

Contents

Research papers

R. FRONTCZAK, A short remark on Horadam identities with binomial coefficients	5
B. KAFLE, F. LUCA, A. TOGBÉ, Corrigendum to “Pentagonal and heptagonal repdigits” [Ann. Math. et Inf. 52 (2020) 137–145]	15
Z. KÁSA, Warshall’s algorithm—survey and applications	17
G. KUSPER, CS. BIRÓ, A. ADAMKÓ, I. BAJÁK, Introducing w -Horn and z -Horn: A generalization of Horn and q -Horn formulae	33
O. LANTANG, GY. TERDIK, A. HAJDU, A. TIBA, Comparison of single and ensemble-based convolutional neural networks for cancerous image classification	45
S. E. RIHANE, A. TOGBÉ, On the intersection of Padovan, Perrin sequences and Pell, Pell-Lucas sequences	57
M. SAHAI, S. F. ANSARI, The structure of the unit group of the group algebra $F(C_3 \times D_{10})$	73
V. SKALA, Efficient Taylor expansion computation of multidimensional vector functions on GPU	83
W. SRIPRAD, S. SRISAWAT, P. NAKLOR, Vieta–Fibonacci-like polynomials and some identities	97
Sz. SVITEK, M. VONTSZEMŰ, On structure of the family of regularly distributed sets with respect to the union	109
N. X. THO, The Diophantine equation $x^2 + 3^a \cdot 5^b \cdot 11^c \cdot 19^d = 4y^n$	121
N. X. THO, What positive integers n can be presented in the form $n = (x + y + z)(1/x + 1/y + 1/z)$?	141
R. TÓTH, Script-aided generation of Mental Cutting Test exercises using Blender	147
A. VÁGNER, Route planning on GTFS using Neo4j	163

Methodological papers

F. LAUDANO, Visual argumentations in teaching trigonometry	183
Z. MATOS, E. KÓNYA, Teaching numeral systems based on history in high school	195
M. RÓZÁNSKI, B. SMOLEŃ-DUDA, R. WITUŁA, M. JOCHLIK, A. SMUDA, On some methods of calculating the integrals of trigonometric rational functions	205
Cs. SZABÓ, Cs. BERECKY-ZÁMBÓ, J. SZENDERÁK, J. SZEIBERT, On a metamathematical question in talent care	215

ANNALES MATHEMATICAE ET INFORMATICAЕ 54. (2021)

ANNALES MATHEMATICAE ET INFORMATICAЕ

TOMUS 54. (2021)



COMMISSIO REDACTORIUM

Sándor Bácsó (Debrecen), Sonja Gorjanc (Zagreb), Tibor Gyimóthy (Szeged),
Miklós Hoffmann (Eger), József Holovács (Eger), Tibor Juhász (Eger),
László Kovács (Miskolc), Gergely Kovásznai (Eger), László Kozma (Budapest),
Kálmán Liptai (Eger), Florian Luca (Mexico), Giuseppe Mastroianni (Potenza),
Ferenc Mátyás (Eger), Ákos Pintér (Debrecen), Miklós Rontó (Miskolc),
László Szalay (Sopron), János Sztrik (Debrecen), Gary Walsh (Ottawa)



HUNGARIA, EGER

ANNALES MATHEMATICAE ET INFORMATICAE

VOLUME 54. (2021)

EDITORIAL BOARD

Sándor Bácsó (Debrecen), Sonja Gorjanc (Zagreb), Tibor Gyimóthy (Szeged),
Miklós Hoffmann (Eger), József Holovács (Eger), Tibor Juhász (Eger),
László Kovács (Miskolc), Gergely Kovásznai (Eger), László Kozma (Budapest),
Kálmán Liptai (Eger), Florian Luca (Mexico), Giuseppe Mastroianni (Potenza),
Ferenc Mátyás (Eger), Ákos Pintér (Debrecen), Miklós Rontó (Miskolc),
László Szalay (Sopron), János Sztrik (Debrecen), Gary Walsh (Ottawa)

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
Felelős szerkesztő: Dr. Domonkosi Ágnes
Műszaki szerkesztő: Dr. Tómacs Tibor
Megjelent: 2021. december

Research papers

A short remark on Horadam identities with binomial coefficients

Robert Frontczak*

Landesbank Baden-Württemberg (LBBW), Stuttgart, Germany
robert.frontczak@lbbw.de

Submitted: February 17, 2021

Accepted: March 17, 2021

Published online: April 27, 2021

Abstract

In this note, we introduce a very simple approach to obtain Horadam identities with binomial coefficients including an additional parameter. Many known Fibonacci identities (as well as polynomial identities) will follow immediately as special cases.

Keywords: Horadam number, Fibonacci number, binomial transform

AMS Subject Classification: 11B37, 11B39

1. Introduction and motivation

Layman [15] recalled the Fibonacci identities

$$F_{2n} = \sum_{k=0}^n \binom{n}{k} F_k, \quad 2^n F_n = \sum_{k=0}^n \binom{n}{k} F_{2k}, \quad 3^n F_n = \sum_{k=0}^n \binom{n}{k} F_{4k},$$

and attributed them to Hoggatt [9]. Here, as usual, F_n is the n th Fibonacci number, defined by $F_0 = 0$, $F_1 = 1$ and $F_{n+2} = F_{n+1} + F_n$, $n \geq 0$. Layman proved more such identities, in particular, the following alternating sums:

$$(-1)^n F_{3n} = \sum_{k=0}^n \binom{n}{k} (-2)^k F_{2k}, \quad (-5)^n F_{3n} = \sum_{k=0}^n \binom{n}{k} (-2)^k F_{5k},$$

*Statements and conclusions made in this article are entirely those of the author. They do not necessarily reflect the views of LBBW.

and

$$(-4)^n F_{3n} = \sum_{k=0}^n \binom{n}{k} (-1)^k F_{6k}.$$

Several additional sums of this kind and generalizations were derived by Carlitz and Ferns [5], Carlitz [4], Haukkanen [7, 8] and Prodinger [16]. More recently, some authors also worked on generalizations and derived expressions for sums with weighted binomial sums, sums with polynomials and sums where only half of the binomial coefficients are used. We refer to [2] and [10–14]. Adegoke [1] generalized many of the above results and derived summation identities involving Horadam numbers and binomial coefficients, some of which we will encounter below.

In this note, we give another type of generalization of some Horadam binomial sums. More precisely, we introduce a very simple approach to obtain Horadam identities with an additional parameter. All results are derived completely routinely using standard methods.

Let $w_n = w_n(a, b; p, q)$ be a general Horadam sequence, i.e., a second order recurrence

$$w_n = pw_{n-1} - qw_{n-2}, \quad n \geq 2,$$

with nonzero constant p, q and initial values $w_0 = a, w_1 = b$. We mention the following instances: $w_n(0, 1; 1, -1) = F_n$ is the Fibonacci sequence, $w_n(0, 1; 2, -1) = P_n$ is the Pell sequence, $w_n(0, 1; 1, -2) = J_n$ is the Jacobsthal sequence, $w_n(0, 1; 3, 2) = M_n$ is the Mersenne sequence, $w_n(0, 1; 6, 1) = B_n$ is the balancing number sequence, $w_n(2, 1; 1, -1) = L_n$ is the Lucas sequence, $w_n(2, 2; 2, -1) = Q_n$ is the Pell-Lucas sequence, $w_n(2, 1; 1, -2) = j_n$ is the Jacobsthal-Lucas sequence, and $w_n(1, 3; 6, 1) = C_n$ is Lucas-balancing number sequence. All sequences are listed in OEIS [17] where additional information and references are available. We also note that the sequence w_n also contains important sequences of polynomials: $w_n(0, 1; x, -1) = F_n(x)$ are the Fibonacci polynomials, $w_n(0, 1; 2x, -1) = P_n(x)$ are the Pell polynomials, $w_n(0, 1; 1, -2x) = J_n(x)$ are the Jacobsthal polynomials, and $w_n(0, 1; 6x, 1) = B_n(x)$ are the balancing polynomials, respectively.

The Binet formula of w_n in the non-degenerated case, $p^2 - 4q > 0$, is

$$w_n = A\alpha^n + B\beta^n,$$

with

$$A = \frac{b - a\beta}{\alpha - \beta}, \quad B = \frac{a\alpha - b}{\alpha - \beta},$$

and where α and β are roots of the equation $x^2 - px + q = 0$, that is

$$\alpha = \frac{p + \sqrt{p^2 - 4q}}{2}, \quad \beta = \frac{p - \sqrt{p^2 - 4q}}{2}.$$

In what follows we will need the following expressions:

$$\alpha + \beta = p, \quad \alpha\beta = q, \quad \alpha - \beta = \sqrt{p^2 - 4q},$$

as well as

$$\alpha^2 = p\alpha - q, \quad \beta^2 = p\beta - q, \quad (1.1)$$

$$\alpha^3 = (p^2 - q)\alpha - pq, \quad \beta^3 = (p^2 - q)\beta - pq, \quad (1.2)$$

$$\alpha^4 = (p^3 - 2pq)\alpha - q(p^2 - q), \quad \beta^4 = (p^3 - 2pq)\beta - q(p^2 - q),$$

and so on.

Finally, we mention the standard fact about sequences and their binomial transforms [3]: Let $(a_n)_{n \geq 0}$ be a sequence of numbers and $(b_n)_{n \geq 0}$ be its binomial transform. Then, we have the following relations:

$$b_n = \sum_{k=0}^n \binom{n}{k} a_k \quad \Leftrightarrow \quad a_n = \sum_{k=0}^n \binom{n}{k} (-1)^{n-k} b_k.$$

2. A simple generalization

The next lemma will be the key ingredient to derive our results.

Lemma 2.1. *Let n and j be integers with $0 \leq j \leq n$. Then, for each $a, x \in \mathbb{C}$ we have the identity*

$$\binom{n}{j} x^j (a \pm x)^{n-j} = \sum_{k=j}^n \binom{k}{j} \binom{n}{k} (\pm 1)^{k-j} x^k a^{n-k}.$$

Proof. From the binomial theorem we have

$$\begin{aligned} \binom{n}{j} x^j (a \pm x)^{n-j} &= \binom{n}{j} \sum_{m=0}^{n-j} \binom{n-j}{m} (\pm 1)^m x^{m+j} a^{n-(j+m)} \\ &= \binom{n}{j} \sum_{k=j}^n \binom{n-j}{k-j} (\pm 1)^{k-j} x^k a^{n-k} \\ &= \sum_{k=j}^n \binom{k}{j} \binom{n}{k} (\pm 1)^{k-j} x^k a^{n-k}, \end{aligned}$$

where in the last step we have used the identity

$$\binom{n}{j} \binom{n-j}{k-j} = \binom{n}{k} \binom{k}{j}, \quad 0 \leq j \leq k \leq n. \quad \square$$

Example 2.2. Setting $(x; a) = (\alpha^2; -q)$, $(x; a) = (\beta^2; -q)$, using (1.1) and the linearity of the Binet form gives the following identity valid for all $0 \leq j \leq n$

$$\binom{n}{j} p^{n-j} w_{n+j+m} = \sum_{k=j}^n \binom{k}{j} \binom{n}{k} q^{n-k} w_{2k+m}, \quad m \geq 0.$$

The case $j = 0$ and the corresponding inverse binomial transform produce immediately

$$\left(\frac{p}{q}\right)^n w_{n+m} = \sum_{k=0}^n \binom{n}{k} q^{-k} w_{2k+m}$$

as well as

$$q^{-n} w_{2n+m} = \sum_{k=0}^n \binom{n}{k} (-1)^{n-k} \left(\frac{p}{q}\right)^k w_{k+m}.$$

Obviously, with $w_n = F_n$ (or L_n) we recover the classical results, which appeared in [5] and [15]. The balancing number counterparts were stated in [6].

Example 2.3. If we set $\Delta = p^2 - 4q$, then a simple computation shows that

$$\alpha^2 - q = \sqrt{\Delta}\alpha \quad \text{and} \quad \beta^2 - q = -\sqrt{\Delta}\alpha.$$

Thus, with $(x; a) = (\alpha^2; -q)$, $(x; a) = (\beta^2; -q)$ and using again (1.1), we see that if n and j have the same parity, then for all $0 \leq j \leq n$

$$\binom{n}{j} \Delta^{(n-j)/2} w_{n+j+m} = \sum_{k=j}^n \binom{k}{j} \binom{n}{k} (-q)^{n-k} w_{2k+m}, \quad m \geq 0.$$

Especially, for $j = 0$ and n even we get

$$q^{-n} \Delta^{n/2} w_{n+m} = \sum_{k=0}^n \binom{n}{k} (-q)^{-k} w_{2k+m}$$

and for $j = 1$ and n odd

$$-q^{-n} \Delta^{(n-1)/2} n w_{n+1+m} = \sum_{k=1}^n \binom{n}{k} k (-q)^{-k} w_{2k+m}.$$

If n and j are of unequal parity (n odd and j even, for instance), then

$$\binom{n}{j} \Delta^{(n-1-j)/2} v_{n+j+m} = \sum_{k=j}^n \binom{k}{j} \binom{n}{k} (-q)^{n-k} u_{2k+m}, \quad m \geq 0,$$

and

$$\binom{n}{j} \Delta^{(n+1-j)/2} u_{n+j+m} = \sum_{k=j}^n \binom{k}{j} \binom{n}{k} (-q)^{n-k} v_{2k+m}, \quad m \geq 0,$$

with $u_n = w_n(0, 1; p, q)$ and $v_n = w_n(2, p; p, q)$.

Example 2.4. Setting $(x; a) = (\alpha^3; -pq)$, $(x; a) = (\beta^3; -pq)$, using (1.2) yields for all $0 \leq j \leq n$

$$\binom{n}{j} (p^2 - q)^{n-j} w_{n+2j+m} = \sum_{k=j}^n \binom{k}{j} \binom{n}{k} (pq)^{n-k} w_{3k+m}, \quad m \geq 0.$$

The case $j = 0$ in combination with the binomial transform produce

$$\left(\frac{p^2 - q}{pq}\right)^n w_{n+m} = \sum_{k=0}^n \binom{n}{k} (pq)^{-k} w_{3k+m}$$

as well as

$$(pq)^{-n} w_{3n+m} = \sum_{k=0}^n \binom{n}{k} (-1)^{n-k} \left(\frac{p^2 - q}{pq}\right)^k w_{k+m}.$$

Example 2.5. Combining the values $(x; a) = (p\alpha^3; q^2)$ and $(x; a) = (p\beta^3; q^2)$ with

$$p\alpha^3 + q^2 = (p^2 - q)\alpha^2, \quad \text{and} \quad p\beta^3 + q^2 = (p^2 - q)\beta^2,$$

Lemma 2.1 gives

$$\binom{n}{j} p^j (p^2 - q)^{n-j} w_{2n+j+m} = \sum_{k=j}^n \binom{k}{j} \binom{n}{k} p^k q^{2n-2k} w_{3k+m}, \quad m \geq 0.$$

Again, from the case $j = 0$ and the binomial transform we get

$$\left(\frac{p^2 - q}{q^2}\right)^n w_{2n+m} = \sum_{k=0}^n \binom{n}{k} \left(\frac{p}{q^2}\right)^k w_{3k+m}$$

as well as

$$\left(\frac{p}{q^2}\right)^n w_{3n+m} = \sum_{k=0}^n \binom{n}{k} (-1)^{n-k} \left(\frac{p^2 - q}{q^2}\right)^k w_{2k+m}.$$

Example 2.6. In this example we combine the values $(x; a) = (\alpha^4; q(p^2 - q))$ and $(x; a) = (\beta^4; q(p^2 - q))$ to get

$$\binom{n}{j} (p(p^2 - 2q))^{n-j} w_{n+3j+m} = \sum_{k=j}^n \binom{k}{j} \binom{n}{k} (q(p^2 - q))^{n-k} w_{4k+m}, \quad m \geq 0.$$

Hence,

$$\left(\frac{p(p^2 - 2q)}{q(p^2 - q)}\right)^n w_{n+m} = \sum_{k=0}^n \binom{n}{k} (q(p^2 - q))^{-k} w_{4k+m}$$

as well as

$$(q(p^2 - q))^{-n} w_{4n+m} = \sum_{k=0}^n \binom{n}{k} (-1)^{n-k} \left(\frac{p(p^2 - 2q)}{q(p^2 - q)}\right)^k w_{k+m}.$$

Example 2.7. An application of Lemma 2.1 with $(x; a) = (\alpha^4; q^2)$ and $(x; a) = (\beta^4; q^2)$ and noting that

$$\alpha^4 + q^2 = (p^2 - 2q)\alpha^2 \quad \text{and} \quad \beta^4 + q^2 = (p^2 - 2q)\beta^2,$$

proves the next identity:

$$\binom{n}{j} (p^2 - 2q)^{n-j} w_{2n+2j+m} = \sum_{k=j}^n \binom{k}{j} \binom{n}{k} q^{2n-2k} w_{4k+m}, \quad m \geq 0.$$

The case $j = 0$ in conjunction with the binomial transform yield

$$\left(\frac{p^2 - 2q}{q^2}\right)^n w_{2n+m} = \sum_{k=0}^n \binom{n}{k} q^{-2k} w_{4k+m}$$

as well as

$$q^{-2n} w_{4n+m} = \sum_{k=0}^n \binom{n}{k} (-1)^{n-k} \left(\frac{p^2 - 2q}{q^2}\right)^k w_{2k+m}.$$

3. Slightly more general identities

Lemma 3.1. *For each $n \geq 1$ we have the relations*

$$\alpha^n = \alpha u_n - q u_{n-1} \quad \text{and} \quad \beta^n = \beta u_n - q u_{n-1}$$

with $u_n = w_n(0, 1; p, q)$.

Proof. We can prove the statements by induction on n . Since, $\alpha^1 = \alpha u_1 - q u_0$, the inductive step is

$$\begin{aligned} \alpha^{n+1} &= \alpha \alpha^n \\ &= \alpha^2 u_n - q \alpha u_{n-1} \\ &= (\alpha u_2 - q u_1) u_n - q \alpha u_{n-1} \\ &= \alpha (p u_n - q u_{n-1}) - q u_n \quad (u_2 = p) \\ &= \alpha u_{n+1} - q u_n. \end{aligned}$$

The proof of the second statement is a copy of the first proof. \square

The next identity is stated as a proposition.

Proposition 3.2. *For integers $m \geq 2$, $r \geq 0$ and $0 \leq j \leq n$ it is true that*

$$\binom{n}{j} u_{m-1}^{-n} u_m^{n-j} w_{j(m-1)+n+r} = \sum_{k=j}^n \binom{k}{j} \binom{n}{k} u_{m-1}^{-k} q^{n-k} w_{mk+r}.$$

Proof. From Lemma 3.1 we see that

$$q + \frac{\alpha^n}{u_{n-1}} = \alpha \frac{u_n}{u_{n-1}} \quad \text{and} \quad q + \frac{\beta^n}{u_{n-1}} = \beta \frac{u_n}{u_{n-1}}.$$

Using Lemma 2.1 with $a = q$ and $x = \frac{\alpha^n}{u_{n-1}}$ and $x = \frac{\beta^n}{u_{n-1}}$ completes the proof. \square

The next two sum identities follow immediately:

$$\left(\frac{u_m}{qu_{m-1}}\right)^n w_{n+r} = \sum_{k=0}^n \binom{n}{k} (qu_{m-1})^{-k} w_{mk+r}$$

as well as

$$(qu_{m-1})^{-n} w_{mn+r} = \sum_{k=0}^n \binom{n}{k} (-1)^{n-k} \left(\frac{u_m}{qu_{m-1}}\right)^k w_{k+r}.$$

We also mention the formula for $j = 1$:

$$nu_{m-1}^{-n} u_m^{n-1} w_{n+m+r-1} = \sum_{k=1}^n \binom{n}{k} k u_{m-1}^{-k} q^{n-k} w_{mk+r}.$$

Lemma 3.3. For each $k, n \geq 1$ we have the relations

$$\alpha^{kn} = \frac{u_{kn}}{u_n} \alpha^n - q^n \frac{u^{(k-1)n}}{u_n} \quad \text{and} \quad \beta^{kn} = \frac{u_{kn}}{u_n} \beta^n - q^n \frac{u^{(k-1)n}}{u_n}$$

with $u_n = w_n(0, 1; p, q)$.

Proof. Both statements can be verified directly by computation working with $u_n \alpha^{kn}$ (respectively $u_n \beta^{kn}$) and $q = \alpha\beta$. \square

Proposition 3.4. For integers $m \geq 2$, $s \geq 1$, $r \geq 0$, and $0 \leq j \leq n$ we have the identity

$$\begin{aligned} & \binom{n}{j} \left(\frac{u_s}{u_{ms}}\right)^j \left(\frac{u_{ms}}{u_{(m-1)s}}\right)^n q^{-sn} w_{sn+sj(m-1)+r} \\ &= \sum_{k=j}^n \binom{k}{j} \binom{n}{k} q^{-sk} \left(\frac{u_s}{u_{(m-1)s}}\right)^k w_{msk+r}. \end{aligned}$$

Proof. The identity follows upon combining Lemma 2.1 with Lemma 3.3 with $a = q^s u_{(m-1)s}/u_s$ and $x = \alpha^{ms}$ and $x = \beta^{ms}$, respectively. \square

The special identities for $j = 0$ are

$$\left(\frac{u_{ms}}{u_{(m-1)s}}\right)^n q^{-sn} w_{sn+r} = \sum_{k=0}^n \binom{n}{k} q^{-sk} \left(\frac{u_s}{u_{(m-1)s}}\right)^k w_{msk+r}$$

and

$$\left(\frac{u_s}{u_{(m-1)s}}\right)^n q^{-sn} w_{msn+r} = \sum_{k=0}^n \binom{n}{k} (-1)^{n-k} q^{-sk} \left(\frac{u_{ms}}{u_{(m-1)s}}\right)^k w_{sk+r}.$$

Corollary 3.5. For integers $s \geq 1$, $r \geq 0$ and $0 \leq j \leq n$ we have the identity

$$\binom{n}{j} v_s^{n-j} q^{-sn} w_{s(n+j)+r} = \sum_{k=j}^n \binom{k}{j} \binom{n}{k} q^{-sk} w_{2sk+r}.$$

In particular,

$$(-1)^n q^{-sn} w_{2sn+r} = \sum_{k=0}^n \binom{n}{k} (-1)^k q^{-sk} v_s^k w_{sk+r}$$

and

$$q^{-sn} v_s^n w_{sn+r} = \sum_{k=0}^n \binom{n}{k} q^{-sk} w_{2sk+r}.$$

Proof. Set $m = 2$ in Proposition 3.4 and use $u_{2n}/u_n = v_n$. □

Some more examples could be stated, but we stop here, as the principle is clear.

Acknowledgements. The author thanks Kunle Adegoke for discussions. He is also grateful to the anonymous referee for a rapid review and for suggesting important references that are directly linked to this research project.

References

- [1] K. ADEGOKE: *A master identity for Horadam numbers*, Preprint (2019), URL: <https://arxiv.org/abs/1903.11057>.
- [2] K. ADEGOKE: *Weighted sums of some second-order sequences*, *Fibonacci Quart.* 56.3 (2018), pp. 252–262.
- [3] K. N. BOYADZHIEV: *Notes on the Binomial Transform: Theory and Table with Appendix on Stirling Transform*, World Scientific, 2018.
- [4] L. CARLITZ: *Some classes of Fibonacci sums*, *Fibonacci Quart.* 16.5 (1978), pp. 411–425.
- [5] L. CARLITZ, H. H. FERNS: *Some Fibonacci and Lucas identities*, *Fibonacci Quart.* 8.1 (1970), pp. 61–73.
- [6] R. FRONTCAK: *Sums of balancing and Lucas-balancing numbers with binomial coefficients*, *Int. J. Math. Anal.* 12.12 (2018), pp. 585–594.
- [7] P. HAUKKANEN: *Formal power series for binomial sums of sequences of numbers*, *Fibonacci Quart.* 31.1 (1993), pp. 28–31.
- [8] P. HAUKKANEN: *On a binomial sum for the Fibonacci and related numbers*, *Fibonacci Quart.* 34.4 (1996), pp. 326–331.
- [9] V. E. HOGGATT, JR.: *Some special Fibonacci and Lucas generating functions*, *Fibonacci Quart.* 9.2 (1971), pp. 121–133.
- [10] E. KILIC, N. IRMAK: *Binomial identities involving the generalized Fibonacci type polynomials*, *Ars Combin.* 98 (2011), pp. 129–134.
- [11] E. KILIÇ: *Some classes of alternating weighted binomial sums*, *An. Ştiinţ. Univ. Al. I. Cuza Iaşi. Mat.* 3.2 (2016), pp. 835–843.

- [12] E. KILIÇ, E. J. IONASCU: *Certain binomial sums with recursive coefficients*, Fibonacci Quart. 48.2 (2010), pp. 161–167.
- [13] E. KILIÇ, N. ÖMÜR, Y. T. ULUTAŞ: *Binomial sums whose coefficients are products of terms of binary sequences*, Util. Math. 84 (2011), pp. 45–52.
- [14] E. KILIÇ, E. TAN: *On binomial sums for the general second order linear recurrence*, Integers 10 (2010), pp. 801–806,
DOI: <https://doi.org/10.1515/INTEGER.2010.057>.
- [15] J. W. LAYMAN: *Certain general binomial-Fibonacci sums*, Fibonacci Quart. 15.3 (1977), pp. 362–366.
- [16] H. PRODINGER: *Some information about the binomial transform*, Fibonacci Quart. 32.5 (1994), pp. 412–415.
- [17] N. J. A. SLOANE: *The On-Line Encyclopedia of Integer Sequences*, Published electronically at <https://oeis.org>.

Corrigendum to
“Pentagonal and heptagonal repdigits”
[Annales Mathematicae et Informaticae
52 (2020) 137–145]

Bir Kafle^a, Florian Luca^b, Alain Togbé^a

^aDepartment of Mathematics and Statistics
Purdue University Northwest
1401 S. U.S. 421, Westville, IN 46391 USA
bkafle@pnw.edu
atogbe@pnw.edu

^bSchool of Mathematics
University of the Witwatersrand
Private Bag X3, Wits 2050, South Africa
florian.luca@wits.ac.za

Submitted: December 28, 2020

Accepted: March 17, 2021

Published online: March 29, 2021

Abstract

Our original paper [1], contains some typos that we would like to fix here. These typos do not affect the final results that we obtained.

Keywords: Pentagonal numbers, heptagonal numbers, repdigits.

AMS Subject Classification: 11A25, 11B39, 11J86

In the proof of Theorem 2.1, we should have multiplied equation (2.2) by $16A^2\ell^2 10^{2r}$ instead of $16\ell^2 10^{2r}$. This gives us

$$Y^2 = X^3 + \bar{A}, \tag{1}$$

where

$$X := 4A\ell 10^{m_1+r}, Y := 12A\ell 10^r(2An + B),$$

and

$$\bar{A} := 16A^2\ell^2 10^{2r} (9(B^2 - 4AC) - 4A\ell).$$

The second typo is that equation (2.6) should have been

$$\ell \left(\frac{10^m - 1}{9} \right) = \frac{n(5n - 3)}{2}. \quad (2)$$

The last typo is that a_3 should have been

$$a_3 := 11979\ell^2 10^{4r} (99 - 24\ell).$$

Except the above typos, all the proofs and computations are correct.

Acknowledgements. We thank Dr. Eric F. Bravo for pointing out to us the typos in our paper.

References

- [1] F. LUCA, B. KAFLE, A. TOGBÉ: *Pentagonal and heptagonal repdigits*, *Annales Mathematicae et Informaticae* 52 (2020), pp. 137–145,
DOI: <https://doi.org/10.33039/ami.2020.09.002>.

Warshall’s algorithm—survey and applications

Zoltán Kása

Sapientia Hungarian University of Transylvania, Cluj-Napoca, Romania

Department of Mathematics and Informatics, Târgu Mureş

kasa@ms.sapientia.ro

Submitted: December 23, 2020

Accepted: August 3, 2021

Published online: August 13, 2021

Abstract

This survey presents the well-known Warshall’s algorithm, a generalization and some interesting applications: transitive closure of relations, distances between vertices in graphs, number of paths in acyclic digraphs, all paths in digraphs, scattered complexity for rainbow words, special walks in finite automata.

Keywords: Warshall’s algorithm, Floyd–Warshall algorithm, paths in graphs, scattered subword complexity, finite automata

AMS Subject Classification: 05C85, 68W05, 68R10, 68R14

1. Introduction

Warshall’s algorithm [14] with its generalization [11] is widely used in graph theory [1, 3–5, 8–10, 12] and in various fields of sciences (e.g. fuzzy [13] and quantum [6] theory, Kleene algebra [7]). In the following, we present the algorithm, its generalization, and the collected applications related to graphs, to be easily accessible together here.

Let R be a binary relation on the set $S = \{s_1, s_2, \dots, s_n\}$, we write $s_i R s_j$ if s_i is in relation with s_j . The relation R can be represented by the so called *relation matrix*, which is

$$A = (a_{ij})_{\substack{i=1, \dots, n \\ j=1, \dots, n}}, \quad \text{where} \quad a_{ij} = \begin{cases} 1, & \text{if } s_i R s_j, \\ 0, & \text{otherwise.} \end{cases}$$

The transitive closure of the relation R is the binary relation R^* defined as: $s_i R^* s_j$ if and only if there exists $s_{p_1}, s_{p_2}, \dots, s_{p_r}, r \geq 2$ such that $s_i = s_{p_1}, s_{p_1} R s_{p_2}, s_{p_2} R s_{p_3}, \dots, s_{p_{r-1}} R s_{p_r}, s_{p_r} = s_j$. The relation matrix of R^* is $A^* = (a_{ij}^*)$.

Let us define the following two operations: i) if $a, b \in \{0, 1\}$ then $a + b = 0$ for $a = 0, b = 0$, and $a + b = 1$ otherwise; ii) $a \cdot b = 1$ for $a = 1, b = 1$, and $a \cdot b = 0$ otherwise. In this case

$$A^* = A + A^2 + \dots + A^n.$$

The transitive closure of a relation can be computed easily by the Warshall's algorithm [2, 14]:

WARSHALL(A, n)

Input: the relation matrix A ; the number of elements n

Output: $W = A^*$

```

1   $W \leftarrow A$ 
2  for  $k \leftarrow 1$  to  $n$ 
3      do for  $i \leftarrow 1$  to  $n$ 
4          do for  $j \leftarrow 1$  to  $n$ 
5              do if  $w_{ik} = 1$  and  $w_{kj} = 1$ 
6                  then  $w_{ij} \leftarrow 1$ 
7  return  $W$ 
```

Listing 1. Warshall's algorithm.

The complexity of this algorithm is $\Theta(n^3)$.

A binary relation can be represented by a directed graph (i.e. digraph) too. The relation matrix is equal to the adjacency matrix of the corresponding graph. See Fig. 1 for an example. Fig. 2 represents the graph of the corresponding transitive closure relation.

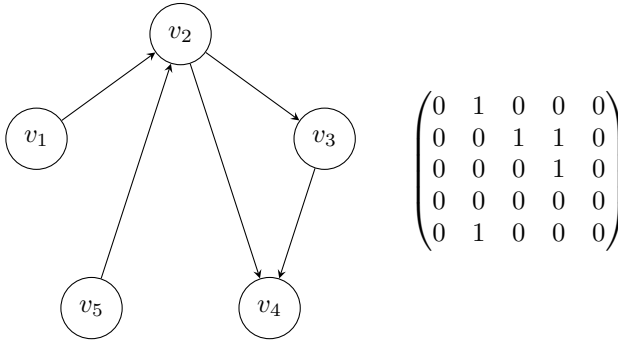


Figure 1. A binary relation represented by a graph with the corresponding adjacency matrix.

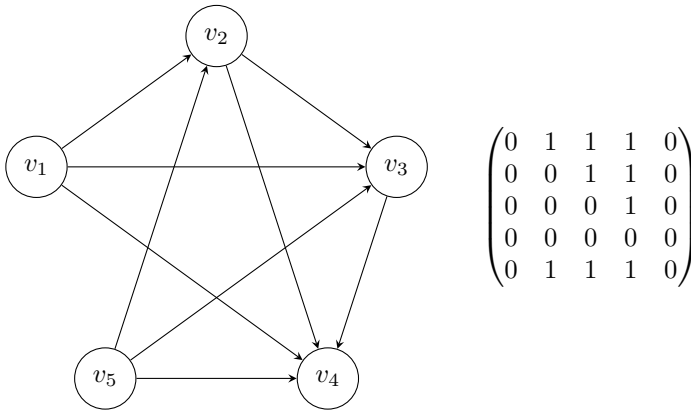


Figure 2. The transitive closure of the relation in Fig. 1.

2. Generalization of Warshall's algorithm

Lines 5 and 6 in the Warshall's algorithm presented in Listing 1 can be expressed as

$$w_{ij} \leftarrow w_{ij} + w_{ik} \cdot w_{kj}$$

using the operations defined above. If instead of the operations $+$ and \cdot we use two operations \oplus and \odot from a semiring, a generalized Warshall's algorithm results [11]:

GENERALIZED-WARSHALL(A, n)

Input: the relation matrix A ; the number of elements n

Output: $W = A^*$

```

1   $W \leftarrow A$ 
2  for  $k \leftarrow 1$  to  $n$ 
3      do for  $i \leftarrow 1$  to  $n$ 
4          do for  $j \leftarrow 1$  to  $n$ 
5              do  $w_{ij} \leftarrow w_{ij} \oplus (w_{ik} \odot w_{kj})$ 
6  return  $W$ 
```

Listing 2. The generalized Warshall's algorithm.

The complexity of this algorithm is also $\Theta(n^3)$. This generalization leads us to a number of interesting applications.

3. Applications

3.1. Distances between vertices. Floyd–Warshall algorithm

Given a (di)graph with positive or negative edge weights (but with no negative cycles) and its modified adjacency matrix $D_0 = (d_{ij}^0)$, we can obtain the distance matrix $D = (d_{ij})$ in which d_{ij} represents the distance between vertices v_i and v_j . The distance between vertices v_i and v_j is the length of the shortest path between them. The modified adjacency matrix $D_0 = (d_{ij}^0)$ is the following:

$$d_{ij}^0 = \begin{cases} 0, & \text{if } i = j, \\ \infty, & \text{if there is no edge from vertex } v_i \text{ to vertex } v_j, i \neq j, \\ w_{ij}, & \text{the weight of the edge from } v_i \text{ to } v_j, i \neq j. \end{cases}$$

Choosing for \oplus the min operation (minimum of two real numbers), and for \odot the real addition (+), we obtain the well-known Floyd–Warshall algorithm as a special case of the generalized Warshall’s algorithm [5, 11, 12] :

FLOYD-WARSHALL(D_0, n)

Input: the adjacency matrix D_0 ; the number of elements n

Output: the distance matrix D

```

1   $D \leftarrow D_0$ 
2  for  $k \leftarrow 1$  to  $n$ 
3      do for  $i \leftarrow 1$  to  $n$ 
4          do for  $j \leftarrow 1$  to  $n$ 
5              do  $d_{ij} \leftarrow \min\{d_{ij}, d_{ik} + d_{kj}\}$ 
6  return  $D$ 
```

Listing 3. Floyd-Warshall algorithm.

An example is presented in Fig. 3. The shortest paths can also be easily obtained by storing the previous vertex v_k on the path, in line 5 of Listing 3. In the case of acyclic digraphs, the algorithm can be easily modified to obtain the longest distances between vertices and consequently the longest paths.

3.2. Number of paths in acyclic digraphs

Here, by path we understand a directed path. In an acyclic digraph the following algorithm counts the number of paths between vertices [4, 9]. The operation \oplus , \odot are the classical add and multiply operations for real numbers and let w_{ij} denote the number of paths from vertex v_i to vertex v_j .

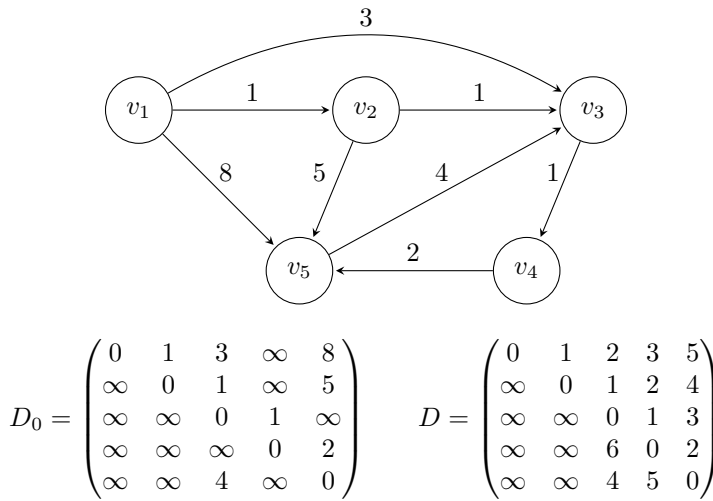


Figure 3. A weighted digraph with the corresponding matrices.

WARSHALL-PATHS(A, n)

Input: the adjacency matrix A ; the number of elements n

Output: W with number of paths between vertices

```

1   $W \leftarrow A$ 
2  for  $k \leftarrow 1$  to  $n$ 
3      do for  $i \leftarrow 1$  to  $n$ 
4          do for  $j \leftarrow 1$  to  $n$ 
5              do  $w_{ij} \leftarrow w_{ij} + w_{ik} \cdot w_{kj}$ 
6  return  $W$ 
    
```

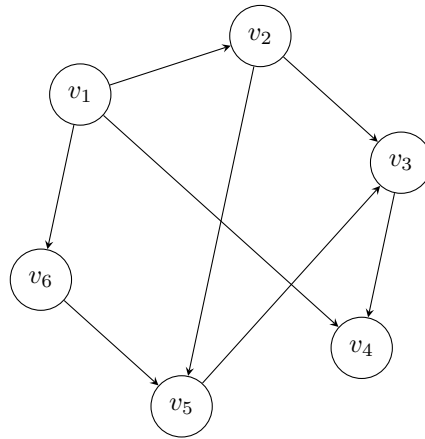
Listing 4. Finding the number of paths between vertices.

The proof is omitted, as it is very similar to the one given in [2] for the original Warshall's algorithm.

An example can be seen in Fig. 4. For example between vertices v_1 and v_3 there are 3 paths: (v_1, v_2, v_3) ; (v_1, v_2, v_5, v_3) and (v_1, v_6, v_5, v_3) .

For the case when the arcs of the graph are colored, we may be interested in the number of monochromatic paths. The generalized algorithm can also be used for monochromatic subgraphs. The following novel algorithm (first described here) solves the problem for all colors at once.

In the adjacency (color) matrix, a_{ij} is equal to the code of the color of the arc (v_i, v_j) , and is equal to 0, if there is no arc from v_i to v_j . In the three-dimensional result matrix W the element w_{ijp} represents the number of the paths from v_i to v_j with p -colored arcs each.



$$A = \begin{pmatrix} 0 & 1 & 0 & 1 & 0 & 1 \\ 0 & 0 & 1 & 0 & 1 & 0 \\ 0 & 0 & 0 & 1 & 0 & 0 \\ 0 & 0 & 0 & 0 & 0 & 0 \\ 0 & 0 & 1 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & 1 & 0 \end{pmatrix} \quad W = \begin{pmatrix} 0 & 1 & 3 & 4 & 2 & 1 \\ 0 & 0 & 2 & 2 & 1 & 0 \\ 0 & 0 & 0 & 1 & 0 & 0 \\ 0 & 0 & 0 & 0 & 0 & 0 \\ 0 & 0 & 1 & 1 & 0 & 0 \\ 0 & 0 & 1 & 1 & 1 & 0 \end{pmatrix}$$

Figure 4. An acyclic digraph and the corresponding matrices.

WARSHALL-MONOCROMATIC-PATHS(A, n, c)

Input: adjacency color matrix A ; number of elements n ; number of colors c

Output: matrix W with number of monochromatic paths between vertices

```

1 Set all elements of  $W$  equal to 0
2 for  $i \leftarrow 1$  to  $n$ 
3   do for  $j \leftarrow 1$  to  $n$ 
4     do if  $a_{ij} \neq 0$ 
5       do  $w_{ija_{ij}} \leftarrow 1$ 
6 for  $p \leftarrow 1$  to  $c$ 
7   do for  $k \leftarrow 1$  to  $n$ 
8     do for  $i \leftarrow 1$  to  $n$ 
9       do for  $j \leftarrow 1$  to  $n$ 
10        do  $w_{ijp} \leftarrow w_{ijp} + w_{ikp} \cdot w_{kjp}$ 
11 return  $W$ 
```

Listing 5. Finding the number of monochromatic paths between vertices.

The sides for $p = 1, \dots, c$ of the three-dimensional matrix W contain the result for the different colors (see Fig. 5).

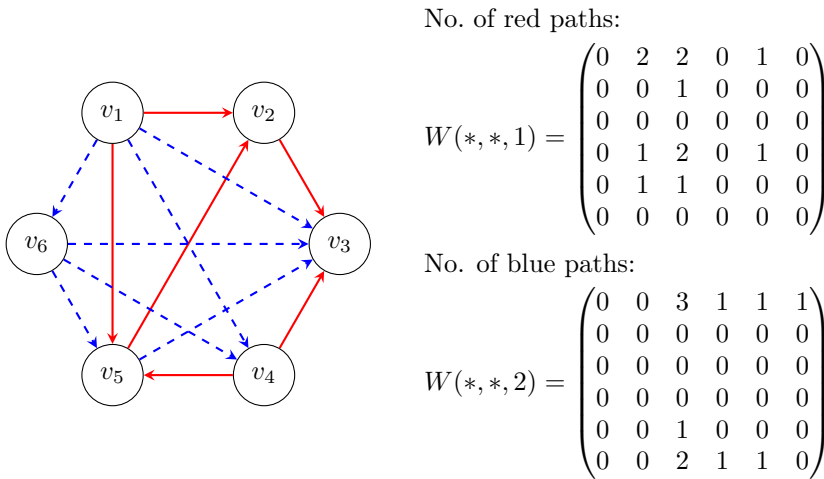


Figure 5. An example of a colored digraph with the two sides of the solution matrix.

3.3. All paths in digraphs

The Warshall's algorithm combined with the Latin square method can be used to obtain all paths in a (not necessarily acyclic) digraph [9]. A path will be denoted by a string formed by its vertices in their natural order in the path.

Let us consider a matrix \mathcal{A} with the elements A_{ij} which are a set of strings. Initially, the elements of this matrix are defined as:

$$A_{ij} = \begin{cases} \{v_i v_j\}, & \text{if } \exists \text{ an arc from } v_i \text{ to } v_j, i \neq j, \\ \emptyset, & \text{otherwise,} \end{cases} \quad \text{for } i, j = 1, 2, \dots, n. \quad (3.1)$$

If \mathcal{A} and \mathcal{B} are sets of strings, \mathcal{AB} will be formed by the set of concatenation of each string from \mathcal{A} with each string from \mathcal{B} , if they have no common letters:

$$\mathcal{AB} = \{ab \mid a \in \mathcal{A}, b \in \mathcal{B}, \text{ if } a \text{ and } b \text{ have no common letters}\}. \quad (3.2)$$

If $s = s_1 s_2 \dots s_p$ is a string, let us denote by $'s$ the string obtained from s by eliminating the first character: $'s = s_2 s_3 \dots s_p$. Let us denote by $'A_{ij}$ the set A_{ij} in which we eliminate from each element the first character. In this case $'\mathcal{A}$ is a matrix with elements $'A_{ij}$.

Operations are: set union and set product defined as before.

Starting with the matrix \mathcal{A} defined as before, the algorithm to obtain all paths is the following, in which W_{ij} represents the set of paths from vertex v_i to v_j .

WARSHALL-LATIN(\mathcal{A}, n)

Input: the adjacency matrix \mathcal{A} defined in (3.1); the number of elements n

Output: \mathcal{W} matrix of the paths between vertices

```

1  $\mathcal{W} \leftarrow \mathcal{A}$ 
2 for  $k \leftarrow 1$  to  $n$ 
3   do for  $i \leftarrow 1$  to  $n$ 
4     do for  $j \leftarrow 1$  to  $n$ 
5       do if  $W_{ik} \neq \emptyset$  and  $W_{kj} \neq \emptyset$ 
6         then  $W_{ij} \leftarrow W_{ij} \cup W_{ik} ' W_{kj}$ 
7 return  $\mathcal{W}$ 

```

Listing 6. Algorithm for finding all paths in digraphs.

An example is presented in Fig. 6: here, for example, between vertices v_1 and v_3 there are two paths: v_1v_3 and $v_1v_2v_3$.

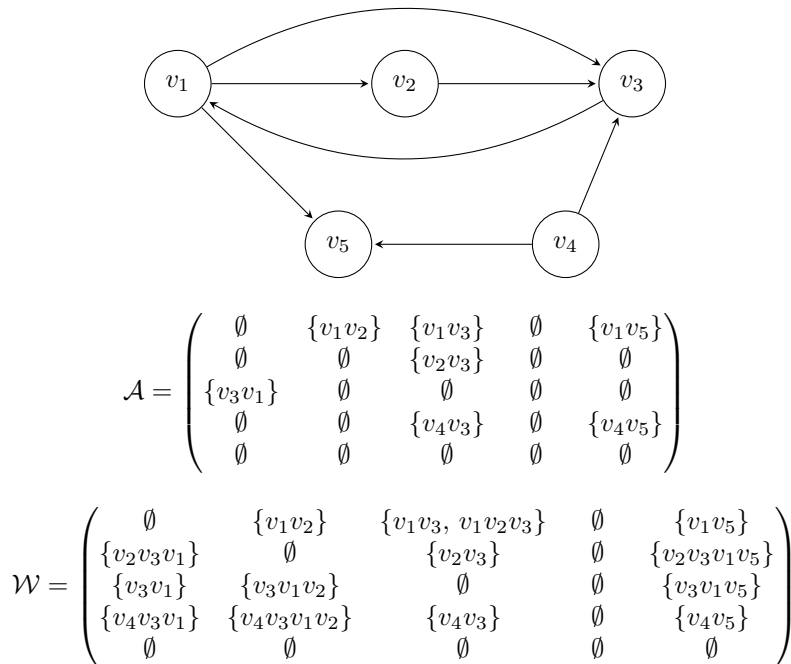


Figure 6. An example of digraph for all paths problem with the corresponding matrices.

Although the algorithm is not polynomial (due to the $W_{ik} ' W_{kj}$ multiplication), the method can be used well in practice for many cases.

3.4. Scattered complexity for rainbow words

The application described in this subsection can be found in [9]. Let Σ be an alphabet, Σ^n the set of all length- n words over Σ , Σ^* the set of all finite word over Σ .

Definition 3.1. Let n and s be positive integers, $M \subseteq \{1, 2, \dots, n - 1\}$ and $u = x_1x_2 \dots x_n \in \Sigma^n$. An M -**subword** of length s of u is defined as $v = x_{i_1}x_{i_2} \dots x_{i_s}$ where

$$\begin{aligned} i_1 &\geq 1, \\ i_{j+1} - i_j &\in M \text{ for } j = 1, 2, \dots, s - 1, \\ i_s &\leq n. \end{aligned}$$

Definition 3.2. The number of M -subwords of a word u for a given set M is the scattered subword complexity, simply M -complexity.

Examples. The word $abcd$ has 11 $\{1, 3\}$ -subwords: $a, ab, abc, abcd, ad, b, bc, bcd, c, cd, d$. The $\{2, 3, 4, 5\}$ -subwords of the word $abcdef$ are the following: $a, ac, ad, ae, af, ace, acf, adf, b, bd, be, bf, bdf, c, ce, cf, d, df, e, f$.

Words with different letters are called *rainbow words*. The M -complexity of a length- n rainbow word does not depend on what letters it contains, and is denoted by $K(n, M)$.

To compute the M -complexity of a rainbow word of length n we will use graph theoretical results. Let us consider the rainbow word $a_1a_2 \dots a_n$ and the corresponding digraph $G = (V, E)$, with

$$\begin{aligned} V &= \{a_1, a_2, \dots, a_n\}, \\ E &= \{(a_i, a_j) \mid j - i \in M, i = 1, 2, \dots, n, j = 1, 2, \dots, n\}. \end{aligned}$$

For $n = 6, M = \{2, 3, 4, 5\}$ see Fig. 7.

The adjacency matrix $A = (a_{ij})_{\substack{i=1, \dots, n \\ j=1, \dots, n}}$ of the graph is defined by:

$$a_{ij} = \begin{cases} 1, & \text{if } j - i \in M, \\ 0, & \text{otherwise,} \end{cases} \quad \text{for } i = 1, 2, \dots, n, j = 1, 2, \dots, n.$$

Because the graph has no directed cycles, the element in row i and column j in A^k (where $A^k = A^{k-1}A$, with $A^1 = A$) will represent the number of length- k directed paths from a_i to a_j . If I is the identity matrix (with elements equal to 1 only on the first diagonal, and 0 otherwise), let us define the matrix $R = (r_{ij})$:

$$R = I + A + A^2 + \dots + A^k, \quad \text{where } k < n, A^{k+1} = O \quad (\text{the null matrix}).$$

The M -complexity of a rainbow word is then

$$K(n, M) = \sum_{i=1}^n \sum_{j=1}^n r_{ij}.$$

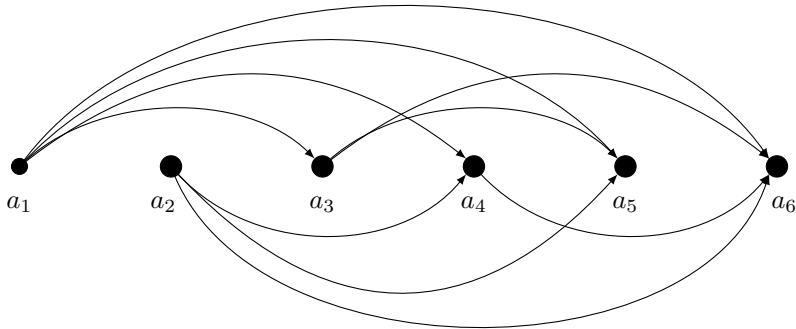


Figure 7. Graph for $\{2, 3, 4, 5\}$ -subwords of the rainbow word of length 6.

Matrix R can be better computed using the WARSHALL-PATHS algorithm described in Listing 4, which gives a matrix W , and therefore $R = I + W$.

For example, let us consider the graph in Fig. 7. [9] The corresponding adjacency matrix is:

$$A = \begin{pmatrix} 0 & 0 & 1 & 1 & 1 & 1 \\ 0 & 0 & 0 & 1 & 1 & 1 \\ 0 & 0 & 0 & 0 & 1 & 1 \\ 0 & 0 & 0 & 0 & 0 & 1 \\ 0 & 0 & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & 0 & 0 \end{pmatrix}.$$

After applying the WARSHALL-PATHS algorithm:

$$W = \begin{pmatrix} 0 & 0 & 1 & 1 & 2 & 3 \\ 0 & 0 & 0 & 1 & 1 & 2 \\ 0 & 0 & 0 & 0 & 1 & 1 \\ 0 & 0 & 0 & 0 & 0 & 1 \\ 0 & 0 & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & 0 & 0 \end{pmatrix}, \quad R = \begin{pmatrix} 1 & 0 & 1 & 1 & 2 & 3 \\ 0 & 1 & 0 & 1 & 1 & 2 \\ 0 & 0 & 1 & 0 & 1 & 1 \\ 0 & 0 & 0 & 1 & 0 & 1 \\ 0 & 0 & 0 & 0 & 1 & 0 \\ 0 & 0 & 0 & 0 & 0 & 1 \end{pmatrix}$$

and then $K(6, \{2, 3, 4, 5\}) = 20$, the sum of elements in R .

Remark 3.3. For this case the WARSHALL-PATHS algorithm can be slightly modified: because of the specific form of the graph the lines 2 and 4 can also be written in the following form:

```
2 for k ← 2 to n - 1
4 do for j ← i + 1 to n
```

□

Using the WARSHALL-LATIN algorithm (Listing 6) we can obtain all nontrivial (with length at least 2) M -subwords of a given length- n rainbow word $a_1 a_2 \cdots a_n$.

3.5. Special walks in finite automata

Let us consider a finite automaton $A = (Q, \Sigma, \delta, \{q_1\}, F)$, where Q is a finite set of states, Σ the input alphabet, $\delta: Q \times \Sigma \rightarrow Q$ the transition function, q_1 the initial state, F the set of final states. In the following, we omit to mark the initial and the final states. The transition function can also be generalized to words: $\delta(q, wa) = \delta(\delta(q, w), a)$, where $q \in Q, a \in \Sigma, w \in \Sigma^*$. A sequence of the form

$$q_1, a_1, q_2, a_2, \dots, a_{n-1}, q_n, \quad n \geq 2,$$

where

$$\delta(q_1, a_1) = q_2, \delta(q_2, a_2) = q_3, \dots, \delta(q_{n-1}, a_{n-1}) = q_n$$

is a walk in the automata labelled by the word $a_1 a_2 \dots a_{n-1}$. This also can be written as:

$$q_1 \xrightarrow{a_1} q_2 \xrightarrow{a_2} q_3 \xrightarrow{a_3} \dots \xrightarrow{a_{n-2}} q_{n-1} \xrightarrow{a_{n-1}} q_n,$$

or shortly: $q_1 \xrightarrow{a_1 a_2 \dots a_{n-1}} q_n$.

We are interested in finding walks with special labels: one-letter power words (power of a single letter) and rainbow words (containing only dissimilar letters).

3.5.1. Walks labeled with one-letter power words

For each pair p, q of states we search for the letters a for which there exists a natural $k \geq 1$ such that we have the transition $\delta(p, a^k) = q$ (see [11]). Let us denote these sets by:

$$W_{ij} = \{a \in \Sigma \mid \exists k \geq 1, \delta(q_i, a^k) = q_j\},$$

where a^k is a length- k one-letter power word.

Here the elements A_{ij} of the adjacency matrix \mathcal{A} initially are defined as:

$$A_{ij} = \{a \mid \delta(q_i, a) = q_j\}, \quad \text{for } i, j = 1, 2, \dots, n.$$

Instead of \oplus we use here set union (\cup) and instead of \odot set intersection (\cap).

WARSHALL-AUTOMATA-1(\mathcal{A}, n)

Input: the adjacency matrix \mathcal{A} ; the number of states n

Output: the matrix \mathcal{W} with sets of letters for one letter power words

```

1   $\mathcal{W} \leftarrow \mathcal{A}$ 
2  for  $k \leftarrow 1$  to  $n$ 
3      do for  $i \leftarrow 1$  to  $n$ 
4          do for  $j \leftarrow 1$  to  $n$ 
5              do  $W_{ij} \leftarrow W_{ij} \cup (W_{ik} \cap W_{kj})$ 
6  return  $\mathcal{W}$ 
```

Listing 7. Finding walks labeled by one-letter power words.

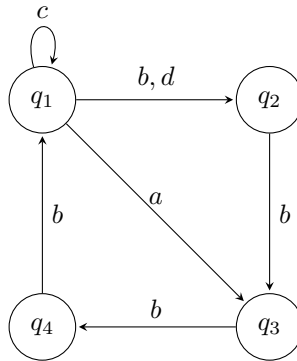


Figure 9. An example of a finite automaton without indicating the initial and final states.

The transition table of the finite automaton in Fig. 9 is:

δ	a	b	c	d
q_1	q_3	q_2	q_1	q_2
q_2	\emptyset	q_3	\emptyset	\emptyset
q_3	\emptyset	q_4	\emptyset	\emptyset
q_4	\emptyset	q_1	\emptyset	\emptyset

Matrices for the graph in Fig. 9 are the following:

$$\mathcal{A} = \begin{pmatrix} \{c\} & \{b, d\} & \{a\} & \emptyset \\ \emptyset & \emptyset & \{b\} & \emptyset \\ \emptyset & \emptyset & \emptyset & \{b\} \\ \{b\} & \emptyset & \emptyset & \emptyset \end{pmatrix}, \quad \mathcal{W} = \begin{pmatrix} \{b, c\} & \{b, d\} & \{a, b\} & \{b\} \\ \{b\} & \{b\} & \{b\} & \{b\} \\ \{b\} & \{b\} & \{b\} & \{b\} \\ \{b\} & \{b\} & \{b\} & \{b\} \end{pmatrix}.$$

For example $\delta(q_2, bb) = q_4$, $\delta(q_2, bbb) = q_1$, $\delta(q_2, bbbb) = q_2$, $\delta(q_1, c^k) = q_1$ for $k \geq 1$.

3.5.2. Walks labeled with rainbow words

To find walks with rainbow labels, we can use the a variant of the WARSHALL-LATIN algorithm (Listing 6), where instead of string of vertices $v_1v_2 \cdots v_k$ we use the corresponding string of labels of the edges $(v_1, v_2), \dots, (v_{k-1}, v_k)$.

Here the elements A_{ij} of the adjacency matrix \mathcal{A} are initially defined as:

$$A_{ij} = \{a \mid \delta(q_i, a) = q_j\}, \quad \text{for } i, j = 1, 2, \dots, n.$$

The concatenation $W_{ik}W_{kj}$ in the following algorithm is defined as in the formula (3.2). Each element of W_{ik} is concatenated with each element of W_{kj} only if these elements (which are strings) have no common letters. If a string appears more than once during concatenation, only one copy is retained. The following algorithm is a new one.

WARSHALL-AUTOMATA-2(\mathcal{A}, n)

Input: the adjacency matrix \mathcal{A} ; the number of states n

Output: the matrix \mathcal{W} of the rainbow words between vertices

```

1  $\mathcal{W} \leftarrow \mathcal{A}$ 
2 for  $k \leftarrow 1$  to  $n$ 
3   do for  $i \leftarrow 1$  to  $n$ 
4     do for  $j \leftarrow 1$  to  $n$ 
5       do if  $W_{ik} \neq \emptyset$  and  $W_{kj} \neq \emptyset$ 
6         then  $W_{ij} \leftarrow W_{ij} \cup W_{ik} W_{kj}$ 
7 return  $\mathcal{W}$ 

```

Listing 8. Finding walks labeled by rainbow words.

For the automaton in Fig. 9 the above algorithm uses the matrix \mathcal{A} :

$$\mathcal{A} = \begin{pmatrix} \{c\} & \{b, d\} & \{a\} & \emptyset \\ \emptyset & \emptyset & \{b\} & \emptyset \\ \emptyset & \emptyset & \emptyset & \{b\} \\ \{b\} & \emptyset & \emptyset & \emptyset \end{pmatrix}$$

and gives the following result:

$$\mathcal{W} = \begin{pmatrix} \{c\} & \{b, d, cb, cd\} & \{a, ca, db, cdb\} & \{ab, cab\} \\ \emptyset & \emptyset & \{b\} & \emptyset \\ \emptyset & \emptyset & \emptyset & \{b\} \\ \{b, bc\} & \{bd, bcd\} & \{ba, bca\} & \emptyset \end{pmatrix}.$$

Conclusions

In 1962 S. Warshall published the algorithm later named after him for computing the transitive closure of a binary relation [14]. R. W. Floyd reported the application of this in the same year to determine the shortest paths in weighted graphs [5]. P. Robert and J. Ferland in their 1968 article [11] gave an interesting generalization that led to the applications discussed in this article [5, 9, 11, 12]. Two algorithms, WARSHALL-MONOCROMATICS-PATHS and WARSHALL-AUTOMATA-2, are new applications firstly described here.

It is amazing how diverse the applications are. And there can be more!

Acknowledgements. The author thanks the anonymous reviewers for their attentive and thorough work in improving the paper with their helpful remarks.

References

- [1] A. AINIA, A. SALEHIPOUR: *Speeding up the Floyd–Warshall algorithm for the cycled shortest path problem*, Applied Mathematics Letters 25.1 (2012), pp. 1–5, DOI: <https://doi.org/10.1016/j.aml.2011.06.008>.
- [2] S. BAASE: *Computer Algorithms: Introduction to Design and Analysis*, Addison-Wesley, 1983, 1988.
- [3] R. BERGHAMMER: *A Functional, Successor List Based Version of Warshall's Algorithm with Applications. Relational and Algebraic Methods in Computer Science. RAMICS, 2011. Lecture Notes in Computer Science, vol 6663*, in: Relational and Algebraic Methods in Computer Science, DOI: https://doi.org/10.1007/978-3-642-21070-9_10.
- [4] C. ELZINGA, H. WANG: *Kernels for acyclic digraphs*, Pattern Recognition Letters 33.16 (2013), pp. 2239–2244, DOI: <https://doi.org/10.1016/j.patrec.2012.07.017>.
- [5] R. W. FLOYD: *Algorithm 97: Shortest Path*, Communications of the ACM 5.6 (1962), p. 345, DOI: <https://doi.org/10.1145/367766.368168>.
- [6] A. S. GUPTA, A. PATHAK: *Quantum Floyd-Warshall algorithm*, arXiv:quant-ph/0502144.
- [7] P. HÖFNER, B. MÖLLER: *Dijkstra, Floyd and Warshall meet Kleene*, Formal Aspect of Computing 24 (2012), pp. 459–476, DOI: <https://doi.org/10.1007/s00165-012-0245-4>.
- [8] S. HOUGARDY: *The Floyd–Warshall algorithm on graphs with negative cycles*, Information Processing Letters 110, pp. 279–281, DOI: <https://doi.org/10.1016/j.ipl.2010.02.001>.
- [9] Z. KÁSA: *On scattered subword complexity*, Acta Univ. Sapientiae Informatica 3.1 (2011), pp. 127–136, URL: acta.sapientia.ro/acta-info/C3-1/info31-6.pdf.
- [10] A. OJO, N. MA, I. WOUNGANG: *Modified Floyd-Warshall algorithm for equal cost multipath in software-defined data center. 2015 IEEE International Conference on Communication Workshop (ICCW), London*, in: pp. 346–351, DOI: <https://doi.org/10.1109/ICCW.2015.7247203>.
- [11] P. ROBERT, J. FERLAND: *Généralisation de l'algorithme de Warshall*, Revue Française d'Informatique et de Recherche Opérationnelle 2.7 (1968), pp. 71–85, URL: www.numdam.org/item/?id=M2AN_1968__2_1_71_0.
- [12] Z. A. VATTAI: *Floyd-Warshall again*, URL: www.ekt.bme.hu/Cikkek/54-Vattai_Floyd-Warshall_Again.pdf.
- [13] Q. WANG, D. ZHANG: *A simple and direct algorithm for computing transitive closure of fuzzy matrix*, Journal of Xi'an University of Technology 3 (2006), URL: en.cnki.com.cn/Article_en/CJFDTOTAL-XALD200603011.htm.
- [14] S. WARSHALL: *A theorem on boolean matrices*, Journal of the ACM 9.1 (1962), pp. 11–12, DOI: <https://doi.org/10.1145/321105.321107>.

Introducing w -Horn and z -Horn: A generalization of Horn and q -Horn formulae

Gábor Kusper^a, Csaba Biró^b, Attila Adamkó^c, Imre Baják^d

^aEszterházy Károly University
kusper.gabor@uni-eszterhazy.hu

^bEszterházy Károly University and Eötvös Lóránd University
biro.csaba@uni-eszterhazy.hu

^cUniversity of Debrecen
adamkoa@inf.unideb.hu

^dBudapest Business School
bajak.imre@uni-bge.hu

Submitted: February 2, 2021

Accepted: March 17, 2021

Published online: March 19, 2021

Abstract

In this paper we generalize the well-known notions of Horn and q -Horn formulae. A Horn clause, by definition, contains at most one positive literal. A Horn formula contains only Horn clauses. We generalize these notions as follows. A clause is a w -Horn clause if and only if it contains at least one negative literal or it is a unit or it is the empty clause. A formula is a w -Horn formula if it contains only w -Horn clauses after exhaustive unit propagation, i.e., after a Boolean Constraint Propagation (BCP) step. We show that the set of w -Horn formulae properly includes the set of Horn formulae. A function $\beta(x)$ is a valuation function if $\beta(x) + \beta(\neg x) = 1$ and $\beta(x) \in \{0, 0.5, 1\}$, where x is a Boolean variable. A formula \mathcal{F} is a q -Horn formula if and only if there is a valuation function $\beta(x)$ such that for each clause C in \mathcal{F} we have that $\sum_{x \in C} \beta(x) \leq 1$. In this case we call $\beta(x)$ a q -feasible valuation for \mathcal{F} . In other words, a formula is q -Horn if and only if each clause in it contains at most one “positive” literal (where $\beta(x) = 1$) or at most two half ones (where $\beta(x) = 0.5$). We generalize these notions as follows. A

formula \mathcal{F} is a z -Horn formula if and only if $\mathcal{F}' = \text{BCP}(\mathcal{F})$ and either \mathcal{F}' is trivially satisfiable or trivially unsatisfiable or there is a valuation function $\gamma(x)$ such that for each clause \mathcal{C} in \mathcal{F}' we have that $\sum_{x \in \mathcal{C} \wedge \gamma(x) \neq 0.5} \gamma(\neg x) \geq 1$ or $\sum_{x \in \mathcal{C} \wedge \gamma(x) = 0.5} \gamma(x) = 1$. In this case we call $\gamma(x)$ to be a z -feasible valuation for \mathcal{F}' . In other words, a formula is z -Horn if and only if each clause in it after a BCP step contains at least one “negative” literal (where $\gamma(x) = 0$) or exactly two half ones (where $\gamma(x) = 0.5$). We show that the set of z -Horn formulae properly includes the set of q -Horn formulae. We also show that the w -Horn SAT problem can be decided in polynomial time. We also show that each satisfiable formula is z -Horn.

Keywords: SAT, Horn, q -Horn, z -Horn, w -Horn.

AMS Subject Classification: 03B05, 03B20, 03B70

1. Introduction

Propositional satisfiability is the problem of determining, for a formula of the propositional calculus, if there is an assignment of truth values to its variables for which that formula evaluates to true. By SAT we mean the problem of propositional satisfiability for formulae in conjunctive normal form (CNF).

SAT is the first, and one of the simplest, of the many problems which have been shown to be \mathcal{NP} -complete [8]. It is the dual of propositional theorem proving, and many practical \mathcal{NP} -hard problems may be transformed efficiently to SAT. Thus, a good SAT algorithm would likely have considerable utility. It seems improbable that a polynomial time algorithm can be found for the general SAT problem unless $\mathcal{N} = \mathcal{NP}$, but we know that there are restricted SAT problems that are solvable in polynomial time. So a “good” SAT algorithm should first check whether the input SAT instance is an instance of such a restricted SAT problem. In this paper we introduce the w -Horn SAT problem, which is solvable in polynomial time. We also introduce the z -Horn SAT problem, but we do not know yet whether it is solvable in polynomial time or not.

We list some polynomial time solvable restricted SAT problems:

1. The restriction of SAT to instances where all clauses have length k is denoted by k -SAT. 2-SAT and 3-SAT are of special interest, because 3 is the smallest value of k for which k -SAT is \mathcal{NP} -complete, while 2-SAT is solvable in linear time [2, 11].
2. Horn SAT is the restriction to instances where each clause contains at most one positive literal. Horn SAT is solvable in linear time [10, 28], as are a number of generalizations such as renamable Horn SAT [1, 23], extended Horn SAT [7] and q -Horn SAT [5, 6]. An interesting variant for us is dual-Horn, or anti-Horn SAT, where in each clause there are at most one negative literal. The dual-Horn SAT is solvable in polynomial time.
3. The hierarchy of tractable satisfiability problems [9], which is based on Horn

- SAT and 2-SAT, is solvable in polynomial time. An instance on the k level of the hierarchy is solvable in $\mathcal{O}(nk + 1)$ time.
4. Nested SAT, in which there is a linear ordering on the variables and no two clauses overlap with respect to the interval defined by the variables they contain, is solvable in linear time. [16].
 5. SAT in which no variable appears more than twice. All such problems are satisfiable in linear time if they contain no unit clauses [32].
 6. r,r -SAT, where r,s -SAT is the class of problems in which every clause has exactly r literals and every variable has at most s occurrences. All r,r -SAT problems are satisfiable in polynomial time [32].
 7. A formula is SLUR (Single Lookahead Unit Resolution) solvable if, for all possible sequences of selected variables, algorithm SLUR does not give up. Algorithm SLUR is a nondeterministic algorithm based on unit propagation. It eventually gives up the search if it starts with, or creates, an unsatisfiable formula with no unit clauses. The class of SLUR solvable formulae was developed as a generalization including Horn SAT, renamable Horn SAT, extended Horn SAT, and the class of CC-balanced formulae [27].
 8. Resolution-Free SAT Problem, where every resolution results in a tautologous clause, is solvable in linear time [21]. And a generalization of it, the Blocked SAT Problem, where in each clause there is a blocked literal (resolution on that literal results in a tautologous clause, or the resolvent together with the blocked literal is subsumed) [19].
 9. Linear autarkies can be found in polynomial time [17]. A partial assignment is an autarky if it satisfies all clauses such that they have a common variable. For example, a pure literal is an autarky. Linear autarkies include q -Horn formulae, and incomparable with the SLUR [33].
 10. Matched expressions are recognized by creating a bipartite graph (V_1, V_2, E) , such that vertices of V_1 represent clauses, vertices of V_2 represent variables, and there is an edge from clause C to variable v if and only if C contains v or $\neg v$. If there is a total matching in this graph, i.e., there is a subset of edges, such that each clause and each variable are present but only once, then we say that the formula is matched. Matched formulae are satisfiable [13]. Total matching can be constructed, if it exists, in polynomial time. The class of matched formulae is incomparable with the q -Horn and SLUR classes.
 11. SAT problems generated from directed graphs are always satisfiable. Two assignments, the one where all variables are true, the so called white assignment, and the one where all variables are false, the so called black assignment, always satisfy them, so such problems are called Black-and-White SAT problems [3, 4, 22].

