Contents

A. Akhmametieva, Steganalysis of digital contents, based on the analysis of unique color triplets . . . 3
J. Caha, J. Dvorský, First derivatives of fuzzy surfaces . . . 19
D. Callan, T. Mansour, M. Shattuck, Twelve subsets of permutations enumerated as maximally clustered permutations . . . 41
P. Dospra, D. Poulakis, Determining special roots of quaternion polynomials . . . 75
D. H. Jin, J. W. Lee, Generic lightlike submanifolds of an indefinite trans-Sasakian manifold with a non-metric ϕ-symmetric connection . . . 85
I. Juhász, Gardener's spline curve . . . 109
N. Kilic, The $h(x)$-Lucas quaternion polynomials . . . 119
R. Klén, Hyperbolic distance between hyperbolic lines . . . 129
V. Laohakosol, P. Tangsupphathawat, A combinatorial generalization of the gcd-sum function using a generalized Möbius function . . . 141
Sz. Márien, Decision structure based object-oriented design principles . . . 149
B. K. Patel, U. K. Dutta P. K. Ray, Period of balancing sequence modulo powers of balancing and Pell numbers . . . 177
Á. Tóth, Comparison and affine combination of generalized barycentric coordinates for convex polygons . . . 185
E. Troll, M. Hoffmann, Caustics of spline curves . . . 201
T. Yamada, Infinitary superperfect numbers . . . 211
O. Zubelevich, On bounded and unbounded curves determined by their curvature and torsion . . . 219

Methodological papers

I. Gerják, Teaching digital image processing – eyes and eyesight . . . 229
Sz. Petz, M. Hoffmann, The development of mathematical competences in Hungarian teacher training education . . . 243

ANNALES MATHEMATICAE ET INFORMATICAE 47. (2017)

ANNALES MATHEMATICAE ET INFORMATICAE

TOMUS 47. (2017)

EGER 1774
ESZTERHÁZY KÁROLY UNIVERSITY

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 University of Applied Sciences 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 Only new and unpublished material can be accepted.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 LaTeX. If you have any problems or questions, please write an e-mail to the managing editor Miklós Hoffmann: `hofi@ektf.hu`

The volumes are available at `http://ami.uni-eszterhazy.hu`

ANNALES MATHEMATICAE ET INFORMATICAE

VOLUME 47. (2017)

INSTITUTE OF MATHEMATICS AND INFORMATICS
ESZTERHÁZY KÁROLY UNIVERSITY OF APPLIED SCIENCES
HUNGARY, EGER

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

A kiadásért felelős az
Eszterházy Károly Egyetem rektora
Megjelent a Líceum Kiadó gondozásában
Kiadóvezető: Nagy Andor
Felelős szerkesztő: Zimányi Árpád
Műszaki szerkesztő: Tómács Tibor
Megjelent: 2017. december Példányszám: 30

Készítette az
Eszterházy Károly Egyetem nyomdája
Felelős vezető: Kérészy László

Annales Mathematicae et Informaticae
47 (2017) pp. 3–18
http://ami.uni-eszterhazy.hu

Steganalysis of digital contents, based on the analysis of unique color triplets

Anna Akhmametieva

Department of Informatics and Management of Information Systems Security,
Odessa National Polytechnic University, Ukraine
anna-odessitka@mail.ru

Submitted June 24, 2016 — Accepted December 22, 2016

Abstract

The new steganalytic algorithm for detection of the presence of additional information that embeds into digital images and digital videos by LSB Matching method with a small hidden capacity (not more than 0.5 bpp) is presented. The proposed steganalytic algorithm analyses digital content in the spatial domain and is based on the accounting of sequential color triads in the matrix of unique colors of the digital content. Steganalytic algorithm has a high effectiveness of detecting the additional information embedded into one arbitrary color component of the container with a small hidden capacity.

Keywords: Steganalysis, LSB Matching, the spatial domain of the container, a digital image, a digital video

MSC: 68U10, 94A08, 68P30

1. Introduction

The rapid development of information and communication technologies leads to their wide distribution in the state, public and household sectors, it is possible easily and quickly to transfer any information to long distances. If in the state activities secure channels of communication applies, then such open channels as e-mail, social networks allows to exchange externally innocuous data, therefore they are often used with criminal intentions. Open access to the Internet and scientific resources allows you to track newest developments in the field of information security, steganography and steganalysis. The use of steganographic methods and algorithms allows to transfer confidential information via open communication channels by hiding the fact of its presence in the transmitted content. In a competitive environment, in the conditions of the spread of terrorism the hidden communication can lead to significant losses for businesses and to the catastrophic consequences of terrorist attacks for the society in general.

To prevent the criminal acts with using steganography it is extremely important to develop steganalysis aimed at the detecting the fact of the presence/absence of hidden information in any digital content (see [1]). Digital images, audio or video sequences can be used as containers in steganography.

