

## Contents

ALVARADO, A., Arithmetic progressions on quartic elliptic curves . . . . .	3
AMRANE, R. A., BELBACHIR, H., Non-integerness of class of hyperharmonic numbers . . . . .	7
BALOGH, Zs., JUHÁSZ, T., Conditions for groups whose group algebras have minimal Lie derived length . . . . .	13
BARRENECHEA, A. L., Regularity of certain Banach valued stochastic processes . . . . .	21
BEG, I., ABBAS, M., AZAM, A., Periodic fixed points of random operators . . . . .	39
BÉRCZES, T., GUTA, G., KUSPER, G. SCHREINER, W., SZTRIK, J., Evaluating a probabilistic model checker for modeling and analyzing retrial queueing systems . . . . .	51
EGRÍ-NAGY, A., NEHANIV, C. L., On the skeleton of a finite transformation semigroup . . . . .	77
FILIP, F., LIPTAI, K., MÁTYÁS, F., TÓTH, J. T., On the best estimations for dispersions of special ratio block sequences . . . . .	85
KRASNIQI, V., MANSOUR, T., SHABANI, A. SH., Some inequalities for $q$ -polygamma function and $\zeta_q$ -Riemann zeta functions . . . . .	95
MÁTYÁS, F., LIPTAI, K., TÓTH, J. T., FILIP, F., Polynomials with special coefficients . . . . .	101
LUCA, F., MEJÍA HUGUET, V. J., On perfect numbers which are ratios of two Fibonacci numbers . . . . .	107
OLAJOS, P., Properties of balancing, cobalancing and generalized balancing numbers . . . . .	125
SHEIKHOESLAMI, S. M., VOLKMANN, L., Signed $(k, k)$ -domatic number of a graph . . . . .	139
SONDOW, J., MARQUES, D., Algebraic and transcendental solutions of some exponential equations . . . . .	151
TROLL, E. M., HOFFMANN, M., Geometric properties and constrained modification of trigonometric spline curves of Han . . . . .	165
XIE, J., TAN, J., LI, S., CTH B-spline curves and its applications . . . . .	177
<b>Methodological papers</b>	
KULKARNI, R. G., Solving certain quintics . . . . .	193
NAGY-KONDOR, R., Spatial Ability, Descriptive Geometry and Dynamic Geometry Systems . . . . .	199
PATAKI, N., SZŰGYI, Z., C++ exam methodology . . . . .	211
SIKET, I., GYIMÓTHY, T., The software developers' view on product metrics — A survey-based experiment . . . . .	225
TÉGLÁSI, I., Mathematical competences examined on secondary school students . . . . .	241

# ANNALES MATHEMATICAE ET INFORMATICAE

TOMUS 37. (2010)



COMMISSIO REDACTORIUM

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



HUNGARIA, EGER

**ANNALES MATHEMATICAE ET INFORMATICAE**

**International journal for mathematics and computer science**

**Referred by  
Zentralblatt für Mathematik  
and  
Mathematical Reviews**

The journal of the Institute of Mathematics and Informatics of Eszterházy Károly College is open for scientific publications in mathematics and computer science, where the field of number theory, group theory, constructive and computer aided geometry as well as theoretical and practical aspects of programming languages receive particular emphasis. Methodological papers are also welcome. Papers submitted to the journal should be written in English. Only new and unpublished material can be accepted.

Authors are kindly asked to write the final form of their manuscript in  $\text{\LaTeX}$ . If you have any problems or questions, please write an e-mail to the managing editor Miklós Hoffmann: [hofi@ektf.hu](mailto:hofi@ektf.hu)

The volumes are available at <http://ami.ektf.hu>

ANNALES  
MATHEMATICAE ET  
INFORMATICAE

VOLUME 37. (2010)

EDITORIAL BOARD

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

INSTITUTE OF MATHEMATICS AND INFORMATICS  
ESZTERHÁZY KÁROLY COLLEGE  
HUNGARY, EGER

HU ISSN 1787-5021 (Print)  
HU ISSN 1787-6117 (Online)

A kiadásért felelős az  
Eszterházy Károly Főiskola rektora  
Megjelent az EKF Líceum Kiadó gondozásában  
Kiadóvezető: Kis-Tóth Lajos  
Felelős szerkesztő: Zimányi Árpád  
Műszaki szerkesztő: Tómacs Tibor  
Megjelent: 2010. december Pédányszám: 30

Készítette az  
Eszterházy Károly Főiskola nyomdája  
Felelős vezető: Kérészy László

# Arithmetic progressions on quartic elliptic curves

Alejandra Alvarado

University of Arizona

*Submitted 19 October 2009; Accepted 20 March 2010*

## Abstract

Consider the curve  $C : y^2 = ax^4 + bx^2 + c$ . MacLeod previously found four curves of the given form, with an arithmetic progression in the  $x$  coordinates, of length 14. By similar methods, we also find the same four curves, and several more.

*Keywords:* Diophantine equations, arithmetic progressions.

## 1. Introduction

Let  $F(x)$  be a quartic polynomial over the rationals, which is not the square of a quadratic. If a rational point exists on  $y^2 = F(x)$ , then this curve is birationally equivalent to an elliptic curve. We will call these curves *quartic elliptic curves* [4].

We will say that points on a curve are in *arithmetic progression* (AP) if their  $x$  coordinates form an arithmetic progression. Previously, Ulas found an infinite family of curves with an AP of length 12 [4]. The author first found a curve

$$C_a : y^2 = f_a(x),$$

where  $f_a$  is degree four and parameter  $a$ , with length ten AP. The AP in  $x$  is  $\{1, 2, \dots, 10\}$ . Ulas then noted that  $f_a(0) = f_a(11)$ . The quartic curve  $Y^2 = f_a(0)$  is birationally equivalent to an elliptic curve of rank three. Thus, points on this rank three elliptic curve map to points on  $Y^2 = f_a(0)$  which give infinitely many values for  $a$ .

By the use of symmetry and methods similar to those found in Ulas, MacLeod [2] found an infinite family of curves with AP length ten. Numerical investigations lead to four examples with AP length 14. In this paper, we follow similar methods as Ulas and MacLeod. We find the same four curves, plus eleven more.

## 2. Arithmetic progressions

MacLeod simplifies Ulas' approach when searching for points in AP. Because the general solution to length ten AP is difficult, Ulas instead considers a curve with symmetry. As noted in MacLeod, it is enough to consider the curve to be symmetric about the  $x$ -axis. In that case, we can write the curve as  $y^2 = ax^4 + bx^2 + c$ , i.e.,  $F(x) = ax^4 + bx^2 + c$  with rational coefficients. In this section, we construct curves with length 14 AP. From these, we will attempt to find examples of length 16.

Suppose

$$F(\pm 1) = p^2$$

$$F(\pm 3) = q^2$$

$$F(\pm 5) = r^2$$

$$F(\pm 7) = s^2$$

This gives us an AP of length eight. The first three equations imply

$$\begin{aligned} a &= \frac{2p^2 - 3q^2 + r^2}{384} \\ b &= -\frac{34p^2 - 39q^2 + 5r^2}{192} \\ c &= \frac{150p^2 - 25q^2 + 3r^2}{128} \end{aligned}$$

which forces  $s^2 = 5p^2 - 9q^2 + 5r^2$ . This last equation, representing a quadric surface, has a parametrization in  $u$  and  $v$ :

$$\begin{aligned} (p : q : r : s) &= (-u^2 - 2uv - 5v^2 - 2uw + w^2 : \\ &u^2 - 5v^2 + w^2 : \\ &u^2 - 5v^2 - 2uw - 2vw - w^2 : \\ &u^2 + 10uv + 5v^2 + 10vw + w^2) \end{aligned}$$

Then  $F(x) = ax^4 + bx^2 + c$  is a polynomial in  $x$  with coefficients in  $(u, v, w)$ . Thus far, we have an infinite family of curves with an arithmetic progression of length eight in the  $x$ -coordinates. Up to this point, we have followed similar techniques as MacLeod, except that our parametrization has smaller coefficients. We now introduce a different approach to this problem. If we want an AP of length at least 14, then we must force

$$F(\pm 9) = t_1^2$$

$$F(\pm 11) = t_2^2$$

$$F(\pm 13) = t_3^2$$

Consider the family of planes  $w = Au + Bv$  in the  $(u, v, w)$  space. These last three homogeneous equations are now quartics in  $(u, v)$  with coefficients in  $(A, B)$ . With

respect to  $u$ , these curves are singular if and only if their discriminant is zero. With the help of Magma [1], we find that  $(2A + 1 - B)^2(A + B + 2)^2$  is a factor of the discriminant of all three. After substituting  $B = 2A + 1$ , we find  $(u + 2v)^2$  is a factor of all three equations. If we substitute  $B = -A - 2$ , then  $(u - v)^2$  is a factor of the three equations.

Let us first consider the case  $B = -A - 2$ . If  $v = 1$ , then

$$\begin{aligned} F(\pm 9) &= (u - 1)^2(u^2 + A^4u^2 + 24Au^2 + 2A^2u^2 - 24A^3u^2 + 124Au + 38u - 2A^4u \\ &\quad + 180uA^2 + 76A^3u + 361 + 14A^2 + A^4 - 52A^3 + 412A) \\ F(\pm 11) &= (u - 1)^2(841 - 2A^4u + 560uA^2 + 216A^3u + 384Au + 972A + 58u \\ &\quad + A^4u^2 + 14A^2 + A^4 - 132A^3 - 84A^3u^2 + u^2 + 84Au^2 + 2A^2u^2) \\ F(\pm 13) &= (u - 1)^2(1681 - 2A^4u + 1280uA^2 + 472A^3u + 872Au + 1952A + 82u \\ &\quad + A^4u^2 + 14A^2 + A^4 - 272A^3 - 200A^3u^2 + u^2 + 200Au^2 + 2A^2u^2) \end{aligned}$$

Write the rational value  $A = a_1/a_2$ , and consider the degree six polynomial

$$f(u) = \frac{F(9)F(11)F(13)}{(u - 1)^6}$$

with coefficients in  $(a_1, a_2)$ . Then the equation  $Y^2 = f(u)$  represents a hyperelliptic curve of degree six. The reason for considering the above curves with discriminant zero, is because it is much more practical to search for points on a hyperelliptic curve of degree six rather than twelve. With the aid of Magma, we found points on this curve by varying values of  $(a_1, a_2)$  up to  $|a_1| + |a_2| = 200$ . We then checked whether  $F(\pm 9)$ ,  $F(\pm 11)$ , and  $F(\pm 13)$  are squares but  $F$  is not a perfect square. We found the same four curves listed in MacLeod:

1.  $y^2 = -17x^4 + 3130x^2 + 8551$
2.  $y^2 = 2002x^4 - 226820x^2 + 18168514$
3.  $y^2 = 3026x^4 - 222836x^2 + 3709234$
4.  $y^2 = 34255x^4 - 1436006x^2 + 447963175$

and seven new curves:

1.  $y^2 = 2753x^4 - 728770x^2 + 59217921$
2.  $y^2 = 627x^4 - 87870x^2 + 3312859$
3.  $y^2 = 3689x^4 - 88994x^2 + 4312441$
4.  $y^2 = -143644199x^4 + 26117509014x^2 - 24973534431$
5.  $y^2 = -15015x^4 + 2758974x^2 + 25050025$

$$6. y^2 = 506363x^4 - 1726486x^2 + 740805923$$

$$7. y^2 = -2219x^4 + 378494x^2 + 19089469$$

If we now look at the case  $B = 2A + 1$ , we find at least four more distinct curves:

$$1. y^2 = 1012726x^4 - 3452972x^2 + 1481611846$$

$$2. y^2 = -308503x^4 + 53324830x^2 + 72961849$$

$$3. y^2 = -31730x^4 + 4968916x^2 + 68267950$$

$$4. y^2 = -18750709x^4 + 5055585994x^2 + 16925811919$$

We end this paper with some final comments. First, none of these curves contain a length 16 AP with  $x$ -coordinates  $\{-13, -11, \dots, 13\}$ . Secondly, the reason we used Magma was because it effectively found rational points on hyperelliptic curves. Although, Michael Stoll's ratpoint package is now supported by Sage [3]. Ratpoints finds rational points of bounded height on hyperelliptic curves.

**Acknowledgements.** Thank you to Dr. Andrew Bremner for suggesting this problem, Dong-Quan Nguyen, and Robert Miller.

## References

- [1] BOSMA, W., CANNON, J., PLAYOUST, C., The Magma algebra system. I. The user language., *J. Symbolic Comput.*, 24 (1997) 235–265.
- [2] MACLEOD, A.J., 14-term arithmetic progressions on quartic elliptic curves, *J. Integer Seq.*, 9 (2006) 1, Article 06.1.2, 4 pp. (electronic).
- [3] STEIN, W.A. et al., *Sage Mathematics Software (Version 4.2.1)*, The Sage Development Team, 2009, <http://www.sagemath.org>.
- [4] ULAS, M., A note on arithmetic progressions on quartic elliptic curves, *J. Integer Seq.*, 8 (2005) 3, Article 05.3.1, 5 pp. (electronic).

**Alejandra Alvarado**

617 N. Santa Rita Ave. Tucson, AZ 85721

e-mail: [alvarado@math.arizona.edu](mailto:alvarado@math.arizona.edu)



# Non-integerness of class of hyperharmonic numbers

Rachid Aït Amrane<sup>a</sup>, Hacène Belbachir<sup>b</sup>

<sup>a</sup>ESI/ Ecole nationale Supérieure d'Informatique, Alger, Algeria

<sup>b</sup>USTHB/ Faculté de Mathématiques, Alger, Algeria

*Submitted 7 September 2009; Accepted 10 February 2010*

## Abstract

Our purpose is to establish that hyperharmonic numbers – successive partial sums of harmonic numbers – satisfy a non-integer property. This gives a partial answer to Mező's conjecture.

*Keywords:* Harmonic numbers; Hyperharmonic numbers; Bertrand's postulate.

*MSC:* 11B65, 11B83.

## 1. Introduction

In 1915, L. Taisinger proved that, except for  $H_1$ , the harmonic number  $H_n := 1 + \frac{1}{2} + \dots + \frac{1}{n}$  is not an integer. More generally, H. Belbachir and A. Khelladi [1] proved that a sum involving negative integral powers of consecutive integers starting with 1 is never an integer.

In [3, p. 258–259], Conway and Guy defined, for a positive integer  $r$ , the hyperharmonic numbers as iterate partial sums of harmonic numbers

$$H_n^{(1)} := H_n \text{ and } H_n^{(r)} = \sum_{k=1}^n H_k^{(r-1)} \quad (r > 1).$$

The number  $H_n^{(r)}$ , called the  $n^{\text{th}}$  hyperharmonic number of order  $r$ , can be expressed by binomial coefficients as follows (see [3])

$$H_n^{(r)} = \binom{n+r-1}{r-1} (H_{n+r-1} - H_{r-1}). \quad (1.1)$$

For other interesting properties of these numbers, see [2].

I. Mezó, see [5], proved that  $H_n^{(r)}$ , for  $r = 2$  and  $3$ , is never an integer except for  $H_1^{(r)}$ . In his proof, he used the reduction modulo the prime  $2$ . He conjectured that  $H_n^{(r)}$  is never an integer for  $r \geq 4$ , except for  $H_1^{(r)}$ .

In our work, we give another proof that  $H_n^{(r)}$  is not an integer for  $r = 2, 3$  when  $n \geq 2$ . We also give an answer to Mezó's conjecture for  $r = 4$  and a partial answer for  $r > 4$ .

Our proof is based on Bertrand's postulate which says that for any  $k \geq 4$ , there is a prime number in  $]k, 2k - 2[$ . See for instance [4, p. 373].

## 2. Results

**Theorem 2.1.** *For any  $n \geq 2$ , the hyperharmonic number  $H_n^{(2)}$  is never an integer.*

**Proof.** Let  $n \geq 2$  and assume  $H_n^{(2)} \in \mathbb{N}$ . We have  $H_n^{(2)} = \binom{n+1}{1} (H_{n+1} - H_1) = (n+1)(H_{n+1} - 1)$ , therefore  $(n+1)H_{n+1} = (n+1)\left(1 + \frac{1}{2} + \dots + \frac{1}{n+1}\right)$  is an integer. Let  $P$  be the greatest prime number less than or equal to  $n$ . We have  $\frac{(n+1)!}{P}H_{n+1} - \frac{(n+1)!}{P} \sum_{k \neq P} \frac{1}{k} = \frac{(n+1)!}{P^2}$ . The left hand side of the equality is an integer while the right hand side is not. Indeed, by Bertrand's postulate, the prime  $P$  is coprime to any  $k$ ,  $k \leq n+1$ , contradiction.  $\square$

**Theorem 2.2.** *For any  $n \geq 2$ , the hyperharmonic number  $H_n^{(3)}$  is never an integer.*

**Proof.** The arguments here are similar to those in the proof of the following theorem.  $\square$

**Theorem 2.3.** *For any  $n \geq 2$ , the hyperharmonic number  $H_n^{(4)}$  is never an integer.*

**Proof.** We have  $H_2^{(4)} = \frac{9}{2} \notin \mathbb{N}$ ,  $H_3^{(4)} = \frac{37}{3} \notin \mathbb{N}$  and  $H_4^{(4)} = \frac{319}{12} \notin \mathbb{N}$ . Let  $n \geq 5$  and assume that  $H_n^{(4)} \in \mathbb{N}$ . With the same arguments as in the proof of Theorem 1 we deduce that  $(n+1)(n+2)(n+3)H_n \in \mathbb{N}$ . Let  $P$  be the greatest prime less than or equal to  $n$ . Then  $\frac{(n+3)!}{P}H_n - \frac{(n+3)!}{P} \left(1 + \frac{1}{2} + \dots + \frac{1}{P-1} + \frac{1}{P+1} + \dots + \frac{1}{n}\right) = \frac{(n+3)!}{P^2}$ . The left hand side of the equality is an integer while the right hand side is not. Again,  $P$  is coprime to any  $k$ ,  $P < k \leq n+3$ . Therefore, if  $P$  divides  $(n+3)!$ , then  $P$  would divide  $(P+1) \cdots (n+3)$ , thus one of the factors would be equal to  $2P$ , consequently  $2P - 2 \leq n+1$ , hence, by Bertrand's postulate, there would exist a prime strictly between  $P$  and  $n+1$ , contradicting the fact that  $P$  is the greatest prime less than or equal to  $n$ . Therefore,  $H_n^{(4)} \notin \mathbb{N}$  for any  $n \geq 2$ .  $\square$

For  $r \geq 5$ , we give a class of hyperharmonic numbers satisfying the non-integer property.

**Theorem 2.4.** *Let  $n \in \mathbb{N}$  such that  $n \geq 2$  and that none of the integers  $n + 1, n + 2, \dots, n + r - 4$  is a prime number, then we have  $H_n^{(r)} \notin \mathbb{N}$ .*

**Proof.** It is easy to see that  $H_2^{(r)} = \frac{r+1}{2} + \frac{r}{2} \notin \mathbb{N}$ ,  $H_3^{(r)} = \frac{(r+1)(r+2)}{6} + \frac{r(r+2)}{6} + \frac{r(r+1)}{6} \notin \mathbb{N}$  and  $H_4^{(r)} = \frac{(r+1)(r+2)(r+3)}{24} + \frac{r(r+2)(r+3)}{24} + \frac{r(r+1)(r+3)}{24} + \frac{r(r+1)(r+2)}{24} \notin \mathbb{N}$ . For any  $n \geq 5$ , we have by relation (1.1)

$$H_n^{(r)} = \frac{(n+1)(n+2) \cdots (n+r-1)}{(r-1)!} \left( H_n + \frac{1}{n+1} + \frac{1}{n+2} + \cdots + \frac{1}{n+r-1} - H_{r-1} \right).$$

Set  $E_{r,n} := (r-1)! \left( H_n^{(r)} - \binom{n+r-1}{r-1} H_{r-1} \right) - (n+1) \cdots (n+r-1) \left( \frac{1}{n+1} + \cdots + \frac{1}{n+r-1} \right)$ . Thus  $E_{r,n} = (n+1)(n+2) \cdots (n+r-1) \left( 1 + \frac{1}{2} + \cdots + \frac{1}{n} \right)$ .

Assume that  $H_n^{(r)}$  is an integer. So  $E_{r,n}$  is an integer as well. Let  $P$  be the greatest prime  $\leq n$ . Then we have

$$\frac{n!}{P} E_{r,n} = \frac{(n+r-1)!}{P} \left( 1 + \cdots + \frac{1}{P} + \cdots + \frac{1}{n} \right),$$

and therefore

$$\frac{(n+r-1)!}{P} E_{r,n} - \frac{(n+r-1)!}{P} \sum_{k \neq P} \frac{1}{k} = \frac{(n+r-1)!}{P^2}.$$

The left side of the equality is an integer. If the right side is an integer, then  $P$  should divide  $(n+2) \cdots (n+r-1)$ , hence one of the integers  $n, \dots, (n+r-3)$  should be equal to  $2P-2$ , so either there exist a prime  $Q$  strictly between  $P$  and  $n+1$  and this is a contradiction with Bertrand's postulate, either one of the integers  $n+k$  with  $1 \leq k \leq r-4$  is prime and this contradicts the assumption of the Theorem.  $\square$

It is well known that we can exhibit an arbitrary long sequence of consecutive composite integers:  $m!+2, m!+3, \dots, m!+m$ , ( $m \geq 3$ ). We will use this fact to establish that for all  $r \geq 5$ , we can find a non integer hyperharmonic number  $H_n^{(r)}$ .

**Corollary 2.5.** *Let  $r \geq 5$ , then the hyperharmonic numbers  $H_{r!+1}^{(r)}, H_{r!+2}^{(r)}, H_{r!+3}^{(r)}$  and  $H_{r!+4}^{(r)}$  satisfy the non-integer property.*

**Proof.** It suffices to use Theorem 2.4.  $\square$

The arguments used in the proof of Theorem 2.4 give more. As an illustration, we treat the case  $r = 5$ .

**Proposition 2.6.** *For any  $n \geq 2$ , the hyperharmonic number  $H_n^{(5)}$  is never an integer when  $n+1 \neq 2Q-3$  is prime with  $Q$  prime.*

**Proof.** For  $n = 2$  or  $3$ ,  $n$  odd, or even with  $n + 1$  composite, see Theorem 2.4. For even  $n \geq 4$  with  $n + 1$  prime, using notations in the proof of Theorem 2.4, if  $H_n^{(5)} \in \mathbb{N}$  then  $P \mid (n + 2)(n + 3)(n + 4)$ . We have  $P \nmid (n + 2)$ , there would be a prime between  $P$  and  $n = 2P - 2$ . We have  $P \nmid (n + 3)$ , otherwise  $n + 3 = 2P$  which contradicts the fact  $n + 3$  is odd. Finally, if  $n + 4 = 2P$  i.e.  $n + 1 = 2P - 3$ , we have a contradiction.  $\square$

**Example 2.7.** For  $n \leq 100$ , we list the values of  $r$ , given by Theorem 2.4, such that  $H_n^{(r)}$  is never an integer.

1.  $H_n^{(5)} \notin \mathbb{N}$  for  $n = 2, 3, \mathbf{4}, 5, 7, 8, 9, 11, \mathbf{12}, 13, 14, 15, \mathbf{16}, 17, 19, 20, 21, 23, 24, 25, 26, 27, \mathbf{28}, 29, 31, 32, 33, 34, 35, \mathbf{36}, 37, 38, 39, \mathbf{40}, 41, 43, 44, 45, \mathbf{46}, 47, 48, 49, 50, 51, \mathbf{52}, 53, 54, 55, 56, 57, 59, \mathbf{60}, 61, 62, 63, 64, 65, \mathbf{66}, 67, 68, 69, 71, \mathbf{72}, 73, 74, 75, 76, 77, 79, 80, 81, 83, 84, 85, 86, 87, \mathbf{88}, 89, 90, 91, 92, 93, 94, 95, \mathbf{96}, 97, 98, 99, \mathbf{100}$ .

The bold numbers are given by Proposition 2.6.

2.  $H_n^{(6)} \notin \mathbb{N}$  for  $n = 2, 3, 7, 8, 13, 14, 19, 20, 23, 24, 25, 26, 31, 32, 33, 34, 37, 38, 43, 44, 47, 48, 49, 50, 53, 54, 55, 56, 61, 62, 63, 64, 67, 68, 73, 74, 75, 76, 79, 80, 83, 84, 85, 86, 89, 90, 91, 92, 93, 94, 97, 98$ .
3.  $H_n^{(7)} \notin \mathbb{N}$  for  $n = 2, 3, 7, 19, 23, 24, 25, 31, 32, 33, 37, 43, 47, 48, 49, 53, 54, 55, 61, 62, 63, 67, 73, 74, 75, 79, 83, 84, 85, 89, 90, 91, 92, 93, 97$ .
4.  $H_n^{(8)} \notin \mathbb{N}$  for  $n = 2, 3, 23, 24, 31, 32, 47, 48, 53, 54, 61, 62, 73, 74, 83, 84, 89, 90, 91, 92$ .
5.  $H_n^{(9)} \notin \mathbb{N}$  for  $n = 2, 3, 23, 31, 47, 73, 83, 89, 90, 91$ .
6.  $H_n^{(10)} \notin \mathbb{N}$  for  $n = 2, 3, 89, 90$ .
7.  $H_n^{(11)} \notin \mathbb{N}$  for  $n = 2, 3, 89$ .

**Acknowledgements.** The second author is grateful to István Mező for sending us a copy of his cited paper and pointing our attention to hyperharmonic numbers. We are also grateful to Yacine Aït Amrane for useful discussions.

## References

- [1] BELBACHIR, H., KHELLADI, A., On a sum involving powers of reciprocals of an arithmetical progression, *Annales Mathematicae et Informaticae*, 34 (2007), 29–31.
- [2] BENJAMIN, A. T., GAEBLER, D., GAEBLER, R., A combinatorial approach to hyperharmonic numbers, *INTEGERS: The Electronic Journal of Combinatorial Number Theory*, 3 (2003), #A15.

- [3] CONWAY, J. H., GUY, R. K., The book of numbers, *New York, Springer-Verlag*, 1996.
- [4] HARDY, G. H., WRIGHT, E. M., An introduction to the theory of numbers, *Oxford at the Clarendon Press*, 1979.
- [5] MEZŐ, I., About the non-integer property of hyperharmonic numbers, *Annales Univ. Sci. Budapest., Sect. Math.*, 50 (2007), 13–20.
- [6] TAEISINGER, L., Bemerkung über die harmonische Reihe, *Monatshefte für Mathematik und Physik*, 26 (1915), 132–134.

**Rachid Aït Amrane**

ESI/ Ecole nationale Supérieure d'Informatique,  
BP 68M, Oued Smar,  
16309, El Harrach,  
Alger, Algeria  
e-mail: [r\\_ait\\_amrane@esi.dz](mailto:r_ait_amrane@esi.dz), [raitamrane@gmail.com](mailto:raitamrane@gmail.com)

**Hacène Belbachir**

USTHB/ Faculté de Mathématiques,  
BP 32, El Alia,  
16111 Bab Ezzouar,  
Alger, Algeria  
e-mail: [hbelbachir@usthb.dz](mailto:hbelbachir@usthb.dz), [hacenebelbachir@gmail.com](mailto:hacenebelbachir@gmail.com)



# Conditions for groups whose group algebras have minimal Lie derived length\*

Zsolt Balogh<sup>a</sup>, Tibor Juhász<sup>b</sup>

<sup>a</sup>Institute of Mathematics and Informatics, College of Nyíregyháza  
e-mail: baloghzs@nyf.hu

<sup>b</sup>Institute of Mathematics and Informatics, Eszterházy Károly College  
e-mail: juhaszti@ektf.hu

*Submitted 19 November 2010; Accepted 17 December 2010*

*Dedicated to professor Béla Pelle on his 80<sup>th</sup> birthday*

## Abstract

Two independent research yielded two different characterizations of groups whose group algebras have minimal Lie derived lengths. In this note we show that the two characterizations are equivalent and we propose a simplified description for these groups.

## 1. Introduction

Let  $FG$  be the group algebra of a group  $G$  over a field  $F$ . As every associative algebra,  $FG$  can be viewed as a Lie algebra with the Lie multiplication defined by  $[x, y] = xy - yx$ , for all  $x, y \in FG$ . Let  $\delta^{[0]}(FG) = \delta^{(0)}(FG) = FG$ , and for  $n \geq 0$  denote by  $\delta^{[n+1]}(FG)$  the  $F$ -subspace of  $FG$  spanned by all elements  $[x, y]$  with  $x, y \in \delta^{[n]}(FG)$ , and by  $\delta^{(n+1)}(FG)$  the associative ideal of  $FG$  generated by all elements  $[x, y]$  with  $x, y \in \delta^{(n)}(FG)$ . We say that  $FG$  is Lie solvable (resp. strongly Lie solvable) if there exists  $n$  such that  $\delta^{[n]}(FG) = 0$  (resp.  $\delta^{(n)}(FG) = 0$ ), and the least such  $n$  is called the Lie derived length (resp. strong Lie derived length) of  $FG$  and denoted by  $dl_L(FG)$  (resp.  $dl^L(FG)$ ).

Sahai [6] proved that

$$\omega(FG')^{2^n-1}FG \subseteq \delta^{(n)}(FG) \subseteq \omega(FG')^{2^{n-1}}FG \text{ for all } n > 0, \quad (1.1)$$

---

\*This research was supported by NKTH-OTKA-EU FP7 (Marie Curie action) co-funded grant No. MB08A-82343

from which it follows that  $FG$  is strongly Lie solvable if and only if either  $G$  is abelian or the augmentation ideal  $\omega(FG')$  of the subalgebra  $FG'$  is nilpotent, that is the derived subgroup  $G'$  of  $G$  is a finite  $p$ -group and  $\text{char}(F) = p$ . Obviously,  $\delta^{[n]}(FG) \subseteq \delta^{(n)}(FG)$  for all  $n$ , thus every strongly Lie solvable group algebra  $FG$  is Lie solvable too, and  $\text{dl}_L(FG) \leq \text{dl}^L(FG)$ . However, according to a result of Passi, Passman and Sehgal (see e.g. in [5]), there exists a Lie solvable group algebra which is not strongly Lie solvable. They proved that a group algebra  $FG$  is Lie solvable if and only if one of the following conditions holds: (i)  $G$  is abelian; (ii)  $G'$  is a finite  $p$ -group and  $\text{char}(F) = p$ ; (iii)  $G$  has a subgroup of index 2 whose derived subgroup is a finite 2-group and  $\text{char}(F) = 2$ . Note that for  $\text{char}(F) = 2$  the values of  $\text{dl}_L(FG)$  and  $\text{dl}^L(FG)$  can be different (see e.g. Corollary 1 of [1]).

Evidently, if  $FG$  is commutative, then  $\text{dl}_L(FG) = \text{dl}^L(FG) = 1$ . Shalev [8] proved that if  $FG$  is a non-commutative Lie solvable group algebra of characteristic  $p$ , then  $\text{dl}_L(FG) \geq \lceil \log_2(p+1) \rceil$ , where  $\lceil \log_2(p+1) \rceil$  denotes the upper integer part of  $\log_2(p+1)$ . Shalev also showed that there is no better lower bound than  $\lceil \log_2(p+1) \rceil$ , emphasizing that the complete characterization of groups for which this lower bound is exact “*may be a delicate task*”. Clearly, for a non-commutative strongly Lie solvable group algebra  $FG$  the value of  $\text{dl}^L(FG)$  can also be estimated from below by the same integer  $\lceil \log_2(p+1) \rceil$ , and the question of characterizing groups for which this bound is achieved can be posed. Since we conjecture there is no group algebra  $FG$  over a field  $F$  of characteristic  $p > 2$  such that  $\text{dl}_L(FG) \neq \text{dl}^L(FG)$ , we may expect that the answer will solve Shalev’s original problem.

Levin and Rosenberger (see e.g. in [5]) described the group algebras of Lie derived length two. Moreover, they also proved that  $\text{dl}_L(FG) = 2$  if and only if  $\text{dl}^L(FG) = 2$ . This answers both questions for the special cases  $p = 2$  and 3. Assume that  $p \geq 5$  and  $G'$  has order  $p^n$ . As it is well-known  $\omega(FG')^{n(p-1)} \neq 0$ , furthermore there exists an integer  $i$  such that  $p < 2^i \leq 2p - 1$ . Hence, for  $n \geq 2$  we have

$$0 \neq \omega(FG')^{n(p-1)} \subseteq \omega(FG')^{2p-2} \subseteq \omega(FG')^{2^i-1},$$

and by (1.1),  $\text{dl}^L(FG) \geq i + 1 > \lceil \log_2(p+1) \rceil$ . Let now  $n = 1$ , that is  $G'$  is of order  $p$ , and denote by  $C_G(G')$  the centralizer of  $G'$  in  $G$ . In view of Theorem 1 of [1] (in which the authors determined the Lie derived length and the strong Lie derived length of group algebras of groups whose derived subgroup is cyclic of odd order) the value of  $\text{dl}^L(FG)$  depends on the order of the factor group  $G/C_G(G')$  as follows. For  $m \geq 0$ , let

$$s(l, m) = \begin{cases} 1 & \text{if } l = 0; \\ 2s(l-1, m) + 1 & \text{if } s(l-1, m) \text{ is divisible by } 2^m; \\ 2s(l-1, m) & \text{otherwise.} \end{cases}$$

If  $G/C_G(G')$  has order  $2^m p^r$ , then  $\text{dl}^L(FG) = d + 1$ , where  $d$  is the minimal integer for which  $s(d, m) \geq p$ ; otherwise  $\text{dl}^L(FG) = \lceil \log_2(2p) \rceil > \lceil \log_2(p+1) \rceil$ . Hence we have obtained the following criterion for groups whose group algebras have minimal strong Lie derived length.



**Theorem 1.1** (Balogh, Juhász [1]). *Let  $FG$  be a strongly Lie solvable group algebra of positive characteristic  $p$ . Then  $dl^L(FG) = \lceil \log_2(p+1) \rceil$  if and only if one of the following conditions holds:*

- (i)  $p = 2$  and  $G'$  is central elementary abelian subgroup of order 4;
- (ii)  $G'$  has order  $p$ ,  $G/C_G(G')$  has order  $2^m p^r$ , and the minimal integer  $d$  such that  $s(d, m) \geq p$  satisfies the inequality  $2^d - 1 < p$ .

An alternative characterization of these groups is obtained independently in [9] by using a different method. For  $m \geq 0$  let

$$g(0, m) = 1, \quad \text{and} \quad g(l, m) = g(l-1, m) \cdot 2^{m+1} + 1 \quad \text{for all } l \in \mathbb{N};$$

further, denote by  $q_{n-m, m}$  and  $\epsilon_{n-m, m}$  the quotient and the remainder of the Euclidean division of  $n - m - 1$  by  $m + 1$ , respectively.

**Theorem 1.2** (Spinelli [9]). *Let  $FG$  be a non-commutative strongly Lie solvable group algebra over a field  $F$  of positive characteristic  $p$ . Let  $n$  be the positive integer such that  $2^n \leq p < 2^{n+1}$  and  $s, q$  ( $q$  odd) the non-negative integers such that  $p - 1 = 2^s q$ . The following statements are equivalent:*

- (i)  $dl^L(FG) = \lceil \log_2(p+1) \rceil$ ;
- (ii)  $p$  and  $G$  satisfy one of the following conditions:
  - (a)  $p = 2$ ,  $G'$  has exponent 2 and an order dividing 4 and  $G'$  is central;
  - (b)  $p \geq 3$  and  $G'$  is central of order  $p$ ;
  - (c)  $5 \leq p < 2^{n+2}/3$ ,  $G'$  is not central of order  $p$  and  $|G/C_G(G')| = 2^m$  with  $m \leq s$  a positive integer such that  $p \leq 2^{\epsilon_{n-m, m}} \cdot g(q_{n-m, m} + 1, m)$ .

In the present paper the authors are going to dispel doubts about that the different conditions of the two above theorems could describe different classes of groups. We give a direct proof of the equivalence between them. According to [10], these same conditions describe completely the groups whose group algebras have minimal Lie derived length. Combining our results with the main theorem of [10], we propose the following simplified answer to Shalev's question.

**Theorem 1.3.** *Let  $FG$  be a Lie solvable group algebra over a field  $F$  of positive characteristic  $p$ . Then the following statements are equivalent:*

- (i)  $dl_L(FG) = \lceil \log_2(p+1) \rceil$ ;
- (ii)  $dl^L(FG) = \lceil \log_2(p+1) \rceil$ ;
- (iii) either  $p = 2$  and  $G'$  is central elementary abelian subgroup of order 2 or 4; or  $G'$  has order  $p > 2$ ,  $|G/C_G(G')| = 2^m$  and  $\lceil \log_2(p+1) \rceil = \lceil \log_2(\frac{2^{m+1}-1}{2^m} p) \rceil$ .

## 2. Proof of the equivalence

In the next lemma we concentrate on the series  $s(l, m)$  and  $g(l, m)$ , and on the connection between them.

**Lemma 2.1.** *For all  $m, n, i \geq 0$ ,*

$$(i) \quad 2^i \leq s(i, m) < 2^{i+1};$$

$$(ii) \quad s(i, m+1) \leq s(i, m);$$

$$(iii) \quad g(i, m) = s((m+1)i, m);$$

$$(iv) \quad s(n, m) = 2^{\epsilon_{n-m,m}} \cdot g(q_{n-m,m} + 1, m);$$

$$(v) \quad s(i, m) = \frac{2^{m+i+1} - 2^{(m+1)\{\frac{i}{m+1}\}}}{2^{m+1} - 1}, \text{ where } \{\frac{i}{m+1}\} \text{ is the fractional part of } \frac{i}{m+1}.$$

**Proof.**

(i) This is obvious for  $i = 0$ , and assume that  $2^i \leq s(i, m) < 2^{i+1}$ , or equivalently,  $2^{i+1} \leq 2s(i, m) < 2^{i+2}$  for some  $i \geq 0$ . Moreover,  $2s(i, m)$  is even, so  $2^{i+1} \leq 2s(i, m) < 2s(i, m) + 1 < 2^{i+2}$ . By definition,

$$2s(i, m) \leq s(i+1, m) \leq 2s(i, m) + 1$$

and the statement (i) is true.

(ii) For a fixed  $m$  assume that  $l$  is the minimal integer for which  $s(l, m+1) > s(l, m)$ . Then we get that  $2s(l-1, m+1) \geq s(l, m) \geq 2s(l-1, m)$ . Being  $l$  minimal  $s(l-1, m) = s(l-1, m+1)$ . Since  $s(l-1, m)$  cannot be divisible by  $2^{m+1}$  so

$$s(l, m) \geq 2s(l-1, m) = 2s(l-1, m+1) = s(l, m+1)$$

which is a contradiction.

(iii) For  $i = 0$  the definitions say that  $g(0, m) = s(0, m) = 1$ . Assume that  $i \geq 0$  and  $g(i, m) = s((m+1)i, m)$ . Then

$$g(i+1, m) = g(i, m) \cdot 2^{m+1} + 1 = s((m+1)i, m) \cdot 2^{m+1} + 1.$$

Since  $g(j, m)$  is odd for all  $j$ , we conclude that  $s((m+1)j, m)$  is also odd. Using the definition we get that

$$s((m+1)i, m) \cdot 2^{m+1} + 1 = s((m+1)(i+1), m)$$

and the proof is complete.

(iv) According to the definition,  $s(i, m)$  is odd whenever  $i$  is divisible by  $m+1$ , and

$$\begin{aligned} s(n, m) &= s((m+1)(q_{n-m,m} + 1) + \epsilon_{n-m,m}, m) \\ &= s((m+1)(q_{n-m,m} + 1), m) \cdot 2^{\epsilon_{n-m,m}}, \end{aligned}$$

and by (iii),

$$2^{\epsilon_{n-m,m}} \cdot s((m+1)(q_{n-m,m}+1), m) = 2^{\epsilon_{n-m,m}} \cdot g(q_{n-m,m}+1, m).$$

(v) Denote by  $q$  and  $r$  the quotient and the remainder of the Euclidean division of  $i$  by  $m+1$ , respectively. It is easy to check that

$$\begin{aligned} s(i, m) &= 2^{q(m+1)+r} + 2^{(q-1)(m+1)+r} + \dots + 2^r \\ &= 2^r \sum_{j=0}^q (2^{m+1})^j = \frac{2^{(m+1)(q+1)+r} - 2^r}{2^{m+1} - 1}. \end{aligned}$$

Using  $i = q(m+1) + r$  and  $r = (m+1)\{\frac{i}{m+1}\}$  we have the desired formula.  $\square$

Let  $G$  be a group with derived subgroup of order  $p$ . As it is well-known, the automorphism group of  $G'$  is isomorphic to the unit group of the field of  $p$  elements. But this unit group is cyclic of order  $p-1$ , so the factor group  $G/C_G(G')$ , which is isomorphic to a subgroup of it, is cyclic and its order divides  $p-1$ .

**Proof of the equivalence.** Denote by  $\mathfrak{A}$  the set of all groups  $G$  which satisfy the conditions (ii) of Theorem 1.2; by  $\mathfrak{B}$  those for which (i) or (ii) of Theorem 1.1 hold. Assume first that  $G \in \mathfrak{A}$ . We distinguish the following cases.

1.  $G'$  is central elementary abelian subgroup of order 4. Then by Theorem 1.1(i),  $G \in \mathfrak{B}$ .
2.  $G'$  is central of order  $p$ . Then the factor group  $G/C_G(G')$  is trivial, and  $s(i, 0) = 2^{i+1} - 1$ . It is clear that the minimal integer  $d$  such that  $2^{d+1} - 1 \geq p$  satisfies the inequality  $2^d - 1 < p$ , therefore  $G \in \mathfrak{B}$  in this case.
3.  $G'$  is not central of order  $p$ . Suppose that  $2^n \leq p < 2^{n+1}$ . Then, by Theorem 1.2(ii/c),  $|G/C_G(G')| = 2^m$  with a positive integer  $m$  such that

$$p \leq 2^{\epsilon_{n-m,m}} \cdot g(q_{n-m,m}+1, m).$$

According to Lemma 2.1(iv),  $s(n, m) = 2^{\epsilon_{n-m,m}} \cdot g(q_{n-m,m}+1, m)$ , hence  $p \leq s(n, m)$ . At the same time,  $2^n \leq p < 2^{n+1}$ , so by Lemma 2.1(i) we have that  $n$  is the minimal integer such that  $p \leq s(n, m)$ , and since  $2^n - 1 < p$ , it follows that  $G \in \mathfrak{B}$ .

We have just shown that  $\mathfrak{A} \subseteq \mathfrak{B}$ . To prove the converse inclusion we consider the following cases.

1.  $G'$  is central elementary abelian subgroup of order 4. Then by part (a) of Theorem 1.2(ii),  $G$  also belongs to  $\mathfrak{A}$ .
2.  $G'$  is cyclic of order  $p$ . Then by the assumption  $|G/C_G(G')| = 2^m p^r$ , but  $|G/C_G(G')|$  must divide  $p-1$ , actually  $r$  is always zero, and if  $s, q$  ( $q$  is odd) are the non-negative integers such that  $p-1 = 2^s q$ , then  $m \leq s$ .

- (a)  $m = 0$ . Then  $G'$  is central, and by parts (a) and (b) of Theorem 1.2, we have  $G \in \mathfrak{A}$ .
- (b)  $m > 0$ . Then  $G'$  is not central and  $p$  is odd. Assume that the minimal integer  $d$  such that  $s(d, m) \geq p$  satisfies the inequality  $2^d - 1 < p$ . It follows that  $2^d \leq p < 2^{d+1}$ , so  $n = d$ . By Lemma 2.1(iv),

$$p \leq s(n, m) = 2^{\epsilon_n - m, m} \cdot g(q_{n-m, m} + 1, m).$$

Furthermore, Lemma 2.1(ii) yields  $p \leq s(n, m) \leq s(n, 1) < 2^{n+2}/3$ . Finally, we show that  $p \geq 5$ . Indeed, if  $p$  was equal to 3, then  $m$  should be equal to 1, and from  $s(d, 1) \geq 3$  it would follow that  $d = 2$ . But  $2^2 - 1 \not\leq 3$ , so this is an impossible case.

The proof is done. □

### 3. Remarks

First we mention that we can get rid of the recursive sequence  $s(l, m)$  in Theorem 1 of [1]. Indeed, assume that  $|G'| = p^n$ , where  $p$  is an odd prime,  $|G/C_G(G')| = 2^m p^r$  and  $d$  is the minimal integer for which  $s(d, m) \geq p^n$ . By Lemma 2.1(v), we have

$$\frac{2^{m+d} - 2^{(m+1)\{\frac{d-1}{m+1}\}}}{2^{m+1} - 1} < p^n \leq \frac{2^{m+d+1} - 2^{(m+1)\{\frac{d}{m+1}\}}}{2^{m+1} - 1}.$$

Since  $(m+1)\{\frac{d-1}{m+1}\}, (m+1)\{\frac{d}{m+1}\} \in \{0, 1, \dots, m\}$ , so

$$\frac{2^{m+d} - 1}{2^{m+1} - 1} < p^n \leq \frac{2^{m+d+1} - 1}{2^{m+1} - 1},$$

and

$$d < \log_2 \left( \frac{2^{m+1} - 1}{2^m} p^n + \frac{1}{2^m} \right) \leq d + 1.$$

Keeping in mind that  $d$  is an integer, we conclude that

$$d + 1 = \lceil \log_2 \left( \frac{2^{m+1} - 1}{2^m} p^n + \frac{1}{2^m} \right) \rceil = \lceil \log_2 \left( \frac{2^{m+1} - 1}{2^m} p^n \right) \rceil.$$

Now, we can restate our Theorem 1 of [1] as follows.

**Theorem 3.1.** *Let  $G$  be a group with cyclic derived subgroup of order  $p^n$ , where  $p$  is an odd prime, and let  $F$  be a field of characteristic  $p$ . If  $G/C_G(G')$  has order  $2^m p^r s$ , where  $(2p, s) = 1$ , then*

$$dl_L(FG) = dl^L(FG) = \lceil \log_2 2p^n \nu_m \rceil,$$

where  $\nu_m = 1$  if  $s > 1$ , otherwise  $\nu_m = 1 - \frac{1}{2^{m+1}}$ .

This implies Theorem 1.3.

Let now  $FG$  be a group algebra over a field  $F$  of positive characteristic  $p$  with Lie (or strong Lie) derived length  $n$ . Then  $p < 2^n$ , furthermore,  $p \geq 2^{n-1}$  if and only if (iii) of Theorem 1.3 holds. Using this fact, we make an attempt to give a characterization of group algebras of Lie derived length 3 over a field of characteristic  $p > 3$ . As we told above,  $p$  must be smaller than 8, and for the cases  $p = 5$  and 7 (iii) of Theorem 1.3 must hold. It is easy to check that only the following  $(p, m)$  pairs are possible:  $(7, 0), (5, 0), (5, 1)$ . This proves the following statement.

**Corollary 3.2.** *Let  $FG$  be the group algebra of a group  $G$  over a field  $F$  of characteristic  $p > 3$ . Then  $\text{dl}_L(FG) = 3$  if and only if one of the following conditions holds: (i)  $p = 7$  and  $G'$  is central of order 7; (ii)  $p = 5$ ,  $G'$  has order 5, and either  $G'$  is central or  $x^g = x^{-1}$  for every  $x \in G'$  and  $g \notin C_G(G')$ .*

For an alternative proof and for the case  $p = 3$  we refer the reader to [6, 7].

Finally, we would like to draw reader's attention to recent articles [2, 3, 4] about Lie derived lengths of group algebras.

## References

- [1] BALOGH, Zs., JUHÁSZ, T., Lie derived lengths of group algebras of groups with cyclic derived subgroup, *Commun. Alg.*, 36 (2008), no. 2, 315–324.
- [2] BALOGH, Zs., JUHÁSZ, T., Derived lengths of symmetric and skew symmetric elements in group algebras, *JP J. Algebra Number Theory Appl.*, 12 (2008), no. 2, 191–203.
- [3] BALOGH, Zs., JUHÁSZ, T., Derived lengths in group algebras, *Proceedings of the International Conference on Modules and Representation Theory, Presa Univ. Clujeană, Cluj-Napoca*, (2009), 17–24.
- [4] BALOGH, Zs., JUHÁSZ, T., Remarks on the Lie derived lengths of group algebras of groups with cyclic derived subgroup, *Ann. Math. Inform.*, 34 (2007), 9–16.
- [5] BOVDI, A., The group of units of a group algebra of characteristic  $p$ , *Publ. Math. (Debrecen)*, 52 (1998), no. 1-2, 193–244.
- [6] SAHAI, M., Lie solvable group algebras of derived length three, *Publ. Mat. (Debrecen)*, 39 (1995), no. 2, 233–240.
- [7] SAHAI, M., Group algebras which are Lie solvable of derived length three. *J. Algebra Appl.*, 9 (2010), no. 2, 257–266.
- [8] SHALEV, A., The derived length of Lie soluble group rings, I. *J. Pure Appl. Algebra*, 78 (1992), no. 3, 291–300.
- [9] SPINELLI, E., Group algebras with minimal strong Lie derived length, *Canad. Math. Bull.*, 51 (2008), no. 2, 291–297.
- [10] SPINELLI, E., Group algebras with minimal Lie derived length, *J. Algebra*, 320 (2008), 1908–1913.

**Zsolt Balogh**

Institute of Mathematics and Informatics  
College of Nyíregyháza  
H-4410 Nyíregyháza  
Sóstói út 31/B  
Hungary

**Tibor Juhász**

Institute of Mathematics and Informatics  
Eszterházy Károly College  
H-3300 Eger  
Leányka út 4  
Hungary

# Regularity of certain Banach valued stochastic processes

A. L. Barrenechea

UNCPBA - FCExactas  
Dpto. de Matemáticas - NUCOMPA, Argentina

*Submitted 9 March 2010; Accepted 15 November 2010*

## Abstract

We consider random processes defined on Banach sequence spaces. Then we seek on conditions of  $\mathcal{M}$ -regularity of bounded linear operators, where  $\mathcal{M}$  denotes any of the usual stochastic modes of convergence.

*Keywords:* Random process on Banach sequence spaces. Stochastic modes of convergence. Locally finite bounded coverings.

*MSC:* 62L10, 65B99.

## 1. Introduction

Non deterministic systems derived from applications of probability theory to a wide real life situations give rise to the investigation of stochastic (or random) processes. This setting allows a quote of indeterminacy that reasonably must be considered according to the way the underlying process evolves in time. Among other basic examples, Markov processes concern to possibly dependent random variables, while Poisson processes concern events that occur continuously and independent of one another (cf. [7]).

Tests or experiments observed in discrete times amount to sequences of random variables. The problematic of convergence acceleration methods has been studied for many years with broad applications to numerical integration, to informatics, in solving differential equations, etc. (cf. [15, 2]). Sequence transformations and extrapolations were applied in order to accelerate the convergence of sequences in some well known statistical procedures, for instance bootstrap or jackknife (cf. [5, 4]).

The notion of stochastic regularity under the action of linear transformations applied to sequences of random elements in a Banach space was introduced by

H. Lavastre in 1995 (see [6]). His approach was very general, considering sequences  $\{X_n\}_{n=1}^{\infty}$  of random variables on a fixed probability space  $(\Omega, \mathcal{A}, \mathbb{P})$  with values in a Banach space  $(\mathbb{E}, \|\cdot\|)$ . Any such sequence induces a map

$$X: w \rightarrow \{X_n(w)\}_{n=1}^{\infty}$$

of  $\Omega$  into the set  $\mathcal{S}(\mathbb{E})$  of all sequences of elements of  $\mathbb{E}$ . Let us suppose that  $\mathcal{S}(\mathbb{E})$  is a normed space and that  $X$  is a *generalized random variable*, i.e.  $X^{-1}(B) \in \mathcal{A}$  if  $B$  is any Borel subset of  $\mathbb{E}$ . Given a linear functional  $T$  on  $\mathcal{S}(\mathbb{E})$  it is natural to ask whether  $T(X): w \rightarrow T[\{X_n(w)\}_{n=1}^{\infty}]$  is still a generalized random variable. If this is the case, the preservation of stochastic modes of convergence led to several notions of *stochastic regularity* of the sequence  $\{X_n\}_{n=1}^{\infty}$  under the action of  $T$ . From a theoretic point of view, besides its applications the determination of conditions of stochastic regularity has its own interest. For the resolution of this problem for  $\mathbb{E} = \mathcal{L}^p(\Omega, \mathbb{F}, \mathbb{P})$ , where  $1 \leq p < \infty$  and  $\mathbb{F}$  is a Banach space, the reader can see [6, Th. III, 3, p. 480]. Further, stochastic regularity under the action of certain linear transformations defined by some infinite triangular matrices of complex numbers is established in [6, Th. III, 6 and Th. III, 7, p. 482].

The purpose of this article is to initiate an extension of Lavastre's research to stochastic processes in other Banach spaces. Nevertheless, we are aware that this goal is easy to state as well as difficult to fulfil. So, we will restrict its generality to the case of bounded linear operators acting on separable Banach sequence spaces. In order to be self-contained in Prop. 2.1 we will show that the set of random variables  $X: \Omega \rightarrow \mathbb{E}$  between a probability space  $(\Omega, \mathcal{A}, \mathbb{P})$  and a separable Banach space  $\mathbb{E}$  admits a complex vector space structure. It is known that if  $\mathbb{E}$  is separable and  $X: \Omega \rightarrow \mathbb{E}$  is a random variable then  $\|X\|: \Omega \rightarrow [0, \infty)$  is a random variable (cf. [8]). Prop. 2.2 and Corollary 2.3 will motivate Definition 3.1 in Section 3, giving a precise meaning to random processes defined by a sequence of random variables on a Banach space  $\mathbb{E}$ . In this section we will analyze some concrete examples constructed on an underlying Hilbert space or on a Banach space of continuous functions (see Ex. 3.3 and Ex. 3.4 below). In Section 4 we consider conditions of stochastic regularity of linear bounded operators acting on a Banach sequence space  $\mathcal{S}(\mathbb{E})$ . In particular, we will observe in Remark 3.2 that our approach is more general than the so *called summation process* defined in [6]. In §4.1 we will establish precise conditions of stochastic regularity related to rather general bounded operators, when  $\mathbb{E} = \mathbb{C}$  and  $\mathcal{S}(\mathbb{E})$  is the uniform Banach space of convergent sequences of complex numbers  $c(\mathbb{C})$ . Finally, in §4.2 we will establish conditions of stochastic regularity of a class of bounded operators for the Banach space  $C[0, 1]$  and the Banach sequence space  $l^p(C[0, 1])$ , with  $1 < p < \infty$ .

Besides some posed questions, we believe that possible ways for further investigations will be open. In order of generality, the former will require some knowledge about the structure of bounded linear operators on Banach sequence spaces. Among a huge literature in this topic we only mention [1, 10, 9].



## 2. Random variables and Banach sequence spaces

Throughout this article  $(\Omega, \mathcal{A}, \mathbb{P})$  will be a probability space,  $(\mathbb{E}, \|\cdot\|)$  will be a separable Banach space and  $\mathfrak{X}$  will be a topological space. By  $\mathcal{M}_{\mathbb{P}}(\Omega, \mathcal{A}, \mathfrak{X})$  we will denote the class of *random variables*  $X: \Omega \rightarrow \mathfrak{X}$ , i.e. those functions so that  $X^{-1}(B) \in \mathcal{A}$  for all sets  $B \in \mathfrak{B}(\mathfrak{X})$ , where  $\mathfrak{B}(\mathfrak{X})$  is the class of Borel subsets of  $\mathfrak{X}$ . Indeed,  $\mathcal{M}_{\mathbb{P}}(\Omega, \mathcal{A}, \mathfrak{X})$  is really the quotient of all such random variables when we identify those that differ on a set of  $\mathbb{P}$ -measure zero.

**Proposition 2.1.** *If the Banach space  $(\mathbb{E}, \|\cdot\|)$  is separable then  $\mathcal{M}_{\mathbb{P}}(\Omega, \mathcal{A}, \mathbb{E})$  is a complex vector space.*

**Proof.** Clearly  $\mathcal{M}_{\mathbb{P}}(\Omega, \mathcal{A}, \mathbb{E})$  is endowed with a natural complex vector space structure, and it only remains to see that this structure is valid. Let  $\{f_n\}_{n=1}^{\infty}$  be a dense sequence of elements of  $\mathbb{E}$ . Then any open subset  $\mathcal{O}$  of  $\mathbb{E} \times \mathbb{E}$  can be written as

$$\mathcal{O} = \bigcup_{(n,m,r) \in \mathbb{N} \times \mathbb{N} \times \mathbb{Q}_{>0}: \mathbb{B}_{\infty}((f_n, f_m), r) \subseteq \mathcal{O}} \mathbb{B}_{\infty}((f_n, f_m), r),$$

where for  $(n, m, r) \in \mathbb{N} \times \mathbb{N} \times \mathbb{Q}_{>0}$  is

$$\mathbb{B}_{\infty}((f_n, f_m), r) = \{(g, h) \in \mathbb{E} \times \mathbb{E} : \max\{\|f_n - g\|, \|f_m - h\|\} < r\}.$$

So, if  $X_1, X_2 \in \mathcal{M}_{\mathbb{P}}(\Omega, \mathcal{A}, \mathbb{E})$  the set  $(X_1, X_2)^{-1}(\mathcal{O})$  is realized as

$$\bigcup_{(n,m,r) \in \mathbb{N} \times \mathbb{N} \times \mathbb{Q}_{>0}: \mathbb{B}_{\infty}((f_n, f_m), r) \subseteq \mathcal{O}} X_1^{-1}(\mathbb{B}(f_n, r)) \cap X_2^{-1}(\mathbb{B}(f_m, r)),$$

i.e.  $(X_1, X_2)^{-1}(\mathcal{O}) \in \mathcal{A}$ . Hence  $(X_1, X_2) \in \mathcal{M}_{\mathbb{P}}(\Omega, \mathcal{A}, \mathbb{E} \times \mathbb{E})$ . Since  $\mathbb{E}$  is a topological vector space the conclusion now follows immediately.  $\square$

**Proposition 2.2.** *Let  $\{X_n\}_{n=1}^{\infty} \subseteq \mathcal{M}_{\mathbb{P}}(\Omega, \mathcal{A}, \mathbb{E})$ .*

(i) *Let us write*

$$\begin{aligned} \Omega_{\mathbb{E}}^{\infty}(\{X_n\}_{n=1}^{\infty}) &\triangleq \{w \in \Omega : \{X_n(w)\}_{n=1}^{\infty} \in l^{\infty}(\mathbb{N}, \mathbb{E})\}, \\ \Omega_{\mathbb{E}}^c(\{X_n\}_{n=1}^{\infty}) &\triangleq \{w \in \Omega : \{X_n(w)\}_{n=1}^{\infty} \in c(\mathbb{N}, \mathbb{E})\}, \\ \Omega_{\mathbb{E}}^{c_0}(\{X_n\}_{n=1}^{\infty}) &\triangleq \{w \in \Omega : \{X_n(w)\}_{n=1}^{\infty} \in c_0(\mathbb{N}, \mathbb{E})\}, \\ \Omega_{\mathbb{E}}^p(\{X_n\}_{n=1}^{\infty}) &\triangleq \{w \in \Omega : \{X_n(w)\}_{n=1}^{\infty} \in l^p(\mathbb{N}, \mathbb{E})\}, \end{aligned}$$

with  $1 \leq p < +\infty$ . The above sets are  $\mathcal{A}$ -measurable and

$$\Omega_{\mathbb{E}}^p(\{X_n\}_{n=1}^{\infty}) \subseteq \Omega_{\mathbb{E}}^{c_0}(\{X_n\}_{n=1}^{\infty}) \subseteq \Omega_{\mathbb{E}}^c(\{X_n\}_{n=1}^{\infty}) \subseteq \Omega_{\mathbb{E}}^{\infty}(\{X_n\}_{n=1}^{\infty}). \quad (2.1)$$

(ii) *If  $X_n \xrightarrow{a.e.} 0$  then  $\mathbb{P}(\Omega_{\mathbb{E}}^{c_0}(\{X_n\}_{n=1}^{\infty})) = 1$ .*

**Proof.** (i) It suffices to observe that

$$\begin{aligned}\Omega_{\mathbb{E}}^{\infty}(\{X_n\}_{n=1}^{\infty}) &= \bigcup_{m=1}^{\infty} \bigcap_{p=1}^{\infty} \{\|X_p\| \leq m\}, \\ \Omega_{\mathbb{E}}^c(\{X_n\}_{n=1}^{\infty}) &= \bigcap_{m=1}^{\infty} \bigcup_{p=1}^{\infty} \bigcap_{q \geq p, r \geq 0} \{\|X_q - X_{q+r}\| \leq 1/m\}, \\ \Omega_{\mathbb{E}}^{c_0}(\{X_n\}_{n=1}^{\infty}) &= \bigcap_{m=1}^{\infty} \liminf_{q \rightarrow \infty} \{\|X_q\| \leq 1/m\}.\end{aligned}$$

Further,

$$\Omega_{\mathbb{E}}^p(\{X_n\}_{n=1}^{\infty}) = \left\{ w \in \Omega : \sup_{m \in \mathbb{N}} \sum_{n=1}^m \|X_n(w)\|^p < +\infty \right\}$$

and  $\{\sum_{n=1}^m \|X_n\|^p\}_{m \in \mathbb{N}} \subseteq \mathcal{M}_P(\Omega, \mathcal{A}, \mathbb{R})$ . Thus  $\Omega_{\mathbb{E}}^p(\{X_n\}_{n=1}^{\infty}) \in \mathcal{A}$ , because  $\mathcal{M}_P(\Omega, \mathcal{A}, \mathbb{R})$  is an order complete vector space and  $\mathcal{A}$  is a  $\sigma$ -algebra. The inclusions (2.1) are trivial.

(ii) It is trivial.  $\square$

**Corollary 2.3.** *Let  $\{X_n\}_{n=1}^{\infty} \subseteq \mathcal{M}_P(\Omega, \mathcal{A}, \mathbb{E})$  so that  $X_n \xrightarrow{a.e.} 0$ . Then there are induced well defined random variables*

$$X^{c_0}(w) = \{X_n(w)\}_{n=1}^{\infty}, \quad X^c(w) = \{X_n(w)\}_{n=1}^{\infty}, \quad X^{\infty}(w) = \{X_n(w)\}_{n=1}^{\infty},$$

where  $w \in \Omega$ , with values in the Banach spaces  $c_0(\mathbb{N}, \mathbb{E})$ ,  $c(\mathbb{N}, \mathbb{E})$  and  $l_{\infty}(\mathbb{N}, \mathbb{E})$  respectively.

**Remark 2.4.** Convergence in probability is not appropriate in general to derive natural random variables with values in classical Banach sequence spaces. For instance, let  $n = k + 2^v$ ,  $0 \leq k < 2^v$ ,  $v \in \mathbb{N}_0$ , and set  $X_n = n\chi_{[k/2^v, (k+1)/2^v]}$ . The sequence  $\{X_n\}_{n=1}^{\infty}$  of random variables on the Lebesgue measure space  $[0, 1]$  converges in probability to zero and  $\Omega_{\mathbb{R}}^{\infty}(\{X_n\}_{n=1}^{\infty}) = \emptyset$ .

**Remark 2.5.** Previously to the main Definition 3.1 of this article, let us remember the usual stochastic modes of convergence:

1. *Convergence in distribution*

$X_n \xrightarrow{d} X$  if and only if given  $B \in \mathfrak{B}(\mathbb{E})$  so that  $P(\{X \in \partial B\}) = 0$  then  $P(\{X_n \in B\}) \rightarrow P(\{X \in B\})$ .

2. *Convergence in probability*

$X_n \xrightarrow{P} X$  if and only if  $\forall \varepsilon > 0$ ,  $P(\{\|X_n - X\| \geq \varepsilon\}) \rightarrow 0$ .

3. *Almost everywhere convergence*

$X_n \xrightarrow{a.e.} X$  if and only if  $P(\{X_n \rightarrow X\}) = 1$ .

4. *Almost complete convergence*

$X_n \xrightarrow{a.c.} X$  if and only if  $\forall \varepsilon > 0, \sum_{n=1}^{\infty} \mathbb{P}(\{\|X_n - X\| \geq \varepsilon\}) < +\infty$ .

5. *Convergence in the  $r$ -th mean*

$X_n \xrightarrow{L^r} X$  if and only if  $\mathbb{E}(\|X_n - X\|^r) \rightarrow 0$ .

6. *Convergence in the mean*

$X_n \xrightarrow{\mathbb{E}} X$  if and only if  $\mathbb{E}(X_n - X) \rightarrow 0$ . (See Remark 2.6 below).

It is well known that almost complete convergence implies almost everywhere convergence, almost everywhere convergence implies convergence in probability and convergence in probability implies convergence in distribution (cf. [12, pp. 240]). Likewise, if  $r > s$  then convergence in the  $r$ -th mean implies convergence in the  $s$ -th mean and the later implies convergence in probability. Further, by Lévy's convergence theorem if  $X_n \xrightarrow{a.e.} X$  in  $\mathcal{M}_{\mathbb{P}}(\Omega, \mathcal{A}, \mathbb{R})$  and there is a random variable  $Y$  so that for all  $n \in \mathbb{N}$  is  $|X_n| \leq Y$  and  $\mathbb{E}(Y) < +\infty$  then  $X_n \xrightarrow{L^r} X$  (see [14, pp. 187–188]).

**Remark 2.6.** If the Banach space  $\mathbb{E}$  is separable the notion of *expected value* of a random variable  $X \in \mathcal{M}_{\mathbb{P}}(\Omega, \mathcal{A}, \mathbb{E})$  is well defined. Precisely, given a random variable  $X$  its expected value is any element  $f \in \mathbb{E}$  so that if  $\varphi \in \mathbb{E}^*$  then

$$\langle f, \varphi \rangle = \int_{\Omega} \langle X(w), \varphi \rangle d\mathbb{P}(w).$$

Since  $\mathbb{E}^*$  becomes a separating family if such an element exists it is necessarily unique and it is denoting as  $\mathbb{E}(X)$ . For instance,  $\mathbb{E}(X)$  exists if  $\mathbb{E}(\|X\|) < +\infty$ . For further information the reader can see [11].

### 3. Random processes on Banach sequence spaces

**Definition 3.1.** A random process of  $\mathcal{M}_{\mathbb{P}}(\Omega, \mathcal{A}, \mathbb{E})$  on a Banach sequence space  $\mathcal{S}(\mathbb{E})$  is a sequence  $\{X_n\}_{n=1}^{\infty} \cup \{X\} \subseteq \mathcal{M}_{\mathbb{P}}(\Omega, \mathcal{A}, \mathbb{E})$  so that:

(i) the set

$$\Omega^{\mathcal{S}(\mathbb{E})}(\{X_n - X\}_{n=1}^{\infty}) \triangleq \{w \in \Omega : \{X_n(w) - X(w)\}_{n=1}^{\infty} \in \mathcal{S}(\mathbb{E})\}$$

belongs to  $\mathcal{A}$ ;

(ii)  $\mathbb{P}(\Omega^{\mathcal{S}(\mathbb{E})}(\{X_n - X\}_{n=1}^{\infty})) = 1$ . By  $[\mathcal{M}_{\mathbb{P}}(\Omega, \mathcal{A}, \mathbb{E}), \mathcal{S}(\mathbb{E})]$  we will denote the class of all such random processes.

**Remark 3.2.** By Prop. 2.2 any almost everywhere convergent sequence of random variables with values in a Banach space  $\mathbb{E}$  defines a random process on the classical Banach sequence spaces  $c_0(\mathbb{N}, \mathbb{E})$ ,  $c(\mathbb{N}, \mathbb{E})$  and  $l^{\infty}(\mathbb{N}, \mathbb{E})$ .

**Example 3.3.** Let  $1 \leq p < \infty$ ,  $T \in \mathcal{B}(\mathcal{L}^p[0, 1])$ . If  $n \in \mathbb{N}$  let  $X_n(t) = T^n(\chi_{[0,t]})$ ,  $0 \leq t \leq 1$ . If  $0 \leq s, t \leq 1$  then

$$\begin{aligned} \|X_n(t) - X_n(s)\|_p &= \|T^n(\chi_{[0,t]} - \chi_{[0,s]})\|_p \\ &\leq \|T^n\| \|\chi_{[0,t] \Delta [0,s]}\|_p \\ &\leq \|T\|^n |s - t|^{1/p}, \end{aligned}$$

i.e.  $X_n: [0, 1] \rightarrow \mathcal{L}^p[0, 1]$  becomes uniformly continuous and

$$\{X_n\}_{n=1}^\infty \subseteq \mathcal{M}_{dx}([0, 1], \mathfrak{L}[0, 1], \mathcal{L}^p[0, 1]),$$

where  $dx$  is the Lebesgue measure on  $[0, 1]$  and  $\mathfrak{L}[0, 1]$  is the Lebesgue  $\sigma$ -algebra of subsets of  $[0, 1]$ . For instance, let  $Tf(t) = \int_0^t f dx$  if  $f \in \mathcal{L}^p[0, 1]$ . It is easy to see that  $T$  is a bounded linear operator and if  $n \in \mathbb{N}$  and  $0 \leq t, \tau \leq 1$  then

$$X_n(t)(\tau) \triangleq T^n(\chi_{[0,t]})(\tau) = \begin{cases} (\tau^n - (\tau - t)^n)/n! & \text{if } 0 \leq t \leq \tau, \\ \tau^n/n! & \text{if } \tau \leq t \leq 1. \end{cases} \quad (3.1)$$

Consequently, if  $t \in [0, 1]$  and  $n \in \mathbb{N}$  the following inequality

$$\|X_n(t)\|_p \leq 1 / \left[ n! (1 + np)^{1/p} \right] \quad (3.2)$$

holds. From (3.2) we infer that  $X_n \xrightarrow{a.c.} 0$  and that  $\{X_n\}_{n=1}^\infty$  defines well random process on any of the classical Banach sequence spaces on  $\mathcal{L}^p[0, 1]$ . Further, if  $n \in \mathbb{N}$  from (3.1) we have that  $X_n: [0, 1] \rightarrow \mathbb{C}[0, 1]$  and

$$\|X_n(s) - X_n(t)\|_\infty = \max\{|s - t|^n, |(1 - t)^n - (1 - s)^n|\} / n!$$

if  $0 \leq s, t \leq 1$ , i.e.  $X_n$  is continuous and  $\{X_n\}_{n=1}^\infty \subseteq \mathcal{M}_{dx}([0, 1], \mathfrak{L}[0, 1], \mathbb{C}[0, 1])$ . Since

$$\|X_n(t)\|_\infty = (1 - (1 - t)^n) / n!$$

the same conclusions are true for the underlying Banach space  $\mathbb{C}[0, 1]$ . In this setting the sequence of random variables  $\{X_n\}_{n=1}^\infty$  converges to zero in the  $r$ -th mean for all  $r \in \mathbb{N}$ . For, if  $n \in \mathbb{N}$  and  $s \in \mathbb{R}$  we have

$$F_n(s) \triangleq \int_{\{\|X_n\|_\infty \leq s\}} dx = \begin{cases} 0 & \text{if } s \leq 0, \\ 1 - (1 - sn!)^{1/n} & \text{if } 0 < s < 1/n!, \\ 1 & \text{if } s \geq 1/n!. \end{cases} \quad (3.3)$$

In particular,  $d\text{-}\lim_{n \rightarrow \infty} \|X_n\|_\infty = \mathbb{H}$ , i.e. the sequence of random variables  $\{\|X_n\|_\infty\}_{n=1}^\infty$  converges in distribution to the Heaviside function. Now, using (3.3) we obtain

$$\mathbb{E}(\|X_n\|_\infty^r) = \int_0^{1/n!} s^r dF_n(s)$$

$$\begin{aligned}
 &= (n-1)! \int_0^{1/n!} s^r (1-sn!)^{1/n-1} ds \\
 &= \frac{1}{nn!^r} \int_0^1 u^r (1-u)^{1/n-1} du \\
 &= \frac{1}{nn!^r} \cdot \text{Be}(r+1, 1/n) \\
 &= \frac{1}{nn!^r} \cdot \frac{\Gamma(r+1)\Gamma(1/n)}{\Gamma(r+1+1/n)} \\
 &= \frac{r!}{nn!^r} \cdot \prod_{j=0}^r (1/n+j)^{-1} \leq \frac{r!}{(n-1)!^r},
 \end{aligned}$$

i.e.  $\lim_{n \rightarrow \infty} \mathbb{E}(\|X_n\|_\infty^r) = 0$ . Further, if  $n \in \mathbb{N}$  then

$$\mathbb{E}(X_n)(\tau) = \frac{\tau^n}{n!} - \frac{\tau^{n+1}}{(n+1)!}. \quad (3.4)$$

For, let  $\phi \in \text{BV}[0, 1]$  be a complex valued function of bounded variation on  $[0, 1]$ . By the Fubini-Tonelli theorem and (3.1) we see that

$$\begin{aligned}
 \iint_{[0,1] \times [0,1]} |X_n(t)(\tau)| d|\phi|(\tau) \times dt &= \int_0^1 \int_0^1 |X_n(t)(\tau)| d|\phi|(\tau) dt \\
 &\leq \int_0^1 d|\phi|(\tau) / n! \leq \|\phi\|_{\text{BV}[0,1]} < +\infty,
 \end{aligned}$$

where  $\|\phi\|_{\text{BV}[0,1]} \triangleq |\phi(0)| + \text{V}_{[0,1]}(\phi)$ . As it is well known  $(\text{BV}[0, 1], \|\cdot\|_{\text{BV}[0,1]})$  becomes a Banach space isometrically isomorphic to  $(\text{C}[0, 1])^*$  (cf. [3, Th. 1.37, p. 16]). Hence,

$$\begin{aligned}
 \left\langle \frac{\tau^n}{n!} - \frac{\tau^{n+1}}{(n+1)!}, d\phi(\tau) \right\rangle &= \int_0^1 \left( \frac{\tau^n}{n!} - \frac{\tau^{n+1}}{(n+1)!} \right) d\phi(\tau) \\
 &= \int_0^1 \left( \int_0^\tau \frac{\tau^n - (\tau-t)^n}{n!} dt + \frac{\tau^n}{n!} (1-\tau) \right) d\phi(\tau) \\
 &= \int_0^1 \int_0^1 X_n(t)(\tau) dt d\phi(\tau) \\
 &= \int_0^1 \int_0^1 X_n(t)(\tau) d\phi(\tau) dt \\
 &= \int_0^1 \langle X_n(t), d\phi \rangle dt.
 \end{aligned}$$

By the uniqueness of the expected value of  $X_n$  as it was pointed in Remark 2.6 we obtain (3.4). In particular,  $\mathbb{E}(X_n) \rightarrow 0$  in  $\text{C}[0, 1]$ .

**Example 3.4.** Let  $\Omega = \{00, 010, 0110, \dots\} \cup \{11, 101, 1001, \dots\}$  and if  $0 < p < 1$  let  $q = 1 - p$ . Given  $w \in \Omega$  we put  $P(w) = p^a q^b$  if  $w$  contains  $a$  zeros and  $b$  ones. Hence  $(\Omega, P)$  becomes a discrete probability space. For instance,  $\Omega$  can be seen as the set of all possible random events in a game consisting in throwing a possible non calibrated coin successively, assuming that the play ends when the first result occurs again. Let us consider a separable Hilbert space  $\mathcal{H}$  endowed with an orthonormal basis  $\{e_n\}_{n=1}^\infty$ . We can represent any element  $w \in \Omega$  as a sequence  $w = \{w_m\}_{m=1}^\infty$ , where  $w_m = 0$  except a possible finite number of indices. For instance, we write  $010 = \{0, 1, 0, 0, 0, \dots\}$ ,  $1001 = \{1, 0, 0, 1, 0, 0, 0, \dots\}$ , etc. Now, for  $w \in \Omega$  and  $n \in \mathbb{N}$  we will write  $Y_n(w) = \sum_{m=1}^n w_m \cdot e_m$ . Then  $\{Y_n\}_{n=1}^\infty \subseteq \mathcal{M}_P(\Omega, \mathcal{P}(\Omega), \mathcal{H})$ . Further, if for  $w \in \Omega$  we set

$$Y(w) = \sum_{m=1}^{\infty} w_m \cdot e_m \quad (3.5)$$

then  $Y: \Omega \rightarrow \mathcal{H}$  is a well defined random variable since any series in (3.5) is reduced to a finite sum. If  $X_n \triangleq Y_n - Y$ ,  $n \in \mathbb{N}$ , clearly  $\Omega_{\mathcal{H}}^P(\{X_n\}_{n \in \mathbb{N}}) = \Omega$ . Indeed,  $\{X_n\}_{n=1}^\infty$  converges to zero in the  $r$ -th mean for all  $r \in \mathbb{N}$ . For, if  $n \in \mathbb{N}$  then

$$\begin{aligned} P(\{\|X_n\| = 0\}) &= P\left(\left\{00, 010, \dots, 01 \dots 1 \overset{(n+1)}{0}, 11, 101, \dots, 10 \dots 0 \overset{(n)}{1}\right\}\right) \quad (3.6) \\ &= p^2 \sum_{j=0}^{n-1} q^j + q^2 \sum_{j=0}^{n-2} p^j \\ &= 1 - pq^n - p^{n-1}q, \\ P(\{\|X_n\| = 1\}) &= P\left(\left\{01 \dots \overset{(n)}{1} 10, 10 \dots \overset{(n)}{0} 1, 10 \dots \overset{(n)}{0} 01, \dots\right\}\right) \\ &= p^2 q^n + p^{n-1} q^2 + p^n q^2 + \dots \\ &= p^2 q^n + p^{n-1} q. \end{aligned}$$

For an integer  $m \geq 2$  we see that

$$P\left(\{\|X_n\| = m^{1/2}\}\right) = P\left(\left\{01 \dots \overset{(n)}{1} 1 \dots \overset{(n+m)}{1}\right\}\right) = p^2 q^{n+m-1}. \quad (3.7)$$

Using the identities (3.6) and (3.7) we evaluate

$$\begin{aligned} E(\|X_n\|^r) &= \sum_{m=0}^{\infty} m^{r/2} P\left(\{\|X_n\| = m^{1/2}\}\right) \quad (3.8) \\ &= p^2 q^n + p^{n-1} q + p^2 q^{n-1} \sum_{m=2}^{\infty} m^{r/2} q^m. \end{aligned}$$

Letting  $n \rightarrow \infty$  in (3.8) the claim follows. With the notation of Ex. 3.4 we will show that

$$\lim_{n \rightarrow \infty} \mathbb{E}(X_n) = 0. \quad (3.9)$$

For, we will prove that if  $n \in \mathbb{N}$  then

$$\mathbb{E}(X_n) = - \sum_{v=n+1}^{\infty} (pq^{v-1} + p^{v-2}q^2) e_v \quad (3.10)$$

and later (3.9) will follow at once. As  $0 < p, q < 1$  the above series is absolutely convergent. If  $g \in \mathcal{H}$  the random variable  $w \rightarrow \langle X_n(w), g \rangle$  maps  $\Omega$  onto the set  $\left\{ \sum_{s=1}^k \langle g, e_{n+s} \rangle \right\}_{k=1}^{\infty}$ . If  $m \in \mathbb{N}$  set  $\Omega_m = \{w \in \Omega : w_v = 0 \text{ if } v > m\}$ . Thus  $\{\Omega_m\}_{m=1}^{\infty}$  is an increasing sequence of sets and  $\Omega = \cup \Omega_m$ . If  $m \in \mathbb{N}$  and  $m > n$  we have

$$\begin{aligned} \int_{\Omega} \langle X_n(w), g \rangle \chi_{\Omega_m}(w) d\mathbb{P}(w) &= - \int_{\Omega_m} \sum_{v=n+1}^m w_v \langle e_v, g \rangle d\mathbb{P}(w) \quad (3.11) \\ &= - \sum_{s=1}^m \left\langle \sum_{t=1}^s e_{n+t}, g \right\rangle p^2 q^{n+s-1} \\ &\quad - \sum_{v=n+1}^m \langle e_v, g \rangle p^{v-2} q^2 \\ &= -p \sum_{t=1}^m \langle e_{n+t}, g \rangle (q^{n+t-1} - q^{n+m}) \\ &\quad - \sum_{v=n+1}^m \langle e_v, g \rangle p^{v-2} q^2. \end{aligned}$$

Since the series  $\sum_{m=1}^{\infty} q^m m^{1/2}$  converges we conclude that

$$0 \leq \limsup_{m \rightarrow \infty} q^{n+m} \sum_{t=1}^m |\langle e_{n+t}, g \rangle| \leq \limsup_{m \rightarrow \infty} q^{n+m} \|g\| m^{1/2} = 0. \quad (3.12)$$

From (3.11) and (3.12) we get

$$\begin{aligned} \lim_{m \rightarrow \infty} \int_{\Omega} \langle X_n(w), g \rangle \chi_{\Omega_m}(w) d\mathbb{P}(w) &= -p \sum_{t=1}^{\infty} \langle e_{n+t}, g \rangle q^{n+t-1} \quad (3.13) \\ &\quad - \sum_{v=n+1}^{\infty} \langle e_v, g \rangle p^{v-2} q^2 \\ &= - \sum_{v=n+1}^{\infty} \langle e_v, g \rangle (pq^{v-1} + p^{v-2}q^2) \end{aligned}$$

$$= \left\langle - \sum_{v=n+1}^{\infty} (pq^{v-1} + p^{v-2}q^2) e_v, g \right\rangle.$$

But for  $m \in \mathbb{N}$  and  $w \in \Omega$  we see that

$$|\langle X_n(w), g \rangle| \chi_{\Omega_m}(w) \leq |\langle X_n(w), g \rangle|. \quad (3.14)$$

Moreover,

$$\begin{aligned} \int_{\Omega} |\langle X_n(w), g \rangle| dP(w) &= |\langle g, e_{n+1} \rangle| P \left( \left\{ 10 \dots 0 \overset{(n+1)}{1}, 01 \dots \overset{(n+1)}{1} 0 \right\} \right) \quad (3.15) \\ &+ \sum_{k=2}^{\infty} \left| \sum_{s=1}^k \langle g, e_{n+s} \rangle \right| P \left( \left\{ 01 \dots \overset{(n+1)}{1} \dots \overset{(n+k)}{1} 0 \right\} \right) \\ &= |\langle g, e_{n+1} \rangle| (p^{n-1}q^2 + pq^n) \\ &+ \sum_{k=2}^{\infty} \left| \sum_{s=1}^k \langle g, e_{n+s} \rangle \right| pq^{n+k-1}. \end{aligned}$$

Further,

$$\sum_{k=1}^{\infty} \left| \sum_{s=1}^k \langle g, e_{n+s} \rangle \right| q^k \leq \|g\| \sum_{k=1}^{\infty} k^{1/2} q^k < +\infty. \quad (3.16)$$

Thus, by (3.15) and (3.16) the random variable  $w \rightarrow \langle X_n(w), g \rangle$  becomes absolutely integrable on  $\Omega$ . Finally, using (3.14) and the Lebesgue dominated convergence theorem in (3.13) we obtain

$$\int_{\Omega} \langle X(w), g \rangle dP(w) = \left\langle - \sum_{v=n+1}^{\infty} (pq^{v-1} + p^{v-2}q^2) e_v, g \right\rangle$$

and (3.10) follows.