12. SAT can be solved efficiently by biology inspired methods. For example, \mathcal{P} systems with active membranes can solve it in linear time [14]. This article presents two solutions. The first solution is a uniform one, but it is not polynomially uniform. The second solution, which is based on the first one, is a polynomially semi-uniform solution. Other membrane based solutions can be found in [25].
13. When a finite fixed set of Boolean variables is used, then n -SAT can be solved by a specific deterministic finite automaton. So n -SAT is polynomial, but the specific deterministic finite automaton uses double exponential memory space [26].

In this paper we generalize the well-known notions of Horn and q -Horn formulae. A Horn clause, by definition, contains at most one positive literal. A Horn formula contains only Horn clauses.

We generalize these notions as follows. A clause is a w -Horn clause if and only if it contains at least one negative literal or it is a unit or it is the empty clause. A formula is a w -Horn formula if it contains only w -Horn clauses after propagating all units in it, i.e., after a BCP step. We show that the set of w -Horn formulae properly includes the set of Horn formulae.

A function $\beta(x)$ is a valuation function if $\beta(x) + \beta(\neg x) = 1$ and $\beta(x) \in \{0, 0.5, 1\}$, where x is a Boolean variable.

A formula is q -Horn if and only if each clause in it contains at most one “positive” literal (where $\beta(x) = 1$) or at most two half ones (where $\beta(x) = 0.5$).

We generalize these notions as follows. A formula is z -Horn if and only if each clause in it after a BCP step contains at least one “negative” literal or exactly two half ones.

We show that the set of z -Horn formulae properly includes the set of q -Horn formulae. We also show that the w -Horn SAT problem can be decided in polynomial time. We also show that each satisfiable formula is z -Horn.

2. Definitions

A literal is a Boolean variable or the negation of a Boolean variable. A clause is a set of literals. A clause set is a set of clauses. An assignment is a set of literals. Clauses are interpreted as disjunction of their literals. Assignments are interpreted as conjunction of their literals.

The negation of a variable v is denoted by \bar{v} . Given a set U of literals, we denote $\bar{U} := \{\bar{u} \mid u \in U\}$ and call it the negation of the set U . If w denotes a negative literal \bar{v} , then \bar{w} denotes the positive literal v . If \mathcal{C} is a clause, then $\bar{\mathcal{C}}$ is an assignment. If \mathcal{A} is an assignment, then $\bar{\mathcal{A}}$ is a clause.

If \mathcal{C} is a clause and its cardinality is k , denoted by $|\mathcal{C}| = k$, then we say that \mathcal{C} is a k -clause. Special cases are unit clauses or units which are 1-clauses, and clear or total clauses which are n -clauses. Note that any unit clause is at the same time a clause and an assignment.

If \mathcal{S} is a clause set and $\{u\}$ is a unit, then we can do unit propagation, for short UP, by $\{u\}$ on \mathcal{S} , denoted by $UP(\mathcal{S}, \{u\})$, as follows: $UP(\mathcal{S}, \{u\}) := \{\mathcal{C} \setminus \{\bar{u}\} \mid \mathcal{C} \in \mathcal{S} \wedge u \notin \mathcal{C}\}$.

By BCP we mean exhaustive unit propagation. To be more formal:

$$BCP(\mathcal{S}) = \begin{cases} BCP(UP(\mathcal{C}, \{u\})), & \text{where } \{u\} \in \mathcal{C}, \\ \mathcal{C}, & \text{if there are no more units in } \mathcal{C}. \end{cases}$$

We say that assignment \mathcal{M} is a model for clause set \mathcal{S} iff for all $\mathcal{C} \in \mathcal{S}$ we have $\mathcal{M} \cap \mathcal{C} \neq \emptyset$.

We say that a clause set is trivially unsatisfiable iff it contains the empty clause. We say that a clause set is trivially satisfiable iff it is the empty set.

We introduce two functions $P(\mathcal{C})$, the number of positive literals in clause \mathcal{C} , and $N(\mathcal{C})$, the number of negative literals in clause \mathcal{C} . Note, that $P(\mathcal{C}) + N(\mathcal{C}) = |\mathcal{C}|$.

The clause \mathcal{C} is a Horn clause iff $P(\mathcal{C}) \leq 1$. Note that the empty clause is a Horn clause. The clause set \mathcal{F} is a Horn formula iff for each clause \mathcal{C} in \mathcal{F} we have that \mathcal{C} is a Horn clause.

We generalize these notions as follows. The clause \mathcal{C} is a w -Horn clause iff $N(\mathcal{C}) \geq 1$ or \mathcal{C} is a unit or \mathcal{C} is the empty clause. The clause set \mathcal{F} is a w -Horn formula iff $\mathcal{F}' = BCP(\mathcal{F})$ and for each clause \mathcal{C} in \mathcal{F}' we have that \mathcal{C} is a w -Horn clause.

Examples for w -Horn formulae:

1. $(\neg a \vee b \vee c)$.
2. $(\neg a \vee \neg b) \wedge (\neg a \vee b) \wedge (a \vee \neg b)$.
3. $(\neg a \vee \neg b \vee \neg c) \wedge (\neg a \vee \neg b \vee c) \wedge (\neg a \vee b \vee \neg c) \wedge (\neg a \vee b \vee c) \wedge (a \vee \neg b \vee \neg c) \wedge (a \vee \neg b \vee c) \wedge (a \vee b \vee \neg c)$, this example shows the great expressiveness of w -Horn.
4. $(a) \wedge (\neg a \vee b)$, because after BCP we obtain the empty clause set.
5. $(\neg a \vee \neg b) \wedge (\neg a \neg b) \wedge (a \vee \neg b) \wedge (a \vee b \vee c) \wedge (\neg c)$, because after BCP we obtain $(\neg a \vee \neg b) \wedge (\neg a \vee b) \wedge (a \vee \neg b)$.
6. $(a) \wedge (\neg a)$ is w -Horn, because after BCP we obtain a clause set which contains the empty clause, and the empty clause is w -Horn.
7. $(\neg a \vee b \vee c)$ is w -Horn, because $N(\mathcal{C}) = 1$, but not Horn, because $P(\mathcal{C}) = 2$.

By w -Horn SAT problem we mean the problem of deciding whether a given w -Horn formula is satisfiable or not.

A function $\beta(x)$ is a valuation function if $\beta(x) + \beta(\neg x) = 1$ and $\beta(x) \in \{0, 0.5, 1\}$, where x is a Boolean variable. Note that if \mathcal{C} is a clause, then $\sum_{x \in \mathcal{C}} (\beta(x) + \beta(\neg x)) = |\mathcal{C}|$.

A formula \mathcal{F} is a q -Horn formula iff there is a valuation function $\beta(x)$ such that for each clause \mathcal{C} in \mathcal{F} we have that $\sum_{x \in \mathcal{C}} \beta(x) \leq 1$. In this case we call $\beta(x)$ a q -feasible valuation for \mathcal{F} .

In other words, a formula is q -Horn if and only if each clause in it contains at most one “positive” literal (where $\beta(x) = 1$) or at most two half ones (where $\beta(x) = 0.5$). We generalize these notions as follows.

A formula \mathcal{F} is a z -Horn formula iff $\mathcal{F}' = BCP(\mathcal{F})$ and either \mathcal{F}' is trivially satisfiable or trivially unsatisfiable or there is a valuation function $\gamma(x)$ such that $\sum_{x \in \mathcal{C} \wedge \gamma(x) \neq 0.5} \gamma(\neg x) \geq 1$ or $\sum_{x \in \mathcal{C} \wedge \gamma(x) = 0.5} \gamma(x) = 1$. In this case we call $\gamma(x)$ to be a z -feasible valuation for \mathcal{F}' .

In other words, a formula is z -Horn if and only if each clause in it after a BCP step contains at least one “negative” literal (where $\gamma(x) = 0$) or exactly two half ones (where $\gamma(x) = 0.5$).

Examples for z -Horn formulae:

1. $(a) \wedge (\neg a)$, because after BCP we obtain a trivially unsatisfiable clause set; this example is also q -Horn, because $\beta(a) = 0.5$ is a q -feasible valuation for it.
2. $(a) \wedge (\neg a \vee b)$, because after BCP we obtain the empty clause set, which is trivially satisfiable.
3. $(a \vee b) \wedge (\neg a \vee c)$, because every 2-SAT problem is a z -Horn formula.
4. $(\neg a \vee b \vee c) \wedge (\neg a \vee \neg b \vee \neg c)$ is z -Horn, because $\gamma(a) = \gamma(b) = \gamma(c) = 0$ is a z -feasible valuation, but it is enough to say that $\gamma(\neg a) = 1$. Note that this formula is said not to be q -Horn, see examples 2.9. and 2.10. in [12], but it is actually q -Horn, because $\beta(\neg a) = 0$, and $\beta(b) = \beta(c) = 0.5$ is a q -feasible valuation for it.
5. $(\neg a \vee b \vee c) \wedge (\neg a \vee \neg b \vee c) \wedge (a \vee \neg b \vee \neg c)$ is z -Horn, because $\gamma(\neg a) = \gamma(\neg b) = \gamma(\neg c) = 1$ is a z -feasible valuation, but not q -Horn. This has also been checked by our q -Horn / z -Horn checker written in Java. This checker can be found on our webpage: <http://fmv.ektf.hu/tools.html> [20].

3. Properties of w -Horn formulae

Lemma 3.1. *The set of w -Horn formulae properly includes the set of Horn formulae.*

Proof. First we show inclusion. Let \mathcal{F} be an arbitrary but fixed Horn formula. Let $\mathcal{F}' = BCP(\mathcal{F})$. Note that \mathcal{F}' does not contain any unit clauses. Note furthermore that \mathcal{F}' is still a Horn formula, because the set of Horn formulae is closed under unit propagation. We show that \mathcal{F}' is a w -Horn formula. There are two cases: \mathcal{F}' is either the empty set or not. In the first case, by definition, \mathcal{F} is w -Horn. In the second case let \mathcal{C} be an arbitrary but fixed clause from \mathcal{F}' . There are two cases, either \mathcal{C} is the empty clause or not. In the first case \mathcal{C} is also a w -Horn clause. In the second case we do the following steps. We know that \mathcal{C} is a Horn clause, so $P(\mathcal{C}) \leq 1$. From this, by multiplying both sides by -1 , we obtain that

$-P(\mathcal{C}) \geq -1$, and by adding $|\mathcal{C}|$ to both sides, we obtain $|\mathcal{C}| - P(\mathcal{C}) \geq |\mathcal{C}| - 1$. From this, by $P(\mathcal{C}) + N(\mathcal{C}) = |\mathcal{C}|$, we know that $N(\mathcal{C}) \geq |\mathcal{C}| - 1$. We know that $\mathcal{C} \in \mathcal{F}'$, so \mathcal{C} is not a unit, we also know that it is not empty clause, so $|\mathcal{C}| - 1 \geq 1$. From these we obtain that $N(\mathcal{C}) \geq 1$. So, by definition, \mathcal{C} is a w -Horn clause. Hence, \mathcal{F} is a w -Horn formula.

As a second step we show that there is a formula which is w -Horn, but not Horn. The formula $\mathcal{C} = (\neg a \vee b \vee c)$ is w -Horn, because $N(\mathcal{C}) = 1$, but not Horn, because $P(\mathcal{C}) = 2$. Hence, the set of w -Horn formulae properly includes the set of Horn formulae. \square

Theorem 3.2. *The w -Horn SAT problem is solvable in polynomial time.*

Proof. Let \mathcal{F} be an arbitrary but fixed w -Horn formula. We show that it is solvable in polynomial time. Let $\mathcal{F}' = BCP(\mathcal{S})$. This step is polynomial since unit propagation is polynomial [34]. If \mathcal{F}' contains the empty clause, then \mathcal{F} is unsatisfiable. Otherwise \mathcal{F} is satisfiable and its model consists of the units propagated in the BCP step, the rest of the variables are negative. \square

4. Properties of z -Horn formulae

Lemma 4.1. *The set of z -Horn formulae properly includes the set of q -Horn formulae.*

Proof. First we show inclusion. Let \mathcal{F} be an arbitrary but fixed q -Horn formula. We show that \mathcal{F} is a z -Horn formula. Let $\mathcal{F}' = BCP(\mathcal{F})$. Note that \mathcal{F}' is still a q -Horn formula, because the set of q -Horn formulae is closed under unit propagation. There are two cases: \mathcal{F}' is either the empty set or not. In the first case, by definition, \mathcal{F} is z -Horn. In the second case let \mathcal{C} be an arbitrary but fixed clause from \mathcal{F}' . Note that \mathcal{C} is not a unit. Since \mathcal{F}' is a q -Horn formula, we know that there exists a q -feasible valuation for \mathcal{F}' , let us call it $\beta(x)$, such that $\sum_{x \in \mathcal{C}} \beta(x) \leq 1$.

There are 4 cases: Either (1) $\sum_{x \in \mathcal{C}} \beta(x) = 0$, or (2) $\sum_{x \in \mathcal{C}} \beta(x) = 0.5$, or (3) $\sum_{x \in \mathcal{C}} \beta(x) = 1$ and $\sum_{x \in \mathcal{C} \wedge \beta(x) \neq 0.5} \beta(x) = 1$, or (4) $\sum_{x \in \mathcal{C}} \beta(x) = 1$ and $\sum_{x \in \mathcal{C} \wedge \beta(x) = 0.5} \beta(x) = 1$.

In case (1) either \mathcal{F}' contains the empty clause or not. In the first case, by definition, \mathcal{F} is z -Horn. In the second case we have that $\sum_{x \in \mathcal{C} \wedge \beta(x) \neq 0.5} \beta(\neg x) = |\mathcal{C}|$. Since \mathcal{C} is not the empty clause, we have that $\sum_{x \in \mathcal{C} \wedge \beta(x) \neq 0.5} \beta(\neg x) \geq 1$. This means that $\beta(x)$ is a q -feasible valuation for \mathcal{F}' . Therefore, \mathcal{F} is, by definition, a z -Horn formula.

In case (2) we have that $\sum_{x \in \mathcal{C} \wedge \beta(x) \neq 0.5} \beta(\neg x) = |\mathcal{C}| - 0.5$. Since \mathcal{C} is not the empty clause and neither a unit, we have that $\sum_{x \in \mathcal{C} \wedge \beta(x) \neq 0.5} \beta(\neg x) \geq 1$. This means that $\beta(x)$ is a q -feasible valuation for \mathcal{F}' . Therefore, \mathcal{F} is, by definition, a z -Horn formula.

In case (3) we have that $\sum_{x \in \mathcal{C} \wedge \beta(x) \neq 0.5} \beta(\neg x) = |\mathcal{C}| - 1$. Since \mathcal{C} is not the empty clause and neither a unit, we have that $\sum_{x \in \mathcal{C} \wedge \beta(x) \neq 0.5} \beta(\neg x) \geq 1$. This

means that $\beta(x)$ is a q -feasible valuation for \mathcal{F}' . Therefore, \mathcal{F} is, by definition, a z -Horn formula.

In case (4) we have that $\sum_{x \in \mathcal{C} \wedge \beta(x)=0.5} \beta(x) = 1$. So $\beta(x)$ is a q -feasible valuation for \mathcal{F}' . Therefore, \mathcal{F} is, by definition, a z -Horn formula.

So in all cases we have that \mathcal{F} is a z -Horn formula. Hence, the set of z -Horn formulae includes the set of q -Horn formulae.

As a second step we show that there is a formula which is z -Horn, but not q -Horn. For example the formula $(\neg a \vee b \vee c) \wedge (\neg a \vee \neg b \vee c) \wedge (a \vee \neg b \vee \neg c)$ is z -Horn but not q -Horn, see the z -Horn examples in section 2. Hence, the set of z -Horn formulae properly includes the set of q -Horn formulae. \square

Theorem 4.2. *Any satisfiable \mathcal{F} formula is z -Horn.*

Proof. Let \mathcal{F} be an arbitrary but fixed satisfiable formula. Let \mathcal{M} be a model for \mathcal{F} , i.e., for each clause \mathcal{C} in \mathcal{F} we have that \mathcal{C} intersection \mathcal{M} is not empty. Let $\gamma(x)$ be a valuation function constructed in the following way: For all m in \mathcal{M} let $\gamma(m) = 0$. It is easy to see that $\gamma(x)$ is a z -feasible valuation for \mathcal{F} . Hence, any satisfiable \mathcal{F} formula is z -Horn. \square

5. Future work

We do not consider in this paper the question of what the relation is between w -Horn and z -Horn and other generalizations of Horn formulae, linear autarky [18, 24], and other polynomial time SAT problems.

Since we allow more than two “half” literals in a z -Horn clause if there is at least one “negative” literal, we can use the so called simulated annealing based methods [15, 29] to find a z -feasible valuation of the input clause set.

According to our current ideas the cooling process work as follows. At the beginning, each literal is a “half” one. Then we cool the system and some literals become “negative”, we repeat this until we obtain the 2-SAT core of the problem, which means that in each clause there is at least one “negative” literal or exactly two “half” ones.

The other way to attack this problem is to use neural networks. The expressive power of z -Horn is great, i.e., almost all SAT problems are z -Horn, but in the worst case, to find the corresponding z -feasible function, we have to solve the input SAT problem. Instead of this expensive method we can use votes like units have “negative” value, any other variables are “half” ones. We can use more elaborated neural networks, which predict which variables are “negative”, “positive”, and “half” one. Then we can combine them to find the z -feasible function by a voting system, like in [30, 31].

Acknowledgment. G. Kusper would like to thank the support of the Complex improvement of research capacities and services at Eszterházy Károly University, project ID: EFOP-3.6.1-16-2016-00001, and also the support of the Implementation

of services that implement and provide secure personal data management and value-based information trade services in healthcare management, project ID: GINOP-2.1.2-8-1-4-16-2017-00176.

Cs. Biró would like to thank the support of the Ministry of Innovation and Technology and the National Research, Development and Innovation Office within the Quantum Information National Laboratory of Hungary.

References

- [1] B. ASPVALL: *Recognizing disguised NR (1) instances of the satisfiability problem*, Journal of Algorithms 1.1 (1980), pp. 97–103, DOI: [https://doi.org/10.1016/0196-6774\(80\)90007-3](https://doi.org/10.1016/0196-6774(80)90007-3).
- [2] B. ASPVALL, M. F. PLASS, R. E. TARJAN: *A linear-time algorithm for testing the truth of certain quantified boolean formulas*, Information Processing Letters 8.3 (1979), pp. 121–123, DOI: [https://doi.org/10.1016/0020-0190\(79\)90002-4](https://doi.org/10.1016/0020-0190(79)90002-4).
- [3] C. BIRÓ, G. KUSPER: *BaW 1.0-A Problem Specific SAT Solver for Effective Strong Connectivity Testing in Sparse Directed Graphs*, in: 2018 IEEE 18th International Symposium on Computational Intelligence and Informatics (CINTI), IEEE, 2018, pp. 000137–000142, DOI: <https://doi.org/10.1109/CINTI.2018.8928191>.
- [4] C. BIRÓ, G. KUSPER: *Equivalence of strongly connected graphs and black-and-white 2-SAT problems*, Miskolc Mathematical Notes 19.2 (2018), pp. 755–768, DOI: <https://doi.org/10.18514/mmn.2018.2140>.
- [5] E. BOROS, Y. CRAMA, P. L. HAMMER, M. SAKS: *A complexity index for satisfiability problems*, SIAM Journal on Computing 23.1 (1994), pp. 45–49, DOI: <https://doi.org/10.1137/S0097539792228629>.
- [6] E. BOROS, P. L. HAMMER, X. SUN: *Recognition of q -Horn formulae in linear time*, Discrete Applied Mathematics 55.1 (1994), pp. 1–13, DOI: [https://doi.org/10.1016/0166-218X\(94\)90033-7](https://doi.org/10.1016/0166-218X(94)90033-7).
- [7] V. CHANDRU, J. N. HOOKER: *Extended Horn sets in propositional logic*, Journal of the ACM (JACM) 38.1 (1991), pp. 205–221, DOI: <https://doi.org/10.1145/102782.102789>.
- [8] S. A. COOK: *The complexity of theorem-proving procedures*, in: Proceedings of the third annual ACM symposium on Theory of computing, 1971, pp. 151–158, DOI: <https://doi.org/10.1145/800157.805047>.
- [9] M. DALAL, D. W. ETHERINGTON: *A hierarchy of tractable satisfiability problems*, Information Processing Letters 44.4 (1992), pp. 173–180, DOI: [https://doi.org/10.1016/0020-0190\(92\)90081-6](https://doi.org/10.1016/0020-0190(92)90081-6).
- [10] W. F. DOWLING, J. H. GALLIER: *Linear-time algorithms for testing the satisfiability of propositional Horn formulae*, The Journal of Logic Programming 1.3 (1984), pp. 267–284, DOI: [https://doi.org/10.1016/0743-1066\(84\)90014-1](https://doi.org/10.1016/0743-1066(84)90014-1).
- [11] S. EVEN, A. ITAI, A. SHAMIR: *On the complexity of time table and multi-commodity flow problems*, in: 16th Annual Symposium on Foundations of Computer Science (sfcs 1975), IEEE, 1975, pp. 184–193, DOI: <https://doi.org/10.1109/SFCS.1975.21>.
- [12] J. FRANCO: *Relative size of certain polynomial time solvable subclasses of satisfiability*, in: Satisfiability Problem: Theory and Applications (DIMACS Workshop March 11-13, 1996), vol. 35, 1997, pp. 211–223, URL: <https://apps.dtic.mil/sti/pdfs/ADA326040.pdf>.

- [13] J. FRANCO, A. VAN GELDER: *A perspective on certain polynomial-time solvable classes of satisfiability*, Discrete Applied Mathematics 125.2 (2003), pp. 177–214, ISSN: 0166-218X, DOI: [https://doi.org/10.1016/S0166-218X\(01\)00358-4](https://doi.org/10.1016/S0166-218X(01)00358-4).
- [14] Z. GAZDAG, G. KOLONITS: *A New Approach for Solving SAT by P Systems with Active Membranes*, Membrane Computing. CMC 2012. Lecture Notes in Computer Science 7762 (2012), pp. 195–207, DOI: https://doi.org/10.1007/978-3-642-36751-9_14.
- [15] S. KIRKPATRICK, C. D. GELATT, M. P. VECCHI: *Optimization by simulated annealing*, science 220.4598 (1983), pp. 671–680.
- [16] D. E. KNUTH: *Nested satisfiability*, Acta Informatica 28.1 (1990), pp. 1–6, DOI: <https://doi.org/10.1007/BF02983372>.
- [17] O. KULLMANN: *Investigations on autark assignments*, Discrete Applied Mathematics 107.1 (2000), SI Boolean Functions, pp. 99–137, ISSN: 0166-218X, DOI: [https://doi.org/10.1016/S0166-218X\(00\)00262-6](https://doi.org/10.1016/S0166-218X(00)00262-6).
- [18] O. KULLMANN: *Investigations on autark assignments*, Discrete Applied Mathematics 107.1-3 (2000), pp. 99–137, DOI: [https://doi.org/10.1016/S0166-218X\(00\)00262-6](https://doi.org/10.1016/S0166-218X(00)00262-6).
- [19] G. KUSPER: *Finding models for blocked 3-SAT problems in linear time by systematical refinement of a sub-model*, in: Annual Conference on Artificial Intelligence, Springer, 2006, pp. 128–142, DOI: https://doi.org/10.1007/978-3-540-69912-5_11.
- [20] G. KUSPER: *q-Horn and z-Horn Checker*, 2021, DOI: <https://doi.org/10.13140/RG.2.2.27575.24482>, URL: http://fmv.ektf.hu/files/q-Horn_and_z-Horn_Checker.zip.
- [21] G. KUSPER: *Solving the resolution-free SAT problem by submodel propagation in linear time*, Annals of Mathematics and Artificial Intelligence 43.1-4 (2005), pp. 129–136, DOI: <https://doi.org/10.1007/s10472-005-0423-7>.
- [22] G. KUSPER, C. BIRÓ: *Convert a Strongly Connected Directed Graph to a Black-and-White 3-SAT Problem by the Balatonboglár Model*, Algorithms 13.12 (2020), p. 321, DOI: <https://doi.org/10.3390/a13120321>.
- [23] H. R. LEWIS: *Renaming a set of clauses as a Horn set*, Journal of the ACM (JACM) 25.1 (1978), pp. 134–135, DOI: <https://doi.org/10.1145/322047.322059>.
- [24] H. VAN MAAREN: *A short note on linear autarkies, q-Horn formulas and the complexity index*, tech. rep., Citeseer, 1999, DOI: <https://doi.org/10.1006/inco.2000.2867>.
- [25] B. NAGY: *On Efficient Algorithms for SAT*, in: Membrane Computing, ed. by E. CSUHAJ-VARJÚ, M. GHEORGHE, G. ROZENBERG, A. SALOMAA, G. VASZIL, Berlin, Heidelberg: Springer Berlin Heidelberg, 2013, pp. 295–310, DOI: https://doi.org/10.1007/978-3-642-36751-9_20.
- [26] B. NAGY: *The languages of SAT and n-SAT over finitely many variables are regular*, Bulletin-European Association for Theoretical Computer Science 82 (2004), pp. 286–297.
- [27] J. S. SCHLIPP, F. S. ANNEXSTEIN, J. V. FRANCO, R. P. SWAMINATHAN: *On finding solutions for extended Horn formulas*, Information Processing Letters 54.3 (1995), pp. 133–137, DOI: [https://doi.org/10.1016/0020-0190\(95\)00019-9](https://doi.org/10.1016/0020-0190(95)00019-9).
- [28] M. G. SCUTELLA: *A note on Dowling and Gallier’s top-down algorithm for propositional Horn satisfiability*, The Journal of Logic Programming 8.3 (1990), pp. 265–273, DOI: [https://doi.org/10.1016/0743-1066\(90\)90026-2](https://doi.org/10.1016/0743-1066(90)90026-2).
- [29] W. M. SPEARS: *Simulated annealing for hard satisfiability problems*. Cliques, Coloring, and Satisfiability 26 (1993), pp. 533–558.

- [30] T. TAJTI: *Fuzzification of training data class membership binary values for neural network algorithms*, *Annales Mathematicae et Informaticae* 52 (2020), pp. 217–228.
- [31] T. TAJTI: *New voting functions for neural network algorithms*, *Annales Mathematicae et Informaticae* 52 (2020), pp. 229–242.
- [32] C. A. TOVEY: *A simplified NP-complete satisfiability problem*. *Discret. Appl. Math.* 8.1 (1984), pp. 85–89,
DOI: [https://doi.org/10.1016/0166-218X\(84\)90081-7](https://doi.org/10.1016/0166-218X(84)90081-7).
- [33] H. VAN MAAREN: *A Short Note on Some Tractable Cases of the Satisfiability Problem*, *Information and Computation* 158.2 (2000), pp. 125–130, ISSN: 0890-5401,
DOI: <https://doi.org/10.1006/inco.2000.2867>.
- [34] H. ZHANG, M. E. STICKELY: *An Efficient Algorithm for Unit Propagation*, *Proc. of AI-MATH* 96 (1996),
URL: <https://citeseerx.ist.psu.edu/viewdoc/download?doi=10.1.1.49.5500&rep=rep1&type=pdf>.

Comparison of single and ensemble-based convolutional neural networks for cancerous image classification

Oktavian Lantang^a, Gyorgy Terdik^a,
Andras Hajdu^b, Attila Tiba^b

^aDepartment of Applied Information Technology and its Theoretical Background,
University of Debrecen
oktavian_lantang@unsrat.ac.id
terdik.gyorgy@inf.unideb.hu

^bDepartment of Computer Graphics and Image Processing,
University of Debrecen
hajdu.andras@inf.unideb.hu
tiba.attila@inf.unideb.hu

Submitted: June 16, 2020

Accepted: March 17, 2021

Published online: March 31, 2021

Abstract

In this work, we investigated the ability of several Convolutional Neural Network (CNN) models for predicting the spread of cancer using medical images. We used a dataset released by the Kaggle, namely PatchCamelyon. The dataset consists of 220,025 pathology images digitized by a tissue scanner. A clinical expert labeled each image as cancerous or non-cancerous. We used 70% of the images as a training set and 30% of them as a validation set. We design three models based on three commonly used modules: VGG, Inception, and Residual Network (ResNet), to develop an ensemble model and implement a voting system to determine the final decision. Then, we compared the performance of this ensemble model to the performance of each single model. Additionally, we used a weighted majority voting system, where the final prediction is equal to the weighted average of the prediction produced by each network. Our results show that the classification of the two ensemble models reaches 96%. Thus these results prove that the ensemble model outperforms single network architectures.

Keywords: Cancer image classification, ensemble-based model, convolutional neural network.

1. Introduction

Currently, non-communicable diseases are the most significant contributor to mortality rates throughout the world. One type of non-communicable disease that plays an essential role in the high number of deaths is cancer. In 2015 WHO estimated that cancer was the leading cause of human death during the productive period, which is below 70 years [2]. By definition, cancer refers to more than one hundred types of diseases with their unique features. Every human being has trillions of body cells that multiply and depend on each other. The body's metabolism automatically controls the development of each cell to maintain its size and shape. However, cancer cells work oppositely. These cells develop regardless the protocol instructed by the human body. And worse, cancer cells can move from one place to another [21].

In the last decade, pathologists used a microscope to predict cancer. Experts are trained to understand clinical symptoms and later diagnose them. The doctor uses these results for decision making. Now routines like this are no longer a priority since the development of the whole slide image scanner documents of the histological images in digital form. By relying on sophisticated imaging and analysis techniques, this tool can record more complex variables that exist in histological images [12]. Furthermore, the images produced by this tool can detect not only the presence of cancer cells in the body but also show biological processes such as apoptosis, angiogenesis, and metastasis [22]. The histological image documentation process massively produces a tremendous amount of data. The availability of a large amount of data can be seen as an opportunity to develop a machine learning system by designing a Convolutional Neural Network (CNN) [17].

The success of CNNs in producing good predictions can be seen in many previous works, among others [7, 11, 18, 19]. Krizhevsky et al. developed a network called Alexnet. This network is designed in eight stack layers. The eight layers are divided into two large blocks, and the first is filled by five convolutional layers and three fully connected layers. While at the last layer, this model has a 1000-way softmax, which refers to multiclass classification problems. They trained it with 1.2 million high-resolution images provided by ImageNet. Using this model, a 16.4% error rate for 5 CNN architectures and a 15.3% error rate for 7 CNN ones in the top five classifications were reported [11].

The VGG module was developed by Simonyan et al. The idea of this network is the definition and the repetition of convolutions blocks. This model also utilizes Max Pooling layers to reduce the dimension and small filter to decrease computation costs. Satisfactory results were reported in this work. Namely, using the same dataset, this study reports a 6.8% error rate for the top five predicted labels and a 23.7% error rate in the top first predicted labels [18].

As we know, CNN is an architecture that was developed to extract features from

images comprehensively. However, one of the problems faced is the high variety of the spatial position of the image information. In a dataset, the information we want to retrieve is not always in the center of the image. Moreover, the desired information may have a small percentage of other details. The large spatial variety of information from an image makes it difficult to determine the suitable filter size for CNN. Using a large filter makes the information more global, thus increasing the cost of computing. On the other hand, if we use a small filter, it will cause the information to be more local and eliminate essential knowledge from the image. For this reason, the Inception architecture was designed by installing multiple different size filters at the same level and concatenate them to reduce computing costs without losing deciding information. This idea will produce architectures that tend to be broad than deep [19].

The above studies showed that a deeper and more complex architecture resulted in a better accuracy and validation score. However, deep and complex architecture can damage the accuracy and validation of the model. He et al. tried to solve this problem by developing a Residual Network (Resnet) model. Resnet's basic concept is to group CNN into several blocks, and each block has a short cut to do a pass. This model architecture is constructed from 34 layers of residual blocks for the smallest architecture to 152 layers for the most complex one. The 152 layers single architecture reported very satisfying results by having a 19.38% error rate for the top first predicted labels and a 4.49% error rate for the top five predicted labels [7].

2. Related works

Classification using deep learning methods has produced excellent works. One of these was the work of Veeling et al. [20]. The suggested model adopts the DenseNet architecture, which uses Dense Block and Transition Block. Dataset was tested on six different single DenseNet models, and the P4M-DenseNet model gave the best results with an accuracy score of 89.8%. Kassani et al. [10] developed a model from three base modules: VGG19, MobileNet, and DenseNet. The model was trained using transfer learning techniques in a CNN ensemble framework utilizing four different datasets, including the PatchCamelyon dataset. Specifically, on the PatchCamelyon dataset, this work reported the accuracy of 94.64% for the CNN ensemble model. Another study from Xia et al. [23] compared two well-known CNN training methods, namely training from scratch and fine-tuning. They used the Camelyon 16 dataset, which is the origin of the PatchCam dataset. This work reported a result of 84.3% accuracy when the GoogleLeNet architecture was trained using a fine-tuned training method.

In this work, we investigated two CNN models' ability, namely single and ensemble, for predicting the spread of cancer using medical images. We used a PatchCamelyon dataset of 220,025 pathology images digitized using a tissue scanner and labeled as cancerous or non-cancerous. 70% of images were used as a training set and the rest as the validation set. To develop an ensemble model, we chose three

commonly used CNN modules, namely VGG, Inception, and Residual ResNet, with a voting system to determine the final decision. Furtherly, we compared the performance of this ensemble model to the performance of each single module. Additionally, we used a weighted majority voting system where the final prediction is equal to the weighted average of the prediction produced by each network.

3. Methodology

3.1. Dataset, hardware and software

The data we use is published in Kaggle, the PatchCamelyon dataset, which is derived from the Camelyon16 Dataset [1, 20]. The dataset consists of pathology images generated from a digital scanner. The whole slide image are broken down into smaller segments of size 92×92 pixels. The dataset contains 220,025 images, then divided 154,018 for the training set and 66,007 for the validation set. To simplify our work, we use a validation set as well as a test set. Next, we show sample images of the data set in Figures 1 and 2. To support this work, we utilize Google Collaboraty with NVIDIA Cuda Compilation Tool V8.0.61 besides that we also use DELL desktop with GEFORCE GTX 1060 6GB.

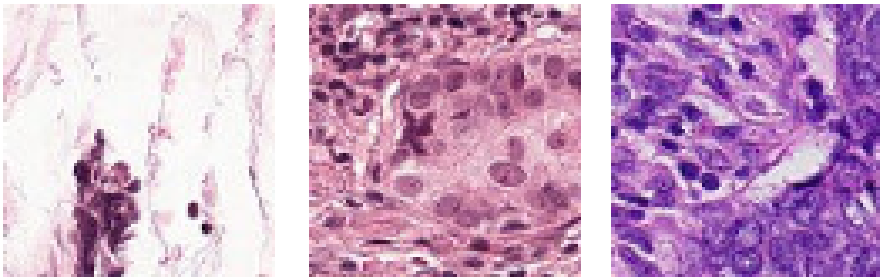


Figure 1. Cancerous images.

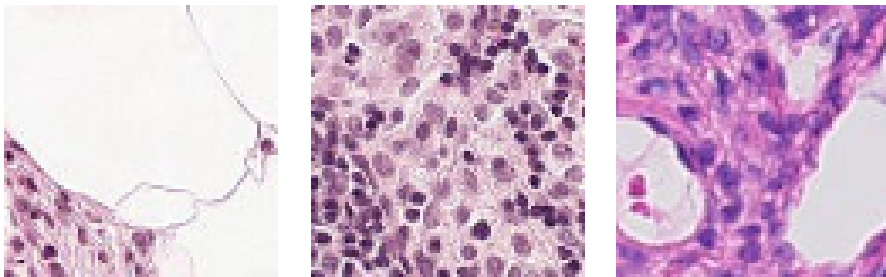


Figure 2. Non-cancerous images.

3.2. Preprocessing and augmentation

We chose 92×92 pixels as the input size. Furthermore, we used an augmentation process as part of image pre-processing to provide a sufficient amount of data and resist the overfitting condition. The technical process involved rotation, shifting, shearing, zoom and flipping as shown in Table 1.

Table 1. Augmentation process.

Rotation	45°
Shifting	0.2
Shearing	0.2
Zoom	0.2
Flipping	Horizontal

3.3. Base model of ensembles

As for the neural network architectures VGG, Inception, ResNet, we did not integrate any existing realizations, we implemented them from scratch to gain less complex models. The first model was the LT-VGG based on the VGG module. We stacked thirteen layers with the following details: ten convolutions layers and three fully connected ones. We inserted a Max Pooling layer after every two convolutions layers to have four pooling layers in total. Before entering fully connected layers, the feature dimensions are changed using the Flatten layer and then passed on to three fully connected layers: two 64-neurons and a Softmax with two-classes at the end of the network.

The second model was LT-Inception based on the Inception module. The modifications performed in this model include twelve convolutions, which are divided into two levels. Each level is filled by six convolutions and one Max Pooling layer. Before going to the next level, the convolutions at level one were concatenated. After the concatenation process at the second level, the dimension was shrunk using the Average Pooling layer. The dimensions were changed using the Flatten layer and finally streamed to three fully connected layers of two 64-neurons and a Softmax for two-classes.

The last model was the LT-ResNet based on the ResNet module. We installed eighteen convolutions layers and also inserted one residual layer for every three convolutional layers. So in total, we used 24 convolutions layers. We also used the Average Pooling layer to reduce the features' dimensions before converting to one dimension using the Flatten Layer. Next, we used two fully connected layers of two 64-neurons and a Softmax two-classes.

Refers to [4, 15], Softmax function $f(s): \mathbb{R}^K \rightarrow \mathbb{R}^K$ is a vector function in the range $[0, 1]$, where K is the number of classes. This function is obtained by calculating the exponential number to the power of s_i , where s_i refers to the score s from class i . Hereafter, numerator divided by the sum of the constant e to the

power of all score in number of classes:

$$f(s)_i = \frac{e^{s_i}}{\sum_{c=1}^K e^{s_c}}. \quad (3.1)$$

3.4. Ensemble model architecture

The ensemble method is one of the popular techniques to improve CNN's accuracy, as described in [9]. The CNN ensemble technique is a combination of several CNNs used to accomplish the same task. In their study, 193 articles were selected in four different databases: ACM, Scopus, IEEE Xplore, and PubMed. Their work reported that the majority voting method is the most widely used in the heterogeneous ensemble type. The most popular type of classifier is Support Vector Machine, beating Artificial Neural Network in fourth place. Nevertheless, the dataset used is mostly extracted from mammograms, not images.

To see more clearly the use of the CNN ensemble method in image datasets, we also studied the work of Savelli et al. [16]. By implementing the CNN Ensemble, they detected minor lesions in medical images. From this work, we can see how the four CNN singles are combined, and then the final decision is taken from the average score of the four single models. This work used the dataset of medical images, namely INbreast, which relates to breast cancer, and E-ophttha, a retinal fundus image.

Furthermore, Haragi's work[6] designed the CNN ensemble for the classification of skin lesions. In this study, we focus on recognizing how the final decision techniques are applied to the ensemble method. We can see that the authors consider several ways, such as Probabilistic, Majority Voting, and Weighting. From the results reported, there is a significant difference in accuracy between single CNN and ensemble one. Meanwhile, ensemble CNN's final decision technique shows that Simple Majority Voting provides the best accuracy score. On the other hand, the weighting method excels in measuring the area under curve (AUC).

We trained three base models separately so that the ensemble model will have three prediction results. We chose two types of voting systems that are used by the ensemble model. The first voting system is majority voting. This system gives each base model equal weight without considering achieving each model's accuracy when trained separately. Whereas the other voting system is that we apply special weights to each model, referring to the accuracy of each training's results. Furtherly, we compared the performance of this ensemble model to the performance of each single model. The architecture of the ensemble model shown in Figure 3.

3.5. Training process

We experimented by gradually increasing the epoch from 10 to 100. The best results were obtained at the epoch of 50. After that, there was an inconsistency in both machine capability and the accuracy score. To save training time, we took

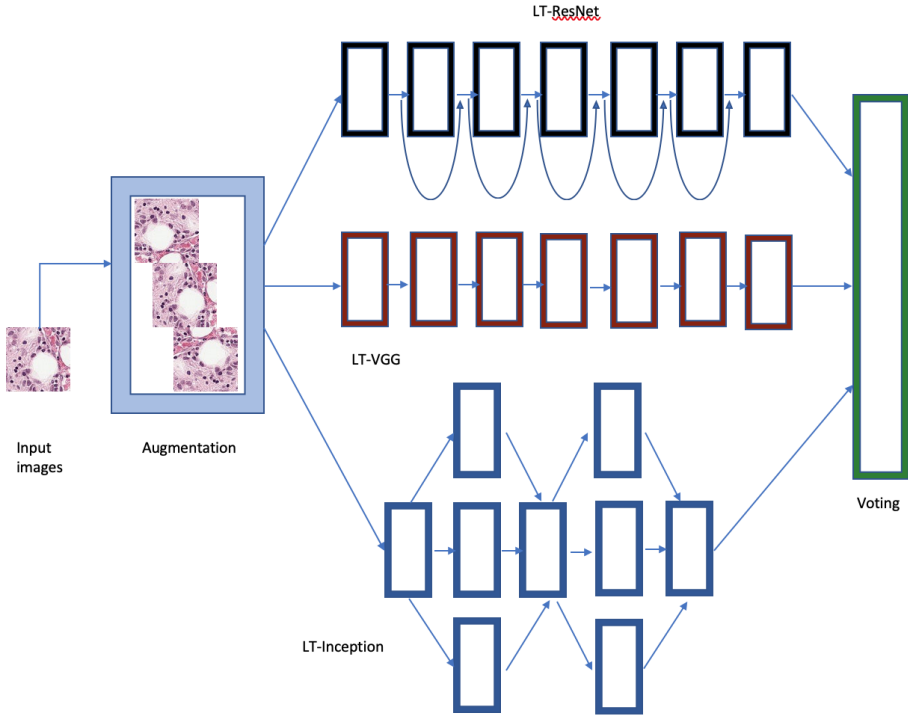


Figure 3. Achitecture of the ensemble model.

advantage of implementing the batch size system in the training process.

Since we used more than eight convolutions with non-linear activation, we decided to use the Normal Distribution developed by He et al. [8] as initial weights during the training process. To optimize the training process, we took advantage of the ADAM optimizer by setting the learning rate at $1e-4$ and reduce by $1e-6$ for each subsequent epoch.

To measure the performance of the model, we have calculated its accuracy, precision and recall score [3, 14]. It can be derived using the following formulas:

$$\text{Accuracy} = \frac{TP}{TP + FP},$$

$$\text{Precision} = \frac{TP + TN}{TP + TN + FP + FN},$$

$$\text{Recall} = \frac{TP}{TP + FN},$$

where TP stands for true positive, and this value was taken from the data in class 0 (no cancer) and predicted to be accurate as class 0. TN is for true negative, that is, data was on class 1 (cancer) and correctly predicted as a member of class 1. Conversely, FP is an abbreviation of false positive, where FP is a member of class

1, which is wrongly predicted as a member of class 0. And lastly, FN is for false negative, which is a member of class 0 that was wrongly predicted as a member of class 1.

We also measured the loss score that represents how far the model is from the target. To calculate the loss score, we used the cross-entropy for the Softmax loss function with two classes target. By having formula (3.1), the softmax loss function will become:

$$\text{CE} = - \sum_i^K t_i \log(f(s)_i). \quad (3.2)$$

Equation (3.2) explains that cross-entropy CE is the sum of ground truth t_i logarithm the CNN score of each class that represents by $f(s)_i$.

The ensemble process is to train the three models separately, then we vote. The first type of voting used is simple majority voting. Here, we do not pay attention to each model's achievement in the training process. In other words, each model gets the same portion in the voting process. The second type of voting is that we provide different portions for each model. We tried some combinations of weights considering the individual accuracies of the ensemble members. The results show that an optimal choice of weights is 0.35 for the two best networks and 0.3 for the third network. So, we set LT-ResNet and LT-VGG having weights 0.35 and LT-Inception 0.30. Voting system itself refers to [5, 13], if we have multiple scores x_1, x_2, \dots, x_n , with corresponding weights w_1, w_2, \dots, w_n , then the weighted mean can be calculated through

$$\bar{x} = \frac{\sum_{i=1}^n w_i x_i}{\sum_{i=1}^n w_i}.$$

4. Results

4.1. Loss score

Figure 4 illustrates the loss score of the three models while training. Graph *a* shows that the LT-ResNet model's loss score has a stable movement, likewise in graph *b*, which displays a decrease in the loss score, which is also stable from the LT-Inception model. Meanwhile, the LT-VGG model shows the unsteady movement of reducing the loss score, as shown in graph *c*. Figure 4 shows the three models' loss scores, respectively, LT-ResNet 0.1324, LT-Inception 0.1937, and LT-VGG 0.2689 at the last epoch.

4.2. Accuracy, precision and recall

Figure 5 describes the training process of the three base models. From this figure, we can see the accuracy and validation score of the models. These three graphs show a significant increase in accuracy from the first epoch to the 50 epochs. The consistently smaller differences between the training and validation accuracies (blue,

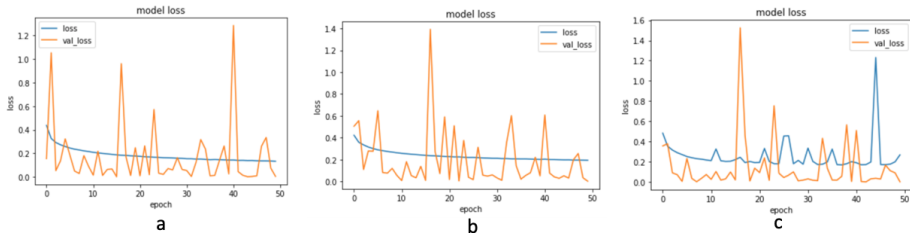


Figure 4. Loss of: (a) LT-ResNet, (b) LT-Inception, (c) LT-VGG.

yellow lines on Figure 5, respectively) prove that the model is not overfitting. The performance of the LT-ResNet model is shown in graph *a*, with an accuracy score of 0.95. Meanwhile, the LT-Inception model's performance is shown in graph *b*, with an accuracy score of 0.93. The LT-VGG model also has a good performance, as shown in graph *c*, with an accuracy score of 0.95. The training process's complete results, which include the accuracy, precision, and recall scores of the three models, are presented in Table 2.

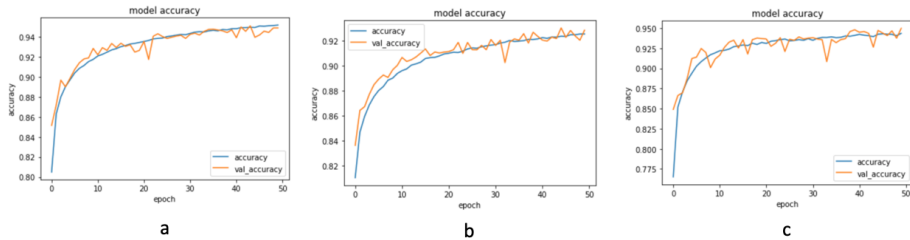


Figure 5. Accuracy of: (a) LT-Resnet, (b) LT-Inception, (c) LT-VGG.

Table 2. Precision, Recall and Accuracy of the investigated models.

x	LT-ResNet	LT-Inception	LT-VGG	MV	WMV
Pre 0	0.95	0.92	0.95	0.95	0.95
Rec 0	0.97	0.96	0.97	0.98	0.98
Pre 1	0.95	0.93	0.96	0.96	0.96
Rec 1	0.92	0.89	0.92	0.93	0.93
Acc	0.95	0.93	0.95	0.96	0.96

After getting the results from these three models, we proceed by using the voting method as an implementation of the ensemble model. The majority voting (MV) results and the weighted majority voting (WMV) results show an equivalent quality in the calculation of each class. We have precision scores of 0.95 and 0.96,

respectively, for Class 0 and Class 1. Recall scores apiece 0.98 and 0.93 for Class 0 and Class 1. Finally accuracy score of these two voting systems corrects the accuracy value of all single models, which is 0.96 for both class.

4.3. Confusion matrix

To see the performance of the models, we present their predictions on the validation set. In Table 3, we report the predicted results of the three base models and two voting systems.

Table 3. Confusion matrix of the investigated models.

x	LT-ResNet	LT-Inception	LT-VGG	MV	WMV
TP	38043	37657	38186	38392	38397
TN	24677	23634	24504	24773	24787
FP	2020	3063	2193	1924	1910
FN	1265	1653	1124	918	913

From Table 3, we can see that if we compare the prediction results of the three base models, LT-ResNet model is superior in predicting class 1 and LT-VGG model in class 0. However, the ensemble model corrects the achievement of the three base models of around 200 to 300 images per class. Overall, the weighted majority voting shows the best result with 38,397 images accurately predicted as class 0 and 24,787 images correctly predicted as class 1. On the other hand, there were 1910 images from class 1 that were mistakenly predicted as class 0, and only 913 images in class 0 were incorrectly predicted as members of class 1.

5. Conclusion

From this study, we can conclude that the ensemble method can be used to improve the model's accuracy. It can be seen from the work of Kassani and ours compared to Veeling and Xia's works in Table 4. In this case, we experienced that the weighting method had no significant impact on the voting process. It can be seen from the equal accuracy score for the two ensemble models. Developing a network from scratch can be leveraged to reduce the complexity and depth of the architecture without compromising the network's quality. This can be seen from the comparison of the accuracy of our work with Kassani's.

From some of our references, several methods might be considered to be used in future work. One of them is the hyperparameter tuning method. The grid search method seems considerable to determine hyperparameters automatically. However, considering machine capability, we cannot use it at this time, and instead, we specify the parameters manually. Another thing that can be considered is the weighting method for the final decision, which can be part of the training

parameters. In other words, the user does not need to determine the weight of each single model, but the training process itself determines which model has the most influence on the ensemble model.

Table 4. Comparison results.

Method	Architecture	Accuracy
Veeling et al.	P4M-DenseNet	89.8%
Xia et al.	GoogleLeNet fine-tuned	84.3%
Kassani et al	Ensemble	94.64%
Proposed method	Ensemble	96%

Acknowledgments. This work was supported in part by the project EFOP-3.6.3-VEKOP-16-2017-00002, supported by the European Union, co-financed by the European Social Fund. Research was also supported by the ÚNKP-19-3-I. New National Excellence Program of the Ministry for Innovation and Technology.

This study was funded by LPDP Indonesia in the form of a doctoral scholarship (<https://www.lpdp.kemenukeu.go.id>)

References

- [1] B. E. BEJNORDI, M. VETA, P. J. VAN DIEST, B. VAN GINNEKEN, N. KARSEMELJER, G. LITJENS, J. A. W. M. VAN DER LAAK, the CAMELYON16 CONSORTIUM: *Diagnostic Assessment of Deep Learning Algorithms for Detection of Lymph Node Metastases in Women With Breast Cancer*, JAMA 318.22 (2017), pp. 2199–2210, DOI: <https://doi.org/10.1001/jama.2017.14585>.
- [2] F. BRAY, J. FERLAY, I. SOERJOMATARAM, R. SIEGEL, L. TORRE, A. JEMAL: *Global Cancer Statistics 2018: GLOBOCAN Estimates of Incidence and Mortality Worldwide for 36 Cancers in 185 Countries*, CA: A Cancer Journal For Clinicians 68.6 (2018), pp. 394–424, DOI: <https://doi.org/10.3322/caac.21492>.
- [3] T. FAWCETT: *An Introduction to ROC Analysis*, Pattern Recognition Letters 27.8 (2006), pp. 861–874, DOI: <https://doi.org/10.1016/j.patrec.2005.10.010>.
- [4] I. GOODFELLOW, Y. BENGIO, A. COURVILLE: *Deep Feedforward Networks*, in: Deep Learning, USA: MIT press, 2016, p. 181.
- [5] J. GROSSMAN, M. GROSSMAN, R. KATZ, in: *The First System of Weighted Differential and Integral Calculus, Non-Newtonian Calculus*, 2006.
- [6] B. HARANGI: *Skin Lesion classification With Ensemble of Deep Convolutional Neural Networks*, Journal of Biomedical Informatics 86 (2018), pp. 25–32, DOI: <https://doi.org/10.1016/j.jbi.2018.08.006>.
- [7] K. HE, X. ZHANG, S. REN, J. SUN: *Deep Residual Learning for Image Recognition*, in: Proceedings of the 2016 IEEE Conference on Computer Vision and Pattern Recognition, Las Vegas, NV, USA: IEEE, 2016, pp. 770–778, DOI: <https://doi.org/10.1109/CVPR.2016.90>.

- [8] K. HE, X. ZHANG, S. REN, J. SUN: *Delving Deep Into Rectifiers: Surpassing Human-Level Performance On Imagenet Classification*, in: Proceedings of the IEEE international conference on computer vision, Santiago, Chile: IEEE, 2015, pp. 1026–1034, DOI: <https://doi.org/10.1109/ICCV.2015.123>.
- [9] M. HOSNI, I. ABNANE, A. IDRI, J. M. C. DE GEA, J. L. F. ALEMAN: *Reviewing Ensemble Classification Methods in Breast Cancer*, Computer Methods and Programs in Biomedicine 177 (2019), pp. 89–112, DOI: <https://doi.org/10.1016/j.cmpb.2019.05.019>.
- [10] S. H. KASSANI, P. H. KASSANI, M. J. WESOLOWSKI, K. A. SCHNEIDER, R. DETERS: *Classification of Hispatology Biopsy Images Using Ensemble of Deep Learning Networks*, arXiv preprint arXiv:1909.11870 (2019).
- [11] A. KRIZHEVSKY, I. SUTSKEVER, G. HINTON: *ImageNet Classification with Deep Convolutional Neural Networks*, Communications of the ACM 60.6 (2017), pp. 1079–1105, DOI: <https://doi.org/10.1145/3065386>.
- [12] A. MADABHUSHI: *Digital Pathology Image Analysis: Opportunities and Challenges*, Imaging In medicine 1.1 (2009), pp. 7–10.
- [13] R. MESIAR, J. SPIRKOVA: *Weighted Means and Weighting Functions*, Kybernetika 42.2 (2006), pp. 151–160.
- [14] D. M. W. POWERS: *Evaluation: From Precision, Recall and F-measure to ROC, Informedness, Markedness and Correlation*, Journal of Machine Learning Technologies 2.1 (2011), pp. 37–63, DOI: <https://doi.org/10.9735/2229-3981>.
- [15] P. SADOWSKI: *Notes on back propagation* (2016), URL: <https://www.ics.uci.edu/pjsadows/notes.pdf>.
- [16] B. SAVELLI, A. BRIA, M. MOLINARA, C. MARROCCO, F. TORTORELLA: *A Multi-context CNN Ensemble For Small Lesion Detection*, Artificial Intelligence in Medicine 103 (2020), pp. 1–13, DOI: <https://doi.org/10.1016/j.artmed.2019.101749>.
- [17] H.-C. SHIN, H. ROTH, M. GAO, L. LU, Z. XU, I. NOGUES, J. YAO, D. MOLLURA, R. SUMMERS: *Deep Convolutional Neural Networks for Computer-Aided Detection: CNN Architectures, Dataset Characteristics and Transfer Learning*, IEEE Transactions on Medical Imaging 35.5 (2016), pp. 1285–1298, DOI: <https://doi.org/10.1109/tmi.2016.2528162>.
- [18] K. SIMONYAN, A. ZISSERMAN: *Very Deep Convolutional Networks for Large-Scale Image Recognition*, arXiv preprint arXiv:1409.1556.
- [19] C. SZEGEDY, W. LIU, Y. JIA, Y. JIA, S. REED, D. ANGUELOV, D. ERHAN, V. VANHOUCHE, A. RABINOVICH: *Going Deeper with Convolutions*, in: Proceedings of the 2015 IEEE Conference on Computer Vision and Pattern Recognition, Boston, MA, USA: IEEE, 2015, pp. 1–9, DOI: <https://doi.org/10.1109/cvpr.2015.7298594>.
- [20] B. S. VEELING, J. LINMANS, J. WINKENS, T. COHEN, M. WELLING: *Rotation Equivariant CNNs for Digital Pathology*, in: Proceedings on the 2018 Medical Image Computing and Computer Assisted Intervention, Spring, Cham, 2018, pp. 210–218, DOI: https://doi.org/10.1007/978-3-030-00934-2_24.
- [21] R. WEINBERG: *How Cancer Arises*, Scientific American 275.3 (1996), pp. 67–70.
- [22] R. WEISSELEDER: *Molecular Imaging in Cancer*, Science 312.5777 (2006), pp. 1168–1171, DOI: <https://doi.org/10.1126/science.1125949>.
- [23] T. XIA, A. KUMAR, D. FENG, J. KIM: *Patch-level Tumor Classification in Digital Hispatology Images with Domain Adapted Deep Learning*, in: 2018 40th Annual International Conference of the IEEE Engineering in Medicine and Biology Society (EMBC), Honolulu, HI, USA: IEEE, 2018, pp. 644–647, DOI: <https://doi.org/10.1109/EMBC.2018.8512353>.

On the intersection of Padovan, Perrin sequences and Pell, Pell-Lucas sequences

Salah Eddine Rihane^a, Alain Togbé^b

^aDepartment of Mathematics and Computer Science,
Abdelhafid Boussouf University, Mila 43000, Algeria
salahrihane@hotmail.fr

^bDepartment of Mathematics and Statistics, Purdue University Northwest,
1401 S, U.S. 421, Westville IN 46391, USA
atogbe@pnw.edu

Submitted: May 19, 2020

Accepted: March 19, 2021

Published online: April 6, 2021

Abstract

In this paper, we find all the Padovan and Perrin numbers which are Pell or Pell-Lucas numbers.

Keywords: Padovan numbers, Perrin numbers, Pell numbers, Pell-Lucas numbers, Linear form in logarithms, reduction method.

AMS Subject Classification: 11B39, 11J86.

1. Introduction

Let (u_n) and (v_n) be two linear recurrent sequences. The problem of finding the common terms of (u_n) and (v_n) was treated in [4, 5, 7–9]. They proved, under some assumption, that the Diophantine equation

$$u_n = v_m$$

has only finitely many integer solutions (m, n) . The aim of this paper is to study the common terms of Padovan, Perrin, Pell and Pell-Lucas sequences that we will recall below.

Let $\{P_m\}_{m \geq 0}$ be the Pell sequence given by

$$P_{m+2} = 2P_{m+1} + P_m,$$

for $m \geq 0$, where $P_0 = 0$ and $P_1 = 1$. This is the sequence A000129 in the OEIS and its first few terms are

$$0, 1, 2, 5, 12, 29, 70, 169, 408, 985, 2378, 5741, 13860, 33461, 80782, 195025, \dots$$

We let $\{Q_m\}_{m \geq 0}$ be the companion Lucas sequence of the Pell sequence also called the sequence of Pell–Lucas numbers. It starts with $Q_0 = 2$, $Q_1 = 2$ and obeys the same recurrence relation

$$Q_{m+2} = 2Q_{m+1} + Q_m, \quad \text{for all } m \geq 0$$

as the Pell sequence. This is the sequence A002203 in the OEIS and its first few terms are

$$2, 2, 6, 14, 34, 82, 198, 478, 1154, 2786, 6726, 16238, 39202, 94642, 228486, 551614, \dots$$

The Padovan sequence $\{\mathcal{P}_n\}_{n \geq 0}$ is defined by

$$\mathcal{P}_{n+3} = \mathcal{P}_{n+1} + \mathcal{P}_n,$$

for $n \geq 0$, where $\mathcal{P}_0 = 0$ and $\mathcal{P}_1 = \mathcal{P}_2 = 1$. This is the sequence A000931 in the OEIS. A few terms of this sequence are

$$0, 1, 1, 1, 2, 2, 3, 4, 5, 7, 9, 12, 16, 21, 28, 37, 49, 65, 86, 114, 151, 200, \dots$$

Let $\{E_n\}_{n \geq 0}$ be the Perrin sequence given by

$$E_{n+3} = E_{n+1} + E_n,$$

for $n \geq 0$, where $E_0 = 3$, $E_1 = 0$ and $E_2 = 2$. Its first few terms are

$$3, 0, 2, 3, 2, 5, 5, 7, 10, 12, 17, 22, 29, 39, 51, 68, 90, 119, 158, 209, 277, \dots$$

It is the sequence A001608 in the OEIS.

The proofs of our main theorems are mainly based on linear forms in logarithms of algebraic numbers and a reduction algorithm originally introduced by Baker and Davenport in [1]. Here, we use a version due to de Weger [3]. We organize this paper as follows. In Section 2, we recall the important results that will be used to prove our main results. Sections 4–6 are devoted to the statements and the proofs of our main results.

2. The tools

In this section, we recall all the tools that we will use to prove our main results.

2.1. Linear forms in logarithms

We need some results from the theory of lower bounds for nonzero linear forms in logarithms of algebraic numbers. We start by recalling Theorem 9.4 of [2], which is a modified version of a result of Matveev [6]. Let \mathbb{L} be an algebraic number field of degree $d_{\mathbb{L}}$. Let $\eta_1, \eta_2, \dots, \eta_l \in \mathbb{L}$ not 0 or 1 and d_1, \dots, d_l be nonzero integers. We put

$$D = \max\{|d_1|, \dots, |d_l|\},$$

and

$$\Gamma = \prod_{i=1}^l \eta_i^{d_i} - 1.$$

Let A_1, \dots, A_l be positive integers such that

$$A_j \geq h'(\eta_j) := \max\{d_{\mathbb{L}}h(\eta_j), |\log \eta_j|, 0.16\}, \quad \text{for } j = 1, \dots, l,$$

where for an algebraic number η of minimal polynomial

$$f(X) = a_0(X - \eta^{(1)}) \cdots (X - \eta^{(k)}) \in \mathbb{Z}[X]$$

over the integers with positive a_0 , we write $h(\eta)$ for its Weil height given by

$$h(\eta) = \frac{1}{k} \left(\log a_0 + \sum_{j=1}^k \max\{0, \log |\eta^{(j)}|\} \right).$$

The following consequence of Matveev's theorem is Theorem 9.4 in [2].

Theorem 2.1. *If $\Gamma \neq 0$ and $\mathbb{L} \subseteq \mathbb{R}$, then*

$$\log |\Gamma| > -1.4 \cdot 30^{l+3} l^{4.5} d_{\mathbb{L}}^2 (1 + \log d_{\mathbb{L}}) (1 + \log D) A_1 A_2 \cdots A_l.$$

2.2. The de Weger reduction

Here, we present a variant of the reduction method of Baker and Davenport due to de Weger [3].

Let $\vartheta_1, \vartheta_2, \beta \in \mathbb{R}$ be given, and let $x_1, x_2 \in \mathbb{Z}$ be unknowns. Let

$$\Lambda = \beta + x_1 \vartheta_1 + x_2 \vartheta_2. \tag{2.1}$$

Let c, μ be positive constants. Set $X = \max\{|x_1|, |x_2|\}$. Let X_0, Y be positive. Assume that

$$|\Lambda| < c \cdot \exp(-\mu \cdot Y), \tag{2.2}$$

$$Y \leq X \leq X_0. \tag{2.3}$$

When $\beta = 0$ in (2.1), we get

$$\Lambda = x_1 \vartheta_1 + x_2 \vartheta_2.$$

Put $\vartheta = -\vartheta_1/\vartheta_2$. We assume that x_1 and x_2 are coprime. Let the continued fraction expansion of ϑ be given by

$$[a_0, a_1, a_2, \dots],$$

and let the k th convergent of ϑ be p_k/q_k for $k = 0, 1, 2, \dots$. We may assume without loss of generality that $|\vartheta_1| < |\vartheta_2|$ and that $x_1 > 0$. We have the following results.

Lemma 2.2 (See Lemma 3.2 in [3]). *Let*

$$A = \max_{0 \leq k \leq Y_0} a_{k+1},$$

where

$$Y_0 = -1 + \frac{\log(\sqrt{5}X_0 + 1)}{\log\left(\frac{1+\sqrt{5}}{2}\right)}.$$

If (2.2) and (2.3) hold for x_1, x_2 and $\beta = 0$, then

$$Y < \frac{1}{\mu} \log\left(\frac{c(A+2)X_0}{|\vartheta_2|}\right).$$

When $\beta \neq 0$ in (2.1), put $\vartheta = -\vartheta_1/\vartheta_2$ and $\psi = \beta/\vartheta_2$. Then, we have

$$\frac{\Lambda}{\vartheta_2} = \psi - x_1\vartheta + x_2.$$

Let p/q be a convergent of ϑ with $q > X_0$. For a real number x , we let $\|x\| = \min\{|x - n|, n \in \mathbb{Z}\}$ be the distance from x to the nearest integer. We have the following result.

Lemma 2.3 (See Lemma 3.3 in [3]). *Suppose that*

$$\|q\psi\| > \frac{2X_0}{q}.$$

Then, the solutions of (2.2) and (2.3) satisfy

$$Y < \frac{1}{\mu} \log\left(\frac{q^2c}{|\vartheta_2|X_0}\right).$$

2.3. Properties of Padovan and Perrin sequences

In this subsection, we recall some facts and properties of the Padovan and the Perrin sequences which will be used later.

The characteristic equation

$$x^3 - x - 1 = 0,$$

has roots $\alpha, \beta, \gamma = \bar{\beta}$, where

$$\alpha = \frac{r_1 + r_2}{6}, \quad \beta = \frac{-r_1 - r_2 + i\sqrt{3}(r_1 - r_2)}{12},$$

and

$$r_1 = \sqrt[3]{108 + 12\sqrt{69}} \quad \text{and} \quad r_2 = \sqrt[3]{108 - 12\sqrt{69}}.$$

Let

$$\begin{aligned} c_\alpha &= \frac{(1 - \beta)(1 - \gamma)}{(\alpha - \beta)(\alpha - \gamma)} = \frac{1 + \alpha}{-\alpha^2 + 3\alpha + 1}, \\ c_\beta &= \frac{(1 - \alpha)(1 - \gamma)}{(\beta - \alpha)(\beta - \gamma)} = \frac{1 + \beta}{-\beta^2 + 3\beta + 1}, \\ c_\gamma &= \frac{(1 - \alpha)(1 - \beta)}{(\gamma - \alpha)(\gamma - \beta)} = \frac{1 + \gamma}{-\gamma^2 + 3\gamma + 1} = \overline{c_\beta}. \end{aligned}$$

The Binet's formula of \mathcal{P}_n is

$$\mathcal{P}_n = c_\alpha \alpha^n + c_\beta \beta^n + c_\gamma \gamma^n, \quad \text{for all } n \geq 0, \tag{2.4}$$

and that of E_n is

$$E_n = \alpha^n + \beta^n + \gamma^n, \quad \text{for all } n \geq 0. \tag{2.5}$$

Numerically, we have

$$\begin{aligned} 1.32 &< \alpha < 1.33, \\ 0.86 &< |\beta| = |\gamma| < 0.87, \\ 0.72 &< c_\alpha < 0.73, \\ 0.24 &< |c_\beta| = |c_\gamma| < 0.25. \end{aligned}$$

It is easy to check that

$$|\beta| = |\gamma| = \alpha^{-1/2}.$$

Further, using induction, we can prove that

$$\alpha^{n-2} \leq \mathcal{P}_n \leq \alpha^{n-1}, \quad \text{holds for all } n \geq 4 \tag{2.6}$$

and

$$\alpha^{n-2} \leq E_n \leq \alpha^{n+1}, \quad \text{holds for all } n \geq 2. \tag{2.7}$$

2.4. Properties of Pell and Pell-Lucas sequences

Let $\delta = 1 + \sqrt{2}$ and $\bar{\delta} := 1 - \sqrt{2}$ be the roots of the characteristic equation $x^2 - 2x - 1$ of P_m and Q_m . The Binet formula of P_m is given by

$$P_m = \frac{\delta^m - \bar{\delta}^m}{2\sqrt{2}}, \quad \text{for all } m \geq 0, \tag{2.8}$$

and that of Q_m is

$$Q_m = \delta^m + \bar{\delta}^m, \quad \text{for all } m \geq 0. \quad (2.9)$$

Moreover, we have

$$\delta^{m-2} < P_m < \delta^{m-1}, \quad \text{for all } m \geq 2, \quad (2.10)$$

and

$$\delta^{m-1} < Q_m < \delta^{m+1}, \quad \text{for all } m \geq 2. \quad (2.11)$$

3. Padovan numbers which are Pell numbers

In this section, we will prove our first main result, which is the following.

Theorem 3.1. *The only solutions of the Diophantine equation*

$$\mathcal{P}_n = P_m \quad (3.1)$$

in positive integers m and n are

$$(n, m) \in \{(0, 0), (1, 1), (2, 1), (3, 1), (4, 2), (5, 2), (8, 3), (11, 4)\}.$$

Hence, $\mathcal{P} \cap P = \{0, 1, 2, 5, 12\}$.

Proof. A quick computation with Maple reveals that the solutions of the Diophantine equation (3.1) in the interval $[0, 60]$ are the solutions cited in Theorem 3.1.

From now, assuming that $n > 60$, then by (2.6) and (2.10), we have

$$\alpha^{n-2} < \delta^{m-1} \quad \text{and} \quad \delta^{m-2} < \alpha^{n-1}.$$

Thus, we get

$$(n-2)c_1 + 1 < m < (n-1)c_1 + 2, \quad \text{where } c_1 := \log \alpha / \log \delta.$$

Particularly, we have $n < 4m$. So to solve equation (3.1), it suffices to get a good upper bound on m .

