reducing class size and improving learning outcomes*, International Journal of Educational Research 43.6 (2005), pp. 387–425, doi: 10.1016/j.ijer.2006.07.002.

[13] E. Kónya, Z. Kovács: *Management of Problem Solving in a Classroom Context*, Center for Educational Policy Studies Journal (2021), doi: 10.26529/cepsj.895.

[14] R. E. Mayer: *Learning and Instruction*, Pearson Merrill Prentice Hall, 2003.

[15] E. Pehkonen: *Use of open-ended problems in mathematics classroom*, Dept. of Teacher Education, University of Helsinki, 1997.

[16] G. Polya: *On Learning, Teaching, and Learning Teaching*, The American Mathematical Monthly 70.6 (1963), pp. 605–619, (visited on 01/30/2024).

[17] L. Y. Sari, M. Adnan, H. Hadiyanto: *Enhancing Students' Active Involvement, Motivation and Learning Outcomes on Mathematical Problem Using Problem-Based Learning*, International Journal of Educational Dynamics (2019), url: https://api.semanticscholar.org/CorpusID:164775110.

[18] A. H. Schoenfeld: *Explicit Heuristic Training as a Variable in Problem-Solving Performance*, Journal for Research in Mathematics Education 10.3 (1979), pp. 173–187.

[19] P. Singh, S. H. Teoh, T. H. Cheong, N. S. Md Rasid, L. K. Kor, N. A. Md Nasir: *The Use of Problem-Solving Heuristics Approach in Enhancing STEM Students Development of Mathematical Thinking*, International Electronic Journal of Mathematics Education 13.3 (2018), doi: 10.12973/iejme/3921.

[20] Z. Usiskin: *Van Hiele Levels and Achievement in Secondary School Geometry*, Chicago: University of Chicago, 1982.

[21] P. M. Van Hiele: *Development and the learning process*, Acta Paedagogica Ultrajectina 17 (1959), pp. 1–31.

[22] P. M. Van Hiele: *Structure and Insight: A Theory of Mathematics Education*, Computer Science and Applied Mathematics, Academic Press, 1986.

[23] P. M. Van Hiele, G. Van Hiele: *A method of initiation into geometry at secondary schools*, in: Report on methods of initiation into geometry, ed. by H. Freudenthal, J. B. Wolters, 1958.

Annales Mathematicae et Informaticae
60 (2024) pp. 190–204
DOI: 10.33039/ami.2024.09.002
URL: https://ami.uni-eszterhazy.hu

Monitoring activities in prospective teachers' mathematics lessons

Márton Kiss[ab], **Eszter Kónya**[ab]

[a]University of Debrecen,

[b]MTA-Renyi-ELTE Research Group in Mathematics Education
kiss.marton@science.unideb.hu
eszter.konya@science.unideb.hu

Abstract. Research shows that teachers rarely or never pay attention to the "how" and "why" of using metacognitive skills. Mathematics teaching is more effective when metacognitive activities (planning, monitoring, and reflection) are implemented, and coherent discourses are developed in the classroom. The question arises as to whether the handling of these activities should be taught in a targeted way in teacher training or whether they are spontaneously integrated into practice by the end of the training. Our research investigated the emergence of monitoring activities in the lessons of prospective mathematics teachers in their secondary school teaching practice during classroom discussions. From the analysis of seven case studies, we conclude that monitoring activities are rare in the classroom despite the potential to trigger them is present. Prospective teachers cannot handle these situations appropriately spontaneously, so they need to be trained to deal with them.

Keywords: metacognition, monitoring, teacher training

AMS Subject Classification: 97C70, 97D40

1. Introduction

The starting point for our research is the experience that metacognitive activities are not emphasized enough in mathematics classroom discussions. The question arises whether the handling of these activities should be taught in a targeted way in teacher training or whether they can be incorporated spontaneously by the end of the training. Our research investigated the emergence of monitoring activities in the lessons of prospective mathematics teachers in their secondary school teaching

Submitted: May 24, 2024
Accepted: September 25, 2024
Published online: October 24, 2024

practice during classroom discussions. We analysed the lessons of 7 prospective teachers from the point of view of how they managed situations that encouraged students to engage in monitoring activities in the lessons we highlighted. These classroom discussions followed the students' independent problem-solving and discussed the results, i.e., they focused primarily on the verification stage of problem-solving [19]. Our observations were not preceded by any targeted intervention or development in this area from the perspective of either the teacher candidates or the students. Our research investigated lessons in which students could search for and analyse errors. Lucangeli and colleagues [21] argue that error analysis can effectively promote the development of both metacognitive activities and mathematical performance.

2. Theoretical background

Schoenfeld [25] divided mathematical knowledge and behaviour in problem-solving into five interrelated components:

- Cognitive resources (knowledge base)
- Heuristic (problem-solving) strategies
- Monitoring and control (metacognition)
- Beliefs
- Practices

For our research, the monitoring and control (see later together as monitoring), or in a broader sense, the metacognitive component, will be highlighted and examined.

According to Flavell [9, 11], metacognition is an individual's knowledge of their cognitive processes and active monitoring, consistent regulation, and control of these processes. In most cases, the interpretation of metacognition presupposes at least a level of awareness that one should be able to report one's thoughts [6]. In this paper, we interpret such awareness in terms of metacognition.

After almost 20 years of work since the emergence of the concept, Flavell, Miller, and Miller [12] identified two main components of metacognition: metacognitive knowledge and metacognitive skills. Metacognitive knowledge refers to an individual's knowledge of their information processing abilities, the nature of cognitive tasks, and coping strategies for such tasks. Metacognitive activities refer to the regulative and executive skills (metacognitive skills) related to planning, monitoring, and reflecting on one's cognitive activities [22]. In this research, we investigate the monitoring activities of secondary school students from a specific perspective. Monitoring refers to the monitoring and controlling of cognitive activities and their outcomes [22].

As a metacognitive activity, monitoring occurs in all phases of problem-solving. It refers to continuously monitoring and evaluating one's cognitive activities to

trigger regulatory processes [26]. Typically, monitoring activities are evaluated retrospectively in the final phases of the problem-solving models immediately after the cognitive task has been performed [13].

The relationship between metacognition and mathematical performance has been investigated in several studies. Among others, the 2003 PISA surveys, based on a sample of 1433 15-year-old students, showed that the relationship between an individual's level of metacognitive development and mathematical performance is significant [24]. All researchers report that metacognitive developmental level is one of the most important predictors of mathematical performance [7, 24, 32]. Verschaffel [31] concluded that metacognition is particularly important in mathematical problem-solving.

Several studies have shown a significant correlation between monitoring activities and mathematical performance in secondary school students [23, 28]. Incorrect monitoring can lead to deficits in activating relevant content knowledge and regulating cognitive processes [14]. Consequently, the quality of monitoring affects performance on a given task in the short term and the accumulation of cognitive and metacognitive knowledge relevant to mathematical problem-solving in the long term [20].

Depaepe and colleagues [7] have shown that teachers rarely or not at all address the "how" and "why" of using metacognitive activities. Dignath and Büttner [10] confirmed that teachers teach mainly cognitive and very few metacognitive strategies, suggesting that teachers may need training and explicit instruction on metacognition. Veenman and colleagues [30] found that teachers are willing to put effort into teaching metacognition within lessons but need the tools to implement metacognition as an integral part of their lessons. A study by Wafubwa and colleagues [33] investigated 213 Kenyan mathematics teachers' perceptions of metacognition and the impact of certain background factors on metacognition. An essential message of the study is that teachers can only teach students to be metacognitive if they are metacognitive themselves. It is necessary that teachers consciously use metacognitive activities and that their use is appropriately demonstrated and taught to students. The results pointed to the need to train teachers in implementing metacognition. Teachers mentioned large class sizes, heavy workloads, and lack of motivation as reasons for not using metacognitive activities. Teachers' use of metacognition was not influenced by teaching experience or educational qualifications.

Most researchers claim that metacognition can be learned and should be taught [3, 15, 21, 24, 27]. Studies have shown that mere observation or time spent on reflection is insufficient [8]. Metacognitive skills must be taught explicitly to develop mathematical skills, as they do not develop spontaneously from implicit exposure [1, 8]. One means of doing so is to expose learners to metacognitive-mathematical discourse while explicitly addressing the "how" and "why" of metacognitive activities. Cohors-Fresenborg and Kaune [4], and Kaune [16] have developed a mathematics curriculum that emphasized metacognitive activities and demonstrated the importance of a discourse-based learning culture for understanding mathematical

problems. The learning process in the classroom can only lead to deep understanding when the associated metacognitive activities are elaborated, and coherent discussion and debate take place. Therefore, the class discussion should include discourse [5, 22].

- Discursivity refers to the activities carried out to support the coherence and accuracy of the discussion. Examples of discursive activities include the accurate (re)formulation and comparison of students' ideas, strategies, concepts, and misconceptions, as well as the linking of these to mathematical concepts or arguments.

- In contrast, negative discursivity refers to activities that negatively affect understanding meaning. Examples include self-answers or self-responsive questions, the use of inappropriate vocabulary or superficial and unclear sentences, the incorrect logical structure of an argument, and the introduction of an alternative idea into the discourse without reference to what has been said before.

The approach just mentioned is not new. In Hungarian mathematics education, similar ideas can be found among the methodological principles of Tamás Varga [29], such as shaping students' oral and written expression, developing independent opinion-making, and encouraging discussion. When students are given the opportunity to think independently in the classroom and to express their observations in their own words, the teacher is put in a situation where his or her cognitive load increases [18]. The cognitive load refers to the load in working memory (short-term memory) generated during the processing of information [2]. This is because the teacher has to immediately understand and evaluate unexpected situations and then decide how to deal with the situation that arises. The teacher can reduce this burden by making the lesson as predictable as possible, i.e., by adopting the students' expressions and suggestions, keeping the discussions between students in the background, and speaking and explaining rather than giving the lead. However, this differs from, or even contrary to, the methodological principles and approach mentioned above [18].

When these aspects are not taken into account in teacher training, it is no wonder that novice and in-service teachers opt for a teacher-dominated style of lesson management, with much less cognitive load, consisting of definitions and routine tasks, rather than a problem-focused approach to the subject matter, simply because it is instinctively less cognitive load [18].

3. Research questions

Q1. Do situations occur that can trigger monitoring activities in the lessons of teacher candidates during classroom discussions?

Q2. If yes, how do the teacher candidates manage these situations?

Q3. What factors might influence the occurrence of monitoring activities in classroom discussion during the teacher candidates' lessons in classroom discussions?

4. Method

Seven teacher candidates in their final year of training were asked to implement a lesson in a secondary school classroom during their coherent teaching practice. The lesson had to include at least one episode where students could actively participate in classroom discussions. The idea behind our request was that the teacher candidates should allow students to express their ideas and arguments. Teacher candidates were also asked to introduce a task that was solved incorrectly or to create a situation where a student's incorrect solution was analysed and discussed together. To meet our request, we suggested that teacher candidates present a problem solved incorrectly or create a situation where a student's incorrect solution is analyzed and discussed. We studied one lesson from each of the seven teacher candidates, focusing on such episodes.

Table 1 shows the identifies the grades of the classes, lesson themes, and episodes' themes with their duration time. The investigated episodes were not of equal duration. The duration of each episode was determined by the situation.

Table 1. Overview of the investigated episodes.

Prospective teacher	**Grade**	**Lesson theme**	**Episode**
A	9.	Proportionality	Discussing a faulty solution given by a group.
B	12.	Square root equations	Analysis of an incorrectly solved square root equation.
C	10.	Second-degree equations	Which of the two different solutions to the four short equations is correct?
D	9.	Solving word problems with equations	Which of the two different solutions is correct?
E	9.	Linear equations	Analysis of an incorrectly solved equation.
F	9.	Algebraic expressions	Which of the three algebraic expressions is the same as the original?
G	10.	Square root equations	Analysis of an incorrectly solved square root equation.

The lessons were audiotaped and transcribed. Furthermore, written notes were taken by an observer.

We used the categorisation system developed by Cohors-Fresenborg and Kaune (2007) to categorise classroom monitoring activities and discourses. Our coding

system addresses the teacher and student questions/responses and the type of discursivity, as shown in Table 2 .

Table 2. The coding system.

Teacher asks	T?
Teacher answers	T!
Student asks	S?
Student answers	S!
Discursivity	D
Negative discursivity	ND

Dialogues are identified as units with the same number after the letter T or S. (For example, T1?, S1! is a dialogue unit, or S2?, T2?, S2! is another dialogue unit).

Manifestations that encourage or suggest monitoring activities are marked in red in the lesson.

In the presentation of the episode, some manifestations have been annotated to give a qualitative analysis of the situation regarding encouraging monitoring activities and involving students. The comments were motivated by the need to learn lessons and provide further advice, not by blaming the teacher candidates.

After coding, the following criteria were used to characterise the episodes:

- How many monitoring situations have been created in the episodes? How was the situation handled?
- How many discursive (D) and negative discursive (ND) manifestations did the teacher candidate have?
- How many Student-Student (S-S), Teacher-Student (T-S), Teacher- Teacher (T-T) dialogues occurred during the episode?

5. Results and discussion

5.1. A case study of a prospective teacher

The prospective teacher (denoted by A) was teaching a lesson to a 9th-grade class in an urban high school in 2021. The students have three mathematics lessons per week. The lesson theme was proportionality and percentages.

The problem statement that was the starting point for the subsequent discussion was the following.

Klári wants to paint the walls of her kitchen purple. The purple paint is mixed for her from three colours: blue, red, and yellow. The ratio of blue, red, and yellow in the mixture is $4 : 5 : 1$. *Six litres of blue,* 9 *litres of red, and* 2 *litres of yellow*

paint were found in the warehouse. What is the maximum number of litres of purple paint that can be mixed from the warehouse stocks?

An expected solution applying the trial-and-error strategy is to check the ratio of the quantities of yellow, blue, and red paints using all the stored yellow, blue, and red paints, respectively. The answer is 6 litres of blue, $5 \cdot 1.5 = 7.5$ litres of red, and $1 \cdot 1.5 = 1.5$ litres of yellow paints, which give 15 litres of purple paint. Starting from the other two colours, we arrive at a contradiction.

Prospective teacher A formed groups of 4-5 students and asked them to solve the problem together and discuss the ideas if they had more than one. However, we observed groups in which only a few members discussed the problem, and the others worked independently. In the meantime, the teacher asked the groups if they needed help, how they were doing, how much time they needed, etc. After a while, he stopped waiting, and the class discussion started.

T: Good. Let's look at it, ... let's look at it together!

S: Teacher, I'm not sure we're going in the right direction ... – S1!

The situation initiated by the student may trigger monitoring activity (1). *The student's certainty about the correctness of his answer is weakened. This is an excellent opportunity for the teacher to initiate a discussion involving as many students as possible.*

T: How did you get started? – T1? **D**

The teacher opens a dialogue. *The question encourages monitoring activities, although this could have been preceded by asking the students what made them think they were not going in the right direction. This could have been a more valuable question from a metacognitive point of view because the students would have had to articulate the reason for the uncertainty.*

S: We started by writing 4, 5, and 1 in x and making an equation. – S1!

T: We know that the paint ratio is $4 : 5 : 1$. – T2! **D**

The student's mathematical language is problematic, and the teacher pointed out the related statement from the problem text to make it more transparent for the audience.

T: And what else do we know? How many litres of paint? – T2? ND

They started solving the problem again together, hoping the mistake would be found.

S: 6 litres blue, 9 litres red, and 2 litres yellow. – S2!

T: Good. We know these things. – T2!

S: Yes, we wrote these down so that $4x + 5x + x$ is $10x$, equal, and we added 6 to 9 and 2, which is 17. – S2!

The opportunity to trigger the monitoring activity is no longer available. A discussion and so a rich manifestation of monitoring activity have not emerged.

T: What is the problem with this? – T3? **D**

The situation initiated by the teacher may trigger monitoring activity (2). *Her question immediately suggested that the equation was wrong. She could have asked the students a less revealing question to get their opinions.*

T: The solution does not consider the proportions. Since you have poured all the paints together, they may not be in the same proportion ... – T3! ND

The teacher did not take advantage of the situation. She answered herself vaguely. He answered for himself, and not clearly. He could have tried a different question (e.g., What does this solution not consider?). She should have given the students more space to think.

S: 17. – S3!

Students did not understand the previous help.

T: You may not get these proportions right when you pour all the paints together. – T4! ND

The opportunity to trigger the monitoring activity is no longer available. *The teacher answered herself again instead of asking the students.*

T: So ... you've come up with 1.7, which others have also whispered. – T4!

S: So, ours is not good for sure, teacher? – S5?

The situation initiated by the student may trigger monitoring activity (3). *The student did not understand what he had not considered in his solution.*

T: Let's have a look! If x is 1.7, then ... – T5! **D**

The teacher opens a dialogue. *It was good that she did not give an immediate answer ("not good") but tried to help students see the contradiction.*

S: $6 \cdot 1.7$, and then we multiplied it, ... but why ... – S5!

The situation initiated by the student may trigger monitoring activity (4). *Why did the student multiply even by 6?*

T: Yes, and then you multiplied it also by 4, ... – T5! ND

The teacher did not take advantage of the situation. *She quickly took control of the situation. She could have reacted to the "but why ..." part of the student's sentence by asking the others to explain. From a metacognitive point of view, it was a valuable moment that could have been elaborated.*

S: ... and by 5, then you've got how many litres of blue you need, ... according to the ratios, yes, if you've got that, then you've got the red, which ... – S5!

T: And did it come out? – T6? ND

The situation initiated by the teacher may trigger monitoring activity (5). *However, the question is too direct and suggestive. (Instead, e.g., What did you get? How do you evaluate the result?)*

S: We need 6.8 litres of blue, which is not good! – S6!

Now, it is clear that the solution is wrong. Although the student's answer refers to some monitoring activity, he did not have to add much work.

T: Yes, it is too much. Unfortunately, it's not good. Has anyone done it differently? – T6? **D**

The teacher did not take advantage of the situation. *It was good that the teacher did not explain the correct solution herself but asked the other students. Moreover, the lesson continued ...*

The episode analysed dealt with a faulty solution that spontaneously emerged and had to be evaluated. One of the groups of students' solutions was discussed. At the end of the discussion, the students realised, with the help of the prospective

teacher, that their solution needed improvement. The prospective teacher left another group of students the opportunity to explain the correct answer.

Table 3 shows that in the episode, we identified mainly teacher–student dialogues (T–S) and examples of the teacher answering her own question (T–T). No examples of student-to-student dialogue (S–S) were found.

Table 3. Types of the dialogues.

Type of the dialogue	**Number**
S–S	0
T–S	4
T–T	2

Table 4 presents that the prospective teacher's manifestations contained as many discursive as negative discursive elements.

Table 4. The quality of the discourse.

Quality of discourse	**Number**
discursivity	5
negative discursivity	5

In Table 5, we can see 5 situations to stimulate monitoring activities have emerged, 2 of which were initiated by the teacher and 3 by the student. In two situations, the prospective teacher successfully encouraged the student to engage in monitoring activities, but she did not take advantage of three situations.

Table 5. The number of promising situations related to monitoring activities.

	Initiated by the prospective teacher	**Initiated by the student**
Promising situations	2	3
Monitoring activity has occurred	1	1

From the metacognitive point of view, we can establish that the episode provided an excellent opportunity for monitoring activities. However, several of these metacognitively valuable situations were left unused. There were some self-answering questions or self-responses from the prospective teacher, which were classified as negative discursivity. Some of the teacher's prompts were too directed without giving space for students to formulate their ideas. Because the prospective teacher

had robust control over the lesson, she missed several opportunities for more inclusive teaching. While she attempted to point out student mistakes, she could have involved more students in the discussion, creating a more conducive environment for dialogue.

5.2. Overall findings of the case studies of the seven prospective teachers

The chart below shows that the dialogues were mainly teacher-student dialogues, which is not surprising Figure 1. It should be added that many of these dialogues were characterised by long teacher utterances, or at least longer than those of the students. Students were often terse in their statements. The phenomenon of "talking to oneself" was observed in five of the seven teacher candidates. That is, when they asked a question, they answered it, or, after having expressed a unit of thought, they started a new one without letting the students reflect. The student-student dialogue was sporadic. An example of this was only in the lessons of two prospective teachers.

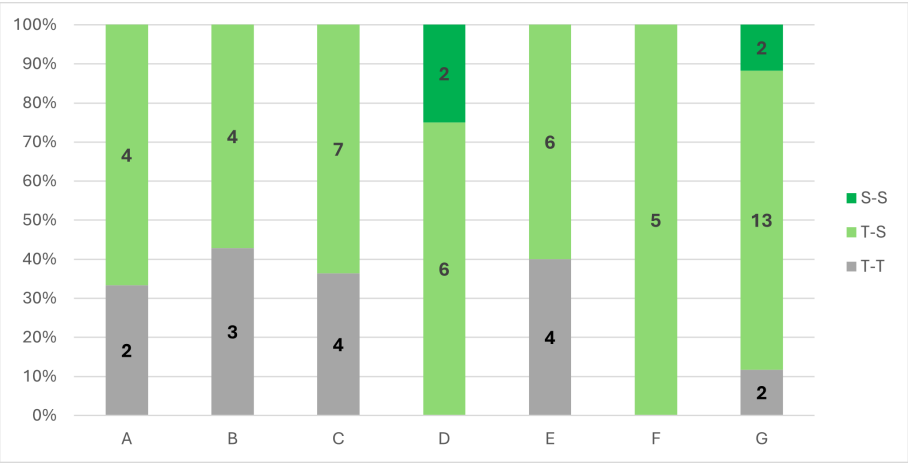


Figure 1. Types of dialogues in lessons.

Figure 2 shows the proportion of prospective teachers' manifestations identified as discursive (D) and negative discursive (ND). Discursivity, i.e., actions taken by the teacher to improve classroom communication, was the majority in the case of three prospective teachers (D, E, and G). The ratio of discursivity and negative discursivity was nearly the same for the two (A and F). For two teacher students (B and C), negative discursivity, i.e., unnecessary, misleading, or self-responsive contributions, was in most of the observed lessons. It should be highlighted that there was one prospective teacher in whom only discursivity and one in whom only negative discursivity was identified.

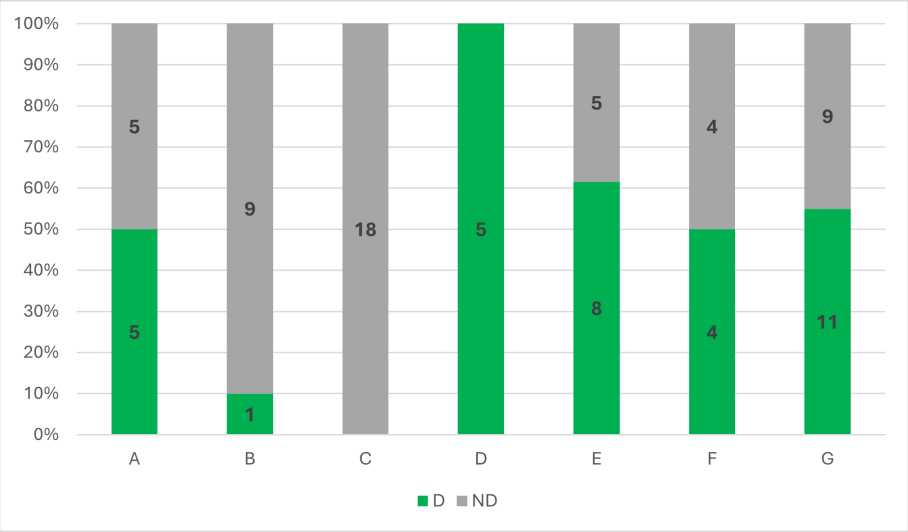


Figure 2. The rates of discursivity and negative discursivity.

Table 6 summarises the number of promising situations related to monitoring activities by prospective teachers. We found such situations implemented successfully for all prospective teachers except C. The relatively small number of such situations does not allow us to rank the prospective teachers; furthermore, the teaching episodes analysed were of different topics, durations, and intensities.

Table 6. Promising situations for monitoring activities in prospective teachers' lessons.

Prospective teacher	**Promising situations**	**Monitoring activity has occurred**
A	5	2
B	2	1
C	0	0
D	6	4
E	4	2
F	3	1
G	7	5

6. Summary and conclusions

The paper examined seven prospective teachers' lessons in seven different high schools. Teacher students were asked to create helpful discussions with students and were given the opportunity to share a faulty solution with the class. We aimed to observe how monitoring activities appear in classrooms under these conditions. Furthermore, we investigated the role of prospective teachers in encouraging monitoring activities [17].

We answer our research questions as follows.

Q1. Do situations occur that can trigger monitoring activities in the lessons of teacher candidates during classroom discussions?

Yes. Promising situations were observed that initiated by both prospective teachers and students.

Q2. How do prospective teachers manage these situations during classroom discussions?

Prospective teachers did not really engage the students or unfold the possibilities. They mostly did not have suitable questions to do so. In about half of the cases, they managed to get students to do monitoring activities, but the situation promised more than was achieved in all cases.

Q3. What factors might influence the occurrence of monitoring activities in classroom discussion during the teacher candidates' lessons in classroom discussions?

The prospective teacher's questioning culture, the degree of teacher guidance, the types of dialogue, and the students' activity can play decisive roles in the correct handling of metacognitively promising situations.

Overall, developing a meaningful discussion, a dynamic dialogue between students or teacher and students was very difficult or almost impossible. The tasks and the occurring situations held much more potential for engaging students in classroom discussion and encouraging monitoring activities than what materialised.

Our results support the findings discussed in the literature that teachers teach very few metacognitive strategies, suggesting that teachers may need training and explicit instruction [10] and tools to implement metacognition as an integral part of their lessons [30]. A study by Wafubwa and colleagues [33] reports similar experiences. Teachers can only teach students to be metacognitive if they themselves are metacognitive. It is necessary that teachers consciously use metacognitive activities and that their use is appropriately demonstrated and taught to students. Teachers should be trained in the implementation of metacognition in secondary schools.

Acknowledgements. This study was funded by the Research Program for Public Education Development of the Hungarian Academy of Sciences (KOZOKT2021-16).

References

[1] E. Ader: *What would you demand beyond mathematics? Teachers' promotion of students' self-regulated learning and metacognition*, ZDM 51.4 (2019), pp. 613–624, doi: 10.1007/s11858-019-01054-8.

[2] A. Ambrus: *A matematika tanulás-tanítás néhány kognitív pszichológiai kérdése Some cognitive psychological question in mathematics education*, GRADUS 2.2 (2015), pp. 63–73.

[3] E. Baten, A. Desoete: *Metacognition and motivation in school-aged children with and without mathematical learning disabilities in Flanders*, ZDM 51.4 (2019), pp. 679–689, doi: 10.1007/s11858-018-01024-6.

[4] E. Cohors-Fresenborg, C. Kaune: *Mechanisms of the taking effect of metacognition in understanding processes in mathematics teaching*, in: Developments in mathematics education in German-speaking countries - Selected papers from the annual conference on didactics of mathematics, Ludwigsburg, March 5–9, 2001, ed. by G. Törner, R. Bruder, A. Peter-Koop, N. Neill, H. G. Weigand, B. Wollring, Göttingen: Gesellschaft für Didaktik der Mathematik (GDM), 2001, pp. 29–38.

[5] E. Cohors-Fresenborg, C. Kaune: *Modelling classroom discussions and categorising discursive and metacognitive activities*, in: Proceedings of the Fifth Congress of the European Society for Research in Mathematics Education, (CERME 5, February 22 – 26, 2007), ed. by D. Pitta–Pantazi, G. Philippou, Larnaca: University of Cyprus and ERME, 2007, pp. 1180–1189.

[6] C. Csíkos: *A gondolkodás stratégiai összetevőinek fejlesztése iskoláskorban [Developing the strategic components of thinking at school-age]*, Szeged, Hungary, 2016.

[7] F. Depaepe, E. D. Corte, L. Verschaffel: *Teachers' metacognitive and heuristic approaches to word problem solving: Analysis and impact on students' beliefs and performance*, ZDM 42.2 (2010), pp. 205–218, doi: 10.1007/s11858-009-0221-5.

[8] A. Desoete, M. Veenman: *Metacognition in mathematics: Critical issues on nature, theory, assessment and treatment*, in: Metacognition in mathematics education, ed. by A. Desoete, Gent: Nova Science, 2006, pp. 1–10.

[9] A. Desoete, M. Veenman: *Metacognitive aspects of problem solving*, in: The nature of intelligence, ed. by B. Resnick, Gent: Nova Science, 1976, pp. 231–236.