Ones of the most widespread steganographic methods are different variations of the method of modification of the least significant bit (LSB Matching, LSB Replacement etc.) due to the simplicity of realization and the possibility of its use both in the spatial domain and in the transformation domain. Nevertheless, continuous improvement of steganographic developments impedes using of LSB method with a high hidden capacity, because it is easy to detect such embedding. Therefore, to ensure concealing of the secret communication LSB method is often used with a small hidden capacity (less than 0.5 bpp), which greatly complicates the process of detecting the presence/absence of the additional information.

A large number of steganalysis developments aimed at the detection of the presence/absence of additional information, embedded by LSB Matching method into digital images. There are enough effective steganalysis methods and algorithms (see [2, 3, 4, 5, 6, 7, 8]), that analyse digital images in the transformation domain (frequency domain, the singular/spectral decompositions of the corresponding matrices, etc.), however transfer of a digital content to transformation domain and back leads to the additional accumulation of computational errors, which considerably complicates the process of detecting presence of additional information that embeds with a small hidden capacity.

The steganalytic methods that analyse spatial domain of digital contents allows to avoid both additional time expenses, and accumulation of computational errors, however existing developments (see [9, 10, 11, 12, 13]) often have low effectiveness when embedding of additional information is carried out by LSB Matching method with a small hidden capacity.

The use of a digital video as a container in steganography allows to transfer a significant amount of data due to the large number of frames, using at the same time a small hidden capacity. It is very problematic to detect the presence of additional information in these conditions. However, despite advantages of using of digital videos, in open access there are not a lot of works devoted to video steganalysis. Widely spread are methods that analyses spatial (see [14]) and temporal (see [15, 16]) domains of digital video, as well as specialized methods directed against steganographic tools such as embedding of additional information to the motion vectors (see [17]) or H.264/AVC (see [18]).

Taking into account the advantages of steganalysis in the spatial domain, the aim is to develop steganalytic algorithm aimed to detect embedding of additional information by LSB Matching with a small hidden capacity (no more than 0.5 bpp) into digital containers, which are color digital images and digital videos.

2. Research essence

As containers we consider color digital images and digital videos stored according to the RGB color scheme. We will use term "cover" (cover-image or cover-video) for the unfilled containers. Each video sequence V consists of frames $F_l(m,n)$, where $l = \overline{1,K}$, K - number of frames, m - frame height, n - frame width. Additional information, which is a binary sequence, embeds by LSB Matching into the spatial domain of the randomly selected color component of the container. Result of embedding of additional information to the container we will call stego (stego-image or stego-video). It should be noted that if as a container the image in a losses format is used, it will be resaved in a lossless format after embedding of additional information.

LSB Matching method is realized according to the formula (see [10]):

$$p_s(i,j) = \begin{cases} p_c(i,j)+1 & \text{if } b \neq LSB(p_c(i,j)) \ \&\ r > 0, \\ p_c(i,j) & \text{if } b = LSB(p_c(i,j)), \\ p_c(i,j)-1 & \text{if } b \neq LSB(p_c(i,j)) \ \&\ r < 0, \end{cases} \tag{2.1}$$

where $p_c(i,j)$, $p_s(i,j)$ - the brightness value of the pixel of the color matrix of the original digital image/frame and stego respectively, b - bit of the secret message, r - a random value in the range $[-1,+1]$, $LSB(p)$ - the least significant bit of p (see [10]). Thus, embedding of additional information or will increase pixel's brightness value of an original matrix on 1 $(+1)$, or will reduce it on 1 (-1), or will remain it unchangeable (0).

Each frame of the video sequence V is an image formed by three color components: red, green and blue matrices of size $m \times n$. Accordingly, each pixel of the frame/image is represented as a triplet of values (R,G,B). All triplets (r_i, g_i, b_i), $i = \overline{1,k}$, $k = m \cdot n$, that occur in a digital image/frame of a digital video form some matrix CT (color triplets) of size $k \times 3$. Triplets can repeat in the matrix CT, depending on that how often they appear in a digital image/frame of a digital video.

Definition 2.1. All of the triplet's various values (R,G,B) we will call unique colors, their number is denoted by U.

Definition 2.2. A matrix of size $U \times 3$ of ordered unique colors (r_j, g_j, b_j), $j = \overline{1,U}$ we will call the matrix of unique colors UCT (unique color triplets).

The matrix of unique colors in MathWorks MatLAB can be received by standard procedure $UCT = unique(CT,'rows')$, where parameter $'rows'$ denotes that the matrix UCT will contain unique rows (triplets) of matrix CT. Thus, the matrix UCT is an ordered sequence of unique colors which at least once occur in the analyzed digital image/frame of a digital video. The matrix of unique colors does not take into account the frequency of appearance of some triplet $(R, G, B) \subset UCT$ in the container.

Consider what changes the matrix of unique colors will be undergone when additional information embeds to the container by LSB Matching according to the formula 2.1.