Equation (3.1) can be expressed as

$$c_\alpha \alpha^n - \frac{\delta^m}{2\sqrt{2}} = -c_\beta \beta^n - c_\gamma \gamma^n - \frac{\bar{\delta}^m}{2\sqrt{2}},$$

by using (2.4) and (2.8). Thus, we get

$$\left| c_\alpha \alpha^n - \frac{\delta^m}{2\sqrt{2}} \right| = \left| c_\beta \beta^n + c_\gamma \gamma^n + \frac{\bar{\delta}^m}{2\sqrt{2}} \right| < 0.85.$$

Multiplying through by $2\sqrt{2}\delta^{-m}$, we obtain

$$\left| (c_\alpha 2\sqrt{2})\alpha^n \delta^{-m} - 1 \right| < 2.41\delta^{-m}. \quad (3.2)$$

Now, we apply Matveev’s theorem by choosing

$$\Lambda_1 = 2\sqrt{2}c_\alpha \alpha^n \delta^{-m} - 1$$

and

$$\eta_1 := 2\sqrt{2}c_\alpha, \quad \eta_2 := \alpha, \quad \eta_3 := \delta, \quad d_1 := 1, \quad d_2 := n, \quad d_3 := -m.$$

The algebraic numbers η_1, η_2 and η_3 belong to $\mathbb{K} := \mathbb{Q}(\alpha, \delta)$ for which $d_{\mathbb{K}} = 6$. Since $n < 4m$, therefore we can take $D := 4m = \max\{1, m, n\}$. Furthermore, we have

$$h(\eta_2) = \frac{\log \alpha}{3} \quad \text{and} \quad h(\eta_3) = \frac{\log \delta}{2},$$

thus, we can take

$$\max\{6h(\eta_2), |\log \eta_2|, 0.16\} < 0.58 := A_2$$

and

$$\max\{6h(\eta_3), |\log \eta_3|, 0.16\} = 2.65 := A_3.$$

On the other hand, the conjugates of η_1 are $\pm 2\sqrt{2}c_\alpha, \pm 2\sqrt{2}c_\beta$ and $\pm 2\sqrt{2}c_\gamma$, so the minimal polynomial of η_1 is

$$(x^2 - 8c_\alpha^2)(x^2 - 8c_\beta^2)(x^2 - 8c_\gamma^2) = \frac{529x^6 - 2024x^4 - 640x^2 - 512}{529}.$$

Since $2\sqrt{2}c_\alpha > 1$ and $|2\sqrt{2}c_\beta| = |2\sqrt{2}c_\gamma| < 1$, then we get

$$h(\eta_1) = \frac{\log 529 + 2 \log(2\sqrt{2}c_\alpha)}{6}.$$

So, we can take

$$\max\{6h(\eta_1), |\log \eta_1|, 0.16\} < 7.8 := A_1.$$

To apply Matveev’s theorem, we still need to prove that $\Lambda_1 \neq 0$. Assume the contrary, i.e. $\Lambda_1 = 0$. So, we get

$$\delta^m = 2\sqrt{2}c_\alpha \alpha^n.$$

Conjugating the above relation using the \mathbb{Q} -automorphism of Galois σ defined by $\sigma = (\alpha\beta)$ and taking the absolute value we obtain

$$1 < \delta^m = 2\sqrt{2}|c_\beta| |\beta|^n < 1,$$

which is a contradiction. Thus $\Lambda_1 \neq 0$.

Matveev’s theorem tells us that

$$\begin{aligned} \log |\Lambda_1| &> -1.4 \times 30^6 \times 3^{4.5} \times 6^2 (1 + \log 6)(1 + \log 4m) \times 7.8 \times 0.58 \times 2.65 \\ &> -1.8 \times 10^{14} \times (1 + \log 4m). \end{aligned}$$

The last inequality together with (3.2) leads to

$$m < 1.99 \times 10^{14}(1 + \log 4m).$$

Thus, we obtain

$$m < 7.52 \times 10^{15}. \quad (3.3)$$

Now, we will use Lemma 2.3 to reduce the upper bound (3.3) on m .

Define

$$\Gamma_1 = n \log \alpha - m \log \delta + \log(2\sqrt{2}c_\alpha).$$

Clearly, we have $e^{\Gamma_1} - 1 = \Lambda_1$. Since $\Lambda_1 \neq 0$, then $\Gamma_1 \neq 0$. If $\Gamma_1 > 0$ then we get

$$0 < \Gamma_1 < e^{\Gamma_1} - 1 = |e^{\Gamma_1} - 1| = |\Lambda_1| < 2.41\delta^{-m}.$$

If $\Gamma_1 < 0$, so we have $1 - e^{\Gamma_1} = |e^{\Gamma_1} - 1| = |\Lambda_1| < 1/2$, because $n > 60$. Then $e^{|\Gamma_1|} < 2$. Thus, one can see that

$$0 < |\Gamma_1| < e^{|\Gamma_1|} - 1 = e^{|\Gamma_1|} |\Lambda_1| < 4.82\delta^{-m}.$$

From both cases, we deduce that

$$0 < \left| n(-\log \alpha) + m \log \delta - \log(2\sqrt{2}c_\alpha) \right| < 4.82 \exp(-0.88 \times m).$$

The inequality (3.3) implies that we can take $X_0 := 3.01 \times 10^{16}$. Furthermore, we can choose

$$c := 4.82, \quad \mu := 0.88, \quad \psi := -\frac{\log(2\sqrt{2}c_\alpha)}{\log \mu},$$

$$\vartheta := \frac{\log \alpha}{\log \delta}, \quad \vartheta_1 := -\log \alpha, \quad \vartheta_2 := \log \delta, \quad \beta := -\log(2\sqrt{2}c_\alpha).$$

With the help of Maple, we find that

$$q_{29} = 3860032780734237233$$

satisfies the hypotheses of Lemma 2.3. Furthermore, Lemma 2.3 tells us

$$m < \frac{1}{0.88} \log \left(\frac{3860032780734237233^2 \times 4.82}{\log \delta \times 3.01 \times 10^{16}} \right) \leq 57.$$

This contradicts the assumption that $n > 60$. Therefore, the theorem is proved. \square

4. Padovan numbers which are Pell-Lucas numbers

Our second result will be stated and proved in this section.

Theorem 4.1. *The only solutions of the Diophantine equation*

$$\mathcal{P}_n = Q_m \tag{4.1}$$

in positive integers m and n are

$$(n, m) \in \{(4, 0), (4, 1), (5, 0), (5, 1)\}.$$

Hence, we deduce that $\mathcal{P} \cap Q = \{2\}$.

Proof. A quick computation with Maple reveals that the solutions of the Diophantine equation (4.1) in the interval $[0, 60]$ are those cited in Theorem 4.1.

From now, we suppose that $n > 60$, then by (2.6) and (2.11), we have

$$\alpha^{n-2} < \delta^{m+1} \quad \text{and} \quad \delta^{m-1} < \alpha^{n-1}.$$

Thus, we get

$$(n - 2)c_1 - 1 < m < (n - 1)c_1 + 1, \quad \text{where } c_1 := \log \alpha / \log \delta.$$

Particularly, we have $n < 4m$. So, to solve equation (4.1), we will determine a good upper bound on m .

By using (2.4) and (2.9), equation (4.1) can be rewritten into the form

$$c_\alpha \alpha^n - \delta^m = -c_\beta \beta^n - c_\gamma \gamma^n - \bar{\delta}^m$$

So, we obtain

$$|c_\alpha \alpha^n - \delta^m| \leq 2|c_\beta \beta^n| + 1 < 1.5.$$

Multiplying both sides by δ^{-m} , we get

$$|c_\alpha \alpha^n \delta^{-m} - 1| < 1.5\delta^{-m}. \tag{4.2}$$

Now, we will apply Matveev's theorem to

$$\Lambda_2 = c_\alpha \alpha^n \delta^{-m} - 1$$

by taking

$$\eta_1 := c_\alpha, \quad \eta_2 := \alpha, \quad \eta_3 := \delta, \quad d_1 := 1, \quad d_2 := n, \quad d_3 := -m.$$

The algebraic numbers η_1 , η_2 and η_3 belong to $\mathbb{K} := \mathbb{Q}(\alpha, \delta)$ with $d_{\mathbb{K}} = 6$. As above, we take

$$D = 4m, \quad A_2 = 0.58, \quad A_3 = 2.65.$$

On the other hand, the minimal polynomial of c_α is

$$23x^3 - 23x^2 - 6x - 1,$$

which has roots c_α , c_β and c_γ . Since $c_\alpha < 1$ and $|c_\beta| = |c_\gamma| < 1$, then we get

$$h(\eta_1) = \frac{\log 23}{3}.$$

So, we can take

$$\max\{6h(\eta_1), |\log \eta_1|, 0.16\} < 6.28 := A_1.$$

To apply Matveev's theorem, we will prove that $\Lambda_2 \neq 0$. Suppose the contrary, i.e $\Lambda_2 = 0$. Thus, we get

$$\delta^m = c_\alpha \alpha^n.$$

Conjugating the above relation using the \mathbb{Q} -automorphism of Galois σ defined by $\sigma = (\alpha\beta)$ and taking the absolute value, we obtain

$$1 < \delta^m = |c_\beta| |\beta|^n < 1,$$

which is a contradiction. Thus, we deduce that $\Lambda_2 \neq 0$.

We use Matveev's theorem to obtain

$$\begin{aligned} \log |\Lambda_2| &> -1.4 \times 30^6 \times 3^{4.5} \times 6^2 (1 + \log 6) (1 + \log 4m) \times 6.28 \times 0.58 \times 2.65 \\ &> -1.39 \times 10^{14} (1 + \log 4m). \end{aligned}$$

The last inequality together with (4.2) leads to

$$m < 1.58 \times 10^{14} (1 + \log 4m).$$

Thus, we obtain

$$m < 6.05 \times 10^{15}. \tag{4.3}$$

Now, we will use Lemma 2.3 to reduce the upper bound (4.3) on m .

Putting

$$\Gamma_2 = n \log \alpha - m \log \delta + \log(c_\alpha),$$

we proceed like in Section 3 to obtain

$$0 < |n(-\log \alpha) + m \log \delta - \log(c_\alpha)| < 3 \exp(-0.88 \times m).$$

Using inequality (4.3), we take $X_0 := 2.42 \times 10^{16}$. Moreover, we choose

$$c := 3, \quad \mu := 0.88, \quad \psi := -\frac{\log(c_\alpha)}{\log \mu},$$

$$\vartheta := \frac{\log \alpha}{\log \delta}, \quad \vartheta_1 := -\log \alpha, \quad \vartheta_2 := \log \delta, \quad \beta := -\log(c_\alpha).$$

We use Maple to find that

$$q_{29} = 3860032780734237233$$

satisfies the hypotheses of Lemma 2.3. Therefore, we get

$$m < \frac{1}{0.88} \log \left(\frac{3860032780734237233^2 \times 3}{\log \delta \times 2.42 \times 10^{16}} \right) \leq 56.$$

This contradicts the assumption that $n > 60$. Therefore, the proof of Theorem 4.1 is complete. \square

5. Perrin numbers which are Pell numbers

In this section, we will state and prove our third main result.

Theorem 5.1. *The only solutions of the Diophantine equation*

$$E_n = P_m \tag{5.1}$$

in positive integers m and n are

$$(n, m) \in \{(0, 1), (2, 2), (4, 2), (5, 3), (6, 3), (9, 4), (8, 3), (12, 5)\}.$$

Hence, this implies that $E \cap P = \{0, 2, 5, 12, 29\}$.

Proof. A quick computation with Maple gives the solutions of the Diophantine equation (5.1) in the interval $[0, 55]$, cited in Theorem 5.1.

From now, assuming that $n > 55$, then by (2.7) and (2.10), we have

$$\alpha^{n-2} < \delta^{m-1} \quad \text{and} \quad \delta^{m-2} < \alpha^{n+1}.$$

Thus, we get

$$(n - 2)c_1 + 1 < m < (n + 1)c_1 + 2, \quad \text{where } c_1 := \log \alpha / \log \delta.$$

Particularly, we have $n < 4m$. So to solve equation (5.1), we will determine a good upper bound on m .

By using (2.5) and (2.8), equation (5.1) can be expressed as

$$\alpha^n - \frac{\delta^m}{2\sqrt{2}} = -\beta^n - \gamma^n - \frac{\bar{\delta}^m}{2\sqrt{2}}.$$

Thus, we get

$$\left| \alpha^n - \frac{\delta^m}{2\sqrt{2}} \right| = \left| \beta^n + \gamma^n + \frac{\bar{\delta}^m}{2\sqrt{2}} \right| < 2.36.$$

Dividing through by $\delta^m/(2\sqrt{2})$, we obtain

$$\left| 2\sqrt{2}\alpha^n\delta^{-m} - 1 \right| < 6.68\delta^{-m}. \tag{5.2}$$

Now, we apply Matveev's theorem to

$$\Lambda_3 = 2\sqrt{2}\alpha^n\delta^{-m} - 1$$

and take

$$\eta_1 := 2\sqrt{2}, \quad \eta_2 := \alpha, \quad \eta_3 := \delta, \quad d_1 := 1, \quad d_2 := n, \quad d_3 := -m.$$

The algebraic numbers η_1 , η_2 and η_3 belong to $\mathbb{K} := \mathbb{Q}(\alpha, \delta)$, with $d_{\mathbb{K}} = 6$. As before we can take

$$D = 4m, \quad A_2 = 0.58 \quad \text{and} \quad A_3 = 2.65$$

Furthermore, since $h(\eta_1) = \log(2\sqrt{2})$, we choose

$$\max\{6h(\eta_1), |\log \eta_1|, 0.16\} < 6.24 := A_1.$$

Similarly to what was done above, one can check that $\Lambda_3 \neq 0$. We deduce from Matveev's theorem that

$$\begin{aligned} \log |\Lambda_3| &> -1.4 \times 30^6 \times 3^{4.5} \times 6^2 (1 + \log 6)(1 + \log 4m) \times 6.24 \times 0.58 \times 2.65 \\ &> -1.39 \times 10^{14} \times (1 + \log 4m). \end{aligned}$$

The last inequality together with (5.2) leads to

$$m < 1.57 \times 10^{14} (1 + \log 4m).$$

Thus, we solve the above inequality to obtain

$$m < 6.1 \times 10^{15}. \tag{5.3}$$

Now, we will use Lemma 2.3 to reduce the upper bound (5.3) on m .

Define

$$\Gamma_3 = n \log \alpha - m \log \delta + \log(2\sqrt{2}).$$

Like above, we use Γ_3 to obtain

$$0 < \left| n(-\log \alpha) + m \log \delta - \log(2\sqrt{2}) \right| < 13.36 \exp(-0.88 \times m)$$

Inequality (5.3) implies $X_0 := 2.44 \times 10^{16}$. Now, we take

$$c := 13.36, \quad \mu := 0.88, \quad \psi := -\frac{\log(2\sqrt{2})}{\log \mu},$$

$$\vartheta := \frac{\log \alpha}{\log \delta}, \quad \vartheta_1 := -\log \alpha, \quad \vartheta_2 := \log \delta, \quad \beta := -\log(2\sqrt{2}).$$

We use Maple to see that

$$q_{28} = 153529568750401532$$

satisfies the hypotheses of Lemma 2.3. Applying Lemma 2.3, we get

$$m < \frac{1}{0.88} \log \left(\frac{153529568750401532^2 \times 13.36}{\log \delta \times 2.44 \times 10^{16}} \right) \leq 51.$$

This contradicts the assumption that $n > 55$. Therefore, This completes the proof of Theorem 5.1. \square

6. Perrin numbers which are Pell-Lucas numbers

In this section, we will state and prove our last main result.

Theorem 6.1. *The only solutions of the Diophantine equation*

$$E_n = Q_m \tag{6.1}$$

in positive integers m and n are

$$(n, m) \in \{(2, 0), (2, 1), (4, 0), (4, 1)\}.$$

Hence, we see that $E \cap Q = \{2\}$.

Proof. A quick computation with Maple in the interval $[0, 50]$ gives the solutions of Diophantine equation (6.1) cited in Theorem 6.1.

We suppose that $n > 50$, then by (2.7) and (2.11), we have

$$\alpha^{n-2} < \delta^{m+1} \quad \text{and} \quad \delta^{m-1} < \alpha^{n+1}.$$

Thus, we get

$$(n - 2)c_1 - 1 < m < (n + 1)c_1 + 1, \quad \text{where } c_1 := \log \alpha / \log \delta.$$

Particularly, we have $n < 4m$. So to solve equation (6.1), We will find a good upper bound on m .

By using (2.5) and (2.9), one can see that equation (6.1) can be rewritten as

$$\alpha^n - \delta^m = -\beta^n - \gamma^n - \bar{\delta}^m.$$

We deduce that

$$|\alpha^n - \delta^m| \leq 2|\beta^n| + 1 < 3.$$

Dividing both sides by δ^m , we get

$$|\alpha^n \delta^{-m} - 1| < 3\delta^{-m}. \tag{6.2}$$

To apply Matveev's theorem to

$$\Lambda_4 = \alpha^n \delta^{-m} - 1,$$

we take

$$\eta_1 := \alpha, \quad \eta_2 := \delta, \quad d_1 := n, \quad d_2 := -m, \quad D = 4m, \quad A_1 = 0.58 \quad \text{and} \quad A_2 = 2.65.$$

Moreover, one can show that $\Lambda_4 \neq 0$. Thus, we apply Matveev's theorem to obtain

$$\begin{aligned} \log |\Lambda_4| &> -1.4 \times 30^5 \times 2^{4.5} \times 6^2 (1 + \log 6)(1 + \log 4m) \times 0.58 \times 2.65 \\ &> -1.19 \times 10^{11} (1 + \log 4m). \end{aligned}$$

The last inequality together with (6.2) implies

$$m < 1.35 \times 10^{11}(1 + \log 4m).$$

Thus, we obtain

$$m < 4.19 \times 10^{12}. \quad (6.3)$$

Now, we will use Lemma 2.2 to reduce the upper bound (6.3) on m .

Put

$$\Gamma_4 = n \log \alpha - m \log \delta.$$

We proceed as above and use Γ_4 to obtain

$$0 < |n(-\log \alpha) + m \log \delta| < 6 \exp(-0.88 \times m).$$

From inequality (6.3), we take $X_0 := 1.68 \times 10^{13}$. So, we have $Y := 63.95005\dots$. Moreover, we choose

$$c := 6, \quad \mu := 0.88, \quad \vartheta := \frac{\log \alpha}{\log \mu}, \quad \vartheta_1 := -\log \alpha, \quad \vartheta_2 := \log \mu.$$

With the help of Maple, we find that

$$\max_{0 \leq k \leq 64} a_{k+1} = 1029.$$

So, Lemma 2.2 gives

$$m < \frac{1}{0.88} \log \left(\frac{6 \times 1031 \times 1.68 \times 10^{13}}{\log \delta} \right) \leq 45.$$

This contradicts the assumption that $n > 50$. Therefore, Theorem 6.1 is completely proved. \square

Acknowledgements. The authors are grateful to the referee for the useful comments that help to improve the quality of the paper.

References

- [1] A. BAKER, H. DAVENPORT: *The equations $3x^2 - 2 = y^2$ and $8x^2 - 7 = z^2$* , Quart. J. Math. Oxford Ser. (2) 20 (1969), pp. 129–137, doi: <https://doi.org/10.1093/qmath/20.1.129>.
- [2] Y. BUGEAUD, M. MIGNOTTE, S. SIKSEK: *Classical and modular approaches to exponential Diophantine equations I. Fibonacci and Lucas perfect powers*, Annals of mathematics 163.3 (2006), pp. 969–1018.
- [3] B. M. DE WEGER: *Algorithms for Diophantine equations*, CWI tracts 65 (1989).
- [4] P. KISS: *On common terms of linear recurrences*, Acta Mathematica Academiae Scientiarum Hungarica 40.1-2 (1982), pp. 119–123.

- [5] M. LAURENT: *Équations exponentielles polynômes et suites récurrentes linéaires*, Astérisque 147.148 (1987), pp. 121–139.
- [6] E. M. MATVEEV: *An explicit lower bound for a homogeneous rational linear form in the logarithms of algebraic numbers. II*, *Izvestiya: Mathematics* 64.6 (2000), pp. 1217–1269.
- [7] M. MIGNOTTE: *Une extension du théorème de Skolem–Mahler*, *CR Acad. Sci. Paris* 288 (1979), pp. 233–235.
- [8] M. MIGNOTTE: *Intersection des images de certaines suites récurrentes linéaires*, *Theoretical Computer Science* 7.1 (1978), pp. 117–121.
- [9] H. P. SCHLICKWEI, W. M. SCHMIDT: *The intersection of recurrence sequences*, *Acta Arithmetica* 72.1 (1995), pp. 1–44.

The structure of the unit group of the group algebra $F(C_3 \times D_{10})$

Meena Sahai, Sheere Farhat Ansari*

Department of Mathematics and Astronomy,
University of Lucknow, Lucknow, U.P. 226007, India

meena_sahai@hotmail.com

sheere_farhat@rediffmail.com

Submitted: February 7, 2021

Accepted: September 7, 2021

Published online: September 19, 2021

Abstract

Let D_n be the dihedral group of order n . The structure of the unit group $U(F(C_3 \times D_{10}))$ of the group algebra $F(C_3 \times D_{10})$ over a finite field F of characteristic 3 is given in [10]. In this article, the structure of $U(F(C_3 \times D_{10}))$ is obtained over any finite field F of characteristic $p \neq 3$.

Keywords: Group rings, unit groups, dihedral groups, cyclic groups

AMS Subject Classification: 16U60, 20C05

1. Introduction

Let $U(FG)$ be the group of invertible elements of the group algebra FG of a group G over a field F . The study of units and their properties is one of the most challenging problems in the theory of group rings. Explicit calculations in $U(FG)$ are usually difficult, even when G is fairly small and F is a finite field. The results obtained in this direction are also useful for the investigation of the Lie properties of group rings, the isomorphism problem and other open questions in this area, see [2].

*The financial assistance provided to the second author in the form of a Senior Research Fellowship from the University Grants Commission, INDIA is gratefully acknowledged.

For a normal subgroup H of G , the natural homomorphism $G \rightarrow G/H$ can be extended to an F -algebra homomorphism from $FG \rightarrow F(G/H)$ defined by $\sum_{g \in G} a_g g \mapsto \sum_{g \in G} a_g gH$, $a_g \in F$. The kernel of this homomorphism, denoted by $\Delta(G, H)$, is the ideal generated by $\{h - 1 : h \in H\}$ in FG and $FG/\Delta(G, H) \cong F(G/H)$.

Let $J(FG)$ be the Jacobson radical of FG and let $V = 1 + J(FG)$. The F -algebra $FG/J(FG)$ is semisimple whenever G is a finite group. It is known from the Wedderburn structure theorem that

$$FG/J(FG) \cong \bigoplus_{i=1}^r M(n_i, K_i)$$

where r is the number of non-isomorphic irreducible FG modules, $n_i \in \mathbb{N}$ and K_i 's are finite dimensional division algebras over F . In this context two results by Ferraz [3, Theorem 1.3 and Prop 1.2] (stated at the end of this section) are very useful in determining the Wedderburn decomposition of $FG/J(FG)$.

If FG is semisimple, then $J(FG) = 0$ and by [8, Prop 3.6.11],

$$FG \cong F(G/G') \oplus \Delta(G, G')$$

where $F(G/G')$ is the sum of all the commutative simple components of FG , whereas $\Delta(G, G')$ is the sum of all the non-commutative simple components of FG . We conclude that, if FG is semisimple, then

$$FG \cong F(G/G') \oplus \bigoplus_{i=1}^l M(n_i, K_i).$$

Now, if $\dim_F(Z(FG)) = r$ and if the number of commutative simple components is s , then $l \leq r - s$.

In what follows, D_n is the dihedral group of order n , C_n is the cyclic group of order n , F^n is the direct sum of n copies of F , F_n is the extension of F of degree n , $M(n, F)$ is the algebra of all $n \times n$ matrices over F , $GL(n, F)$ is the general linear group of degree n over F , $Z(FG)$ is the center of FG , $[g]$ is the conjugacy class of $g \in G$ and T_p is the set of all p -elements of G including 1.

Let F be a field of characteristic $p > 0$ and let G be a finite group. An element $g \in G$ is p -regular, if $p \nmid o(g)$. Let t be the l.c.m. of the orders of p -regular elements of G and let ω be a primitive t -th root of unity over the field F . Then

$$A = \{r \mid \omega \rightarrow \omega^r \text{ is an automorphism of } F(\omega) \text{ over } F\}.$$

Let γ_g be the sum of all conjugates of $g \in G$. If g is a p -regular element, then the cyclotomic F -class of γ_g is

$$S_F(\gamma_g) = \{\gamma_{g^r} \mid r \in A\}.$$

Many authors [1, 4, 5, 7, 9–12] have studied the structure of $U(FG)$ for a finite group G and for a finite field F . The structure of $U(F(C_3 \times D_{10}))$ for $p = 3$ is

given in [10]. In this article, we provide an explicit description for the Wedderburn decomposition of $FG/J(FG)$, $G = C_3 \times D_{10}$ and F a finite field of characteristic $p \neq 3$, using the theory developed by Ferraz [3] and with the help of this description we obtain the structure of $U(F(C_3 \times D_{10}))$.

Lemma 1.1 ([3, Proposition 1.2]). *Let K be a field and let G be a finite group. The number of simple components of $KG/J(KG)$ is equal to the number of cyclotomic K -classes in G .*

Lemma 1.2 ([3, Theorem 1.3]). *Let K be a field and let G be a finite group. Suppose that $\text{Gal}(K(\omega)/K)$ is cyclic. Let s be the number of cyclotomic K -classes in G . If R_1, R_2, \dots, R_s are the simple components of $Z(KG/J(KG))$ and P_1, P_2, \dots, P_s are the cyclotomic K -classes of G , then with a suitable re-ordering of indices, $|P_i| = [R_i : K]$.*

2. Structure of $U(F(C_3 \times D_{10}))$

Theorem 2.1. *Let F be a finite field of characteristic p with $|F| = q = p^n$ and let $G = C_3 \times D_{10}$.*

1. *If $p = 2$, then $U(FG) \cong$*

$$\begin{cases} C_2^{3n} \times (C_{2^{2n-1}}^3 \times GL(2, F)^6), & \text{if } q \equiv 1, 4 \pmod{15}; \\ C_2^{3n} \times (C_{2^{2n-1}} \times C_{2^{2n-1}} \times GL(2, F_2)^3), & \text{if } q \equiv 2, -7 \pmod{15}. \end{cases}$$

2. *If $p = 5$, then*

$$U(FG) \cong V \times \begin{cases} C_5^{6n}, & \text{if } q \equiv 1 \pmod{6}; \\ C_5^{2n} \times C_{5^{2n-1}}^2, & \text{if } q \equiv -1 \pmod{6}. \end{cases}$$

$$\text{where } V \cong (C_5^{15n} \times C_5^{6n}) \times C_5^{3n} \text{ and } Z(V) \cong C_5^{9n}.$$

3. *If $p > 5$, then $U(FG) \cong$*

$$\begin{cases} C_{p^{2n-1}}^6 \times GL(2, F)^6, & \text{if } q \equiv 1, -11 \pmod{30}; \\ C_{p^{2n-1}}^2 \times C_{p^{2n-1}}^2 \times GL(2, F)^2 \times GL(2, F_2)^2, & \text{if } q \equiv -1, 11 \pmod{30}; \\ C_{p^{2n-1}}^6 \times GL(2, F_2)^3, & \text{if } q \equiv 7, 13 \pmod{30}; \\ C_{p^{2n-1}}^2 \times C_{p^{2n-1}}^2 \times GL(2, F_2)^3, & \text{if } q \equiv -7, -13 \pmod{30}. \end{cases}$$

Proof. Let $G = \langle x, y, z \mid x^2 = y^5 = z^3 = 1, xy = y^4x, xz = zx, yz = zy \rangle$. The conjugacy classes in G are:

$$\begin{aligned} [z^i] &= \{z^i\} \text{ for } i = 0, 1, 2; \\ [yz^i] &= \{y^{\pm 1}z^i\} \text{ for } i = 0, 1, 2; \\ [y^2z^i] &= \{y^{\pm 2}z^i\} \text{ for } i = 0, 1, 2; \\ [xz^i] &= \{xz^i, xy^{\pm 1}z^i, xy^{\pm 2}z^i\} \text{ for } i = 0, 1, 2. \end{aligned}$$

1. $p = 2$. Clearly, $\widehat{T}_2 = 1 + x\widehat{y}$.

Let $\alpha = \sum_{k=0}^1 \sum_{j=0}^2 \sum_{i=5(j+3k)}^{5(j+3k)+4} a_i x^k y^{i-5(j+3k)} z^j$. If $\alpha \widehat{T}_2 = 0$, then we have

$$\alpha + \sum_{k=0}^1 \sum_{j=0}^2 \sum_{i=5(j+3k)}^{5(j+3k)+4} a_i x^{k+1} \widehat{y} z^j = 0.$$

For $k = 0, 1, 2$ and $i = 0, 1, 2, 3, 4$ this yields the following equations:

$$a_{5k+i} + \sum_{j=0}^4 a_{5k+j+15} = 0,$$

$$a_{5k+15+i} + \sum_{j=0}^4 a_{5k+j} = 0.$$

After simplification we get, $a_{5k} = a_{5k+i} = a_{5k+i+15}$ for $i = 0, 1, 2, 3, 4$ and $k = 0, 1, 2$. Hence

$$\text{Ann}(\widehat{T}_2) = \left\{ \sum_{i=0}^2 \beta_i (1+x) \widehat{y} z^i \mid \beta_i \in F \right\}.$$

Since $z, \widehat{y} \in Z(FG)$, $\text{Ann}^2(\widehat{T}_2) = 0$ and $\text{Ann}(\widehat{T}_2) \subseteq J(FG)$. Thus by [12, Lemma 2.2], $J(FG) = \text{Ann}(\widehat{T}_2)$ and $\dim_F(J(FG)) = 3$. Hence $V \cong C_2^{3n}$ and by [6, Lemma 2.1],

$$U(FG) \cong C_2^{3n} \rtimes U(FG/J(FG)).$$

Now it only remains to find the Wedderburn decomposition of $FG/J(FG)$.

As [1], $[y]$, $[y^2]$, $[z]$, $[z^2]$, $[yz]$, $[yz^2]$, $[y^2z]$, and $[y^2z^2]$ are the 2-regular conjugacy classes of G , $t = 15$ and $\dim_F(FG/J(FG)) = 27$. Now the following cases occur:

- (a) If $q \equiv 1, 4 \pmod{15}$, then $|S_F(\gamma_g)| = 1$ for $g = 1, y, y^2, z, z^2, yz, yz^2, y^2z, y^2z^2$. Consequently, [3, Theorem 1.3], yields nine components in the decomposition of $FG/J(FG)$. In view of the dimension requirements, the only possibility is:

$$FG/J(FG) \cong F^3 \oplus M(2, F)^6.$$

- (b) If $q \equiv 2, -7 \pmod{15}$, then $|S_F(\gamma_g)| = 1$ for $g = 1$ and $|S_F(\gamma_g)| = 2$ for $g = y, z, yz, yz^2$. So, due to the dimension restrictions, we have

$$FG/J(FG) \cong F \oplus F_2 \oplus M(2, F_2)^3.$$

2. $p = 5$. If $K = \langle y \rangle$, then $G/K \cong H \cong \langle x, z \rangle \cong C_6$. Thus from the ring epimorphism $\eta : FG \rightarrow FH$, given by

$$\eta \left(\sum_{j=0}^2 \sum_{i=0}^4 y^i z^j (a_{i+5j} + a_{i+5j+15x}) \right) = \sum_{j=0}^2 \sum_{i=0}^4 z^j (a_{i+5j} + a_{i+5j+15x}),$$

we get a group epimorphism $\phi : U(FG) \rightarrow U(FH)$ and $\ker \phi \cong 1 + J(FG) = V$. Further, we have the inclusion map $i : U(FH) \rightarrow U(FG)$ such that $\phi i = 1_{U(FH)}$. Thus $U(FG) \cong V \rtimes U(FC_6)$.

The structure of $U(FC_6)$ is given in [11, Theorem 4.1].

If $v = \sum_{j=0}^2 \sum_{i=0}^4 y^i z^j (a_{i+5j} + a_{i+5j+15x}) \in U(FG)$, then $v \in V$ if and only if $\sum_{i=0}^4 a_i = 1$ and $\sum_{i=0}^4 a_{i+5k} = 0$ for $k = 1, 2, 3, 4, 5$. Hence

$$V = \left\{ 1 + \sum_{j=0}^2 \sum_{i=1}^4 (y^i - 1) z^j (b_{i+4j} + b_{i+4j+12x}) \mid b_i \in F \right\}$$

and $|V| = 5^{24n}$. Since, $J(FG)^5 = 0$, $V^5 = 1$.

Now we show that $V \cong (C_5^{15n} \times C_5^{6n}) \times C_5^{3n}$. The proof is split into the following steps:

Step 1: Let $R = \{1 + ay(1 - y)^3x \mid a \in F\} \subseteq V$. Then $R \cong C_5^n$.

If

$$r_1 = 1 + ay(1 - y)^3x \in R$$

and

$$r_2 = 1 + by(1 - y)^3x \in R$$

where $a, b \in F$, then

$$r_1 r_2 = 1 + (a + b)y(1 - y)^3x \in R.$$

Therefore, R is an abelian subgroup of V of order 5^n . Hence $R \cong C_5^n$.

Step 2: $|C_V(R)| = 5^{21n}$, where $C_V(R) = \{v \in V \mid r^v = r \text{ for all } r \in R\}$.

Let

$$r = 1 + ay(1 - y)^3x \in R$$

and

$$v = 1 + \sum_{j=0}^2 \sum_{i=1}^4 (y^i - 1) z^j (b_{i+4j} + b_{i+4j+12x}) \in V$$

where $a, b_i \in F$. Then $v = 1 + v_1 + v_2x$, $v_1 = \sum_{j=0}^2 \sum_{i=1}^4 b_{i+4j} (y^i - 1) z^j$ and $v_2 = \sum_{j=0}^2 \sum_{i=1}^4 b_{i+4j+12} (y^i - 1) z^j$. So $v^{-1} = v^4 = 1 + 4v_1 + 4v_2x \pmod{(y - 1)^2 FG}$. Thus

$$r^v = 1 + v^{-1} ay(1 - y)^3 xv = r + 2a\hat{y} \sum_{j=0}^2 \sum_{i=1}^4 i b_{i+4j} z^j x.$$

Thus $r^v = r$ if and only if $\sum_{i=1}^4 ib_{i+4j} = 0$ for $j = 0, 1, 2$. Hence

$$C_V(R) = \left\{ 1 + \sum_{j=0}^2 \sum_{i=1}^3 [(y^i - 1) + i(y^4 - 1)]c_{i+3j}z^j \right. \\ \left. + \sum_{j=0}^2 \sum_{i=1}^4 (y^i - 1)c_{i+4j+9}z^j x \mid c_i \in F \right\}$$

and $|C_V(R)| = 5^{21n}$.

Step 3: $C_V(R) \cong C_5^{15n} \rtimes C_5^{6n}$.

Consider the sets

$$S = \{1 + y^3(y-1)^2[yb_1 + y(y+2)b_2 + b_3 + (yb_4 + (y+1)^2b_5)x]\}$$

and

$$T = \{1 + y^3(y-1)[(y-1)(yc_1 + (y+1)^2c_2) + (yc_3 + (y^2 + y + 1)c_4)x]\}$$

where $b_{1+j} = \sum_{i=0}^2 p_{i+3j}z^i$ for $j = 0, 1, 2, 3, 4$ and $c_{1+j} = \sum_{i=0}^2 q_{i+3j}z^i$ for $j = 0, 1, 2, 3$. With some computation it can be shown that S and T are abelian subgroups of $C_V(R)$. So $S \cong C_5^{15n}$ and $T \cong C_5^{12n}$.

Now, let

$$s = 1 + y^3(y-1)^2[yb_1 + y(y+2)b_2 + b_3 + (yb_4 + (y+1)^2b_5)x] \in S$$

and

$$t = 1 + y^3(y-1)[(y-1)(yc_1 + (y+1)^2c_2) + (yc_3 + (y^2 + y + 1)c_4)x] \in T.$$

Then

$$s^t = 1 + y^3(y-1)^2\{yb_1 + y(y+2)b_2 + b_3 + k_1y^3(1-y) \\ + [yb_4 + (y+1)^2b_5 + (y-1)^2(k_2 + k_3)]x\} \in S$$

where

$$k_1 = (c_4 + 2c_3)(b_4 - b_5), k_2 = (c_4 + 2c_3)(b_2 - b_3) \\ k_3 = 2(c_4^2 - c_3c_4 - c_3^2)(b_4 - b_5).$$

Let

$$U = S \cap T = \{1 + y^3(y-1)^2[yc_1 + (y+1)^2c_2]\}$$

where $c_{1+j} = \sum_{i=0}^2 q_{i+3j}z^i$ for $j = 0, 1$. Thus $U \cong C_5^{6n}$. So for some subgroup $W \cong C_5^{6n}$ of T , $T = U \times W$ and $W \cap S = 1$. Hence $C_V(R) \cong S \rtimes W \cong C_5^{15n} \rtimes C_5^{6n}$.

Step 4: Let $M = \{1 + \sum_{j=0}^2 r_j z^j y(y+1)^2(1-y)(1+x) \mid r_i \in F\} \subseteq V$. Then $M \cong C_5^{3n}$.

Let

$$m_1 = 1 + \sum_{j=0}^2 r_j z^j y(y+1)^2(1-y)(1+x) \in M$$

and

$$m_2 = 1 + \sum_{j=0}^2 s_j z^j y(y+1)^2(1-y)(1+x) \in M$$

where $r_j, s_j \in F$. Then

$$m_1 m_2 = 1 + \sum_{j=0}^2 (r_j + s_j) z^j y(y+1)^2(1-y)(1+x) \in M.$$

Therefore, M is an abelian subgroup of V of order 5^{3n} . Hence, $M \cong C_5^{3n}$.

Step 5: $V \cong C_V(R) \rtimes M$.

Let

$$\begin{aligned} a &= 1 + \sum_{j=0}^2 \sum_{i=1}^3 [(y^i - 1) + i(y^4 - 1)] c_{i+3j} z^j \\ &\quad + \sum_{j=0}^2 \sum_{i=1}^4 (y^i - 1) c_{i+4j+9} z^j x \in C_V(R) \end{aligned}$$

and let

$$b = 1 + \sum_{j=0}^2 r_j z^j y(y+1)^2(1-y)(1+x) \in M$$

where $c_i, r_i \in F$. Then

$$\begin{aligned} a^b &= 1 + \sum_{j=0}^2 \sum_{i=1}^3 [(y^i - 1) + i(y^4 - 1)] c_{i+3j} z^j \\ &\quad + \sum_{j=0}^2 \sum_{i=1}^4 (y^i - 1) c_{i+4j+9} z^j x + (k_1 + k_2 x) \in C_V(R) \end{aligned}$$

where

$$\begin{aligned} k_1 &= \sum_{j=0}^2 r_j z^j \left\{ \sum_{k=0}^2 (c_{10+4k} - c_{11+4k} - c_{12+4k} + c_{13+4k}) z^k \right. \\ &\quad \left. + 3 \sum_{j=0}^2 r_j z^j \sum_{i=1}^4 \sum_{k=0}^2 i(c_{i+4k+9} z^k) \right\} y(1-y)^3 \end{aligned}$$

and

$$k_2 = 2 \sum_{j=0}^2 r_j z^j \left\{ \sum_{k=0}^2 (c_{2+3k} - c_{3+3k}) z^k (1-y) \right. \\ \left. - \sum_{j=0}^2 r_j z^j \sum_{i=1}^4 \sum_{k=0}^2 i c_{i+4k+9} z^k \right\} y (1-y)^3 - 2 \sum_{j=0}^2 r_j z^j \sum_{i=0}^4 d_i y^i$$

with

$$d_0 = \sum_{j=0}^2 (4c_{10+4j} + 4c_{11+4j} + c_{12+4j} + c_{13+4j}) z^j, \\ d_1 = \sum_{j=0}^2 (4c_{10+4j} + 3c_{11+4j} + 3c_{12+4j}) z^j, \\ d_2 = \sum_{j=0}^2 (4c_{11+4j} + 3c_{12+4j} + 3c_{13+4j}) z^j, \\ d_3 = \sum_{j=0}^2 (2c_{10+4j} + 2c_{11+4j} + c_{12+4j}) z^j, \\ d_4 = \sum_{j=0}^2 (2c_{11+4j} + 2c_{12+4j} + c_{13+4j}) z^j.$$

Clearly, $C_V(R) \cap M = 1$. Therefore, $V = C_V(R) \rtimes M$.

In the sequel, we show that $Z(V) \cong C_5^{9n}$.

If $v = 1 + \sum_{j=0}^2 \sum_{i=1}^4 (y^i - 1) z^j (b_{i+4j} + b_{i+4j+12} x) \in C_V(y) = \{v \in V \mid vy = yv\}$, then

$$vy - yv = \sum_{i=1}^4 \sum_{j=0}^2 y(1-y^i)(y^3-1)b_{i+4j+12} z^j x.$$

Thus $v \in C_V(y)$ if and only if $b_i = b_{i+j}$ for $j = 1, 2, 3$ and $i = 13, 17, 21$. Hence

$$C_V(y) = \left\{ 1 + \sum_{j=0}^2 \sum_{i=1}^4 (y^i - 1) c_{i+4j} z^j + \hat{y} \sum_{j=0}^2 c_{j+13} z^j x \mid c_i \in F \right\}.$$

Since $Z(V) \subseteq C_V(y)$,

$$Z(V) = \{s \in C_V(y) \mid sv = vs \text{ for all } v \in V\}.$$

Let $u = 1 + \sum_{j=0}^2 \sum_{i=1}^4 (y^i - 1)c_{i+4j}z^j + \widehat{y}x \sum_{j=0}^2 c_{j+13}z^j \in C_V(y)$. Since $v = 1 + (y - 1)zx \in V$ and $\widehat{y} \in Z(FG)$, $vu - uv = 0$ yields

$$(y - 1) \sum_{i=1}^4 \sum_{j=0}^2 (y^i - y^{-i})c_{i+4j}z^{j+1}x = 0.$$

Thus $c_i = c_{i+3}$ for $i = 1, 5, 9$ and $c_j = c_{j+1}$ for $j = 2, 6, 10$ and $u = 1 + y^4(y - 1)^2 \sum_{j=0}^2 d_{1+j}z^j + y^3(y^2 - 1)^2 \sum_{j=0}^2 d_{4+j}z^j + \widehat{y} \sum_{j=0}^2 d_{7+j}z^j x$. Clearly $u \in Z(V)$.

We conclude that $Z(V) = \{1 + y^4(y - 1)^2 \sum_{j=0}^2 d_{1+j}z^j + y^3(y^2 - 1)^2 \sum_{j=0}^2 d_{4+j}z^j + \widehat{y} \sum_{j=0}^2 d_{7+j}z^j x \mid d_i \in F\} \cong C_5^{9n}$.

3. If $p > 5$, then $J(FG) = 0$. Thus FG is semisimple and $t = 30$. As $G/G' \cong C_6$, we have

$$FG \cong FC_6 \oplus \bigoplus_{i=1}^l M(n_i, K_i).$$

Since $\dim_F(Z(FG)) = 12$, $l \leq 6$. Now we have the following cases:

- (a) If $q \equiv 1, -11 \pmod{30}$, then $|S_F(\gamma_g)| = 1$ for all $g \in G$. Therefore by [11, Theorem 4.1] and [3, Prop 1.2 and Theorem 1.3],

$$FG \cong F^6 \oplus \bigoplus_{i=1}^6 M(n_i, F)$$

and $\sum_{i=1}^6 n_i^2 = 24$. Clearly $n_i = 2$ for $i \in \{1, 2, 3, 4, 5, 6\}$. Hence,

$$FG \cong F^6 \oplus M(2, F)^6.$$

- (b) If $q \equiv -1, 11 \pmod{30}$, then $|S_F(\gamma_g)| = 1$ for $g = 1, x, y, y^2$ and $|S_F(\gamma_g)| = 2$ for $g = z, xz, yz, y^2z$. In this case $FC_6 \cong F^2 \oplus F_2^2$, thus dimension constraints yield

$$n_1^2 + n_2^2 + 2n_3^2 + 2n_4^2 = 24.$$

We get $n_1 = n_2 = n_3 = n_4 = 2$. Hence,

$$FG \cong F^2 \oplus F_2^2 \oplus M(2, F)^2 \oplus M(2, F_2)^2.$$

- (c) If $q \equiv 7, 13 \pmod{30}$, then $T = \{1, 7, 13, 19\} \pmod{30}$. Thus $|S_F(\gamma_g)| = 1$ for $g = 1, x, z, z^2, xz, xz^2$ and $|S_F(\gamma_g)| = 2$ for $g = y, yz, yz^2$. Therefore,

$$2(n_1^2 + n_2^2 + n_3^2) = 24.$$

We get $n_1 = n_2 = n_3 = 2$. Hence,

$$FG \cong F^6 \oplus M(2, F_2)^3.$$

- (d) If $q \equiv -7, -13 \pmod{30}$, then $T = \{1, 17, 19, 23\} \pmod{30}$. Thus $|S_F(\gamma_g)| = 1$ for $g = 1, x$ and $|S_F(\gamma_g)| = 2$ for $g = y, z, xz, yz, yz^2$. Hence,

$$FG \cong F^2 \oplus F_2^2 \oplus M(2, F_2)^3. \quad \square$$

References

- [1] S. F. ANSARI, M. SAHAI: *Unit Groups of Group Algebras of Groups of Order 20*, *Quaestiones Mathematicae* 44.4 (2021), pp. 503–511, DOI: <https://doi.org/10.2989/16073606.2020.1727583>.
- [2] A. A. BOVDI, J. KURDICS: *Lie properties of the Group Algebra and the Nilpotency Class of the Group of Units*, *Journal of Algebra* 212.1 (1999), pp. 28–64, DOI: <https://doi.org/10.1006/jabr.1998.7617>.
- [3] R. A. FERRAZ: *Simple Components of the Center of $FG/J(FG)$* , *Communications in Algebra* 36.9 (2008), pp. 3191–3199, DOI: <https://doi.org/10.1080/00927870802103503>.
- [4] J. GILDEA: *The Structure of $U(F_{5^k}D_{20})$* , *International Electronic Journal of Algebra* 8 (2010), pp. 153–160.
- [5] J. GILDEA: *Units of $F_{5^k}D_{10}$* , *Serdica Mathematical Journal* 36 (2010), pp. 247–254.
- [6] N. MAKHIJANI, R. K. SHARMA, J. B. SRIVASTAVA: *A Note on Units in $F_p^m D_{2p^n}$* , *Acta Mathematica Academiae Paedagogicae Nyiregyhaziensis* 30.1 (2014), pp. 17–25.
- [7] N. MAKHIJANI, R. K. SHARMA, J. B. SRIVASTAVA: *The Unit Groups of Some Special Semisimple Group Algebras*, *Journal of Algebra* 39.1 (2016), pp. 9–28, DOI: <https://doi.org/10.2989/16073606.2015.1024410>.
- [8] C. P. MILIES, S. K. SEHGAL: *An Introduction to Group Rings*, in: Dordrecht: Kluwer Academic Publishers, 2002.
- [9] F. MONAGHAN: *Units of Some Group Algebras of Non Abelian Groups of Order 24 Over Any Finite Field of Characteristic 3*, *International Electronic Journal of Algebra* 12 (2012), pp. 133–161.
- [10] M. SAHAI, S. F. ANSARI: *The group of units of the group algebras of groups D_{30} and $C_3 \times D_{10}$ over a finite field of characteristic 3*, *International Electronic Journal of Algebra* 29 (2021), pp. 165–174, DOI: <https://doi.org/10.24330/ieja.852146>.
- [11] M. SAHAI, S. F. ANSARI: *Unit Groups of Finite Group Algebras of Abelian Groups of Order At Most 16*, *Asian-European Journal of Mathematics* 14.3 (2021), 2150030 (17 pages), DOI: <https://doi.org/10.1142/S1793557121500303>.
- [12] G. TANG, Y. WEI, Y. LI: *Unit Groups of Group Algebras of Some Small Groups*, *Czechoslovak Mathematical Journal* 64.1 (2014), pp. 149–157, DOI: <https://doi.org/10.1007/s10587-014-0090-0>.

Efficient Taylor expansion computation of multidimensional vector functions on GPU*

Vaclav Skala

University of West Bohemia, Pilsen, Czech Republic

skala@kiv.zcu.cz

Submitted: September 4, 2020

Accepted: March 12, 2021

Published online: March 17, 2021

Abstract

The Taylor expansion [19] is used in many applications for a value estimation of scalar functions of one or two variables in the neighbour point. Usually, only the first two elements of the Taylor expansion are used, i.e. a value in the given point and derivatives estimation. The Taylor expansion can be also used for vector functions, too. The usual formulae are well known, but if the second element of the expansion, i.e. with the second derivatives are to be used, mathematical formulations are getting too complex for efficient programming, as it leads to the use of multi-dimensional matrices.

This contribution describes a new form of the Taylor expansion for multidimensional vector functions. The proposed approach uses “standard” formalism of linear algebra, i.e. using vectors and matrices, which is simple, easy to implement. It leads to efficient computation on the GPU in the three dimensional case, as the GPU offers fast vector-vector computation and many parts can be done in parallel.

Keywords: Taylor expansion, vector functions, vector-vector operations, approximation, GPU and SSE instructions, parallel computation, radial basis functions.

AMS Subject Classification: 41A58, 65D15, 26B05, 65D05

*This research was supported by the Czech Science Foundation (GACR), project No. GA 17-05534S.

1. Introduction

The Taylor expansion was introduced by the English mathematician Brook Taylor in 1715. However, closely related methods were given by Madhava of Sangamagrama in the 14th century [19]. It is used in many applications and used to approximate evaluation of many functions. In particular, the first two elements of the Taylor expansion are used as a linearization of a function behaviour at the given point and its surroundings [20]. The Taylor expansion is used in solutions of partial differential equations (PDE) [1, 7, 17], ordinary differential equations (ODE) [2, 6, 21], integral equations (IE) integro-differential equations (IDE) [1, 11], approximation of inverse functions (AIF) [8], control theory [4], fluid flow visualization of 3D flow using radial basis functions [12, 13, 15], computer vision [5, 18], in statistical mechanics [10], antenna design [6, 9], operator theory [3], etc.

2. Taylor expansion of scalar functions

The Taylor expansion of a scalar function is defined as successive derivatives, generally called tensors. In the one-dimensional case, i.e. scalar functions, the first derivative is actually the gradient $\nabla f(x)$, the second derivative has the form of a Hessian matrix, the third form leads to three-dimensional matrix $\mathbf{H}(x)$, i.e. triples of vectors etc. In the following, the Taylor expansion for scalar and for a vector functions are described.

2.1. One-dimensional case

The Taylor expansion of a continuous scalar function of a one dimensional variable is given as

$$f(x) = f(x_0) + \sum_{k=1}^{\infty} \frac{1}{k!} \frac{\partial^k f(x_0)}{\partial x^k} (x - x_0)^k,$$

or as

$$f(x) = f(x_0) + \sum_{k=0}^{\infty} \frac{1}{k!} \frac{\partial^k f(x_0)}{\partial x^k} \Delta^k,$$

where $\Delta = x - x_0$. Generally, the Taylor expansion can be described as

$$f(x) = T_0 + T_1 + T_2 + T_3 + \dots,$$

where T_k can be expressed as

$$T_k = \frac{1}{k!} \frac{\partial^k f(x_0)}{\partial x^k} \Delta^k.$$

It can be seen, that the Taylor expansion of a scalar function of a one dimensional variable can be described as

$$f(x) = f(x_0) + \frac{\partial^1 f(x_0)}{\partial x} \Delta + \frac{1}{2} \frac{\partial^2 f(x_0)}{\partial x^2} \Delta^2 + \frac{1}{6} \frac{\partial^3 f(x_0)}{\partial x^3} \Delta^3 + \dots$$

However, the Taylor expansion is also used for a scalar function of m -dimensional variables, i.e. $f(\mathbf{x}) = f(x_1, \dots, x_m)$. In this case, the expanded version of the Taylor expansion gets a little bit more complicated.

2.2. Multi-dimensional case

In the case of the scalar function with the multidimensional argument, i.e. $f(\mathbf{x}) = f(x_1, \dots, x_m)$, the Taylor expansion is more complicated as

$$f(\mathbf{x}) = T_0 + T_1 + T_2 + T_3 + \dots,$$

where T_k can be expressed as

$$T_k = \frac{1}{k!} [D^k f(\mathbf{x}_0)] [\Delta^k],$$

where

$$D^k f(\mathbf{x}) = \frac{\partial^k f(\mathbf{x})}{\partial x_1^{k_1} \dots \partial x_m^{k_m}}, \quad [\Delta^k] = [\Delta_1^{k_1}, \dots, \Delta_m^{k_m}]^T,$$

$$k = \sum_{i=1}^m k_i, \quad k_i \geq 0.$$

Now, the Taylor expansion is defined as

$$f(\mathbf{x}) = f(\mathbf{x}_0) + \nabla f(\mathbf{x}_0)(\mathbf{x} - \mathbf{x}_0) + \frac{1}{2}(\mathbf{x} - \mathbf{x}_0)^T \mathbf{H}(\mathbf{x}_0)(\mathbf{x} - \mathbf{x}_0) + T_3 + \dots,$$

or as

$$f(\mathbf{x}) = f(\mathbf{x}_0) + \nabla f(\mathbf{x}_0) [\Delta_i] + \frac{1}{2} [\Delta_i]^T \mathbf{H}(\mathbf{x}_0) [\Delta_i] + T_3 + \dots, \quad (2.1)$$

where $[\Delta_i] = [\Delta_1, \dots, \Delta_m]^T$, $\nabla f(\mathbf{x}_0)$ is a gradient of the function $f(\mathbf{x})$ at the point \mathbf{x}_0 , $\mathbf{H}(\mathbf{x}_0)$ is the Hessian matrix of the given function, i.e.

$$\mathbf{H}(\mathbf{x}_0) = \left[\frac{\partial^2 f(\mathbf{x}_0)}{\partial x_i \partial x_j} \right], \quad i, j = 1, \dots, m. \quad (2.2)$$

In the majority of cases

$$\frac{\partial^2 f(\mathbf{x}_0)}{\partial x_i \partial x_j} = \frac{\partial^2 f(\mathbf{x}_0)}{\partial x_j \partial x_i}, \quad i, j = 1, \dots, m. \quad (2.3)$$

The element T_3 of the Taylor expansion for a scalar function of m -dimensional variable is

$$T_3 = \frac{1}{6} \sum_{i,j,k=1,1,1}^{m,m,m} \frac{\partial^3 f(\mathbf{x}_0)}{\partial x_i \partial x_j \partial x_k} \Delta_i \Delta_j \Delta_k. \quad (2.4)$$

This is quite complex form leading to higher computational requirements. Similarly to the case of the Hessian matrix, it can be expected that the order of the function derivations is independent.

It can be seen that the element T_3 of the Taylor expansion consists of a “three dimensional matrix”. It leads to the tensor notation, which is usually not part of the engineering education. If this notation is used directly in a program implementation, it leads to redundant computations due to the symmetry of higher order partial derivatives, see (2.2) and (2.3). Also handling with indexes might be too complicated.

Furthermore, in the physically oriented applications, it is necessary to use the Taylor expansion also for vector functions, i.e. for n -dimensional functions with m -dimensional arguments, in general.

3. Taylor expansion of vector functions

Vector functions are used in many physically oriented computations, e.g. fluid mechanics, electromagnetic field computation etc. The Taylor expansion for vector functions is more complicated.

Let us consider a vector function

$$\mathbf{f}(\mathbf{x}) = \begin{bmatrix} f_1(\mathbf{x}) \\ \vdots \\ f_n(\mathbf{x}) \end{bmatrix}, \quad \mathbf{x} = [x_1, \dots, x_m].$$

The Taylor expansion of a vector function can be expressed as

$$\mathbf{f}(\mathbf{x}) = \sum_{i=0}^{\infty} \mathbf{T}_i(\mathbf{x}_0),$$

where $\mathbf{T}_i(\mathbf{x}_0)$ are vectors, now. Explicitly, it is possible to write

$$\mathbf{f}(\mathbf{x}) = \mathbf{f}(\mathbf{x}_0) + \mathbf{J}(\mathbf{x}_0) [\Delta_i] + \frac{1}{2} \begin{bmatrix} [\Delta_i]^T \mathbf{H}^1(\mathbf{x}_0) [\Delta_i] \\ \vdots \\ [\Delta_i]^T \mathbf{H}^n(\mathbf{x}_0) [\Delta_i] \end{bmatrix} + \mathbf{T}_3 + \dots,$$

where $\mathbf{J}(\mathbf{x}_0) = \left[\frac{\partial f_i(\mathbf{x}_0)}{\partial x_j} \right]$ is the Jacobi matrix ($n \times m$) and $\mathbf{H}^k(\mathbf{x}_0)$ are the Hessian matrices ($m \times m$) with the second derivatives of the function $f_k(\mathbf{x})$, $k = 1, \dots, n$, in general.

It can be seen, that the element \mathbf{T}_2 of the Taylor expansion is not expressed by standard linear algebra formalism as its result must be a vector, i.e. a “three-dimensional matrix” would have to be used containing elements $\left[\frac{\partial^2 f_i(\mathbf{x}_0)}{\partial x_j \partial x_k} \right]$. Also, it is necessary to point out that memory requirements can be estimated as $O(nm^2)$, as the matrix \mathbf{H}^k is of the size ($m \times m$) and $k = 1, \dots, n$.

4. Re-formulation of the Taylor expansion

A short summaries of the Taylor expansion for scalar and vector functions have been given in sections 2 and 3. If higher degree elements than the linear ones are to be used, e.g. T_2 or T_3 , the efficient representation and implementation gets more complex and computationally time consuming.

In the following, a modification of the Taylor expansion for the case $n = m = 3$ is presented. It uses only standard matrix-vector multiplication and also allows simpler symbolic manipulation of it. However, the given approach can be extended for higher dimensions, i.e. $n > 3$ and $m > 3$.

4.1. Scalar functions

In the case of a scalar function with a multidimensional argument the Taylor expansion is defined as

$$f(\mathbf{x}) = f(\mathbf{x}_0) + \nabla f(\mathbf{x}_0) [\Delta_i] + \frac{1}{2} [\Delta_i]^T \mathbf{H}(\mathbf{x}_0) [\Delta_i] + T_3 + \dots,$$

where $[\Delta_i] = [\Delta_1, \dots, \Delta_m]^T$, $[\Delta_i^2] = [\Delta_1^2, \dots, \Delta_m^2]^T$ and $\Delta_i = x_i - x_{i0}$, $i = 1, \dots, m$.

The T_2 element is formed by a quadratic form and the T_3 element is formed by a three-dimensional matrix, see (2.1). It causes several complications in formal manipulation and implementation as well. However, in the majority of cases

$$\frac{\partial^2 f(\mathbf{x}_0)}{\partial x_i \partial x_j} = \frac{\partial^2 f(\mathbf{x}_0)}{\partial x_j \partial x_i}, \quad i, j = 1, \dots, m.$$

Therefore only $m(m+1)/2$ values are needed for evaluation of the T_2 element of the Taylor expansion. It means that the T_2 element of the Taylor expansion, i.e. the element with the Hessian matrix, can be split to two parts using the inner product (dot product) as follows

$$T_2 = \frac{1}{2} \left[\frac{\partial^2 f(\mathbf{x}_0)}{(\partial x_i)^2} \right] [\Delta_i^2] + \sum_{i,j \& i > j}^{m,m} \frac{\partial^2 f(\mathbf{x}_0)}{\partial x_i \partial x_j} [\Delta_i \Delta_j],$$

where $[\Delta] = [\Delta_1, \dots, \Delta_m]^T$ and $[\Delta^2] = [\Delta_1^2, \dots, \Delta_m^2]^T$, in general.

It means, that in the three-dimensional case, i.e. $f(\mathbf{x}) = f(x_1, \dots, x_3)$, the Taylor expansion gets quite simple as the element T_2 has the form

$$T_2 = \frac{1}{2} \nabla^2 f(\mathbf{x}_0) \begin{bmatrix} \Delta_1^2 \\ \Delta_2^2 \\ \Delta_3^2 \end{bmatrix} + \begin{bmatrix} \frac{\partial^2 f(\mathbf{x}_0)}{\partial x_1 \partial x_2} & \frac{\partial^2 f(\mathbf{x}_0)}{\partial x_2 \partial x_3} & \frac{\partial^2 f(\mathbf{x}_0)}{\partial x_3 \partial x_1} \end{bmatrix} \begin{bmatrix} \Delta_1 \Delta_2 \\ \Delta_2 \Delta_3 \\ \Delta_3 \Delta_1 \end{bmatrix}.$$

Now, using the matrix notation, the Taylor expansion can be rewritten as

$$f(\mathbf{x}) = f(\mathbf{x}_0) + \nabla f(\mathbf{x}_0) [\Delta_i] + \frac{1}{2} \mathbf{D} [\Delta_i^2] + \mathbf{R} [\Delta_i \Delta_j] + T_3 + \dots,$$

where

$$\mathbf{D} = \left[\frac{\partial^2 f(\mathbf{x}_0)}{(\partial x_i)^2} \right]^T, \quad \mathbf{R} = \left[\frac{\partial^2 f(\mathbf{x}_0)}{\partial x_i \partial x_j} \right], \quad i \neq j,$$

and \mathbf{D} is a vector, \mathbf{R} is a matrix.

The above given formulation uses just inner products (dot products) instead of matrix multiplications, which leads to significantly faster computation especially on GPU (requires just only one clock) or if SSE instructions are used.

In some cases, it is useful to use the element T_3 of the Taylor expansion, as it enables to represent ‘‘inflections’’ of a function and increase precision of approximation. It leads to a necessity to replace ‘‘three-dimensional matrix’’ used in the T_3 element, see (2.4) by more simple formulation. Originally, the 3D matrix contains 27 values of partial derivatives. However, using the algebraic operations the T_3 element can be expressed as

$$\begin{aligned} T_3 = \frac{1}{6} \left\{ \sum_{i=1}^3 \frac{\partial^3 f(\mathbf{x}_0)}{\partial x_i^3} \Delta_i^3 + 6 \frac{\partial^3 f(\mathbf{x}_0)}{\partial x_1 \partial x_2 \partial x_3} \Delta_1 \Delta_2 \Delta_3 \right. \\ + 3 \frac{\partial^3 f(\mathbf{x}_0)}{\partial x_1^2 \partial x_2} \Delta_1^2 \Delta_2 + 3 \frac{\partial^3 f(\mathbf{x}_0)}{\partial x_1^2 \partial x_3} \Delta_1^2 \Delta_3 \\ + 3 \frac{\partial^3 f(\mathbf{x}_0)}{\partial x_1 \partial x_3^2} \Delta_1 \Delta_3^2 + 3 \frac{\partial^3 f(\mathbf{x}_0)}{\partial x_1 \partial x_2^2} \Delta_1 \Delta_2^2 \\ \left. + 3 \frac{\partial^3 f(\mathbf{x}_0)}{\partial x_2 \partial x_3^2} \Delta_2 \Delta_3^2 + 3 \frac{\partial^3 f(\mathbf{x}_0)}{\partial x_2^2 \partial x_3} \Delta_2^2 \Delta_3 \right\}. \end{aligned}$$

It means, that in the case of a scalar function of three dimensional variables, the T_3 term can be easily evaluated as only 10 values of partial derivatives are computed instead of 27 in the original formulation.

The T_3 element can be formally expressed as

$$\begin{aligned} T_3 = \frac{\partial^3 f(\mathbf{x}_0)}{\partial x_1 \partial x_2 \partial x_3} \Delta_1 \Delta_2 \Delta_3 + \frac{1}{6} \sum_{i=1}^3 \frac{\partial^3 f(\mathbf{x}_0)}{\partial x_i^3} \Delta_i^3 \\ + \frac{1}{2} \left\{ \frac{\partial^3 f(\mathbf{x}_0)}{\partial x_1^2 \partial x_2} \Delta_1^2 \Delta_2 + \frac{\partial^3 f(\mathbf{x}_0)}{\partial x_1^2 \partial x_3} \Delta_1^2 \Delta_3 \right. \\ + \frac{\partial^3 f(\mathbf{x}_0)}{\partial x_1 \partial x_3^2} \Delta_1 \Delta_3^2 + \frac{\partial^3 f(\mathbf{x}_0)}{\partial x_1 \partial x_2^2} \Delta_1 \Delta_2^2 \\ \left. + \frac{\partial^3 f(\mathbf{x}_0)}{\partial x_2 \partial x_3^2} \Delta_2 \Delta_3^2 + \frac{\partial^3 f(\mathbf{x}_0)}{\partial x_2^2 \partial x_3} \Delta_2^2 \Delta_3 \right\}. \end{aligned}$$

However, in physically oriented applications, it is necessary to use the Taylor expansion also for vector functions, i.e. n -dimensional functions with m -dimensional arguments.

For the vector-vector operations, i.e. if GPU or SSE instructions are used, the

T_3 element can be expressed as

$$\begin{aligned}
 T_3 &= \frac{\partial^3 f(\mathbf{x}_0)}{\partial x_1 \partial x_2 \partial x_3} \Delta_1 \Delta_2 \Delta_3 \\
 &+ \frac{1}{6} \begin{bmatrix} \frac{\partial^3 f(\mathbf{x}_0)}{\partial x_1^3} & \frac{\partial^3 f(\mathbf{x}_0)}{\partial x_2^3} & \frac{\partial^3 f(\mathbf{x}_0)}{\partial x_3^3} \end{bmatrix} \begin{bmatrix} \Delta_1^3 \\ \Delta_2^3 \\ \Delta_3^3 \end{bmatrix} \\
 &+ \frac{1}{2} \begin{bmatrix} \frac{\partial^3 f(\mathbf{x}_0)}{\partial x_1^2 \partial x_2} & \frac{\partial^3 f(\mathbf{x}_0)}{\partial x_1^2 \partial x_3} & \frac{\partial^3 f(\mathbf{x}_0)}{\partial x_1 \partial x_2^2} & \frac{\partial^3 f(\mathbf{x}_0)}{\partial x_1 \partial x_3^2} & \frac{\partial^3 f(\mathbf{x}_0)}{\partial x_2 \partial x_3^2} & \frac{\partial^3 f(\mathbf{x}_0)}{\partial x_2^2 \partial x_3} \end{bmatrix} \begin{bmatrix} \Delta_1^2 \Delta_2 \\ \Delta_1^2 \Delta_3 \\ \Delta_1 \Delta_2^2 \\ \Delta_1 \Delta_3^2 \\ \Delta_2 \Delta_3^2 \\ \Delta_2^2 \Delta_3 \end{bmatrix}.
 \end{aligned}$$

It means, that the T_3 element of the Taylor expansion can be implemented using the inner product (dot product) and therefore, it is possible to extend this approach for the Taylor expansion of vector functions.

4.2. Vector functions

The Taylor expansion can be easily extended for vector functions, i.e.

$$\mathbf{f}(\mathbf{x}) = [f_1(x_1, \dots, x_m), \dots, f_n(x_1, \dots, x_m)]^T.$$

However, the formulae get more complex in the general case. As there are many applications using three-dimensional representation, i.e. $n = m = 3$, the re-formulation of the Taylor expansion can be simplified using the analogy of the Taylor expansion for scalar functions as follows

$$\mathbf{f}(\mathbf{x}) = \sum_{i=0}^{\infty} \mathbf{T}_i(\mathbf{x}_0),$$

where $\mathbf{T}_i(\mathbf{x}_0)$ are vectors, now. Using the explicit notation

$$\mathbf{f}(\mathbf{x}) = \mathbf{f}(\mathbf{x}_0) + \mathbf{J}(\mathbf{x}_0) [\Delta_i] + \frac{1}{2} \mathbf{D} [\Delta_i^2] + \mathbf{R} [\Delta_i \Delta_j] + T_3 + \dots$$

where

$$\begin{aligned}
 \begin{bmatrix} f_1(\mathbf{x}) \\ \vdots \\ f_3(\mathbf{x}) \end{bmatrix} &= \begin{bmatrix} f_1(\mathbf{x}_0) \\ \vdots \\ f_3(\mathbf{x}_0) \end{bmatrix} + \begin{bmatrix} \frac{\partial f_1(\mathbf{x}_0)}{\partial x_1} & \dots & \frac{\partial f_1(\mathbf{x}_0)}{\partial x_3} \\ \vdots & \ddots & \vdots \\ \frac{\partial f_3(\mathbf{x}_0)}{\partial x_1} & \dots & \frac{\partial f_3(\mathbf{x}_0)}{\partial x_3} \end{bmatrix} \begin{bmatrix} \Delta_1 \\ \vdots \\ \Delta_3 \end{bmatrix} \\
 &+ \frac{1}{2} \begin{bmatrix} \frac{\partial^2 f_1(\mathbf{x}_0)}{\partial x_1^2} & \dots & \frac{\partial^2 f_1(\mathbf{x}_0)}{\partial x_3^2} \\ \vdots & \ddots & \vdots \\ \frac{\partial^2 f_3(\mathbf{x}_0)}{\partial x_1^2} & \dots & \frac{\partial^2 f_3(\mathbf{x}_0)}{\partial x_3^2} \end{bmatrix} \begin{bmatrix} \Delta_1^2 \\ \vdots \\ \Delta_3^2 \end{bmatrix}
 \end{aligned}$$

$$+ \begin{bmatrix} \frac{\partial^2 f_1(\mathbf{x}_0)}{\partial x_1 \partial x_2} & \frac{\partial^2 f_1(\mathbf{x}_0)}{\partial x_2 \partial x_3} & \frac{\partial^2 f_1(\mathbf{x}_0)}{\partial x_3 \partial x_1} \\ \frac{\partial^2 f_2(\mathbf{x}_0)}{\partial x_1 \partial x_2} & \frac{\partial^2 f_2(\mathbf{x}_0)}{\partial x_2 \partial x_3} & \frac{\partial^2 f_2(\mathbf{x}_0)}{\partial x_3 \partial x_1} \\ \frac{\partial^2 f_3(\mathbf{x}_0)}{\partial x_1 \partial x_2} & \frac{\partial^2 f_3(\mathbf{x}_0)}{\partial x_2 \partial x_3} & \frac{\partial^2 f_3(\mathbf{x}_0)}{\partial x_3 \partial x_1} \end{bmatrix} \begin{bmatrix} \Delta_1 \Delta_2 \\ \Delta_2 \Delta_3 \\ \Delta_3 \Delta_1 \end{bmatrix} + T_3 + \dots,$$

where $[\Delta_i] = [\Delta_1, \dots, \Delta_m]^T$ and $[\Delta_i^2] = [\Delta_1^2, \dots, \Delta_m^2]^T$.