[10] C. Dignath, G. F. Büttner: *Teachers' direct and indirect promotion of self-regulated learning in primary and secondary school mathematics classes – insights from video-based classroom observations and teacher interviews*, Metacognition and Learning 13.2 (2018), pp. 127–157, doi: 10.1007/s11409-018-9181-x.

[11] J. H. Flavell: *Metacognition and cognitive monitoring: A new area of cognitive-developmental inquiry*, American Psychologist 34.10 (1979), pp. 906–911, doi: 10.1037/0003-066X.34.10.906.

[12] J. H. Flavell, P. H. Miller, S. A. Miller: *Cognitive Development (4th Edition)*, Pearson, 2001.

[13] J. Garofalo, F. K. Lester: *Metacognition, cognitive monitoring, and mathematical performance*, Journal for Research in Mathematics Education 16.3 (1985), pp. 163–176, doi: 10.2307/748391.

[14] D. J. Hacker, L. Bol, M. C. Keener: *Metacognition in education: A focus on calibration*, in: Handbook of metamemory and memory, ed. by J. Dunlosky, R. Bjork, Mahwah, NJ: Lawrence Erlbaum Associates, 2008, pp. 411–455.

[15] D. J. Hacker, S. A. Kiuhara, J. R. Levin: *A metacognitive intervention for teaching fractions to students with or at-risk for learning disabilities in mathematics*, ZDM 51.4 (2019), pp. 601–612, doi: 10.1007/s11858-019-01040-0.

[16] C. Kaune: *Reflection and metacognition in mathematics education – Tools for the improvement of teaching quality*, ZDM 38.4 (2006), pp. 350–360, doi: 10.1007/BF02652795.

[17] M. Kiss, E. Kónya: *Analysis of metacognitive activities in pre-service teachers' lessons – case study*, in: Proceedings of the Thirteenth Congress of the European Society for Research in Mathematics Education, (CERME 13, July 10–14, 2023), ed. by P. Drijvers, C. Csapodi, H. Palmér, K. Gosztonyi, E. Kónya, Budapest: Alfréd Rényi Institute of Mathematics and ERME, 2023, pp. 3602–3603.

[18] E. Kónya, Z. Kovács: *Kognitív terhelés a problémaközpontú matematikaórákon [Cognitive load in problem-centered mathematics classroom]*, in: Módszerek, művek, teóriák – A X. Tantárgy-pedagógiai Nemzetközi Tudományos Konferencia előadásai, ed. by S. Bordás, Baja: Eötvös József Főiskolai Kiadó, 2020, pp. 279–288.

[19] F. K. Lester: *Methodological consideration in research on mathematical problem-solving instruction*, in: Teaching and learning mathematical problem solving: Multiple research perspectives, ed. by E. A. Silver, New York: Routledge, 1985, pp. 41–69, doi: 10.4324/9780203063545.

[20] K. Lingel, J. Lenhart, W.Schneider: *Metacognition in mathematics: Do different metacognitive monitoring measures make a difference?*, ZDM 51.4 (2019), pp. 587–600, doi: 10.1007/s11858-019-01062-8.

[21] D. Lucangeli, M. C. Fastame, M. Pedron, A. Porru, V. Duca, P. K. Hitchcott, M. P. Penna: *Metacognition and errors: The impact of self-regulatory trainings in children with specific learning disabilities*, ZDM 51.4 (2019), pp. 577–585, doi: 10.1007/s11858-019-01044-w.

[22] E. Nowińska: *Assessing how teachers promote students' metacognition when teaching mathematical concepts and methods*, in: Proceedings of the Eleventh Congress of the European Society for Research in Mathematics Education, (CERME 11, February 05-10, 2019), ed. by U. T. Jankvist, M. van den Heuvel-Panhuizen, M. Veldhuis, Utrecht: Freudenthal Group & Freudenthal Institute, Utrecht University and ERME, 2019, pp. 3720–3727.

[23] T. Roderer, C. M. Roebers: *Children's performance estimation in mathematics and science tests over a school year: A pilot study*, Electronic Journal of Research in Educational Psychology 11.1 (2013), pp. 5–24.

[24] W. Schneider, C. Artelt: *Metacognition and mathematics education*, ZDM 42.2 (2010), pp. 149–161, doi: 10.1007/s11858-010-0240-2.

[25] A. H. Schoenfeld: *Learning to think mathematically: Problem solving, metacognition, and sense making in mathematics*, in: Handbook of research on mathematics teaching and learning, ed. by D. A. Grows, New York: Macmillan, 1992, pp. 334–370.

[26] A. H. Schoenfeld: *Mathematical Problem Solving*, Academic Press, 1985.

[27] A. Shilo, B. Kramarski: *Mathematical-metacognitive discourse: How can it be developed among teachers and their students? Empirical evidence from a videotaped lesson and two case studies*, ZDM 51.4 (2019), pp. 625–640, doi: 10.1007/s11858-018-01016-6.

[28] S. Tobias, H. T. Everson: *The importance of knowing what you know: A knowledge monitoring framework for studying metacognition in education*, in: Handbook of metacognition in education, ed. by D. J. Hacker, J. Dunlosky, A. C. Graesser, New York: Routledge, 2009, pp. 107–127.

[29] T. Varga: *Mathematics education in Hungary today*, Educational Studies in Mathematics 19.3 (1988), pp. 291–298.

[30] M. V. J. Veenman, B. H. A. M. V. Hout-Wolters, P. Aferbach: *Metacognition and learning: Conceptual and methodological considerations*, Metacognition Learning 1 (2006), pp. 3–14, doi: 10.1007/s11409-006-6893-0.

[31] L. Verschaffel: *Realistic mathematical modelling and problem solving in the upper elementary school: Analysis and improvement*, in: Teaching and learning thinking skills. Contexts of learning, ed. by J. H. M. Hamers, J. E. H. V. Luit, B. Csapo, Lisse: Swets & Zeitlinger, 1999, pp. 215–240.

[32] L. Verschaffel, W. V. Dooren, J. Star: *Applying cognitive psychology based instructional design principles in mathematics teaching and learning: Introduction*, ZDM 49.4 (2017), pp. 491–496, doi: 10.1007/s11858-017-0861-9.

[33] R. N. Wafubwa, C. Csíkos, R. Opoku-Sarkodie: *In-service mathematics teachers' conception and perceptions of metacognition in their teaching experience*, SN Social Sciences 2.2 (2022), p. 21, doi: 10.1007/s43545-022-00321-y.

Annales Mathematicae et Informaticae
60 (2024) pp. 205–217
DOI: 10.33039/ami.2024.04.001
URL: https://ami.uni-eszterhazy.hu

Discovering epitrochoid curves with STEAM-based learning methods*

Attila Körei[a], **Szilvia Szilágyi**[b]

[a]University of Miskolc, Department of Applied Mathemathics
attila.korei@uni-miskolc.hu

[b]University of Miskolc, Department of Analysis
szilvia.szilagyi@uni-miskolc.hu

Abstract. In this paper, we present a new teaching-learning technique for drawing and identifying several members of an important family of parametric curves based on educational robotics supported by dynamic geometry software. Epitrochoid curves are essential in teaching first-year computer science and engineering students mathematics. From a methodological point of view, these are usually attractive and aesthetic curves suitable to capture students' interest and give first-hand experiences and activities. To implement a new STEAM-based learning method in this field, we created a drawing LEGO robot to visualise the epitrochoid curves and improved a virtual Epitrochoid Tracker with the Desmos graphing calculator to check the parametric equations for the drawn curve. The paper focuses on the two pillars of the developed STEAM-based learning method and their pedagogical aspects.

Keywords: STEAM-based education, dynamic geometry software, educational robotics, parametric curves, epitrochoid curves

AMS Subject Classification: 97I20, 14H50

1. Introduction

Nowadays, innovative teaching and learning methods can be applied in higher education thanks to technological advances and the spread of STEAM education. A trend in STEAM education is the integration of cutting-edge technologies [4, 17]. The necessity of integrating technology into education and using educational tech-

*The research was carried out in the framework of the RRF-2.3.1-21-2022-00013 National Laboratory for Social Innovation project.

Submitted: January 27, 2023
Accepted: April 9, 2024
Published online: April 11, 2024

nologies in the process of teaching-learning is a widely accepted idea in the field of education science [23]. Many studies demonstrate that integrating technology into mathematics education and its use in the teaching-learning process increases students' academic success and motivation, positively affects their attitudes towards learning, supports the development of students' problem-solving and cooperative learning skills, and provides teachers with more opportunities to guide their students [2, 7, 18, 21]. In today's highly digital age, portable info communication tools (ICT) have become essential to students' daily lives. Dynamic geometry software (DGS) that can also run on mobile devices helps to increase focus by capturing students' attention by providing visual elements [24]. However, digital content alone is insufficient for Generation Z students to sustain attention. Several studies have shown that first-hand experiences are the best way to achieve optimal learning outcomes for today's university students [1, 5, 10, 16]. This explains the re-emergence of hands-on and explorable models in universities and increased projects involving physical activity in mathematics courses. The 19^{th} century descriptive and differential geometry models, such as Schilling's kinematic models, are again used in university practice [19]. The limited accessibility makes it worth considering the possibilities of creating kinematic models that demonstrate a given problem well in the classroom environment. One of these possibilities is the innovative use of educational robot kits to provide didactic tools for specific chapters of mathematics.

Students at technical universities study curves of the Euclidean two-dimensional plane during their first semester, including trochoidal curves. These curves can be drawn with physical tools or virtually. We can draw by hand or with an easy-to-build device if we choose the first option. Drawing trochoids by hand requires tools; for example, we can use LEGO Technic gears to model the non-slip rolling, as demonstrated in [13].

As for drawing a virtual curve, maths teachers have plenty of possibilities to illustrate the current curriculum. In recent years, DGS has become increasingly popular because it effectively helps students understand basic and advanced geometrical concepts at all levels of education. However, because of its interactive nature, a complex DGS has many features beyond the simple representation of curves. Such a program can be used, among other things, to create spectacular animations that demonstrate the generation of curves in the process, providing a deeper understanding of this topic.

This paper focuses on one of the families of trochoidal curves, epitrochoids. We introduce a new STEAM-based method by integrating practical approaches to support the teaching-learning process.

2. Mathematical background

A roulette is a plane curve considered as the trajectory of a point rigidly connected with some curve rolling upon another fixed curve without slipping. Let us denote the fixed curve by F and the moving curve by M. The point P, which traces the roulette, is said to be the pole or the generator. If F and M are both circles, and

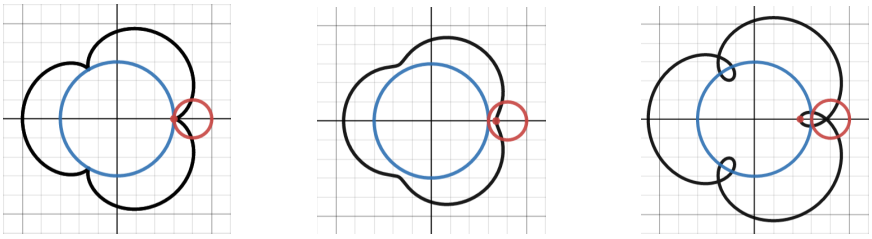
M is rolling outside of F, the obtained roulette is called an epitrochoid from the Greek words 'epi' (over) and 'trokhos' (wheel).

Suppose that a circle of radius r is rolling around the outside of a fixed circle of radius R, and the pole P is attached to the moving circle a distance d from its centre. Then, the parametric equations of the epitrochoid traced by P are

$$\begin{aligned} x(t) &= (R+r)\cos t - d\cos\left(\frac{R+r}{r}t\right), \\ y(t) &= (R+r)\sin t - d\sin\left(\frac{R+r}{r}t\right), \end{aligned} \tag{2.1}$$

where the independent variable $t \in \mathbb{R}$ denotes the angle between a line through the centre of both circles and the x-axis [15]. In the sequel, the quantities R, r and d in (2.1) will be referred to as parametric constants and d will be called the pole distance.

One of the ways of classifying epitrochoids is based on the relation between the radius of the moving circle and the pole distance. If $d = r$, then the obtained special curve is called an epicycloid (Figure 1a). We speak about curtate or contracted epitrochoid if $d > r$ (Figure 1b), and in case of $d < r$, we obtain prolate or protracted epitrochoid (Figure 1c).



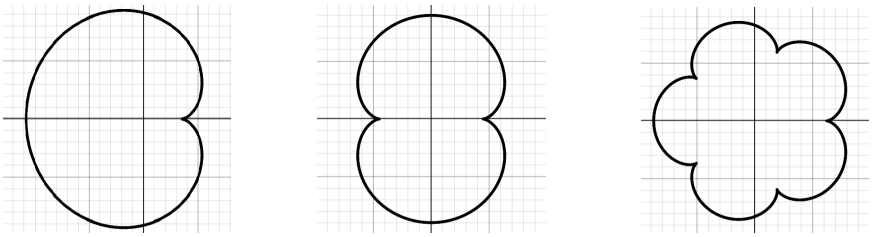
(a) Epicycloid. **(b)** Curtate epitrochoid. **(c)** Prolate epitrochoid.

Figure 1. Representative example of epitrochoid curves. An epicycloid touches the fixed circle; a curtate epitrochoid does not touch the fixed circle, while a prolate epitrochoid crosses it.

Remark 2.1. It should be noted that there is no complete agreement in the literature on the naming of cycloid-type curves. For example, in Hungary, the term epitrochoid is not used; instead, the dimpled and looped epicycloids' names are common for the curtate and prolate epitrochoids [8].

If the parametric constants of the epitrochoid are chosen specifically, we obtain special plane curves. Clearly, if $d = 0$, the trace of the pole is a circle of radius $R+r$. If the radius of the fixed and the moving circle are equal, we have a limaçon. In particular, a limaçon is called cardioid if $d = r$ also holds. For STEAM-based teaching of the cardioid curve combined with educational robotics, see [12, 14]. Some other epicycloids have been given their specific name depending on the ratio

of the radii of the circles. For example, a nephroid is an epicycloid with $\frac{R}{r} = 2$ and we have a ranunculoid when $\frac{R}{r} = 5$ (Figure 2). In general, if R and r are relatively prime numbers, the curve closes on itself and has R cusps, where a cusp is defined as a point where the epicycloid meets the fixed circle. If $\frac{R}{r}$ is irrational, the curve will never return to the initial starting point.



(a) Cardioid ($R = r$). **(b)** Nephroid ($R = 2r$). **(c)** Ranunculoid ($R = 5r$).

Figure 2. Special epicycloids.

Remark 2.2. Hypotrochoids can be generated similarly to epitrochoids, but in this case, the rolling circle is inside the fixed one. For the parametric equations of hypotrochoids, see [11], where we outline a learning project to study these curves.

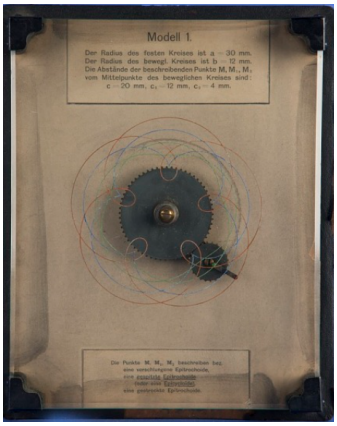
3. An innovative STEAM-based method for teaching epitrochoid curves

The approach proposed in this paper to teach and learn epitrochoid curves is based on the combination of educational robotics and dynamic geometry software.

3.1. Epitrochoid drawing devices

At the end of the 19$^{\text{th}}$ and the beginning of the last century, using physical models to illustrate certain mathematical phenomena was relatively common among mathematics teachers. These models were used mainly to represent curves and surfaces and other mathematical concepts useful to mathematicians, engineers and scientists. Most models were created in Europe, particularly in Germany, for use in school and college teaching. One of the most prominent manufacturers of the models was Martin Schilling's company. The kinematic models were designed by Frederick Schilling, a professor of mathematics at Göttingen. The 1911 Schilling Catalog lists 377 items divided into forty series. Series XXIV consists of kinematic models, of which the first group of 4 models is suitable for representing trochoids. The first model is a device for plotting epitrochoids, where a gear can be rolled along another fixed gear (Figure 3a).

While Schilling's kinematic models were designed primarily for educational purposes, the Spirograph toy was intended to entertain a wider audience. The Spiro-



(a) Kinematic model by Martin Schilling, Series XXIV, model 1, number 329.

(b) Original Spirograph set.

Figure 3. Devices for drawing trochoids.

graph kits contain serrated-edge plastic parts and can be used to create various patterns, including several trochoidal curves (Figure 3b). Spirograph parts contain small holes into which the tip of a pen can be inserted, and the pattern can be drawn by rolling the part over another fixed part. Obviously, the pole distance can only be less than the radius of the rolled circle, so only curtate trochoids can be drawn with the Spirograph toy [22].

3.2. Robot model implementation

Epitrochoid curves are derived by the non-slip rolling of a circle along another fixed circle. Non-slip rolling is achieved in most epitrochoidal drawing tools by the use of gears. Since LEGO sets contain a variety of different sizes of gears, they are suitable for building a drawing device. The prototype of our drawing robot (Figure 4) was built using elements from the LEGO SPIKE Prime set, but of course, elements from other robot kits can also be used. The factor that most determines which epitrochoids can be drawn with the robot is the size of the suitable gears in the set. The SPIKE Prime core set includes four double bevel gears with 36, 28, 20 and 12 teeth. These gears can be used to model the rolling circle in epitrochoid generation. The fixed circle was modelled with a different type of part. For this purpose, we applied LEGO Technic turntables, which are available in two sizes.

In the robot model, a motor was used to roll the moving gear along the fixed gear. The motor actually turns a lever to which the gear is attached. The connecting lever was designed to be adjustable in length to easily assemble the possible gear configurations. A drawing head is mounted on the moving gear, where the writing instrument can be placed, also at adjustable distances from the centre of the gear. This means that the pole distance can be equal to the radius of the

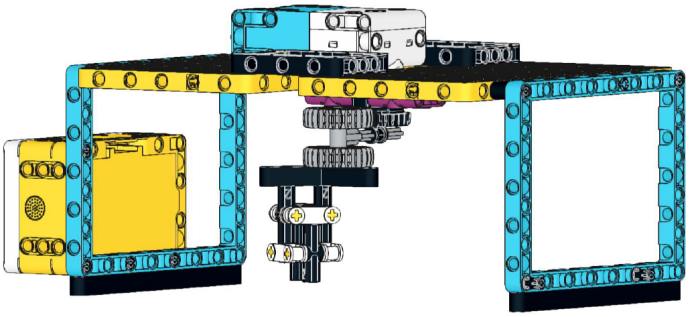
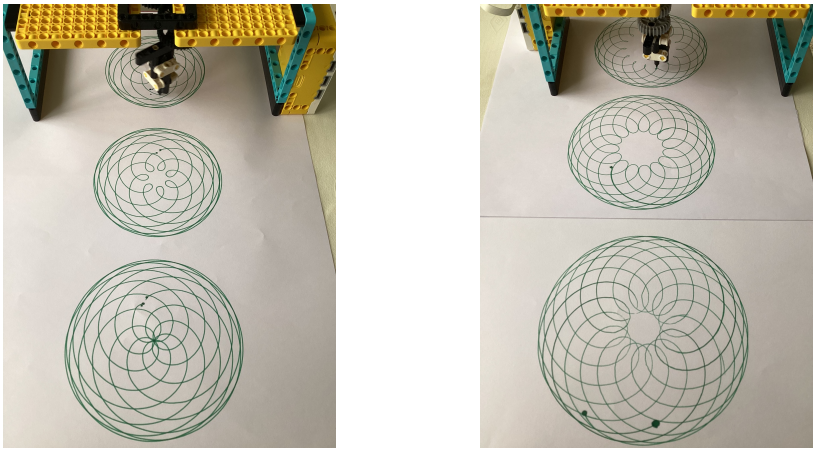


Figure 4. Epitrochoid drawing LEGO robot.

moving circle but can also be larger or smaller. This range of variability allows all three types of epitrochoids to be drawn. In Figure 5, the robot draws different curtate and prolate epitrochoids with pole distances of $d = 16\,\text{mm}$, $d = 24\,\text{mm}$ and $d = 32\,\text{mm}$.



(a) $R = 14\,\text{mm}$, $r = 18\,\text{mm}$. **(b)** $R = 30\,\text{mm}$, $r = 14\,\text{mm}$.

Figure 5. Drawing epitrochoids by different gear configurations.

The Spirograph kit contains parts of many different sizes, but of the three main types of epitrochoid, only the curtate ones can be represented. Schilling's teaching model shows a single representative of all three types because the gears built into the structure are not interchangeable, and the pole distances are fixed values. Conversely, the robot model can incorporate multiple gear combinations and variable pole distances, combining the advantages of the previous devices in a single device and far exceeding their capabilities.

3.3. Principles of STEAM-based learning method

In the Mathematical Analysis courses, first-year students have both theoretical and practical training. After a theoretical grounding, they can work on projects in small groups of students in practical sessions. Education experts say STEAM education is about more than developing practical skills alone. It also helps students develop the capacity to take thoughtful risks, engage in meaningful learning activities, become resilient problem solvers, embrace and appreciate collaboration and work through the creative process [3]. For teaching epitrochoid curves, a LEGO robot should not be used solely as a teacher demonstration tool but is best used as a drawing robot to be built, programmed and operated by the students. In this spirit, our STEAM-based learning method can be represented as a cyclic chain. Repeating segments of the chain consists of five main stages:

(1) Description of the problem, theoretical background, setting goals and tasks.

(2) Practical problem solving, i.e. build (or rebuild) and test a LEGO SPIKE Prime drawing robot.

(3) Determination of the parametric equations for the drawn epitrochoid curve.

(4) Checking the written parametric equations using dynamic geometry software.

(5) Summary of experiences, conclusions of the project, making a presentation.

The chain representing the method consists of repetitive parts because the robot needs to be modified in a targeted way to draw the different epitrochoid curves. The above method provides an opportunity to engage students in cooperative learning and teamwork, where they can share their ideas and apply new knowledge to gain a deeper understanding of the problem. A considerable advantage of epitrochoids is the wide range of curves that can be drawn, making it smooth and easy to perform cyclic steps. In terms of the teaching-learning process, this cyclicality has a positive effect as students become more proficient with both the drawing robot and the geometrical application. In this way, we can integrate educational robotics into the practices of Mathematical Analysis courses.

Building and rebuilding a drawing robot is one of the pillars of the method developed. It was important to collect the requirements that the robot has to meet. Our LEGO drawing robot was designed with the following goals in mind:

(1) to be as simple as possible for quick construction,

(2) to be precisely steerable and highly accurate,

(3) to be able to be equipped with many pairs of suitable gears to draw a variety of curves,

(4) the pole distance should be varied so that all three types of epitrochoid can be drawn,

(5) and the parametric equations of the curve drawn by the robot should be given immediately, knowing the parameters of the current configuration, i.e. the diameter of the mounted gears and the position of the pole.

3.4. Using a dynamic geometry software

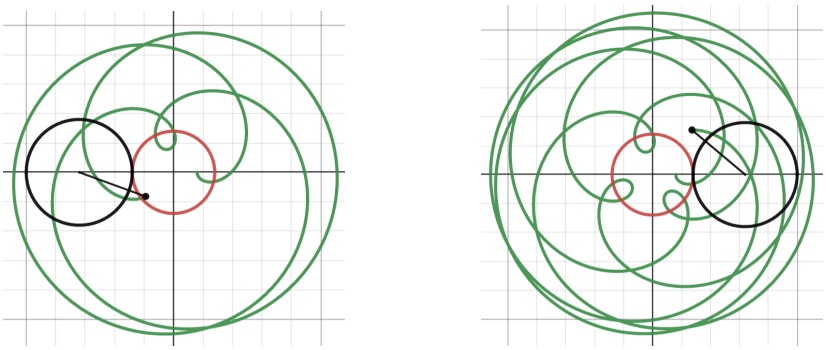
The other pillar of the method is a virtual app for the verification step. There are many applications for drawing and animating epitrochoids on the web, but integrating them into our method has, in general, posed some problems, so we decided to create a new application that can be used for robot-drawn curves during the identification step.

The teaching of parametric curves, particularly roulettes, can be made highly visual by incorporating moving animations that follow the formation of curves from point to point, showing how they are derived. We chose the Desmos graphing calculator among the numerous available options to support our work. Desmos is an easy-to-use free online application because it runs in a browser. Compatible with a wide range of devices, it is also available for smartphones and tablets, so students can use their preferred personal devices when completing assignments.

To plot a parametric curve using Desmos, we need to give the parametric equations $(x(t), y(t))$ and specify an interval for the variable t. By default, the domain of t is the interval $[0, 1]$, which can be modified according to the given curve. If the formula contains additional parametric constants, their values can be changed with a slider so that the effect of the change on the plotted curve is immediately visible. The appearance of the graph can also be modified by changing the colours, adding labels or customising the coordinate system.

In order to quickly identify the curves drawn by the robot, we created an application that takes advantage of the dynamic options offered by Desmos. This little program called Epitrochoid Tracker animates the process of how the curve is formed. It draws not only the trajectory of the pole but, according to the current parameter value, the position of the moving circle and the line segment connecting its centre with the pole are also demonstrated. In Figure 6, the Epitrochoid Tracker displays precisely the same curve that was previously drawn by the LEGO robot (see the middle image in Figure 5a).

We propose two different ways of using dynamic geometry software in the classroom. One is the post-check function. By this, we mean that, given the dimensions of the gears mounted on the robot as well as the pole distance, we can plot with Desmos the curve that we have drawn previously with the robot. This is very easy to do with the Epitrochoid Tracker; after entering the values of the parametric constants and an interval for the variable t, the actual curve is plotted on the screen point by point. The question arises: How do we specify the interval for t to plot the full curve? It depends on the $\frac{r}{R}$ ratio. If $\frac{r}{R} = \frac{p}{q}$, where p and q are relative prime numbers, then the moving circle needs p complete turns to return the pole to its initial position. Thus, a closed curve is obtained if the domain of t is set to the interval $[0, p \cdot 2\pi]$. In case of the epitrochoid shown in Figure 6, $\frac{r}{R} = \frac{9}{7}$, so its



(a) At $t = 5\pi$. **(b)** At $t = 10\pi$.

Figure 6. Two phases of the process when Epitrochoid Tracker plots a curve with parametric constants $R = 14$, $r = 18$, $d = 24$.

parametric equations are

$$x(t) = \frac{16}{5}\cos t - \frac{12}{5}\cos\left(\frac{16}{9}t\right),$$
$$y(t) = \frac{16}{5}\sin t - \frac{12}{5}\sin\left(\frac{16}{9}t\right),$$

where $t \in [0, 18\pi]$.

Another application of Desmos is for pre-planning. This means that the Epitrochoid Tracker can visualise in advance the curves that the robot can draw, given the available gears and pole distances. With this simulation, the robot can be set to draw the curves considered most exciting or important. The Tracker also has a feature that allows one or more parametric constants to be run with a given step size. With this function, students can explore additional interesting and spectacular epitrochoids.

4. Advantages and disadvantages

Robotic mathematical models are an effective tool to promote learning in mathematics education. The success and impact of STEAM learning, skills and competencies may not easily be captured with existing assessment and evaluation tools [4]. Educational theorists believe that robotic activities have tremendous potential to improve classroom teaching [6, 9]. However, studies in which robots are integrated into the curriculum in university practice as a pedagogical tool and examine its effect on twenty-first-century skills are limited. In the spirit of [6], we have gathered our experiences on the advantages and disadvantages of STEAM-based education for drawing robots related to developing 21$^{\text{st}}$-century skills. Our observations are summarised in Table 1.

Table 1. Advantages and disadvantages of the work with drawing robot models related to developing 21st-century skills.

Skills	Advantages	Disadvantages
Knowledge construction	Subject-specific skills development can be achieved by integrating educational robotics.	Integration of educational robotics is tool-dependent.
Interdisciplinarity	Mathematics is combined with engineering and computer science.	A level of knowledge should be reached where the student is able to make connections between the knowledge constructs of different disciplines.
Critical thinking and problem-solving	Visual thinking and modelling enrich the mathematical and technical imagination.	Preparing and executing robotics projects can be time-consuming, impacting the scheduling of teachers.
Using new technologies	The LEGO SPIKE Prime kit is one of the latest educational robot sets. It can be programmed in Scratch or Python.	Due to rapid technological advancements, robotics tools can quickly become outdated.
Innovation and creativity	Educational drawing robots support students' creativity by offering the possibility to develop new and innovative solutions through their versatile re-building.	No structured guidance is available for innovative projects, as they are based on students' initiative.
Cooperation and flexibility	Participation in robotics projects allow students to work in teams, share ideas, and collaborate.	Students can only work with robots in a classroom environment.