## 4. Random processes and stochastic regularity

**Definition 4.1.** With the notation of Definition 3.1, let  $A \in \mathcal{B}[\mathcal{S}(\mathbb{E})]$ . Then  $A$  is called  $\mathcal{M}$ -regular for  $\{X_n - X\}_{n=1}^{\infty}$  on the Banach sequence space  $\mathcal{S}(\mathbb{E})$  if it preserves its  $\mathcal{M}$ -stochastic mode of convergence, i.e. if  $\mathcal{M}\text{-}\lim_{n \rightarrow \infty} X_n = X$  then  $\mathcal{M}\text{-}\lim_{m \rightarrow \infty} \|A_m(\{X_n - X\}_{n=1}^{\infty})\| = 0$ . A subset  $\mathcal{R}$  of  $\mathcal{S}(\mathbb{E})$  is called  $\mathcal{M}$ -regular for the sequence  $\{X_n - X\}_{n=1}^{\infty}$  on  $\mathcal{S}(\mathbb{E})$  if each element of  $\mathcal{R}$  is  $\mathcal{M}$ -regular for it. Indeed,  $\mathcal{R}$  will be called simply  $\mathcal{M}$ -regular on  $\mathcal{M}_P(\Omega, \mathcal{A}, \mathbb{E})$  and  $\mathcal{S}(\mathbb{E})$  if each element of  $\mathcal{R}$  preserves the  $\mathcal{M}$ -stochastic mode of convergence of any random process of  $[\mathcal{M}_P(\Omega, \mathcal{A}, \mathbb{E}), \mathcal{S}(\mathbb{E})]$ .

**Remark 4.2.** The well known shift operator  $W((f_n)_{n=1}^{\infty}) = (f_{n+1})_{n=1}^{\infty}$  is linear and bounded on any of the classical Banach sequences spaces  $\ell^p(\mathbb{N}, \mathbb{C})$ ,  $c_0(\mathbb{N}, \mathbb{C})$ ,



$c(\mathbb{N}, \mathbb{C})$  and  $l^\infty(\mathbb{N}, \mathbb{C})$ . For conditions concerning to the  $\mathcal{M}$ -regularity of  $p(W)$  when  $p$  is any polynomial the reader can see [6]. That approach could be improved in various directions, for instance: (1st) What can be said about the  $\mathcal{M}$ -regularity of general bounded operators on Banach sequence spaces over  $\mathbb{C}$ ? (2nd) What happens if we state the same problem replacing  $\mathbb{C}$  by any other Banach space? The first question already has its own interest since Banach sequence spaces of complex or real numbers offer a natural frame to modeling a huge variety of statistical and numerical analysis processes. Even in this case the determination of the structure and characterization of bounded operators sometimes constitute a difficult matter. In particular, the characterization of bounded operators on  $c(\mathbb{N}, \mathbb{C})$  is a celebrated result of I. Schur (cf. [13]). For more information on these topics the reader can see [9], [10]. For a proof of Schur's theorem and the characterization of bounded operators on Banach sequence spaces of complex series see [1].

#### 4.1. $\mathcal{M}$ -regularity on $[\mathcal{M}_P(\Omega, \mathcal{A}, \mathbb{C}), c(\mathbb{N}, \mathbb{C})]$

If  $A \in \mathcal{B}(c(\mathbb{N}, \mathbb{C}))$  there is a unique complex matrix  $\{a_{n,m}\}_{n,m=0}^\infty$  so that for  $z \in c(\mathbb{N}, \mathbb{C})$  we have

$$A(z) = \left\{ a_{n,0}\lambda(z) + \sum_{m=1}^\infty a_{n,m} \cdot z_m \right\}_{n=1}^\infty,$$

where  $\lambda(z) = \lim_{n \rightarrow \infty} z_n$ . Further,

$$\begin{aligned} \|A\| &= \sup_{n \in \mathbb{N}} \sum_{m=0}^\infty |a_{n,m}|, \\ a_{0,0} &= \lim_{n \rightarrow \infty} \sum_{m=1}^\infty a_{n,m}, \\ a_{0,m} &= \lim_{n \rightarrow \infty} a_{n,m} \text{ if } m \in \mathbb{N} \end{aligned} \tag{4.1}$$

and  $\{a_{0,m}\}_{m=1}^\infty \in l^1(\mathbb{N}, \mathbb{C})$  (cf. [1], Corollary 2, p. 20). Let us consider the random process on  $c(\mathbb{N}, \mathbb{C})$  induced by  $X_n = \chi_{[n, +\infty)}$ ,  $n \in \mathbb{N}$  on the probability space  $(\mathbb{R}, \mathcal{L}(\mathbb{R}), P)$ , where  $\mathcal{L}(\mathbb{R})$  is the class of Lebesgue measurable subsets of  $\mathbb{R}$  and  $P(E) = \int_{E \cap (0, +\infty)} \exp(-x) dx$  if  $E \in \mathcal{L}(\mathbb{R})$ . Let  $A \in \mathcal{B}(c(\mathbb{N}, \mathbb{C}))$  be defined by the infinite matrix whose nm-entry is

$$a_{n,m} = \begin{cases} 1 & \text{if } n = m = 0, \\ 0 & \text{if } n = 0, m \in \mathbb{N}, \\ (1+n)^{-m} & \text{if } n, m \in \mathbb{N}. \end{cases}$$

Then  $A$  is ac-regular for the sequence  $\{X_n\}_{n=1}^\infty$ . For, let  $\varepsilon > 0$ ,  $m \in \mathbb{N}$ . Then

$$\sum_{n=1}^m P(\{|X_n| \geq \varepsilon\}) = \sum_{n=1}^m \int_n^{+\infty} \exp(-x) dx = \sum_{n=1}^m \exp(-n)$$

i.e.  $\sum_{n=1}^{\infty} \mathbb{P}(\{|X_n| \geq \varepsilon\}) = 1/(e-1)$  and  $X_n \xrightarrow{a.c.} 0$ . If  $A(\{X_n\}_{n=1}^{\infty}) = \{Y_n\}_{n=1}^{\infty}$  then

$$Y_n = \sum_{m=1}^{\infty} (1+n)^{-m} \chi_{[m,+\infty)} \text{ if } n \in \mathbb{N}. \quad (4.2)$$

Consequently, for  $n \in \mathbb{N}$  and  $w \in \mathbb{R}$  it is easy to see that

$$Y_n(w) = \frac{1}{n} \left( 1 - \frac{1}{(1+n)^{\lfloor w \rfloor}} \right) \chi_{[0,+\infty)}(w).$$

Thus  $\{|Y_n| \geq \varepsilon\} = \emptyset$  if  $n > 1/\varepsilon$  and so  $\text{ac-lim}_{n \rightarrow \infty} Y_n = 0$ . However,  $c(\mathbb{N}, \mathbb{C})$  is not ac-regular for  $\{X_n\}_{n=1}^{\infty}$ . For, if  $B \in \mathcal{B}(c(\mathbb{N}, \mathbb{C}))$  is defined by the infinite matrix whose nm-entry is  $2^{-m-1}$  we write  $B(\{X_n\}_{n=1}^{\infty}) = \{Z_n\}_{n=1}^{\infty}$ . For  $w \in \mathbb{R}$  we now evaluate that  $Z_n(w) = (1 - 2^{-\lfloor w \rfloor})/2$  for all  $n \in \mathbb{N}$ . If  $0 < \varepsilon < 1/2$  let us choose  $v \in \mathbb{N}$  so that  $\varepsilon < (1 - 2^{-v})/2$ . Then,

$$\{|Z_n| \geq \varepsilon\} \supseteq \{Z_n \geq 2^{-1} - 2^{-v-1}\} = [v, +\infty),$$

i.e.  $\mathbb{P}(\{|Z_n| \geq \varepsilon\}) \geq \exp(-v)$ . Therefore  $\text{ac-lim}_{n \rightarrow \infty} |Z_n| \neq 0$  and  $B$  is not ac-regular for the sequence  $\{X_n\}_{n=1}^{\infty}$ . Since obviously  $B$  is not a d-regular operator for  $\{X_n\}_{n=1}^{\infty}$  it is also not p-regular nor not ae-regular for it. Finally, if  $r > 0$  then  $A$  becomes  $L^r$ -regular for  $\{X_n\}_{n=1}^{\infty}$ . For,

$$L^r\text{-}\lim_{n \rightarrow \infty} X_n = \lim_{n \rightarrow \infty} \mathbb{E}(|X_n|^r) = \lim_{n \rightarrow \infty} \exp(-m) = 0.$$

If  $n \in \mathbb{N}$  using (4.2)  $Y_n$  becomes a discrete random variable and

$$\begin{aligned} \mathbb{E}(|Y_n|^r) &= \frac{1}{n^r} \sum_{m=1}^{\infty} \left( 1 - \frac{1}{(n+1)^m} \right)^r \mathbb{P}([m-1, m)) \\ &\leq \frac{1}{n^r} \sum_{m=1}^{\infty} [\exp(-m) - \exp(-m-1)] = \frac{1}{e n^r}, \end{aligned}$$

i.e.  $L^r\text{-}\lim_{n \rightarrow \infty} Y_n = 0$ . However, it is evident that  $B$  is not  $L^r$ -regular for  $\{X_n\}_{n=1}^{\infty}$ .

**Problem 4.3.** Is it possible to characterize the subclasses of  $\mathcal{M}$ -regular operators of  $\mathcal{B}(c(\mathbb{N}, \mathbb{C}))$  for the sequence  $\{X_n\}_{n=1}^{\infty}$ ? In the general case, what relevant properties can be developed concerning to those classes? Can be determinated some subsets of  $\mathcal{B}(c(\mathbb{N}, \mathbb{C}))$  that are  $\mathcal{M}$ -regular for all random process on any unrestricted probability space  $(\Omega, \mathcal{A}, \mathcal{P})$ ? A partial answer to the last question is given in the following Th. 4.5. To this end remember the following.

**Definition 4.4.** A covering of a non empty set  $X$  is a subset  $\mathcal{U}$  of  $\mathcal{P}(X)$  so that  $X = \bigcup \mathcal{U}$ . It is said that the covering  $\mathcal{U}$  of  $X$  is locally finite if any element of  $X$  belongs to a finite number of elements of  $\mathcal{U}$ . Further, a locally finite covering  $\mathcal{U}$  of  $X$  is called bounded if

$$\eta = \sup \{ \text{card} \{ U \in \mathcal{U} : x \in U \} : x \in X \} < \infty.$$

Then  $\eta \in \mathbb{N}$  and we will say that  $\eta$  is the least upper bound of  $\mathcal{U}$ .

**Theorem 4.5.** (i) Let  $\mathcal{U} = \{U_n\}_{n=1}^{\infty}$  be a locally finite bounded covering of  $\mathbb{N}$  with a least upper bound  $\eta$ . If  $A \in \mathcal{B}(c(\mathbb{N}, \mathbb{C}))$  is defined by any infinite matrix  $\{a_{n,m}\}_{n,m=0}^{\infty}$  so that  $a_{n,m} = 0$  if  $m \notin U_n$  then  $A$  is ac-regular for any random process on the Banach space sequence  $c(\mathbb{N}, \mathbb{C})$ .

(ii) Let  $A \in \mathcal{B}(c(\mathbb{N}, \mathbb{C}))$  induced by an infinite matrix of non negative coefficients  $\{a_{n,m}\}_{n,m=0}^{\infty}$  with  $a_{0,0} = 0$ . Then  $A$  is  $L^r$ -regular if  $1 \leq r < +\infty$ .

**Proof.** (i) If  $\{X_n\}_{n=1}^{\infty} \cup \{X\} \subseteq \mathcal{M}_{\mathbb{P}}(\Omega, \mathcal{A}, \mathbb{E})$  and  $X = \text{ac-lim}_{n \rightarrow \infty} X_n$  we know that  $X = \text{ae-lim}_{n \rightarrow \infty} X_n$  and by Corollary 2.3 it is defined a random process on  $c_0(\mathbb{N}, \mathbb{C})$ . If  $n \in \mathbb{N}$  let  $Y_n \triangleq \sum_{m=1}^{\infty} a_{n,m} (X_m - X)$ . So, if  $\varepsilon > 0$  then  $\{|Y_n| \geq \varepsilon\} = \emptyset$  or

$$\begin{aligned} \{|Y_n| \geq \varepsilon\} &\subseteq \left\{ \sum_{m \in U_n} |a_{n,m} (X_m - X)| \geq \varepsilon \right\} \\ &\subseteq \left\{ \sup_{m \in U_n} |X_m - X| \sum_{m \in U_n} |a_{n,m}| \geq \varepsilon \right\} \\ &\subseteq \left\{ \sup_{m \in U_n} |X_m - X| \geq \varepsilon / \|A\| \right\} \\ &\subseteq \bigcup_{m \in U_n} \{|X_m - X| \geq \varepsilon / \|A\|\}. \end{aligned}$$

Consequently, if  $N \in \mathbb{N}$  we estimate

$$\begin{aligned} \sum_{n=1}^N \mathbb{P}(\{|Y_n| \geq \varepsilon\}) &\leq \sum_{n=1}^N \sum_{m \in U_n} \mathbb{P}(\{|X_m - X| \geq \varepsilon / \|A\|\}) \\ &\leq \sum_{m \in \bigcup_{n=1}^N U_n} \mathbb{P}(\{|X_m - X| \geq \varepsilon / \|A\|\}) \text{card}\{n : m \in U_n\} \\ &\leq \eta \sum_{m=1}^{\infty} \mathbb{P}(\{|X_m - X| \geq \varepsilon / \|A\|\}). \end{aligned}$$

Therefore,

$$\sum_{n=1}^{\infty} \mathbb{P}(\{|Y_n| \geq \varepsilon\}) \leq \eta \sum_{m=1}^{\infty} \mathbb{P}(\{|X_m - X| \geq \varepsilon / \|A\|\}) < \infty$$

and our claim follows.

(ii) Let  $A \in \mathcal{B}(c(\mathbb{N}, \mathbb{C}))$  defined by an infinite matrix  $\{a_{n,m}\}_{n,m \in \mathbb{N}}$  with non negative coefficients and  $a_{0,0} = 0$ . Let  $\{Z_m\}_{m=1}^{\infty} \cup \{Z\}$  be a sequence of random variables defining a Banach random process on  $c(\mathbb{N}, \mathbb{C})$  so that  $Z_m \xrightarrow{L^r} Z$ . Giving  $n \in \mathbb{N}$  set  $W_n \triangleq A_n(\{Z_m - Z\}_{m=1}^{\infty})$ . Of course we may assume that  $A \neq 0$ . Consider the measure space  $(\mathbb{N}, \mathcal{P}(\mathbb{N}), \mu_n)$  so that  $\mu_n(S) \triangleq \|A\|^{-1} \sum_{m \in S} a_{n,m}$ . Let us

consider the function

$$F: \mathbb{N} \times \Omega \rightarrow \mathbb{C}, F(m, w) \triangleq Z_m(w) - Z(w).$$

Giving  $\zeta \in \mathbb{C}$  and  $r > 0$  it is easy to see that

$$\{|F - \zeta| < r\} = \bigcup_{m=1}^{\infty} \{m\} \times \{|Z_m - Z| < r\},$$

i.e.  $\{|F - \zeta| < r\}$  is clearly a measurable subset of  $\mathbb{N} \times \Omega$  and since  $\zeta$  and  $r$  are arbitrary  $F$  is measurable. Indeed, for almost all  $w \in \Omega$  and  $m \in \mathbb{N}$  there is a positive constant  $K(w)$  so that  $|Z_v(w)| \leq K(w)$  if  $v \in \mathbb{N}$  and we have

$$\begin{aligned} \int_{\{1, \dots, m\}} |F(v, w)| d\mu_n(v) &= \|A\|^{-1} \sum_{v=1}^m a_{n,v} |Z_v(w) - Z(w)| \quad (4.3) \\ &\leq 2K(w) \|A\|^{-1} \sum_{v=1}^m a_{n,v} \\ &\leq 2K(w). \end{aligned}$$

By an easy application of the monotone convergence theorem in (4.3) we deduce that  $F(\circ, w) \in \mathcal{L}^1(\mathbb{N}, \mu_n)$ . Further,

$$F(\circ, w) = \lim_{m \rightarrow \infty} \sum_{v=1}^m (Z_v(w) - Z(w)) \chi_{\{v\}}(\circ)$$

and if  $m \in \mathbb{N}$  we have that

$$\left| \sum_{v=1}^m (Z_v(w) - Z(w)) \chi_{\{v\}}(\circ) \right| \leq |F(\circ, w)|$$

on  $\mathbb{N}$ . By Lebesgue's dominated convergence theorem for almost all  $w \in \Omega$  we get

$$\begin{aligned} W_n(w) &= \sum_{m=1}^{\infty} a_{n,m} (Z_m(w) - Z(w)) \quad (4.4) \\ &= \|A\| \sum_{m=1}^{\infty} (Z_m(w) - Z(w)) \mu_n(\{m\}) \\ &= \|A\| \sum_{m=1}^{\infty} (Z_m(w) - Z(w)) \int_{\mathbb{N}} \chi_{\{m\}}(v) d\mu_n(v) \\ &= \|A\| \int_{\mathbb{N}} F(v, w) d\mu_n(v). \end{aligned}$$

Using (4.4) and applying the Minkowski's integral inequality we now write

$$\mathbb{E}(|W_n|^r)^{1/r} = \left( \int_{\Omega} |W_n(w)|^r dP(w) \right)^{1/r} \quad (4.5)$$

$$\begin{aligned}
 &= \|A\| \left( \int_{\Omega} \left| \int_{\mathbb{N}} F(m, w) d\mu_n(m) \right|^r d\mathbb{P}(w) \right)^{1/r} \\
 &\leq \|A\| \int_{\mathbb{N}} \left( \int_{\Omega} |F(m, w)|^r d\mathbb{P}(w) \right)^{1/r} d\mu_n(m) \\
 &= \|A\| \int_{\mathbb{N}} \left( \int_{\Omega} |Z_m(w) - Z(w)|^r d\mathbb{P}(w) \right)^{1/r} d\mu_n(m) \\
 &= \|A\| \int_{\mathbb{N}} \mathbb{E}(|Z_m - Z|^r)^{1/r} d\mu_n(m) \\
 &= \sum_{m=1}^{\infty} a_{n,m} \mathbb{E}(|Z_m - Z|^r)^{1/r}.
 \end{aligned}$$

Finally, the sequence  $\{\mathbb{E}(|Z_m - Z|^r)\}_{m=1}^{\infty}$  is bounded and the claim follows letting  $n \rightarrow \infty$  in (4.5), using (4.1) and that  $a_{0,0} = 0$ .  $\square$

#### 4.2. $\mathcal{M}$ -regularity on $[\mathcal{M}_{dt}([0, 1], \mathcal{L}[0, 1], \mathbb{C}[0, 1]), \mathbb{I}^p(\mathbb{C}[0, 1])]$

**Theorem 4.6.** *Let  $\mathcal{U} = \{U_n\}_{n \in \mathbb{N}}$  be a disjoint bounded covering of  $\mathbb{N}$  with a least upper bound  $\eta$ . Given  $m \in \mathbb{N}$  let  $n(m)$  be the unique positive integer so that  $m \in U_{n(m)}$ . Let  $1 < p, q < \infty$  so that  $1/p + 1/q = 1$  and let  $a \triangleq \{a_{n,m}\}_{n,m=1}^{\infty}$  be a set of complex numbers so that the series  $\sigma(a) \triangleq \sum_{m=1}^{\infty} |a_{n(m),m}|^q$  is finite. Given  $x \in \mathbb{I}^p(\mathbb{C}[0, 1])$  set*

$$A^a(x) = \left\{ \sum_{m \in U_n} a_{n,m} \cdot x_m \right\}_{n=1}^{\infty}.$$

Then

(i)  $A^a(x) \in \mathbb{I}^p(\mathbb{C}[0, 1])$ .

(ii)  $A^a \in \mathcal{B}[\mathbb{I}^p(\mathbb{C}[0, 1])]$ .

(iii) The class  $\mathcal{R} \triangleq \{A^a : \sigma(a) < \infty\}$  is simply almost completely regular on

$$[\mathcal{M}_{dx}([0, 1], \mathcal{L}[0, 1], \mathbb{C}[0, 1]), \mathbb{I}^p(\mathbb{C}[0, 1])].$$

(iv) The class  $\mathcal{R} \triangleq \{A^a : \sigma(a) < \infty\}$  is regular in the mean on any random process  $\{X_n\}_{n=1}^{\infty} \cup \{X\}$  so that  $\sum_{n=1}^{\infty} \|\mathbb{E}(X_n - X)\|_{\infty}^p < \infty$ .

**Proof.** (i) Since  $\mathcal{U}$  is a bounded covering of  $\mathbb{N}$  then  $A^a(x) \hookrightarrow \mathbb{C}[0, 1]$  if  $x \in \mathbb{I}^p(\mathbb{C}[0, 1])$ . Indeed, if  $a \in \mathcal{R}$  and  $N \in \mathbb{N}$  we obtain

$$\begin{aligned}
 \left[ \sum_{n=1}^N \|A_n^a(x)\|_{\infty}^p \right]^{1/p} &\leq \left[ \sum_{n=1}^N \left( \sum_{m \in U_n} |a_{n,m}| \|x_m\|_{\infty} \right)^p \right]^{1/p} \\
 &\leq \sum_{m \in U_1 \cup \dots \cup U_N} \|x_m\|_{\infty} \left( \sum_{n \in \mathbb{N}: m \in U_n} |a_{n,m}|^p \right)^{1/p}
 \end{aligned} \tag{4.6}$$

$$\begin{aligned}
&= \sum_{m \in U_1 \cup \dots \cup U_N} \|x_m\|_\infty \cdot |a_{n(m),m}| \\
&\leq \sigma(a)^{1/q} \cdot \|x\|_{\mathbb{P}(\mathbb{C}[0,1])}
\end{aligned}$$

Letting  $N \rightarrow \infty$  from (4.6) we see that  $A^a(x) \in \mathbb{P}(\mathbb{C}[0,1])$  and

$$\|A^a(x)\|_{\mathbb{P}(\mathbb{C}[0,1])} \leq \sigma(a)^{1/q} \cdot \|x\|_{\mathbb{P}(\mathbb{C}[0,1])}.$$

(ii) It is now clear that  $A^a$  is linear and that  $\|A^a\| \leq \sigma(a)^{1/q}$ .

(iii) Let  $\{X_m\}_{m=1}^\infty \cup \{X\}$  be a random process of  $\mathcal{M}_{dx}([0,1], \mathcal{L}[0,1], \mathbb{C}[0,1])$  on the Banach sequence space  $\mathbb{P}(\mathbb{C}[0,1])$  so that  $X_m \xrightarrow{a.c.} X$ . Given  $a \in \mathcal{R}$  we will show that  $A_n^a(\{X_m - X\}_{m=1}^\infty) \xrightarrow{a.c.} 0$ . For, evidently we can assume  $\sigma(a) > 0$ . If  $\varepsilon > 0$  and  $n \in \mathbb{N}$  we write

$$\begin{aligned}
\{\|A_n^a(\{X_m - X\}_{m=1}^\infty)\|_\infty \geq \varepsilon\} &= \left\{ \left\| \sum_{m \in U_n} a_{n,m} \cdot (X_m - X) \right\|_\infty \geq \varepsilon \right\} \\
&\subseteq \left\{ \sigma(a)^{1/q} \sum_{m \in U_n} \|X_m - X\|_\infty \geq \varepsilon \right\} \\
&\subseteq \bigcup_{m \in U_n} \left\{ \|X_m - X\|_\infty \geq \frac{\varepsilon}{\sigma(a)^{1/q} \cdot \text{card}(U_n)} \right\} \\
&\subseteq \bigcup_{m \in U_n} \left\{ \|X_m - X\|_\infty \geq \frac{\varepsilon}{\sigma(a)^{1/q} \cdot \eta(a)} \right\}.
\end{aligned}$$

Consequently, if  $N \in \mathbb{N}$  we see that

$$\begin{aligned}
\sum_{n=1}^N \int_0^1 \chi_{\{\|A_n^a(\{X_m - X\}_{m=1}^\infty)\|_\infty \geq \varepsilon\}} dt &\leq \sum_{n=1}^N \sum_{m \in U_n} \int_0^1 \chi_{\{\|X_m - X\|_\infty \geq \frac{\varepsilon}{\sigma(a)^{1/q} \cdot \eta(a)}\}} dt \\
&\leq \sum_{m=1}^\infty \int_0^1 \chi_{\{\|X_m - X\|_\infty \geq \frac{\varepsilon}{\sigma(a)^{1/q} \cdot \eta(a)}\}} dt < \infty,
\end{aligned}$$

and our claim follows.

(iv) Let  $X_n \xrightarrow{\mathbb{E}} X$ ,  $a \in \mathcal{R}$ . If  $n \in \mathbb{N}$  and

$$Y_n \triangleq A_n^a(\{X_m - X\}_{m=1}^\infty) \triangleq \sum_{m \in U_n} a_{n,m} \cdot (X_m - X)$$

it will suffice to show that

$$\sum_{n=1}^\infty \|\mathbb{E}(Y_n)\|_\infty^p < \infty. \quad (4.7)$$

Indeed, we can assume  $X = 0$  a.e. Thus, if  $v \in \mathbb{N}$  and

$$\|\phi_1\|_{\text{BV}[0,1]} = \dots = \|\phi_v\|_{\text{BV}[0,1]} = 1$$

we have

$$\begin{aligned}
 \left| \sum_{n=1}^v \langle \mathbf{E}(Y_n), \phi_n \rangle \right| &= \left| \sum_{n=1}^v \int_0^1 \left( \int_0^1 Y_n(t)(s) d\phi_n(s) \right) dt \right| \\
 &= \left| \sum_{n=1}^v \int_0^1 \left( \int_0^1 \sum_{m \in U_n} a_{n,m} X_m(t) d\phi_n(s) \right) dt \right| \\
 &= \left| \sum_{n=1}^v \sum_{m \in U_n} a_{n,m} \langle \mathbf{E}(X_m), \phi_n \rangle \right| \\
 &\leq \sum_{n=1}^v \sum_{m \in U_n} |a_{n,m}| \|\mathbf{E}(X_m)\|_\infty \\
 &\leq \sum_{n=1}^v \left( \sum_{m \in U_n} \|\mathbf{E}(X_m)\|_\infty^p \right)^{1/p} \left( \sum_{m \in U_n} |a_{n,m}|^q \right)^{1/q} \\
 &\leq \left( \sum_{n=1}^\infty \|\mathbf{E}(X_n)\|_\infty^p \right)^{1/p} \sigma(a)^{1/q}.
 \end{aligned}$$

But  $l^p(C[0, 1])^* \approx l^q(BV[0, 1])$ , where  $\approx$  denotes an isometric isomorphism of Banach spaces. Therefore,

$$\begin{aligned}
 \left( \sum_{n=1}^v \|\mathbf{E}(Y_n)\|_\infty^p \right)^{1/p} &= \sup_{\|\phi_1\|_{BV[0,1]} = \dots = \|\phi_v\|_{BV[0,1]} = 1} \left| \sum_{n=1}^v \langle \mathbf{E}(Y_n), \phi_n \rangle \right| \\
 &\leq \left( \sum_{n=1}^\infty \|\mathbf{E}(X_n)\|_\infty^p \right)^{1/p} \sigma(a)^{1/q},
 \end{aligned}$$

and (4.7) follows since  $v$  is arbitrary. □

## References

- [1] BARRENECHEA, A.L., PEÑA, C.C., *Compactness and Radon-Nikodym pro-perties on the Banach space of convergent series*, An. Șt. Univ. Ovidius Constanța. Vol. 16, (1), (2008), 19–30.
- [2] BÖTTCHER, A., GRUDSKY, S.M., *Toeplitz matrices, asymptotic linear algebra and functional analysis*, Birkhäuser Verlag, Basel - Boston - Berlin, ISBN 3-7643-6290-1, (2000).
- [3] DOUGLAS, R.D., *Banach algebra techniques in operator theory*, Graduate Texts in Maths., 179. Springer-Verlag, N. Y., ISBN 0-387-98377-5, (1988).
- [4] EFROM, B., *The jackknife, the bootstrap, and other resampling plans*, So-ciety of Industrial and Applied Mathematics CBMS-NSF Monographs, 38, (1982).

- 
- [5] GRAY, H.L., *On a unification of bias reduction and numerical approximation*, Probability and Statistics. J. N. Srivastance Ed., North-Holland, Amsterdam, (1991), 105–116.
- [6] LAVASTRE, H., *On the stochastic regularity of sequence transformations o-perating in a Banach space*, Appl. Mathematicae. 22, 4, (1995), 477–484.
- [7] LAWLER, G.F., *Introduction to stochastic processes*, Chapman & Hall / CRC, U.S.A, ISBN: 0-41299-511-5, (2006).
- [8] LEDOUX, M., TALAGRAND, M., *Probability in Banach spaces*, 1st Edition, Springer ISBN: 978-3-540-52013-9, (1991).
- [9] LINDENSTRAUSS, J., TZAFRIRI, L., *Classical Banach spaces I*, Springer-Verlag, Germany, ISBN 3-540-60628-9, (1977).
- [10] MADDOX, I. J., *Infinite matrices of operators*, Lect. Notes in Maths., 786, Springer-Verlag, Germany, ISBN 3-540-09764-3, (1980).
- [11] PAGGETT, W.J., TAYLOR, R.L., *Laws of large number for normed linear spaces and certain Fréchet spaces*, Lect. Notes in Maths., Springer-Verlag, ISBN: 3540065857, (1973).
- [12] ROHATGI, V.K., *An introduction to probability theory and mathematical statistics*, John Wiley & Sons, ISBN-10: 0471731358, (1976).
- [13] SCHUR, I., *Über lineare Transformationen in der Theorie der unendlichen Reihen*, J. f. reine u. angew. Math, 151, (1921), 79–111.
- [14] SHIRYAEV, A.N., *Probability*, 2nd Edition, Springer-Verlag, N.Y., ISBN-13: 978-0387945491, (1995).
- [15] WIMP, V., *Sequence transformations and their applications*, Academic Press, N.Y, ISBN-13: 978-3540152835, (1981).

**A. L. Barrenechea**

UNCPBA - FCExactas - Dpto. de Matemáticas - NUCOMPA

Pinto 399 - Tandil - Argentina

e-mail: [analucia@exa.unicen.edu.ar](mailto:analucia@exa.unicen.edu.ar)



# Periodic fixed points of random operators

Ismat Beg<sup>a</sup>, Mujahid Abbas<sup>a</sup>, Akbar Azam<sup>b</sup>

<sup>a</sup>Center for Advanced Studies in Mathematics  
Lahore University of Management Sciences

<sup>b</sup>Department of Mathematics  
COMSATS Institute of Information Technology

*Submitted 10 February 2010; Accepted 19 April 2010*

## Abstract

Sufficient conditions for existence of random fixed point of a nonexpansive rotative random operator are obtained and existence of random periodic points of a random operator is proved. We also derive random periodic point theorem for  $\epsilon$ -expansive random operator.

*Keywords:* Random periodic point; random fixed point;  $\epsilon$ -contractive random operator;  $\epsilon$ -expansive random operator; rotative random operator; metric space; Banach space; measurable space.

*MSC:* 47H09, 47H10, 47H40, 54H25, 60H25

## 1. Introduction

Random nonlinear analysis has grown into an active research area closely associated with the study of random nonlinear operators and their properties needed in solving nonlinear random operator equations (see [7, 18, 21]). The study of random fixed point theory was initiated by the Prague school of probabilists in the 1950's ([15, 24]). Random fixed point theorems are of tremendous importance in probabilistic functional analysis as they provide a convenient way of modelling many real life problems and random methods have also revolutionized the financial markets. The survey article by Bharucha -Reid [8] in 1976 attracted the attention of several mathematician and gave wings to this theory. Itoh [17] extended Spacek's and Hans's theorems to random multivalued contraction mappings. In recent years, a lot of efforts have been made ([2, 3, 4, 5, 6, 16, 22, 23], and references therein) to show the existence of random fixed points of certain random single valued and multivalued operators and various applications in diverse area from pure mathematics

to applied sciences have been explored. The aim of this paper is to establish the existence of random fixed point of nonexpansive rotative random operator in the setting of Banach spaces. A random analogue of Edelstein theorem to establish the existence of random periodic points for random single valued  $\epsilon$ - contractive operator is proved. These results are then used to obtain the random periodic point of  $\epsilon$ - expansive random operators. The results proved in this paper improve and generalize several well known results in the literature [9, 12, 17].

## 2. Preliminaries

We begin with some definitions and state the notations used throughout this paper. Let  $(\Omega, \Sigma)$  be a measurable space ( $\Sigma$ - sigma algebra) and  $F$  be a nonempty subset of a separable metric space  $(X, d)$ . A single valued mapping  $T: \Omega \rightarrow X$  is *measurable* if  $T^{-1}(U) \in \Sigma$  for each open subset  $U$  of  $X$ , where  $T^{-1}(U) = \{\omega \in \Omega : T(\omega) \cap U \neq \emptyset\}$ . A mapping  $T: \Omega \times X \rightarrow X$  is a *random operator* if and only if for each fixed  $x \in X$ , the mapping  $T(\cdot, x): \Omega \rightarrow X$  is measurable and it is *continuous* if for each  $\omega \in \Omega$ , the mapping  $T(\omega, \cdot): X \rightarrow X$  is continuous. A measurable mapping  $\xi: \Omega \rightarrow X$  is a *random fixed point* of a random operator  $T: \Omega \times X \rightarrow X$  if and only if  $\xi(\omega) = T(\omega, \xi(\omega))$  for each  $\omega \in \Omega$ . We denote the set of random fixed points of a random operator  $T$  by  $RF(T)$  and the set of all measurable mappings from  $\Omega$  into  $X$  by  $M(\Omega, X)$ . For the random operator  $f: \Omega \times X \rightarrow X$ , the map  $f_\omega^{-1}: X \rightarrow X$  is defined by  $f_\omega^{-1}(y) = x$  if and only if  $f(\omega, x) = y$ .

We denote the  $n$ th iterate  $T(\omega, T(\omega, T(\omega, \dots, T(\omega, x) \dots)))$  of random operator  $T: \Omega \times X \rightarrow X$  by  $T^n(\omega, x)$ . The letter  $I$  denotes the random operator  $I: \Omega \times X \rightarrow X$  defined by  $I(\omega, x) = x$  and  $T^0 = I$ . The random operator  $T: \Omega \times X \rightarrow X$  is called *random periodic operator* with period  $p \in \mathbb{N}$ , if for each  $x \in X$  and  $\omega \in \Omega$  we obtain  $T^p(\omega, x) = I(\omega, x)$ . Let  $B(x_0, r)$  denotes the spherical ball centred at  $x_0$  with radius  $r$ , defined as the set  $\{x \in X : d(x, x_0) \leq r\}$ .

**Definition 2.1.** Let  $F$  be a nonempty subset of a separable metric space  $X$ . The random operator  $T: \Omega \times F \rightarrow F$  is said to be:

- (a)  $k(\omega)$ - *contraction random operator* if for any  $x, y \in F$  and  $\omega \in \Omega$ , we have

$$d(T(\omega, x), T(\omega, y)) \leq k(\omega)d(x, y),$$

where  $k: \Omega \rightarrow [0, 1)$  is a measurable map. If  $k(\omega) = 1$  for any  $\omega \in \Omega$ , then  $T$  is called *nonexpansive random operator*.

- (b) *contractive random operator* if for any  $x, y \in F$  and  $\omega \in \Omega$ , we have

$$d(T(\omega, x), T(\omega, y)) < d(x, y).$$

- (c)  $\epsilon$ -*contractive random operator* if for  $\epsilon > 0$  and  $x, y \in F$  with  $x \neq y$  and  $d(x, y) < \epsilon$ , we have,

$$d(T(\omega, x), T(\omega, y)) < d(x, y),$$

for every  $\omega \in \Omega$ . Obviously, every contractive random operator is  $\epsilon$ - contractive random operator for any  $\epsilon > 0$ .

- (d)  $\epsilon$ -expansive random operator if for  $\epsilon > 0$  and  $x, y \in F$  with  $x \neq y$  and  $d(x, y) < \epsilon$ , we have

$$d(T(\omega, x), T(\omega, y)) > d(x, y), \tag{2.1}$$

for every  $\omega \in \Omega$ . If inequality (2.1) holds for every  $x, y \in X$  with  $x \neq y$  then  $T$  is called an *expansive random operator*.

Obviously, every expansive random operator is  $\epsilon$ - expansive random operator for any  $\epsilon > 0$ .

**Definition 2.2.** Let  $T: \Omega \times F \rightarrow F$  be a random operator, where  $F$  is a nonempty subset of a separable complete metric space  $X$ . A measurable mapping  $\xi: \Omega \rightarrow F$  is called a *random periodic point* of  $T$  there exists  $n \geq 1$  such that  $T^n(\omega, \xi(\omega)) = \xi(\omega)$ , for every  $\omega \in \Omega$ . That is, random periodic point is random fixed point of  $n$ th iterate of  $T$  for some  $n \geq 1$ . The least such positive integer  $n$  is called *period* of random periodic point  $\xi$ .

Note that random fixed point of  $T$  is also random periodic point of  $T$  of period 1 but there exists a random periodic point of  $T$  which fails to be the random fixed point of  $T$  as shown in the examples presented below. It is also shown that there exists a random operator having random periodic point of period 5 but does not posses the random periodic point of period 3.

**Example 2.3.** Let  $\Omega = [0, 1]$  and  $\Sigma$  be the sigma algebra of Lebesgue’s measurable subsets of  $\Omega$ . Take  $X = R$  with  $d(x, y) = |x - y|$ , for  $x, y \in R$ . Define random operator  $T$  from  $\Omega \times X$  to  $X$  as,

$$T(\omega, x) = \begin{cases} \omega^2 - x, & \text{if } (\omega, x) \in \Omega \times [0, 1] \\ \omega^2 - x - 1, & \text{otherwise.} \end{cases}$$

Define the measurable mapping  $\xi: \Omega \rightarrow X$  as  $\xi(\omega) = \frac{1}{2}(3\omega^2 - 1)$ , for every  $\omega \in \Omega$ . Now  $\xi$  is a random periodic point of  $T$  with period 2 but it fails to be a random fixed point of  $T$ .

**Example 2.4.** Let  $\Omega = [0, 1]$  and  $\Sigma$  be the sigma algebra of Lebesgue’s measurable subsets of  $\Omega$ . Take  $X = R$  with  $d(x, y) = |x - y|$ , for  $x, y \in R$ . Define random operator  $T$  from  $\Omega \times X$  to  $X$  as,  $T(\omega, 1) = 3, T(\omega, 2) = 5, T(\omega, 3) = 4, T(\omega, 4) = 2, T(\omega, 5) = 1$  and  $T(\omega, x) = x - \omega$ , when  $x \notin \{1, 2, 3, 4, 5\}$ .

Define measurable mapping  $\xi: \Omega \rightarrow X$  as  $\xi(\omega) = 1$ , for every  $\omega \in \Omega$ . Note that  $\xi$  is a random periodic point of period 5. It is also noted that random operator  $T$  in this example does not posses random fixed point because for any  $\xi$  to be the random fixed point, we must have  $T(\omega, \xi(\omega)) = \xi(\omega)$ , for every  $\omega \in \Omega$ . But this random operator equation holds only for  $\omega = 0$ .

**Remark 2.5.** Let  $F$  be a closed subset of a complete separable metric space  $X$  and the sequence of measurable mappings  $\{\xi_n\}$  from  $\Omega$  to  $F$  be point wise convergent, that is,  $\xi_n(\omega) \rightarrow q := \xi(\omega)$  for each  $\omega \in \Omega$ . Then  $\xi$  being the limit of the sequence of measurable mappings is measurable and closedness of  $F$  implies  $\xi$  is a mapping from  $\Omega$  to  $F$ . Since  $F$  is a subset of a complete separable metric space  $X$ , also if  $T$  is a continuous random operator from  $\Omega \times F$  to  $F$  then by the lemma 8.2.3 of [1], the map  $\omega \rightarrow T^n(\omega, f(\omega))$  is measurable for any measurable mapping  $f$  from  $\Omega$  to  $F$ .

**Definition 2.6.** Let  $F$  be a nonempty subset of a Banach space  $X$ . The random operator  $T: \Omega \times F \rightarrow F$  is said to be  $(k, n)$ -rotative random operator for  $k < n$ , if for each  $\omega \in \Omega$ ,

$$\|\xi(\omega) - T^n(\omega, \xi(\omega))\| \leq k \|\xi(\omega) - T(\omega, \xi(\omega))\|,$$

where  $\xi$  is a mapping from  $\Omega$  to  $F$  and  $n \in \mathbb{N}$ . The operator  $T$  is said to be  $n$ -rotative random operator if it  $(k, n)$ -rotative random operator for some  $k < n$  and  $T$  is called *rotative random operator* if it is an  $n$ -rotative random operator for some  $n \in \mathbb{N}$ . Note that any random periodic operator with period  $p$  is  $(0, p)$ -rotative random operator.

**Remark 2.7.** If  $T: \Omega \times F \rightarrow F$  is  $k(\omega)$  contraction random operator where  $F$  is a closed subset of Banach space  $X$  and  $n > 1$ . For any  $\xi: \Omega \rightarrow F$ , consider,

$$\begin{aligned} \|\xi(\omega) - T^n(\omega, \xi(\omega))\| &\leq \sum_{k=1}^n \|T^{k-1}(\omega, \xi(\omega)) - T^k(\omega, \xi(\omega))\| \\ &\leq (1 + k(\omega) + (k(\omega))^2 + \dots \\ &\quad + (k(\omega))^{n-1}) \|\xi(\omega) - T(\omega, \xi(\omega))\| \\ &< n \|\xi(\omega) - T(\omega, \xi(\omega))\|, \end{aligned}$$

for every  $\omega \in \Omega$ . Thus  $T$  is a rotative random operator.

### 3. Periodic and fixed points of rotative random operators

In this section, we first show an existence of a random fixed point of a nonexpansive rotative random operator which not only provides a random analogue of theorem 17.1 of [11] (see also [12]) but also improves theorem 2.1 of [17] in the sense that it does not require the boundedness of  $T(\omega, F)$  for any  $\omega \in \Omega$ . Moreover we replace continuous condensing random operator by nonexpansive rotative random operator.

Periodic point problems were systematically studied since the beginning of fifties (see [9, 10, 13, 14, 19, 20]). We show some results on the existence of random periodic points of random single valued  $\epsilon$ -contractive operator in the setting of a separable metric space.

**Theorem 3.1.** *Let  $F$  be a nonempty closed and convex subset of a separable Banach space  $X$  and  $T: \Omega \times F \rightarrow F$  be a nonexpansive rotative random operator. Then  $T$  has a random fixed point.*

**Proof.** Let  $\xi: \Omega \rightarrow F$  be any fixed measurable mapping. For  $0 < \alpha < 1$  and any arbitrary measurable mapping  $\eta: \Omega \rightarrow F$ , define  $T_\alpha: \Omega \times F \rightarrow F$  as,

$$T_\alpha(\omega, \eta(\omega)) = (1 - \alpha)\xi(\omega) + \alpha T(\omega, \eta(\omega)).$$

Note that for each  $\alpha$ , the random operator  $T_\alpha$  has Lipschitz constant  $\alpha$ . we may apply [8] to obtain the sequence of random operators  $F_\alpha: \Omega \times F \rightarrow F$  such that  $T_\alpha(\omega, F_\alpha(\omega, \xi(\omega))) = F_\alpha(\omega, \xi(\omega))$ , for every  $\omega \in \Omega$ . Consequently, we have

$$F_\alpha(\omega, \xi(\omega)) = (1 - \alpha)\xi(\omega) + \alpha T(\omega, F_\alpha(\omega, \xi(\omega))).$$

It can be verified that each  $F_\alpha$  is nonexpansive random operator. By iterating  $F_\alpha$  we obtain

$$F_\alpha^k(\omega, \xi(\omega)) = (1 - \alpha)F_\alpha^{k-1}(\omega, \xi(\omega)) + \alpha T(\omega, F_\alpha^k(\omega, \xi(\omega))), \quad k \in N. \quad (3.1)$$

Note that,

$$\begin{aligned} & (1 - \alpha)F_\alpha(\omega, \xi(\omega)) \\ &= (1 - \alpha)\xi(\omega) + \alpha T(\omega, F_\alpha(\omega, \xi(\omega))) - \alpha F_\alpha(\omega, \xi(\omega)) \\ &= (1 - \alpha)\xi(\omega) + \alpha(T(\omega, F_\alpha(\omega, \xi(\omega))) - F_\alpha(\omega, \xi(\omega))). \end{aligned}$$

Thus for each  $\omega \in \Omega$

$$\begin{aligned} & (1 - \alpha)(\xi(\omega) - F_\alpha(\omega, \xi(\omega))) \\ &= \alpha(F_\alpha(\omega, \xi(\omega)) - T(\omega, F_\alpha(\omega, \xi(\omega))). \end{aligned} \quad (3.2)$$

Now suppose  $T$  is a  $(a, n)$ -rotative random operator, that is

$$\|\xi(\omega) - T^n(\omega, \xi(\omega))\| \leq a \|\xi(\omega) - T(\omega, \xi(\omega))\|,$$

for every  $\omega \in \Omega$ . Now,

$$\begin{aligned} & \|F_\alpha(\omega, \xi(\omega)) - F_\alpha^2(\omega, \xi(\omega))\| \\ &= \left\| (1 - \alpha)\xi(\omega) + \alpha T(\omega, F_\alpha(\omega, \xi(\omega))) - (1 - \alpha)F_\alpha(\omega, \xi(\omega)) \right. \\ & \quad \left. - \alpha T(\omega, F_\alpha^2(\omega, \xi(\omega))) \right\| \\ &= \left\| (1 - \alpha)(\xi(\omega) - F_\alpha(\omega, \xi(\omega))) + \alpha(T(\omega, F_\alpha(\omega, \xi(\omega))) \right. \\ & \quad \left. - \alpha T(\omega, F_\alpha^2(\omega, \xi(\omega)))) \right\| \\ &= \left\| \alpha(F_\alpha(\omega, \xi(\omega)) - T(\omega, F_\alpha(\omega, \xi(\omega)))) + \alpha(T(\omega, F_\alpha(\omega, \xi(\omega))) \right. \\ & \quad \left. - \alpha T(\omega, F_\alpha^2(\omega, \xi(\omega)))) \right\| \\ &= \alpha \|F_\alpha(\omega, \xi(\omega)) - T(\omega, F_\alpha^2(\omega, \xi(\omega)))\| \end{aligned}$$

$$\begin{aligned}
&\leq \alpha \|F_\alpha(\omega, \xi(\omega)) - T^m(\omega, F_\alpha(\omega, \xi(\omega)))\| \\
&\quad + \alpha \|T^m(\omega, F_\alpha(\omega, \xi(\omega))) - T(\omega, F_\alpha^2(\omega, \xi(\omega)))\| \\
&\leq \alpha a \|F_\alpha(\omega, \xi(\omega)) - T(\omega, F_\alpha(\omega, \xi(\omega)))\| \\
&\quad + \alpha \|T^{m-1}(\omega, F_\alpha(\omega, \xi(\omega))) - F_\alpha^2(\omega, \xi(\omega))\| \\
&= (1 - \alpha)a \|F_\alpha(\omega, \xi(\omega)) - \xi(\omega)\| \\
&\quad + \alpha \|T^{m-1}(\omega, F_\alpha(\omega, \xi(\omega))) - F_\alpha^2(\omega, \xi(\omega))\|,
\end{aligned}$$

for every  $\omega \in \Omega$ . Now we claim that the following inequality holds for every  $\omega \in \Omega$  and  $m \geq 2$ .

$$\begin{aligned}
&\alpha \|T^{m-1}(\omega, F_\alpha(\omega, \xi(\omega))) - F_\alpha^2(\omega, \xi(\omega))\| \\
&\leq (m - 1) - m\alpha + \alpha^m \|\xi(\omega) - F_\alpha(\omega, \xi(\omega))\| \\
&\quad + \alpha^m \|F_\alpha(\omega, \xi(\omega)) - F_\alpha^2(\omega, \xi(\omega))\|. \tag{3.3}
\end{aligned}$$

For this consider,

$$\begin{aligned}
&\alpha \|T(\omega, F_\alpha(\omega, \xi(\omega))) - F_\alpha^2(\omega, \xi(\omega))\| \\
&= \alpha \|T(\omega, F_\alpha(\omega, \xi(\omega))) - (1 - \alpha)F_\alpha(\omega, \xi(\omega)) - \alpha T(\omega, F_\alpha^2(\omega, \xi(\omega)))\| \\
&= \alpha \left\| \begin{aligned} &(1 - \alpha)(T(\omega, F_\alpha(\omega, \xi(\omega))) - F_\alpha(\omega, \xi(\omega))) - \alpha(T(\omega, F_\alpha^2(\omega, \xi(\omega))) \\ &\quad - T(\omega, F_\alpha(\omega, \xi(\omega)))) \end{aligned} \right\| \\
&\leq (1 - \alpha) \|\alpha(T(\omega, F_\alpha(\omega, \xi(\omega))) - F_\alpha(\omega, \xi(\omega)))\| \\
&\quad + \alpha^2 \|T(\omega, F_\alpha^2(\omega, \xi(\omega))) - T(\omega, F_\alpha(\omega, \xi(\omega)))\| \\
&= (1 - \alpha)^2 \|\xi(\omega) - F_\alpha(\omega, \xi(\omega))\| + \alpha^2 \|T(\omega, F_\alpha^2(\omega, \xi(\omega))) - T(\omega, F_\alpha(\omega, \xi(\omega)))\| \\
&\leq (1 - \alpha)^2 \|\xi(\omega) - F_\alpha(\omega, \xi(\omega))\| + \alpha^2 \|F_\alpha^2(\omega, \xi(\omega)) - F_\alpha(\omega, \xi(\omega))\|.
\end{aligned}$$

So (3.3) is valid for  $m = 2$  and for any  $\omega \in \Omega$ .

Assuming the validity of (3.3) for  $m = j$  and for any  $\omega \in \Omega$ , consider

$$\begin{aligned}
&\alpha \|T^j(\omega, F_\alpha(\omega, \xi(\omega))) - F_\alpha^2(\omega, \xi(\omega))\| \\
&= \alpha \|T^j(\omega, F_\alpha(\omega, \xi(\omega))) - (1 - \alpha)F_\alpha(\omega, \xi(\omega)) - \alpha T(\omega, F_\alpha^2(\omega, \xi(\omega)))\| \\
&= \alpha \left\| \begin{aligned} &(1 - \alpha)(T^j(\omega, F_\alpha(\omega, \xi(\omega))) - F_\alpha(\omega, \xi(\omega))) + \alpha(T^j(\omega, F_\alpha(\omega, \xi(\omega))) \\ &\quad - T(\omega, F_\alpha^2(\omega, \xi(\omega)))) \end{aligned} \right\| \\
&\leq \alpha(1 - \alpha) \|T^j(\omega, F_\alpha(\omega, \xi(\omega))) - F_\alpha(\omega, \xi(\omega))\| \\
&\quad + \alpha^2 \|T^j(\omega, F_\alpha(\omega, \xi(\omega))) - T(\omega, F_\alpha^2(\omega, \xi(\omega)))\| \\
&\leq j\alpha(1 - \alpha) \|F_\alpha(\omega, \xi(\omega)) - T(\omega, F_\alpha(\omega, \xi(\omega)))\| \\
&\quad + \alpha^2 \|T^{j-1}(\omega, F_\alpha(\omega, \xi(\omega))) - F_\alpha^2(\omega, \xi(\omega))\| \\
&\leq j\alpha(1 - \alpha) \|F_\alpha(\omega, \xi(\omega)) - T(\omega, F_\alpha(\omega, \xi(\omega)))\| \\
&\quad + \alpha[(j - 1) - j\alpha + \alpha^j] \|\xi(\omega) - F_\alpha(\omega, \xi(\omega))\| \\
&\quad + \alpha^{j+1} \|F_\alpha(\omega, \xi(\omega)) - F_\alpha^2(\omega, \xi(\omega))\| \\
&\leq j(1 - \alpha)^2 + \alpha^2[(j - 1) - j\alpha + \alpha^j] \|\xi(\omega) - F_\alpha(\omega, \xi(\omega))\|
\end{aligned}$$

$$\begin{aligned}
 & + \alpha^{j+1} \|F_\alpha(\omega, \xi(\omega)) - F_\alpha^2(\omega, \xi(\omega))\| \\
 \leq & [j - (j + 1)\alpha + \alpha^{j+1}] \|\xi(\omega) - F_\alpha(\omega, \xi(\omega))\| \\
 & + \alpha^{j+1} \|F_\alpha(\omega, \xi(\omega)) - F_\alpha^2(\omega, \xi(\omega))\|.
 \end{aligned}$$

So by induction inequality (3.3) is valid for every  $\omega \in \Omega$  and  $m \geq 2$ .

Now consider, for  $\omega \in \Omega$

$$\begin{aligned}
 & \|F_\alpha(\omega, \xi(\omega)) - F_\alpha^2(\omega, \xi(\omega))\| \\
 \leq & (1 - \alpha)a \|F_\alpha(\omega, \xi(\omega)) - \xi(\omega)\| \\
 & + \alpha \|T^{n-1}(\omega, F_\alpha(\omega, \xi(\omega))) - F_\alpha^2(\omega, \xi(\omega))\| \\
 \leq & (1 - \alpha)a \|F_\alpha(\omega, \xi(\omega)) - \xi(\omega)\| \\
 & + [(n - 1) - n\alpha + \alpha^n] \|\xi(\omega) - F_\alpha(\omega, \xi(\omega))\| \\
 & + \alpha^n \|F_\alpha(\omega, \xi(\omega)) - F_\alpha^2(\omega, \xi(\omega))\|.
 \end{aligned}$$

It further implies that

$$\begin{aligned}
 & (1 - \alpha^n) \|F_\alpha(\omega, \xi(\omega)) - F_\alpha^2(\omega, \xi(\omega))\| \\
 \leq & [(1 - \alpha)a + (n - 1) - n\alpha + \alpha^n] \|\xi(\omega) - F_\alpha(\omega, \xi(\omega))\|,
 \end{aligned}$$

for every  $\omega \in \Omega$ . Now we arrive at

$$\begin{aligned}
 & \|F_\alpha(\omega, \xi(\omega)) - F_\alpha^2(\omega, \xi(\omega))\| \\
 \leq & (1 - \alpha^n)^{-1} [(1 - \alpha)a + (n - 1) - n\alpha + \alpha^n] \|\xi(\omega) - F_\alpha(\omega, \xi(\omega))\| \\
 \leq & (a + n)(1 - \alpha)(1 - \alpha^n)^{-1} - 1 \|\xi(\omega) - F_\alpha(\omega, \xi(\omega))\| \\
 = & [(a + n) \left( \sum_{i=0}^{n-1} \alpha^i \right)^{-1} - 1] \|\xi(\omega) - F_\alpha(\omega, \xi(\omega))\| \\
 = & g(\alpha) \|\xi(\omega) - F_\alpha(\omega, \xi(\omega))\|,
 \end{aligned}$$

for every  $\omega \in \Omega$ , where  $g(\alpha) = [(a + n) \left( \sum_{i=0}^{n-1} \alpha^i \right)^{-1} - 1]$ . Since  $g$  is continuous and decreasing for  $\alpha \in (0, 1]$  with  $g(1) = \frac{a}{n} < 1$ , there exists  $b \in (0, 1]$  such that  $g(1) < 1$  for  $\alpha \in (b, 1]$ . For such  $\alpha$ , the sequence of measurable mappings defined by  $\eta_m(\omega) = F_\alpha^m(\omega, \xi(\omega)) \rightarrow \eta(\omega)$ , for each  $\omega \in \Omega$ ,  $\eta: \Omega \rightarrow F$ , being the limit of the sequence of measurable functions, is also measurable (see remark 2.6). From (3.1) it follows that  $\eta$  is a random fixed point of  $T$ .  $\square$

**Example 3.2.** Let  $\Omega = [0, 1]$  and  $\Sigma$  be the sigma algebra of Lebesgue's measurable subsets of  $\Omega$ . Take  $X = R$  with  $d(x, y) = |x - y|$ , for  $x, y \in R$ . Define random operator  $T$  from  $\Omega \times X$  to  $X$  as,  $T(\omega, x) = \omega - x$ .

Define a fixed measurable mapping  $\xi: \Omega \rightarrow X$  as  $\xi(\omega) = \frac{\omega}{3}$ , for every  $\omega \in \Omega$ . Note that  $T$  is nonexpansive random operator. Since random operator equation  $T^2(\omega, \xi(\omega)) = \xi(\omega)$  holds for every  $\omega \in \Omega$ , therefore it is  $(2, 1)$ -rotative random operator. Thus the conditions of Theorem 3.1 are satisfied. Moreover a measurable mapping  $\eta: \Omega \rightarrow X$  defined as  $\eta(\omega) = \frac{\omega}{2}$ , for every  $\omega \in \Omega$ , serve as a unique random fixed point of  $T$ .

**Theorem 3.3.** *Let  $X$  be a separable metric space and  $T: \Omega \times X \rightarrow X$  be a  $\epsilon$ -contractive random operator. Let  $\xi_0: \Omega \rightarrow X$  be any measurable mapping such that a sequence  $\{T^n(\omega, \xi_0(\omega))\}$  has a point wise convergent subsequence of measurable mappings. Then  $T$  has a random periodic point.*

**Proof.** Let  $\{T^{n_i}(\omega, \xi_0(\omega))\}$  be a subsequence of  $\{T^n(\omega, \xi_0(\omega))\}$  such that  $T^{n_i}(\omega, \xi_0(\omega)) \rightarrow \xi(\omega)$  for each  $\omega \in \Omega$  as  $n_i \rightarrow \infty$  where  $\{n_i\}$  is a strictly increasing sequence of positive integers. The mapping  $\xi: \Omega \rightarrow X$  being point wise limit of sequence of measurable mappings is measurable. Define sequence of measurable mappings  $\xi_i: \Omega \rightarrow X$  as  $\xi_i(\omega) = T^{n_i}(\omega, \xi_0(\omega))$ . Given  $\epsilon > 0$ , there exists an integer  $n_0$  such that

$$d(\xi_i(\omega), \xi(\omega)) < \frac{\epsilon}{4}, \text{ for } i \geq n_0 \text{ and } \omega \in \Omega.$$

Put  $k = n_{i+1} - n_i$ . Consider,

$$\begin{aligned} d(\xi_{i+1}(\omega), T^k(\omega, \xi(\omega))) &= d(T^k(\omega, \xi_i(\omega)), T^k(\omega, \xi(\omega))) \\ &< d(\xi_i(\omega), \xi(\omega)) < \frac{\epsilon}{4}, \text{ for each } \omega \in \Omega. \end{aligned}$$

Now,

$$\begin{aligned} &d(\xi(\omega), T^k(\omega, \xi(\omega))) \\ &\leq d(\xi_{i+1}(\omega), T^k(\omega, \xi(\omega))) + d(\xi_{i+1}(\omega), \xi(\omega)) \\ &< \frac{\epsilon}{4} + \frac{\epsilon}{4} = \frac{\epsilon}{2}, \text{ for every } \omega \in \Omega. \end{aligned}$$

Now we claim that  $\xi$  is a random periodic point of  $T$ . To prove this, assume that  $\eta: \Omega \rightarrow X$  be any measurable mapping such that  $\eta(\omega) = T^k(\omega, \xi(\omega))$  but

$$\eta(\omega) \neq \xi(\omega), \text{ for some } \omega \in \Omega. \quad (3.4)$$

Which implies that  $0 < d(\eta(\omega), \xi(\omega)) < \epsilon$ . As  $T$  is a  $\epsilon$ -contractive random operator therefore for  $\omega \in \Omega$  for which (3.4) holds, we have

$$d(T(\omega, \xi(\omega)), T(\omega, \eta(\omega))) < d(\xi(\omega), \eta(\omega)).$$

Define  $h: \Omega \times X^2 \rightarrow R$  as,  $h(\omega, x(\omega), y(\omega)) = \frac{d(T(\omega, x(\omega)), T(\omega, y(\omega)))}{d(x(\omega), y(\omega))}$ , where  $x(\omega) \neq y(\omega) \in X$  for each  $\omega \in \Omega$ . Now  $h(\omega, \cdot, \cdot)$  is continuous at  $(\xi(\omega), \eta(\omega))$  for every  $\omega \in \Omega$  for which (3.4) is valid.

Take  $0 < \alpha < 1$ , then there exists  $\delta > 0$  such that  $x(\omega) \in B(\xi(\omega), \delta)$  and  $y(\omega) \in B(\eta(\omega), \delta)$  gives

$$d(T(\omega, x(\omega)), T(\omega, y(\omega))) < \alpha d(x(\omega), y(\omega)).$$

As,  $\lim_{r \rightarrow \infty} T^k(\omega, \xi_r(\omega)) = T^k(\omega, \xi(\omega)) = \eta(\omega)$ , for every  $\omega \in \Omega$ . So there exists  $n_1 \geq n_0$  such that

$$d(\xi_r(\omega), \xi(\omega)) < \delta$$



and

$$d(T^k(\omega, \xi_r(\omega)), \eta(\omega)) < \delta,$$

for  $r \geq n_1$  and  $\omega \in \Omega$ . Hence we have

$$d(T(\omega, \xi_r(\omega)), T(\omega, T^k(\omega, \xi_r(\omega)))) < \alpha d(\xi_r(\omega), T^k(\omega, \xi_r(\omega))). \quad (3.5)$$

Consider,

$$\begin{aligned} & d(\xi_r(\omega), T^k(\omega, \xi_r(\omega))) \\ & \leq d(\xi_r(\omega), \xi(\omega)) + d(\xi(\omega), T^k(\omega, \xi(\omega))) + d(T^k(\omega, \xi(\omega)), T^k(\omega, \xi_r(\omega))) \\ & < \frac{\epsilon}{4} + \frac{\epsilon}{2} + \frac{\epsilon}{4} = \epsilon, \end{aligned} \quad (3.6)$$

for  $r \geq n_1 \geq n_0$  and  $\omega \in \Omega$  for which (3.4) holds. Now using (3.5) and (3.6), we have

$$\begin{aligned} & d(T(\omega, \xi_r(\omega)), T(\omega, T^k(\omega, \xi_r(\omega)))) \\ & < \alpha d(\xi_r(\omega), T^k(\omega, \xi_r(\omega))) < d(\xi_r(\omega), T^k(\omega, \xi_r(\omega))) < \epsilon, \end{aligned}$$

for  $r \geq n_1$ . Since  $T$  is a  $\epsilon$ - contractive random operator so for  $r \geq n_1$  and  $q > 0$ , we have

$$\begin{aligned} & d(T^q(\omega, \xi_r(\omega)), T^q(\omega, T^k(\omega, \xi_r(\omega)))) \\ & < d(\xi_r(\omega), T^k(\omega, \xi_r(\omega))) < \frac{\epsilon}{\alpha}. \end{aligned}$$

Put  $q = n_{r+1} - n_r$ , we have  $d(\xi_{r+1}(\omega), T^k(\omega, \xi_{r+1}(\omega))) < \frac{\epsilon}{\alpha}$ . Hence,

$$d(\xi_s(\omega), T^k(\omega, \xi_s(\omega))) < \epsilon \alpha^{s-r}.$$

Now,

$$\begin{aligned} & d(\xi(\omega), \eta(\omega)) \leq d(\xi(\omega), \xi_s(\omega)) + d(\xi_s(\omega), T^k(\omega, \xi_s(\omega))) \\ & + d(T^k(\omega, \xi_s(\omega)), \eta(\omega)) \rightarrow 0, \text{ as } s \rightarrow \infty. \end{aligned}$$

for those  $\omega \in \Omega$  for which (3.4) holds. This contradiction concludes the result.  $\square$

**Corollary 3.4.** *If in theorem 3.2, the random periodic point  $\xi$  (say) of  $T$  satisfies*

$$d(\xi(\omega), T(\omega, \xi(\omega))) < \epsilon, \text{ for every } \omega \in \Omega. \quad (3.7)$$

*Then  $\xi$  is a random fixed point of  $T$ .*

**Proof.** Let  $k$  be the positive integer such that  $T^k(\omega, \xi(\omega)) = \xi(\omega)$ , for every  $\omega \in \Omega$ . If  $\xi$  is not a random fixed point of  $T$ , then  $\xi(\omega) \neq T(\omega, \xi(\omega))$  for some  $\omega \in \Omega$ . Since  $T$  is  $\epsilon$ - contractive random operator, using (3.7) we have

$$\begin{aligned} & d(\xi(\omega), T(\omega, \xi(\omega))) = d(T^k(\omega, \xi(\omega)), T^{k+1}(\omega, \xi(\omega))) \\ & < d(\xi(\omega), T(\omega, \xi(\omega))). \end{aligned}$$

This contradiction concludes the proof.  $\square$

**Remark 3.5.** If  $X$  is a separable compact metric space and  $T: \Omega \times X \rightarrow X$  is an  $\epsilon$ - contractive random operator. Then applying theorem 3.3, we conclude that  $T$  has a random periodic point.

**Theorem 3.6.** Let  $X$  be a separable compact metric space and  $T: \Omega \times X \rightarrow X$  be an  $\epsilon$ - contractive random operator. Then  $T$  has finitely many random periodic points.

**Proof.** Let  $\xi, \zeta: \Omega \rightarrow X$  be two random periodic points of  $T$  with  $\xi(\omega) \neq \zeta(\omega)$  and  $d(\xi(\omega), \zeta(\omega)) < \epsilon$  for some  $\omega \in \Omega$ . Let  $m, n \geq 1$  be two integers such that  $T^m(\omega, \xi(\omega)) = \xi(\omega)$  and  $T^n(\omega, \zeta(\omega)) = \zeta(\omega)$  for every  $\omega \in \Omega$ . Obviously  $T^{mn}(\omega, \xi(\omega)) = \xi(\omega)$  and  $T^{mn}(\omega, \zeta(\omega)) = \zeta(\omega)$  for each  $\omega \in \Omega$ . Now consider,

$$\begin{aligned} d(\xi(\omega), \zeta(\omega)) &= d(T^{mn}(\omega, \xi(\omega)), T^{mn}(\omega, \zeta(\omega))) \\ &< d(\xi(\omega), \zeta(\omega)), \end{aligned}$$

which is contradiction. Therefore any two random periodic point of  $T$  must be at least  $\epsilon$ - apart. Compactness of  $X$  prevents us defining infinitely many random periodic points from  $\Omega \times X$  to  $X$ .  $\square$

**Acknowledgement.** The authors are thankful to referee for precise remarks to improve the presentation of the paper.

## References

- [1] AUBIN, J. P., and FRANKOWSKA, H., *Set-Valued Analysis*, Birkhauser, Boston, 2009.
- [2] BEG, I., Random fixed points of random operators satisfying semicontractivity conditions, *Math. Japan.*, 46 (1) (1997), 151–155.
- [3] BEG, I., Random Edelstein theorem, *Bull. Greek Math. Soc.*, 45 (2001), 31–41.
- [4] BEG, I., Minimal displacement of random variables under Lipschitz random maps, *Topological Methods in Nonlinear Anal.*, 19 (2002), 391–397.
- [5] BEG, I., and ABBAS, M., Iterative procedures for solutions of random operator equations in Banach spaces, *J. Math. Anal. Appl.*, 315(1)(2006), 181–201.
- [6] BEG, I., and THAKUR, B.S., Solution of random operator equations using general composite implicit iteration process, *Int. J. Modern Math.*, 4(1)(2009), 19–34.
- [7] BHARUCHA-REID, A.T., *Random Integral Equations*, Academic Press, New York and London, 1972.
- [8] BHARUCHA-REID, A.T., Fixed point theorems in probabilistic analysis, *Bull. Amer. Math. Soc.*, 82 (1976), 641–645.
- [9] EDELSTEIN, M., On fixed and periodic points under contractive mapping, *J. London Math. Soc.*, 37 (1962), 74–79.
- [10] FOURNIER, G., and GÓRNIOWICZ, L., The Lefschetz fixed point theorem for some non compact multivalued maps, *Fund. Math.*, 94 (1977), 245–254.

- [11] GOEBEL, K., and KIRK, W.A., *Topics in Metric Fixed Point Theory*, Cambridge University Press, Cambridge 1990.
- [12] GOEBEL, K., and KOTER, M., A remark on nonexpansive mappings, *Canadian Math. Bull.*, 24 (1981), 113–115.
- [13] GÓRNIWICZ, L., *Topological Fixed Point Theory of Multivalued Mappings*, Kluwer, Dordrecht, 1999.
- [14] HALPERN, B., Periodic points on tori, *Pacific J. Math.*, 83 (1979), 117–133.
- [15] HANŠ, O., Reduzierende zulliällige transformaten, *Czechoslovak Math. J.*, 7 (1957), 154–158.
- [16] HANŠ, O., Random operator equations, in: *Proceedings of the Fourth Berkeley Symposium on Mathematical Statistics and Probability*, Vol. II, Part I, University of California Press, Berkeley, 1961, 185–202.
- [17] ITOH, S., Random fixed point theorems with an application to random differential equations in Banach spaces, *J. Math. Anal. Appl.*, 67 (1979), 261–273.
- [18] KALMOUN, E.M., Some deterministic and random vector equilibrium problems, *J. Math. Anal. Appl.*, 267 (2002), 62–75.
- [19] KAMPEN, J., On fixed points of maps and iterated maps and applications, *Nonlinear Anal.*, 42(3) (2000), 509–532.
- [20] MATSUOKA, T., The number of periodic points of smooth maps, *Ergod. Th. and Dynam. Sys.*, 8 (1989), 153–163.
- [21] O'REGAN, D., A continuation type result for random operators, *Proc. Amer. Math. Soc.*, 126, 7(1998), 1963–1971.
- [22] PAPAGEORGIOU, N.S., Random fixed point theorems for measurable multifunctions in Banach spaces, *Proc. Amer. Math. Soc.*, 97 (1986), 507–514.
- [23] PAPAGEORGIOU, N.S., On measurable multifunctions with stochastic domain, *J. Austral. Math. Soc. (Ser. A)*, 45 (1988), 204–216.
- [24] SPACEK, A., Zufällige Gleichungen, *Czechoslovak Math. J.*, 5 (1955), 462–466.

**Ismat Beg, Mujahid Abbas**

Center for Advanced Studies in Mathematics,  
Lahore University of Management Sciences,  
Lahore-54792, Pakistan  
Phone: 0092-42-35608229  
Fax: 0092-42-35722591  
e-mail: [ibeg@lums.edu.pk](mailto:ibeg@lums.edu.pk)  
[mujahid@lums.edu.pk](mailto:mujahid@lums.edu.pk)

**Akbar Azam**

Department of Mathematics,  
COMSATS Institute of Information Technology,  
Islamabad, Pakistan



# Evaluating a probabilistic model checker for modeling and analyzing retrial queueing systems\*

Tamás Bérczes<sup>a</sup>, Gábor Guta<sup>b</sup>, Gábor Kusper<sup>c</sup>  
Wolfgang Schreiner<sup>b</sup>, János Sztrik<sup>a</sup>

<sup>a</sup>Faculty of Informatics, University of Debrecen, Hungary

<sup>b</sup>Research Institute for Symbolic Computation (RISC), Johannes Kepler University,  
Linz, Austria

<sup>c</sup>Eszterházy Károly College, Eger, Hungary

*Submitted 6 April 2009; Accepted 27 May 2010*

*Dedicated to professor Béla Pelle on his 80<sup>th</sup> birthday*

## Abstract

We describe the results of analyzing the performance model of a finite-source retrial queueing system with the probabilistic model checker PRISM. The system has been previously investigated with the help of the performance modeling environment MOSEL; we are able to accurately reproduce the results reported in literature. The present paper compares PRISM and MOSEL with respect to their modeling languages and ways of specifying performance queries and benchmark the executions of the tools.

## 1. Introduction

The *performance analysis* of computing and communicating systems has always been an important subject of computer science. The goal of this analysis is to make predictions about the quantitative behavior of a system under varying conditions, e.g., the expected response time of a server under varying numbers of

---

\*The work is supported by the TÁMOP 4.2.1./B-09/1/KONV-2010-0007 project. The project is implemented through the New Hungary Development Plan, co-financed by the European Social Fund and the European Regional Development Fund.

service requests, the average utilization of a communication channel under varying numbers of communication requests, and so on.

To perform such an analysis, however, first an adequate mathematical model of the system has to be developed which comprises the interesting aspects of the system but abstracts away from details that are irrelevant to the questions addressed. Originally, these models were developed purely by manual efforts, typically in formal frameworks based on queuing theory, stochastic Petri networks, and the like, which can be ultimately translated into continuous time Markov chains (CTMCs) as the fundamental mathematical basis [18]. Since the manual creation of complex models is tedious and error-prone, specification languages and corresponding tools were developed that automated the model creation from high-level system descriptions. Since the generated models cannot typically be solved analytically, simulation-based techniques were applied in order to predict their quantitative behavior from a large number of sampled system runs. Latter on, however, the underlying systems of equations were solved (for fixed parameter values) by iterative numerical calculations, thus deriving (mathematically exact but numerically approximated) solutions for the long-term (steady state) behavior of the system.

One tool of this kind is MOSEL (Modeling, Specification, and Evaluation Language) [14, 3] with its latest incarnation MOSEL-2 [15]. The software has a high-level specification language for modeling interconnected queue networks where transitions execute at certain rates to move entities across queues. The environment supports various back ends for simulating the model system or for computing numerical solutions of the derived system of steady-state equations. In particular, it may construct a stochastic Petri net model as input to the SPNP solver [10].

While above developments emerged in the *performance modeling and evaluation* community, also the *formal methods* community has produced theoretical frameworks and supporting tools that are, while coming from a different direction, nevertheless applicable to performance analysis problems. Originally, the only goal of formal methods was to determine qualitative properties of systems, i.e., properties that can be expressed by formal specifications (typically in the language of temporal logic).

In the last couple of years, however, the formal methods community also got more and more interested in systems that exhibit stochastic behavior, i.e., systems whose transitions are executed according to specific rates (respectively probabilities); this gives rise to continuous time (respectively discrete time) Markov chains like those used by the performance modeling community and to questions about quantitative rather than qualitative system properties. To pursue this new direction of *quantitative verification* [12], model checking techniques were correspondingly extended to *stochastic/probabilistic model checking* [13].

A prominent tool in this category is the probabilistic model checker PRISM [16, 9] which provides a high-level modeling language for describing systems that exhibit probabilistic behavior, with models based on continuous-time Markov chains (CTMCs) as well as discrete-time Markov chains (DTMCs) and Markov decision procedures (MDPs). For specifying system properties, PRISM uses the probabilis-

tic logics CSL (continuous stochastic logic) for CTMCs and PCTL (probabilistic computation tree logic) for DTMCs and MDPs, both logics being extensions of CTL (computation tree logic), a temporal logic that is used in various classical model checkers for specifying properties [7]. While some probabilistic model checkers are faster, PRISM provides a comparatively comfortable modeling language; for a more detailed comparison, see [11].

The fact that the previously disjoint areas of performance evaluation and formal methods have become overlapping is recognized by both communities. While originally only individual authors hailed this convergence [8], today various conferences and workshops are intended to make both communities more aware of each others' achievements [5, 21]. One attempt towards this goal is to compare techniques and tools from both communities by concrete application studies. The present paper is aimed at exactly this direction.

The starting point of our investigation is the paper [19] which discusses various performance modeling tools; in particular, it presents the application of MOSEL to the modeling and analysis of a retrieval queuing system previously described in [1] and latter refined in [17]. The goal of the present paper is to construct PRISM models analogous to the MOSEL models presented in [19] for computing the performance measures presented in the above paper, to compare the results derived by PRISM with those from MOSEL, to evaluate the usability and expressiveness of both frameworks with respect to these tasks, to benchmark the tools with respect to their efficiency (time and memory consumption), and finally to draw some overall conclusions about the suitability of PRISM to performance modeling compared with classical tools in this area.

The rest of the paper (which is based on the more detailed technical report [4]) is structured as follows: Section 2 describes the application to be modeled and the questions to be asked about the model; Section 3 summarizes the previously presented MOSEL solution; Section 4 presents the newly developed PRISM solution; Section 5 gives the experimental results computed by PRISM in comparison to those computed by MOSEL and also gives benchmarks of both tools; Section 6 concludes and gives an outlook on further work.

## 2. Problem description

### 2.1. Problem overview

In this section we give a brief overview on the model of the retrieval queueing system presented in [19]. The variable names used latter in the model are indicated in italics in the textual description. The dynamic behavior of the model is illustrated by UML state machine diagrams [20].

The system contains a single server and  $NT$  terminals. Their behavior is as follows:

- Intuitively, terminals send requests to the server for processing. If the server is busy, the terminals retry to send the request latter. More precisely, the

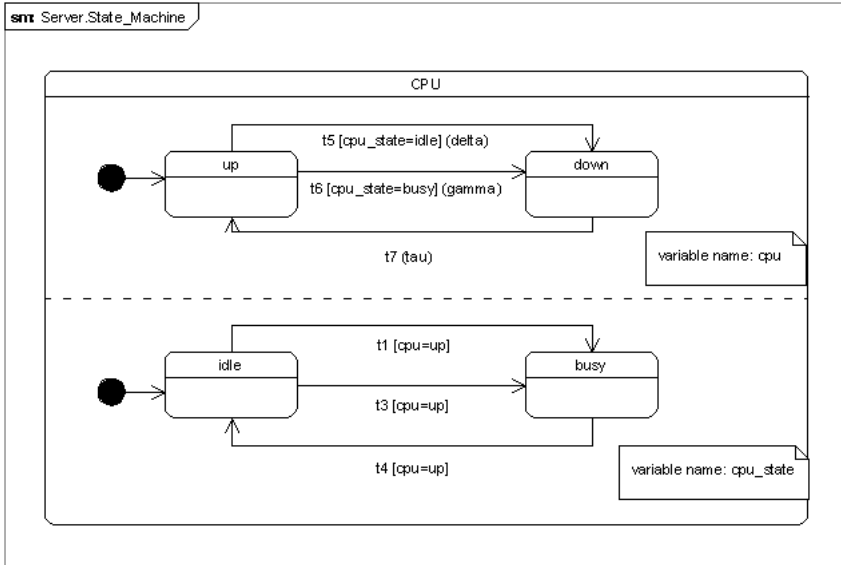


Figure 1: State machine representation of the server

terminals can be in three different states (which are named in parentheses):

1. ready to generate a primary call (*busy*),
  2. sending repeated calls (*retrying*) and
  3. under service by the server (*waiting*).
- The server according to its CPU state (*cpu*) can be operational ( $cpu=cpu\_up$ ) or non-operational ( $cpu=cpu\_down$ ): if it is operational we distinguish between two further states (*cpu\_state*): idle ( $cpu\_state=cpu\_idle$ ) and busy ( $cpu\_state=cpu\_busy$ ).
  - In the initial state of the system, the server is operational ( $cpu=cpu\_up$ ) waiting for requests ( $cpu\_state=cpu\_idle$ ) and all terminals are ready to generate a primary call.

## 2.2. Finite state model

The behavior of the system can be described by the state transitions of the terminals and the server, which occur at different rates.

We extend the standard UML [20] state machine diagram notation and semantics to present our model in an easy-to-read way. According to the standard, the diagram contains states and transitions; the transitions in different swim-lanes can occur independently. Our extensions are the following:



- Every comment of a swim-lane contains a variable name which is changed by the transition of that lane.
- Each transition is associated with a triple of a label, a guard (in square brackets) and a rate(in parentheses); if there is no rate indicated, then the rate equals 1.
- A parallel composition semantics: the set of the states of the composed system is the Cartesian product of the state sets of the two swim-lanes or state machines. The composed state machines can make a transition whenever one of the original state machines can make one, except if multiple transitions in different original state machines have the same label: in that case, they must be taken simultaneously.

In Figure 1 we show the state transitions of the server:

- t1 (The server starts to serve a primary call)** If the server is in operational state and idle, it can receive a primary call and become busy.
- t2 (The server rejects to serve a primary call)** If the server is operational and busy, it can reject a primary call.
- t3 (The server starts to serve a retried call)** If the server is in operational state and idle, it can start to serve a repeated call.
- t4 (The server finishes a call)** If the server is operational and busy, it can finish the processing of the call.
- t5 (An idle server becomes inoperable)** If the server is in operational state and idle, it can become inoperable with rate  $\delta$ .
- t6 (A busy server becomes inoperable)** If the server is in operational state and busy, it can become inoperable with rate  $\gamma$ .
- t7 (A server gets repaired)** If the server is inoperable, it can become operable again with rate  $\tau$ .

The state transitions of the terminal are described in Figure 2:

- t1 (The server starts to serve a primary call)** The call of a terminal which issues a primary call is accepted and it becomes a waiting terminal with probability  $\lambda$ .
- t2 (The server rejects a primary call)** The call of a terminal which issues a primary call is rejected and it becomes a retrying terminal with probability  $\lambda$ .
- t3 (The server starts to serve a retried call)** The call of a terminal which retries a call is accepted and it becomes a waiting terminal with probability  $\nu$ .
- t4 (The server finishes a call)** The call of a terminal is finished and it becomes ready to generate a new primary call again with rate  $\mu$ .

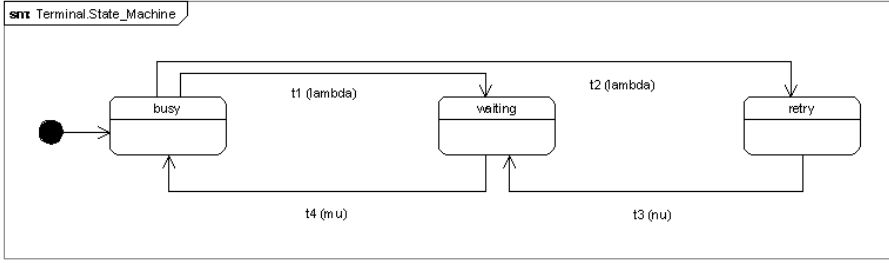


Figure 2: State machine representation of the terminals

The system can be represented alternatively by merging the server and the terminals into a single system as modelled in the original MOSEL model [19]: the guard conditions of all transitions with the same label are logically conjoined and their probabilities are multiplied.