Now, similar approach can be taken as in the Taylor expansion for scalar functions. It means, that in the three-dimensional case, i.e. $f(\mathbf{x}) = f(x_1, \dots, x_3)$, the Taylor expansion gets quite simple as the element \mathbf{T}_2 , which is a vector, has the form

$$\mathbf{T}_2 = \frac{1}{2} \begin{bmatrix} \nabla^2 f_1(\mathbf{x}_0) \\ \nabla^2 f_2(\mathbf{x}_0) \\ \nabla^2 f_3(\mathbf{x}_0) \end{bmatrix} \begin{bmatrix} \Delta_1^2 \\ \Delta_2^2 \\ \Delta_3^2 \end{bmatrix} + \begin{bmatrix} \frac{\partial^2 f_1(\mathbf{x}_0)}{\partial x_1 \partial x_2} & \frac{\partial^2 f_1(\mathbf{x}_0)}{\partial x_2 \partial x_3} & \frac{\partial^2 f_1(\mathbf{x}_0)}{\partial x_3 \partial x_1} \\ \frac{\partial^2 f_2(\mathbf{x}_0)}{\partial x_1 \partial x_2} & \frac{\partial^2 f_2(\mathbf{x}_0)}{\partial x_2 \partial x_3} & \frac{\partial^2 f_2(\mathbf{x}_0)}{\partial x_3 \partial x_1} \\ \frac{\partial^2 f_3(\mathbf{x}_0)}{\partial x_1 \partial x_2} & \frac{\partial^2 f_3(\mathbf{x}_0)}{\partial x_2 \partial x_3} & \frac{\partial^2 f_3(\mathbf{x}_0)}{\partial x_3 \partial x_1} \end{bmatrix} \begin{bmatrix} \Delta_1 \Delta_2 \\ \Delta_2 \Delta_3 \\ \Delta_3 \Delta_1 \end{bmatrix}.$$

The above given formulation uses just three inner products (dot products) instead of matrix multiplications, which leads to significantly faster computation especially on GPU (requires just only one clock) or if SSE instructions are used.

In some cases, it is useful to use the element \mathbf{T}_3 of the Taylor expansion, as it enables to represent “inflections” of a function and increase precision of approximation. It leads to a necessity to replace “three-dimensional matrix” used in the \mathbf{T}_3 element, see (2.4) by simpler formulation. In the original formulation, the 3D matrix contains 27 values of partial derivatives.

However, using the algebraic operations the \mathbf{T}_3 , which is a vector, the k^{th} element, $k = 1, \dots, 3$ can be expressed as

$$T_{3k} = \frac{1}{6} \left\{ \sum_{i=1}^3 \frac{\partial^3 f_k(\mathbf{x}_0)}{\partial x_i^3} \Delta_i^3 + 6 \frac{\partial^3 f_k(\mathbf{x}_0)}{\partial x_1 \partial x_2 \partial x_3} \Delta_1 \Delta_2 \Delta_3 \right. \\ + 3 \frac{\partial^3 f_k(\mathbf{x}_0)}{\partial x_1^2 \partial x_2} \Delta_1^2 \Delta_2 + 3 \frac{\partial^3 f_k(\mathbf{x}_0)}{\partial x_1^2 \partial x_3} \Delta_1^2 \Delta_3 \\ + 3 \frac{\partial^3 f_k(\mathbf{x}_0)}{\partial x_1 \partial x_3^2} \Delta_1 \Delta_3^2 + 3 \frac{\partial^3 f_k(\mathbf{x}_0)}{\partial x_1 \partial x_2^2} \Delta_1 \Delta_2^2 \\ \left. + 3 \frac{\partial^3 f_k(\mathbf{x}_0)}{\partial x_2 \partial x_3^2} \Delta_2 \Delta_3^2 + 3 \frac{\partial^3 f_k(\mathbf{x}_0)}{\partial x_2^2 \partial x_3} \Delta_2^2 \Delta_3 \right\}.$$

In the case of a vector function of three dimensional variables, the \mathbf{T}_3 term can be easily evaluated as only 3×10 values of partial derivatives are computed instead of 3×27 in the original formulation.

The k^{th} element, $k = 1, \dots, 3$, of the \mathbf{T}_3 vector element can be formally ex-

pressed as

$$\begin{aligned}
T_{3_k} = & \frac{\partial^3 f_k(\mathbf{x}_0)}{\partial x_1 \partial x_2 \partial x_3} \Delta_1 \Delta_2 \Delta_3 + \frac{1}{6} \sum_{i=1}^3 \frac{\partial^3 f_k(\mathbf{x}_0)}{\partial x_i^3} \Delta_i^3 \\
& + \frac{1}{2} \left\{ \frac{\partial^3 f_k(\mathbf{x}_0)}{\partial x_1^2 \partial x_2} \Delta_1^2 \Delta_2 + \frac{\partial^3 f_k(\mathbf{x}_0)}{\partial x_1^2 \partial x_3} \Delta_1^2 \Delta_3 \right. \\
& \quad + \frac{\partial^3 f_k(\mathbf{x}_0)}{\partial x_1 \partial x_3^2} \Delta_1 \Delta_3^2 + \frac{\partial^3 f_k(\mathbf{x}_0)}{\partial x_1 \partial x_2^2} \Delta_1 \Delta_2^2 \\
& \quad \left. + \frac{\partial^3 f_k(\mathbf{x}_0)}{\partial x_2 \partial x_3^2} \Delta_2 \Delta_3^2 + \frac{\partial^3 f_k(\mathbf{x}_0)}{\partial x_2^2 \partial x_3} \Delta_2^2 \Delta_3 \right\}.
\end{aligned}$$

The vector \mathbf{T}_3 of the Taylor expansion can be expressed using standard linear algebra notation, instead of using three dimensional matrix notation, as

$$\begin{aligned}
\mathbf{T}_3 = & \begin{bmatrix} \frac{\partial^3 f_1(\mathbf{x}_0)}{\partial x_1 \partial x_2 \partial x_3} & \frac{\partial^3 f_2(\mathbf{x}_0)}{\partial x_1 \partial x_2 \partial x_3} & \frac{\partial^3 f_3(\mathbf{x}_0)}{\partial x_1 \partial x_2 \partial x_3} \end{bmatrix} \Delta_1 \Delta_2 \Delta_3 \\
& + \frac{1}{6} \begin{bmatrix} \frac{\partial^3 f_1(\mathbf{x}_0)}{\partial x_1^3} & \frac{\partial^3 f_1(\mathbf{x}_0)}{\partial x_2^3} & \frac{\partial^3 f_1(\mathbf{x}_0)}{\partial x_3^3} \\ \frac{\partial^3 f_2(\mathbf{x}_0)}{\partial x_1^3} & \frac{\partial^3 f_2(\mathbf{x}_0)}{\partial x_2^3} & \frac{\partial^3 f_2(\mathbf{x}_0)}{\partial x_3^3} \\ \frac{\partial^3 f_3(\mathbf{x}_0)}{\partial x_1^3} & \frac{\partial^3 f_3(\mathbf{x}_0)}{\partial x_2^3} & \frac{\partial^3 f_3(\mathbf{x}_0)}{\partial x_3^3} \end{bmatrix} \begin{bmatrix} \Delta_1^3 \\ \Delta_2^3 \\ \Delta_3^3 \end{bmatrix} \\
& + \frac{1}{2} \begin{bmatrix} \frac{\partial^3 f_1(\mathbf{x}_0)}{\partial x_1^2 \partial x_2} & \frac{\partial^3 f_1(\mathbf{x}_0)}{\partial x_1^2 \partial x_3} & \frac{\partial^3 f_1(\mathbf{x}_0)}{\partial x_1 \partial x_2^2} & \frac{\partial^3 f_1(\mathbf{x}_0)}{\partial x_1 \partial x_3^2} & \frac{\partial^3 f_1(\mathbf{x}_0)}{\partial x_2 \partial x_3^2} & \frac{\partial^3 f_1(\mathbf{x}_0)}{\partial x_2^2 \partial x_3} \\ \frac{\partial^3 f_2(\mathbf{x}_0)}{\partial x_1^2 \partial x_2} & \frac{\partial^3 f_2(\mathbf{x}_0)}{\partial x_1^2 \partial x_3} & \frac{\partial^3 f_2(\mathbf{x}_0)}{\partial x_1 \partial x_2^2} & \frac{\partial^3 f_2(\mathbf{x}_0)}{\partial x_1 \partial x_3^2} & \frac{\partial^3 f_2(\mathbf{x}_0)}{\partial x_2 \partial x_3^2} & \frac{\partial^3 f_2(\mathbf{x}_0)}{\partial x_2^2 \partial x_3} \\ \frac{\partial^3 f_3(\mathbf{x}_0)}{\partial x_1^2 \partial x_2} & \frac{\partial^3 f_3(\mathbf{x}_0)}{\partial x_1^2 \partial x_3} & \frac{\partial^3 f_3(\mathbf{x}_0)}{\partial x_1 \partial x_2^2} & \frac{\partial^3 f_3(\mathbf{x}_0)}{\partial x_1 \partial x_3^2} & \frac{\partial^3 f_3(\mathbf{x}_0)}{\partial x_2 \partial x_3^2} & \frac{\partial^3 f_3(\mathbf{x}_0)}{\partial x_2^2 \partial x_3} \end{bmatrix} \begin{bmatrix} \Delta_1^2 \Delta_2 \\ \Delta_1^2 \Delta_3 \\ \Delta_1 \Delta_2^2 \\ \Delta_1 \Delta_3^2 \\ \Delta_2 \Delta_3^2 \\ \Delta_2^2 \Delta_3 \end{bmatrix}.
\end{aligned}$$

If matrix notation is used, the \mathbf{T}_3 element can be expressed as

$$\mathbf{T}_3 = \mathbf{U} (\Delta_1 \Delta_2 \Delta_3) + \mathbf{V} \begin{bmatrix} \Delta_1^3 \\ \Delta_2^3 \\ \Delta_3^3 \end{bmatrix} + \mathbf{W} \begin{bmatrix} \Delta_1^2 \Delta_2 \\ \Delta_1^2 \Delta_3 \\ \Delta_1 \Delta_2^2 \\ \Delta_1 \Delta_3^2 \\ \Delta_2 \Delta_3^2 \\ \Delta_2^2 \Delta_3 \end{bmatrix},$$

where

$$\begin{aligned}
\mathbf{U} = & \begin{bmatrix} \frac{\partial^3 f_1(\mathbf{x}_0)}{\partial x_1 \partial x_2 \partial x_3} & \frac{\partial^3 f_2(\mathbf{x}_0)}{\partial x_1 \partial x_2 \partial x_3} & \frac{\partial^3 f_3(\mathbf{x}_0)}{\partial x_1 \partial x_2 \partial x_3} \end{bmatrix}, \\
\mathbf{V} = & \begin{bmatrix} \frac{\partial^3 f_1(\mathbf{x}_0)}{\partial x_1^3} & \frac{\partial^3 f_1(\mathbf{x}_0)}{\partial x_2^3} & \frac{\partial^3 f_1(\mathbf{x}_0)}{\partial x_3^3} \\ \frac{\partial^3 f_2(\mathbf{x}_0)}{\partial x_1^3} & \frac{\partial^3 f_2(\mathbf{x}_0)}{\partial x_2^3} & \frac{\partial^3 f_2(\mathbf{x}_0)}{\partial x_3^3} \\ \frac{\partial^3 f_3(\mathbf{x}_0)}{\partial x_1^3} & \frac{\partial^3 f_3(\mathbf{x}_0)}{\partial x_2^3} & \frac{\partial^3 f_3(\mathbf{x}_0)}{\partial x_3^3} \end{bmatrix},
\end{aligned}$$

$$\mathbf{W} = \begin{bmatrix} \frac{\partial^3 f_1(\mathbf{x}_0)}{\partial x_1^2 \partial x_2} & \frac{\partial^3 f_1(\mathbf{x}_0)}{\partial x_1^2 \partial x_3} & \frac{\partial^3 f_1(\mathbf{x}_0)}{\partial x_1 \partial x_2^2} & \frac{\partial^3 f_1(\mathbf{x}_0)}{\partial x_1 \partial x_3^2} & \frac{\partial^3 f_1(\mathbf{x}_0)}{\partial x_2 \partial x_3^2} & \frac{\partial^3 f_1(\mathbf{x}_0)}{\partial x_2^2 \partial x_3} \\ \frac{\partial^3 f_2(\mathbf{x}_0)}{\partial x_1^2 \partial x_2} & \frac{\partial^3 f_2(\mathbf{x}_0)}{\partial x_1^2 \partial x_3} & \frac{\partial^3 f_2(\mathbf{x}_0)}{\partial x_1 \partial x_2^2} & \frac{\partial^3 f_2(\mathbf{x}_0)}{\partial x_1 \partial x_3^2} & \frac{\partial^3 f_2(\mathbf{x}_0)}{\partial x_2 \partial x_3^2} & \frac{\partial^3 f_2(\mathbf{x}_0)}{\partial x_2^2 \partial x_3} \\ \frac{\partial^3 f_3(\mathbf{x}_0)}{\partial x_1^2 \partial x_2} & \frac{\partial^3 f_3(\mathbf{x}_0)}{\partial x_1^2 \partial x_3} & \frac{\partial^3 f_3(\mathbf{x}_0)}{\partial x_1 \partial x_2^2} & \frac{\partial^3 f_3(\mathbf{x}_0)}{\partial x_1 \partial x_3^2} & \frac{\partial^3 f_3(\mathbf{x}_0)}{\partial x_2 \partial x_3^2} & \frac{\partial^3 f_3(\mathbf{x}_0)}{\partial x_2^2 \partial x_3} \end{bmatrix}.$$

It means, that the Taylor expansion can be written in the form containing only vectors and matrices.

$$\begin{aligned} \mathbf{f}(\mathbf{x}) &= \mathbf{f}(\mathbf{x}_0) + \mathbf{J}(\mathbf{x}_0) [\Delta_i] && \# \text{ linear case} \\ &+ \frac{1}{2} \mathbf{D} [\Delta_i^2] + \mathbf{R} [\Delta_i \Delta_j] && \# \text{ quadratic case} \\ &+ \mathbf{U} (\Delta_1 \Delta_2 \Delta_3) + \mathbf{V} \begin{bmatrix} \Delta_1^3 \\ \Delta_2^3 \\ \Delta_3^3 \end{bmatrix} + \mathbf{W} \begin{bmatrix} \Delta_1^2 \Delta_2 \\ \Delta_1^2 \Delta_3 \\ \Delta_1 \Delta_2^2 \\ \Delta_1 \Delta_3^2 \\ \Delta_2 \Delta_3^2 \\ \Delta_2^2 \Delta_3 \end{bmatrix} && \# \text{ cubic case} + \dots \end{aligned}$$

It can be seen, that the above given formulae are simple, easy to implement efficiently, especially if GPU or SSE instructions are used. The presented approach can be applied also for the case, when $n \neq m$, in general. However, it should be noted that size of some vectors and matrices grows quadratic.

5. Application

Visualization of 3D vector fields, i.e. fluid flow and electromagnetic fields, uses the Taylor expansion to approximate acquired data (measured or obtained from a simulation). If the data are large, the approximation is also used for data reduction, while keeping the important features of the vector data [14]. If the Taylor expansion is used with Radial Basis Functions (RBF) [12], it is possible to obtain analytical function describing the given vector field data respecting critical points, vector field second derivatives [13, 15].

The Taylor expansion was used for radial basis function (RBF) approximation of the EF5 Tornado data¹ using second derivatives of the Taylor expansion [16]. This led to high compression ratio, see illustrative images in the Figure 1 [16], and analytical form describing the tornado fluid flow in the analytical form for the flow speed as $\mathbf{v} = \mathbf{f}(\mathbf{x})$.

As the second derivatives were used, the proposed new formulation of the Taylor expansion offers simple formal structure, efficient computation and significant speed-up of computation. The formulation is convenient for GPU implementation which offers high speed-up due to parallelism available.

¹Data set of EF5 tornado courtesy of Leigh Orf from Cooperative Institute for Meteorological Satellite Studies, University of Wisconsin, Madison, WI, USA.

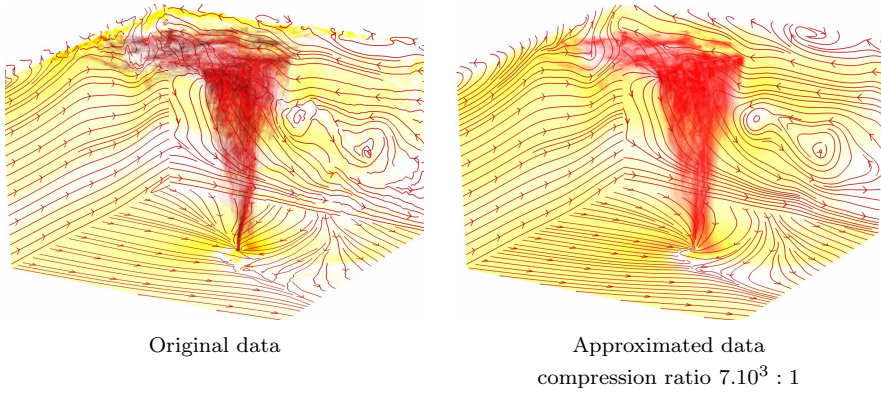


Figure 1. Tornado data and its approximation using second derivatives (taken from [16]).

6. Conclusion

This paper describes a new re-formulation of the Taylor expansion for scalar and vector functions for the multidimensional case and its optimization for the 3D case. This new re-formulation enables representation of the third order of approximation using standard linear algebra formalism, without tensor notation use. The proposed approach leads to significant speed-up of computation, see chapter 4. In the case of the GPU or SSE implementation additional speed up can be expected, especially due to fast vector-vector operations and native parallelism on GPU. Specialized version for the three dimensional case is presented, which is simple to implement as well.

The presented approach can be directly applied to 3D flow or electromagnetic fields computation and simulation. It can be extended to higher dimensions, however, the complexity of formulae grows quadratic. However, the expected speed up will grow with a dimension against “*standard*” implementation.

The influence of the second derivations was explored in [16]. It led to significant improvements for vector fields approximation, i.e. compression ratio and precision. In future, the influence of the *cubic part* of the Taylor expansion is to be studied, as inclusion of points of inflections and curvatures of vector fields should lead to higher compression ratio.

Acknowledgment. The author would like to thank to colleagues, especially Michal Smolik for image generation, and to students, especially to Jakub Vasta, Filip Hacha and Martin Cervenka, at the University of West Bohemia (Czech Republic) for hints and suggestions. Thanks belong also to colleagues at Shandong University and Zhejiang University (China) for their critical comments and constructive suggestions, and to anonymous reviewers for their valuable comments and hints provided.

References

- [1] A. ALVANDI, M. PARIPOUR: *The combined reproducing kernel method and Taylor series for handling nonlinear Volterra integro-differential equations with derivative type kernel*, Applied Mathematics and Computation 355 (2019), pp. 151–160, ISSN: 0096-3003, DOI: <https://doi.org/10.1016/j.amc.2019.02.023>.
- [2] A. BAEZA, S. BOSCARINO, P. MULET, G. RUSSO, D. ZORÍO: *Approximate Taylor methods for ODEs*, Computers and Fluids 159 (2017), pp. 156–166, ISSN: 0045-7930, DOI: <https://doi.org/10.1016/j.compfluid.2017.10.001>.
- [3] L. BERNAL-GONZALEZ, H. CABANA-MENDEZ, G. MUNOZ-FERNANDEZ, J. SEOANE-SEPULVEDA: *Universality of sequences of operators related to Taylor series*, Journal of Mathematical Analysis and Applications 474.1 (2019), pp. 480–491, ISSN: 0022-247X, DOI: <https://doi.org/10.1016/j.jmaa.2019.01.056>.
- [4] T. BREITEN, K. KUNISCH, L. PFEIFFER: *Taylor expansions of the value function associated with a bilinear optimal control problem*, Annales de l'Institut Henri Poincaré (C) Analyse Non Linéaire 36.5 (2019), pp. 1361–1399, ISSN: 0294-1449, DOI: <https://doi.org/10.1016/j.anihpc.2019.01.001>.
- [5] C.-Z. DONG, O. CELIK, F. CATBAS, E. OBRIEN, S. TAYLOR: *A robust vision-based method for displacement measurement under adverse environmental factors using spatio-temporal context learning and Taylor approximation*, Sensors (Switzerland) 19.14 (2019), ISSN: 1424-8220, DOI: <https://doi.org/10.3390/s19143197>.
- [6] N. HU, B. DUAN, W. XU, J. ZHOU: *A new interval pattern analysis method of array antennas based on Taylor expansion*, IEEE Transactions on Antennas and Propagation 65.11 (2017), pp. 6151–6156, ISSN: 0018-926X, DOI: <https://doi.org/10.1109/TAP.2017.2754458>.
- [7] W. HU, Y. GU, C.-M. FAN: *A meshless collocation scheme for inverse heat conduction problem in three-dimensional functionally graded materials*, Engineering Analysis with Boundary Elements 114 (2020), pp. 1–7, ISSN: 0955-7997, DOI: <https://doi.org/10.1016/j.enganabound.2020.02.001>.
- [8] R. JEDYNAK: *New facts concerning the approximation of the inverse Langevin function*, Journal of Non-Newtonian Fluid Mechanics 249 (2017), pp. 8–25, ISSN: 0377-0257, DOI: <https://doi.org/10.1016/j.jnnfm.2017.09.003>.
- [9] K. KONNO, Q. CHEN, R. BURKHOLDER: *Efficiency improvement with a recursive Taylor expansion of Bessel functions for layered media Green's function*, 2017 IEEE Antennas and Propagation Society International Symposium, Proceedings 2017-January (2017), pp. 1355–1356, DOI: <https://doi.org/10.1109/APUSNCURSINRSM.2017.8072720>.
- [10] Y. LIU, S. LIU, F. ZHAN, X. ZHANG: *Firing patterns of the modified Hodgkin–Huxley models subject to Taylor's formula*, Physica A: Statistical Mechanics and its Applications 547 (2020), ISSN: 0378-4371, DOI: <https://doi.org/10.1016/j.physa.2020.124405>.
- [11] K. NEAMPREM, A. KLANGRAK, H. KANEKO: *Taylor-series expansion methods for multivariate hammerstein integral equations*, IAENG International Journal of Applied Mathematics 47.4 (2017), pp. 437–441, ISSN: 1992-9978.
- [12] V. SKALA, M. SMOLIK: *A new approach to vector field interpolation, classification and robust critical points detection using radial basis functions*, Advances in Intelligent Systems and Computing 765 (2019), pp. 109–115, ISSN: 2194-5357, DOI: https://doi.org/10.1007/978-3-319-91192-2_12.
- [13] M. SMOLIK, V. SKALA: *Classification of Critical Points Using a Second Order Derivative*, Procedia Computer Science 108 (2017), pp. 2373–2377, DOI: <https://doi.org/10.1016/j.procs.2017.05.271>.

- [14] M. SMOLIK, V. SKALA: *Radial basis function and multi-level 2D vector field approximation*, Mathematics and Computers in Simulation 181 (2021), pp. 522–538, ISSN: 0378-4754, DOI: <https://doi.org/10.1016/j.matcom.2020.10.009>.
- [15] M. SMOLIK, V. SKALA: *Vector field second order derivative approximation and geometrical characteristics*, LNCS 10404 (2017), pp. 148–158, ISSN: 0302-9743, DOI: https://doi.org/10.1007/978-3-319-62392-4_11.
- [16] M. SMOLIK, V. SKALA, Z. MAJDISOVA: *3D vector field approximation and critical points reduction using radial basis functions*, English, International Journal of Mechanics 13 (2019), pp. 100–103, ISSN: 1998-4448.
- [17] X. WANG, Y. LIU, J. OUYANG: *A meshfree collocation method based on moving Taylor polynomial approximation for high order partial differential equations*, Engineering Analysis with Boundary Elements 116 (2020), pp. 77–92, ISSN: 0955-7997, DOI: <https://doi.org/10.1016/j.enganabound.2020.04.002>.
- [18] Q. WEN, F. XU, J.-H. YONG: *Real-time 3D eye performance reconstruction for RGBD cameras*, IEEE Transactions on Visualization and Computer Graphics 23.12 (2017), pp. 2586–2598, ISSN: 1077-2626, DOI: <https://doi.org/10.1109/TVCG.2016.2641442>.
- [19] WIKIPEDIA: *Taylor series*, Accessed: 2020-05-04, 2020, URL: https://en.wikipedia.org/wiki/Taylor_series.
- [20] E. WOBBS, M. MÖLLER, V. GALAVI, C. VUIK: *Taylor least squares reconstruction technique for material point methods*, Proceedings of the 6th European Conference on Computational Mechanics: Solids, Structures and Coupled Problems, ECCM 2018 and 7th European Conference on Computational Fluid Dynamics, ECFD 2018 (2020), pp. 806–817.
- [21] D. ZÉZÉ, M. POTIER-FERRY, Y. TAMPANGO: *Multi-point Taylor series to solve differential equations*, Discrete and Continuous Dynamical Systems - Series S 12.6 (2019), pp. 1791–1806, ISSN: 1937-1632, DOI: <https://doi.org/10.3934/dcdss.2019118>.

Vieta–Fibonacci-like polynomials and some identities*

Wanna Sriprad, Somnuk Srisawat, Peesiri Naklor

Department of Mathematics and computer science,
Faculty of Science and Technology,
Rajamangala University of Technology Thanyaburi,
Pathum Thani 12110, Thailand
wanna_sriprad@rmutt.ac.th
sommuk_s@rmutt.ac.th
1160109010257@mail.rmutt.ac.th

Submitted: April 22, 2021

Accepted: September 7, 2021

Published online: September 9, 2021

Abstract

In this paper, we introduce a new type of the Vieta polynomial, which is Vieta–Fibonacci-like polynomial. After that, we establish the Binet formula, the generating function, the well-known identities, and the sum formula of this polynomial. Finally, we present the relationship between this polynomial and the previous well-known Vieta polynomials.

Keywords: Vieta–Fibonacci polynomial, Vieta–Lucas polynomial, Vieta–Fibonacci-like polynomial

AMS Subject Classification: 11C08, 11B39, 33C45

1. Introduction

In 2002, Horadam [1] introduced the new types of second order recursive sequences of polynomials which are called Vieta–Fibonacci and Vieta–Lucas polynomials respectively. The definition of Vieta–Fibonacci and Vieta–Lucas polynomials are defined as follows:

*This research was supported by the Faculty of Science and Technology, Rajamangala University of Technology Thanyaburi

Definition 1.1 ([1]). For any natural number n the Vieta–Fibonacci polynomials sequence $\{V_n(x)\}_{n=0}^{\infty}$ and the Vieta–Lucas polynomials sequence $\{v_n(x)\}_{n=0}^{\infty}$ are defined by

$$\begin{aligned} V_n(x) &= xV_{n-1}(x) - V_{n-2}(x), & \text{for } n \geq 2, \\ v_n(x) &= xv_{n-1}(x) - v_{n-2}(x), & \text{for } n \geq 2, \end{aligned}$$

respectively, where $V_0(x) = 0$, $V_1(x) = 1$ and $v_0(x) = 2$, $v_1(x) = x$.

The first few terms of the Vieta–Fibonacci polynomials sequence are $0, 1, x, x^2 - 1, x^3 - 2x, x^4 - 3x^2 + 1$ and the first few terms of the Vieta–Lucas polynomials sequence are $2, x, x^2 - 2, x^3 - 3x, x^4 - 4x^2 + 2, x^5 - 5x^3 + 5x$. The Binet formulas of the Vieta–Fibonacci and Vieta–Lucas polynomials are given by

$$\begin{aligned} V_n(x) &= \frac{\alpha^n(x) - \beta^n(x)}{\alpha(x) - \beta(x)}, \\ v_n(x) &= \alpha^n(x) + \beta^n(x), \end{aligned}$$

respectively. Where $\alpha(x) = \frac{x + \sqrt{x^2 - 4}}{2}$ and $\beta(x) = \frac{x - \sqrt{x^2 - 4}}{2}$ are the roots the characteristic equation $r^2 - xr + 1 = 0$. We also note that $\alpha(x) + \beta(x) = x$, $\alpha(x)\beta(x) = 1$, and $\alpha(x) - \beta(x) = \sqrt{x^2 - 4}$.

Recall that the Chebyshev polynomials are a sequence of orthogonal polynomials which can be defined recursively. The n^{th} Chebyshev polynomials of the first and second kinds are denoted by $\{T_n(x)\}_{n=0}^{\infty}$ and $\{U_n(x)\}_{n=0}^{\infty}$ and are defined respectively by $T_0(x) = 1$, $T_1(x) = x$, $T_n(x) = 2xT_{n-1}(x) - T_{n-2}(x)$, for $n \geq 2$, and $U_0(x) = 1$, $U_1(x) = 2x$, $U_n(x) = 2xU_{n-1}(x) - U_{n-2}(x)$, for $n \geq 2$. These polynomials are of great importance in many areas of mathematics, particularly approximation theory. It is well known that the Chebyshev polynomials of the first kind and second kind are closely related to Vieta–Fibonacci and Vieta–Lucas polynomials. So, in [4] Vitula and Slota redefined Vieta polynomials as modified Chebyshev polynomials. The related features of Vieta and Chebyshev polynomials are given as $V_n(x) = U_n(\frac{1}{2}x)$ and $v_n(x) = 2T_n(\frac{1}{2}x)$ (see [1, 2, 5]).

In 2013, Tasci and Yalcin [6] introduced the recurrence relation of Vieta–Pell and Vieta–Pell–Lucas polynomials as follows:

Definition 1.2 ([6]). For $|x| > 1$ and for any natural number n the Vieta–Pell polynomials sequence $\{t_n(x)\}_{n=0}^{\infty}$ and the Vieta–Pell–Lucas polynomials sequence $\{s_n(x)\}_{n=0}^{\infty}$ are defined by

$$\begin{aligned} t_n(x) &= 2xt_{n-1}(x) - t_{n-2}(x), & \text{for } n \geq 2, \\ s_n(x) &= 2xs_{n-1}(x) - s_{n-2}(x), & \text{for } n \geq 2. \end{aligned}$$

respectively, where $t_0(x) = 0$, $t_1(x) = 1$ and $s_0(x) = 2$, $s_1(x) = 2x$.

The $t_n(x)$ and $s_n(x)$ are called the n^{th} Vieta–Pell polynomial and the n^{th} Vieta–Pell–Lucas polynomial respectively. Tasci and Yalcin [6] obtained the Binet form

and generating functions of Vieta–Pell and Vieta–Pell–Lucas polynomials. Also, they obtained some differentiation rules and the finite summation formulas. Moreover, the following relations are obtained

$$s_n(x) = 2T_n(x), \quad \text{and} \quad t_{n+1}(x) = U_n(x).$$

In 2015, Yalcin et al. [8], introduced and studied the Vieta–Jacobsthal and Vieta–Jacobsthal–Lucas polynomials which defined as follows:

Definition 1.3 ([8]). For any natural number n the Vieta–Jacobsthal polynomials sequence $\{G_n(x)\}_{n=0}^{\infty}$ and the Vieta–Jacobsthal–Lucas polynomials sequence $\{g_n(x)\}_{n=0}^{\infty}$ are defined by

$$\begin{aligned} G_n(x) &= G_{n-1}(x) - 2xG_{n-2}(x), \quad \text{for } n \geq 2, \\ g_n(x) &= g_{n-1}(x) - 2xg_{n-2}(x), \quad \text{for } n \geq 2, \end{aligned}$$

respectively, where $G_0(x) = 0$, $G_1(x) = 1$ and $g_0(x) = 2$, $g_1(x) = 1$.

Moreover, for any nonnegative integer k with $1 - 2^{k+2}x \neq 0$, Yalcin et al. [8] also considered the generalized Vieta–Jacobsthal polynomials sequences $\{G_{k,n}(x)\}_{n=0}^{\infty}$ and Vieta–Jacobsthal–Lucas polynomials sequences $\{g_{k,n}(x)\}_{n=0}^{\infty}$ by the following recurrence relations

$$\begin{aligned} G_{k,n}(x) &= G_{k,n-1}(x) - 2^k x G_{k,n-2}(x), \quad \text{for } n \geq 2, \\ g_{k,n}(x) &= g_{k,n-1}(x) - 2^k x g_{k,n-2}(x), \quad \text{for } n \geq 2, \end{aligned}$$

respectively, where $G_{k,0}(x) = 0$, $G_{k,1}(x) = 1$ and $g_{k,0}(x) = 2$, $g_{k,1}(x) = 1$. If $k = 1$, then $G_{1,n}(x) = G_n(x)$ and $g_{1,n}(x) = g_n(x)$. In [8], the Binet form and generating functions for these polynomials are derived. Furthermore, some special cases of the results are presented.

Recently, the generalization of Vieta–Fibonacci, Vieta–Lucas, Vieta–Pell, Vieta–Pell–Lucas, Vieta–Jacobsthal, and Vieta–Jacobsthal–Lucas polynomials have been studied by many authors.

In 2016 Kocer [3], considered the bivariate Vieta–Fibonacci and bivariate Vieta–Lucas polynomials which are generalized of Vieta–Fibonacci, Vieta–Lucas, Vieta–Pell, Vieta–Pell–Lucas polynomials. She also gave some properties. Afterward, she obtained some identities for the bivariate Vieta–Fibonacci and bivariate Vieta–Lucas polynomials by using the known properties of bivariate Vieta–Fibonacci and bivariate Vieta–Lucas polynomials.

In 2020 Uygun et al. [7], introduced the generalized Vieta–Pell and Vieta–Pell–Lucas polynomial sequences. They also gave the Binet formula, generating functions, sum formulas, differentiation rules, and some important properties for these sequences. And then they generated a matrix whose elements are of generalized Vieta–Pell terms. By using this matrix they derived some properties for generalized Vieta–Pell and generalized Vieta–Pell–Lucas polynomial sequences.

Inspired by the research going on in this direction, in this paper, we introduce a new type of Vieta polynomial, which is called Vieta–Fibonacci-like polynomial.

We also give the Binet form, the generating function, the well-known identities, and the sum formula for this polynomial. Furthermore, the relationship between this polynomial and the previous well-known Vieta polynomials are given in this study.

2. Vieta–Fibonacci-like polynomials

In this section, we introduce a new type of Vieta polynomial, called the Vieta–Fibonacci-like polynomials, as the following definition.

Definition 2.1. For any natural number n the Vieta–Fibonacci-like polynomials sequence $\{S_n(x)\}_{n=0}^{\infty}$ is defined by

$$S_n(x) = xS_{n-1}(x) - S_{n-2}(x), \quad \text{for } n \geq 2, \quad (2.1)$$

with the initial conditions $S_0(x) = 2$ and $S_1(x) = 2x$.

The first few terms of $\{S_n(x)\}_{n=0}^{\infty}$ are $2, 2x, 2x^2 - 2, 2x^3 - 4x, 2x^4 - 6x^2 + 2, 2x^5 - 8x^3 + 6x, 2x^6 - 10x^4 + 12x^2 - 2, 2x^7 - 12x^5 + 20x^3 - 8x$ and so on. The n^{th} terms of this sequence are called Vieta–Fibonacci-like polynomials.

First, we give the generating function for the Vieta–Fibonacci-like polynomials as follows.

Theorem 2.2 (The generating function). *The generating function of the Vieta–Fibonacci-like polynomials sequence is given by*

$$g(x, t) = \frac{2}{1 - xt + t^2}.$$

Proof. The generating function $g(x, t)$ can be written as $g(x, t) = \sum_{n=0}^{\infty} S_n(x)t^n$. Consider,

$$g(x, t) = \sum_{n=0}^{\infty} S_n(x)t^n = S_0(x) + S_1(x)t + S_2(x)t^2 + \cdots + S_n(x)t^n + \cdots$$

Then, we get

$$\begin{aligned} -xtg(x, t) &= -xS_0(x)t - xS_1(x)t^2 - xS_2(x)t^3 - \cdots - xS_{n-1}(x)t^n - \cdots \\ t^2g(x, t) &= S_0(x)t^2 + S_1(x)t^3 + S_2(x)t^4 + \cdots + S_{n-2}(x)t^n + \cdots \end{aligned}$$

Thus,

$$\begin{aligned} g(x, t)(1 - xt + t^2) &= S_0(x) + (S_1(x) - xS_0(x))t \\ &\quad + \sum_{n=2}^{\infty} (S_n(x) - xS_{n-1}(x) + S_{n-2}(x))t^n \\ &= 2, \end{aligned}$$

$$g(x, t) = \frac{2}{1 - xt + t^2}.$$

This completes the proof. \square

Next, we give the explicit formula for the n^{th} Vieta–Fibonacci-like polynomials.

Theorem 2.3 (Binet’s formula). *Let $\{S_n(x)\}_{n=0}^{\infty}$ be the sequence of Vieta–Fibonacci-like polynomials, then*

$$S_n(x) = A\alpha^n(x) + B\beta^n(x), \quad (2.2)$$

where $A = \frac{2(x-\beta(x))}{\alpha(x)-\beta(x)}$, $B = \frac{2(\alpha(x)-x)}{\alpha(x)-\beta(x)}$ and $\alpha(x) = \frac{x+\sqrt{x^2-4}}{2}$, $\beta(x) = \frac{x-\sqrt{x^2-4}}{2}$ are the roots of the characteristic equation $r^2 - xr + 1 = 0$.

Proof. The characteristic equation of the recurrence relation (2.1) is $r^2 - xr + 1 = 0$ and the roots of this equation are $\alpha(x) = \frac{x+\sqrt{x^2-4}}{2}$ and $\beta(x) = \frac{x-\sqrt{x^2-4}}{2}$.

It follows that

$$S_n(x) = d_1\alpha^n(x) + d_2\beta^n(x),$$

for some real numbers d_1 and d_2 . Putting $n = 0$, $n = 1$, and then solving the system of linear equations, we obtain that

$$S_n(x) = \frac{2(x-\beta(x))}{\alpha(x)-\beta(x)}\alpha^n(x) + \frac{2(\alpha(x)-x)}{\alpha(x)-\beta(x)}\beta^n(x).$$

Setting $A = \frac{2(x-\beta(x))}{\alpha(x)-\beta(x)}$ and $B = \frac{2(\alpha(x)-x)}{\alpha(x)-\beta(x)}$, we get

$$S_n(x) = A\alpha^n(x) + B\beta^n(x).$$

This completes the proof. \square

We note that $A + B = 2$, $AB = -\frac{4}{(\alpha(x)-\beta(x))^2}$, and $A\beta(x) + B\alpha(x) = 0$.

The other explicit forms of Vieta–Fibonacci-like polynomials are given in the following two theorems.

Theorem 2.4 (Explicit form). *Let $\{S_n(x)\}_{n=0}^{\infty}$ be the sequence of Vieta–Fibonacci-like polynomials. Then*

$$S_n(x) = 2 \sum_{i=0}^{\lfloor \frac{n}{2} \rfloor} (-1)^i \binom{n-i}{i} x^{n-2i}, \quad \text{for } n \geq 1.$$

Proof. From Theorem 2.2, we obtain

$$\begin{aligned} \sum_{n=0}^{\infty} S_n(x)t^n &= \frac{2}{1 - (xt - t^2)} \\ &= 2 \sum_{n=0}^{\infty} (xt - t^2)^n \end{aligned}$$

$$\begin{aligned}
&= 2 \sum_{n=0}^{\infty} \sum_{i=0}^n \binom{n}{i} (xt)^{n-i} (-t^2)^i \\
&= 2 \sum_{n=0}^{\infty} \sum_{i=0}^n \binom{n}{i} (-1)^i x^{n-i} t^{n+i} \\
&= \sum_{n=0}^{\infty} \left[2 \sum_{i=0}^{\lfloor \frac{n}{2} \rfloor} (-1)^i \binom{n-i}{i} x^{n-2i} \right] t^n.
\end{aligned}$$

From the equality of both sides, the desired result is obtained. This complete the proof. \square

Theorem 2.5 (Explicit form). *Let $\{S_n(x)\}_{n=0}^{\infty}$ be the sequence of Vieta-Fibonacci-like polynomials. Then*

$$S_n(x) = 2^{-n+1} \sum_{i=0}^{\lfloor \frac{n}{2} \rfloor} (-1)^i \binom{n+1}{2i+1} x^{n-2i} (x^2 - 4)^i, \quad \text{for } n \geq 1.$$

Proof. Consider,

$$\begin{aligned}
\alpha^{n+1}(x) - \beta^{n+1}(x) &= 2^{-(n+1)} [(x + \sqrt{x^2 - 4})^{n+1} - (x - \sqrt{x^2 - 4})^{n+1}] \\
&= 2^{-(n+1)} \left[\sum_{i=0}^{n+1} \binom{n+1}{i} x^{n-i+1} (\sqrt{x^2 - 4})^i \right. \\
&\quad \left. - \sum_{i=0}^{n+1} \binom{n+1}{i} x^{n-i+1} (-\sqrt{x^2 - 4})^i \right] \\
&= 2^{-n} \left[\sum_{i=0}^{\lfloor \frac{n}{2} \rfloor} \binom{n+1}{2i+1} x^{n-2i} (\sqrt{x^2 - 4})^{2i+1} \right].
\end{aligned}$$

Thus,

$$\begin{aligned}
S_n(x) &= A\alpha^n(x) + B\beta^n(x) \\
&= 2 \frac{\alpha^{n+1}(x) - \beta^{n+1}(x)}{\alpha(x) - \beta(x)} \\
&= 2 \frac{\alpha^{n+1}(x) - \beta^{n+1}(x)}{\sqrt{x^2 - 4}} \\
&= 2^{-n+1} \sum_{i=0}^{\lfloor \frac{n}{2} \rfloor} \binom{n+1}{2i+1} x^{n-2i} (x^2 - 4)^i.
\end{aligned}$$

This completes the proof. \square

Theorem 2.6 (Sum formula). *Let $\{S_n(x)\}_{n=0}^{\infty}$ be the sequence of Vieta-Fibonacci-like polynomials. Then*

$$\sum_{k=0}^{n-1} S_k(x) = \frac{2 - S_n(x) + S_{n-1}(x)}{2 - x}, \quad \text{for } n \geq 1.$$

Proof. By using Binet formula (2.2), we get

$$\begin{aligned} \sum_{k=0}^{n-1} S_k(x) &= \sum_{k=0}^{n-1} (A\alpha^k(x) + B\beta^k(x)) \\ &= A \frac{1 - \alpha^n(x)}{1 - \alpha(x)} + B \frac{1 - \beta^n(x)}{1 - \beta(x)} \\ &= \frac{A + B - (A\beta(x) + B\alpha(x)) - (A\alpha^n(x) + B\beta^n(x))}{1 - x + 1} \\ &\quad + \frac{A\alpha^{n-1}(x) + B\beta^{n-1}(x)}{1 - x + 1} \\ &= \frac{2 - S_n(x) + S_{n-1}(x)}{2 - x}. \end{aligned}$$

This completes the proof. \square

Since the derivative of the polynomials is always exists, we can give the following formula.

Theorem 2.7 (Differentiation formula). *The derivative of $S_n(x)$ is obtained as the follows.*

$$\frac{d}{dx} S_n(x) = \frac{(n+1)v_{n+1}(x) - xV_{n+1}(x)}{2(x^2 - 4)},$$

where $V_n(x)$ and $v_n(x)$ are the n^{th} Vieta-Fibonacci and Vieta-Lucas polynomials, respectively.

Proof. The result is obtained by using Binet formula (2.2). \square

Again, by using Binet formula (2.2), we obtain some well-known identities as follows.

Theorem 2.8 (Catalan's identity or Simson identities). *Let $\{S_n(x)\}_{n=0}^{\infty}$ be the sequence of Vieta-Fibonacci-like polynomials. Then*

$$S_n^2(x) - S_{n+r}(x)S_{n-r}(x) = S_{r-1}^2(x), \quad \text{for } n \geq r \geq 1. \quad (2.3)$$

Proof. By using Binet formula (2.2), we obtain

$$\begin{aligned} S_n^2(x) - S_{n+r}(x)S_{n-r}(x) \\ = (A\alpha^n(x) + B\beta^n(x))^2 - (A\alpha^{n+r}(x) + B\beta^{n+r}(x))(A\alpha^{n-r}(x) + B\beta^{n-r}(x)) \end{aligned}$$

$$\begin{aligned}
&= -AB(\alpha(x)\beta(x))^{n-r}(\alpha^{2r}(x) - 2(\alpha(x)\beta(x))^r + \beta^{2r}(x)) \\
&= \frac{4}{(\alpha(x) - \beta(x))^2}(\alpha^r(x) - \beta^r(x))^2 \\
&= (A\alpha^{r-1}(x) + B\beta^{r-1}(x))^2 \\
&= S_{r-1}^2(x).
\end{aligned}$$

Thus,

$$S_n^2(x) - S_{n+r}(x)S_{n-r}(x) = S_{r-1}^2(x).$$

This completes the proof. \square

Take $r = 1$ in Catalan's identity (2.3), then we get the following corollary.

Corollary 2.9 (Cassini's identity). *Let $\{S_n(x)\}_{n=0}^\infty$ be the sequence of Vieta-Fibonacci-like polynomials. Then*

$$S_n^2(x) - S_{n+1}(x)S_{n-1}(x) = 4, \quad \text{for } n \geq 1.$$

Theorem 2.10 (d' Ocagne's identity). *Let $\{S_n(x)\}_{n=0}^\infty$ be the sequence of Vieta-Fibonacci-like polynomials. Then*

$$S_m(x)S_{n+1}(x) - S_{m+1}(x)S_n(x) = 2S_{m-n-1}(x), \quad \text{for } m \geq n \geq 1. \quad (2.4)$$

Proof. We will prove d' Ocagne's identity (2.4) by using Binet formula (2.2). Consider,

$$\begin{aligned}
&S_m(x)S_{n+1}(x) - S_{m+1}(x)S_n(x) \\
&= (A\alpha^m(x) + B\beta^m(x))(A\alpha^{n+1}(x) + B\beta^{n+1}(x)) \\
&\quad - (A\alpha^{m+1}(x) + B\beta^{m+1}(x))(A\alpha^n(x) + B\beta^n(x)) \\
&= -AB(\alpha(x)\beta(x))^n(\alpha(x) - \beta(x))(\alpha^{m-n}(x) - \beta^{m-n}(x)) \\
&= \frac{4}{(\alpha(x) - \beta(x))^2}(\alpha(x) - \beta(x))(\alpha^{m-n}(x) - \beta^{m-n}(x)) \\
&= 2(A\alpha^{m-n-1}(x) + B\beta^{m-n-1}(x)) \\
&= 2S_{m-n-1}(x).
\end{aligned}$$

This completes the proof. \square

Theorem 2.11 (Honsberger identity). *Let $\{S_n(x)\}_{n=0}^\infty$ be the sequence of Vieta-Fibonacci-like polynomials. Then*

$$S_{m+1}(x)S_{n+1}(x) + S_m(x)S_n(x) = \frac{4xv_{m+n+3}(x) - 8v_{m-n}(x)}{x^2 - 4}, \quad \text{for } m \geq n \geq 1,$$

where $v_n(x)$ is the n^{th} Vieta-Lucas polynomials.

Proof. By using Binet formula (2.2), we obtain

$$\begin{aligned}
 S_{m+1}(x)S_{n+1}(x) + S_m(x)S_n(x) &= (A\alpha^{m+1}(x) + B\beta^{m+1}(x))(A\alpha^{n+1}(x) + B\beta^{n+1}(x)) \\
 &\quad + (A\alpha^m(x) + B\beta^m(x))(A\alpha^n(x) + B\beta^n(x)) \\
 &= xA^2\alpha^{m+n+1}(x) + xB^2\beta^{m+n+1}(x) + 2AB(\alpha^{m-n}(x) + \beta^{m-n}(x)) \\
 &= \frac{4x(\alpha^{m+n+3}(x) + \beta^{m+n+3}(x)) - 8(\alpha^{m-n}(x) + \beta^{m-n}(x))}{(\alpha(x) - \beta(x))^2} \\
 &= \frac{4xv_{m+n+3}(x) - 8v_{m-n}(x)}{x^2 - 4}.
 \end{aligned}$$

This completes the proof. \square

In the next theorem, we obtain the relation between the Vieta–Fibonacci-like, Vieta–Fibonacci and the Vieta–Lucas polynomials by using Binet formula (2.2).

Theorem 2.12. *Let $\{S_n(x)\}_{n=0}^\infty$, $\{V_n(x)\}_{n=0}^\infty$ and $\{v_n(x)\}_{n=0}^\infty$ be the sequences of Vieta–Fibonacci-like, Vieta–Fibonacci and Vieta–Lucas polynomials, respectively. Then*

- (1) $S_n(x) = 2V_{n+1}(x)$, for $n \geq 0$,
- (2) $S_n(x) = v_n(x) + xV_n(x)$, for $n \geq 0$,
- (3) $S_n(x)v_{n+1}(x) = 2V_{2n+2}(x)$, for $n \geq 0$,
- (4) $S_{n+1}(x) + S_{n-1}(x) = 2xV_{n+1}(x)$, for $n \geq 1$,
- (5) $S_{n+1}(x) - S_{n-1}(x) = 2v_{n+1}(x)$, for $n \geq 1$,
- (6) $S_{n+2}^2(x) - S_{n-1}^2(x) = 4xV_{2n+2}(x)$, for $n \geq 1$,
- (7) $2S_n(x) - xS_{n-1}(x) = 2v_n(x)$, for $n \geq 1$,
- (8) $S_{n+2}(x) + S_{n-2}(x) = (2x^2 - 4)V_{n+1}(x)$, for $n \geq 2$,
- (9) $S_{n+2}^2(x) - S_{n-2}^2(x) = 4x(x^2 - 2)V_{2n+2}(x)$, for $n \geq 2$,
- (10) $v_{n+1}(x) - v_n(x) = \frac{1}{2}(x^2 - 4)S_{n-1}(x)$, for $n \geq 1$,
- (11) $2v_{n+1}(x) - xv_n(x) = \frac{1}{2}(x^2 - 4)S_{n-1}(x)$, for $n \geq 1$,
- (12) $4v_n^2(x) + (x^2 - 4)S_{n-1}^2(x) = 8v_n(x)$, for $n \geq 1$,
- (13) $4v_n^2(x) - (x^2 - 4)S_{n-1}^2(x) = 16$, for $n \geq 1$.

Proof. The results (1)–(13) are easily obtained by using Binet formula (2.2). \square

3. Matrix Form of Vieta–Fibonacci-like polynomials

In this section, we establish some identities of Vieta–Fibonacci-like and Vieta–Fibonacci polynomials by using elementary matrix methods.

Let Q_s be 2×2 matrix defined by

$$Q_S = \begin{bmatrix} 2x^2 - 2 & 2x \\ -2x & -2 \end{bmatrix}. \quad (3.1)$$

Then by using this matrix we can deduce some identities of Vieta–Fibonacci-like and Vieta–Fibonacci polynomials.

Theorem 3.1. *Let $\{S_n(x)\}_{n=0}^\infty$ be the sequence of Vieta–Fibonacci-like polynomials and Q_s be 2×2 matrix defined by (3.1). Then*

$$Q_S^n = 2^{n-1} \begin{bmatrix} S_{2n}(x) & S_{2n-1}(x) \\ -S_{2n-1}(x) & -S_{2n-2}(x) \end{bmatrix}, \quad \text{for } n \geq 1.$$

Proof. For the proof, mathematical induction method is used. It obvious that the statement is true for $n = 1$. Suppose that the result is true for any positive integer k , then we also have the result is true for $k + 1$. Because

$$\begin{aligned} Q_S^{k+1} &= Q_S^k \cdot Q_S \\ &= 2^{k-1} \begin{bmatrix} S_{2k}(x) & S_{2k-1}(x) \\ -S_{2k-1}(x) & -S_{2k-2}(x) \end{bmatrix} \begin{bmatrix} 2x^2 - 2 & 2x \\ -2x & -2 \end{bmatrix} \\ &= 2^{(k+1)-1} \begin{bmatrix} S_{2(k+1)}(x) & S_{2(k+1)-1}(x) \\ -S_{2(k+1)-1}(x) & -S_{2(k+1)-2}(x) \end{bmatrix}. \end{aligned}$$

By Mathematical induction, we have that the result is true for each $n \in \mathbb{N}$, that is

$$Q_S^n = 2^{n-1} \begin{bmatrix} S_{2n}(x) & S_{2n-1}(x) \\ -S_{2n-1}(x) & -S_{2n-2}(x) \end{bmatrix}, \quad \text{for } n \geq 1. \quad \square$$

Theorem 3.2. *Let $\{S_n(x)\}_{n=0}^\infty$ be the sequence of Vieta–Fibonacci-like polynomials. Then for all integers $m \geq 1$, $n \geq 1$, the following statements hold.*

- (1) $2S_{2(m+n)}(x) = S_{2m}(x)S_{2n}(x) - S_{2m-1}(x)S_{2n-1}(x)$,
- (2) $2S_{2(m+n)-1}(x) = S_{2m}(x)S_{2n-1}(x) - S_{2m-1}(x)S_{2n-2}(x)$,
- (3) $2S_{2(m+n)-1}(x) = S_{2m-1}(x)S_{2n}(x) - S_{2m-2}(x)S_{2n-1}(x)$,
- (4) $2S_{2(m+n)-2}(x) = S_{2m-1}(x)S_{2n-1}(x) - S_{2m-2}(x)S_{2n-2}(x)$.

Proof. By Theorem 3.1 and the property of power matrix $Q_s^{m+n} = Q_s^m \cdot Q_s^n$, then we obtained the results. \square

By Theorem 3.1 and $S_n(x) = 2V_{n+1}(x)$, we get the following Corollary.

Corollary 3.3. Let $\{V_n(x)\}_{n=0}^\infty$ be the sequence of Vieta–Fibonacci polynomials and Q_s be 2×2 matrix defined by (3.1). Then

$$Q_S^n = 2^n \begin{bmatrix} V_{2n+1}(x) & V_{2n}(x) \\ -V_{2n}(x) & -V_{2n-1}(x) \end{bmatrix}, \quad \text{for } n \geq 1.$$

Proof. From Theorem 3.1, we get

$$Q_S^n = 2^{n-1} \begin{bmatrix} S_{2n}(x) & S_{2n-1}(x) \\ -S_{2n-1}(x) & -S_{2n-2}(x) \end{bmatrix}, \quad \text{for } n \geq 1.$$

Since $S_n(x) = 2V_{n+1}(x)$, we get that

$$\begin{aligned} Q_S^n &= 2^{n-1} \begin{bmatrix} 2V_{2n+1}(x) & 2V_{2n}(x) \\ -2V_{2n}(x) & -2V_{2n-1}(x) \end{bmatrix} \\ &= 2^n \begin{bmatrix} V_{2n+1}(x) & V_{2n}(x) \\ -V_{2n}(x) & -V_{2n-1}(x) \end{bmatrix}, \quad \text{for } n \geq 1. \end{aligned}$$

This completes the proof. \square

By Theorem 3.2 and $S_n(x) = 2V_{n+1}(x)$, we get the following Corollary.

Corollary 3.4. Let $\{V_n(x)\}_{n=0}^\infty$ be the sequence of Vieta–Fibonacci polynomials. Then for all integers $m \geq 1$, $n \geq 1$, the following statements hold.

- (1) $V_{2(m+n)+1}(x) = V_{2m+1}(x)V_{2n+1}(x) - V_{2m}(x)V_{2n}(x)$,
- (2) $V_{2(m+n)}(x) = V_{2m+1}(x)V_{2n}(x) - V_{2m}(x)V_{2n-1}(x)$,
- (3) $V_{2(m+n)}(x) = V_{2m}(x)V_{2n+1}(x) - V_{2m-1}(x)V_{2n}(x)$,
- (4) $V_{2(m+n)-1}(x) = V_{2m}(x)V_{2n}(x) - V_{2m-1}(x)V_{2n-1}(x)$.

Proof. From Theorem 3.2 and $S_n(x) = 2V_{n+1}(x)$, we get that

$$\begin{aligned} V_{2(m+n)+1}(x) &= \frac{1}{2} S_{2(m+n)}(x) \\ &= \frac{1}{4} (S_{2m}(x)S_{2n}(x) - S_{2m-1}(x)S_{2n-1}(x)) \\ &= \frac{1}{4} (2V_{2m+1}(x)2V_{2n+1}(x) - 2V_{2m}(x)2V_{2n}(x)) \\ &= V_{2m+1}(x)V_{2n+1}(x) - V_{2m}(x)V_{2n}(x). \end{aligned}$$

Thus, we get that (1) holds. By the same argument as above, we get that (2), (3), and (4) holds. This completes the proof. \square

By Corollary 3.4 and $S_n(x) = 2V_{n+1}(x)$, we get the following corollary.

Corollary 3.5. Let $\{S_n(x)\}_{n=0}^\infty$ and $\{V_n(x)\}_{n=0}^\infty$ be the sequences of Vieta–Fibonacci-like polynomials and Vieta–Fibonacci polynomials, respectively. Then for all integers $m \geq 1$, $n \geq 1$, the following statements hold.

$$(1) S_{2(m+n)}(x) = 2(V_{2m+1}(x)V_{2n+1}(x) - V_{2m}(x)V_{2n}(x)),$$

$$(2) S_{2(m+n)-1}(x) = 2(V_{2m+1}(x)V_{2n}(x) - V_{2m}(x)V_{2n-1}(x)),$$

$$(3) S_{2(m+n)-1}(x) = 2(V_{2m}(x)V_{2n+1}(x) + V_{2m-1}(x)V_{2n}(x)),$$

$$(4) S_{2(m+n)-2}(x) = 2(V_{2m}(x)V_{2n}(x) + V_{2m-1}(x)V_{2n-1}(x)).$$

Proof. From Corollary 3.4 and $S_n(x) = 2V_{n+1}(x)$, we get that

$$\begin{aligned} S_{2(m+n)}(x) &= 2V_{2(m+n)+1}(x) \\ &= 2(V_{2m+1}(x)V_{2n+1}(x) - V_{2m}(x)V_{2n}(x)). \end{aligned}$$

Thus, we get that (1) holds. By the same argument as above, we get that (2), (3), and (4) holds. This completes the proof. \square

Acknowledgements. The authors would like to thank the faculty of science and technology, Rajamangala University of Technology Thanyaburi (RMUTT), Thailand for the financial support. Moreover, the authors would like to thank the referees for their valuable suggestions and comments which helped to improve the quality and readability of the paper.

References

- [1] A. F. HORADAM: *Vieta polynomials*, Fibonacci Q 40.3 (2002), pp. 223–232.
- [2] E. JACOBSTHA: *Über vertauschbare polynome*, Math. Z. 63 (1955), pp. 244–276, DOI: <https://doi.org/10.1007/BF01187936>.
- [3] E. G. KOCER: *Bivariate Vieta–Fibonacci and Bivariate Vieta–Lucas Polynomials*, IOSR Journal of Mathematics 20.4 (2016), pp. 44–50, DOI: <https://doi.org/10.9790/5728-1204024450>.
- [4] *On modified Chebyshev polynomials*, J. Math. Anal. App 324.1 (2006), pp. 321–343, DOI: <https://doi.org/10.1016/j.jmaa.2005.12.020>.
- [5] N. ROBBINS: *Vieta’s triangular array and a related family of polynomials*, Int. J. Math. Math. Sci 14 (1991), pp. 239–244, DOI: <https://doi.org/10.1155/S0161171291000261>.
- [6] D. TASCI, F. YALCIN: *Vieta–Pell and Vieta–Pell–Lucas polynomials*, Adv. Difference Equ 224 (2013), pp. 1–8, DOI: <https://doi.org/10.1186/1687-1847-2013-224>.
- [7] S. UYGUN, H. KARATAS, H. AYTAZ: *Notes on generalization of Vieta–Pell and Vieta–Pell Lucas polynomials*, International Journal of Mathematics Research 12.1 (2020), pp. 5–22, DOI: <https://doi.org/10.37624/IJMR/12.1.2020.5-22>.
- [8] F. YALCIN, D. TASCI, E. ERKUS-DUMAN: *Generalized Vieta–Jacobsthal and Vieta–Jacobsthal Lucas Polynomials*, Mathematical Communications 20 (2015), pp. 241–251.

On structure of the family of regularly distributed sets with respect to the union*

Szilárd Svitek, Miklós Vontszemű

Department of Mathematics, J. Selye University,
Komárno, Slovakia
sviteks@ujs.sk
vontszemum@ujs.sk

Submitted: April 23, 2021

Accepted: October 11, 2021

Published online: October 20, 2021

Abstract

Let $0 \leq q \leq 1$ and \mathbb{N} denotes the set of all positive integers. In this paper we will be interested in the family $\mathcal{U}(x^q)$ of all regularly distributed set $X \subset \mathbb{N}$ whose ratio block sequence is asymptotically distributed with distribution function $g(x) = x^q$; $x \in (0, 1]$, and we will study the structure of this family with respect to the union.

Keywords: Ideals of sets of positive integers, distribution functions, block sequences, exponent of convergence

AMS Subject Classification: 40A05, 40A35, 11J71

1. Introduction

In the whole paper we assume $X = \{x_1 < x_2 < \dots < x_n < \dots\} \subset \mathbb{N}$ where \mathbb{N} denotes the set of all positive integers.

The following sequence derived from X

$$\frac{x_1}{x_1}, \frac{x_1}{x_2}, \frac{x_2}{x_2}, \frac{x_1}{x_3}, \frac{x_2}{x_3}, \frac{x_3}{x_3}, \dots, \frac{x_1}{x_n}, \frac{x_2}{x_n}, \dots, \frac{x_n}{x_n}, \dots \quad (1.1)$$

*This research was supported by The Slovak Research and Development Agency under the grant VEGA No. 1/0776/21.

is called *the ratio block sequence* of the set (sequence) X .

It is formed by the blocks $X_1, X_2, \dots, X_n, \dots$ where

$$X_n = \left(\frac{x_1}{x_n}, \frac{x_2}{x_n}, \dots, \frac{x_n}{x_n} \right), \quad n = 1, 2, \dots$$

is called the n -th *block*. This kind of block sequences was introduced by O. Strauch and J. T. Tóth [12] and they studied the set $G(X_n)$ of its distribution functions. Further, we will be interested in ratio block sequences of type (1.1) possessing an asymptotic distribution function, i.e. $G(X_n)$ is a singleton (see definitions in the next section).

By means of these distribution functions in [13] was defined the next families of subsets of \mathbb{N} . For $0 \leq q \leq 1$ we denote $\mathcal{U}(x^q)$ the family of all regularly distributed set $X \subset \mathbb{N}$ whose ratio block sequence is asymptotically distributed with distribution function $g(x) = x^q$; $x \in (0, 1]$.

Further in [13] the following interesting results can be seen, that λ the exponent of convergence is closely related to distributional properties of sets of positive integers. More precisely, for each $q \in [0, 1]$ the family $\mathcal{I}_{\leq q}$ of all sets $A \subset \mathbb{N}$ such that $\lambda(A) \leq q$ is identical with the family $\mathcal{I}(x^q)$ of all sets $A \subset \mathbb{N}$ which are covered by some regularly distributed set $X \in \mathcal{U}(x^q)$.

The *exponent of convergence* of a set $A \subset \mathbb{N}$ is defined by

$$\lambda(A) = \inf \left\{ s \in (0, \infty) : \sum_{n \in \mathbb{N}} a_n^{-s} < \infty \right\},$$

where $A = \{a_1 < a_2 < \dots\} \subset \mathbb{N}$.

In this paper we will be interested in the family $\mathcal{U}(x^q)$ and study the structure of this family respect to the union.

The rest of our paper is organized as follows. In Section 2 and Section 3 we recall some known definitions, notations and theorems, which will be used and extended. In Section 4 our new results are presented.

2. Definitions

The following basic definitions are from papers [9, 12, 14].

- For each $n \in \mathbb{N}$ consider the *step distribution function*

$$F(X_n, x) = \frac{\#\{i \leq n; \frac{x_i}{x_n} < x\}}{n},$$

for $x \in [0, 1)$, and for $x = 1$ we define $F(X_n, 1) = 1$.

- A non-decreasing function $g: [0, 1] \rightarrow [0, 1]$, $g(0) = 0$, $g(1) = 1$ is called a *distribution function* (abbreviated d.f.). We shall identify any two d.f.s coinciding at common points of continuity.

- A d.f. $g(x)$ is a d.f. of the sequence of blocks $X_n, n = 1, 2, \dots$, if there exists an increasing sequence $n_1 < n_2 < \dots$ of positive integers such that

$$\lim_{k \rightarrow \infty} F(X_{n_k}, x) = g(x)$$

a.e. on $[0, 1]$. This is equivalent to the weak convergence, i.e., the preceding limit holds for every point $x \in [0, 1]$ of continuity of $g(x)$.

- Denote by $G(X_n)$ the set of all d.f.s of $X_n, n = 1, 2, \dots$. The set of distribution functions of ratio block sequences was studied in [1–7, 9–12].

If $G(X_n) = \{g(x)\}$ is a singleton, the d.f. $g(x)$ is also called the *asymptotic distribution function* of X_n .

- Let λ be the convergence exponent function on the power set $2^{\mathbb{N}}$ of \mathbb{N} , i.e. for $A \subset \mathbb{N}$ put

$$\lambda(A) = \inf \left\{ t > 0 : \sum_{a \in A} \frac{1}{a^t} < \infty \right\}.$$

If $q > \lambda(A)$ then $\sum_{a \in A} \frac{1}{a^q} < \infty$ and if $q < \lambda(A)$ then $\sum_{a \in A} \frac{1}{a^q} = \infty$. In the case when $q = \lambda(A)$, the series $\sum_{a \in A} \frac{1}{a^q}$ can be either convergent or divergent.

From [8, p. 26, Exercises 113, 114], it follows that the set of all possible values of λ forms the whole interval $[0, 1]$, i.e. $\{\lambda(A) : A \subset \mathbb{N}\} = [0, 1]$ and if $A = \{a_1 < a_2 < \dots < a_n < \dots\}$ then $\lambda(A)$ can be calculated by

$$\lambda(A) = \limsup_{n \rightarrow \infty} \frac{\log n}{\log a_n}.$$

Evidently the exponent of convergence λ is a monotone set function, i.e. $\lambda(A) \leq \lambda(B)$ for $A \subset B \subset \mathbb{N}$ and also $\lambda(A \cup B) = \max\{\lambda(A), \lambda(B)\}$ holds for all $A, B \subset \mathbb{N}$.

- By means of λ the following sets were defined (see [14]):

$$\begin{aligned} \mathcal{I}_{<q} &= \{A \subset \mathbb{N} : \lambda(A) < q\} \quad \text{for } 0 < q \leq 1, \\ \mathcal{I}_{\leq q} &= \{A \subset \mathbb{N} : \lambda(A) \leq q\} \quad \text{for } 0 \leq q \leq 1 \quad \text{and} \\ \mathcal{I}_0 &= \{A \subset \mathbb{N} : \lambda(A) = 0\}. \end{aligned}$$

Obviously $\mathcal{I}_{\leq 0} = \mathcal{I}_0$ and $\mathcal{I}_{\leq 1} = 2^{\mathbb{N}}$.

For a finite set $A \subset \mathbb{N}$ we have $\lambda(A) = 0$. Consequently, $\mathcal{F}in = \{A \subset \mathbb{N} : A \text{ is finite}\} \subset \mathcal{I}_0$. Families $\mathcal{I}_{<q}, \mathcal{I}_{\leq q}$ are related for $0 < q < q' < 1$ by following inclusions (see [14, Theorem 1]),

$$\mathcal{F}in \subsetneq \mathcal{I}_0 \subsetneq \mathcal{I}_{<q} \subsetneq \mathcal{I}_{\leq q} \subsetneq \mathcal{I}_{<q'} \subsetneq \mathcal{I}_{<1},$$

and the difference of successive sets is infinite, so equality does not hold in any of the inclusions.

- Let $\mathcal{I} \subset 2^{\mathbb{N}}$. Then \mathcal{I} is called an *ideal* of subsets of positive integers, if \mathcal{I} is additive (if $A, B \in \mathcal{I}$ then $A \cup B \in \mathcal{I}$), hereditary (if $A \in \mathcal{I}$ and $B \subset A$ then $B \in \mathcal{I}$), $\mathcal{I} \supseteq \mathcal{F}in$ and $\mathbb{N} \notin \mathcal{I}$.

3. Overview of known results

In this section we mention known results related to the topic of this paper and some other ones we use in the proofs of our theorems. In the whole part in (S1)–(S7) we assume $X = \{x_1 < x_2 < \dots < x_n < \dots\} \subset \mathbb{N}$.