5. Conclusions and future work

Methods of teaching mathematics in higher education must align with the possibilities offered by new didactic tools. Examples of such new tools are educational robots. In our paper, we presented a complex teaching-learning technique that can be used with first-year engineering and IT students when discussing one of the notable families of parametric curves, the epitrochoids. Using available technologies

and tools, we have effectively combined educational robotics and dynamic geometry software to make STEAM-based learning of this highly visual and exciting subject. Applying the developed method allows students to improve their cognitive and creative activities, enhance their performance, and confirm their mathematical competencies. The main challenge was to create a STEAM-based method for epitrochoids that could be embedded in the classical university education framework while using non-traditional teaching aids to achieve the subject objectives. The relatively large number of curves that can be drawn with the LEGO robot allows groups or pairs of students to be assigned different problems, which is a significant advantage when defining project-based tasks.

When using a STEAM-based method, the proportions of the components are not constant. Also, when using our method, it is evident that for engineering students, for example, the construction of a robot is discussed in more detail, i.e. its technological and engineering aspects are more thoroughly covered than for computer science students. One of the crucial features of STEAM education is that the digital world is added to the analogue object creation. In our method, the same epitrochoid curves are created in two ways: the LEGO robot and a DGS. So, students will experience the joy of creating with LEGO elements, learn various useful building tricks, gain theoretical knowledge of epitrochoid curves, and gain up-to-date digital knowledge through a DGS. We have shown that the Epitrochoid Tracker can be used easily to check the parametric equations of the drawn curve and is also helpful in designing new curves.

In addition to providing an aesthetic experience, epitrochoid curves are also helpful in practical life: they can be used, for example, in surveillance or spatial coverage applications, as periodic motion primitives for human dancers or for performing complex choreographic patterns in small autonomous vehicles [20]. Real-life applications allow us to formulate exciting and meaningful projects related to epitrochoids for first-year engineering and IT students. In our further work, we will collect and systematise concrete practical applications for student projects where the developed method can be used.

References

[1] M. Abdulwahed, B. Jaworski, A. R. Crawford: *Innovative approaches to teaching mathematics in higher education: a review and critique.* Nordic Studies in Mathematics Education 17.2 (2012), pp. 49–68.

[2] A. Bray, B. Tangney: *Technology usage in mathematics education research – A systematic review of recent trends*, Computers & Education 114 (2017), pp. 255–273, issn: 0360-1315, doi: 10.1016/j.compedu.2017.07.004.

[3] R. W. Bybee: *What is STEAM Education?*, Science 329.(5995) (2010), p. 995, doi: 10.1126/science.1194998.

[4] C. Carter, H. Barnett, K. Burns, N. Cohen, E. Durall, D. Lordick, F. Nack, A. Newman, S. Ussher: *Defining STEAM Approaches for Higher Education*, European Journal of STEM Education 6.1 (2021), doi: 10.20897/ejsteme/11354.

[5] M. Charalambides, R. Panaoura, E. Tsolaki, S. Pericleous: *First Year Engineering Students'; Difficulties with Math Courses- What Is the Starting Point for Academic Teachers?*, Education Sciences 13.8 (2023), doi: 10.3390/educsci13080835.

[6] T. Coşkun, O. Filiz: *The impact of twenty-first century skills on university students' robotic achievements*, Education and Information Technologies 28 (2023), pp. 16255–16283, doi: 10.1007/s10639-023-11850-1.

[7] P. Drijvers, C. Kieran, M.-A. Mariotti, J. Ainley, M. Andresen, Y. C. Chan, T. Dana-Picard, G. Gueudet, I. Kidron, A. Leung, M. Meagher: *Integrating Technology into Mathematics Education: Theoretical Perspectives*, in: Mathematics Education and Technology-Rethinking the Terrain: The 17th ICMI Study, ed. by C. Hoyles, J.-B. Lagrange, Boston, MA: Springer US, 2010, pp. 89–132, isbn: 978-1-4419-0146-0, doi: 10.1007/978-1-4419-0146-0_7.

[8] G. Gáspár, I. Raisz, L. Huszthy: *Műszaki matematika, I. kötet*, Tankönyvkiadó, 1968.

[9] E. B. B. Gyebi, M. Hanheide, G. Cielniak: *The Effectiveness of Integrating Educational Robotic Activities into Higher Education Computer Science Curricula: A Case Study in a Developing Country*, in: Educational Robotics in the Makers Era, Advances in Intelligent Systems and Computing, vol 560, ed. by D. Alimisis, M. Moro, E. Menegatti, Cham: Springer International Publishing, 2017, pp. 73–87, doi: 10.1007/978-3-319-55553-9_6.

[10] C. J. Herrera Castrillo: *Aprendizaje de ecuaciones diferenciales aplicadas en física utilizando tecnología*, Revista Torreón Universitario 11.31 (2022), pp. 26–35, doi: 10.5377/rtu.v11i31.14223.

[11] A. Körei, S. Szilágyi: *Displaying parametric curves with virtual and physical tools*, The Teaching of Mathematics XXV.2 (2022), pp. 61–73, doi: 10.57016/TM-EHGC7743.

[12] A. Körei, S. Szilágyi: *How to Draw Cardioids with LEGO Robots: A Technical-Mathematical Project in Higher Education*, in: Robotics in Education, Lecture Notes in Networks and Systems, vol. 747, ed. by R. Balogh, D. Obdržálek, E. Christoforou, Cham: Springer Nature Switzerland, 2023, pp. 37–48, doi: 10.1007/978-3-031-38454-7_4.

[13] A. Körei, S. Szilágyi: *Parametric Graph Project - Using LEGO Gears for Drawing Curves*, in: Advanced Research in Technologies, Information, Innovation and Sustainability, ed. by T. Guarda, F. Portela, M. F. Augusto, Springer Nature Switzerland, 2022, pp. 101–114, doi: 10.1007/978-3-031-20319-0_8.

[14] A. Körei, S. Szilágyi, I. Vaičiulyte: *Task design for teaching cardioid curve with dynamic geometry software and educational robotics in university practice*, Problems of Education in the 21st Century 81 (2023), pp. 840–860, doi: 10.33225/pec/23.81.840.

[15] J. D. Lawrence: *A Catalog of Special Plane Curves*, Dover Publications, 1972.

[16] M. McCarthy: *Experiential Learning Theory: From Theory To Practice*, Journal of Business & Economics Research 14.3 (2016), pp. 91–100, doi: 10.19030/jber.v14i3.9749.

[17] M. Meletiou-Mavrotheris, E. Paparistodemou, L. Dick, A. Leavy, E. Stylianou: *Editorial: New and emerging technologies for STEAM teaching and learning*, Front. Educ. 7 (2022), pp. 1–3, doi: 10.3389/feduc.2022.971287.

[18] H. Serin: *The Integration of Technological Devices in Mathematics Education: A Literature Review*, International Journal of Social Sciences and Educational Studies 10.3 (2023), pp. 54–59, doi: 10.23918/ijsses.v10i3p54.

[19] A. Shell-Gellasch: *The Spirograph and Mathematical Models from 19th-Century Germany*, Math Horizons 22.4 (2015), pp. 22–25, doi: 10.4169/mathhorizons.22.4.22.

[20] P. Tsiotras, L. Castro: *The Artistic Geometry of Consensus Protocols*, in: Controls and Art, Inquiries at the Intersection of the Subjective and the Objective, Springer Nature Switzerland, 2014, pp. 129–153, doi: 10.1007/978-3-319-03904-6_6.

[21] O. Viberg, Å. Grönlund, A. Andersson: *Integrating digital technology in mathematics education: a Swedish case study*, Interactive Learning Environments 31.1 (2023), pp. 232–243, doi: 10.1080/10494820.2020.1770801.

[22] R. Whitaker: *Mathematics of the Spirograph*, School Science and Mathematics 88 (2010), pp. 554–564, doi: 10.1111/j.1949-8594.1988.tb11854.x.

[23] E. Yalcin İncik, T. Incik: *Generation Z Students' Views on Technology in Education: What They Want What They Get*, Malaysian Online Journal of Educational Technology 10.2 (2022), pp. 109–124, doi: 10.52380/mojet.2022.10.2.275.

[24] R. Ziatdinov, J. R. Valles: *Synthesis of Modeling, Visualization, and Programming in GeoGebra as an Effective Approach for Teaching and Learning STEM Topics*, Mathematics 10.3 (2022), doi: 10.3390/math10030398.

Annales Mathematicae et Informaticae
60 (2024) pp. 218–227
DOI: 10.33039/ami.2024.05.002
URL: https://ami.uni-eszterhazy.hu

Retrieval practice – a tool to be able to retain higher mathematics even 3 months after the exam*

Anna Muzsnay[ae], **Csaba Szabó**[be], **Janka Szeibert**[cd]

[a]University of Debrecen
muzsnay.anna@science.unideb.hu

[b]Eötvös Loránd University, Faculty of Science
szabo.csaba.mathdid@ttk.elte.hu

[c]Eötvös Loránd University, Faculty of Primary and Pre-School Education
szeibert.janka@gmail.com

[d]Edutus University

[e]MTA-ELTE Theory of Learning Mathematics Research Group

Abstract. It is a common phenomenon that students forget the learned material within a few days after their exam. A considerable part of university students do not gain long-term knowledge. Aiming to reduce forgetting and increase further retention in a first-year mathematics course for mathematics pre-service teachers, we applied a special kind of retrieval practice in their lessons. The positive effects of retrieval practice – the strategic use of retrieval to enhance memory – have been shown in the medium term in learning university mathematics. In this paper, we investigate the potential benefit of the applied retrieval practice in learning Number Theory at the university level, focusing on knowledge lasting for 3 months. $N = 42$ first-year pre-service mathematics teacher students wrote a post-test on the material they learned in the course Number Theory three months after their exam. According to our results, those, who learned Number Theory by retrieval practice, performed significantly better than those who learned on the traditional way. Our findings suggest that retrieval practice can have a powerful, long-lasting effect on learning and solving complex mathematical problems.

*This research was supported by the Research Programme for Public Education Development of the Hungarian Academy of Sciences and by Digital Education Development Competence Center at Eötvös Loránd University, Budapest, with project number 2022-1.1.1-KK-2022-00003.

Submitted: December 8, 2023
Accepted: May 8, 2024
Published online: May 10, 2024

Keywords: retrieval practice, test-enhanced learning, active recall, university mathematics, Number Theory

1. Introduction

Many studies showed that a significant part of university students chose learning strategies that are not beneficial in creating long-term knowledge [8]. Without applying powerful learning techniques, students forget the learned material within a few days after their exam. However, gaining long-term knowledge is crucial in learning and teaching mathematics [11]. This is especially true for future mathematics teachers since they are the ones who will teach future students [18]. A possible tool to increase further retention is retrieval practice, the strategic use of retrieval to enhance memory. In this study, we aim to show that by using a simple intervention – implementing a special kind of retrieval practice – educators can support their incoming students and facilitate the creation and retention of knowledge and problem-solving skills in a given higher mathematical content.

2. Literature review

Retrieval practice – the strategic use of retrieval to enhance memory – has been proven to be an effective learning method in many cases [7, 29]. Retrieval practice, also known as testing or test-enhanced learning, can refer to any activity (such as questions during class, quizzes, flashcards, brain dump, and examination questions) that requires retrieving information from memory without the help of any external sources. In the past two decades, numerous studies have shown that by actively recalling information from memory one can get more durable knowledge than by rereading the material [28]. Information learned by practicing recalling is retained significantly better than information learned by non-retrieval-based strategies, such as copying [4], re-reading [24, 28], or organizing the information in a new way [10].

Although retrieval practice is not a new concept – it was first studied more than a hundred years ago (for early research see [1]) – it has received more attention in psychological and educational research only in the last 20 years. Test-enhanced learning has been proven to be a successful learning strategy in many different areas as it contributes to consolidating the constructed mental representations in memory (e.g., [14]). The testing effect – the phenomenon that retrieving information from short- or long-term memory can strengthen one's memory of the retrieved information – has been shown in many different areas such as memorizing texts, foreign language vocabulary, general knowledge facts, learning materials that include visual or spatial information, and also in skill learning [6, 26, 28]. It was also demonstrated in laboratory circumstances and real educational environments [7, 21, 25]. Even though the research on retrieval practice indicates that it is a powerful way to promote learning, some researchers have suggested that the success of retrieval practice depends on the complexity of the to-be-learned material.

2.1. Retrieval practice and the complexity of the learning material

It is not evident if the testing effect is present when learning "complex" materials. On the one hand, retrieval works against forgetting; knowledge acquired by retrieval practice leads to a lower forgetting rate [15, 23, 31]. On the other hand, since students rarely remember all content items of the to-be-learned material perfectly (e.g., [28]) it can reduce the amount of successfully executed knowledge construction activities [27]. In their study, Roelle and Berthold examined whether the effects of incorporating retrieval into learning tasks depend on the learning tasks' complexity. They argue that the benefit of incorporating retrieval into learning tasks depends on the complexity of the tasks – the degree to which learning tasks require learners to combine content items that are included in the learning material ([27, p. 142]). They found that the net benefit of incorporating retrieval was higher for the low-complexity tasks, and as the complexity of the learning task increases the benefit decreases.

Also, Van Gog and Sweller claim that retrieval practice is not beneficial for complex materials [9]. According to them complex materials, the testing effect either diminishes or completely disappears. In their study "complex material" is referred to as "high in element interactivity, containing various information elements that are related and must therefore be processed simultaneously in working memory" ([9, p. 248]). However, the findings of van Gog and Sweller were criticized by Karpicke et al. [13] for the lack of an objective measure of complexity. Also, they listed a series of studies where the retrieval effect was present when learning complex materials such as the research of McDaniel et al. [20], Chan [5], and Butler [3].

2.2. Retrieval practice and mathematics

In recent years there has been a growing interest in investigating the testing effect in mathematics learning, in mathematical problem-solving. Since mathematical problems require developed deductive and problem-solving skills, and the problems themselves are quite complex, it is not obvious whether or not incorporating retrieval practice is a powerful way to enhance learning in this field. Developing problem-solving skills in mathematics requires not only the memorization of facts and procedures but also a deep conceptual understanding. Although more applied research is needed in this field [2], recent studies suggest that increasing retrieval practice may be an effective way of learning mathematics [12, 16, 17, 19, 30, 32]. The experiments of Yeo and Fazio [32] and Lyle et al. [16] are particularly relevant for this study.

The study of Yeo and Fazio [32] investigated the effect of retrieval practice versus (re)studying worked examples in mathematical problem-solving five minutes and one week after a one-session learning phase. They found that the optimal learning strategy depends on the learning objectives, the retention interval, and the kind of knowledge being learned (stable facts or flexible procedures) as well.

Among other things, they showed that when the goal was to learn a novel math procedure, the effectiveness of the two methods depended on the retention interval. When they tested participants' knowledge five minutes after the learning phase using nonidentical learning problems in the test, repeated studying was more effective than repeated testing. However, one week later, the group that learned by repeated testing performed as well as the group that learned by (re)studying worked examples. With identical learning problems repeated testing was more advantageous than repeated studying.

Lyle et al. [16] investigated the effect of retrieval practice in a genuine educational setting. They measured the impact of spaced versus massed retrieval and their impact on long-term retention in a precalculus course for engineering students. They found that increasing the spacing of practice – even though it significantly reduced quiz performance – resulted in better retention at the end of a precalculus course and also 1 month later.

3. Research focus and research question

In a former study, the implementation of retrieval practice in university mathematics was investigated [30]. Their results showed that in a first-year Number-Theory course, students who learned by retrieval performed better on the final test. This result strengthens the results of Lyle et al. [16] in their precalculus course in the sense that the exam included a major part of the material learned at the beginning of the semester. In the present study, we concentrate on pre-service mathematics teachers' knowledge in a given Number Theory material three moths after they took the Number Theory exam. Our research question is the following:

Can the retrieval effect be demonstrated in a longer term, a few months after finishing the course in university mathematics (Number Theory)?

4. Methods

4.1. Sample and study design

The authors conducted a quasi-experimental study to figure out whether applying a certain form of retrieval practice leads to better knowledge in the long term in a first-year mathematics course (Algebra and Number Theory 1.). Participants of the study were first-year mathematics pre-service teacher students from University Name who took the compulsory course Algebra and Number Theory 1. in their first semester. Altogether 114 students attended the course. Among the 114 students, 46 dropped out, and 68 wrote the post-test. Their ages were between 18 and 23.

The course consisted of a 60-minute lecture and a 90-minute problem session and lasted 13 weeks. The students attended the same lectures, while their practice sessions were taught in groups of 15 students on average. All together there were six practice groups. Students learned the same material and solved the same

problems during practice sessions. Three of the six groups were randomly selected as the experimental group, and the other three were the control group. We tried to eliminate the teacher's effect as much as possible: the teachers in the control and experimental groups were matched in the sense that there was one experienced teacher, one demonstrator, and one doctoral student in the control and experimental groups. Furthermore, teachers of the practice groups had a short meeting each week where they discussed the main issues related to the course. The structure of the problem-solving sessions was similar in each group.

The control group started the practice sessions by writing a short test on the material from the previous week's lecture (as it is traditional in the case of this subject). After the short test, they discussed the homework which was followed by the main part of the session: the problem-solving part with the aid of the professors.

The experimental group's sessions started by discussing the homework, then they had the problem-solving part which was followed by an end-of-class test on the material of the given practice session. So the structure of the lesson was almost identical, the difference between the two types of groups was that in the experimental group, there was no test at the beginning of the lesson, instead, they had a test at the end of the class. The end-of-class tests consisted of two open-ended problems, similar to those encountered during the practice session. We tried to ask desirably difficult questions for the students: neither too easy nor too hard for them. Students had to solve it on their own, without any help. This way, they had to retrieve what they just learned. They did not get any feedback about the solution to the problems asked in the end-of-class test, only if they explicitly asked for it.

In both groups, the tests were evaluated, and students could gain 2 points on each test. To make sure that students took these tests seriously, they needed to score 50 % by the end of the semester to pass the course.

4.2. The material covered by the course

The regular course materials for the Algebra and Number Theory lectures and problem-solving seminars were used based on the textbook by [22]. Topics covered by the course were:

- Divisibility, the greatest common divisor, the Euclidean algorithm, prime numbers, and the fundamental theorem of arithmetic.

- Special arithmetic functions, additive and multiplicative arithmetic functions. Divisor sum of multiplicative functions. The Möbius function. Perfect numbers.

- Congruences. The Euler-Fermat theorem. Linear congruences and diophantine equations. Linear congruence systems. Applications in computational number theory.

- Congruences of higher degree. Reduction to prime power, resp. prime moduli. The number of solutions, the reduction of degree in case of prime moduli. Wilson's theorem. Binomial congruences, order, primitive roots, index.

4.3. Instruments

Students' knowledge was measured at the beginning of their studies, and three months after their final exam. Before the course started, each student completed a competence-level test, which served as an input test. This test is obligatory for every pre-service mathematics teacher student and serves as a general competence and knowledge test on the high-school curriculum. Also, students who passed the course wrote a "surprise" test three months after the final exam. The test consisted of four problems and 20 minutes were given to complete it. Two problems, Problem 1 and Problem 2 involve tasks that can be solved procedurally, Problems 3 and Problem 4 are more complex. Here, we present the four problems.

Problem 1. Find the remainder of $2346235^{226688442}$ mod 23.

In this problem the only necessary knowledge is the Euler-Fermat theorem. In the solution you need to take the base mod 23 and the exponent mod $\varphi(23)$. As 23 is a prime, $\varphi(23) = 22$. At first sight, it seems slow to find the remainders. Since the test test took only 20 minutes, there was no time to use the division algorithm. But, if you look carefully at the actual numbers, pairing the digits there are numbers divisible by 23 and 22, respectively, so it can be done fast.

Problem 2. Solve the following system of congruences: $2x \equiv 8 \bmod 14$ and $x \equiv 7 \bmod 11$.

The standard solution of this problem is not very fast either. The first step is to reduce $2x \equiv 8 \bmod 14$ to $x \equiv 4 \bmod 7$, and then observe that 18 is a solution. Then, referring to the Chinese remainder theorem one can see, that 18 is the unique solution mod 77.

Problem 3. Find all solutions of the following equation over the integers:

$$3x^{16} - 4y^{48} + 17z^{2012} = 34172.$$

This problem is a challenging one. There are 4 terms, one of them is constant, and students have an arsenal of tricks for handling high-degree Diophantine equations. Here, the find a prime p and take the equation mod p trick works. The question is, which prime? This problem is a routine problem when we learn the trick, but rather tricky in this environment.

Problem 4. For which a, b digits can $\overline{a97531ba}$ be divided by 55?

This problem requires to know the divisibility rules mod 5 and 11. Since it is divisible by 5, $a = 0$ or $a = 5$. Then by taking the alternating sum of the digits b can be given. Although this is high-school knowledge, surprisingly, quite a few people got low scores on this problem.

5. Results

During the statistical analysis, we analyzed data from students who attended Algebra and Number Theory 1. and wrote the input and output tests. Altogether 79 passed the course, 68 of them took the post test. Out of the 68 students 51 wrote the pre test. The remaining 17 students can be divided into two groups. The first group took Number Theory 1. second time, they did not complete the course in their earlier studies. The second group did not write the pre-test, and this way they had to take a bridging course. We excluded the data of 4 students who had entered the university at least two years before the course. We have also excluded those 3 students who scored 0. They did not take seriously the test and handed in empty sheets. We excluded 4 excellent students who scored above 90 % on both tests and also on the midterms during the semester. Considering their test scores in the statistics would have given a false result. They scored well independently of the testing effect. In this study, we analyze the remaining 42 students' test results: 21 from the experimental group and 21 from the control group. The data analysis was conducted using R 4.2.3 software.

When analyzing the results of the input test we found that the variances of the two groups differed ($M_{control} = 69.2$, $SD_{control} = 11.0$, $M_{experimental} = 58.9$, $SD_{experimental} = 21.1$), so we applied the Welch's t-test. The test showed that there was no difference between the two groups, Welch two-way sample $t(31.5) = 1.99$, $p = 0.11$ at the beginning of the course, with effect size $d = 0.60$.

Then, we analyzed the results of the post-test. The variances were different by the F-test ($p = 0.008$), so we applied the Welch test. There was a significant difference between the two groups, $t(43.4) = 3.15$, $p < 0.001$, with effect size $d = 0.88$. The average of the experimental group was 75 % and the average of the control group was 60 % (see Figure 1).

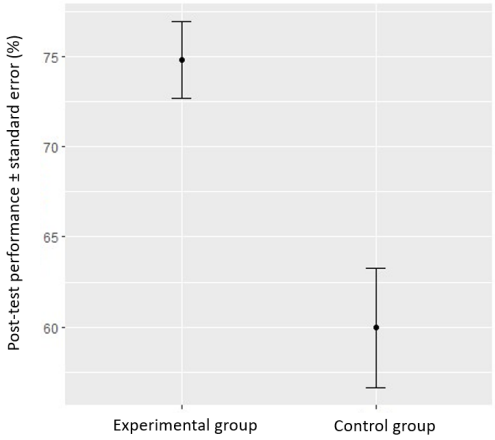


Figure 1. The results of the post-test in the experimental and the control groups.

6. Discussion

Many studies showed that a significant part of university students chose learning strategies that are not beneficial in creating long-term knowledge [8]. Without applying powerful learning techniques, students forget the learned material within a few days after their exam. However, gaining long-term knowledge is crucial in learning and teaching mathematics [11]. Retrieval practice, the strategic use of retrieval to enhance memory is recommended as an effective method for improving learning [6, 7]. It is not obvious whether the retrieval effect is present when learning "complex" materials. Furthermore, the long-term effect of retrieval practice when learning higher mathematics is still unknown.

In this case study, we investigated the effectiveness of a particular type of retrieval practice in a university mathematics course concentrating on preserving the konwledge for three months. Participants of the study were 44 first-year mathematics pre-service teacher students who took the compulsory course Algebra and Number Theory 1. Students' input level was measured by a competence-level test at the beginning of the course. The effects of retrieval practice and traditional learning were measured by a "surprise" post-test on the material they learned in the course three months after their final exam. According to our results, those, who learned Number Theory by retrieval practice, performed significantly better than those who learned on the traditional way. Our findings suggest that retrieval practice can have a powerful, fairly long-lasting effect on learning and solving complex mathematical problems.

We believe that the strength of this particular study is that we could measure the effect of a learning method in a real educational environment lasting for a whole semester. Tracking university students' knowledge can be challenging in a real school environment since reaching students after we no longer teach them is difficult and nearly impossible.

Although according to the results of this case study, this type of retrieval practice seems to be an effective way to create rather long-lasting knowledge in university mathematics, further research is needed in this area to draw far-reaching conclusions. It would be important to test the method in different school settings: with different students, in different universities, and other mathematics courses. Finally, it would be interesting to explore more deeply the effect of retrieval practice on higher mathematical knowledge in an even longer term.

References

[1] E. E. Abbott: *On the analysis of the factor of recall in the learning process*, The Psychological Review: Monograph Supplements 11.3 (1909), pp. 159–177, doi: 10.1037/h0093018.

[2] P. Agarwal, L. Nunes, J. Blunt: *Retrieval Practice Consistently Benefits Student Learning: a Systematic Review of Applied Research in Schools and Classrooms*, Educational Psychology Review 33 (Dec. 2021), pp. 1–45, doi: 10.1007/s10648-021-09595-9.

[3] A. Butler: *Repeated Testing Produces Superior Transfer of Learning Relative to Repeated Studying*, Journal of experimental psychology. Learning, memory, and cognition 36 (Sept. 2010), pp. 1118–33, doi: 10.1037/a0019902.

[4] S. Carpenter, T. Lund, C. Coffman, P. Armstrong, M. Lamm, R. Reason: *A Classroom Study on the Relationship Between Student Achievement and Retrieval-Enhanced Learning*, Educational Psychology Review 28 (May 2015), doi: 10.1007/s10648-015-9311-9.

[5] J. C. K. Chan: *Long-term effects of testing on the recall of nontested materials*, Memory 18.1 (2010), pp. 49–57, doi: 10.1080/09658210903405737.

[6] G. M. Donoghue, J. A. C. Hattie: *A Meta-Analysis of Ten Learning Techniques*, Frontiers in Education 6 (2021), issn: 2504-284X, doi: 10.3389/feduc.2021.581216.

[7] J. Dunlosky, K. Rawson, E. Marsh, M. Nathan, D. Willingham: *Improving Students' Learning With Effective Learning Techniques*, Psychological Science in the Public Interest 14 (Jan. 2013), pp. 4–58, doi: 10.1177/1529100612453266.

[8] J. Dunlosky, K. A. Rawson, E. J. Marsh, M. J. Nathan, D. T. Willingham: *What Works, What Doesn't*, Scientific American Mind 24 (2013), pp. 46–53.

[9] T. van Gog, J. Sweller: *Not New, but Nearly Forgotten: the Testing Effect Decreases or even Disappears as the Complexity of Learning Materials Increases*, Educational Psychology Review 27.2 (2015), pp. 247–264, issn: 1040726X, 1573336X, (visited on 12/04/2023).

[10] N. Goossens, G. Camp, P. Verkoeijen, H. Tabbers, S. Bouwmeester, R. Zwaan: *Distributed Practice and Retrieval Practice in Primary School Vocabulary Learning: A Multi-classroom Study: Distributed Practice and Retrieval Practice*, Applied Cognitive Psychology 30 (Oct. 2016), doi: 10.1002/acp.3245.

[11] R. F. Hopkins, K. B. Lyle, J. L. Hieb, P. A. S. Ralston: *Spaced Retrieval Practice Increases College Students' Short- and Long-Term Retention of Mathematics Knowledge*, Educational Psychology Review 28.4 (2016), pp. 853–873, (visited on 12/07/2023).

[12] R. F. Hopkins, K. B. Lyle, J. L. Hieb, P. A. S. Ralston: *Spaced Retrieval Practice Increases College Students' Short- and Long-Term Retention of Mathematics Knowledge*, Educational Psychology Review 28.4 (2016), pp. 853–873.

[13] J. D. Karpicke, W. R. Aue: *The Testing Effect Is Alive and Well with Complex Materials*, Educational Psychology Review 27 (2015), pp. 317–326, doi: 10.1007/s10648-015-9309-3.

[14] J. D. Karpicke: *2.27 - Retrieval-Based Learning: A Decade of Progress*, in: Learning and Memory: A Comprehensive Reference (Second Edition), ed. by J. H. Byrne, Second Edition, Oxford: Academic Press, 2017, pp. 487–514, isbn: 978-0-12-805291-4, doi: 10.1016/B978-0-12-809324-5.21055-9, url: https://www.sciencedirect.com/science/article/pii/B9780128093245210559.