### 2.3. Mathematical model

In this section we describe the mathematical formulation of the queries. The state of the system at time  $t$  can be described by the process  $X(t)=(cpu(t), cpu\_state(t), retrying\_terminals(t))$ , where  $cpu(t)=0$  ( $cpu\_up$ ) if the server is operable,  $cpu(t)=1$  ( $cpu\_down$ ) if the server is not operable,  $cpu\_state(t)=0$  ( $cpu\_idle$ ) if the server is idle and  $cpu\_state(t)=1$  ( $cpu\_busy$ ) if the server is busy and  $retrying\_terminals(t)$  describe the number of repeated calls at time  $t$ . The number of waiting terminals and busy terminals are not expressed explicitly in the mathematical model. Their values can be calculated according to the following equations:

- $waiting\_terminals=0$  if  $cpu\_state=cpu\_idle$ ,
- $waiting\_terminals=1$  if  $cpu\_state=cpu\_busy$ ,
- $busy\_terminals=NT-(waiting\_terminals+retrying\_terminals)$ ,

Because of the exponentiality of the involved random variables and the finite number of sources, this process is a Markov chain with a finite state space. Since the state space of the process  $X(t)$ ,  $t \geq 0$  is finite, the process is ergodic for all reasonable values of the rates involved in the model construction. From now on, we assume that the system is in the steady-state.

We define the stationary probabilities by:

$$P(q, r, j) = \lim_{t \rightarrow \infty} P(cpu(t), cpu\_state(t), retrying\_terminals(t)),$$

$$q = 0, 1, r = 0, 1, j = 0, \dots, NT - 1,$$

The main steady-state system performance measures can be derived as follows:

- Utilization of the servers

$$cpuutil = \sum_{j=0}^{NT-1} P(0, 1, j)$$

- Availability of the servers

$$goodcpu = \sum_{r=0}^1 \sum_{j=0}^{NT-1} P(0, r, j)$$

- Utilization of the repairman

$$repairutil = \sum_{r=0}^1 \sum_{j=0}^{NT-1} P(1, r, j) = 1 - goodcpu$$

- Mean rate of generation of primary calls

$$\begin{aligned} busyterm &= E[NT - cpu\_state(t) - retrying\_terminals(t); cpu(t) = 0] \\ &= \sum_{r=0}^1 \sum_{j=0}^{NT-1} (NT - r - j) P(0, r, j) \end{aligned}$$

- Utilization of the sources

$$termutil = \frac{busyterm}{NT}$$

- Mean rate of generation of repeated calls

$$retravg = E[retrying\_terminals(t); cpu(t) = 0] = \sum_{r=0}^1 \sum_{j=0}^{NT-1} j P(0, r, j)$$

- Mean number of calls staying in the server

$$waitall = E[cpu\_state(t)] = \sum_{q=0}^1 \sum_{j=0}^{NT-1} P(q, 1, j)$$

- Mean number of calls staying in the orbit

$$retrall = E[retrying\_terminals(t)] = \sum_{q=0}^1 \sum_{r=0}^1 \sum_{j=0}^{NT-1} j P(q, r, j)$$

- Overall utilization

$$\text{overallutil} = \text{cpuutil} + \text{repairutil} + NT * \text{termutil}$$

- Mean number of calls staying in the orbit or in the server

$$\text{meanorbit} = \text{waitall} + \text{retrall}$$

- Mean response times

$$E[T] = \frac{E[\text{retrying\_terminals}(t)] + E[\text{cpu\_state}(t)]}{\lambda * \text{busyterm}}$$

The last equation is essentially a consequence of *Little's Theorem*, a classical result in queuing theory [6], which describes for a queuing system in equilibrium by the equation  $T = L/\lambda$  the relationship between the long-term average waiting time  $T$  of a request, the long-term average number of requests  $L$  pending in the system, and the long-term average request arrival rate  $\lambda$ . Furthermore, according to *Jackson's Theorem*, a network of  $N$  queues with arrival rates  $\lambda$  may (under rather loose assumptions) be considered as a single queue with arrival rate  $\bar{\lambda} = \lambda N$ . This relationship will become crucial in the use of MOSEL and PRISM described in the following sections because it allows us to reduce questions about average timing properties of a system to questions about quantities which can be deduced from the (long-term) observation of states.

## 2.4. Questions about the system

Our goal is to study various quantitative properties of the presented models to get a deeper understanding of the modelled systems. The following properties are analyzed:

**cpuutil** The ratio of the time the server spends serving calls compared to the total execution time ( $0 \leq \text{cpuutil} \leq 1$ ).

**goodcpu** The ratio of the time when the server is operable compared to the total execution time ( $0 \leq \text{goodcpu} \leq 1$ ).

**repairutil** The ratio of the time when the server is inoperable compared to the total execution time ( $0 \leq \text{repairutil} \leq 1$ ).

**busyterm** The average number of served terminals while the system is operable ( $0 \leq \text{busyterm} \leq NT$ ).

**termutil** The ratio of served terminals while the system is operable to the total number of terminals ( $0 \leq \text{termutil} \leq 1$ ).

**retravg** The average number of retrying terminals while the system is operable ( $0 \leq \text{retravg} \leq NT - 1$ ).

**waitall** The average number of waiting terminals during the total system execution time ( $0 \leq \text{waitall} \leq 1$ ).

**retrall** The average number of retrying terminals during the total system execution time ( $0 \leq \text{retrall} \leq NT-1$ ).

**overallutil** The sum of the system average utilization, i.e., the sum of *cpuutil*, *repairutil* and  $NT * \text{termutil}$  ( $0 \leq \text{overallutil} \leq NT+1$ ).

**meanorbit** The average number of retrying terminals and waiting terminals during the total system execution time ( $0 \leq \text{retrall} \leq NT$ ).

**resptime** The mean response time, i.e., the average waiting time till a call of a terminal is successfully accepted.

## 2.5. Different versions of the system

In [19], actually four slightly different systems were described:

**continuous** The presented model.

**non-continuous** If the server becomes inoperable, then the call has to be retried (the waiting terminal becomes retrying).

**continuous, intelligent** It can also reject a call if the server is inoperable (the original model cannot handle a call if the server is inoperable).

**non-continuous, intelligent** The combination of the non-continuous and intelligent model.

The latter three variants are not formally described in the present paper. However, they have been implemented and have been used for the experiments in Section 5.

## 3. Modeling and analyzing in MOSEL

The MOSEL language (Modeling Specification and Evaluation Language) was developed at the University of Erlangen. The MOSEL system uses a macro-like language to model communication networks and computer systems, like stochastic Petri nets. The MOSEL tool contains some language features, like variables and functions in the style of the C programming language. The MOSEL system calls an external tool after having translated the MOSEL code into the respective tool's format. For example the Petri net analysis tool SPNP and the state analysis tool MOSES can be used. Because of page limitation the interested reader is referred to [4] where the source codes and technical details of our MOSEL model can be found.

## 4. Modeling and analyzing in PRISM

In this section we describe how we translate the model described in Section 2 into a PRISM model. Further information about the PRISM system can be found in [16]. In the first subsection we show the source-code of the PRISM model; in the second subsection we formulate questions in the model.

### 4.1. Translating the model to PRISM

In this subsection and the following ones, we present the full source code of the PRISM model (in *verbatim*) surrounded by detailed comments. The model description has 4 main parts:

- the type of the model,
- the constant declarations,
- the module declarations and
- the reward specifications.

In our case, all models are represented in Continuous-time Markov chains model, which is indicated by the keyword `stochastic`.

```
stochastic
```

Constants can be used in two manners:

- uninitialized constants denote parameters of the model and,
- initialized constants denote fixed values.

The parameters of the model are the following constants:

```
const int NT; // number of terminals
const double lambda; // the rate of primary call generation
const double mu; // the rate of the call servicing
const double nu; // the rate of repeated call generation
const double delta; // the failure rate in idle state of the server
const double gamma; // the failure rate in busy state of the server
const double tau; // the repair rate of the server
```

In our simulation we do not distinguish between the failure rate in idle and busy state, so we equal `gamma` with `delta`.

We define two pairs of constants to represent the state of the server to make the model human-readable:

```
const int cpu_up = 0; // the server is operable
const int cpu_down = 1; // or not
const int cpu_busy = 0; // the server is busy serving a call
const int cpu_idle = 1; // or idle waiting for a call
```

The next fragment are the module definitions. A module definition is started with the `module` keyword and is closed with the `endmodule` keyword. All modules contain state variables and state transitions. We have two modules `TERMINALS` and `SERVER` described in the following subsections

## 4.2. Terminals

The module `TERMINALS` represents the set of the terminals. We keep track of the number of terminals in specific states, because in PRISM it is not possible to have multiple instances of a module. Thus all variables range from 0 to the maximal number of terminals, which is denoted by the range indicator within square brackets in the source code.

```
module TERMINALS
busyTs      : [0..NT] init NT;
retryingTs  : [0..NT] init 0;
waitingTs   : [0..NT] init 0;
```

We have the following variables in the model :

- `busyTs` is the number of terminals, which are capable to generate primary calls (they are busy with local tasks and may generate calls to the server);
- `retryingTs` is the number of retrying terminals, i.e., terminals which have generated an unsuccessful call and are retrying the same call;
- `waitingTs` is the number of waiting terminals, i.e., terminals which have issued a successful call to the server and wait for the answer of the call.

In the current model, we have only one server, therefore the number of waiting terminals never be more than 1. Initially all terminals are busy terminals.

The transitions are represented in form `[l] g -> r : u`. The transition with label `l` occurs if the guard `g` evaluates to true; the rate of the transition is `r`, the values of the state variables are updated according to `u`. The labels serve as synchronization identifiers for parallel composition. Transitions with the same label in different modules execute together, i.e., all guards of the transition must be true and the total transition rate is the product of the individual transition rates. We also have to notice that the transitions of the terminals have their counterparts on the server side, which make the transition guards unique.

The transition with label `t1` describes the scenario of a successful primary call:

```
[t1] busyTs > 0 & waitingTs < NT -> lambda*busyTs :
      (busyTs' = busyTs-1) & (waitingTs' = waitingTs+1);
```

The transition occurs if there are some busy terminals and the number of waiting terminals is lower than the number of terminals. The second part of the guard condition is purely technical to explicitly state that the value of `waitingTs` is not

greater than the maximally allowed value. (According to the model semantics we know that it never becomes greater than one, because the server serves only one call at once.) All busy terminals produce that call with rate  $\lambda$ , so the rate is  $\lambda$  multiplied by the number of busy terminals. After that transition, the number of busy terminals decreases by one and the number of busy terminals increases by one.

The transition with label `t2` describes the scenario of an unsuccessful primary call:

```
[t2] busyTs > 0 & retryingTs < NT -> lambda*busyTs :
    (busyTs' = busyTs-1) & (retryingTs' = retryingTs+1);
```

The transition occurs if there are some busy terminals and the number of retrying terminals is lower than the number of terminals. The second part of the guard condition is also purely technical to explicitly state that the value of `waitingTs` is not greater than the maximally allowed value. (According the model semantics we know that it never becomes grater than maximal number, because the sum of the terminal variables equals the number of terminals.) All busy terminals produce that call with rate  $\lambda$ , so the rate is  $\lambda$  multiplied by the number of busy terminals. After that transition the number of busy terminals decreases by one and the number of busy terminal increases by one.

The transition with label `t3` describes the scenario of a successfully repeated call:

```
[t3] retryingTs > 0 & waitingTs < NT -> nu*retryingTs :
    (retryingTs' = retryingTs-1) & (waitingTs' = waitingTs+1);
```

The transition occurs if there are some retrying terminals and the number of waiting terminals is smaller than the number of terminals. All retrying terminals produce the calls with rate  $\nu$ , so the rate is  $\nu$  multiplied by the number of busy terminals. After that transition, the number of retrying terminals decreases by one and the number of waiting terminals increases by one.

The transition with label `t4` describes the scenario of an answer for a waiting terminal:

```
[t4] waitingTs > 0 & busyTs < NT -> 1 :
    (waitingTs' = waitingTs-1) & (busyTs' = busyTs+1);
```

The transition occurs if there are some waiting terminals and the number of busy terminals smaller than the number of terminals. Its rate is determined by the call serving rate on the server side (see below). After that transition, the number of retrying terminals decreases by one and the number of waiting terminals increases by one.

endmodule



### 4.3. Server

The second module represents the server by two binary state variables. The variable `cpu` expresses the operability of the server by the values 0 and 1, which are denoted by the constants `cpu_up` and `cpu_down`, respectively. The variable `cpu_state` the state of the server by values 0 and 1, which are denoted by the constants `cpu_busy` and `cpu_idle`, respectively.

```
module SERVER
cpu : [cpu_up..cpu_down] init cpu_up;
cpu_state : [cpu_busy..cpu_idle] init cpu_idle;
```

The transition with label `t1` describes the server side scenario of a successful primary call. It occurs, if the server is operable and idle. After the transition, the server becomes busy.

```
[t1] cpu = cpu_up & cpu_state = cpu_idle -> 1 :
  (cpu_state' = cpu_busy);
```

The transition with label `t2` describes the server side scenario of an unsuccessful primary call. It occurs, if the server is operable and busy. After the transition, the state of the server doesn't change.

```
[t2] cpu = cpu_up & cpu_state = cpu_busy -> 1 :
  (cpu' = cpu) & (cpu_state' = cpu_state);
```

The transition with label `t3` describes the server side scenario of a successful primary call. It is the same as the transition `t1`, because the server can't distinguish between a primary and a repeated call.

```
[t3] cpu = cpu_up & cpu_state = cpu_idle -> 1 :
  (cpu_state' = cpu_busy);
```

The transition with label `t4` describes the server side scenario of finishing a call (a successful call served). It occurs with rate  $\mu$  and the server becomes idle after the transition.

```
[t4] cpu = cpu_up & cpu_state = cpu_busy & mu > 0 -> mu :
  (cpu_state' = cpu_idle);
```

The transition with label `t5` describes the scenario when an idle server becomes inoperable. It occurs, if the server is operable and idle with rate  $\gamma$ . If a server becomes inoperable, it keeps its state. After it gets repaired, it continues the processing, if it was busy at the time of the failure.

```
[t5] cpu_state = cpu_idle & cpu = cpu_up & delta > 0 -> delta :
  (cpu' = cpu_down);
```

The transition with label `t6` describes the scenario when a busy server becomes inoperable. It occurs, if the server is operable and busy with rate  $\delta$ .

```
[t6] cpu_state = cpu_busy & cpu = cpu_up & gamma > 0 -> gamma :
    (cpu' = cpu_down);
```

The transition with label `t7` describes the scenario when a server gets repaired. It occurs, if the server is inoperable with rate  $\tau$ .

```
[t7] cpu = cpu_down & tau > 0 -> tau : (cpu' = cpu_up);
```

```
endmodule
```

#### 4.4. Rewards

The last section of a model description is the declaration of rewards. Rewards are numerical values assigned to states or to transitions. Arbitrary many reward structures can be defined over the model and they can be referenced by a label. We use rewards to define the various questions defined in Section 2.4.

The first reward is the server utilization (*cpuutil*). It assigns a value 1 to all states where the server is operable and busy.

```
rewards "cpuutil"
    cpu = cpu_up & cpu_state = cpu_busy : 1;
endrewards
```

The reward *goodcpu* assigns 1 to all states where the server is operable.

```
rewards "goodcpu"  cpu = cpu_up : 1; endrewards
```

The reward *repairutil* assigns 1 to all states where the server is inoperable.

```
rewards "repairutil"  cpu = cpu_down : 1; endrewards
```

The reward *busyterm* assigns the number of busy terminals to all states where the server is operable.

```
rewards "busyterm"  cpu = cpu_up : busyTs; endrewards
```

The reward *termutil* assigns the ratio of the busy terminals over the total number of terminals to all states where the server is operable.

```
rewards "termutil"  cpu = cpu_up : busyTs/NT; endrewards
```

The reward *retravg* assigns the number of retrying terminals to all states where the server is operable.

```
rewards "retravg"  cpu = cpu_up : retryingTs; endrewards
```

The reward *waitall* assigns the number of waiting terminals to states with such terminals.

```
rewards "waitall"  waitingTs > 0 : waitingTs; endrewards
```

The reward *waitall* assigns the number of retrying terminals to states with such terminals.

```
rewards "retrall"  retryingTs > 0 : retryingTs; endrewards
```

The reward *meanorbit* assigns the number of retrying and waiting terminals to states with such terminals.

```
rewards "meanorbit"
  retryingTs > 0 : retryingTs;
  waitingTs > 0 : waitingTs;
endrewards
```

The reward *pending* computes the number of pending calls (calls by terminals that are waiting or retrying); the relevance of this reward for computing the mean response time *response* will be explained in the next subsection.

```
rewards "pending"
  retryingTs > 0 : retryingTs;
  waitingTs > 0 : waitingTs;
endrewards
```

The reward *overallutil* assigns to the all states the total number of all busy elements, i.e., the server, if it is busy or is under repair (a repair unit is busy with its repair), and all busy terminals.

```
rewards "overallutil"
  cpu = cpu_up & cpu_state = cpu_busy : 1;
  cpu = cpu_down : 1 ;
  cpu = cpu_up : busyTs;
endrewards
```

## 4.5. Questions about the System in PRISM

As we mentioned in the introduction, in PRISM the queries about the CTMC models can be formulated in CSL (Continuous Stochastic Logic). CSL is a branching-time logic similar to CTL or PCTL [2]. It is capable to express queries about both transient and steady-state properties. Transient properties refers to the values of the rewards at certain times and the steady-state properties refer to long-run rewards.

The PRISM system support not only evaluating predicates about the rewards, but also queries about the rewards. In our experiments we used only the following one CSL construction:  $R\{ "1" \}=? [ S ]$ . This query ask for the expected long-run reward of the structure labelled with 1. Most questions about the model described in Section 2.4 can be formulated as CSL expressions.

```

R{"cpuutil"}=? [ S ]
R{"goodcpu"}=? [ S ]
R{"repairuti"}=? [ S ]
R{"busyterm"}=? [ S ]
R{"termutil"}=? [ S ]
R{"retravg"}=? [ S ]
R{"waitall"}=? [ S ]
R{"retrall"}=? [ S ]
R{"overallutil"}=? [ S ]
R{"meanorbit"}=? [ S ]

```

The response time (`resptime`) cannot be directly calculated from a CSL query, because CSL does not allow us to ask questions about execution times (rather than say probabilities or long-term average rewards). We rather resort to queuing theory and apply the definition of  $E[T]$  stated in Section 2 which can be expressed as

$$\text{resptime} = \text{pending} / (\lambda * \text{busyterm})$$

Since this calculation is not directly expressible as a CSL query, we apply a post-processor to compute `resptime` from the values for `pending` and `busyterm` generated by PRISM from above CSL queries. Similar to MOSEL, we can thus reduce questions about timing properties of a system to the computation of quantities that can be derived from system states and are thus amenable to CSL queries in PRISM.

## 5. Experimental results

In this section, we show the result of the experiments carried through with PRISM. The parameters used for the experiments are listed in Figure 4; they are the same as published in [19]. The results of the experiments with PRISM are presented in diagrams Figure 5, 6, 7, 8, 9, 10, whereas the raw results can be seen in Tables in [4].

The experiments was performed in two main steps: the execution of the experiments through the GUI of PRISM and the post-processing of the results. We selected the appropriate CSL query according the Figure 3 and set up the parameters according the Figure 4; after the execution of PRISM the results were exported to CSV files for further processing. The post processing happened with a help of Python scripts.

### 5.1. Analysis results

The diagrams compared with the ones presented in [19] clearly show that the two models (MOSEL and PRISM) produce identical results for the same parameters. Comparing the raw results of the experiments, it shows that they are differ only after the 5th decimal digit. The quality of the results produced with PRISM is this the same as the ones produced in MOSEL.

Nr. of the experiment	used reward(s)
1	pending and termutil
2	overallutil
3	meanorbit
4	pending and termutil
5	overallutil
6	meanorbit

Figure 3: Rewards calculated in the experiments

Exp. Nr.	NT	$\lambda$	$\mu$	$\nu$	$\gamma/\delta$	$\tau$	X axis
1	6	0.8	4	0.5	X axis	0.1	0. 0.01. ..., 0.12
2	6	0.1	0.5	0.5	X axis	0.1	0. 0.01. ..., 0.12
3	6	0.1	0.5	0.05	X axis	0.1	0. 0.01. ..., 0.12
4	6	0.8	4	0.5	0.05	X axis	0.5. 1.0. ..., 4.0
5	6	0.05	0.3	0.2	0.05	X axis	0.5. 1.0. ..., 4.0
6	6	0.1	0.5	0.05	0.05	X axis	0.5. 1.0. ..., 4.0

Figure 4: Parameters of the experiments

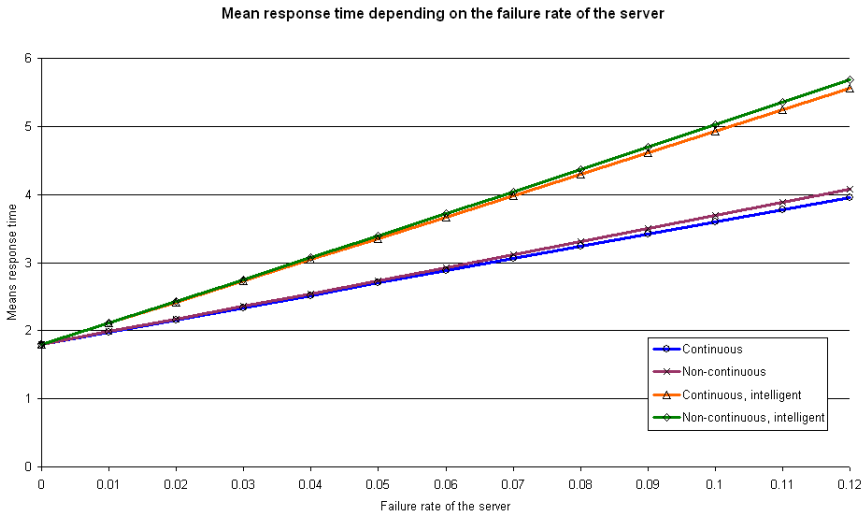


Figure 5: Results of the 1st experiment

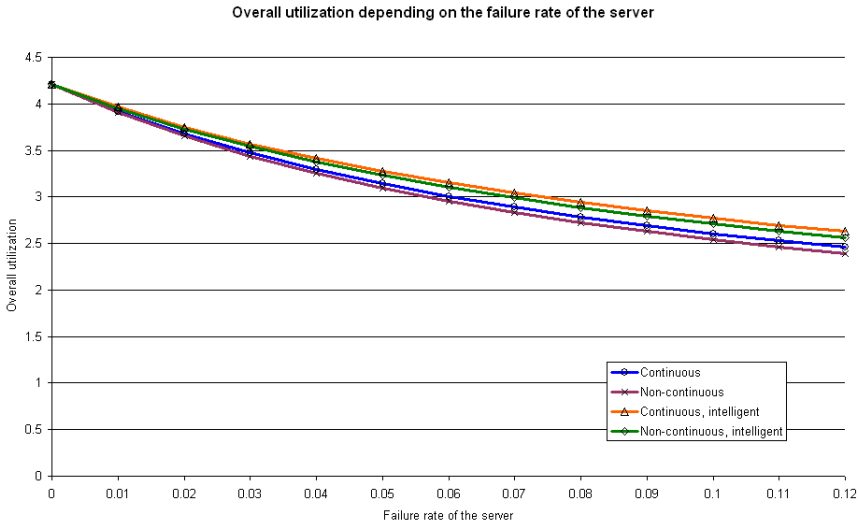


Figure 6: Results of the 2nd experiment

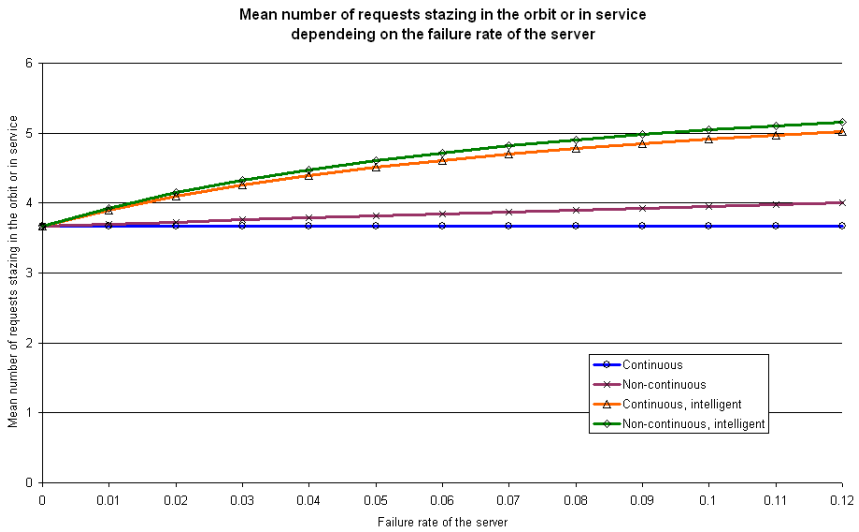


Figure 7: Results of the 3rd experiment

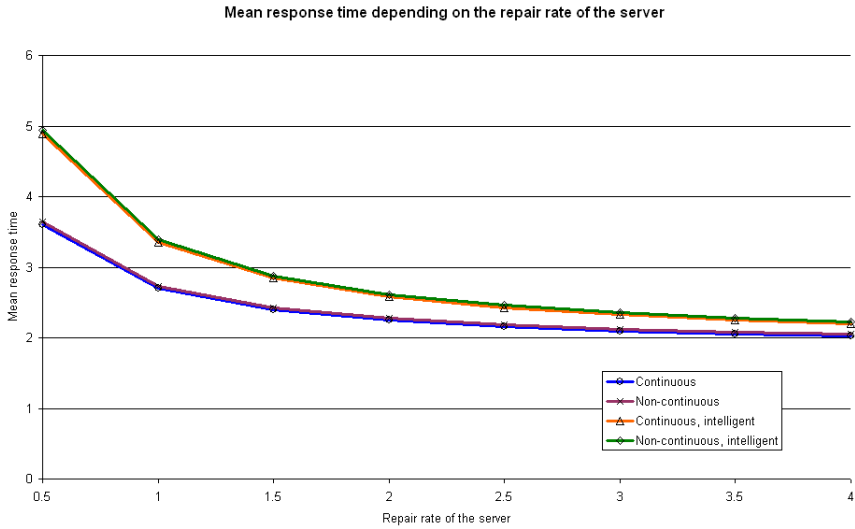


Figure 8: Results of the 4th experiment

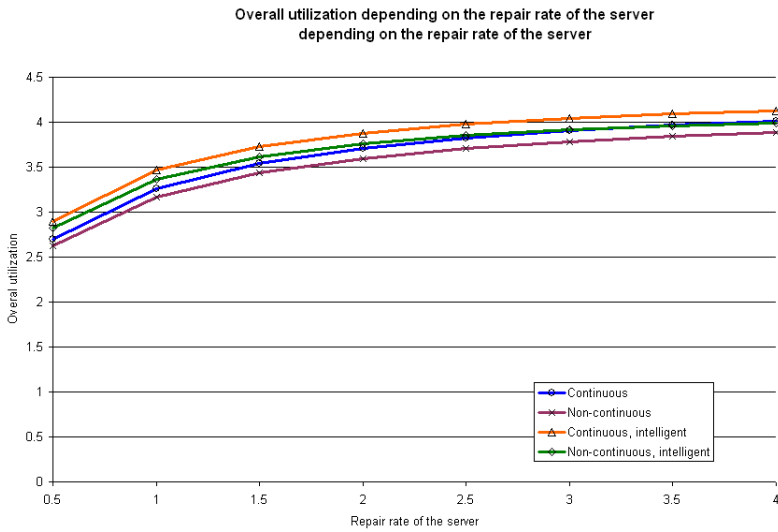


Figure 9: Results of the 5th experiment

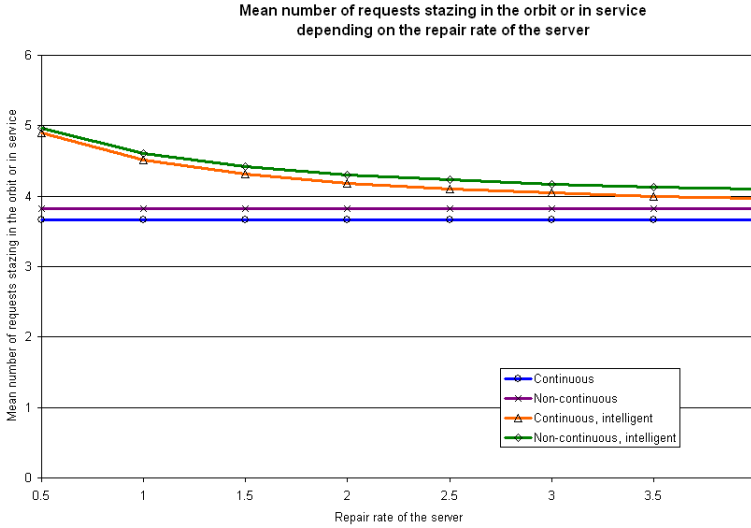


Figure 10: Results of the 6th experiment

## 5.2. Tool benchmarks

A benchmark was carried through to compare the efficiency of the two tools. The parameters of the machine that was used for the benchmark: P4 2.6GHz with 512KB L2 Cache and 512MB of main memory. Unfortunately MOSEL is not capable to handle models where the number of terminals ( $NT$ ) is greater than 126, such that the runtime of the benchmarks (which in PRISM especially depend on  $NT$ ) remain rather small.

Both of the tools were tested with the described model using the following parameters:  $\lambda = 0.05$ ,  $\mu = 0.3$ ,  $\nu = 0.2$ ,  $\gamma = \delta = 0.05$ ,  $\tau = 0.1$ . The comparison of the two tools can be seen in Figure 11 and Figure 12. In Figure 13, we can see a more detailed description of the PRISM benchmark (the times of the model construction and model checking are indicated separately).

The following preliminary conclusions can be drawn from benchmark:

- The execution times of the MOSEL system almost stay constant independently of  $NT$ ;
- The execution times of the PRISM system increase rapidly with the increase of  $NT$ .
- The model construction time in PRISM dominates the execution time rather than the model checking time (also [11] reports on the overhead of PRISM for model generation).



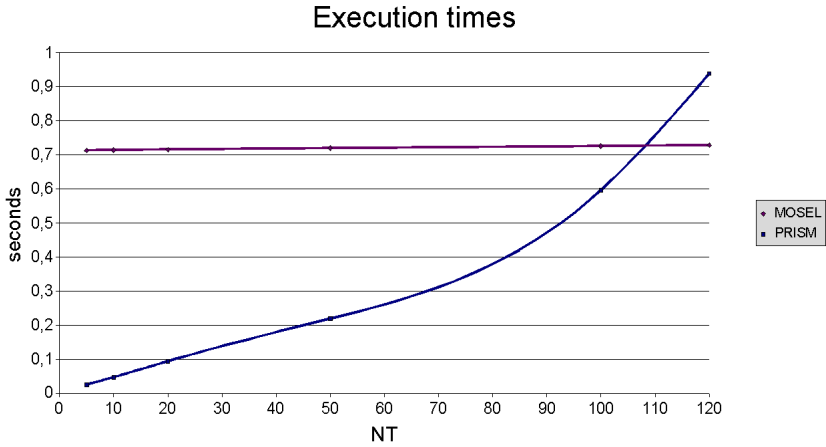


Figure 11: Results of the 2nd experiment

NT	MOSEL	PRISM
5	0.7125	0.025
10	0.7135	0.047
20	0.715	0.094
50	0.719	0.219
100	0.725	0.596
120	0.728	0.938
150	-	1.550
200	-	2.377

Figure 12: Total execution times of the MOSEL and the PRISM in seconds

NT	Model const.	Model checking	Total
5	0.015	0.01	0.025
10	0.031	0.016	0.047
20	0.047	0.047	0.094
50	0.141	0.078	0.219
100	0.391	0.205	0.596
120	0.594	0.344	0.938
150	1.071	0.479	1.550
200	1.609	0.768	2.377

Figure 13: Execution times in seconds

While MOSEL is thus more efficient for smaller models, with PRISM also larger models can be analyzed. Furthermore, once a PRISM model is constructed, it can be arbitrarily often model checked with different parameter values (the PRISM “Experiments” feature). For such scenarios, the model checking time is more relevant than the model construction time.

## 6. Conclusions

Probabilistic model checkers like PRISM are nowadays able to analyze quantitative behaviors of concurrent systems in a similar way that classical performance analysis tools like MOSEL are. In this paper, we reproduced for the particular example of a retrial queuing system the results of an analysis that were previously generated with the help of MOSEL. The numerical results were virtually identical such that we can put confidence on the quality of the analysis. The construction of the models and the benchmarks of the tools demonstrate the following differences between both tools:

- The PRISM modeling language allows us to decompose a system into multiple components whose execution can be synchronized by combined state transitions; this makes the model more manageable than the monolithic MOSEL model. However, the decomposition can be only based on a fixed number of components such that  $NT$  terminals must be still represented by a single PRISM module.
- The state transitions in PRISM are described on a lower level than those in MOSEL: all guard conditions have to be made explicit (while the MOSEL FROM part of a rule imposes implicit conditions on the applicability of the rule) and all effects have to be exposed (while the MOSEL TO part of a rule imposes implicit effects); on the other side, this makes the PRISM rules more transparent than the MOSEL rules. In any case, the difference is syntactic rather than fundamental.
- Several kinds of analysis can be expressed in the property specification language of PRISM (by the definition of “rewards” and CSL queries for the long-term values of rewards) on a higher level than in MOSEL (where explicit calculations have to be written). Like in MOSEL, not every kind of analysis can be directly expressed in PRISM; especially the average execution times can be only computed indirectly from the combination of reward values by external calculations.
- PRISM is also able to answer questions about qualitative system properties such as safety or liveness properties that are beyond the scope of MOSEL.
- The time for an analysis depends in PRISM on the size of the state space of the system while it essentially remains constant in MOSEL (which on the other side puts a rather small limit on the ranges of state variables); the time

growth factor in PRISM is significantly super-linear. While we were thus able to analyze larger systems with PRISM than with MOSEL, it is thus not yet clear whether the analysis will really scale to very large systems.

- As documented by the PRISM web page, the tool is actively used by a large community in various application areas; PRISM is actively supported and further developed (the current release version 3.1.1 is from April 2006, the current development version is from December 2007). On the other hand, the latest version 2.0 of MOSEL-2 is from 2003; the MOSEL web page has not been updated since then.

The use of PRISM for the performance analysis of systems thus seems a promising direction; we plan to further investigate its applicability by analyzing more systems with respect to various kinds of features. While there may be still certain advantages of using dedicated performance evaluation tools like MOSEL, we believe that probabilistic model checking tools are quickly catching up; on the long term, it is very likely that the more general capabilities of these systems and their ever growing popularity will make them also the tools of choice in the performance evaluation community.

## References

- [1] ALMÁSI, B., ROSZIK, J., SZTRIK, J., Homogeneous Finite-Source Retrial Queues with Server Subject to Breakdowns and Repairs, *Mathematical and Computer Modelling*, (2005) 42, 673–682.
- [2] BAIER, C., HAVERKORT, B., HERMANN, H., KATOEN, J., Model Checking Continuous-time Markov chains by transient analysis, In *12th annual Symposium on Computer Aided Verification (CAV 2000)*, volume 1855 of *Lecture Notes in Computer Science*, Springer, (2000) 358–372.
- [3] BARNER, J., BEGAIN, K., BOLCH, G., HEROLD, H., MOSEL — MOdeling, Specification and Evaluation Language, In *2001 Aachen International Multiconference on Measurement, Modelling and Evaluation of Computer and Communication Systems*, Aachen, Germany, September 11–14, 2001.
- [4] BERCZES, T., GUTA, G., KUSPER, G., SCHREINER, W., SZTRIK, J., Comparing the Performance Modeling Environment MOSEL and the Probabilistic Model Checker PRISM for Modeling and Analyzing Retrial Queueing Systems, Technical Report 07-17, Research Institute for Symbolic Computation (RISC), Johannes Kepler University, Linz, Austria, December 2007.
- [5] BERNARDO, M., HILLSTON, J. (editors), *Formal Methods for Performance Evaluation: 7th International School on Formal Methods for the Design of Computer, Communication, and Software Systems, SFM 2007*, volume 4486 of *Lecture Notes in Computer Science*, Bertinoro, Italy, May 28 – June 2, 2007. Springer.
- [6] COOPER, R. B., *Introduction to Queueing Theory*, North Holland, 2nd edition, 1981.

- 
- [7] CLARKE, E. M., GRUMBERG, O., PELED, D. A., *Model checking*, MIT Press, Cambridge, MA, USA, 1999.
- [8] HERZOG, U., Formal Methods for Performance Evaluation, In Ed Brinksma, Holger Hermanns, and Joost-Pieter Katoen, editors, *European Educational Forum: School on Formal Methods and Performance Analysis*, volume 2090 of *Lecture Notes in Computer Science*, pages 1–37, Lectures on Formal Methods and Performance Analysis, First EEF/Euro Summer School on Trends in Computer Science, Berg en Dal, The Netherlands, July 3–7, 2000, Revised Lectures, 2001. Springer.
- [9] HINTON, A., KWIATKOWSKA, M. Z., NORMAN, G., PARKER, D., PRISM: A Tool for Automatic Verification of Probabilistic Systems, In Holger Hermanns and Jens Palsberg, editors, *Tools and Algorithms for the Construction and Analysis of Systems, 12th International Conference, TACAS 2006, Vienna, Austria, March 27–30, 2006*, volume 3920 of *Lecture Notes in Computer Science*, Springer, (2006) 441–444.
- [10] HIREL, C., TUFFIN, B., TRIVEDI, K. S., SPNP: Stochastic Petri Nets. Version 6.0, In Boudewijn R. Haverkort, Henrik C. Bohnenkamp, and Connie U. Smith, editors, *Computer Performance Evaluation: Modelling Techniques and Tools, 11th International Conference, TOOLS 2000, Schaumburg, IL, USA, March 27–31, 2000, Proceedings*, volume 1786 of *Lecture Notes in Computer Science*, Springer, (2000) 354–357.
- [11] JANSEN, D. N., KATOEN, J.-P., OLDENKAMP, M., STOELINGA, M., ZAPREEV, I., How Fast and Fat Is Your Probabilistic Model Checker? An Experimental Performance Comparison, In *Hardware and Software: Verification and Testing*, volume 4899 of *Lecture Notes in Computer Science*, pages 69–85, Proceedings of the Third International Haifa Verification Conference, HVC 2007, Haifa, Israel, October 23–25, 2007, 2008, Springer.
- [12] KWIATKOWSKA, M., Quantitative Verification: Models, Techniques and Tools. In *6th joint meeting of the European Software Engineering Conference and the ACM SIGSOFT Symposium on the Foundations of Software Engineering (ESEC/FSE)*, Cavtat near Dubrovnik, Croatia, September 3–7, 2007, ACM Press.
- [13] NORMAN, G., KWIATKOWSKA, M., PARKER, D., Stochastic Model Checking, In M. Bernardo and J. Hillston, editors, *Formal Methods for Performance Evaluation: 7th International School on Formal Methods for the Design of Computer, Communication, and Software Systems, SFM 2007*, volume 4486 of *Lecture Notes in Computer Science*, pages 220–270, Bertinoro, Italy, May 28 – June 2, 2007, Springer.
- [14] MOSEL — Modeling, Specification, and Evaluation Language, June 2003. <http://www4.informatik.uni-erlangen.de/Projects/MOSEL>.
- [15] MOSEL-2, September 2007. <http://www.net.fmi.uni-passau.de/hp/projects-overview/mosel-2.html>.
- [16] PRISM — Probabilistic Symbolic Model Checker, September 2007. <http://www.prismmodelchecker.org>.
- [17] ROSZIK, J., SZTRIK, J., VIRTAMO, J., Performance Analysis of Finite-Source Retrieval Queues Operating in Random Environments, *International Journal of Operational Research*, (2007) 2, 254–268.
- [18] STEWART, W. J., Performance Modelling and Markov Chains, In *Formal Methods for Performance Evaluation: 7th International School on Formal Methods for the*

*Design of Computer, Communication, and Software Systems, SFM 2007*, volume 4486 of *Lecture Notes in Computer Science*, pages 1–33, Bertinoro, Italy, May 28 – June 2, 2007, Springer.

- [19] SZTRIK, J., KIM, C. S., Performance Modeling Tools with Applications, *Annales Mathematicae et Informaticae*, (2006) 33, 125–140.
- [20] Unified Modeling Language (UML), version 2.1.1, 2007.  
<http://www.omg.org/technology/documents/formal/uml.htm>.
- [21] WOLTER, K. (editor), *Formal Methods and Stochastic Models for Performance Evaluation*, number 4748 in *Lecture Notes in Computer Science*, Fourth European Performance Engineering Workshop, EPEW 2007, Berlin, Germany, September 27–28, 2007.

**Tamás Bérczes, János Sztrik**

Faculty of Informatics, University of Debrecen  
Hungary  
e-mail: {tberczes, jsztrik}@inf.unideb.hu

**Gábor Guta, Wolfgang Schreiner**

Research Institute for Symbolic Computation (RISC)  
Johannes Kepler University  
Linz, Austria  
e-mail: {Gabor.Guta,Wolfgang.Schreiner}@risc.uni-linz.ac.at

**Gábor Kusper**

Eszterházy Károly College, Eger, Hungary  
e-mail: gkusper@aries.ektf.hu



# On the skeleton of a finite transformation semigroup

Attila Egri-Nagy<sup>ab</sup>, Chrystopher L. Nehaniv<sup>b</sup>

<sup>a</sup>Department of Computing Science, Eszterházy Károly College, Hungary

<sup>b</sup>School of Computer Science, University of Hertfordshire, United Kingdom

*Submitted 5 October 2010; Accepted 13 December 2010*

*Dedicated to professor Béla Pelle on his 80<sup>th</sup> birthday*

## Abstract

There are many ways to construct hierarchical decompositions of transformation semigroups. The holonomy algorithm is especially suitable for computational implementations and it is used in our software package. The structure of the holonomy decomposition is determined by the action of the semigroup on certain subsets of the state set. Here we focus on this structure, the skeleton, and investigate some of its properties that are crucial for understanding and for efficient calculations.

*Keywords:* transformation semigroup, Krohn-Rhodes decomposition, holonomy algorithm

*MSC:* 20M20, 20M35, 06A06

## 1. Introduction

The holonomy decomposition [11, 12, 6, 8, 9, 3] is an important proof technique for the Krohn-Rhodes theory [1, Chapter 5], as it works with transformation semigroups, instead of abstract ones, and it is relatively close to the computer scientist's way of thinking. Our computer algebra package, **SgpDec** [5] is now a mature piece of software, so we can study the holonomy decompositions of semigroups with tens of thousands of elements. Here we concentrate on simpler examples and study the underlying structure of the holonomy decomposition, namely the *skeleton* of the transformation semigroup [6, 9]. It is important to note that this notion is different from the skeleton of an abstract semigroup (bordered set of idempotents) and from the topological concept with the same name.

## Mathematical preliminaries

A *transformation semigroup*  $(X, S)$  is a finite nonempty set  $X$  (the state set) together with a set  $S$  of total transformations of  $X$  closed under function composition. A semigroup is a *monoid* if it contains the identity element, the identity map in case of transformations. The action on the points (states)  $x \in X$  naturally extends to set of points:  $P \cdot s = \{p \cdot s \mid p \in P\}$ ,  $P \subseteq X$ ,  $s \in S$ . The set  $\mathcal{O}(X) = \{X \cdot s \mid s \in S\}$  is the *orbit* of the state set. For finite transformations we use two different notations. The traditional matrix notation uses two rows, one for the elements of  $X$  and the second for their corresponding images. We also use the linear (one-line) notation defined in [7] with slight modifications described in [4]. The linear notation is a generalization of the cyclic notation for permutations, therefore the cycle decomposition works as usual. However, for collapsing states we use  $[x_{i_1}, \dots, x_{i_k}; x_i]$  meaning that  $x_{i_j} \mapsto x_i$  for all  $j \in \{1, \dots, k\}$ . These expressions can be nested recursively and blended with the cycle notation. This mirrors the fact that graphically a finite transformation is a bunch of cycles decorated with trees (incoming collapses). Examples are abundant in Section 3. The linear notation is proved to be very useful in software implementations and it is expected to soon have widespread use.

## 2. The skeleton

From now on we consider transformation monoids instead of transformation semigroups. From a categorical viewpoint this is a dangerous step (see [10, p22]), but in a computational setting it is natural. The *augmented orbit* of the state set under the action of the semigroup is  $\mathcal{O}'(X) = \mathcal{O}(X) \cup \{X\} \cup \{\{x\} \mid x \in X\}$ , i.e. we add the state set itself and the singletons. In case of a monoid,  $X$  is already in the orbit.

**Definition 2.1** ([6, 9]). The *skeleton* of a transformation monoid  $(X, M)$  is the augmented orbit equipped with a preorder relation  $(\mathcal{O}'(X), \subseteq_M)$ . This relation is the generalized inclusion defined by

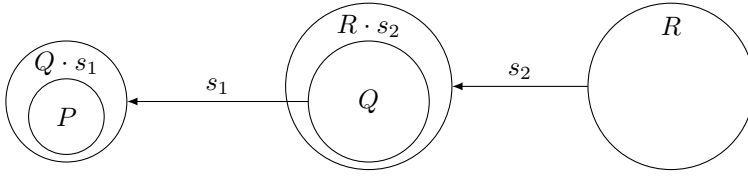
$$P \subseteq_M Q \iff \exists s \in M \text{ such that } P \subseteq Q \cdot s \quad P, Q \in \mathcal{O}'(X), \quad (2.1)$$

i.e. we can transform  $Q$  to include  $P$  under the action of  $M$ .

The skeleton is a feature of the monoid action, it does not depend on the actual generating set, therefore it is justified to talk about *the* skeleton of the transformation monoid.

It is easy to see that  $\subseteq_M$  is a preorder: it is reflexive, since  $P \subseteq P \cdot 1$ , and it is transitive, since if  $P \subseteq Q \cdot s_1$  and  $Q \subseteq R \cdot s_2$  then  $P \subseteq R \cdot s_2 s_1$ , thus  $P \subseteq_M R$ .





We also define an equivalence relation on  $\mathcal{O}'(X)$  by taking the generalized inclusion in both directions:  $P \equiv_M Q \iff P \subseteq_M Q$  and  $Q \subseteq_M P$ . These equivalence classes are the *strong orbits* of the transformation monoid and are denoted by  $O_1, \dots, O_m$ . For each equivalence class there will be a component in the hierarchical decomposition.

### Height and depth of sets

The *height* of a set  $Q \in \mathcal{O}'(X)$  is given by the function  $h : \mathcal{O}'(X) \rightarrow \mathbb{N}$ , which is defined by  $h(Q) = 0$  if  $Q$  is a singleton, and for  $|Q| > 1$ ,  $h(Q)$  is defined by the length of the longest strict generalized inclusion chain(s) in the skeleton starting from a non-singleton set and ending in  $Q$ :

$$h(Q) = \max_i (Q_1 \subset_M \dots \subset_M Q_i = Q),$$

where  $|Q_1| > 1$ . The height of  $(X, M)$  is  $h = h(X)$ .

It is also useful to speak of *depth* values, which are derived from the height values:

$$d(Q) = h(X) - h(Q) + 1.$$

The top level is depth 1.

Calculating the height values establishes the hierarchical levels in the decomposition, i.e. the number of coordinate positions in the holonomy decomposition is  $h(X)$ .

### Covers

Considering the inclusion relation  $(\mathcal{O}'(X), \subseteq)$ , the set of (*lower*) *covers* of a subset  $P \in \mathcal{O}'(X)$  is denoted by  $\mathcal{C}(P)$ . These are the maximal subsets of  $P$ . The component of the holonomy decomposition corresponding to a set  $P$  is derived from those elements of  $M$  that act on  $\mathcal{C}(P)$ , given that  $P$  is a chosen representative of some equivalence class. This action is a restriction of the action of  $M$  on  $\mathcal{O}'(X)$ . Obvious properties of covers are:

$$P = \bigcup_{i=1}^k P_i, \quad P_i \subseteq P_j \implies P_i = P_j$$

where  $P_i \in \mathcal{C}(P)$  and  $k = |\mathcal{C}(P)|$ .

### 3. Skeletons with salient features

#### Nonimage covers

Generalized inclusion by definition allows for the existence of (lower) covers of a set that are not images of the set, i.e.  $P_i \in \mathcal{C}(P)$  but there is no  $s \in M$  such that  $P_i = P \cdot s$ . However, we still have to show that these nonreachable maximal subsets are indeed possible. Let's consider the following generator set:

$a = \begin{pmatrix} 1 & 2 & 3 & 4 & 5 & 6 \\ 1 & 2 & 3 & 1 & 1 & 1 \end{pmatrix} = [4, 5, 6; 1]$  has the image  $\{1, 2, 3\}$ ,

$b = \begin{pmatrix} 1 & 2 & 3 & 4 & 5 & 6 \\ 4 & 4 & 4 & 5 & 4 & 6 \end{pmatrix} = ([1, 2, 3; 4], 5)$  and  $c = \begin{pmatrix} 1 & 2 & 3 & 4 & 5 & 6 \\ 4 & 4 & 4 & 5 & 6 & 4 \end{pmatrix} = ([1, 2, 3; 4], 5, 6)$  produce the image  $\{4, 5, 6\}$  and form a generator set (a transposition and a cycle) for the symmetric group  $S_3$  acting on the image,

$d = \begin{pmatrix} 1 & 2 & 3 & 4 & 5 & 6 \\ 4 & 4 & 4 & 4 & 5 & 5 \end{pmatrix} = [1, 2, 3; 4][6; 5]$  together with these point collapsings  $S_3$  produce the images with cardinality 2,

$e = \begin{pmatrix} 1 & 2 & 3 & 4 & 5 & 6 \\ 4 & 4 & 4 & 1 & 2 & 3 \end{pmatrix} = (1, [[5; 2], [6; 3]; 4])$  maps  $\{4, 5, 6\}$  to  $\{1, 2, 3\}$  (and permutes 1 and 4),

$f = \begin{pmatrix} 1 & 2 & 3 & 4 & 5 & 6 \\ 2 & 3 & 1 & 4 & 4 & 4 \end{pmatrix} = (1, 2, 3)[5, 6; 4]$  is just a cycle on  $\{1, 2, 3\}$ .

The skeleton of the monoid they generate contains a set  $\{1, 2, 3\}$  which has nonimage covers, see Fig. 1.

Unfortunately, the existence of nonimage covers makes a computational implementation slightly more complicated, as we really have to calculate with the generalized inclusion, which is the same as dealing with two relations (inclusion, and 'image of' relation).

#### Width

It is important to know the bound for the number of states in a component of a decomposition. These states are determined by the number of covering sets of the component's underlying set.

**Proposition 3.1.** *Let  $\mathcal{C}(Q)$  be the set of covers of  $Q$  and  $|Q| = m$ , then*

$$|\mathcal{C}(Q)| \leq \binom{m}{\lfloor \frac{m}{2} \rfloor}.$$

**Proof.**  $(2^Q, \subseteq)$  has a maximal antichain (a set of mutually incomparable elements) consisting of all subsets with  $\lfloor \frac{m}{2} \rfloor$  elements. We then apply Dilworth's Theorem [2], which says that the width (the size of a largest antichain) of a partially ordered set is the same as the minimum number of chains whose union is the partially ordered set itself. This theorem implies that the number of chains needed to cover  $(2^Q, \subseteq)$  is equal to  $\binom{m}{\lfloor \frac{m}{2} \rfloor}$ . Since  $\mathcal{O}(X)$  does not necessarily equal  $2^Q$  (it is a subset of it), we need the same number of or less chains to cover the elements of  $\mathcal{O}(X)$  below  $Q$  in the inclusion relation, i.e. the subsets of  $Q$ . The number of chains covering

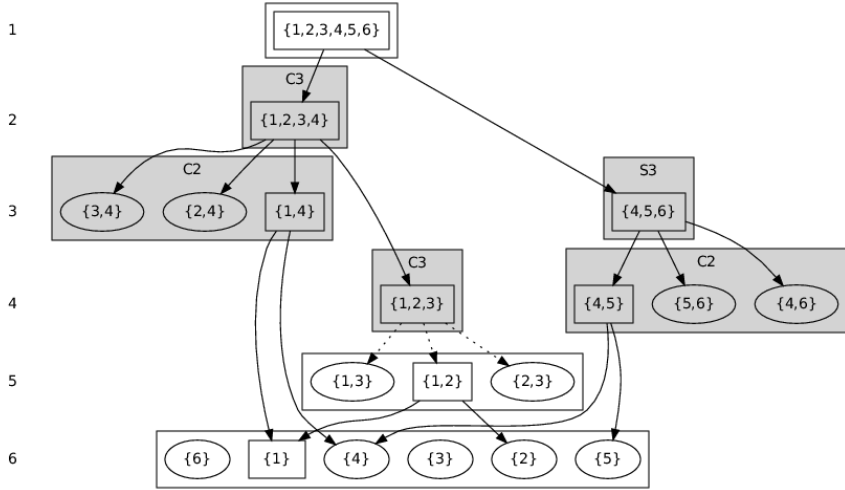


Figure 1: The skeleton of a monoid acting on 6 points (see text for the generators). The nodes are the elements of the augmented orbit. The boxes are the equivalence classes, the rectangular nodes the chosen representatives of a class. The box of the equivalence class is grey if there is a nontrivial subgroup of the monoid acting on the elements of the equivalence class (these groups are isomorphic on equivalent elements). The arrows point to the covers of a set. Dotted arrows indicate nonimage covers. On the side depth values are indicated.

$\mathcal{O}(X)$  below  $Q$  is at least the number of the maximal subsets of  $Q$ , which are the covers of  $Q$  by definition. □

We show that the maximum value can be achieved, so we have a sharp bound. We need the generators of the symmetric group  $S_n$ :

$$(2\ 3\ \dots\ n-1\ 1), (2\ 1\ 3\ \dots\ n)$$

and an arbitrary transformation  $t$  which collapses  $\lceil \frac{n}{2} \rceil$  states, thus its rank is  $\lfloor \frac{n}{2} \rfloor$ . For instance a transformation  $t$  given by:

$$t(i) = \begin{cases} i & t \leq \lceil \frac{n}{2} \rceil \\ \lceil \frac{n}{2} \rceil & \text{otherwise.} \end{cases}$$

For a concrete example see Fig. 2.

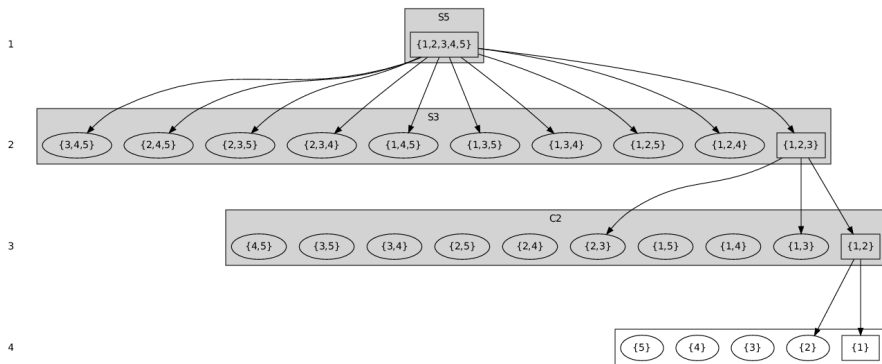


Figure 2: The skeleton of the monoid generated by  $\{(1, 2), (1, 2, 3, 4, 5), [4, 5; 3] = \begin{pmatrix} 1 & 2 & 3 & 4 & 5 \\ 1 & 2 & 3 & 3 & 3 \end{pmatrix}\}$ . The top node 5-element set has 10 covering sets.

### Maximum height skeletons

Previous examples may suggest that height could be bounded by the size of the state set. This is far from being true. For instance the semigroup generated by  $\{(\begin{smallmatrix} 1 & 2 & 3 & 4 & 5 & 6 & 7 \\ 2 & 4 & 1 & 6 & 6 & 7 & 7 \end{smallmatrix}) = [ [ [ [ [3; 1]; 2]; 4]; 5; 6]; 7], (\begin{smallmatrix} 1 & 2 & 3 & 4 & 5 & 6 & 7 \\ 7 & 2 & 3 & 4 & 5 & 6 & 5 \end{smallmatrix}) = [[1; 7]; 5], (\begin{smallmatrix} 1 & 2 & 3 & 4 & 5 & 6 & 7 \\ 1 & 2 & 3 & 3 & 5 & 6 & 7 \end{smallmatrix}) = [4; 3]\}$  gives rise to a skeleton with height 21. It is easy to see why these high skeletons exist: it is possible to have strict generalized inclusion between sets of the same cardinality. For instance  $\{3, 4\} \subset_M \{1, 2\}$  if  $M$  is generated by  $s_1 = (\begin{smallmatrix} 1 & 2 & 3 & 4 \\ 1 & 2 & 1 & 1 \end{smallmatrix}) = [3, 4; 1]$  and  $s_2 = (\begin{smallmatrix} 1 & 2 & 3 & 4 \\ 3 & 4 & 3 & 4 \end{smallmatrix}) = [1; 3][2; 4]$ , where  $s_1$  produces the image  $\{1, 2\}$ ,  $s_2$  takes it to  $\{3, 4\}$ , but there is no transformation for the reverse direction.

We do not know an exact bound for the length of the holonomy decompositions yet, but we can summarize the observations of computational experiments.

**Experimental observation 3.2.** High skeletons tend to have a low number of nontrivial holonomy group components with small cardinality.

It seems that in order to build a high skeleton, we need sufficiently many elements in  $\mathcal{O}(X)$ , and that is provided by the nontrivial group components' permutations. But on the other hand, if we have a group component with high order, then its subgroups might also be components on lower levels, thus collapsing the hierarchy.

It has been shown [6, Chapter XI by Bret Tilson, pp. 287–312] that the length of the longest essential (containing a nontrivial group)  $\mathcal{J}$ -class chain in the semigroup (see cited reference for detailed definitions) is a lower bound for the length of the holonomy decomposition. Then the obvious guess would be that it is the same as the length of the longest  $\mathcal{J}$ -class chains in the semigroups. Again, computational experiments show that this is not the case. The length of the longest  $\mathcal{J}$ -chain can

be smaller, equal to or bigger than the levels of the holonomy decomposition. This is due to the fact that in general we do not act on the semigroup itself but on another set.

## 4. Conclusions and future work

We carried out an initial analysis of hierarchical decompositions of transformation semigroups using the holonomy algorithm. We showed that when working with the components' state sets we have to deal with covers that are not images of the covered set. We also found a sharp upper bound for the width of the decomposition. However, other properties of the holonomy decomposition, including its height, still need further investigation.

## References

- [1] ARBIB, M. A., editor, Algebraic Theory of Machines, Languages, and Semigroups, *Academic Press*, 1968.
- [2] DILWORTH, R. P., A decomposition theorem for partially ordered sets, *Annals of Mathematics*, 51 (1950) 161–166.
- [3] DÖMÖSI, P., NEHANIV, C. L., Algebraic Theory of Finite Automata Networks: An Introduction, volume 11. *SIAM Series on Discrete Mathematics and Applications*, 2005.
- [4] EGRI-NAGY, A., NEHANIV, C. L., On straight words and minimal permutators in finite transformation semigroups. *LNCS Lecture Notes in Computer Science*, 2010. Proceedings of the 15th International Conference on Implementation and Application of Automata CIAA, in press.
- [5] EGRI-NAGY, A., NEHANIV, C. L., **SgpDec** – software package for hierarchical coordinatization of groups and semigroups, implemented in the **GAP** computer algebra system, Version 0.5.25+, 2010. <http://sgpdec.sf.net>.
- [6] EILENBERG, S., Automata, Languages and Machines, volume B, *Academic Press*, 1976.
- [7] GANYUSHKIN, O., MAZORCHUK, V., Classical Transformation Semigroups, *Algebra and Applications*, Springer, 2009.
- [8] GINZBURG, A., Algebraic Theory of Automata, *Academic Press*, 1968.
- [9] HOLCOMBE, W. M. L. Algebraic Automata Theory, *Cambridge University Press*, 1982.
- [10] RHODES, J., STEINBERG, B., The q-theory of Finite Semigroups, *Springer*, 2008.
- [11] ZEIGER, H. P., Cascade synthesis of finite state machines, *Information and Control*, 10 (1967) 419–433, plus erratum.
- [12] ZEIGER, H. P., Yet another proof of the cascade decomposition theorem for finite automata, *Math. Systems Theory*, 1 (1967) 225–228, plus erratum.

**Attila Egri-Nagy**

Eszterházy Károly College  
Institute of Mathematics and Informatics  
Department of Computing Science  
Eger, Leányka út 4, Hungary  
e-mail: [attila@egri-nagy.hu](mailto:attila@egri-nagy.hu)

**Chrystopher L. Nehaniv**

Royal Society Wolfson BioComputation Research Lab  
Centre for Computer Science & Informatics Research, University of Hertfordshire  
Hatfield, Hertfordshire AL10 9AB, United Kingdom  
e-mail: [C.L.Nehaniv@herts.ac.uk](mailto:C.L.Nehaniv@herts.ac.uk)

# On the best estimations for dispersions of special ratio block sequences\*

Ferdinánd Filip<sup>a</sup>, Kálmán Liptai<sup>b</sup>  
Ferenc Mátyás<sup>b</sup>, János T. Tóth<sup>a</sup>

<sup>a</sup>Department of Mathematics, J. Selye University

<sup>b</sup>Institute of Mathematics and Informatics, Eszterházy Károly College

*Submitted 4 October 2010; Accepted 24 November 2010*

*Dedicated to professor Béla Pelle on his 80<sup>th</sup> birthday*

## Abstract

Properties of dispersion of block sequences were investigated by J. T. Tóth, L. Mišík, F. Filip [20]. The present paper is a continuation of the study of relations between the density of the block sequence and so called dispersion of the block sequence.

*Keywords:* dispersion, block sequence,  $(R)$ -density.

*MSC:* Primary 11B05.

## 1. Introduction

In this part we recall some basic definitions. Denote by  $\mathbb{N}$  and  $\mathbb{R}^+$  the set of all positive integers and positive real numbers, respectively. For  $X \subset \mathbb{N}$  let  $X(n) = \#\{x \in X; x \leq n\}$ . In the whole paper we will assume that  $X$  is infinite. Denote by  $R(X) = \{\frac{x}{y}; x \in X, y \in X\}$  the *ratio set of  $X$*  and say that a set  $X$  is  $(R)$ -dense if  $R(X)$  is (topologically) dense in the set  $\mathbb{R}^+$ . Let us notice that the concept of  $(R)$ -density was defined and first studied in papers [17] and [18].

Now let  $X = \{x_1, x_2, \dots\}$  where  $x_n < x_{n+1}$  are positive integers. The sequence

$$\frac{x_1}{x_1}, \frac{x_1}{x_2}, \frac{x_2}{x_2}, \frac{x_1}{x_3}, \frac{x_2}{x_3}, \frac{x_3}{x_3}, \dots, \frac{x_1}{x_n}, \frac{x_2}{x_n}, \dots, \frac{x_n}{x_n}, \dots \quad (1.1)$$

---

\*Supported by grants APVV SK-HU-0009-08 and VEGA Grant no. 1/0753/10.

of finite sequences derived from  $X$  is called *ratio block sequence* of the set  $X$ . Thus the block sequence is formed by blocks  $X_1, X_2, \dots, X_n, \dots$  where

$$X_n = \left( \frac{x_1}{x_n}, \frac{x_2}{x_n}, \dots, \frac{x_n}{x_n} \right); \quad n = 1, 2, \dots$$

This kind of block sequences were studied in papers, [1], [3], [4], [16] and [20]. Also other kinds of block sequences were studied by several authors, see [2], [6], [8], [12] and [19]. Let  $Y = (y_n)$  be an increasing sequence of positive integers. A sequence of blocks of type

$$Y_n = \left( \frac{1}{y_n}, \frac{2}{y_n}, \dots, \frac{y_n}{y_n} \right)$$

was investigated in [11] which extends a result of [5]. Authors obtained a complete theory for the uniform distribution of the related block sequence  $(Y_n)$ .

For every  $n \in \mathbb{N}$  let

$$D(X_n) = \max \left\{ \frac{x_1}{x_n}, \frac{x_2 - x_1}{x_n}, \dots, \frac{x_{i+1} - x_i}{x_n}, \dots, \frac{x_n - x_{n-1}}{x_n} \right\},$$

the maximum distance between two consecutive terms in the  $n$ -th block.

In this paper we will consider the characteristics (see [20])

$$\underline{D}(X) = \liminf_{n \rightarrow \infty} D(X_n),$$

called the *dispersion* of the block sequence (1.1) derived from  $X$ , and its relations to the previously mentioned asymptotic density of the original set  $X$ .

At the end of this section, let us mention the concept of a dispersion of a general sequence of numbers in the interval  $\langle 0, 1 \rangle$ . Let  $(x_n)_{n=0}^{\infty}$  be a sequence in  $\langle 0, 1 \rangle$ . For every  $N \in \mathbb{N}$  let  $x_{i_1} \leq x_{i_2} \leq \dots \leq x_{i_N}$  be reordering of its first  $N$  terms into a nondecreasing sequence and denote

$$d_N = \frac{1}{2} \max \left\{ \max \{ x_{i_{j+1}} - x_{i_j}; j = 1, 2, \dots, N-1 \}, x_{i_1}, 1 - x_{i_N} \right\}$$

the dispersion of the finite sequence  $x_0, x_1, x_2, \dots, x_N$ . Properties of this concept can be found for example in [10] where it is also proved that

$$\limsup_{N \rightarrow \infty} N d_N \geq \frac{1}{\log 4}$$

holds for every one-to-one infinite sequence  $x_n \in \langle 0, 1 \rangle$ . Also notice that the density of the whole sequence  $(x_n)_{n=0}^{\infty}$  is equivalent to  $\lim_{N \rightarrow \infty} d_N = 0$ . Also notice that the analogy of this property for the concept of dispersion of block sequences defined in the present paper does not hold.

Much more on these and related topics can be found in monograph [13].



## 2. Results

When calculating the value  $\underline{D}(X)$ , the following theorems are often useful (See [20], Theorem 1, Corollary 1, respectively).

(A1) Let

$$X = \{x_1, x_2, \dots\} = \bigcup_{n=1}^{\infty} (c_n, d_n) \cap \mathbb{N},$$

where  $x_n < x_{n+1}$  and let  $c_n < d_n < c_{n+1}$ , for  $n \in \mathbb{N}$ , be positive integers. Then

$$\underline{D}(X) = \liminf_{n \rightarrow \infty} \frac{\max\{c_{i+1} - d_i : i = 1, 2, \dots, n\}}{d_{n+1}}.$$

(A2) Let  $X$  be identical to the form of  $X$  in (A1). Suppose that there exists a positive integer  $n_0$  such that for all integers  $n > n_0$

$$c_{n+1} - d_n \leq c_{n+2} - d_{n+1}.$$

Then

$$\underline{D}(X) = \liminf_{n \rightarrow \infty} \frac{c_{n+1} - d_n}{d_{n+1}}.$$

The basic properties of the dispersion  $\underline{D}(X)$  and the relations between dispersion and  $(R)$ -density are investigated in the paper [TMF]. The next theorem states the upper bound for dispersions  $\underline{D}(X)$  of  $(R)$ -dense sets where  $1 \leq a = \lim_{n \rightarrow \infty} \frac{d_n}{c_n} < \infty$  (See [20], Theorem 10).

(A3) Let  $X = \bigcup_{n=1}^{\infty} (c_n, d_n) \cap \mathbb{N}$  be an  $(R)$ -dense set where  $c_n < d_n < c_{n+1}$  for all  $n \in \mathbb{N}$  and suppose that the limit  $\lim_{n \rightarrow \infty} \frac{d_n}{c_n} = a$  exists. Then

$$\underline{D}(X) \leq \min \left\{ \frac{1}{a+1}, \max \left\{ \frac{a-1}{a^2}, \frac{1}{a^2} \right\} \right\},$$

more precisely,

$$\underline{D}(X) \leq \begin{cases} \frac{1}{1+a} & \text{if } a \in \langle 1, \frac{1+\sqrt{5}}{2} \rangle \\ \frac{1}{a^2} & \text{if } a \in \langle \frac{1+\sqrt{5}}{2}, 2 \rangle \\ \frac{a-1}{a^2} & \text{if } a \in \langle 2, \infty \rangle. \end{cases}$$

The following theorem shows that in the third case (if  $a \geq 2$ ), that the dispersion  $\underline{D}(X)$  can be any number in the interval  $\langle 0, \frac{a-1}{a^2} \rangle$ , where  $X = \bigcup_{n=1}^{\infty} (c_n, d_n) \cap \mathbb{N}$  is  $(R)$ -dense and  $\lim_{n \rightarrow \infty} \frac{d_n}{c_n} = a$ . Thus the upper bound for  $\underline{D}(X)$  is the best possible in the case  $a \geq 2$  (See [4], Theorem 2).

(A4) Let  $a \geq 1$  be a real number and  $k$  be an arbitrary natural number. Then for every  $\alpha \in \langle 0, \frac{a^k - 1}{a^{2k}} \rangle$  there exists an  $(R)$ -dense set

$$X = \bigcup_{n=1}^{\infty} (c_n, d_n) \cap \mathbb{N}$$

where  $c_n < d_n < c_{n+1}$  are positive integers for every  $n \in \mathbb{N}$ , such that  $\lim_{n \rightarrow \infty} \frac{d_n}{c_n} = a$  and  $\underline{D}(X) = \alpha$ .

In this paper we prove that in the second case (if  $a \in \langle \frac{1+\sqrt{5}}{2}, 2 \rangle$ ), the dispersion  $\underline{D}(X)$  can be any number in the interval  $\langle 0, \frac{1}{a^2} \rangle$ , where  $X = \bigcup_{n=1}^{\infty} (c_n, d_n) \cap \mathbb{N}$  is  $(R)$ -dense and  $\lim_{n \rightarrow \infty} \frac{d_n}{c_n} = a$ . Thus the upper bound for  $\underline{D}(X)$  is the best possible in the case  $a \in \langle \frac{1+\sqrt{5}}{2}, 2 \rangle$ . The following lemma will be useful.

**Lemma 2.1.** *Let the set*

$$\begin{aligned} M(X) &= \{n \in \mathbb{N} : c_{n+1} - d_n = \max\{c_{i+1} - d_i : i = 1, 2, \dots, n\}\} = \\ &= \{m_1 < m_2 < \dots < m_k < \dots\} \end{aligned}$$

be infinite. Then

$$\underline{D}(X) = \liminf_{k \rightarrow \infty} \frac{c_{m_k+1} - d_{m_k}}{d_{m_k+1}}.$$

**Proof.** Let  $n \in \mathbb{N}$  be an arbitrary integer such that  $n \geq m_1$ . Then there is unique  $k \in \mathbb{N}$  with  $m_k \leq n < m_{k+1}$ . From the definition of the set  $M(X)$  we obtain

$$\frac{\max\{c_{i+1} - d_i : i = 1, 2, \dots, n\}}{d_{n+1}} = \frac{c_{m_k+1} - d_{m_k}}{d_{n+1}} \geq \frac{c_{m_k+1} - d_{m_k}}{d_{m_k+1}}.$$

Then obviously

$$\underline{D}(X) = \liminf_{n \rightarrow \infty} \frac{\max\{c_{i+1} - d_i : i = 1, 2, \dots, n\}}{d_{n+1}} \geq \liminf_{k \rightarrow \infty} \frac{c_{m_k+1} - d_{m_k}}{d_{m_k+1}}.$$

On the other hand, the sequence  $\left(\frac{c_{m_k+1} - d_{m_k}}{d_{m_k+1}}\right)_{k=1}^{\infty}$  is a subsequence of the sequence  $\left(\frac{\max\{c_{i+1} - d_i : i = 1, 2, \dots, n\}}{d_{n+1}}\right)_{n \in \mathbb{N}}$ , hence

$$\underline{D}(X) = \liminf_{n \rightarrow \infty} \frac{\max\{c_{i+1} - d_i : i = 1, 2, \dots, n\}}{d_{n+1}} \leq \liminf_{k \rightarrow \infty} \frac{c_{m_k+1} - d_{m_k}}{d_{m_k+1}}.$$

The last two inequalities imply

$$\underline{D}(X) = \liminf_{k \rightarrow \infty} \frac{c_{m_k+1} - d_{m_k}}{d_{m_k+1}}.$$

□

**Theorem 2.2.** Let  $a \in \langle \frac{1+\sqrt{5}}{2}, 2 \rangle$  be an arbitrary real number. Then for every  $\alpha \in \langle 0, \frac{1}{a^2} \rangle$  there is an  $(R)$ -dense set

$$X = \bigcup_{n=1}^{\infty} (c_n, d_n) \cap \mathbb{N},$$

where  $c_n < d_n < c_{n+1}$  are positive integers for every  $n \in \mathbb{N}$  such that  $\lim_{n \rightarrow \infty} \frac{d_n}{c_n} = a$  and  $\underline{D}(X) = \alpha$ .

**Proof.** Let  $a \in \langle \frac{1+\sqrt{5}}{2}, 2 \rangle$ . According to (A4), it is sufficient to prove Theorem 2.2 for  $\frac{a-1}{a^2} < \alpha \leq \frac{1}{a^2}$ . Define function  $f(b) = \frac{b-1}{ab}$ . Clearly  $f$  is continuous and increasing on the interval  $\langle a, \infty \rangle$ . Moreover

$$f(a) = \frac{a-1}{a^2} \quad \text{and} \quad f(a^2) = \frac{a^2-1}{a^3}.$$

We have  $\frac{a^2-1}{a^3} \geq \frac{1}{a^2}$  if  $a \geq \frac{1+\sqrt{5}}{2}$ . Thus there exists a real number  $a < b \leq a^2$  such that

$$\frac{b-1}{ab} = \alpha.$$

Define a set  $X \subset \mathbb{N}$  by

$$X = \bigcup_{n=1}^{\infty} (A_n \cup B_n) \cap \mathbb{N},$$

where for every  $n \in \mathbb{N}$

$$A_n = (a_{n,1}, b_{n,1}) \cup (a_{n,2}, b_{n,2}) \quad \text{a} \quad B_n = \bigcup_{k=1}^n (c_{n,k}, d_{n,k}).$$

Put  $a_{1,1} = 1$  and for every  $n \in \mathbb{N}$  and  $k = 2, 3, \dots, n$

$$b_{n,1} = [aa_{n,1}] + 1, \quad a_{n,2} = b_{n,1} + 1, \quad b_{n,2} = [aa_{n,2}] + 1,$$

$$c_{n,1} = [bb_{n,2}] + 1, \quad d_{n,1} = [ac_{n,1}] + 1, \quad c_{n,k} = [bd_{n,k-1}] + 1, \quad d_{n,k} = [ac_{n,k}] + 1,$$

and  $a_{n+1,1} = (n+1)d_{n,n}$ .

Obviously for every  $n \in \mathbb{N}$

$$a < \frac{b_{n,1}}{a_{n,1}} \leq a + \frac{1}{a_{n,1}} \quad \text{and} \quad a < \frac{b_{n,2}}{a_{n,1}} \leq a + \frac{1}{a_{n,1}},$$

and for  $k = 1, 2, \dots, n$

$$a < \frac{d_{n,k}}{c_{n,k}} \leq a + \frac{1}{a_{n,1}}.$$

First we prove that  $\underline{D}(X) = \alpha$ . We have the following inequalities:

$$\begin{aligned} c_{n+1,1} - b_{n+1,2} &\geq bb_{n+1,2} - b_{n+1,2} \geq (b-1)b_{n+1,2} \geq (b-1)a^2a_{n+1,1} \geq \\ &\geq (a-1)a^2a_{n+1,1} \geq aa_{n+1,1} > a_{n+1,1} > a_{n+1,1} - d_{n,n} \end{aligned}$$

The inequality  $a^2(a-1) \geq a$  follows from  $a \geq \frac{1+\sqrt{5}}{2}$ . Then

$$c_{n,2} - d_{n,1} \geq bd_{n,1} - d_{n,1} = (b-1)d_{n,1} \geq (b-1)ac_{n,1} \geq (a-1)ac_{n,1} \geq c_{n,1} > c_{n,1} - b_{n,2}$$

and for every  $k = 2, 3, \dots, n-1$

$$\begin{aligned} c_{n,k+1} - d_{n,k} &\geq bd_{n,k} - d_{n,k} = (b-1)d_{n,k} \geq (b-1)ac_{n,k} \geq \\ &\geq (a-1)ac_{n,k} \geq c_{n,k} > c_{n,k} - d_{n,k-1}. \end{aligned}$$

Finally

$$\begin{aligned} a_{n+2,1} - d_{n+1,n+1} &= (n+2)d_{n+1,n+1} - d_{n+1,n+1} > \\ &> d_{n+1,n+1} > c_{n+1,n+1} > c_{n+1,n+1} - d_{n+1,n}. \end{aligned}$$

From the above inequalities we have for a sufficiently large  $n \in \mathbb{N}$  the following inequalities:

$$\begin{aligned} 1 = a_{n,2} - b_{n,1} &< a_{n,1} - d_{n-1,n-1} < c_{n,1} - b_{n,2} < c_{n,2} - d_{n,1} < \dots \\ &\dots < c_{n,n} - d_{n,n-1} < a_{n+1,1} - d_{n,n}. \end{aligned} \quad (2.1)$$

Now we use Lemma 2.1. From (2.1) one can see that it is sufficient to study the quotients:

a)  $\frac{a_{n+1,1} - d_{n,n}}{b_{n+1,2}},$

b)  $\frac{c_{n,1} - b_{n,2}}{d_{n,1}},$

c)  $\frac{c_{n,k} - d_{n,k-1}}{d_{n,k}}$  for  $k = 2, 3, \dots, n$ .

In case a)

$$\liminf_{n \rightarrow \infty} \frac{a_{n+1,1} - d_{n,n}}{b_{n+1,2}} = \liminf_{n \rightarrow \infty} \frac{(n-1)d_{n,n}}{na^2d_{n,n}} = \frac{1}{a^2} \geq \alpha,$$

in case b)

$$\liminf_{n \rightarrow \infty} \frac{c_{n,1} - b_{n,2}}{d_{n,1}} = \liminf_{n \rightarrow \infty} \frac{(b-1)b_{n,2}}{abb_{n,2}} = \frac{b-1}{ab} = \alpha$$

and in case c)

$$\frac{c_{n,k} - d_{n,k-1}}{d_{n,k}} \leq \frac{(b-1)d_{n,k-1} + 1}{abd_{n,k-1}} \leq \frac{b-1}{ab} + \frac{1}{abd_{n,k-1}} \leq \alpha + \frac{1}{abd_{n,1}}$$

and

$$\begin{aligned} \frac{c_{n,k} - d_{n,k-1}}{d_{n,k}} &\geq \frac{(b-1)d_{n,k-1}}{abd_{n,k-1} + b + 1} \geq \\ &\geq \frac{b-1}{ab} - \frac{b-1}{ab} \frac{b+1}{abd_{n,k-1} + b + 1} \geq \alpha - \frac{b^2 - 1}{d_{n,1}}. \end{aligned}$$

From this it is obvious that  $\underline{D}(X) = \alpha$ .

It remains to prove that the set  $X$  is  $(R)$ -dense. We have  $\frac{1}{a^2} \leq \frac{1}{b}$  and  $\frac{1}{b^{l+2}} \leq \frac{1}{b^{l+1}a}$  for every  $l = 1, 2, \dots$ , hence

$$\left(\frac{1}{a^2}, 1\right) \cup \bigcup_{l=1}^{\infty} \left(\frac{1}{b^l a^{l+2}}, \frac{1}{b^l a^{l-1}}\right) = (0, 1)$$

and it is sufficient to prove that the ratio set of the set  $X$  is dense on intervals

$$\left(\frac{1}{a^2}, 1\right) \quad \text{and} \quad \left(\frac{1}{b^l a^{l+2}}, \frac{1}{b^l a^{l-1}}\right)$$

for every  $l = 1, 2, \dots$ .

Now we prove that the ratio set of  $X$  is dense on  $\left(\frac{1}{a^2}, 1\right)$ . Let  $(e, f) \subset \left(\frac{1}{a^2}, 1\right)$ . Put  $\varepsilon = f - e$ . Consider the set

$$\left\{ \begin{aligned} \frac{a_{n,1} + 1}{b_{n,2}} &< \frac{a_{n,1} + 2}{b_{n,2}} < \dots < \frac{b_{n,1}}{b_{n,2}} < \\ &< \frac{a_{n,2} + 1}{b_{n,2}} < \frac{a_{n,2} + 2}{b_{n,2}} < \dots < \frac{b_{n,2} - 1}{b_{n,2}} < \frac{b_{n,2}}{b_{n,2}} = 1 \end{aligned} \right\}, \quad (2.2)$$

which is obviously a subset of the ratio set of  $X$ . The largest difference between consecutive terms of (2.2) is  $\frac{2}{b_{n,2}}$ . Then

$$\frac{a_{n,1} + 1}{b_{n,2}} = \frac{a_{n,1}}{b_{n,2}} + \frac{1}{b_{n,2}} \leq \frac{a_{n,1}}{a^2 a_{n,1}} + \frac{1}{b_{n,2}} = \frac{1}{a^2} + \frac{1}{b_{n,2}}.$$

If we choose  $n \in \mathbb{N}$  so that  $\frac{2}{b_{n,2}} < \varepsilon$ , then the interval  $(e, f)$  is not disjoint with (2.2), hence the ratio set of  $X$  is dense in the interval  $\left(\frac{1}{a^2}, 1\right)$ .