(S1) We will use step function

$$c_0(x) = \begin{cases} 0, & \text{if } x = 0, \\ 1, & \text{if } 0 < x \leq 1. \end{cases}$$

Assume that $G(X_n)$ is singleton, i.e., $G(X_n) = \{g(x)\}$. Then either $g(x) = c_0(x)$ for $x \in [0, 1]$; or $g(x) = x^q$ for $x \in [0, 1]$ and some fixed $0 < q \leq 1$.

[12, Theorem 8.2]

The result (S1) provides motivation to introduce the following families of subsets of \mathbb{N} (see [13]):

$$\begin{aligned} \mathcal{U}(c_0(x)) &= \{X \subset \mathbb{N} : G(X_n) = \{c_0(x)\}\}, \\ \mathcal{I}(c_0(x)) &= \{A \subset \mathbb{N} : \exists X \in \mathcal{U}(c_0(x)), A \subset X\}, \end{aligned}$$

and for $0 < q \leq 1$

$$\begin{aligned} \mathcal{U}(x^q) &= \{X \subset \mathbb{N} : G(X_n) = \{x^q\}\}, \\ \mathcal{I}(x^q) &= \{A \subset \mathbb{N} : \exists X \in \mathcal{U}(x^q), A \subset X\}. \end{aligned}$$

Obviously,

$$\mathcal{U}(c_0(x)) \subsetneq \mathcal{I}(c_0(x)), \quad \mathcal{U}(x^q) \subsetneq \mathcal{I}(x^q).$$

Sets X from $\mathcal{U}(c_0(x))$ are characterized by (S4) and sets belonging to $\mathcal{U}(x^q)$ are characterized by (S2) and (S5). In [13, Theorem 1 and Example 1] is proved that the family $\mathcal{U}(c_0(x))$ is additive, i.e. it is closed with respect to finite unions and does not form an ideal as it is not hereditary, i.e. there exists sets $C \in \mathcal{U}(c_0(x))$ and $B \subset C$ such that $B \notin \mathcal{U}(c_0(x))$. On the other hand the family $\mathcal{I}(c_0(x))$ is an ideal (see [13, Theorem 2]). For these families the following statements hold.

(S2) Let $0 < q \leq 1$ be a real number. Then

$$X \in \mathcal{U}(x^q) \iff \forall k \in \mathbb{N} : \lim_{n \rightarrow \infty} \frac{x_{kn}}{x_n} = k^{\frac{1}{q}}.$$

[6, Theorem 1]

(S3) Let $0 < q \leq 1$ be a real number and $X \in \mathcal{U}(x^q)$. Then

$$\lim_{n \rightarrow \infty} \frac{x_{n+1}}{x_n} = 1.$$

[4, Remark 3]

(S4) We have

$$X \in \mathcal{U}(c_0(x)) \iff \lim_{n \rightarrow \infty} \frac{1}{nx_n} \sum_{i=1}^n x_i = 0.$$

[12, Theorem 7.1]

(S5) Let $0 < q \leq 1$ be a real number. Then

$$X \in \mathcal{U}(x^q) \iff \lim_{n \rightarrow \infty} \frac{1}{nx_n} \sum_{i=1}^n x_i = \frac{q}{q+1}.$$

[3, Theorem 1]

(S6) Let $X \in \mathcal{U}(c_0(x))$. Then

$$\lim_{n \rightarrow \infty} \frac{\log n}{\log x_n} = 0 \text{ (i.e. } \lambda(X) = 0\text{)}.$$

[3, Theorem 2]

(S7) Let $0 < q \leq 1$ be a real number and $X \in \mathcal{U}(x^q)$. Then

$$\lim_{n \rightarrow \infty} \frac{\log n}{\log x_n} = q \text{ (therefore } \lambda(X) = q\text{)}.$$

[3, Theorem 3]

(S8) Let $0 < q \leq 1$. Then each of the families \mathcal{I}_0 , $\mathcal{I}_{<q}$ and $\mathcal{I}_{\leq q}$ forms an admissible ideal, except for $\mathcal{I}_{<1}$.

[14, Theorem 1]

(S9) Let $0 < q \leq 1$. Then each of the families $\mathcal{I}(c_0(x))$, $\mathcal{I}(x^q)$ forms an admissible ideal and $\mathcal{I}(c_0(x)) = \mathcal{I}_0$, $\mathcal{I}(x^q) = \mathcal{I}_{\leq q}$.

[13, Theorem 5 and Theorem 7]

Given $t \geq 1$, define the counting function of $X \subset \mathbb{N}$ as

$$X(t) = \#\{x \leq t : x \in X\}.$$

(S10) Let $0 < q \leq 1$, $X = \{x_1 < x_2 < \dots\} \subset \mathbb{N}$ and $Y = \{y_1 < y_2 < \dots\} \subset \mathbb{N}$.

Let $g(x) \in \{c_0(x), x^q\}$ be fixed and assume that

$$Y \in \mathcal{U}(g(x)) \quad \text{and} \quad \lim_{t \rightarrow \infty} \frac{X(t)}{Y(t)} = 0.$$

Then

$$X \cup Y \in \mathcal{U}(g(x)).$$

[13, Theorem 4]

4. Results

In this section we will study the structure of the family $\mathcal{U}(x^q)$ respect to the union of its elements. We show that there exist such sets $X, Y \in \mathcal{U}(x^q)$ that $X \cup Y \notin \mathcal{U}(x^q)$, but on the other hand, if $X, Y \in \mathcal{U}(x^q)$ (hence $\lambda(X) = q$ and $\lambda(Y) = q$) then necessary $\lambda(X \cup Y) = q$, thus

$$X \cup Y \in \mathcal{I}_{\leq q} \setminus \mathcal{I}_{< q} = \mathcal{I}(x^q) \setminus \mathcal{I}_{< q} \subsetneq \mathcal{I}(x^q).$$

This follows from the (S7), (S9) and the fact that $\lambda(X \cup Y) = \max\{\lambda(X), \lambda(Y)\}$.

Theorem 4.1. *Let $0 < q \leq 1$. Then the family $\mathcal{U}(x^q)$ does not form an ideal as it is not additive, i.e. it is not closed with respect to finite unions.*

Proof. It is sufficient to show that there exist sets $X, Y \in \mathcal{U}(x^q)$ such that $X \cup Y \notin \mathcal{U}(x^q)$. Let $0 < q \leq 1$ and $X = \{x_1 < x_2 < \dots < x_n < \dots\} \subset \mathbb{N}$ be such that $x_{n+1} > x_n + 1$ for every $n \in \mathbb{N}$ and $X \in \mathcal{U}(x^q)$. For example, it will be like that $x_n = \lfloor 2n^{\frac{1}{q}} \rfloor$ (as usual, $\lfloor x \rfloor$ is the integer part of the real x). From (S2) it is clear that $X \in \mathcal{U}(x^q)$.

Then $x_n = 2n^{\frac{1}{q}} - \varepsilon(n)$ for some $0 \leq \varepsilon(n) < 1$, and by Lagrange's Mean Value Theorem for $f(x) = 2x^{\frac{1}{q}}$ on $[n, n+1]$ we get that $x_{n+1} > x_n + 1$ for all n .

Define the set $Y = \{y_1 < y_2 < \dots < y_n < \dots\}$ such that $y_1 = x_1$ and for $n \geq 2$

$$y_n = \begin{cases} x_n - 1, & \text{if } n \in (2^{2k}, 2^{2k+1}], \quad k = 0, 1, 2, \dots, \\ x_n, & \text{if } n \in (2^{2k+1}, 2^{2k+2}], \quad k = 0, 1, 2, \dots \end{cases}$$

We show that $Y \in \mathcal{U}(x^q)$. Since $x_n - 1 \leq y_n \leq x_n$ then for every $k \in \mathbb{N}$

$$\frac{x_{kn} - 1}{x_{kn}} \frac{x_{kn}}{x_n} = \frac{x_{kn} - 1}{x_n} \leq \frac{y_{kn}}{y_n} \leq \frac{x_{kn}}{x_n - 1} = \frac{x_n}{x_n - 1} \frac{x_{kn}}{x_n}.$$

From this according to (S2) for each $k \in \mathbb{N}$ we have

$$\lim_{n \rightarrow \infty} \frac{y_{kn}}{y_n} = \lim_{n \rightarrow \infty} \frac{x_{kn}}{x_n} = k^{\frac{1}{q}},$$

thus $Y \in \mathcal{U}(x^q)$.

Further let

$$X \cup Y = \{z_1 < z_2 < \dots < z_n < \dots\}.$$

We now show that $X \cup Y \notin \mathcal{U}(x^q)$, i.e. according to (S5)

$$\lim_{n \rightarrow \infty} \frac{1}{nz_n} \sum_{i=1}^n z_i \neq \frac{q}{q+1}.$$

Let n_k ($k = 1, 2, \dots$) be such that $z_{n_k} = x_{2^{2k+1}}$. Then

$$n_k = 2^{2k+1} + \sum_{i=0}^k (2^{2i+1} - 2^{2i}) = 2^{2k+1} + \sum_{i=0}^k 2^{2i}$$

$$= 2^{2k+1} + \frac{2^{2k+2} - 1}{2^2 - 1} = \frac{5}{3}2^{2k+1} - \frac{1}{3}. \tag{4.1}$$

We estimate the following means

$$\begin{aligned} \frac{1}{n_k z_{n_k}} \sum_{i=1}^{n_k} z_i &\geq \frac{1}{n_k z_{n_k}} \left(\sum_{i=1}^{2^{2k+1}} x_i + \sum_{i=2^{2k}+1}^{2^{2k+1}} y_i \right) \\ &= \frac{1}{n_k x_{2^{2k+1}}} \left(\sum_{i=1}^{2^{2k+1}} x_i + \sum_{i=1}^{2^{2k+1}} y_i - \sum_{i=1}^{2^{2k}} y_i \right) \\ &= \frac{2^{2k+1}}{n_k} \frac{1}{2^{2k+1} x_{2^{2k+1}}} \sum_{i=1}^{2^{2k+1}} x_i \\ &\quad + \frac{2^{2k+1}}{n_k} \frac{y_{2^{2k+1}}}{x_{2^{2k+1}}} \frac{1}{2^{2k+1} y_{2^{2k+1}}} \sum_{i=1}^{2^{2k+1}} y_i \\ &\quad - \frac{2^{2k}}{n_k} \frac{y_{2^{2k}}}{x_{2^{2k+1}}} \frac{1}{2^{2k} y_{2^{2k}}} \sum_{i=1}^{2^{2k}} y_i. \end{aligned} \tag{4.2}$$

Since $X, Y \in \mathcal{U}(x^q)$ then by (S5) we give

$$\begin{aligned} \lim_{k \rightarrow \infty} \frac{1}{2^{2k+1} x_{2^{2k+1}}} \sum_{i=1}^{2^{2k+1}} x_i &= \lim_{k \rightarrow \infty} \frac{1}{2^{2k+1} y_{2^{2k+1}}} \sum_{i=1}^{2^{2k+1}} y_i \\ &= \lim_{k \rightarrow \infty} \frac{1}{2^{2k} y_{2^{2k}}} \sum_{i=1}^{2^{2k}} y_i = \frac{q}{q+1}. \end{aligned}$$

From definition of the set Y and (S2) it follows

$$\lim_{k \rightarrow \infty} \frac{y_{2^{2k}}}{x_{2^{2k+1}}} = \lim_{k \rightarrow \infty} \frac{x_{2^{2k}}}{x_{2^{2k+1}}} = \lim_{k \rightarrow \infty} \frac{x_{2^{2k}}}{x_{2 \cdot 2^{2k}}} = \frac{1}{2^{\frac{1}{q}}} \leq \frac{1}{2}.$$

Furthermore we have

$$\lim_{k \rightarrow \infty} \frac{y_{2^{2k+1}}}{x_{2^{2k+1}}} = \lim_{k \rightarrow \infty} \frac{x_{2^{2k+1}-1}}{x_{2^{2k+1}}} = 1,$$

and (4.1) implies

$$\lim_{k \rightarrow \infty} \frac{2^{2k+1}}{n_k} = \frac{3}{5}, \quad \lim_{k \rightarrow \infty} \frac{2^{2k}}{n_k} = \frac{3}{10}.$$

Then from estimation (4.2) by previously statements we obtain

$$\liminf_{k \rightarrow \infty} \frac{1}{n_k z_{n_k}} \sum_{i=1}^{n_k} z_i \geq \left(\frac{3}{5} + \frac{3}{5} \cdot 1 - \frac{3}{10} \cdot \frac{1}{2} \right) \frac{q}{q+1} = \frac{21}{20} \frac{q}{q+1} > \frac{q}{q+1},$$

which it means that $X \cup Y \notin \mathcal{U}(x^q)$. □

However, if we choose such sets $X, Y \in \mathcal{U}(x^q)$ that $X \cap Y \in \mathcal{I}_0$, then holds already the following.

Theorem 4.2. *Let $0 < q \leq 1$ and sets $X, Y \in \mathcal{U}(x^q)$ are such that $X \cap Y \in \mathcal{I}_0$. Then $X \cup Y \in \mathcal{U}(x^q)$.*

Proof. Let $0 < q \leq 1$, $X = \{x_1 < x_2 < \dots\} \subset \mathbb{N}$, $Y = \{y_1 < y_2 < \dots\} \subset \mathbb{N}$. Assume that $X, Y \in \mathcal{U}(x^q)$. According to (S5) and (S3) we have

$$\frac{1}{nx_n} \sum_{i=1}^n x_i \rightarrow \frac{q}{q+1} \quad \text{and} \quad \frac{1}{ny_n} \sum_{i=1}^n y_i \rightarrow \frac{q}{q+1} \quad \text{as } n \rightarrow \infty, \quad (4.3)$$

and

$$\frac{x_{k+1}}{x_k} \rightarrow 1 \quad \text{and} \quad \frac{y_{k+1}}{y_k} \rightarrow 1 \quad \text{as } n \rightarrow \infty. \quad (4.4)$$

Let $X \cap Y = \{y_{i_1}, y_{i_2}, \dots, y_{i_n}, \dots\}$. We denote

$$A(X \cap Y, y_n) = \sum_{y_{n_i} \in [1, y_n]} y_{n_i}.$$

Further, let $X \cup Y = \{z_1 < z_2 < \dots < z_m < \dots\}$ and choose sufficiently large $m \in \mathbb{N}$. Let $z_m \in X \cup Y$. If $z_m = y_n$ then

$$x_k \leq y_n < x_{k+1} \quad \text{and} \quad y_{i_l} \leq y_n < y_{i_{l+1}},$$

for some $k, l \in \mathbb{N}$.

Thus $m = X \cup Y(y_n)$, $X \cap Y(y_n) = l$ and $m = k + n - l$. Then we estimate the value

$$\begin{aligned} \frac{1}{mz_m} \sum_{i=1}^m z_i &= \frac{1}{k+n-l} \frac{1}{y_n} \left(\sum_{i=1}^n y_i + \sum_{i=1}^k x_i - A(X \cap Y, y_n) \right) \\ &= \frac{n}{k+n-l} \frac{1}{ny_n} \sum_{i=1}^n y_i + \frac{k}{k+n-l} \frac{x_k}{y_n} \frac{1}{kx_k} \sum_{i=1}^k x_i - \frac{A(X \cap Y, y_n)}{(k+n-l)y_n} \\ &= \frac{k+n}{k+n-l} \frac{1}{ny_n} \sum_{i=1}^n y_i + \frac{k}{k+n-l} \left(\frac{x_k}{y_n} \frac{1}{kx_k} \sum_{i=1}^k x_i - \frac{1}{ny_n} \sum_{i=1}^n y_i \right) - \frac{A(X \cap Y, y_n)}{(k+n-l)y_n}. \end{aligned} \quad (4.5)$$

On the other hand

$$\begin{aligned} \frac{k+n}{k+n-l} &= 1 - \frac{X \cap Y(y_n)}{X \cup Y(y_n)}, \\ 0 &\leq \frac{A(X \cap Y, y_n)}{(k+n-l)y_n} \leq \frac{X \cap Y(y_n) \cdot y_n}{(k+n-l)y_n} = \frac{X \cap Y(y_n)}{X \cup Y(y_n)} \leq \frac{X \cap Y(y_n)}{X(y_n)}, \end{aligned}$$

and as $m \rightarrow \infty$, also $k \rightarrow \infty$ and $n \rightarrow \infty$. Since from Theorem 4.3 we have

$$\frac{X \cap Y(n)}{X(n)} \rightarrow 0 \quad \text{as } n \rightarrow \infty,$$

then holds

$$\frac{k+n}{k+n-l} \rightarrow 1, \quad \frac{A(X \cap Y, y_n)}{(k+n-l)y_n} \rightarrow 0 \quad \text{as } m \rightarrow \infty.$$

Furthermore from (4.4) and condition $x_k \leq y_n < x_{k+1}$ we obtain

$$\frac{x_k}{y_n} \rightarrow 1 \quad \text{as } m \rightarrow \infty.$$

Then by (4.3), (4.5) and from the fact, that $\frac{k}{k+n-l}$ is bounded we have

$$\frac{1}{mz_m} \sum_{i=1}^m z_i \rightarrow \frac{q}{q+1} \quad \text{as } m \rightarrow \infty,$$

thus $X \cup Y \in \mathcal{U}(x^q)$.

The proof in the case $z_m = x_k$ and $y_n \leq x_k \leq y_{n+1}$ is similar. □

In the following theorems we will deal with sets X, Y for which $X \in \mathcal{U}(g_1(x))$ $Y \in \mathcal{U}(g_2(x))$ where $g_1(x) \neq g_2(x)$ and $g_1(x), g_2(x) \in \{c_0(x), x^q\}$.

Theorem 4.3. *Let $0 < q \leq 1$ and sets $X \in \mathcal{U}(c_0(x))$ (it can also be $X \in \mathcal{I}_0$), $Y \in \mathcal{U}(x^q)$. Then*

$$\lim_{n \rightarrow \infty} \frac{X(n)}{Y(n)} = 0.$$

Proof. Let $0 < q \leq 1$, $X = \{x_1 < x_2 < \dots\} \subset \mathbb{N}$, $Y = \{y_1 < y_2 < \dots\} \subset \mathbb{N}$. Assume that $X \in \mathcal{U}(c_0(x))$ and $Y \in \mathcal{U}(x^q)$. Then by (S6) and (S7) for sufficiently large $k \in \mathbb{N}$ there exists $n_0 \in \mathbb{N}$ such that for every $n \geq n_0$ we have

$$x_n > n^k \quad \text{and} \quad y_n < n^{\frac{1}{q} + \frac{1}{k}}.$$

Therefore

$$0 \leq \frac{X(n)}{Y(n)} < \frac{n^{\frac{1}{k}}}{n^{\frac{qk}{q+k}}} = n^{\frac{1}{k} - \frac{qk}{q+k}},$$

where the exponent for sufficiently large k is negative, since $\frac{1}{k} - \frac{qk}{q+k} \rightarrow -q$ as $k \rightarrow \infty$. From this and previous estimation follows $\frac{X(n)}{Y(n)} \rightarrow 0$ as $n \rightarrow \infty$. □

Note that the previous Theorem 4.3 holds even if for the sets $X = \{x_1 < x_2 < \dots\} \subset \mathbb{N}$, $Y = \{y_1 < y_2 < \dots\} \subset \mathbb{N}$ we assume that

$$\lim_{n \rightarrow \infty} \frac{\log n}{\log x_n} = 0 \quad (\text{i.e. } X \in \mathcal{I}_0) \quad \text{and} \quad \lim_{n \rightarrow \infty} \frac{\log n}{\log y_n} = q.$$

On the other hand we have.

Corollary 4.4. *Let $0 < q \leq 1$ and sets $X \in \mathcal{U}(c_0(x))$, $Y \in \mathcal{U}(x^q)$. Then*

$$X \cup Y \in \mathcal{U}(x^q).$$

Proof. This is a direct corollary of Theorem 4.3 and (S10). □

Theorem 4.5. *Let $0 < q_1 < q_2 \leq 1$ and sets $X \in \mathcal{U}(x^{q_1})$, $Y \in \mathcal{U}(x^{q_2})$. Then*

$$\lim_{n \rightarrow \infty} \frac{X(n)}{Y(n)} = 0.$$

Proof. Let $0 < q_1 < q_2 \leq 1$, $X = \{x_1 < x_2 < \dots\} \subset \mathbb{N}$, $Y = \{y_1 < y_2 < \dots\} \subset \mathbb{N}$. Assume that $X \in \mathcal{U}(x^{q_1})$ and $Y \in \mathcal{U}(x^{q_2})$. Then by (S7) for sufficiently large $k \in \mathbb{N}$ there exists $n_0 \in \mathbb{N}$ such that for every $n \geq n_0$ we have

$$x_n > n^{\frac{1}{q_1} - \frac{1}{k}} \quad \text{and} \quad y_n < n^{\frac{1}{q_2} + \frac{1}{k}}.$$

Therefore

$$0 \leq \frac{X(n)}{Y(n)} < \frac{n^{\frac{q_1 k}{q_1 + k}}}{n^{\frac{q_2 k}{q_2 + k}}} = n^{\frac{q_1 k}{q_1 + k} - \frac{q_2 k}{q_2 + k}},$$

where the exponent for sufficiently large k is negative, since $\frac{q_1 k}{q_1 + k} - \frac{q_2 k}{q_2 + k} \rightarrow q_1 - q_2$ as $k \rightarrow \infty$. From this and previous estimation follows $\frac{X(n)}{Y(n)} \rightarrow 0$ as $n \rightarrow \infty$. □

Note that the previous Theorem 4.5 holds even if for the sets $X = \{x_1 < x_2 < \dots\} \subset \mathbb{N}$, $Y = \{y_1 < y_2 < \dots\} \subset \mathbb{N}$ we assume that

$$\lim_{n \rightarrow \infty} \frac{\log n}{\log x_n} = q_1 \quad \text{and} \quad \lim_{n \rightarrow \infty} \frac{\log n}{\log y_n} = q_2.$$

Corollary 4.6. *Let $0 < q_1 < q_2 \leq 1$ and sets $X \in \mathcal{U}(x^{q_1})$, $Y \in \mathcal{U}(x^{q_2})$. Then*

$$X \cup Y \in \mathcal{U}(x^{q_2}).$$

Proof. This is a direct corollary of Theorem 4.5 and result (S10). □

References

- [1] V. BALÁŽ, L. MIŠÍK, O. STRAUCH, J. T. TÓTH: *Distribution functions of ratio sequences, III*, Publ. Math. Debrecen 82 (2013), pp. 511–529, DOI: <https://doi.org/10.5486/PMD.2013.4770>.
- [2] V. BALÁŽ, L. MIŠÍK, O. STRAUCH, J. T. TÓTH: *Distribution functions of ratio sequences, IV*, Period. Math. Hung. 66 (2013), pp. 1–22, DOI: <https://doi.org/10.1007/s10998-013-4116-4>.
- [3] J. BUKOR, F. FILIP, J. T. TÓTH: *On properties derived from different types of asymptotic distribution functions of ratio sequences*, Publ. Math. Debrecen 95.1-2 (2019), pp. 219–230, DOI: <https://doi.org/10.5486/PMD.2019.8498>.
- [4] F. FILIP, L. MIŠÍK, J. T. TÓTH: *On distribution function of certain block sequences*, Unif. Distrib. Theory 2 (2007), pp. 115–126.
- [5] F. FILIP, L. MIŠÍK, J. T. TÓTH: *On ratio block sequences with extreme distribution function*, Math. Slovaca 59 (2009), pp. 275–282, DOI: <https://doi.org/10.2478/s12175-009-0123-6>.

- [6] F. FILIP, J. T. TÓTH: *Characterization of asymptotic distribution functions of ratio block sequences*, Period. Math. Hung. 60.2 (2010), pp. 115–126,
DOI: <https://doi.org/10.1007/s10998-010-2115-2>.
- [7] G. GREKOS, O. STRAUCH: *Distribution functions of ratio sequences, II*, Unif. Distrib. Theory 2 (2007), pp. 53–77.
- [8] G. PÓLYA, G. SZEGŐ: *Problems and Theorems in Analysis I*. Berlin, Heidelberg, New York: Springer-Verlag, 1978.
- [9] O. STRAUCH: *Distribution functions of ratio sequences. An expository paper*, Tatra Mt. Math. Publ. 64 (2015), pp. 133–185,
DOI: <https://doi.org/10.1515/tmmp-2015-0047>.
- [10] O. STRAUCH: *Distribution of Sequences: A Theory*, VEDA and Academia, 2019.
- [11] O. STRAUCH, Š. PORUBSKÝ: *Distribution of Sequences: A Sampler*, Frankfurt am Main: Peter Lang, 2005.
- [12] O. STRAUCH, J. T. TÓTH: *Distribution functions of ratio sequences*, Publ. Math. Debrecen 58 (2001), pp. 751–778.
- [13] J. T. TÓTH, J. BUKOR, F. FILIP, L. MIŠÍK: *On ideals defined by asymptotic distribution functions of ratio block sequences*, Filomat (2021), to appear.
- [14] J. T. TÓTH, F. FILIP, J. BUKOR, L. ZSILINSZKY: $\mathcal{I}_{<q}$ - and $\mathcal{I}_{\leq q}$ -convergence of arithmetic functions, Period. Math. Hung. 82.2 (2021), pp. 125–135,
DOI: <https://doi.org/10.1007/s10998-020-00345-y>.

The Diophantine equation

$$x^2 + 3^a \cdot 5^b \cdot 11^c \cdot 19^d = 4y^n$$

Nguyen Xuan Tho

School of Applied Mathematics and Informatics,
Hanoi University of Science and Technology
tho.nguyensexuan1@hust.edu.vn

Submitted: April 4, 2021

Accepted: August 19, 2021

Published online: August 23, 2021

Abstract

We investigate the Diophantine equation $x^2 + 3^a \cdot 5^b \cdot 11^c \cdot 19^d = 4y^n$ with $n \geq 3$, $x, y, a, b, c, d \in \mathbb{N}$, $x, y > 0$, and $\gcd(x, y) = 1$.

Keywords: Diophantine equations, Lesbegue–Ramanujan–Nagell equations, primitive divisors of Lucas numbers

AMS Subject Classification: 11D61, 11D72

1. Introduction

Let D be a positive integer. The equation

$$x^2 + D = 4y^n \tag{1.1}$$

is called a Lesbegue-Ramanujan-Nagell equation. It has been studied by several authors. Luca, Tengely, and Togbé [7] studied (1.1) when $1 \leq D \leq 100$ and $D \not\equiv 1 \pmod{4}$, $D = 7^a \cdot 11^b$, or $D = 7^a \cdot 13^b$, where $a, b \in \mathbb{N}$. Bhattar, Hoque, and Sharma [1] studied (1.1) when $D = 19^{2k+1}$, where $k \in \mathbb{N}$. Chakraborty, Hoque, and Sharma [4] studied (1.1) when $D = p^m$, where $p \in \{1, 2, 3, 7, 11, 19, 43, 67, 163\}$ and $m \in \mathbb{N}$. For a comprehensive survey of equation (1.1) and other Lesbegue-Ramanujan-Nagell type equations, see Le and Soydan [6] with over 350 references. In this paper, we study (1.1) when $D = 3^a \cdot 5^b \cdot 11^c \cdot 19^d$. It can be deduced from our work all solutions to (1.1) when the set of prime divisors of D is a *proper* subset of $\{3, 5, 11, 19\}$. The main result is the following.

Theorem 1.1. *All integer solutions (n, a, b, c, d, x, y) to the equation*

$$x^2 + 3^a \cdot 5^b \cdot 11^c \cdot 19^d = 4y^n$$

with

(i) $n \geq 3, a, b, c, d \geq 0, x, y > 0, \gcd(x, y) = 1,$

(ii) $(a, b, c, d) \not\equiv (1, 1, 1, 1) \pmod{2}$ if $5 \mid n,$

are given in Tables 1, 4, 5, 7, and 8.

Our main tool is the so-called primitive divisor theorem of Lucas numbers by Bilu, Hanrot, and Voutier [2].

2. Preliminaries

Let α and β be two algebraic integers such that $\alpha + \beta$ and $\alpha\beta$ are nonzero coprime integers, and $\frac{\alpha}{\beta}$ is not a root of unity. The Lucas sequence $(L_n)_{n \geq 1}$ is defined by

$$L_n = \frac{\alpha^n - \beta^n}{\alpha - \beta} \quad \text{for all } n \geq 1.$$

A prime number p is called a primitive divisor of L_n if

$$p \mid L_n \quad \text{but} \quad p \nmid (\alpha - \beta)^2 L_1 \cdots L_{n-1}.$$

From the work of Bilu, Hanrot, and Voutier's [2] we know

(i) if q is a primitive divisor of L_n , then $n \mid q - \left(\frac{(\alpha - \beta)^2}{q}\right),$

(ii) if $n > 30$, then L_n has a primitive divisor,

(iii) for all $4 < n \leq 30$, if L_n does not have a primitive divisor, then (n, α, β) can be derived from Table 1 in [2].

3. Proof of Theorem 1.1

From

$$x^2 + 3^a \cdot 5^b \cdot 11^c \cdot 19^d = 4y^n \tag{3.1}$$

we have $2 \nmid x$. Reducing (3.1) mod 4 gives $1 + (-1)^{a+c+d} \equiv 0 \pmod{4}$. Hence, $2 \nmid a+c+d$. Note that $x, y > 0, \gcd(x, y) = 1$, and $n \geq 3$. Write $3^a \cdot 5^b \cdot 11^c \cdot 19^d = AB^2$, where $A, B \in \mathbb{Z}^+$ and A is square-free. Here $A \in \{3, 11, 15, 19, 55, 95, 627, 3135\}$. Let $K = \mathbb{Q}(\sqrt{-A})$. Let $h(K)$ and \mathcal{O}_K be the class number and the ring of integers of K respectively. Then $h(K) \in \{1, 2, 4, 8, 40\}$ and $K = \mathbb{Z} \left[\frac{1 + \sqrt{-A}}{2} \right]$.

Assume now that n is an odd prime not dividing $h(K)$. Then

$$\left(\frac{x + B\sqrt{-A}}{2}\right) \left(\frac{x - B\sqrt{-A}}{2}\right) = (y)^n. \tag{3.2}$$

Since x and AB^2 are odd, the two ideals $\left(\frac{x+B\sqrt{-A}}{2}\right)$ and $\left(\frac{x-B\sqrt{-A}}{2}\right)$ are coprime. We also have $n \nmid h(A)$, so (3.2) implies that

$$\frac{x + B\sqrt{-A}}{2} = u\alpha^n, \tag{3.3}$$

where u is a unit in \mathcal{O}_K and $\alpha \in \mathcal{O}_K$. Since the order of the unit group of \mathcal{O}_K is a power of 2, it is coprime to n . Therefore, in (3.3) u can be absorbed into α . So we can assume $u = 1$. Let $\alpha = \frac{r+s\sqrt{-A}}{2}$ and $\beta = \frac{r-s\sqrt{-A}}{2}$, where $r, s \in \mathbb{Z}$ and $r \equiv s \pmod{2}$. We claim r and s are coprime odd integers. If r and s are even, let $r_1 = \frac{r}{2}$ and $s_1 = \frac{s}{2}$. Then

$$x = \frac{\alpha^n + \beta^n}{2} = 2 \sum_{k=0}^{\frac{n-1}{2}} \binom{n}{2k} r_1^{n-2k} (-A)^k s_1^{2k},$$

impossible since $2 \nmid x$. Therefore r and s are odd. Then

$$x = \sum_{k=0}^{\frac{n-1}{2}} \binom{n}{2k} r^{n-2k} (-A)^k s^{2k}.$$

Let $l = \gcd(r, s)$. Then $l \mid x$ and $l \mid \frac{r^2+As^2}{4}$. Hence, $l \mid \gcd(x, y)$. Therefore $l = 1$. So $\gcd(r, s) = 1$. Let $q = \gcd(r, A)$. Since $|y| = \frac{r^2+As^2}{4}$, we have $q \mid y$. Since $x^2 + AB^2 = 4y^n$, we have $q \mid x^2$. Since $\gcd(x, y) = 1$, we have $q = 1$. Since $\alpha + \beta = r$ and $\alpha\beta = \frac{r^2+As^2}{4}$, we have $\alpha + \beta$ and $\alpha\beta$ are coprime integers.

The proof of Theorem 1.1 is now achieved by means of the following four lemmas. We only require the condition $(a, b, c, d) \not\equiv (1, 1, 1, 1) \pmod{2}$ in the Lemma 3.5. So Lemmas 3.1, 3.2, 3.3, 3.4 give all solutions to (1.1) in each case of n with $\gcd(x, y) = 1$.

Lemma 3.1. *All solutions (n, a, b, c, d, x, y) to (3.1) with $n = 3$ are given in Table 1.*

Table 1. Solutions to (3.1) with $n = 3$ and $\gcd(x, y) = 1$.

(n, a, b, c, d, x, y)	(n, a, b, c, d, x, y)
$(3, 1, 0, 0, 0, 1, 1)$	$(3, 1, 0, 0, 0, 37, 7)$
$(3, 1, 0, 0, 2, 17, 7)$	$(3, 1, 0, 0, 4, 719, 61)$
$(3, 7, 1, 0, 4, 19307, 766)$	$(3, 7, 1, 2, 0, 15599, 394)$
$(3, 7, 1, 4, 2, 111946687, 146326)$	$(3, 7, 3, 1, 2, 2043331, 10144)$

(3, 7, 3, 4, 0, 2073287, 10246)	(3, 7, 3, 4, 2, 2495189, 12424)
(3, 3, 1, 0, 0, 11, 4)	(3, 3, 1, 6, 0, 96433, 1336)
(3, 9, 1, 0, 2, 443531, 3664)	(3, 3, 1, 2, 0, 7, 16)
(3, 3, 7, 14, 2, 380377270937, 47690296)	(3, 3, 1, 2, 2, 5771, 214)
(3, 3, 9, 1, 2, 2, 397447, 3436)	(3, 3, 1, 2, 2, 28267, 586)
(3, 3, 1, 2, 2, 154757, 1816)	(3, 3, 7, 2, 2, 43847521, 78334)
(3, 3, 3, 0, 0, 2761, 124)	(3, 3, 3, 0, 2, 1883, 106)
(3, 3, 3, 2, 0, 3107, 136)	(3, 3, 3, 2, 2, 1271, 334)
(3, 3, 5, 0, 4, 271051, 2764)	(3, 27, 5, 1, 1, 1291606603, 1184566)
(3, 3, 5, 10, 2, 10684962781, 3063094)	(3, 4, 1, 0, 1, 9673, 286)
(3, 5, 1, 0, 0, 623, 46)	(3, 5, 1, 0, 2, 781, 64)
(3, 11, 1, 1, 1, 74333, 1126)	(3, 5, 1, 1, 1, 1824473, 9406)
(3, 5, 1, 2, 0, 101, 34)	(3, 11, 1, 2, 0, 11877401, 32794)
(3, 5, 1, 2, 4, 873907, 5806)	(3, 5, 7, 2, 4, 1169073209, 699154)
(3, 5, 1, 4, 0, 713, 166)	(3, 11, 1, 4, 6, 1399486399, 862744)
(3, 5, 3, 0, 6, 778921, 7984)	(3, 5, 3, 2, 2, 41803, 916)
(3, 5, 5, 6, 0, 8694731, 26794)	

Proof. Write $a = 6a_1 + \epsilon_1$, $b = 6b_1 + \epsilon_2$, $c = 6c_1 + \epsilon_3$, and $d = 6d_1 + \epsilon_4$, where $a_1, b_1, c_1, d_1 \in \mathbb{N}$ and $\epsilon_1, \epsilon_2, \epsilon_3, \epsilon_4 \in \{0, 1, \dots, 5\}$. Let $D_1 = 3^{\epsilon_1} \cdot 5^{\epsilon_2} \cdot 11^{\epsilon_3} \cdot 19^{\epsilon_4}$. From (3.1) we have

$$Y^2 = X^3 - 16D_1, \tag{3.4}$$

where $X = \frac{4y}{3^{2a_1} \cdot 5^{2b_1} \cdot 11^{2c_1} \cdot 19^{2d_1}}$ and $Y = \frac{4x}{3^{3a_1} \cdot 5^{3b_1} \cdot 11^{3c_1} \cdot 19^{3d_1}}$. Since $2 \nmid a + c + d$, we have $2 \nmid \epsilon_1 + \epsilon_3 + \epsilon_4$. We use Magma [3] to search for S -integral points on (3.4), where $S = \{3, 5, 11, 19\}$. Solutions to (3.1) deduced from these S -integral points are listed in Table 1. We are able to find S -integral points on (3.4) for all but the cases of $(\epsilon_1, \epsilon_2, \epsilon_3, \epsilon_4)$ listed in Table 2.

Table 2

$(\epsilon_1, \epsilon_2, \epsilon_3, \epsilon_4)$	$(\epsilon_1, \epsilon_2, \epsilon_3, \epsilon_4)$	$(\epsilon_1, \epsilon_2, \epsilon_3, \epsilon_4)$	$(\epsilon_1, \epsilon_2, \epsilon_3, \epsilon_4)$
(0, 1, 5, 4)	(0, 4, 5, 4)	(1, 1, 5, 5)	(1, 2, 1, 5)
(1, 2, 3, 5)	(1, 2, 5, 3)	(1, 2, 5, 5)	(1, 3, 3, 5)
(1, 3, 5, 3)	(1, 3, 5, 5)	(1, 4, 1, 5)	(1, 4, 3, 5)
(1, 4, 5, 1)	(1, 4, 5, 5)	(1, 5, 3, 3)	(1, 5, 5, 3)
(1, 5, 5, 5)	(3, 1, 5, 3)	(3, 1, 5, 5)	(3, 3, 1, 5)
(3, 3, 5, 3)	(3, 4, 5, 3)	(3, 5, 3, 5)	(4, 1, 3, 4)
(4, 1, 5, 4)	(4, 3, 5, 4)	(4, 4, 4, 5)	(4, 5, 1, 4)
(4, 5, 3, 4)	(4, 5, 4, 5)	(4, 5, 5, 2)	(4, 5, 5, 4)
(5, 0, 3, 5)	(5, 0, 5, 5)		

We will show that (3.1) has no solutions for these cases of $(\epsilon_1, \epsilon_2, \epsilon_3, \epsilon_4)$. Since

$3 \nmid h(K)$, there exist coprime odd integers r, s such that

$$\frac{x + B\sqrt{-A}}{2} = \left(\frac{r + s\sqrt{-A}}{2} \right)^3.$$

Comparing the imaginary parts gives

$$4B = s(3r^2 - As^2). \tag{3.5}$$

Notice that $B = 3^{3a_1+u_1} \cdot 5^{3b_1+u_2} \cdot 11^{3c_1+u_3} \cdot 19^{3d_1+u_4}$, where $u_i = \lfloor \frac{\epsilon_i}{2} \rfloor$ for $i = 1, 2, 3, 4$. Hence,

$$4 \cdot 3^{3a_1+u_1} \cdot 5^{3b_1+u_2} \cdot 11^{3c_1+u_3} \cdot 19^{3d_1+u_4} = s(3r^2 - As^2). \tag{3.6}$$

Case 1: $A = 11$. Then $(\epsilon_1, \epsilon_2, \epsilon_3, \epsilon_4) = (0, 4, 5, 4)$. Hence, (3.6) reduces to

$$4 \cdot 3^{3a_1} \cdot 5^{3b_1+2} \cdot 11^{3c_1+2} \cdot 19^{3d_1+2} = s(3r^2 - 11s^2). \tag{3.7}$$

If $11 \mid r$, then $11 \nmid s$. Hence, $11^2 \nmid s(3r^2 - 11s^2)$. Thus, (3.7) is impossible. So $11 \nmid r$. Hence, $11^{3c_1+2} \mid s$. Since $\left(\frac{3 \cdot 11}{5}\right) = -1$ and $\gcd(r, s) = 1$, we have $5 \nmid 3r^2 - 11s^2$. Hence, $5^{3b_1+2} \mid s$. Since $\left(\frac{3 \cdot 11}{19}\right) = -1$, we have $19 \nmid 3r^2 - 11s^2$. Therefore $19^{3d_1+2} \mid s$.

Case 1.1: $a_1 > 0$. Reducing (3.7) mod 3 gives $3 \mid s$. Hence, $3^{3a_1} \mid s$. Since $2 \nmid s$, we have $s = 3^{3a_1-1} \cdot 5^{3b_1+2} \cdot 11^{3c_1+2} \cdot 19^{3d_1+2} \cdot s_1$, where $s_1 \in \{\pm 1\}$. Then (3.7) reduces to

$$4 = s_1(r^2 - 3^{6a_1-3} \cdot 5^{6b_1+4} \cdot 11^{6c_1+5} \cdot 19^{6d_1+4}). \tag{3.8}$$

Since $a_1 > 0$, we have $6a_1 - 3 > 0$. Reducing (3.8) mod 3 shows $s_1 = 1$. Then (3.8) reduces to

$$4 = r^2 - 3^{6a_1-3} \cdot 5^{6b_1+4} \cdot 11^{6c_1+5} \cdot 19^{6d_1+4}. \tag{3.9}$$

Reducing mod 7 shows

$$4 \equiv r^2 - 6 \pmod{7},$$

impossible mod 7 since $\left(\frac{10}{7}\right) = -1$.

Case 1.2: $a_1 = 0$. Since $2 \nmid s$, we have $s = \pm 5^{3b_1+2} \cdot 11^{3c_1+2} \cdot 19^{3d_1+2}$. Then (3.7) reduces to

$$4 = \pm(3r^2 - 11s^2),$$

impossible mod 5 since $5 \mid s$, $5 \nmid r$, and $\left(\frac{\pm 3}{5}\right) = -1$.

Case 2: $A = 19$. Then $(\epsilon_1, \epsilon_2, \epsilon_3, \epsilon_4) = (4, 4, 4, 5)$. Hence, (3.6) reduces to

$$4 \cdot 3^{3a_1+2} \cdot 5^{3b_1+2} \cdot 11^{3c_1+2} \cdot 19^{3d_1+2} = s(3r^2 - 19s^2). \tag{3.10}$$

If $19 \mid r$, then $19 \nmid s$. Hence, $19^2 \nmid s(3r^2 - 19s^2)$, so (3.8) is impossible mod 19^2 . Therefore $19 \nmid r$. Hence, $19^{3d_1+2} \mid s$. Since $\left(\frac{3 \cdot 19}{5}\right) = \left(\frac{3 \cdot 19}{11}\right) = -1$, we have $5 \nmid 3r^2 - 19s^2$ and $11 \nmid 3r^2 - 19s^2$. Hence, $5^{3b_1+2} \cdot 11^{3c_1+2} \mid s$. Reducing (3.8) mod 3

shows that $3 \mid s$. Hence, $3^{3a_1+1} \mid s$. Therefore $s = 3^{3a_1+1} \cdot 5^{3b_1+2} \cdot 11^{3c_1+2} \cdot 19^{3d_1+2} \cdot s_1$, where $s_1 \in \{\pm 1\}$. Then (3.10) reduces to

$$4 = s_1(r^2 - 3^{6a_1+1} \cdot 5^{6b_1+2} \cdot 11^{6c_1+4} \cdot 19^{6d_1+5} \cdot s_1^2). \quad (3.11)$$

Reducing (3.11) mod 3 shows $s_1 \equiv 1 \pmod{3}$. Hence, $s_1 = 1$. Then

$$4 = r^2 - 3^{6a_1+1} \cdot 5^{6b_1+2} \cdot 11^{6c_1+4} \cdot 19^{6d_1+5}. \quad (3.12)$$

Write (3.12) as

$$4 = Y^2 - 3 \cdot 5^2 \cdot 11 \cdot 19^2 \cdot X^3, \quad (3.13)$$

where $Y = r$ and $X = 3^{2a_1} \cdot 5^{2b_1+1} \cdot 11^{2c_1+1} \cdot 19^{2d_1+1}$.

Magma [3] shows (3.13) only has integer solutions $(X, Y) = (0, \pm 2)$. Hence, (3.12) has no solutions.

Case 3: $A = 55$. Then $(\epsilon_1, \epsilon_2, \epsilon_3, \epsilon_4) = (0, 1, 5, 4), (4, 1, 3, 4), (4, 5, 3, 4), (4, 5, 5, 2), (4, 5, 5, 4)$. Equation (3.6) reduces to

$$4 \cdot 3^{3a_1+u_1} \cdot 5^{3b_1+u_2} \cdot 11^{3c_1+u_3} \cdot 19^{3d_1+u_4} = s(3r^2 - 55s^2). \quad (3.14)$$

Since $\left(\frac{3 \cdot 55}{19}\right) = -1$, we have $19 \nmid 3r^2 - 55s^2$.

Case 3.1: $(\epsilon_1, \epsilon_2, \epsilon_3, \epsilon_4) = (0, 1, 5, 4)$. Equation (3.14) reduces to

$$4 \cdot 5^{3b_1} \cdot 11^{3c_1+2} \cdot 19^{3d_1+2} = s(3r^2 - 55s^2). \quad (3.15)$$

Since $3b_1 = 0$ or $3b_1 \geq 3$, from (3.15) have $5^{3b_1} \mid s$. From (3.15) we also have $11^{3c_1+2} \mid s$. Therefore $s = 5^{3b_1} \cdot 11^{3c_1+2} \cdot 19^{3d_1+2} s_1$, where $s_1 \in \{\pm 1\}$. Equation (3.15) reduces to

$$4 = \pm 3r^2 - 55s^2,$$

impossible mod 5 since $5 \nmid r$ and $\left(\frac{\pm 3}{5}\right) = -1$.

Case 3.2: $3a_1 + u_1 > 0$.

Case 3.2.1: $(\epsilon_1, \epsilon_2, \epsilon_3, \epsilon_4) = (4, 1, 3, 4)$. Then (3.14) reduces to

$$4 \cdot 3^{3a_1+2} \cdot 5^{3b_1} \cdot 11^{3c_1+1} \cdot 19^{3d_1+2} = s(3r^2 - 55s^2). \quad (3.16)$$

Reducing (3.16) mod 3 gives $3 \mid s$. Hence, $3^{3a_1+1} \mid s$. If $5 \mid r$, then $5 \nmid s$. Hence, $5^2 \nmid s(3r^2 - 55s^2)$. Therefore, (3.16) is impossible mod 5^{3b_1} . Hence, $5 \nmid r$. Thus $5^{3b_1} \mid s$.

• $11 \nmid s$. Then $11 \mid r$. Hence, $11^2 \nmid s(3r^2 - 55s^2)$. From (3.16) we have $3c_1 + 1 = 1$. Let $s = 3^{3a_1+1} \cdot 5^{3b_1} \cdot 19^{3d_1+u_4} s_1$, where $s_1 \in \mathbb{Z}$ and $r = 11r_1$, where $r_1 \in \mathbb{Z}$. Then (3.16) reduces to

$$4 = s_2(11r_1^2 - 3^{6a_1+2} \cdot 5^{6b_1+1} \cdot 19^{6d_1+4} \cdot s_1^2). \quad (3.17)$$

Reducing (3.17) mod 3 shows that $s_1 \equiv -1 \pmod{3}$. Hence, $s_1 = -1$. Then (3.17) reduces to

$$4 = 3^{6a_1+2} \cdot 5^{6b_1+1} \cdot 19^{6d_1+4} - 11r_1^2,$$

impossible mod 19 since $\left(\frac{-11}{19}\right) = -1$.

• $11 \mid s$. Then $11^{3c_1+1} \mid s$. Let $s = 3^{3a_1+2} \cdot 5^{3b_1} \cdot 11^{3c_1+1} \cdot 19^{3d_1+2} \cdot s_1$, where $s_1 \in \{\pm 1\}$. Then (3.16) reduces to

$$4 = s_1(r^2 - 3^{6a_1+3} \cdot 5^{6b_1+1} \cdot 11^{6c_1+1} \cdot 19^{6d_1+4} \cdot s_1^2). \quad (3.18)$$

Reducing (3.18) mod 3 gives $s_1 \equiv 1 \pmod{3}$. Hence, $s_1 = 1$. Then (3.18) reduces to

$$4 = r^2 - 3^{6a_1+3} \cdot 5^{6b_1+1} \cdot 11^{6c_1+1} \cdot 19^{6d_1+4}. \quad (3.19)$$

Reducing mod 13 shows

$$4 \equiv r^2 - 1 \pmod{13}$$

impossible since $\left(\frac{5}{13}\right) = -1$.

Case 3.3.2: $(\epsilon_1, \epsilon_2, \epsilon_3, \epsilon_4) = (4, 5, 3, 4), (4, 5, 5, 2), (4, 5, 5, 4)$. Then (3.16) reduces to

$$4 \cdot 3^{3a_1+2} \cdot 5^{3b_1+2} \cdot 11^{3c_1+u_3} \cdot 19^{3d_1+u_4} = s(3r^2 - 55s^2). \quad (3.20)$$

Then $3^{3a_1+1} \cdot 5^{3b_1+2} \cdot 19^{3d_1+u_4} \mid s$.

• $11 \mid s$. Then $11^{3c_1+u_3} \mid s$. Hence, $s = 3^{3a_1+1} \cdot 5^{3b_1+2} \cdot 11^{3c_1+u_3} \cdot 19^{3d_1+u_4} \cdot s_1$, where $s_1 \in \mathbb{Z}$. Then (3.20) reduces to

$$4 = s_1(r^2 - 3^{6a_1+1} \cdot 5^{6b_1+4} \cdot 11^{6c_1+2u_3+1} \cdot 19^{6d_1+2u_4} \cdot s_1^2). \quad (3.21)$$

Reducing (3.21) mod 3 gives $s_1 \equiv 1 \pmod{3}$. Hence, $s_1 = 1$. Then

$$4 = r^2 - 3^{6a_1+1} \cdot 5^{6b_1+4} \cdot 11^{6c_1+\epsilon_3} \cdot 19^{6d_1+\epsilon_4}. \quad (3.22)$$

Write (3.22) as a cubic

$$Y^2 = 4 + 3 \cdot 5 \cdot 11^{v_1} \cdot 19^{v_2} \cdot X^3, \quad (3.23)$$

where $Y = r$, X only has prime divisors 5, 11, 19, and $(v_1, v_2) = (0, 1), (2, 2), (2, 1)$. Equation (3.23) only has integer solutions $(X, Y) = (0, \pm 2), (1, 17)$ as

$$\begin{aligned} 2^2 &= 4 + 3 \cdot 5 \cdot 11^{v_1} \cdot 19^{v_2} \cdot 0^3, \\ 17^2 &= 4 + 3 \cdot 5 \cdot 19 \cdot 1^2. \end{aligned}$$

None of these solutions gives solutions to (3.22).

• $11 \nmid s$. Reducing (3.20) mod 11 shows $11 \mid r$. Since $11^2 \nmid s(3r^2 - 55s^2)$, in (3.20) we must have $3c_1 + u_3 = 1$. Hence, $(\epsilon_1, \epsilon_2, \epsilon_3) = (4, 5, 3, 4)$. Then (3.16) reduces to

$$4 \cdot 3^{3a_1+2} \cdot 5^{3b_1+2} \cdot 11 \cdot 19^{3d_1+2} = s(3r^2 - 55s^2). \quad (3.24)$$

Let $s = 3^{3a_1+1} \cdot 5^{3b_1+2} \cdot 19^{3d_1+2} \cdot s_1$ and $r = 11r_1$, where $s_1, r_1 \in \mathbb{Z}$. Then (3.24) reduces to

$$4 = s_2(11r_1^2 - 3^{6a_1+1} \cdot 5^{6b_1+4} \cdot 19^{6d_1+4} \cdot s_2^2). \quad (3.25)$$

Reducing (3.25) mod 3 shows $s_2 \equiv -1 \pmod{3}$. Hence, $s_2 = -1$. Therefore

$$4 = 3^{6a_1+1} \cdot 5^{6b_1+4} \cdot 19^{6d_1+4} \cdot s_2^2 - 11r_1^2,$$

impossible mod 19 since $\left(\frac{-11}{19}\right) = -1$.

Case 4: $A = 95$. Then $(\epsilon_1, \epsilon_2, \epsilon_3, \epsilon_4) = (4, 5, 4, 5)$. Then (3.16) reduces to

$$4 \cdot 3^{3a_1+2} \cdot 5^{3b_1+2} \cdot 11^{3c_1+2} \cdot 19^{3d_1+2} = s(3r^2 - 95s^2). \quad (3.26)$$

Then $3^{3a_1+1} \mid s$, $5^{3b_1+2} \mid s$, $19^{3d_1+2} \mid s$. Since $\left(\frac{3 \cdot 93}{11}\right) = -1$, (3.26) implies $11^{3c_1+2} \mid s$. Let $s = 3^{3a_1+1} \cdot 5^{3b_1+2} \cdot 11^{3c_1+2} \cdot 19^{3d_1+2} \cdot s_1$, where $s_1 = \pm 1$. Then (3.26) reduces to

$$4 = s_1(r^2 - 3^{6a_1+1} \cdot 5^{6b_1+5} \cdot 11^{6c_1+4} \cdot 19^{6d_1+5} \cdot s_1^2).$$

Reducing mod 3 shows $s_1 \equiv 1 \pmod{3}$. Hence, $s_1 = 1$. Then

$$4 = r^2 - 3^{6a_1+1} \cdot 5^{6b_1+5} \cdot 11^{6c_1+4} \cdot 19^{6d_1+5}. \quad (3.27)$$

Write (3.27) as a cubic curve

$$Y^2 = 4 + 3 \cdot 5^2 \cdot 11 \cdot 19^2 \cdot X^3, \quad (3.28)$$

where $Y = r$ and $X = 3^{2a_1} \cdot 5^{2b_1+1} \cdot 11^{2c_1+1} \cdot 19^{2d_1+1}$. Magma shows that equation (3.28) only has integer solutions $(X, Y) = (0, \pm 2)$. Hence, (3.27) has no solutions.

Case 5: $A = 3 \cdot 11 \cdot 19$. Then $(\epsilon_1, \epsilon_2, \epsilon_3, \epsilon_4) = (1, 2, 1, 5), (1, 2, 3, 5), (1, 2, 5, 3), (1, 2, 5, 5), (1, 4, 1, 5), (1, 4, 3, 5), (1, 4, 5, 1), (1, 4, 5, 5), (3, 4, 5, 3), (5, 0, 3, 5), (5, 0, 5, 5)$. Then (3.16) reduces to

$$4 \cdot 3^{3a_1+u_1-1} \cdot 5^{3b_1+u_2} \cdot 11^{3c_1+u_3} \cdot 19^{3d_1+u_4} = s(r^2 - 209s^2). \quad (3.29)$$

Since r and s is odd, we have $8 \mid r^2 - 209s^2$. Therefore equation (3.29) is impossible mod 8.

Case 6: $A = 3 \cdot 5 \cdot 11 \cdot 19$. Then $(\epsilon_1, \epsilon_2, \epsilon_3, \epsilon_4) = (1, 1, 1, 5), (1, 3, 3, 5), (1, 3, 5, 3), (1, 3, 5, 5), (1, 5, 3, 3), (1, 5, 5, 3), (1, 5, 5, 5), (3, 1, 5, 3), (3, 1, 5, 5), (3, 3, 1, 5), (3, 3, 5, 3), (3, 5, 3, 5)$. Hence, (3.16) reduces to

$$4 \cdot 3^{3a_1+u_1-1} \cdot 5^{3b_1+u_2} \cdot 11^{3c_1+u_3} \cdot 19^{3d_1+u_4} = s(r^2 - 5 \cdot 11 \cdot 19 \cdot s^2). \quad (3.30)$$

Notice that s can only have prime factors 3, 5, 11, 19. Dividing both sides of (3.30) by s^3 gives a quartic equation of the form

$$Y^2 = 5 \cdot 11 \cdot 19 + 4 \cdot 3^{\gamma_1} \cdot 5^{u_2} \cdot 11^{u_3} \cdot 19^{u_4} \cdot X^3, \quad (3.31)$$

where $Y = \frac{r}{s}$, X can only have prime factors 3, 5, 11, 19, and $\gamma_1 = u_1$ if $u_1 \geq 1$, $\gamma_1 = 2$ if $u_1 = 0$. We use Magma to search for S -integral points on (3.31), where $S = \{3, 5, 11, 19\}$. The result is given in Table 4, where UD means Magma is not able to find S -integral points.

Table 3. Solutions to (3.31).

$(\epsilon_1, \epsilon_2, \epsilon_3, \epsilon_4)$	$(\gamma_1, u_2, u_3, u_4)$	(X, Y)
(1, 1, 1, 5)	(2, 0, 0, 2)	\emptyset
(1, 3, 3, 5)	(2, 1, 1, 2)	\emptyset
(1, 3, 5, 3)	(2, 1, 2, 1)	\emptyset
(1, 3, 5, 5)	(2, 1, 2, 2)	UD
(1, 5, 3, 3)	(2, 2, 1, 1)	\emptyset
(1, 5, 5, 3)	(2, 2, 2, 1)	UD
(1, 5, 5, 5)	(2, 2, 2, 2)	UD
(3, 1, 5, 3)	(0, 0, 2, 1)	\emptyset
(3, 1, 5, 5)	(0, 0, 2, 2)	\emptyset
(3, 3, 1, 5)	(0, 1, 0, 2)	\emptyset
(3, 3, 5, 3)	(0, 1, 2, 1)	\emptyset
(3, 5, 3, 5)	(0, 2, 1, 2)	UD

Case 6.1: $(\epsilon_1, \epsilon_2, \epsilon_3, \epsilon_4) = (3, 5, 3, 5)$. Equation (3.16) reduces to

$$4 \cdot 3^{3a_1} \cdot 5^{3b_1+2} \cdot 11^{3c_1+1} \cdot 19^{3d_1+2} = s(r^2 - 5 \cdot 11 \cdot 19 \cdot s^2). \tag{3.32}$$

Case 6.1.1: $11 \mid r$. Then $3c_1 + 1 = 1$. Let $r = 11r_1$ and $s = 5^{3b_1+2} \cdot 19^{3d_1+2} \cdot s_1$, where $r_1, s_1 \in \mathbb{Z}$. Then (3.32) reduces to

$$4 \cdot 3^{3a_1} = s_1(11r_1^2 - 5^{6b_1+5} \cdot 19^{6d_1+5} \cdot s_1^2). \tag{3.33}$$

• $s_1 = 1$. Then (3.33) reduces to

$$4 \cdot 3^{3a_1} = 11r_1^2 - 5^{6b_1+5} \cdot 19^{6d_1+5}.$$

Since $(\frac{3}{19}) = -1$ and $(\frac{11}{19}) = 1$, we have $2 \mid 3a_1$. Let $a_1 = 2a_2$, where $a_2 \in \mathbb{N}$. Then

$$4 \cdot 3^{6a_2} = 11r_1^2 - 5^{6b_1+5} \cdot 19^{6d_1+5}.$$

Reducing mod 7 gives

$$4 \equiv 4r_1^2 - 2 \pmod{7},$$

impossible mod 7 since $(\frac{6}{7}) = -1$.

• $s_1 = -1$. Then (3.33) reduces to

$$4 \cdot 3^{3a_1} = 11r_1^2 - 5^{6b_1+5} \cdot 19^{6d_1+5}.$$

Since $(\frac{3}{19}) = (\frac{-11}{19}) = -1$, we have $3 \nmid a_1$. Hence, $a_1 = 2a_2 + 1$, where $a_2 \in \mathbb{N}$. Then

$$4 \cdot 3^{6a_2+3} = 11r_1^2 - 5^{6b_1+5} \cdot 19^{6d_1+5}.$$

Reducing mod 7 gives

$$3 \equiv 4r_1^2 - 2 \pmod{7},$$

impossible mod 7 since $\left(\frac{5}{7}\right) = -1$.

Case 6.1.2: $11 \mid s$. Then $11^{3c_1+1} \mid s$. Let $s = 5^{3b_1+2} \cdot 11^{3c_1+1} \cdot 19^{3d_1+2} \cdot s_1$, where $s_1 \in \mathbb{Z}$. Then (3.33) reduces to

$$4 \cdot 3^{3a_1} = s_1(r^2 - 5^{6b_1+5} \cdot 11^{6c_1+3} \cdot 19^{6d_1+5} \cdot s_1^2). \quad (3.34)$$

• $3 \nmid s_1$. Since $11 \nmid r$ and $\left(\frac{3}{11}\right) = 1$, from (3.34) we have

$$\left(\frac{s_1}{11}\right) = \left(\frac{s_1 r^2}{11}\right) = \left(\frac{4 \cdot 3^{3a_1}}{11}\right) = 1.$$

Since $s_1 \in \{-1, 1\}$, we have $s_1 = 1$. Then (3.34) reduces to

$$4 \cdot 3^{3a_1} = r^2 - 5^{6b_1+5} \cdot 11^{6c_1+3} \cdot 19^{6d_1+5}.$$

Hence, $\left(\frac{4 \cdot 3^{3a_1}}{19}\right) = 1$. Since $\left(\frac{-3}{19}\right) = -1$, we have $2 \mid a_1$. Let $a_1 = 2a_2$, where $a_2 \in \mathbb{N}$. Then

$$4 \cdot 3^{6a_2} = r^2 - 5^{6b_1+5} \cdot 11^{6c_1+3} \cdot 19^{6d_1+5}.$$

Reducing mod 7 gives

$$4 \equiv r^2 - 2 \pmod{7},$$

impossible since $\left(\frac{6}{7}\right) = -1$.

Case 6.2: $(\epsilon_1, \epsilon_2, \epsilon_3, \epsilon_4) = (1, 3, 5, 5)$. Then (3.33) reduces to

$$4 \cdot 3^{3a_1-1} \cdot 5^{3b_1+1} \cdot 11^{3d_1+2} \cdot 19^{3d_1+2} = s(r^2 - 5 \cdot 11 \cdot 19 \cdot s^2). \quad (3.35)$$

Case 6.2.1: $5 \mid r$. Since $5^2 \nmid s(r^2 - 5 \cdot 11 \cdot 19 \cdot s^2)$, we have $3b_1 + 1 = 1$. Let $r = 5r_1$ and $s = 11^{3c_1+2} \cdot 19^{3d_1+2} \cdot s_1$, where $r_1, s_1 \in \mathbb{Z}$. Then (3.35) reduces to

$$4 \cdot 3^{3a_1-1} = s_1(5r_1^2 - 11^{6c_1+5} \cdot 19^{6d_1+5} \cdot s_1^2). \quad (3.36)$$

Notice that $\left(\frac{3}{11}\right) = \left(\frac{5}{11}\right) = 1$. Hence, (3.36) gives $\left(\frac{s_1}{11}\right) = 1$.

• $3 \nmid s_1$. Since $\left(\frac{-1}{11}\right) = -1$, we have $s_1 = 1$. Then (3.36) reduces to

$$4 \cdot 3^{3a_1-1} = 5r_1^2 - 11^{6c_1+5} \cdot 19^{6d_1+5}.$$

Hence,

$$\left(\frac{4 \cdot 3^{3a_1-1}}{19}\right) = \left(\frac{5r_1^2}{19}\right) = 1.$$

Since $\left(\frac{3}{19}\right) = -1$, we have $2 \mid 3a_1 - 1$. Hence, $2 \nmid a_1$. Let $a_1 = 2a_2 + 1$, where $a_2 \in \mathbb{N}$. Then

$$4 \cdot 3^{6a_2+2} = 5r_1^2 - 11^{6c_1+5} \cdot 19^{6d_1+5}.$$

Reducing mod 5 gives $4(-1)^{3a_2+1} \equiv 1 \pmod{5}$. Hence, $2 \mid a_2$. Let $a_2 = 2a_3$, where $a_3 \in \mathbb{N}$. Then

$$4 \cdot 3^{12a_3+2} = 5r_1^2 - 11^{6c_1+5} \cdot 19^{6d_1+5}. \quad (3.37)$$

Let $c_1 = 2c_2 + i_1$ and $d_1 = 2d_2 + i_2$ where $i_1, i_2 \in \{0, 1\}$. From (3.37) we have

$$Y^2 = X(X^2 + 5^3 \cdot 6^4 \cdot 11^{5+6i_1} \cdot 19^{5+6i_2}), \quad (3.38)$$

where $X = \frac{20 \cdot 3^{6a_1+2}}{11^{6c_2} \cdot 19^{6d_2}}$, $Y = \frac{100 \cdot 3^{3a_1+2} \cdot r_1}{11^{12c_2} \cdot 19^{12d_2}}$. Magma [3] shows that the only $\{11, 19\}$ -integral point on (3.38) is $(0, 0)$. Hence, (3.37) has no solutions.

• $3 \mid s_1$. Since $\left(\frac{s_1}{11}\right) = 1$, we have $s_1 = 3^{3a_1-1}$, then (3.36) reduces to

$$4 = 5r_1^2 - 3^{6a_1-2} \cdot 11^{6c_1+5} \cdot 19^{6d_1+5},$$

impossible mod 3 since $\left(\frac{5}{3}\right) = -1$.

Case 6.2.2: $5 \mid s$. Then $s = 5^{3b_1+1} \cdot 11^{3c_1+2} \cdot 19^{3d_1+2} \cdot s_1$, where $s_1 \in \mathbb{Z}$. Then (3.35) reduces to

$$4 \cdot 3^{3a_1-1} = s_1(r^2 - 5^{6b_1+3} \cdot 11^{6c_1+5} \cdot 19^{6d_1+5} \cdot s_1^2). \quad (3.39)$$

• $3 \nmid s_1$. If $s_1 = 1$, then (3.39) reduces to

$$4 \cdot 3^{3a_1-1} = r^2 - 5^{6b_1+3} \cdot 11^{6c_1+5} \cdot 19^{6d_1+5}. \quad (3.40)$$

Hence, $\left(\frac{4 \cdot 3^{3a_1-1}}{19}\right) = \left(\frac{r^2}{19}\right) = 1$. Since $\left(\frac{3}{19}\right) = -1$, we have $2 \mid 3a_1 - 1$. Let $a_1 = 2a_2 + 1$, where $a_2 \in \mathbb{N}$. Then (3.40) reduces to

$$4 \cdot 3^{6a_2+2} = r^2 - 5^{6b_1+3} \cdot 11^{6c_1+5} \cdot 19^{6d_1+5}.$$

Reducing mod 13 gives

$$10 \equiv r^2 - 8 \pmod{13},$$

impossible since $\left(\frac{18}{13}\right) = -1$.

If $s_1 = -1$, then (3.39) reduces to

$$4 \cdot 3^{3a_1-1} = 5^{6b_1+3} \cdot 11^{6c_1+5} \cdot 19^{6d_1+5} - r^2. \quad (3.41)$$

Hence, $\left(\frac{4 \cdot 3^{3a_1-1}}{19}\right) = \left(\frac{-1}{19}\right) = -1$. Since $\left(\frac{3}{19}\right) = -1$, we have $2 \nmid 3a_1 - 1$. Let $a_1 = 2a_2$, where $a_2 \in \mathbb{N}$. Then (3.41) reduces to

$$4 \cdot 3^{6a_2-1} = 5^{6b_1+3} \cdot 11^{6c_1+5} \cdot 19^{6d_1+5} - r^2,$$

impossible mod 5 since $\left(\frac{-3}{5}\right) = -1$.

• $3 \mid s_1$. Then $s_1 = 3^{3a_1-1} \cdot s_2$, where $s_2 \in \mathbb{Z}$. Hence, (3.39) reduces to

$$4 = s_2(r^2 - 3^{6a_1-2} \cdot 5^{6b_1+3} \cdot 11^{6c_1+5} \cdot 19^{6d_1+5} \cdot s_2^2).$$

Hence, $s_2 r^2 \equiv 4 \pmod{19}$. Therefore $\left(\frac{s_2}{19}\right) = 1$. Thus, $s_2 = 1$. Then

$$4 = r^2 - 3^{6a_1-2} \cdot 5^{6b_1+3} \cdot 11^{6c_1+5} \cdot 19^{6d_1+5}.$$

Reducing mod 13 gives

$$4 \equiv r^2 - 11 \pmod{13},$$

impossible since $\left(\frac{15}{13}\right) = -1$.

Case 6.3: $(\epsilon_1, \epsilon_2, \epsilon_3, \epsilon_4) = (1, 5, 5, 3)$. Then (3.33) reduces to

$$4 \cdot 3^{3a_1-1} \cdot 5^{3b_1+2} \cdot 11^{3c_1+2} \cdot 19^{3d_1+1} = s(r^2 - 5 \cdot 11 \cdot 19 \cdot s^2). \quad (3.42)$$

Case 6.3.1: $19 \mid r$. Then $19 \nmid s$. Thus, $19^2 \nmid s(r^2 - 5 \cdot 11 \cdot 19 \cdot s^2)$. Thus, in (3.42), we must have $d_1 = 0$. So (3.42) reduces to

$$4 \cdot 3^{3a_1-1} = s_1(19r_1^2 - 5^{6b_1+5} \cdot 11^{6c_1+5}). \quad (3.43)$$

Since $\left(\frac{3}{11}\right) = 1$ and $\left(\frac{19}{11}\right) = -1$, we have from (3.43) that $\left(\frac{s_1}{11}\right) = -1$.

• $3 \nmid s_1$. Then $s_1 \in \{\pm 1\}$. Since $\left(\frac{s_1}{11}\right) = -1$, we have $s_1 = -1$. Therefore (3.43) reduces to

$$4 \cdot 3^{3a_1-1} = 5^{6b_1+5} \cdot 11^{6c_1+5} - 19 \cdot r_1^2. \quad (3.44)$$

Thus, $\left(\frac{4 \cdot 3^{3a_1-1}}{19}\right) = \left(\frac{5 \cdot 11}{19}\right) = 1$. Since $\left(\frac{3}{19}\right) = -1$, we have $2 \mid 3a_1 - 1$. Thus, $a_1 = 2a_2 + 1$, where $a_2 \in \mathbb{N}$. Then (3.44) reduces to

$$4 \cdot 3^{6a_1+2} = 5^{6b_1+5} \cdot 11^{6c_1+5} - 19 \cdot r_1^2.$$

Reducing mod 13 gives

$$10 \equiv 9 - 6 \cdot r_1^2 \pmod{13},$$

impossible since $\left(\frac{-6}{13}\right) = -1$

• $3 \mid s_1$. Then $s_1 \in \{\pm 3^{3a_1+1}\}$. Since $\left(\frac{s_1}{11}\right) = -1$, we have $s_1 = -3^{3a_1+1}$. Therefore (3.42) reduces to

$$4 = 3^{6a_1-2} \cdot 5^{6b_1+5} \cdot 11^{6c_1+5} - 19 \cdot r_1^2,$$

impossible mod 3 since $\left(\frac{-19}{3}\right) = -1$.