[15] O. Kliegl, K.-H. Bäuml: *Retrieval Practice Can Insulate Items Against Intralist Interference: Evidence From the List-Length Effect, Output Interference, and Retrieval-Induced Forgetting*, Journal of Experimental Psychology: Learning, Memory, and Cognition 42 (Aug. 2015), pp. 202–214, doi: 10.1037/xlm0000172.

[16] K. Lyle, C. Bego, R. Hopkins, J. Hieb, P. Ralston: *How the Amount and Spacing of Retrieval Practice Affect the Short- and Long-Term Retention of Mathematics Knowledge*, Educational Psychology Review 32 (Mar. 2020), pp. 277–295, doi: 10.1007/s10648-019-09489-x.

[17] K. B. Lyle, N. A. Crawford: *Retrieving Essential Material at the End of Lectures Improves Performance on Statistics Exams*, Teaching of Psychology 38.2 (2011), pp. 94–97, doi: 10.1177/0098628311401587.

[18] L. Ma: *Knowing and Teaching Elementary Mathematics: Teachers' Understanding of Fundamental Mathematics in China and the United States*, Jan. 2020, isbn: 9781003009443, doi: 10.4324/9781003009443.

[19] B. M. May: *Effects of spaced, repeated retrieval practice and test-potentiated learning on mathematical knowledge and reasoning*, International Journal of Mathematical Education in Science and Technology 53.1 (2022), pp. 92–107, DOI: 10.1080/0020739X.2021.1961034.

[20] M. Mcdaniel, D. Howard, G. Einstein: *The Read-Recite-Review Study Strategy*, Psychological science 20 (Apr. 2009), pp. 516–22, DOI: 10.1111/j.1467-9280.2009.02325.x.

[21] K. McDermott, P. Agarwal, L. D'Antonio, H. Roediger, M. Mcdaniel: *Both Multiple-Choice and Short-Answer Quizzes Enhance Later Exam Performance in Middle and High School Classes*, Journal of Experimental Psychology Applied 20 (Nov. 2013), DOI: 10.1037/xap0000004.

[22] I. Niven, H. S. Zuckerman, H. L. Montgomery: *An Introduction to the Theory of Numbers, 5th edition*, Wiley, 1991, ISBN: 0471625469.

[23] M. Racsmány, A. Kereszтes: *Initial Retrieval Shields Against Retrieval-Induced Forgetting*, Frontiers in psychology 6 (May 2015), p. 657, DOI: 10.3389/fpsyg.2015.00657.

[24] H. Roediger, P. Agarwal, M. Mcdaniel, K. McDermott: *Test-Enhanced Learning in the Classroom: Long-Term Improvements From Quizzing*, Journal of experimental psychology. Applied 17 (Nov. 2011), pp. 382–95, DOI: 10.1037/a0026252.

[25] H. Roediger, A. Putnam, M. Sumeracki: *Ten Benefits of Testing and Their Applications to Educational Practice*, in: vol. 55, Jan. 2011, pp. 1–36, ISBN: 9780123876911, DOI: 10.1016/B978-0-12-387691-1.00001-6.

[26] H. L. Roediger, A. C. Butler: *The critical role of retrieval practice in long-term retention*, Trends in Cognitive Sciences 15.1 (2011), pp. 20–27, ISSN: 1364-6613, DOI: 10.1016/j.tics.2010.09.003.

[27] J. Roelle, K. Berthold: *Effects of Incorporating Retrieval Into Learning Tasks: The Complexity of the Tasks Matters*, Learning and Instruction 49 (Jan. 2017), pp. 142–156, DOI: 10.1016/j.learninstruc.2017.01.008.

[28] C. Rowland: *The Effect of Testing Versus Restudy on Retention: A Meta-Analytic Review of the Testing Effect*, Psychological bulletin 140 (Aug. 2014), DOI: 10.1037/a0037559.

[29] A. Saidat, P. Baker: *Effects of Worked Example on Students' Learning Outcomes in Complex Algebraic Problems*, International Journal of Instruction 16 (Apr. 2023), pp. 229–246, DOI: 10.29333/iji.2023.16214a.

[30] C. Szabo, C. G. Zámbó, A. Muzsnay, J. Szeibert, L. Bernáth: *Investigación de la eficacia de práctica de recuperación en matemáticas universitarias*, Revista de Educación Superior 401 (July 2023), pp. 79–96, DOI: 10.4438/1988-592X-RE-2023-401-584.

[31] K. Szpunar, K. McDermott, H. Roediger: *Testing During Study Insulates Against the Buildup of Proactive Interference*, Journal of experimental psychology. Learning, memory, and cognition 34 (Nov. 2008), pp. 1392–9, DOI: 10.1037/a0013082.

[32] D. J. Yeo, L. K. Fazio: *The Optimal Learning Strategy Depends on Learning Goals and Processes: Retrieval Practice Versus Worked Examples*, Journal of Educational Psychology 111 (2019), pp. 73–90.

Annales Mathematicae et Informaticae
60 (2024) pp. 228–236
DOI: 10.33039/ami.2024.03.001
URL: https://ami.uni-eszterhazy.hu

Spatial intelligence: Why do we measure?

Rita Nagy-Kondor

University of Debrecen, Faculty of Engineering, Hungary
rita@eng.unideb.hu

Abstract. Intelligence tests are widely used to measure intellectual performance and prophesy the academic, professional achievement in the future, or to select successful employees. There are correlations between various measures of spatial intelligence, problem solving and Science, Technology, Engineering and Mathematics (STEM) performance. Spatial visualization skills are essential for an expert to be successful in numerous disciplines. Spatial intelligence has an important role in learning and teaching of engineering studies. This report investigated the spatial visualization skills of international engineering students at the University of Debrecen – in unique way in our university, simultaneously examining four types of sub-abilities in plane and space – in comparison with the international results.

Keywords: problem solving, spatial ability, mathematics education, engineering education

AMS Subject Classification: C30, G20, G30, G40

1. Introduction

The numbers define and make visible what is real. These numbers can be produced by measurement. There are some questions about measurement [30], which are worth considering: How can we measure our skills and competences? How can we keep measures useful? Do these measures create information that increases our developmental ability? Will this information help individuals, the organization grow?

Adequate mastery of cognitive and mathematical skills essential so, that children can clearly understand the world that around them [8].

We can not develop cognitive skills, intelligence, especially spatial intelligence without measuring. Intelligence tests are widely assumed to measure max intellectual performance and associations between IQ scores and later-life outcomes are

Submitted: June 12, 2023
Accepted: March 4, 2024
Published online: March 6, 2024

typically interpreted as estimates of the effect of intellectual ability on academic and professional accomplishment [5]. Furthermore, pre-employment testing is a method to select successful employees within organizations. Cognitive tests measure for an applicant's ability to apply learned concepts to new situations during day-to-day work activities [4].

According to Gardner [6] there are seven different types of intelligence: linguistic, logical-mathematical, spatial, musical, physical-kinesthetic, interpersonal and intrapersonal intelligence. "Spatial intelligence is the ability of forming a mental model of the spatial world and maneuvering and working with this model" [6, p. 9].

According to previous studies spatial intelligence, spatial abilities are predictors of success in technical education and have a high importance in engineering education, computer graphics, architecture, arts and cartography [2, 10, 14–16, 24, 27]. Many studies have shown that there are correlations between various measures of spatial intelligence, problem solving and performance in Science, Technology, Engineering and Mathematics (STEM) [1, 11, 12, 15, 20, 29, 31]. Students with higher ability in mental rotation in regular score higher on anatomy examinations [28]. According toBennett-Pierre, Gunderson [3] fiber arts may be particularly relevant for understanding critically understudied non-rigid spatial skills. Spatial visualization skills are essential for an expert to be successful in several disciplines. Spatial intelligence has an important role in learning and teaching of engineering studies. The skill of imaginative manipulation of the object is particularly important for engineering students. So, we can define spatial ability as the complex system of cognitive component, consisting the ability to connect constructed and perceived images of 3D world [15], that is essential for success in many scientific fields [14].

Several different methods are used to test the spatial intelligence, among which Mental Rotation Test, Mental Cutting Test (MCT) and Purdue Spatial Visualization Test are widely used. During the last years researchers started developing Virtual Reality (VR) aided applications [7] to generate MCT exercises [22] with the use of Blender and its Python API [21, 23, 24]. However, in some cases we may need a more comprehensive measurement.

There are various factors effecting spatial ability; one of these is gender. Female students may not have the same spatial ability skills as male students, which can partly explain gender differences in spatial ability test; also female students choose typical mistake in some tasks more frequently, than male students [17]. Males have a higher spatial ability than females in Mental Cutting Test [16, 18]. Some articles show a significant difference on mental rotation tasks at every age [19]. There are conflicting results in the reviewed studies. Some articles found no significant difference between male and female groups in spatial intelligence of student in mathematics [25, 26]. Due to these contradictory results gender difference of spatial intelligence is still an intensively researched topic.

In the light of the existing literature, this report investigated the spatial intelligence of international engineering students in Hungarian higher education –

especially the imaginary manipulation of an object, which is essential for training.

2. Research questions and hypotheses

The goal of this article was to measure spatial intelligence – especially the imaginary manipulation of an object – freshman engineering students.

During the research, the research questions are the following:

RQ1: Is there a significant relationship between first-year international engineering students' performance in four different task types of spatial ability?

RQ2: Is there a significant difference between male and female first-year international engineering students in terms of spatial intelligence in four difficulty levels?

RQ3: What are the types of spatial intelligence tasks on which first-year international engineering students perform less well?

The hypotheses are the following:

H1: There is significant relationship between engineering students' performance in four different task types of spatial ability.

H2: There is a difference between male and female first-year international engineering students in terms of spatial intelligence, but not significant.

H3: There is lower performance of engineering students' in task of imaginary manipulation of three-dimensional object with hard mental cutting (the plane section is not triangle, quadrilateral, circle or ellipse).

3. Methodology

At the University of Debrecen, Faculty of Engineering 43 (4 female 9%, 39 male 91%) first-year international mechatronics engineering students took the tests, who came from 17 different countries (from Africa 33%, from Asia 67%). This group represents well the population of foreign students of the Faculty of Engineering.

The students came from 17 different countries, with different levels of spatial geometry preknowledge. There are few female students in the engineering training, despite this, we consider it important to examine the difference between gender. Subjects of the study are volunteered to participate and confidential feedbacks were given to those participants who are interested in. Standard instructions were given to tasks.

The instrument used in this study is a test of imaginary manipulation of an object – in four difficulty levels –, considering the literary background. This test was used to imaginary following of the phases of the objective activity. Our test includes four sections:

1. plane geometry task: imaginary manipulation of two-dimensional object (rotation, translation, reflection);

2. spatial geometry tasks:

a) imaginary manipulation of three-dimensional object with simple mental cutting task (the plane is parallel to base of the object);

b) imaginary manipulation of three-dimensional object with hard mental cutting task (the plane is not parallel to base of the object);

c) imaginary manipulation of three-dimensional object with surface development task.

This is a descriptive-analytic study. Data were analysed using the SPSS statistical analysis program in order to data analysis.

4. Results

The frequency of performance of test of imaginary manipulation of an object is presented in Figure 1. Few students scored between 0–20% (1 student) and between 81–100% (4 students) in total test. More students scored between 21–40% (9 students) and between 61–80% (12 students). Most students scored between 41–60% (14 students).

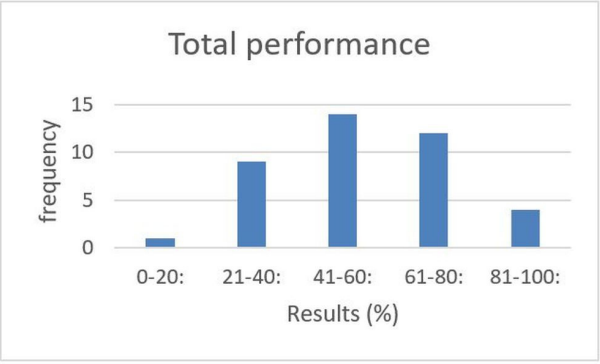


Figure 1. Frequency of total performance.

Data analysis in Table 1 showed that there was not a significant correlation between task 1 scores and task 2a scores of first-year students ($p = 0.532, r = 0.100$). There was not a significant correlation between task 1 and task 2b ($p = 0.105, r = 0.257$).

Data analysis in Table 2 showed that there was not a significant correlation between task 1 scores and task 2c scores ($p = 0.510, r = 0.106$). Similarly, there was not a significant correlation between task 2a and task 2c ($p = 0.054, r = 0.303$), and between task 2b and task 2c ($p = 0.242, r = 0.187$).

Data analysis in Table 3 showed that there was a significant correlation between task 2a scores and task 2b scores of first-year students ($p = 0.003, r = 0.448$). H1 was only partially fulfilled: So, there is significant relationship between international engineering students' performance in different task types of spatial ability,

Table 1. Relationship between task 1 scores and task 2a scores.

		task 1	task 2a
task 1	Pearson Correlation	1	,100
	Sig. (2-tailed)		,532
	N	41	41
task 2a	Pearson Correlation	,100	1
	Sig. (2-tailed)	,532	
	N	41	41

Table 2. Relationship between task 1 scores and task 2c scores.

		task 1	task 2c
task 1	Pearson Correlation	1	,106
	Sig. (2-tailed)		,510
	N	41	41
task 2c	Pearson Correlation	,106	1
	Sig. (2-tailed)	,510	
	N	41	41

Table 3. Relationship between task 2a scores and task 2b scores.

		task 2a	task 2b
task 2a	Pearson Correlation	1	,448**
	Sig. (2-tailed)		,003
	N	41	41
task 2b	Pearson Correlation	,448**	1
	Sig. (2-tailed)	,003	
	N	41	41

**. Correlation is significant at the 0.01 level (2-tailed).

but just between imaginary manipulation of three-dimensional object with mental cutting task if the plane is parallel to base of the object and imaginary manipulation of three-dimensional object with mental cutting task if the plane is not parallel to base of the object. That is why it is worth examining the components that are important to us separately.

The performance of test of imaginary manipulation of an object – in four difficulty levels – is presented in Figure 2, answers given by male and female students are compared:

task 1 (81% male students and 75% female with difference of 6%), task 2a (49% male and 38% female with the difference of 11%), task 2b (46% male and 35%

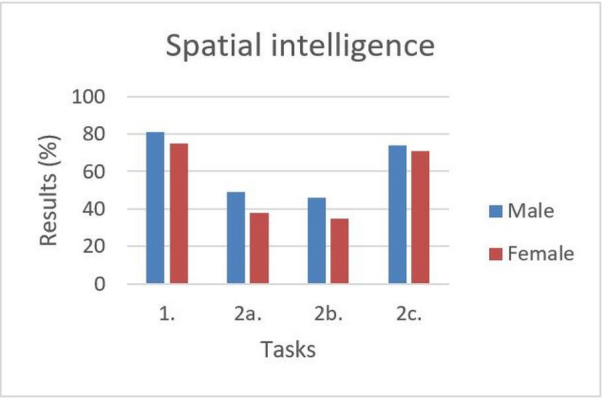


Figure 2. Results of test of imaginary manipulation of an object – in four difficulty levels.

female with the difference of 11%), task 2c (74% male and 71% female with the difference of 3%). There is a difference between male and female students in terms of spatial intelligence in all task types.

H2 was fulfilled: Independent T-test results indicated that there was no significant difference between gender and scores of imaginary manipulation of an object in all tasks (task 1: $T = 0.285, p = 0.777$; task 2a: $T = 0.420, p = 0.677$; task 2b: $T = 0.864, p = 0.393$; task 2c: $T = -0.040, p = 0.968$).

Female and male students achieved the lowest score in task 2b, this is imaginary manipulation of three-dimensional object with mental cutting – if the plane is not parallel to base of the object –, so we examine this task separately. Frequency of performance task 2b – imaginary manipulation of a three-dimensional object with mental cutting if the plane is not parallel to base of the object – is presented in Figure 3.

Few students scored between 61–80% (4 students) and between 81–100% (1 student) in total test. More students scored between 0–20% (12 students) and between 21–40% (10 students). Most students scored between 41–60% (14 students).

Results of engineering students in MCT are comparable with an average around 60% in Australia, in the US, in Europe [1] and in Hungary [14].

H3 was fulfilled, but performance of engineering students' are lower not only in task 2b, but 2a as well. Spatial skills of first-year international mechatronics engineering students are lower, we can see it in task 2a and 2b. This can be explained by the fact that Jansen et al.[9] examined cultural differences in the performance of spatial skills in mental rotation and the performance of some Asian countries such as Thailand and the Philippines was lower than that of participants with Western cultures. Spatial ability of freshman engineering students in Africa at Polytechnic of Namibia is significant lower [1]. According to Ault, John [1] the cause of these differences are the factors of previous experience and educational background.

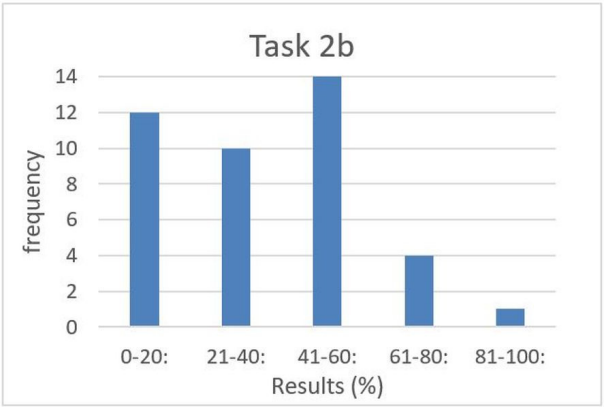


Figure 3. Frequency of task 2b.

5. Conclusion

Why Do We Measure spatial intelligence? Performance in engineering studies is related to spatial intelligence, furthermore spatial skills shows close relationship with STEM skills, so it is worth paying attention to examining several components of spatial intelligence at the same time.

The spatial intelligence – imaginary manipulation of an object – of first-year international engineering students have been studied in this paper, in four difficulty levels. The students came from different countries, with different levels of previous knowledge of spatial geometry.

Based on our survey, we can conclude that many engineering students had problems with the imaginary manipulation of spatial objects, and the mental image is incorrect in many cases. Students found the imaginary manipulation of three-dimensional object with mental cutting tasks more difficult than imaginary manipulation of two-dimensional object and imaginary manipulation of three-dimensional object with surface development. Therefore, the students achieved better results in the plane geometry task and the spatial geometry task which requires less mental manipulation (with surface development), than in imaginary manipulation of an object with mental cutting spatial geometry tasks.

We have observed gender differences in spatial abilities between male and female in our research. If teachers devote more time to improve a students' spatial ability, it help to reach better skills, regardless of students' gender.

Spatial intelligence is deemed o high priority for STEM education, so this finding is essential. It is therefore necessary to develop spatial skills at the university level as well, so that students do not have problems with this deficiency in their university studies. This is not only necessary in engineering education, but in all majors that require this ability.

We can achieve better results in understanding of spatial relationships with the

use of Dynamic Geometry Systems, interactive animations and traditional paper models [15]. Studies showed [13] that use of VR and navigate in space are positively affects spatial skills. Our results can help in choosing the appropriate educational aids.

References

[1] H. K. Ault, S. John: *Assessing and Enhancing Visualization Skills of Engineering Students in Africa: A Comprehensive Study*, Engineering Design Graphics Journal 74.2 (2010), pp. 12–20.

[2] L. Baranova, I. Katrenicova: *Role of Descriptive geometry course in development of students' spatial visualization skills*, Annales Mathematicae et Informaticae 49 (2018), pp. 21–32, doi: 10.33039/ami.2018.04.001.

[3] G. Bennett-Pierre, E. Gunderson: *Fiber Arts Require Spatial Skills: How a Stereotypically Feminine Practice Can Help Us Understand Spatial Skills and Improve Spatial Learning*, Sex Roles 88 (2023), pp. 1–16, doi: 10.1007/s11199-022-01340-y.

[4] M. Carrigan: *Pre-employment testing prediction of employee success and legal issues: A revisitation of Griggs V. Duke Power*, Journal of Business & Economics Research (JBER) 5.8 (2007), doi: 10.19030/jber.v5i8.2567.

[5] A. L. Duckworth, P. D. Quinn, D. R. Lynam, R. Loeber, M. Stouthamer-Loeber: *Role of test motivation in intelligence testing*, Proceedings of the National Academy of Sciences 108.19 (2011), pp. 7716–7720, doi: 10.1073/pnas.1018601108.

[6] H. Gardner: *Frames of mind: the theory of multiple intelligences*, New York: Basic Books, 1983.

[7] T. Guzsvinecz, M. Szabó, B. Halmosi, C. Sik-Lanyi: *The Virtual Reality Sound-based Spatial Orientation Project*, in: 12th IEEE International Conference on Cognitive Infocommunications – CogInfoCom. IEEE, 2021, pp. 11–16.

[8] B. R. J. Jansen, E. A. Schmitz, H. L. J. van der Maas: *Affective and motivational factors mediate the relation between math skills and use of math in everyday life*, Frontiers in Psychology 7 (2016), doi: 10.3389/fpsyg.2016.00513.

[9] P. Jansen, F. Paes, S. Hoja, S. Machado: *Mental Rotation Test Performance in Brazilian and German Adolescents: The Role of Sex, Processing Speed, and Physical Activity in Two Different Cultures*, Frontiers in Psychology 10 (2019), doi: 10.3389/fpsyg.2019.00945.

[10] R. Kiss-György: *The education and development of mathematical space concept and space representation through fine arts*, Annales Mathematicae et Informaticae 52 (2020), pp. 299–307, doi: 10.33039/ami.2020.12.001.

[11] E. N. Laricheva, A. Ilikchyan: *Exploring the Effect of Virtual Reality on Learning in General Chemistry Students with Low Visual-Spatial Skills*, Journal of Chemical Education 100.2 (2023), pp. 589–596, doi: 10.1021/acs.jchemed.2c00732.

[12] D. Lubinski: *Spatial ability and STEM: A sleeping giant for talent identification and development*, Personality and Individual Differences 49.4 (2010), pp. 344–351, doi: 10.1016/j.paid.2010.03.022.

[13] L. Mejia-Puig, T. Chandrasekera: *The virtual body in a design exercise: a conceptual framework for embodied cognition*, International Journal of Technology and Design Education (2022), doi: 10.1007/s10798-022-09793-8.

[14] R. Nagy-Kondor: *Gender Differences in Spatial Visualization Skills of Engineering Students*, Annales Mathematicae et Informaticae 46 (2016), pp. 265–276, doi: 10.1007/s10798-022-09793-8.

[15] R. Nagy-Kondor: *Spatial ability: Measurement and development*, in: Khine, M. S. (ed.): Visual-Spatial Ability in STEM Education: Transforming Research into Practice, Switzerland: Springer, 2017, pp. 35–58, doi: 10.1007/978-3-319-44385-0_3.

[16] R. Nagy-Kondor, S. Esmailnia: *Polyhedrons vs. Curved Surfaces with Mental Cutting: Impact of Spatial Ability*, Acta Polytechnica Hungarica 18.6 (2021), pp. 71–83, doi: 10.12700/APH.18.6.2021.6.4.

[17] R. Nagy-Kondor, C. Sörös: *Engineering students' Spatial Abilities in Budapest and Debrecen*, Annales Mathematicae et Informaticae 40 (2012), pp. 187–201.

[18] B. Németh, M. Hoffmann: *Gender differences in spatial visualization among engineering students*, Annales Mathematicae et Informaticae 33 (2006), pp. 169–174.

[19] S. Pietsch, P. Jansen: *Different Mental Rotation Performance in Students of Music, Sport and Education*, Learning and Individual Differences 22 (2012), pp. 159–163.

[20] D. L. Shea, D. Lubinski, C. P. Benbow: *Importance of assessing spatial ability in intellectually talented young adolescents: A 20-year longitudinal study*, Journal of Educational Psychology 93 (2001), pp. 604–614.

[21] R. Tóth: *Script-aided generation of Mental Cutting Test exercises using Blender*, Annales Mathematicae et Informaticae 54 (2021), pp. 147–161, doi: 10.33039/ami.2021.03.011.

[22] R. Tóth, M. Hoffmann, M. Zichar: *Lossless Encoding of Mental Cutting Test Scenarios for Efficient Development of Spatial Skills*, Education Sciences 13.2 (2023), doi: 10.3390/educsci13020101.

[23] R. Tóth, B. Tóth, M. Hoffmann, M. Zichar: *viskillz-blender — A Python package to generate assets of Mental Cutting Test exercises using Blender*, SoftwareX 22 (2023), doi: 10.1016/j.softx.2023.101328.

[24] R. Tóth, B. Tóth, M. Zichar, A. Fazekas, M. Hoffmann: *Educational Applications to Support the Teaching and Learning of Mental Cutting Test Exercises*, in: Cheng, LY. (eds) ICGG 2022 - Proceedings of the 20th International Conference on Geometry and Graphics. ICGG 2022. Lecture Notes on Data Engineering and Communications Technologies, vol. 146, Springer, 2023, pp. 928–938, doi: 10.1007/978-3-031-13588-0_81.

[25] M. Turgut, R. Nagy-Kondor: *Spatial Visualisation Skills of Hungarian and Turkish prospective mathematics teachers*, International Journal for Studies in Mathematics Education 6.1 (2013), pp. 168–183.

[26] M. Turgut, K. Yenilmez: *Spatial visualization abilities of preservice mathematics teachers*, Journal of Research in Education and Teaching 1.2 (2012), pp. 243–252.

[27] Z. Varga, J. Lóki, H. Czédli, C. Kézi, Á. Fekete, J. Biró: *Evaluating the Accuracy of Orthophotos and Satellite Images in the Context of Road Centerlines in Test Sites in Hungary*, Research Journal of Applied Sciences 10 (2015), pp. 568–573.

[28] M. A. Vorstenbosch, T. P. Klaassen, A. R. T. Donders, J. G. Kooloos, S. M. Bolhuis, R. F. Laan: *Learning anatomy enhances spatial ability*, Anatomical sciences education 6.4 (2013), pp. 257–262.

[29] J. Wai, D. Lubinski, C. P. Benbow: *Spatial ability for STEM domains: Aligning over 50 years of cumulative psychological knowledge solidifies its importance*, Journal of Educational Psychology 101.4 (2009), pp. 817–835, doi: 10.1037/a0016127.

[30] M. Wheatley, M. Kellner-Rogers: *What Do We Measure and Why? Questions About The Uses of Measurement*, Journal for Strategic Performance Measurement June (1999).

[31] C. B. Williams, J. Gero, Y. Lee, M. Paretti: *Exploring spatial reasoning ability and design cognition in undergraduate engineering students*, in: Proceedings of the ASME 2010 International Design Engineering Technical Conf. and Comp. and Information in Engineering Conference, 2010, pp. 1–8, doi: 10.1115/DETC2010-28925.

Annales Mathematicae et Informaticae
60 (2024) pp. 237–261
DOI: 10.33039/ami.2024.12.001
URL: https://ami.uni-eszterhazy.hu

Simulation-driven optimisation in didactic game design*

Enikő Palencsár[a], **Szilvia Szilágyi**[b]

[a]Faculty of Mechanical Engineering and Informatics,
University of Miskolc
palencsar.eniko@student.uni-miskolc.hu

[b]Department of Analysis, Institute of Mathematics,
University of Miskolc
szilvia.szilagyi@uni-miskolc.hu

Abstract. In educational settings, the utilisation of didactic games is a growing trend. Modifying existing games often proves inadequate for addressing certain complex course materials, necessitating the development of original didactic games with novel sets of rules. The design of such games is a challenging endeavour fraught with potential pitfalls. The fine-tuning of new games typically involves extensive trial and error, a process that is both time-consuming and labour-intensive. However, Monte Carlo simulations offer a time-efficient alternative for determining the optimal values of numerical game parameters. This paper illustrates this approach through the example of the YETI cooperative board game, which has the direct comparison test of infinite series at its didactic focus. First, the three levels of game design are briefly introduced as defined by the MDA framework. Drawing from this model, the tuning process of the game is presented, involving a systematic, cluster-based approach to marking the infinite sums featured in the YETI card decks, several strategies to mitigate the impact of the Alpha Player Problem in the game, and the configuration of key game parameters via simulations.

Keywords: game design, MDA model, Monte Carlo simulation, didactic game, infinite series

AMS Subject Classification: 97D40, 65C05, 40A05

*This research was supported by the ÚNKP-23-1 New National Excellence Program of the Ministry for Culture and Innovation from the Source of the National Research, Development and Innovation Fund.