Let  $l \in \mathbb{N}$  be arbitrary. We prove that the ratio set of  $X$  is dense in the interval  $\left(\frac{1}{b^l a^{l+2}}, \frac{1}{b^l a^{l-1}}\right)$ . Let  $(e, f) \subset \left(\frac{1}{b^l a^{l+2}}, \frac{1}{b^l a^{l-1}}\right)$ . Put  $\varepsilon = f - e$ . Choose  $n_1 \in \mathbb{N}$  so that  $n_1 > l$  and  $a_{n,1} + 1 > \frac{2}{\varepsilon}$  for every  $n > n_1$ . Consider the set

$$\left\{ \begin{aligned} \frac{b_{n,2}}{c_{n,l} + 1} &> \frac{b_{n,2} - 1}{c_{n,l} + 1} > \dots > \frac{a_{n,2} + 1}{c_{n,l} + 1} > \frac{b_{n,1}}{c_{n,l} + 1} > \\ &> \frac{b_{n,1} - 1}{c_{n,l} + 1} > \dots > \frac{a_{n,1} + 1}{c_{n,l} + 1} > \frac{a_{n,1} + 1}{c_{n,l} + 2} > \dots > \frac{a_{n,1} + 1}{d_{n,l}} \end{aligned} \right\}, \quad (2.3)$$

which is obviously a subset of the ratio set of  $X$ . The largest difference between consecutive terms of (2.3) is  $\leq \frac{2}{a_{n,1} + 1}$ . On the other hand,

$$\lim_{n \rightarrow \infty} \frac{b_{n,2}}{c_{n,l} + 1} = \frac{1}{b^l a^{l-1}} \quad \text{and} \quad \lim_{n \rightarrow \infty} \frac{a_{n,1} + 1}{d_{n,l}} = \frac{1}{b^l a^{l+2}}.$$

Then there exists  $n_2 \in \mathbb{N}$ , such that for every  $n > n_2$

$$\frac{b_{n,2}}{c_{n,l} + 1} > \frac{1}{b^l a^{l-1}} - \varepsilon \quad \text{and} \quad \frac{a_{n,1} + 1}{d_{n,l}} < \frac{1}{b^l a^{l+2}} + \varepsilon.$$

If we choose  $n > \max\{n_1, n_2\}$ , then the interval  $(e, f)$  is not disjoint with (2.3), hence the ratio set of  $X$  is dense in the interval  $\left(\frac{1}{b^l a^{l+2}}, \frac{1}{b^l a^{l-1}}\right)$ . This concludes the proof.  $\square$

## References

- [1] BUKOR, J., CSIBA, P., On estimations of dispersion of ratio block sequences, *Math. Slovaca*, 59 (2009), 283–290.
- [2] HŁAWKA, E., The theory of uniform distribution, *AB Academic publishers*, London, 1984.
- [3] FILIP, F., MIŠÍK, L., TÓTH, J. T., Dispersion of ratio block sequences and asymptotic density, *Acta Arith.*, 131 (2008), 183–191.
- [4] FILIP, F., TÓTH, J. T., On estimations of dispersions of certain dense block sequences, *Tatra Mt. Math. Publ.*, 31 (2005), 65–74.
- [5] KNAPOWSKI, S., Über ein Problem der Gleichverteilung, *Colloq. Math.*, 5 (1958), 8–10.
- [6] KUIPERS, L., NIEDERREITER, H., Uniform distribution of sequences, *John Wiley & Sons*, New York, 1974.
- [7] MIŠÍK, L., Sets of positive integers with prescribed values of densities, *Math. Slovaca*, 52 (2002), 289–296.
- [8] MYERSON, G., A sampler of recent developments in the distribution of sequences, Number theory with an emphasis on the Markoff spectrum (Provo, UT, 1991) vol.147, *Marcel Dekker*, New York, (1993) 163–190.
- [9] MIŠÍK, L., TÓTH, J. T., Logarithmic density of sequence of integers and density of its ratio set, *Journal de Théorie des Nombres de Bordeaux*, 15 (2003), 309–318.
- [10] NIEDERREITER, H., *On a mesure of denseness for sequences*, in: Topics in Classical Number Theory, Vol. I, II., (G. Halász Ed.), (Budapest 1981), *Colloq. Math. Soc. János Bolyai*, Vol. 34, Nort-Holland, Amsterdam, (1984) 1163–1208.
- [11] PORUBSKÝ, Š., ŠALÁT, T. AND STRAUCH, O., On a class of uniform distributed sequences, *Math. Slovaca*, 40 (1990), 143–170.
- [12] SCHOENBERG, I. J., Über die asymptotische Vertaeilung reeler Zahlen mod 1, *Math. Z.*, 28 (1928), 171–199.
- [13] STRAUCH, O., PORUBSKÝ, Š., Distribution of Sequences: A Sampler, *Peter Lang, Frankfurt am Main*, 2005.
- [14] STRAUCH, O., TÓTH, J. T., Asymptotic density of  $A \subset \mathbb{N}$  and density of the ratio set  $R(A)$ , *Acta Arith.*, 87 (1998), 67–78.
- [15] STRAUCH, O., TÓTH, J. T., Corrigendum to Theorem 5 of the paper “Asymptotic density of  $A \subset \mathbb{N}$  and density of the ratio set  $R(A)$ ”, *Acta Arith.*, 87 (1998), 67–78, *Acta Arith.*, 103.2 (2002), 191–200.
- [16] STRAUCH, O., TÓTH, J. T., Distribution functions of ratio sequences, *Publ. Math. Debrecen*, 58 (2001), 751–778.
- [17] ŠALÁT, T., On ratio sets of sets of natural numbers, *Acta Arith.*, 15 (1969), 273–278.
- [18] ŠALÁT, T., Quotientbasen und (R)-dichte Mengen, *Acta Arith.*, 19 (1971), 63–78.
- [19] TICHY, R. F., Three examples of triangular arrays with optimal discrepancy and linear recurrences, *Applications of Fibonacci numbers*, 7 (1998), 415–423.
- [20] TÓTH, J. T., MIŠÍK, L., FILIP, F., On some properties of dispersion of block sequences of positive integers, *Math. Slovaca*, 54 (2004), 453–464.

**Ferdinánd Filip, János T. Tóth**

Department of Mathematics

J. Selye University

Bratislavská cesta 3322

945 01 Komárno

Slovakia

e-mail: [filip.ferdinand@selyeuni.sk](mailto:filip.ferdinand@selyeuni.sk)

[toth.janos@selyeuni.sk](mailto:toth.janos@selyeuni.sk)

**Kálmán Liptai, Ferenc Mátyás**

Institute of Mathematics and Informatics

Eszterházy Károly College

H-3300 Eger

Leányka út 4.

Hungary

e-mail: [liptaik@ektf.hu](mailto:liptaik@ektf.hu)

[matyas@ektf.hu](mailto:matyas@ektf.hu)





# Some inequalities for $q$ -polygamma function and $\zeta_q$ -Riemann zeta functions

Valmir Krasniqi<sup>a</sup>, Toufik Mansour<sup>b</sup>  
Armend Sh. Shabani<sup>a</sup>

<sup>a</sup>Department of Mathematics, University of Prishtina, Prishtinë 10000, Republic of  
Kosova

<sup>b</sup>Department of Mathematics, University of Haifa, 31905 Haifa, Israel

*Submitted 26 February 2010; Accepted 10 June 2010*

## Abstract

In this paper, we present some inequalities for  $q$ -polygamma functions and  $\zeta_q$ -Riemann Zeta functions, using a  $q$ -analogue of Holder type inequality.

*Keywords:*  $q$ -polygamma functions,  $q$ -zeta function.

*MSC:* 33D05, 11S40, 26D15.

## 1. Introduction and preliminaries

In this section, we provide a summary of notations and definitions used in this paper. For details, one may refer to [3, 5].

For  $n = 1, 2, \dots$  we denote by  $\psi_n(x) = \psi^{(n)}(x)$  the polygamma functions as the  $n$ -th derivative of the psi function  $\psi(x) = \frac{\Gamma'(x)}{\Gamma(x)}$ ,  $x > 0$ , where  $\Gamma(x)$  denotes the usual gamma function.

Throughout this paper we will fix  $q \in (0, 1)$ . Let  $a$  be a complex number. The  $q$ -shifted factorials are defined by

$$(a; q)_n = \prod_{k=0}^{n-1} (1 - aq^k), \quad n = 1, 2, \dots,$$
$$(a; q)_\infty = \lim_{n \rightarrow \infty} (a; q)_n = \prod_{k \geq 0} (1 - aq^k).$$

Jackson [4] defined the  $q$ -gamma function as

$$\Gamma_q(x) = \frac{(q; q)_\infty}{(q^x; q)_\infty} (1 - q)^{1-x}, \quad x \neq 0, -1, \dots \tag{1.1}$$

It satisfies the functional equation

$$\Gamma_q(x + 1) = [x]_q \Gamma_q(x), \quad \Gamma_q(1) = 1, \tag{1.2}$$

where for  $x$  complex  $[x]_q = \frac{1 - q^x}{1 - q}$ .

The  $q$ -gamma function has the following integral representation (see [2])

$$\Gamma_q(x) = \int_0^{\frac{1}{1-q}} t^{x-1} E_q^{-qt} d_q t = \int_0^{\frac{\infty}{1-q}} t^{x-1} E_q^{-qt} d_q t, \quad x > 0.$$

where  $E_q^x = \sum_{j=0}^\infty q^{\frac{j(j-1)}{2}} \frac{x^j}{[j]_q!} = (1 + (1 - q)x)_q^\infty$ , which is the  $q$ -analogue of the classical exponential function.

The  $q$ -analogue of the  $\psi$  function is defined as the logarithmic derivative of the  $q$ -gamma function

$$\psi_q(x) = \frac{\Gamma'_q(x)}{\Gamma_q(x)}, \quad x > 0. \tag{1.3}$$

The  $q$ -Jackson integral from 0 to  $a$  is defined by (see [4, 5])

$$\int_0^a f(x) d_q x = (1 - q)a \sum_{n=0}^\infty f(aq^n) q^n. \tag{1.4}$$

For  $a = \infty$  the  $q$ -Jackson integral is defined by (see [4, 5])

$$\int_0^\infty f(x) d_q x = (1 - q) \sum_{n=-\infty}^\infty f(q^n) q^n \tag{1.5}$$

provided that sums in (1.4) and (1.5) converge absolutely.

In [2] the  $q$ -Riemman zeta function is defined as follows (see Section 2.3 for the definitions)

$$\zeta_q(s) = \sum_{n=1}^\infty \frac{1}{\{n\}_q^s} = \sum_{n=1}^\infty \frac{q^{(n+\alpha([n]_q))s}}{[n]_q^s}. \tag{1.6}$$

In relation to (1.3) and (1.6), K. Brahim [1], using a  $q$ -analogue of the generalized Schwarz inequality, proved the following Theorems.

**Theorem 1.1.** For  $n = 1, 2, \dots$ ,

$$\psi_{q,n}(x) \psi_{q,m}(x) \geq \psi_{q, \frac{m+n}{2}}^2(x),$$

where  $\psi_{q,n} = \psi_q^{(n)}$  is  $n$ -th derivative of  $\psi_q$  and  $\frac{m+n}{2}$  is an integer.

**Theorem 1.2.** For all  $s > 1$ ,

$$[s + 1]_q \frac{\zeta_q(s)}{\zeta_q(s + 1)} \geq q[s]_q \frac{\zeta_q(s + 1)}{\zeta_q(s + 2)}.$$

The aim of this paper is to present some inequalities for  $q$ -polygamma functions and  $q$ -zeta functions by using a  $q$ -analogue of Holder type inequality, similar to those in [1].

## 2. Main results

### 2.1. A lemma

In order to prove our main results, we need the following lemma.

**Lemma 2.1.** Let  $a \in \mathbf{R}_+ \cup \{\infty\}$ , let  $f$  and  $g$  be two nonnegative functions and let  $p, t > 1$  such that  $p^{-1} + t^{-1} = 1$ . The following inequality holds

$$\int_0^a f(x)g(x)d_qx \leq \left( \int_0^a f^p(x)d_qx \right)^{\frac{1}{p}} \left( \int_0^a g^t(x)d_qx \right)^{\frac{1}{t}}.$$

**Proof.** Let  $a > 0$ . By (1.4) we have that

$$\int_0^a f(x)g(x)d_qx = (1 - q)a \sum_{n=0}^{\infty} f(aq^n)g(aq^n)q^n. \quad (2.1)$$

By the use of the Holder's inequality for infinite sums, we obtain

$$\left( \sum_{n=0}^{\infty} f(aq^n)g(aq^n)q^n \right) \leq \left( \sum_{n=0}^{\infty} f^p(aq^n)q^n \right)^{\frac{1}{p}} \cdot \left( \sum_{n=0}^{\infty} g^t(aq^n)q^n \right)^{\frac{1}{t}}. \quad (2.2)$$

Hence

$$\begin{aligned} & (1 - q)a \left( \sum_{n=0}^{\infty} f(aq^n)g(aq^n)q^n \right) \\ & \leq ((1 - q)a)^{\frac{1}{p}} \left( \sum_{n=0}^{\infty} f^p(aq^n)q^n \right)^{\frac{1}{p}} \cdot ((1 - q)a)^{\frac{1}{t}} \left( \sum_{n=0}^{\infty} g^t(aq^n)q^n \right)^{\frac{1}{t}}. \end{aligned} \quad (2.3)$$

The result then follows from (2.1), (2.2) and (2.3).  $\square$

### 2.2. The $q$ -polygamma function

From (1.1) one can derive the following series representation for the function  $\psi_q(x) = \frac{\Gamma'_q(x)}{\Gamma_q(x)}$ :

$$\psi_q(x) = -\log(1 - q) + \log q \sum_{n \geq 1} \frac{q^{nx}}{1 - q^n}, \quad x > 0, \quad (2.4)$$

which implies that

$$\psi_q(x) = -\log(1-q) + \frac{\log q}{1-q} \int_0^q \frac{t^{x-1}}{1-t} d_q t. \quad (2.5)$$

**Theorem 2.2.** For  $n = 2, 4, 6 \dots$  set  $\psi_{q,n}(x) = \psi_q^{(n)}(x)$  the  $n$ -th derivative of the function  $\psi_q$ . Then for  $p, t > 1$  such that  $\frac{1}{p} + \frac{1}{t} = 1$  the following inequality holds

$$\psi_{q,n}\left(\frac{x}{p} + \frac{y}{t}\right) \leq \psi_{q,n}(x)^{\frac{1}{p}} \cdot \psi_{q,n}(y)^{\frac{1}{t}}. \quad (2.6)$$

**Proof.** From (2.5) we deduce that

$$\psi_{q,n}(x) = \frac{\log q}{1-q} \int_0^q \frac{(\log u)^n u^{x-1}}{1-u} d_q u, \quad (2.7)$$

hence

$$\psi_{q,n}\left(\frac{x}{p} + \frac{y}{t}\right) = \frac{\log q}{1-q} \int_0^q \frac{(\log u)^n u^{\frac{x}{p} + \frac{y}{t} - 1}}{1-u} d_q u.$$

By Lemma 2.1 with  $a = q$  we have

$$\begin{aligned} \psi_{q,n}\left(\frac{x}{p} + \frac{y}{t}\right) &= \frac{\log q}{1-q} \int_0^q \left[\frac{(\log u)^n}{1-u}\right]^{\frac{1}{p}} u^{\frac{x-1}{p}} \left[\frac{(\log u)^n}{1-u}\right]^{\frac{1}{t}} u^{\frac{y-1}{t}} d_q u \\ &\leq \left(\frac{\log q}{1-q} \int_0^q \frac{(\log u)^n u^{x-1}}{1-u} d_q u\right)^{\frac{1}{p}} \left(\frac{\log q}{1-q} \int_0^q \frac{(\log u)^n u^{y-1}}{1-u} d_q u\right)^{\frac{1}{t}} \\ &= (\psi_{q,n}(x))^{\frac{1}{p}} (\psi_{q,n}(y))^{\frac{1}{t}} \end{aligned}$$

where  $f(u) = \left(\frac{(\log u)^n}{1-u}\right)^p u^{\frac{x-1}{p}}$  and  $g(u) = \left(\frac{(\log u)^n}{1-u}\right)^t u^{\frac{y-1}{t}}$ . □

For  $p = t = 2$  in (2.6) one has the following result.

**Corollary 2.3.** We have

$$\psi_{q,n}\left(\frac{x+y}{2}\right) \leq \sqrt{\psi_{q,n}(x) \cdot \psi_{q,n}(y)}.$$

### 2.3. $q$ -zeta function

For  $x > 0$  we set  $\alpha(x) = \frac{\log x}{\log q} - E\left(\frac{\log x}{\log q}\right)$  and  $\{x\}_q = \frac{[x]_q}{q^{x+\alpha(\{x\}_q)}}$ , where  $E\left(\frac{\log x}{\log q}\right)$  is the integer part of  $\frac{\log x}{\log q}$ .

In [2] the  $q$ -zeta function is defined as follows

$$\zeta_q(s) = \sum_{n=1}^{\infty} \frac{1}{\{n\}_q^s} = \sum_{n=1}^{\infty} \frac{q^{(n+\alpha([n]_q))s}}{[n]_q^s}.$$

There ([2]) it is proved that  $\zeta_q$  is a  $q$ -analogue of the classical Riemann Zeta function, and for all  $s \in \mathbf{C}$  such that  $\Re(s) > 1$ , and for all  $u > 0$  one has

$$\zeta_q(s) = \frac{1}{\tilde{\Gamma}_q(s)} \int_0^\infty u^{s-1} Z_q(u) d_q u,$$

where  $Z_q(t) = \sum_{n=1}^\infty e_q^{-\{n\}_q t}$ ,  $\tilde{\Gamma}_q(t) = \frac{\Gamma_q(t)}{K_q(t)}$ , and

$$K_q(t) = \frac{(1-q)^{-s}}{1+(1-q)^{-1}} \cdot \frac{(-1-q; q)_\infty (-1-q)^{-1}; q)_\infty}{(-1-q)q^s; q)_\infty (-1-q)^{-1}q^{1-s}; q)_\infty}.$$

**Theorem 2.4.** For  $\frac{1}{p} + \frac{1}{t} = 1$  and  $\frac{x}{p} + \frac{y}{t} > 1$ ,

$$\frac{\tilde{\Gamma}_q\left(\frac{x}{p} + \frac{y}{t}\right)}{\tilde{\Gamma}_q^{\frac{1}{p}}(x) \cdot \tilde{\Gamma}_q^{\frac{1}{t}}(y)} \leq \frac{\zeta_q^{\frac{1}{p}}(x) \cdot \zeta_q^{\frac{1}{t}}(y)}{\zeta_q\left(\frac{x}{p} + \frac{y}{t}\right)}.$$

**Proof.** From Lemma 2.1 we have that

$$\begin{aligned} \int_0^\infty u^{\frac{x}{p} + \frac{y}{t} - 1} Z_q(u) d_q u &= \int_0^\infty u^{\frac{x-1}{p}} \cdot (Z_q(u))^{\frac{1}{p}} u^{\frac{y-1}{t}} \cdot (Z_q(u))^{\frac{1}{t}} d_q u \\ &\leq \left( \int_0^\infty u^{x-1} \cdot (Z_q(u)) d_q u \right)^{\frac{1}{p}} \cdot \left( \int_0^\infty u^{y-1} \cdot (Z_q(u)) d_q u \right)^{\frac{1}{t}}. \end{aligned}$$

For  $f(u) = u^{\frac{x-1}{p}} \cdot (Z_q(u))^{\frac{1}{p}}$  and  $g(u) = u^{\frac{y-1}{t}} \cdot (Z_q(u))^{\frac{1}{t}}$  we obtain that

$$\tilde{\Gamma}_q\left(\frac{x}{p} + \frac{y}{t}\right) \cdot \zeta_q\left(\frac{x}{p} + \frac{y}{t}\right) \leq \tilde{\Gamma}_q^{\frac{1}{p}}(x) \cdot \tilde{\Gamma}_q^{\frac{1}{t}}(y) \cdot \zeta_q^{\frac{1}{p}}(x) \cdot \zeta_q^{\frac{1}{t}}(y),$$

which completes the proof.  $\square$

**Acknowledgements.** The authors would like to thank the anonymous referees for their comments and suggestions.

## References

- [1] BRAHIM, K., Turán-Type Inequalities for some  $q$ -special functions, *J. Ineq. Pure Appl. Math.*, 10(2) (2009) Art. 50.
- [2] FITOUHI, A., BETTAIBI, N., BRAHIM, K., The Mellin transform in quantum calculus, *Constructive Approximation*, 23(3) (2006) 305–323.
- [3] GASPER, G., RAHMAN, M., Basic Hypergeometric Series, 2nd Edition, (2004), Encyclopedia of Mathematics and Applications, 96, Cambridge University Press, Cambridge.
- [4] JACKSON, F.H., On a  $q$ -definite integrals, *Quart. J. Pure and Appl. Math.*, 41 (1910) 193–203.
- [5] KAC, V.G., CHEUNG, P., Quantum Calculus, Universitext, Springer-Verlag, New York, (2002).

**Valmir Krasniqi**

Department of Mathematics  
University of Prishtina  
Prishtinë 10000, Republic of Kosova  
e-mail: [vali.99@hotmail.com](mailto:vali.99@hotmail.com)

**Toufik Mansour**

Department of Mathematics  
University of Haifa  
31905 Haifa, Israel  
e-mail: [toufik@math.haifa.ac.il](mailto:toufik@math.haifa.ac.il)

**Armend Sh. Shabani**

Department of Mathematics  
University of Prishtina  
Prishtinë 10000, Republic of Kosova  
e-mail: [armend\\_shabani@hotmail.com](mailto:armend_shabani@hotmail.com)

# Polynomials with special coefficients\*

Ferenc Mátyás<sup>a</sup>, Kálmán Liptai<sup>a</sup>  
János T. Tóth<sup>b</sup>, Ferdinánd Filip<sup>b</sup>

<sup>a</sup>Institute of Mathematics and Informatics  
Eszterházy Károly College, Eger, Hungary

<sup>b</sup>Department of Mathematics  
Selye János University, Komarno, Slovakia

*Submitted 3 October 2009; Accepted 22 November 2010*

*Dedicated to professor Béla Pelle on his 80<sup>th</sup> birthday*

## Abstract

The aim of this paper is to investigate the zeros of polynomials

$$P_{n,k}(x) = K_{k-1}x^n + K_kx^{n-1} + \cdots + K_{n+k-2}x + K_{n+k-1},$$

where the coefficients  $K_i$ 's are terms of a linear recursive sequence of  $k$ -order ( $k \geq 2$ ).

*Keywords:* linear recurrences, zeros of polynomials with special coefficients

*MSC:* 11C08, 13B25

## 1. Introduction

Let the linear recursive sequence  $K = \{K_n\}_{n=0}^\infty$  of order  $k$  ( $k \geq 2$ ) be defined by the initial values  $K_0 = K_1 = \cdots = K_{k-2} = 0$  and  $K_{k-1} = 1$ , the nonnegative integral weights  $A_1, A_2, \dots, A_k \neq 0$  and the linear recursion

$$K_n = A_1K_{n-1} + A_2K_{n-2} + A_3K_{n-3} + \cdots + A_kK_{n-k} \quad (n \geq k). \quad (1.1)$$

According to the explicit form for  $K_n$  we can write that

$$K_n = p_1(n)\alpha_{1,k}^n + p_2(n)\alpha_{2,k}^n + \cdots + p_t(n)\alpha_{t,k}^n, \quad (1.2)$$

---

\*Research has been supported by the Hungarian-Slovakian Foundation No. SK-8/2008.

where  $\alpha_{1,k}, \alpha_{2,k}, \dots, \alpha_{t,k}$  are the distinct zeros of the characteristic polynomial

$$f_k(x) = x^k - A_1x^{k-1} - A_2x^{k-2} - \dots - A_{k-1}x - A_k \quad (1.3)$$

of the sequence  $K$ , while  $p_i(n)$ 's ( $1 \leq i \leq t \leq k$ ) are polynomials of  $n$  with at most degree  $m_i - 1$ , where  $m_i$  is the multiplicity of  $\alpha_{i,k}$  ( $\sum_{i=1}^t m_i = k$ ).

In the particular case  $k = 2, K_0 = 0, K_1 = 1, A_1 = A_2 = 1$  we can get the Fibonacci-sequence  $F = \{F_n\}_{n=0}^\infty$ , while if  $k = 3, A_1 = A_2 = A_3 = 1$  the sequence  $K$  is known as the Tribonacci-sequence  $T = \{T_n\}_{n=0}^\infty$ .

D. Garth, D. Mills and P. Mitchell [1] introduced the definition of the Fibonacci-coefficient polynomials  $p_n(x) = F_1x^n + F_2x^{n-1} + \dots + F_nx + F_{n+1}$  and – among others – determined the number of the real zeros of  $p_n(x)$ . In [2] we investigated the zeros of the much more general polynomials

$$q_{n,i}(x) = R_i x^n + R_{i+t} x^{n-1} + R_{i+2t} x^{n-2} \dots + R_{i+(n-1)t} x + R_{i+nt},$$

where the sequence  $R = \{R_n\}_{n=0}^\infty$  can be obtained from (1.1) if  $k = 2$  and  $i \geq 1, t \geq 1$  are fixed integers.

The aim of this paper is to investigate the number of the real zeros of the polynomials

$$P_{n,k}(x) = K_{k-1}x^n + K_kx^{n-1} + \dots + K_{n+k-2}x + K_{n+k-1}. \quad (1.4)$$

It is worth mentioning that the problem investigated in this paper can be extended for much more general sequences than  $K$ , which can be the topic of a further paper, as it was suggested by the anonymous referee. The authors would like to express their gratitude to the referee for his/her valuable comments.

## 2. Preliminary and known results

At first we are going to introduce the following notation. Using (1.3) and (1.4) put

$$Q_{n,k}(x) := f_k(x) \cdot P_{n,k}(x). \quad (2.1)$$

**Lemma 2.1.** *The polynomial  $Q_{n,k}(x)$  has the following much more suitable form:*

$$\begin{aligned} Q_{n,k}(x) = & K_{k-1}x^{n+k} - K_{n+k}x^{k-1} - \\ & - (A_k K_{n+1} + A_{k-1} K_{n+2} + \dots + A_2 K_{n+k-1})x^{k-2} - \\ & - \dots - (A_k K_{n+k-2} + A_{k-1} K_{n+k-1})x - A_k K_{n+k-1}. \end{aligned}$$

**Proof.** After the multiplication in (2.1)  $Q_{n,k}(x)$  can be written as

$$\begin{aligned} Q_{n,k}(x) = & K_{k-1}x^{n+k} + (K_k - A_1 K_{k-1})x^{n+k-1} \\ & + (K_{k+1} - A_1 K_k - A_2 K_{k-1})x^{n+k-2} + \\ & \vdots \end{aligned}$$



$$\begin{aligned}
 &+ (K_{2k-2} - A_1K_{2k-3} - A_2K_{2k-4} - \dots - A_{k-1}K_{k-1})x^{n+1} \\
 &+ (K_{2k-1} - A_1K_{2k-2} - A_2K_{2k-3} - \dots - A_{k-1}K_k - A_kK_{k-1})x^n + \\
 &\vdots \\
 &+ (K_{n+k-1} - A_1K_{n+k-2} - A_2K_{n+k-3} - \dots - A_{k-1}K_n - A_kK_{n-1})x^k \\
 &- (A_1K_{n+k-1} + A_2K_{n+k-2} + \dots + A_{k-1}K_{n+1} + A_kK_n)x^{k-1} \\
 &- (A_2K_{n+k-1} + A_3K_{n+k-2} + \dots + A_{k-1}K_{n+2} + A_kK_{n+1})x^{k-2} - \\
 &\vdots \\
 &- (A_{k-1}K_{n+k-1} + A_kK_{n+k-2})x - A_kK_{n+k-1}.
 \end{aligned}$$

But, due to the definition (1.1) the coefficients of the terms  $x^j$  are 0 if  $n + k - 1 \geq j \geq k$ , thus we get that

$$\begin{aligned}
 Q_{n,k}(x) &= K_{k-1}x^{n+k} - K_{n+k}x^{k-1} \\
 &- (A_kK_{n+1} + A_{k-1}K_{n+2} + \dots + A_2K_{n+k-1})x^{k-2} \\
 &- \dots - (A_kK_{n+k-2} + A_{k-1}K_{n+k-1})x - A_kK_{n+k-1},
 \end{aligned}$$

which matches the statement of Lemma 2.1. □

Let us consider the distinct zeros  $\alpha_{1,k}, \alpha_{2,k}, \dots, \alpha_{t,k}$  of the characteristic polynomial  $f_k(x)$  from (1.3). The root  $\alpha_{1,k}$  is said to be the dominant root of  $f_k(x)$  if  $\alpha_{1,k} > |\alpha_{j,k}|$  for every  $2 \leq j \leq t$  and the multiplicity of  $\alpha_{1,k}$  is equal to 1, that is  $m_1 = 1, \alpha_{1,k} \in \mathbf{R}$  and since  $A_k \geq 1$  therefore  $\alpha_{1,k} > 1$ .

**Lemma 2.2.** *Let  $\alpha_{1,k}$  be the dominant root of  $f_k(x)$ . Then*

$$\lim_{n \rightarrow \infty} \frac{K_n}{K_{n-1}} = \alpha_{1,k}.$$

**Proof.** This is a known result, or it can easily be proven if one uses (1.2), where now  $p_1(n)$  is a nonzero real number. □

When the weights  $A_1 = A_2 = \dots = A_k = 1$  in (1.1), that is, when

$$f_k(x) = x^k - x^{k-1} - x^{k-2} - \dots - x - 1, \tag{2.2}$$

then we prove the following result about the real zeros of this  $f_k(x)$ .

**Lemma 2.3.** *If  $f_k(x)$  is of form (2.2), then*

- (i) *the polynomial  $f_k(x)$  has only one positive zero, e.g.  $\alpha_{1,k}$ ,*
- (ii)  *$\alpha_{1,k}$  strictly increasingly tends to 2, if  $k$  tends to infinity,*
- (iii) *if  $k$  is even, then the polynomial  $f_k(x)$  has exactly one negative zero, e.g.  $\alpha_{2,k}$ ,*
- (iv) *if  $k$  is even, then  $\alpha_{2,k}$  strictly decreasingly tends to  $-1$ , if  $k$  tends to infinity,*
- (v) *if  $k$  is odd, then the polynomial  $f_k(x)$  has no negative zero.*

**Proof.** Since  $x = 1$  and  $x = 0$  are not roots of the equation  $x^k - x^{k-1} - x^{k-2} - \dots - x - 1 = 0$ , therefore it can be rewritten into the following equivalent forms:

$$\begin{aligned} x^k &= x^{k-1} + x^{k-2} + \dots + x + 1, \\ x^k &= \frac{x^k - 1}{x - 1}, \\ x^{k+1} &= 2x^k - 1, \\ 2 - x &= x^{-k}. \end{aligned} \tag{2.3}$$

Drawing the graphs of both sides of (2.3) in the same Descartes' coordinate system, one can obtain the desired statements (i)–(v).  $\square$

**Remark 2.4.** In the case of Tribonacci sequence the polynomial  $f_3(x) = x^3 - x^2 - x - 1$  has dominant root, namely  $\alpha_{1,3} = 1,839286755\dots$ , the two other zeros of  $f_3(x)$  are non-real conjugate complex numbers of absolute value  $0.737353\dots$ . While the characteristic polynomial of the Fibonacci sequence is  $f_2(x) = x^2 - x - 1$ , its positive and negative zeros are  $\alpha_{1,2} = \frac{1+\sqrt{5}}{2}$  and  $\alpha_{2,2} = \frac{1-\sqrt{5}}{2}$ , respectively.

It will be suitable to apply the following lemma if we want to give bounds for the absolute value of (real and complex) zeros of the polynomial

$$P_{n,k}(x) = K_{k-1}x^n + K_kx^{n-1} + \dots + K_{n+k-2}x + K_{n+k-1}.$$

**Lemma 2.5.** *If every coefficients of the polynomial  $g(x) = a_0 + a_1x + \dots + a_nx^n$  are positive numbers and the roots of equation  $g(x) = 0$  are denoted by  $z_1, z_2, \dots, z_n$ , then*

$$\gamma \leq |z_i| \leq \delta$$

hold for every  $1 \leq i \leq n$ , where  $\gamma$  is the minimal, while  $\delta$  is the maximal value in the sequence

$$\frac{a_0}{a_1}, \frac{a_1}{a_2}, \dots, \frac{a_{n-1}}{a_n}.$$

**Proof.** This lemma is known as Theorem of S. Kakeya [3].  $\square$

### 3. Results and proofs

At first we deal with the number of the real zeros of the polynomial defined in (1.4), that is

$$P_{n,k}(x) = K_{k-1}x^n + K_kx^{n-1} + \dots + K_{n+k-2}x + K_{n+k-1}.$$

Clearly, positive real zeros of  $P_{n,k}(x)$  do not exist, since – under our conditions – all of the coefficients are positive. Thus we can restrict our investigation on the existence of negative real zeros.

**Theorem 3.1.** *Let  $d$  and  $h$  denote the number of the negative real zeros of the characteristic polynomial  $f_k(x)$  defined in (1.3), and the polynomial  $P_{n,k}(x)$  defined in (1.4), respectively. Then*

- (i)  $k - 1 - 2j = h + d$  for some  $j = 0, 1, 2, \dots, (k - 2)/2$ , if  $k$  and  $n$  are even,
- (ii)  $k - 2j = h + d$  for some  $j = 0, 1, 2, \dots, (k - 2)/2$ , if  $k$  is even and  $n$  is odd,
- (iii)  $k - 1 - 2j = h + d$  for some  $j = 0, 1, 2, \dots, (k - 1)/2$ , if  $k$  is odd and  $n$  is even,
- (iv)  $k - 2j = h + d$  for some  $j = 0, 1, 2, \dots, (k - 1)/2$ , if  $k$  and  $n$  are odd.

**Proof.** We will prove only the case (i), since the other three cases can similarly be proven. Let us consider the polynomial  $Q_{n,k}(x)$  from (2.1). According to Lemma 2.1

$$\begin{aligned} Q_{n,k}(x) &= f_k(x)\dot{P}_{n,k}(x) \\ &= K_{k-1}x^{n+k} - K_{n+k}x^{k-1} \\ &\quad - (A_kK_{n+1} + A_{k-1}K_{n+2} + \dots + A_2K_{n+k-1})x^{k-2} - \dots \\ &\quad - (A_kK_{n+k-2} + A_{k-1}K_{n+k-1})x - A_kK_{n+k-1}. \end{aligned}$$

For using the Descartes' rule of signs we create the the polynomial  $Q_{n,k}(-x)$ , which – with the assumption  $k$  and  $n$  are even – is:

$$\begin{aligned} Q_{n,k}(-x) &= K_{k-1}x^{n+k} + K_{n+k}x^{k-1} \\ &\quad - (A_kK_{n+1} + A_{k-1}K_{n+2} + \dots + A_2K_{n+k-1})x^{k-2} + \dots \\ &\quad + (A_kK_{n+k-2} + A_{k-1}K_{n+k-1})x - A_kK_{n+k-1}. \end{aligned}$$

Since the number of changes of signs in the polynomial  $Q_{n,k}(-x)$  is  $k - 1$  (which is odd), therefore the number of the negative real zeros of the polynomial  $Q_{n,k}(x)$  may be  $1, 3, 5, \dots, k - 1$ . From these negative real zeros  $d$  zeros belong to the polynomial  $f_k(x)$ , while the other  $h$  to the polynomial  $P_{n,k}(x)$ . This proves the statement of Theorem 3.1 (i). □

**Corollary 3.2.** *If the polynomial  $f_k(x)$  is defined as in (2.2), that is when  $A_1 = A_2 = \dots = A_k = 1$ , then – according to Lemma 2.3 –  $d = 1$ , if  $k$  is even, while  $d = 0$ , if  $k$  is odd. This implies that in this case the number of the negative real zeros of the polynomial  $P_{n,k}(x)$  is:*

- (i)  $h = k - 2 - 2j$  for some  $j = 0, 1, 2, \dots, (k - 2)/2$ , if  $k$  and  $n$  are even,
- (ii)  $h = k - 1 - 2j$  for some  $j = 0, 1, 2, \dots, (k - 2)/2$ , if  $k$  is even and  $n$  is odd,
- (iii)  $h = k - 1 - 2j$  for some  $j = 0, 1, 2, \dots, (k - 1)/2$ , if  $k$  is odd and  $n$  is even,
- (iv)  $h = k - 2j$  for some  $j = 0, 1, 2, \dots, (k - 1)/2$ , if  $k$  and  $n$  are odd.

**Corollary 3.3.** *In the case of Tribonacci sequence , for  $f_k(x) = f_3(x) = x^3 - x^2 - x - 1$  we get the following result. The number of the negative real zeros of the polynomial  $P_{n,3}(x)$  is*

- (i) 0 or 2, if  $n$  is even,
- (ii) 1 or 3, if  $n$  is odd.

For the absolute value of zeros of polynomial  $P_{n,k}(x)$  defined in (1.4) we prove the next theorem:

**Theorem 3.4.** *Let  $z$  be any zero of polynomial  $P_{n,k}(x)$  and let  $a$  and  $b$  denote the minimum and the maximum of the set*

$$\left\{ \frac{K_{n+k-1}}{K_{n+k-2}}, \frac{K_{n+k-2}}{K_{n+k-3}}, \frac{K_{n+k-3}}{K_{n+k-4}}, \dots, \frac{K_{k+1}}{K_k}, \frac{K_k}{K_{k-1}} \right\},$$

respectively. Then

$$a \leq |z| \leq b.$$

**Proof.** Applying Lemma 2.5 one can obtain the statement.  $\square$

**Remark 3.5.** According to Lemma 2.2 if  $\alpha_{1,k}$  denotes the dominant root of  $f_k(x)$  then

$$\lim_{n \rightarrow \infty} \frac{K_n}{K_{n-1}} = \alpha_{1,k}.$$

E.g. for the Tribonacci sequence the above quotients of consecutive coefficients tend to 1,83928675 in an alternating way, where  $a = 1$ , and  $b = 2$ .

## References

- [1] GARTH, D., MILLS, D., P. MITCHELL, P., Polynomials Generated by the Fibonacci Sequence, *Journal of Integer Sequences*, Vol. 10 (2007), Article 07.6.8
- [2] MÁTYÁS, F., Further generalization of the Fibonacci-coefficient polynomials, *Annales Mathematicae et Informaticae*, 35 (2008), 123–128.
- [3] ZEMYAN, S.M., On the zeros of the  $n$ th partial sum of the exponential series, *The American Mathematical Monthly*, 112 (2005), No. 10, 891–909.

**Ferenc Mátyás, Kálmán Liptai**

Institute of Mathematics and Informatics

Eszterházy Károly College

P.O. Box 43

H-3301 Eger

Hungary

e-mail: matyas@ektf.hu

liptaik@ektf.hu

**János T. Tóth, Ferdinánd Filip**

Department of Mathematics

Selye János University

P.O. Box 54

94501 Komarno

Slovakia

e-mail: tothj@selyeuni.sk

filipf@selyeuni.sk

# On perfect numbers which are ratios of two Fibonacci numbers\*

Florian Luca<sup>a</sup>, V. Janitzio Mejía Huguet<sup>b</sup>

<sup>a</sup>Instituto de Matemáticas, Universidad Nacional Autónoma de México

<sup>b</sup>Universidad Autónoma Metropolitana

*Submitted 23 August 2010; Accepted 29 October 2010*

## Abstract

Here, we prove that there is no perfect number of the form  $F_{mn}/F_m$ , where  $F_k$  is the  $k$ th Fibonacci number.

*Keywords:* Perfect numbers, Fibonacci numbers.

*MSC:* 11Axx, 11B39, 11Dxx.

## 1. Introduction

For a positive integer  $n$  let  $\sigma(n)$  be the sum of its divisors. A number  $n$  is called perfect if  $\sigma(n) = 2n$  and multiperfect if  $n \mid \sigma(n)$ . Let  $(F_k)_{k \geq 0}$  be the Fibonacci sequence given by  $F_0 = 0$ ,  $F_1 = 1$  and  $F_{k+2} = F_{k+1} + F_k$  for all  $k \geq 0$ .

In [6], it was shown that there is no perfect Fibonacci number. More generally, in [1], it was shown that in fact  $F_n$  is not multiperfect for any  $n \geq 3$ .

In [8], it is was shown that the set  $\{F_{mn}/F_m : m, n \in \mathbf{N}\}$  contains no perfect number. The proof of this result from [8] uses in a fundamental way the claim that if  $N$  is odd and perfect, then

$$N = p^a q_1^{a_1} \cdots q_s^{a_s} \tag{1.1}$$

for some distinct primes  $p$  and  $q_1, \dots, q_s$ , with  $p \equiv a \equiv 1 \pmod{4}$ ,  $a_i$  even for  $i = 1, \dots, s$  and  $q_i \equiv 3 \pmod{4}$  for  $i = 1, \dots, s$ . We could not find neither a reference nor a proof for the fact that the primes  $q_i$  must necessarily be congruent

---

\*F. L. was supported in part by Grants SEP-CONACyT 79685 and PAPIIT 100508, and V. J. M. H. was supported by Grant UAM-A 2232508.

to 3 (mod 4). The remaining assertions about  $p$ ,  $a$  and the exponents  $a_i$  for  $i = 1, \dots, s$  were proved by Euler.

In this paper, we revisit the question of perfect numbers of the shape  $F_{mn}/F_m$  and give a proof of the fact that there are indeed no such perfect numbers. We record our result as follows.

**Theorem 1.1.** *There are no perfect numbers of the form  $F_{mn}/F_m$  for natural numbers  $m$  and  $n$ .*

Our proof avoids the information about the congruence classes of the primes  $q_i$  for  $i = 1, \dots, s$  from (1.1). Ingredients of the proof are Ribenboim's description of square-classes for Fibonacci and Lucas numbers [9], as well as an effective version of Runge's theorem from Diophantine equations due to Gary Walsh [11].

In what follows, for a positive integer  $n$  we use  $\Omega(n)$ ,  $\omega(n)$  and  $\tau(n)$  for the number of prime divisors of  $n$  (counted with and without multiplicities) and the total numbers of divisors of  $n$ , respectively.

From now on, we put  $N := F_{mn}/F_m$  for some positive integers  $m$  and  $n$ , and assume that  $N$  is perfect. Clearly,  $n > 1$ , and by the result from [6] we may assume that  $m > 1$  also. A quick computation with Mathematica confirmed that there is no such example with  $mn \leq 100$ . So, from now on, we also suppose that  $mn > 100$ .

## 2. The even perfect number case

While there is no problem with the treatment of the even perfect number case from [8], we include it here for the convenience of the reader.

For every positive integer  $m$ , let  $z(m)$  be the minimal positive integer  $k$  such that  $m \mid F_k$ . This always exists and it is called the *index of appearance* of  $m$  in the Fibonacci sequence. Indices of appearance have important properties. For example,  $m$  divides  $F_k$  if and only if  $z(m)$  divides  $k$ . Furthermore, if  $p$  is prime, then

$$p \equiv \left(\frac{p}{5}\right) \pmod{z(p)}, \quad (2.1)$$

where for an odd prime  $q$  and an integer  $a$  we write  $\left(\frac{a}{q}\right)$  for the Legendre symbol of  $a$  with respect to  $q$ . In particular, from congruence (2.1), we deduce that  $p \equiv 1 \pmod{z(p)}$  if  $p \equiv \pm 1 \pmod{5}$ , and  $p \equiv -1 \pmod{z(p)}$  provided that  $p \equiv \pm 2 \pmod{5}$ . Clearly,  $z(5) = 5$ .

So, if  $p$  is a prime factor of  $F_n$ , then  $z(p)$  divides  $n$ . If  $z(p) = n$ , then  $p$  is called *primitive* for  $F_n$ . Equivalently,  $p$  is a primitive prime factor of  $F_n$  if  $p$  does not divide  $F_m$  for any positive integer  $m < n$ . An important result of Carmichael [2] asserts that  $F_n$  has a primitive prime factor for all  $n \notin \{1, 2, 6, 12\}$ . From congruence (2.1), we have that if  $p$  is primitive for  $F_n$ , then  $p \equiv \pm 1 \pmod{n}$  unless  $p = n = 5$ .

So, let us now suppose that  $N = F_{mn}/F_m$  is even and perfect. By the structure

theorem of even perfect numbers, we have that

$$\frac{F_{mn}}{F_m} = 2^{p-1}(2^p - 1), \tag{2.2}$$

where  $p$  and  $2^p - 1$  are both primes. If  $p \in \{2, 3\}$ , then  $F_{mn} = 2 \times 3 \times F_m$ , or  $2^2 \times 7 \times F_m$ . However, since  $mn > 100$ , it follows that  $F_{mn}$  has a primitive prime factor  $q$ . The prime  $q$  does not divide  $F_m$  and since  $q \equiv \pm 1 \pmod{mn}$ , it follows that  $q \geq mn - 1 > 99$ . Thus,  $q$  cannot be one of the primes 2, 3, or 7, and we have obtained a contradiction.

Suppose now that  $p \geq 5$ . Then  $16 \mid F_{mn}/F_m$ . Assume first that  $3 \nmid m$ . Since  $z(2) = 3$  and  $3 \nmid m$ , it follows that  $F_m$  is odd, therefore  $16 \mid F_{mn}$ . Hence,  $12 = z(16) \mid mn$ . However, since 9 divides  $F_{12}$ , we get that  $9 \mid F_{12} \mid F_{mn}$ . Relation (2.2) together with the fact that  $p \geq 5$  implies that  $N$  is coprime to 3, therefore  $9 \mid F_m$ . Hence,  $12 = z(9) \mid m$ , contradicting our assumption that  $3 \nmid m$ . Thus,  $3 \mid m$ . In particular,  $2 \mid F_m$ , therefore  $2^5 \mid F_{mn}$ . Write  $mn = 2^s \times 3 \times \lambda$  for some odd positive integer  $\lambda$ . Since  $2^5 \mid F_{mn}$ , we get that  $2^3 \times 3 = z(2^5) \mid mn$ , therefore  $s \geq 3$ . Next we show that  $m \mid 2^{s-3} \times 3 \times \lambda$ . Indeed, for is not, since  $m$  is a multiple of 3, it would follow that  $2^{s-2} \times 3 \mid m$ . It is known that if  $a$  is positive then the exponent of 2 in the factorization of  $F_{2^a \times 3 \times b}$  is exactly  $a + 2$  for all odd integers  $b$ . Hence, the exponent of 2 in  $F_{mn}$  is precisely  $s + 2$ , while since  $2^{s-2} \times 3$  divides  $m$ , we get that the exponent of 2 in  $F_m$  is at least  $s$ . Thus, the exponent of 2 in  $F_{mn}/F_m$  cannot exceed  $(s + 2) - s = 2$ , a contradiction. We conclude that indeed  $m \mid 2^{s-3} \times 3 \times \lambda$ .

Hence,  $mn$  has at least

$$\tau(2^s \times 3 \times \lambda) - \tau(2^{s-3} \times 3 \times \lambda) = (s + 1)\tau(3\lambda) - (s - 2)\tau(3\lambda) = 3\tau(3\lambda) \geq 6$$

divisors  $d$  which do not divide  $m$ . These divisors are of the form  $2^\alpha d_1$ , where  $\alpha \in \{s - 2, s - 1, s\}$ , and  $d_1$  is odd. Since these numbers are all even, it follows that for a most three of them (namely, for  $d \in \{2, 6, 12\}$ ), the number  $F_d$  might not have a primitive prime factor. Thus, for the remaining even divisors  $d$  of  $mn$  which do not divide  $m$  (at least three of them in number), we have that  $F_d$  has a primitive prime factor  $p_d$ . The primes  $p_d$  for such values of  $d$  are distinct and do not divide  $F_m$ , therefore they appear in the factorization of  $N = F_{mn}/F_m$ . Hence,  $\omega(N) \geq 3$ , which contradicts relation (2.2) according to which  $\omega(N) = 2$ .

Hence,  $N$  cannot be even and perfect.

### 3. The odd perfect number case

Here, we use a result of Ribenboim [9] concerning square-classes of Fibonacci and Lucas numbers. We say that positive integers  $a$  and  $b$  are in the same *Fibonacci square-class* if  $F_a F_b$  is a square. The Fibonacci square-class of  $a$  is called trivial if  $F_a F_b$  is a square only for  $b = a$ . Then Ribenboim's result is the following.

**Theorem 3.1.** *If  $a \neq 1, 2, 3, 6, 12$ , then the Fibonacci square-class of  $a$  is trivial.*

In the same paper [9], Ribenboim also found the square-classes of the Lucas numbers. Recall that the Lucas sequence  $(L_k)_{k \geq 0}$  is given by  $L_0 = 2$ ,  $L_1 = 1$  and  $L_{k+2} = L_{k+1} + L_k$  for all  $k \geq 0$ . We say that positive integers  $a$  and  $b$  are in the same *Lucas square-class* if  $L_a L_b$  is a square. As previously, the Lucas square-class of  $a$  is called trivial if  $L_a L_b$  is a square only for  $b = a$ . Then Ribenboim's result is the following.

**Theorem 3.2.** *If  $a \neq 0, 1, 3, 6$ , then the Lucas square-class of  $a$  is trivial.*

We deal with the case of the odd perfect number  $N = F_{mn}/F_m$  through a sequence of lemmas. We write  $N$  as in (1.1) with odd distinct primes  $p$  and  $q_1, \dots, q_s$  and integer exponents  $a$  and  $a_1, \dots, a_s$  such that  $p \equiv a \equiv 1 \pmod{4}$  and  $a_i$  are even for  $i = 1, \dots, s$ . We use  $\square$  to denote a perfect square.

**Lemma 3.3.** *Both  $m$  and  $n$  are odd.*

**Proof.** Assume that  $n$  is even. Then  $F_{mn} = F_{mn/2} L_{mn/2}$  and  $F_m \mid F_{mn/2}$ . Thus,

$$N = \frac{F_{mn}}{F_m} = \left( \frac{F_{mn/2}}{F_m} \right) L_{mn/2} = p \square. \quad (3.1)$$

Now it is well-known that  $\gcd(F_\ell, L_\ell) \in \{1, 2\}$  and since  $N$  is odd, we get that  $\gcd(F_{mn/2}, L_{mn/2}) = 1$ . Hence, the two factors on the left hand side of equation (3.1) above are coprime, and we conclude that either

$$\begin{cases} \frac{F_{mn/2}}{F_m} = p \square \\ L_{mn/2} = \square \end{cases}, \quad \text{or} \quad \begin{cases} \frac{F_{mn/2}}{F_m} = \square \\ L_{mn/2} = p \square \end{cases}.$$

In the first case, since  $L_1 = 1$ , we get that  $mn/2$  is in the same Lucas square-class as 1, which is impossible by Theorem 3.2 because  $mn/2 > 50$ . In the second case, we get that  $mn/2$  and  $m$  are in the same Fibonacci square-class, which is impossible by Theorem 3.1 for  $mn/2 > 50$  unless  $mn/2 = m$ , which happens when  $n = 2$ . But if  $n = 2$ , we then get that

$$N = \frac{F_{2m}}{F_m} = L_m,$$

and the fact that  $L_m$  is not perfect was proved in [6]. The proof of the lemma is complete.  $\square$

**Lemma 3.4.** *We have  $a_i \equiv 0 \pmod{4}$  for all  $i = 1, \dots, s$ .*

**Proof.** It is well-known that if  $\ell$  is odd then every odd prime factor of  $F_\ell$  is congruent to 1 modulo 4. One of the simplest way of seeing this is via the formula  $F_{2\ell+1} = F_\ell^2 + F_{\ell+1}^2$  valid for all  $\ell \geq 0$ , together with the fact that  $F_\ell$  and  $F_{\ell+1}$  are coprime. Since  $mn$  is odd (by Lemma 3.3), it follows that  $q_i \equiv 1 \pmod{4}$  for all  $i = 1, \dots, s$ . Now

$$\sigma(q_i^{a_i}) = 1 + q_i + \dots + q_i^{a_i} \equiv a_i + 1 \pmod{4}.$$



If  $a_i$  is not a multiple of 4 for some  $i \in \{1, \dots, s\}$ , then  $a_i \equiv 2 \pmod{4}$ , therefore  $\sigma(q_i^{a_i}) \equiv 3 \pmod{4}$ . Hence,  $\sigma(q_i^{a_i})$  has a prime factor  $q \equiv 3 \pmod{4}$ . However, since  $q \mid \sigma(q_i^{a_i}) \mid \sigma(N) = 2N$ , it follows that  $q$  is a divisor of  $N$ , which is false because from what we have said above all prime factors of  $N$  are congruent to 1 modulo 4.  $\square$

**Lemma 3.5.** *The number  $n$  is prime.*

**Proof.** Say  $n = r_1^{b_1} \cdots r_\ell^{b_\ell}$ , where  $3 \leq r_1 < \cdots < r_\ell$  are primes and  $b_1, \dots, b_\ell$  are positive integers. Then

$$\frac{F_{mn}}{F_m} = \left(\frac{F_{mn/r_1}}{F_m}\right) \left(\frac{F_{mn}}{F_{mn/r_1}}\right) = p \square. \tag{3.2}$$

It is well-known that the relation

$$\gcd\left(F_a, \frac{F_{ar}}{F_a}\right) = \begin{cases} r & \text{if } r \mid F_a \\ 1 & \text{otherwise} \end{cases} \tag{3.3}$$

holds for all positive integers  $a$  and primes  $r$ . Furthermore, if the above greatest common divisor is not 1, then  $r \parallel F_{ar}/F_a$ . We apply this with  $a := mn/r_1$  and  $r := r_1$  distinguishing two different cases.

The first case is when  $F_{mn/r_1}$  and  $F_{mn}/F_{mn/r_1}$  are coprime. In this case, (3.2) implies that

$$\text{either } \frac{F_{mn/r_1}}{F_m} = \square, \quad \text{or} \quad \frac{F_{mn}}{F_{mn/r_1}} = \square.$$

The second instance is impossible by Theorem 3.1 since  $mn > 100$ . By the same theorem, the first instance is also impossible unless  $mn/r_1 = m$ , which happens when  $n = r_1$ , which is what we want to prove.

So, let us analyze the second case. Then  $r_1 \mid F_{mn/r_1}$ . Since  $r_1 \mid F_{z(r_1)}$ , we get that  $r_1 \mid \gcd(F_{mn/r_1}, F_{z(r_1)}) = F_{\gcd(mn/r_1, z(r_1))}$ . We know that  $r_1 \geq 3$  by Lemma 3.3. If  $r_1 = 3$ , then  $z(r_1) = 4$  and  $r_1 \mid F_{\gcd(mn/3, 4)} = F_1 = 1$ , where the fact that  $\gcd(mn/r_1, 4) = 1$  follows from Lemma 3.3 which tells us that the number  $mn$  is odd. We have reached a contradiction, so it must be the case that  $r_1 \geq 5$ . Let us observe that if  $r_1 \geq 7$ , then  $z(r_1) \mid r_1 \pm 1$ . Hence, in this case

$$r_1 \mid F_{\gcd(mn/r_1, r_1 \pm 1)}.$$

Since  $r_1$  is the smallest prime in  $n$ , it follows that  $n/r_1$  is coprime to  $r_1 \pm 1$ , therefore  $\gcd(mn/r_1, r_1 \pm 1) = \gcd(m, r_1 \pm 1) \mid m$ . Consequently,  $r_1 \mid F_m$  if  $r_1 \geq 7$ . We now return to equation (3.2) and use the fact that  $r_1 \parallel F_{mn}/F_{mn/r_1}$  and  $r_1 = \gcd(F_{mn/r_1}, F_{mn}/F_{mn/r_1})$ .

We distinguish two instances.

The first instance is when  $r_1 = p$ . We then get that

$$\frac{F_{mn/r_1}}{F_m} = \square, \quad \text{and} \quad \frac{F_{mn}}{F_{mn/r_1}} = p \square.$$

By Theorem 3.1, the first equation is not possible unless  $n = r_1$ , which is what we want.

The second instance is when  $r_1 \neq p$ . Then, by Lemma 3.4, we have that  $r_1^4 \mid N$ , and since  $r_1 \parallel F_{mn}/F_{mn/r_1}$ , we get that  $r_1^3 \mid F_{mn/r_1}/F_m$ . If  $r_1 = 5$ , this implies that  $r_1^3 \mid n/r_1$ , because it is well-known that the exponent of 5 in the factorization of  $F_\ell$  is the same as the exponent of 5 in the factorization of  $\ell$ . If  $r_1 \geq 7$ , then  $r_1 \mid F_m$ , so  $z(r_1) \mid m$ . It is then well-known that if  $r_1^e$  denotes the exponent of  $r_1$  in the factorization of  $F_{z(r_1)}$ , then for every nonzero multiple  $\ell$  of  $z(r_1)$ , the exponent of  $r_1$  in  $F_\ell$  is  $f$  ( $\geq e$ ), where  $f - e$  is the precise exponent of  $r_1$  in  $\ell/z(r_1)$ . It then follows again that the divisibility relation  $r_1^3 \mid F_{mn/r_1}/F_m$  together with the fact that  $r_1 \mid F_m$  imply that  $r_1^3 \mid n/r_1$ . Hence, in all cases ( $r_1 = 5$ , or  $r_1 \geq 7$ ), we have that  $r_1^4 \mid n$ . Now we write

$$N = \frac{F_{mn}}{F_m} = \left( \frac{F_{mn/r_1^2}}{F_m} \right) \left( \frac{F_{mn}}{F_{mn/r_1^2}} \right) = p\Box. \tag{3.4}$$

Using (3.3), one proves easily that the greatest common divisor of the two factors on the right above is  $r_1^2$  and that  $r_1^2 \parallel F_{mn}/F_{mn/r_1^2}$ . The above equation (3.4) then leads to

$$\text{either } \frac{F_{mn/r_1^2}}{F_m} = \Box, \quad \text{or } \frac{F_{mn}}{F_{mn/r_1^2}} = \Box.$$

Theorem 3.1 implies that the second instance is impossible and that the first instance is possible only when  $n = r_1^2$ . However, we have already seen that  $r_1^4$  must divide  $n$ . Thus, the first instance cannot appear either. The proof of this lemma is complete.  $\square$

From now on, we shall assume that  $n$  is prime and we shall denote  $n$  by  $q$ .

**Lemma 3.6.** *We have  $q \nmid m$ .*

**Proof.** Say  $q \mid m$ . Then

$$\frac{F_{mq}}{F_m} = \left( \frac{F_m}{F_{m/q}} \right) \left( \frac{F_{mq}/F_m}{F_m/F_{m/q}} \right) = p\Box. \tag{3.5}$$

Both factors above are integers.

Suppose first that the two factors above are coprime. Then

$$\text{either } \frac{F_m}{F_{m/q}} = \Box, \quad \text{or } \frac{F_{mq}/F_m}{F_m/F_{m/q}} = \Box.$$

The first instance is impossible by Theorem 3.1. The second instance leads to  $F_{mq}/F_{m/q} = \Box$ , which is again impossible by the same Theorem 3.1.

Suppose now that the two factors appearing in the right hand side in relation (3.5) are not coprime. But then if  $r$  is a prime such that

$$r \mid \gcd \left( \frac{F_m}{F_{m/q}}, \frac{F_{mq}/F_m}{F_m/F_{m/q}} \right), \quad \text{then} \quad r \mid \gcd \left( F_m, \frac{F_{mq}}{F_m} \right),$$

therefore  $r = q$  by (3.3). Since  $q \mid F_m/F_{m/q}$ , we get that  $q \mid F_{m/q}$  and  $q \parallel F_m/F_{m/q}$ , and also  $q \parallel F_{mq}/F_m = N$ . Thus,  $q = p$ , and now equation (3.5) implies

$$\frac{F_m}{F_{m/q}} = p\Box, \quad \text{and} \quad \frac{F_{mq}/F_m}{F_m/F_{m/q}} = \Box.$$

The second relation leads again to  $F_{mq}/F_{m/q} = \Box$ , which is impossible by Theorem 3.1. Hence, indeed  $q \nmid m$ .  $\square$

**Lemma 3.7.** *We have  $q \geq 7$ .*

**Proof.** We have  $q \geq 3$  by Lemma 3.3. If  $q = 3$ , then since  $3 \nmid m$  (by Lemma 3.6), it follows that  $F_m$  is odd. But then  $N = F_{3m}/F_m$  is even, which is a contradiction. If  $q = 5$ , then  $N = F_{5m}/F_m$  has the property that  $5 \parallel N$ . Thus,  $p = 5$ , and we get the equation

$$\frac{F_{5m}}{F_m} = 5\Box,$$

which has no solution (see equation (8) in [1]). The lemma is proved.  $\square$

**Lemma 3.8.** *(i) All primes  $p$  and  $q_1, \dots, q_s$  have their orders of appearance divisible by  $q$ . In particular, they are all congruent to  $\pm 1 \pmod{q}$ ;*

*(ii)  $p \equiv 1 \pmod{5}$  and  $p \equiv 1 \pmod{q}$ . Furthermore,  $N \equiv 1 \pmod{5}$  and  $N \equiv 1 \pmod{q}$ ;*

*(iii) If  $q_i \equiv 1 \pmod{q}$  for some  $i = 1, \dots, s$ , then  $a_i \geq 2q - 2$ ;*

*(iv) We have  $q \equiv \pm 1 \pmod{20}$ . In particular,  $F_q \equiv 1 \pmod{5}$ ;*

*(v)  $F_q \neq p$ .*

**Proof.** (i) Observe first that all primes  $p$  and  $q_1, \dots, q_s$  are  $\geq 7$ . Indeed, it is clear that they are all odd. If one of them is 3, then  $3 \mid F_{mq}$ , so that  $4 = z(3) \mid mq$ , which is impossible by Lemma 3.3, while if one of them is 5, then  $5 \mid F_{mq}/F_m$ , which implies that  $q = 5$ , contradicting Lemma 3.7. Thus,  $p$  and  $q_i$  are congruent to  $\pm 1 \pmod{z(p)}$  and  $\pm 1 \pmod{z(q_i)}$  for  $i = 1, \dots, s$ , respectively. If  $q \mid z(p)$  and  $q \mid z(q_i)$  for  $i = 1, \dots, s$ , we are through. So, assume that for some prime number  $r$  in  $\{p, q_1, \dots, q_s\}$  we have that  $q \nmid z(r)$ . Then  $r \mid F_{mq}$  and  $r \mid F_{z(r)}$ , so that  $r \mid \gcd(F_{mq}, F_{z(r)}) = F_{\gcd(mq, z(r))} \mid F_m$ . Thus,  $r \mid F_m$  and  $r \mid N = F_{mq}/F_m$ , therefore  $r \mid \gcd(F_m, F_{mq}/F_m)$ , so  $r = q$  by (3.3). In this case,  $q \parallel F_{mq}/F_m$ , therefore  $q = p$ . The above argument shows, up to now, that all prime factors of  $N$  are either congruent to  $\pm 1 \pmod{q}$ , or the prime  $q$  itself, but if this occurs, then  $p = q$ . But with  $p = q$ , we have that  $(q + 1) = (p + 1) \mid \sigma(N) = 2N$ , therefore  $(q + 1)/2$  is a divisor of  $N$ . Thus, all prime factors of  $(q + 1)/2$  are either  $q$ , which is not possible, or primes which are congruent to  $\pm 1 \pmod{q}$ , which is not possible either. This contradiction shows that in fact  $q \nmid N$ , therefore indeed all prime factors of  $N$  have

their orders of appearance divisible by  $q$  and, in particular, they are all congruent to  $\pm 1 \pmod{q}$  by (2.1).

(ii) Clearly,  $(p+1) \mid \sigma(N) = 2N$ . By (i),  $p \equiv \pm 1 \pmod{q}$ , and by relation (2.1), we have that  $p \equiv \left(\frac{p}{5}\right) \pmod{q}$ . If  $p \equiv -1 \pmod{q}$ , then  $q \mid (p+1) \mid 2N$ , so that  $q \mid N$ , which is impossible by (i). So,  $p \equiv 1 \pmod{q}$ , showing that  $\left(\frac{p}{5}\right) \equiv 1 \pmod{5}$ , therefore  $p \equiv \pm 1 \pmod{5}$ . Finally, if  $p \equiv -1 \pmod{5}$ , then  $5 \mid (p+1) \mid \sigma(N) = 2N$ , so  $5 \mid N$ , which is impossible by (i). Thus, indeed  $p \equiv 1 \pmod{5}$  and  $p \equiv 1 \pmod{q}$ . The fact that  $N \equiv 1 \pmod{q}$  is now a consequence of the fact that  $p \equiv 1 \pmod{5}$ ,  $q_i > 5$  and  $a_i$  is a multiple of 4 for all  $i = 1, \dots, s$  (see Lemma 3.4), therefore  $q_i^{a_i} \equiv 1 \pmod{5}$  for all  $i = 1, \dots, s$ . The fact that  $N \equiv 1 \pmod{q}$  follows because by (i)  $p \equiv 1 \pmod{q}$ ,  $q_i \equiv \pm 1 \pmod{q}$ , and  $a_i$  is even for all  $i = 1, \dots, s$ .

(iii) Assume that  $q_i \equiv 1 \pmod{q}$  for some  $i = 1, \dots, s$ . Then

$$\sigma(q_i^{a_i}) = 1 + q_i + \dots + q_i^{a_i} \equiv a_i + 1 \pmod{q}.$$

Since  $\sigma(q_i^{a_i})$  is an odd divisor of  $\sigma(N) = 2N$ , we get that  $\sigma(q_i^{a_i})$  is a divisor of  $N$ , so, by (i), all its prime factors are congruent to  $\pm 1 \pmod{q}$ . Hence,  $\sigma(q_i^{a_i}) \equiv \pm 1 \pmod{q}$ , showing that  $a_i \equiv -2, 0 \pmod{q}$ . Since  $a_i$  is also even, we get that  $a_i \equiv -2, 0 \pmod{2q}$ . In particular,  $a_i \geq 2q - 2$ , which is what we wanted.

(iv) We use the formula

$$F_{qm} = \frac{1}{2^{q-1}} \sum_{i=0}^{(q-1)/2} \binom{q}{2i+1} 5^i F_m^{2i+1} L_m^{q-1-2i}. \quad (3.6)$$

Assume that  $5^b \parallel m$  with some integer  $b \geq 0$ . We then see that all the terms in the sum appearing on the right hand side of formula (3.6) above are multiples of  $5^{b+1}$ , whereas the first term (with  $i = 0$ ) is  $qF_m L_m^{q-1}$ , which is divisible by  $5^b$ , but not by  $5^{b+1}$ . It then follows that

$$\frac{F_{qm}}{F_m} \equiv \frac{q}{2^{q-1}} L_m^{q-1} \pmod{5}. \quad (3.7)$$

Since  $m$  is odd, the sequence  $(L_k)_{k \geq 0}$  is periodic modulo 5 with period 4, and  $L_1 = 1$ ,  $L_3 = 4 \equiv -1 \pmod{5}$ , it follows that  $L_m \equiv \pm 1 \pmod{5}$ , so that  $L_m^{q-1} \equiv 1 \pmod{5}$ . Hence, from congruence (3.7), we get  $N \equiv q/2^{q-1} \pmod{5}$ . Since also  $N \equiv 1 \pmod{5}$  (see (ii)), we get that  $q \equiv 2^{q-1} \pmod{5}$ . In particular,  $q$  is a quadratic residue modulo 5, therefore  $q \equiv \pm 1 \pmod{5}$ . If  $q \equiv 1 \pmod{5}$ , we then get that the congruence  $2^{q-1} \equiv 1 \pmod{5}$  holds, so that  $q \equiv 1 \pmod{4}$  as well. If  $q \equiv -1 \pmod{5}$ , we then get that the congruence  $2^{q-1} \equiv -1 \pmod{5}$  holds, so that  $q \equiv -1 \pmod{4}$  as well. Summarizing, we get that  $q \equiv \pm 1 \pmod{20}$ , and, in particular,  $F_q \equiv 1 \pmod{5}$ .

(v) Assume that  $F_q = p$ . Then  $F_q + 1 = p + 1$  divides  $\sigma(N) = 2N$ . Now let us recall that if  $a > b$  are odd numbers, then

$$F_a + F_b = F_{(a+\delta b)/2} L_{(a-\delta b)/2},$$

where  $\delta \in \{\pm 1\}$  is such that  $a \equiv \delta b \pmod{4}$ . Applying this with  $a := q$  and  $b := 1$ , we get that  $5 \mid F_{(q+\delta)/2} L_{(q-\delta)/2}$  divides  $2F_{qm}$ . Observe that since  $q \equiv \delta \pmod{4}$ , it follows that  $(q-\delta)/2$  is even. Now it is well-known and easy to prove that if  $u$  is even and  $v$  is odd, then  $\gcd(L_u, F_v) = 1$ , or  $2$ . Thus,  $L_{(q-\delta)/2}$  cannot divide  $2F_{mq}$ , unless  $L_{(q-\delta)/2} \leq 4$ , which is not possible for  $q \geq 7$ .  $\square$

From now on, we write  $r$  for the minimal prime factor dividing  $m$ .

**Lemma 3.9.** *There exists a divisor  $d \in \{r, r^2\}$  of  $m$  such that*

$$\frac{F_{mq}/F_{mq/d}}{F_m/F_{m/d}} = \square. \tag{3.8}$$

Furthermore, the case  $d = r^2$  can occur only when  $r \mid F_q$ .

**Proof.** Write again, as often we did before,

$$N = \frac{F_{mq}}{F_m} = \left(\frac{F_{mq/r}}{F_{m/r}}\right) \left(\frac{F_{mq}/F_{mq/r}}{F_m/F_{m/r}}\right) = p\square. \tag{3.9}$$

Suppose first that the two factors appearing in the left hand side of equation (3.9) above are coprime. Then

$$\text{either } \frac{F_{mq/r}}{F_{m/r}} = \square, \quad \text{or } \frac{F_{mq}/F_{mq/r}}{F_m/F_{m/r}} = \square.$$

The first instance is impossible by Theorem 3.1, while the second instance is the conclusion of our lemma with  $d := r$ .

So, from now on let's assume that the two factors appearing in the left hand side of equation (3.9) are not coprime. Let  $\lambda$  be any prime dividing both numbers  $F_{mq/r}/F_{m/r}$  and  $(F_{mq}/F_{mq/r})/(F_m/F_{m/r})$ . Then  $\lambda \mid \gcd(F_{mq/r}, F_{mq}/F_{mq/r})$ . By (3.3), we get that  $\lambda = r$ . In this last case,  $r = \gcd(F_{mq/r}, F_{mq}/F_{mq/r})$ ,  $r \parallel F_{mq}/F_{mq/r}$ , and also  $r \mid F_{mq/r}/F_{m/r}$ . If  $r \mid F_{m/r}$ , it then follows that  $r \mid \gcd(F_{m/r}, F_{mq/r}/F_{m/r})$ , so, by (3.3), we get that  $r = q$ , which contradicts Lemma 3.6. Hence,  $r \nmid F_{m/r}$ . Thus,  $r \mid F_{mq/r}$  and  $r \nmid F_{m/r}$ . Now if  $r \mid F_m$ , then  $r \mid \gcd(F_m, F_{mq/r}) = F_{\gcd(m, mq/r)} = F_{m/r}$ , which is impossible. Thus,  $r \nmid F_m$ , so that  $r \nmid F_m/F_{m/r}$ . Since  $r \parallel F_{mq}/F_{mq/r}$ , we get that  $r \parallel (F_{mq}/F_{mq/r})/(F_m/F_{m/r})$ .

We now distinguish two instances.

The first instance is when  $r = p$ , case in which equation (3.9) leads to

$$\frac{F_{mq/r}}{F_{m/r}} = \square, \quad \text{and} \quad \frac{F_{mq}/F_{mq/r}}{F_m/F_{m/r}} = p\square. \tag{3.10}$$

The first relation in (3.10) above is impossible by Theorem 3.1.

The second instance is when  $r \neq p$ .

Let  $r = q_i$  for some  $i = 1, \dots, s$ , and suppose first that  $r \parallel m$ . Then  $r^{a_i-1} \mid F_{mq/r}$ . Furthermore, since  $r \nmid mq/r$ , we also get that  $r^{a_i-1} \parallel F_{z(r)}$ . Hence,  $r^{a_i-1} \mid$

$\gcd(F_{mq/r}, F_{z(r)}) = F_{\gcd(mq/r, z(r))}$ . Since  $r \mid N$ , we have that  $r \geq 7$  (by (i) of Lemma 6, for example), therefore  $z(r) \mid r \pm 1$ . Since  $r$  is the smallest prime in  $m$  and  $r \parallel m$ , we get that  $\gcd(mq/r, z(r)) \mid \gcd(mq/r, r \pm 1) \mid q$ . Thus, either  $\gcd(mq/r, z(r)) = 1$ , leading to  $r^{a_i-1} \mid F_1$ , which is of course impossible, or  $\gcd(mq/r, z(r)) = q$ , leading to  $r^{a_i-1} \mid F_q$ .

Next, we get from equation (3.9) that

$$\text{either } \frac{F_{mq}/F_{mq/r}}{F_m/F_{m/r}} = r\Box, \quad \text{or } \frac{F_{mq}/F_{mq/r}}{F_m/F_{m/r}} = pr\Box. \tag{3.11}$$

By (v) of Lemma 3.8, we have that  $q \equiv \pm 1 \pmod{20}$ . Hence,  $mq \equiv \pm m \pmod{20}$ , therefore  $F_{mq} \equiv F_{\pm m} \equiv F_m \pmod{5}$ . The last relation, namely  $F_m \equiv F_{-m} \pmod{5}$ , holds because  $m$  is odd. Similarly,  $mq/r \equiv \pm m/r \pmod{20}$ , so that  $F_{mq/r} \equiv F_{m/r} \pmod{5}$ . Since  $F_{m/r}, F_{mq/r}, F_m$  and  $F_{mq}$  are all invertible modulo 5 (because the smallest prime factor of  $m$  which is  $r$  divides  $F_q$ , therefore  $r \geq 2q - 1 > 5$ ), it follows that  $(F_{mq}/F_{mq/r})/(F_m/F_{m/r}) \equiv 1 \pmod{5}$ . Relation (3.11) together with the fact that  $p \equiv 1 \pmod{5}$ , which is (ii) of Lemma 3.8, now shows that  $1 \equiv r\Box \pmod{5}$ , therefore  $\left(\frac{r}{5}\right) = 1$ , so, by (2.1), we have  $r \equiv 1 \pmod{q}$ . Hence, by (iii) of Lemma 3.8, we have that  $a_i \geq 2q - 2$ , therefore  $a_i - 1 \geq 2q - 3$ . Since  $r^{a_i-1} \mid F_q$  and  $r \geq 2q - 1$ , we get the inequality

$$(2q - 1)^{2q-3} \leq F_q,$$

which is false for all primes  $q \geq 7$ .

This contradiction shows that in this case it is not possible that  $r \parallel m$ . Thus,  $r^2 \mid m$ , and then we can write

$$N = \frac{F_{mq}}{F_m} = \left(\frac{F_{mq/r^2}}{F_{m/r^2}}\right) \left(\frac{F_{mq}/F_{mq/r^2}}{F_m/F_{m/r^2}}\right) = p\Box. \tag{3.12}$$

Furthermore, one shows easily that  $r^2 \parallel (F_{mq}/F_{mq/r^2})/(F_m/F_{m/r^2})$  by applying (3.3) twice. Since  $r = q_i$  for some  $i \in \{1, \dots, s\}$  and  $a_i$  is even, it follows that the exponent of  $r$  in the factorization of  $F_{mq/r^2}/F_{m/r^2}$  is also even. We now get from equation (3.12) that

$$\text{either } \frac{F_{mq/r^2}}{F_{m/r^2}} = \Box, \quad \text{or } \frac{F_{mq}/F_{mq/r^2}}{F_m/F_{m/r^2}} = \Box.$$

The first instance is impossible by Theorem 3.1, while the second instance is the conclusion of our lemma for  $d := r^2$ . Notice that along the way we also saw that this case is possible only when  $r \mid F_q$ . The lemma is therefore proved.  $\square$

**Lemma 3.10.** *Let  $q$  and  $d \in \{r, r^2\}$ , where  $q$  and  $r$  are two distinct odd primes. Then the coefficients of the polynomial*

$$f_{q,d}(X) = \frac{(X^{qd} - 1)(X - 1)}{(X^q - 1)(X^d - 1)}$$

are in the set  $\{0, \pm 1\}$ .

**Proof.** When  $d := r$ , the given polynomial is  $\Phi_{qr}(X)$ , where  $\Phi_\ell(X)$  stands for the  $\ell$ th cyclotomic polynomial, and the fact that all its coefficients are in  $\{0, \pm 1\}$  has appeared in many papers (see, for example, [4] and [5]). When  $d := r^2$ , we have  $f_{q,d}(X) = \Phi_{qr}(X)\Phi_{qr^2}(X)$ , and the fact that the coefficients of this polynomial are also in  $\{0, \pm 1\}$  was proved in Proposition 4 in [3].  $\square$

**Lemma 3.11.** *The inequality  $m < 2d^3q^2$  holds.*

**Proof.** We start with the Diophantine equation (3.8). Recall that if we put  $\alpha := (1 + \sqrt{5})/2$  and  $\beta := (1 - \sqrt{5})/2$  for the two roots of the characteristic polynomial  $x^2 - x - 1$  of the Fibonacci and Lucas sequences, then the Binet formulas

$$F_n = \frac{\alpha^n - \beta^n}{\alpha - \beta} \quad \text{and} \quad L_n = \alpha^n + \beta^n \quad \text{hold for all} \quad n \geq 0.$$

Putting  $d \in \{r, r^2\}$ , Lemma 3.9 tells us that

$$\frac{(\alpha^{mq} - \beta^{mq})(\alpha^{m/d} - \beta^{m/d})}{(\alpha^m - \beta^m)(\alpha^{mq/d} - \beta^{mq/d})} = \square. \tag{3.13}$$

We recognize the expression on the left of (3.13) above as  $f_{q,d}^*(\alpha^{m/d}, \beta^{m/d})$ , where for a polynomial  $P(X)$  we write  $P^*(X, Y)$  for its homogenization, and  $f_{q,d}(X)$  is the polynomial appearing in Lemma 3.10. It is clear that  $f_{q,d}^*(X, Y)$  is monic and symmetric since it is the homogenization of either the cyclotomic polynomial  $\Phi_{qr}(X)$ , or of the product  $\Phi_{qr^2}(X)\Phi_{qr}(X)$ , and both these polynomials have the property that they are monic, their last coefficient is 1, and they are reciprocal, meaning that if  $\zeta$  is a root of one of these polynomials, so is  $1/\zeta$ . These conditions lead easily to the conclusion that their homogenizations are symmetric. By the fundamental theorem of symmetric polynomials, we have that  $f_{q,d}^*(X, Y) = F_{q,d}(X+Y, XY)$  is a monic polynomial with integer coefficients in the basic symmetric polynomials  $X+Y$  and  $XY$ . Specializing  $X := \alpha^{m/d}$ ,  $Y := \beta^{m/d}$ , we have that  $X+Y = \alpha^{m/d} + \beta^{m/d} = L_{m/d}$ , and  $XY = (\alpha\beta)^{m/d} = -1$ , where the last equality holds because  $m$  is odd. Hence,  $f_{q,d}^*(\alpha^{m/d}, \beta^{m/d}) = G_{q,d}(L_{m/d})$  is a monic polynomial in  $L_{m/d}$ . Its degree is obviously  $D := (q-1)(d-1)$ , which is even. Hence, equation (3.13) can be written as

$$G_{q,d}(x) = y^2, \tag{3.14}$$

where  $x := L_{m/d}$ ,  $y$  is an integer, and  $G_{q,d}(X)$  is a monic polynomial of even degree  $D$ . The finitely many integer solutions  $(x, y)$  of this equation can be easily bounded using Runge’s method. This has been done in great generality by Gary Walsh [11]. Here is a particular case of Gary Walsh’s theorem.

**Lemma 3.12.** *Let  $F(X) \in \mathbf{Z}[X]$  be a monic polynomial of even degree without double roots. Then all integer solutions  $(x, y)$  of the Diophantine equation*

$$F(x) = y^2$$

satisfy

$$|x| < 2^{2D-2} \left( \frac{D}{2} + 2 \right)^2 (h(F) + 2)^{D+2},$$

where  $h(F)$  denotes the maximum absolute value of the coefficients of the polynomial  $F(X)$ .

From Lemma 3.12, we read that all integer solutions  $(x, y)$  of the Diophantine equation (3.14) satisfy

$$|x| \leq 2^{2D-2} \left( \frac{D}{2} + 2 \right)^2 (h(G_{q,d}) + 2)^{D+2}, \tag{3.15}$$

where  $h(G_{q,d})$  is the maximum absolute value of all the coefficients of  $G_{q,d}(X)$ . Theorem 3.12 requires that the polynomial  $G_{q,d}(X)$  has only simple roots. Let's prove that this is indeed the case.

Let us take a closer look at how we got  $G_{q,d}(X)$  from  $f_{q,d}^*(X, Y)$ . Note that the roots of  $f_{q,d}(X)$  are the roots of unity  $\zeta$  of order  $dq$ , which are neither of order  $d$ , nor of order  $q$ . Let  $\zeta$  and  $\eta$  stand for such roots of unity. Then  $G_{q,d}(X)$  is obtained from  $f_{q,d}(X)$  first by homogenizing, next by replacing  $Y$  by  $-X^{-1}$ , and finally by rewriting the resulting expression as a polynomial in  $X + Y = X - X^{-1}$ . Thus,  $G_{q,d}(X)$  is a polynomial whose roots are  $\zeta - \zeta^{-1}$ . To see that they are all distinct, note that if  $\zeta - \zeta^{-1} = \eta - \eta^{-1}$ , then either  $\zeta = \eta$ , or  $\zeta = -1/\eta$ . However, the second option is not possible when both  $\zeta$  and  $\eta$  are roots of unity of odd orders  $qd$  (to see why, raise the equality  $\zeta = -1/\eta$  to the odd exponent  $dq$  to get the contradiction  $1 = -1$ ). Thus, the numbers  $\zeta - \zeta^{-1}$  remain distinct when  $\zeta$  runs through roots of unity of order  $dq$  which are neither of order  $d$  nor of order  $q$ , showing that  $G_{d,q}(X)$  has only simple roots, and therefore inequality (3.15) applies in our instance.