Case 6.3.2: $19 \mid s$. Then $s = 5^{3b_1+2} \cdot 11^{3b_1+2} \cdot 19^{3d_1+1} \cdot s_1$, where $s_1 \in \mathbb{Z}$. Then (3.42) reduces to

$$4 \cdot 3^{3a_1-1} = s_1(r^2 - 5^{6b_1+5} \cdot 11^{6c_1+5} \cdot 19^{6d_1+5} \cdot s_1^2). \quad (3.45)$$

Since $\left(\frac{3}{11}\right) = 1$, we have $\left(\frac{s_1}{11}\right) = 1$.

• $3 \nmid s_1$. Then $s_1 = 1$. Hence, (3.45) reduces to

$$4 \cdot 3^{3a_1-1} = r^2 - 5^{6b_1+5} \cdot 11^{6c_1+5} \cdot 19^{6d_1+5}.$$

Since $\left(\frac{3}{19}\right) = -1$, we have $2 \mid 3a_1 - 1$. Let $a_1 = 2a_2 + 1$, where $a_2 \in \mathbb{N}$. Then

$$4 \cdot 3^{6a_2+2} = r^2 - 5^{6b_1+5} \cdot 11^{6c_1+5} \cdot 19^{6d_1+5}.$$

Reducing mod 7 gives

$$1 \equiv r^2 - 4 \pmod{7},$$

impossible mod 7 since $\left(\frac{5}{7}\right) = -1$.

• $3 \mid s_1$. Then $s_1 = 3^{3a_1-1}$. Hence, (3.45) reduces to

$$4 = r^2 - 3^{6a_1-2} \cdot 5^{6b_1+5} \cdot 11^{6c_1+5} \cdot 19^{6d_1+5}.$$

Reducing mod 7 gives

$$4 \equiv r^2 - 2 \pmod{7},$$

impossible since $\left(\frac{6}{7}\right) = -1$

Case 6.4: $(\epsilon_1, \epsilon_2, \epsilon_3, \epsilon_4) = (1, 5, 5, 5)$. Then (3.33) reduces to

$$4 \cdot 3^{3a_1-1} \cdot 5^{3b_1+2} \cdot 11^{3c_1+2} \cdot 19^{3d_1+2} = s(r^2 - 5 \cdot 11 \cdot 19 \cdot s^2). \tag{3.46}$$

Thus, $s = 5^{3b_1+2} \cdot 11^{3c_1+2} \cdot 19^{3d_1+2} \cdot s_1$, where $s_1 \in \mathbb{Z}$. Therefore

$$4 \cdot 3^{3a_1-1} = s_1(r^2 - 5^{6b_1+5} \cdot 11^{6c_1+5} \cdot 19^{6d_1+5} \cdot s_1^2). \tag{3.47}$$

Since $\left(\frac{3}{11}\right) = 1$, we have $\left(\frac{s_1}{11}\right) = 1$. Notice that $\left(\frac{-1}{11}\right) = -1$.

• $3 \nmid s_1$. Then $s_1 = 1$. Hence, (3.47) reduces to

$$4 \cdot 3^{3a_1-1} = r^2 - 5^{6b_1+5} \cdot 11^{6c_1+5} \cdot 19^{6d_1+5}.$$

Reducing mod 7 gives

$$(-1)^{1+a_1} \equiv r^2 - 4 \pmod{7},$$

impossible since $\left(\frac{4 \pm 1}{7}\right) = -1$.

• $3 \mid s_1$. Then $s_1 = 3^{3a_1-1}$. Hence, (3.47) reduces to

$$4 = r^2 - 3^{6a_1-2} \cdot 5^{6b_1+5} \cdot 11^{6c_1+5} \cdot 19^{6d_1+5}.$$

Reducing mod 7 gives

$$4 \equiv r^2 - 2 \pmod{7},$$

impossible since $\left(\frac{6}{7}\right) = -1$. □

Lemma 3.2. All solutions (n, a, b, c, d, x, y) with $3 \mid n$ and $n > 3$ to (3.1) are given in Table 4.

Table 4. Solutions to (3.1) with $3 \mid n$, $n > 3$, and $\gcd(x, y) = 1$.

(n, a, b, c, d, x, y)
$(n, 1, 0, 0, 0, 1, 1)$
$(6, 3, 1, 0, 0, 11, 2)$
$(6, 3, 1, 2, 0, 7, 4)$
$(12, 3, 1, 2, 0, 7, 2)$
$(6, 5, 1, 0, 2, 781, 8)$
$(9, 5, 1, 0, 2, 781, 4)$
$(18, 5, 1, 0, 2, 781, 2)$

Proof. Let $n = 3k$, where $k \in \mathbb{Z}^+$ and $k > 1$. Let $y_1 = y^k$. Then (3.1) reduces to

$$x^2 + 3^a \cdot 5^b \cdot 11^c \cdot 19^d = 4y_1^3. \tag{3.48}$$

We apply Lemma 3.1 to equation (3.48). Notice that solutions in Table 2 are deduced from solutions in Table 1. For example, solution

$$(n, a, b, c, d, x, y) = (3, 3, 1, 0, 0, 11, 4)$$

from Table 1 gives us a solution

$$(n, a, b, c, d, x, y) = (6, 3, 1, 0, 0, 11, 2)$$

in Table 2. □

Lemma 3.3. All solutions (n, a, b, c, d, x, y) to (3.1) with $n = 4$ are list in Table 5.

Table 5. Solutions to (3.1) with $n = 4$.

(n, a, b, c, d, x, y)	(n, a, b, c, d, x, y)	(n, a, b, c, d, x, y)	(n, a, b, c, d, x, y)
(4, 4, 0, 0, 1, 31, 5)	(4, 4, 8, 3, 0, 141407, 353)	(4, 0, 1, 1, 0, 3, 2)	(4, 0, 5, 1, 0, 1557, 28)
(4, 4, 1, 2, 1, 947, 26)	(4, 0, 1, 2, 1, 1147, 24)	(4, 0, 1, 3, 2, 237, 28)	(4, 0, 1, 1, 0, 7, 3)
(4, 8, 3, 4, 1, 270973, 524)	(4, 0, 3, 0, 1, 53, 6)	(4, 0, 3, 1, 2, 1923, 32)	(4, 1, 0, 0, 0, 1, 1)
(4, 5, 0, 4, 0, 7199, 61)	(4, 1, 4, 1, 1, 195937, 313)	(4, 1, 0, 2, 2, 65521, 181)	(4, 1, 1, 0, 0, 7, 2)
(4, 1, 2, 2, 0, 23, 7)	(4, 2, 0, 1, 0, 49, 5)	(4, 2, 4, 2, 1, 10033, 73)	(4, 2, 1, 0, 1, 13, 4)
(4, 2, 1, 1, 0, 23, 4)	(4, 6, 1, 1, 0, 337, 14)	(4, 2, 2, 0, 1, 73, 7)	(4, 2, 2, 0, 1, 233, 11)
(4, 2, 3, 1, 2, 937, 34)	(4, 3, 1, 2, 0, 7, 8)	(4, 3, 2, 0, 0, 337, 13)	

Proof. Let $a = 4a_1 + i_1, b = 4b_1 + i_2, c = 4c_1 + i_3, d = 4d_1 + i_4$, where $a_1, b_1, c_1, d_1 \in \mathbb{N}$ and $0 \leq i_1, i_2, i_3, i_4 \leq 3$. From (3.1) we have

$$Y^2 = 4X^4 - 3^{i_1} \cdot 5^{i_2} \cdot 11^{i_3} \cdot 19^{i_4}, \tag{3.49}$$

where $X = \frac{y}{3^{a_1} \cdot 5^{b_1} \cdot 11^{c_1} \cdot 19^{d_1}}, Y = \frac{x}{3^{2a_1} \cdot 5^{2b_1} \cdot 11^{2c_1} \cdot 19^{2d_1}}, a_1, b_1, c_1, d_1 \in \mathbb{N}, 0 \leq i_1, i_2, i_3, i_4 \leq 3$, and $2 \nmid i_1 + i_3 + i_4$. Magma [3] is able to find S -integral points on (3.49) for all but the case $(i_1, i_2, i_3, i_4) = (3, 3, 3, 3)$, where $S = \{3, 5, 11, 19\}$. We list all cases of (i_1, i_2, i_3, i_4) where (3.49) has solutions in Table 6, the case $(i_1, i_2, i_3, i_4) = (3, 3, 3, 3)$ is undetermined (or UD).

Table 6. Solutions to (3.49).

(i_1, i_2, i_3, i_4)	(X, Y)	(n, a, b, c, x, y)
(0, 0, 0, 1)	$(\pm 5/3, \pm 31/9)$	(4, 4, 0, 0, 1, 31, 5)
(0, 0, 3, 0)	$(\pm 353/75, \pm 141407/5625)$	(4, 4, 8, 3, 0, 141407, 353)
(0, 1, 1, 0)	$(\pm 2, \pm 3)$	(4, 0, 1, 1, 0, 3, 2)
(0, 1, 1, 0)	$(\pm 28/5, \pm 1557/25)$	(4, 0, 5, 1, 0, 1557, 28)
(0, 1, 2, 1)	$(\pm 22/3, \pm 77/9)$	\emptyset
(0, 1, 2, 1)	$(\pm 26/3, \pm 947/9)$	(4, 4, 1, 2, 1, 947, 26)

(0, 1, 2, 1)	(±24, ±1147)	(4, 0, 1, 2, 1, 1147, 24)
(0, 1, 3, 2)	(±28, ±237)	(4, 0, 1, 3, 2, 237, 28)
(0, 2, 0, 1)	(±5, ±45)	∅
(0, 2, 1, 0)	(±3, ±7)	(4, 0, 2, 1, 0, 7, 3)
(0, 2, 2, 1)	(±11, ±33)	∅
(0, 3, 0, 1)	(±524/99, ±270973/9801)	(4, 8, 3, 4, 1, 270973, 524)
(0, 3, 0, 1)	(±6, ±53)	(4, 0, 3, 0, 1, 53, 6)
(0, 3, 1, 2)	(±32, ±1923)	(4, 0, 3, 1, 2, 1923, 32)
(1, 0, 0, 0)	(±1, ±1)	(4, 1, 0, 0, 0, 1, 1)
(1, 0, 0, 0)	(±61/33, ±7199/1089)	(4, 5, 0, 4, 0, 7199, 61)
(1, 0, 1, 1)	(±313/5, ±195937/25)	(4, 1, 4, 1, 1, 195937, 313)
(1, 0, 2, 2)	(±181, ±65521)	(4, 1, 0, 2, 2, 65521, 181)
(1, 1, 0, 0)	(±2, ±7)	(4, 1, 1, 0, 0, 7, 2)
(1, 1, 0, 2)	(±76/3, ±11533/9)	∅
(1, 1, 1, 1)	(±28, ±1567)	∅
(1, 2, 0, 2)	(±19, ±703)	∅
(1, 2, 2, 0)	(±7, ±23)	(4, 1, 2, 2, 0, 23, 7)
(2, 0, 1, 0)	(±3, ±15)	∅
(2, 0, 1, 0)	(±5, ±49)	(4, 2, 0, 1, 0, 49, 5)
(2, 0, 2, 1)	(±73/5, ±10033/25)	(4, 2, 4, 2, 1, 10033, 73)
(2, 1, 0, 1)	(±4, ±13)	(4, 2, 1, 0, 1, 13, 4)
(2, 1, 1, 0)	(±4, ±23)	(4, 2, 1, 1, 0, 23, 4)
(2, 1, 0, 1)	(±14/3, ±337/9)	(4, 6, 1, 1, 0, 337, 14)
(2, 1, 2, 1)	(±22, ±913)	∅
(2, 2, 0, 1)	(±7, ±73)	(4, 2, 2, 0, 1, 73, 7)
(2, 2, 0, 1)	(±11, ±233)	(4, 2, 2, 0, 1, 233, 11)
(2, 2, 1, 0)	(±5, ±5)	∅
(2, 2, 3, 2)	(±575/3, ±654595/9)	∅
(2, 2, 3, 2)	(±775, ±1201205)	∅
(2, 3, 1, 2)	(±34, ±937)	(4, 2, 3, 1, 2, 937, 34)
(3, 1, 2, 0)	(±8, ±7)	(4, 3, 1, 2, 0, 7, 8)
(3, 2, 0, 0)	(±13, ±337)	(4, 3, 2, 0, 0, 337, 13)
(3, 3, 3, 3)	UD	UD

We consider the case $(i_1, i_2, i_3, i_4) = (3, 3, 3, 3)$. Then (3.1) reduces to

$$(2y^2 - x)(2y^2 + x) = 3^{4a_1+3} \cdot 5^{4b_1+3} \cdot 11^{4c_1+3} \cdot 19^{4d_1+3}.$$

Hence,

$$4y^2 = A_1 + B_1, \tag{3.50}$$

where $A_1, B_1 \in \mathbb{Z}^+$ ad $A_1 B_1 = 3^{4a_1+3} \cdot 5^{4b_1+3} \cdot 11^{4c_1+3} \cdot 19^{4d_1+3}$. Without loss of generality, we can assume that $3 \mid A_1$.

Case 1: $3 \mid A_1$ and $19 \mid B_1$. Then $3^{4a_1+3} \mid A_1$. Since $(\frac{5}{19}) = (\frac{11}{19}) = 1$ and $(\frac{3}{19}) = -1$, we have $(\frac{A_1}{19}) = -1$. Hence, equation (3.50) is impossible mod 19.

Case 2: $3 \cdot 19 \mid A_1$ and $5 \mid B_1$. Since $3^{4a_1+3} \cdot 19^{4b_1+3} \mid A_1$, $\left(\frac{11}{5}\right) = \left(\frac{19}{5}\right) = 1$ and $\left(\frac{3}{5}\right) = -1$, we have $\left(\frac{A_1}{5}\right) = -1$, impossible since we deduce from (3.50) that

$$\left(\frac{A_1}{5}\right) = \left(\frac{4y^2}{5}\right) = 1.$$

Case 3: $3 \cdot 5 \cdot 19 \mid A_1$ and $11 \mid B_1$. Then $A_1 = 3^{3a_1+3} \cdot 5^{3b_1+3} \cdot 19^{3d_1+3}$ and $B_1 = 11^{3b_1+3}$. Equation (3.50) becomes

$$4y^2 = 3^{4a_1+3} \cdot 5^{4b_1+3} \cdot 19^{4d_1+3} + 11^{4c_1+3},$$

impossible mod 11 since $\left(\frac{19}{11}\right) = -1$ and $\left(\frac{3}{11}\right) = \left(\frac{5}{11}\right) = 1$.

Case 4: $3 \cdot 5 \cdot 11 \cdot 19 \mid A_1$. Then $A_1 = 3^{4a_1+3} \cdot 5^{4b_1+3} \cdot 11^{4c_1+3} \cdot 19^{4d_1+3}$ and $B_1 = 1$. Equation (3.50) reduces to

$$4y^2 = 1 + 3^{4a_1+3} \cdot 5^{4b_1+3} \cdot 11^{4c_1+3} \cdot 19^{4d_1+3}.$$

Then $(2y, 3^{2a_1+1} \cdot 5^{2b_1+1} \cdot 11^{2c_1+1} \cdot 19^{2d_1+1})$ is a solution to the Pell equation

$$X^2 - 3 \cdot 5 \cdot 11 \cdot 19 \cdot Y^2 = 1. \tag{3.51}$$

The fundamental solution to (3.51) is $(X, Y) = (56, 1)$. We look for $k \in \mathbb{Z}^+$ such that

$$3^{2a_1+1} \cdot 5^{2b_1+1} \cdot 11^{2c_1+1} \cdot 19^{2d_1+1} = Y_k = \frac{\lambda_1^k - \lambda_2^k}{\lambda_1 - \lambda_2}, \tag{3.52}$$

where $\lambda_1 = 56 + \sqrt{3135}$ and $\lambda_2 = 56 - \sqrt{3135}$.

If $k > 30$, then from the work of Bilu, Hanrot, and Voutier [2] we know that Y_k has a primitive divisor q such that $k \mid q - \left(\frac{(\lambda_1 - \lambda_2)^2}{q}\right)$, impossible since $q \in \{3, 5, 11, 19\}$ and $k > 30$.

Therefore $k \leq 30$. Checking the values of k in the range $1 \leq k \leq 30$ shows that (3.52) is impossible for all $1 \leq k \leq 30$.

We conclude that all solutions to (3.1) with $n = 4$ is given in Table 5. □

Lemma 3.4. *Solutions (n, a, b, c, d, x, y) to (3.1) with $4 \mid n$ and $n > 4$ are listed in Table 7.*

Table 7. Solutions to (3.1) with $n4 \mid n$, $n > 4$, and $\gcd(x, y) = 1$.

(n, a, b, c, d, x, y)
$(n, 10, 0, 0, 1, 1)$
$(8, 2, 1, 0, 1, 13, 2)$
$(8, 2, 1, 1, 0, 23, 2)$
$(12, 3, 1, 2, 0, 7, 2)$

Proof. Let $n = 4k$, where $k \in \mathbb{Z}^+$ and $k > 1$. Let $y_1 = y^k$. Then

$$x^2 + 3^a \cdot 5^b \cdot 11^c \cdot 19^d = 4y_1^4. \tag{3.53}$$

We use Table 5 in Lemma 3.3 to find solutions to (3.53) and get Table 7. □

Lemma 3.5. All solutions to (3.1) with $n \geq 5$, $3 \nmid n$, $4 \nmid n$, $(a, b, c, d) \neq (1, 1, 1, 1) \pmod{2}$, and $\gcd(x, y) = 1$ are given in Table 8.

Table 8. Solutions to (3.1) with $n \geq 5$, $3 \nmid n$, $4 \nmid n$, and $\gcd(x, y) = 1$.

(n, a, b, c, d, x, y)
$(n, 1, 0, 0, 0, 1, 1)$
$(5, 2, 0, 5, 2, 38599, 55)$
$(5, 0, 4, 5, 2, 41261, 99)$
$(5, 0, 3, 5, 0, 25289, 44)$

Proof. We can assume that n is an odd prime ≥ 5 . Then

$$\frac{x + B\sqrt{-A}}{2} = \left(\frac{r + s\sqrt{-A}}{2} \right)^n,$$

where r, s are odd coprime integers. Therefore

$$\frac{B}{s} = \frac{\alpha^n - \beta^n}{\alpha - \beta} = L_n,$$

where $\alpha = \frac{r+s\sqrt{-A}}{2}$ and $\beta = \frac{r-s\sqrt{-A}}{2}$. If $\frac{\alpha}{\beta}$ is a root of unity, then $\frac{\alpha}{\beta} = \zeta_m$, a primitive m -root of unity. Since $|\mathbb{Q}(\zeta_m) : \mathbb{Q}| = \phi(m)$ and $\mathbb{Q}(\frac{\alpha}{\beta}) = 2$, we have $\phi(m) = 2$. Therefore $m \in \{3, 4\}$. Hence, $\zeta_m \in \left\{ \pm\sqrt{-1}, \frac{\pm 1 \pm \sqrt{-3}}{2} \right\}$. Therefore $A = 3$ and $\alpha = \pm \frac{1 \pm \sqrt{-3}}{2}$. Hence, $y = 1$. We deduce that $(n, a, b, c, d, x, y) = (n, 1, 0, 0, 0, 1, 1)$.

We consider the case $\frac{\alpha}{\beta}$ is not a root of unity. If L_n has a primitive divisor q , then $n \mid q - \left(\frac{\alpha - \beta}{q} \right)^2$. Since $q \in \{3, 5, 11, 19\}$ and $n \geq 5$, we deduce that $n = 5$, $q = 19$, and $\left(\frac{\alpha - \beta}{q} \right)^2 = -1$. Since $(\alpha - \beta)^2 = -As^2$, we have $\left(\frac{-A}{19} \right) = -1$. Since $19 \nmid A$ and $A \in \{3, 11, 15, 19, 55, 95, 627\}$, we have $A \in \{11, 55\}$. Let $B = 3^i \cdot 5^j \cdot 11^k \cdot 19^l$, where $i, j, k, l \in \mathbb{N}$. Since

$$B = \frac{\alpha^5 - \beta^5}{\sqrt{-A}},$$

we have

$$16B = s(5r^4 - 10Ar^2s^2 + A^2s^4). \tag{3.54}$$

Notice that $s \mid B$, so s only has prime divisors $3, 5, 11, 19$. Dividing both sides of (3.54) by s^5 gives a quartic curve

$$\gamma Y^2 = 5X^4 - 10AX^2 + A^2, \tag{3.55}$$

where $\gamma = \pm 3^{i_1} \cdot 5^{i_2} \cdot 11^{i_3} \cdot 19^{i_4}$, $i_1, i_2, i_3, i_4 \in \{0, 1\}$, $X = \frac{r}{s}$, $Y \in \mathbb{Q}$ and Y only has prime divisors $3, 5, 11, 19$. We use Magma to search for S -integral points on (3.55), where $S = \{3, 5, 11, 19\}$. We list the value of γ where (3.55) has S -integral points and the corresponding tuples (n, a, b, c, d, x, y) in Table 9. \square

Table 9. Solutions to (3.55).

A	(3.55)	(X, Y)	(r, s)	(n, a, b, c, d, x, y)
11	$-19Y^2 = 5X^4 - 110X^2 + 121$	$(\pm 11/3, \pm 44/9)$	$(\pm 11, \pm 3)$	$(5, 2, 0, 5, 2, 38599, 55)$
11	$-95Y^2 = 5X^4 - 110X^2 + 121$	$(\pm 78/25, \pm 1399/625)$ $(\pm 11/5, \pm 44/25)$ $(\pm 8/5, \pm 29/25)$	$(\pm 78, \pm 25)$ $(\pm 11, \pm 5)$ $(\pm 8, \pm 5)$	\emptyset $(5, 0, 4, 5, 2, 41261, 99)$ \emptyset
55	$Y^2 = X^4 - 110X^2 + 605$	$(\pm 11, \pm 40)$	$(\pm 11, \pm 1)$	$(5, 0, 3, 5, 0, 25289, 44)$

Remark 3.6. We need the condition $5 \nmid h(\mathbb{Q}(\sqrt{-A}))$ in the proof of Lemma 3.5. Since the class number of $\mathbb{Q}(\sqrt{-3 \cdot 5 \cdot 11 \cdot 19})$ is 40, the condition $(a, b, c, d) \not\equiv (1, 1, 1, 1) \pmod{2}$ in Lemma 3.5 and Theorem 1.1 is indispensable.

When $(a, b, c, d) \equiv (1, 1, 1, 1) \pmod{2}$, then $a = 10a_1 + i_1$, $b = 10b_1 + i_2$, $c = 10c_1 + i_3$, and $d = 10d_1 + i_4$, where $a_1, b_1, c_1, d_1 \in \mathbb{N}$ and $i_1, i_2, i_3, i_4 \in \{1, 3, 5, 7, 9\}$. Then (3.1) reduces to

$$Y^2 = 4X^5 + 3^{i_1} \cdot 5^{i_2} \cdot 11^{i_3} \cdot 19^{i_4}, \tag{3.56}$$

where

$$Y = \frac{x}{3^{5i_1} \cdot 5^{5i_2} \cdot 11^{5i_3} \cdot 19^{5i_4}} \quad \text{and} \quad X = \frac{y}{3^{2i_1} \cdot 5^{2i_2} \cdot 11^{2i_3} \cdot 19^{2i_4}}.$$

Equation (3.56) represents a curve of genus 2, and we need to find $\{3, 5, 11, 19\}$ -integral points on this curve. It might be possible to attack (3.56) using the method in [5] but the author of this paper has not been able to proceed in this way.

Acknowledgements. The author is supported by Vietnam Institute of Advanced Study in Mathematics. The author would like to thank the anonymous referee for many value comments and suggestions.

References

- [1] S. BHATTER, A. HOQUE, R. SHARMA: *On the solutions of a Lebesgue–Nagell type equation*, Acta Mathematica Hungarica 158.1 (2019), pp. 17–26, DOI: <https://doi.org/10.1007/s10474-018-00901-6>.
- [2] Y. BILU, G. HANROT, P. M. VOUTIER: *Existence of primitive divisors of Lucas and Lehmer numbers. With an appendix by M. Mignotte*, Journal für die reine und angewandte Mathematik 539.3 (2001), pp. 75–122, DOI: <https://doi.org/10.1515/crll.2001.080>.
- [3] W. BOSMA, J. CANNON, C. PLAYOUST: *The Magma algebra system. I. The user language*, Journal of Symbolic Computation 24.3-4 (1997), pp. 235–265, DOI: <https://doi.org/10.1006/jsco.1996.0125>.
- [4] K. CHAKRABORTY, A. HOQUE, R. SHARMA: *Complete solutions of certain Lebesgue–Ramanujan–Nagell type equations*, Publicationes Mathematicae Debrecen 97.3-4 (2020), URL: http://publi.math.unideb.hu/load_jpg.php?p=2405.
- [5] H. R. GALLEGOS-RUIZ: *S-integral points on hyperelliptic curves*, International Journal of Number Theory 7.3 (2011), pp. 803–824, DOI: <https://doi.org/10.1142/S1793042111004435>.

- [6] M. LE, G. SOYDAN: *A brief survey on the generalized Lebesgue–Ramanujan–Nagell equation*, Surveys in Mathematics and its Applications 15 (2020), pp. 473–523,
URL: https://www.utgjiu.ro/math/sma/v15/p15_20.pdf.
- [7] F. LUCA, S. TENGELY, A. TOGBÉ: *On the Diophantine equation $x^2 + C = 4y^n$* , Annales mathématiques du Québec 33.2 (2009), pp. 171–184,
URL: <http://www.labmath.uqam.ca/~Annales/volumes/33-2/PDF/171-184.pdf>.

What positive integers n can be presented in the form $n = (x + y + z)(1/x + 1/y + 1/z)$?

Nguyen Xuan Tho

School of Applied Mathematics and Informatics,
Hanoi University of Science and Technology
`tho.nguyenxuan1@hust.edu.vn`

Submitted: November 9, 2020

Accepted: April 19, 2021

Published online: April 27, 2021

Abstract

This paper shows that the equation in the title does not have positive integer solutions when n is divisible by 4. This gives a partial answer to a question by Melvyn Knight. The proof is a mixture of elementary p -adic analysis and elliptic curve theory.

Keywords: Elliptic curves, p -adic numbers

AMS Subject Classification: 11G05, 11D88

1. Introduction

According to Bremner, Guy, and Nowakowski [1], Melvyn Knight asked what integers n can be represented in the form

$$n = (x + y + z) \left(\frac{1}{x} + \frac{1}{y} + \frac{1}{z} \right), \quad (1.1)$$

where x, y, z are integers. In the same paper [1], the authors made an extension study of (1.1) in integers when n is in the range $|n| \leq 1000$. Integer solutions are found except for 99 values of n . The question becomes more interesting if we ask for positive integer solutions, which was also briefly discussed in [1, Section 2]. In this paper, we will prove the following theorem:

Theorem 1.1. *Let n be a positive integer. Then equation*

$$(x + y + z) \left(\frac{1}{x} + \frac{1}{y} + \frac{1}{z} \right) = n$$

does not have positive integer solutions if $4 \mid n$.

This theorem gives the first parametric family when (1.1) does not have positive integer solutions. The proof technique is a nice combination of p -adic analysis and elliptic curve theory, which was successfully applied to prove the insolubility of the equation

$$(x + y + z + w) \left(\frac{1}{x} + \frac{1}{y} + \frac{1}{z} + \frac{1}{w} \right) = n$$

for the families $n = 4m^2$, $4m^2 + 4$, $m \in \mathbb{Z}$ and $m \not\equiv 2 \pmod{4}$, see [2].

2. The Hilbert symbol

Let p be a prime number, and let $a, b \in \mathbb{Q}_p$. The Hilbert symbol $(a, b)_p$ is defined as

$$(a, b)_p = \begin{cases} 1, & \text{if the equation } aX^2 + bY^2 = Z^2 \text{ has a solution} \\ & (X, Y, Z) \neq (0, 0, 0) \text{ in } \mathbb{Q}_p^3, \\ -1, & \text{otherwise.} \end{cases}$$

The symbol $(a, b)_\infty$ is defined similarly but \mathbb{Q}_p is replaced by \mathbb{R} . The following properties of the Hilbert symbol are true, see Serre [3, Chapter III]:

(i) For all a, b , and c in \mathbb{Q}_p^* , then

$$\begin{aligned} (a, bc)_p &= (a, b)_p (a, c)_p, \\ (a, b^2)_p &= 1. \end{aligned}$$

(ii) For all a and b in \mathbb{Q}_p^* , then

$$(a, b)_\infty \prod_{p \text{ prime}} (a, b)_p = 1.$$

(iii) Let p be a prime number, and let a and b in \mathbb{Q}_p^* . Write $a = p^\alpha u$ and $b = p^\beta v$, where $\alpha = v_p(a)$ and $\beta = v_p(b)$. Then

$$\begin{aligned} (a, b)_p &= (-1)^{\frac{\alpha\beta(p-1)}{2}} \left(\frac{u}{p} \right)^\beta \left(\frac{v}{p} \right)^\alpha & \text{if } p \neq 2, \\ (a, b)_p &= (-1)^{\frac{(u-1)(v-1)}{4} + \frac{\alpha(v^2-1)}{8} + \frac{\beta(u^2-1)}{8}} & \text{if } p = 2, \end{aligned}$$

where $\left(\frac{u}{p} \right)$ denotes the Legendre symbol.

3. A main theorem

Theorem 3.1. *Let n be a positive integer divisible by 4. Let u and v be nonzero rational numbers such that*

$$v^2 = u(u^2 + (n^2 - 6n - 3)u + 16n).$$

Then

$$u > 0.$$

Let $A = n^2 - 6n - 3$, $B = 16n$ and $D = n - 1$. Then

$$A^2 - 4B = (n - 9)(n - 1)^2.$$

Now

$$v^2 = u(u^2 + Au + B). \tag{3.1}$$

The proof of Theorem 3.1 is achieved by means of the following three lemmas.

Lemma 3.2. *If p is an odd prime number, then*

$$(u, -D)_p = 1.$$

Proof. Let $r = v_p(u)$. Then $u = p^r s$, where $s \in \mathbb{Z}_p$, and $p \nmid s$.

Case 1: $r < 0$. Then from (3.1), we have

$$v^2 = p^{3r} s(s^2 + p^{-r}A + Bp^{-2r}).$$

Therefore $2v_p(v) = 3r$, hence $2 \mid r$. Now

$$(p^{-3r/2}v)^2 = s(s^2 + p^{-r}A + Bp^{-2r}). \tag{3.2}$$

Note that $p \nmid s$. Taking (3.2) modulo p gives s is a square modulo p . Hence $s \in \mathbb{Z}_p^2$. We also have $2 \mid r$, so $u = 2^r s \in \mathbb{Q}_p^2$. Therefore $(u, -D)_p = 1$.

Case 2: $r = 0$.

If $p \nmid D$, then both u and $-D$ are units in \mathbb{Z}_p . Therefore $(u, -D)_p = 1$.

If $p \mid D$, then $n \equiv 1 \pmod{p}$. Hence $A = n^2 - 6n - 3 \equiv -8 \pmod{p}$ and $B = 16n \equiv 16 \pmod{p}$. Thus

$$\begin{aligned} v^2 &\equiv u(u^2 - 8u + 16) \pmod{p} \\ &\equiv u(u - 4)^2 \pmod{p}. \end{aligned} \tag{3.3}$$

If $u \equiv 4 \pmod{p}$, then $u \in \mathbb{Z}_p^2$. Hence $(u, -D)_p = 1$. If $u \not\equiv 4 \pmod{p}$, then from (3.3), we have

$$u \equiv \left(\frac{v}{u - 4} \right)^2 \pmod{p}.$$

Therefore $u \in \mathbb{Z}_p^2$. Hence $(u, -D)_p = 1$.

Case 3: $r > 0$. Then (3.1) becomes

$$v^2 = p^r s(p^{2r} s^2 + Ap^r s + B). \quad (3.4)$$

If $p \mid B$, then $p \mid n$. Therefore $-D = 1 - n \equiv 1 \pmod{p}$. Hence $-D \in \mathbb{Z}_p^2$. Thus $(u, -D)_p = 1$.

If $p \nmid B$, then from (3.4) we have $r = 2v_p(v)$. Thus $2 \mid r$.

If $p \nmid D$, then both s and $-D$ are units in \mathbb{Z}_p . Therefore $(s, -D)_p = 1$. Hence

$$(u, -D)_p = (p^r s, -D)_p = (s, -D)_p = 1.$$

If $p \mid D$, then $n \equiv 1 \pmod{p}$. Therefore $A = n^2 - 6n - 3 \equiv -8 \pmod{p}$ and $B = 16n \equiv 16 \pmod{p}$. Let $\omega = p^{\frac{-r}{2}} v$. Because $r = 2v_p(v)$, we have $p \nmid \omega$. From (3.4) we have

$$\omega^2 \equiv s(p^{2r} s^2 - 8p^r s + 16) \equiv 16s \pmod{p},$$

so that

$$s \equiv (\omega/4)^2 \pmod{p}.$$

Thus $s \in \mathbb{Z}_p^2$. Hence $(s, -D)_p = 1$. Note that $2 \mid r$, therefore

$$(u, -D)_p = (p^r s, -D)_p = (s, -D)_p = 1. \quad \square$$

Lemma 3.3. *We have*

$$(u, -D)_2 = 1.$$

Proof. Let $n = 4k$, where $k \in \mathbb{Z}^+$.

If $2 \mid k$, then $-D = 1 - 4k \equiv 1 \pmod{8}$. Therefore $-D \in \mathbb{Z}_2^2$. Hence $(u, -D)_2 = 1$. So we only need to consider the case $2 \nmid k$. Let $r = v_2(u)$. Then $u = 2^r s$, where $2 \nmid s$.

Case 1: $2 \mid r$. Then

$$\begin{aligned} (u, -D)_2 &= (2^r s, 1 - 4k)_2 \\ &= (s, 1 - 4k)_2 \\ &= (-1)^{\frac{(s-1)(1-4k-1)}{4}} \\ &= 1. \end{aligned}$$

Case 2: $2 \nmid r$. We show that this case is not possible.

If $r < 0$, then from (3.1), we have

$$v^2 = 2^{3r} s(s^2 + 2^{-r} A s + 2^{-2r} B).$$

Therefore $3r = 2v_2(v)$. Hence $2 \mid r$, a contradiction.

If $r \geq 0$, then (3.1) becomes

$$v^2 = 2^r s(2^{2r} s^2 + 2^r(16k^2 - 24k - 3)s + 2^6 k). \quad (3.5)$$

If $r \geq 7$, then

$$v^2 = 2^{r+6}s(2^{2r-6}s^2 + (16k^2 - 24k - 3)2^{r-6}s + k).$$

Therefore $r + 6 = 2v_2(v)$. Hence $2 \mid r$, a contradiction.

If $r < 7$, then $r \leq 5$. Let $\phi = \frac{v}{2^r}$. Then from (3.5), we have

$$\phi^2 = s(2^r s^2 + (16k^2 - 24k - 3)s + 2^{6-r}k). \tag{3.6}$$

If $r = 5$, then taking (3.6) modulo 8 gives $\phi^2 \equiv s(-3s + 2k) \pmod{8}$. Hence $2sk \equiv \phi^2 + 3s^2 \equiv 4 \pmod{8}$, which is not possible because $2 \nmid sk$.

If $r = 3$, then taking (3.6) modulo 8 gives $\phi^2 \equiv -3s^2 \pmod{8}$. Hence $0 \equiv \phi^2 + 3s^2 \equiv 4 \pmod{8}$, a contradiction.

If $r = 1$, then taking (3.6) modulo 8 gives $\phi^2 \equiv s(2 - 3s) \pmod{8}$. So $2s \equiv 3s^2 + \phi^2 \equiv 4 \pmod{8}$, which is not possible because $2 \nmid s$. \square

Lemma 3.4.

$$(u, -D)_\infty = 1.$$

Proof. From the product formula for the Hilbert symbol, we have

$$(u, -D)_\infty \prod_{p \text{ prime}, p < \infty} (u, -D)_p = 1. \tag{3.7}$$

By Lemma 3.2, Lemma 3.3, and (3.7), we have $(u, -D)_\infty = 1$. \square

To complete the proof of Theorem 3.1, we see that Lemma 3.4 shows that the equation $uX^2 + (1 - n)Y^2 = Z^2$ has a solution $(X, Y, Z) \neq (0, 0, 0)$ in \mathbb{R}^3 . Because $1 - n < 0$, we have $u > 0$. Hence Theorem 3.1 is proved.

4. A proof of Theorem 1.1

We follow [1, Section 2]. Write (1.1) as

$$x^2(y + z) + x(y^2 + (3 - n)yz + z^2) + yz = 0. \tag{4.1}$$

Hence

$$x = \frac{-y^2 + (n - 3)yz - z^2 \pm \Delta}{2(y + z)},$$

where Δ satisfies

$$\Delta^2 = y^4 - 2(n - 1)yz(y^2 + z^2) + (n^2 - 6n - 3)y^2z^2 + z^4. \tag{4.2}$$

Then (4.2) is birationally equivalent to the elliptic curve

$$v^2 = u(u^2 + (n^2 - 6n - 3)u + 16n), \tag{4.3}$$

and we can write out the maps between (4.1) and (4.3):

$$u = \frac{-4(xy + yz + zx)}{z^2}, \quad v = \frac{2(u - 4n)y}{z} - (n - 1)u,$$

and

$$\frac{x, y}{z} = \frac{\pm v - (n - 1)u}{2(4n - u)}.$$

Then the following is true.

Proposition 4.1. *The necessary and sufficient conditions for (4.1) to have positive integer solutions (x, y, z) are $n > 0$ and $u < 0$.*

Proof. See Bremner, Guy, and Nowakowski [1, Section 2]. □

Now, let $n = 4k$, where $k \in \mathbb{Z}^+$. Assume there exists a positive integer solution (x, y, z) to (1.1). Then Proposition 4.1 shows that $u < 0$. If $v = 0$, then (4.3) implies $u^2 + (n^2 - 6n - 3)u + 16n = 0$. Therefore $(n - 9)(n - 1)^3 = (n^2 - 6n - 3)^2 - 4 \times 16n$ is a perfect square. Hence $(n - 9)(n - 1)$ is a perfect square. Let $(n - 9)(n - 1) = m^2$. Then $(n - 5)^2 - 16 = m^2$. The equation $X^2 - 16 = Y^2$ only has integer solutions $(X, Y) = (\pm 5, \pm 3)$. Thus $n - 5 = \pm 5$, giving no solutions $n = 4k$. Therefore $v \neq 0$. Hence $u, v \neq 0$. From Theorem 3.1, we have $u > 0$, contradicting Proposition 4.1. Hence (1.1) does not have solutions in positive integers.

Acknowledgement. The author would like to thank the referee for his careful reading and valuable comments.

References

- [1] A. BREMNER, R. K. GUY, R. J. NOWAKOWSKI: *Which integers are representable as the product of the sum of three integers with the sum of their reciprocals?*, Math. Compt. 61 (1993), pp. 117–130, DOI: <https://doi.org/10.1090/S0025-5718-1993-1189516-5>.
- [2] A. BREMNER, N. X. THO: *The equation $(w + x + y + z)(1/w + 1/x + 1/y + 1/z) = n$* , Int. J. Number Theory 14.5 (2018), pp. 1–18, DOI: <https://doi.org/10.1142/S1793042118500756>.
- [3] J.-P. SERRE: *Local Fields*, New York: Springer, 1973.

Script-aided generation of Mental Cutting Test exercises using Blender

Róbert Tóth

University of Debrecen, Faculty of Informatics

`toth.robert@inf.unideb.hu`

Submitted: November 15, 2020

Accepted: March 12, 2021

Published online: March 29, 2021

Abstract

This paper presents a possible generation process how to efficiently model, export and render resources of *Mental Cutting Test* exercises with the use of Blender.

Keywords: spatial skills, Mental Cutting Test, Blender

1. Introduction

The spatial availability of people can be measured by using different types of tests such as *Mental Cutting Test* and *Mental Rotation Test*. Offering exercises of *Mental Cutting Test* (MCT) is a widely used method in which people have to determine the shape of the intersection of a 3D shape and a plane. The test was introduced in the USA in the late 1930s [4], while researches still deal with this topic nowadays [12]. *Virtual Reality (VR)* has become a widely supported and popular method to extend our reality. During the last decade, researchers started developing VR aided applications to extend the functionality of classic MCT exercises [7]. On the other hand, the use of VR requires head gears which results in strong limitations on the researches. In the meantime, *Augmented Reality (AR)* has also become a popular way to extend our reality and interact with the users in a more efficient way without any investments – we only need to use the camera of our mobile devices to place objects onto any surfaces. While this paper focuses on the technical description

of a recently designed, efficient procedure to generate MCT exercises, we have to mention that many Hungarian and East European researchers deal with various approaches of developing spatial skills [1, 10, 11, 14, 15]. Thus, the proposed procedure has the opportunity to enhance the methods of a popular and actively researched topic both in national and international scientific context.

In 2019 the development of a new mobile application has started at the Faculty of Informatics, University of Debrecen with the aim of improving the spatial skills of its users by offering MCT exercises [16]. The exercises are derived from the classic MCT, but Augmented Reality and the principle of gamification [8] are also being used to support, engage and motivate the users.

1.1. Terminology

Another important key of a practicing application is the number of different exercises. To efficiently construct the dataset, three abstraction levels have been introduced (see Figure 1).

1. **Scenarios:** A scenario consists of the basic resources that are needed to represent an exercise: the isometric projection of a 3D shape that is intersected by a 2D plane (so called *2D model*), and the shape of their intersection (so called *answer* or *shape of intersection*). To support the users with the AR function, a new component has been introduced (so called *3D model*) which is a 3D object both containing the 3D shape and a 3D frame that describes the intersection plane (so called *intersection*).
2. **Exercise:** Classic MCT exercises contain four additional shapes as wrong answers. Thus, we must construct alternative answers to form a real exercise. In our terminology, an *exercise* is a composition of a scenario and answers of another scenarios. Thus, multiple exercises can be constructed of a scenario by permutating the wrong answers. The method of choosing the wrong answers can depend on several factors: to form a practice exercise for beginner users, we should minimize the similarity between the alternatives; on the other hand, MCT exercises that are used to measure skills usually contain very similar alternatives.
3. **Assignment:** An assignment is an instance of an exercise that is assigned to a user. Each assignment is generated from the exercises that are defined by the instructor – thus, the users' choices and spatial skills can be easily compared by processing the assignments.

Files that are rendered or exported from Blender [5] and used to display or post-process assignments (2D models, 3D models and intersections) are called *resources*.

1.2. Resources

As we described in Section 1.1, to form a scenario three resources are required:

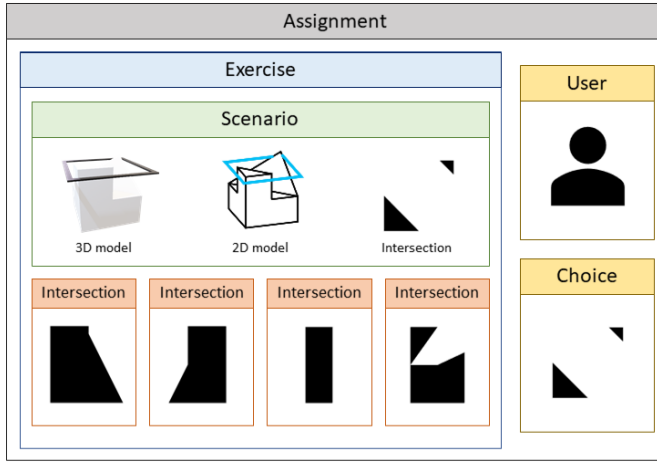


Figure 1. Entities and resources that form an assignment.

1. A 2D image that contains the 3D shape and the 2D plane applying the special isometric projection of MCT exercises.
2. A 3D model that contains the 3D shape and the 2D plane and can be instanced and displayed with the AR function.
3. A 2D image that contains the shape of the intersection of the 3D shape and the intersection plane. This resource is the solution of the classic MCT exercises.

Several file formats can be selected for these resources. Two requirements have been constructed towards the file formats:

1. Image format should be scalable.
2. The size of the resources must be minimized .

To fulfil the requirements, SVG [3] was chosen for image resources, while GLB [9]) was chosen for 3D models. On the other hand, a rastergraphic input format is also needed for post-processing methods. For that purpose, PNG [2] files are used. GLB models can be exported and PNG images can be rendered in Blender without any add-ons, while *FreeStyle* and its *FreeStyle SVG Exporter* are used to render SVG files.

1.3. Scenarios

As the focus is on the generation of assignments, further sections will describe this process. The aim is to give a well detailed description of the methods that are used

to prepare, render and postprocess all the resources that are needed to construct a scenario. During this research, Blender is used to design our models, then generate the required output files in SVG and GLB formats as well as additional PNG images are also being generated to support the image processing that is used to determine similarity between the shapes of intersection. Blender provides various methods to design and manipulate models while it also offers a well-detailed API in Python language [6]. Thus, a library has been designed and implemented that supports the instructor through the steps of the scenario development. As a result, multiple scenarios can be generated starting from a single 3D shape by applying rotation vectors and predefined intersection planes. However, the steps still need some human contribution and configuration that can be given in JSON markup language.

1.4. The process

Based on the API of Blender an iterative workflow of development was designed to generate the required resources from the modeller software. The algorithm contains the following steps:

1. Prepare scenarios.
 - (a) Design 3D shapes.
 - (b) Design intersection planes.
 - (c) Rotate shapes.
2. Render and export resources.
 - (a) Render 2D models.
 - (b) Export 3D models.
 - (c) Create and render the shapes and intersections.
3. Post-process scenarios.
 - (a) Filter scenarios.
 - (b) Substitute answers.
 - (c) Rotate and scale answers.
 - (d) Edit, compress and style SVG resources.

2. Preparation

2.1. Design 3D shapes

MCT exercises consist of various models, but two basic approaches can be used to design them:

1. Start with a basic shape, then transform its vertices, edges and faces (mostly with subdivision and extrusion).
2. Start with multiple basic shapes, then construct compositions of them by calculating their unions, intersections or differences. This method can be supported by Blender's *boolean modifier*. However, the limitations of this operation will be described in Section 3.3.

We have to keep in mind that only the 2D projection of the model is displayed in classic MCT exercises to the users and they can only examine three faces of the shape without limitations. Thus, most of the details should be concentrated into these sides – otherwise people will not be able to determine the shape of all the intersections.

All shapes that are scaled to have a maximum size of 2 in each dimensions. Thus, the size of shapes are uniformed and they fill the viewport of the camera that is being used to render 2D models.

2.2. Design intersection planes

Due to the previously mentioned limitations of the applied projection, most of the MCT exercises only use a few intersection planes. However, our goal is to offer various scenarios. Thus, one pillar of the permutation is the various usage of intersection planes. Table 1 shows the 19 planes can be used in the generation process, mostly resulting a non-empty intersection. Each intersection plane is described with vectors C , N and R ($C, N, R \in \mathcal{R}^3$), where C contains the coordinates of a point that is on the plane, N is the normal of the intersection plane, while R is a rotation vector and will be used in the further transformations.




















Consider a cube with dimensions (2, 2, 2) which geometry origin is in location (0, 0, 0). Intersections 0-9 are parallel with one of the faces of the shape, intersections 10-15 contain the diagonals of the faces, while intersections 16-19 contain the diagonals of the shape.

2.3. Rotate shapes

Scenarios can be permuted easily by applying rotation on the shapes around their origins while intersection planes are still the same. This method has two advantages:






1. Firstly, multiple scenarios containing the same shape and intersection plane can be constructed whose intersection is still the same. As a result, users can practice on the same scenarios using different perspectives.
2. Secondly, the shape of the intersection also depends on the features of the 3D shape. Thus, different types of intersections can be generated by just simply rotating the object and still using the same intersection planes.

Table 1. Intersection planes with their C , N and R vectors demonstrated with the cube.

ID	Example	Features	ID	Example	Features
01		$C=(0, 0, 0)$ $N=(0, 0, 1)$ $R=(0, 0, 0)$	10		$C=(0, 0, 0)$ $N=(0, 1, -1)$ $R=(-45, 0, 0)$
02		$C=(0, 0, -0.8)$ $N=(0, 0, 1)$ $R=(0, 0, 0)$	11		$C=(0, 0, 0)$ $N=(-1, 0, -1)$ $R=(0, -45, 0)$
03		$C=(0, 0, 0.8)$ $N=(0, 0, 1)$ $R=(0, 0, 0)$	12		$C=(0, 0, 0)$ $N=(0, -1, -1)$ $R=(45, 0, 0)$
04		$C=(0, 0, 0)$ $N=(0, 1, 0)$ $R=(90, 0, 0)$	13		$C=(0, 0, 0)$ $N=(1, 0, -1)$ $R=(0, 45, 0)$
05		$C=(0, 0.8, 0)$ $N=(0, 1, 0)$ $R=(90, 0, 0)$	14		$C=(0, 0, 0)$ $N=(1, 1, 0)$ $R=(90, 0, 45)$
06		$C=(0, -0.8, 0)$ $N=(0, 1, 0)$ $R=(90, 0, 0)$	15		$C=(0, 0, 0)$ $N=(1, -1, 0)$ $R=(0, 90, 45)$
07		$C=(0, 0, 0)$ $N=(1, 0, 0)$ $R=(0, 90, 0)$	16		$C=(0, 0, 0)$ $N=(-0.5, 0.5, 1)$ $R=(0, -35, -135)$
08		$C=(-0.8, 0, 0)$ $N=(1, 0, 0)$ $R=(0, 90, 0)$	17		$C=(0, 0, 0)$ $N=(-0.5, -0.5, 1)$ $R=(0, -35, -225)$
09		$C=(0.8, 0, 0)$ $N=(1, 0, 0)$ $R=(0, 90, 0)$	18		$C=(0, 0, 0)$ $N=(0.5, -0.5, 1)$ $R=(0, -35, -315)$
			19		$C=(0, 0, 0)$ $N=(0.5, 0.5, 1)$ $R=(0, -35, -45)$

The number of possible rotations – consequently the number of different permutations – depends on the features of the shape. Table 2 shows the five orientations that can be used in most of the cases.

Table 2. A shape with its intersection plane, rotated by different R vectors.

				
$(0, 0, 0)$	$(0, 0, 270)$	$(0, 0, 90)$	$(270, 0, 0)$	$(270, 0, 270)$

3. Generate resources

3.1. 2D model

The camera that is used to render 2D models is created with *orthographic projection* at *location* $(3.3, -1.25, 1.4)$, *rotation* $(68.5, 0, 72)$, *scale* $(1.0, 0.772, 1.0)$ and *orthographic scale* 4.50. These properties result the same isometric perspective as the classic MCT exercises. The algorithm of the rendering contains the following steps:

1. The shape is moved into location $(0, 0, 0)$ (see Figure 3).
2. The frame which represents the intersection plane is displayed. All 19 frames are pre-modelled and subdivided into $200 * 200$ subfaces in its local X and Y dimensions. Also, the outer edges of the frame are marked as *FreeStyle edges*.
3. The SVG resource is being rendered with FreeStyle by selecting all the *contour* and *FreeStyle edges*. The output will be modified and styled in the post-processing steps.

3.2. 3D model

Secondly, the 3D model is being exported in GLB format. Thus, a white material is applied to the shapes, while one of the 19 thick black frames is also added to represent the intersection planes. The model is exported with Blender's export function.

3.3. Intersection

Thirdly, the most complex step of the rendering process is to create and render the shape of the intersection.

In the first phase of the research, sample scenarios have been designed by using the *boolean modifier* with its *intersection* option [17]. However, several issues could be found. As the operator deals with non-zero volume meshes, the intersection planes were created with constant, non-zero heights (values were chosen from range $[0.001, 0.02]$). To render the shape of intersection, a camera was added to location

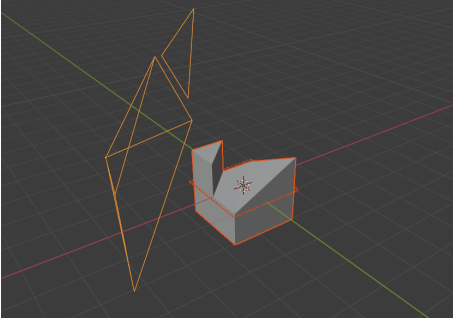


Figure 2. The shape with its frame representing the intersection plane with the camera that is used to render 2D models.

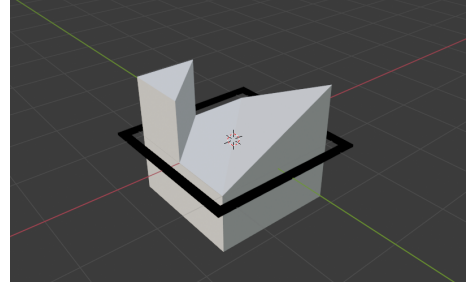


Figure 3. A thick black frame is being exported with the 3D model.

$(0, 0, 3.2)$ with rotation $(0, 0, 90)$ and the shapes were moved to location $(0, 0, 0)$. Contour edges were marked as *FreeStyle edges* and the scene was rendered with *FreeStyle*. Unfortunately, the operator regularly miscalculated the edges of the shapes by adding extra edges and vertices the shape. As the documentation points out, only manifold meshes are guaranteed to give proper results, as well as we should also avoid any co-planar faces or co-linear edges of shapes. The first criterion can be avoided in most of the scenarios, but the second criterion cannot be satisfied since most of the MCT scenarios contain co-linear edges. Thus, we could not use this modifier anymore because extra vertices and edges were added in various cases which were not detectable with a deterministic algorithm.

After that, the usage of Blender's *bisection* operator started. This operator is accessible from the user interface, while it is also supported by the Python API. The algorithm is the following:

1. Create an independent duplicate of the 3D shape, then clear its *FreeStyle edges*.
2. Bisect the shape by applying the `bisect()` function. The plane is described by coordinate C , and normal N of the selected intersection, described in Table 1.
3. The edges of the object that are in the intersection plane are automatically selected by the `bisect` function. Flip the selection and delete all the unselected vertices and edges to eliminate all the features except the intersection. Mark the border edges of the intersection as *Freestyle edges*.
4. Apply a global rotation on the shape that is described by vector R of the intersection plane to flatten it. Now the Z coordinates of all vertices and edges are 0. Calculate the geometry origin, and move the intersection to location $(0, 0, 0)$.

5. Determine the dimensions of the intersection and scale the object to fill the viewport of the camera with the maximum size of $(2, 2, 0)$.
6. Render the SVG and PNG resources that contain the shape of the intersection. The SVG file is being used as the output resource, while the PNG format supports the post-processing.

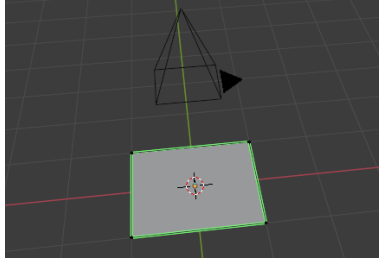


Figure 4. The shape of the answer under the camera being used to render the resource files.

4. Post-processing

The steps of the modelling process and the resource files that are needed to construct a scenario have already been presented. In this section the post-processing steps that are used to improve the quality of our resources as well as can be used to modify the configuration file of the scenario will be described.

4.1. Filter scenarios

As all the required combinations of shapes, rotations and intersection planes have been generated, at this point the generated resources must be examined. Different combinations of objects, rotations and intersection planes can result unusable scenarios. Consider the following cases:

1. The intersection plane and the 3D shape do not have an intersection. Thus, the intersections are empty.
2. An important detail of the shape is hidden from the actual perspective, thus users cannot give deterministic answer.
3. Two or more rotations result the same scenario because of the symmetry of the object.

In all cases, the scenario should be omitted from the database and be skipped in the further executions of rendering. Case (1) and (3) could be determined automatically by processing the PNG formats of intersections and 2D models. However,

case (2) should be detected by the instructor. In the current implementation, case (1) is detected automatically. Our configuration file will store all the combinations that should not be applied anymore.

4.2. Substitute answers

Of course, two or more scenarios can result the same shape of intersection. Furthermore, in many cases they only differ from each other in their rotation. Thus, all the similarities must be discovered and the same shapes must be substituted with only one, uniformed instance. Thus, a simple image processing algorithm was developed which reads and compares the PNG resources of intersections by determining the ratio of their common pixels. This algorithm can easily discover that intersections are the same except their rotation by rotating them around the Z axis with 90, 180 and 270 degrees. On the other hand, we can flip the images both vertically and horizontally to transform them into the same state. Similarities that are detected by this algorithm are added to the configuration file. Of course, there are several cases in which this solution cannot determine the similarity (e.g. one shape is rotated with 45 degrees). Thus, feature detection and pairing algorithms were applied to the resources such as OpenCV's SIFT and ORB [13]. Unfortunately, intersections are very simple shapes that do not have enough features. Thus, these algorithms cannot determine similarities with acceptable precision, the instructor must detect and configure these similarities manually.

4.3. Rotate and scale answers

As all the different shapes of intersections have been selected and the mapping has been defined between them, the selected shapes must be uniformed. The instructor can extend the configuration by describing a rotation that should be performed on the shape of intersection using axis Z. Then, all the intersections will be rotated and automatically scaled to the uniform size during further executions of the rendering script.

4.4. Edit, compress and style SVG resources

The SVG format which is being generated by the *FreeStyle SVG* plugin can be transformed into a more optimized format. The plugin generates unnecessary attributes (e. g. attributes of the *Inkscape* namespace), inline style attributes and adds unnecessary points on the edges of the shapes. Thus, the intersections are being manipulated using the Python *minidom* package to transform our vector graphics into a more compressed format without any loss of important data. Based on the features of SVG files, we can easily compress sizes by applying the following steps:

1. Dissolve all the points of paths if their neighbors are on the same line.

2. Merge or bisect `g` elements to better organize the path elements and make them able to apply CSS rules easier.
3. Replace the inline style attributes with CSS rules.
4. Remove redundant or unnecessary white spaces and commas from the document (e. g. from the attribute values of `path` elements and the indentation).

Unfortunately, several processors are not able to correctly cascade CSS rules to elements even though the recommendation supports this feature. As the Android package *Pixplicity's Sharp* being used during this research does not support cascading, we must skip step (3).

Table 3. An intersection in its original (a), transformed (b) and filled (c) state.

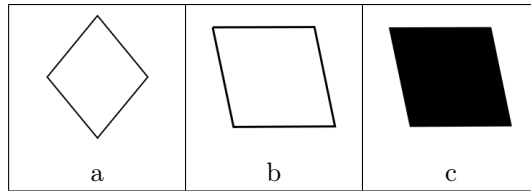
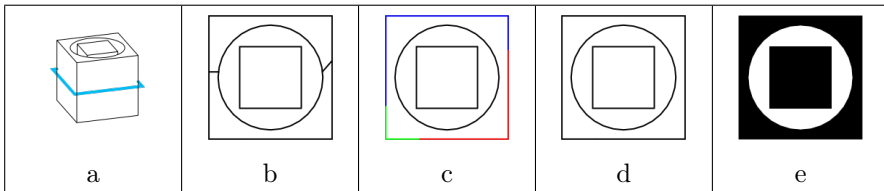


Table 4. A 2D scenario (a), its intersection (b) with rendering all of the edges, and its intersection (c) with rendering only border edges. The post-processing algorithm joins the separated paths (d) and the shapes are being filled (e).





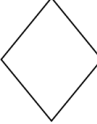





On the other hand, `path` elements should be modified in this step. Of course, some of the operations (e.g. to fill the `path` elements) could be done by the *FreeStyle* plugin in most of the cases, but an independent method is preferred to in-built solutions of Blender and several glitches can be caused by *FreeStyle*. Table 4 shows a scenario in which a hole is cut into a cube, which contains a prism. The `bisect` function subdivides the intersection, which results non-border edges. They can be skipped easily by marking only the border edges as *FreeStyle edges*. The SVG should contain three shapes: a square with a nested circle and another square. Unfortunately, *FreeStyle* constructs three paths instead of one, representing correctly the outer square.

Thus, the non-looped edges are collected by the post-processing algorithm and joined into a loop which correctly form a shape. Then, the paths of intersections are

filled with black or white color alternately, depending on whether a path forms an outer shape or contained in a black or white shape. Thus, holes in the intersections are handled by the algorithm.

Table 5. Original and post-processed SVG resources with their sizes in bytes, compressed with $\varepsilon = 0.2$.

Original		Post-processed	
Shape	Size	Shape	Size
	27 739		1 401 (5 %)
	31 490		2 876 (9 %)
	1 751		467 (26 %)
	2 212		532 (24 %)

5. Configuration file

As all the important steps of the generation algorithm have already been described, the schema of the configuration file can be introduced. The goal was to design a lightweight document that can contain all the needed metadata that should be used in any step of the generation process (see Figure 5):

1. **rotations:** The array of rotation vectors that can be applied to the 3D shape during the rendering.
2. **skip:** The array of scenario IDs (containing the ID of intersection plane and rotation). These combinations should be skipped due to any issue, e.g. empty intersection, or invisible details of the 3D shape.
3. **compress-ratio:** The value of ε that is used during the elimination of inner points of edges.

4. **similarity-ratio**: The value of ε that is used during the detection of similar answers.
5. **answers**: The list of objects that describes intersections selected as uniformed answers. Each instance is described by its scenario ID, rotation and the assigned id.
6. **substitutes**: The mapping between the original and the uniformed, replacement intersections.

```

{
  "rotations": [
    [0, 0, 0], [0, 0, 270], [0, 0, 90],
    [270, 0, 0], [270, 0, 270]
  ],
  "skip": ["17.300", "17.303"],
  "compression-ratio": 0.2,
  "similarity-ratio": 0.9,
  "answers": [
    { "scenario": "01.000", "rotation": 35, "id": "01"},
    { "scenario": "01.100", "rotation": -90, "id": "02"}
    ...
  ],
  "substitutes": {
    "01.000": "01",
    "01.003": "01",
    "01.100": "02",
    "01.300": "02"
    ...
  }
}

```

Figure 5. A sample configuration.

Configuration files can be manually edited as well as the values of properties `skip`, `answers` and `substitutions` can be automatically created or modified. As a result, configurations can be iterated during multiple executions of the process. Thus, further executions can be executed without any manual contributions. The interpretation of the configuration of Figure 5:

1. Permute scenarios with five rotations, skipping the combination of intersection plane with rotation (270, 0, 0), then intersection plane 17 with rotation (270, 0, 270).
2. Use $\varepsilon = 0.2$ in the compression of SVGs, and 0.9 to discover similar intersections.

3. Rotate intersection of scenario 01.000 with 35 degrees as uniformed answer 01, intersection of scenario 01.100 with -90 degrees as uniformed answer 02.
4. Map the answers of scenarios 01.000 and 01.003 to uniformed answer 01, and answers of scenarios 01.100 and 01.300 to uniformed answer 02.

6. Conclusion

This paper presented a script-aided process that was designed to support the design and rendering process of MCT exercises with the use of Blender and its Python API. With the combination of different intersection planes and rotations, almost 100 scenarios can be generated from a single model in most of the cases. The large number of scenarios lets us offer practicing exercises to improve the spatial skills of people. The shapes of intersections are being generated correctly in most cases; only the combination of overlapping intersection planes and faces can lead to non-deterministic results.

Acknowledgements. This work was supported by the construction EFOP-3.6.3-VEKOP-16-2017-00002. The project was supported by the European Union, co-financed by the European Social Fund.

References

- [1] L. BARANOVÁ, I. KATRENIČOVÁ: *Role of Descriptive geometry course in development of students' spatial visualization skills*, in: vol. 49, 2018, pp. 21–32, DOI: <https://doi.org/10.33039/ami.2018.04.001>.
- [2] T. BOUTELL: *PNG (Portable Network Graphics) Specification Version 1.0*, RFC 2083, Mar. 1997, DOI: <https://doi.org/10.17487/RFC2083>, URL: <https://rfc-editor.org/rfc/rfc2083.txt> (visited on 11/13/2020).
- [3] N. BROWNLEE, IAB: *SVG Drawings for RFCs: SVG 1.2 RFC*, RFC 7996, Dec. 2016, DOI: <https://doi.org/10.17487/RFC7996>, URL: <https://rfc-editor.org/rfc/rfc7996.txt> (visited on 11/13/2020).
- [4] *CEEB Special Aptitude Test in Spatial Relations (MCT)*. Developed by the College Entrance Examination Board, 1939.
- [5] B. O. COMMUNITY: *Blender - a 3D modelling and rendering package*, Blender Foundation, Stichting Blender Foundation, Amsterdam, 2018, URL: <http://www.blender.org>.
- [6] B. O. COMMUNITY: *Blender 2.90.1 Python API Documentation*, Blender Foundation, Stichting Blender Foundation, Amsterdam, 2020, URL: <https://docs.blender.org/api/current/> (visited on 11/13/2020).
- [7] T. GUZSVINECZ, M. SZELES, E. PERGE, C. SIK-LANYI: *Preparing spatial ability tests in a virtual reality application*, in: 2019 10th IEEE International Conference on Cognitive Informatics and Communications (CogInfoCom), 2019, pp. 363–368, DOI: <https://doi.org/10.1109/CogInfoCom47531.2019.9089919>.

- [8] K. M. KAPP: *The Gamification of Learning and Instruction: Game-based Methods and Strategies for Training and Education*, 1st, Pfeiffer & Company, 2012, pp. 9–13, ISBN: 978-1118096345.
- [9] KHROSNOGROUP: *glTF Specification, 2.0*, 2016,
URL: <https://github.com/KhronosGroup/glTF/blob/master/specification/2.0/README.md#binary-gltf-layout> (visited on 11/13/2020).
- [10] R. NAGY-KONDOR: *Gender differencies in spatial visualization skills of engineering students*, in: *Annales Mathematicae et Informaticae*, vol. 46, 2016, pp. 265–276.
- [11] R. NAGY-KONDOR: *Spatial Ability, Descriptive Geometry and Dynamic Geometry Systems*, in: *Annales Mathematicae et Informaticae*, vol. 37, 2010, pp. 199–210.
- [12] B. NEMETH, M. HOFFMANN: *Gender differences in spatial visualization among engineering students*, in: *Annales Mathematicae et Informaticae*, vol. 33, 2016, pp. 169–174.
- [13] E. RUBLEE, V. RABAUD, K. KONOLIGE, G. BRADSKI: *ORB: An efficient alternative to SIFT or SURF*, in: 2011 International Conference on Computer Vision, 2011, pp. 2564–2571, DOI: <https://doi.org/10.1109/ICCV.2011.6126544>.
- [14] Z. ŠIPUŠ, A. ČIŽMEŠIJA: *Spatial ability of students of mathematics education in Croatia evaluated by the Mental Cutting Test*, in: *Annales Mathematicae et Informaticae*, vol. 40, 2012, pp. 203–316.
- [15] C. SÖRÖS, B. NÉMETH, M. HOFFMANN: *Typical mistakes in Mental Cutting Test and their consequences in gender differences*, in: *Teaching Mathematics and Computer Science* 5, vol. 5, 2016, pp. 385–392.
- [16] R. TÓTH, M. ZICHAR, M. HOFFMANN: *Gamified Mental Cutting Test for enhancing spatial skills*, in: 2020 11th IEEE International Conference on Cognitive Infocommunications (CogInfoCom), 2020, pp. 000299–000304, DOI: <https://doi.org/10.1109/CogInfoCom50765.2020.9237888>.
- [17] R. TÓTH, M. ZICHAR, M. HOFFMANN: *Improving and Measuring Spatial Skills with Augmented Reality and Gamification*, in: ICGG 2020 - Proceedings of the 19th International Conference on Geometry and Graphics, Springer, 2021, chap. 68, pp. 1–10, DOI: https://doi.org/10.1007/978-3-030-63403-2_68.

Route planning on GTFS using Neo4j*

Anikó Vágner

Department of Information Technology,
Faculty of Informatics,
University of Debrecen, Hungary
`vagner.aniko@inf.unideb.hu`

Submitted: February 1, 2021

Accepted: July 15, 2021

Published online: July 23, 2021

Abstract

GTFS (General Transit Feed Specification) is a standard of Google for public transportation schedules. The specification describes stops, routes, dates, trips, etc. of one or more public transportation company for a city or a country. Examining a GTFS feed it can be considered as a graph. In addition in the last decades new database management systems was born in order to support the big data era and/or help to write program codes. Their collective name is the NoSQL databases, which covers many types of database systems. One type of them is the graph databases, from which the Neo4j is the most widespread. In this paper I try to find the answer for the question how the Neo4j can support the usage of the GTFS. The most obvious usage of the GTFS is the route planning for which the Neo4j offers some algorithms. I built some storage structures on which the tools provided by Neo4j can be effectively used to plan routes on GTFS data.