Submitted: June 22, 2024
Accepted: December 12, 2024
Published online: December 13, 2024

1. Introduction

Didactic game design plays an important role in developing innovative approaches for teaching mathematics. Game-based learning is a widespread educational method that is gaining more and more prominence due to the preference of younger generations for hands-on learning experiences [11, 20, 30]. Games can engage students, help them maintain their attention, boost their motivation, and improve their skills while simultaneously promoting teamwork and communication, which are essential skills in adulthood. Researchers often regard digital games as the primary medium for educational purposes. Nonetheless, traditional tabletop games with tangible components offer considerable advantages as well, such as fostering collaboration, or allowing for smooth adjustments and overall control of the gaming experience [3, 40]. Non-digital games also have the potential to stimulate learners' creativity, memory, empathy, problem-solving skills, and self-confidence [31].

The design of educational games is guided by several key principles and shaped by the creators' decisions throughout the process. Didactic board game design can be challenging yet rewarding. At times, it entails creating entirely new games, while on other occasions, it involves adapting existing games for educational purposes. In the context of games, modding is defined as "using existent commercial games and adapting their mechanisms, narratives, rules, and components to achieve specific goals" [32]. For instance, LimStorm is a card game that follows the rules of SOLO, designed as an educational tool for practising the topic of limits [34]. The original game was modified to integrate the intended didactic content by replacing the numbers on traditional SOLO cards with limit expressions. Another example of an adaptation is GemHunters, which originated from the well-known board game Saboteur and aims to help players master the use of certain trigonometric functions and identities [7]. In the modded version of Saboteur, path cards include additional trigonometric formulas, the values of which must form monotonically increasing or decreasing sequences on each assembled path. Generally speaking, modded games can often be constructed more quickly than original games. Their design process typically involves fewer logical errors, though didactic flaws can still occur. However, adaptations are constrained by the structure and rules of their source material, offering limited flexibility compared to the invention of original games.

On the other hand, when no suitable game exists for adaptation, educators can design original games tailored to the needs and preferences of their students. This intricate process is prone to errors and miscalculations, necessitating a methodology that allows for continuous and thorough testing from the beginning. For more complex games, separate tests for different components may be required. However, the additional effort is often worthwhile, as original games offer significantly greater flexibility and wider customisation options compared to adaptations. One such didactic game is YETI, designed to familiarise university students with the concept of infinite series and the application of their direct comparison test.

This paper presents the application of Monte Carlo simulations in game design as a time-efficient method for fine-tuning numerical game parameters, using the

example of YETI as a case study. In the next section, a literature review is provided of relevant research on game design and the application of the Monte Carlo method. Subsequently, the board game YETI is briefly introduced, followed by the presentation of its tuning process, which encompasses the selection of the infinite sums featured in the YETI card decks, and the calibration of the most important game parameter values, as well as the exploration of different game variants for addressing the Alpha Player Problem.

2. Literature review

Tabletop game design presents a myriad of challenges that must be addressed. Unlike other forms of entertainment such as books, music, or movies, the consumption of games is unpredictable [14]. When a game is being developed, the specific events and outcomes that will unfold during gameplay are unknown to the designer, which complicates the design process. An important aspect of game development is user-centred design, where the game must provide an interactive experience in which the player exercises agency and autonomy [41]. For a game design attempt to be considered successful, it must harmonise the three key elements of the gaming experience: the designer's intentions, the game artifact providing said experience, and the impressions of the player who engages with the game [41].

Hunicke, Leblanc, Zubek [14] identify three levels of abstraction in their MDA model for game design. Mechanics refer to the fundamental activities and mechanisms available to the player, while dynamics describe the real-time behaviour of these mechanics during gameplay. The third level, aesthetics, encompasses the emotional responses elicited in the player during interactions with the game system. From the designer's viewpoint, mechanics define the boundaries of dynamic behaviour, which in turn shapes the aesthetic experience. On the other hand, the players first connect with the aesthetics of the game, which are expressed through observable dynamics and mechanics. Mechanics consist of manipulable game components, permissible actions, and the rules governing these actions. For instance, the mechanics of card games include shuffling and betting, from which dynamics like bluffing can originate [14]. Adjusting game mechanics has a direct impact on game dynamics. Thus, developing models that predict and describe dynamics can help avoid common design errors.

In a revision of the MDA model, Zubek [41] makes a distinction between mechanics, gameplay – the process of players interacting with game mechanics – and player experience, noting that the term aesthetics is often misunderstood to refer solely to visual appeal rather than the entire experience. Zubek [41] also points out that positioning the designer and the player at opposite ends of the MDA chain may be misleading, as the iterative design process involves both top-down and bottom-up approaches, simultaneously addressing game design from both perspectives. Top-down design begins with the desired experience and deconstructs it into various components, identifying the type of gameplay needed to generate this experience and the mechanics required to produce this gameplay. Each game

aims to achieve multiple player experience goals to varying extents, such as providing enjoyable sensory experiences (sensation), transporting players into a different world where they can engage in activities that are impossible or unlikely in real life (fantasy), inducing challenge through time pressure and competition, fostering fellowship through teamwork and information sharing, and creating dramatic tension through a compelling narrative [14].

On the other hand, bottom-up design consists of developing game mechanics, testing them with real players, and continuously evaluating the resulting player experience. Upton [37] provides six important heuristics for assessing and improving game mechanics. The initial two heuristics, choice and variety, highlight that players must perceive a diverse array of possible actions. A limited range of actions can lead to boredom due to the repetitive nature of the game, while an excess of choices can result in confusion and frustration. The third heuristic, consequence, emphasises the importance of player actions having tangible outcomes, given that players may feel a loss of agency within the game in the absence of direct consequences. The fourth heuristic is predictability, followed by uncertainty as the fifth, indicating that players must be able to foresee the outcomes of their actions to a certain extent, but these consequences must not be completely predetermined. The final heuristic, satisfaction, suggests that in a game of high quality, desirable outcomes must always be within reach. An unwinnable game can lead to frustration and disinterest, whereas a game that always delivers desirable outcomes with minimal effort can lack challenge and excitement. Hence, balancing difficulty and attainability is essential for sustaining players' interest and engagement [37].

Balancing, which aims to boost game quality through numerous iterations of fine-tuning and playtesting [17], is among the most challenging phases of game design. During this process, designers rely on adjustable game parameters to reach the intended player experience. Each unique combination of game parameter values can be regarded as a distinct vector in a multidimensional space of game variants, called the game space [16, 36]. It is important to note that this iterative tuning approach can be both costly and time-consuming. In game-based learning, the need to harmonise educational content with gameplay mechanics introduces an additional layer of complexity to the balancing process [4].

The ability to explore the game space without relying entirely on human testers can significantly accelerate the tuning process, as automated methods that facilitate the identification of the most promising configurations reduce the need for extensive playtesting. Jaffe et al. [17] raise the possibility of automating some aspects of game evaluation using AI-driven simulation tools. Similarly, Nelson [21] suggests that metrics derived directly from the game itself, rather than empirical playtests, can offer valuable insights. Hypothetical player-testing, as described by Nelson [21], is an evaluation strategy that uses simplified player models to analyse how the game behaves in idealised or extreme scenarios. Prior studies imply that this approach can effectively reduce the need for player-intensive testing [2, 18, 19]. Chaslot et al. [6] argue that game AI often requires extensive domain knowledge and long development cycles. However, Monte Carlo-based solutions can counter

these challenges by using simulated playouts instead of domain-specific heuristics.

The Monte Carlo method, named after the Monte Carlo Casino in Monaco, relies on probabilistic models to estimate the average characteristics of real-world processes [27]. This approach is particularly valuable if direct experimentation is too time-consuming, impractical, or costly [10]. John von Neumann and Stanislaw Ulam pioneered Monte Carlo simulations during the Manhattan Project to model the random diffusion of neutrons in nuclear materials [28]. The application of the method consists of repeated random sampling and statistical analysis. It approximates the value of an unknown quantity using the principles of inferential statistics, which assert that a random sample – a proper subset of a population – tends to reflect the properties of the population.

Monte Carlo simulation typically involves several steps. First, a deterministic model, closely resembling the real scenario, is generated using the most likely values of input parameters, and mathematical relationships to transform the input values into the desired outputs. Once the deterministic model is satisfactory, risk components are added by identifying the underlying distributions of the input variables, based on historical data. Subsequently, random samples are drawn from these distributions representing various sets of input values, which are used in the deterministic model to generate sets of output values. This process is repeated to collect a range of possible outputs. Finally, statistical analysis is performed on these output values to provide a basis for decision-making with statistical confidence [26]. Confidence in the estimate depends on both the size and the variance of the sample: higher variance necessitates larger samples to attain the same level of confidence. By leveraging simulations, researchers can investigate complex systems, repeat experiments, and make modifications as needed. However, there are limitations to using simulations as a modelling methodology. Instead of providing exact measurements, simulations yield statistical estimates, leading to uncertain results prone to experimental errors [28]. Additionally, simulations can be computationally intensive, and the accuracy of the results depends heavily on the quality of the model and the inputs used. Furthermore, like any software, simulation programs can have bugs [10].

In the context of game design, according to Nummenmaa, Kuittinen, Holopainen [24], simulations aim to abstract models to a degree where designers can focus on the core dynamics of a system without being overwhelmed by small details. This approach is particularly advantageous, as simulations can uncover potential issues in the long-term dynamics of a game. Beyond diagnostics, simulations can assist in the tuning of game parameter values during the balancing process. They also serve as a powerful tool for analysing existing games, contributing to the invention of optimal strategies, the discovery of new game variants, and the development of AI-driven players. Additionally, Schell [29] suggests that Monte Carlo simulations can be valuable in assessing the role of chance within a non-deterministic game.

Numerous examples highlight the versatility of Monte Carlo methods in game design. Browne, Maire [5] employed simulations to assess the quality of turn-

based, deterministic games and to discover new, high-quality game variants. Similarly, Isaksen et al. [15] applied Monte Carlo simulations to explore the game space in *Flappy Bird*, uncovering a wide range of playable configurations. A specific implementation of Monte Carlo methods, Monte Carlo Tree Search (MCTS), is a best-first search method designed to pinpoint the most promising moves in a given game scenario [39]. In addition to its applications in persona-based player modelling [12] and skill-based automated playtesting [13], MCTS has proven to be a valuable tool in research related to deterministic games, such as *Go* [9] and *Hex* [1], as well as games involving multiplayer interactions or elements of uncertainty, such as *Chinese Checkers* [33], *Magic: The Gathering* [38], and *Scotland Yard* [23]. Monte Carlo simulations are also instrumental in the design and analysis of educational games. For instance, Fitrianawati et al. [8] leveraged Monte Carlo simulations to enumerate all possible solutions for each draw in the arithmetic card game 24. This enabled the assignment of difficulty levels to different card combinations, contributing to the mitigation of students' mathematics anxiety.

3. Presentation of the board game YETI

YETI is a collaborative board game designed for application in a higher education context, with a focus on the topic of infinite series and their direct comparison test. Infinite series play a significant role in various disciplines, including finance, physics, statistics, and engineering. Determining the convergence property of an infinite sum presents a significant challenge which can be tackled using convergence tests. The n^{th} term test for divergence is a straightforward technique stating that if the terms of an infinite sum do not approach zero, the sum diverges. Cauchy's test, also known as the root test, involves taking the n^{th} root of the absolute value of the n^{th} term, and evaluating its limit as n approaches infinity. Similarly, the application of D'Alembert's test entails calculating the limit of the ratio of the series' successive terms. The direct comparison test asserts that if an infinite sum of non-negative terms has a majorant that converges, the original sum also converges. Conversely, an infinite sum having a divergent minorant of non-negative terms indicates that the original sum diverges. The application of the direct comparison test requires an intuitive initial guess, making its usage considerably less algorithmic than the application of the other tests mentioned above. Depending on whether the examined infinite sum is believed to converge or diverge, one must find a sufficiently simple convergent majorant or divergent minorant to support this belief. The development of this intuitive approach necessitates extensive practice, the process of which can be supported by playing the board game YETI.

At the outset of the development process of YETI, it was essential to identify the primary aesthetic goals of the game to enable the use of a top-down design approach. The new tabletop board game aimed, most of all, to present a challenge for learners through its main didactic content, create a collaborative environment, offer sensory experiences through tactile and visual elements, and feature a simple yet effective narrative. A decision was made early on to design a cooperative rather

than a competitive game, encouraging students with a strong grasp of the given topic to help their peers understand the course materials. Parallel to the top-down development, a bottom-up design approach was also employed, focusing on the key mechanics of the new game, with all six heuristics presented by Upton [37] thoroughly considered.

A common challenge educators face when designing didactic games is the fragility of the strategy-luck equilibrium in the game mechanics. For a game to serve as an effective learning tool, it must incorporate a substantial portion of the course materials. In addition, a degree of uncertainty is critical for maintaining the game's enjoyment factor, as Upton [37] suggests. Introducing elements of randomness, such as cards, dice, or spinners, is necessary to prevent the game from becoming a mere exercise sheet in an unusual format rather than a genuinely engaging educational experience. However, if winning is perceived to be based more on luck than on skill and comprehension, players may lose motivation to engage with the course materials, thus undermining the educational objectives of a didactic game. Therefore, although the inclusion of a card deck is pivotal in the game YETI, it must be ensured that the role of luck does not overshadow the importance of actual knowledge. Given the wide variance in potential players' skills, differentiation, achieved by developing card decks of three levels of difficulty, was key to maintaining the game's satisfaction aspect through adapting the gameplay experience to the prior knowledge of participants.

Beyond these considerations, it is essential to acknowledge that the complexity of a game strongly impacts its effectiveness as a teaching tool. Based on his extensive research on the topic, Sousa [32] states that didactic games that students are likely to play only once should have a low level of complexity. Unknown, original games necessitate debriefing and the continuous presence of facilitators to support gameplay. Given that the rules of YETI are not derived from another well-known game, additional time must be allocated for concise explanations of the game's narrative and rules before each session. Therefore, it is essential to keep the rules simple and intuitive, facilitating an immediate understanding of the game logic. Additionally, it is important to consider the time constraints and repeatability of the game. Since a university practical lesson has a duration of around 90 minutes, the gameplay must not exceed 35-40 minutes.

With these guiding principles in mind, the main components of YETI were developed, featuring an immersive narrative, an original set of rules, a versatile game mat, and three card decks of varying levels of difficulty, ranging from beginner to expert. The game encourages players to engage with the direct comparison test in a hands-on, interactive manner, promoting a deeper understanding of the subject material through repeated practice and collaborative problem-solving.

3.1. Narrative

The game's narrative revolves around a Yeti attack in a small mountain village. Players are tasked with repelling the invading beasts before they destroy the entire village. Players can gain territory by identifying valid pairs among the infinite

sums on the cards in the YETI deck. A valid pair consists of a divergent sum and its divergent minorant, or a convergent sum and its convergent majorant. Within the game's narrative, these sums act as identifiers for different Yeti hunters, who must be paired into squads of two based on their fighting styles, determined by the divergence or convergence of their identifiers. If at least two squads are present in the same neighbourhood of the village, any Yeti previously based there is successfully driven away, and no further invaders dare to enter the given area. However, if two Yetis occupy the same neighbourhood, they soon engage in a territorial fight and destroy the entire village as a consequence, an unfavourable outcome which players must prevent.

3.2. Game mat

The structure of the YETI game mat, as seen in Figure 1, is clean and simple, with designated slots for Yeti or series cards. The illustrations on the mat depict a small valley village with tiny houses and meandering paths, surrounded by snowy mountain peaks. Four card slots of the same colour constitute a field, symbolising a neighbourhood within the fictional village. The relational operators between neighbouring card slots indicate the direction of relations between the pairs of infinite sums that players must respect when placing cards onto the mat. Three distinct field colours are used: dark blue indicates convergence and light blue signifies divergence, while grey card slots serve as wild cards, allowing players to play both their convergent and divergent pairs [35]. An additional rule, reinforced by the relational operators on the game mat, restricts the placement of infinite sums with equal n^{th} terms as pairs to the grey fields.

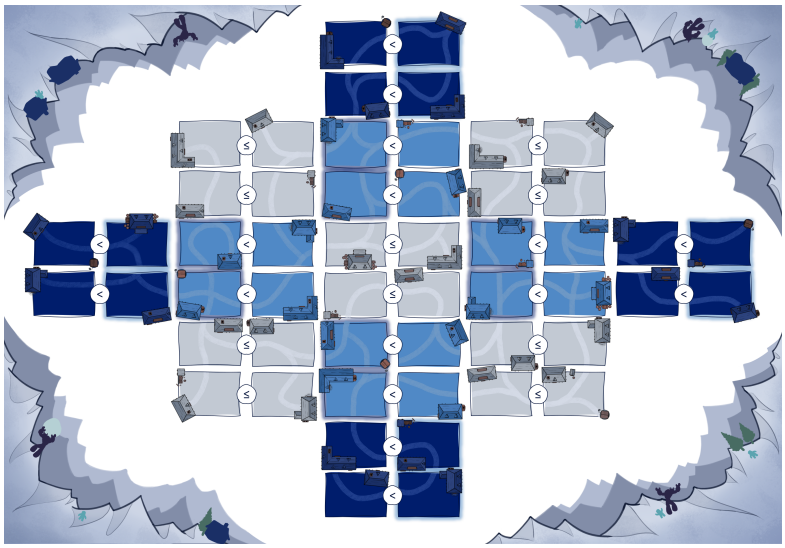


Figure 1. Game mat graphics for the board game YETI.

3.3. Set of rules

The beginning of the game YETI is marked by players strategically positioning five Yeti cards onto five different fields of their choosing on the game mat. Subsequently, the remaining deck comprising 95 cards is shuffled, and players start drawing cards. Upon encountering a Yeti card, it must be promptly laid down on any unoccupied field. If no such field is available, resulting in multiple Yeti cards being placed onto the same field, the game is lost immediately. Otherwise, drawing continues until players collectively have exactly 12 series cards in their possession, signalling the onset of the pairing phase. During the pairing phase, participants engage in discussions regarding potential card pairings, and lay down the pairs of infinite sums they identify. Successfully filling a field of four card slots results in the expulsion of any Yeti previously occupying the given field, enabling players to discard the corresponding Yeti card. The game culminates in victory if all Yeti cards are removed from the mat and more than four fields are filled, or if at least eight fields are filled. However, defeat ensues if players possess the maximum of 12 series cards but are unable to form valid pairs.

4. The tuning process of the game

The final stage of game design typically involves playtesting and tuning [14]. During the development of the board game YETI, dynamics originating from proposed sets of mechanics were continuously tested and refined using computer simulations. Through the iterative improvement of game mechanics and parameters, specific learning objectives and aesthetic goals could be reached. The tuning process consisted of two major steps: marking the exact infinite series featured in the card decks and establishing the optimal values for all numerical game parameters. In addition to these endeavours, it was imperative to address certain concerns raised by the cooperative nature of the game, such as the Alpha Player Problem.

4.1. The composition of the card decks

The development of the YETI card decks on three different levels posed a considerable challenge for multiple reasons. Primarily, the 80 infinite sums of non-negative terms featured in a deck must represent the various series types and solution methods covered in the course materials. Moreover, determining an optimal ratio of Yeti cards to series cards, achieved through extensive playtesting, is essential to align with Upton's satisfaction heuristic mentioned in Section 2. Another critical demand is guaranteeing the pairability of randomly drawn cards, increasing the likelihood of identifying pairs among the cards dealt at the beginning of each round, thus maintaining the game's dynamic flow. This requires each series card to be compatible with multiple other cards, ideally between three to five, within a deck.

To meet these criteria, a systematic methodology was employed for selecting

the series included in the card decks, leading to the establishment of clusters of infinite sums with the following properties:

1. Infinite sums within the same cluster either all converge or all diverge, allowing the entire cluster to be classified as either convergent or divergent.
2. Half of the infinite sums in each cluster are simpler majorant or minorant series (depending on the cluster's convergence property).
3. Majorants or minorants within a cluster are equal, differing only in the form their n^{th} term is written in.
4. The other half of sums in a cluster are more intricate infinite series to which the direct comparison test needs to be applied after finding a suitable majorant or minorant.
5. All of the intricate series within a cluster can be paired with any majorant or minorant in the same cluster.
6. Majorants or minorants within a cluster can also be paired with each other but only on the grey fields of the game mat due to their equality.
7. Valid pairings can also be established between infinite sums in different clusters. The clustering approach only serves as a guarantee for a minimum number of possible pairings within a deck.

Figure 2 provides an illustrative example of clusters.

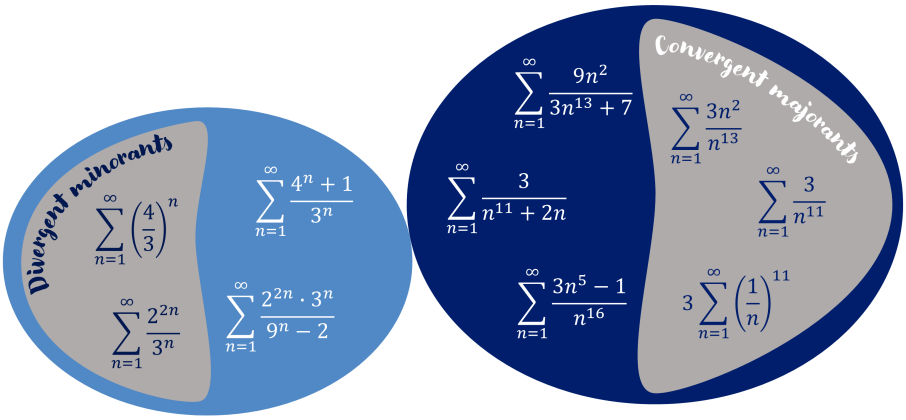


Figure 2. An illustrative example of the structure of clusters.

The deliberate construction of clusters with the specified properties ensured an adequate number of valid pairings within the game decks. Consider, for instance, a cluster of size $s \in \{2k \mid k \in \mathbb{Z}^+\}$, bearing in mind that such a cluster must always contain an even number of cards. This cluster consists of $\frac{s}{2}$ minorants/majorants and the same number of more intricate series. The intricate series are guaranteed to form valid pairs with each majorant or minorant within the cluster, while the

minorants and majorants can form valid pairs with all other members of the cluster, including other minorants and majorants, except for themselves. Thus, the total number of possible pairings within a cluster is given by:

$$\frac{s}{2} \cdot \frac{s}{2} + \frac{\left(\frac{s}{2} \cdot \left(\frac{s}{2} - 1\right)\right)}{2} = \frac{3s^2}{8} - \frac{s}{4}$$

Table 1. The number of pairing possibilities within a cluster depending on cluster size.

Cluster size	Number of pairing possibilities
2 cards	1
4 cards	5
6 cards	12
8 cards	22
10 cards	35

Table 1 presents the number of guaranteed pairing possibilities within a cluster of a given size for all five cluster sizes appearing in the YETI decks. Table 2 details the distribution of clusters by size and convergence property across the various levels of YETI. Each deck, whether beginner, intermediate, or expert, contains exactly six divergent and eight convergent clusters, resulting in a total of 14 clusters. On average, each cluster comprises approximately $80/14 \approx 5.71$ cards, leading to a mean value of pairing possibilities between 5 and 12 per cluster, and, consequently, a total of roughly $5 \cdot 14 = 70$ to $12 \cdot 14 = 168$ guaranteed pairs per deck.

Table 2. The distribution of clusters by size and convergence property in the three YETI decks. The abbreviation 'div.' denotes divergent and 'conv.' denotes convergent.

Cluster size	Number of clusters					
	Beginner		Intermediate		Expert	
	Div.	Conv.	Div.	Conv.	Div.	Conv.
2 cards	2	0	2	0	2	0
4 cards	1	1	1	3	1	2
6 cards	1	5	1	2	1	3
8 cards	2	2	2	2	2	3
10 cards	0	0	0	1	0	0
all	6	8	6	8	6	8

Following the selection of the series featured in the game, abstract models of the card decks were constructed as directed graphs, representing the relationships between the infinite sums within a deck. Each infinite series in a deck, associated

with a vertex in a directed graph, was assigned a unique identifier. The edges of the graphs were defined by the relationships between the n^{th} terms of the sums, pointing from the smaller n^{th} term to the larger, and from the smaller identifier to the larger in cases of equality. It is crucial to note that these models do not encompass all possible relationships between the infinite series showcased on the cards, since the primary objective of YETI is to help students build an intuition for recognising valid pairings that are easily identifiable without using a calculator. Thus, in the graphs, only these straightforward pairing possibilities are represented by edges.

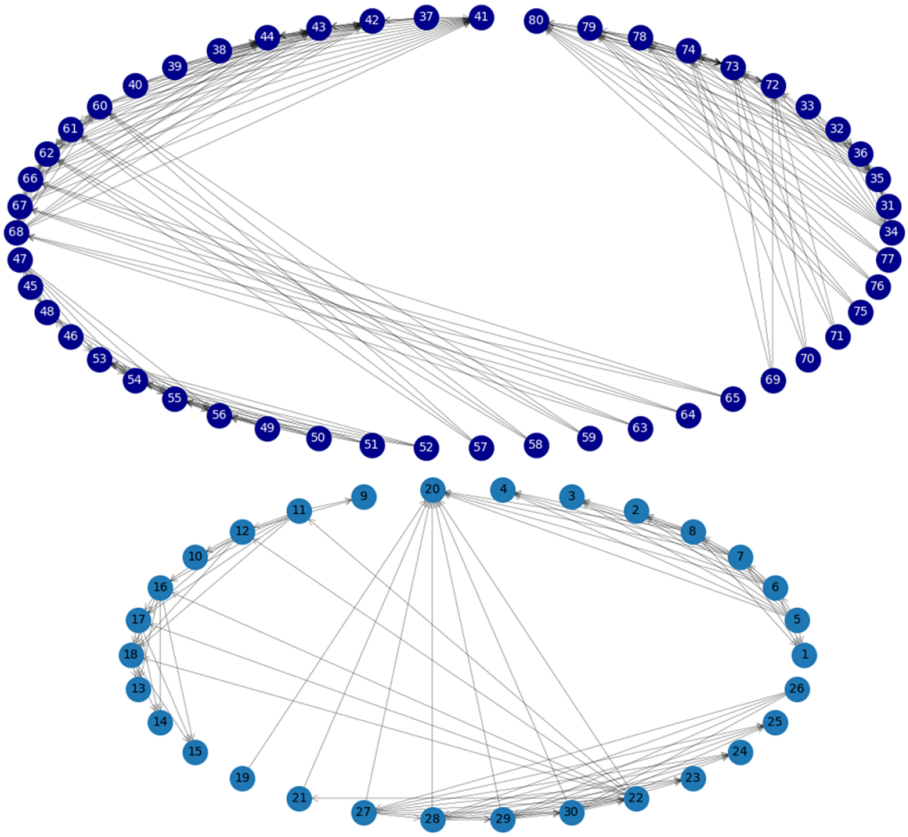


Figure 3. Directed graph $G_1 = (V_1, E_1)$ of the relations between the infinite sums in the beginner deck, limited to the relations easily noticeable to the naked eye. The vertices representing divergent sums are light blue, while the convergent ones are dark blue. $|V_1| = 50 + 30 = 80,\ |E_1| = 177 + 88 = 265.$

Initially, the abstract representations of the decks were restricted to the most easily identifiable relations. However, more advanced players might uncover addi-

tional relationships between the infinite series on the cards, including those stemming from the transitive property of inequalities. Both the graphs depicting only the most easily identifiable relations (Figure 3) and those illustrating single transitive relations (Figure 4), the recognition of which requires more practice and insight, were generated using the *NetworkX* Python package for the creation and manipulation of complex networks [22]. In the context of the subsequent game simulations, these two slightly different model types served as approximations for the knowledge of complete beginners and more experienced players.