It remains to bound  $h(G_{q,d})$ . For this, let us start with

$$f_{q,d}^*(X, Y) = \sum_{t=0}^D c_t X^t Y^{D-t},$$

where  $c_t \in \{0, \pm 1\}$  by Lemma 3.10. Since  $f_{q,d}^*(X, Y)$  is symmetric, we have  $c_t = c_{D-t}$  for all  $t = 0, \dots, D$ , therefore

$$f_{q,d}^*(\alpha^{mt/d}, \beta^{mt/d}) = \sum_{\substack{0 \leq t \leq D \\ t \equiv 0 \pmod{2}}} c_t (\alpha^{mt/d} + \beta^{mt/d}) (\alpha\beta)^{(D-t)/2}.$$

Now for even  $t$  we have

$$\alpha^{mt/d} + \beta^{mt/d} = L_{mt/d} = \sum_{i=0}^{t/2} \frac{t}{t-i} \binom{t-i}{i} (-1)^i L_{m/d}^{t-2i}. \tag{3.16}$$

The knowledgeable reader would recognize the expression on the right as the Dickson polynomial  $D_t(Z, -1)$  specialized in  $Z := L_{m/d}$ . Thus,

$$G_{q,d}(L_{m/d}) = f_{q,d}^*(\alpha^{mt/d}, \beta^{mt/d})$$



$$\begin{aligned}
 &= \sum_{\substack{0 \leq t \leq D \\ t \equiv 0 \pmod{2}}} c_t (-1)^{(D-t)/2} \sum_{i=0}^{t/2} \frac{t}{t-i} \binom{t-i}{i} (-1)^i L_{m/d}^{t-2i}, \\
 &= \sum_{\substack{0 \leq u \leq D \\ u \equiv 0 \pmod{2}}} b_u L_{m/r}^u,
 \end{aligned}$$

where

$$b_u := \sum_{\substack{u \leq t \leq D \\ t \equiv 0 \pmod{2}}} c_t (-1)^{(D-t)/2+(t-u)/2} \frac{2t}{t+u} \binom{\frac{t+u}{2}}{\frac{t-u}{2}}. \tag{3.17}$$

Hence,

$$G_{q,d}(X) = \sum_{\substack{0 \leq u \leq D \\ u \equiv 0 \pmod{2}}} b_u X^u,$$

where  $b_u$  is given by (3.17). Since  $|c_t| \leq 1$ ,  $2t/(t+u) \leq 2$  and  $(t+u)/2 \leq D$ , we get that

$$|b_u| \leq 2 \sum_{t=0}^D \binom{D}{t} = 2^{D+1} \quad \text{for all } u = 0, 1, \dots, D,$$

therefore  $h(G_{q,d}) \leq 2^{D+1}$ . Inserting this into (3.15) and using the fact that  $D > q > 4$ , therefore  $D > D/2 + 2$ , we get

$$L_{m/d} \leq 2^{2D-2} \left( \frac{D}{2} + 2 \right)^2 (2^{D+1} + 1)^{D+2} < 2^{2D} D^2 2^{(D+2)^2}. \tag{3.18}$$

Since both sides of the inequality (3.18) are integers, we get that

$$L_{m/d} \leq 2^{(D+2)^2} 2^{2D} D^2 - 1,$$

and since  $L_{m/d} = \alpha^{m/d} + \beta^{m/d} > \alpha^{m/d} - 1$ , we get that

$$\alpha^{m/d} < 2^{(D+2)^2} 2^{2D} D^2,$$

which is equivalent to

$$\frac{m}{d} < \left( \frac{\log 2}{\log \alpha} \right) (D+2)^2 \left( 1 + \frac{2D}{(D+2)^2} + \frac{2 \log D}{(D+2)^2 \log 2} \right).$$

Since  $q \geq 7$  and  $r \geq 3$ , we get that  $D \geq 12$ . The functions  $D \mapsto D/(D+2)^2$  and  $\log D/(D+2)^2$  are decreasing for  $D \geq 12$ , so the expression in parenthesis is

$$\leq 1 + \frac{2 \times 12}{(12+2)^2} + \frac{2 \log 12}{(12+2)^2 \log 2} < 1.2.$$

Since  $\log 2 / \log \alpha < 1.5$ , it follows that

$$\frac{m}{d} < 1.5 \times 1.2(D+2)^2 < 2(D+2)^2.$$

Since  $D = (q-1)(d-1)$ , it follows that  $D+2 = qd - q - d + 3 < qd$ , so that

$$m < 2d(qd)^2 = 2d^3q^2,$$

which is what we wanted to prove.  $\square$

**Lemma 3.13.** *The number  $N$  has at most three distinct prime factors  $< 10^{14}$ .*

**Proof.** Assume that this is not so and that  $N$  has at least four distinct primes  $< 10^{14}$ . One of them might be  $p$ , but the other three, let's call them  $r_i$  for  $i = 1, 2, 3$ , have the property that  $r_i^4 \mid N$  (see Lemma 3.4). A calculation of McIntosh and Roettger [7] showed that the divisibility relation  $r \parallel F_{z(r)}$  holds for all primes  $r < 10^{14}$ . In particular,  $r_i \parallel F_{z(r_i)}$  for  $i = 1, 2, 3$ . Since  $r_i^4 \mid N$  for  $i = 1, 2, 3$ , we get that  $r_i^3 \mid m$  for  $i = 1, 2, 3$ . Hence,

$$r_1^3 r_2^2 r_3^3 \leq m \leq 2d^3q^2 \leq 2r^6q^2.$$

Clearly,  $r_1 \geq r$  and  $r_2 \geq r$ , since  $r$  is the smallest prime factor of  $m$ , therefore  $r_3^3 \leq 2q^2$ . Since  $r_3 \equiv \pm 1 \pmod{q}$  (see Lemma 6 (i)), we get that  $r_3 \geq 2q - 1$ . Thus, we have arrived at the inequality

$$(2q-1)^3 < 2q^2,$$

which is false for any prime  $q \geq 7$ . Thus, the conclusion of the lemma must hold.  $\square$

We are now ready to finally show that there is no such  $N$ . By Lemma 3.13, it can have at most three prime factors  $< 10^{14}$ . Since  $q \geq 7$  and all prime factors of  $N$  are congruent to  $\pm 1 \pmod{q}$ , it follows that the smallest three such primes are at least 13, 17, and 19, respectively. Thus,

$$2 = \frac{\sigma(N)}{N} < \frac{N}{\phi(N)} \leq \left(1 + \frac{1}{12}\right) \left(1 + \frac{1}{16}\right) \left(1 + \frac{1}{18}\right) \prod_{\substack{p \mid N \\ p > 10^{14}}} \left(1 + \frac{1}{p-1}\right),$$

which, after taking logarithms and using the fact that the inequality  $\log(1+x) < x$  holds for all positive real numbers  $x$ , leads to

$$0.494 < \log(1.64) < \sum_{\substack{p \mid N \\ p > 10^{14}}} \log\left(1 + \frac{1}{p-1}\right) < \sum_{\substack{p \mid N \\ p > 10^{14}}} \frac{1}{p-1}. \quad (3.19)$$

Let's call a prime *good* if  $p < z(p)^3$  and *bad* otherwise. We record the following result.

**Lemma 3.14.** *We have*

$$\sum_{\substack{p > 10^{14} \\ p \text{ bad}}} \frac{1}{p-1} < 0.002. \tag{3.20}$$

**Proof.** Observe first that since  $p > 10^{14}$ , it follows that  $z(p) \geq 69$ . For a positive number  $u$  let  $\mathcal{P}_u := \{p : z(p) = u\}$ . Let  $u \geq 69$  be any integer and put  $\ell_u := \#\mathcal{P}_u$ . Then, since  $p \equiv \pm 1 \pmod{u}$  for all  $p \in \mathcal{P}_u$ , we have that

$$(u-1)^{\ell_u} \leq \prod_{p \in \mathcal{P}_u} p \leq F_u < \alpha^{u-1},$$

therefore

$$\ell_u < \frac{(u-1) \log \alpha}{\log(u-1)}.$$

Thus, for a fixed  $u$ , we have

$$\sum_{\substack{p \in \mathcal{P}_u \\ p \text{ bad}}} \frac{1}{p-1} < \frac{\ell_u}{u^3-1} < \frac{\log \alpha}{(u^2+u+1) \log(u-1)} < \frac{\log \alpha}{u^2 \log(u-1)},$$

which leads to

$$\sum_{\substack{p > 10^{14} \\ p \text{ bad}}} \frac{1}{p-1} < \sum_{u \geq 69} \frac{\log \alpha}{u^2 \log(u-1)} < \frac{\log \alpha}{\log 68} \sum_{u \geq 69} \frac{1}{u^2} < \frac{\log \alpha}{68 \log 68} < 0.002.$$

□

Returning to inequality (3.19), we get

$$0.49 < \sum_{\substack{p > 10^{14} \\ p|N \\ p \text{ good}}} \frac{1}{p-1}. \tag{3.21}$$

The following result is Lemma 8 in [1].

**Lemma 3.15.** *The estimate*

$$\sum_{p \in \mathcal{P}_u} \frac{1}{p-1} < \frac{12 + 2 \log \log u}{\phi(u)} \quad \text{holds for all } u \geq 3. \tag{3.22}$$

Let  $\mathcal{U}$  be the set of divisors  $u$  of  $mq$  of the form  $u := z(p)$  for some good prime factor  $p$  of  $N$  with  $p > 10^{14}$ . Observe that all elements of  $\mathcal{U}$  exceed  $10^{14/3} > 46415$ . Inserting the estimate (3.22) of Lemma 3.15 into estimate (3.21), we get

$$0.49 < \sum_{u \in \mathcal{U}} \frac{12 + 2 \log \log u}{\phi(u)}. \tag{3.23}$$

Let  $u_1$  be the smallest element in  $\mathcal{U}$ . We distinguish two cases.

**Case 1.**  $q < r/\sqrt{2}$ .

By Lemma 3.11, we have that  $m < 2r^6q^2 < r^8$ , therefore  $\Omega(m) \leq 7$ , so  $\omega(m) \leq 7$ , and  $\tau(m) \leq 2^7$ . Observe that  $\mathcal{U}$  is contained in the set of divisors of  $qm$  which are not divisors of  $m$ , and this last set has cardinality  $\tau(qm) - \tau(m) = \tau(m) \leq 2^7$ . Here, we used the fact that  $\tau(qm) = 2\tau(m)$ , which holds because  $q \nmid m$  (see Lemma 3.6). Hence,  $\#\mathcal{U} \leq 2^7$ . Furthermore, since  $\omega(m) \leq 7$ , we get that  $\omega(qm) \leq 8$  and

$$\frac{qm}{\phi(qm)} \leq \prod_{i=1}^8 \left(1 + \frac{1}{p_i - 1}\right) < 5.9,$$

where we used the notation  $p_i$  for the  $i$ th prime number. Hence, the inequality

$$\frac{1}{\phi(u)} \leq \frac{6}{u}$$

holds for all divisors  $u$  of  $mq$ . Using also the fact that the functions  $u \mapsto 1/u$  and  $u \mapsto \log \log u/u$  are decreasing for  $u \geq q \geq 7$ , we arrive at the conclusion that inequality (3.23) implies

$$\begin{aligned} 0.49 &< \sum_{u \in \mathcal{U}} \frac{12 + 2 \log \log u}{\phi(u)} < 6 \sum_{u \in \mathcal{U}} \frac{12 + 2 \log \log u}{u} \\ &< 6\#\mathcal{U} \left( \frac{12 + 2 \log \log u_1}{u_1} \right) \leq 6 \times 2^7 \left( \frac{12 + 2 \log \log u_1}{u_1} \right). \end{aligned}$$

Since  $6 \times 2^7 \times 0.49^{-1} < 1600$ , we get that

$$u_1 < 1600(12 + 2 \log \log u_1). \tag{3.24}$$

Inequality (3.24) yields  $u_1 < 27000 < 46415$ , which is a contradiction.

**Case 2.**  $q > r/\sqrt{2}$ .

Note that in this case we necessarily have  $d = r$ , for otherwise we would have  $d = r^2$ , but by Lemma 3.9 this situation occurs only when  $r$  is a prime factor of  $F_q$ . If this were so, we would get that  $r \geq 2q - 1$ , therefore  $q > r/\sqrt{2} > (2q - 1)/\sqrt{2}$ , but this last inequality is not possible for any  $q \geq 7$ . Hence,  $d = r$  and  $m < 2r^4q^2 < 8q^6$ . Since members  $u$  of  $\mathcal{U}$  are the product between  $q$  and some divisor  $v$  of  $m$  (see Lemma 3.8 (i)), we deduce from inequality (3.23) that

$$0.49 < \frac{12 + 2 \log \log(8q^7)}{q - 1} \sum_{v|m} \frac{1}{\phi(v)}. \tag{3.25}$$

It is easy to prove that the inequality

$$\sum_{v|\ell} \frac{1}{\phi(v)} < \frac{\zeta(2)\zeta(3)}{\zeta(6)} \frac{\ell}{\phi(\ell)} \quad \text{holds for all positive integers } \ell. \tag{3.26}$$

Inserting inequality (3.26) for  $\ell := m$  into inequality (3.25), we get that

$$q - 1 < \left( \frac{\zeta(2)\zeta(3)}{\zeta(6) \cdot 0.49} \right) (12 + 2 \log \log(8q^7)) \frac{m}{\phi(m)}. \tag{3.27}$$

The constant in parenthesis in the right hand side of inequality (3.27) above is  $< 4$ . Furthermore, Theorem 15 in [10] says that the inequality

$$\frac{\ell}{\phi(\ell)} < 1.8 \log \log \ell + 2.51/\log \log \ell \quad \text{holds for all } \ell \geq 3. \tag{3.28}$$

The function  $\ell \mapsto 1.8 \log \log \ell + 2.51/\log \log \ell$  is increasing for  $\ell \geq 26$ , and since  $m < 8q^6$ , we get, by inserting inequality (3.28) with  $\ell := m$  into inequality (3.27), that the inequality

$$q - 1 < 4 (12 + 2 \log \log(8q^7)) (1.8 \log \log(8q^6) + 2.51/\log \log(8q^6)), \tag{3.29}$$

holds whenever  $m \geq 26$ . Inequality (3.29) yields  $q \leq 577$ . This was if  $m \geq 26$ . On the other hand, if  $m < 26$ , then  $m/\phi(m) \leq 15/8 < 2$ , so we get

$$q - 1 < 8 (12 + 2 \log \log(8q^7)),$$

which yields  $q \leq 151$ . So, we always have  $q \leq 577$ .

Let us now get the final contradiction. The factorizations of all Fibonacci numbers  $F_\ell$  with  $\ell \leq 1000$  are known. A quick look at this table convinces us that  $F_q$  is square-free for all primes  $q \leq 577$ .

If  $F_q$  is prime, then  $F_q \neq p$  by Lemma 3.8 (v). Furthermore, by Lemma 6 (iv), putting  $q_i = F_q$  for some  $i = 1, \dots, s$ , we get that  $q_i \equiv 1 \pmod{q}$ , therefore  $a_i \geq 2q - 2$ . So  $q_i^{2q-3}$  divides  $m$ , leading to

$$(2q - 1)^{2q-3} \leq q_i^{2q-3} \leq m \leq 8q^6, \tag{3.30}$$

and this last inequality is false for any  $q \geq 7$ .

If  $F_q$  is divisible by at least three primes, it follows that at least two of them, let's call them  $q_i$  and  $q_j$ , are not  $p$ . By Lemma 3.4, we get that  $q_i^3$  and  $q_j^3$  divide  $m$ . Thus,

$$(2q - 1)^6 \leq q_i^3 q_j^3 \leq m \leq 8q^6, \tag{3.31}$$

and again this last inequality is again false for any  $q \geq 7$ .

Finally, if  $F_q$  has precisely two prime factors, then either both of them are distinct from  $p$ , and then we get a contradiction as in (3.31), or  $F_q = pq_i$  for some  $i \in \{1, \dots, s\}$ . But in this case, by Lemma 3.8 (ii) and (iv), we get that  $q_i \equiv 1 \pmod{5}$ , therefore  $q_i \equiv 1 \pmod{q}$ , so  $q_i^{2q-3}$  divides  $m$  by Lemma 3.8 (iii), and we get a contradiction as in (3.30).

This completes the proof of our main result.

## References

- [1] BROUGHAN, K. A., GONZÁLEZ, M., LEWIS, R., LUCA, F., MEJÍA HUGUET, V. J., TOGBÉ, A., There are no multiply perfect Fibonacci numbers, *INTEGERS*, to appear.
- [2] CARMICHAEL, R. D., On the numerical factors of the arithmetic forms  $\alpha^n \pm \beta^n$ , *Ann. Math. (2)*, 15 (1913), 3–70.
- [3] KAPLAN, N., Bounds on the maximal height of divisors of  $x^n - 1$ , *J. Number Theory*, 129 (2009), 2673–2688.
- [4] LAM, T. Y., LEUNG, K. H., On the cyclotomic polynomial  $\Phi_{pq}(X)$ , *Amer. Math. Monthly*, 103 (1996) 562–564.
- [5] LENSTRA, H. W., Vanishing sums of roots of unity, *Proceedings, Bicentennial Congress Wiskundig Genootschap (Vrije Univ., Amsterdam, 1978)*, Part II, (1979) 249–268.
- [6] LUCA, F., Perfect Fibonacci and Lucas numbers, *Rend. Circ. Mat. Palermo (2)*, 49 (2000), 313–318.
- [7] MCINTOSH, R., ROETTGER, E. L., A search for Fibonacci-Wieferich and Wolstenholme primes, *Math. Comp.*, 76 (2007), 2087–2094.
- [8] PHONG, B. M., Perfect numbers concerning the Fibonacci sequence, *Acta Acad. Paed. Agriensis, Sectio Math.*, 26 (1999), 3–8.
- [9] RIBENBOIM, P., Square-classes of Fibonacci and Lucas numbers, *Portugaliae Math.*, 46 (1989), 159–175.
- [10] ROSSER, J. B., SCHOENFELD, L., Approximate formulas for some functions of prime numbers, *Illinois J. Math.*, 6 (1962), 64–94.
- [11] WALSH, P. G., A quantitative version of Runge’s theorem on Diophantine equations, *Acta Arith.*, 62 (1992), 157–172; ‘Correction to: A quantitative version of Runge’s theorem on Diophantine equations’, *Acta Arith.*, 73 (1995), 397–398.

### Florian Luca

C. P. 58089, Morelia Michoacán, México

e-mail: fluca@matmor.unam.mx

### V. Janitzio Mejía Hugué

Av. San Pablo # 180

Col. Reynosa Tamaulipas

Azcapozalco, 02200, México DF, México

e-mail: vjanitzio@gmail.com

# Properties of balancing, cobalancing and generalized balancing numbers\*

Péter Olajos

Department of Applied Mathematics, University of Miskolc

*Submitted 30 June 2010; Accepted 20 November 2010*

*Dedicated to professor Béla Pelle on his 80<sup>th</sup> birthday*

## Abstract

A positive integer  $n$  is called a balancing number if

$$1 + 2 + \cdots + (n - 1) = (n + 1) + (n + 2) + \cdots + (n + r)$$

for some positive integer  $r$ .

Several authors investigated balancing numbers and their various generalizations.

The goal of this paper is to survey some interesting properties and results on balancing, cobalancing and all types of generalized balancing numbers.

*Keywords:* balancing and cobalancing number, recurrence relation, sequence balancing number, power numerical center,  $(a, b)$ -type balancing number

*MSC:* 11D25, 11D41

## 1. Introduction

The sequence  $R = \{R_i\}_{i=0}^{\infty} = R(A, B, R_0, R_1)$  is called a second order linear recurrence if the recurrence relation

$$R_i = AR_{i-1} + BR_{i-2} \quad (i > 1)$$

holds for its terms, where  $A, B \neq 0, R_0$  and  $R_1$  are fixed rational integers and  $|R_0| + |R_1| > 0$ . The polynomial  $f(x) = x^2 - Ax - B$  is called the companion polynomial

---

\*Supported in part by Grant T-48945 and T-48791 from the Hungarian National Foundation for Scientific Research.

of the sequence  $R = R(A, B, R_0, R_1)$ . Let  $D = A^2 + 4B$  be the discriminant of  $f$ . The roots of the companion polynomial will be denoted by  $\alpha$  and  $\beta$ . As it is well-known, if  $D > 0$  then sequence can be written in the form

$$R_i = \frac{a\alpha^i - b\beta^i}{\alpha - \beta}, \quad (i \geq 2),$$

where  $a = R_1 - R_0\beta$  and  $b = R_1 - R_0\alpha$ .

In [3] A. Behera and G. K. Panda gave the notion of balancing number.

**Definition 1.1** ([3]). A positive integer  $n$  is called a balancing number if

$$1 + 2 + \cdots + (n - 1) = (n + 1) + (n + 2) + \cdots + (n + r)$$

for some positive integer  $r$ . This number is called the balancer corresponding to the balancing number  $n$ . The  $m$ th term of the sequence of balancing numbers is denoted by  $B_m$ .

**Remark 1.2.** It can be derived from Definition 1.1 that the following statements are equivalent to each other (see also [3]):

- $n$  is a balancing number,
- $n^2$  is a triangular number (i.e.  $n^2 = 1 + 2 + \cdots + k$  for some  $k \in \mathbb{N}$ ),
- $8n^2 + 1$  is a perfect square.

It is easy to see that 6, 35, and 204 are balancing numbers with balancers 2, 14 and 84, respectively.

## 2. Properties of balancing numbers

### 2.1. Generating balancing numbers

In [3] A. Behera and G. K. Panda proved other interesting properties about balancing numbers.

Let us consider the following functions:

$$F(x) = 2x\sqrt{8x^2 + 1} \tag{2.1}$$

$$G(x) = 3x + \sqrt{8x^2 + 1} \tag{2.2}$$

$$H(x) = 17x + 6\sqrt{8x^2 + 1} \tag{2.3}$$

They proved that these functions always generate balancing numbers.

**Theorem 2.1** (Theorem 2.1 in [3]). *For any balancing number  $n$ ,  $F(n)$ ,  $G(n)$ , and  $H(n)$  are also balancing numbers.*



**Remark 2.2.** Using the theorem above we get that if  $n$  is a balancing number, then  $G(F(n)) = 6n\sqrt{8n^2 + 1} + 16n^2 + 1$  is an odd balancing number, because  $F(n)$  is always even and  $G(n)$  is odd when  $n$  is even.

For generating balancing numbers they proved the following theorems:

**Theorem 2.3** (Theorem 3.1 in [3]). *If  $n$  is any balancing number, then there is no balancing number  $k$  such that  $n < k < 3n + \sqrt{8n^2 + 1}$ .*

Its corollary is the following:

**Corollary 2.4** (Corollary 3.2 in [3]). *If  $n = B_m$  is a balancing number with  $m > 1$ , then we have  $B_{m-1} = 3n - \sqrt{8n^2 + 1}$ .*

They proved that a balancing number can also be generated by two balancing numbers.

**Theorem 2.5** (Theorem 4.1 in [3]). *If  $n$  and  $k$  are balancing numbers, then*

$$f(n, k) = n\sqrt{8k^2 + 1} + k\sqrt{8n^2 + 1} \tag{2.4}$$

*is also a balancing number.*

## 2.2. A recurrence relation and other properties

In [3] Behera and Panda proved that the balancing numbers fulfill the following recurrence relation

$$B_{m+1} = 6B_m - B_{m-1} \quad (m \geq 1)$$

where  $B_0 = 1$  and  $B_1 = 6$ . Using this recurrence relation they get interesting relations between balancing numbers.

**Theorem 2.6** (Theorem 5.1 in [3]). *For any  $m > 1$  we have*

- $B_{m+1} \cdot B_{m-1} = (B_m + 1)(B_m - 1)$ ,
- $B_m = B_k \cdot B_{m-k} - B_{k-1} \cdot B_{m-k-1}$  for any positive integer  $k < m$ ,
- $B_{2m} = B_m^2 - B_{m-1}^2$ ,
- $B_{2m+1} = B_m(B_{m+1} - B_{m-1})$ .

In [26] G. K. Panda established other interesting arithmetic-type, de-Moivre's-type and trigonometric-type properties of balancing numbers.

**Theorem 2.7** (Theorem 2.1 in [26]). *If  $m$  and  $k$  are natural numbers and  $m > k$ , then  $(B_m + B_k)(B_m - B_k) = B_{m+k} \cdot B_{m-k}$ .*

**Remark 2.8.** The Fibonacci numbers  $F_m$  satisfy a similar property (see [16] p. 59)

$$F_{m+k} \cdot F_{m-k} = F_m^2 - (-1)^{m+k} F_k^2.$$

We know that if  $m$  is natural number, then  $1 + 3 + \dots + (2m - 1) = m^2$ . In [26] G. K. Panda proved three properties of balancing numbers similar to the identity above. For balancing numbers we get:

**Theorem 2.9** (Theorem 2.2 in [26]).

- $B_1 + B_3 + \dots + B_{2m-1} = B_m^2$ ,
- $B_2 + B_4 + \dots + B_{2m} = B_m B_{m+1}$ ,
- $B_1 + B_2 + \dots + B_{2m} = B_m(B_m + B_{m+1})$ .

The identity  $(\cos x + i \sin x)^n = \cos nx + i \sin nx$  for complex numbers is known as the de-Moivre's formula. The following theorem gives a de-Moivre's-type property of balancing numbers. Let  $C_m = \sqrt{8B_m^2 + 1}$ .

**Theorem 2.10** (Theorem 2.3 in [26]). *If  $m$  and  $k$  are natural numbers, then*

$$(C_m + \sqrt{8}B_m)^k = C_{mk} + \sqrt{8}B_{mk}.$$

**Remark 2.11.** The Fibonacci ( $F_m$ ) and Lucas ( $L_m$ ) numbers satisfy a similar property

$$\left[ \frac{L_m + \sqrt{5}F_m}{2} \right]^r = \frac{L_{mr} + \sqrt{5}F_{mr}}{2}.$$

Panda proved another interesting result about the greatest common divisor of balancing numbers.

**Theorem 2.12** (Theorem 2.5 in [26]). *If  $m$  and  $k$  are natural numbers then*

$$\gcd(B_m, B_k) = B_{(m,k)}.$$

In [3] we can find nonrecursive forms to obtain balancing numbers. One of these results is the following:

**Theorem 2.13** (Theorem 7.1 in [3]). *If  $B_m$  is the  $m$ th balancing number then*

$$B_m = \frac{\lambda_1^{m+1} - \lambda_2^{m+1}}{\lambda_1 - \lambda_2}, \quad m = 0, 1, 2, \dots,$$

where  $\lambda_1 = 3 + \sqrt{8}$  and  $\lambda_2 = 3 - \sqrt{8}$ .

**Remark 2.14.** We get this formula easily using the companion polynomial of the recurrence relation of  $B_m$ .

### 2.3. Fibonacci and Lucas balancing numbers

In [21] K. Liptai obtained several results about special type of balancing numbers. Let us consider the definition below:

**Definition 2.15** ([21] and [22]). We call a balancing number a *Fibonacci* or a *Lucas balancing number* if it is a Fibonacci or a Lucas number, too.

Using this definition and companion polynomial of  $B_m$  K. Liptai proved that the balancing numbers are solutions of a Pell's equation.

**Theorem 2.16** (Theorem 1 in [21]). *The terms of the second order linear recurrence  $R(6, -1, 1, 6)$  are the solutions of the equation*

$$x^2 - 8y^2 = 1$$

for some integer  $y$ .

There is also a connection between Fibonacci or Lucas numbers and Pell's equation. The following theorem is due to D. E. Ferguson:

**Theorem 2.17** (Theorem in [7]). *The only solutions of the equation*

$$x^2 - 5y^2 = \pm 4$$

are  $x = \pm L_m$ ,  $y = \pm F_m$  ( $n = 0, 1, 2 \dots$ ), where  $L_m$  and  $F_m$  are the  $m$ th terms of the Lucas and Fibonacci sequences, respectively.

To find all Fibonacci or Lucas balancing numbers K. Liptai proved that there are finitely many common solutions of the Pell's equations above using a method of A. Baker and H. Davenport.

The main theorem in [21] and [22] are the following:

**Theorem 2.18** (Theorem 4 in [21] and [22]). *There is no Fibonacci or Lucas balancing number.*

**Remark 2.19.** Using another method L. Szalay got the same result for the solutions of simultaneous Pell equations in [35]. In this method he converted simultaneous Pell's equations into a family of Thue equations which could be solved completely.

## 3. Properties of cobalancing numbers

### 3.1. Introduction

By slightly modifying the definition 1.1 we get:

**Definition 3.1** ([27]). We call  $n \in \mathbb{N}$  a cobalancing number if

$$1 + 2 + \cdots + n = (n + 1) + (n + 2) + \cdots + (n + r^c)$$

for some  $r^c \in \mathbb{N}$ . Here we call  $r^c$  the cobalancer corresponding to the cobalancing number  $n$ . Denote  $n$  by  $B_m^c$  if  $n$  is the  $m$ th term of the sequence of cobalancing numbers.

**Remark 3.2.** The first three cobalancing numbers are 2, 14 and 84 with cobalancers 1, 6, 35, respectively.

### 3.2. Properties of cobalancing numbers

Cobalancing numbers  $B_m^c$  have similar properties to balancing numbers  $B_m$ . In [27] G. K. Panda and P. K. Ray proved the following properties:

**Theorem 3.3** (Theorem 2.2 in [27]). *If  $n = B_m^c$  is a cobalancing number with  $m > 1$  then  $B_{m+1}^c = 3n + \sqrt{8n^2 + 8n + 1} + 1$  and  $B_{m-1}^c = 3n - \sqrt{8n^2 + 8n + 1} + 1$ .*

By Theorem 3.3 they get a recurrence relation for cobalancing numbers that is

$$B_{m+1}^c = 6B_m^c - B_{m-1}^c + 2, \quad (m = 2, 3, \dots)$$

where they set  $B_1^c = 0$ . The following theorem is a consequence of the relation above.

**Theorem 3.4** (Theorem 3.1 in [27]). *Every cobalancing number is even.*

We also denote by  $r_m$  the balancer belonging to  $B_m$  and  $r_m^c$  the cobalancer belongig to  $B_m^c$ . Then by using the definition 1.1 and 3.1 the following theorems are valid:

**Theorem 3.5** (Theorem 6.1 in [27]). *Every balancer is a cobalancing number and every cobalancer is a balancing number.*

Using our notation we get:

**Theorem 3.6** (Theorem 6.2 in [27]). *We have  $r_m = B_m^c$  and  $r_{m+1}^c = B_m$  for every  $m = 1, 2, \dots$*

Panda and Ray got a corollary from the theorems above.

**Corollary 3.7** (Corollary 6.4 in [27]).  *$r_{m+1} = r_m + 2B_m$ .*

### 3.3. Connection between (co)balancing and Pell numbers

In [28] we can find interesting results about the connection of Pell, balancing or cobalancing numbers. Let  $P_m$  be the  $m$ th Pell number ( $m = 1, 2, \dots$ ). It is well known that

$$P_1 = 1, P_2 = 2, P_{m+1} = 2P_m + P_{m-1}.$$

The authors call  $C_m = \sqrt{8B_m^2 + 1}$  the  $m$ th Lucas-balancing number and  $c_m = \sqrt{8(B^c)_m^2 + 8B_m^c + 1}$  the  $m$ th Lucas-cobalancing number. The first result of them is the following:

**Theorem 3.8** (Theorem 2.2 in [28]). *The sequences of Lucas-balancing and Lucas-cobalancing numbers satisfy recurrence relations with identical balancing numbers. More precisely,  $C_1 = 3, C_2 = 17, C_{m+1} = 6C_m - C_{m-1}$  and  $c_1 = 1, c_2 = 7, c_{m+1} = 6c_m - c_{m-1}$  for  $m = 2, 3, \dots$*

In [28] the authors get a formula how to calculate balancing or cobalancing numbers from Pell numbers.

**Theorem 3.9** (Theorem 3.2 in [28]). *If  $P$  is a Pell number then  $\lceil P/2 \rceil$  is either a balancing number or a cobalancing number. More precisely  $P_{2m}/2 = B_m$  and  $\lceil P_{2m-1}/2 \rceil = B_m^c$  ( $m = 1, 2, \dots$ ).*

There is another result for calculating balancing number and its balancer, too.

**Theorem 3.10** (Theorem 3.4 in [28]). *The sum of the first  $2m - 1$  Pell numbers is equal to the sum of the  $m$ th balancing number and its balancer.*

## 4. Generalizations

### 4.1. Sequence balancing and cobalancing numbers

In [25] G. K. Panda defined sequence balancing and sequence cobalancing numbers.

**Definition 4.1** ([25]). Let  $\{s_m\}_{m=1}^\infty$  be a sequence of real numbers. We call a number  $s_m$  of this sequence a sequence balancing number if

$$s_1 + s_2 + \dots + s_{m-1} = s_{m+1} + s_{m+2} + \dots + s_{m+r}$$

for some natural number  $r$ . Similarly, we call  $s_m$  a sequence cobalancing number if

$$s_1 + s_2 + \dots + s_m = s_{m+1} + s_{m+2} + \dots + s_{m+r}$$

for some natural number  $r$ .

**Remark 4.2.** For example, if we take  $s_m = 2m$  then the sequence balancing numbers of this sequence are 12, 70, 408, ... which are twice the balancing numbers. It is also true for sequence cobalancing numbers and similarly in the case when  $s_m = \frac{m}{2}$ .

In [25] the author investigated the existence of sequence balancing or cobalancing numbers in the sequence of odd natural numbers. So, let  $s_m = 2m - 1$ . Using simple technics he got that the sequence of sequence balancing numbers in the sequence of odd natural numbers is given by  $\{2B_{m+1}^c + r_{m+1}^c + 1\}_{m=1}^\infty$  (see Theorem 2.1.4 in [25]). So, let the  $m$ th sequence balancing number in the sequence of odd natural numbers be denoted by  $x_m$ . Then by this fact above G. K. Panda got the following recurrence relation for these solutions.

**Theorem 4.3** (Theorem 2.1.5 in [25]). *The sequence  $\{x_m\}_{m=1}^\infty$  satisfies the recurrence relation  $x_{m+1} = 6x_m - x_{m-1}$  for  $m \geq 2$ .*

**Remark 4.4.** The author in [25] investigated also the existance of sequence balancing or cobalancing numbers in the cases when  $a_m = m+1$  and  $a_m = F_m$  (among Fibonacci numbers). In the first case the sequence balancing numbers among the numbers  $a_m = m+1$  can be given by a linear combination of balancing numbers.

In the second one he gets that the only sequence cobalancing number in the Fibonacci sequence is  $F_2 = 1$ .

## 4.2. Generalized balancing sequences

In [4] A. Bérczes, K. Liptai and I. Pink generalized the definition 4.1 due to G. K. Panda.

**Definition 4.5** ([4]). We call a binary recurrence  $R_i = R(A, B, R_0, R_1)$  a balancing sequence if

$$R_1 + R_2 + \cdots + R_{m-1} = R_{m+1} + R_{m+2} + \cdots + R_{m+k} \quad (4.1)$$

holds for some  $k \geq 1$  and  $m \geq 2$ .

In that paper they proved that any sequence  $R_i = R(A, B, 0, R_1)$  with conditions  $D = A^2 + 4B > 0$ ,  $(A, B) \neq (0, 1)$  is not a balancing sequence.

**Theorem 4.6** (Theorem 1 in [4]). *There is no balancing sequence of the form  $R_i = R(A, B, 0, R_1)$  with  $D = A^2 + 4B > 0$  except for  $(A, B) = (0, 1)$  in which case (4.1) has infinitely many solutions  $(m, k) = (m, m - 1)$  and  $(m, k) = (m, m)$  for  $m \geq 2$ .*

By this theorem they got the following corollary.

**Corollary 4.7** (Corollary 1 in [4]). *Let  $R_i = R(A, B, 0, 1)$  be a Lucas-sequence with  $A^2 + 4B > 0$ . Then  $R_i$  is not a balancing sequence.*

## 4.3. $(k, l)$ -numerical centers

**Definition 4.8** ([23]). Let  $y, k$  and  $l$  be fixed positive integers with  $y \geq 4$ . A positive integer  $x$  ( $x \leq y - 2$ ) is called a  $(k, l)$ -power numerical center for  $y$ , or a  $(k, l)$ -balancing number for  $y$  if

$$1^k + 2^k + \cdots + (x - 1)^k = (x + 1)^l + \cdots + (y - 1)^l.$$

**Remark 4.9.** In [8] R. Finkelstein studied "The house problem" and introduced the notion of first-power numerical center which coincides with the notion of balancing number  $B_m$ . He proved that infinitely many integers  $y$  possess  $(1, 1)$ -power centers and there is no integer  $y > 1$  with a  $(2, 2)$ -power numerical center. In his paper, he conjectured that if  $k > 1$  then there is no integer  $y > 1$  with  $(k, k)$ -power numerical center. Later in [33] his conjecture was confirmed for  $k = 3$ . Recently, Ingram in [17] proved Finkelstein's conjecture for  $k = 5$ .

In [23] the authors proved a general result about  $(k, l)$ -balancing numbers, but they could not deal with Finkelstein's conjecture in its full generality. Their main results are the following theorems.

**Theorem 4.10** (Theorem 1 in [23]). *For any fixed positive integer  $k > 1$ , there are only finitely many positive pairs of integers  $(y, l)$  such that  $y$  possesses a  $(k, l)$ -power numerical center.*

For the proof of this theorem they used a result from [31]. Thus Theorem 4.10 is ineffective in case  $l \leq k$  in the sense that no upper bound was made for possible numerical centers except for the cases when  $l = 1$  or  $l = 3$ .

**Theorem 4.11** (Theorem 2 in [23]). *Let  $k$  be a fixed positive integer with  $k \geq 1$  and  $l \in \{1, 3\}$ . If  $(k, l) \neq (1, 1)$ , then there are only finitely many  $(k, l)$ -balancing numbers, and these balancing numbers are bounded by an effectively computable constant depending only on  $k$ .*

**Remark 4.12.** In [23] the authors gave an example for numerical centers in the case when  $(k, l) = (2, 1)$ . After solving an elliptic equation by MAGMA [24] they got three  $(2, 1)$ -power numerical centers  $x$ , namely 5, 13 and 36.

#### 4.4. $(a, b)$ -type balancing numbers

Another generalization is the following by T. Kovács, K. Liptai and P. Olajos:

**Definition 4.13** ([20]). Let  $a, b$  be nonnegative coprime integers. We call a positive integer  $an + b$  an  $(a, b)$ -type balancing number if

$$(a + b) + (2a + b) + \dots + (a(n - 1) + b) = (a(n + 1) + b) + \dots + (a(n + r) + b)$$

for some  $r \in \mathbb{N}$ . Here  $r$  is called the balancer corresponding to the balancing number. We denote the positive integer  $an + b$  by  $B_m^{(a,b)}$  if this number is the  $m$ th among the  $(a, b)$ -type balancing numbers.

**Remark 4.14.** We have to mention that if we use notation  $a_n = an + b$  then we get sequence balancing numbers and if  $a = 1$  and  $b = 0$  for  $(a, b)$ -type balancing numbers than we get balancing numbers  $B_m$ .

Using the definition the authors in [20] get the following proposition:

**Lemma 4.15** (Proposition 1 in [20]). *If  $B_m^{(a,b)}$  is an  $(a, b)$ -type balancing number then the following equation*

$$z^2 - 8 \left( B_m^{(a,b)} \right)^2 = a^2 - 4ab - 4b^2 \quad (4.2)$$

*is valid for some  $z \in \mathbb{Z}$ .*

#### 4.4.1. Polynomial values among balancing numbers

Let us consider the following equation for  $(a, b)$ -type balancing numbers

$$B_m^{(a,b)} = f(x) \quad (4.3)$$

where  $f(x)$  is a monic polynomial with integer coefficients. By Proposition 4.15 and the result from Brindza [5] Kovács, Liptai and Olajos proved the following theorem:

**Theorem 4.16** (Theorem 1 in [20]). *Let  $f(x)$  be a monic polynomial with integer coefficients, of degree  $\geq 2$ . If  $a$  is odd, then for the solutions of (4.3) we have  $\max(m, |x|) < c_0(f, a, b)$ , where  $c_0(f, a, b)$  is an effectively computable constant depending only on  $a, b$  and  $f$ .*

Let us consider a special case of Theorem 4.16 with  $f(x) = x^l$ . Using one of the results from Bennett [1] the authors in [20] get the following theorem:

**Theorem 4.17** (Theorem 2 in [20]). *If  $a^2 - 4ab - 4b^2 = 1$ , then there is no perfect power  $(a, b)$ -balancing number.*

**Remark 4.18.** There are infinitely many integer solutions of the equation  $a^2 - 4ab - 4b^2 = 1$ .

The authors are interested in combinatorial numbers (see also Kovács [19]), that is binomial coefficients, power sums, alternating power sums and products of consecutive integers. For all  $k, x \in \mathbb{N}$  let

$$\begin{aligned} S_k(x) &= 1^k + 2^k + \dots + (x-1)^k, \\ T_k(x) &= -1^k + 2^k - \dots + (-1)^{x-1}(x-1)^k, \\ \Pi_k(x) &= x(x+1)\dots(x+k-1). \end{aligned}$$

We mention that the degree of  $S_k(x)$ ,  $T_k(x)$  and  $\Pi_k(x)$  are  $k+1$ ,  $k$  and  $k$ , respectively and  $\binom{x}{k}$ ,  $S_k(x)$ ,  $T_k(x)$  are polynomials with non-integer coefficients. Moreover, in the case when  $f(x) = \Pi_k(x)$  Theorem 4.16 is valid but the parameter  $a$  is odd.

Let us consider the following equation

$$B_m^{(a,b)} = p(x), \quad (4.4)$$

where  $p(x)$  is a polynomial with rational integer coefficients. In this case Kovács, Liptai and Olajos gave effective results for the solutions of equation (4.4).



**Theorem 4.19** (Theorem 3 in [20]). *Let  $k \geq 2$  and  $p(x)$  be one of the polynomials  $\binom{x}{k}$ ,  $\Pi_k(x)$ ,  $S_{k-1}(x)$ ,  $T_k(x)$ . Then the solutions of equation (4.3) satisfy  $\max(m, |x|) < c_1(a, b, k)$ , where  $c_1(a, b, k)$  is an effectively computable constant depending only on  $a$ ,  $b$  and  $k$ .*

#### 4.4.2. Numerical results

In [20] T. Kovács, K. Liptai and the author completely solve the above type equations for some small values of  $k$  that lead to genus 1 or genus 2 equations. In this case the equation can be written as

$$y^2 = 8f(x)^2 + 1, \tag{4.5}$$

where  $f(x)$  is one of the following polynomials. Beside binomial coefficients  $\binom{x}{k}$ , we consider power sums and products of consecutive integers, as well. We mention that in their results, for the sake of completeness, they provide all integral (even the negative) solutions to equation (4.5).

**Genus 1 and 2 equations** They completely solve equation (4.5) for all parameter values  $k$  in case when they can reduce the equation to an equation of genus 1. We have to mention that a similar argument has been used to solve several combinatorial Diophantine equations of different types, for example in [9], [10], [12], [13], [18], [19], [29], [30], [34], [37], [38]. Further they also solved a particular case ( $f(x) = S_5(x)$ ) when equation (4.3) can be reduced to the resolution of a genus 2 equation. To solve this equation, they used the so-called Chabauty method. We have to note that the Chabauty method has already been successfully used to solve certain combinatorial Diophantine equations, see e.g. the corresponding results in the papers [6], [11], [14], [15], [32], [36] and the references given there.

**Theorem 4.20** (Theorem 4 in [20]). *Suppose that  $a^2 - 4ab - 4b^2 = 1$ . Let  $f(x) \in \{\binom{x}{2}, \binom{x}{3}, \binom{x}{4}, \Pi_2(x), \Pi_3(x), \Pi_4(x), S_1(x), S_2(x), S_3(x), S_5(x)\}$ . Then the solutions  $(m, x)$  of equation (4.3) are those contained in Table 1. For the corresponding parameter values we have  $(a, b) = (1, 0)$  in all cases.*

**Remark 4.21.** In [20] the authors considered some other related equations that led to genus 2 equations. However, because of certain technical problems, they could not solve them by the Chabauty method. They determined the "small" solutions (i.e.  $|x| \leq 10000$ ) of equation (4.5) in cases

$$f(x) \in \left\{ \binom{x}{6}, \binom{x}{8}, \Pi_6(x), \Pi_8(x), S_7(x) \right\}.$$

Their conjecture is that that there is no solution for these equations.

$f(x)$	Solutions $(m, x)$ of (4.3)
$\binom{x}{2}$	$(1, -3), (1, 4)$
$\binom{x}{3}$	$(2, -5), (2, 7)$
$\binom{x}{4}$	$(2, -4), (2, 7)$
$\Pi_2(x)$	$(1, -3), (1, 2)$
$\Pi_3(x)$	$(1, -3), (1, 1)$
$\Pi_4(x)$	$\emptyset$
$S_1(x)$	$(1, -4), (1, 3)$
$S_2(x)$	$(3, -8), (3, 9), (5, -27), (5, 28)$
$S_3(x)$	$\emptyset$
$S_5(x)$	$\emptyset$

Table 1

## References

- [1] BENNETT, M. A., Rational approximation to algebraic numbers of small height: the Diophantine equation  $|ax^n - by^n| = 1$ , *J. Reine Angew. Math.*, 535 (2001) 1–49.
- [2] BAKER, A., WÜSTHOLZ, G., Logarithmic forms and group varieties, *J. Reine Angew. Math.*, 442 (1993) 19–62.
- [3] BEHERA, A., PANDA, G. K., On the square roots of triangular numbers, *Fibonacci Quarterly*, 37 No. 2 (1999) 98–105.
- [4] BÉRCZES, A., LIPTAI, K., PINK, I., On generalized balancing numbers, *Fibonacci Quarterly*, 48 No. 2 (2010) 121–128.
- [5] BRINDZA, B., On  $S$ -integral solutions of the equation  $y^m = f(x)$ , *Acta Math. Hungar.* 44 (1984) 133–139.
- [6] BRUIN, N., GYÖRY, K., HAJDU, L., TENGELY T., Arithmetic progressions consisting of unlike powers, *Indag. Math.* 17 (2006) 539–555.
- [7] FERGUSON, D. E., Letter to the editor, *Fibonacci Quarterly*, 8 (1970) 88–89.
- [8] FINKELSTEIN, R. P., The House Problem, *American Math. Monthly*, 72 (1965) 1082–1088.
- [9] HAJDU, L., On a diophantine equation concerning the number of integer points in special domains II, *Publ. Math. Debrecen*, 51 (1997) 331–342.
- [10] HAJDU, L., On a diophantine equation concerning the number of integer points in special domains, *Acta Math. Hungar.*, 78 (1998) 59–70.
- [11] HAJDU, L., Powerful arithmetic progressions, *Indeg. Math.*, 19 (2008) 547–561.
- [12] HAJDU, L., PINTÉR, Á., Square product of three integers in short intervals, *Math. Comp.*, 68 (1999) 1299–1301.
- [13] HAJDU, L., PINTÉR, Á., Combinatorial diophantine equations, *Publ. Math. Debrecen*, 56 (2000) 391–403.
- [14] HAJDU, L., TENGELY SZ., Arithmetic progressions of squares, cubes and  $n$ -th powers, *J. Functiones es Approximatio*, 41 No. 2 (2009) 129–138.

- [15] HAJDU, L., TENGYEL, SZ., TIJDEMAN, R., Cubes in products of terms in arithmetic progression, *Publ. Math. Debrecen*, 74 (2009) 215–232.
- [16] HOGGATT JR., V. E., Fibonacci and Lucas numbers, *Houghton Mifflin Company IV*, (1969) 92 p.
- [17] INGRAM, P., On the  $k$ -th power numerical centres, *C. R. Math. Acad. Sci. R. Can.*, 27 (2005) 105–110.
- [18] KOVÁCS, T., Combinatorial diophantine equations - the genus 1 case, *Publ. Math. Debrecen*, 72 (2008) 243–255.
- [19] KOVÁCS, T., Combinatorial numbers in binary recurrences, *Period. Math. Hungar.*, 58 No. 1 (2009) 83–98.
- [20] KOVÁCS, T., LIPTAI, K., OLAJOS, P., About  $(a, b)$ -type balancing numbers, *Publ. Math. Debrecen*, 77 No. 3-4 (2010), 485–498.
- [21] LIPTAI, K., Fibonacci balancing numbers, *Fibonacci Quarterly*, 42 No. 4 (2004) 330–340.
- [22] LIPTAI, K., Lucas balancing numbers, *Acta Math. Univ. Ostrav.*, 14 No. 1 (2006) 43–47.
- [23] LIPTAI, K., LUCA F., PINTÉR, Á., SZALAY L., Generalized balancing numbers, *Indagationes Math. N. S.*, 20 (2009) 87–100.
- [24] MAGMA Computational Algebra System, Computational Algebra Group School of Mathematics and Statistics, University of Sydney, NSW 2006, Australia, <http://magma.maths.usyd.edu.au/magma/>
- [25] PANDA, G. K., Sequence balancing and cobalancing numbers, *Fibonacci Quarterly*, 45 (2007) 265–271.
- [26] PANDA, G. K., Some fascinating properties of balancing numbers, *Proceedings of the Eleventh International Conference on Fibonacci Numbers and their Applications*, Cong. Numer. 194 (2009) 185–189.
- [27] PANDA, G. K., RAY, P. K., Cobalancing numbers and cobalancers, *Int. J. Math. Sci.*, No. 8 (2005) 1189–1200.
- [28] PANDA, G. K., RAY, P. K., Some links of balancing and cobalancing numbers and with Pell and associated Pell numbers, (oral communicated).
- [29] PINTÉR, Á., A note on the Diophantine equation  $\binom{x}{4} = \binom{y}{2}$ , *Publ. Math. Debrecen*, 47 (1995) 411–415.
- [30] PINTÉR, Á, DE WEGER, B. M. M.,  $210 = 14 \times 15 = 5 \times 6 \times 7 = \binom{21}{2} = \binom{10}{4}$ , *Publ. Math. Debrecen*, 51 (1997) 175–189.
- [31] RAKACZKI, CS., On the diophantine equation  $S_m(x) = g(y)$ , *Publ. Math. Debrecen*, 65 (2004) 439–460.
- [32] SHOREY, T. N., LAISHRAM, S., TENGYEL, SZ., Squares in products in arithmetic progression with at most one term omitted and common difference a prime power, *Acta Arith.*, 135 (2008) 143–158.
- [33] STEINER, R., On the  $k$ -th power numerical centers, *Fibonacci Quarterly*, 16 (1978) 470–471.
- [34] STROEKER, R. J., DE WEGER, B. M. M., Elliptic binomial diophantine equations, *Math. Comp.*, 68 (1999) 1257–1281.

- [35] SZALAY, L., On the resolution of simultaneous Pell equations, *Annales Mathematicae et Informaticae*, 34 (2007) 77–87.
- [36] TENGELY, Sz., Note on a paper "An extension of a theorem of Euler" by Hirata-Kohno et al., *Acta Arith.*, 134 (2008) 329–335.
- [37] DE WEGER, B. M. M., A binomial Diophantine equation, *Quart. J. Math. Oxford Ser. (2)*, 47 (1996) 221–231.
- [38] DE WEGER, B. M. M., Equal binomial coefficients: some elementary considerations, *J. Number Theory*, 63 (1997) 373–386.

**Péter Olajos**

Department of Applied Mathematics,  
University of Miskolc,  
H-3515 Miskolc-Egyetemváros,  
Hungary  
e-mail: [matolaj@uni-miskolc.hu](mailto:matolaj@uni-miskolc.hu)

# Signed $(k, k)$ -domatic number of a graph

S. M. Sheikholeslami<sup>a</sup>, L. Volkmann<sup>b</sup>

<sup>a</sup>Azərbaycan University of Tarbiat Moallem Tabriz  
Department of Mathematics

<sup>b</sup>RWTH-Aachen University, Lehrstuhl II für Mathematik

*Submitted 19 March 2010; Accepted 4 October 2010*

## Abstract

Let  $G$  be a finite and simple graph with vertex set  $V(G)$ , and let  $f: V(G) \rightarrow \{-1, 1\}$  be a two-valued function. If  $k \geq 1$  is an integer and  $\sum_{x \in N[v]} f(x) \geq k$  for each  $v \in V(G)$ , where  $N[v]$  is the closed neighborhood of  $v$ , then  $f$  is a signed  $k$ -dominating function on  $G$ . A set  $\{f_1, f_2, \dots, f_d\}$  of signed  $k$ -dominating functions on  $G$  with the property that  $\sum_{i=1}^d f_i(x) \leq k$  for each  $x \in V(G)$ , is called a signed  $(k, k)$ -dominating family (of functions) on  $G$ . The maximum number of functions in a signed  $(k, k)$ -dominating family on  $G$  is the signed  $(k, k)$ -domatic number on  $G$ , denoted by  $d_S^k(G)$ .

In this paper we initiate the study of the signed  $(k, k)$ -domatic number, and we present different bounds on  $d_S^k(G)$ . Some of our results are extensions of well-known properties of the signed domatic number  $d_S(G) = d_S^1(G)$ .

*Keywords:* Signed  $(k, k)$ -domatic number, signed  $k$ -dominating function, signed  $k$ -domination number

*MSC:* 05C69

## 1. Terminology and introduction

Various numerical invariants of graphs concerning domination were introduced by means of dominating functions and their variants. In this paper we define the *signed  $(k, k)$ -domatic number* in an analogous way as Volkmann and Zelinka [6] have introduced the signed domatic number.

We consider finite, undirected and simple graphs  $G$  with vertex set  $V(G)$  and edge set  $E(G)$ . The cardinality of the vertex set of a graph  $G$  is called the *order* of  $G$  and is denoted by  $n(G)$ . If  $v \in V(G)$ , then  $N(v)$  is the *open neighborhood* of  $v$ , i.e., the set of all vertices adjacent to  $v$ . The *closed neighborhood*  $N[v]$  of a vertex

$v$  consists of the vertex set  $N(v) \cup \{v\}$ . The number  $d(v) = |N(v)|$  is the *degree* of the vertex  $v$ . The *minimum* and *maximum degree* of a graph  $G$  are denoted by  $\delta(G)$  and  $\Delta(G)$ . The *complement* of a graph  $G$  is denoted by  $\overline{G}$ . We write  $K_n$  for the *complete graph* of order  $n$  and  $C_n$  for a *cycle* of length  $n$ . A *fan* and a *wheel* is a graph obtained from a path and a cycle by adding a new vertex and edges joining it to all the vertices of the path and cycle, respectively. If  $A \subseteq V(G)$  and  $f$  is a mapping from  $V(G)$  into some set of numbers, then  $f(A) = \sum_{x \in A} f(x)$ .

If  $k \geq 1$  is an integer, then the *signed  $k$ -dominating function* is defined in [7] as a two-valued function  $f: V(G) \rightarrow \{-1, 1\}$  such that  $\sum_{x \in N[v]} f(x) \geq k$  for each  $v \in V(G)$ . The sum  $f(V(G))$  is called the weight  $w(f)$  of  $f$ . The minimum of weights  $w(f)$ , taken over all signed  $k$ -dominating functions  $f$  on  $G$ , is called the *signed  $k$ -domination number* of  $G$ , denoted by  $\gamma_{kS}(G)$ . As the assumption  $\delta(G) \geq k - 1$  is necessary, we always assume that when we discuss  $\gamma_{kS}(G)$ , all graphs involved satisfy  $\delta(G) \geq k - 1$  and thus  $n(G) \geq k$ . The special case  $k = 1$  was defined and investigated in [1]. Further information on  $\gamma_{1S}(G) = \gamma_S(G)$  can be found in the monographs [2] and [3] by Haynes, Hedetniemi, and Slater.

Rall [4] has defined a variant of the domatic number of  $G$ , namely the fractional domatic number of  $G$ , using functions on  $V(G)$ . Analogous to the fractional domatic number we may define the signed  $(k, k)$ -domatic number.

A set  $\{f_1, f_2, \dots, f_d\}$  of signed  $k$ -dominating functions on  $G$  with the property that  $\sum_{i=1}^d f_i(x) \leq k$  for each  $x \in V(G)$ , is called a *signed  $(k, k)$ -dominating family* on  $G$ . The maximum number of functions in a signed  $(k, k)$ -dominating family on  $G$  is the *signed  $(k, k)$ -domatic number* of  $G$ , denoted by  $d_S^k(G)$ .

First we study basic properties of  $d_S^k(G)$ . Some of them are extensions of well-known results on the signed domatic number  $d_S(G) = d_S^1(G)$  given in [6]. Using these results, we determine the signed  $(k, k)$ -domatic numbers of fans, wheels and grids.

## 2. Basic properties of the signed $(k, k)$ -domatic number

**Theorem 2.1.** *The signed  $(k, k)$ -domatic number  $d_S^k(G)$  is well-defined for each graph  $G$  with  $\delta(G) \geq k - 1$ .*

**Proof.** Since  $\delta(G) \geq k - 1$ , the function  $f: V(G) \rightarrow \{-1, 1\}$  with  $f(v) = 1$  for each  $v \in V(G)$  is a signed  $k$ -dominating function on  $G$ . Thus the family  $\{f\}$  is a signed  $(k, k)$ -dominating family on  $G$ . Therefore the set of signed  $k$ -dominating functions on  $G$  is non-empty and there exists the maximum of their cardinalities, which is the signed  $(k, k)$ -domatic number of  $G$ .  $\square$

**Theorem 2.2.** *If  $G$  is a graph of order  $n$ , then*

$$\gamma_{kS}(G)d_S^k(G) \leq kn.$$

**Proof.** If  $\{f_1, f_2, \dots, f_d\}$  is a signed  $(k, k)$ -dominating family on  $G$  such that  $d = d_S^k(G)$ , then the definitions imply

$$\begin{aligned} d\gamma_{kS}(G) &= \sum_{i=1}^d \gamma_{kS}(G) \leq \sum_{i=1}^d \sum_{x \in V(G)} f_i(x) \\ &= \sum_{x \in V(G)} \sum_{i=1}^d f_i(x) \leq \sum_{x \in V(G)} k = kn. \end{aligned}$$

□

**Theorem 2.3.** If  $G$  is a graph with minimum degree  $\delta(G) \geq k - 1$ , then

$$d_S^k(G) \leq \delta(G) + 1.$$

**Proof.** Let  $\{f_1, f_2, \dots, f_d\}$  be a signed  $(k, k)$ -dominating family on  $G$  such that  $d = d_S^k(G)$ . If  $v \in V(G)$  is a vertex of minimum degree  $\delta(G)$ , then it follows that

$$\begin{aligned} dk &= \sum_{i=1}^d k \leq \sum_{i=1}^d \sum_{x \in N[v]} f_i(x) \\ &= \sum_{x \in N[v]} \sum_{i=1}^d f_i(x) \\ &\leq \sum_{x \in N[v]} k = k(\delta(G) + 1), \end{aligned}$$

and this implies the desired upper bound on the signed  $(k, k)$ -domatic number. □

The special case  $k = 1$  in Theorems 2.2 and 2.3 can be found in [6]. As an application of Theorem 2.3, we will prove the following Nordhaus-Gaddum type result.

**Theorem 2.4.** If  $k \geq 1$  is an integer and  $G$  a graph of order  $n$  such that  $\delta(G) \geq k - 1$  and  $\delta(\overline{G}) \geq k - 1$ , then

$$d_S^k(G) + d_S^k(\overline{G}) \leq n + 1.$$

If  $d_S^k(G) + d_S^k(\overline{G}) = n + 1$ , then  $G$  is regular.

**Proof.** Since  $\delta(G) \geq k - 1$  and  $\delta(\overline{G}) \geq k - 1$ , it follows from Theorem 2.3 that

$$\begin{aligned} d_S^k(G) + d_S^k(\overline{G}) &\leq (\delta(G) + 1) + (\delta(\overline{G}) + 1) \\ &= (\delta(G) + 1) + (n - \Delta(G) - 1 + 1) \\ &\leq n + 1, \end{aligned}$$

and this is the desired Nordhaus-Gaddum inequality. If  $G$  is not regular, then  $\Delta(G) - \delta(G) \geq 1$ , and the above inequality chain leads to the better bound  $d_S^k(G) + d_S^k(\overline{G}) \leq n$ . This completes the proof. □

**Theorem 2.5.** *If  $v$  is a vertex of a graph  $G$  such that  $d(v)$  is odd and  $k$  is odd or  $d(v)$  is even and  $k$  is even, then*

$$d_S^k(G) \leq \frac{k}{k+1}(d(v) + 1).$$

**Proof.** Let  $\{f_1, f_2, \dots, f_d\}$  be a signed  $(k, k)$ -dominating family on  $G$  such that  $d = d_S^k(G)$ . Assume first that  $d(v)$  and  $k$  are odd. The definition yields to  $\sum_{x \in N[v]} f_i(x) \geq k$  for each  $i \in \{1, 2, \dots, d\}$ . On the left-hand side of this inequality a sum of an even number of odd summands occurs. Therefore it is an even number, and as  $k$  is odd, we obtain  $\sum_{x \in N[v]} f_i(x) \geq k+1$  for each  $i \in \{1, 2, \dots, d\}$ . It follows that

$$\begin{aligned} k(d(v) + 1) &= \sum_{x \in N[v]} k \geq \sum_{x \in N[v]} \sum_{i=1}^d f_i(x) \\ &= \sum_{i=1}^d \sum_{x \in N[v]} f_i(x) \\ &\geq \sum_{i=1}^d (k+1) = d(k+1), \end{aligned}$$

and this leads to the desired bound. Assume next that  $d(v)$  and  $k$  are even. Note that  $\sum_{x \in N[v]} f_i(x) \geq k$  for each  $i \in \{1, 2, \dots, d\}$ . On the left-hand side of this inequality a sum of an odd number of odd summands occurs. Therefore it is an odd number, and as  $k$  is even, we obtain  $\sum_{x \in N[v]} f_i(x) \geq k+1$  for each  $i \in \{1, 2, \dots, d\}$ . Now the desired bound follows as above, and the proof is complete.  $\square$

The next result is an immediate consequence of Theorem 2.5.

**Corollary 2.6.** *If  $G$  is a graph such that  $\delta(G)$  and  $k$  are odd or  $\delta(G)$  and  $k$  are even, then*

$$d_S^k(G) \leq \frac{k}{k+1}(\delta(G) + 1).$$

As an Application of Corollary 2.6, we will improve the Nordhaus-Gaddum bound in Theorem 2.4 for many cases.

**Theorem 2.7.** *Let  $k \geq 1$  be an integer, and let  $G$  be a graph of order  $n$  such that  $\delta(G) \geq k-1$  and  $\delta(\overline{G}) \geq k-1$ . If  $\Delta(G) - \delta(G) \geq 1$  or  $k$  is even or  $k$  and  $\delta(G)$  are odd or  $k$  is odd and  $\delta(G)$  and  $n$  are even, then*

$$d_S^k(G) + d_S^k(\overline{G}) \leq n.$$

**Proof.** If  $\Delta(G) - \delta(G) \geq 1$ , then Theorem 2.4 implies the desired bound. Thus assume now that  $G$  is  $\delta(G)$ -regular.



*Case 1:* Assume that  $k$  is even. If  $\delta(G)$  is even, then it follows from Theorem 2.3 and Corollary 2.6 that

$$\begin{aligned} d_S^k(G) + d_S^k(\overline{G}) &\leq \frac{k}{k+1}(\delta(G) + 1) + (\delta(\overline{G}) + 1) \\ &= \frac{k}{k+1}(\delta(G) + 1) + (n - \delta(G) - 1 + 1) \\ &< n + 1, \end{aligned}$$

and we obtain the desired bound. If  $\delta(G)$  is odd, then  $n$  is even and thus  $\delta(\overline{G}) = n - \delta(G) - 1$  is even. Combining Theorem 2.3 and Corollary 2.6, we find that

$$\begin{aligned} d_S^k(G) + d_S^k(\overline{G}) &\leq (\delta(G) + 1) + \frac{k}{k+1}(\delta(\overline{G}) + 1) \\ &= (n - \delta(\overline{G})) + \frac{k}{k+1}(\delta(\overline{G}) + 1) \\ &< n + 1, \end{aligned}$$

and this completes the proof of Case 1.

*Case 2:* Assume that  $k$  is odd. If  $\delta(G)$  is odd, then it follows from Theorem 2.3 and Corollary 2.6 that

$$d_S^k(G) + d_S^k(\overline{G}) \leq \frac{k}{k+1}(\delta(G) + 1) + (n - \delta(G)) < n + 1.$$

If  $\delta(G)$  is even and  $n$  is even, then  $\delta(\overline{G}) = n - \delta(G) - 1$  is odd, and we obtain the desired bound as above.  $\square$

**Theorem 2.8.** *If  $G$  is a graph such that  $k$  is odd and  $d_S^k(G)$  is even or  $k$  is even and  $d_S^k(G)$  is odd, then*

$$d_S^k(G) \leq \frac{k-1}{k}(\delta(G) + 1).$$

**Proof.** Let  $\{f_1, f_2, \dots, f_d\}$  be a signed  $(k, k)$ -dominating family on  $G$  such that  $d = d_S^k(G)$ . Assume first that  $k$  is odd and  $d$  is even. If  $x \in V(G)$  is an arbitrary vertex, then  $\sum_{i=1}^d f_i(x) \leq k$ . On the left-hand side of this inequality a sum of an even number of odd summands occurs. Therefore it is an even number, and as  $k$  is odd, we obtain  $\sum_{i=1}^d f_i(x) \leq k - 1$  for each  $x \in V(G)$ . If  $v$  is a vertex of minimum degree, then it follows that

$$\begin{aligned} dk &= \sum_{i=1}^d k \leq \sum_{i=1}^d \sum_{x \in N[v]} f_i(x) \\ &= \sum_{x \in N[v]} \sum_{i=1}^d f_i(x) \\ &\leq \sum_{x \in N[v]} (k - 1) = (\delta(G) + 1)(k - 1), \end{aligned}$$

and this yields to the desired bound. Assume second that  $k$  is even and  $d$  is odd. If  $x \in V(G)$  is an arbitrary vertex, then  $\sum_{i=1}^d f_i(x) \leq k$ . On the left-hand side of this inequality a sum of an odd number of odd summands occurs. Therefore it is an odd number, and as  $k$  is even, we obtain  $\sum_{i=1}^d f_i(x) \leq k - 1$  for each  $x \in V(G)$ . Now the desired bound follows as above, and the proof is complete.  $\square$

According to Theorem 2.1,  $d_S^k(G)$  is a positive integer. If we suppose in the case  $k = 1$  that  $d_S(G) = d_S^1(G)$  is an even integer, then Theorem 2.8 leads to the contradiction  $d_S(G) \leq 0$ . Consequently, we obtain the next known result.

**Corollary 2.9** (Volkmann, Zelinka [6] 2005). *The signed domatic number  $d_S(G)$  is an odd integer.*

**Corollary 2.10.** *If  $T$  is a nontrivial tree, then  $d_S(T) = 1$  and  $d_S^2(T) \leq 2$ . In addition, if the diameter of  $T$  is at most three, then  $d_S^2(T) = 1$ .*

**Proof.** Theorem 2.3 implies that  $d_S(T) \leq 2$  and  $d_S^2(T) \leq 2$ . Applying Corollary 2.9, we obtain  $d_S(T) = 1$ . Now let  $f$  be a signed 2-dominating function on  $T$ . Then we observe that  $f(x) = 1$  if  $x$  is a leaf or  $x$  is neighbor of a leaf. However, if the diameter of  $T$  is at most three, then each vertex of  $T$  is a leaf or a neighbor of a leaf and thus  $f(x) = 1$  for every vertex  $x \in V(T)$ . This shows that  $d_S^2(T) = 1$  in that case, and the proof is complete.  $\square$

The following example demonstrates that the bound  $d_S^2(T) \leq 2$  in Corollary 2.10 is sharp.

Let  $T'$  be a tree of order 10 with the leaves  $u_1, u_2, v_1, v_2, w_1, w_2$  and the vertices  $u_3, v_3, w_3$  and  $z$  such that  $u_3$  is adjacent to  $u_1$  and  $u_2$ ,  $v_3$  is adjacent to  $v_1$  and  $v_2$ ,  $w_3$  is adjacent to  $w_1$  and  $w_2$  and  $z$  is adjacent to  $u_3, v_3$  and  $w_3$ . Then the functions  $f_i: V(T') \rightarrow \{-1, 1\}$  such that  $f_1(x) = 1$  for each  $x \in V(T')$  and  $f_2(z) = -1$  and  $f_2(x) = 1$  for each vertex  $x \in V(T') \setminus \{z\}$  are signed 2-dominating functions on  $T'$  such that  $f_1(x) + f_2(x) \leq 2$  for each vertex  $x \in V(T')$ . Using Corollary 2.10, we conclude that  $d_S^2(T') = 2$ .

**Theorem 2.11.** *Let  $k \geq 2$  be an integer, and let  $G$  be a graph with minimum degree  $\delta(G) \geq k - 1$ . Then  $d_S^k(G) = 1$  if and only if for every vertex  $v \in V(G)$  the closed neighborhood  $N[v]$  contains a vertex of degree at most  $k$ .*

**Proof.** Assume that  $N[v]$  contains a vertex of degree at most  $k$  for every vertex  $v \in V(G)$ , and let  $f$  be a signed  $k$ -dominating function on  $G$ . If  $d(v) \leq k$ , then it follows that  $f(v) = 1$ . If  $d(x) \leq k$  for a neighbor  $x$  of  $v$ , then we observe  $f(v) = 1$  too. Hence  $f(v) = 1$  for each  $v \in V(G)$  and thus  $d_S^k(G) = 1$ .

Conversely, assume that  $d_S^k(G) = 1$ . If  $G$  contains a vertex  $w$  such  $d(x) \geq k + 1$  for each  $x \in N[w]$ , then the functions  $f_i: V(G) \rightarrow \{-1, 1\}$  such that  $f_1(x) = 1$  for each  $x \in V(G)$  and  $f_2(w) = -1$  and  $f_2(x) = 1$  for each vertex  $x \in V(G) \setminus \{w\}$  are signed  $k$ -dominating functions on  $G$  such that  $f_1(x) + f_2(x) \leq 2 \leq k$  for each vertex  $x \in V(G)$ . Thus  $\{f_1, f_2\}$  is a signed  $(2, 2)$ -dominating family on  $G$ , a contradiction to  $d_S^k(G) = 1$ .  $\square$

Next we present a lower bound on the signed  $(k, k)$ -domatic number.

**Theorem 2.12.** *Let  $k \geq 1$  be an integer, and let  $G$  be a graph with minimum degree  $\delta(G) \geq k - 1$ . If  $G$  contains a vertex  $v \in V(G)$  such that all vertices of  $N[N[v]]$  have degree at least  $k + 1$ , then  $d_S^k(G) \geq k$ .*

**Proof.** Let  $\{u_1, u_2, \dots, u_k\} \subset N(v)$ . The hypothesis that all vertices of  $N[N[v]]$  have degree at least  $k + 1$  implies that the functions  $f_i: V(G) \rightarrow \{-1, 1\}$  such that  $f_i(u_i) = -1$  and  $f_i(x) = 1$  for each vertex  $x \in V(G) \setminus \{u_i\}$  are signed  $k$ -dominating functions on  $G$  for  $i \in \{1, 2, \dots, k\}$ . Since  $f_1(x) + f_2(x) + \dots + f_k(x) \leq k$  for each vertex  $x \in V(G)$ , we observe that  $\{f_1, f_2, \dots, f_k\}$  is a signed  $(k, k)$ -dominating family on  $G$ , and Theorem 2.12 is proved.  $\square$

**Corollary 2.13.** *If  $G$  is a graph of minimum degree  $\delta(G) \geq k + 1$ , then  $d_S^k(G) \geq k$ .*

**Theorem 2.14.** *Let  $k \geq 1$  be an integer, and let  $G$  be a  $(k + 1)$ -regular graph of order  $n$ . If  $n \not\equiv 0 \pmod{k + 2}$ , then  $d_S^k(G) = k$ .*

**Proof.** Let  $f$  be an arbitrary signed  $k$ -dominating function on  $G$ . If we define the sets  $P = \{v \in V(G) \mid f(v) = 1\}$  and  $M = \{v \in V(G) \mid f(v) = -1\}$ , then we firstly show that

$$|P| \geq \left\lceil \frac{n(k+1)}{k+2} \right\rceil. \quad (2.1)$$

Because of  $\sum_{x \in N[y]} f(x) \geq k$  for each vertex  $y \in V(G)$ , the  $(k + 1)$ -regularity of  $G$  implies that each vertex  $u \in P$  is adjacent to at most one vertex in  $M$  and each vertex  $v \in M$  is adjacent to exactly  $k + 1$  vertices in  $P$ . Therefore we obtain

$$|P| \geq |M|(k+1) = (n - |P|)(k+1),$$

and this leads to (2.1) immediately.

Now let  $\{f_1, f_2, \dots, f_d\}$  be a signed  $(k, k)$ -dominating family on  $G$  with  $d = d_S^k(G)$ . Since  $\sum_{i=1}^d f_i(u) \leq k$  for every vertex  $u \in V(G)$ , each of these sums contains at least  $\lceil (d-k)/2 \rceil$  summands of value -1. Using this and inequality (2.1), we see that the sum

$$\sum_{x \in V(G)} \sum_{i=1}^d f_i(x) = \sum_{i=1}^d \sum_{x \in V(G)} f_i(x) \quad (2.2)$$

contains at least  $n \lceil (d-k)/2 \rceil$  summands of value -1 and at least  $d \lceil n(k+1)/(k+2) \rceil$  summands of value 1. As the sum (2.2) consists of exactly  $dn$  summands, it follows that

$$n \left\lceil \frac{d-k}{2} \right\rceil + d \left\lceil \frac{n(k+1)}{k+2} \right\rceil \leq dn. \quad (2.3)$$

It follows from the hypothesis  $n \not\equiv 0 \pmod{k+2}$  that

$$\left\lceil \frac{n(k+1)}{k+2} \right\rceil > \frac{n(k+1)}{k+2},$$

and thus (2.3) leads to

$$\frac{n(d-k)}{2} + \frac{dn(k+1)}{k+2} < dn.$$

A simple calculation shows that this inequality implies  $d < k+2$  and so  $d \leq k+1$ . If we suppose that  $d = k+1$ , then we observe that  $d$  and  $k$  of different parity. Applying Theorem 2.8, we obtain the contradiction

$$k+1 = d \leq \frac{k-1}{k}(k+2) < k+1.$$

Therefore  $d \leq k$ , and Corollary 2.13 yields to the desired result  $d = k$ .  $\square$

On the one hand Theorem 2.14 demonstrates that the bound in Corollary 2.13 is sharp, on the other hand the following example shows that Theorem 2.14 is not valid in general when  $n \equiv 0 \pmod{k+2}$ .

Let  $v_1, v_2, \dots, v_{k+2}$  be the vertex set of the complete graph  $K_{k+2}$ . We define the functions  $f_i: V(G) \rightarrow \{-1, 1\}$  such that  $f_i(v_i) = -1$  and  $f_i(x) = 1$  for each vertex  $x \in V(G) \setminus \{v_i\}$  and each  $i \in \{1, 2, \dots, k+2\}$ . Then we observe that  $f_i$  is a signed  $k$ -dominating function on  $K_{k+2}$  for each  $i \in \{1, 2, \dots, k+2\}$  and  $\sum_{i=1}^{k+2} f_i(x) = k$  for each vertex  $x \in V(K_{k+2})$ . Therefore  $\{f_1, f_2, \dots, f_{k+2}\}$  is a signed  $(k, k)$ -dominating family on  $G$  and thus  $d_S^k(K_{k+2}) \geq k+2$ . Using Theorem 2.3, we obtain  $d_S^k(K_{k+2}) = k+2$ .

### 3. Signed $(k, k)$ -domatic number of fans, wheels and grids

Volkmann and Zelinka [6] have proved that  $d_S(G) = 1$  when  $G$  is a fan or a wheel of order  $n \geq 4$ . If a graph  $G$  has a vertex of degree 3, then Volkmann [5] showed that  $d_S(G) = 1$ . Therefore  $d_S(G) = 1$  for each grid. Using the results of Section 2, we now determine the signed  $(k, k)$ -domatic numbers of fans, wheels and grids for  $k \geq 2$ .

**Theorem 3.1.** *Let  $G$  be a fan of order  $n \geq 3$ . Then  $d_S^3(G) = 1$ ,  $d_S^2(G) = 1$  when  $3 \leq n \leq 5$  and  $d_S^2(G) = 2$  when  $n \geq 6$ .*

**Proof.** Since  $N[v]$  contains a vertex of degree at most 3 for every vertex  $v \in V(G)$ , it follows from Theorem 2.11 that  $d_S^3(G) = 1$ .

Let now  $x_1, x_2, \dots, x_n$  be the vertex set of the fan  $G$  such that  $x_1 x_2 \dots x_n x_1$  is a cycle of length  $n$  and  $x_n$  is adjacent to  $x_i$  for each  $i = 2, 3, \dots, n-2$ .

If  $n \leq 5$ , then  $N[v]$  contains a vertex of degree at most 2 for every vertex  $v \in V(G)$ , and Theorem 2.11 implies  $d_S^2(G) = 1$ .

If  $n \geq 6$ , then the functions  $f_i: V(G) \rightarrow \{-1, 1\}$  such that  $f_1(x) = 1$  for each  $x \in V(G)$  and  $f_2(x_3) = -1$  and  $f_2(x) = 1$  for each vertex  $x \in V(G) \setminus \{x_3\}$  are

signed 2-dominating functions on  $G$  such that  $f_1(x) + f_2(x) \leq 2$  for each vertex  $x \in V(G)$ . Thus  $d_S^2(G) \geq 2$ . In view of Corollary 2.6, we see that

$$d_S^2(G) \leq \frac{2}{3}(\delta(G) + 1) = 2,$$

and therefore  $d_S^2(G) = 2$ . □

**Theorem 3.2.** *Let  $G$  be a wheel of order  $n \geq 5$ . Then  $d_S^4(G) = d_S^3(G) = 1$ ,  $d_S^2(G) = 4$  when  $n - 1 \equiv 0 \pmod{3}$  and  $d_S^2(G) = 2$  when  $n - 1 \not\equiv 0 \pmod{3}$ .*

**Proof.** Since  $N[v]$  contains a vertex of degree at most 3 for every vertex  $v \in V(G)$ , it follows from Theorem 2.11 that  $d_S^4(G) = d_S^3(G) = 1$ .

Now let  $x_1, x_2, \dots, x_n$  be the vertex set of the wheel  $G$  such that  $x_1x_2 \dots x_{n-1}x_1$  is a cycle of length  $n - 1$  and  $x_n$  is adjacent to  $x_i$  for each  $i = 1, 2, \dots, n - 1$ . It follows from Theorem 2.3 that  $d_S^2(G) \leq 4$ . Since  $\delta(G) = 3$ , Corollary 2.13 implies that  $d_S^2(G) \geq 2$ . Let  $\{f_1, f_2, \dots, f_d\}$  be a signed  $(2, 2)$ -dominating family on  $G$  with  $d = d_S^2(G)$ . If  $d = 3$ , then Theorem 2.8 leads to the contradiction

$$3 = d \leq \frac{1}{2}(\delta(G) + 1) = 2.$$

Consequently,  $d = 2$  or  $d = 4$ . Assume that  $d = 4$ . Since  $f_1(x) + f_2(x) + f_3(x) + f_4(x) \leq 2$  for each vertex  $x \in V(G)$ , there exists at least one number  $j \in \{1, 2, 3, 4\}$  such that  $f_j(x) = -1$  for each  $x \in V(G)$ . Assume, without loss of generality, that  $f_1(x_n) = -1$ . Because of  $\sum_{x \in N[v]} f_1(x) \geq 2$  for each vertex  $v$ , we deduce that  $f_1(x_1) = f_1(x_2) = \dots = f_1(x_{n-1}) = 1$ . If we assume, without loss of generality, that  $f_2(x_1) = -1$ , then it follows that  $f_2(x_2) = f_2(x_3) = 1$ . If we assume next, without loss of generality, that  $f_3(x_2) = -1$ , then we observe that  $f_3(x_3) = f_3(x_4) = 1$  and therefore  $f_4(x_3) = -1$  and thus  $f_4(x_4) = f_4(x_5) = 1$ . This leads to  $f_2(x_4) = -1$  and so  $f_2(x_5) = f_2(x_6) = 1$ . Inductively, we see that  $f_2(x_i) = -1$  if and only if  $i \equiv 1 \pmod{3}$ ,  $f_3(x_i) = -1$  if and only if  $i \equiv 2 \pmod{3}$  and  $f_4(x_i) = -1$  if and only if  $i \equiv 0 \pmod{3}$ . This can be realized if and only if  $n - 1 \equiv 0 \pmod{3}$ , and this completes the proof. □

The *cartesian product*  $G = G_1 \times G_2$  of two vertex disjoint graphs  $G_1$  and  $G_2$  has  $V(G) = V(G_1) \times V(G_2)$  and two vertices  $(u_1, u_2)$  and  $(v_1, v_2)$  of  $G$  are adjacent if and only if either  $u_1 = v_1$  and  $u_2v_2 \in E(G_2)$  or  $u_2 = v_2$  and  $u_1v_1 \in E(G_1)$ . The cartesian product of two paths  $P_r = x_1x_2 \dots x_r$  and  $P_t = y_1y_2 \dots y_t$  is called a *grid*.

**Theorem 3.3.** *Let  $G = P_r \times P_t$  be a grid of order  $n = rt \geq 2$  such that  $r \leq t$ . Then*

(1) *If  $r = 1$ , then  $d_S^2(G) = 1$ .*

(2) *If  $r = 2$ , then  $d_S^3(G) = 1$ ,  $d_S^2(G) = 1$  when  $t \leq 4$  and  $d_S^2(G) = 2$  when  $t \geq 5$ .*

(3) If  $r \geq 3$ , then  $d_S^2(G) = 2$ .

(4) If  $3 \leq r \leq 4$ , then  $d_S^3(G) = 1$ .

(5) If  $r = 5$  and  $t = 5$ , then  $d_S^3(G) = 2$ .

(6) If  $r = 5$  and  $t \geq 6$  or  $r \geq 6$ , then  $d_S^3(G) = 3$ .

**Proof.** (1) Assume that  $r = 1$ . Then  $G$  is a path and it follows from Theorem 2.11 that  $d_S^2(G) = 1$ .

(2) Assume that  $r = 2$ . Then  $2 \leq d(v) \leq 3$  for every  $v \in V(G)$ , and hence Theorem 2.11 implies that  $d_S^3(G) = 1$ . If  $t \leq 4$ , then  $N[v]$  contains a vertex of degree at most 2 for every vertex  $v \in V(G)$ , and so  $d_S^2(G) = 1$ , by Theorem 2.11. If  $t \geq 5$ , then all vertices of  $N[(x_1, y_3)]$  are of degree 3, and thus it follows from Theorem 2.11 that  $d_S^2(G) \geq 2$ . Since  $\delta(G) = 2$ , we deduce from Corollary 2.6 that  $d_S^2(G) \leq 2$  and so  $d_S^2(G) = 2$ .

(3) Assume that  $r \geq 3$ . Then all vertices of  $N[(x_2, y_2)]$  are of degree at least 3, and thus it follows from Theorem 2.11 that  $d_S^2(G) \geq 2$ . Since  $\delta(G) = 2$ , we deduce from Corollary 2.6 that  $d_S^2(G) \leq 2$  and so  $d_S^2(G) = 2$ .

(4) Assume that  $3 \leq r \leq 4$ . This condition shows that  $N[v]$  contains a vertex of degree at most 3 for every vertex  $v \in V(G)$ , and so Theorem 2.11 implies that  $d_S^3(G) = 1$ .

(5) Assume that  $r = t = 5$ . Then all vertices of  $N[(x_3, y_3)]$  are of degree 4, and thus it follows from Theorem 2.11 that  $d_S^3(G) \geq 2$ . Since  $N[v]$  contains a vertex of degree at most 3 for every vertex  $v \in V(G) \setminus \{(x_3, y_3)\}$ , we deduce that  $f(v) = 1$  for every signed 3-dominating function  $f$  on  $G$  and every vertex  $v \neq (x_3, y_3)$ . This implies that  $d_S^3(G) \leq 2$  and thus  $d_S^3(G) = 2$ .