Keywords: Graph database, GTFS, route planning

AMS Subject Classification: 68-04, 68P20

1. Introduction

Nowadays the smart city concept is very fashionable. Despite the fact that it is not well defined, it can be said that smart city concept appears when some technologies

*The work is supported by the EFOP-3.6.1-16-2016-00022 project. The project is co-nanced by the European Union and the European Social Fund.

are used to develop the life quality of the citizens who inhabit in the cities which populations are increasing [4, 5]. The smart city concept addresses many domains like transport, health, homes, buildings and environment. The focus of this paper is on the public transport of a city, in the narrow sense on the route planning for public transport.

There are a lot of working solutions to support the route planning for public transport in a big town, for example Google Maps Transit¹ for many towns around the world; Traveline² for Great Britain; BKK Futár³ for Budapest, Hungary; Journey Planner of Public Transport Victoria⁴ for Victoria, Australia; Rejseplanen⁵ for Danmark and Journey Planner of Transport for London⁶ for London. Tuaycharoen [18] also introduced one in his paper.

These information can be reached by web applications or mobile applications, but on the back-end side there are comprehensive solutions to store the schedule, run the necessary algorithms and delivers the information about the planned routes. Regarding the well-known 3-tier architecture [7], we can suppose that each of these applications comprises 3 parts: presentation tier, logic tier and data tier. The presentation tier interacts with the users, gets the departure and arrival information, shows the maps, and the resulted routes. The data tier stores the schedule information and provides access to the database. And the logic tier calculates the routes itself from the database and deliver the resulted information to the presentation tier.

The database world has had a big change in the last few years, namely beside the relational database systems a lot of new database management systems have been born to answer the problems of the big data and the application development where the in-memory data structures did not fit to the relational data model [17]. The collective name of these databases is NoSQL and it comprises many types, like key-value, document, graph and column-family.

The well-known and world-wide used source of the public transport schedule is the GTFS databases [10]. Examining a GTFS feed it can be found that it is a graph. My goal is put the content of GTFS sources into a graph database. Considering the database ranking [6] the Neo4j database system is the most popular graph database.

In this paper my goal is to analyse how the Neo4j as a database management system can support the route planning systems for public transport based on GTFS sources. In the paper I didn't consider the presentation tier, only the database and the logic tier. Additionally I examine only the tools and opportunities of the Neo4j for both the storage and the algorithm.

¹<https://www.google.com/transit>

²traveline.info

³<http://futar.bkk.hu>

⁴<https://www.ptv.vic.gov.au/journey>

⁵<https://journeyplanner.dk/>

⁶<https://tfl.gov.uk/plan-a-journey/>

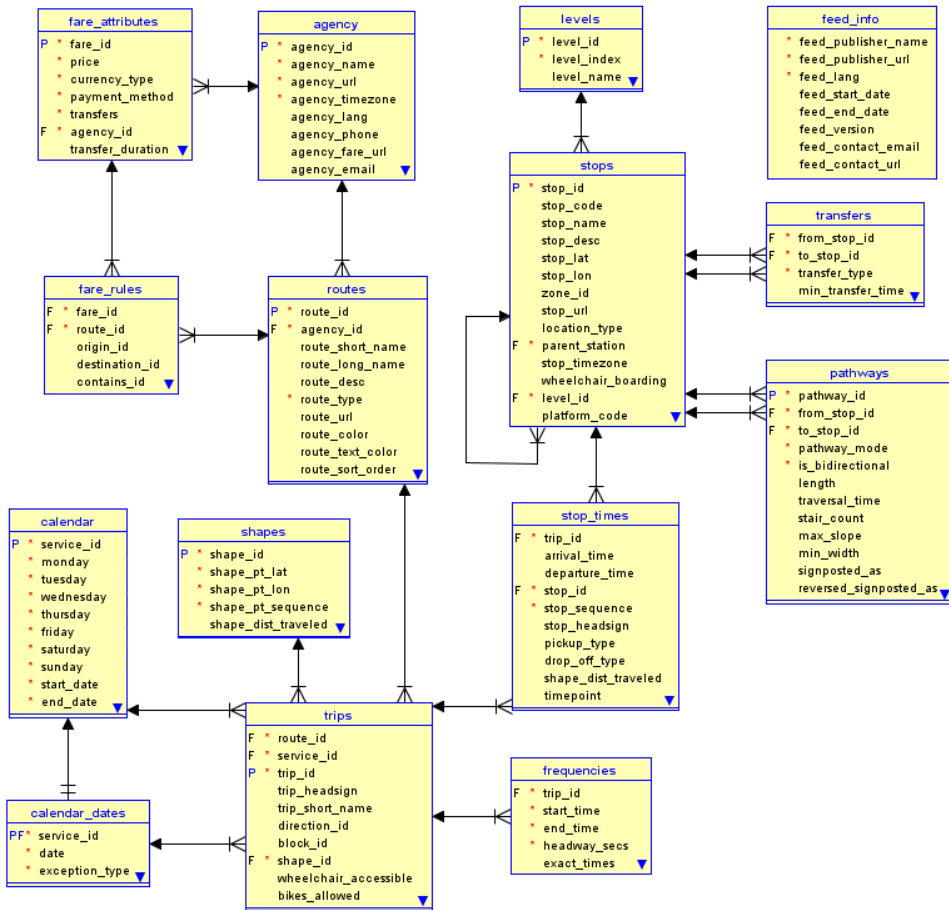


Figure 1. Model of GTFS.

2. GTFS

“The General Transit Feed Specification defines a common format for public transportation schedules and associated geographic information. GTFS ‘feeds’ let public transit agencies publish their transit data and allow developers write applications that consume the data in an interoperable way.” [10] It was introduced by Google in 2005. [9]

The GTFS contains 15 text files in which the fields are separated by commas. The Figure 1 shows a diagram of the GTFS. I modelled it with Oracle SQL Developer Data Modeller, at the same time it is known that the GTFS is not satisfies the relational requirements. [20, 21]

5 files of the 15 are compulsory: agency, routes, trips, stop_times and stops.

Additionally at least one file is required out of the `calendar_dates` and `calendar`.

Many GTFS feeds can be downloaded from various websites. I preferred the <https://transitfeeds.com/> website, where the GTFS feeds are organized based on their location. During my work I recognized that I need local knowledge, and as I live in Debrecen I asked the local GTFS feed from Debrecen Regional Transport Association⁷.

3. Neo4j

The first version of Neo4j was developed in 2002 [15]. Neo4j is a graph database management system, it can store and manage property graphs, which means that the database contains nodes and directed relationships, additionally each node and relationship can have some properties. Each node can have labels which show the roles of the nodes in the database. The relationships can have type, and each of them connects the start node to the end node. [15]

Beside the structure it is very important what kind of tools a database management system can offer for searching data in the database. The Neo4j documentation [15] states “Neo4j was built to efficiently store, handle, and query highly-connected data in your data model” and “accessing nodes and relationships in a native graph database is an efficient, constant-time operation and allows you to quickly traverse millions of connections per second per core.” So I supposed that searching routes in a graph database contained data from a GTFS feed using built-in tools of Neo4j will be very easy.

4. Related work

There are research papers about storing and/or managing GTFS feed data in graph database with more or less success. In the following paragraphs I introduce a few important of them.

Fortin [9] analysed transit networks. They used GTFS feeds as source information and loaded the data into a Neo4j database. They found that their structure didn't support the route planning with Neo4j tools so in their future research they need to change the structure or use other tools for route planning.

Miler [13] stated that “graph database management systems are not routing engines and are not suitable for full graph traversal, which is used in the shortest path calculations”. They also said that “if the memory is not an issue, then graph database is the right choice for the shortest path calculation.”

Kaltenrieder [12] also worked with Neo4j but they enhanced it with their own program code to realize route planning. Falco [8] used Neo4j to store GTFS data but they didn't apply route planning in their system, they provide only transport information to their users. Similary Abbeyquaye [2] employed Neo4j database to store GTFS data, but he built his route engine in JavaScript.

⁷www.derke.hu

Gao [11] found that their relational approach for graph search queries such as the shortest path discovery is more efficient than the algorithms of Neo4j on large graphs.

All in all, many researchers wanted to store the GTFS data in Neo4j. And as you see, it is not easy to apply the Neo4j tools for route planning on this data.

5. Load GTFS into Neo4j

My initial idea was to load GTFS data into the Neo4j as I can in the easiest way. My goal was to write a common loader which can process any GTFS feeds.

I wrote a Python program to load the data. First I used `py2neo`⁸, but it turned out very early that it doesn't support my work, so I had to change to Neo4j Bolt Driver for Python⁹.

The agency, stop, route, trip and stop_times files were load into the database a way that each row of a file become a node in the database. The labels of the nodes showed from which file they come, moreover the values in the rows became the properties of the nodes, where the optional attributes were not loaded into the database. I used the headlines of the files to name the properties of the nodes. I worked the same way on the optional files paying attention that they are optional. The optional files are: level, shape, fare_attributes, fare_rules, frequencies, transfers, pathways and feed_info. In Neo4j I followed the name convention of the Neo4j [15], so I used Camel case, beginning with an upper-case character and I didn't use plural for the nodes.

I created relationships between nodes based on the relationships between the files introduced on Figure 1. So I created the relationship listed in Table 1. When a relationship was made I deleted the appropriate property of the node which stored the connection information. The names of the relationship were followed the name convention of Neo4j, all of them are upper case, using underscore to separate words. I named the relationships in a way that I used the labels of the start and end nodes, supplementing with other information (like to or from) if it is needed.

In the shapes file many lines with the same shape_id describe a shape. So I created a ShapeID node for each shape_id and I created relationships between the Shape nodes and the ShapeID nodes, and between Trip and ShapeID. See Table 2 for the new relationships.

I found a similar problem with block_id in the case of Trip, where more than one trips can have the same block_id. So I created Block nodes, and made relationships between Trip and Block with block_id. See Table 2 for the new relationship.

The last problem was caused by the calendar and calendar_dates. I decided that I use preprocessing for these two files, so I generated a new calendar_tmp file to assign dates to service_ids. First I went through the calendar file and from the start_date to the end_date I generated all adequate dates to the service_id. Then, I went through on the calendar_dates and I deleted or inserted the {service_id,

⁸<https://py2neo.org/v4/>

⁹<https://neo4j.com/docs/api/python-driver/current/>

date pair} from the `calendar_tmp` if a row of `calendar_dates` showed it. Finally I created Service and Date labelled nodes based on the `calendar_tmp` and made the relationships between them.

Then I created the relationship between Trip and Service nodes with the `service_ids` of the Trip.

Table 1. Additional relationships of nodes in Neo4j.

Relationship name	From node	To node	Source GTFS file	Column name in GTFS file
SHAPE_SHAPEID	Shape	ShapeID	shapes	shape_id
TRIP_SHAPEID	Trip	ShapeID	trips	shape_id
TRIP_BLOCK	Trip	Block	trips	block_id
SERVICE_DATE	Service	Date	calendar_tmp	date
TRIP_SERVICE	Trip	Service	trips	service_id

At this point I loaded every information of the GTFS structure to the Neo4j and the loaded data structure follows the logic of the GTFS structure. Now I have a graph structure on which I can try out the route planning tools of the Neo4j. See the model of the loaded data structure drawn in Neo4j at Figure 2.

I tested my data loader with the GTFS feeds for Debrecen¹⁰; Budapest, Debrecen, Miskolc, Pécs, Tampere and Szeged¹¹; and the sample feed provided by the Google¹². I found that there are many GTFS feeds which don't contain all optional files of the GTFS.

6. Route planning tasks

In route planning the traveller wants to reach another place if they are in a given place additionally they want to leave now, so the time and the date is also important. As my goal was to examine how the Neo4j route planning tools work with GTFS data loaded into the Neo4j, it was simpler if I used stops instead of GPS coordinates. With the Harvesine formula [3] I can easily find the near stops to a given place of which the GPS coordinates are given.

The easiest routing task is to find a route between two stops if it exists, I mean when the traveller doesn't need to change the line.

A more difficult task is when the stops are not on the same route, so the traveller should change the route. The problem can be easier if the traveller can change the route only in the stop. More difficult solution if the transfers and the pathways of the GTFS should be used. Since the GTFS feeds that I used miss these optional

¹⁰derke.hu

¹¹transitfeeds.com/

¹²developers.google.com/transit/gtfs/examples/gtfs-feed

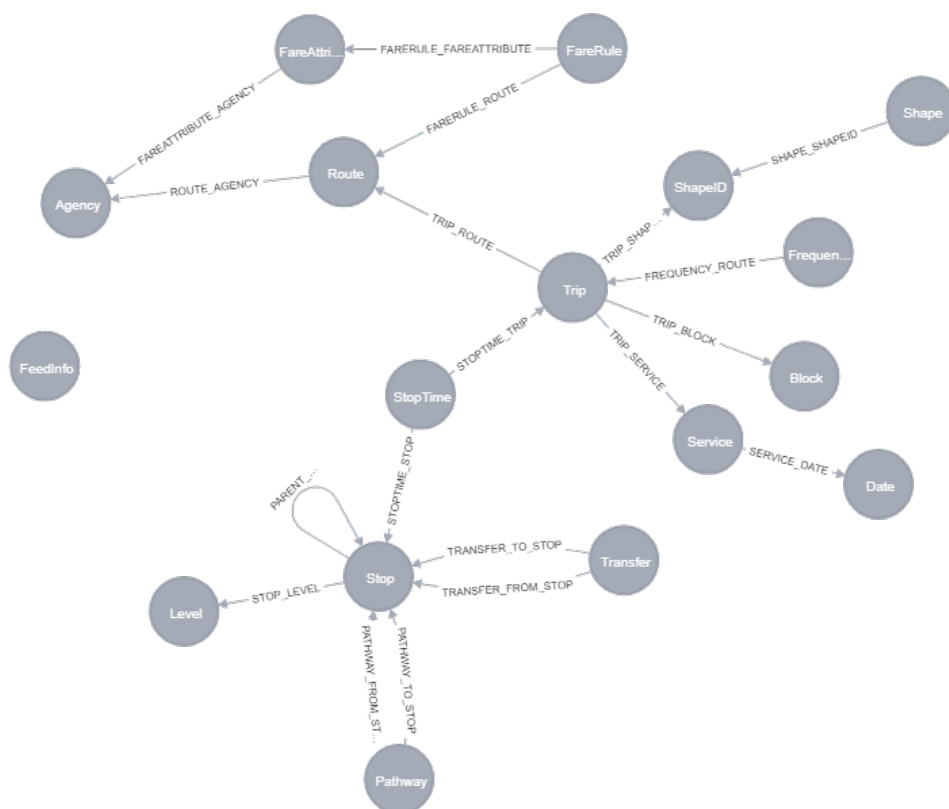


Figure 2. Model of the loaded data structure drawn in Neo4j.

files, I didn't consider this case. I didn't consider the parent station relationship also to make the problem easier.

The most route planning algorithms offer some modes which set of routes the traveller needs: all, the shortest, the k-shortest and the applications can offer some other modes also. Of course the first that I need is the shortest, but as I use the public transport in Debrecen with the help of an app by Szincsák [19, 20], I see that it is not enough. The buses are late many times, so I prefer choosing a more frequent line with more walking than the shortest route. All routes is also not the good choice, since it will contain the routes which goes first time to the border of the city, than comes back so it takes 2 hours instead of the 10 minutes which is offered by the shortest path. So I prefer the k-shortest path. Another idea to limit the route changes in a way since nobody wants to change routes many times.

7. Introduction Neo4j opportunities for route planning on the loaded data structure

Neo4j uses the Cypher graph query language to query the graph. Its basic tool is the MATCH clause with WHERE and RETURN clauses with which the developer can find nodes and relationships in the graph.

To find the route between two stops without changing the line, the following Cypher statement can be executed:

```
match p= (startStop:Stop)-[:STOPTIME_STOP]-(st1:StopTime)-
  [:STOPTIME_TRIP]-(t:Trip)-[:STOPTIME_TRIP]-(st2:StopTime)-
  [:STOPTIME_STOP]-(endStop:Stop)
where endStop.stop_name='Laktanya utca'
  and startStop.stop_name='Vezér utca'
  and toInt(st1.stop_sequence)<toInt(st2.stop_sequence)
return p;
```

Figure 2 helps to follow the names of nodes and relationships. In this first Cypher statement I didn't use time and date for the searching, so the result contains many path between the startStop and endStop. Since the direction is important and the traveller cannot travel to the opposite direction than the vehicle goes, I should put the condition about the stop_sequence to the statement.

If I consider the date and the time also and if I want to show the route information, the following Cypher statement can be used:

```
match p= (startStop:Stop)-[:STOPTIME_STOP]-(st1:StopTime) -
  [:STOPTIME_TRIP]-(t:Trip)-[:STOPTIME_TRIP]-(st2:StopTime)-
  [:STOPTIME_STOP]-(endStop:Stop),
p2=(t:Trip)-[:TRIP_ROUTE]-(r:Route),
p3=(t:Trip)-[:TRIP_SERVICE]-(ser:Service)-[:SERVICE_DATE]-(d:Date)
where endStop.stop_name='Laktanya utca'
  and startStop.stop_name='Vezér utca'
  and toInt(st1.stop_sequence)<toInt(st2.stop_sequence)
  and d.date=" 20190503"
  and st1.departure_time>'10:00:00'
  and st1.departure_time<'11:00:00'
  and st2.arrival_time<'12:00:00'
return p,p2,p3;
```

On the Debrecen data this Cypher query works well.

In the case when the traveller should change the route, and they have this opportunity only in Stops, based on Figure 2 you can see that in the statement we should start from a Stop, than go to a StopTime as in the previous case and we should end again with a StopTime and a Stop. Between the starting StopTime and ending StopTime we should move on the STOPTIME_TRIP or STOPTIME_STOP

relationships several times. If I want to get back all solutions, a Cypher statement can be used which is a modified version of the previous one, namely I have put an asterisk to the proper relationship, like this:

```

match p= (startStop:Stop)-[:STOPTIME_STOP]-(st1:StopTime)-
  [:STOPTIME_TRIP]-(t:Trip)-[:STOPTIME_TRIP|:STOPTIME_STOP*]-
  (st2:StopTime) - [:STOPTIME_STOP]-(endStop:Stop)
where endStop.stop_name='Laktanya utca'
  and startStop.stop_name='Gyepusor utca'
  and toInt(st1.stop_sequence)<toInt(st2.stop_sequence)
  and st1.departure_time>'10:00:00'
  and st1.departure_time<'11:00:00'
  and st2.arrival_time<'12:00:00'
with p, nodes(p) as nodes
where all(x in nodes where not(labels(x)='Trip')
  or exists((x)-[:TRIP_SERVICE]-(:Service)-
    [:SERVICE_DATE]-(:Date{date:" 20190503"}) ))
return p;

```

The query stores the time and date information. I solved the date information a way that every Trip in the path should run on the given day.

On the Debrecen data this Cypher query doesn't work, it causes out of memory error. I could change the memory size for the Neo4j, but to find all routes with line changes between two stops is so many solutions that it is not worth to search.

If I want to limit the route changes, the previous Cypher statement can be changed a way, that `[:STOPTIME_TRIP|:STOPTIME_STOP*]` part gets a limit. Following Figure 2 I found that the multiplication number of the relationships can be calculated the following way: $(\text{changes}+1)*4-3$. With 0 changes this number is 1, with 5 changes this number is 21. I tried out with more numbers from 5 to 21:

```

match p= (startStop:Stop)-[:STOPTIME_STOP]-(st1:StopTime)-
  [:STOPTIME_TRIP]-(t:Trip)-[:STOPTIME_TRIP|:STOPTIME_STOP*1..21]-
  (st2:StopTime)-[:STOPTIME_STOP]-(endStop:Stop)
where endStop.stop_name='Laktanya utca'
  and startStop.stop_name='Gyepusor utca'
  and st1.departure_time>'10:00:00'
  and st1.departure_time<'11:00:00'
  and st2.arrival_time<'12:00:00'
with p, nodes(p) as nodes
where all(x in nodes where not(labels(x)='Trip')
  or exists((x)-[:TRIP_SERVICE]-(:Service)-
    [:SERVICE_DATE]-(:Date{date:" 20190503"}) ))
return p;

```

In all cases the execution time of this query is very long, I don't think that it could be used in an application. Moreover, in the where clause there is a complex

condition, and we should use more complex conditions than this to find the routes that we need. I tried to give hints to the execution plan, but it didn't help the query.

Then, I considered the built-in functions of Neo4j which support the route planning. Neo4j offers path finding algorithm, namely Minimum Weight Spanning Tree, Shortest Path, Single Source Shortest Path, All Pairs Shortest Path, A*, Yen's K-shortest paths and Random Walk [15].

In my work the startStop and the endStop are known, so the Minimum Weight Spanning Tree is useless for this problem. Similarly, the the All Pairs Shortest Path is not the solution at this situation since it find the shortest paths between all pairs of nodes. Than as well the Single Source Shortest Path (SSSP) algorithm is not useful in this situation as it calculates the shortest (weighted) path from a node to all other nodes in the graph. The Random Walk provides random paths in a graph, but in my case the route is not random.

The Shortest Path algorithm uses Dijkstra algorithm. To see the nodes of the route between the startStop and endStop its stream version should be used.

```
match p=(startStop:Stop{stop_name:'"Laktanya utca"'}),
      (endStop:Stop{stop_name:'"Gyepusor utca"'})
call algo.shortestPath.stream(startStop, endStop)
yield nodeId
return algo.getNodeById(nodeId);
```

The function and so the statements works well, and its execution time is also good. However, the travel wants to move in a given date and time, so some restrictions is needed for the statement. The function doesn't offer such conditions in this form.

The next algorithm is the Yen's K-shortest paths algorithm, which computes single-source k-shortest loopless paths for a graph with non-negative relationship weights.

```
match (startStop:Stop{stop_name:'"Laktanya utca"'}),
      (endStop:Stop{stop_name:'"Gyepusor utca"'})
call algo.kShortestPaths.stream(startStop, endStop, 5, 'cost' ,{})
yield index, nodeIds, costs
return algo.getNodesById(nodeIds)
```

Similarly as the shortestPath function, the kShortestPath function and so the statements works well, the execution time is also good, but the problems are also similar as in the case of the shortestPath algorithm, namely some restrictions is needed because of the date and time.

The A* algorithm improves the shortest path algorithm that way that the user can add some information to the algorithm in order that the algorithm could make better choices over which paths to take through the graph. The syntax and the usage of the algorithm can be read in the documentation of the Neo4j [15]. It needs a kind of heuristic, the weight is compulsory, which can be the time between the

stops, but we don't have in our data structure. The algorithm also needs longitude and latitude, which are given in the Stop but for the other nodes they are not given, so we cannot use this algorithm in this form on our structure.

To sum up this section the data structure introduced on Figure 2 is very good to store the data but it doesn't support the route planning in the given city. So in the next chapter I modified it to support the route planning.

8. The modified version of the data structure

Pyrga [16] compared the time-expanded and the time-dependent graphs for transport information and they find that the time-dependent graph was more compact and offers a clearly better performance. The time-expanded approach means that every event at a station is modelled as a node in the graph. In my data structure the events are modelled as StopTimes, so my data structure corresponds in some features of the time-expanded approach. The time-dependent approach means that the graph contain only one node per station and there is an edge between two stations if there is an elementary connection from the one station to the other.

Following the ideas of the time-dependent graph approach to build simpler routes in the graph, I should connect the Stops in the graph, and I should equip the relationships with information in order to find the routes easier.

I made a REACH relationship from a start node to an end node if there is a route with which a traveller can reach the end stop from the start stop. This means that if a route goes from A Stop to D Stop and between the two stops the route stops in B and C stops, I created the relationships which are introduced in Figure 3, so D can be reached from A, B and C, C can be reached from A and B, and B can be reached from A.

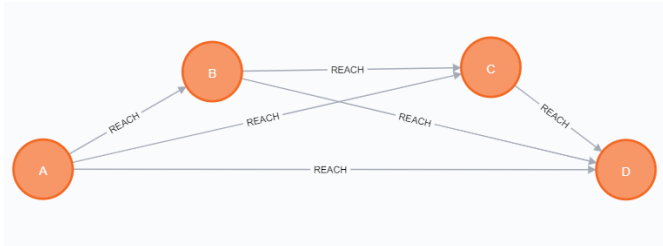


Figure 3. Routes between A and D.

Additionally I stored some information on the relationships: the `route_id`, the minimum and maximum duration, `max_shape_distance_traveled_diff` and `min_shape_distance_traveled_diff`, `max_stop_sequence_diff` and `min_stop_sequence_diff`. I wanted to store the `trip_id` list also, but the Neo4j doesn't allowed it since the list was too long.

I calculated the minimum and maximum duration based on the arrival and departure times of the StopTimes. Similarly the StopTimes stores

shape_distance_traveled information, so the shape_distance_traveled_diff is the difference of the shape_distance_traveled of the two Stops. I calculated similarly the stop_sequence_diff from the stop_sequence property of the StopTimes.

A REACH relationship refers one route and not a trip. A StopTime belongs to a Trip, but I wanted to make as simple the relationship as I can, so I wanted to store information for a Route. Since a Route covers many Trips, I used the maximum and the minimum values of each calculated properties. There are only a few REACH relationships for Debrecen where the min and the max for shape_distance_traveled_diff is not the same. The duration depends on the time of the day, so there are many REACH relationships where they are not the same. And finally the min and max of stop_sequence_diff-s are equal in the case of all REACH relationships for Debrecen.

With this data structure first I wanted to find the route between two stops without changing the line. It's an easy Cypher statement:

```
match (sf:Stop)-[:REACH]->(st:Stop)
where sf.stop_name="Laktanya utca"
      and st.stop_name="Segner tér"
return sf, st
```

This statement is to find all route opportunity between the two stops without any limitations, so it would run for a long time, and we know that the result would be so much that it is not worth to work with it. The following statement is for one line change:

```
match p=(sf:Stop)-[:REACH*1..2]->(st:Stop)
where sf.stop_name="Laktanya utca"
      and st.stop_name="Segner tér"
return p
```

In the result there are many Stops, since between the start and end stop there are many stops where the traveller can change the "line". If you sit on a bus, you don't want to get off and get on again, so we need to filter the result. In the filter I tried to write down that the route_id is distinct during a solution. I tried it a way that the number of the relationships in the path is the same as the number of the distinct route_ids. Since the route_id is a property of a relationship the Neo4j doesn't allow to write such a statement, additionally the count also doesn't work in this conditions. Then I tried the all predictive function in the where condition, but it doesn't allow to put two variable before its where part.

The next opportunity to use the graph algorithms of Neo4j. The first is the shortestPath algorithm. In this case the REACH relationship can be used, and the direction can be given. Moreover it is important to use a weight, since a REACH relationship can mean only 1 sec or even 1 hour. The first time I used the max_duration as weight. My first trying was the following:

```
match p= (startStop:Stop{stop_name:'Laktanya utca'}),
```

```
(endStop:Stop{stop_name:'"Gyepusor utca"'})
call algo.shortestPath.stream( startStop, endStop, 'max_duration',
  {relationshipQuery:'REACH',direction:'OUTGOING'})
yield nodeId
return algo.getNodeById(nodeId)
```

The algorithm works well, but in this form the statement doesn't know anything about the time and date. So the resulted shortest path may not exist in the necessary time interval. The same problem arises if I use the `max_shape_dist_traveled` as weight.

The next examined algorithm is the A* algorithm:

```
match (startStop:Stop{stop_name:'"Laktanya utca"'}),
(endStop:Stop{stop_name:'"Gyepusor utca"'})
call algo.shortestPath.astar.stream(startStop, endStop,
  'max_duration','flat', 'flon',
  {nodequery:'Stop',relationshipQuery:'REACH',
  direction:'OUTGOING', defaultvalue:1.0})
yield nodeId, cost
return algo.getNodeById(nodeId)
```

It works well, but it also gives back only one solution which is not consider the date and the time, so we may not find route in a given time to reach our goal.

The last algorithm is the k shortest path algorithm:

```
match
  (startStop:Stop{stop_name:'"Laktanya utca"',stop_id: '2400205'}),
  (endStop:Stop{stop_name:'"Gyepusor utca"', stop_id: '2300507'})
call algo.kShortestPaths.stream(startStop, endStop,
  10, 'max_duration2',
  {nodequery:'Stop', relationshipQuery:'REACH',
  direction:'OUTGOING'})
yield index, nodeIds, costs
return index, nodeIds, costs
```

As we get a lot of solution, this algorithm can be useful to plan route between two stops.

I also used the `shape_distance_diff` as the weight to find the k-shortest path, but the function give back out of memory error. Since the previous statement worked well, at this point I changed the memory size from 0.5 GB to 4GB (the size of the database was 0.5GB), but the result was the same. I tried out with the `max_duration` as a weight, and it was surprising to me, since with the other weight the algorithm worked well. Additionally this weight would be better for the route planning.

I worked further with the kshortest path using the `max_duration`, since it works. I examined each routes whether they have appropriate date and time. If yes, it is a solution, if no, we can through this route away.

The following code is an example how we can examine the route. This statement results only a travelling between two stops without change the line. We should examine all the route, each of which are in the route list resulted the kshortest path. Since the statement was slow, I used some hints to make faster, so the speed can be acceptable.

```

match (s1:Stop{stop_id:"2400205"})-[rch:REACH]->
  (s2:Stop{stop_id:"1001605"}),
  (r:Route)-[]-(t:Trip),
  (t)-[]-(st1:StopTime)-[]-(s1),
  (t)-[]-(st2:StopTime)-[]-(s2),
  (t)-[]-(se:Service)-[]-(d:Date)
using join on s1,s2
where id(r)=rch.route_id
  and st1.departure_time>"16:00:00"
  and st1.departure_time<"18:00:00"
  and d.date=" 20190510"
return r,t,d, se, s1,s2,st1,st2

```

So the kShortestPath algorithm on this data structure with post processing to filter the date and time can be used for route planning in Neo4j on GTFS data, but as we see here, it gives error for some circumstances.

9. Other data structures

It is obvious that the stop should be nodes in the Neo4j. Additionally nodes and/or relationships should be put between two stops if the end stop can be reached from the start stop somehow. The modified structure was the one extremity where there is a relationship between two nodes if there is a route between them. The basic structure first was the other extremity, two stops are connected through the stop_times and trips. Between these two extremity I tried more opportunity, I connected the stop_times if there is a trip between them, then I connected the stops if there is a trip between them, I connected only the neighbour stops or stop_times, where the neighbour means that there are no other stops between the neighbour stops in a route or a trip. I found the same problems everywhere: I cannot write statements which can find the routes that I need, or I find something, but it doesn't work since out of memory error, or it is very slow and I cannot tune the query to be good.

10. Conclusion

Several researchers find that Neo4j doesn't support the route planning for transit transport [9, 13]. Others used Neo4j only to store transit data [2, 8, 12]. But Miller [14] states that a graph database is "the best solution if there is a need for

a dynamic data model that represents highly connected data”. Abay [1] says that “the graph database model is particularly useful when data connectivity of the data is as important as the data itself”.

In my work I loaded GTFS data into the Neo4j and then I tried to apply the route planning algorithms or statements of Neo4j on it. I used the match-where-return cypher statements, than the shortest path, the k-shortest path and the A* algorithm. I changed the data structure to support the route planning. Finally I found a solution with the kShortestPath built-in function, but this function cause out of memory error in some cases.

Despite these facts I liked to work with Neo4j, it offers a browser, which can effortlessly be used, the cypher statements can easily be understood, the Neo4j Bolt Driver for Python can be used simply. But I found that the documentation doesn’t contain complex examples, I had to browse the internet for many solutions, and as I found it at the kShortestPath algorithm it contains some programming mistakes.

The Neo4j is about 17 years old, and the goal of this database management system was not the route planning. Even so I nearly found a solution for route planning on GTFS. I hope that Neo4j will be improved and after a few years it can be used also for route planning on GTFS data also.

References

- [1] N. C. ABAY, A. MUTLU, P. KARAGOZ: *A path-finding based method for concept discovery in graphs*, in: 2015 6th International Conference on Information, Intelligence, Systems and Applications (IISA), 2015 6th International Conference on Information, Intelligence, Systems and Applications (IISA), Corfu, Greece: IEEE, July 2015, pp. 1–6, ISBN: 978-1-4673-9311-9, DOI: <https://doi.org/10.1109/IISA.2015.7388092>, URL: <http://ieeexplore.ieee.org/document/7388092/> (visited on 02/01/2021).
- [2] A. ABBEYQUAYE: *Building a map-based transit planner for the tro-tro system in Accra, Applied Project*, Ashesi University College, 2017.
- [3] N. R. CHOPDE, M. K. NICHAT: *Landmark based shortest path detection by using A* Algorithm and Haversine Formula*, International Journal of Innovative Research in Computer and Communication Engineering 1.2 (2013), pp. 298–302.
- [4] R. DAMERI, A. COCCHIA: *Smart city and digital city: Twenty years of terminology evolution*, in: X Conference of the Italian Chapter of AIS, ITAIS2013, 2013, pp. 1–8.
- [5] R. P. DAMERI: *Searching for Smart City definition: a comprehensive proposal*, INTERNATIONAL JOURNAL OF COMPUTERS & TECHNOLOGY 11.5 (Oct. 30, 2013), pp. 2544–2551, ISSN: 2277-3061, DOI: <https://doi.org/10.24297/ijct.v11i5.1142>, URL: <https://cirworld.com/index.php/ijct/article/view/1142ijct> (visited on 02/01/2021).
- [6] *DB-Engines Ranking*, 2021, URL: <https://db-engines.com/en/ranking>.
- [7] R. ELMASRI, S. NAVATHE: *Fundamentals of database systems*, 6th ed, OCLC: ocn586123196, Boston: Addison-Wesley, 2011, 1172 pp., ISBN: 978-0-13-608620-8.

- [8] E. FALCO, I. MALAVOLTA, A. RADZIMSKI, S. RUBERTO, L. IOVINO, F. GALLO: *Smart City L'Aquila: An Application of the "Infostructure" Approach to Public Urban Mobility in a Post-Disaster Context*, Journal of Urban Technology 25.1 (Jan. 2, 2018), pp. 99–121, ISSN: 1063-0732, 1466-1853,
DOI: <https://doi.org/10.1080/10630732.2017.1362901>,
URL: <https://www.tandfonline.com/doi/full/10.1080/10630732.2017.1362901> (visited on 02/01/2021).
- [9] P. FORTIN, C. MORENCY, M. TRÉPANIÉ: *Innovative GTFS Data Application for Transit Network Analysis Using a Graph-Oriented Method*, Journal of Public Transportation 19.4 (Dec. 2016), pp. 18–37, ISSN: 1077-291X, 2375-0901,
DOI: <https://doi.org/10.5038/2375-0901.19.4.2>,
URL: <http://scholarcommons.usf.edu/jpt/vol19/iss4/2/> (visited on 02/01/2021).
- [10] *GTFS Static Overview*, 2021,
URL: <https://developers.google.com/transit/gtfs/>.
- [11] JUN GAO, JIASHUAI ZHOU, J. X. YU, TENGJIAO WANG: *Shortest Path Computing in Relational DBMSs*, IEEE Transactions on Knowledge and Data Engineering 26.4 (Apr. 2014), pp. 997–1011, ISSN: 1041-4347,
DOI: <https://doi.org/10.1109/TKDE.2013.43>,
URL: <http://ieeexplore.ieee.org/document/6475943/> (visited on 02/01/2021).
- [12] P. KALTENRIEDER, J. PARRA, T. KREBS, N. ZURLINDEN, E. PORTMANN, T. MYRACH: *A Dynamic Route Planning Prototype for Cognitive Cities*, in: Designing Cognitive Cities, ed. by E. PORTMANN, M. E. TABACCHI, R. SEISING, A. HABENSTEIN, vol. 176, Series Title: Studies in Systems, Decision and Control, Cham: Springer International Publishing, 2019, pp. 235–257, ISBN: 978-3-030-00316-6 978-3-030-00317-3,
DOI: https://doi.org/10.1007/978-3-030-00317-3_10,
URL: http://link.springer.com/10.1007/978-3-030-00317-3_10 (visited on 02/01/2021).
- [13] M. MILER, D. MEDAK, D. ODOBAŠIĆ: *The shortest path algorithm performance comparison in graph and relational database on a transportation network*, PROMET - Traffic&Transportation 26.1 (Feb. 28, 2014), pp. 75–82, ISSN: 1848-4069, 0353-5320,
DOI: <https://doi.org/10.7307/ptt.v26i1.1268>,
URL: <https://traffic.fpz.hr/index.php/PROMTT/article/view/1268> (visited on 02/01/2021).
- [14] J. J. MILLER: *Graph Database Applications and Concepts with Neo4j*, in: SAIS 2013 Proceedings, 2013, pp. 1–24.
- [15] *Neo4j*, 2021,
URL: <https://www.neo4j.com>.
- [16] E. PYRGA, F. SCHULZ, D. WAGNER, C. ZAROLIAGIS: *Efficient models for timetable information in public transportation systems*, ACM Journal of Experimental Algorithmics 12 (June 2008), pp. 1–39, ISSN: 1084-6654, 1084-6654,
DOI: <https://doi.org/10.1145/1227161.1227166>,
URL: <https://dl.acm.org/doi/10.1145/1227161.1227166> (visited on 02/01/2021).
- [17] P. J. SADALAGE, M. FOWLER: *NoSQL distilled: a brief guide to the emerging world of polyglot persistence*, Upper Saddle River, NJ: Addison-Wesley, 2013, 164 pp., ISBN: 978-0-321-82662-6.
- [18] N. TUAYCHAROEN, A. SAKCHAROEN, W. CHA-AIM: *Bangkok Bus Route Planning API*, Procedia Computer Science 86 (2016), pp. 441–444, ISSN: 18770509,
DOI: <https://doi.org/10.1016/j.procs.2016.05.075>,
URL: <https://linkinghub.elsevier.com/retrieve/pii/S1877050916304112> (visited on 02/01/2021).
- [19] A. VÁGNER, T. SZINCÁSÁK: *Data structure to store GTFS data efficiently on mobile devices*, Journal of Computer Science and Software Application 1.1 (2014), pp. 27–41.

-
- [20] A. VÁGNER, T. SZINCSÁK: *Public transit schedule and route planner application for mobile devices*, in: Proceedings of the 9th International Conference on Applied Informatics, Volume 2, The 9th International Conference on Applied Informatics, Eger, Hungary: Eszterházy Károly College, 2015, pp. 153–161, ISBN: 978-615-5297-19-9,
DOI: <https://doi.org/10.14794/ICAI.9.2014.2.153>,
URL: <http://icai.ektf.hu/icai2014/papers/ICAI.9.2014.2.153.pdf> (visited on 02/01/2021).
- [21] J. C. WONG: *Use of the general transit feed specification (GTFS) in transit performance measurement*, Georgia Institute of Technology, 2013.

Methodological papers

Visual argumentations in teaching trigonometry

Francesco Laudano

Università degli studi del Molise

francesco.laudano@unimol.it

Submitted: April 13, 2021

Accepted: October 13, 2021

Published online: October 22, 2021

Abstract

In this paper, we study the possibility of building a learning path that allows students to develop trigonometric knowledge and skills by the end of Grade 10 of secondary science-based schools. In particular, we describe an action research experiment, in part done through distance learning, aimed at incorporating all trigonometry topics within the framework of the study of Euclidean geometry. The inquiry-based learning methodology and the support of dynamic geometry software with a laboratory teaching approach were used. The learning path is based on several “visual/dynamic proof” and is explained by an example lesson on the Cosines Law.

The experiment could be extended by teachers into physical/virtual classrooms and could offer practical strategies and tools for teaching trigonometry.

Keywords: Trigonometric path, Euclidean geometry, trigonometric relationships, visual proofs

AMS Subject Classification: 97G10, 97G60

1. Introduction

The multiple applications of trigonometry have made it a fundamental milestone in the training of students, which has led many governments to introduce the study of this topic since the first year of secondary science-based schools. This practice can be seen, for example, in Canada ([13, 14]), France ([8–10]), Hungary ([6]),

Italy ([11]), the UK ([5]), and the USA ([1]). However, the study of trigonometry is performed mainly in the third and fourth years, probably because some trigonometric reasoning seems too abstract for 14- to 15-year-old students. Moreover, as highlighted in [2], the learning of trigonometry is a highly difficult area of mathematics for both students and teachers. Furthermore, learning outcomes are strongly influenced by the teaching approach, which in some cases points towards the application of trigonometric relations, almost entirely omitting proofs, and in other cases, in contrast, emphasizes algebraic-formal aspects at the risk of losing students' attention. In the following, without neglecting these two important aspects of trigonometry training, we present a geometric path that allows to develop the applications of trigonometry without overlooking the educational potential of the proofs. The core idea, which is developed in the following, consists of incorporating trigonometry into the elementary geometry path, replacing the classic formal arguments with visual and dynamic proofs, which are simpler to build, especially in distance learning. This is in order to improve students' demonstrative skills, since, as pointed out in [18] and [7], even students who are particularly talented in mathematics, they still have difficulty in solving problems that require proof.

The experiment, in part done through distance learning, was built on the inquiry-based learning (IBL) methodology (see [4] for a literature review). The use of educational technologies in the teaching of trigonometry has been analyzed by Ross et al. [16] and can provide decisive help to teachers. Therefore, in the development of the path, dynamic geometry software (GDS) with a laboratory teaching approach was used. The construction of mathematical ideas is based upon real problems and follows the different phases of inquiry until students' knowledge is deeply rooted through the relevant proofs.

2. The didactic project

Since the IBL is founded on concrete problems and stimulates questions and actions to solve them, it seemed as the most suitable methodology to use in order to implement an innovative learning activity for students in G9–G10. The fundamental ideas of IBL can be found in Dewey's thought [3]. According to him, students build their knowledge and skills through a sequence of research stages, in which they formulate hypothesis, verify them and discuss the results of their investigations [17]. The IBL activities are based on a workshop format: the teacher acts as a facilitator by guiding students' development and exchange of ideas by asking them appropriate questions; students work in small groups and take active part in their learning process [15]. In the teaching of geometry, we start from a concrete problem and, after having identified its essence, we search for its solution. In the activity that we describe, the basic idea consists of making students observe and manipulate through GDS some geometric constructions which had been appropriately prepared, and then gradually direct students towards the discovery of the formulas and trigonometric theorems. In the next phase, the teacher raises

students' awareness to the need to prove the relations that have been discovered, and guides them to the search of the proofs. Later, in order to establish the effectiveness of the activity, the initial problem from which the activity started is solved by using the discovered properties. In the course of the didactic activities, we have tried to chain the evolution of geometric knowledge and students' gratification in the awareness of their mutual influence. In particular, efforts have been made to concatenate the two aspects in order to trigger a virtuous circle in the learning process. In fact, we believe that the possibility of discovering "something new" provides pupils with a "motivational lever" that can be decisive for the development of skills. What is presented in Table 1 is the path that incorporates the study of trigonometry in the development of geometry, thus valuing also this part of Mathematics that is considered by many authors fundamental in the training of students' logical abilities. The trigonometry Units are developed in greater detail in Table 2.

Table 1. Units, General outcomes and Chapters.

Unit	General Outcome	Chapter
Isometry G9	Solve problems involving congruence between polygons, using both measurements and geometric proofs	Congruences Triangles Quadrilaterals Circles
Similarity G9	Solve problems involving similitude between polygons, using both measurements and geometric proofs	Similarities Triangles Circle
Trigonometry 1 G9	Solve triangles using trigonometric theorems	Trigonometric Ratios Right Triangles Any Triangles
Equivalence G10	Solve problems using trigonometric theorems	Equivalences Triangles and Trapezoids Special n -gons Euclidean Theorems
Trigonometry 2 G10	Solve problems involving trigonometric functions using trigonometric identities and trigonometric equations	Trigonometric Functions Trigonometric Identities Equations and Problems

Table 2. Detail of the Trigonometry Chapters.

Unit	Chapter	Lesson
Trigonometry 1	Trigonometric Ratios	Sine, Cosine, Tangent ratio Basic trigonometric identities The ratio of 30° , 60° and 45°
	Right Triangle	Pythagoren Theorem Right Triangle Theorems Solving righth triangles
	Any Triangle	Triangle Area Formula Length of a cord, Sines Law Cosines Law Solving any triangles
Trigonometry 2	Trigonometric Functions	Graph of Trigonometric Functions Periodicity of trigonometric functions Associated angles relations
	Trigonometric Identities	Ptolemy's identities Double angle formulas Half angle formulas Product-sum identities Parametric formulas
	Equations and Problems	Basic equations Some non linear equation Application problems

3. A lesson plan on the Cosines Law

In this section, we describe an example lesson on Cosines Law directed to G9 students. In particular, we get an overview of how students discovered the cosine law while using dynamic geometry software in the classroom.

The activity, which lasts approximately 90 minutes, is conceived as a DGS lab and entailed the use of worksheets to guide the students to start the inquiry.

Objectives:

- Discover the Cosines Law.
- Proof Cosines Law.
- Solve problems using Cosines Law.

For this activity it seemed appropriate to give students many suggestions in order to guide their search in the right direction. Instead, for the activities that followed, suggestions were gradually reduced in order to make students become more and more autonomous in the management of the inquiry.

If we get an overview of how students discovered the cosine law while using dynamic geometry software in the classroom, there is no need to mention the test.

Step 1. Problem posing

Everyone knows the Pythagorean Theorem and how it allows to solve problems related to right triangles. But what happens when the problem concerns non-right triangles? I.e. consider the following text:

“An aircraft tracking station determines the distance from a common point O to each aircraft and the angle between the aircraft. If angle O between two aircraft is equal to 49° and the distances from point O to the two aircraft are 50 km and 72 km, find distance between the two aircraft (round answers to 1 decimal place).”

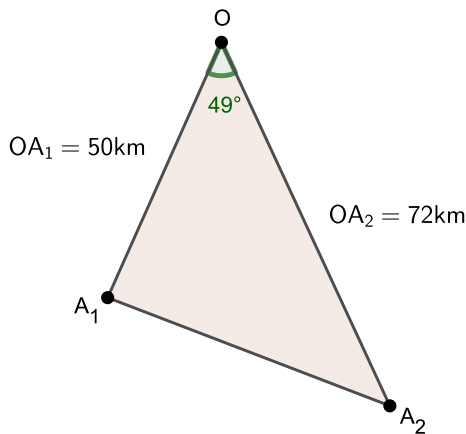


Figure 1. Sketch of the problem.

After a few minutes of reflection followed by a teacher-led discussion, the students propose two possible solution strategies. The first starts by drawing a perpendicular line from point A_1 to OA_2 and applies the Pythagorean formula twice. The second one starts from the GDS construction of a triangle B_1PB_2 (similar to A_1OA_2), with $B_1P = 5$ cm, $PB_2 = 7.2$ cm and $B_1\hat{P}B_2 = 49^\circ$ and find the distance between the two aircraft multiplying the length of B_1B_2 for the ratio 10^6 . The teacher points out that both strategies are valid and general. However, a formula that expresses the length of one side of a triangle as a function of the other sides and the angle between them would make it possible to solve the problem more effectively. Therefore the teacher invites the students to reflect and discuss to answer the following questions.

Question 1. *In principle, could there be a general formula that expresses the length of one side of a triangle as a function of the other sides and the angle between them?*

Some students point out that two sides and the angle between them uniquely determine all the triangle's elements, so they concur that such a formula might exist. Then, the teacher propose an investigation aimed at an extension of Pythagoras' theorem that would be useful to solve the aircraft problem.

Step 2. Working with DGS

The search for the formula involves several stages. The first phase is aimed at discovering the role played by the angle and, in particular, at understanding which of the functions $\sin \alpha$ and $\cos \alpha$ could intervene in the formula. The teacher invites students to work on a file suitably prepared with DGS (see Figure 2) so that they can change the width of the angle \widehat{BCA} by acting on the slider, while BC and DC are segments with a given length equal to 10 cm.

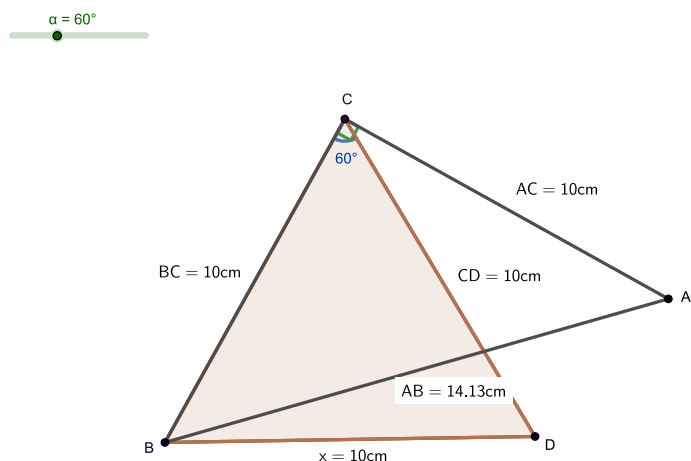


Figure 2. Working with isosceles triangles.

Using the slider, students immediately realize that increasing the width of the angle \widehat{BCA} also increases the length of AB . Then the teacher guides them through the following questions and the related discussion.

Question 2. *How does the value of AB^2 change as the sine and cosine of the angle α ?*

To answer this question, the teacher invites the students to note down on a worksheet the sine and cosine values of the 30° , 45° , 60° , 90° , 120° , 150° angles and the corresponding values of AB^2 .

Table 3. Relations between $\sin \alpha$, $\cos \alpha$ and AB^2 .

α	$\sin \alpha$	$\cos \alpha$	AB^2
30°	0.5	0.866	...
45°
60°	0.866	0.5	14.13 cm
90°
120°
150°

Comparing the relationships shown in the table, the students note that the value of the cosine constantly decreases as the angle increases. Instead that of the sine increases for $0^\circ < \alpha \leq 90^\circ$ while it decreases for $90^\circ \leq \alpha < 180^\circ$. Therefore the length of AB does not seem to be related to $\sin \alpha$.

Moreover, they observe that as α increases, the quantity $-\cos \alpha$ increases, and so does the length of AB .

Therefore the teacher advises students to look for the more simple formula verifying the conditions that emerged with the use of the DSG. Then, since the relation sought must extend the Pythagorean theorem, it guides them through the following questions and related discussion.

Question 3. *Can the formula be obtained from the Pythagorean relation by subtracting an appropriate quantity that depends on $\cos \alpha$?*

The teacher leads the discussion to obtain AB^2 by subtracting a term proportional to $\cos \alpha$ from $BC^2 + AC^2$. Then he invites the students to calculate the proportionality factor in various cases using DGS from Figure 2, to answer the following question.

Question 4. *What is the relationship between AB^2 and α in the previous cases?*

The teacher checks that students, by calculating the quantities $(AB^2 - AC^2 - BC^2)/\cos \alpha$, always get the value 200, as α changes. Then it check to see if students really discover that $AB^2 = 100 + 100 - 2 \cdot 100 \cos \alpha$.

Therefore, the teacher invites students to modify the length of AC and BC to check if the above relation can be extended to other isosceles triangles. By carrying out these tests, students discover the formula $AB^2 = AC^2 + BC^2 - 2 \cdot AC \cdot BC \cdot \cos \alpha$ for isosceles triangles.

In the next phase, the teacher makes the students check if the formula they have found can be extended to the scalene triangle BCD in Figure 3. The related file has been prepared so that the students can change both the width of the angle \widehat{BCD} and the length of BC and DC , using the dynamism of the software. As for isosceles triangles, the teacher leads the discussion to obtain BD^2 by subtracting from $BC^2 + CD^2$ a term proportional to $\cos \alpha$. Then, he invites the students to

calculate the proportionality factor in various cases through DGS to answer the following question.

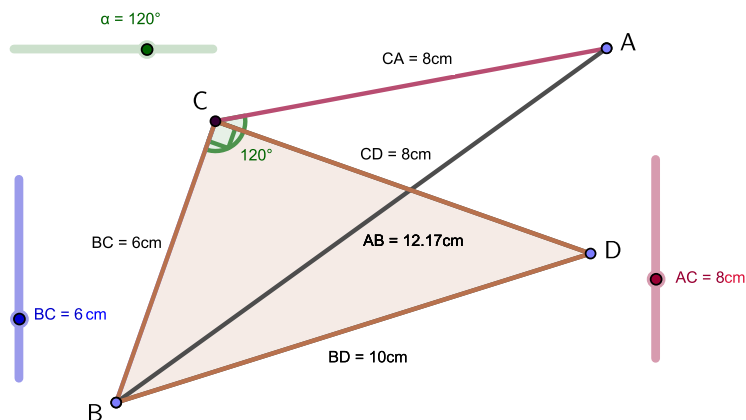


Figure 3. Working with scalene triangles.

Question 5. *Is the above relation still valid for the triangle BCD in Figure 3?*

Students verify that as α changes, the amount $(BD^2 - BC^2 - CD^2)/\cos \alpha$ is constant. Then the teacher invites them to identify the algebraic relationship between AB , BC and this factor, and check to see if students really discover that $AB^2 = 36 + 64 - 2 \cdot 6 \cdot 8 \cos \alpha$. At this point, the teacher invites the students to modify the length of sides AC and BC using the related sliders, to extend to any scalene triangles the relationship between the proportionality constant and their lengths. In this way the students discover that the above formula could hold for every triangle.

Step 3. Formulating a conjecture

The teacher proposes to students to build more triangles, using the prepared DGS file, to test the formula we seem to have discovered. The questions posed by the teacher guide the students in formulating the conjecture.

Question 6. *Is the above relation still valid for any triangle, in particular for obtuse triangles?*

Question 7. *At this point, can we formulate a conjecture to generalize the Pythagorean Theorem?*

From the previous observations the students obtain the formula

$$AB^2 = AC^2 + BC^2 - 2 \cdot AB \cdot BC \cdot \cos \alpha.$$

Through the next question and the peer discussion that follows, the teacher lets emerge the need for a proof.

Question 8. *Can we be sure that the previous formula holds for all triangles?*

Step 4. Proving the law

Since the formula to prove is an extension of Pythagorean Theorem, the teacher proposes to prove it by trying to generalize a proof of this Theorem. In particular, he suggests starting from the recently studied proof, which showed in the left of Figure 4, based on Tangent-Secant Theorem.

Question 9. *Can the reasoning showed in the left side of Figure 4 be extended to non-right triangles?*

If necessary, the teacher suggests to use the Two Secant Theorem, since it naturally extends the Tangent-Secant Theorem, and guide students in understanding proof. In the right side of the Figure 4 we show a sketch of the proof, that will only be shown to pupils if they can't find out for themselves.

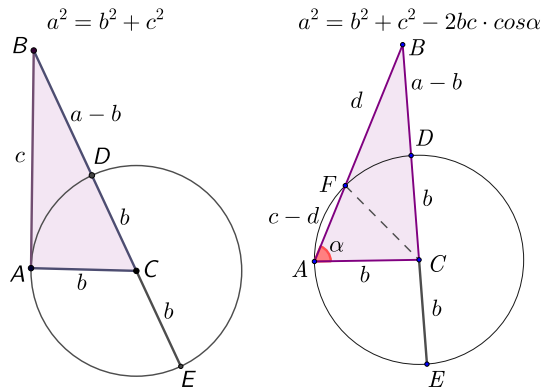


Figure 4. In the left we have $(a + b) : c = c : (a - b)$, i.e. the Pythagorean Theorem. In the right we have $(a + b) : c = d : (a - b)$ and $c - d = 2b \cdot \cos \alpha$, hence the Cosines Law.

Step 5. Solving the problem

At this point the teacher points out to the students that they have the tools to solve the initial problem and asks them to find the solution. He also invites them to reflect and discuss by asking the following questions.

Question 10. *What kind of problems can we solve by using the Cosines Law?*

Question 11. *Can we find another way to solve the aircraft problem?*

Question 12. *Can we find another way to prove the Cosines Law?*

4. Conclusions

The experimentation of the path that we have presented involved approximately 100 pupils from different geographical and socio-cultural backgrounds. At the end

of each unit the efficiency of the activity was assessed by considering the following aspects:

- students' ability to complete the research path by discovering and formulating the theorems that have been studied, and in particular the trigonometric relations;
- students' ability to use the acquired knowledge to solve applied problems.

The data collection was of an exploratory research; therefore, it could be considered only as a prelude to a more in-depth research plan aimed at comparing the results obtained using more traditional teaching approaches. In particular, we consider it appropriate to analyze the levels reached by students, according to van Hiele classification [19].

With this work we want also to highlight the importance of geometric reasoning in the didactic field, also due to the fact that in the last decade the teaching of Euclidean geometry in secondary schools seems to have lost its vigor. Yet, this part of mathematics is considered fundamental in the development of students' logical abilities by many authors [12].

While taking into account the limitations of our research, the initial results obtained by the students, if confirmed by further and more in-depth experiments, seem to indicate the achievement of the objectives of the experimentation. In particular, the visual/dynamic proofs seemed to be useful in the DL, where it is more difficult to maintain students' attention on long formal argumentations, whereas it would be more effective to stimulate them to think about geometric figures.

References

- [1] COMMON CORE STATE STANDARDS INITIATIVE: *Mathematics Standards for Mathematics*, Common Core State Standards Initiative, 2020, URL: <http://www.corestandards.org/Math/>.
- [2] A. T. DELICE, T. ROPER: *Implications of a comparative study for mathematics education in the English education system*, Teaching Mathematics and its applications 25.3 (2006), pp. 64–72, DOI: <https://doi.org/10.1145/1073204.1073229>.
- [3] J. DEWEY: *Experience and education*, New York: Macmillan, 1983, DOI: <https://doi.org/10.1007/978-3-642-56432-1>.
- [4] J. DREYE, D. M. LARSEN, M. D. HJELMBORG, C. MICHELSEN, M. MISFELDTAND: *Inquiry-based learning in mathematics education: important themes in the literature*. In *Research of Department of Mathematics and Science Education*, Stockholm: Stockholm University, 2016, DOI: <https://doi.org/10.1007/978-3-642-56432-1>.
- [5] ENGLISH DEPARTMENT FOR EDUCATION: *Mathematics GCSE subject content and assessment objectives. 2013*, London: Department for Education, 2013, URL: https://assets.publishing.service.gov.uk/government/uploads/system/uploads/attachment_data/file/254441/GCSE_mathematics_subject_content_and_assessment_objectives.pdf.

- [6] GOVERNMENT OF HUNGARY: *The Hungarian National Core Curriculum*, Magyar közlöny (official journal of Hungary): Government of Hungary, 2014, URL: https://ofi.oh.gov.hu/sites/default/files/ofipast/2014/04/NAT_2012_EN_final_2014marc14.pdf.
- [7] Á. GYÖRY, E. KÓNYAB: *Proving skills in geometry of secondary grammar school leavers specialized in mathematics*, *Annales Mathematicae et Informaticae* 50 (2019), pp. 217–236, DOI: <https://doi.org/10.33039/ami.2019.11.003>.
- [8] MINISTÈRE DE L'ÉDUCATION NATIONALE ET DE LA JEUNESSE: *Bulletin officiel n^o,30 du 26-7-2018 - Cycle 4. 2018*, Paris: Ministère de l'Éducation nationale et de la Jeunesse, 2018, URL: https://cache.media.eduscol.education.fr/file/30/62/8/ensel169_annexe3_985628.pdf.
- [9] MINISTÈRE DE L'ÉDUCATION NATIONALE ET DE LA JEUNESSE: *Programme de mathématiques de première générale*, Paris: Department for Education, 2019, URL: https://cache.media.education.gouv.fr/file/SP1-MEN-22-1-2019/16/8/spe632_annexe_1063168.pdf.
- [10] MINISTÈRE DE L'ÉDUCATION NATIONALE ET DE LA JEUNESSE: *Programme de mathématiques de seconde générale et technologique*, Paris: Department for Education, 2019, URL: https://cache.media.education.gouv.fr/file/SP1-MEN-22-1-2019/95/7/spe631_annexe_1062957.pdf.
- [11] MINISTERO DELL'ISTRUZIONE DELL'UNIVERSITÀ E DELLA RICERCA: *Schema di regolamento recante 'Indicazioni nazionali riguardanti gli obiettivi specifici di apprendimento concernenti le attività' e gli insegnamenti compresi nei piani degli studi previsti per i percorsi liceali*, Roma: Ministero dell'Istruzione dell'Università e della Ricerca, 2010, URL: <https://www.gazzettaufficiale.it/eli/id/2010/12/14/010G0232/sg>.
- [12] E. E. MOISE: *The meaning of Euclidean Geometry in school Mathematics*, *The Mathematics Teacher* 68.6 (1975), pp. 472–477.
- [13] ONTARIO MINISTRY OF EDUCATION: *The Ontario Curriculum Grades 11 and 12 Mathematics 2007*, Toronto: Queen's Printer for Ontario, 2007, URL: <http://www.edu.gov.on.ca/eng/curriculum/secondary/math1112currb.pdf>.
- [14] ONTARIO MINISTRY OF EDUCATION: *The Ontario Curriculum Grades 9 and 10 Mathematics 2005*, Toronto: Queen's Printer for Ontario, 2005, URL: <http://www.edu.gov.on.ca/eng/curriculum/secondary/math910curr.pdf>.
- [15] C. RASMUSSEN, K. MARRONGELLE, O. N. KWON, A. HODGE: *Four goal for instructors using inquiry-based learning*, *Not Am Math Soc.* 64.11 (2017), pp. 1308–1311.
- [16] J. A. ROSS, C. D. BRUCE, T. M. SIBBALD: *Sequencing computer-assisted learning of transformations of trigonometric functions*, *Teaching Mathematics and its applications* 30 (2011), pp. 120–137, DOI: <https://doi.org/10.1080/0020739X.2019.1565453>.
- [17] M. SANTOS-TRIGO: *An inquiry approach to construct instructional trajectories based on the use of digital technologies*, *Eurasia J Math Sci Technol Educ.* 4.4 (2008), pp. 347–357, DOI: <https://doi.org/10.12973/ejmste/75361>.
- [18] C. SZABÓ, C. BERECKZY-ZÁMBÓ, A. MUZSNAY, J. SZEIBERT: *Students' non-development in high school geometry*, *Annales Mathematicae et Informaticae* 52 (2020), pp. 309–319, DOI: <https://doi.org/10.33039/ami.2020.12.004>.
- [19] Z. USISKIN: *Van Hiele Levels and Achievement in Secondary School Geometry. CDASSG Project*, Chicago, Illinois: Chicago University, 1982.

Teaching numeral systems based on history in high school

Zoltán Matos^a, Eszter Kónya^b

^aElementary and Grammar School of University of Szeged

matos@freemail.hu

^bUniversity of Debrecen, Institute of Mathematics

eszter.konya@science.unideb.hu

Submitted: June 7, 2021

Accepted: August 19, 2021

Published online: August 23, 2021

Abstract

In the first decade after the turn of the millennium, previous doubts about the inclusion of the history of mathematics in education also received more attention. Several researchers point to the difficulties of teachers enthusiastic on the topic, to the research methodological difficulties of such studies, and the need to increase the number of empirical researches. In addition to increasing the amount of such empirical evidence, this paper seeks to contribute to the continuously developing answers to the basic questions (what and how?) of integrating the history of mathematics into public education in the recent decades by presenting a given topic, the teaching of numeral systems based on history, and the results of the related surveys. In the course of our research, we examined the question of whether the use of the history of mathematics as a tool, as opposed to teaching by focusing solely on routine tasks, helps to fix the curriculum into the long-term memory.

Keywords: History of mathematics, teaching of numeral systems

AMS Subject Classification: 97A30, 97F30

1. Introduction

From the beginning of the 1960s, more and more researchers have turned to the use of the history of mathematics in education. In 1972, at the second International

Congress on Mathematical Education (ICME) in Exeter, an international research group, The international study group on the relationship between History and Pedagogy of Mathematics (HPM), was set up on the subject, which has regularly organized international conferences and published since then.

Research on the use of the history of mathematics in education has gained new momentum since the 1990s, and more and more researchers have given arguments about incorporating the history of mathematics into education (e.g., [2]). Although Lefebvre warns in his article summarizing this topic that “all forms of categorization carry risky and arbitrary parts” ([7], p. 24), by the 1990s, intensified efforts could be noticed regarding answering the questions why and how and categorizing the answers. For example, Fried [3] grouped the 15 arguments about the history of mathematics given by Fauvel [1] around a total of three themes: (1) making mathematics more human; (2) making mathematics more interesting, understandable, and approachable; (3) to allow a deeper insight into problems and problem-solving.

Jankvist’s [5] article, written on this topic and quoted extensively, seeks to categorize the used methods (the how) and, separated from them, the arguments for use (the why). He classified the methods into three categories: the illumination approach, the modules approach, and the history-based approach. The first tries to spice up mathematics teaching mostly with isolated stories and anecdotes. This includes pictures that appear in the margins of textbooks as well as stories at the beginning or end of the chapters. The second category includes, for example, the study of a problem based on a topic in the history of mathematics and thus the way in which numeral systems are introduced in this paper. The third category means the presentation of the development, progression of the mathematical material covering a given part of the curriculum. Jankvist divided the answers to the question why into two categories. On the one hand, motivational factors that aid teaching and learning, on the other hand, tools that display the “soul” and development of mathematics, allowing the student not to see mathematics as a finished thing that “descended from heaven” in its perfection, axiomatically constructed.

In the first decade after the turn of the millennium, previous doubts about the inclusion of the history of mathematics in education also received more attention. Several researchers point to the difficulties of teachers enthusiastic on the topic (e.g., Fried [3]; Siu [9]), to the research methodological difficulties of such studies (e.g., Guillemette [4]), and the need to increase the number of empirical researches (e.g., Jankvist [6]).

In addition to increasing the amount of such empirical evidence, this paper seeks to contribute to the continuously developing answers to the basic questions (what and how?) of integrating the history of mathematics into public education in the recent decades by presenting a given topic, the teaching of numeral systems based on history, and the results of the related surveys.

2. The circumstances of the teaching experiment and the research question

The teaching experiment took place in Hungary, where public education is divided into two parts: an eight-grade primary school, after which students can continue their studies in a 3-year vocational school, a 4-year vocational grammar school teaching vocational and general subjects, finishing by graduation, or a 4-year grammar school (5-year for bilingual classes) teaching only general subjects, finishing by graduation, preparing for higher education. In high school, a class often has a profile of some kind, i.e., it studies the subjects related to that profile in a higher number of hours. During the admission procedure, most secondary schools select from the applying students on the basis of the admission points obtained on the uniform mathematics and Hungarian aptitude test worksheet.

The teaching experiment and then the related surveys took place in a grammar school in a big city in two consecutive school years in the ninth grade, in the first year in three groups of different profiles, and in the second year in two groups of the same profile.

During the experiment, the topic of numeral systems was taught to one group based on the history of mathematics, while to the control group, by focusing only on routine tasks, both studied for the same amount of time. This topic was not completely unknown to any of the groups. In primary school, all students had already learned about it in connection with the notion of sign-value and place-value notation when writing numbers.

In the course of our research, we examined the question of whether the use of the history of mathematics as a tool, as opposed to teaching by focusing solely on routine tasks, helps to fix the curriculum into the long-term memory.

Accordingly, we formulated the following principles for the curriculum constructed for the experimental groups:

- the inclusion of the history of mathematics in the classroom serves to teach the compulsory curriculum,
- as far as possible, the history of mathematics should be included in the teaching lesson as an integral part of the curriculum and not as a separate unit (for example, in the form of a student presentation or as mentioning interesting facts at the end of the lesson).

Following these principles, the lessons of the experimental group were about:

- E1 To introduce the topic, to raise awareness, to describe that there are peoples (such as the Piraha of Amazon) whose language lacks the concept of numbers. And the question was, although numbers are important, whether it matters how we describe them?
- E2 Introduction of the two characteristics of the numeral system we most commonly use, the decimal base and the place-value notation, by examples (see Figure 1).

a mi számírásunk

tízalapú helyértékes

$$\text{pl. } \begin{array}{c} 10^3 \quad 10^2 \quad 10^1 \quad 1 \\ \hline 1 \quad 9 \quad 3 \quad 8 \end{array} = 1 \cdot 1000 + 9 \cdot 100 + 3 \cdot 10 + 8 \cdot 1$$

egyiptomi számírás : 10-alapú, de nem helyértékes

1-1		34	
III		23	
II-10			+ =

12-31	1-31		
	2-62		
	4-124	} → 372	
	8-248		

14-15	1-15		
	2-30	} ⇒ 210	
	4-60		
	8-120		

Figure 1. Our numeral system and the Egyptian numeral system (from the notebook of one of the students).

E3 A counterexample to one of these qualities: the ancient Egyptian numeral system is decimal but not place-valued. After a few examples, a description of addition and multiplication performed on integers. In the meantime, we highlight the following two problems: in theory, infinitely many different characters are needed to describe numbers, and sometimes a large number of characters are needed to describe numbers that are used in everyday life. The teacher examples presented were always followed by shorter or longer student work, so students had to independently multiply 14 by 15 in an Egyptian way (see Figure 1, “a mi számírásunk – our numeral system”, “egyiptomi számírás – Egyptian numeral system”, “tíz alapú – decimal”. “helyértékes – place-value”).

E4 Example of the non-decimal and non-place-value numeral system: Roman

numerals. This type of numeral system, while solving the previous two problems, raises another issue of the basic operations. Although students had already learned Roman numerals in the lower grades of elementary school, the number of knots and the logic of the system structure were revived.

E5 An example of the non-decimal but place-value numeral system. Noticing that the number of characters usable here is finite, and depending on the base number of the numeral system, a given number can be written in a longer or shorter form. Converting back and forth between decimal and other numeral systems having other bases. Adding in the non-decimal numeral system.

E6 Traces of numeral systems having other bases in our lives (watch, angles, notation of numbers in foreign languages.)

During the lesson, the teacher tried to guide the students to the key points by questions (e.g., How many characters are needed in the Egyptian numeral system to describe the number 798? Answer: 24.) or to bring to the surface the students' existing knowledge on this topic. (Are they familiar with any numeral system that is neither decimal nor place-value? Answer: Roman numeral system.) Because of the time constraints of the lessons, the teacher did not make the students rediscover the customs of other ages but introduced them (e.g., How did Egyptians multiply two integers?), thus maintaining their traditional "source of knowledge" role. The teacher in the control group viewed their own role as a teacher similarly.