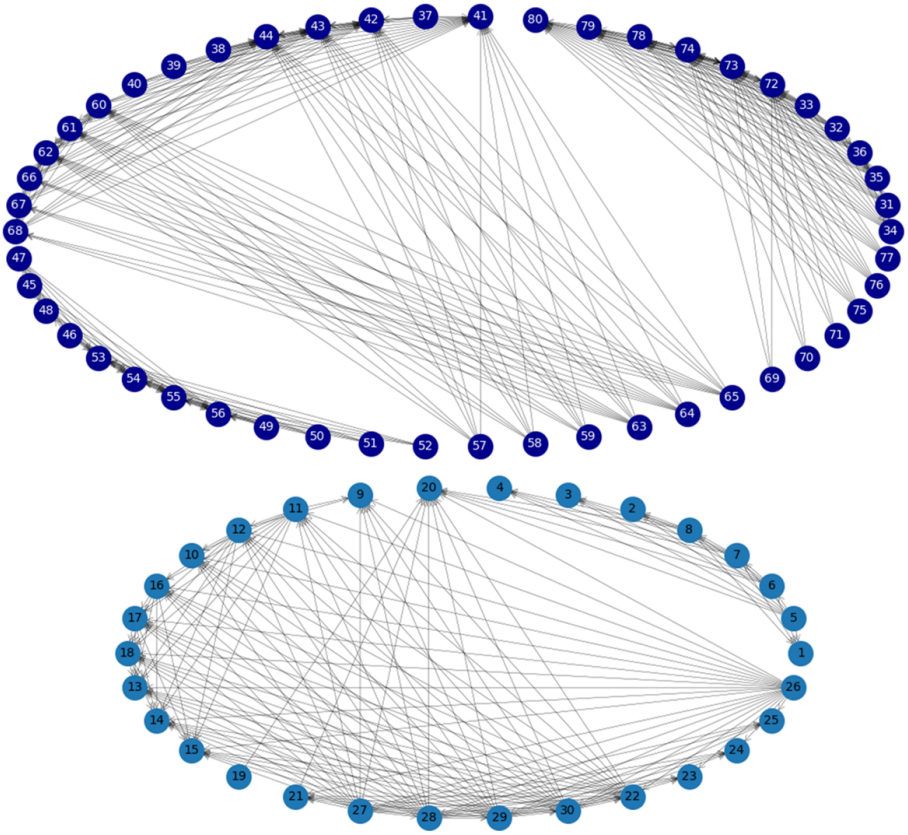


Figure 4. Directed graph $G_2 = (V_2, E_2)$ of the relations between the infinite sums in the beginner deck, including the relations formed using the transitive property of inequalities. The vertices representing divergent sums are light blue, while the convergent ones are coloured dark blue.
$|V_2| = 50 + 30 = 80,\ |E_2| = 248 + 159 = 407.$

For instance, consider the following convergent sums of non-negative terms in

the beginner deck:

$$\sum_{n=1}^{\infty} \frac{6^n}{5^{2n}+3n} \quad \sum_{n=1}^{\infty} \left(\frac{3}{4}\right)^n \quad \sum_{n=1}^{\infty} \left(\frac{6}{25}\right)^n$$

For all $n \in \mathbb{Z}^+$, the easily identifiable relations between the n^{th} terms are:

$$1) \quad \frac{6^n}{5^{2n}+3n} < \frac{6^n}{5^{2n}} = \frac{6^n}{25^n} = \left(\frac{6}{25}\right)^n$$

$$2) \quad \left(\frac{6}{25}\right)^n < \left(\frac{6}{8}\right)^n = \left(\frac{3}{4}\right)^n$$

With a bit of practice, the two inequalities above can be combined forming the following relation:

$$3) \quad \frac{6^n}{5^{2n}+3n} < \left(\frac{6}{25}\right)^n < \left(\frac{3}{4}\right)^n$$

Thus, three possible pairings exist among the given infinite series:

$$P = \left\{ \begin{array}{c} \left(\sum_{n=1}^{\infty} \frac{6^n}{5^{2n}+3n}, \sum_{n=1}^{\infty} \left(\frac{6}{25}\right)^n\right), \left(\sum_{n=1}^{\infty} \left(\frac{6}{25}\right)^n, \sum_{n=1}^{\infty} \left(\frac{3}{4}\right)^n\right), \\ \left(\sum_{n=1}^{\infty} \frac{6^n}{5^{2n}+3n}, \sum_{n=1}^{\infty} \left(\frac{3}{4}\right)^n\right) \end{array} \right\}$$

4.2. Monte Carlo simulations in the design process

During the tuning process of the game YETI, Monte Carlo simulations were employed to determine the optimal values of key numerical game parameters, such as the number of Yeti cards in a deck. Considering that a deck of 100 cards can be shuffled in 100! different ways, exhaustive playtesting was deemed unfeasible, necessitating a complex, multi-parameter optimisation process, which would have been not only impractical but also highly susceptible to error if performed manually. Consequently, computer simulations were used for experimental purposes. Despite the series cards in the YETI deck forming an intricate network of interconnections, in an ideal game scenario, where player decision-making is optimised, the outcomes depend solely on the initial shuffle of the card deck, which introduces the only element of uncertainty into the game. Therefore, computer simulations focused on card shuffling were used to conduct a statistical analysis of the game, generating thousands of randomised sample game setups.

4.2.1. Optimisation targets

The primary objective of the optimisation process was to achieve a balance between the strategic and stochastic elements of the game, while minimising the occurrence

of unfavourable scenarios such as premature victories, stalemates, and inevitable losses. Based on this principle and the didactic goals of the game, the following optimisation targets were defined:

1. *Coverage*: Players should be incentivised to seek both convergent and divergent pairs, primarily pairing the simpler infinite series, intended as majorants or minorants, with more intricate series.
2. *Seamless start*: There should always be easily identifiable pairs among the dealt infinite series at the beginning of the game. The deck shuffle should result in stalemates in no more than 1% of cases.
3. *Incentive to think and act*: Players should be motivated to play a minimum of 2 pairs of infinite sums on average per round.
4. *Balance of knowledge and luck*: As per Upton's insights detailed in Section 2, the uncertainty introduced by the card deck should be offset by ensuring that optimal decisions lead to victories. Experienced players should lose no more than 10-15% of their games when making optimal decisions, aligning with Upton's satisfaction heuristic.

4.2.2. Methodology

The law of large numbers asserts that as the number of trials in random experiments increases, the average of the results obtained from these trials will converge to the expected value of the underlying probability distribution. Essentially, with an infinite number of trials, empirical results will match theoretical probabilities. Guided by this principle, a large number of simulations were carried out for different potential values of the key YETI game parameters, where the relative frequency of certain designated adverse cases was recorded and analysed in order to identify an optimal setup. These simulations, implemented in Python and interfacing with a database of the infinite sums featured in the card decks, ranged from simple card shuffling to comprehensive game simulations, with algorithms replacing the mid-game decision-making of human players.

In the game simulations, two distinct decision-making models were employed: simplified and optimal moves. In each round, either the first identified pair was immediately placed on the game mat (simplified move), or the pairing process was optimised to maximise the number of cards played (optimal move). In the latter case, all potential pairs among the dealt series cards were aggregated into a set P. Next, it was determined whether the elements in each subset of P had any overlap. The optimal move for the given round was selected as the largest subset containing independent pairs of infinite series, meaning all pairs in the subset could be placed onto the game mat simultaneously. While the optimal move simulations naturally yielded more realistic outcomes, they significantly increased the program's runtime due to the exhaustive process of generating and evaluating these subsets.

For instance, suppose that the following pairs can be played in a round:

$$P = \left\{ \begin{array}{c} \left(\sum_{n=1}^{\infty}\left(\frac{6}{25}\right)^n, \sum_{n=1}^{\infty}\left(\frac{3}{4}\right)^n\right), \left(\sum_{n=1}^{\infty}\frac{6^n}{5^{2n}+3n}, \sum_{n=1}^{\infty}\left(\frac{6}{25}\right)^n\right), \\ \left(\sum_{n=1}^{\infty}\frac{3^n-1}{4^n+1}, \sum_{n=1}^{\infty}\left(\frac{3}{4}\right)^n\right), \left(\sum_{n=1}^{\infty}\frac{6^n}{5^{2n}+3n}, \sum_{n=1}^{\infty}\left(\frac{3}{4}\right)^n\right) \end{array} \right\}$$

In a simplified move, M_1 represents the set of pairs to be placed. The remaining potential pairs cannot be utilised since each infinite sum appears only once per deck.

$$M_1 = \left\{ \left(\sum_{n=1}^{\infty}\left(\frac{6}{25}\right)^n, \sum_{n=1}^{\infty}\left(\frac{3}{4}\right)^n\right) \right\}$$

In contrast, the optimal move simulation, by generating all subsets of P, can produce a more favourable, larger set of pairs to be placed, denoted as M_2:

$$M_2 = \left\{ \left(\sum_{n=1}^{\infty}\frac{6^n}{5^{2n}+3n}, \sum_{n=1}^{\infty}\left(\frac{6}{25}\right)^n\right), \left(\sum_{n=1}^{\infty}\frac{3^n-1}{4^n+1}, \sum_{n=1}^{\infty}\left(\frac{3}{4}\right)^n\right) \right\}$$

4.2.3. Avalanche cards

The rules of YETI create an imbalance between the fields on the game mat, with the grey card slots becoming overpowered compared to the blue ones. While blue card slots can only accommodate either convergent or divergent pairs, grey fields can house both types, effectively doubling the number of playable pairs. Additionally, pairs with identical n^{th} terms can only be placed onto grey fields. Consequently, players are incentivised to prioritise filling the grey card slots first, where even the simplest of pairs can be laid down, leading to a potential imbalance in the game's didactic content.

To address this issue, avalanche cards were introduced. An avalanche card can clear the contents of a field of a specific colour, providing a means to discard Yeti cards effectively. There are three sorts of avalanche cards, corresponding to the three primary field types: convergent, divergent, and mixed avalanches. These are distinguished by the colour of the avalanche card's borders and are only applicable to a field of the matching colour. Moreover, avalanche cards must be played during the round in which they are drawn. Therefore, to use avalanches to their advantage, players must distribute their Yeti cards across fields of different colours early in the game, as the timing and the appearance order of avalanche cards are unpredictable. All things considered, the introduction of avalanches encourages players to keep each distinct field type in the game, seeking both divergent and convergent pairs of infinite sums, as well as both simpler and more intricate pairings throughout the game, meeting the requirements of the *coverage* optimisation target.

4.2.4. The number of cards drawn per round

The number of cards dealt per round, denoted by k, is one of the most significant game parameters. In setting this parameter, the second optimisation target, *seamless start*, was considered, ensuring that players could begin their games in at least 99% of cases after drawing the first k series cards. To prevent a stalemate, it is evident that there must be at least one valid pair among the dealt cards, suggesting that a higher value of k might be reasonable. However, to avoid overwhelming players, the count of infinite sums they need to manage each round must be minimised. Consequently, the value of k must be set to the smallest possible value that meets the *seamless start* target.

To assess all potential values of k, a straightforward deck shuffling simulation was implemented. The `shuffle()` method from the random module in Python was invoked on a list comprising all the infinite sums featured in a deck. This list served as an approximation for the shuffled deck of cards. Through numerous shuffling simulations for different values of k, it was examined how frequently the initial shuffle resulted in unfavourable starting scenarios, defined as instances where a stalemate ensued at the game's onset. The shuffling experiment was conducted $N = 100{,}000$ times for each examined value of k, the results of which are detailed in Table 3, where f denotes the ratio of experiments resulting in a stalemate.

Table 3. Results of the simulations for different values of k out of $N = 100{,}000$ experiments per parameter value, where f denotes the percentage of simulations with unfavourable outcomes.

k	5	6	7	8	9	10	11	12	13
f (%)	41.85	27.41	16.34	9.14	4.85	2.32	1.07	0.47	0.18

As illustrated by Table 3, a lower value of k correlates with a higher incidence of adverse cases. For instance, when only five cards were dealt per round, the relative frequency of stalemates at the beginning of the game exceeded 40%. The lowest possible value of k where the percentage of stalemates was deemed acceptable is 12, with the ratio of adverse cases falling below 1%. Thus, the value of the parameter k has been finalised: in each round of the game YETI, players must draw enough new cards to ensure that the total number of series cards in front of them amounts to 12.

4.2.5. The number of Yeti cards per deck

To meet the third optimisation target, it was essential for YETI to encourage participants to identify at least two valid pairs of infinite series per round. The Yeti cards featured in the game deck serve as the primary incentive for this, as an excess of Yetis on the game mat can quickly lead to defeat. Evidently, a lack of Yeti cards would render the gameplay monotonous, depriving it of excitement. In contrast, an excess of Yeti cards would make victory nearly unattainable, even for

the most experienced players. Therefore, to promote an engaging and motivating player experience, the appearance of one new Yeti per round was deemed ideal, inspiring the placement of two pairs of infinite sums in each round to avoid an increase in the number of Yeti cards on the game mat. Nevertheless, any scenarios requiring players to play more than three Yeti cards in quick succession were to be prevented. Hence, fine-tuning the number of Yeti cards per deck, denoted by y, became a critical step in the optimisation process.

In the shuffling simulation, the distribution of Yeti cards in a deck was modelled by adding $y-5$ Yeti elements, with the five initially placed Yeti cards substracted, to the list representing the dealer deck. The YETI game mat accommodates a total of 13 fields. To motivate players to make their pairing decisions as quickly and efficiently as possible, it was necessary to maintain the possibility of two Yetis being forced onto the same field, which constitutes a key condition for defeat. Consequently, the minimum value considered for y had to be $13+1=14$. The main optimisation goal regarding the number of Yeti cards was to minimise the frequency of scenarios where numerous consecutive Yetis appear in a deck. The upper limit for an acceptable cluster size of Yeti cards was fixed at three; any cluster exceeding this size was considered an adverse case. Thus, the probability of encountering at least one group of four or more consecutive Yeti cards in the shuffled deck was estimated as a function of the discrete parameter y.

The shuffling simulation was run $N=100.000$ times for each examined value of y. The results are presented in Table 4, where f denotes the percentage of experiments where clusters of four or more Yeti cards were present in the shuffled deck. The table indicates that the values of y for which the ratio of unfavourable shuffle outcomes did not exceed or barely exceeded 1% fell between 14 and 16. However, further experiments were necessary to adjust the value of y, ensuring that approximately one Yeti would appear in each round.

Table 4. Results of the simulations for different values of y out of $N=100{,}000$ experiments per parameter value, where f denotes the percentage of simulations with unfavourable outcomes.

y	14	15	16	17	18	19	20
f (%)	0.44	0.66	1.00	1.43	1.96	2.62	3.43

In addition to modelling card shuffling, a comprehensive simulation of YETI was also implemented to draw conclusions about the balance between uncertainty and predictability in the game. Two simulation variations were distinguished based on the simulated players' experience level. For modelling beginner gameplay, only the most easily identifiable relations, illustrated by Figure 3, were taken into account when pairing infinite sums. In contrast, for approximating the decisions of experienced players, pairs generated through transitivity, as seen in Figure 4, were also included. Experiments were conducted using both the simplified and the optimal move types detailed in Subsection 4.2.2.

The first task to be addressed using comprehensive game simulations was setting

the number of Yeti cards in a deck. The range of possible values had previously been narrowed down to the set $\{14, 15, 16\}$ based on the deck shuffling simulations. It had also been stated that, ideally, approximately one new yeti would emerge in each round. The fluctuation of the average number of Yetis appearing per round was examined as a function of y, the count of Yeti cards in a deck. The data from simulations encompassing $N = 10{,}000$ game iterations are detailed in Table 5. The results indicate that the average value of one Yeti per round is most closely approached with $y = 16$, marking 16 as the optimal number of Yeti cards in a deck.

Table 5. The average number of Yeti cards emerging in a round as a function of y, out of $N = 10{,}000$ simulations per parameter value.

Player experience	Move type	y		
		14	15	16
Beginner	simplified	0.69	0.77	0.86
	optimal	0.72	0.81	0.90
Experienced	simplified	0.69	0.78	0.85
	optimal	0.72	0.81	0.89

4.2.6. Comprehensive game simulations

Following the establishment of key game parameter values, additional simulations were executed to examine various aspects of YETI's game dynamics. These experiments aimed to determine the total number of rounds played and pairings made in a game, the percentage of games resulting in a win, and the proportion of losses attributable to either an inability to make a move or an excess of Yeti cards on the game mat. Arithmetic means were calculated based on $N = 10{,}000$ comprehensive game simulations, summarised in Table 6.

Table 6. Sample averages for $N = 10{,}000$ comprehensive game simulations without the *reinforcement* option.

Player experience	Move type	Number of rounds	Number of pairs	Victory (%)	Loss	
					Stalemate (%)	Multiple yetis per field (%)
Beginner	simpl.	5.53	9.75	42.06	57.18	0.76
	opt.	5.46	9.94	44.77	54.55	0.68
Experienced	simpl.	5.17	12.03	71.38	27.95	0.67
	opt.	4.99	12.19	73.68	25.53	0.79

The data collected indicate that the most common cause of defeat for both novice and experienced players in the game YETI is a stalemate, an inability to

make a move. Only a negligible percentage of losses results from multiple Yetis occupying a single field. On average, the game concludes after 5-6 rounds, with experienced players playing approximately two more pairs of cards than beginners. Regrettably, the ratio of games won turns out to be notably low, falling below 50% for novices and below 75% for experienced players, with a high number of losses attributable to stalemates. This suggests that even players making optimal decisions might lose a significant portion of their games, implying an imbalance between strategy and luck in the established rules and mechanics of YETI, which may negatively impact player motivation.

Table 7. Sample averages for $N = 10{,}000$ comprehensive game simulations integrating the *reinforcement* option.

Player experience	Move type	Number of rounds	Number of pairs	Victory (%)	Loss	
					Stalemate (%)	Multiple yetis per field (%)
Beginner	simpl.	7.31	11.88	69.23	29.71	1.06
	opt.	7.12	11.96	70.83	27.99	1.18
Experienced	simpl.	6.02	13.15	89.89	9.32	0.79
	opt.	5.69	13.28	91.24	7.74	1.02

To mitigate this problem, a new optional action named reinforcement was added to the game mechanics. Once per game, this rule allows players to replace three of their series cards with new cards from the dealer deck. Should a Yeti card be dealt during this process, it can be shuffled back into the deck, and another card can be drawn in its place. The simulation results for $N = 10{,}000$ games integrating the reinforcement option are detailed in Table 7. In the simulation of the reinforcement option, the series cards selected for discarding are chosen randomly without any optimisation. Still, its integration results in a considerable increase in both the number of rounds and the number of pairs played. More significantly, a shift in the victory rate reflects a substantial improvement: leveraging the reinforcement option, beginners, when making optimal decisions, can win approximately 70% of their games, while experienced players can benefit from a victory rate of around 90%.

It is evident that the YETI game variant with the parameter values $k = 12$ and $y = 16$, along with the addition of avalanche cards and the reinforcement option, fully meets the optimisation targets outlined in Subsection 4.2.1. However, it is important to emphasise that the parameter optimisation process was focused solely on game mechanics. To validate the effectiveness and assess the educational content of the game, further experiments with human participants, who are subject to errors and learning curves, are necessary. Additionally, potential issues arising from the collaborative nature of the game warrant further investigation.

4.3. Cooperativity variants of the game

Drawing from their research on mathematical board games, Nurnberger-Haag, Wernet, Benjamin [25] propose that an effective didactic board game should encompass competitiveness and asynchronicity, integrate elements of luck, strategy, and mathematical knowledge, and be suitable for both introducing and revising a specific topic. However, their study exclusively examines competitive games, and the authors themselves acknowledge that their recommendation is not definitive regarding the question of competition versus cooperation. Despite this limitation, YETI, with an added feature supporting asynchronous play, aligns with the enumerated recommendations. The proposed modification can be implemented following the drawing phase, when players cooperatively identify pairs of infinite series from the pool of dealt cards. During this phase, each player is encouraged to choose a series card and formulate a hypothesis regarding the convergence property of the infinite sum showcased on it. If the others agree with the hypothesis, the selected card is set aside into a group labelled as "convergent" or "divergent". In contrast, disagreements result in the chosen card remaining in the center, indicating that its convergence property remains undetermined. This individual guessing exercise not only introduces an asynchronous element into the game, but also simplifies the pairing process by narrowing down the pool of possible combinations, as only infinite sums with matching convergence properties can be paired up in the game.

The collaborative aspect of the game YETI prompts several key considerations. A common issue in cooperative games is referred to as the Alpha Player Problem, where a dominant player dictates the group's actions and strategies, reducing the input and control of others, thus diminishing their engagement and enjoyment. YETI addresses this by allowing players to take turns hypothesising the convergence of the dealt series at the beginning of each round, ensuring active participation from all. Additionally, all pairing decisions must be explained, providing an opportunity for less experienced players to learn from their more skilled counterparts. Since YETI is recommended for supervised play, educators also have the opportunity to address the Alpha Player Problem directly by either reassigning dominant players to more suitable groups or creating new groups equipped with more challenging decks.

The game's structure is also adaptable to mitigate the Alpha Player Problem. In collaborative game design, the most widely implemented solutions to prevent a single player's dominance include rotating leadership roles, realising exceptionally quick gameplay, introducing hidden information, or incorporating a traitor mechanic. The latter two can be applied to the game YETI as well. The traitor can be narratively presented as a spy from a rival village, aiming to mislead players into making at least five pairing mistakes. If successful, the traitor wins the game, and the others lose. Naturally, the impact of the traitor role requires thorough evaluation regarding its effect on players, its impact on the learning process, and the potential confusion it may cause for less experienced players. Alternatively, hidden information can be introduced by dealing each player two hidden series cards at the beginning of the game. In this case, players have the option to exchange one card

with another player each round, gradually becoming acquainted with all hidden cards. If a player obtains pairable infinite sums as hidden cards, they can place them on the game mat at the beginning of the round. This modification prevents the dominance of alpha players, maintaining a balanced and engaging gameplay experience for each participant.

5. Conclusion

This paper presented the intricate process of fine-tuning the didactic board game YETI, which aims to familiarise students with the direct comparison test of infinite series. Key aspects addressed by the tuning process include building a systematic approach to marking the infinite sums featured in the card decks, tackling concerns related to cooperativity, and optimising the key numerical parameters of the game. In this endeavour, important targets such as balancing strategy and luck, achieving seamless starts, and encouraging player engagement were reached. Monte Carlo simulations were employed to resolve potential gameplay issues such as early victories, stalemates, and guaranteed losses. Additionally, new gameplay features like the reinforcement option and avalanche cards were introduced as part of the optimisation process to maintain a challenging yet enjoyable player experience.

In board game research, Monte Carlo simulations are a well-established tool, primarily used for developing AI-driven game agents and simulating opponents via Monte Carlo Tree Search (MCTS) [1, 9, 23, 33, 38]. While there have been attempts to leverage these simulations for fine-tuning mechanics or exploring game configurations, such uses have largely been limited to computer games [12, 13, 15, 16]. It is also important to note that, so far, the Monte Carlo method has predominantly been used as a technique for analysing adversarial, deterministic, turn-based games [1, 5, 9]. Using YETI as a case study, our research illustrates how Monte Carlo simulations can contribute to optimising game mechanics and achieving balanced gameplay in the context of a non-deterministic, non-digital, cooperative didactic board game.

The vast number of possible scenarios in game design makes manual testing impractical and time-consuming, covering only a small area of the game space. This is where Monte Carlo simulations prove valuable, bridging the gap between the invention of mechanics and the assessment of the resulting dynamics. The utilisation of simulations in the case of YETI enabled a thorough examination of various game parameters, which significantly expedited the design process by reducing the reliance on time-consuming manual testing. This efficiency allowed for a comprehensive analysis of potential gameplay scenarios within a shorter time frame, ensuring that the design process was concluded promptly and effectively. Nevertheless, it is essential to acknowledge the limitations of such simulations. While they are highly effective for setting game parameter values, they are not a substitute for real-world testing. The behaviour of virtual players in simulations does not accurately reflect the playing style, learning curve, and strategic diversity of human players. Therefore, further testing involving human participants is

necessary to validate the effectiveness and educational value of the game.

As evidenced by our research, tuning is an indispensable phase of didactic board game design, where the finer details of the game are established to achieve the intended player experience. This process can be approached in various ways, but it consistently demands rigorous, critical thinking, a problem-solving mindset, and a systematic approach. Monte Carlo simulations are a powerful tool in didactic game design, particularly for games where the structure can be accurately simulated. The successful application of these simulations in the development of the board game YETI highlights their potential to facilitate the game design process and provide well-balanced, engaging gameplay.

References

[1] B. Arneson, R. B. Hayward, P. Henderson: *Monte Carlo Tree Search in Hex*, IEEE Transactions on Computational Intelligence and AI in Games 2.4 (2010), pp. 251–258, doi: 10.1109/TCIAIG.2010.2067212.

[2] S. Bakkes, C. T. Tan, Y. Pisan: *Personalised gaming: a motivation and overview of literature*, in: Proceedings of The 8th Australasian Conference on Interactive Entertainment: Playing the System, IE '12, Auckland, New Zealand: Association for Computing Machinery, 2012, doi: 10.1145/2336727.2336731.

[3] C. Balakrishna: *The Impact of In-Classroom Non-Digital Game-Based Learning Activities on Students Transitioning to Higher Education*, Education Sciences 13.4, 328 (2023), doi: 10.3390/educsci13040328.

[4] C. Benedikte Søgaard Hansen, T. Bjørner: *Designing an Educational Game: Design Principles from a Holistic Perspective*, The International Journal of Learning: Annual Review 17.10 (2011), pp. 279–290, doi: 10.18848/1447-9494/cgp/v17i10/47275.

[5] C. Browne, F. Maire: *Evolutionary Game Design*, IEEE Transactions on Computational Intelligence and AI in Games 2.1 (2010), pp. 1–16, doi: 10.1109/TCIAIG.2010.2041928.

[6] G. Chaslot, S. Bakkes, I. Szita, P. Spronck: *Monte-Carlo Tree Search: A New Framework for Game AI*, Proceedings of the AAAI Conference on Artificial Intelligence and Interactive Digital Entertainment 4.1 (2021), pp. 216–217, doi: 10.1609/aiide.v4i1.18700.

[7] M. Dudás, S. Lengyelné Szilágyi, I. Piller: *Card Deck Designer Software for the Mathematical Game Called Ékkővadászok*, Gradus 6.4 (2019), pp. 17–27.

[8] M. Fitrianawati, Z. Aliansyah, N. R. N. Peni, I. W. Farid, L. Hakim: *Monte Carlo method at the 24 game and its application for mathematics education*, Journal of Honai Math 5.2 (2022), pp. 83–94, doi: 10.30862/jhm.v5i2.250.

[9] S. Gelly, L. Kocsis, M. Schoenauer, M. Sebag, D. Silver, C. Szepesvári, O. Teytaud: *The grand challenge of computer Go: Monte Carlo tree search and extensions*, Commun. ACM 55.3 (2012), pp. 106–113, doi: 10.1145/2093548.2093574.

[10] R. L. Harrison: *Introduction to Monte Carlo Simulation*, AIP Conference Proceedings 1204.1 (2010), pp. 17–21, doi: 10.1063/1.3295638.

[11] M. Hernandez-de-Menendez, C. A. Escobar Díaz, R. Morales-Menendez: *Educational experiences with Generation Z*, International Journal on Interactive Design and Manufacturing (IJIDeM) 14.3 (2020), pp. 847–859, doi: 10.1007/s12008-020-00674-9.

[12] C. Holmgård, A. Liapis, J. Togelius, G. Yannakakis: *Monte-Carlo Tree Search for Persona Based Player Modeling*, Proceedings of the AAAI Conference on Artificial Intelligence and Interactive Digital Entertainment 11.5 (2021), pp. 8–14, doi: 10.1609/aiide.v11i5.12849.

[13] B. Horn, J. A. Miller, G. Smith, S. Cooper: *A Monte Carlo approach to skill-based automated playtesting*, Proceedings of the 14th AAAI Conference on Artificial Intelligence and Interactive Digital Entertainment (2018).

[14] R. Hunicke, M. Leblanc, R. Zubek: *MDA: A Formal Approach to Game Design and Game Research*, AAAI Workshop - Technical Report 1 (2004).

[15] A. Isaksen, D. Gopstein, J. Togelius, A. Nealen: *Discovering Unique Game Variants*, in: Computational Creativity and Games Workshop at the 2015 International Conference on Computational Creativity, 2015.

[16] A. Isaksen, D. Gopstein, J. Togelius, A. Nealen: *Exploring Game Space of Minimal Action Games via Parameter Tuning and Survival Analysis*, IEEE Transactions on Games 10.2 (2018), pp. 182–194, doi: 10.1109/TCIAIG.2017.2750181.

[17] A. Jaffe, A. Miller, E. Andersen, Y.-E. Liu, A. Karlin, Z. Popovic: *Evaluating Competitive Game Balance with Restricted Play*, Proceedings of the AAAI Conference on Artificial Intelligence and Interactive Digital Entertainment 8.1 (2021), pp. 26–31, doi: 10.1609/aiide.v8i1.12513.

[18] R. Khoshkangini, G. Valetto, A. Marconi, M. Pistore: *Automatic generation and recommendation of personalized challenges for gamification*, User Modeling and User-Adapted Interaction 31.1 (2020), pp. 1–34, doi: 10.1007/s11257-019-09255-2.

[19] A. Liapis, C. Holmgård, G. N. Yannakakis, J. Togelius: *Procedural Personas as Critics for Dungeon Generation*, in: Applications of Evolutionary Computation, ed. by A. M. Mora, G. Squillero, Cham: Springer International Publishing, 2015, pp. 331–343.