(6) Assume that  $r = 5$  and  $t \geq 6$  or  $r \geq 6$ . In view of Theorem 2.3, we have  $d_S^3(G) \leq 3$ . Define now the functions  $f_i: V(G) \rightarrow \{-1, 1\}$  such that  $f_1(v) = 1$  for each vertex  $v \in V(G)$ ,  $f_2((x_3, y_3)) = -1$  and  $f_2(v) = 1$  for each  $v \in V(G) \setminus \{(x_3, y_3)\}$  and  $f_3((x_3, y_4)) = -1$  and  $f_3(v) = 1$  for each  $v \in V(G) \setminus \{(x_3, y_4)\}$ . Then  $\{f_1, f_2, f_3\}$  is a family of signed 3-dominating functions on  $G$  such that  $f_1(v) + f_2(v) + f_3(v) \leq 3$  for each vertex  $v \in V(G)$ . Therefore  $d_S^3(G) = 3$ , and the proof is complete.  $\square$

## References

- [1] DUNBAR, J. E., HEDETNIEMI, S. T., HENNING, M. A., SLATER, P. J., Signed domination in graphs. *Graph Theory, Combinatorics, and Applications*, John Wiley and Sons, Inc. 1 (1995), 311–322.
- [2] HAYNES, T. W., HEDETNIEMI, S. T., SLATER, P. J., *Fundamentals of Domination in Graphs*, Marcel Dekker, Inc., New York (1998).
- [3] HAYNES, T. W., HEDETNIEMI, S. T., SLATER, P. J., editors, *Domination in Graphs, Advanced Topics*, Marcel Dekker, Inc., New York (1998).
- [4] RALL, D. F., A fractional version of domatic number, *Congr. Numer.* **74** (1990), 100–106.

- [5] VOLKMANN, L., Some remarks on the signed domatic number of graphs with small minimum degree, *Appl. Math. Letters* **22** (2009), 1166-1169.
- [6] VOLKMANN, L., ZELINKA, B., Signed domatic number of a graph, *Discrete Appl. Math.* **150** (2005), 261-267.
- [7] WANG, C. P., The signed  $k$ -domination numbers in graphs, *Ars Combin.*, to appear.

**S. M. Sheikholeslami**

Azərbaycan University of Tarbiat Moallem Tabriz, I.R. Iran

e-mail: [s.m.sheikholeslami@azaruniv.edu](mailto:s.m.sheikholeslami@azaruniv.edu)

**L. Volkmann**

RWTH-Aachen University, 52056 Aachen, Germany

e-mail: [volkm@math2.rwth-aachen.de](mailto:volkm@math2.rwth-aachen.de)





# Algebraic and transcendental solutions of some exponential equations

Jonathan Sondow<sup>a</sup>, Diego Marques<sup>b</sup>

<sup>a</sup>209 West 97th Street, New York, NY 10025 USA

<sup>b</sup>Departamento de Matemática, Universidade de Brasília, Brazil

*Submitted 16 December 2009; Accepted 30 May 2010*

## Abstract

We study algebraic and transcendental powers of positive real numbers, including solutions of each of the equations  $x^x = y$ ,  $x^y = y^x$ ,  $x^x = y^y$ ,  $x^y = y$ , and  $x^{x^y} = y$ . Applications to values of the iterated exponential functions are given. The main tools used are classical theorems of Hermite-Lindemann and Gelfond-Schneider, together with solutions of exponential Diophantine equations.

*Keywords:* Algebraic, irrational, transcendental, Gelfond-Schneider Theorem, Hermite-Lindemann Theorem, iterated exponential.

*MSC:* Primary 11J91, Secondary 11D61.

## 1. Introduction

Transcendental number theory began in 1844 with Liouville's explicit construction of the first transcendental numbers. In 1872 Hermite proved that  $e$  is transcendental, and in 1884 Lindemann extended Hermite's method to prove that  $\pi$  is also transcendental. In fact, Lindemann proved a more general result.

**Theorem 1.1** (Hermite-Lindemann). *The number  $e^\alpha$  is transcendental for any nonzero algebraic number  $\alpha$ .*

As a consequence, the numbers  $e^2$ ,  $e^{\sqrt{2}}$ , and  $e^i$  are transcendental, as are  $\log 2$  and  $\pi$ , since  $e^{\log 2} = 2$  and  $e^{\pi i} = -1$  are algebraic.

At the 1900 International Congress of Mathematicians in Paris, as the seventh in his famous list of 23 problems, Hilbert raised the question of the arithmetic

nature of the power  $\alpha^\beta$  of two algebraic numbers  $\alpha$  and  $\beta$ . In 1934, Gelfond and Schneider, independently, completely solved the problem (see [2, p. 9]).

**Theorem 1.2** (Gelfond-Schneider). *Assume  $\alpha$  and  $\beta$  are algebraic numbers, with  $\alpha \neq 0$  or 1, and  $\beta$  irrational. Then  $\alpha^\beta$  is transcendental.*

In particular,  $2^{\sqrt{2}}$ ,  $\sqrt{2}^{\sqrt{2}}$ , and  $e^\pi = i^{-2i}$  are all transcendental.

Since transcendental numbers are more “complicated” than algebraic irrational ones, we might think that the power of two transcendental numbers is also transcendental, like  $e^\pi$ . However, that is not always the case, as the last two examples for Theorem 1.1 show. In fact, there is no known classification of the power of two transcendental numbers analogous to the Gelfond-Schneider Theorem on the power of two algebraic numbers.

In this paper, we first explore a related question (a sort of converse to one raised by the second author in [14, Apêndice B]).

**Question 1.3.** Given positive real numbers  $X \neq 1$  and  $Y \neq 1$ , with  $X^Y$  algebraic, under which conditions will at least one of the numbers  $X, Y$  be transcendental?

Theorem 1.2 gives one such condition, namely,  $Y$  irrational. In Sections 2 and 3, we give other conditions for Question 1.3, in the case  $X^Y = Y^X$ . To do this, we use the Gelfond-Schneider Theorem to find algebraic and transcendental solutions to each of the exponential equations  $y = x^x$ ,  $y = x^{1/x}$ , and  $x^y = y^x$  with  $x \neq y$ .

In the Appendix, we study the arithmetic nature of values of three classical infinite power tower functions. We do this by using the Gelfond-Schneider and Hermite-Lindemann Theorems to classify solutions to the equations  $y = x^y$  and  $y = x^{x^y}$ .

A general reference is Knoebel’s Chauvenet Prize-winning article [12]. Consult its very extensive annotated bibliography for additional references and history.

**Notation.** We denote by  $\mathbb{N}$  the natural numbers,  $\mathbb{Z}$  the integers,  $\mathbb{Q}$  the rationals,  $\mathbb{R}$  the reals,  $\mathbb{A}$  the algebraic numbers, and  $\mathbb{T}$  the transcendental numbers. For any set  $S$  of complex numbers,  $S^+ := S \cap (0, \infty)$  denotes the subset of positive real numbers in  $S$ . The Fundamental Theorem of Arithmetic is abbreviated FTA.

## 2. The case $X = Y$ : algebraic numbers $T^T$ with $T$ transcendental

In this section, we give answers to Question 1.3 in the case  $X = Y$ . For this we need a result on the arithmetic nature of  $Q^Q$  when  $Q$  is rational.

**Lemma 2.1.** *If  $Q \in \mathbb{Q} \setminus \mathbb{Z}$ , then  $Q^Q$  is irrational.*

**Proof.** If  $Q > 0$ , write  $Q = a/b$ , where  $a, b \in \mathbb{N}$  and  $\gcd(a, b) = 1$ . Set  $a_1 = a^a$  and  $b_1 = b^a$ . Then  $\gcd(a_1, b_1) = 1$  and  $(a_1/b_1)^{1/b} = Q^Q \in \mathbb{Q}^+$ . Using the FTA,

we deduce that  $b_1^{1/b} \in \mathbb{N}$ . We must show that  $b = 1$ . Suppose on the contrary that some prime  $p \mid b$ . Let  $p^n$  be the largest power of  $p$  that divides  $b$ . Using  $b^{a/b} = b_1^{1/b} \in \mathbb{N}$  and the FTA again, we deduce that  $p^{na/b} \in \mathbb{N}$ . Hence  $b \mid na$ . Since  $\gcd(a, b) = 1$ , we get  $b \mid n$ . But then  $p^n \mid n$ , contradicting  $p^n > n$ . Therefore,  $b = 1$ .

If  $Q < 0$ , write  $Q = -a/b$ , where  $a, b \in \mathbb{N}$  and  $\gcd(a, b) = 1$ . If  $b$  is odd, then by the previous case,  $Q^Q = (-1)^a(a/b)^{-a/b} \notin \mathbb{Q}$ . If  $b$  is even, then  $a$  is odd and  $(-1)^{1/b} \notin \mathbb{R}$ ; hence  $Q^Q = (-1)^{1/b}(a/b)^{-a/b} \notin \mathbb{Q}$ . This completes the proof.  $\square$

As an application, using Theorem 1.2 we obtain that  $Q^{Q^Q}$  is transcendental if  $Q \in \mathbb{Q} \setminus \mathbb{Z}$ .

Consider now the equation  $x^x = y$ . When  $0 < y < e^{-1/e} = 0.69220\dots$ , there is no solution  $x > 0$ . If  $y = e^{-1/e}$ , then  $x = e^{-1} = 0.36787\dots$ . For  $y \in (e^{-1/e}, 1)$ , there are exactly two solutions  $x_0$  and  $x_1$ , with  $0 < x_0 < e^{-1} < x_1 < 1$ . (See Figure 1, which shows the case  $y = 1/\sqrt{2}$ ,  $x_0 = 1/4$ ,  $x_1 = 1/2$ .) Finally, given  $y \in [1, \infty)$ , there is a unique solution  $x \in [1, \infty)$ .

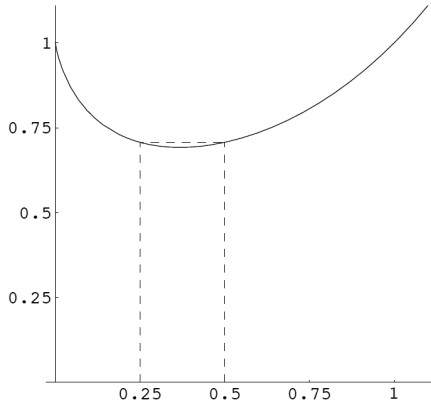


Figure 1:  $y = x^x$

Turning to the case  $X = Y$  of Question 1.3, we give two classes of algebraic numbers  $A$  such that  $T^T = A$  implies  $T$  is transcendental.

**Proposition 2.2.** *Given  $A \in [e^{-1/e}, \infty)$ , let  $T \in \mathbb{R}^+$  satisfy  $T^T = A$ . If either*

- (i)  $A^n \in \mathbb{A} \setminus \mathbb{Q}$  for all  $n \in \mathbb{N}$ , or
- (ii)  $A \in \mathbb{Q} \setminus \{n^n : n \in \mathbb{N}\}$ ,

*then  $T$  is transcendental. In particular,  $T \in \mathbb{T}$  if  $T^T \in \mathbb{Q} \cap (e^{-1/e}, 1)$ .*

**Proof.** (i) Suppose  $T \in \mathbb{A}$ . Since  $T > 0$  and  $T^T = A \in \mathbb{A}$ , Theorem 1.2 implies  $T \in \mathbb{Q}$ , say  $T = m/n$  with  $m, n \in \mathbb{N}$ . But then  $A^n = T^m \in \mathbb{Q}$ , contradicting (i). Therefore,  $T \in \mathbb{T}$ .

(ii) Since  $T^T = A \in \mathbb{Q} \setminus \{n^n : n \in \mathbb{N}\}$ , Lemma 2.1 implies  $T$  is irrational. Then Theorem 1.2 yields  $T \in \mathbb{T}$ , and the proposition follows.  $\square$

To illustrate case (i), take  $A = \sqrt{3} - 1 \in (e^{-1/e}, 1)$ . Using a computer algebra system, such as *Mathematica* with its `FindRoot` command, we solve the equation  $x^x = A$  with starting values of  $x$  near 0 and 1, obtaining the solutions  $T_0 := 0.15351\dots$  and  $T_1 := 0.63626\dots$ . Similarly, for case (ii), setting  $A = 2$  leads to the solution  $T_2 := 1.68644\dots$ . Then

$$T_0^{T_0} = T_1^{T_1} = \sqrt{3} - 1, \quad T_0 < e^{-1} < T_1; \quad T_2^{T_2} = 2; \quad T_0, T_1, T_2 \in \mathbb{T}.$$

**Problem 2.3.** In Proposition 2.2, replace the two sufficient conditions (i), (ii) with a necessary and sufficient condition that includes them.

We will return to the case  $X = Y$  of Question 1.3 at the end of the next section (see Corollary 3.8).

### 3. The case $X^Y = Y^X$ , with $X \neq Y$

In this section, we give answers to Question 1.3 by finding algebraic and transcendental solutions of the equation  $x^y = y^x$ , for positive real numbers  $x \neq y$ . (Compare Figure 2. Moulton [16] gives a graph for both positive and negative values of  $x$  and  $y$ , and discusses solutions in the complex numbers.)

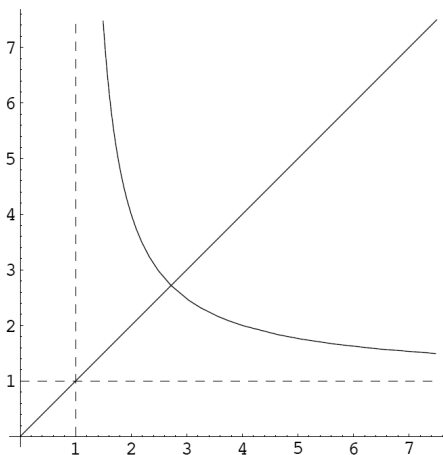


Figure 2:  $x^y = y^x$

Consider now Question 1.3 in the case  $X^Y = Y^X = A \in \mathbb{A}$ , with  $X \neq Y$ . We give a condition on  $A$  which guarantees that at least one of  $X, Y$  is transcendental.

**Proposition 3.1.** *Assume that*

$$T, R \in \mathbb{R}^+, \quad A := T^R = R^T, \quad T \neq R. \quad (3.1)$$

*If  $A^n \in \mathbb{A} \setminus \mathbb{Q}$  for all  $n \in \mathbb{N}$ , then at least one of the numbers  $T, R$ , say  $T$ , is transcendental.*

**Proof.** Suppose on the contrary that  $T, R \in \mathbb{A}$ . Since  $T^R = R^T = A \in \mathbb{A}$  and (3.1) implies  $T, R \neq 0$  or  $1$ , Theorem 1.2 yields  $T, R \in \mathbb{Q}$ , say  $T = a/b$  and  $R = m/n$ , where  $a, b, m, n \in \mathbb{N}$ . But then  $A^n = (a/b)^m \in \mathbb{Q}$ , contradicting the hypothesis. Therefore,  $\{T, R\} \cap \mathbb{T} \neq \emptyset$ .  $\square$

In order to give an example of Proposition 3.1, we need the following classical result, which is related to a problem posed in 1728 by D. Bernoulli [4, p. 262]. (In [12], see Sections 1 and 3 and the notes to the bibliography.)

**Lemma 3.2.** *Given  $z \in \mathbb{R}^+$ , there exist  $x$  and  $y$  such that*

$$x^y = y^x = z, \quad 0 < x < y,$$

*if and only if  $z > e^e = 15.15426\dots$ . In that case,  $1 < x < e < y$  and  $x, y$  are given parametrically by*

$$x = x(t) := \left(1 + \frac{1}{t}\right)^t, \quad y = y(t) := \left(1 + \frac{1}{t}\right)^{t+1} \quad (3.2)$$

*for  $t > 0$ . Moreover,  $x(t)^{y(t)}$  is decreasing, and any one of the numbers  $x \in (1, e)$ ,  $y \in (e, \infty)$ ,  $z \in (e^e, \infty)$ , and  $t \in (0, \infty)$  determines the other three uniquely.*

**Proof.** Given  $x, y \in \mathbb{R}^+$  with  $x < y$ , denote the slope of the line from the origin to the point  $(x, y)$  by  $s := y/x$ . Then  $s > 1$ , and  $y = sx$  gives the equivalences

$$\begin{aligned} x^y = y^x &\iff x^{sx} = (sx)^x \iff x^s = sx \\ &\iff x = x_1(s) := s^{1/(s-1)} \iff y = y_1(s) := s^{s/(s-1)}. \end{aligned}$$

The substitution  $s = 1 + t^{-1}$  then produces (3.2), implying  $1 < x < e < y$ . Using L'Hopital's rule, we get

$$\lim_{t \rightarrow 0^+} x(t) = 1, \quad \lim_{t \rightarrow 0^+} y(t) = \infty \implies \lim_{t \rightarrow 0^+} y(t)^{x(t)} = \infty.$$

By calculus,  $x(t)$  is increasing,  $y(t)$  is decreasing, and  $y(t)^{x(t)} \rightarrow e^e$  as  $t \rightarrow \infty$  (see Figure 3). Anderson [1, Lemma 4.3] proves that the function  $y_1(s)^{-x_1(s)}$  is decreasing on the interval  $1 < s < \infty$ , and we infer that  $y(t)^{x(t)}$  is decreasing on  $0 < t < \infty$  (see Figure 4). The lemma follows.  $\square$

For instance, taking  $t = 1$  in (3.2) leads to  $2^4 = 4^2 = 16$ . To parameterize the part of the curve  $x^y = y^x$  with  $x > y > 0$ , replace  $t$  with  $-t - 1$  in (3.2) (or replace  $s$  with  $1/s$  in the parameterization  $x = x_1(s)$ ,  $y = y_1(s)$ , which is due to Goldbach [11, pp. 280-281]). For example, setting  $t = -2$  in (3.2) yields  $(x, y) = (4, 2)$ .

Euler [8, pp. 293-295] described a different way to find solutions of  $x^y = y^x$  with  $0 < x < y$ . Namely, the equivalence

$$x^y = y^x \iff x^{1/x} = y^{1/y}$$

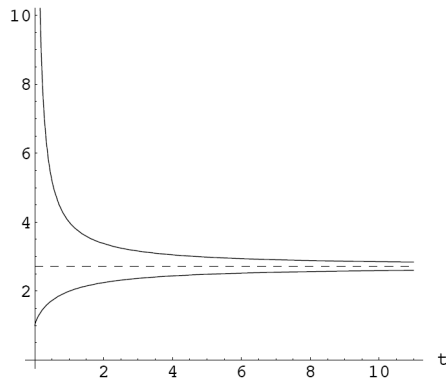


Figure 3: The graphs of  $x(t)$  (bottom) and  $y(t)$

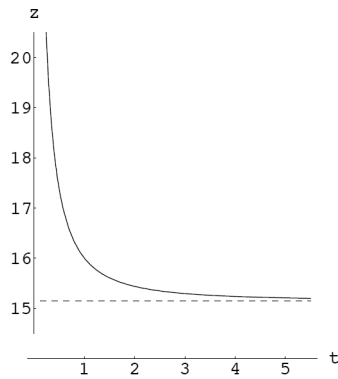


Figure 4:  $z = x(t)y(t) = y(t)^{x(t)}$

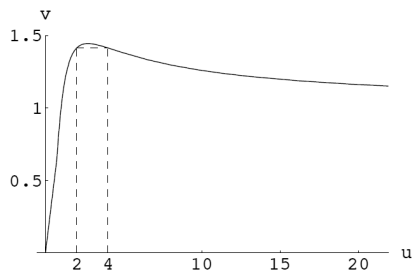


Figure 5:  $v = g(u) = u^{1/u}$

shows that a solution is determined by equal values of the function  $g(u) = u^{1/u}$  at  $u = x$  and  $u = y$ . (Figure 5 exhibits the case  $x = 2, y = 4$ .) From the properties of  $g(u)$ , including its maximum at  $u = e$  and the bound  $g(u) > 1$  for  $u \in (1, \infty)$ ,

we see again that  $1 < x < e < y$ .

We can now give an example for Proposition 3.1.

**Example 3.3.** Set  $A = 14 + \sqrt{2}$ . Since  $A > e^e$ , the equation  $x(t)^{y(t)} = A$  has a (unique) solution  $t = t_1 > 0$ . (Computing  $t_1$ , we find that  $x(t_1) = 2.26748\dots$  and  $y(t_1) = 3.34112\dots$ ) Then  $(T, R) := (x(t_1), y(t_1))$  or  $(y(t_1), x(t_1))$  satisfies

$$T^R = R^T = 14 + \sqrt{2}, \quad T \neq R, \quad T \in \mathbb{T}.$$

In the next proposition, we characterize the algebraic and rational solutions of  $x^y = y^x$  with  $0 < x < y$ . (Part (i) is due to Mahler and Breusch [13]. For other references, as well as all rational solutions to the more general equation  $x^y = y^{mx}$ , where  $m \in \mathbb{N}$ , see Bennett and Reznick [3].)

**Proposition 3.4.** *Assume  $0 < A_1 < A_2$ . Define  $x(t)$  and  $y(t)$  as in (3.2).*

(i) *Then  $A_1^{A_2} = A_2^{A_1}$  and  $A_1, A_2 \in \mathbb{A}$  if and only if  $A_1 = x(t)$  and  $A_2 = y(t)$ , with  $t \in \mathbb{Q}^+$ .*

(ii) *In that case, if  $t \in \mathbb{N}$ , then  $A_1^{A_2} = A_2^{A_1} \in \mathbb{A}$  and  $A_1, A_2 \in \mathbb{Q}$ , while if  $t \notin \mathbb{N}$ , then  $A_1^{A_2} = A_2^{A_1} \in \mathbb{T}$  and  $A_1, A_2 \notin \mathbb{Q}$ .*

**Proof.** (i) By Lemma 3.2, it suffices to prove that  $t \in \mathbb{Q}$  if  $x(t), y(t) \in \mathbb{A}$ . Formulas (3.2) show that  $x(t)^{(t+1)/t} = y(t)$ . As  $x(t) \neq 0$  or 1, Theorem 1.2 implies  $t \notin \mathbb{A} \setminus \mathbb{Q}$ . From (3.2) we also see that  $y(t)/x(t) = 1 + t^{-1}$ , and hence  $t \in \mathbb{A}$ . Therefore,  $t \in \mathbb{Q}$ . (ii) It suffices to show that if  $A_1^{A_2} = A_2^{A_1} \in \mathbb{A}$ , where  $A_1 = x(a/b)$  and  $A_2 = y(a/b)$ , with  $a, b \in \mathbb{N}$  and  $\gcd(a, b) = 1$ , then  $b = 1$ . Theorem 1.2 implies  $A_1, A_2 \in \mathbb{Q}$ . It follows, using (3.2) and the FTA, that  $a + b$  and  $a$  are  $b$ th powers, say  $a + b = m^b$  and  $a = n^b$ , where  $m, n \in \mathbb{N}$ . Then  $d := m - n \geq 1$  and  $b = (n + d)^b - n^b = bn^{b-1}d + \dots + d^b$ . Hence  $b = 1$ .  $\square$

For example, taking  $t = 2$  and  $1/2$  yields

$$(9/4)^{27/8} = (27/8)^{9/4} \in \mathbb{A}, \quad \sqrt{3}^{\sqrt{27}} = \sqrt{27}^{\sqrt{3}} \in \mathbb{T}.$$

Here is another sufficient condition for Question 1.3 in the case  $X^Y = Y^X$  with  $X \neq Y$ .

**Corollary 3.5.** *Let  $T, R \in \mathbb{R}^+$  satisfy  $T^R = R^T = N \in \mathbb{N}$  and  $T \neq R$ . If  $N \neq 16$ , then at least one of the numbers  $T, R$ , say  $T$ , is transcendental.*

**Proof.** If on the contrary  $T, R \in \mathbb{A}$ , then Proposition 3.4 implies  $(T, R) = (x(n), y(n))$  or  $(y(n), x(n))$ , for some  $n \in \mathbb{N}$ . Thus  $x(n)^{y(n)} = N \neq 16$ . But a glance at Figure 4 (or at Lemma 3.2) shows that is impossible.  $\square$

For instance, the equation  $x(t)^{y(t)} = 17$  has a (unique) solution  $t = t_1 > 0$  (computing  $t_1$ , we get  $(x(t_1), y(t_1)) = (1.78381\dots, 4.89536\dots)$ ), and for  $(T, R) = (x(t_1), y(t_1))$  or  $(y(t_1), x(t_1))$  we have

$$T^R = R^T = 17, \quad T \neq R, \quad T \in \mathbb{T}.$$

We make the following prediction.

**Conjecture 3.6.** In Proposition 3.1 and Corollary 3.5 a stronger conclusion holds, namely, that both  $T$  and  $R$  are transcendental.

We can give a conditional proof of Conjecture 3.6, assuming a conjecture of Schanuel [2, p. 120]. Namely, in view of Proposition 3.1 and Corollary 3.5, Conjecture 3.6 is an immediate consequence of the following conditional result [15, Theorem 3].

**Theorem 3.7.** *Assume Schanuel's conjecture and let  $z$  and  $w$  be complex numbers, not 0 or 1. If  $z^w$  and  $w^z$  are algebraic, then  $z$  and  $w$  are either both rational or both transcendental.*

We now give an application of Proposition 3.4 to Question 1.3 in the case  $X = Y$ .

**Corollary 3.8.** *Let  $T, Q \in (0, 1)$  satisfy  $T^T = Q^Q$  and  $T \neq Q \in \mathbb{Q}$ . Then  $T \in \mathbb{T}$  if and only if  $x(n) \neq 1/Q \neq y(n)$  for all  $n \in \mathbb{N}$ . In particular,  $T \in \mathbb{T}$  if  $1/Q \in \mathbb{N} \setminus \{1, 2, 4\}$ .*

**Proof.** It is easy to see the equivalences

$$T^T = Q^Q \iff (1/T)^{1/Q} = (1/Q)^{1/T}$$

and, as  $\mathbb{A}$  is a field,  $T \in \mathbb{T} \iff 1/T \in \mathbb{T}$ . Using Proposition 3.4, the “if and only if” statement follows. Since  $n \in \mathbb{N}$  and  $1/Q \in \mathbb{N} \setminus \{2, 4\}$  imply  $x(n) \neq 1/Q \neq y(n)$ , the final statement also holds.  $\square$

For example, taking  $Q = 4/9 = 1/x(2)$  leads to  $(4/9)^{4/9} = (8/27)^{8/27} \in \mathbb{A}$ , while  $Q = 1/3$  and  $2/3$  give

$$(1/3)^{1/3} = T_1^{T_1}, \quad T_1 \in \mathbb{T}; \quad (2/3)^{2/3} = T_2^{T_2}, \quad T_2 \in \mathbb{T}.$$

Here  $T_1 = 0.40354\dots$  and  $T_2 = 0.13497\dots$  can be calculated by computing solutions to the equations  $x^x = (1/3)^{1/3}$  and  $x^x = (2/3)^{2/3}$ , using starting values of  $x$  in the intervals  $(e^{-1}, 1)$  and  $(0, e^{-1})$ , respectively.

## 4. Appendix: The infinite power tower functions

We use the Gelfond-Schneider and Hermite-Lindemann Theorems to find algebraic, irrational, and transcendental values of three classical functions, whose analytic properties were studied by Euler [9], Eisenstein [7], and many others.

**Definition 4.1.** The *infinite power tower* (or *iterated exponential*) function  $h(x)$  is the limit of the sequence of *finite power towers* (or *hyperpowers*)  $x, x^x, x^{x^x}, \dots$ . For  $x > 0$ , the sequence converges if and only if (see [1], Cho and Park [5], De Villiers and Robinson [6], Finch [10, p. 448], and [12])

$$0.06598\dots = e^{-e} \leq x \leq e^{1/e} = 1.44466\dots,$$



and in that case we write

$$h(x) = x^{x^{x^{\dots}}}$$

By substitution, we see that  $h$  satisfies the identity

$$x^{h(x)} = h(x). \tag{4.1}$$

Thus  $y = h(x)$  is a solution of the equations  $x^y = y$  and, hence,  $x = y^{1/y}$ . In other words,  $g(h(x)) = x$ , where  $g(u) = u^{1/u}$  for  $u > 0$ . Replacing  $x$  with  $g(x)$ , we get  $g(h(g(x))) = g(x)$  if  $g(x) \in [e^{-e}, e^{1/e}]$ . Since  $g$  is one-to-one on  $(0, e]$ , and since  $h$  is bounded above by  $e$  (see [12] for a proof) and  $g([e, \infty)) \subset (1, e^{1/e}]$  (see Figure 5), it follows that

$$h(g(x)) = x \quad (e^{-1} \leq x \leq e), \quad h(g(x)) < x \quad (e < x < \infty). \tag{4.2}$$

Therefore,  $h$  is a partial inverse of  $g$ , and is a bijection (see Figure 6)

$$h : [e^{-e}, e^{1/e}] \rightarrow [e^{-1}, e].$$

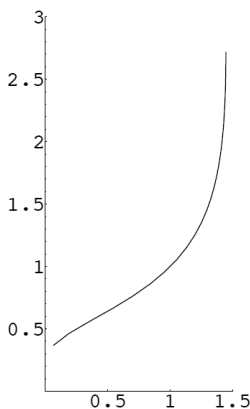


Figure 6:  $y = h(x) = x^{x^{x^{\dots}}}$

For example, taking  $x = 1/2$  and 2 in (4.2) gives

$$(1/4)^{(1/4)^{(1/4)^{\dots}}} = \frac{1}{2}, \quad \sqrt{2}^{\sqrt{2}^{\sqrt{2}^{\dots}}} = 2, \tag{4.3}$$

while choosing  $x = 3$  yields

$$\sqrt[3]{3}^{\sqrt[3]{3}^{\sqrt[3]{3}^{\dots}}} < 3.$$

Recall that the Hermite-Lindemann Theorem says that if  $A$  is any nonzero algebraic number, then  $e^A$  is transcendental. We claim that *if in addition  $A$  lies*

in the interval  $(-e, e^{-1})$ , then  $h(e^A)$  is also transcendental. To see this, set  $x = e^A$  and  $y = h(x)$ . Then (4.1) yields  $e^{Ay} = y$ , and Theorem 1.1 implies  $y \in \mathbb{T}$ , proving the claim. For instance,

$$\sqrt[3]{e^{\sqrt[3]{e^{\sqrt[3]{e^{\dots}}}}} = 1.85718\dots \in \mathbb{T}, \quad (4.4)$$

where the value of  $h(\sqrt[3]{e})$  can be obtained by computing a solution to  $x^{1/x} = \sqrt[3]{e}$ , using a starting value of  $x$  between  $e^{-1}$  and  $e$ .

Here is an application of Proposition 2.2.

**Corollary 4.2.** *Given  $A \in [e^{-e}, e^{1/e}]$ , if either  $A^n \in \mathbb{A} \setminus \mathbb{Q}$  for all  $n \in \mathbb{N}$ , or  $A \in \mathbb{Q} \setminus \{1/4, 1\}$ , then*

$$A^{A^{A^{\dots}}} \in \mathbb{T}. \quad (4.5)$$

**Proof.** From (4.1), we have  $A_1 := 1/A = (1/h(A))^{1/h(A)}$ . The hypotheses imply that  $A_1$  satisfies condition (i) or (ii) of Proposition 2.2. Thus  $1/h(A)$  and, hence,  $h(A)$  are transcendental.  $\square$

For example,  $h((\sqrt{2} + 1)/2) = 1.27005\dots \in \mathbb{T}$  and

$$(1/2)^{(1/2)^{(1/2)^{\dots}}} = 0.64118\dots \in \mathbb{T}.$$

It is easy to give an infinite power tower analog to the examples in Section 2 of powers  $T^T \in \mathbb{A}$  with  $T \in \mathbb{T}$ . Indeed, Theorem 1.2 and relation (4.1) imply that if  $A \in (\mathbb{A} \setminus \mathbb{Q}) \cap (e^{-1}, e)$ , then

$$T := 1/A^A \in \mathbb{T}, \quad T^{T^{T^{\dots}}} = 1/A \in \mathbb{A}. \quad (4.6)$$

Notice that (4.3), (4.4), (4.5), (4.6) represent the four possible cases  $(x, h(x)) \in \mathbb{A} \times \mathbb{A}, \mathbb{T} \times \mathbb{T}, \mathbb{A} \times \mathbb{T}, \mathbb{T} \times \mathbb{A}$ , respectively.

We now define two functions each of which extends  $h$  to a larger domain.

**Definition 4.3.** The *odd infinite power tower function*  $h_o(x)$  is the limit of the sequence of finite power towers of odd height:

$$x, x^{x^x}, x^{x^{x^{x^x}}}, \dots \longrightarrow h_o(x).$$

Similarly, the *even infinite power tower function*  $h_e(x)$  is defined as the limit of the sequence of finite power towers of even height:

$$x^x, x^{x^{x^x}}, x^{x^{x^{x^{x^x}}}}, \dots \longrightarrow h_e(x).$$

Both sequences converge on the interval  $0 < x \leq e^{1/e}$  (for a proof, see [1] or [12]).

It follows from Definition 4.3 that  $h_o$  and  $h_e$  satisfy the identities

$$x^{x^{h_o(x)}} = h_o(x), \quad x^{x^{h_e(x)}} = h_e(x) \tag{4.7}$$

and the relations

$$x^{h_e(x)} = h_o(x), \quad x^{h_o(x)} = h_e(x) \tag{4.8}$$

on  $(0, e^{1/e}]$ . From (4.7), we see that  $y = h_o(x)$  and  $y = h_e(x)$  are solutions of the equation  $y = x^{x^y}$ . So is  $y = h(x)$ , since  $y = x^y$  implies  $y = x^{x^y}$ .

It is proved in [1] and [12] that on the subinterval  $[e^{-e}, e^{1/e}] \subset (0, e^{1/e}]$  the three infinite power tower functions  $h, h_o, h_e$  are all defined and are equal, but on the subinterval  $(0, e^{-e})$  only  $h_o$  and  $h_e$  are defined, and they satisfy the inequality

$$h_o(x) < h_e(x) \quad (0 < x < e^{-e}) \tag{4.9}$$

and are surjections (see Figure 7)

$$h_o : (0, e^{1/e}] \rightarrow (0, e], \quad h_e : (0, e^{1/e}] \rightarrow [e^{-1}, e].$$

In order to give an analog for  $h_o$  and  $h_e$  to Corollary 4.2 on  $h$ , we require a lemma.

**Lemma 4.4.** *Assume  $Q, Q_1 \in \mathbb{Q}^+$ . Then*

$$Q^{Q^{Q_1}} = Q_1 \tag{4.10}$$

*if and only if  $(Q, Q_1)$  is equal to either  $(1/16, 1/2)$  or  $(1/16, 1/4)$  or  $(1/n^n, 1/n)$ , for some  $n \in \mathbb{N}$ .*

**Proof.** The “if” part is easily verified. To prove the “only if” part, note first that (4.10) and Theorem 1.2 imply  $Q^{Q^{Q_1}} \in \mathbb{Q}$ . Then, writing  $Q = a/b$  and  $Q_1 = m/n$ , where  $a, b, m, n \in \mathbb{N}$  and  $\gcd(a, b) = \gcd(m, n) = 1$ , the FTA implies  $a = a_1^n$  and  $b = b_1^n$ , for some  $a_1, b_1 \in \mathbb{N}$ . From (4.10) we infer that  $m^{b_1^m} = a_1^{na_1^m}$  and  $n^{b_1^m} = b_1^{na_1^m}$ .

We show that  $m = 1$ . If  $m \neq 1$ , then some prime  $p \mid m$ , and hence  $p \mid a_1$ . Write  $m = m'p^r$  and  $a_1 = a_2p^s$ , where  $r, s \in \mathbb{N}$  and  $\gcd(m', p) = \gcd(a_2, p) = 1$ . Substituting into  $m^{b_1^m} = a_1^{na_1^m}$ , we deduce that  $rb_1^m = sna_1^m$ . Since  $\gcd(a_1, b_1) = 1$ , we have  $a_1^m \mid r$ . But  $a_1^m = a_1^{m'p^r} > r$ , a contradiction. Therefore,  $m = 1$ .

It follows that  $a_1 = 1$ , and hence  $n^{b_1} = b_1^n$ . Proposition 3.4 then implies that  $(n, b_1) = (2, 4)$  or  $(n, b_1) = (4, 2)$  or  $n = b_1$ . The lemma follows.  $\square$

**Proposition 4.5.** *We have  $h_o(1/16) = 1/4$  and  $h_e(1/16) = 1/2$ . On the other hand, if  $Q \in \mathbb{Q} \cap (0, e^{-e}]$  but  $Q \neq 1/16$ , then  $h_o(Q)$  and  $h_e(Q)$  are both irrational, and at least one of them is transcendental.*

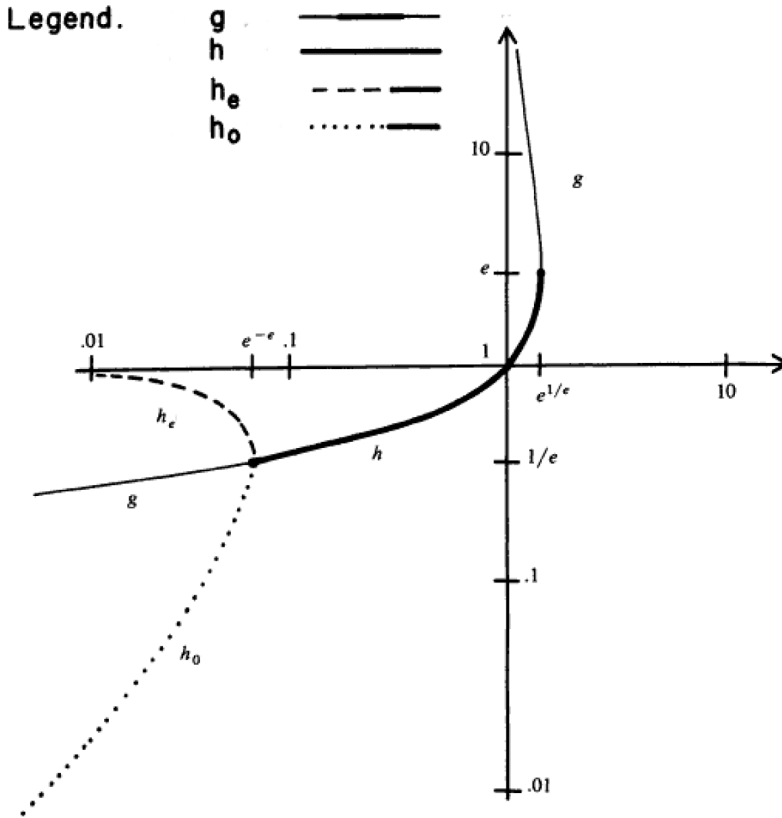


Figure 7: (from [12])  $x = g(y)$ ,  $y = h(x)$ ,  $y = h_e(x)$ ,  $y = h_o(x)$

**Proof.** Since  $1/16 < e^{-e}$ , the equation

$$(1/16)^{(1/16)^y} = y$$

has exactly three solutions (see [12] and Figure 7), namely,  $y = 1/4$ ,  $1/2$ , and  $y_0$ , say, where  $1/4 < y_0 < 1/2$ . By (4.7) and (4.9), two of the solutions are  $y = h_o(1/16)$  and  $h_e(1/16)$ . In view of (4.9), either  $h_o(1/16) = 1/4$  or  $h_o(1/16) = y_0$ . But the latter would imply that  $h_e(1/16) = 1/2$ , which leads by (4.8) to  $y_0 = (1/16)^{1/2} = 1/4$ , a contradiction. Therefore  $h_o(1/16) = 1/4$ . Then (4.8) implies  $h_e(1/16) = (1/16)^{1/4} = 1/2$ , proving the first statement.

To prove the second, suppose  $Q_1 := h_o(Q)$  is rational. Then (4.7) and Lemma 4.4 imply  $(Q, Q_1) = (1/n^n, 1/n)$ , for some  $n \in \mathbb{N}$ . Hence  $Q^{Q_1} = Q_1$ . But from (4.8) and (4.9) we see that  $Q^{h_o(Q)} = h_e(Q) > h_o(Q)$ , so that  $Q^{Q_1} > Q_1$ , a contradiction. Therefore,  $h_o(Q)$  is irrational. The proof that  $h_e(Q) \notin \mathbb{Q}$  is similar. Now (4.8) and Theorem 1.2 imply that  $\{h_o(Q), h_e(Q)\} \cap \mathbb{T} \neq \emptyset$ .  $\square$

For example, the numbers  $h_o(1/17) = 0.20427\dots$  and  $h_e(1/17) = 0.56059\dots$  are both irrational, and at least one is transcendental. The values were computed directly from Definition 4.3.

**Conjecture 4.6.** In the second part of Proposition 4.5 a stronger conclusion holds, namely, that both  $h_o(Q)$  and  $h_e(Q)$  are transcendental.

As with Conjecture 3.6, we can give a conditional proof of Conjecture 4.6. Namely, in view of Proposition 4.5 and the identities (4.7), Conjecture 4.6 is a special case of the following conditional result [15, Theorem 4].

**Theorem 4.7.** Assume Schanuel's conjecture and let  $\alpha \neq 0$  and  $z$  be complex numbers, with  $\alpha$  algebraic and  $z$  irrational. If  $\alpha^{\alpha^z} = z$ , then  $z$  is transcendental.

Some of our results on the arithmetic nature of values of  $h$ ,  $h_o$ , and  $h_e$  can be extended to other positive solutions to the equations  $y = x^y$  and  $y = x^{x^y}$ . As with the rest of the paper, an extension to negative and complex solutions is an open problem (compare [12, Section 4] and [16]).

**Acknowledgments.** We are grateful to Florian Luca, Wadim Zudilin, and the anonymous referee for valuable comments.

## References

- [1] ANDERSON, J., Iterated exponentials, *Amer. Math. Monthly* **111** (2004) 668–679.
- [2] BAKER, A., *Transcendental Number Theory*, Cambridge Mathematical Library, Cambridge University Press, Cambridge, 1990.
- [3] BENNETT, M. A., REZNICK, B., Positive rational solutions to  $x^y = y^{mx}$ : a number-theoretic excursion, *Amer. Math. Monthly* **111** (2004) 13–21.
- [4] BERNOULLI, D., Letter to Goldbach, June 29, 1728, *Correspond. Math. Phys.*, vol. 2, P. H. von Fuss, ed., Imperial Academy of Sciences, St. Petersburg, 1843.
- [5] CHO, Y., PARK, K., Inverse functions of  $y = x^{1/x}$ , *Amer. Math. Monthly* **108** (2001) 963–967.
- [6] DE VILLIERS, J. M., ROBINSON, P. N., The interval of convergence and limiting functions of a hyperpower sequence, *Amer. Math. Monthly* **93** (1986) 13–23.
- [7] EISENSTEIN, G., Entwicklung von  $\alpha^{\alpha^{\alpha^{\dots}}}$ , *J. Reine Angew. Math.* **28** (1844) 49–52.
- [8] EULER, L., *Introductio in analysin infinitorum*, vol. 2, 1748, reprinted by Culture et Civilization, Brussels, 1967.
- [9] EULER, L., De formulis exponentialibus replicatis, *Acta Academiae Scientiarum Petropolitanae* **1** (1778) 38–60; also in *Opera Omnia, Series Prima*, vol. 15, G. Faber, ed., Teubner, Leipzig, 1927, pp. 268–297.
- [10] FINCH, S. R., *Mathematical Constants*, Encyclopedia of Mathematics and its Applications, 94, Cambridge University Press, Cambridge, 2003.

- [11] GOLDBACH, C., Reply to Daniel Bernoulli, *Correspond. Math. Phys.*, vol. 2, P. H. von Fuss, ed., Imperial Academy of Sciences, St. Petersburg, 1843.
- [12] KNOEBEL, R. A., Exponentials reiterated, *Amer. Math. Monthly* **88** (1981) 235–252.
- [13] MAHLER, K., BREUSCH, R., Problem 5101: Solutions of an old equation, *Amer. Math. Monthly* **70** (1963) 571 (proposal); **71** (1964) 564 (solution).
- [14] MARQUES, D., Alguns resultados que geram números transcendentos, Master's thesis, Universidade Federal do Ceará, Brazil, 2007.
- [15] MARQUES, D., SONDOW, J., Schanuel's conjecture and algebraic powers  $z^w$  and  $w^z$  with  $z$  and  $w$  transcendental, *East-West J. Math.*, 12 (2010) 75–84; available at <http://arxiv.org/abs/1010.6216>.
- [16] MOULTON, E. J., The real function defined by  $x^y = y^x$ , *Amer. Math. Monthly* **23** (1916) 233–237.

**Jonathan Sondow**

209 West 97th Street, New York, NY 10025 USA

e-mail: [jsondow@alumni.princeton.edu](mailto:jsondow@alumni.princeton.edu)**Diego Marques**

Departamento de Matemática, Universidade de Brasília, DF, Brazil

e-mail: [diego@mat.unb.br](mailto:diego@mat.unb.br)

# Geometric properties and constrained modification of trigonometric spline curves of Han

Ede M. Troll, Miklós Hoffmann\*

Institute of Mathematics and Computer Science  
Eszterházy Károly College

*Submitted 25 March 2010; Accepted 10 August 2010*

*Dedicated to professor Béla Pelle on his 80<sup>th</sup> birthday*

## Abstract

New types of quadratic and cubic trigonometric polynomial curves have been introduced in [2] and [3]. These trigonometric curves have a global shape parameter  $\lambda$ . In this paper the geometric effect of this shape parameter on the curves is discussed. We prove that this effect is linear. Moreover we show that the quadratic curve can interpolate the control points at  $\lambda = \sqrt{2}$ . Constrained modification of these curves is also studied. A curve passing through a given point is computed by an algorithm which includes numerical computations. These issues are generalized for surfaces with two shape parameters. We show that a point of the surface can move along a hyperbolic paraboloid.

*Keywords:* trigonometric curve, spline curve, constrained modification

*MSC:* 68U07, 65D17

## 1. Introduction

In Computer Aided Geometric Design the most prevalently used curves are B-Spline and NURBS curves. Besides the quadratic and the cubic B-Spline and NURBS curves the trigonometric spline curves are another way to define curves above a new function space. Among the first ones, C-Bézier and uniform CB-spline

---

\*The second author was supported by the János Bolyai Fellowship of the Hungarian Academy of Science.

curves are defined by means of the basis  $\{\sin t, \cos t, t, 1\}$ , which was generalized to  $\{\sin t, \cos t, t^{k-3}, t^{k-4}, \dots, t, 1\}$  (cf. [14, 15, 16, 1]). Wang et al. introduced NUAT B-spline curves ([13]) that are the non-uniform generalizations of CB-spline curves. The other basic type is the HB-spline curve, the basis of which is  $\{\sinh t, \cosh t, t, 1\}$ , and  $\{\sinh t, \cosh t, t^{k-3}, t^{k-4}, \dots, t, 1\}$  in higher order ([12, 9]). Li and Wang developed its non-uniform generalization ([8]). Trigonometric curves can produce several kinds of classical important curves explicitly due to their trigonometric basis functions, including circle and circular cylinder [14], ellipse [16], surfaces of revolution [11], cycloid [10], helix [12], hyperbola and catenary [9]. Two recently defined trigonometric curves are the quadratic [2] and cubic [3] trigonometric curves of Xuli Han. The aim of this paper is to discuss the geometric properties of these curves, including the effect of their shape parameters and the possibility of constrained modification of the curves by these parameters. The method of our study is similar to the papers discussing geometric properties and modification of other types of spline curves. Such research has been done e.g. for C-Bézier curves [6], FB-spline curves [4], GB-spline curves [7] and another quartic curve of Han [5].

In this paper the definition of quadratic and cubic trigonometric polynomial curves are presented in Section 2. The geometric effects of the shape parameter as well as its application for constrained modification are discussed for quadratic curve in Section 3. These results are extended for the cubic case in Section 4, and for quadratic trigonometric surfaces in Section 5.

## 2. Definition of the basis functions

The construction of the basis functions of the quadratic trigonometric curve is the following [2].

**Definition 2.1.** Given knots  $u_0 < u_1 < \dots < u_{n+3}$ , let

$$\begin{aligned} \Delta u_i &:= u_{i+1} - u_i, & t_i(u) &:= \frac{\pi}{2} \left( \frac{u - u_i}{\Delta u_i} \right), \\ \alpha_i &:= \frac{\Delta u_i}{\Delta u_{i-1} + \Delta u_i}, & c(t) &:= (1 - \sin(t))(1 - \lambda \sin(t)), \\ \beta_i &:= \frac{\Delta u_i}{\Delta u_i - \Delta u_{i+1}}, & d(t) &:= (1 - \cos(t))(1 - \lambda \cos(t)), \end{aligned}$$

where  $-1 < \lambda < 1$ . Then the associated trigonometric polynomial basis functions are defined to be the following functions:

$$b_i(u) = \begin{cases} \beta_i d(t_i), & u \in [u_i, u_{i+1}), \\ 1 - \alpha_{i+1} c(t_{i+1}) + \beta_{i+1} d(t_{i+1}), & u \in [u_{i+1}, u_{i+2}), \\ \alpha_{i+2} c(t_{i+2}), & u \in [u_{i+2}, u_{i+3}), \\ 0, & u \notin [u_i, u_{i+3}), \end{cases}$$

for  $i = 0, 1, \dots, n$ .



In [2] the author proves several theorems in terms of this new curve, but the most important ones are the following: if  $\lambda = 0$ , then the trigonometric polynomial curve is an arc of an ellipse and the basis function  $b_i(u)$  has  $C^1$  continuity at each of the knots.

The definition of the cubic trigonometric polynomial curve in [3] is as follows.

**Definition 2.2.** Given knots  $u_0 < u_1 < \dots < u_{n+4}$  and refer to  $U = (u_0, u_1, \dots, u_{n+4})$  as a knot vector. For  $\lambda \in \mathbb{R}$  and all possible  $i \in \mathbb{Z}^+$ , let  $\Delta_i = u_{i+1} - u_i$ ,

$$\alpha_i = \frac{\Delta_i}{\Delta_{i-1} + \Delta_i}, \quad \beta_i = \frac{\Delta_i}{\Delta_i + \Delta_{i+1}}, \quad \gamma_i = \frac{1}{\Delta_{i-1} + (2\lambda + 1)\Delta_i + \Delta_{i+1}},$$

$$\begin{aligned} a_i &= \Delta_i \alpha_i \gamma_{i-1}, & d_i &= \Delta_i \beta_i \gamma_{i+1}, \\ b_{i2} &= \frac{1}{4\lambda + 6} (\Delta_{i+1} - \Delta_i) \gamma_i, & c_{i1} &= \frac{1}{4\lambda + 6} (\Delta_{i-1} - \Delta_i) \gamma_i, \\ b_{i0} &= \Delta_{i-1} \alpha_i \gamma_i - a_i + b_{i2}, & c_{i0} &= [1 - (2\lambda + 3)\alpha_i] c_{i1}, \\ b_{i1} &= \frac{1}{2} \Delta_i \gamma_i + b_{i2}, & c_{i2} &= \frac{1}{2} \Delta_i \gamma_i + c_{i1}, \\ b_{i3} &= [1 - (2\lambda + 3)\beta_i] b_{i2}, & c_{i3} &= \Delta_{i+1} \beta_i \gamma_i + c_{i1} - d_i, \\ f_0(t) &= (1 - \sin(t))^2 (1 - \lambda \sin(t)), & f_1(t) &= (1 + \cos(t))^2 (1 + \lambda \cos(t)), \\ f_2(t) &= (1 + \sin(t))^2 (1 + \lambda \sin(t)), & f_3(t) &= (1 - \cos(t))^2 (1 - \lambda \cos(t)). \end{aligned}$$

Given a knot vector  $U$ , let  $t_j(u) = \frac{\pi}{2} \frac{u - u_j}{\Delta_j}$  ( $j = 0, 1, \dots, n + 3$ ), the associated trigonometric polynomial basis functions are defined to be the following functions:

$$B_i(u) = \begin{cases} d_i f_3(t_i), & u \in [u_i, u_{i+1}), \\ \sum_{j=0}^3 c_{i+1,j} f_j(t_{i+1}), & u \in [u_{i+1}, u_{i+2}), \\ \sum_{j=0}^3 b_{i+2,j} f_j(t_{i+2}), & u \in [u_{i+2}, u_{i+3}), \\ a_{i+3} f_0(t_{i+3}), & u \in [u_{i+3}, u_{i+4}), \\ 0, & u \notin [u_i, u_{i+4}), \end{cases}$$

for  $i = 0, 1, \dots, n$ .

### 3. The quadratic case

#### 3.1. Geometric effect of the shape parameter

The basis functions of the quadratic trigonometric polynomial curve have a shape parameter  $\lambda \in (-1, 1)$ , (see Section 1). For our works we restrict the domain of definition of the basis functions to one span. Let  $u \in [u_i, u_{i+1})$ , then

$$\begin{aligned} b_i(u) &= \left( \frac{u_{i+1} - u_i}{u_{i+2} + u_i} \right) \left( 1 - \cos \left( \frac{\pi}{2} \frac{u - u_i}{u_{i+1} - u_i} \right) \right) \left( 1 - \lambda \cos \left( \frac{\pi}{2} \frac{u - u_i}{u_{i+1} - u_i} \right) \right), \\ b_{i-1}(u) &= 1 - \left( \frac{u_{i+1} - u_i}{u_{i-1} + u_{i+1}} \right) \left( 1 - \sin \left( \frac{\pi}{2} \frac{u - u_i}{u_{i+1} - u_i} \right) \right) \left( 1 - \lambda \sin \left( \frac{\pi}{2} \frac{u - u_i}{u_{i+1} - u_i} \right) \right) \end{aligned}$$

$$\begin{aligned}
 & - \left( \frac{u_{i+1} - u_i}{u_{i+2} + u_i} \right) \left( 1 - \cos \left( \frac{\pi}{2} \frac{u - u_i}{u_{i+1} - u_i} \right) \right) \left( 1 - \lambda \cos \left( \frac{\pi}{2} \frac{u - u_i}{u_{i+1} - u_i} \right) \right), \\
 b_{i-2}(u) & = \left( \frac{u_{i+1} - u_i}{u_{i-1} + u_{i+1}} \right) \left( 1 - \sin \left( \frac{\pi}{2} \frac{u - u_i}{u_{i+1} - u_i} \right) \right) \left( 1 - \lambda \sin \left( \frac{\pi}{2} \frac{u - u_i}{u_{i+1} - u_i} \right) \right),
 \end{aligned}$$

hence one arc of the quadratic trigonometric polynomial curve is

$$\mathbf{T}_i(u, \lambda) = b_{i-2}(u)\mathbf{p}_{i-2} + b_{i-1}(u)\mathbf{p}_{i-1} + b_i(u)\mathbf{p}_i,$$

where  $\mathbf{p}_i$  are the control points.

**Theorem 3.1.** *The geometric effect of the shape parameter is linear, that is if  $u_0 \in [u_i, u_{i+1}]$  is fixed, then the curve point  $\mathbf{T}_i(u_0, \lambda)$  moves along a line segment.*

**Proof.** Parts of basis functions not containing the shape parameter  $\lambda$  can be considered as constants:

$$\begin{aligned}
 k_1 & = \left( \frac{u_{i+1} - u_i}{u_{i+2} + u_i} \right) \left( 1 - \cos \left( \frac{\pi}{2} \frac{u_0 - u_i}{u_{i-1} - u_i} \right) \right), & k_2 & = \cos \left( \frac{\pi}{2} \frac{u_0 - u_i}{u_{i-1} - u_i} \right), \\
 k_3 & = \left( \frac{u_{i+1} - u_i}{u_{i+1} + u_{i-1}} \right) \left( 1 - \sin \left( \frac{\pi}{2} \frac{u_0 - u_i}{u_{i-1} - u_i} \right) \right), & k_4 & = \sin \left( \frac{\pi}{2} \frac{u_0 - u_i}{u_{i-1} - u_i} \right).
 \end{aligned}$$

By these constants the basis functions of the quadratic trigonometric polynomial curve can be expressed as:

$$\begin{aligned}
 b_{i-2}(u_0, \lambda) & = k_3 - \lambda k_3 k_4, \\
 b_{i-1}(u_0, \lambda) & = 1 - k_3 + \lambda k_3 k_4 - k_1 + \lambda k_1 k_2, \\
 b_i(u_0, \lambda) & = k_1 - \lambda k_1 k_2.
 \end{aligned}$$

Thus the path of the curve point at parameter  $u_0$  is

$$\mathbf{T}(u_0, \lambda) = (k_3 - \lambda k_3 k_4)\mathbf{p}_{i-2} + (1 - k_3 + \lambda k_3 k_4 - k_1 + \lambda k_1 k_2)\mathbf{p}_{i-1} + (k_1 - \lambda k_1 k_2)\mathbf{p}_i,$$

which yields an equation of a line segment

$$\begin{aligned}
 \mathbf{T}(u_0, \lambda) & = (k_3\mathbf{p}_{i-2} + (1 - k_1 - k_3)\mathbf{p}_{i-2} + k_1\mathbf{p}_i) \\
 & \quad + \lambda (k_3 k_4(\mathbf{p}_{i-1} - \mathbf{p}_{i-2}) + k_1 k_2(\mathbf{p}_{i-1} - \mathbf{p}_i)).
 \end{aligned}$$

The theorem follows. □

### 3.2. Common interpolation

Interpolation of points in general is possible by any of the spline curves by a reverse algorithm: considering the points  $\mathbf{p}_0, \mathbf{p}_1, \dots, \mathbf{p}_n$  to be interpolated by the curve, the new control points can be computed. This algorithm, however, generally requires time-consuming computation solving a system of equations. Thus it can be useful to study if the curve (if it includes a shape parameter) can directly interpolate the given points at a proportional value of the shape parameter.

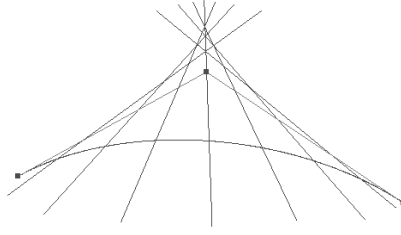


Figure 1: The geometric effect of the shape parameter with fixed parameter  $u_0$ . Straight lines show the  $\lambda$ -paths of the quadratic trigonometric curve.

**Theorem 3.2.** *The quadratic trigonometric polynomial curve interpolates the control points at  $\lambda = \sqrt{2}$ .*

**Proof.** Let the control points  $\mathbf{p}_0, \mathbf{p}_1, \mathbf{p}_2$  so  $u_{i-1} = 0, u_i = 0, u_{i+1} = 1, u_{i+2} = 1$  and the parameter where we search the interpolation is  $u = \frac{u_i + u_{i+1}}{2}$  hence

$$\begin{aligned}
 b_i(u) &= \left(1 - \cos\left(\frac{\pi}{4}\right)\right) \left(1 - \sqrt{2}\cos\left(\frac{\pi}{4}\right)\right) = 0, \\
 b_{i-1}(u) &= 1 - \left(1 - \cos\left(\frac{\pi}{4}\right)\right) \left(1 - \sqrt{2}\cos\left(\frac{\pi}{4}\right)\right) \\
 &\quad - \left(1 - \sin\left(\frac{\pi}{4}\right)\right) \left(1 - \sqrt{2}\sin\left(\frac{\pi}{4}\right)\right) = 1, \\
 b_{i-2}(u) &= \left(1 - \sin\left(\frac{\pi}{4}\right)\right) \left(1 - \sqrt{2}\sin\left(\frac{\pi}{4}\right)\right) = 0.
 \end{aligned}$$

The theorem follows. □



Figure 2: The quadratic trigonometric polynomial curve with  $\lambda = \sqrt{2}$  interpolates the control points, having end tangents parallel to the sides of the control polygon.

### 3.3. Constrained modification

In Computer Aided Geometric Design a frequent problem is the constrained modification, when control points are given and we have another point  $\mathbf{p}$  what we need to interpolate by the curve, possibly by altering the parameters of the curve. Our aim is to modify the given curve  $\mathbf{T}(u, \lambda)$  by altering exclusively the shape parameter in a way, that the modified curve will pass through the given point, that is, for some parameters  $\mathbf{T}(\bar{u}, \bar{\lambda}) = \mathbf{p}$ . The first observation to create this interpolation with the quadratic trigonometric polynomial curve is to show how we could produce a segment with this curve.

**Lemma 3.3.** *Let the control points be  $\mathbf{p}_{i-2}, \mathbf{p}_{i-1}, \mathbf{p}_i$ . If  $\lambda = -1$ , then the quadratic trigonometric curve is a line segment between  $\mathbf{p}_{i-2}$  and  $\mathbf{p}_i$ .*

**Proof.** Let the curve segment be

$$\mathbf{T}_i(u_0, -1) = \mathbf{p}_{i-1} + A(\mathbf{p}_{i-2} - \mathbf{p}_{i-1}) + B(\mathbf{p}_i - \mathbf{p}_{i-1}),$$

where  $A = k_3 - \lambda k_3 k_4$ ,  $B = k_1 - \lambda k_1 k_2$ , and  $A + B = 1$ , therefore

$$\mathbf{T}_i(u_0, -1) = A\mathbf{p}_{i-2} + (1 - A)\mathbf{p}_i.$$

The theorem follows. □

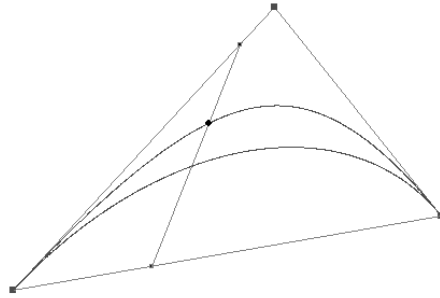


Figure 3: Constrained modification on the quadratic trigonometric polynomial curve.

For a fixed parameter  $u_0$  one can also find the intersection of the  $\lambda$ -path associated to  $u_0$  and the control polygon. In the first case, when  $0 \leq u_0 \leq 0,5$  we can find the parameter value  $\lambda_0$  at which  $B = 0$ . Since  $B = k_1 - \lambda k_1 k_2 = 0$ , therefore  $\lambda_0 = \frac{1}{k_2}$ . So the intersection point is  $\mathbf{T}(u_0, \frac{1}{k_2})$ . In the second case, when  $0,5 < u_0 \leq 1$  then  $A = 0$  should hold.  $A = k_3 - \lambda k_3 k_4$  yields  $\lambda_0 = \frac{1}{k_4}$ . Thus the point what we are looking for is  $\mathbf{T}(u_0, \frac{1}{k_4})$ . Since lambda paths are line segments, they can be described by the two points  $\mathbf{T}_i(u_0, -1)$  and  $\mathbf{T}_i(u_0, \lambda_0)$ . Our next task

is to find the value  $\bar{u}$  for which the path  $\mathbf{T}(\bar{u}, \lambda)$  passes through  $\mathbf{p}$ , by elementary incidence computation. Finally the value of  $\bar{\lambda}$  can be found by a numerical algorithm for which

$$\mathbf{T}(\bar{u}, \bar{\lambda}) = \mathbf{p}.$$

holds. From the algorithm it is clear that the admissible positions of  $\mathbf{p}$  can be inside the convex hull of the control points.

### 4. Extension to the cubic case

The cubic trigonometric polynomial curve also has a shape parameter  $\lambda \in \mathbb{R}$  (see Section 1), and we need the following remark from [3].

**Remark 4.1.** If  $u_i \neq u_{i+1}$  ( $3 \leq i \leq n$ ), then for  $u \in [u_i, u_{i+1}]$ , the curve  $T(u)$  can be represented by curve segment

$$\mathbf{T}(u) = B_{i-3}\mathbf{p}_{i-3} + B_{i-2}\mathbf{p}_{i-2} + B_{i-1}\mathbf{p}_{i-1} + B_i\mathbf{p}_i.$$

With a uniform knot vector, we have

$$B_{i-3}(u) = \frac{f_0(t)}{4\lambda+6}, \quad B_{i-2}(u) = \frac{f_1(t)}{4\lambda+6}, \quad B_{i-1}(u) = \frac{f_2(t)}{4\lambda+6}, \quad B_i(u) = \frac{f_3(t)}{4\lambda+6},$$

where  $t = \pi(u - u_i)/(2\Delta_i)$ .

**Theorem 4.2.** *With a uniform knot vector if  $u_0 \in [u_i, u_{i+1}]$  is fixed, then the geometric effect of the shape parameter is linear, i.e. the  $\lambda$ -path of the point  $\mathbf{T}(u_0, \lambda)$  of the curve are straight line segments.*

**Proof.** The derivatives of the basis functions with respect to  $\lambda$  are

$$\begin{aligned} \frac{\delta B_{i-3}}{\delta \lambda} &= \frac{[-(1 - \sin(t))^2 \sin(t)] (4\lambda + 6) - 4(1 - \sin(t))^2 (1 - \lambda \sin(t))}{(4\lambda + 6)^2} = \\ &= -\frac{(-1 + \sin(t))^2 (2 + 3 \sin(t))}{2(3 + 2\lambda)^2}, \\ \frac{\delta B_{i-2}}{\delta \lambda} &= \frac{[(1 + \cos(t))^2 \cos(t)] (4\lambda + 6) - 4(1 + \cos(t))^2 (1 + \lambda \cos(t))}{(4\lambda + 6)^2} = \\ &= \frac{(1 + \cos(t))^2 (-2 + 3 \cos(t))}{2(3 + 2\lambda)^2}, \\ \frac{\delta B_{i-1}}{\delta \lambda} &= \frac{[(1 + \sin(t))^2 \sin(t)] (4\lambda + 6) - 4(1 + \sin(t))^2 (1 + \lambda \sin(t))}{(4\lambda + 6)^2} = \\ &= \frac{(1 + \sin(t))^2 (-2 + 3 \sin(t))}{2(3 + 2\lambda)^2}, \\ \frac{\delta B_i}{\delta \lambda} &= \frac{[-(1 - \cos(t))^2 \cos(t)] (4\lambda + 6) - 4(1 - \cos(t))^2 (1 - \lambda \cos(t))}{(4\lambda + 6)^2} = \end{aligned}$$

$$= -\frac{(-1 + \cos(t))^2(2 + 3 \cos(t))}{2(3 + 2\lambda)^2},$$

hence the derivative of the cubic trigonometric polynomial curve with respect to  $\lambda$  is

$$\begin{aligned} \frac{\delta \mathbf{T}}{\delta \lambda} &= \frac{1}{2(3 + 2\lambda)^2} \left( (-1 + \sin(t))^2(2 + 3 \sin(t)) \mathbf{p}_0 \right. \\ &\quad \left. + (1 + \cos(t))^2(-2 + 3 \cos(t)) \mathbf{p}_1 + (1 + \sin(t))^2(-2 + 3 \sin(t)) \mathbf{p}_2 \right. \\ &\quad \left. + (-1 + \cos(t))^2(2 + 3 \cos(t)) \mathbf{p}_3 \right). \end{aligned}$$

In the domain of  $\mathbf{T}(u_0, \lambda)$  all the derivative vectors of the  $\lambda$ -paths with respect to  $\lambda$  point to the same direction,  $\lambda$  alters only the derivatives lengths. This proves the statement.  $\square$

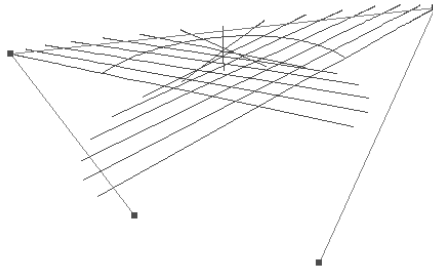


Figure 4: The geometric effect of the shape parameter is linear in the cubic case as well,  $\lambda$ -paths are line segments.

Since  $\lambda$ -paths belong to a fixed  $u_0$  are line segments in the cubic case as well, constrained modification can be computed analogously to the quadratic case (cf. Fig. 5).

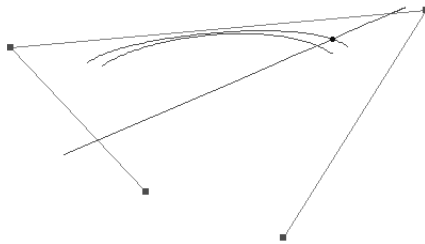


Figure 5: Constrained modification on the cubic trigonometric polynomial curve.

### 5. Extension to surfaces

In this section we present the quadratic trigonometric polynomial surface which is a tensor product surface obtained by the quadratic curve. The basis functions are as follows

$$\begin{aligned}
 b_i(u) &= \left(\frac{u_{i+1} - u_i}{u_{i+2} + u_i}\right) \left(1 - \cos\left(\frac{\pi}{2} \frac{u - u_i}{u_{i+1} - u_i}\right)\right) \left(1 - \lambda_u \cos\left(\frac{\pi}{2} \frac{u - u_i}{u_{i+1} - u_i}\right)\right), \\
 b_{i-1}(u) &= 1 - \left(\frac{u_{i+1} - u_i}{u_{i-1} + u_{i+1}}\right) \left(1 - \sin\left(\frac{\pi}{2} \frac{u - u_i}{u_{i+1} - u_i}\right)\right) \left(1 - \lambda_u \sin\left(\frac{\pi}{2} \frac{u - u_i}{u_{i+1} - u_i}\right)\right) \\
 &\quad - \left(\frac{u_{i+1} - u_i}{u_{i+2} + u_i}\right) \left(1 - \cos\left(\frac{\pi}{2} \frac{u - u_i}{u_{i+1} - u_i}\right)\right) \left(1 - \lambda_u \cos\left(\frac{\pi}{2} \frac{u - u_i}{u_{i+1} - u_i}\right)\right), \\
 b_{i-2}(u) &= \left(\frac{u_{i+1} - u_i}{u_{i-1} + u_{i+1}}\right) \left(1 - \sin\left(\frac{\pi}{2} \frac{u - u_i}{u_{i+1} - u_i}\right)\right) \left(1 - \lambda_u \sin\left(\frac{\pi}{2} \frac{u - u_i}{u_{i+1} - u_i}\right)\right), \\
 b_i(v) &= \left(\frac{v_{i+1} - v_i}{v_{i+2} + v_i}\right) \left(1 - \cos\left(\frac{\pi}{2} \frac{v - v_i}{v_{i+1} - v_i}\right)\right) \left(1 - \lambda_v \cos\left(\frac{\pi}{2} \frac{v - v_i}{v_{i+1} - v_i}\right)\right), \\
 b_{i-1}(v) &= 1 - \left(\frac{v_{i+1} - v_i}{v_{i-1} + v_{i+1}}\right) \left(1 - \sin\left(\frac{\pi}{2} \frac{v - v_i}{v_{i+1} - v_i}\right)\right) \left(1 - \lambda_v \sin\left(\frac{\pi}{2} \frac{v - v_i}{v_{i+1} - v_i}\right)\right) \\
 &\quad - \left(\frac{v_{i+1} - v_i}{v_{i+2} + v_i}\right) \left(1 - \cos\left(\frac{\pi}{2} \frac{v - v_i}{v_{i+1} - v_i}\right)\right) \left(1 - \lambda_v \cos\left(\frac{\pi}{2} \frac{v - v_i}{v_{i+1} - v_i}\right)\right), \\
 b_{i-2}(v) &= \left(\frac{v_{i+1} - v_i}{v_{i-1} + v_{i+1}}\right) \left(1 - \sin\left(\frac{\pi}{2} \frac{v - v_i}{v_{i+1} - v_i}\right)\right) \left(1 - \lambda_v \sin\left(\frac{\pi}{2} \frac{v - v_i}{v_{i+1} - v_i}\right)\right),
 \end{aligned}$$

where  $-1 < \lambda_u, \lambda_v < 1$  are shape parameters, and the surface patch is defined as

$$\begin{aligned}
 \mathbf{T}(u, v) &= b_{i-2}(u)b_{i-2}(v)\mathbf{p}_j + b_{i-2}(u)b_{i-1}(v)\mathbf{p}_{j+1} + b_{i-2}(u)b_i(v)\mathbf{p}_{j+2} + \\
 &\quad + b_{i-1}(u)b_{i-2}(v)\mathbf{p}_{j+3} + b_{i-1}(u)b_{i-1}(v)\mathbf{p}_{j+4} + b_{i-1}(u)b_i(v)\mathbf{p}_{j+5} + \\
 &\quad + b_i(u)b_{i-2}(v)\mathbf{p}_{j+6} + b_i(u)b_{i-1}(v)\mathbf{p}_{j+7} + b_i(u)b_i(v)\mathbf{p}_{j+8}.
 \end{aligned}$$

The shape parameters  $\lambda_u, \lambda_v$  are independent so they modify the surface in separated ways. As we have seen for  $\lambda = -1$  the quadratic curve is a line segment passing through  $\mathbf{p}_{i-2}$  and  $\mathbf{p}_i$ . Consequently the quadratic trigonometric polynomial surface at  $\lambda_u = -1$  and  $\lambda_v = -1$  is a plane interpolating control points  $\mathbf{p}_j, \mathbf{p}_{j+2}, \mathbf{p}_{j+6}, \mathbf{p}_{j+8}$ .

**Theorem 5.1.** *The quadratic trigonometric polynomial surface interpolates the control points when  $\lambda_u = \sqrt{2}$  and  $\lambda_v = \sqrt{2}$ .*

**Proof.** Considering the control points  $\mathbf{p}_k$  ( $k = 1, 2, \dots, 9$ ), knots  $u_{i-1} = 0, u_i = 0, u_{i+1} = 1, u_{i+2} = 1, v_{i-1} = 0, v_i = 0, v_{i+1} = 1, v_{i+2} = 1$ , the parameters at which

the interpolation holds are  $u = \frac{u_i + u_{i+1}}{2}$  and  $v = \frac{v_i + v_{i+1}}{2}$  hence the statement follows from Theorem 3.2.  $\square$

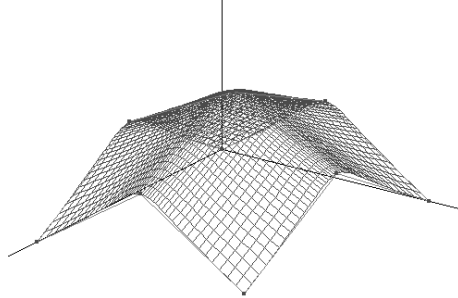


Figure 6: The quadratic trigonometric polynomial surface with  $\lambda_u = \sqrt{2}$  and  $\lambda_v = \sqrt{2}$ .

### 5.1. Geometric effect of the shape parameters $\lambda_u$ and $\lambda_v$ on the quadratic trigonometric surface

Since shape parameters acts independently on the surface, if we fix either of them, the geometric effect of the other shape parameter is the same as in the curve case (see Section 3.1). Consequently in the case when both of the shape parameters are changing simultaneously, points of the surface can move on a doubly ruled surface, an example of which can be seen in Fig.7.

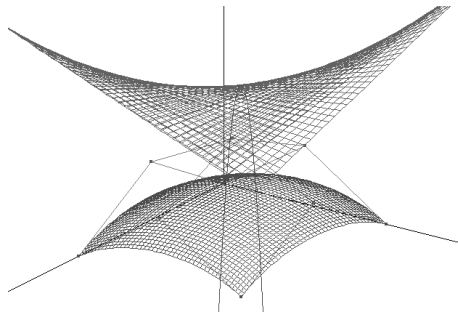


Figure 7:  $\lambda$ -paths of a surface point associated to  $u_0 = 0, 2$  and  $v_0 = 0, 2$  are on a hyperbolic paraboloid (one of its section curves is also shown).



## References

- [1] CHEN, Q., WANG, G., A class of Bézier-like curves, *Computer Aided Geometric Design*, 20, (2003) 29–39.
- [2] HAN, X., Quadratic trigonometric polynomial curves with a shape parameter, *Computer Aided Geometric Design*, 19 (2002) 503–512.
- [3] HAN, X., Cubic trigonometric polynomial curves with a shape parameter, *Computer Aided Geometric Design*, 21 (2004) 535–548.
- [4] HOFFMANN, M., JUHÁSZ, I., Modifying the shape of FB-spline curves, *Journal of Applied Mathematics and Computing*, 27, (2008) 257–269.
- [5] HOFFMANN, M., JUHÁSZ, I., On the quartic curve of Han, *Journal of Computational and Applied Mathematics*, 223, (2009) 124–132.
- [6] HOFFMANN, M., LI, Y., WANG, G., Paths of C-Bézier and C-B-spline curves, *Computer Aided Geometric Design*, 23, (2006) 463–475.
- [7] LI, Y., HOFFMANN, M., WANG, G., On the shape parameter and constrained modification of GB-spline curves, *Annales Mathematicae et Informaticae*, 34, (2007) 51–59.
- [8] LI, Y., WANG, G., Two kinds of B-basis of the algebraic hyperbolic space, *Journal of Zhejiang Univ. Sci.*, 6, (2005) 750–759.
- [9] LÜ, Y., WANG, G., YANG, X., Uniform hyperbolic polynomial B-spline curves, *Computer Aided Geometric Design*, 19, (2002) 379–393.
- [10] MAINAR, M., PENA, J.M., SANCHEZ-REYES, J., Shape preserving alternatives to the rational Bézier model. *Computer Aided Geometric Design*, 18, (2001) 37–60.
- [11] MORIN, G., WARREN, J., WEIMER, H., A subdivision scheme for surface of revolution, *Computer Aided Geometric Design*, 18, (2001) 483–502.
- [12] POTTMANN, H., The geometry of Tchebycheffian splines, *Computer Aided Geometric Design*, 10, (1993) 181–210.
- [13] WANG, G., CHEN, Q., ZHOU, M., NUAT B-spline curves, *Computer Aided Geometric Design*, 21, (2004) 193–205.
- [14] ZHANG, J.W., C-curves, an extension of cubic curves, *Computer Aided Geometric Design*, 13, (1996) 199–217.
- [15] ZHANG, J.W., Two different forms of CB-splines, *Computer Aided Geometric Design*, 14, (1997) 31–41.
- [16] ZHANG, J.W., C-Bézier curves and surfaces, *Graph. Models Image Process.*, 61, (1999) 2–15.

**Ede M. Troll, Miklós Hoffmann**

Institute of Mathematics and Computer Science

Eszterházy Károly College

Leányka str. 4

Eger

Hungary

e-mail: [ede.troll@gmail.com](mailto:ede.troll@gmail.com)

[hofi@ektf.hu](mailto:hofi@ektf.hu)



# CTH B-spline curves and its applications\*

Jin Xie<sup>ab</sup>, Jieqing Tan<sup>ac</sup>, Shengfeng Li<sup>ad</sup>

<sup>a</sup>School of Computer and Information, Hefei University of Technology, Hefei, China

<sup>b</sup>Department of Mathematics and Physics, Hefei University, Hefei, China

<sup>c</sup>School of Mathematics, Hefei University of Technology, Hefei, China

<sup>d</sup>Department of Mathematics and Physics, Bengbu College, Bengbu, China

*Submitted 21 April 2010; Accepted 26 August 2010*

## Abstract

A method of generating cubic blending spline curves based on weighted trigonometric and hyperbolic polynomial is presented in this paper. The curves inherit nearly all properties of cubic B-splines and enjoy some other advantageous properties for modeling. They can represent some conics and some transcendental curves exactly. Here weight coefficients are also shape parameters, which are called weight parameters. The interval  $[0,1]$  of weight parameter values can be extended to  $[\frac{e-1}{(e-1)^2-\pi}, \frac{e-1}{(e-1)^2\pi^2-8e}]$ . Not only can the shape of the curves be adjusted globally or locally, but also the type of some segments of a blending curve can be switched by taking different values of the weight parameters. Without solving system of equations and letting certain weight parameter be  $\frac{(e-1)^2(2-\pi)}{2(e-1)^2-2\pi}$ , the curves can interpolate corresponding control points directly.

*Keywords:* cubic uniform B-spline, CTH B-spline, weight parameter, local and global interpolation, local and global adjustment, transcendental curve

*MSC:* 68U05

## 1. Introduction

B-spline curves and surfaces are well known geometric modeling tools in Computer Aided Geometric Design (CAGD). Due to their several limitations in practical applications[1], several new forms of curve and surface schemes have been proposed

---

\*Reaserch supported by the National Nature Science Foundation of China (No.61070227), the Doctoral Program Foundation of Ministry of Education of China (No. 20070359014) and the Key Project Foundation of Scientific Research for Hefei university (No.11KY02ZD).

for geometric modeling in CAGD[2-12]. C-curves are introduced in [2,3] by using the basis  $\{1, t, \text{cost}, \text{sint}\}$  instead of  $\{1, t, t^2, t^3\}$  in cubic spline curves, which can represent some transcendental curves such as the ellipse, the helix and the cycloid. Further properties of C-curves have been studied in [4]. Hoffmann et al. [5] investigated a geometric interpretation of the change of parameter  $\alpha$  for C-B-spline curves. Similarly, using the hyperbolic basis  $\{1, t, \text{cosht}, \text{sinht}\}$  instead of  $\{1, t, t^2, t^3\}$  in cubic uniform B-splines, one can construct a curve family too. This has been studied as exponential B-splines [6,7,8]. Just for convenience, we call them HB-splines. Koch and Lyche[6] presented a kind of exponential splines in tension in the space spanned by  $\{1, t, \text{cosht}, \text{sinht}\}$ . Lü et al.[7] gave the explicit expressions for uniform splines. Li and Wang[8] generalized the curves and surfaces of exponential forms to algebraic hyperbolic spline forms of any degree, which can represent exactly some remarkable curves and surfaces such as the hyperbola, the catenary, the hyperbolic spiral and the hyperbolic paraboloid.

CB-splines and HB-splines are the same in structure and their shapes are adjustable. However, after comparing CB-splines and HB-splines, we found that a CB-spline is located on one side of the B-spline, and the HB-spline is located on the other side of the B-spline, see Figure 1. Therefore, one thinks whether the two different curves can be unified. If we can unify them, then the new curve will have more plentiful modeling power. In order to construct more flexible curves for the surface modeling, Zhang et al. [9,10] proposed a curve family, named FB-spline, that is the unification of CB-spline and HB-spline. However, the formulas for the FB-splines were rather complicated. Hoffmann et al. [11] introduced practical shape modification algorithms of FB-spline curves and the geometrical effects of the alteration of shape parameters, which are essential from the users' point of view. Wang and Fang[12] unified and extended three types of splines by a new kind of spline (UE-spline for short) defined over the space  $\{\text{coswt}, \text{sinwt}, 1, t, \dots, t^l, \dots\}$ , where the type of a curve can be switched by a frequency sequence  $\{\omega_i\}$ .

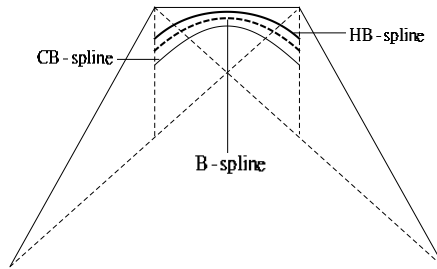


Figure 1: CB-spline and HB-spline are located on the different sides of B-spline

In this paper, we present a set of new bases by unifying the trigonometric basis and the hyperbolic basis using weight method, which inherits the most properties of cubic uniform B-spline bases. Based on those bases, we introduce a new spline curve, named CTH B-spline curve. This approach has the following features:

- The introduced curves can cross the B-splines and reach the both sides of cubic B-splines.
- The shape of the curves can be adjusted globally or locally.
- Without solving system of equations and letting weight parameters be  $(e - 1)^2(2 - \pi)/(2(e - 1)^2 - 2\pi)$ , the curves can interpolate certain control points directly.
- With the weight parameters and control points chosen properly, the CTH B-spline curves can be used to represent some conics and transcendental curves.
- The type of the curves can be switched by letting weight parameters  $\lambda_i = 0$  or 1 easily. And, a blending curve can be composed of different type curve segments.