In contrast, the lesson of the control group consisted of the following main sections.

K1 The two basic features of our numeral system are decimal and place-value, recalling the concepts learned in primary school (place value, sign value). (K1=E2)

K2 Examining what if the place values are not the powers of 10 but those of other positive integers other than 1. Converting back and forth between decimal and numeral systems having other bases. (See Figure 2) (This part corresponded to E5.)

$$4023_6 = 1 \cdot 6^3 + 2 \cdot 6^2 + 3 \cdot 6^1 = 879_{10}$$

$$1301_{10} \rightarrow 5032_6$$

$$1301_{10} = 20201_5$$

$$\begin{array}{r} 1301 \\ 5 \overline{) 1301} \\ \underline{10} \\ 30 \\ \underline{25} \\ 50 \\ \underline{50} \\ 0 \end{array}$$

$$\begin{array}{r} 2602 \\ 5 \overline{) 1301} \\ \underline{10} \\ 30 \\ \underline{25} \\ 50 \\ \underline{50} \\ 0 \end{array}$$

Figure 2. Converting between numeral system (from the notebook of a student from the control group).

3. Findings

The two teaching methods, followed by the related survey, were conducted in three groups with different profiles in the first year. Two of them had an experimental role, integrating historical elements, and the third was a control group. Among the experimental groups, group A consisted of students with a similar profile and similar ability to the control group; the other experimental group B consisted of students significantly different from the control group, with a weaker ability in mathematics. The latter group studied mathematics in French as a French bilingual group. The difference in their math skills is indicated by the average score on the high school admission. The data for the groups are given in Table 1. The lessons were held in a good atmosphere in all three groups; most students understood the curriculum, and no students indicated a problem during the homework check of the next lesson.

Table 1. Groups in the first experiment.

	Experimental group A	Experimental group B	Control group
Orientation of the group	Physics	French bilingual	Biology-Physics-Chemistry
Number of students in the group	14	13	14
Number of lessons in Math	4 per week	3 per week	4 per week
Average points on high school Math admission (50 points maximum)	37.00	31.00	35.56

The final tests written two to three weeks after the experimental lesson included a conversion task between numeral systems (“in both directions”) for all three groups. In the case of these tests, there was no significant difference regarding the results of the students, at least 80% of the students in all three groups solved the task correctly.

However, two months after the experimental classes (including the two-week winter break in the meantime), at the beginning of a math class, students from all three groups were “unexpectedly” asked to answer the following two questions anonymously:

1. *Write the five-digit form of 473 .*
2. *Write the decimal form of 431_6 .*

The following statistics were compiled on the students’ answers to the two tasks (Table 2). For Question 1, students in both experimental groups had more than 20% higher rates of correct answers than students in the control group. In Question 2, the proportion of correct respondents was 50% higher among the students

of group A and 25% higher among the students of group B than in the control group.

Table 2. Summary of the students' answers (first experiment).

		Experimental group A	Experimental group B	Control group
From base 10 to base 5	Number of correct answers	13	12	10
	Number of incorrect answers	1	1	4
From base 6 to base 10	Number of correct answers	11	6	3
	Number of incorrect answers	3	7	11

In the second school year, we repeated the above experiment in groups with the same profile as the first year's control group (biology-chemistry-physics). In the experimental group, the average of the results of the 8th-grade math ability test was 35.06, while in the control group, it was 36.53, i.e., there was no significant difference between the math skills of the two groups.

As in the previous year, there was no problem either in the lessons or with the homework: More than 80% of the students solved the conversion tasks in a final survey written two weeks after the experimental lesson. As in the previous school year, both groups were "unexpectedly" given the next task two months after the experimental lesson at the beginning of a math class, which had to be answered anonymously:

1. Write the binary form of 345.
2. Write the decimal form of 1221_3 .

The results of the two groups are summarized in Table 3. Question 1 had a 35%, Question 2 has 55% higher rate of correct answers among students in the experimental group than in the control group.

Table 3. Summary of the students' answers (second experiment).

		Experimental group	Control group
From base 10 to base 2	Number of correct answers	12	5
	Number of incorrect answers	5	9
From base 3 to base 10	Number of correct answers	13	3
	Number of incorrect answers	4	11

The percentage distribution of the summary of the responses of the students participating in the experiment during the two academic years two months after the lesson is shown in Table 4 by task type.

Table 4. Summary of the responses of the students participating in the two experiments.

		Experimental groups (44 students)	Control group (28 students)
Converting from decimal numeral system	Proportion of correct answers	84.10%	53.57%
	Proportion of incorrect answers	15.90%	46.43%
Converting to decimal numeral system	Proportion of correct answers	68.18%	21.43%
	Proportion of incorrect answers	31.82%	78.57%

4. Conclusion

At the beginning of our research, we sought to answer the question of whether teaching numeral systems in a historical framework is more helpful to fix information into the long-term memory than teaching focusing solely on solving routine tasks.

In response, we can state that although the control and experimental groups spent the same amount of time studying numeral systems in both school years, students who did not merely practice the same type of task again and again but learned about the topic more comprehensively, embedded in history, the tests written two months later definitely showed better results. It is in line with Revuz's idea, who stated as early as in the 1970s, "In the initial period of teaching, the most serious, almost irreparable damage can be done by replacing the true understanding with the mechanical practice of what has been learned". ([8], p. 14.)

Examining the reasons, the question should be asked: what role did the history of mathematics play in the classes of the experimental groups?

- It provided a framework for the lesson that roughly followed the stations through which humanity came to the decimal, place-value numeral system used today.
- It provided a logical transition between the different numeral systems (e.g., the Roman numeral system addresses the problem of the number of characters required to describe numbers in the Egyptian numeral system).
- Mobilized the student's pre-existing knowledge (e.g., Roman numerals).

- In several cases, it gave a counterexample (e.g., decimal but not place-value system).
- It made numeral systems interesting and related to everyday life, while the other group mastered routine procedures that were impractical to life.
- It highlighted the fact that it is not the base number but the place-value notation that is important when performing basic operations (e.g., in the Egyptian decimal numeral system addition was completely different, but in the three-based place-value notation the principle of written addition does not differ from the decimal system), so this is the key concept of this topic. This was also manifested in the fact that during the surveys, the place value chart appears in the work of the students of the experimental groups, even in the case of the incorrect respondents (as shown in Figure 3 of the notebook of one of the students of the second-year experimental group).

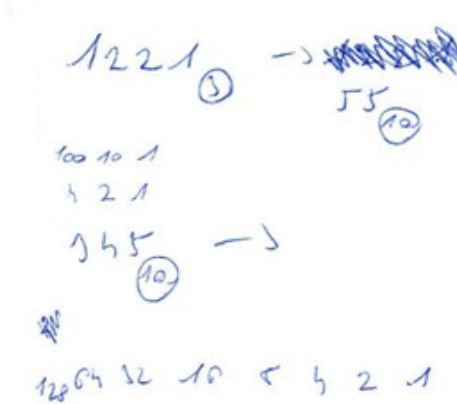


Figure 3. Appearance of the place-value chart in the incorrect answer.

The occurrence of all these may explain why the students of the experimental groups were able to recall what they had learned even after two months.

References

- [1] J. FAUVEL: *Using History in mathematics education*, For the Learning of Mathematics 11.2 (1991), pp. 3–6.
- [2] J. FAUVEL, J. VAN MAANEN (EDS.): *History in Mathematics Education - The ICMI Study*, Dordrecht: Springer, 2002, doi: <https://doi.org/10.1007/0-306-47220-1>.

-
- [3] M. N. FRIED: *Can Mathematics Education and History of Mathematics Coexist?*, Science & Education 10 (2001), pp. 391–408, DOI: <https://doi.org/10.1023/A:1011205014608>.
- [4] D. GUILLEMETTE: *L'histoire dans l'enseignement des mathématiques: sur la méthodologie de recherche*, Petit x 86 (2011), pp. 5–26.
- [5] U. T. JANKVIST: *A categorization of the “whys” and “hows” of using history in mathematics education*, Educational Studies in Mathematics 71.3 (2009), pp. 235–261, DOI: <https://doi.org/10.1007/s10649-008-9174-9>.
- [6] U. T. JANKVIST: *On empirical research in the field of using history in mathematics education*, Revista Latinoamericana de Investigación en Matemática Educativa 12.1 (2009), pp. 67–101.
- [7] J. LEFEBVRE: *Utilisation de l'histoire dans l'enseignement des mathématiques*, Bulletin de l'AMQ 33.3 (1993), pp. 22–27.
- [8] A. REVUZ: *Modern matematika - élő matematika*, Budapest: Gondolat, 1973.
- [9] M.-K. SIU: *No, I don't use history of mathematics in my class. Why?*, in: Proceedings HPM2004 & ESU4 (revised edition), ed. by F. FURINGHETTI, S. KAIJSER, C. TZANAKIS, Uppsala: Uppsala Universitet, 2007, pp. 368–382.

On some methods of calculating the integrals of trigonometric rational functions

Michał Róžański^a, Barbara Smoleń-Duda^a, Roman Wituła^a,
Marcin Jochlik^{b*}, Adrian Smuda^{b†}

^aDepartment of Mathematics, Silesian University of Technology,
Kaszubska 23, 44-100 Gliwice, Poland,
michal.rozanski@polsl.pl, barbara.smolen-duda@polsl.pl, roman.witula@polsl.pl

^bFaculty of Applied Mathematics, Silesian University of Technology,
Kaszubska 23, 44-100 Gliwice, Poland,
marcin.jochlik@onet.pl, adrians91@tlen.pl

Submitted: May 31, 2021

Accepted: October 13, 2021

Published online: October 24, 2021

Abstract

The paper presents original methods of calculating integrals of selected trigonometric rational functions.

Keywords: Integrals of trigonometric rational functions, Darboux property, integration by parts

AMS Subject Classification: 26A36, 26A42

1. Introduction

The aim of this work is to present an “original” methods of determining the integrals of the form

$$\int \frac{p \sin^2 x + q \sin x \cos x + r \cos^2 x}{(a \sin x + b \cos x)^n} dx, \quad (1.1)$$

where $p, q, r, a, b \in \mathbb{R}$, $n \in \mathbb{N}$. Presented methods are useful for manual as well as machine symbolic calculations.

*Marcin Jochlik is a first-year undergraduate student.

†Adrian Smuda is a first-year graduate student.

In the era of omnipotent and, above all, commonly available symbolic calculations, including integration, this article may seem archaic. But the reason for creating this paper is neither complicated nor artificial. The article was initiated during classes in mathematical analysis in the second semester of undergraduate studies in mathematics, which were led by the third author in the March of this year. The way from a simple task – a special case of the integral described by (1.1) – to creative and gripping generalizations turned out to be easy and very fast. It resulted in the presented article that is an effect of pure creative passion.

Let us emphasize that from the very beginning we were looking for alternative sources of the presented methods and computational techniques [1–5]. Only in [5] a solution was found, rather a run-of-the-mill solution, for a certain special case of integral (1.1). Besides, we did not come across any at least promising traces of similar or comparable methods. Therefore, we can confidently say that the proving methods and technical tricks presented in the paper are original. It is worth pointing out that the obtained formulae, e.g. (5.10) and (5.12), may be used in both numerical and symbolic applications.

Notation. We will denote by $\alpha \times (k) \pm \beta \times (l)$, over all numbered identities (k) , (l) in this paper, the following operation: identity (k) is multiplied by α , identity (l) is multiplied by β and then the obtained identities are summed (subtracted, respectively).

First, we describe the method for the simple case of $n = 1$ basing on a 3-step reduction in computation.

2. The first step of the method

We reduce the numerator in (1.1) to

$$p \sin^2 x + q \sin x \cos x + r \cos^2 x = \alpha f(x) + \beta g(x) + \gamma h(x), \quad (2.1)$$

where

$$f(x) = (M(x))^2, \quad g(x) = M(x)M'(x), \quad h(x) = \sin x \cos x,$$

and $M(x)$ denotes denominator in (1.1) in the case $n = 1$, i.e.

$$M(x) := a \sin x + b \cos x. \quad (2.2)$$

By solving the appropriate system of equations (created by comparing the coefficients at $\cos^2 x$, $\sin^2 x$ and $\sin x \cos x$)

$$\begin{cases} \alpha b^2 + \beta ab = r, \\ \alpha a^2 - \beta ab = p, \\ (a^2 - b^2)\beta + 2ab\alpha + \gamma = q, \end{cases}$$

we get

$$\alpha = \frac{p+r}{a^2+b^2}, \quad \beta = \frac{-b^2p+a^2r}{ab(a^2+b^2)}, \quad \gamma = q - \frac{b}{a}p - \frac{a}{b}r.$$

Then integral (1.1) takes the form

$$\begin{aligned} \int \frac{p \sin^2 x + q \sin x \cos x + r \cos^2 x}{a \sin x + b \cos x} dx \\ = \alpha \int M(x) dx + \beta \int M'(x) dx + \gamma \int \frac{\sin x \cos x}{a \sin x + b \cos x} dx. \end{aligned}$$

We only need to calculate the integral

$$\int \frac{\sin x \cos x}{a \sin x + b \cos x} dx.$$

3. The second step of the method

Integrating by parts in two ways, we find

$$\begin{aligned} \int \frac{\sin x \cos x}{a \sin x + b \cos x} dx &= \int (\sin x)' \frac{\sin x}{a \sin x + b \cos x} dx \\ &= \frac{\sin^2 x}{a \sin x + b \cos x} - \int \frac{b \sin x}{(a \sin x + b \cos x)^2} dx \end{aligned} \quad (3.1)$$

and

$$\begin{aligned} \int \frac{\sin x \cos x}{a \sin x + b \cos x} dx &= \int (-\cos x)' \frac{\cos x}{a \sin x + b \cos x} dx \\ &= \frac{-\cos^2 x}{a \sin x + b \cos x} - \int \frac{a \cos x}{(a \sin x + b \cos x)^2} dx. \end{aligned} \quad (3.2)$$

Moreover, by $\frac{a^2}{a^2+b^2} \times (3.1) + \frac{b^2}{a^2+b^2} \times (3.2)$, we obtain

$$\begin{aligned} \int \frac{\sin x \cos x}{a \sin x + b \cos x} dx \\ = \frac{1}{a^2 + b^2} \cdot \frac{a^2 \sin^2 x - b^2 \cos^2 x}{a \sin x + b \cos x} - \frac{ab}{a^2 + b^2} \int \frac{a \sin x + b \cos x}{(a \sin x + b \cos x)^2} dx \\ = \frac{1}{a^2 + b^2} (a \sin x - b \cos x) - \frac{ab}{a^2 + b^2} \int \frac{dx}{a \sin x + b \cos x}. \end{aligned} \quad (3.3)$$

Moreover, from (3.1) and (3.2), we get

$$\int \frac{\sin x \cos x}{a \sin x + b \cos x} dx = \frac{v \sin^2 x - u \cos^2 x}{a \sin x + b \cos x} - \int \frac{au \cos x + bv \sin x}{(a \sin x + b \cos x)^2} dx \quad (3.4)$$

whenever $u, v \in \mathbb{R}$, $u + v = 1$.

4. The third step of the method (supplementary reminder)

We still have to determine the integral $\int \frac{dx}{a \sin x + b \cos x}$. We calculate it as follows

$$\begin{aligned} \int \frac{dx}{a \sin x + b \cos x} &= \int \frac{dx}{\sqrt{a^2 + b^2} \sin(x + \varphi)} = \left| \begin{array}{l} \text{where} \\ \cos \varphi = \frac{a}{\sqrt{a^2 + b^2}} \\ \sin \varphi = \frac{b}{\sqrt{a^2 + b^2}} \end{array} \right| \\ &= \frac{1}{\sqrt{a^2 + b^2}} \int \frac{dx}{2 \cos^2 \frac{x+\varphi}{2} \tan \frac{x+\varphi}{2}} = \frac{1}{\sqrt{a^2 + b^2}} \ln \left| \tan \frac{x + \varphi}{2} \right| + C, \end{aligned}$$

where, after applying the identity

$$\begin{aligned} \tan \frac{x + \varphi}{2} &= \frac{\cos \frac{\varphi}{2} \sin \frac{x}{2} + \sin \frac{\varphi}{2} \cos \frac{x}{2}}{\cos \frac{\varphi}{2} \cos \frac{x}{2} - \sin \frac{\varphi}{2} \sin \frac{x}{2}} = \frac{2 \cos^2 \frac{\varphi}{2} \sin \frac{x}{2} + 2 \cos \frac{\varphi}{2} \sin \frac{\varphi}{2} \cos \frac{x}{2}}{2 \cos^2 \frac{\varphi}{2} \cos \frac{x}{2} - 2 \cos \frac{\varphi}{2} \sin \frac{\varphi}{2} \sin \frac{x}{2}} \\ &= \frac{(1 + \cos \varphi) \sin \frac{x}{2} + \sin \varphi \cos \frac{x}{2}}{(1 + \cos \varphi) \cos \frac{x}{2} - \sin \varphi \sin \frac{x}{2}} = \frac{(a + \sqrt{a^2 + b^2}) \sin \frac{x}{2} + b \cos \frac{x}{2}}{(a + \sqrt{a^2 + b^2}) \cos \frac{x}{2} - b \sin \frac{x}{2}}, \end{aligned}$$

we get

$$\int \frac{dx}{a \sin x + b \cos x} = \frac{1}{\sqrt{a^2 + b^2}} \ln \left| \frac{(a + \sqrt{a^2 + b^2}) \sin \frac{x}{2} + b \cos \frac{x}{2}}{(a + \sqrt{a^2 + b^2}) \cos \frac{x}{2} - b \sin \frac{x}{2}} \right| + C.$$

Another method of calculating the discussed integral, without using the half-angle formula, is presented below

$$\begin{aligned} \int \frac{dx}{a \sin x + b \cos x} &= \int \frac{a \sin x - b \cos x}{a^2 \sin^2 x - b^2 \cos^2 x} dx \\ &= \int \frac{a \sin x}{a^2 - (a^2 + b^2) \cos^2 x} dx + \int \frac{b \cos x}{b^2 - (a^2 + b^2) \sin^2 x} dx \\ &= \frac{1}{2} \int \left(\frac{\sin x}{a - \sqrt{a^2 + b^2} \cos x} + \frac{\sin x}{a + \sqrt{a^2 + b^2} \cos x} \right) dx \\ &\quad + \frac{1}{2} \int \left(\frac{\cos x}{b - \sqrt{a^2 + b^2} \sin x} + \frac{\cos x}{b + \sqrt{a^2 + b^2} \sin x} \right) dx \\ &= \frac{1}{2\sqrt{a^2 + b^2}} \left(\ln \left| \frac{a - \sqrt{a^2 + b^2} \cos x}{a + \sqrt{a^2 + b^2} \cos x} \right| + \ln \left| \frac{b + \sqrt{a^2 + b^2} \sin x}{b - \sqrt{a^2 + b^2} \sin x} \right| \right) + C. \end{aligned}$$

At the end of this section, we present the conventional method of calculating $\int \frac{dx}{a \sin x + b \cos x}$ using the Weierstrass substitution. We consider it in a more general case with an additional constant in the denominator.

$$\begin{aligned}
 & \int \frac{dx}{a \sin x + b \cos x + c} \\
 &= \int \frac{dx}{a(2 \sin \frac{x}{2} \cos \frac{x}{2}) + b(\cos^2 \frac{x}{2} - \sin^2 \frac{x}{2}) + c(\cos^2 \frac{x}{2} + \sin^2 \frac{x}{2})} \\
 &= \int \frac{dx}{((c-b) \tan^2 \frac{x}{2} + 2a \tan \frac{x}{2} + b+c) \cos^2 \frac{x}{2}} = \left| \begin{array}{l} \text{substitution} \\ t = \tan \frac{x}{2} \end{array} \right| \\
 &= \int \frac{2 dt}{(c-b)t^2 + 2at + b+c} = \left| \begin{array}{l} \text{we only give} \\ \text{the final result} \end{array} \right| \\
 &= \begin{cases} \frac{1}{\sqrt{a^2 + b^2 - c^2}} \ln \frac{\sqrt{a^2 + b^2 - c^2} - a - (c-b) \tan \frac{x}{2}}{\sqrt{a^2 + b^2 - c^2} + a + (c-b) \tan \frac{x}{2}}, & \text{when } a^2 + b^2 > c^2, \\ \frac{2}{\sqrt{c^2 - a^2 - b^2}} \arctan \frac{a + (c-b) \tan \frac{x}{2}}{\sqrt{c^2 - a^2 - b^2}}, & \text{when } a^2 + b^2 < c^2, \\ -\frac{1}{a + (c-b) \tan \frac{x}{2}}, & \text{when } a^2 + b^2 = c^2. \end{cases}
 \end{aligned}$$

5. A generalization due to the power of the denominator

In the case of the integrals

$$\int \frac{\sin x \cos x}{(a \sin x + b \cos x)^k} dx, \quad k \in \mathbb{N}, k \geq 2, \tag{5.1}$$

our attempts of the finding of a generalization of formula (3.3) did not provide desired results. Following the discussed methods, we generated only

$$\int \frac{b \sin x}{(a \sin x + b \cos x)^3} dx = \frac{\sin^2 x}{2(a \sin x + b \cos x)^2} + C, \tag{5.2}$$

$$\int \frac{a \cos x}{(a \sin x + b \cos x)^3} dx = \frac{-\cos^2 x}{2(a \sin x + b \cos x)^2} + C. \tag{5.3}$$

Hence, by $\frac{a}{b} \times (5.2) + \frac{b}{a} \times (5.3)$ we get

$$\int \frac{dx}{(a \sin x + b \cos x)^2} = \frac{\frac{a}{b} \sin^2 x - \frac{b}{a} \cos^2 x}{2(a \sin x + b \cos x)^2} + C = \frac{1}{2ab} \cdot \frac{a \sin x - b \cos x}{a \sin x + b \cos x} + C \tag{5.4}$$

and generally

$$\int \frac{au \cos x + bv \sin x}{(a \sin x + b \cos x)^3} dx = \frac{v \sin^2 x - u \cos^2 x}{2(a \sin x + b \cos x)^2} + C \tag{5.5}$$

for any $u, v \in \mathbb{R}$. So, we got certain analogues of formulae (3.3) and (3.4). Formulae (5.2) and (5.3) can be easily verified directly and they prompted us to calculate the following derivatives (and it was a bull's-eye)

$$\begin{aligned} \left(\frac{\cos^2 x}{(a \sin x + b \cos x)^k} \right)' &= \frac{-2a \cos x - (k-2) \cos^2 x (a \cos x - b \sin x)}{(a \sin x + b \cos x)^{k+1}}, \\ \left(\frac{\sin^2 x}{(a \sin x + b \cos x)^k} \right)' &= \frac{2b \sin x - (k-2) \sin^2 x (a \cos x - b \sin x)}{(a \sin x + b \cos x)^{k+1}}. \end{aligned}$$

After integrating the above identities, we get

$$\begin{aligned} &\frac{\cos^2 x}{(a \sin x + b \cos x)^k} \\ &= - \int \frac{2a \cos x}{(a \sin x + b \cos x)^{k+1}} dx - (k-2) \int \cos^2 x \left(\frac{-\frac{1}{k}}{(a \sin x + b \cos x)^k} \right)' dx \\ &\text{(integrating by parts)} \\ &= - \int \frac{2a \cos x}{(a \sin x + b \cos x)^{k+1}} dx + \frac{k-2}{k} \cdot \frac{\cos^2 x}{(a \sin x + b \cos x)^k} \\ &\quad + \frac{2(k-2)}{k} \int \frac{\sin x \cos x}{(a \sin x + b \cos x)^k} dx, \end{aligned}$$

which implies

$$\begin{aligned} &(k-2) \int \frac{\sin x \cos x}{(a \sin x + b \cos x)^k} dx \\ &= \frac{\cos^2 x}{(a \sin x + b \cos x)^k} + k \int \frac{a \cos x}{(a \sin x + b \cos x)^{k+1}} dx \end{aligned} \quad (5.6)$$

for $k \in \mathbb{N}$, $k \geq 3$. Similarly

$$\begin{aligned} &\frac{\sin^2 x}{(a \sin x + b \cos x)^k} \\ &= \int \frac{2b \sin x}{(a \sin x + b \cos x)^{k+1}} dx - (k-2) \int \sin^2 x \left(\frac{-\frac{1}{k}}{(a \sin x + b \cos x)^k} \right)' dx \\ &= \int \frac{2b \sin x}{(a \sin x + b \cos x)^{k+1}} dx + \frac{k-2}{k} \cdot \frac{\sin^2 x}{(a \sin x + b \cos x)^k} \\ &\quad - \frac{2(k-2)}{k} \int \frac{\sin x \cos x}{(a \sin x + b \cos x)^k} dx, \end{aligned}$$

which implies

$$(k-2) \int \frac{\sin x \cos x}{(a \sin x + b \cos x)^k} dx$$

$$= \frac{-\sin^2 x}{(a \sin x + b \cos x)^k} + k \int \frac{b \sin x}{(a \sin x + b \cos x)^{k+1}} dx \tag{5.7}$$

for $k \in \mathbb{N}, k \geq 3$. Additionally, by $\frac{b^2}{a^2+b^2} \times (5.6) + \frac{a^2}{a^2+b^2} \times (5.7)$, we obtain

$$(k-2) \int \frac{\sin x \cos x}{(a \sin x + b \cos x)^k} dx = \frac{1}{a^2 + b^2} \cdot \frac{b \cos x - a \sin x}{(a \sin x + b \cos x)^{k-1}} + \frac{abk}{a^2 + b^2} \int \frac{1}{(a \sin x + b \cos x)^k} dx \tag{5.8}$$

for $k \in \mathbb{N}, k \geq 3$. Moreover, by $u \times (5.6) + v \times (5.7)$ we get

$$(k-2) \int \frac{\sin x \cos x}{(a \sin x + b \cos x)^k} dx = \frac{u \cos^2 x - v \sin^2 x}{(a \sin x + b \cos x)^k} + k \int \frac{au \cos x + bv \sin x}{(a \sin x + b \cos x)^{k+1}} dx \tag{5.9}$$

whenever $u, v \in \mathbb{R}, u + v = 1$, and $k \in \mathbb{N}, k \geq 3$. For $k = 1$, from (5.8) and (5.9), we obtain (3.3) and (3.4), respectively. Furthermore, for any $k \in \mathbb{N}, k \geq 3$, formulae (5.8) and (5.9) are generalizations of formulae (3.3) and (3.4), respectively, for any $k \in \mathbb{N}, k \geq 3$. Let us recall that the case $k = 2$ is not covered by these formulae and it is described by identities (5.4) and (5.5) - we can obtain them also from (5.8) and (5.9) after substitution $k = 2$. In this way, we also obtain a solution to problem (5.1), previously unsuccessfully investigated with the method from Section 3.

Corollary 5.1. *Suppose $ab > 0$ and let $x_0 \in (-\frac{\pi}{2}, 0)$ be such that $\tan x_0 = -\frac{b}{a}$. Then for each $\varphi \in (0, \frac{\pi}{2})$, there is $x(\varphi) \in (x_0, 0)$ that satisfies the condition*

$$\int_{x(\varphi)}^0 \frac{\sin x \cos x}{(a \sin x + b \cos x)^k} dx = - \int_0^\varphi \frac{\sin x \cos x}{(a \sin x + b \cos x)^k} dx.$$

Hence, based on formula (5.8), we get the formulae

$$\int_{x(\varphi)}^\varphi \frac{dx}{(a \sin x + b \cos x)^k} = \frac{1}{k} \cdot \frac{1}{ab} \cdot \frac{a \sin x - b \cos x}{(a \sin x + b \cos x)^{k-1}} \Bigg|_{x(\varphi)}^\varphi,$$

$$\int_{x(\varphi)}^\varphi \frac{au \cos x + bv \sin x}{(a \sin x + b \cos x)^{k+1}} dx = -\frac{1}{k} \cdot \frac{u \cos^2 x - v \sin^2 x}{(a \sin x + b \cos x)^k} \Bigg|_{x(\varphi)}^\varphi,$$

whenever $u, v \in \mathbb{R}, u + v = 1$, and $k \in \mathbb{N}, k \geq 3$.

Proof. Note that

$$\frac{\sin x \cos x}{(a \sin x + b \cos x)^k} < 0$$

in the interval $(x_0, 0)$ and

$$\int_{x_0}^0 \frac{\sin x \cos x}{(a \sin x + b \cos x)^k} dx = -\infty.$$

The function

$$(x_0, 0] \ni x \mapsto \int_x^0 \frac{\sin \tau \cos \tau}{(a \sin \tau + b \cos \tau)^k} d\tau$$

is continuous. It remains to use the Darboux property. \square

Remark 5.2. In according to identity (5.8), we propose to derive a recurrent identity for the integrals

$$I_k = \int \frac{dx}{(a \sin x + b \cos x)^k}, \quad k \in \mathbb{N}.$$

So, we have

$$\begin{aligned} (a^2 + b^2)I_k &= \int \frac{(a \sin x + b \cos x)^2 + (a \cos x - b \sin x)^2}{(a \sin x + b \cos x)^k} dx \\ &= I_{k-2} + \int (a \cos x - b \sin x) \cdot \left(\frac{-\frac{1}{k-1}}{(a \sin x + b \cos x)^{k-1}} \right)' dx \\ &= I_{k-2} + \frac{1}{k-1} \cdot \frac{b \sin x - a \cos x}{(a \sin x + b \cos x)^{k-1}} - \frac{1}{k-1} I_{k-2} \\ &= \frac{k-2}{k-1} I_{k-2} + \frac{1}{k-1} \cdot \frac{b \sin x - a \cos x}{(a \sin x + b \cos x)^{k-1}}. \end{aligned} \quad (5.10)$$

Hence, for example, by (5.4) we obtain

$$3(a^2 + b^2)I_4 = \frac{1}{ab} \cdot \frac{a \sin x - b \cos x}{a \sin x + b \cos x} + \frac{b \sin x - a \cos x}{(a \sin x + b \cos x)^3} + C.$$

Remark 5.3. In according to identities (5.8) and (5.4), it is worth pointing out that

$$\begin{aligned} &\int \frac{\sin x \cos x}{(a \sin x + b \cos x)^2} dx \\ &= \frac{a^2 - b^2}{(a^2 + b^2)^2} \ln |a \sin x + b \cos x| \\ &\quad + \frac{b}{a^2 + b^2} \cdot \frac{\cos x}{a \sin x + b \cos x} + \frac{2abx}{(a^2 + b^2)^2} + C, \end{aligned} \quad (5.11)$$

where the calculations were done using the following decomposition in an ingenious way

$$\int \frac{\sin x \cos x}{(a \sin x + b \cos x)^2} dx = \int \frac{\tan x}{(a \tan x + b)^2 (\tan^2 x + 1)} d(\tan x)$$

(where $u = \tan x$)

$$= \int \frac{u}{(au + b)^2(u^2 + 1)} du = \int \left(\frac{\alpha}{au + b} + \frac{\beta}{(au + b)^2} + \frac{\gamma u + \delta}{u^2 + 1} \right) du$$

(after an observation of the obtained integrals)

$$= A \ln |a \sin x + b \cos x| + B \frac{\cos x}{a \sin x + b \cos x} + Dx + C$$

(we have only 3 unknown constants A, B, D), which, after differentiation, easily implies formula (5.11). Therefore, from (5.4) and (5.11) results that a simple functional identity, as for example formula (5.8), between the integrals

$$\int \frac{dx}{(a \sin x + b \cos x)^2} \quad \text{and} \quad \int \frac{\sin x \cos x}{(a \sin x + b \cos x)^2} dx$$

does not exist. But there exists such a connection between the discussed integral

$$\int \frac{\sin x \cos x}{(a \sin x + b \cos x)^2} dx$$

and the other surprising integral

$$\int \frac{a \cos x + b \sin x}{a \sin x + b \cos x} dx.$$

Based on the identity

$$\begin{aligned} (b^2 + a^2) \sin x \cos x + ab &= (b^2 + a^2) \sin x \cos x + ab(\sin^2 x + \cos^2 x) \\ &= b \sin x(b \cos x + a \sin x) + a \cos x(a \sin x + b \cos x) \\ &= (a \sin x + b \cos x)(a \cos x + b \sin x) \end{aligned}$$

we get

$$\begin{aligned} &\int \frac{\sin x \cos x}{(a \sin x + b \cos x)^2} dx \\ &= \frac{1}{a^2 + b^2} \int \frac{(b^2 + a^2) \sin x \cos x + ab}{(a \sin x + b \cos x)^2} dx - \frac{ab}{a^2 + b^2} \int \frac{dx}{(a \sin x + b \cos x)^2} \\ &= \frac{1}{a^2 + b^2} \int \frac{a \cos x + b \sin x}{a \sin x + b \cos x} dx - \frac{ab}{a^2 + b^2} \int \frac{dx}{(a \sin x + b \cos x)^2} \\ &\stackrel{(5.4)}{=} \frac{1}{a^2 + b^2} \int \frac{a \cos x + b \sin x}{a \sin x + b \cos x} dx - \frac{1}{2(a^2 + b^2)} \cdot \frac{a \sin x - b \cos x}{a \sin x + b \cos x}. \end{aligned}$$

Remark 5.4. Using formulae (2.1), (5.7) and (5.9) we obtain a generalization of the identities presented in Section 2

$$\int \frac{p \sin^2 x + q \sin x \cos x + r \cos^2 x}{(a \sin x + b \cos x)^n} dx$$

$$\begin{aligned}
& \stackrel{n \geq 2}{=} \alpha I_{n-2} + \beta \int \frac{M'(x)}{M^n(x)} dx + \gamma \int \frac{\sin x \cos x}{M^n(x)} dx \\
& \stackrel{n \geq 3}{=} \alpha I_{n-2} - \frac{\beta}{n-1} \cdot \frac{1}{M^{n-1}(x)} + \frac{\gamma(b \cos x - a \sin x)}{(n-2)(a^2 + b^2)M^{n-1}(x)} + \frac{abn\gamma}{(n-2)(a^2 + b^2)} I_n \\
& = \left(\alpha + \frac{abn\gamma}{(n-1)(a^2 + b^2)^2} \right) I_{n-2} + \left(-\frac{\beta}{n-1} + \frac{\gamma}{(n-2)(a^2 + b^2)} (b \cos x - a \sin x) \right. \\
& \quad \left. + \frac{abn\gamma}{(n-2)(n-1)(a^2 + b^2)^2} (b \sin x - a \cos x) \right) \frac{1}{M^{n-1}(x)}, \tag{5.12}
\end{aligned}$$

where $M(x)$ is defined in (2.2) and I_n is discussed in Remark 5.2.

References

- [1] I. N. BRONSTEIN, H. MÜHLIG, G. MUSIOL, K. A. SEMENDJAJEW: *Taschenbuch der Mathematik*, Verlag Harri Deutsch GmbH, 2001.
- [2] I. S. GRADSHTEYN, I. M. RYZHIK: *Table of Integrals, Series, and Products*, ed. by A. JEFFREY, D. ZWILLINGER, New York: Academic Press, 2000, doi: <https://doi.org/10.1016/B978-0-12-294757-5.X5000-4>.
- [3] A. P. PRUDNIKOV, J. A. BRYCHKOV, O. I. MARICHEV: *Integrals and Series, vol. 1: Elementary Functions*, New York: Gordon & Breach, 1986.
- [4] A. F. TIMOFEEV: *Integration of functions*, Russian, Moscow: OGIZ, 1948.
- [5] I. A. VINOGRADOVA, S. N. OLEKHNİK, V. A. SADOVNICHYI: *Problems and exercises in Mathematical Analysis*, Russian, Moscow: Moscow University Press, 1999.

On a metamathematical question in talent care*

Csaba Szabó[†], Csilla Bereczky-Zámbó[‡],
Júlia Szenderák, Janka Szeibert

Eötvös Loránd University
csaba@cs.elte.hu
csilla95@gmail.com
szenderak.julia@gmail.com
szeibert.janka@gmail.com

Submitted: March 20, 2021

Accepted: April 18, 2021

Published online: April 19, 2021

Abstract

Recently more and more ethical issues arise in several sciences. We think that didactics of mathematics is not an exception. In this paper we investigate the question whether we can allow from mathematical precision in talent care. We suggest that these questions origin even from the formulation of a problem. The formulation of three well-known math problem is analyzed.

Keywords: talent care, ethical issues, problem posing

1. Ethical questions in science

“Mathematics is useful because we can find things to do with it. With this utility ethical issues arise relating to how mathematics impacts the world. (...) We study one of the most abstract areas of human knowledge: mathematics, the pursuit of

*This research was supported by the ÚNKP-19-2 New National Excellence Program of the Ministry for Innovation and Technology and by the ELTE Tehetség gondozási Tanács.

[†]The research of the first author was supported by the National Research, Development and Innovation Fund of Hungary, financed under the FK 124814 funding scheme.

[‡]The second author thanks the fund Mészáros Alapítvány for their support.

absolute truth. It has unquestionable authority. Indeed, it is clear that mathematics is one of the most useful and refined tools ever developed. When something is useful, however, it can often also be harmful; this can be either through deliberate misuse or ignorance.” [3] Although the system of mathematical thinking is a closed system, the results of mathematics are widely used in real life. Usually, a mathematician cannot be held responsible for the applications of their mathematical findings as those theorems are purely theoretical and have a well-defined system of conditions. At the same time in real life the same findings are often applied without checking the conditions. Still, these “uncontrolled” deductions are usually true. The opposite case is also possible. For example, in modeling problems, it is almost never possible to give a precise model to the task, and very often it is also impossible to translate the practical model to a mathematical one. Applying mathematical theorems without due prudence (e.g. leaving out conditions) can cause serious problems. Take the global financial crisis of 2007–2008 as an example, as so did [3]. The causes of the GFC are complex; however, there is consensus that mathematical work played a vital role. Unfortunately, the mathematical model and pricing of Collateralised Debt Obligations were based on several assumptions some of which did not hold. In the end, it led to the write-down of \$700 billion of CDO value from 2007 to 2008. The rest is history.

It is worth considering whether a mathematician should care about the aim of their mathematical task, i.e. what will their results be used for. Should they solve a problem if they know that the solution can be used to cause harm? It is not a specialty of mathematics; similar dilemmas appear in other branches of science. A classic example is that of the physicists taking part in the Manhattan plan. Their findings are revolutionary as scientific innovations, still, their work leads to the creation of a weapon capable of destroying humanity. Similar ethical dilemmas arose concerning the work of Ede Teller. Let us quote the famous scientist himself about the issue: “The scientist is not responsible for the laws of nature. It is his job to find out how these laws operate. It is the scientist’s job to find the ways in which these laws can serve the human will. However, it is not the scientist’s job to determine whether a hydrogen bomb should be constructed, whether it should be used, or how it should be used.” He reinforces this point of view later [6] stating that the scientist’s responsibility extends to work and to explain their findings along with the possible consequences – and no further.

Are any of these ethical issues relevant for a pure mathematician, say, a number theorist working in academia? Suppose they develop an algorithm for fast factorization. Should they publish it? If so, when, where, and how? If not, what should they do? Should they have thought about it beforehand? – asks Chiodo and Clifton [3]. Based on interviews, the typical answer would be that they would publish it immediately as they have the right to do so. But the consequences would be problematic – for instance, the breaking of RSA encryption in a chaotic manner could result a collapse of internet commerce and the global economy.

The following question also arises: Are there ethical dilemmas concerning the teaching of mathematics? Let us give some examples. Only a small part of the wide

range and great depth of known mathematical ideas can be shown during maths lessons. The yet limited competencies of students often make giving clear definitions and exact proofs impossible. Instead of proofs, it is not rare to demonstrate only trains of thoughts. When teaching the definition of a prime number in high school, we give a definition that is mathematically incorrect. It is an important question whether pupils are deceived when they are given incomplete definitions or is they are given trains of thoughts instead of proofs. Do we do them wrong by giving a false image of mathematics and mathematical thinking? Luckily this question is already well-handled in the education of the methodology of mathematics and there is a classical saying of Éva Vásárhelyi addressing this issue: “We have to grant some mathematical inaccuracies in favour of comprehensibility due to the level of proficiency of the students” [5]. We can see many occurrences of this kind of inaccuracies in primary and secondary level mathematics education, mainly when working on developing concepts. As a concept develops, in time, it becomes clearer. For example, when introducing exponential functions, understanding precisely why they make sense is out of reach for the students. Then, most of the concepts which were initially sloppy and loose, become exact by the time of final exams. These initial inaccuracies or gracious lies are serious errors from the aspect of mathematics but they are unavoidable because of the spiral structure of the curricula (key concepts are presented repeatedly throughout the curriculum, but with deepening layers of complexity, or in different applications) [2]. However, spiral curricula are well-reasonable from a developmental cognitive psychological point of view (e.g. the information is reinforced and solidified each time the student revisits the subject matter) [2].

Probably the most obvious ethical dilemma of teaching mathematics is what should appear in the National Core Curriculum (NCC), the law which regulates the official learning material in Hungary [26]. Thus, the first question that should arise in those who are preparing the NCC is which topics and competencies to include and whether these topics and competencies reflect the mathematical education that we want to mean by mathematical education. A row of ethical questions can be posed concerning the transitions from NCC to the framework curricula, then the local curricula and the syllabus, and at last the practice of teachers. This latter one includes an already debated issue, namely that teachers tend to teach students towards the maximum percentage on the school-leaving exam by endless mechanical practicing rather than fulfilling the aims set in NCC [11].

2. Ethical dilemmas in talent care

In this paper, we focus on ethical issues concerning mathematical talent care, which has long traditions in Hungary. One of the first mathematical journals was established and published in Hungary: Arany Dániel founded the “Középiskolai Matematikai Lapok” in 1893, the first issue was published in 1894. The journal has been functioning since. In the aspect of mathematics, Hungary belongs to the elite of the world. This fact is strongly related to talent care. Children partici-

pating in talent care programs or optional math classes (e.g. after-school classes) face the concept (and challenge) of giving arguments and proofs much more than their peers. As they need to give more and more accurate proofs (on talent care lessons and competitions), they reach a deeper level of understanding in mathematics. The mathematical development of students is largely affected by problems and problem compilations posed on talent care classes. Some well-known examples of problem books are: Szakköri feladatok matematikából 7–8. osztály (Problems for special math classes grade 7-8.) [20], Szakköri füzetek – Számelmélet (Optative math booklets – Number Theory) [21], Prímszámok (Prime numbers) [19], Négyzetszámok (Perfect squares) [18], Kombinatorika (Combinatorics) [16]. These booklets are well-known and used on paper or online by many students interested in mathematics. What ethical questions can be posed concerning talent care? Can we grant mathematical inaccuracies in favour of comprehensibility even in talent care? Can we correct inaccuracies and gracious lies that have occurred in normal mathematics class? How can we communicate so that neither knowledge nor authority is hurt? We have picked one of these issues and have transformed it to the following research question: In contrast to the inaccuracy that is accepted and often necessary in regular mathematics classes, can we give an inaccurate, mathematically incorrect answer or solution to a question or problem in talent care? Some problems can be considered typical in talent care as they appear often. We deal with three branches of problems which we will name “balance scale-” “statements-” and “camel-” problems. After analyzing the problems, we will also discuss the ethical questions arising concerning them.

3. Problems in talent care

In this section, we show three families of problems appearing in talent care. Two of these problems are of current interest among mathematicians, too. We try to analyze to what extent these problems can and should be posed to high-school students.

3.1. Balance scale-problems

Consider the following problem: Given 9 coins, one of them fake and lighter, find the fake coin in two weighings on a balance scale.[17]

The official solution the booklet shows that it can be done with three weighings, and does not show that two weighings are not enough. The following problem is handled similarly:

Which of the 8 coins is the fake one? There are 8 coins; one of them is fake. All real coins weigh the same. The fake coin is either lighter or heavier than the real coins. Find the fake coin and figure out whether it is heavier or lighter than the others, in the minimum number of weighings on a balance scale. [17]

One might think that the balance-scale problems are traditional and ancient. Yet, this type of problems is quite novel. Surprisingly, its first publication was

by E. D. Schell in the January 1945 issue of the American Mathematical Monthly [22]. The solution for n coins can be found for example in [24], a rather popular high-school problem book. The problem can be posed for several fake coins, and the complete solution is not known. The best known construction for n coins and unknown many fake coins has $7/11n$ many weighings [12] and the best known lower bound is [13] $\log_3(2^n + 2^{n-5} + 2^{n-6} + 2^{n-7} + 2^{n-9} + 2^{n-10} + 2^{n-12} + 2^{n-13})$. We can see from this formula that the general solution of this problem does not only look hard but is not even known. We might still think that for a small, given amount of coins we could pose the problem for students and after they tried cases and experimented, it is easier to show them that the lower and upper bounds are equal. Unfortunately, this is not a viable option either. For example, 11 and 2 are small numbers, so based on the assumption above, finding 2 fake coins out of 11 would be a problem suitable for secondary school students. In this problem, the number of the necessary weighings is 5 which was found out in 2015 and the proof uses the ternary Virtakallio–Golay code [4].

3.2. Camel-problems

We have camels and water and we want to cross a desert. The camels can carry a given amount of water and they can pass water to each other. They also consume water continually. The first question is if we have a given number of camels and all of them need to return to the starting point, except one. Then how far this exceptional camel can go. The second question is how many camels do we need if we want to get to a given distance. The following two analogous problems can be found in [17].

Peregrination in the desert. Ali ben Yusuf works far from his hometown, with a hundred-kilometer-wide desert between his workplace and his parents' house. He wants to visit his parents and starts planning the trip. It turns out that one can travel 20 km a day and the maximum weight to carry is three days' food and water. For simplicity, let us suppose that he can make dumps only after a whole day-route. How many days does he need to cross the desert?

The exact solution of the camel-problems is not known. The problems about desert-crossing were first introduced in 1947 [7]. The second version of the camel problem is formulated with jeeps instead of Yusuf, and is known under the name of the Jeep-problem. An analysis can be found in the article Gale's Round-Trip Jeep Problem [10]. The paper contains the proof of the following theorems. Let us suppose that we have enough fuel for n days. Then the maximum distance reachable with one jeep is $D_1 = 1 + \frac{1}{3} + \frac{1}{5} + \dots + \frac{1}{2n-1}$. Numerous versions of the jeep problem were solved [1, 8, 9]. In each work it is in common that both the readers and the authors themselves have a feeling of deficiency. Although every paper solves some problems, none of them reaches the goal it aims at.

3.3. Problems about statements

Two hundred statements. “On one side of a sheet of paper the following list of statements can be found:

1. At least one of the statements on this paper is true.
2. At least two of the statements on this paper are true.
3. At least three of the statements on this paper are true.
- ...
99. At least ninety-nine of the statements on this paper are true.
100. At least a hundred of the statements on this paper are true.

If we turn the sheet over, the following can be read:

1. At least one of the statements on this paper is false.
2. At least two of the statements on this paper are false.
3. At least three of the statements on this paper are false.
- ...
99. At least ninety-nine of the statements on this paper are false.
100. At least a hundred of the statements on this paper are false.

The text is continuous, we have left out some sentences (marked by three dots). How many true statements are there on the paper?” [17]

It is easy to ascertain that all hundred statements on the first side of the paper are true and on the second side statements 1–50 are true and statements 51–100 are false. Then the answer is that there are one hundred and fifty true statements on the paper.

Eight statements. “The following statements can be read on a paper:

1. At least one of the statements on this paper is false.
2. At least two of the statements on this paper are false.
3. At least three of the statements on this paper are false.
4. At least four of the statements on this paper are false.
5. At least five of the statements on this paper are false.
6. At least six of the statements on this paper are false.
7. At least seven of the statements on this paper are false.
8. ...

Unfortunately, the eighth statement is illegible. Is statement eight true or false?” [17]

Let us examine the first statement problem for three statements. Then the statements are the following:

1. At least one of the statements on this paper is false.
2. At least two of the statements on this paper are false.
3. At least three of the statements on this paper are false.

This problem has no “solution.” – the problem itself is not even a proper problem as it has no exact mathematical sense.

3.4. Students' possible dilemmas concerning the problems

The foundations of the methodology of problem-solving have been laid by György Pólya [14]. Since then problem-solving became an autonomous branch of didactics [23] with numerous aspects out of which we now highlight only one: Problem-solving as a thinking activity involves re-formulation, analysis, generalization, and extension of problems. This is how the idea of solving the three-statement-problem can appear after solving the hundred-statement-problem. Students working on generalizations of the problem probably see that they cannot give a solution for any odd number of statements. At this point they might become frustrated, think that they misunderstood something or made a mistake and start developing mathematical anxiety. They might also get confused not being able to solve a problem that is similar to one that they have already solved. They might even think that their previous solution might have been wrong.

The eight-statement problem can cause dilemmas already when interpreting the problem. One might try to analyze what happens if they write a specific sentence to the eighth place. In case of different sentences the conclusion can be different. If we write " $2+2 = 5$ over integers." the problem is solvable, but if we write " $2+2 = 4$ over integers.", we get a contradiction – this causes confusion concerning what to say about the original problem. If we write the sentence "John eats soup." as the eighth statement, the case is even more problematic: Why would seven sentences on a paper make John eat soup?

The sample solutions presented for the balance scale-problems are not complete. In both cases, a construction for finding the fake coin by a certain number of weighings is presented. But why do we solve the second problem by three weighings, not five? It is easy, the task told us to use the least possible number of measures. But can we manage to find the fake coin with two weighings? To give an exact solution, we need to show two things. First, we need to prove that less than three measures are not enough and that three measures are enough. The sample solution only shows the latter by giving a construction, it does not even mention that fewer weighings are not enough, let alone reasoning why. We can draw the conclusion that the sample solution does not answer fully the question – while making a convincing impression in students of doing so. Even some of the authors of this article were fully convinced by this impression before changing to "teacher's view" and, after careful analysis, noticing that there are mathematical, and what is more, metamathematical errors here. Let us imagine how a student might think. They know that these problems are dedicated to their age group. They try and try and can or cannot solve the problem. Probably they get a weaker result by themselves than the sample solution. Then they read it and see that it gives a thoughtful construction which suggests that it is the best possible. Talented students are likely to think that they are not entitled to judge the correctness of the book, so if they are not convinced, they blame themselves for not understanding the book which is written by clever professionals, therefore it must be perfect. This false impression is reinforced by the fact that these problem compilations contain several similar problems of one type in one block – for didactical reasons and for

making a greater impression.

Usually, it is normal and is also in accord with the theory of the zone of proximal development as it divides the chains of thoughts into steps, advancing from concrete to abstract, from small to large. This structure supports that the solution of one problem gives ideas that help to solve the next problem without taking away the joy of challenge and while providing a good experience, improving thinking as well.

The problem is that here a partial proof of the sample solution makes the impression of being a full one. These dilemmas and similar ones can appear in the case of the camel-problems, too.

4. Ethical dilemmas in talent care

The dilemmas appearing concerning the balance scale, the camel and the statement problems can be of different types: mathematical, teacher's, author's, and poser's dilemmas.

Mathematical dilemmas can be: What are the exact meanings of these problems? What are their exact solutions? Are they at all solvable? Are the conditions unambiguous? Does the problem use mathematical terms correctly? Are we sure that we do not try to see information in the text of the problem that actually is not included?

Teachers' dilemmas can be: Are we able to solve the problem? Are we able to solve with secondary-school methods? When we read the sample solution, we might also think that it is correct. But can we decide whether the sample solutions are real solutions to the problem? We must be careful when posing the problem so that we do not leave any questions unanswered or make any student frustrated. Finding a construction in case of the camel and the balance scale problems is already an exciting problem. Can we pose the problem in order to show a nice construction and at the same time without giving a full solution? If we only ask students to solve a balance scale-problem with a particular number of weighings, then we pose the problem ethically but less elegantly. But we do not address the inaccessible question: "is it possible with fewer measures?". Here the chance of students trying to generalize the problem and find the minimum number of measures still exists. We have already seen that this requires strong mathematical background. At this point we, as teachers, have an important task: we have to tell our students that there exists a minimal number of weighings, but in a lot of cases the full clarification would need a stronger mathematical background than they have. This raises even more questions: Should we show our students the solution even if we know that they will not understand everything? This way we make them see that the problem is solvable. We can also tell them that similar problems are subjects of great interest among mathematicians, too. We can mention that some of the problems are already solved but some of them are still unsolved. We can also give information on the specific problem we posed: who solved it and when.

Problem posers' first dilemma is how to pose these problems: To which age group can we show the problem? To which age group can we pose the problem so

that we can tell the solution, too? How much experience do students have with making proofs? Will they feel the need for proof? For those who feel, we might cause momentary frustration (if they do not have the competence to give correct proof). For those who do not realize the need for proof, we do not cause frustration. Their problem can appear later on when they see other types of thoughts and think that they are proofs. Another question is whether to pose only the easily solvable part of the problem. Here the dilemmas are similar as in case of teachers.

Author's dilemmas appear when someone starts to write a book or a compilation of problems. The first "author's dilemma" is whether I can include a problem in my book knowing that I must provide an incorrect or incomplete solution. We can see a lot of examples of this in books. The exact solutions of these problems cannot be presented in books for primary or secondary school students as they require higher mathematical knowledge. Another possibility is posing only a part of the problem. This makes it less appealing, maybe it will not even fit into the book. But if I omit it, I might deprive the readers of getting to know a nice, deep thought.

5. Interviews

To resolve as many dilemmas as possible, we conducted two interviews. One with PhD students to see how they solve the problems analyzed above. Are the solutions of these problems adequate to discuss on PhD level? Do doctoral students need directing questions to answer their own arising questions?

For the second interview we asked Sándor Róka, the author of [16–18] and many other problem books. He is one of the outstanding characters in Hungarian mathematical talent care. The idea of conducting an interview with him is thrilling and frightening at the same time. He gladly answered our questions and even after the interview he continued to share his thoughts with us via email. These thoughts are based on decades of experience in leading talent care courses for students from different age groups, starting with primary school, up until university level. Among his several widely used booklets and books, probably the most well-known is "2000 problems from the field of elementary mathematics" [15] which is part of the recommended literature for pre-service teachers. Sándor Róka was an educator in teacher training for a long time at the University of Nyíregyháza, now he focuses on talent care for upper primary and secondary students, mainly within the Erdős Pál Talent Care Center.

5.1. Interview with PhD students

To disclose the dilemmas concerning these problems and reveal the mathematical disorders in them we conducted an interview. The interviewees were three PhD students, all of them having a master's degree in mathematics education and participating in the Didactics of Mathematics PhD-program. Besides their PhD studies and research, they also teach mathematics at primary or secondary school. The interview consisted of open questions in connection with five tasks. The five tasks

were consecutive problems from [17]. The whole interview was videotaped. In the structure of the interview, the consecutiveness of the five tasks and the fact that each problem is built (in some sense) on the previous ones held a key role. At the beginning of the interview the interviewees were asked to analyze the problems one by one along with their solutions both from a student's and a teacher's point of view, parallelly. We summarize the interview. If necessary, we quote participants, they will be denoted by A (interviewer), and P_1, P_2, P_3 the PhD students.

The first two problems in the interview were the "Ninety-nine statements" and the "One hundred statements" problems.

Ninety-nine statements. The following statements are written on a paper:

1. Exactly one statement is true on this paper.
2. Exactly two statements are true on this paper.
3. Exactly three statements are true on this paper.

...

99. Exactly ninety-nine statements are true on this paper.

The text is continuous, we have left out some sentences (marked by three dots). Find out which statements on the paper are true.

One hundred statements. The following statements are written on a paper:

1. Exactly one statement is false on this paper.
2. Exactly two statements are false on this paper.

...

99. Exactly ninety-nine statements are false on this paper.
100. Exactly one hundred statements are false on this paper.

The text is continuous, we have left out some sentences (marked by three dots). Find out which statements on the paper are true.

The students solved the first two problems without spending too much time. The third problem of the interview was the "Two hundred statements" problem from Section 3.3.

At this problem, the answer is not so obvious. After considering the possibilities a linear ordering of the statements was proposed.

P1: "From at least x true statement, at least $x - 1$ is implied."

Then "open problem", doubts arose considering about what do we call a statement:

P1: "It might be an open problem with multiple solutions. The definition of a statement is: a sentence is a statement if it is clearly decidable whether or not it is true."

P2: „The text of the problem tells that they are statements so we cannot say that they are not... or that the poser of the problem is not right.”

After thinking a short while, they proved that the statements on the first page must be true, and they started to think about how many false statements have to be on the second page to get an adequate number of false statements. Then they quickly finished the analysis and the solution of the problem. The last two

problems were “The mysterious rock” and “The eight statements” problem from Section 3.3.

The mysterious rock. Once upon a time, two kings were fighting against each other. When one of them won and occupied the other’s castle, he found a strange rock in the castle yard. On the top side of the rock, there was the exact same statement engraved 77 times: “There are at least 77 false sentences engraved on this rock.” Next to the rock, there was a small table with the following explanation: “On the bottom side of the rock there are as many statements as on the top side, but these statements cannot be seen by any human.” How many true statements are there on the rock?

In the first part of their search for a solution they reached again the question of what we call a statement in mathematics. They started to understand that they had not considered the exact definition of a statement.

P1: “Because of the other statements it is necessary that statement 8 is false. This way, there will not be any contradiction in the system.”

P2: “It should be stated that the problem has a solution. Because . . . for example, let us take “The sky is blue” as statement 8. Will “The sky is blue” be false because it should be false based on the other statements?”

P1: “Of course not . . . We only have to decide, whether or not the 8 statement on the paper was true.”

A: “What is the matter with the 8 statements problem? I mean, the main problem. . . didactically and mathematically.”

P2: “Didactically the main error that it is like the task: Follow the sequence: 5, 10, 15, 20,” (meaning, that these kind of problems are not well defined, you need to figure out, what the problem poser thought.)

A: “What is the mathematical error?”

P1: “What is a statement?”

After discussing and clarifying the notion of statement, the next dilemma arose:

P2: “Is it not the resolution that we know from the formulation of the problem that these sentences are statements?”

At some point the students reached the conclusion that in these problems there is no given frame system (axioms) based on which we could decide whether these statements can be formally deduced or not. Firstly, there is no way to give true or false values to these statements such that they become consistent in the classical human language. Secondly, there is no base of knowledge that would tell us which sentences are statements. Thirdly, students, especially high-school students are not supposed to be aware of these ideas. So it is a kind of cheating not to tell students that here we have (or might have) paradoxes.

During the interview the three PhD students wanted to solve the problems in the first place instead of interpreting them. Right after the beginning, although P2 mentions that they should analyze the notion of statement, they soon got back to searching for the solution. The timespan of the interview was nearly an hour and it was only towards the end when PhD students started to have doubts about the

sense of the problems – or we should rather say the senselessness. None of them could tell the correct, exact definition of a statement.

One needs a very strong mathematical background and intelligence to start having doubts concerning the problem itself and the way of its posing. At PhD level this is attainable. The interview shows that the clarification of the problem at secondary school level would be very hard. A nice example of the resolution of a contradictory problem is the next one about knights and knaves. “Suppose A says, ‘Either I am a knave or else two plus two equals five.’ What would you conclude?” [25] The official solution presented in the book is the following: “The only valid conclusion is that the author of this problem is not a knight. The fact is that neither a knight nor a knave could possibly make such a statement. If A were a knight, then the statement that either A is a knave or that two plus two equals five would be false, since it is neither the case that A is a knave nor that two plus two equals five. Thus A, a knight, would have made a false statement, which is impossible. On the other hand, if A were a knave, then the statement that either A is a knave or that two plus two equals five would be true since the first clause that A is a knave is true. Thus A, a knave, would have made a true statement, which is equally impossible. Therefore the conditions of the problem are contradictory (...) Therefore, I, the author of the problem, was either mistaken or lying. I can assure you I wasn’t mistaken. Hence it follows that I am not a knight. For the sake of the records, I would like to testify that I have told the truth at least once in my life, hence I am not a knave either.” [25] So it can mean a resolution if we admit that the problem-poser either made a mistake or lied.

5.2. Interview with Sándor Róka

When preparing for the interview we chose to focus our questions on his problem compilations and beliefs and principles as a problem poser. He was very open towards us and started sharing his views as a problem poser and talent nurturer almost without having to ask concrete questions. The online interview lasted for 80 minutes. We quote some highlights from the interview that are interesting to our paper.

A: I found several problems in your books where the solution is practically a construction (e.g. the balance-scale problems when we have to find a fake coin). In this case we provide some inaccuracy, since with these constructions we can only show that a certain number of weighings is enough. What is your opinion in general about providing inaccuracies and demanding less mathematical preciseness even in talent care in order to make the problem more understandable?

RS: We do not look mathematics as “definition, theorem, proof, and then everything is complete”. This is not the way mathematics was explored. This would be an exaggeration. Precise axiomatic mathematics is important only for very few people. I am not sure that it should be introduced to children. They are not really interested in this more abstract part, they don’t understand why it is important. But problems asking for constructions appear a lot of times in competitions. Several times both directions are asked. If they ask to demonstrate that you can solve

the problem with a given amount of weighings – you just show a construction. In case of a lot of balance-scale problems you can argue that fewer weighings are not enough, so our construction reaches the optimum. Of course, this latter problem can be incredibly hard, I know. For example, take the problem of finding the two heaviest among 8 different balls – telling how many weighings we need is hard. But it is clear that to find the heaviest one we need seven weighings, we can give reasoning and also show a construction how to do it.

A: Do you think that there is a level in talent care when (or starting from when) things can be clarified? Do you have a solid opinion on who should support students in this clarification if later they feel the need to make everything precise? Whose task or role is to clarify the problems already seen without losing preciseness?

RS: I think these questions only bother specialists and maybe very few “gourmands” . . .

A: As you have already mentioned there are problems which can be approached from several aspects and can make the solver think about numerous related problems. To what extent is the phrasing of the problem important in this case? As you have already mentioned, the way how we pose the problems can be important. For example, take a “crossing a desert”-type problem. We can ask “how far can XY go” or we can ask “how XY can reach the furthest possible”. According to you, what is or what can be the role of phrasing?

RS: In this certain case it is quite random which option of phrasing I choose – sometimes this, sometimes that. But if I ask “which far can XY go” – then when someone tells a numerical answer, they also have to explain how that is possible. . . I guess the answers to these two questions somehow go together. But, surely, the way I ask questions, the way I compose them is important.

During the interview Sándor Róka made clear that it was an important aim of all problems in his books to raise students’ interest. It is important that the phrasing is catchy, so students start to solve the problem because of its exciting and interesting nature. In talent care classes it is the problem poser’s task to care with the problem profoundly and give as precise solution as possible. Earlier when dealing with balance scale-problems, Sándor Róka himself also prioritized preciseness so he asked for example “Can you solve with five weighings?”. Nowadays, he rather tells them to try to solve with as few weighings as possible. This way he can achieve that everyone can feel themselves successful – even those who did not find the optimum solution. During the discussion students see that there are several ways to solve the problem and there might be more effective solutions (in this case, with fewer weighings). With the most interested students the professor addresses a problem from several aspects and with several questions. During this process the details that were inaccurate at the beginning can be clarified. This clarification is the task of the teacher, along with giving answers to the questions that arise about the problems. This spirit appears in his books also.

6. Resolution and summary

In our paper we addressed an ethical question related to mathematics education. We analyzed whether we can give an inaccurate, mathematically incorrect answer or solution to a question or problem in talent care in contrast to the inaccuracy that is accepted and often necessary in regular mathematics classes. While inaccuracies and white lies are normal in regular math lessons as part of the spiral method of concept building, are they acceptable in talent care, too? When analyzing the statement-, balance scale- and camel problems, two types of questions arose. The first question is whether it is acceptable – and if yes, then in what content – to play with concepts that students do not know exactly. Distinguishing between the everyday sense and the mathematical concepts of a statement is not easy. The difficulty of the other two types of problems comes from the difference between a construction and an extremum. In both cases it is a challenge for a student to find an optimal construction, but giving a proof that the construction is optimal is beyond a secondary school student's competence. Concerning posing and presenting of problems, several dilemmas and metamathematical questions arise. After analyzing these questions we presented the extract of two interviews. In the first one participants are PhD students and the interview focuses on the mathematical background of a problem. In the second one the interviewee is Sándor Róka, one of the most well-known problem creators in Hungary, author of multiple books and problem booklets. In the first interview it turned out that the solution and resolution of the problems is a challenge even for PhD students. It requires serious mathematical background and intelligence to doubt the posing of the problem itself. This interview enlightened that clarifying these problems would be very difficult at secondary school level. The short conclusion of the second interview is that a problem poser needs to pay attention to a lot of aspects at once. The borderline of these aspects is sharp and we have to cross at least some of them. According to the *Ars Poetica* of Sándor Róka, the first border to keep us inside is that of the attention and the interest of students. Without the mathematical commitment of the students, without gaining their attention and raising their interest none of the further questions can even appear. Later, when students and their teacher deal with the problems together and some doubts arise at border-crossings, they need to resolve the doubts together.

Talent care is a significant task. It requires profound planning, a lot of preparation, and continuous attention.

When Ptolemaios the first asked Euclid whether there is a way of learning geometry that is easier and shorter than the one presented in *Elements*, the great geometer answered: There is no royal road to geometry. In the footsteps of Euclid, we can say: There is no royal road to talent care.

References

- [1] G. G. ALWAY: *Crossing the desert*, The Mathematical Gazette 41.337 (1957), pp. 209–209, DOI: <https://doi.org/10.2307/3609199>.
- [2] J. S. BRUNER: *The process of education*, Harvard University Press (1960).
- [3] M. CHIODO, T. CLIFTON: *The importance of Ethics in Mathematics*, EMS Newsletter 12.114 (2019), pp. 34–37, DOI: <https://doi.org/10.4171/NEWS/114/9>.
- [4] A. M. CHUDNOV: *Weighing algorithms of classification and identification of situations*, Discrete Mathematics and Applications 25.2 (2015), pp. 69–81, DOI: <https://doi.org/10.1515/dma-2015-0007>.
- [5] P. CSÁNYI, K. FÁBIÁN, C. SZABÓ, Z. SZABÓ: *Number theory vs. Hungarian high school textbooks : the fundamental theorem of arithmetic*, Teaching Mathematics and Computer Science 13.2 (2015), pp. 209–223, DOI: <https://doi.org/10.5485/TMCS.2015.0397>.
- [6] T. EDE: *Légitposta*, Budapest: Háttér Lap- és Könyvkiadó, 1990.
- [7] N. J. FINE: *The Jeep Problem*, The American Mathematical Monthly 54.1 (1947), pp. 24–31, DOI: <https://doi.org/10.2307/2304923>.
- [8] D. GALE: *The Jeep Once More or Jeeper by the Dozen*, The American Mathematical Monthly 77.5 (1970), pp. 493–501.
- [9] D. GALE: *Tracking the Automatic ANT*, New York: Springer, 1998, DOI: <https://doi.org/10.1007/978-1-4612-2192-0>.
- [10] A. HAUSRATH, B. JACKSON, J. MITCHEM, E. SCHMIECHEL: *Gale’s Round-Trip Jeep Problem*, The American Mathematical Monthly 102.4 (1995), pp. 299–309, DOI: <https://doi.org/10.2307/2974949>.
- [11] V. KOVÁCS: *Gráfok modern bevezetése a középiskolában*, Training and Practice 15.1-2 (2017), pp. 275–294, DOI: <https://doi.org/10.17165/TP.2017.1-2.16>.
- [12] L. AN-PING, M. AIGNER: *Searching for counterfeit coins*, Graphs and Combinatorics 13 (1997), pp. 9–20.
- [13] L. AN-PING, H. VON EITZEN: *An improved lower-bound on the counterfeit coins problem*, arXiv (2009).
- [14] G. PÓLYA: *A gondolkodás iskolája*, Budapest: Gondolat Könyvkiadó, 1971.
- [15] R. SÁNDOR: *2000 feladat az elemi matematika köréből*, Budapest: Typotex Kiadó, 2010.
- [16] R. SÁNDOR: *Kombinatorika*, Debrecen: Tóth Könyvkereskedés és Kiadó Kft., 2001.
- [17] R. SÁNDOR: *Logika-land*, Budapest: Typotex Kiadó, 2013.
- [18] R. SÁNDOR: *Négyzetszámok*, Debrecen: Tóth Könyvkereskedés és Kiadó Kft., 2001.
- [19] R. SÁNDOR: *Prímszámok*, Debrecen: Tóth Könyvkereskedés és Kiadó Kft., 2002.
- [20] R. SÁNDOR: *Szakköri feladatok matematikából 7-8. osztály*, Debrecen: Tóth Könyvkereskedés és Kiadó Kft., 2002.
- [21] R. SÁNDOR: *Szakköri füzetek – Számelmélet*, Debrecen: Tóth Könyvkereskedés és Kiadó Kft., 2007.
- [22] E. D. SCHELL: *Problem E651—Weighed and found wanting*, The American Mathematical Monthly 52.1 (1945), p. 42.
- [23] A. H. SCHOENFELD: *Mathematical Problem Solving*, Orlando, Florida: Academic Press INC., 1985, DOI: <https://doi.org/10.1016/B978-0-12-628870-4.50001-3>.

- [24] D. O. SKLJARSZKIJ, N. N. CSENCOV, I. M. JAGLOM: *The USSR Olympiad problem book : selected problems and theorems of elementary mathematics*, San Francisco, California: W.H. Freeman and Company, 1962.
- [25] R. SMULLYAN: *What is the Name of this Book?*, New Jersey: Prentice Hall, Englewood Cliffs, NJ, 1978.
- [26] WWW.OFI.HU: *The Hungarian National Core Curriculum*, Teaching Mathematics and Computer Science (2012).