[20] K. Moore, C. Jones, R. S. Frazier: *Engineering Education For Generation Z*, American Journal of Engineering Education (AJEE) 8.2 (2017), pp. 111–126, doi: 10.19030/ajee.v8i2.10067.

[21] M. Nelson: *Game Metrics Without Players: Strategies for Understanding Game Artifacts*, Proceedings of the AAAI Conference on Artificial Intelligence and Interactive Digital Entertainment 7.3 (2011), pp. 19–24, doi: 10.1609/aiide.v7i3.12479.

[22] *NetworkX-NetworkX documentation*, https://networkx.org/ [Accessed: 13 Jun 2024].

[23] P. Nijssen, M. H. M. Winands: *Monte Carlo Tree Search for the Hide-and-Seek Game Scotland Yard*, IEEE Transactions on Computational Intelligence and AI in Games 4.4 (2012), pp. 282–294, doi: 10.1109/TCIAIG.2012.2210424.

[24] T. Nummenmaa, J. Kuittinen, J. Holopainen: *Simulation as a game design tool*, in: Proceedings of the International Conference on Advances in Computer Entertainment Technology, ACE '09, Athens, Greece: Association for Computing Machinery, 2009, pp. 232–239, doi: 10.1145/1690388.1690427.

[25] J. Nurnberger-Haag, J. L. Wernet, J. I. Benjamin: *Gameplay in Perspective: Applications of a Conceptual Framework to Analyze Features of Mathematics Classroom Games in Consideration of Students' Experiences*, International Journal of Education in Mathematics, Science and Technology 11.1 (2022), pp. 267–303, doi: 10.46328/ijemst.2328.

[26] S. Raychaudhuri: *Introduction to Monte Carlo simulation*, in: Proceedings of the 2008 Winter Simulation Conference, Miami, FL, USA: IEEE, 2008, pp. 91–100, doi: 10.1109/WSC.2008.4736059.

[27] J. Ross, A. Marshak: *Monte Carlo Methods*, in: Photon-Vegetation Interactions: Applications in Optical Remote Sensing and Plant Ecology, ed. by R. B. Myneni, J. Ross, Berlin, Heidelberg: Springer, 1991, pp. 441–467, doi: 10.1007/978-3-642-75389-3_14.

[28] R. Y. Rubinstein, D. P. Kroese: *Simulation of Discrete-Event Systems*, in: Simulation and the Monte Carlo Method, Hoboken, NJ, USA: John Wiley & Sons, Inc., 2016, chap. 3, pp. 91–106, doi: 10.1002/9781118631980.ch3.

[29] J. Schell: *The Art of Game Design: A Book of Lenses*, 3rd, Boca Raton, FL: CRC Press, 2019, doi: 10.1201/b22101.

[30] C. Seemiller, M. Grace: *Generation Z goes to college*, San Francisco, CA: Jossey-Bass, 2016.

[31] C. Sousa, S. Rye, M. Sousa, P. J. Torres, C. Perim, S. A. Mansuklal, F. Ennami: *Playing at the School Table: Systematic Literature Review of Board, Tabletop, and other analog game-based learning approaches*, Frontiers in Psychology 14, 1160591 (2023), doi: 10.3389/fpsyg.2023.1160591.

[32] M. Sousa: *Mastering Modern Board Game Design to build new learning experiences: The MBGTOTEACH framework*, International Journal of Games and Social Impact 1.1 (2013), pp. 68–93, doi: 10.24140/ijgsi.v1.n1.04.

[33] N. R. Sturtevant: *An Analysis of UCT in Multi-player Games*, in: Computers and Games, ed. by H. J. van den Herik, X. Xu, Z. Ma, M. H. M. Winands, Berlin, Heidelberg: Springer Berlin Heidelberg, 2008, pp. 37–49.

[34] S. Szilágyi, A. Körei: *Using a Math Card Game in Several Ways for Teaching the Concept of Limit*, in: Mobility for Smart Cities and Regional Development - Challenges for Higher Education, ed. by M. E. Auer, H. Hortsch, O. Michler, T. Köhler, Cham: Springer International Publishing, 2022, pp. 865–877, doi: 10.1007/978-3-030-93904-5_85.

[35] S. Szilágyi, E. Palencsár: *Board Games in Mathematics Education: Presentation of the PDCA-based Graphic Design Process of the YETI Didactic Framework*, Gradus 10.2 (2023), doi: 10.47833/2023.2.csc.002.

[36] J. Togelius, G. N. Yannakakis, K. O. Stanley, C. Browne: *Search-Based Procedural Content Generation*, in: Applications of Evolutionary Computation, ed. by C. Di Chio, S. Cagnoni, C. Cotta, M. Ebner, A. Ekárt, A. I. Esparcia-Alcazar, C.-K. Goh, J. J. Merelo, F. Neri, M. Preuss, J. Togelius, G. N. Yannakakis, Berlin, Heidelberg: Springer Berlin Heidelberg, 2010, pp. 141–150.

[37] B. Upton: *The Aesthetic of Play*, Cambridge, MA: MIT Press, 2015, doi: 10.7551/mitpress/9251.001.0001.

[38] C. D. Ward, P. I. Cowling: *Monte Carlo search applied to card selection in Magic: The Gathering*, in: 2009 IEEE Symposium on Computational Intelligence and Games, 2009, pp. 9–16, doi: 10.1109/CIG.2009.5286501.

[39] M. H. M. Winands: *Monte-Carlo Tree Search in Board Games*, in: Handbook of Digital Games and Entertainment Technologies, ed. by R. Nakatsu, M. Rauterberg, P. Ciancarini, Singapore: Springer Singapore, 2017, pp. 47–76, doi: 10.1007/978-981-4560-50-4_27.

[40] T. Zhang, J. Liu, Y. Shi: *Enhancing collaboration in tabletop board game*, in: Proceedings of the 10th Asia Pacific Conference on Computer Human Interaction, Matsue-city, Shimane, Japan: Association for Computing Machinery, 2012, pp. 7–10, doi: 10.1145/2350046.2350050.

[41] R. Zubek: *Elements of Game Design*, Cambridge, MA: MIT Press, 2020.

Annales Mathematicae et Informaticae
60 (2024) pp. 262–277
DOI: 10.33039/ami.2024.03.003
URL: https://ami.uni-eszterhazy.hu

On a method for measuring the effectiveness of mathematics teaching using delayed testing in technical contexts in engineering education

Dóra Sipos[a], **Imre Kocsis**[b]

[a]University of Debrecen, Doctoral School of Mathematical and Computational Sciences
dorasipos@eng.unideb.hu

[b]University of Debrecen, Faculty of Engineering, Department of Basic Technical Studies
kocsisi@eng.unideb.hu

Abstract. In this research, engineering topics with different levels of modeling are included in engineering mathematics courses and the effect of the method is tested through delayed tests in a technical context independent of the mathematics classroom. To address the difficulty of problem solving through modeling, three types (levels) of mathematical problems motivated by engineering tasks are defined. A library of problems was collected and the problems were systematically integrated into the classroom work. The long-term effect of the targeted application of professional problems was investigated by means of delayed tests in the context of statics (mechanics), which the students studied in the following semester. In our approach, the key efficiency factor is the extent to which students can apply the mathematical concepts and methods they have learned while studying professional subjects and later in their engineering work.

Keywords: teaching efficiency, delayed test, engineering mathematics, technical learning environment

1. Introduction

Since our teaching practice focuses on key engineering competencies in general, we consider engineering mathematics as a professional subject rather than a separate course from other modules in the curriculum, and our goal is to create synergy be-

Submitted: August 3, 2023
Accepted: March 13, 2024
Published online: March 28, 2024

tween mathematical and professional subjects. The desired synergy can be achieved by linking topics when teaching new mathematical and technical materials, engaging in joint projects. Furthermore, the mathematical knowledge is required to be developed and evaluated in as many professional courses as possible during the training.

In this paper, we address the specific concept of teaching efficiency and the role of mathematics in engineering education. We also examine the relationship between the teaching method used in engineering mathematics courses and the ability to recall mathematical concepts and methods later in engineering courses.

Based on our several decades of experience, interviews, and daily communication with engineering students, two of the most important aspects to consider when discussing the efficiency of a methodology development in engineering mathematics are

- the expectations of engineering students (What motivates them?) and
- the desired role of Engineering Mathematics courses in modern engineering training.

The expectations of engineering students – After studying the attitudes of students in our engineering programs, we concluded that expectations (motivating factors) are changing rapidly. The current generations of engineering students increasingly prefer to learn things that are immediately applicable rather than focusing on studying for the future. Students need to feel that the material is useful to them, and this fact is more important than the difficulty of the material to reach our educational goals. We believe that any method of engineering education that fails to meet these expectations is inefficient. A few decades ago, it was quite natural for our engineering students to study pure mathematics for its beauty, regardless of its applications. In our experience, however, it now motivates only a few percent of them. In this research, the level of motivation was not studied directly, but indirectly by assessing the efficiency with the application of delayed test in a professional context.

The desired role of Engineering Mathematics courses in modern engineering training – We believe that engineering education is one of those areas where being a student is (or should be) an integral part of a professional career. In several ways students must already think as engineers and form professional opinions. Thus, the engineering approach must naturally be present in engineering education, where constant evaluation of efficiency should play a crucial role in the educational process. Our research presented in this paper is based on the idea that each step of education is as successful as it may serve the subsequent steps built upon it. Ultimately, the effectiveness of any engineering education depends on the application of the acquired knowledge in engineering work, at the desired time and under the desired circumstances. Accepting this principle, the efficiency of acquiring mathematics cannot be examined independently of the success in the application of the knowledge in the future.

The circumstances of engineering education have changed in the last decade to such an extent that it is necessary to respond to them by revising and developing didactic tools. The difficulties of fundamental courses also referred to as "barriers" to STEM degrees are discussed e.g. in [16]. The difficulties observed in the engineering mathematics courses replicate other work based on them according to [13]. Methodological issues of mathematics education for non-mathematics students, including engineering students, need to be brought to the forefront. The low level of achievements in mathematics subjects can be partly attributed to inappropriate teaching methods. It seems inevitable to broaden the range of didactic methods and to apply them regularly and in a varied way in engineering education. The learning process in basic courses (especially in mathematics and physics) should be brought closer to that used in secondary education in certain respects, such as interactivity, progress monitoring and regular assessment. In addition, the competencies to be acquired need to be identified more precisely and integrated into the curriculum through specific professional tasks. The way we teach is highly dependent on the level of mathematical knowledge of the incoming students, as discussed in a study by the Dublin Institute of Technology. It shows that the level of mathematics at entry is the strongest predictor of successful completion of the first year for an engineering student [7].

The Hungarian government document entitled "Expected Learning Outcomes", which regulates engineering education in Hungary, places emphasis on defining the competencies to be acquired in engineering education, as well as the method of their assessment and control. Although this provides the framework for the competency-based methodology, a more specific and detailed system of competencies is needed to organize the educational process. The general mathematical competences are extensively discussed in the literature and an overview is provided in [1].

In our research, we focus on the concept of efficiency, in which a specific competence plays a central role: the ability of students to recall and use mathematical concepts and methods when solving professional problems or working as engineers. We emphasize the ability to recall the relevant mathematical topics rather than their advanced application. This competence can be studied primarily in engineering courses that follow the engineering mathematics course. In this paper we introduce a three-level database of mathematical tasks motivated by engineering problems that require different levels of model building. We present the result of a delayed test carried out in Statics to show the results of the systematic inclusion of selected engineering tasks. This test involved the application of various concepts and methods such as vector algebra, linear algebra, and differential and integral calculus.

2. Materials and methods

In order to assess the efficiency of the teaching process specific criteria for success must be established that may vary across different levels and fields. In our investigations in engineering education, we use a special concept of efficiency and a

methodology that we employ to enhance efficiency in this context.

2.1. Literature review

Assessment and improvement of the effectiveness of higher education has become a widely discussed research topic from various perspectives. The peculiarities of vocational training of engineering personnel are discussed in [15]. The necessity of active attitude of engineering students and their motivated participation in the educational process is studied in [4]. A study on the introduction of research tasks into mathematics education in a bachelor program of Applied Mathematics and Computer Science is presented in [6]. In the experimental group the methodology of teaching higher mathematics was based on the introduction of research tasks to establish integrative connections, while in the control group, the mathematical disciplines were taught using traditional teaching methods. It was found that it was possible to develop the research potential of students effectively through the consistent organization of the educational process, including a holistic integrative construct in the mathematics curriculum.

Responding to the challenges caused by the rapidly changing expectations and circumstances, several papers have addressed the measurement and improvement of the learning process in engineering education, see e.g. [8, 10]. There is no question that engineering education needs to adapt to the radically changing needs of the engineering profession. There is a large body of research investigating the impact on effectiveness of different teaching methods, such as differentiated teaching, the inclusion of project work, increased student activity, and the integration of practical tasks into class work. These studies provide the theoretical background and motivation for the present research.

The level of professional competence of engineering students is studied in [15]. The research concludes that engineering education should be a system of educational activities that enables students to be professionally prepared for their future work. Therefore, the education should be oriented to the professional requirements, while the professional competence should be in the foreground in order to ensure efficient work performance.

In [18] the need for a more practical mathematical education in engineering is discussed. The MathePraxis project links the mathematical methods taught in the first semesters and practical problems from engineering applications. Within the project, first-year engineering students demonstrate clearly and convincingly where they will need mathematics in their later working life. In [2], the practice of spaced retrieval was investigated in nine introductory Science, Technology, Engineering, and Mathematics (STEM) courses. This practice involves repeatedly revising the same topics over time with intermittent delays.

Since success in mathematics is highly dependent on the initial level of knowledge and may be described in terms of the change in thinking and application skills, the question of efficiency may not be discussed without examining the mathematical knowledge of incoming students and how we can improve it through catch-up courses. Engineering relies heavily on mathematics, and a lack of basic math

skills significantly hampers students' success. It has been observed that students who lack basic mathematical skills are more likely to perform poorly not only in mathematical modules but also in engineering modules such as thermodynamics, mechanics, and dynamics [12]. Their approach to improve educational efficiency involves assessing the need of individual support through online surveys and using the expertise of talented students to mentor their peers.

According to a UK study, a lack of adequate math skills not only affects students' performance in courses but also leads to higher dropout rates in the first two years of study. Many universities offer math support systems to address these issues, but the success of these programs varies. The research conducted by Gallimore and Stewart (see [9]) presents a novel approach to mathematics support developed and implemented at the School of Engineering, University of Lincoln. This approach provides students with a transition to bridge the gap between secondary school and university level mathematics, offers ongoing support through learning assessment and individual learning plans, and ultimately improves students' achievements, engagement, and retention.

In a 2001 study (see [11]), a total of 95 UK universities were surveyed about the provision of mathematics support, and 46 reported that they provided support for their students. An update in 2004 found that 35 out of 106 UK universities still did not provide mathematics support (see [14]). However, a study [3] published in 2012 found that 88 out of 133 institutions had implemented mathematics learning support programs. A teaching model aimed to improve the quality of mathematics education is introduced and experimentally tested in [20].

2.2. The efficiency concept

In our practice, we measure the effectiveness of an educational activity by the extent to which knowledge is available when it needs to be used in professional subjects or in engineering work. In contrast, the usual assessments (tests, exams, class work) measure the success of learning mathematics only from a mathematical perspective; they say little about how successful engineering students are in applying their knowledge in a non-mathematical environment.

One of the most important tools for process control and improvement in engineering is feedback [5]. Although surveys and student evaluations are regularly conducted in higher education to obtain feedback without a precise formulation of the method and purpose of feedback. These are only formal activities and can provide rather general statements. In order to regulate the educational process, a deeper analysis of knowledge is required. A typical bad example of assessing conformity to expectations is when students are asked how useful they find the mathematical topic they are currently studying. This makes sense if the students has already been studying a subject based on acquired knowledge.

Mathematics is a subject that shapes one's perspective and increases one's professional intelligence, and it is also a preparation for learning engineering subjects. Assessments within subjects that focus on learnable and algorithmic knowledge do not show anything about the real usefulness and the ability of students to apply

the knowledge in a long-term and creative way. A project at the Norwegian University of Science and Technology is presented in [17]. The aim of the project was for students to develop a deep understanding of mathematical concepts and processes, making them better equipped to use mathematics in applications. Digital technology was applied to free up teachers' resources and improve contact with students.

Our goal in teaching engineering mathematics is to help students recall the necessary knowledge in the context of engineering subjects. For this purpose, we use several tools and regularly check how successful our students are in professional applications of mathematics. In this study, we present a special approach to test the level of applicable mathematical knowledge in subsequent semesters in a professional context. In addition, we study the effect of integrating professional content into engineering mathematics courses on the results of delayed mathematics tests.

Our hypothesis was that the way we integrated engineering problems into the classroom as an element of our toolset would result students' better recognition of the necessary mathematical tools and better remembering the computational methods when they solve professional tasks, thus improving efficiency as we define it.

2.3. The framework of teaching Engineering Mathematics and the categories of engineering mathematics tasks

Due to the continuous improvement of the mathematics teaching methodology, a new didactic environment has been established at the Faculty of Engineering University of Debrecen in recent years. This environment includes the presentation of theoretical knowledge through methods such as the use of blackboards and data projectors accompanied by visual aids such as numerous figures and animations. In addition, examples of applications in science and engineering are shown, and related mathematical problems are solved interactively by the instructors or with the participation of students in practical classes. However, our experience in the classroom has shown that this commonly accepted and used method of covering the mathematics curriculum is no longer motivating for most engineering students. As a result, the absorption of new knowledge is not successful enough.

Furthermore, based on the entrance tests of first-year students and the experience of catch-up courses, it should be assumed when planning mathematics courses that students have incomplete knowledge of basic concepts, relationships, and computational methods. They have difficulties in recognizing the relationships between different mathematical topics and lack experience in the process of solving problems by creating and evaluating of models. Therefore, it is necessary to use a teaching method that can simultaneously convey new knowledge, applying it in an illustrative form, and can provide creative and motivating activities that prepare students for the application of mathematical knowledge in technical problem solving.

In our method we integrate engineering problems that require different levels of modeling, which is done in conjunction with simultaneous discussion of analytical

and numerical methods, assigning team project tasks in mathematics and professional courses, and project-based learning. Based on our experience, these tools facilitate a deeper understanding and long-term retention of mathematical knowledge, as well as the development of the ability to apply it in learning technical subjects based on mathematics.

Our observations show that an intensive discussion of engineering applications related to each mathematical topic simultaneously with classical mathematical tasks is uncommon for most Hungarian engineering students, as several of them have not encountered mathematical modeling before.

Although there are tasks in secondary schools that would be suitable for introducing the steps of modeling, students are often unaware of them. The authors regularly offer "mini-courses" for high school students and have the opportunity to study the students' competencies and attitudes [19]. We found that high school students generally prefer solving application-oriented problems to purely mathematical ones. But the success rate is still higher when solving purely mathematical problems. University students also express a desire to solve application-oriented tasks, but their modeling skills are quite low and need to be developed.

As part of our investigation, a task database was prepared, in which the tasks were divided into three categories:

- purely mathematical questions motivated by technical applications;
- technical questions with the model provided and only mathematical knowledge is needed for the solution;
- technical tasks formulated in a professional context requiring model creation and higher-level, complex mathematical knowledge.

For most engineering students, it is difficult to identify the appropriate mathematical concept or method related to the professional problem they need to solve. Similar to our observations in high schools, although more university students prefer solving real-world problems to purely mathematical problems, they are less successful in the former one. We believe that this phenomenon is due to the lack of experience with mathematical modeling in secondary education.

In order to prepare students to use mathematics as a tool, we must create a synergy between mathematical and professional subjects, emphasizing as many points of connection as possible. The gradual introduction of practice from the beginning of the study program is essential to develop the ability to use mathematical tools.

Our hypothesis was that regular discussion of professional problems from our three-level database in engineering mathematics classes would result students' better recognizing the mathematical tools needed and remembering the computational methods when they have to solve professional problems.

It is obvious that some students can recall the examples they studied even several semesters later when asked in the same context. Therefore, the delayed tests formally included questions on the professional topic that were different from both standard mathematical texts and engineering problems discussed in engineering

mathematics classes. In the tests, they had to solve simple exercises related to the professional topic with emphasis on the mathematical content to assess the current mathematical skills.

In the experimental group, during the discussion of each mathematical topic, students were introduced to mathematical concepts in the usual way and solved typical mathematical examples in 75% of the time. After that, every week they were asked to solve technical problems from all categories using the mathematical tools they had just learned.

The goal is to simultaneously develop mathematical and professional intelligence by improving the ability to build models and demonstrate connections to professional topics. As a result, students can acquire deeper knowledge, and find answers to questions such as "What's the point of all this?" and "Why do I need to study mathematics?"

Below are examples of all three categories of tasks in the database.

Category 1: Purely mathematical questions motivated by technical applications

In the first part of the task collection, there are exercises that are purely mathematical problems. In some cases, they are formulated as technical questions, and in all cases, they touch on technical applications during the solution.

Example 1.1. A precise approximation of the curve of a corner of a Formula 1 racetrack is described by the graph $f(x) = x^2 + 2x$. If the car moving on the track drifts off at $x = 1$ along the tangent of the track, does it hit the column at the coordinate point $P = (2; 5)$?

Example 1.2. Engineers are planning a straight tunnel with an inverted parabolic cross-section under a mountain. The tunnel is 9 meters high and 6 meters wide at the bottom. What is the largest rectangular cross-section (width and height) of the truck that can still drive through the tunnel?

Example 1.3. The widths of two orthogonally intersecting corridors are 2.4 meters and 1.6 meters, respectively. How long is the ladder that can be taken from one corridor to another?

Category 2: technical questions for which the model is provided and only mathematical knowledge is needed for the solution

In the second category, we classified tasks that are technical or physical in nature but require mathematical knowledge to solve. We believe that it is important to provide students with practical tasks in the mathematics course that include technical examples beyond traditional mathematics education.

Example 2.1. An elevator whose motor is on the top floor is held up by a wire rope. We also know that a 1-meter piece of wire rope weighs 45 [N]. When the cabin is on the ground floor, 60 [m] of cable hangs down. By the time the elevator

reaches the top floor, the cable is fully rolled up. How much work is required just to pull up the cable?

Example 2.2. The length of a spring in the unstretched state is 20 [cm]. To stretch it to a length of 30 [cm] requires a force of 40 [N]. How much work is required to stretch the spring from 35 [cm] to 38 [cm]?

Example 2.3. A lawnmower manual says to tighten the spark plug by a torque of 20.4 [Nm]. If the force is applied to the spark plug wrench from a distance of 25 [cm] from the spark plug, how much force is required to achieve the required torque?

Category 3: Technical tasks formulated in a professional context, requiring model creation and higher-level, complex mathematical knowledge

In the third category we have classified tasks that are no longer purely mathematical, but technical tasks that appear in other subjects. These tasks require the use of higher-level mathematical tools for their solution. Our goal was for students to develop a comprehensive understanding of the mathematical knowledge that appears in other subjects during their studies. We wanted them to be able to apply the methods they had learned in their mathematics courses, rather than just focusing on the process of solving problems.

Example 3.1. Regarding the DC circuit given in Figure 1 solve the problem listed below. Data: $U_{b_1} = 20q\,[\mathrm{V}]$, $U_{b_2} = 10\,[\mathrm{V}]$, $U_{b_3} = 5\,[\mathrm{V}]$, $R_1 = 2\,[\Omega]$, $R_2 = 4\,[\Omega]$, $R_{b_1} = 7\,[\Omega]$, $R_{b_2} = 6\,[\Omega]$, $R_{b_3} = 4\,[\Omega]$. Apply Kirchhoff's first rule for node B. Apply Kirchhoff's second rule for loops $A-B-E-F-A$ and $B-C-D-E-B$. Give the matrix of the obtained system of linear equations. Calculate the unknown current intensities.

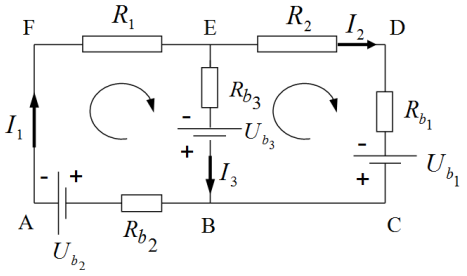


Figure 1. Electric circuit referred to in Example 3.1.

Example 3.2. The stress tensor elements at point P of a structure are demonstrated in the elementary cube in Figure 2. Data: $\sigma_x = 50\,[\mathrm{MPa}]$, $\sigma_y = -30\,[\mathrm{MPa}]$, $\sigma_z = 25\,[\mathrm{MPa}]$, $\tau_{xy} = \tau_{yx} = 30\,[\mathrm{MPa}]$, $\alpha = 30°$. Give the coordinates of unit normal vector $\bar{n}$ if it is in the $x-z$ plane and its angle with the x axis is α. Give the

matrix of the stress tensor at point P. Calculate the stress vector $\bar{\rho}_n$, the normal stress σ_n and the shear stress. Determine the magnitude and the direction of the principal stresses.

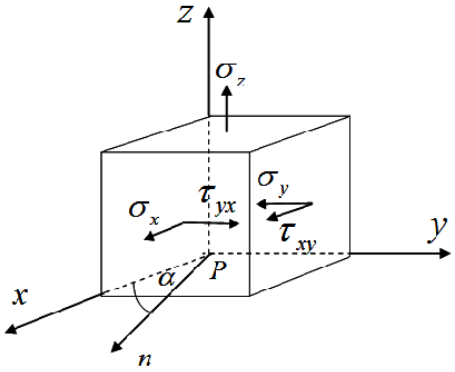


Figure 2. Stress state of a body referred to in Example 3.2.

Example 3.3. Suppose that there are three rotating parts in a machine generating harmonic vibration of the machine structure. The rotational speed values of the three parts are 600 rpm, 720 rpm and 1100 rpm, respectively. The effective velocity values of the three harmonic vibrations are 5.4 mm/s, 3.9 mm/s, 6.0 mm/s. Give the vibration state of the machine in the time domain and in the frequency domain with the velocity-time and the velocity-frequency diagrams.

2.4. Delayed tests

In this research, the delayed test consisted of mathematical questions based on the material covered in Engineering Mathematics I, but formulated as technical problems using the terminology of Statics. The students were not informed about the nature of the questions either before or during the test; therefore, they had to interpret the situations themselves. Although minimal knowledge of the subject Statics was required to provide answers, the presence of this knowledge was a prerequisite for passing the course. It was therefore safe to assume that the students had this knowledge. Once the questions were interpreted, solving them required only the use of purely mathematical tools. The test questions were as follows.

Question 1 (Q1)**.** Force F acts on a material point which is on the surface of the incline in Figure 3. The angle between the horizontal plane and the incline is 25°. The coordinates of force F in the blue coordinate system are $x = 2$ and $y = 5$.

Give the coordinates of F in the red coordinate system. Consider that the red coordinate system can be obtained by the rotation of the blue one.

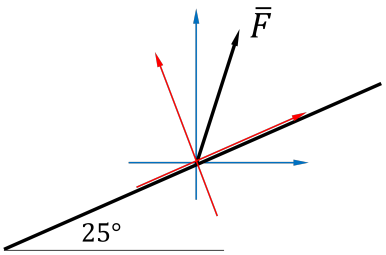


Figure 3. Force system referred to in Question 1.

Question 2 (Q2)**.** A distributed force system given by the intensity $f(x) = x\sin\left(\frac{\pi}{2}x - 2\pi\right)\left[\frac{\mathrm{N}}{\mathrm{m}}\right]$, $4 \le x \le 6$ acts on a 2 meters long segment of the supported beam in Figure 4. Calculate the resultant of a distributed force system.

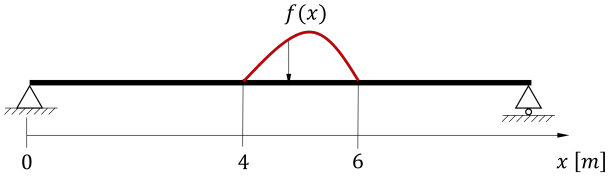


Figure 4. Beam loaded by a distributed force system referred to in Question 2.

Question 3 (Q3)**.** The bending moment acting on a prismatic beam is given as a function of coordinate x as $M_b(x) = -x\sqrt{100 - 4x^2}\,[\mathrm{Nm}]$, $0 \le x \le 5$. Calculate the value of the shear force at $x = 3$.