The rest of this paper is organized as follows. In Section 2, the basis functions unified by the trigonometric basis and the hyperbolic basis using weight method are established and the properties of the basis functions are shown. In Section 3, the CTH B-spline curves are given and some properties are discussed. It is pointed out in Section 4 that some transcendental curves can be represented precisely with the CTH B-spline curves and the applications of the curves are shown in Section 5. Finally, we conclude the paper in Section 6.

## 2. The construction of CTH B-spline basis functions

In order to construct CTH B-spline basis functions, we give two classes of basis functions as follows.

**Definition 2.1.** The following functions,

$$\begin{cases} T_{0,3}(t) = \frac{1-t}{2} - \frac{1}{\pi} \cos \frac{\pi}{2}t, \\ T_{1,3}(t) = \frac{t}{2} + \frac{2}{\pi} \cos \frac{\pi}{2}t - \frac{1}{\pi} \sin \frac{\pi}{2}t, \\ T_{2,3}(t) = \frac{1-t}{2} + \frac{2}{\pi} \sin \frac{\pi}{2}t - \frac{1}{\pi} \cos \frac{\pi}{2}t, \\ T_{3,3}(t) = \frac{t}{2} - \frac{1}{\pi} \sin \frac{\pi}{2}t, \end{cases}$$

are called CT B-spline basis functions.

**Remark 2.2.** The CT B-spline basis functions are the CB-spline basis functions with  $\alpha = \pi/2$ , see[3].

**Definition 2.3.** The following functions,

$$\begin{cases} H_{0,3}(t) = -\frac{e}{(e-1)^2}(1-t) + \frac{e}{(e-1)^2} \sinh(1-t), \\ H_{1,3}(t) = -\frac{e}{(e-1)^2} + \frac{1+e+e^2}{(e-1)^2}(1-t) + \frac{e+1}{2(e-1)} \cosh(1-t) \\ \quad + \frac{1+4e+e^2}{(e-1)^2\pi} \sinh(1-t), \\ H_{2,3}(t) = -\frac{e}{(e-1)^2} + \frac{1+e+e^2}{(e-1)^2}t + \frac{e+1}{2(e-1)} \cosh t + \frac{1+4e+e^2}{(e-1)^2\pi} \sinht, \\ H_{3,3}(t) = -\frac{e}{(e-1)^2}t + \frac{e}{(e-1)^2} \sinht, \end{cases}$$

are called CH B-spline basis functions.

**Remark 2.4.** The CH B-spline basis functions are the AH B-spline basis functions of order 4 with  $\alpha = 1$  for a uniform knot vector, see[7].

Obviously, the CT B-spline basis functions and CH B-spline basis functions share the properties similar to cubic B spline basis functions, such as nonnegativity, partition of unity and symmetry.

Note that shape of the CT B-spline curves and CH B-spline curves based on the CT B-spline basis functions and CH B-spline basis functions are fixed relative to their control polygons respectively, which is inconvenient to the user.

Next, we construct a set of new basis functions by unifying the CT B-spline basis functions and CH B-spline basis functions using weight method.

**Definition 2.5.** The following functions,

$$\left\{ \begin{aligned} TH_{0,3}(t) &= \frac{1}{\pi}(\lambda_i - 1)\cos\frac{\pi}{2}t + \frac{1}{(e-1)^2}((1 - e)^2 - (1 + e^2)\lambda_i)(1 - t) \\ &\quad + 2e\lambda_i \sinh(1 - t), \\ TH_{1,3}(t) &= \frac{1}{2}t + \frac{e^2+1}{2(e-1)^2}((\lambda_{i+1} + 2\lambda_i)t - \lambda_{i+1}) + \frac{2}{\pi}(1 - \lambda_i)\cos\frac{\pi}{2}t - \\ &\quad \frac{1}{\pi}(1 - \lambda_{i+1})\sin\frac{\pi}{2}t + \frac{(1+e)\lambda_{i+1}}{2e-2}\cosh(1 - t) - \frac{(1+e^2)\lambda_{i+1}+4e\lambda_i}{(e-1)^2\pi}\sinh(1 - t), \\ TH_{2,3}(t) &= \frac{1}{2}(1 - t) + \frac{e^2+1}{2(e-1)^2}((\lambda_i + 2\lambda_{i+1})t - \lambda_i) + \frac{2}{\pi}(1 - \lambda_{i+1})\sin\frac{\pi}{2}t \\ &\quad - \frac{1}{\pi}(1 - \lambda_i)\cos\frac{\pi}{2}t + \frac{(1+e)\lambda_i}{2e-2}\cosht - \frac{(1+e^2)\lambda_i+4e\lambda_{i+1}}{(e-1)^2\pi}\sinht, \\ TH_{3,3}(t) &= \frac{1}{\pi}(\lambda_{i+1} - 1)\sin\frac{\pi}{2}t + \frac{1}{(e-1)^2}((1 - e)^2 - (1 + e^2)\lambda_{i+1})t \\ &\quad + 2e\lambda_{i+1}\sinht, \end{aligned} \right. \tag{2.1}$$

are called CTH B-spline basis functions with weight parameter sequence  $\{\lambda_k\}$ .

Straightforward computation testifies that these CTH B-spline basis functions possess the properties similar to the cubic B-Spline basis functions as follows.

(a)Partition of unity:

$$\sum_{j=0}^3 TH_{j,3}(t) = 1. \tag{2.2}$$

(b) Nonnegativity:

$$TH_{j,3}(t) \geq 0, j = 0, 1, 2, 3. \tag{2.3}$$

(c) Symmetry:

$$TH_{0,3}(t; \lambda_i) = TH_{3,3}(1 - t; \lambda_i), TH_{1,3}(t; \lambda_i, \lambda_{i+1}) = TH_{2,3}(1 - t; \lambda_{i+1}, \lambda_i). \tag{2.4}$$

According to the method of extending definition interval of C-curves in Ref. [13], The interval  $[0, 1]$  of weight parameter values can be extended to  $[\frac{e-1}{(e-1)^2-\pi}, \frac{e-1}{(e-1)^2-\pi}]$ , where  $\frac{e-1}{(e-1)^2-\pi} \approx -15.6134$  and  $\frac{e-1}{(e-1)^2-\pi} \approx 3.9412$ .

For a uniform knot vector, Figure 2 shows cubic uniform B-spline basis functions (dashed lines) and the CTH B-spline basis functions with all parameters being the same (left)and with all parameters different from one another (right).

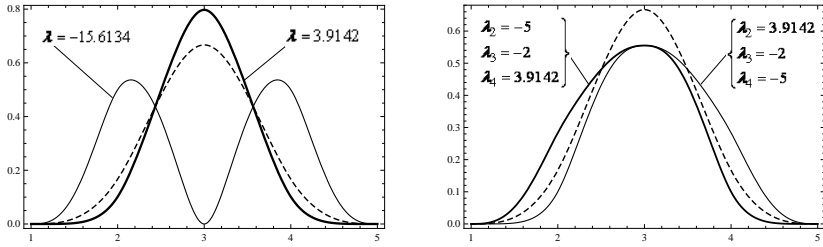


Figure 2: CTH B-spline basis functions

### 3. CTH B-spline curves

#### 3.1. Construction of the curves

**Definition 3.1.** Given control points  $P_i \in R^d (d = 2, 3, i = 0, 1, \dots, n)$  and knots  $u_1 < u_2 < \dots < u_{n-1}$ , for  $u \in [u_i, u_{i+1}]$ ,  $i = 0, 1, \dots, n$ , the curves

$$r(u) = \sum_{j=0}^3 P_{i+j-1} TH_{j,3}(t) \tag{3.1}$$

are defined to be piecewise CTH-B-spline curves, where  $\Delta_i = u_{i+1} - u_i, u = \frac{u-u_i}{\Delta_i}$ .

We can construct the open and closed curves similar to the cubic B-Spline curves.

For open curves, we can expand the curve segment by setting  $\frac{e-1}{(e-1)^2-\pi} \leq \lambda_0, \lambda_n \leq \frac{e-1}{(e-1)^2-\pi^2-8e}, u_0 < u_1, u_{n-1} < u_n, P_{-1} = 2P_0 - P_1, P_{n+1} = 2P_n - P_{n-1}$ . This assures that original points  $P_0$  and  $P_n$  are the points on the curves, i.e.,  $r(u_0) = P_0, r(u_n) = P_n$ . For closed curves, we can periodically assign control points by setting  $P_{n+1} = P_0, P_{n+2} = P_1, P_{n+3} = P_2$ , and expand the knots by setting  $u_{n-1} < u_n < u_{n+1} < u_{n+2}$  and let  $\lambda_i \in [\frac{e-1}{(e-1)^2-\pi}, \frac{e-1}{(e-1)^2-\pi^2-8e}], i = n, n + 1, n + 2, \lambda_1 = \lambda_{n+2}$ . Thus, the parametric formulae for closed curves are defined on the interval  $[u_1, u_{n+1}]$ .

#### 3.2. Properties of the curves

##### 3.2.1. Parametric continuity

Curves (3.1) are piecewise trigonometric hyperbolic polynomial curves. We need to show the continuity of the curves.

**Theorem 3.2.** For  $[u_1, u_{n-1}]$ , curves (3.1) are  $GC^2$  continuous. The uniform curves (3.1) are  $C^2$  continuous.

**Proof.** For  $i = 0, 1, \dots, n$ , We have

$$r(u_i^+) = \left(\frac{\pi - 2}{2\pi} + \frac{\lambda_i}{\pi} - \frac{\lambda_i}{(e - 1)^2}\right)(P_{i-1} + P_{i+1}) + \left(\frac{2}{\pi} - \frac{2\lambda_i}{\pi} + \frac{2\lambda_i}{(e - 1)^2}\right)P_i, \tag{3.2}$$

$$r(u_{i+1}^-) = \left(\frac{\pi - 2}{2\pi} + \frac{\lambda_{i+1}}{\pi} - \frac{\lambda_{i+1}}{(e - 1)^2}\right)(P_i + P_{i+2}) + \left(\frac{2}{\pi} - \frac{2\lambda_{i+1}}{\pi} + \frac{2\lambda_{i+1}}{(e - 1)^2}\right)P_{i+1} \tag{3.3}$$

$$r'(u_i^+) = \frac{1}{2\Delta_i}(P_{i+1} - P_{i-1}), \tag{3.4}$$

$$r'(u_{i+1}^-) = \frac{1}{2\Delta_i}(P_{i+2} - P_i), \tag{3.5}$$

$$r''(u_i^+) = \frac{(e - 1)\pi + ((e - 1)\pi - 2(e + 1))\lambda_i}{4(e - 1)\Delta_i^2}(P_{i-1} - 2P_i + P_{i+1}), \tag{3.6}$$

$$r''(u_{i+1}^-) = \frac{(e - 1)\pi + ((e - 1)\pi - 2(e + 1))\lambda_{i+1}}{4(e - 1)\Delta_i^2}(P_i - 2P_{i+1} + P_{i+2}), \tag{3.7}$$

Thus, we obtain

$$r^{(k)}(u_i^-) = \left(\frac{\Delta_i}{\Delta_{i-1}}\right)^k r^{(k)}(u_i^+), k = 2, 3, i = 0, 1, \dots, n - 2. \tag{3.8}$$

This implies the theorem. □

From (3.4) and (3.5), we know that the tangent line of curves  $r(u)$  at the point  $r(u_i)$  is parallel to the line segment  $P_{i-1}P_{i+1}$  (for any  $\lambda_i$ ). This property corresponds to the property of the cubic uniform B-spline curves.

**Theorem 3.3.** *The curvature of the curves at  $u = u_i$  is*

$$K(u_i) = \frac{|(e - 1)\pi + ((e - 1)\pi - 2(e + 1))\lambda_i| |(P_i - P_{i-1}) \times (P_{i+1} - P_i)|}{e - 1 \|P_{i+1} - P_{i-1}\|^3} \tag{3.9}$$

**Proof.** According to (3.4) and (3.6), the curvature of the curves at  $u = u_i$  is

$$\begin{aligned} K(u_i) &= \frac{|r'(u_i) \times r''(u_i)|}{\|r'(u_i)\|^3} \\ &= \frac{|(e - 1)\pi + ((e - 1)\pi - 2(e + 1))\lambda_i| |(P_{i+1} - P_{i-1}) \times (P_{i-1} - 2P_i + P_{i+1})|}{e - 1 \|P_{i+1} - P_{i-1}\|^3} \\ &= \frac{|(e - 1)\pi + ((e - 1)\pi - 2(e + 1))\lambda_i| |(P_i - P_{i-1}) \times (P_{i+1} - P_i)|}{e - 1 \|P_{i+1} - P_{i-1}\|^3}. \end{aligned}$$

□



According to (3.9), the local parameter  $\lambda_i$  controls the curvature of the curves  $r(u)$  at the end of the curve segments. When  $\lambda_i > \frac{(e-1)\pi}{2(e+1)-(e-1)\pi}$ , the curvature of the curves at  $u = u_i$  increases with the increase of  $\lambda_i$ . When  $\lambda_i < \frac{(e-1)\pi}{2(e+1)-(e-1)\pi}$ , the curvature of the curves at  $u = u_i$  increases with the decrease of  $\lambda_i$ .

**3.2.2. Local and global adjustable properties**

By rewriting (3.1), for  $u \in [u_{i-1}, u_i]$ , we have

$$r_{i-1}(u) = TH_{0,3}(t; \lambda_{i-1})P_{i-2} + TH_{1,3}(t; \lambda_{i-1}, \lambda_i)P_{i-1} + TH_{2,3}(t; \lambda_{i-1}, \lambda_i)P_i + TH_{3,3}(t; \lambda_i)P_{i+1}. \tag{3.10}$$

For  $u \in [u_i, u_{i+1}]$ , we have

$$r_i(u) = TH_{0,3}(t; \lambda_i)P_{i-1} + TH_{1,3}(t; \lambda_i, \lambda_{i+1})P_i + TH_{2,3}(t; \lambda_i, \lambda_{i+1})P_{i+1} + TH_{3,3}(t; \lambda_{i+1})P_{i+2}. \tag{3.11}$$

Obviously, weight parameter  $\lambda_i$  only affect two curve segments  $r_{i-1}(u)$  and  $r_i(u)$  without altering the remainder, namely, weight parameter  $\lambda_i$  only affect control polygon  $\widehat{P_{i-1}P_iP_{i+1}}$ . So we can adjust the curves locally by changing certain  $\lambda_i$ . From Figure 3(a), we can see that increasing  $\lambda_i$  moves locally the curves  $r(u) u \in [u_{i-1}, u_{i+1}]$  towards the control polygon  $\widehat{P_{i-1}P_iP_{i+1}}$ , or decreasing  $\lambda_i$  moves locally the curves  $r(u) u \in [u_{i-1}, u_{i+1}]$  away the control polygon  $\widehat{P_{i-1}P_iP_{i+1}}$ .

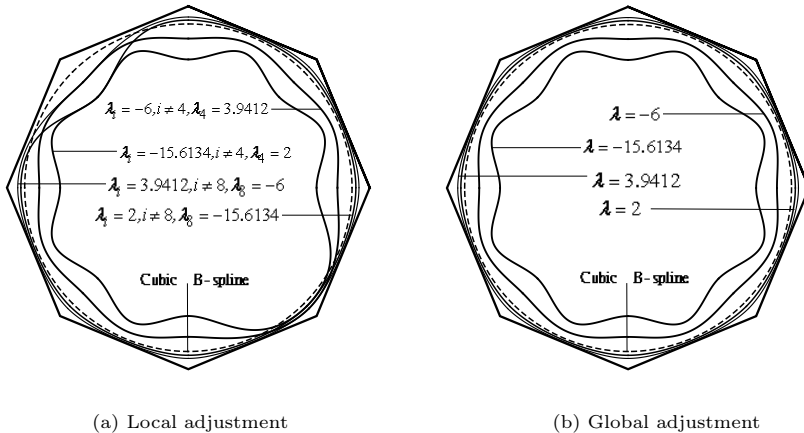


Figure 3: Adjusting the shape of the curves

When all  $\lambda_i$  are the same, the curves can be adjusted globally. From Figure 3(b), we can see that when the control polygon is fixed, adjusting the value of the weight parameters from -15.6134 to 3.9412, the CTH B-spline curves can cross the

cubic B-spline curves (dashed lines) and reach the both sides of cubic B-splines, in other words, the CTH B-spline curves can range from inside the cubic B-spline curves to outside the cubic B-spline curves. And, the weight parameters are of the property that the larger the weight parameter is, the more closely the curves approximate the control polygon.

### 3.2.3. Local and global interpolation

Curve (3.1) can also be used for local interpolation. Let  $\lambda_i = \frac{(e-1)^2(2-\pi)}{2(e-1)^2-2\pi} \approx 8.91206$ , from (3.2) and (3.3), we have  $r(u_i) = P_i$ . This means that curve  $r(u)$  interpolates point  $P_i$  at  $u = u_i$  locally. Thus, we provide a  $GC^2$ continuous local interpolation method without solving a linear system or any additional control points. The given piecewise CTH B-spline curves unify the representation of the curves for interpolating and approximating the control polygons.

Obviously, when all  $\lambda_i = \frac{(e-1)^2(2-\pi)}{2(e-1)^2-2\pi}$ , the curve can interpolate the control polygon globally. Figure 4 shows global interpolation curves with all  $\lambda_i = \frac{(e-1)^2(2-\pi)}{2(e-1)^2-2\pi}$  (red lines) and local interpolation curves with all  $\lambda_i = -1$  except  $\lambda_5 = \frac{(e-1)^2(2-\pi)}{2(e-1)^2-2\pi}$  (blue lines).

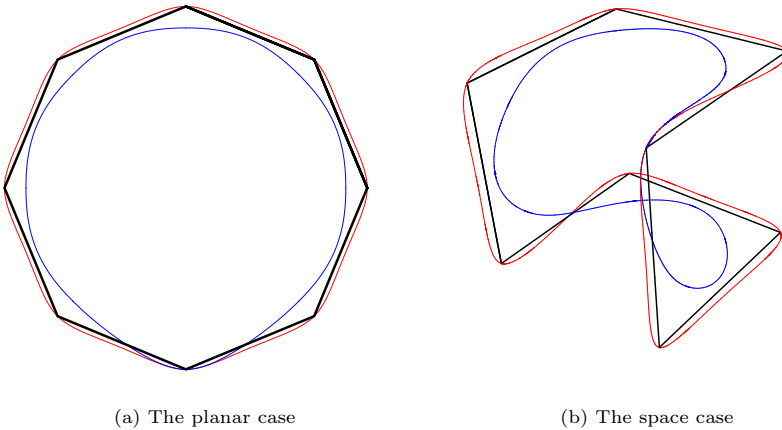


Figure 4: The local and global interpolation curves

## 4. The representations of cycloid, helix and catenary

Given uniform knots, when all  $\lambda_i = 0$ , curves  $r(u)$  are piecewise trigonometric polynomial curves. In this case, for  $u \in [u_i, u_{i+1}]$ , if we take  $P_{i-1} = (\frac{\pi-2}{2}a, a), P_i = (0, \frac{2-\pi}{2}a), P_{i+1} = (\frac{2-\pi}{2}a, a), P_{i+2} = (2a, \frac{2+\pi}{2}a)$  ( $a \neq 0$ ), then the coordinates of  $r(u)$  are

$$\begin{cases} x = a(t_i - \sin\frac{\pi}{2}t_i), \\ y = a(1 - \cos\frac{\pi}{2}t_i). \end{cases}$$

This gives the parametric equation of cycloid. Hence  $r(u)$  is an arc of a cycloid, see Figure 5.

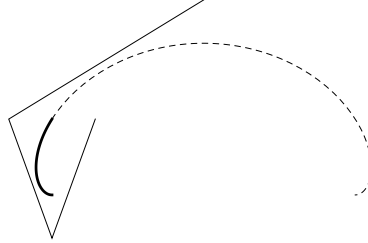


Figure 5: The representation of cycloid by the CTH B-spline curves

If we take  $P_{i-1} = (m, n - \frac{\pi}{2}a, -b)$ ,  $P_i = (m + \frac{\pi}{2}a, n, 0)$ ,  $P_{i+1} = (m, n + \frac{\pi}{2}a, b)$ ,  $P_{i+2} = (m - \frac{\pi}{2}a, n, 2b)$  ( $ab \neq 0$ ), the coordinates of  $r(u)$  are

$$\begin{cases} x = m + a\cos\frac{\pi}{2}t_i, \\ y = n + a\sin\frac{\pi}{2}t_i, \\ z = bt_i, \end{cases}$$

which is parametric equation of a helix. Hence  $r(u)$  is a helix segment, see Figure 6.

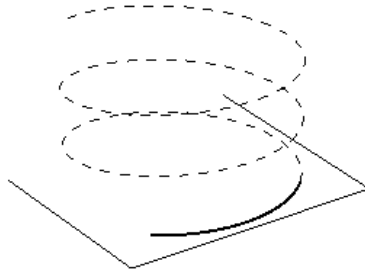


Figure 6: The representation of helix by the CTH B-spline curves

On the other hand, given uniform knots, when all  $\lambda_i = 1$ , curves  $r(u)$  are piecewise hyperbolic polynomial curves. In this case, for  $u \in [u_i, u_{i+1}]$ , if we take  $P_{i-1} = (2a, \frac{e^4+1}{e^3-e}a)$ ,  $P_i = (a, \frac{e^2+1}{e^2-1}a)$ ,  $P_{i+1} = (0, \frac{2e}{e^2-1}a)$ ,  $P_{i+2} = (-a, \frac{e^2+1}{e^2-1}a)$  ( $a \neq 0$ ), then the coordinates of  $r(u)$  are

$$\begin{cases} x = at_i, \\ y = a\cosht_i. \end{cases}$$

This gives the parametric equation of catenary. Hence  $r(u)$  is an arc of a catenary, see Figure 7.

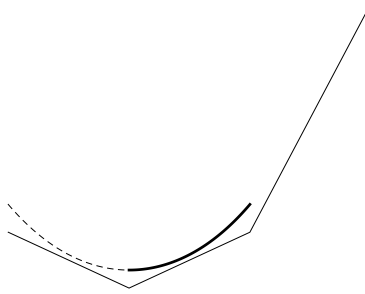


Figure 7: The representation of catenary by the CTH B-spline curves

**Remark 4.1.** By selecting proper control points and weight parameters, some conics such as hyperbola, ellipse and some transcendental curves such as sine curve, cosine curve and hyperbolic sine curves can also be represented via CTH B-spline curves.

## 5. Application of the curves

As mentioned in section 4, the types of the curves can be changed by selecting control points and parameters properly. So, as an application, we can construct a blending curve using different type curve segments flexibly. For example, given a uniform knot vector, let control points as follows,  $P_0 = (-2, \frac{\pi}{4})$ ,  $P_1 = (\frac{\pi-4}{2}, 0)$ ,  $P_2 = (-2, -\frac{\pi}{4})$ ,  $P_3 = (-\frac{\pi+4}{4}, 0)$ ,  $P_4 = (-\frac{e^2+1}{2}, \frac{e^4+e^3-e+1}{e^3-e})$ ,  $P_5 = (-1, \frac{2e^2}{e^2-1})$ ,  $P_6 = (0, \frac{e^2+2e-1}{e^2-1})$ ,  $P_7 = (1, \frac{2e^2}{e^2-1})$ ,  $P_8 = (2, \frac{e^4+e^3-e+1}{e^3-e})$ ,  $P_9 = (1, 6)$ ,  $P_{10} = (2, \frac{\pi+12}{2})$ ,  $P_{11} = (3, 6)$ ,  $P_{12} = (4, \frac{12-\pi}{2})$ ,  $P_{13} = (4, \frac{e^2+1}{e})$ ,  $P_{14} = (3, 1)$ ,  $P_{15} = (2, 0)$ ,  $P_{16} = (1, -1)$ ,  $P_{17} = (\frac{\pi-2}{2}, 1)$ ,  $P_{18} = (0, \frac{2-\pi}{2})$ ,  $P_{19} = (\frac{2-\pi}{2}, 1)$ ,  $P_{20} = (2, \frac{2+\pi}{2})$ . so we obtain a blending curve composed of different type curve segments, which is  $C^2$  continuous, see Figure 8.

## 6. Conclusions

CTH B-spline curves inherited nearly all the properties that CB-spline curves and CH-spline curves and cubic B-spline curves have, such as variation diminishing property, convex hull property, geometric invariance and so on. In this paper, we focus on some special properties of the introduced curves. For example, the shape of the curves can be adjusted globally or locally without adjusting the corresponding control polygon. Without solving system of equations, the curves can interpolate certain control points with proper parameter values. Also, the types of the curves can be switched by weight parameters  $\lambda_i = 0$  or  $1$ , which are easier to determine than the FB-spline or the UE-spline.

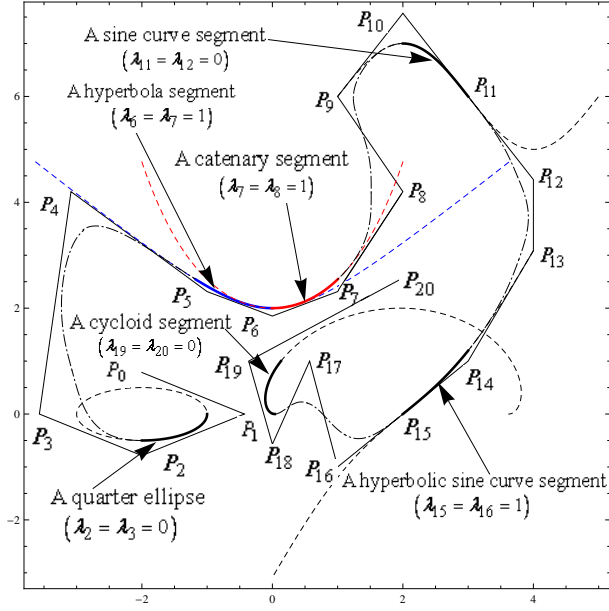


Figure 8: A  $C^2$  continuous blending curve

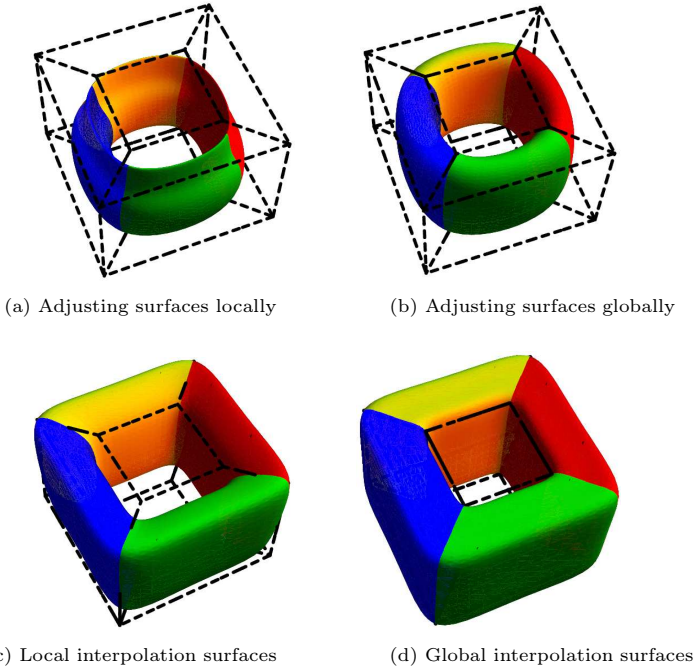


Figure 9: CTH B-spline surfaces

Both rational methods (NURBS or Rational Bézier curves) [15] and CTH B-spline curves can deal with both free form curves and most important analytical shapes for the engineering. However, CTH B-spline curves are simpler in structure and more stable in calculation. The weight parameters of CTH B-spline curves have geometric meaning and are easier to determine than the rational weights in rational methods. Also, CTH B-spline curves can represent the helix, the cycloid, and the catenary precisely, but NURBS can not. Therefore, CTH B-spline curves would be useful for engineering.

Just as in the construction of cubic B-spline tensor product surfaces from cubic B-spline curves, CTH B-spline surfaces can be constructed from CTH B-spline curves easily. And many properties of the curves can be extended to the surfaces. Figure 9 shows an example of the CTH B-spline tensor product surfaces, where surface shapes are adjusted locally and globally (see (a) and (b)), and surfaces can also interpolate the control mesh locally and globally (see(c) and (d)).

## References

- [1] MAINAR, E., PEÑA, J.M., SÁNCHEZ-REYES, J., Shape preserving alternatives to the rational Bézier model, *Computer Aided Geometric Design*, 18 (2001), 37–60.
- [2] ZHANG, J.W., C-curves: An extension of cubic curves, *Computer Aided Geometric Design*, 13 (1996), 199–217.
- [3] ZHANG, J.W., Two different forms of C-B-splines, *Computer Aided Geometric Design*, 14 (1997), 31–41.
- [4] MAINAR, E., PEÑA, J.M., A basis of C-Bézier splines with optimal properties, *Computer Aided Geometric Design*, 19 (2002), 291–295.
- [5] HOFFMANN, M., LI, Y.J., WANG, G.Z., Paths of C-Bézier and C-B-spline curves, *Computer Aided Geometric Design*, 23 (2006), 463–475.
- [6] KOCH, P.E., LYCHE, T., Exponential B-splines in Tension. In: Chui, C.K., *Schumaker, L.L., Ward, J.D. (Eds.), Approximation Theory VI. Academic Press, New York*, 1989, 361–364.
- [7] LÜ, Y.G., WANG, G.Z., YANG, X.N., Uniform hyperbolic polynomial B-spline curves, *Computer Aided Geometric Design*, 19 (2002), 379–393.
- [8] LI, Y.J., WANG, G.Z., Two kinds of B-basis of the algebraic hyperbolic space, *Journal of Zhejiang University Science A*, 6 (2005), 750–759.
- [9] ZHANG, J.W., KRAUSE, F.-L., ZHANG, H.Y., Unifying C-curves and H-curves by extending the calculation to complex numbers, *Computer Aided Geometric Design*, 22 (2005), 865–883.
- [10] ZHANG, J.W., KRAUSE, F.-L., Extend cubic uniform B-splines by unified trigonometric and hyperbolic basis, *Graphic Models*, 67 (2005), 100–119.
- [11] HOFFMANN, M., JUHÁSZ, I., Modifying the shape of FB-spline curves, *Journal of Applied Mathematics and Computing*, 27 (2008), 257–269.
- [12] WANG, G.Z. FANG, M.E., Unified and extended form of three types of splines, *Journal of Computational and Applied Mathematics*, 216 (2008), 498–508.

- 
- [13] LIN, S.H., WANG, G.Z., Extension of definition interval for C-curves, *Journal of Computer Aided Design and Computer Graphics*, 17 (2005), 2281–2285 (in Chinese).
- [14] HAN, X.L., Piecewise quartic polynomial curves with a local weight parameter, *Journal of Computational and Applied Mathematics*, 195 (2006), 34–45.
- [15] FARIN, G., *Curves and Surfaces for Computer Aided Geometric Design*, 4th ed, Academic Press, San Diego, 1997, CA.

**Jin Xie**

Department of Mathematics and Physics, Hefei University, Hefei 230601, China  
e-mail: hfuuxiejin@126.com

**Jieqing Tan**

School of Mathematics, Hefei University of Technology, Hefei 230009, China  
e-mail: jieqingtan@yahoo.com.cn

**Shengfeng Li**

Department of Mathematics and Physics, Bengbu College, Bengbu 233000, China  
e-mail: lsf7679@yahoo.com.cn





# Methodological papers



# Solving certain quintics

Raghavendra G. Kulkarni

Bharat Electronics Ltd., India

*Submitted 1 July 2010; Accepted 26 July 2010*

## Abstract

In this paper we present a simple method for factoring a quintic equation into quadratic and cubic polynomial factors by using a novel decomposition technique, wherein the given quintic is compared with the another, which deceptively appears like a sextic equation.

## 1. Introduction

From the works of Abel (1826) and Galois (1832), we know that a general quintic equation can not be solved in radicals [1, 2]. With some condition imposed on it, the quintic becomes solvable in radicals, and is aptly called solvable quintic equation. In this paper we present a very simple method for solving certain type of solvable quintic equations. The method converts given quintic equation into a decomposable quintic equation in an elegant fashion. The condition to be satisfied by the coefficients of the quintic so that it becomes solvable is derived. We discuss the behavior of roots of such quintic equations. A procedure to synthesize these quintics is given. We solve one numerical example using the proposed method at the end of the paper.

## 2. The proposed method

We know that in an  $N$ -th degree polynomial equation, the  $(N - 1)$ -th term can be eliminated by suitable change of variable. Therefore without loss of generality, we consider the following reduced quintic equation:

$$x^5 + a_3x^3 + a_2x^2 + a_1x + a_0 = 0, \quad (2.1)$$

for solving by the proposed method, where the coefficients,  $a_0$ ,  $a_1$ ,  $a_2$ , and  $a_3$ , are real. Let us consider another quintic equation (which deceptively appears like a sextic equation!) as shown below:

$$\frac{1}{4b_2} [(x^3 + b_2x^2 + b_1x + b_0)^2 - (x^3 - b_2x^2 + c_1x + c_0)^2] = 0, \quad (2.2)$$

where  $b_0$ ,  $b_1$ ,  $b_2$ ,  $c_0$ , and  $c_1$  are unknowns to be determined, and  $b_2 \neq 0$ . Notice that the term inside the square bracket in the above expression is in the form of  $A^2 - B^2$ , hence the expression (2.2) can be split into two factors (quadratic and cubic) as shown below.

$$\left[ x^2 + \left( \frac{b_1 - c_1}{2b_2} \right) x + \left( \frac{b_0 - c_0}{2b_2} \right) \right] \left[ x^3 + \left( \frac{b_1 + c_1}{2} \right) x + \left( \frac{b_0 + c_0}{2} \right) \right] = 0 \quad (2.3)$$

Therefore our aim is to represent the given quintic (2.1) in the form of (2.2), so that it can be easily decomposed as shown in (2.3). To achieve this, the coefficients of quintic (2.1) are to be equated with that of quintic (2.2). However since the coefficients of (2.2) are not explicitly written, we expand and rearrange the the expression (2.2) in the descending powers of  $x$ , as shown below.

$$x^5 + \left[ \frac{b_1 - c_1}{2b_2} \right] x^4 + \left( \frac{b_0 - c_0 + b_2(b_1 + c_1)}{2b_2} \right) x^3 + \left[ \frac{b_1^2 - c_1^2 + 2b_2(b_0 + c_0)}{4b_2} \right] x^2 + \left( \frac{b_0b_1 - c_0c_1}{2b_2} \right) x + \left[ \frac{b_0^2 - c_0^2}{4b_2} \right] = 0 \quad (2.4)$$

Now, equating the coefficients of (2.1) and (2.4), we obtain five equations in five unknowns,  $b_0$ ,  $b_1$ ,  $b_2$ ,  $c_0$ , and  $c_1$ , as shown below.

$$b_1 - c_1 = 0 \quad (2.5)$$

$$b_0 - c_0 + b_2(b_1 + c_1) = 2a_3b_2 \quad (2.6)$$

$$b_1^2 - c_1^2 + 2b_2(b_0 + c_0) = 4a_2b_2 \quad (2.7)$$

$$b_0b_1 - c_0c_1 = 2a_1b_2 \quad (2.8)$$

$$b_0^2 - c_0^2 = 4a_0b_2 \quad (2.9)$$

Employing the elimination method, we attempt to determine the unknowns using above equations (2.5)–(2.9). Using (2.5) we eliminate  $c_1$  from equations (2.6), (2.7), and (2.8) leading to following new equations respectively.

$$b_0 - c_0 + 2b_1b_2 = 2a_3b_2 \quad (2.10)$$

$$b_0 = 2a_2 - c_0 \quad (2.11)$$

$$b_1(b_0 - c_0) = 2a_1b_2 \quad (2.12)$$

Using (2.11) we eliminate  $b_0$  from (2.9), (2.10), and (2.12) resulting in the following expressions.

$$c_0 = a_2 - \frac{a_0 b_2}{a_2} \quad (2.13)$$

$$a_2 - c_0 + b_1 b_2 = a_3 b_2 \quad (2.14)$$

$$b_1(a_2 - c_0) = a_1 b_2 \quad (2.15)$$

Now, we use (2.13) to eliminate  $c_0$  from (2.14) and (2.15) and obtain the following new expressions.

$$b_1 = a_3 - \frac{a_0}{a_2} \quad (2.16)$$

$$b_1 = \frac{a_1 a_2}{a_0} \quad (2.17)$$

Notice an interesting situation here! We are now left with two equations (2.16) and (2.17), and both are expressions for the unknown  $b_1$ . Eliminating  $b_1$  from (2.16) using (2.17) leaves us with an expression, which contains only the coefficients of given quintic (2.1) as shown below.

$$a_1 = \frac{a_0 a_3}{a_2} - \frac{a_0^2}{a_2^2} \quad (2.18)$$

Note that at this stage we have exhausted all the equations, and the unknown  $b_2$  is yet to be determined. It appears that we have hit a dead end in the pursuit of decomposition of quintic. After thinking a while, we note that what really required to be determined are the coefficients of quadratic and cubic polynomial factors in (2.3), and therefore we attempt to find expressions for these coefficients. For this purpose, the expression (2.3) is rewritten as,

$$(x^2 + d_1 x + d_0)(x^3 + e_1 x + e_0) = 0, \quad (2.19)$$

where  $d_0$ ,  $d_1$ ,  $e_0$ , and  $e_1$  are given by,

$$d_0 = \frac{b_0 - c_0}{2b_2}, \quad (2.20)$$

$$d_1 = \frac{b_1 - c_1}{2b_2}, \quad (2.21)$$

$$e_0 = \frac{b_0 + c_0}{2}, \quad (2.22)$$

$$e_1 = \frac{b_1 + c_1}{2}. \quad (2.23)$$

Using (2.12) and (2.17) we evaluate  $d_0$  as:  $d_0 = a_0/a_2$ . From (2.5) we note that  $d_1 = 0$ . From (2.11),  $e_0$  is determined as:  $e_0 = a_2$ . Again using (2.5) we determine  $e_1$  as:  $e_1 = a_1 a_2/a_0$ . Thus all the coefficients in (2.19) are determined and there is no need to determine the unknown  $b_2$ . When each of the polynomial factors in (2.19) is equated to zero and solved, we obtain all the five roots of the given quintic equation (2.1).

### 3. A discussion on such solvable quintic

The expression (2.18) is the condition for the coefficients of quintic (2.1) to satisfy in order that the quintic becomes solvable. Such solvable quintics can be synthesized by determining the coefficient  $a_1$  using expression (2.18) from the remaining real coefficients,  $a_0$ ,  $a_2$ , and  $a_3$ , which are chosen arbitrarily. In the numerical example given (at the end of the paper) we first synthesize the quintic equation and then solve it to determine the roots. How do the roots of such quintic behave? To find out the answer, we express the decomposed quintic (2.19) as below (using the expressions for the coefficients of quadratic and cubic polynomial factors).

$$[x^2 + (a_0/a_2)][x^3 + (a_1a_2/a_0)x + a_2] = 0. \quad (3.1)$$

From the above expression it is clear that the sum of roots of quadratic factor is zero. This automatically sets the sum of roots of cubic factor to zero since the sum of roots of quintic (2.1) is zero as  $x^4$  term is missing in (2.1).

### 4. Numerical example

Let us synthesize solvable quintic proposed in this paper. Consider the quintic equation as below.

$$x^5 - 18x^3 + 30x^2 + a_1x + 30 = 0 \quad (4.1)$$

The coefficient  $a_1$  is determined from (2.18) as:  $-19$ . The coefficient in the quadratic factor  $d_0$  is evaluated as 1, and the coefficients in the cubic factor,  $e_0$  and  $e_1$ , are determined as 30 and  $-19$  [see expression (3.1) for the factored quintic]. Thus the factored quintic is expressed as:

$$(x^2 + 1)(x^3 - 19x + 30) = 0.$$

Equating each factor in the above quintic to zero and solving, we determine the roots of quintic (4.1) as:  $\pm i, 2, 3, -5$ , where  $i = \sqrt{-1}$ .

**Acknowledgements.** The author thanks the management of Bharat Electronics Ltd., Bangalore for supporting this work.

### References

- [1] R. BRUCE KING, Beyond quartic equation.
- [2] ERIC W. WEISSTEIN, Quintic Equation, From MathWorld-A Wolfram Web Resource, <http://mathworld.wolfram.com/QuinticEquation.html>.

**Raghavendra G. Kulkarni**

Senior Deputy General Manager (HMC-D & E)

Bharat Electronics Ltd., Jalahalli Post, Bangalore-560013, India

Member-MAA, Life Member-Ramanujan Mathematical Society

Senior Member-IEEE.

Phone: +91-80-22195270, Fax: +91-80-28382738.

e-mail: [rgkulkarni@ieee.org](mailto:rgkulkarni@ieee.org)





# Spatial Ability, Descriptive Geometry and Dynamic Geometry Systems

Rita Nagy-Kondor

Faculty of Engineering  
University of Debrecen

*Submitted 25 November 2009; Accepted 22 August 2010*

## Abstract

Dynamic Geometry Systems allow new opportunities for the teaching of geometry and descriptive geometry. These systems make possible to create dynamic drawings quickly and flexibly. In the University of Debrecen Faculty of Engineering we executed a controlgrouped developing research for two years, one of them was at Descriptive geometry with participating first year full-time Mechanical engineer students and the other one was at Technical representation practice, in two-two practical groups, for trying out a teaching-learning strategy. We taught one of the groups with the help of Dynamic Geometry System, the other one traditionally, with the paper-and-pencil method. In this paper, I report on our experiences of this course.

*Keywords:* Spatial ability, descriptive geometry, dynamic geometry.

## 1. Introduction

Descriptive Geometry provides training for students' intellectual capacity for spatial perception and it is therefore important for all engineers, physicians and natural scientists. "Descriptive Geometry is a method to study 3D geometry through 2D images thus offering insight into structure and metrical properties of spatial objects, processes and principles" [19]. Moreover some basic differential-geometric properties of curves and surfaces and some analytic geometry are included and one aim is also to develop the students' problem solving ability [20].

The most important ability in working with Descriptive Geometry are the ability to perform operations on the basis of definitions and the spatial ability. We get most of our knowledge in a visual way so it is very important for us how much we are aware of the language of vision.

Spatial ability for engineering students is very important, which decides of the future career. These abilities are not determined genetically, but rather a result of a long learning process. The definition of spatial ability according to Séra and his colleagues [18] “the ability of solving spatial problems by using the perception of two and three dimensional shapes and the understanding of the perceived information and relations” - relying on the ideas of Haanstra and others [4].

Séra and his colleagues [18] are approaching the spatial problems from the side of the activity. The types of exercises:

- projection illustration and projection reading: establishing and drawing two dimensional projection pictures of three dimensional configurations;
- reconstruction: creating the axonometric image of an object based on projection images;
- the transparency of the structure: developing the inner expressive image through visualizing relations and proportions;
- two-dimensional visual spatial conception: the imaginary cutting up and piecing together of two-dimensional figures;
- the recognition and visualization of a spatial figure: the identification and visualization of the object and its position based on incomplete visual information;
- recognition and combination of the cohesive parts of three-dimensional figures: the recognition and combination of the cohesive parts of simple spatial figures that were cut into two or more pieces with the help of their axonometric drawings;
- imaginary rotation of a three-dimensional figure: the identification of the figure with the help of its images depicted from two different viewpoints by the manipulation of mental representations;
- imaginary manipulation of an object: the imaginary following of the phases of the objective activity;
- spatial constructional ability: the interpretation of the position of three-dimensional configurations correlated to each other based on the manipulation of the spatial representations;
- dynamic vision: the imaginary following of the motion of the sections of spatial configuration.

The link between engineering students’ spatial ability and their success in a range of engineering courses is very important. Mental Cutting Test (MCT) is one of the most widely used evaluation method for spatial abilities. Németh and Hoffmann [14] presented an analysis of MCT results of first-year engineering students, with emphasis on gender differences. They used the classical MCT test for

first-year engineering students of Szent István University. Németh, Sörös and Hoffmann [15] attempted to find possible reasons of gender difference, concluding, that typical mistakes play central role in it. They show typical mistakes can be one of the possible reasons, since female students made typical mistakes in some cases more frequently than males. In accordance with the international experiences, they observed relevant improvement after descriptive geometry courses. Williams and his colleagues' paper [24] and others [10] report on research into the spatial abilities of engineering students, too. MCT and similar tests have been widely studied in the following papers: [3, 5, 17, 21, 22, 23].

One of the programs, that supports computer-aided descriptive geometry was developed by a Hungarian expert and helps the teacher to explain the theory and practice of the Monge projection, the reconstruction of the spatial objects in the mind and, with the help of interactive feature, to understand spatial relationships [8]. Designs can be saved in BMP format.

At the University of Debrecen, Faculty of Engineering, we can experience that the basic studies have their difficulties: there are huge differences among the pre-education level of the students, the number of lessons is continuously decreasing and education becomes multitudinous. In our college, full time engineer students have a 2 hour seminar and a 2 or 1 hour lecture in every course from descriptive geometry. During that period of time they should pick up the elements of Monge-projection to the interpenetration of flat bodies and the curvilinear surfaces. (The syllabus differs according to their major.)

The interest, the pre-knowledge and motivation of the students are very different. One of the problems of the traditional teaching is that these problems can not be easily managed. But the use of computer tools makes it possible that each and every student can proceed in his own speed, so they do not lag behind and they do not get bored. The student can plan his/her own pace of learning and the speed of development.

This article reports about our experiences and results of descriptive geometry course.

## 2. Tasks with Dynamic Geometry Systems

Literature suggests that Dynamic Geometry Systems (DGS) is a valuable tool to teach geometry in schools [1, 2, 6, 7, 9, 16]. These systems are not only complement static geometrical figures, but also the software stores construction steps throughout its use and objects can be treated as dynamic figures. In this way when parts of figures are altered then this change also modify the entire figure structure. Thus, students can follow how elements of figures are built on one another.

Laborde [10] classified these tasks according to their role that the designer of the task attributes to Cabri (another type of DGS) and to the expected degree of change. The four type of roles:

- DGS is used mainly as a facilitating material, while aspects of the task are

not changed conceptually.

Our example: Figure 1 shows the construction of a worksheet and Figure 2 shows the right solution. (Figure 1 and Figure 2 - Created with *Cinderella*.) (Interactive worksheet 1 - in our phrasing.)

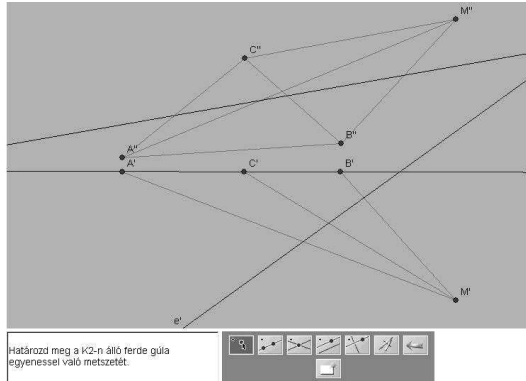


Figure 1: Construction of a worksheet

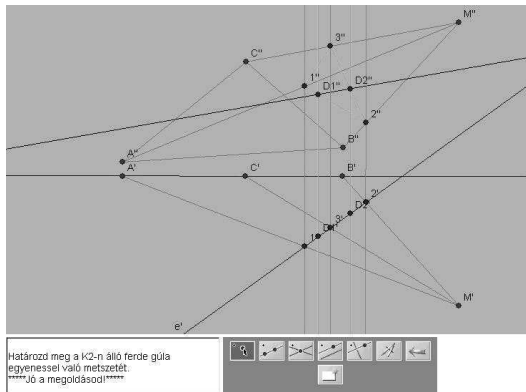


Figure 2: The right solution

- The task itself takes its meaning from DGS (for example Black-Box tasks), with DGS construction tools and dynamic features.

Our example is Pyramid's plane section. (Figure 3 - Created with *Cinderella*.) (Interactive worksheet 2 - in our phrasing.)

The pictures of the Figure 4 show the use of the program's dynamic features in descriptive geometry. On the left side moving the point P to the right side's projection picture we can trace back the representation of the picture

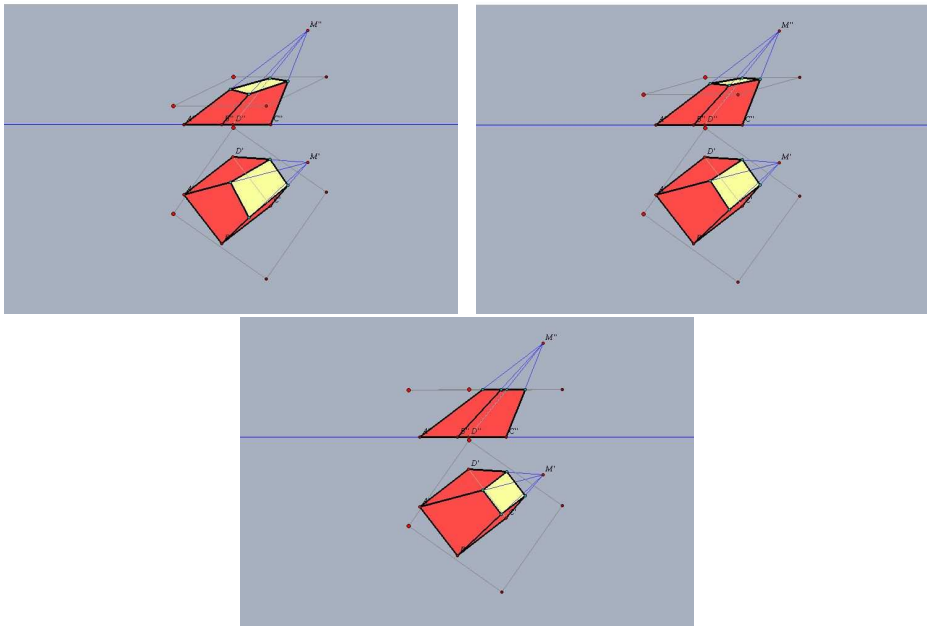


Figure 3: Pyramid's plane section

if our point is at the I., II., III. or IV. spatial quarter. (Figure 4 - Created with *GeoGebra*.)

- DGS is used as a visual amplifier (static figures).

Our example: There are two planes, the first with the ABC triangle, the second with the 123 triangle, from which we cut out the 456 triangle. Construct the intersection of the two planes. To state the visibility you should consider the holes. (Figure 5 - Created with *Cinderella*.)

- DGS is supposed to modify the solving strategies of the task, with some construction tools of DGS.

So by means of movement it can be observed how figures are constructed upon each other, as well as the construction process itself [7]. If a constructed figure in the drag mode does not keep the shape that was expected, it means that the construction process must be wrong [10]. Looking at how students used dragging provided an insight into their cognitive processes [1].

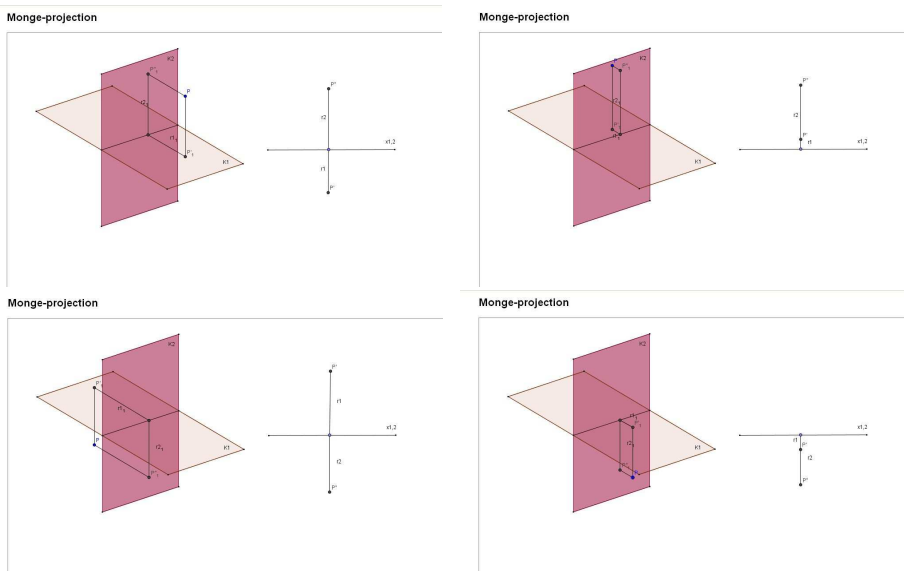


Figure 4: Representation of a point

### 3. Experiments

In the University of Debrecen Faculty of Engineering we executed a controlgrouped developing research in two semesters, it was at Descriptive geometry with participating first year full-time Mechanical engineer students, for trying out a teaching-learning strategy. We taught one of the groups with the help of DGS, the other one traditionally, with the paper-and-pencil method. We carried out the educational research with 80 first year full-time Mechanical engineer students at Descriptive geometry practice, in two-two practical groups.

In the University of Debrecen, Faculty of Engineering the students selected for the engineering programme acquire the basics of the Descriptive geometry - the elements of the Monge projection - in the course of a 2-hour lecture and a 2-hour seminar each week, which they use later in their professional subjects. From the two seminars we held, one group worked with DGS and interactive whiteboard, while the other group did constructions in the traditional way with paper and pencil for two years. The tests were paper-and-pencil tests, even for members of the group that had been working with the computers. The tasks are the traditional paper and pencil tasks. It does not contain theoretical question but practical ones.

Our goals:

- To meet the curriculum requirements.
- Increasing the understanding of the Descriptive geometry.

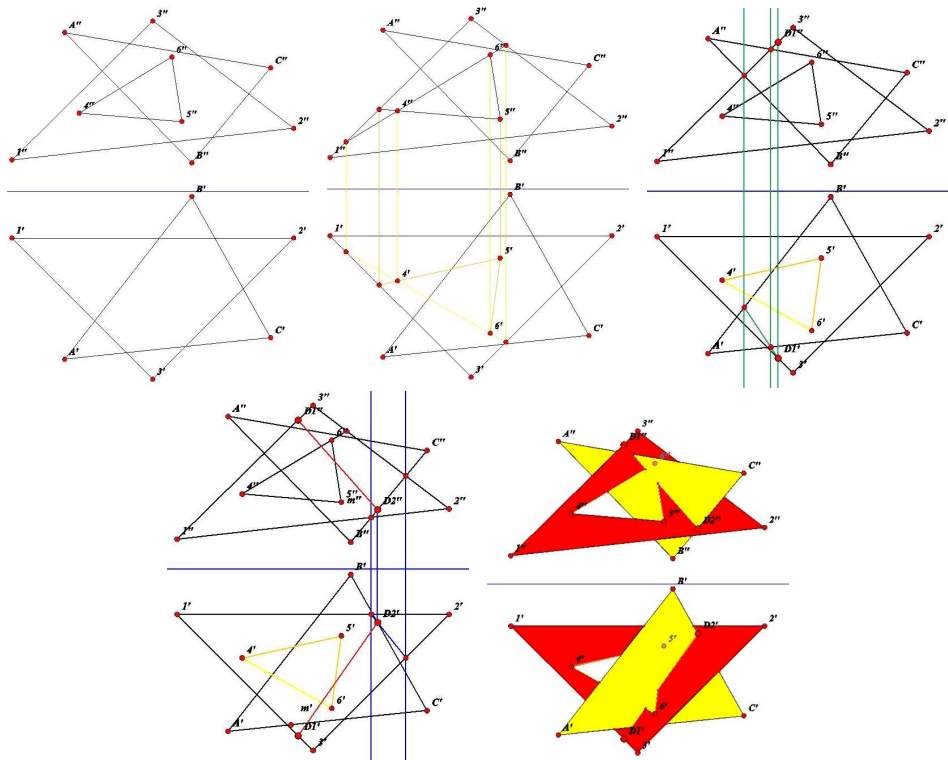


Figure 5: Intersection of planes

- To meet the mass education demands.
- To get prepared one part of our students to get into the university level.

During the semester the lecture went on without using a computer. The two seminar groups had only one computer room, so the first group worked with DGS while the second worked with the traditional paper and pencil methods. We paid special attention to make sure that all the two groups have the same tasks and they got the same paper and pencil homework. The difference is in the drawing opportunities of the program. The solution strategies of both tasks do not differ essentially. These are the peculiarities of the first type of exercises made by Laborde [10]. We tried to organize the practises in such a way that neither approach—whether teaching aided by the DGS, or teaching using paper-and-pencil had an advantage. Thus we hoped to achieve reliable measurements of relative ability. DGS that we chose for practice is able to save all the constructions as an interactive webpage. So there is an opportunity to make worksheets to practice constructions. The teacher can adjust the set of starting objects and desired objects. After that he/she can save

the task with a limited use of geometric tools. Since we do not have to adjust the proceedings of the solution, we just have to adjust the set of the desired objects, so the program accepts more than one approach to the good solution. You can attach guidance or construct help to the worksheet. There is no need to install the program, to make the interactive worksheet; you should only attach a special file to the web page besides the tasks. Using this opportunity the students could work online on the seminars. To the education of the Descriptive geometry that is using DGS we made a webpage made up of Cinderella worksheets, involves the material of the practises. We tried out this curriculum system throughout two years and we continuously examined its efficiency. The curriculum system processed by us, which was suitable for teaching the Descriptive geometry according to the experiences of the 2004/2005 school year we modified and revised it in the 2006/2007 school year.

In the preliminary phase the students' levels of knowledge was examined by measurement of spatial ability. At the beginning of the semesters we examined whether there is a significant difference between the spatial ability of students in their preliminary basic knowledge of descriptive geometry. The measurement of the students' preliminary knowledge took place in the first teaching week. The exercises can be categorised under the following headings [18]:

- imaginary manipulation of an object,
- imaginary rotation of a solid,
- projection description and projection reading,
- reconstruction.

By the preliminary survey it can be seen that the two groups achieved nearly the same. The results of this test are presented in [11].

## 4. Results

To measure the efficiency of the teaching-learning process during the semester, there were two tests and one delayed test, consisting of practical exercises, which was taken by the students four months after the semester; all of which we rated with a score [12, 13]. The comparative survey of the results is based on these tests.

The test of the computer group was more punctual, a little more precise and it was better in both years. But the determination of transparency in 2004 was wrong more often by them than by the students of the paper-and-pencil group. In 2006 from learning this we paid larger attention for practising the determination of transparency in the computer group. Based on both tests we can say that the computer-aided group carried out the acquirement of the legally given educational requirement better than the control group. Figure 6 shows the result of the two tests and the delayed test.

In the traditional, paper-and-pencil first and second tests can be observed that the students of the computer-aided group perform better than the students of the



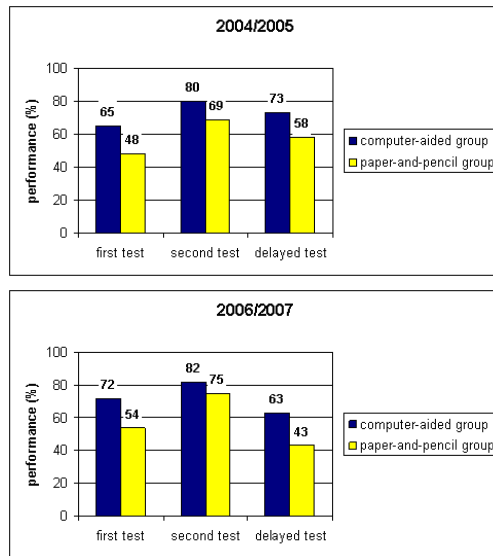


Figure 6: Results of two tests and delayed test

paper-and-pencil group. The difference between the performances in the first test was 17% in 2004, 18% in 2006. The difference between the performances in the second test was 11% in 2004, 7% in 2006.

As we compare the result of the delayed test with the tests we see that the performance of the paper-and-pencil group by all three tests of both years stayed under the performance of the computer group. In 2007 on the delayed test both groups performed worse than in 2005. The reason for this can be that the weekly number of the lecture decreased from two to one and also the fact that in 2007 they wrote the delayed test more than a month after ending the education comparing to 2005.

It cleared out from the answers to the questionnaire, and also during the conversations with the students that they liked that they could work with computer, they found constructions easier in this way. In average they visited the webpage from home used at practice once a week.

Based on the tests we can say that we can reach quality improving with using DGS. Organizing the education by computer takes much more time of the teacher, the effective usage of DGS requires continuous developing work, but the results of the tests show that the invested work returns.

In the computer group it was more typical that the students helped each other, corrected their mistakes. Experimentation was more typical for them as well, as the faulty elements could be hidden without any sign with a mouse click. The members of the paper-and-pencil group waited for the teacher's help, instruction when they stuck in their work. So the computer inspired the students for separateness. We

found that the testing phase at the traditional, paper-and-pencil group was often missing.

As an effect of introducing the worksheets made by DGS into education the motivation level of learning increased, the worksheets are helpful for large percentage of the students.

## 5. Summary

The aim of our educational research was to introduce the worksheets made by DGS into the education of the Descriptive geometry of the mechanical engineer students. To reach this aim, after surveying of the literature of the spatial ability, the computer-aided education, the DGS and Mathematics Didactics, we executed a controlgrouped developing research for trying out the new educational method. At the creating and trying out of the curriculum made up of the DGS generated worksheets that include the material of the practises we took into consideration the offers of the literature and we modified and corrected the curriculumsystem according to the experiences of the first school year.

We may assert on the basis of these results that use of the computer and the use of interactive worksheets provided by DGS increases success and helps to create a proper conceptual structure. The computer-aided seminar helps the effectiveness of teaching, with the help of the interactive worksheets we can improve the student's problem-solving abilities and improvement in the field of creativity was observed among the students. On the seminar we could more easily trace the thoughts of the students [12].

Direction by the teacher is very important even in case of using DGS. If the software is simply made available, the program might become an obstacle to the transition from empirical to theoretical thinking, as it allows the validating of a proposition without the need to use a theory [1].

According to the current experiments, task of the future is to improve, develop and correct worksheets and with the help of all of these sheets we could make the tasks more efficient in the future.

## References

- [1] ARZARELLO, F., OLIVERO, F., PAOLA, D., ROBUTTI, O., A cognitive analysis of dragging practises in Cabri environments, *ZDM*, 34(3) (2002) 66-72.
- [2] BOKAN, N., LJUCOVIC, M., VUKMIROVIC, S., Computer-Aided Teaching of Descriptive Geometry, *Journal for Geometry and Graphics*, 13 (2009) 221-229.
- [3] GORSKA, R., Spatial imagination - an overview of the longitudinal research at Cracow University of Technology, *Journal for Geometry and Graphics*, 9 (2005) 201-208.
- [4] HAANSTRA, F. H., Effects of art education on visual-spatial and aesthetic perception: two meta-analysis, *Rijksuniversiteit Groningen*, Groningen (1994)

- [5] JUSCAKOVA, Z., GORSKA, R., A pilot study of a new testing method for spatial abilities evaluation, *Journal for Geometry and Graphics*, 7 (2003) 237-246.
- [6] KAUFMANN, H., Dynamic Differential Geometry in Education, *Journal for Geometry and Graphics*, 13 (2009) 231-244.
- [7] KORTENKAMP, U. H., Foundations of Dynamic Geometry, Ph.D. thesis, *Swiss Federal Institute of Technology Zürich*, (1999)
- [8] KOVÁCS, E., HOFFMANN, M., Computer Aided Teaching of Descriptive Geometry *Conference on Applied Informatics, Eger-Noszvaj*, (1997) 179-183.
- [9] LEOPOLD, C., GORSKA, R. A., SORBY, S. A., International Experiences in Developing the Spatial Visualization Abilities of Engineering Students, *Journal for Geometry and Graphics*, 5(1) (2001) 81-91.
- [10] LABORDE, C., Integration of technology in the design of geometry tasks with Cabri-geometry, *International Journal of Computers for Mathematical Learning*, 6 (2001) 283-317.
- [11] NAGY-KONDOR, R., Spatial ability of engineering students, *Annales Mathematicae et Informaticae*, 34 (2007) 113-122.
- [12] NAGY-KONDOR, R., The results of a delayed test in Descriptive Geometry, *The International Journal for Technology in Mathematics Education*, 15(3) (2008) 119-128.
- [13] NAGY-KONDOR, R., Using dynamic geometry software at technical college, *Mathematics and Computer Education*, Fall, (2008) 249-257.
- [14] NÉMETH, B., HOFFMANN, M., Gender differences in spatial visualization among engineering students, *Annales Mathematicae et Informaticae*, Vol. 33 (2006), 169-174.
- [15] NÉMETH, B., SÖRÖS, Cs., HOFFMANN, M., Typical mistakes in Mental Cutting Test and their consequences in gender differences, *Teaching Mathematics and Computer Science*, 5(2) (2007) 385-392.
- [16] PRIETO, G., VELASCO, A. D., ARIAS-BARAHONA, R., ANIDO, M., NÚÑEZ, A.-M., CÓ, P., Training of Spatial Visualization Using Computer Exercises, *Journal for Geometry and Graphics*, 14 (2010) 105-115.
- [17] SAITO, T., SHIINA, K., SUZUKI, K., JINGU, T., Spatial Ability Evaluated by a Mental Cutting Test, *Proc. 7th International Conference on Engineering Computer Graphics and Descriptive Geometry, Cracow, Poland*, (1996) 569-573.
- [18] SÉRA, L., KÁRPÁTI, A., GULYÁS, J., A térszemlélet, *Comenius Kiadó, Pécs* (2002)
- [19] STACHEL, H., Descriptive Geometry, the Art of Grasping Spatial Relations, *In K. Suzuki and K. Yoshida (eds), Proc. 6th Internat. Conf. on Engineering Computer Graphics and Descriptive Geometry, Tokyo*, 2 (1994) 533-535.
- [20] STACHEL, H., What is Descriptive Geometry for? <http://citeseer.ist.psu.edu/642381.html>, (2004)
- [21] TAKEYAMA, K., MAEGUCHI, R., CHIBANA, K., YOSHIDA, K., Evaluation of Objective Test using a pair of orthographic projections for descriptive geometry, *Journal for Geometry and Graphics*, 3 (1999) 99-109.
- [22] TSUTSUMI, E., Mental Cutting Test using drawings of intersections, *Journal for Geometry and Graphics*, 8 (2004) 117-126.

- [23] TSUTSUMI, E., SHINA, K., SUZAKI, A., YAMANOUCHI, K., TAKAAKI, S., SUZUKI, K., A Mental Cutting Test on female students using a stereographic system, *Journal for Geometry and Graphics*, 3 (1999) 111-119.
- [24] WILLIAMS, A., SUTTON, K., ALLEN, R., Spatial Ability: Issues Associated with Engineering and Gender, *Proceedings of the 2008 AaeE Conference, Australia*, (2008) 1-6.

**Rita Nagy-Kondor**

H-4028 Debrecen

Ótetmető u. 2-4.

Hungary

e-mail: [rita@mk.unideb.hu](mailto:rita@mk.unideb.hu)

# C++ exam methodology

Norbert Pataki, Zalán Szűgyi

Department of Programming Languages and Compilers  
Eötvös Loránd University

*Submitted 20 October 2009; Accepted 9 March 2010*

## Abstract

The C++ programming language supports multiparadigm programming. We can write programs in procedural, object-oriented, generic way at the same time.

However, it is difficult to figure out exercises for the terminal examinations since not easy to separate the algorithmic cogitation from the knowledge of the programming language. There are some basic elements that programmer students have to know: constructors, parameter passing, objects, inheritance, standard library, handling constants, copying objects, functions and member functions, etc. Exercises must be multiparadigm according to the C++ language. Using only one paradigm in C++ is not enough. This results in that we have to distinguish the different linguistic constructs on the basis of its complexity.

Many questions are arisen in connection with the exercises of terminal examinations. How can we gauge the procedural, the object-oriented, and the generic paradigms at the same time? How can we gauge students' C++ knowledge when we do not lay stress on the algorithmic cogitation? What kind of exercises may be interesting by the Standard Template Library? Which C++ constructs are reckoned to be more difficult and which ones considered to be easier? What are the most important ones? In this paper we give answers to the previous questions, we describe our methodology to assessment of students' C++ knowledge in a semi-automatic grading way. We also present exercise examples that worked out according to our methodology. We take stock of students' results in the paper.

*Keywords:* C++, exam, teaching, multiparadigm programming

*MSC:* 68N19

# 1. Introduction

In software technology a *paradigm* represents the directives in creating abstractions [21]. The paradigm is the principle by which a problem can be comprehended and decomposed into manageable components. A paradigm directs us in identifying the elements in which a problem will be decomposed. The paradigm sets up the rules and properties, but also offers tools for developing applications.

C++ is usually considered as an object-oriented programming language, but it is not completely true. C++ supports *multiparadigm programming* [24]. Structured programming features come from C legacies with better parameter-passing opportunities and features of overloading. Classes may be created in a sophisticated way, for example the C++ programming language distinguishes between three different variants of inheritance based on access control. Templates are also supported. Generic and generative programming have become available with C++'s template construct. The *C++ Standard Template Library* (STL) was the very first library based on generic programming and its usage is similar to the functional programming approach [2], [18].

C++ is considered as a language that hard to teach. C legacies must be known because of their hazard but Stroustrup argues for a use of C++ as a higher-level language that relies on abstraction to provide elegance without loss of efficiency compared to lower-level styles [23]. Many paradigms and approaches should be taught at the same time. By the way, C++'s standard library is roomy. Standard Template Library (STL) includes more than sixty algorithms and seven actual containers and three adaptors. STL is just a part of the standard library.

Multiparadigm software design and its implementation in the C++ programming language are deeply investigated by James Coplien [7]. One of his most important conclusions is that different kind of domain problems should be targeted using different programming paradigms. The domain analysis, especially identifying positive and negative variability helps to select the most appropriate paradigm.

We work with only standard C++, so we do not deal with multithreaded C++ programs, sockets, graphical user interfaces. This can ease the teaching process as well as the examination. For example, we do not have to work with graphical forms, inputs and outputs which could not be integrated to our framework easily.

However, gauging students' C++ knowledge is much more harder. Students' attainments must be examined from many aspects.

Many elementary constructs can be found in the C++ programming language. All students should know these features: functions, classes, methods, templates. Students must use these constructs in a sophisticated way. Constructs like parameter passing, constructors, constants, copying objects, inheritance are also very important. Contrarily, importance of algorithmic cogitation should be minimalized because we gauge the C++ knowledge.

Teaching the standard library is important in a C++ programming language course [25]. Therefore the students can use the STL when writing exams. Many exercises are unusable, like lists, maps, vectors, etc. without significant modification

in their specifications.

In Hungary, a five-point grade system is used. 1 is the failing grade and 5 is the best possible grade. C++'s constructs should be reflected in this grade system. Which constructs the students must know, and which ones are more difficult? Which constructs are the most weighty ones?

In this paper we describe our methodology to assessment of students' C++ knowledge. We present the structure of former exercises that able to grading students in a semi-automatic way and an archetypal exercise is detailed.

This paper is organized as follows. In section 2 we describe the general conditions and introduce the frame of exercises. In section 3 we detail a specific exercise. Other ideas are presented briefly in section 4. We give a brief overview about our experiences in section 5. We analyze the students' results in section 6. Finally, we conclude our issues in section 7.

## 2. Exams in a nutshell

In this section we describe the general circumstances in connection with the exams.

Students have to write exams in a computer lab. They may use their books, notes and the world wide web, but they have to work alone. The exams last about 3 and a half hours. An experienced C++ programmer is able to solve these problems within half an hour.

Five different grades are distinguished in the Hungarian education. The grades denote in numbers from 1 to 5, where 1 means unsatisfactory and 5 means first.

The exercise is typically the implementation of a class template, with many member functions. Students receive the client code that instantiates the template. The client describes the specification of the class template as a sequence of use cases. For the pass grade the students have to implement some base functionality of class template. For better marks they need to implement more and more functions. At the beginning all the code of functional tests are commented. When the students implement all the necessary functions for a given mark they can uncomment the corresponding part of client code to see whether their work were correct or not. On the other hand, these functional and semantical tests are not given proof of the correctness of implementation but usually a very good feedback to the students [15].

The program always prints to display the student's mark, if the program can be compiled. When students download the exercise program, it displays the unsatisfactory mark. In our case students must progress linearly.

The main goal of exercise is usually a template container similar to STL's containers. The representation of the class is not determined. Students can freely choose the representation. The effectiveness is not a primary goal here, however extremely poor design is rejected.

Students have to write a template class with proper template parameters, a trivial constructor. Inserting elements must be supported and a basic information should be obtained from an object and a constant object. Copying object via copy

constructor and assignment operator is usually also needed to the pass mark. We reckon that these constructs are essential ones.

Usually the class must be extended for a better grade. The usual constructs for a fair mark are more difficult methods, like erasing elements, etc..

For a good or an excellent grade `iterator` or `const_iterator` inner type is often required. Iterator objects must work together with the STL algorithms. Students should use the STL containers and iterators to overcome this exercise, because they do not know all necessary members for iterator types.

For a good mark operators that *overloaded on const* are fine. Sometimes usage of polymorphism appears here.

For excellent grade clearly many constructs can be gauged. Special template constructors for any iterator types or template copy constructors are ideal. Sometimes *generic algorithms* are required that are not in the standard, like `copy_if`. These template algorithms must be similar to STL's algorithm. Introduction of a new template parameter with default value is reasonable. Basic template metaprogramming features (like overloading on the returning values according to a template argument) is also proper. We assume that students can take advantage of the STL.

Our method can be applied when the marking conditions are more complex.

The following excerpt describes the general schema of our exercises:

```
#include "work.h"
#include <iostream>

// necessary classes and functions

int main()
{
    int yourMark = 1;
    /* 2
    here we use the basic methods of the class: some use cases
    ...
    if the implementation suits the use cases, variable
    yourMark is increased
    */

    /* 3
    here we use more methods: more uses cases
    ...
    after some basic functional test, value of yourMark is 3
    */

    /* 4
    More difficult methods tested at this block:
    ...
    inasmuch as implementation passes the test, variable
```



```
yourMark is increased
*/

/* 5
Quite difficult methods required in this block:
...
after successful test cases, value of yourMark is 5
*/

std::cout << "Your mark is " << yourMark << std::endl;
return 0;
}
```

Students must use this schema, they are only allowed to uncomment the different parts. However, the linearity is not necessary, but we apply it. The different parts could be independent and at the end of parts variable can be increased one by one.

Students present their solution at the end of the exams. One of the teachers analyzes someone's code and asks the student to make sure cheatless. The teacher gives the grade based on the program's output, but he can give different grade. So, the students do not achieve the program's output as a grade automatically. This is important because the tests are not all-inclusive ones and it could be eluded. The structure of the exam makes much more easier the teacher's work and he or she can focus on the details and it is also a good feedback to the students.

In this section we introduced the general frame of exercises. We categorized the different linguistic constructs to gauge. Hereinafter we paraphrase a specific example that describes an exercise of sorted list template.

### 3. A detailed example

In this example a sorted list container must be implemented which is template. It keeps its elements ordered. Its public interface is quite similar to STL's list container, but STL's list container is not ordered.

The test file includes a functor class that called Compare for a user-defined comparison:

```
#include "sl.h"
#include <deque>
#include <iostream>
#include <string>
#include <numeric>
#include <functional>
#include <vector>
```

```

struct Compare: std::binary_function<int, int, bool>
{
    bool operator()(int a, int b) const
    {
        return a > b;
    }
};

```

Students must working in the file called sl.h. They must know that templates do not compose compilation units.

The following part must work to pass the exam. If this part does not work, then the student fails.

```

/* 2
SortedList<int> li;
SortedList<double> ls;
ls.insert(5.6);
ls.insert(3.2);
li.insert(7);
li.insert(2);
li.insert(5);

const SortedList<int> cli = li;
if (3 == cli.size())
    yourMark = cli.front();
*/

```

Default constructor must be callable. Inserting elements is required, and it should be an actual template: insert must work proper according to the template argument. Creating copy via copy constructor must be supported. This not a problem, if the standard list container is used for representation, because the default copy constructor calls the members' copy constructors to create copies. Furthermore. two more methods must be implemented: the size method that returns how many elements are in the list, and the front method the returns the list's very first element. These two methods called on constant list, so these are const methods according to the features of C++'s constant correctness. The very first element is least element in the ordered list, therefore value of `yourMark` variable will be 2. This part should not be a real challenge for prepared students: it is a very basic linked list. Of course, some students present worser accomplishment because of jitter. We try to help these students with more help or comment.

```

/* 3
li.insert(8);
li.remove(5);
if (7 == cli.back())
    yourMark = cli.size();

```

```
*/
```

Two more methods needed for fair grade: a remove method that erases a given element from the list, and a back method that returns list's last element.

Overload on const is not good in this specific example because an overloaded function would violate the constraint of orderness.

```
/* 4
const int N = std::accumulate(cli.begin(), cli.end(), 0);
yourMark += (14 == N);
*/
```

An iterator type is required for good grade. We call STL's accumulate algorithm with SortedList's iterator. Accumulate adds together the elements in the container. Therefore iterator's proper work is needed, because students cannot modify the accumulate algorithm. If accumulate returns 14, `yourMark` variable is increased. Implementation of an iterator class is not easy because many operators must be implemented and many special members are needed. But when STL's list is used for the representation this implementation is unnecessary, because we can use list's iterator instead of a handcrafted one.

```
/* 5
std::deque<int> d;
d.push_back(2);
d.push_back(1);
d.push_back(3);

const SortedList<int, Compare> lc1(d.begin(), d.end());

std::vector<int> v;
v.push_back(3);
v.push_back(7);
const SortedList<int, Compare> lc2(v.begin(), v.end());

if (7 == lc2.front())
    yourMark = lc1.front() + lc2.size();
*/
```

Two special features needed for the best grade. Arbitrary ordering can be passed as template argument by functor class. All previous code must be compiled with this feature, therefore the new template parameter needs default parameter. With the introduction of this parameter the list's behaviour must remain the same. `std::less<T>` is a standard functor class template to describe the normal behaviour of list ordering. The `operator()` of this template functor class calls the `operator<` of T. Implementation of `less` is quite easy, but can be found in the

STL. This functor class must be the default argument to the new template parameter. Another feature is the special template constructor for arbitrary iterator types. In the example we use this constructor with `vector<int>::iterator` and `deque<int>::iterator`. All standard containers offer this kind of constructor. Nevertheless, overloading is not allowed to overcome this situation.

The functional tests in the previous code fragments do not ensure the correctness of the implementation. However, most problems can be discovered by this method.

This example is not too difficult from the view of algorithmic cogitation, but it is more and more difficult from the view of C++ language. This example presents our conception aright.

## 4. Other examples

We create our exercises according to our methodology. The previous example presents our ideas. We expect an implementation of a template class with ever more difficult features. We keep track the student's grade in a variable. This variable depends on the correct implementation of the exercise.

Many ideas can be found in [18]. The usual exercise is based on STL's flaws. Containers for pointers are not supported by the standard library. Containers of pointers cause many problems (for example, copying is not trivial and avoiding memory leaks).

Another flaw is STL's multimap container does not define the relative order of elements at the same key. A multimap container that defines the relative order of element at the same key is a fine exercise.

STL does not include hashing containers. Hashtables, hashmaps are also ideal containers for exams.

Caching associative containers are similar to the standard associative containers. They are sorted, they can take advantage of sortedness, ensure iterators, but they have a special invariant, their size is limited. If the container would be oversized, it erases the oldest element from the container. Any kind of these containers is good for exams.

Graph types also cannot be found in the standard library. Graphs are worth considering, because they can be gauged in many different ways.

Union of akin containers (for instance set and multiset, or stack and queue) can be worked out. The behaviour of union's container based on a bool template argument.

In this section we present some more examples in a nutshell. These ideas were the basis of former exams.

## 5. Experiences in general

In this section we present our general experiences in connection with students, exercises and C++ itself.

Typically, every kind of grades is achieved. The grades are harmonized with the students' capability. The main approach (selection of the representing object) determines the obtainable grades considerably.

One of the major experiences that STL mightily makes the examination's solution easier. STL allows students concentrating on linguistic challenges. This is the very same experience when STL allows professional programmers concentrating on runtime complexity and different optimizations. The better grades are reached almost only with the STL. For the best mark we assume that students use the STL, and no student can solve the last part without the library.

Strictly speaking, some of the students do not use the standard library, and implement a handcrafted linked list class (or other node-based container). These approaches often fail on small pitfalls. Special analysis tools (like valgrind) are not necessary to avoid memory related bugs. Typically handcrafted containers are makeshift and should be avoided in this situation.

## 6. Quantitative Results

In this section we present the results of students. First, we give an overview about the results of given semester chronologically regardless of resits. We also present the results in graphical way (see Figure 1, Figure 2 and Figure 3). These charts present the number of students who achieved the given grade.

7–10 students failed on every examinations in the examined semester. In addition there were some students who applied for an exam, but did not come.

Twenty-five students gained rather good (excellent and good) grades and eight students failed on the first occasion from forty-six students. Presumably we claimed typical constructs for these marks. Generally, the more talented students come on the very first occasions. However, when a new series of datastructures are introduced we work out a lighter exam (easier member methods with easier algorithms) to focus on the new features. We keep our exams available on the local network.

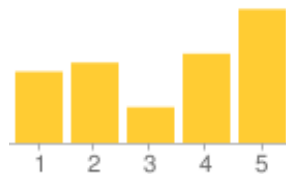


Figure 1: The results on the first time in a semester

The results of second occasion differ from the first one. Fifty-five students came to the exam, nine students failed. Twenty students reached the best grade. The differences became sharper, because the students could prepare for the examination

on the grounds of the previous exam and some of the students practiced with the previous exam, and some of them do not. These two exercises were similar, but the second one was more difficult. This approach results in this far cry.

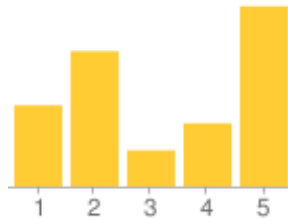


Figure 2: The results on the second time in a semester

Completely different result is yielded on the third time. Figure 3 looks like normal distribution, this result denotes correct exercise: only nine students were able to carry through the exam of fifty-five people, and ten of them cannot do the examination. Most of them – fourteen to be exactly – reached the better grade. The main reason of this incident is that we cannot continue of the previous series but we worked out a new exercise, that had a good difficulty level.

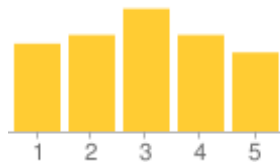


Figure 3: The results on the third time in a semester

Next, we consider results of more than thousand students from the last four years. We compare students' results to their results of *Ada programming language*. This course is similar to ours, but students get the grade in a more classical way, teachers read through the student's code in the course of Ada programming language. Fails are not taken into account.

We divide the students into six groups. The first group is the students, who have rather good grades (excellent or good marks) from C++ as well as from Ada. This group includes 281 students. The second group is the students, who have rather bad grades (pass or fair marks). There are 421 people in this group. The third group includes the students who have rather good grades from C++ but have rather bad grades from Ada. This group includes 192 students. The fourth group is just the opposite of the third one, this group contains the students, who have rather good grades from Ada but have rather bad grades from C++. 119 students are in this group. However, the third and fourth group contain students who have

fair mark from the either of the courses and better from the other one. This is not a significant difference. The students with significant difference can be found in the fifth and sixth group. Students who have excellent mark from C++ but have only pass mark from Ada are in the fifth group. This group contains 32 people. Students who have excellent mark from Ada but have pass grade from C++ are in the sixth group. 18 people belong to this group. The following chart presents these numbers in a graphical format.

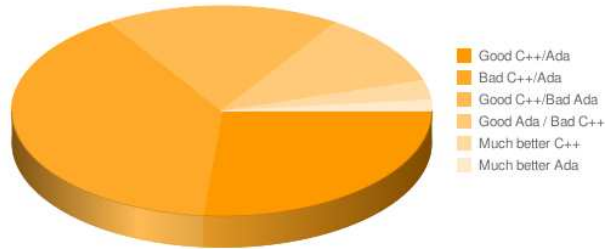


Figure 4: Connection between Ada and C++ grades

These numbers ensure our methodology is fair. About half of the students get similar grades from C++ and Ada but the exam methodology is quite different. Only few students have achieved completely different grades from the two courses. Special prolegomena (e.g. industrial experience) may causes these incidents.

In this section we argued for our methodology in a quantitative manner. We counted how many students achieved different marks in an entire semester. We compare our methodology to an other similar course's methodology. Our methodology has been confirmed by the computation.

## 7. Conclusion

The C++ programming language is difficult to teach and to learn. C++ supports multiparadigm programming. Functions, classes, generative constructs can be used in an orthogonal way.

However, contrive exams is much more harder process. In Hungary a five-grade system is used. We present our methodology based on this grading system. Our methodology supports multiparadigm programming. Our examples take advantage of STL's flaws and supports a semi-automatic grading system. This semi-automatic system means that our client code offers a mark that we check at the end of examination. The offered mark is based on test cases. The test cases are not all-inclusive but we give a feedback to the student as well to the teachers.

We presented our general framework to gauging students' knowledge and a specific example is detailed. We defined classification of C++ constructs based on

their difficulty and essentiality. We outlined students' results in a given semester. We compared the results of our exercises to the results of the course Ada programming language that applies a more classical method. Our charts confirm that our framework is fair.

## References

- [1] ASSASSA, G., MATHKOUR, H., AL-GHAFFEES, B., Automated Software Testing in Educational Environment: A Design of Testing Framework for Extreme Programming, *First National IT Symposium (NITS2006), Bridging the Digital Divide: Challenges and Solutions* (2006).
- [2] AUSTERN, M. H., Generic Programming and the STL, *Addison-Wesley* (1999).
- [3] CARDELLI, L., WEGNER, P., On Understanding Types, Data Abstraction, and Polymorphism, *ACM Computing Surveys*, 17(4), (1985) 471–522.
- [4] CIFUENTES, C., BRANNAN, B., Teaching C/C++ to Computer Science Students with Pascal Programming Experience, *Proceedings of the 1st Australasian conference on Computer science education* (1996) 189–196.
- [5] COLTON, D., FIFE, L., THOMPSON, A., A Web-based Automatic Program Grader, *Information Systems Education Journal*, Vol 4, Number, 114, (2006).
- [6] COLTON, D., FIFE, L., THOMPSON, A., Building a Computer Program Grader, *Information Systems Education Journal*, Vol 3, Number, 6, (2005).
- [7] COPLIEN, J. O., Multi-Paradigm Design for C++, *Addison-Wesley* (1998).
- [8] HARRIS, J. A., ADAMS, E. S., HARRIS, N. L., Making program grading easier: but not totally automatic, *Journal of Computing Sciences in Colleges* Vol 20, Issue 1, (2004) 248–261.
- [9] HELMICK, M. T., Interface-based Programming Assignments and Automatic Grading of Java Programs, *Proceedings of the 12th annual SIGCSE conference on Innovation and technology in computer science education*, (2007) 63–67.
- [10] HERNYÁK, Z., KIRÁLY, R., Teaching programming language in grammar schools, *Annales Mathematicae et Informaticae*, Vol 36, (2009) 163–174,
- [11] HEXT, J. B., WININGS, J. W., An automatic grading scheme for simple programming exercises, *Commun. ACM*, 12(5), (1969) 272–275.
- [12] HITCHNER, L. E., An automatic testing and grading method for a C++ list class, *ACM SIGCSE Bulletin* Vol. 2, Issue 2, (1999) 48–50.
- [13] HITZ, M., KÖGELER, S., Teaching C++ on the WWW, *Proceedings of the 2nd conference on Integrating technology into computer science education*, (1997) 11–13.
- [14] HORWITZ, S., Addison-Wesley's Review for the Computer Science AP Exam in C++, *Addison-Wesley* (1999).
- [15] JUHÁSZ, Z., JUHÁS, M., SAMUELIS, L., SZABÓ, Cs., Teaching Java programming using case studies, *Teaching Mathematics and Computer Science* 6/2, pp. 245–256, 2008
- [16] KARLSSON, B., Beyond the C++ Standard Library: An Introduction to Boost, *Addison-Wesley Professional* (2005).



- [17] KOZMA, L., FROHNER, Á., KOZSIK, T., PORKOLÁB, Z., Beyond 2000, Beyond Object-Orientation, *Proceedings of 5th International Conference on Applied Informatics*, (2001) 125–134
- [18] MEYERS, S., Effective STL, 3rd Edition, *Addison-Wesley* (2001).
- [19] NORDQUIST, P., Providing accurate and timely feedback by automatically grading student programming labs, *Journal of Computing Sciences in Colleges* Vol 23, Issue 2, (2007) 16–23
- [20] PLACER, J., The Promise of Multiparadigm Languages as Pedagogical Tools, *Proceedings of the ACM conference on Comp. Sci.*, (1993) 81–86.
- [21] PORKOLÁB, Z., ZSÓK, V., Teaching Multiparadigm Programming Based on Object-Oriented Experiences, *Tenth Workshop on Pedagogies and Tools for the Teaching and Learning of Object Oriented Concepts (TLOOC)* (2006).
- [22] SAIKONNEN, R., MALMI, L., KORHONEN, A., Fully Automatic Assessment of Programming Exercises, *Proceedings of the 6th annual conference on Innovation and technology in computer science education*, (2001) 133–136.
- [23] STROUSTRUP, B., Learning Standard C++ as a New Language, *C/C++ Users Journal*, (May 1999) 43–54.
- [24] STROUSTRUP, B., The C++ Programming Language, Special Edition, *Addison-Wesley* (2000).
- [25] STROUSTRUP, B., Programming, Principles and Practice Using C++, *Addison-Wesley* (2008).
- [26] TREMBLAY, G., LABONTE, E., Semi-automatic marking of java programs using junit, *Proceedings of International Conference on Education and Information Systems: Technologies and Applications (EISTA '03)*, (2003) 42–47.
- [27] WESTBROOK, D. S., A Multiparadigm Language Approach to Teaching Principles of Programming Languages, *29th ASEE/IEEE Frontiers in Education Conference*, (1999) 11b3–14.
- [28] ZAVE, P., A Compositional Approach to Multiparadigm Programming, *IEEE Software* VI(5), (1989) 15–25.

**Norbert Pataki**

**Zalán Szűgyi**

Department of Programming Languages and Compilers

Eötvös Loránd University

Pázmány Péter sétány 1/C H-1117 Budapest, Hungary

e-mail:

`patakino@elte.hu`

`lupin@ludens.elte.hu`



# The software developers' view on product metrics — A survey-based experiment\*

István Siket, Tibor Gyimóthy

Department of Software Engineering  
University of Szeged, Hungary

*Submitted 12 November 2009; Accepted 14 February 2010*

## Abstract

Object-oriented metrics are becoming evermore popular and they are used in many different areas of software development. Many researchers have showed in practice that object-oriented metrics can be efficiently used for quality assurance. For example, a lot of experimental results confirm that some of the object-oriented metrics (like *coupling*, *size*, and *complexity*) are able to predict the fault-proneness of classes. Quality assurance experts usually accept that actively applying metrics can help their work. On the other hand, developers tend not to use metrics because they do not know about them, or if they do know about them, they do not really know how to use them. Hence we devised a *Survey* to ask developers with different levels of experience about the use of metrics. Our hypothesis was that developers with different levels of experience might have significantly different views about the usefulness of metrics.

In the *Survey* four metrics (*size*, *complexity*, *coupling*, and *code duplication*) were examined. The *Survey* asked questions about the participants' experience and skills, then it asked questions about how the participants would probably use these metrics for software testing or program comprehension, and at the end the relative importance of the metrics was assessed.

The main result of the *Survey* is a list which contains those cases where the views about the metrics from developers having different experience significantly differ. We think that getting to know the developers' views better can help us to create better quality models based on object-oriented metrics.

*Keywords:* Survey, object-oriented metrics, program comprehension, software testing.

---

\*This study was supported in part by the Hungarian national grants OTKA K-73688 and TECH\_08-A2/2-2008-0089-SZOMIN08.

# 1. Introduction

Quite a lot of object-oriented metrics have been defined and published (for example, Brito e Abreu's MOOD metrics [5]) since Chidamber and Kemerer published the first notable article in this area, which discussed 6 object-oriented design metrics [4]. Besides their "simple presentation", they investigated how metrics could be applied for quality assurance. Other surveys looked at the relationship between the object-oriented metrics and the number of bugs found and corrected in software products. For example, Basili et al. [1] examined the relationship between Chidamber and Kemerer metrics and the fault density on a small/medium-sized software system. We repeated Basili's experiment on Mozilla [8], while Olague et al. [9] carried out a similar experiment on six different versions of Rhino, but they examined more metrics. The common conclusion of these studies was that metrics could be used to predict bugs, hence they can be used to measure the quality aspect of a piece of software.

In general we can say that experts very familiar with metrics accept that metrics can be used efficiently in different areas of software development. On the other hand, developers hardly use metrics in their everyday work because they do not know the metrics well enough, or they know about the metrics but they do not know how they can apply them. Therefore we devised a *Survey* to get to learn about the developers' knowledge and views of object-oriented metrics and also to see how experience influences the assessment of their practical worth. We asked 50 software engineers working at our department on industrial and R&D projects to take part in our experiment and to fill out an online Survey. The participants' experience was wide ranging because there were both very experienced programmers and students with very little experience among them. Our hypothesis was that there was a significant difference between the views of senior programmers experienced in different areas of software development and junior developers about the usefulness of metrics. This experiment validated these suspicions in many cases. For example, the senior and junior programmers often judged generated classes with bad metric values quite differently, regardless of the metric they were asked about. On the other hand, we did not find any significant difference in certain situations. One example might be that the senior and junior participants' opinions did not differ significantly from the point of view of program comprehension. Hence, one of the Survey results is a set of hypotheses. The aim of a further investigation is to validate these results by involving some of our project partners. If we can reliably characterize the views of senior and junior developers about the usefulness of metrics, then we could develop the kind of metric-based tools which support development and run more efficiently.

In this paper we will proceed as follows. In the next section we will introduce the Survey and the main results will be discussed in detail. In Section 3 we will discuss several other articles which addressed the same problems. Then in Section 4 we will present our main conclusions, and outline our plans for future study.

## 2. Survey

In this section we will present the Survey and our main findings. It contained over 50 questions, so due to lack of space we cannot present all the questions and results. Therefore we will only describe the Survey in general, and only the most interesting questions and most important results will be elaborated on.

The Survey can be divided into three parts. The first part (Section 2.1) contains several general questions about the participants' experience and skills. From the responses we were able to get a general picture about the participants.

The rest of the questions examined the participants' views about the object-oriented metrics and about the connection between these metrics and program comprehension & testing. Since object-oriented programming is class-based, we examined only class-level metrics. We could have examined many different metric-categories and specific metrics but in that case the Survey would have been too long. Therefore only four general categories (*size*, *complexity*, *coupling*, and *code duplications*) and only one metric per category were selected for the Survey.

- The *size* metric we chose was *Lines of Code* (LOC), which counts all non-empty and non-comment lines of the class and all its methods implemented outside the class definition.
- *Weighted Methods for Class* (WMC), which measures the *complexity* of a class, is defined as the sum of the complexity of its methods where the McCabe cyclomatic complexity is used to measure their complexity.
- *Coupling* metrics measure the interactions between the program elements and *Coupling Between Object classes* (CBO), the chosen metric from this category, counts the number of other classes “used” by the given class.
- In the case of *code duplications* (later we will refer to this category as *clones* as well), the *Clone Instances* (CI) metric was chosen which counts the number of duplicated code instances which are located inside the class.

In the second part (Section 2.2) the metrics were examined one by one; more precisely, we asked exactly the same questions about all four metrics to see what the participants thought about them. The third part can be found at the end of the Survey (Subsection 2.3) where the metrics were examined together in the questions and the participants had to rank the metrics by their importance.

The 50 participants who filled out the questionnaire at our Software Engineering Department all work on industrial and R&D projects. They ranged from beginner students to experienced programmers so the participants' experience and skills differed greatly. Consequently we examined how the different levels of experience influenced their assessment on the practical worth of the metrics examined in the Survey. This meant that besides the presentation of the answers and their distribution, statistical methods were applied to see whether experience affected the participants' responses or not.

In the following we will present the most important parts of the Survey and the conclusions drawn from it in the following way: after each question or group of questions (if they belong together) the possible answers and the set of participants

who indicated the given answer are presented in percentage terms. In addition, after each question we discuss the results and conclusions drawn.

## 2.1. Questions about the participants' skills

The first questions measured the participants' experience and skills. The participants had to rank their experience and skills from 1 (least experienced) to 4 or 5 (most experienced).<sup>1</sup> Since there was no point in drawing any conclusion from the individual questions, the following questions were examined together and the conclusions drawn are presented after them. So, first the questions and the distributions of the answers are presented in Table 1.

Question	1	2	3	4	5
How much programming experience do you have?	8%	14%	24%	54%	–
How much do you know about software metrics?	10%	58%	16%	16%	–
How experienced are you in using the C language?	2%	20%	32%	30%	16%
How experienced are you in using the C++ language?	0%	20%	14%	32%	34%
How experienced are you in using the Java language?	0%	8%	34%	24%	34%
How experienced are you in using the C# language?	32%	32%	22%	8%	6%
How experienced are you in using the SQL language?	6%	34%	20%	30%	10%
How experienced are you in open-source development?	22%	52%	16%	10%	–
How experienced are you in software testing?	10%	34%	40%	16%	–

Table 1: The general questions and the distributions of the replies

First, we examined whether there was any connection between the different experience and skills mentioned above. In spite of the fact that the results inferred from these questions cannot really be generalized to any other group of developers because they are greatly influenced by the group structure of our department, we will present them and briefly explain them. We applied the Kendall tau rank correlation [3] with a 0.05 significance level to see whether there was any connection between the participants' experience and skills or not. Table 2 contains only the significant correlation coefficients. These results highlight some of the typical features of our department. For example, our most experienced programmers use C/C++ (the corresponding correlation coefficients are 0.494 and 0.252) and many of them took part in open source projects as well (0.344), where all the project were written in C/C++ (0.534 and 0.281). Java and C# are less frequently used in our department (there are no significant correlations) but our applications written in Java also use databases, which indicates a correlation between Java and SQL (0.351). And finally, object-oriented metrics are one of our research areas, hence many of us working here are very familiar with them.

Now, we will examine how experience and skills acquired in different areas influenced the responses. The results of 5 out of the 9 questions listed above were

<sup>1</sup>The answers for these questions were full sentences expressing different levels of experience and skills, so the programmers could easily and accurately rank themselves on the given scale. However due to lack of space we cannot present these answers here but only their distributions.

	Exp. in prog.	Metrics knowl.	Exp. C	Exp. C++	Exp. Java	Exp. C#	Exp. SQL	Exp. in os
Exp. in prog.	1.000							
Metrics knowl.	0.425	1.000						
Exp. in C	0.494	0.242	1.000					
Exp. in C++	0.252	0.324	0.393	1.000				
Exp. in Java					1.000			
Exp. in C#						1.000		
Exp. in SQL	0.256				0.351		1.000	
Exp. in os	0.344		0.534	0.281				1.000
Exp. in testing								

Table 2: Correlation between experience in different areas and skills

included (experience in C, Java, C# and SQL were not taken into account). In the rest of this study we applied only two categories, *senior* (experienced) and *junior* (inexperienced), thus the 4 or 5 possible answers of a question had to be placed into one of the two categories. Since there was no exact definition about a person's amount of experience in a particular area, we drew the borderline between the categories ourselves. We categorized the responses for the 5 questions and the results can be seen in Table 3.

Question	No. of junior part.	No. of senior part.
Experienced in programming	23 (46%)	27 (54%)
Metrics knowledge	34 (68%)	16 (32%)
Experienced in C++	17 (34%)	33 (66%)
Experienced in open-source	37 (74%)	13 (26%)
Experienced in testing	22 (44%)	28 (56%)

Table 3: The scores obtained for the senior and junior participants

## 2.2. Questions about metrics separately

The following questions examined the metrics separately, which means that only one metric was considered in each question. Besides the usefulness of metrics, we examined whether there was any significant difference between the responses of the senior and the junior participants. We applied Pearson's  $\chi^2$  test with a 0.1 significance level to see whether there was statistical correlation between the experience level and the judgment of metrics. The *null hypothesis* is that a participant's judgment of a metric does not depend on experience. The *alternative hypothesis* is that experience influences a participant's judgment of metrics. In each case we carried out a test and either accepted the null hypothesis or rejected it (then, we accepted the alternative hypothesis).

We know that the size of the sample is small, hence the test is less reliable and since the sample was collected from our department the results cannot really be generalized to any other software engineering team. In spite of this, the results presented in this article show that it is worth investigating this topic in greater

depth because the results reveal a potential problem about the usage of metrics in practice.

In the following, each question was asked for each metric category (although there was one case where it was no use asking about code duplications). For all the questions and for each experience and skill group defined in the previous subsection we examined the connection between the experience and skills and the replies given to the questions. These results are also presented after the questions.

### 2.2.1. Metrics used for program comprehension and testing

The first two kinds of questions examined how metrics can help in understanding or testing an unknown part of a familiar program. The question was posed for *understanding* with a *size* metric only, but the same one was asked with *complexity*, *coupling*, and *clone* metrics, and all four questions were repeated with *testing*.

Question<sub>1</sub>: *Suppose that you have to become familiar with (or to test) a system whose development you did not take part in. Does the size (complexity, coupling, and clones) of the classes in the system influence your understanding (testing approach)?*

- $A_1$ : Yes, it is easier to understand them if the system consists of more, but smaller classes
- $A_2$ : Yes, it is easier to understand them if the system consists of fewer, but bigger classes
- $A_3$ : No, the size of the classes does not influence my understanding
- $A_4$ : I am not sure
- $A_5$ : In my opinion, size itself is not enough for this and I suggest using other metrics as well

Metric	Understanding					Testing				
	$A_1$	$A_2$	$A_3$	$A_4$	$A_5$	$A_1$	$A_2$	$A_3$	$A_4$	$A_5$
Size	28%	6%	10%	6%	<b>50%</b>	<b>36%</b>	10%	16%	6%	32%
Complexity	<b>68%</b>	6%	6%	0%	20%	<b>80%</b>	2%	2%	2%	14%
Coupling	<b>56%</b>	12%	6%	4%	22%	<b>80%</b>	2%	2%	0%	16%
Clones	<b>64%</b>	8%	14%	4%	10%	<b>64%</b>	8%	14%	4%	10%

Table 4: The distributions of the replies for Question<sub>1</sub>

The distributions of the responses for the four key questions are summarized in Table 4. The figures in bold represent the answers which were selected by most participants. From the point of view of understanding half of the participants would have chosen other metrics than size ( $A_5$ ), while 28% of them said that programs containing more but smaller classes were more understandable ( $A_1$ ). In the case of testing the scores changed a lot because 36% of them indicated that it was easier to test programs that had more but smaller classes ( $A_1$ ), but only slightly fewer participants (32%) wanted to choose another metric than size ( $A_5$ ). In the case of complexity, coupling and clones more than the half of the participants said that it was easier to understand programs containing more but less complex or less



strongly coupled classes, or classes containing fewer clones ( $A_1$ ). This score is even more remarkable in the case of testing.

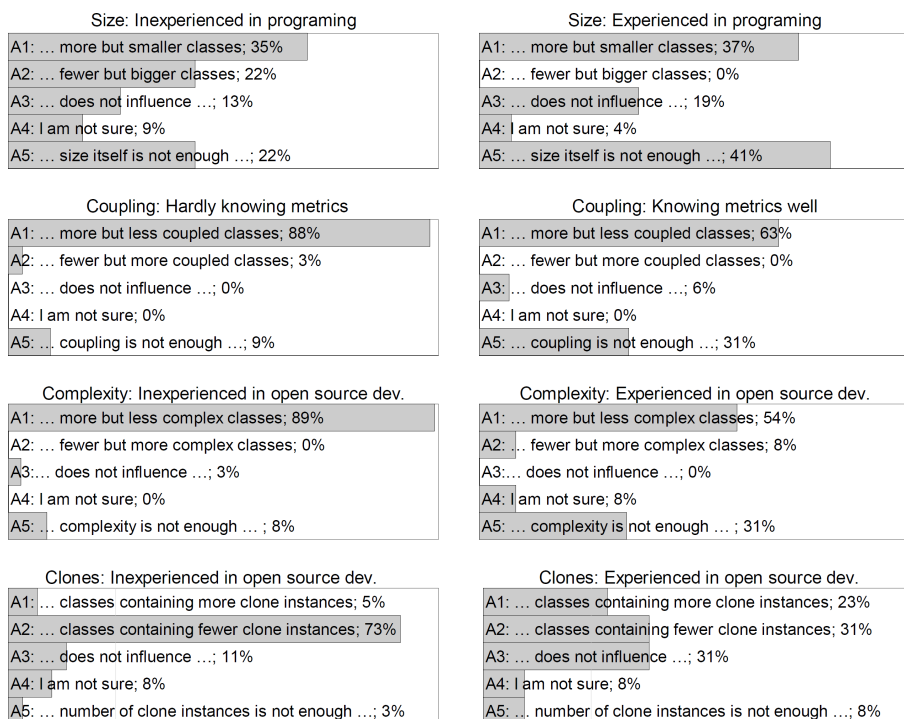


Figure 1: The distributions of the senior and junior participants' answers for Question<sub>1</sub> when testing was considered

We found that from the point of view of understanding there was no significant difference between the responses of the senior and junior participants. On the other hand, with testing we found that, in 4 out of the 20 cases, experience and skills significantly influenced the person's assessment of the metrics. Figure 1 shows the distributions of the responses of the senior and junior participants for the questions where the difference is significant and their justifications are the following:

**Experience in programming and size:** 22% of the participants inexperienced in programming thought that it was easier to test fewer but bigger classes while experienced ones rejected this answer. On the other hand, many more experienced programmers (41% versus 22%) thought that size itself was not enough to decide this question.

**Metric knowledge and coupling:** 25% fewer people quite familiar with metrics thought that low coupling was better for testing but significantly more of them (8% versus 31%) thought that coupling itself was not enough to assess the testing aspect.

**Experience in open source and complexity:** Most of the programmers

not experienced in open source systems (89%) thought that more but less complex classes could be tested more easily while only 54% of open source developers, which is 35% less than the other group, had the same opinion. On the other hand, almost one third (31%) of open source developers said that complexity was not enough while only 8% of the other group marked this option.

**Experience in open source systems and clones:** Most of the inexperienced open source programmers (73%) said that fewer clones were easier to test. However, the experienced group was divided on this point since 3 possible answers were chosen with more or less the same frequency (from 23% to 31%).

### 2.2.2. Acceptable reasons for bad metric values

When a part of a given source code has bad metric values (e.g. due to strong coupling), it is suggested that the code be refactored so as to improve its quality. But in some cases bad metric values may be accepted. For example, we will not refactor a well-known design pattern just because of its bad metric values. The next questions examined what kind of reasons the participants can accept for this.

Question<sub>2</sub>: *What reasons would you accept for a class being too large? (Several answers can be marked.)*

- A<sub>1</sub>: No reason at all
- A<sub>2</sub>: A well-known design pattern
- A<sub>3</sub>: The implemented functionality requires a large size
- A<sub>4</sub>: The source code of the class is generated from some other file
- A<sub>5</sub>: The class must fit a given API
- A<sub>6</sub>: If the large size does not make understanding difficult
- A<sub>7</sub>: It has been tested and works properly
- A<sub>8</sub>: I cannot decide
- A<sub>9</sub>: Any other reason (with a justification)

Metric	A <sub>1</sub>	A <sub>2</sub>	A <sub>3</sub>	A <sub>4</sub>	A <sub>5</sub>	A <sub>6</sub>	A <sub>7</sub>	A <sub>8</sub>	A <sub>9</sub>
Size	2%	34%	52%	56%	56%	24%	28%	2%	4%
Complexity	2%	36%	80%	46%	30%	36%	28%	2%	0%
Coupling	4%	36%	56%	38%	52%	26%	26%	8%	0%
Clones	18%	42%	22%	68%	24%	12%	14%	4%	4%
Average	6.5%	37%	52.6%	52%	40.5%	24.5%	24%	4%	2%

Table 5: The acceptance rates of the different reasons (Question<sub>2</sub>)

Table 5 shows the distributions of the replies, expressed in percentage terms.<sup>2</sup> It is interesting that though many participants said that bad metric values made understanding and testing difficult, only a few of those questioned indicated that they did not accept any reason (A<sub>1</sub>) or that they could not decide (A<sub>8</sub>) based on a large size, complexity or coupling. More participants (18%) rejected the clones option but this percentage is still not very high, so we can say that to some extent

<sup>2</sup>Since an arbitrary number of replies could be given for these questions, the sum of the scores for a question is not 100%.

bad metrics values can be accepted. The answers from  $A_2$  to  $A_7$  that reflect some special excuses got notable scores. The most widely accepted reasons were *the implemented functionality* ( $A_3$ ) and *the generated code* ( $A_4$ ) but *the design patterns* ( $A_2$ ) and *the given API* ( $A_5$ ) also had high scores. The remaining two reasons ( $A_6$  and  $A_7$ ) are still worth examining but they got much lower scores.

We also examined the difference between the senior and junior participants' responses in this case. More answers could be marked for this question, therefore every answer was handled separately and we examined whether there was a significant difference between the answers of the senior and junior participants who accepted the given reason. Since answers  $A_1$ ,  $A_8$ , and  $A_9$  were rarely marked, we decided to exclude them from any further investigation. On the other hand, the excluded answers are just synonyms of "I do not know", therefore all real excuses will be discussed. Table 6 shows the results where the possible answers can be found in the rows and the given metric categories are presented in the columns. Where a significant difference was found between the answers of the two groups, an abbreviation of the participant's experience or skill was written in that table cell. For example, *test* in the second row ( $A_3$ ) and in the second column (Complexity) means that there was a significant difference between the replies of participants experienced in testing well and the replies of participants with little experience in testing.

	Size	Complexity	Coupling	Clones
$A_2$				
$A_3$		test	o.s., test	
$A_4$	exp, met, C++, test	exp, met, C++, test	exp, met	exp, met, C++, test
$A_5$	exp, met			
$A_6$				
$A_7$	exp, met, C++, o.s.	met	o.s.	met, test

Table 6: Significant correlations between the participant's experience and skills and the different excuses

Since too many significant cases were found, we will discuss them in general and only one example will be presented. First, we will analyze the results from the point of view of experience groups. Metric knowledge (met) influenced the judgment of metrics in 8 out of the 24 cases, which is 33.3%. Experience in programming (exp) and experience in testing (test) influenced their judgment 6 times (25%), while experience in C++ (C++) influenced their judgment 4 times (16.7%). Experience with open source systems had the smallest effect because the replies of the two groups differed only in 3 cases (12.5%). We may conclude from these questions that the participants' experience or skills have a notable influence on the judgment of metrics.

Next, we examined how the opinions of the participants varied based on the given replies. The judgment of *generated classes* ( $A_4$ ) affected the opinions of the senior and junior participants in most cases: programming experience and metric knowledge always divided their opinions and having experience in C++

and experience in testing influenced them significantly in 3 cases. On the other hand, experience in open source development had no effect in this case. A typical opinion about the *tested code* ( $A_7$ ) was the other possible reason that had different judgments. In this case, the judgments of *size* and *clones* differed in four and two cases, respectively, but with the other two metric categories the answers of the two groups were very similar in four out of the five cases. This means that here the difference between the opinions of the senior and junior participants was significant in 8 out of the 20 cases examined. After these observations it was interesting that the judgments of *understandable source code* ( $A_6$ ) and *design pattern* ( $A_2$ ) were the same. Only several significant differences were found when the other two reasons ( $A_3$  and  $A_5$ ) were investigated.

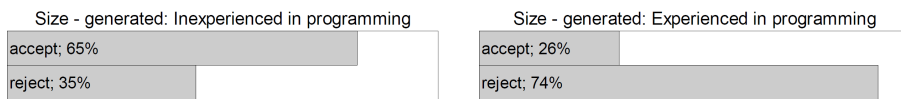


Figure 2: The distributions of the senior and junior participants' replies (Question<sub>2</sub>)

Due to lack of space we cannot discuss all 27 significant differences one by one, hence we present only one example. Figure 2 shows that 74% of the participants experienced in programming accept large generated classes and only 26% of them reject such classes. On the other hand, only 35% of the participants inexperienced in programming accept it and 65% of them reject large generated classes. This is a good example because it shows how much the senior and junior participants' judgment of metrics can differ.

### 2.2.3. Sharing testing resources based on metrics

Testing is a very important phase of software development. Its aim is to reveal all the bugs in the source code, but for large software packages this is impossible because the testing resources (testers, time, etc.) are limited. Hence, we have to share testing resources among the parts of a program and it is important how we do it. The better the testing resources are shared, the more effective the testing phase is, which means that more bugs can be found. We examined how the participants would probably share testing resources if they knew the metric values of the elements in advance. A very simple example (consisting of two classes) was chosen to see how the participants would share testing resources.

Question<sub>3</sub>: *Suppose that there are two classes in an unknown system where the size of class A is 1000 lines (LOC) and the size of class B is 5000 lines. The quality of the two classes is almost the same. During the development the size of class A increased by 10 percent and the size of class B increased by 2 percent. How would you share your testing resources?*

- $A_1$ : I would test only class A
- $A_2$ : I would spend 90% of the testing res. on class A and 10% on class B
- $A_3$ : I would spend 75% of the testing res. on class A and 25% on class B

- $A_4$ : I would spend the testing resources equally on the two classes
- $A_5$ : I would spend 25% of the testing res. on class A and 75% on class B
- $A_6$ : I would spend 10% of the testing res. on class A and 90% on class B
- $A_7$ : I would test only class B
- $A_8$ : I would not determine it based on size
- $A_9$ : I cannot decide

The same question was asked for complexity where the complexity (WMC) values of class A and B were 100 and 500, and for coupling where the coupling (CBO) values of class A and class B were 20 and 100, respectively. We did not ask about code duplications here because it would not have made any sense.

Metric	$A_1$	$A_2$	$A_3$	$A_4$	$A_5$	$A_6$	$A_7$	$A_8$	$A_9$
Size	0%	0%	8%	12%	30%	0%	0%	48%	2%
Complexity	0%	4%	10%	34%	30%	6%	0%	12%	4%
Coupling	0%	2%	10%	28%	28%	6%	4%	20%	2%

Table 7: The distributions of the replies for Question<sub>3</sub>

The results of the responses are listed in Table 7. Almost half (48%) of the participants said that they would not share testing resources based on size ( $A_8$ ) while 30% of them said they would spend 75% of the testing effort on class A ( $A_5$ ). With the complexity issue, most participants (34%) said they would share testing resources equally between the two classes ( $A_4$ ) but only slightly fewer (30%) said that they would spend 75% of the testing resources on class A ( $A_5$ ). In the case of coupling answers  $A_4$  and  $A_5$  got the same response (28%), which is very similar to what we got with complexity.

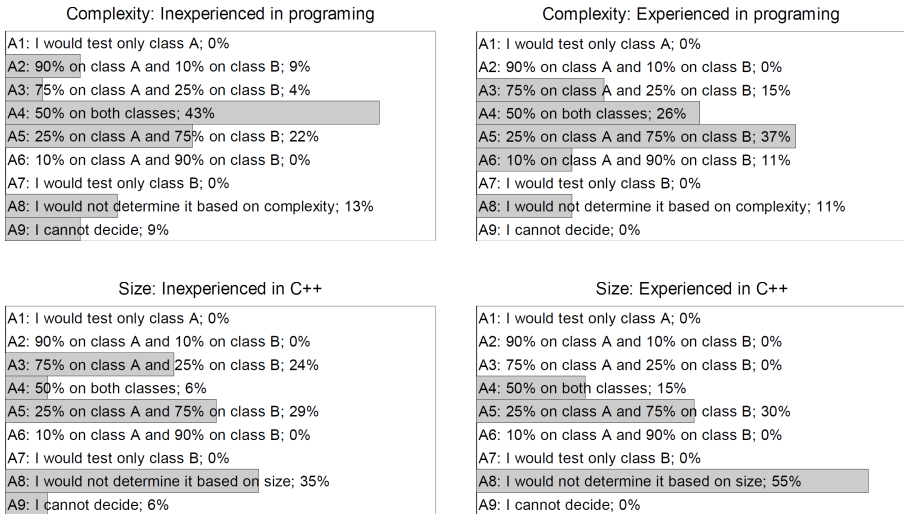


Figure 3: The distributions of the senior and junior participants' replies for Question<sub>3</sub>

We found in two cases that experience and skills had a significant effect on the kind of answers of Question<sub>3</sub> (see Figure 3). These two cases are the following:

**Experience in programming and complexity:** 48% of the (37% and 11%) participants experienced in programming thought that it was the absolute complexity of the classes that matters from a testing perspective and not the increment after a change ( $A_5$  and  $A_6$ ). In contrast, participants inexperienced in programming said they would share testing resources equally ( $A_4$ ).

**Experience in C++ and size:** More than half of the senior C++ programmers (55%) would not rely on size for test design ( $A_8$ ), and they (30%) thought that absolute size was more important than any increment ( $A_5$ ). The opinions of the junior C++ programmers were heterogeneous, which meant that two answers conflicting with each other ( $A_3$  and  $A_5$ ) had significant scores and they were almost the same (24% and 29%).

Besides these four questions (actually, there are a lot of questions but they can be classified into four basic categories), there were other questions which examined the metrics individually. Due to lack of space they will not be presented in detail, but will be mentioned only briefly.

We analyzed ten systems<sup>3</sup> and calculated the averages of the metrics and what percentage of the classes exceeded the triple of the average of the metric values in question. For a given metric both values of the systems were presented anonymously on a diagram and the participants were asked to classify the systems into 7 quality categories using the diagrams. The categories ranged from *very bad quality* to *very good quality*. We gave the same task for all four metrics mentioned previously.

We investigated the participants' opinions about what the optimal size (complexity and coupling) for a class in an object-oriented system was (minimum and maximum values could be given) and what the code size was, above which the clone instances should be eliminated (a limit could be given).

### 2.3. Questions about the importance of metrics

In the third part of the Survey the importance of each metric was examined. In these questions more than one metric was used at the same time and the participants had to select those they thought were the better ones, and they also had to rank them.

Question<sub>4</sub>, the only question delineated from this part of the Survey, examined the importance of the metrics from a testing point of view. The participants had to weight the four metrics (size, complexity, coupling, and code duplications) when deciding how useful the four metrics were. The weight ranged from 1 (the least useful) to 10 (the most useful).

Figure 4 gives a histogram representation of the responses. According to the participants surveyed, *complexity* is the most relevant metric because the two high-

---

<sup>3</sup>We analyzed 6 industrial and 4 open source systems. Among the industrial ones there were telecommunications, a graphical application, and a code analysis system, while the four open source systems were Tamarin, WebKit, Mozilla, and OpenOffice.org.

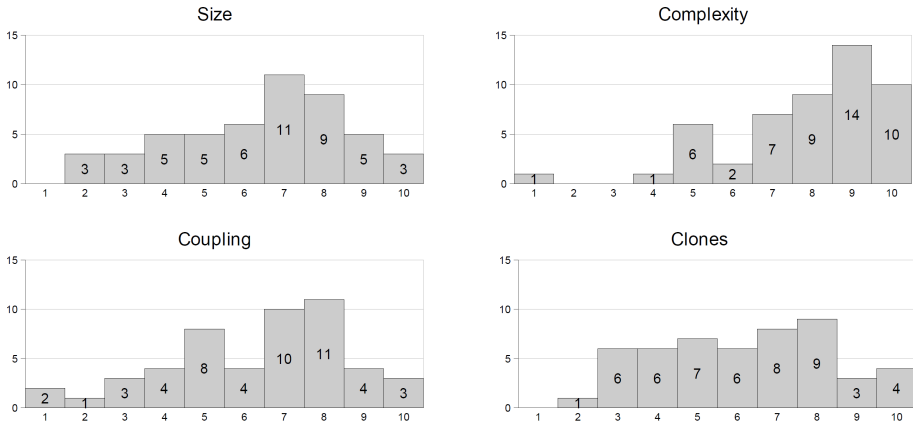


Figure 4: The distributions of the senior and junior participants' replies for Question<sub>4</sub>

est weights got the biggest response and its average (7.88) is the largest. The two most frequent weights of the other three metrics are the same (weights 7 and 8) and their averages (size 6.40, coupling 6.34, and clones 6.20) are almost the same. Despite the fact that the distributions of the three metrics differ, we can say that their degree of importance is very similar, but they seem less important than complexity. This result is slightly surprising because in an earlier paper [8] we investigated which metrics could be used to predict the fault-proneness of the classes and we found that CBO (coupling) and LOC (size) metrics came out top while WMC (complexity) got a lower score, which seems to contradict these new findings.

### 2.3.1. An experiment on Mozilla

In an earlier paper [8] we examined the fault-proneness property of eight object-oriented class level metrics. We calculated the metrics for seven different versions of Mozilla [12] (from version 1.0 to version 1.6), collected the reported and corrected bugs from the bug tracking system called Bugzilla [2] and associated them with the classes. This way we knew the metric values and the number of bugs for each class in each version so we could examine how well the different metrics could predict the fault-proneness property of classes. Although we had all the necessary information for all the Mozilla versions, we chose version 1.6, which contained 3,209 classes, and carried out the experiment on this version. We applied a statistical method (logistic regression) and machine learning (neural networks and decision tree) to predict whether a given class was bug-free (containing no bug) or faulty (containing at least one bug). We examined the metrics one by one with each method and the results of the three methods were very similar. We found that CBO was the best metric but LOC was only a slightly worse and WMC also gave good results. On the other hand, code duplication was not examined, so we have no information about the usefulness of the CI metric.

We carried out an experiment to investigate the result of Question<sub>4</sub> on Mozilla version 1.6. We examined which weight combination given by the participants could find the most classes which contained at least 10 bugs. For this we defined a simple *model* in the following way: for each class we calculated the weighted sum of its four normalized metrics (LOC, CBO, WMC and CI) where the weights were the answers of Question<sub>4</sub>. Then, the classes were sorted by their weighted sum and the top 177 were selected as faulty classes. We selected only the top 177 because there were 177 classes in Mozilla version 1.6 which contained at least 10 bugs. We examined how many of the 177 classes selected by the given model really contained at least 10 bugs so we could compare the “quality” of the models. The more such classes the model found, the better it was. We examined all 50 models and discovered that the best one found 113 out of the 177 classes, which means that it found 63.8% of the worst classes. Here, the weights of the model were the following: size = 3, complexity = 8, coupling = 9, and clones = 3.

### 3. Related works

Our earlier experiment on Mozilla [8] was described in Section 2.3.1 above. In this summary we can see that metrics can be used for fault-proneness but at a different level. However, instead of presenting other similar empirical validations (e.g. Yu et al. [13], Fioravanti and Nesi [6], Basili et al. [1], and Olague et al. [9]), we will summarize another survey.

The ISO/IEC 9126 international standard [10] defines the relationship between the system quality and ISO/IEC 9126 subcharacteristic. The Software Improvement Group (SIG) introduced another level below the subcharacteristics which consists of system properties and they defined a binary mapping between ISO/IEC 9126 subcharacteristics and system properties. José et al. [11] carried out a survey to examine the connection between system properties and quality characteristics for maintainability. 22 software quality experts of SIG were asked to take part in their experiment. The participants had to compare the 4 maintainability subcharacteristics with each other (6 comparisons) and the 9 systems properties with each other for each subcharacteristic (4 times 36 comparisons), so a participant had to make 150 comparisons. They used a scale of 1 (equal importance) to 5 (extreme importance) to rate the relative importance. There were three main questions that they wanted to answer with their survey.

*Does the weighted mapping represent agreement among experts?* The result was that, at the level of subcharacteristics, 2 out of 4 relations were non-consensual, while at the level of system properties 7 out of the 36 relations were non-consensual.

*How similar are the weighted mapping and binary mapping?* After the evaluation they found that in 7 cases the result had to be excluded because there was no consensus; in 21 cases the result confirmed the earlier definitions; in 2 cases new relations were found; and 6 were not presented.

*Can the difference between the mappings be somehow used to refine the quality model?* In the case of mapping from subcharacteristics to maintainability, the re-



sults suggested that the relative weight of testability should be increased and the weight of stability should be decreased. However, the consensus among the experts was too small to warrant the change. On the other hand, there was a better consensus among the experts for mapping from system properties to subcharacteristics. After excluding non-consensual relations, several changes were recommended.

## 4. Conclusions and future work

The main motivation for this Survey was to learn more about the developers' expertise and opinions concerning object-oriented metrics and to investigate how experience influenced their assessment of metrics.

The main contributions of this paper are the following. Firstly, we listed a set of interesting questions of our Survey which examined the software engineers' opinions about four object-oriented metrics. Secondly, we presented the distributions of the replies and drew some conclusions about them. Thirdly, we examined the relationship between the experience and skills of our programmers. Fourthly, we applied a statistical method to see how experience affects the assessment of these metrics and we devised hypotheses based on them. And finally, we carried out an experiment on Mozilla to see which metrics were important in bug prediction.

Our main observations are the following. First, we did learn more about the participants' opinions concerning the four metrics in different situations. Second, we found that in certain cases experience in different areas significantly affects the assessment of the metrics. Third, the importance of metrics in testing is not in accordance with the results of experiments [8, 1] which examined the relationship between metrics and fault-proneness. Fourthly, we devised several hypotheses which asserted that there were significant differences between the senior and junior programmers' assessments of metrics (for example, generated code with wrong metric values was judged differently in 14 out of the 20 cases). The main conclusion is that we need to investigate this topic in greater depth because some of the results here are quite surprising.

In the future we plan to repeat this experiment with our industrial partners to survey the same questions, but in different circumstances. This way we can verify our observations and we should have more reliable conclusions. Furthermore, we will refine the Survey by using the experience gained during this experiment. We examined only four metric categories here, but there are many other interesting issues which are worth investigating. Hence, we plan to incorporate other kinds of metrics (e.g. *cohesion metrics*) and design issues (e.g. *bad smells* [7]) into our next Survey.

## References

- [1] BASILI, V. R., BRIAND, L. C., MELO, W. L., A Validation of Object-Oriented Design Metrics as Quality Indicators, In *IEEE Transactions on Software Engineering*,

- volume 22, (1996) 751–761.
- [2] Bugzilla for Mozilla, <http://bugzilla.mozilla.org>.
  - [3] CHALMER, B. J., *Understanding Statistics*, CRC Press, 1986.
  - [4] CHIDAMBER, S. R., KEMERER, C. F., A Metrics Suite for Object-Oriented Design, In *IEEE Transactions on Software Engineering* 20,6(1994), (1994) 476–493.
  - [5] E ABREU, F. B., MELO, W., Evaluating the Impact of Object-Oriented Design on Software Quality, In *Proceedings of the Third International Software Metrics Symposium*, IEEE Computer Society, March 1996, 90–99.
  - [6] FIORAVANTI, F., NESI, P., A Study on Fault-Proneness Detection of Object-Oriented Systems, In *Fifth European Conference on Software Maintenance and Reengineering (CSMR 2001)*, March 2001, 121–130.
  - [7] FOWLER, M., BECK, K., BRANT, J., OPDYKE, W., ROBERTS, D., *Refactoring: Improving the Design of Existing Code*, Addison-Wesley Pub Co, 1999.
  - [8] GYIMÓTHY, T., FERENC, R., SIKET, I., Empirical Validation of Object-Oriented Metrics on Open Source Software for Fault Prediction, In *IEEE Transactions on Software Engineering*, volume 31, IEEE Computer Society, October 2005, 897–910.
  - [9] OLAGUE, H. M., ETZKORN, L. H., GHOLSTON, S., QUATTLEBAUM, S., Empirical Validation of Three Software Metrics Suites to Predict Fault-Proneness of Object-Oriented Classes Developed Using Highly Iterative or Agile Software Development Processes, In *IEEE Transactions on Software Engineering*, volume 33, June 2007, 402–419.
  - [10] International Standards Organization. *Software engineering - product quality - part 1: Quality model*, ISO/IEC 9126-1 edition, 2001.
  - [11] KANELLOPOULOS, Y., CORREIA, J. P., VISSER, J., A Survey-based Study of the Mapping of System Properties to ISO/IEC 9126 Maintainability Characteristics, In *The International Conference on Software Maintenance (ICSM'09)*, IEEE Computer Society, September 2009, 61–70.
  - [12] The Mozilla Homepage, <http://www.mozilla.org>.
  - [13] YU, P., SYSTÄ, T., MÜLLER, H., Predicting Fault-Proneness using OO Metrics: An Industrial Case Study, In *Sixth European Conference on Software Maintenance and Reengineering (CSMR 2002)*, March 2002, 99–107.

**István Siket, Tibor Gyimóthy**

Department of Software Engineering

University of Szeged

H-6720 Szeged

Árpád tér 2.

Hungary

e-mail: [siket@inf.u-szeged.hu](mailto:siket@inf.u-szeged.hu)

[gyimi@inf.u-szeged.hu](mailto:gyimi@inf.u-szeged.hu)

# Mathematical competences examined on secondary school students

**Ilona Téglási**

Institute of Mathematics and Computer Science  
Eszterházy Károly College

*Submitted 6 October 2010; Accepted 17 December 2010*

*Dedicated to professor Béla Pelle on his 80<sup>th</sup> birthday*

## **Abstract**

First I'd like to write about the idea of mathematical competence and make clear its components, according to the 2000 Lisbon Resolution of the European Union. That, and the results of the first PISA examination implicated a development work in our country, and as a result we now have new types of competence-based school-books, and the methodological culture of teachers is under reformation too. But these changes are not uniformly welcome among teachers. There is a kind of contradiction between traditional and competence-based education and evaluation – so I tried to match the two types with measuring the present level of competence of secondary school students, and upon that work out development methods, in which the defects of the knowledge can be supplied.

The students solved an exercise-paper with 5 exercises (practical, playful, needs many different abilities and skills, but less knowledge, both simple and complex) in 45 minutes. I made two types of measurements – a traditionally used method which measures the mathematical achievement, and another, which measures the level of skills and competences. I analysed the results with comparing the achievements with the competences to show how the mathematical skills prevail among achievements. I show this analysis with graphs in my paper. According to the results of the measurement, we can declare the main areas of development. At the end of my paper I'd like to show how I plan the follow-up of this examination.

## 1. Introduction

Nowadays we can hear a lot about different competences and competence-based teaching. The European Union in 2000 Lisbon Declaration gave recommendation to its member states to develop their educational system. Moreover the results of international measurements on the knowledge of Hungarian students also required reformation. According to the measurements, it seems that our students' knowledge lags behind the knowledge of other countries', it isn't modern enough for the challenges of present days. On these effects new projects were realized to develop education in our country too. In 2002 within the confines of the National Development Plan the key competences were defined, which the education has to take attention on. Through the Human Resource Development Operative Programme started curricular and methodological generative works, and after that the tests of the developed programmes in schools through tenders. Now we have tangible results of it in school-books and other learning materials, programme packs. But the spread of using these new methods is delayed by the doubtfulness of teachers, sometimes negative behaviour, the slothfulness of the changes of educational procedures, and also the negative reactions of parents. It's hard to accept every newness, but it's easier if well prepared and supported by examinations. It's a pity, that it was not like this in Hungary, so that's why we have this resistance against these changes.

The aim of my assessment is to see the applicable mathematical knowledge of secondary school students learning according to traditional curriculum with a pretest. After the evaluation of the test – comparing competences and mathematical achievement too – I would like to look for development strategies and methods, which are valid for the general school system to appraise mathematical competences within the maths lessons, and not in extraordinary time.

## 2. Mathematical competences – theoretical considerations

The key competences are indispensable for the flexible accommodation to changes, the acceptance of changes and the forming of own life. Mathematical competences belong to them. The OECD in connection with the PISA-examinations drew up the competence like this in 2000: "Mathematical competence is a preparedness, which qualifies the person to identify mathematical problems, understand and handle them, and also to form a valid opinion on the act of mathematics in present and future professional and private life, and the role in personal and social connections." [2]. According to this the following components were defined:

1. mathematical thought, conclusion
2. mathematical argument, proof
3. mathematical communication

4. mathematical modelling
5. problem posing and solving
6. representation
7. symbolic and formal language, operations
8. tool using skills.

These components can be on three development levels:

- a. reproductive – executing routine, standard exercises, using definitions,
- b. connective – executing complex, but still standard exercises, using integration,
- c. reflective – handling complex problems, genuine approach, generalization.

Psychologists and theoreticians in pedagogy divided these components to more skills and abilities according to factor and matter analysing [2]. Let’s see such a division:

Skills	Thinking abilities	Communication abilities	Cognitive abilities	Learning abilities
numbering, counting, quantitative conclusion, estimation, measuring, changing units, solving textual exercises	systematisation, combinativity, deductive conclusion, inductive conclusion, probability conclusion, argument, proof	relation vocabulary, textual understanding and explanation, spatial sight, perception of spatial relations, representation, presentation	problem sensitivity, problem representation, originality, creativity, problem solving, metacognition	listening, perception of parts-whole, memory, exercise-keeping attitude, exercise solving velocity

If we think it over, these components are really important parts of the competitive mathematical knowledge, but they mean so multiple and diverse knowledge system, that even some mathematics teachers don’t dispose all of these skills and abilities on the needed level. Such a division could hardly be used in everyday school-life during maths lessons, because of its complexity.

In the 2006 PISA report we can find another description of mathematical competence: “The applied mathematical literacy means, that the person recognizes and understands the act of mathematics in real world, forms valid decisions and his/her mathematical knowledge helps in solving own life’s real problems, and become a constructive, enquirer, moderate member of the society.” [2]. This description is more general, than the previous.

If we look after in different sources, we can find many different definitions. This multiple approach shows that the research of the theme is not closed, its special literature is under evolution, and the educational methods based on it are under

developing yet. On the other hand, it's hard to accept such a diversity. If we lay down a few (6-8) guiding principles to go along during a development programme, it's easier to follow. That's why I use the above mentioned 8 points with the 3 levels to define mathematical competences, as it is in English usage. These points draw up handable abilities, don't touch on different parts of mathematical literacy, general enough to cover the whole spectrum of mathematics, and all the skills and abilities mentioned in the tablet could be classified under one of the 8 points.

### 3. Traditional education – competence based education

Before we compare the traditional and competence based education we have to make clear, why we teach mathematics. After defining and deciding the goals can the materials and methods be determined. The materials of education are assigned by social expectations, we have to cultivate such skills and abilities, that are essential to social integration and forming own life. And of course, in working out materials and methods pedagogical and psychological viewpoints must be taken into consideration too [3]. The acceleration of changes in society and technology gives new challenges to the members of education, and if we react too late, we'll lag behind. After the weak results of the first PISA assessment in 2000, Germany started a development in education, accepted and supported by wide rates of the society, which caused a significant growth in the 2003 and 2006 results [2]. We'd need a similar quick development too. In teaching mathematics we can develop many psychic attributes, that are important parts of social integration and lifestyle. This is one of the most important pedagogical goals. As an educational goal, we can declare to give such experiences, images, ideas and knowledge during teaching mathematics which works up the ability of simple and complex using of abstractions, connections, terminology, operations, cognitive actions. The most important qualification goal is to make the students able to use their knowledge creatively, and work up routines and such personal abilities that help them in lifelong learning process [3].

I think, that the traditional, matter-concentrated teaching has got many valuable moments, which shouldn't be dropped out, but should be saved among new forms and methods. If we look into old school-books and exercise-collections, we can find many practical textual exercises, that could be used as nowadays as 20-30-50 years ago, only we have to make the text current sometimes. Our secondary school mathematics curriculum aspired to give knowledge from many different parts of mathematics. The revision of the National Basis Curriculum and the frame curriculum in 2003 dropped out such parts, that were articular in maths curriculum and maturity earlier, but the needth was querying (e.g. trigonometric equations, additional theorems, etc.) These changes showed the signs of modernisation, and also some new parts were put in, such as statistical counts. These changes and emphasis shifts to new themes are the results of the 2000 Lisbon Declaration, which

started the reforms [7]. I think, the material is still too wide, so that the teachers don't have enough time for practising and problem solving. The decrease of number of lessons, the increase of number of students in one class, and the traditional circumstances of teaching frames hardly give enough space to use new methods efficiently. The structural changes in secondary school system from the 1990's on also makes the efficient work hard: though the number of children weakens year by year, more students learn in grammar schools now, where traditionally more theoretical knowledge is expected than in the previous practical, vocational training schools. So, many of those children has to learn theoretical knowledge who don't really need and claim it – because the practical schools disappeared or changed. This also raise the question of restructuring the teaching materials and the learning methods. But to achieve a real and efficient competence based education we have to change the frames of school timing too.

#### **4. The participants and the test problems of the assessment – research question**

The aim of my assessment is to see the applicable mathematical knowledge of secondary school students learning according to traditional curriculum. This is a pretest of a development strategy on mathematical competences (mainly modelling and problem solving). How can we simply assess the level of abilities together with the mathematical achievement? We used to assess only the achievement in school, but for such a development work I'd need a combined evaluation of the two things. I made my assessment in the Practising Secondary School of Eszterházy Károly College in Eger, in 12 grammar classes (3 classes per grade) with test papers laboured for 45 minutes. Altogether 278 students did the tests (77 on 9th grade, 48 on 10th grade, 82 on 11th grade, 71 on 12th grade). These grammar classes learn standard mathematical curriculum, 3 lessons per week, no special maths classes among them. The main profile of the school gives high level education to students in drawing and visual communication, music, foreign languages, and information technology, with more lessons per week in these subjects than the average. So in mathematics they have got average preparedness. At least half of the students come from other places, many of them live in student hostel, and the rate is even higher among the special classes. The students didn't get special preparation for this assessment. They solved the exercises during a normal maths lesson, and I would like to thank to my colleges, Mrs. Gyözóné Erdős, Mrs. Zoltánné Pelbárt and Mrs. Katalin Lénártné Pintér for their helpfulness.

I got the test problems from competence based school-books made by Educatio Kht., partly from “Secondary school mathematical exercise collection” from National School-book Publisher, and also from the collection of the MA students of “teacher of Mathematics” in Eszterházy Károly College. When collecting and selecting the exercises I took into consideration, that they should:

- be practical,

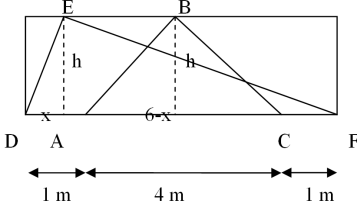
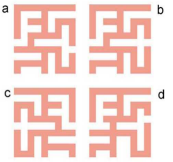

- measure varies of skills and abilities,
- need only such knowledge, that the students already learnt in previous school years,
- be playful, diversified,
- be both simple and complex exercises among them.

I gave 5 different exercises to every grade, which the students had to solve in 45 minutes. The time was far enough for those ones, who read slowly too. The exercises were all textual, so the weakness of reading and understanding ability naturally influenced the results. But, unfortunately, it is not enough to make only one test, to find out, how this weakness effects on mathematical achievement, just during continuous work with a student – so I couldn't take this into account. Every papers had exercises to measure counting skill, logical thinking skill, combinativity, functional, algorithmical thinking, spatial sight, perception of spatial relations, problem solving abilities. In assessing the exercises I took into consideration which thinking or counting units would lead to the result (but of course, different ways could be right) [1]. I made two types of assessment to each papers: a mathematical achievement evaluation, traditionally used in school practise by teachers, and a new type of skills – ability level assessment [4]. I would like to choose the main path of development strategy by comparing these two, in order to develop skills and abilities of students, so as to result growth in mathematical achievement in school too. An example of the test (made for the 11th grade) and the two types of assessments can be seen below. The paper of other grades contains similar exercises.

Table 1: Exercise paper for 11th grade

Exercise	Solution and achievement points	Competence points												
1. Five friends noticed, that their telephone numbers are such 7-digit numbers, which's first digit is 3, every digit is different, and the buttons on the mobile phone pushed one after the other goes in the order of the move of horse in chess. Which are the five numbers? Can there be more numbers like these? <table border="1" style="margin-left: auto; margin-right: auto;"> <tr><td>1</td><td>2</td><td>3</td></tr> <tr><td>4</td><td>5</td><td>6</td></tr> <tr><td>7</td><td>8</td><td>9</td></tr> <tr><td></td><td>0</td><td></td></tr> </table>	1	2	3	4	5	6	7	8	9		0		Possibilities according to the conditions: 3-4-9-2-7-6-0 (1p.) 3-4-9-2-7-6-1 (1p.) 3-4-0-6-7-2-9 (1p.) 3-8-1-6-7-2-9 (1p.) 3-8-1-6-0-4-9 (1p.) There can't be more, because keeping the rule of chess move, the numbers would repeat. Argument: tree graph. (1p.)	a, b, c, g, h, i, j, k, m
1	2	3												
4	5	6												
7	8	9												
	0													



<p>2. Feri's father would like to hang two 90° angled halogene lamps in their cellar. One into the middle of the 6 m long ceiling, lighting straight down, but in this case 1-1 m long stripes would left unlighted. The other lamp would left unlighted. The other lamp obliquely in another place, to lighten the whole floor. How far is the second lamp from the middle? See the figure below!</p> 	<p>Interpretation, notation on figure: (2p.)  <math>ABC_{\Delta}</math> isosceles, rectangular, so the altitude <math>h = AC/2 = 2</math> m (2p.)  <math>DEF_{\Delta}</math> rectangular, the altitude belonging to <math>DF</math> hypotenuse is also 2 m, the two parts of the hypotenuse are: <math>x</math> and <math>6 - x</math> long. (2p.)          According to altitude theorem: <math>x(6 - x) = 2^2</math> (2p.), from this we get: <math>x^2 - 6x + 4 = 0</math>.          Using the solution formule, we get <math>x_1 = 3 - \sqrt{5} \cong 0.76</math> and <math>x_2 = 3 + \sqrt{5} \cong 5.24</math>. (2p.)          Conclusion: he has to put the second lamp (E) in <math>3 - 0.76 = 2.24</math> m distance from the first one (B). (2p.)</p>	<p>a, b, c, f, g, h, i, k, l, m</p>
<p>3. One day Barbara, Bea, Bori and Balázs travelled by train with their friends, and for passing time, they played. At first every member of the company had to think of a 3-digit positive number, which's digits are bigger then 4 and less then 7. When they told their numbers one by one, they realized, that all numbers were different.</p> <p>a) How many were they at most? An other day Barbara, Bea, Bori, Balázs and their 4 friends (Attila, András, Ali and Anna) went to cinema together. All the 8 places, on the tickets, were in one row, next to each other.</p> <p>b) In how many different orders can the 8 friends sit, if non of those, who's name begins with the same letter, can sit near each other?</p>	<p>a) The digits can be 5 or 6, and can be repeated. (1p.) The number of variations of 2 digits, in 3 places are <math>2^3 = 8</math>. (555, 556, 565, 566, 655, 656, 665, 666) So the company's got 8 members at most. (2p.)          b) The sitting order can be ABABABAB or BABABABA patterned. (1p.) The number of orders with letter "A" are <math>4!</math> altogether, and with letter "B" the same. (1p.) All the orders can be the multiplication of these: <math>2 \cdot 4! \cdot 4! = 1152</math>. (2p.)</p>	<p>a, b, c, d, e, f, g, h, i, j, k, l, m</p>
<p>4. Choose which figure is the imprint of the postmark on the picture!</p>  	<p>The "b" is the right imprint. (3p.)</p>	<p>a, b, c, i, m</p>

<p>5. A 130 m long freight train goes 42 kilometres per hour. What time does it go through a 220 m long tunnel?</p>	<p>The length of the train and the tunnel together is:  <math>130\text{ m} + 220\text{ m} = 350\text{ m} = 0.35\text{ km}</math>. That's the way to go if it wants to go through the tunnel. (2p.) According to relation between the way (<math>s</math>), time (<math>t</math>) and velocity (<math>v</math>):  <math>t = s/v = 0.35/42 = 0.00833\text{ h} = 0.5\text{ min}</math>. So it takes the train 0,5 minutes to go through the tunnel. (3p.)</p>	<p>a, b,  c, d,  e, f,  g, h,  i, j,  k, m</p>
---	--	--

## 5. The assessment of mathematical achievement

I evaluated the completed papers with giving 1-1 points to each thinking or counting units, similar to the pointing system of maths maturity. I put the points per student and per exercise into an Excel tablet. I used the Excel to summarize the points of students, giving the achievement in percentage too, the average and deviation of each exercises, and each grades. I made graphs from the points of students, compared with the average, the minimal and optimal level. It caused difficulties that some of the students didn't get the test serious enough, because it didn't have any "stake" – it shows, that in our schools students used to work for marks, not really for knowledge. I tried to strain off these papers, because they wouldn't show the real knowledge and abilities, just the motivation (which is an interesting topic too, but it wasn't the main in my examination).

I appointed the minimal level of mathematical achievement at 20%, and the optimal level above 60%. Here are the summarized graphs of each grades. They show the points of the students in growing order, the minimal level, the average of the grade and the optimal level. Below the graph the tablets show the evaluation of each exercises per grade.

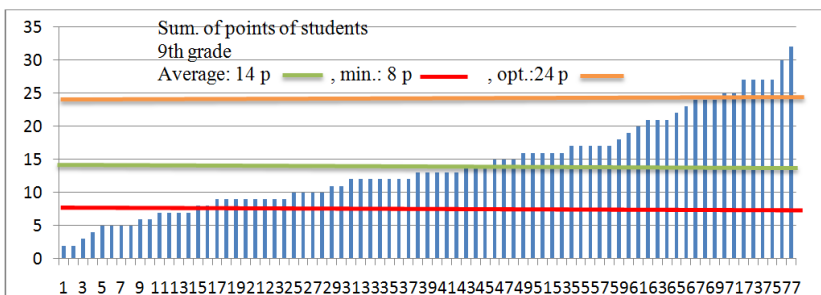


Figure 1: Assessment of achievement, 9th grade

Exercise	1. (8p.)	2. (12p.)	3. (12p.)	4. (3p.)	5. (5p.)	$\sum$ (40p.)	%
Average	3.30	4.71	3.75	1.44	0.81	14.01	35.03
Deviation	2.79	3.84	3.03	1.51	1.34	7.05	17.63

Table 2: Average of points and deviation on 9th grade

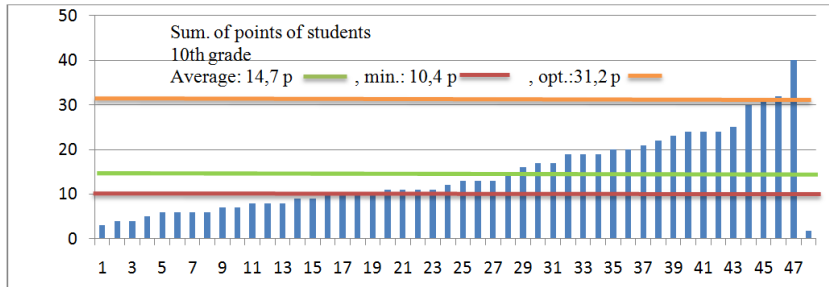


Figure 2: Assessment of achievement, 10th grade

Exercise	1. (6p.)	2. (12p.)	3. (10p.)	4. (12p.)	5. (12p.)	$\sum$ (52p.)	%
Average	5.2	1.7	2.3	4.0	1.4	14.7	28.2
Deviation	1.69	2.85	3.61	4.47	2.43	8.51	16.36

Table 3: Average of points and deviation on 10th grade

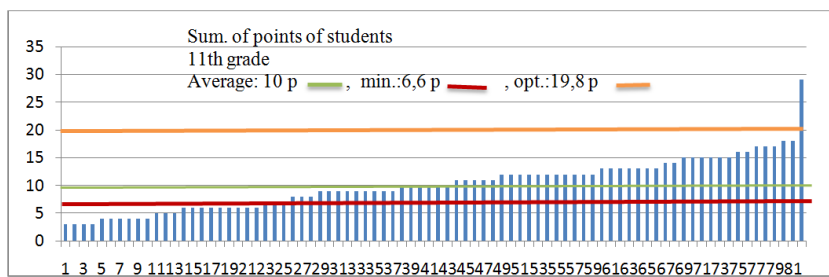


Figure 3: Assessment of achievement, 11th grade

Exercise	1. (6p.)	2. (12p.)	3. (7p.)	4. (3p.)	5. (5p.)	$\sum$ (33p.)	%
Average	3.39	0.56	2.98	1.76	1.50	10.18	30.86
Deviation	2.37	1.51	2.27	1.49	1.97	4.55	13.79

Table 4: Average of points and deviation on 11th grade

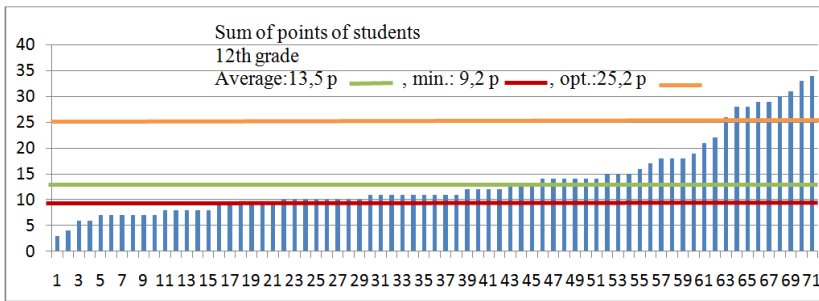


Figure 4: Assessment of achievement, 12th grade

Exercise	1. (4p.)	2. (12p.)	3. (8p.)	4. (6p.)	5. (16p.)	$\sum$ (46p.)	%
Average	3.45	1.18	3.44	2.23	3.32	13.62	29.61
Deviation	1.01	1.45	2.71	1.31	4.77	7.26	15.77

Table 5: Average of points and deviation on 12th grade

On every grade the students showed weak average achievement. The best average was on 9th grade, and the least on 10th grade (see tablet 2 and 3). If we look at the graphs (figures 1–4), we can see, that the averages of all grades are closer to the minimal level, than to the optimal. I could evaluate the work of 278 students, so we can consider the sample is representative for a normal grammar school, learning mathematics according to general curriculum. Altogether 74 students got points under the minimal level, that's the 26.5% of all. Their school achievement is rather weak too, we can declare. Above the optimal level were only 23 students, that is the 8.3% of the sample (see figures 1–4). I think one reason for the weak achievement is, that these exercises were unconventional, strange for most of them, and they weren't prepared for the test. On the other hand, I think the lack of motivation also weakened the results.

In general we can say, that the playfully interpreted, simple exercises, which didn't need too much counting, were more successful, than the complex ones. The result of those exercises, needed only spatial sight were much better, then those ones in which they had to count something from a given figure. The result of the exercises needed logical thinking and combinativity were better then the average as well. The weakest results were the complex exercises, which needed problem solving abilities, planning and more relations (see tables 2–5). I think, in this case the difficulty was the translation of the real problem into the language of mathematics. In my opinion, another reason of this weak result is a kind of laziness, because this generation prefers the easily, quickly available results to the ones demands patience and endurance. I think this problem is more related on present social problems than mathematics teaching, but in school we have to take this also into consideration, and it can be developed with traditional methods as well as competence based methods.

## 6. The assessment of skills and abilities

Still we rarely meet the measuring of skills and abilities in public education. After the first PISA assessment, started the Country Competence Measurement, which measures the abilities of students in textual understanding and mathematical tool usage. The system of this measurement shows into good ways of developing, because it measures the same grades (4th, 6th, 8th, 10th) year by year, among the same circumstances. It also followed by a social survey too. From this, we can follow up the development of these abilities of children year by year, and the performance of schools too. The problem is, that most of the students, parents, and even much of the teachers don't know, what this measurement is really for, and what the results really mean. It is hard to compare the ability points to the traditional school marks. That's why I tried to make a kind of comparison of the two types of evaluations. Because whatever we say, the achievement (marks) is important in school life (and further too), and we would like to see, how skills and abilities come out in achievement.

For the above mentioned purposes I wanted to examine the skills and abilities, so I made another kind of assessment, according to dr. István Czeglédy, used in a survey in Miskolc, among elementary school students. [4] I identified which of the following items can appear in the solution of each exercises (in a more complex one, all of them, in a simple, some), and gave simply 1 or 0 points, if an item appeared or not:

- a) does he/she begins the exercise?
- b) does he/she interprets the exercise well?
- c) is there any valuable in his/her work?
- d) does he/she make figure, tablet, systematize data?
- e) is the figure, tablet, systematization valid for the solution?
- f) does he/she use notations?
- g) does he/she make plan?
- h) is his/her solution purposive (even if there was no written plan)?
- i) is he/she motivated to solve the exercise?
- j) does he/she explain statements?
- k) does he/she look for causal connections, relationships?
- l) does he/she try for visualize in short forms?
- m) does he/she try for whole solution, give full answer?

The competence points identified for the exercises are the following:

	1st exercise	2nd exercise	3rd exercise	4th exercise	5th exercise
9th grade	Balance of scales	Eggs for Easter	Combinatory	Imprint	Train
Altogether: 54 points	a,b,c,d,e,i,j,k,m	a,b,c,d,e,f,g,h,i,j,k,l,m	a,b,c,d,e,f,g,h,i,j,k,l,m	a,b,c,i,m	a,b,c,d,e,f,g,h,i,j,k,m
10th grade	Spinning dice	Selling shirts	How old is the captain?	PIN-code (comb.)	Place of the well
Altogether: 50 points	a,b,c,i,m	a,b,c,f,g,h,i,j,k,l,m	a,b,c,g,h,i,j,k,l,m	a,b,c,f,g,h,i,j,k,l,m	a,b,c,d,e,f,g,h,i,j,k,l,m
11th grade	Phone numbers	Lamps in the cellar	Combinatory	Imprint	Train in tunnel
Altogether: 49 points	a,b,c,g,h,i,j,k,m	a,b,c,f,g,h,i,j,k,l,m	a,b,c,d,e,f,g,h,i,j,k,l,m	a,b,c,i,m	a,b,c,d,e,f,g,h,i,j,k,m
12th grade	Filling dishes	Taxing in Zed	Cost of the horse	Queer money	Tangential trapezoid
Altogether: 54 points	a,b,c,i,k,m	a,b,c,f,g,h,i,j,k,l,m	a,b,c,d,e,f,g,h,i,j,k,l,m	a,b,c,d,e,g,h,i,j,k,m	a,b,c,d,e,f,g,h,i,j,k,l,m

Table 6: Competence points to each exercises

Evaluating the papers according to these viewpoints I summarized the competence points of all students per grade in an Excel tablet. The next step was to compare the achievement and competence points of students. How could I show, which students achieved above, according to, or below the level of their competences? For examining this I took the quotient of the students' competence points (C) and achievement points (A). My hypothesis was, that there can be a special relation between the two kind of assessments, and I tried to show and analyze it. I adjusted the students (per grades) into order of growing achievement points, and represented the C/A quotient on graphs. On the graphs we can see how many students did the test per grades, and the value of the C/A. I identified the "ideal" value of the quotient (maximal achievement points/maximal competence points) for every grade, the average and deviation of grades too (see figures 5-8). The graphs show interesting coherence.

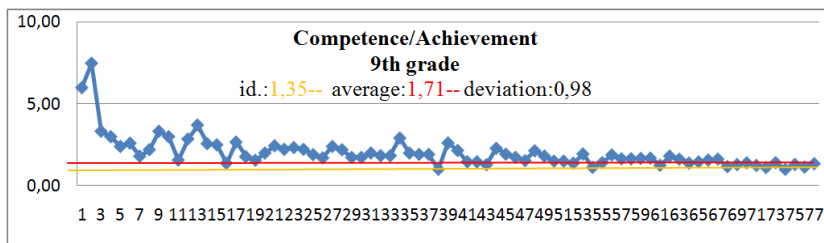


Figure 5: C/A in order of growing mathematical achievement, 9th grade

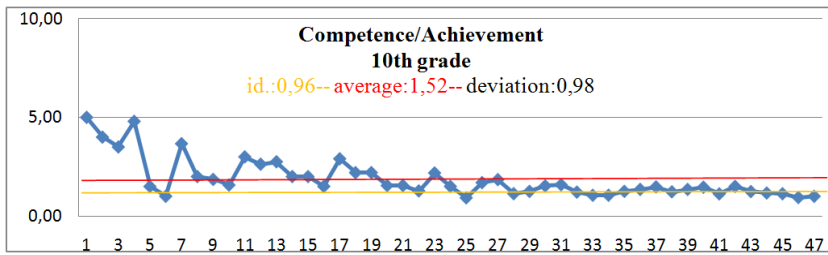


Figure 6: C/A in order of growing mathematical achievement, 10th grade

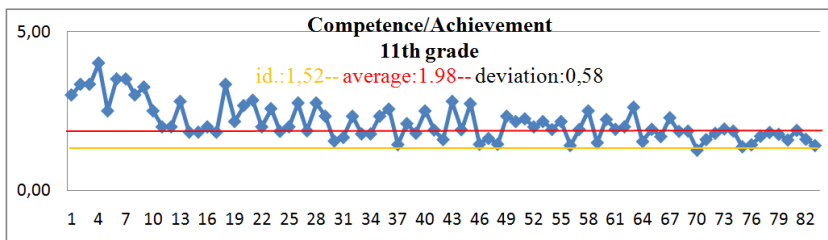


Figure 7: C/A in order of growing mathematical achievement, 11th grade

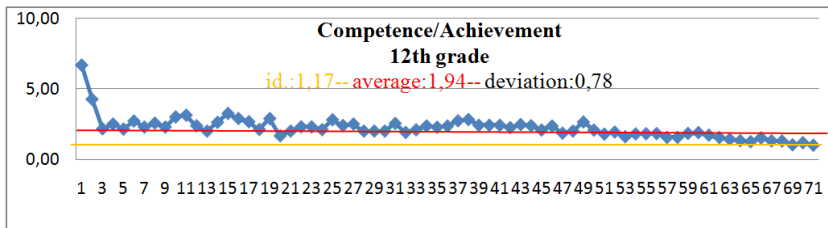


Figure 8: C/A in order of growing mathematical achievement, 12th grade

The graphs show the ideal level of C/A with orange line, and the average of the grade with red line (see figures 5–8). How can the C/A quotient be interpreted? It means that more the level of the quotient is above the ideal line, the student’s mathematical achievement is more below his/her competences. If it is close to the ideal value, it shows, the student’s achievement adequates to his/her competences. And if it is below the ideal line, the student’s achievement is beyond his/her competences.

On every grades the graphs show an exponential-like decrease. This means that if one’s achievement is better, his/her competences come up better in achievement. This also shows, that worse achievement doesn’t always mean worse skills and abilities, but other factors too: lack of motivation, deficiency of previous knowledge, reading or understanding problems, etc. Mapping this needs a longer examination.

But such an analysis would help the maths teachers a lot in evaluating how the students can show up their mathematical competences in solving exercises. On the other hand it would help to decide which way to develop the students: some skills, abilities are weak, or the knowledge. We can see students achieving above and below the competences on every achievement levels, but if we look at the right end of the graphs, (figures 5–8) we can see, that those students, whose achievement points were better, all “brought out” their skills, abilities. In all, more students achieved below his/her competences, than above (see table 7).

<i>Achievement</i>	<i>below competences</i>	<i>adequate to competences</i>	<i>above competences</i>
9th grade	43 students (56%)	22 students (28%)	12 students (16%)
10th grade	26 students (55%)	19 students (40%)	2 students (5%)
11th grade	55 students (67%)	23 students (27%)	5 students (6%)
12th grade	56 students (79%)	12 students (17%)	3 students (4%)

Table 7: Number of students and percentage of the level of C/A

These numbers show that the main problem is, that for much students their skills and abilities don’t come up in solving exercises as mathematical achievement – this could be developed by thorough knowledge, well structured and deliberate matters. I think, we’d need less, but in practical problems better applicable knowledge in our maths education. Nowadays after acquiring new matters (70–75% of all lessons) we have little time, only the 15–20% of the lessons, to practice and deepen the knowledge. I think, we’d have to change this rate into the growth of practising to get better results in problem solving achievements.

## 7. Further examinations, working out development methods

Beyond the examination of the present level of skills and abilities the statistical analysis of competence points is also helpful to find out which of them accrued the least during solving the exercises. From the summary of the competence points I pried those points out, which appeared in less than one third of the students’ solutions – I think these are the competences principally needs to be developed. There are some exercises, which were solved by more than one third of the students (4th exercise on 9th grade, the first exercise on 10th grade, the 4th exercise on 11th grade and the first exercise on 12th grade), so I didn’t mention them – these were short, mainly “choosing from given answers”-type exercises. The next tablet shows the weaknesses (table 8):



9th grade	1st ex.: d, e, k, m 2nd ex.: f, g, h, j, l, m 3rd ex.: f, g, h, j, l, m 5th ex.: d, e, f, g, h, j, l, m	11th grade	1st ex.: g, j, 2nd ex.: b, c, f, g, h, i, k, l, m 3rd ex.: f, g, j, l 5th ex.: d, e, g, h, j, l, m
10th grade	2nd ex.:h, j, l, m 3rd ex.:g, h, i, j, l, m 4th ex.: f, g, h, j, l 5th ex.: f, g, h, i, j, k, l, m	12th grade	2nd ex.: f, g, h, i, j, l, m 3rd ex.: d, e, f, j 4th ex.: d, e, g, h, j, m 5th ex.: g, h, i, j, l, m

Table 8: The worst competence points in exercises

The above statistics show that most defects are

- in planning (g),
- in purposive solution (h),
- in explaining statements (j),
- in visualizing results in short forms (l),
- in trying to give full answer (m).

The present examination was the first step of a longer research, and would be followed by further comparing assessments according to my aims. The results show that we should start development on the following topics:

- “translating” exercises, problems from vernacular words to mathematical symbols,
- working out typical exercises, patterns for using mathematical tools,
- planning solutions of complex exercises, and working out solutions,
- developing exercise keeping ability, patience, extended attention,
- developing argumental and proofing abilities.

The next step of my research is to work out exercise papers on the above mentioned topics, which are applicable for a 45 minutes long lessons and a whole class of 30-35 students. In point of methods, I think, both cooperative and individual learning forms have got their own place in learning mathematics, so as in development work too. My opinion is, that variable usage of different methods could be the most efficient, because none of these forms result increase on its own – we have to find the right rates. The most difficult is to develop those psychical abilities, which are needed for long lasting concentrations, and dividing a complex exercise to parts. But these abilities are essential to social integration and lifelong learning and development. This also poses the question of too much materials – we’d need less, to get time for deepen knowledge. It’s a pity, that the teenagers nowadays see, that these attributes aren’t respected. The world of media and internet, from which the students get most of their information, prefers quick, easy-to-get, “instant” things,

and don't relay the hard work behind the commanding achievements. I think in education we have to make stress not only to adapt ourselves to changing technical circumstances and matters, but also to put the negative changes of society to right path.

After setting the development papers into the course of maths lessons and testing the usage I'd like to assess the students once more. After comparing the results with the first one comes the re-working of exercise papers, or the following of usage. My aim is to work out such methods for developing competences, that can be fitted into curriculum, don't take extra time and efforts to use, don't "set back" the execution of given matters, can be used in general secondary school classes within the present circumstances, the working forms accuring variedly, and results increase in mathematical achievement too.

For motto I chose the words of a great educator, old, but still valid: "From wherever we see, the aim of our didactics must be to ferret out and hunt up the practise of education, so as to teachers should teach less, in the same time the students should learn more. In terms of this didactics let there be less confusion in schools, but more freedom, pleasure, and impresses a real development on all." (Comenius, 1657)

At the end I'd like to thank for the constructive critic I got from the anonym reviewers, they made my paper better, and I got good ideas and inspirations from them for my future work.

## References

- [1] AMBRUS, A., Introduction to mathematics didactics, *ELTE Eötvös Kiadó*, Budapest 2004.
- [2] BALÁZSI, I., OSTORICS, L., SZALAY, B., PISA summary report 2006, [www.oecd-pisa.hu/PISA2006Jelentes.pdf](http://www.oecd-pisa.hu/PISA2006Jelentes.pdf), 2007.
- [3] CZEGLÉDY, I., OROSZ, GY., SZALONTAI, T., SZILÁK, A., Mathematics subject pedagogy, *Bessenyei György Könyvkiadó*, Nyíregyháza, 2005.
- [4] CZEGLÉDY, I., Total mathematics ability assessment of 5th grade elementary students of Miskolc, *Miskolci Pedagógus*, 2006/41.
- [5] CSAPÓ, B., Growth of abilities and development in school, *Akadémiai Kiadó*, Budapest 2003.
- [6] SKEMP, R., Psychology of learning mathematics, *Gondolat Kiadó*, Budapest 1975.
- [7] National Basis Curriculum  
[www.nefmi.gov.hu/kozoktatas/.../nemzeti-alaptanterv-nat](http://www.nefmi.gov.hu/kozoktatas/.../nemzeti-alaptanterv-nat)
- [8] OECD: Measuring Student Knowledge and Skills – A new Framework for Assessment, OECD, *Programme for International Student Assessment (PISA)*, 1–104, Paris, France, 1999.
- [9] NISS, M., Quantitative Literacy and Mathematical Competencies, [www.maa.org/ql/pgs215\\_220.pdf](http://www.maa.org/ql/pgs215_220.pdf)

**Ilona Téglási**

Institute of Mathematics and Computer Science

Eszterházy Károly College

Leányka str. 4

Eger

Hungary

e-mail: [olahneti@ektf.hu](mailto:olahneti@ektf.hu)