Question 4 (Q4)**.** Calculate the moment vectors of forces $\bar{F}_1$ and $\bar{F}_2$ (Figure 5) relative to point O, and calculate the angle between the two-moment vectors. Data:

$$\bar{F}_1 = \begin{pmatrix} -2 \\ 5 \\ 1 \end{pmatrix},\ \bar{F}_2 = \begin{pmatrix} 2 \\ 8 \\ -4 \end{pmatrix},\ \bar{r}_1 = \begin{pmatrix} 2 \\ -1 \\ 1 \end{pmatrix},\ \bar{r}_2 = \begin{pmatrix} 3 \\ 0 \\ 2 \end{pmatrix}.$$

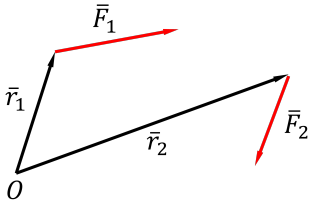


Figure 5. Forces referred in Question 4.

The mathematical knowledge needed to answer the questions:

Q1: linear transforms of the plane; the required coordinates can be obtained by rotating the force vector $-25°$, to get it we need the matrix of the rotation.

Q2: calculation of integrals; the resultant of a distributed force system with force density f can be calculated as $F_{\text{res}} = \int_a^b f$.

Q3: differentiation; the shear force function can be given as the negative derivative of the bending moment function.

Q4: vector operations; moment vectors can be given as vector product of force vectors and position vectors; the angle can be calculated with scalar production.

2.5. The experiment

80 students majoring in vehicle and mechanical engineering participated in the study: 40 students in the experimental group and 40 students in the control group. In these majors, the course Engineering Mathematics I consists of 4 hours of lecture and 4 hours of practical. The inclusion of technical examples of different levels serves to increase the efficiency according to our concept.

At the Faculty, the level of knowledge of the incoming students is checked every year with an entrance test consisting of high school exercises. The students in the two groups achieved almost identical results in this entrance test. The two-sample t-test indicated that there was no significant difference between the scores of the Experimental group ($M = 44.95$, $SD = 24.96$) and the Control group ($M = 49.18$, $SD = 25.70$), $t(78) = 0.75$, $p = 0.458$ (two-tail, $d = 0.17$).

Both the experimental group and the control group studied Engineering Mathematics I according to the same curriculum and for the same number of hours. However, the students in the experimental group spent 1 hour of the 4-hour practical class each week studying models and solving engineering problems, while the students in the control group only solved classical mathematical problems.

The subject Engineering Mathematics I covers the following topics of linear algebra and mathematical analysis: matrix algebra, linear spaces, linear functions; real functions, properties, elementary functions, composition, and inverse of functions; continuity, limit, derivative, linear approximation; Taylor polynomials, analysis of differentiable functions; Riemann integral; anti-derivative; Newton-Leibniz formula; numerical integration; applications of integral calculus.

Our two hypotheses in this research were as follows:

H1: Incorporating engineering problems of different categories into classroom work helps students understand the course material resulting better performance in Engineering Mathematics I for students in the experimental group.

H2: The students in the experimental group, for whom engineering problems are systematically integrated into the Engineering Mathematics I course according to our methodology, perform better on the delayed mathematics tests in the Statics course than the students in the control group.

3. Results

Although our main goal was to examine our teaching methodology in terms of our efficiency concept, we were also interested in its effect on the test scores of the math course.

The results of the two tests written in the "standard" mathematical context indicate that there is no significant difference between the two groups based on the total scores obtained. The two-sample t-test indicated that there was no significant difference between the scores of the Experimental group ($M = 41.85$, $SD = 23.00$) and the Control group ($M = 47.03$, $SD = 23.70$), $t(78) = 0.995$, $p = 0.325$ (two-tail, $d = 0.22$). Thus hypothesis H1 was rejected.

It should be noted that this is not particularly surprising, as we have observed that the mathematics test scores are mostly correlated with the amount of time spent practising computational steps, rather than with the specific knowledge required to study engineering subjects.

The post-measurement was conducted in the frame of the Statics subject, which is based on Engineering Mathematics I and takes place one semester later. The experimental and control groups studied Statics under identical conditions. The post-measurement was conducted with the first test of Statics subject and it was called "extra test" (hiding the research purpose) for extra points. Students were allowed to earn 10% of the total points with this part.

Regarding the second hypothesis (H2), it should be emphasized that both groups received the same mathematical knowledge in the first semester and the same professional knowledge in Statics in the second semester. Both groups had to answer the same questions in the delayed test. The questions were not covered in Engineering Mathematics I for either the experimental group or the control group, thus preventing the students from recalling the answers.

The scores were compared in the two groups; the result of the two sample t-test confirmed our second hypothesis. The students who studied mathematics in a way that regularly involved solving technical problems of different modeling levels during a part of the lessons (Experimental group) ($M = 54.15$, $SD = 24.17$) achieved significantly better results in the subsequent assessment of their mathematical knowledge in the Statics subject than students in he Control group ($M = 37.03$, $SD = 21.24$), $t(78) = 3.36$, $p = 0.001$ (two-tail, $d = 0.75$).

4. Conclusions

In our study, we formulated our definition of the efficiency of teaching engineering mathematics and presented our teaching method aimed to improve efficiency in this sense. Among the three main elements of our methodology, we analyzed the effect of the targeted application of different categories of engineering tasks using a delayed test conducted in the context of a professional subject to be studied in the following semester.

In cooperation with the lecturer of the engineering course we prepared a special delayed test with new types of questions for this study. The test focused on mathematical knowledge but the questions were presented as technical texts. The test questions differed from the exercises discussed in the mathematics classes of both the experimental and control groups, as well as from the questions in the regular mathematics and statics tests. Some of the delayed test questions asked in the Statics course are presented in Subsection 2.4.

To implement the method of "Integration of engineering problems into classwork" we created a collection of tasks consisting of three groups: purely mathematical questions motivated by technical applications, professional questions with given models requiring only mathematical knowledge for their solution, and engineering tasks presented as professional texts requiring model building and higher-level, complex mathematical knowledge. While various collections of engineering problems for discussion in mathematics classes are mentioned in the literature, we also categorized the problems according to the level of modeling required and prepared a unique collection of tasks organized by topic and modeling difficulties.

In the experimental group, we specifically involved professional tasks, dedicating 1 hour of each 4-hour practical class to them. We compared the test results of the two groups within the Engineering mathematics course (normal test) and the Statics course (special post-test). Our results showed that, although there was no significant difference between the two groups in terms of the regular tests of the Engineering Mathematics I course, the experimental group performed significantly better in the mathematics survey of the Statics course that took place one semester later.

Based on our results, to improve the effectiveness of engineering mathematics education, we recommend to conduct mathematics post-tests in the context of professional. If the effectiveness of educational activity is measured by the extent to which knowledge is available for practical use, our study suggests that dedicating a portion of class time to posing and solving professional problems with a focus on modeling significantly increases the ability to recognize the necessary mathematical tools and the effectiveness of knowledge retrieval in the professional environment.

For a more in-depth analysis of the impact of our methodology on the effectiveness of teaching mathematics we are preparing post-tests for further engineering courses and we plan to request more detailed derivations and explanations to allow for a qualitative analysis of mathematical knowledge one or more semesters after learning the subject. Although the result of the t-test and our subjective anal-

ysis confirmed our second hypothesis, larger groups would be involved in further research to increase the reliability of our findings.

References

[1] B. A. Alpers, M. Demlova, C. H. Fant, T. Gustafsson, D. Lawson, L. Mustoe, D. Velichova: *A framework for mathematics curricula in engineering education: a report of the mathematics working group* (2013).

[2] C. R. Bego, K. B. Lyle, J. C. Immekus, P. A. Ralston: *Introducing desirable difficulty in STEM barrier courses with spaced retrieval practice*, IEEE Frontiers in Education Conference (FIE) (2021), pp. 1–6.

[3] M. an Bhaird, D. C. Lawson: *How to set up a Mathematics and Statistics Support Provision*, Sigma – Centre of Excellence in Mathematics and Statistics Support (2012).

[4] E. Bitay, G. Bagyinszki: *Didactic and methodological aspects of technical higher education*, Műszaki Tudományos Közlemények 17.1 (2022), pp. 1–5, doi: 10.33894/mtk-2022.17.0.

[5] I. K. D. Sipos: *Supporting the education of engineering mathematics using the immeadiate feedback method*, Teaching Mathematics and Computer Sciences 21.1 (2023), pp. 49–61.

[6] S. N. Dvoryatkina, R. A. Melnikov, V. E. Shcherbatykh: *Identification of the Research Potential of Students in the Process of Revealing Integrative Connections of the Subject Content of Mathematical Courses*, European Journal of Contemporary Education 11.3 (2022), pp. 718–726, doi: 10.13187/ejced.2022.3.718.

[7] E. N. Fhloinn, M. Carr: *What do they really need to know? Mathematics requirements for incoming engineering undergraduates* (2010).

[8] S. Freeman, S. L. Eddy, M. McDonough, M. K. Smith, N. Okoroafor, H. Jordt, M. P. Wenderoth: *Active learning increases student performance in science, engineering, and mathematics*, Proceedings of the national academy of sciences 111.23 (2014), pp. 8410–8415.

[9] M. Gallimore, J. Stewart: *Increasing the impact of mathematics support on aiding student transition in higher education*, Teaching Mathematics and its Applications: An International Journal of the IMA 33.2 (2014), pp. 98–109, doi: 10.1093/teamat/hru008.

[10] S. Henderson, P. Broadbridge: *Engineering mathematics education in Australia*, MSOR Connections 9.1 (2009), pp. 12–17.

[11] D. Lawson, T. Croft, M. Halpin: *Evaluating and Enhancing the Effectiveness of Mathematics Support Centres*, Final report of a project funded by LTSN MSOR Centre. (2001).

[12] D. Lawson, T. Croft, M. Halpin: *Second edition of a guide for those interested in the establishment and development of Mathematics Support Centres in institutes of higher education* (2003).

[13] K. B. Lyle, C. R. Bego, R. F. Hopkins, J. L. Hieb, P. A. Ralston: *How the amount and spacing of retrieval practice affect the short-and long-term retention of mathematics knowledge*, Educational Psychology Review 32 (2020), pp. 277–295, doi: 10.1007/s10648-019-09489-x..

[14] G. Perkin, A. C. Croft: *Mathematics Support Centres–the extent of current provision*, MSOR Connections 4.2 (2004), pp. 14–18.

[15] A. D. Plutenko, A. V. Leyfa, A. V. Kozyr, T. V. Haletskaya: *Specific Features of Vocational Education and Training of Engineering Personnel for High-Tech Businesses*, European Journal of Contemporary Education 7.2 (2018), pp. 360–371, doi: 10.13187/ejced.2018.2.360.

[16] A. Redmond-Sanogo, J. Angle, E. Davis: *Kinks in the STEM pipeline: Tracking STEM graduation rates using science and mathematics performance*, School Science and Mathematics 116.7 (2016), pp. 378–388, doi: 10.1111/ssm.12195.

[17] F. Rønning: *Future teaching of mathematics for engineers*, in: Proceedings from 42nd SEFI Annual Conference, Belgium: European Society for Engineering Education, 2014.

[18] A. Rooch, P. Junker, J. Härterich, K. Hackl: *Linking mathematics with engineering applications at an early stage–implementation, experimental set-up and evaluation of a pilot project*, European Journal of Engineering Education 41.2 (2016), pp. 172–191.

[19] D. Sipos: *A numerikus számítások szerepe a műszaki modellekben*, International Journal of Engineering and Management Sciences 3.5 (2018), pp. 76–83.

[20] E. V. Soboleva, T. N. Suvorova, M. I. Bocharov, T. I. Bocharova: *Development of the Personalized Model of Teaching Mathematics by Means of Interactive Short Stories to Improve the Quality of Educational Results of Schoolchildren*, European Journal of Contemporary Education 11.1 (2022), pp. 241–257, doi: 10.13187/ejced.2022.1.241.

Annales Mathematicae et Informaticae
60 (2024) pp. 278–284
DOI: 10.33039/ami.2024.12.002
URL: https://ami.uni-eszterhazy.hu

On elementary representations of $\cos 75°$ and $\cos 15°$*

Anna Stirling[a], **Csaba Szabó**[bc], **Sára Szörényi**[a], **Éva Vásárhelyi**[a], **Janka Szeibert**[cd]

[a]MTA-ELTE Theory of Learning Mathematics Research Group

[b]Edutus University

[c]ELTE Eötvös Loránd University

[d]HUN-REN Alfréd Rényi Institute of Mathematics

szeibert.janka@tok.elte.hu, stirling.anna@gmail.com, szabo.csaba.mathdid@ttk.elte.hu, vasareva@gmail.com, szorenyi.sara@aquilone.hu

Abstract. In this paper we present geometric and algebraic representation of values of $\cos 75°$ and $\sin 75°$ without using trigonometric identities or iterated applications of taking roots.

Keywords: values of sine and cosine, representation of real numbers, roots

AMS Subject Classification: 97F40, 97G40

1. Introduction

Trigonometry undoubtedly plays an important role in higher mathematics, physics, engineering and various other fields of science. Nevertheless, students fail to understand and apply trigonometry when it is needed, and the reason for this is quite complex. A semantic analysis shows that only fifty-three per cent of students can solve problems where the unknowns are a distance or an angle and for which knowledge it is necessary to calculate the sine or cosine of an angle [6]. Even preservice teacher students with prior knowledge do not manage to connect the visual representation and the symbolic or verbal representation of sine and cosine, although they are successful in solving problems in all three ways [5]. The rea-

*This work is supported by the Digital Education Development Competence Center at Eötvös Loránd University, Budapest, with project number 2022-1.1.1-KK-2022-00003.

Submitted: June 13, 2024
Accepted: December 12, 2024
Published online: December 20, 2024

son for this could be that they either do not have enough experience in solving problems or that they do not have enough visual experience with sine and cosine values. Understanding trigonometric functions is not a straightforward process and can be affected by many different variables – such as understanding different units for angle measures, the ability to associate triangles with numerical relationships, or understanding functions for which there is no explicit algebraic formula to determine their values[4, 7, 14]. However, several studies suggest that appropriately designed learning environments can improve students' geometric understanding. In particular, various visual representations, including dynamic ones, are especially important to support the learning process [4, 8]. The effectiveness of processing is improved when we work on multiple levels of representation at each stage [2, 3, 9]:

- we physically construct, fold, etc. the corresponding construction;
- we record our observations in drawings, tables, etc.
- we also express our experiences, assumptions and arguments in words

"Consequently, there is a need for more activities that can be used to assess students' understanding of concepts that are integral to the learning of trigonometry." (Arsalan Wares) ([11] pp. 141.)

In this note we present geometric ways to find the values of $\cos 75^\circ$ and $\sin 75^\circ$ without using trigonometric identities or iterated applications of taking roots. The topic is particularly topical as trigonometric identities are no longer included in the 2020 national curriculum. The introduction of trigonometric functions as functions and the associated calculations are not supported on the intermediate level. Trigonometric identities, like expressing $\sin(\alpha+\beta)$ or $\cos 2\alpha$ are totally missing.

The current National Core Curriculum from 2020 specifies the following:
Grades 9–10.: No trigonometry.
Grades 11–12.:

- Sine, cosine, tangent of acute angles.
- Calculations in right triangles using angle functions in practical situations.
- Sine, cosine, tangent of obtuse angles.
- Understanding relationships between different angle functions of a given angle: Pythagorean identity, co-function identities, and supplementary angles.
- Determining an angle using a calculator given the value of an angle function.
- Calculating the area of a triangle knowing two sides and the included angle.
- Knowledge and application of the sine and cosine rules.
- Proof of the sine rule.
- Calculations in quadrilaterals and polygons using angle functions.

We can see that the curriculum does not extend angle functions to arbitrary rotation angles, so they are essentially not treated as functions. There is no need to know how to graph them, transform them, and certainly not to understand trigonometric identities, addition formulas, and other operations with them. Thus, we are spared from interesting problems like finding $\sin 75^\circ$ by expanding $\sin(45^\circ + 30^\circ)$. However, the National Core Curriculum does not set an upper limit on the material that can be taught to students. Therefore, those studying mathematics on an advanced level can still solve problems requiring more complex algebraic and trigonometric knowledge, and the advanced-level textbook does indeed include such content. Still, this knowledge is no longer commonly expected.

Values of $\cos 75^\circ$ and $\sin 75^\circ$ can be calculated with the Ailles-square [1], or with paper-folding [12]. In [13] a tricky way is applied to find these values with the use of two diagonals of a regular dodecagon. Another construction can be found in [11], but it is using iterated square roots.

In this paper we give different ways to find the values of $\cos 75^\circ$ and $\sin 75^\circ$ on a high-school level. Besides trigonometric identities we avoid iterated roots, as well. This work is a continuation of a previous study in which we presented and examined several geometric constructions for upper elementary and high school levels [10].

2. Elementary calculation methods

In this chapter, we show methods of finding the algebraic form of

$$a = 2\cos 15^\circ \quad \text{and} \quad b = 2\cos 75^\circ = 2\sin 15^\circ$$

via geometric construction. We shall always aim at finding a degree two polynomial (equation) such that $a = 2\cos 15^\circ$ is a root (solution) of it. The most convinient polynomials are

$$\begin{aligned}
&x^2 + \sqrt{3} - 2 \quad \text{with roots} \quad a = \sqrt{2-\sqrt{3}} \quad \text{and} \quad -a = -\sqrt{2-\sqrt{3}},\\
&x^2 - \sqrt{6}x + 1 \quad \text{with roots} \quad a = \frac{\sqrt{6}+\sqrt{2}}{2} \quad \text{and} \quad b = \frac{\sqrt{6}-\sqrt{2}}{2},\\
&x^2 - \sqrt{2}x - 1 \quad \text{with roots} \quad a = \frac{\sqrt{6}+\sqrt{2}}{2} \quad \text{and} \quad -b = \frac{\sqrt{2}-\sqrt{6}}{2}.
\end{aligned}$$

We would like to avoid nested roots so that we will arrive at either $x^2 - \sqrt{2}x - 1$ or $x^2 - \sqrt{6}x + 1$. We can see that

$$a + b = \sqrt{6} \quad \text{and} \quad a - b = \sqrt{2}$$

If we also consider

$$a^2 + b^2 = 4,$$

then after substituting $a = b + \sqrt{2}$ or $a = \sqrt{6} - b$ we obtain the desired degree 2 polynomials. We show a method to find $a - b$ and two constructions to find $a + b$.

2.1. Aiming for $a = b + \sqrt{2}$: The house-shaped construction

If we fit isosceles triangles, $AEH\Delta$, $BHC\Delta$ and $DCE\Delta$ with sides 1 and with angles 90° and 30° and 150°, then we obtain Figure 1. Then $HAB\Delta$ is an isosceles triangle with $AHB\sphericalangle = 60°$, hence $HAB\Delta$ is equilateral. To reverse the construction, let's take a square with unit-length sides and construct an equilateral triangle with unit-length sides on one of the square's sides. Then $|EC| = a$. If we draw the $ABH\Delta$ equilateral triangle, point H will be on the segment EC, and the notations and values in Figure 1 will hold.

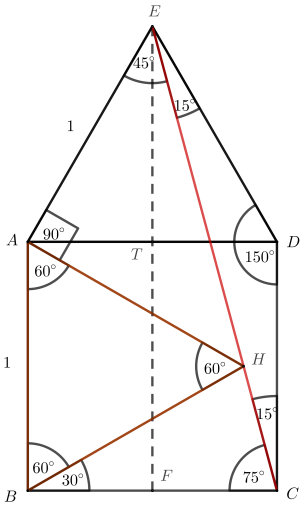


Figure 1. With additional segments $2\sin 15^\circ$ can be represented.

If we draw the common axis of symmetry for the square $ABCD$ and the equilateral triangle $ADE\Delta$ we obtain points T and F. The segment ET is the height of an equilateral triangle with unit sides, thus its length is $\frac{\sqrt{3}}{2}$. Then

$$FC = \frac{1}{2}BC = \frac{1}{2},$$
$$EF = ET + TF = \frac{\sqrt{3}}{2} + 1.$$

Using the Pythagorean theorem, we can calculate the length of the hypotenuse CE:

$$\left(\frac{\sqrt{3}}{2} + 1\right)^2 + \left(\frac{1}{2}\right)^2 = (CE)^2,$$
$$\frac{\left(\sqrt{3}+2\right)^2}{4} + \frac{1}{4} = \frac{8 + 4\sqrt{3}}{4} = 2 + \sqrt{3} = (CE)^2,$$

$$CE = \sqrt{2 + \sqrt{3}}.$$

This is an elementary method to calculate $2\cos 15^\circ$

$$2\cos 15^\circ = \sqrt{2 + \sqrt{3}}.$$

We obtain the value with doubly nested square roots, so this is a dead end.

We want to calculate the lengths of segments EH and CH in Figure 1. We know that the angles of the original triangle $CDE\Delta$, which is isosceles, are $DEC\sphericalangle = ECD\sphericalangle = 15^\circ$ and $EDC\sphericalangle = 150^\circ$. Triangle $AHE\Delta$ is an isosceles right triangle, and triangle $BCH\Delta$ is an isosceles triangle with base angles $BCH\sphericalangle = CHB\sphericalangle = 75^\circ$ and sides of unit-lengths. Hence

$$EH = \sqrt{2}, \quad CH = b, \quad EH = a.$$

Segment EH can be given as the sum of segments CH and EH:

$$b = a + \sqrt{2}. \tag{2.1}$$

Using

$$a^2 + b^2 = 4 \tag{2.2}$$

Substituting Equation (2.1) into Equation (2.2) we get

$$x^2 - \sqrt{2}a - 1 = 0$$

and the quadratic formula gives

$$\frac{\sqrt{2} \pm \sqrt{2 - 4 \cdot 1 \cdot (-1)}}{2} = \frac{\sqrt{2} \pm \sqrt{6}}{2}.$$

Thus, without resorting to algebraic manipulations, we can determine $2\cos 15^\circ$.

2.2. Aming for $a + b$: Cyclic trapezoid-shaped construction

The first idea of constructing $\sqrt{6}$ is to construct an isosceles triangle $CKD\Delta$ with angle 120° and sides $\sqrt{2}$. Let it complete with two isosceles right triangles with sides of unit length to a trapezoid. Or, from another point of view with an isosceles triangle $AKF\Delta$ with angle 150° and sides 1 as in Figure 2. The angles on the longer base of the trapezoid $CDFA$ are 75°, and the angles on the shorter base are 105°. Let CD be parallel to DF. Then

$$AF = \sqrt{6}, \quad AB = b, \quad CD = BF = a. \tag{2.3}$$

Thus

$$\sqrt{6} = AF = AB + CD = a + b.$$

Substituting Equation (2.3) into $a^2+b^2 = 4$ we get the following quadratic equation:

$$a^2 - \sqrt{6}a + 1 = 0$$

from which the segment lengths can be obtained without the use of nested roots form the quadratic formula

$$a = \frac{\sqrt{6} \pm \sqrt{6 - 4 \cdot 1 \cdot 1}}{2} = \frac{\sqrt{6} \pm \sqrt{2}}{2}.$$

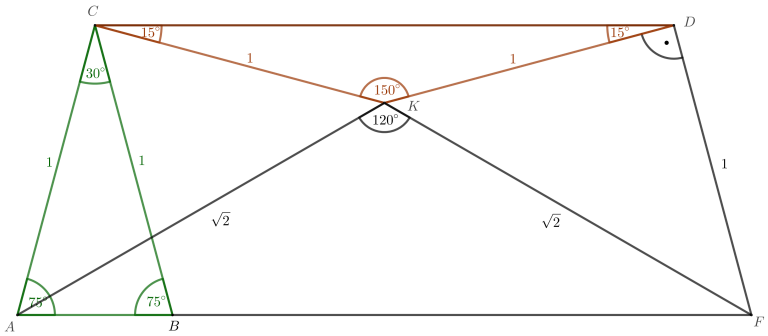


Figure 2. Symmetrical trapezoid-shaped construction.

2.3. Finding $a + b$: Pentagon shaped construction

Another way to construct $\sqrt{6}$ is by drawing an isosceles triangle with leg $\sqrt{3}$, triangle $EAF\Delta$ in Figure 3. Draw the isosceles triangles $ECA\Delta$ and $EDF\Delta$ with angle 120° of unit size lengths. Again we obtain the trapezoid of Figure 2, putting the $ECD\Delta$ on the "top". From here the same calculations as in the previous section lead to a.

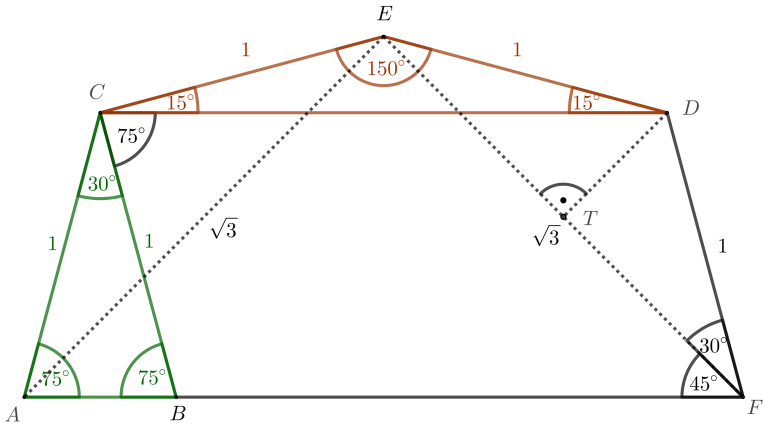


Figure 3. Symmetric pentagon-shaped construction.

Acknowledgements. The authors would like to thank the reviewer for his suggestions and clarifying the scope of the paper.

References

[1] D. Ailles: *Trinagles and trigonometry*, NCTM 64.6 (1971), p. 562, doi: 10.5951/mathteacher.112.5.0400.

[2] J. Bruner: *Toward a Theory of Instruction*, Harvard University Press, 1974, url: https://books.google.hu/books?id=28bmEAAAQBAJ.

[3] J. S. Bruner: *"The Process of Education" Revisited*, The Phi Delta Kappan 53.1 (1971), pp. 18–21, issn: 00317217, url: http://www.jstor.org/stable/20373062 (visited on 06/11/2024).

[4] E. Cekmez: *What generalizations do students achieve with respect to trigonometric functions in the transition from angles in degrees to real numbers?*, The Journal of Mathematical Behavior 58 (June 2020), p. 100778, doi: 10.1016/j.jmathb.2020.100778.

[5] S. Dündar: *Mathematics Teacher-Candidates' Performance in Solving Problems with Different Representation Styles: The Trigonometry Example.* EURASIA J Math Sci Tech Ed. 11.6 (2015), pp. 1379–1397, doi: 10.12973/eurasia.2015.1396a.

[6] R. L. Martín-Fernández E Ruiz-Hidalgo JF: *Meaning and Understanding of School Mathematical Concepts by Secondary Students: The Study of Sine and Cosine.* EURASIA J Math Sci Tech Ed. 15.12 (2019), EM1782, doi: 10.29333/ejmste/110490.

[7] K. Moore: *Making sense by measuring arcs: A teaching experiment in angle measure*, Educational Studies in Mathematics 83 (June 2013), pp. 225–245, doi: 10.1007/s10649-012-9450-6.

[8] K. Moore: *Trigonometry, technology, and didactic objects*, in: In:S. L. Swars, D. W. Stinson and S. Lemons-Smith (Eds.), Proceedings of the 31st annual meeting of the North American Chapter of the International Group for the Psychology of Mathematics Education, 2009, pp. 1480–1488.

[9] S. R. W.: *Hemispheric specialization of mental faculties in the brain of man*, in: Claremont: Claremont Graduate School, 1972, pp. 126–136.

[10] C. Szabó, C. Zámbó, A. Stirling, J. Szenderák, S. Szörényi: *Geometric representations of irrational algebraic numbers in Hungarian high school mathematics education*, in: In: Éva, Vásárhelyi; Johann, Sjuts (szerk.) Theoretische und empi- rische Analysen zum geometrischen Denken, 2021, pp. 323–340.

[11] A. Wares: *Geometry and trigonometry interplay*, NCTM 119.5 (2019), p. 400, doi: 10.5951/mathteacher.112.5.0400.

[12] A. Wares: *Paper folding and trigonometric ratios*, International Journal of Mathematical Education in Science and Technology 50.4 (2019), pp. 636–641, doi: 10.1080/0020739X.2018.1500655.

[13] A. Wares: *Reasoning and proof in trigonometry*, International Journal of Mathematical Education in Science and Technology 54.1 (2023), pp. 141–144, doi: 10.1080/0020739X.2022.2035001.

[14] K. Weber: *Connecting Research to Teaching: Teaching Trigonometric Functions: Lessons Learned from Research*, The Mathematics Teacher 102 (Sept. 2008), pp. 144–150, doi: 10.5951/MT.102.2.0144.