Let to some pixel of the container (for example, a digital image) forming the triplet $(95, 116, 68)$ additional information embeds into a green color component. This triplet will be included in the matrix of unique colors of the container as it at least once occurs in the digital image. Embedding of the bit of information into a green color component of the pixel can change it value or on $(95, 115, 68)$, or on $(95, 117, 68)$, or to leave it without change. Let the pixel value will take $(95, 117, 68)$. If the triplet is founded in the digital image only once, then matrix of unique colors of stego after embedding of additional information will not contain original triplet $(95, 116, 68)$, but its modification $(95, 117, 68)$ will appear. However, many different pixels in the container can have the same color, i.e. the triplet occurs in the container as a rule repeatedly. Therefore, after embedding of additional information into different pixels of the same color, they can be modified by all three ways: $+1, -1, 0$, that will lead to appearance of all three modifications of triplets in the matrix of unique colors of stego (in this example, $(95, 115, 68)$, $(95, 116, 68)$ and $(95, 117, 68)$).

Similar modifications occur when additional information embeds into red or blue color components.

Thus, in case of embedding of additional information into container in the matrix of unique colors there will be additional triplets differing from original on ±1 in that color component where embedding was carried out.

Introduce the following definitions.

Definition 2.3. Under a sequential Red-triad for the current triplet (r_j, g_j, b_j), $j = \overline{1, U}$, in the matrix of unique colors we will understand execution of the condition:

$$(r_j, g_j, b_j) \subset UCT \text{ AND } (r_j - 1, g_j, b_j) \subset UCT \text{ AND } (r_j + 1, g_j, b_j) \subset UCT,\\ j = \overline{1, U}.$$

The Red-triad corresponds to the red color component of the container.

Definition 2.4. Under a sequential Green-triad for the current triplet (r_j, g_j, b_j), $j = \overline{1, U}$, in the matrix of unique colors we will understand execution of the condition:

$$(r_j, g_j, b_j) \subset UCT \text{ AND } (r_j, g_j - 1, b_j) \subset UCT \text{ AND } (r_j, g_j + 1, b_j) \subset UCT,\\ j = \overline{1, U}.$$

The Green-triad corresponds to the green color component of the container.

Definition 2.5. Under a sequential Blue-triad for the current triplet (r_j, g_j, b_j), $j = \overline{1, U}$, in the matrix of unique colors we will understand execution of the condition:

$$(r_j, g_j, b_j) \subset UCT \text{ AND } (r_j, g_j, b_j - 1) \subset UCT \text{ AND } (r_j, g_j, b_j + 1) \subset UCT,$$
$$j = \overline{1, U}.$$

The Blue-triad corresponds to the blue color component of the container.

Definition 2.6. The basic triad is a sequential triad corresponding to that color component of the container, into which embedding of additional information was carried out.

Definition 2.7. Concomitant triads are that sequential triads that corresponds to unfilled color components of the container.

I.e. after embedding of additional information into blue color component the basic triad is Blue-triad and concomitant triads are Red- and Green-triads.

The computational experiment that analyses the quantity of Red-, Green- and Blue-triads in unfilled digital containers was carried out.

There are the following digital images as containers:

1. Set 1: 203 color digital images from [19] in JPG format;

2. Set 2: 201 high-quality digital images from [20] in JPG format;

3. Set 3: 215 images received by non-professional photo cameras in JPG format;

4. Set 4: 200 color digital images from [19] in TIFF format;

5. Set 5: 200 images received by non-professional photo cameras in TIFF format.

The quantity of sequential triads in the matrix of unique colors of unfilled digital containers stored in losses format (Set 1, 2, 3) does not exceed 3% of total number of unique colors, unlike containers stored in lossless format (Set 4, 5) where even in the absence of additional information the quantity of Red-, Green- and Blue-triads reaches 40-60% due to the lack of compression and, consequently, a large variety of unique colors. At the same time it is noted that the relative quantity of Red-, Green- and Blue-triads (relative to the number of unique colors) in unfilled digital images is comparable by values, i.e. the difference is not more than 1-1.5%.

Analyze how the number of sequential triads in the matrix of unique colors will change when additional information embeds into one arbitrary color component of the container with different values of hidden capacity. Consider as a container the digital image in losses format (Figure 1, a), into which additional information is embedded into a red color component by LSB Matching with different values of hidden capacity. Counting of Red-, Green- and Blue-triads in the matrix of unique colors of the formed stego was carried out. Counting of each kind of sequential triads dictated by the fact that in the process of steganalysis it is unknown, into what kind of color components additional information has been embedded. The percentage of Red-, Green- and Blue-triads in relation to the total number of unique colors of stego formed by embedding of additional information to the red color component with different values of hidden capacity is shown in Figure 1, b. Similarly, the quantity of Red-, Green- and Blue-triads in matrices of unique colors of stegos formed by embedding of additional information only into green color component (Figure 1, c) and only into blue color component (Figure 1, d) is determined.

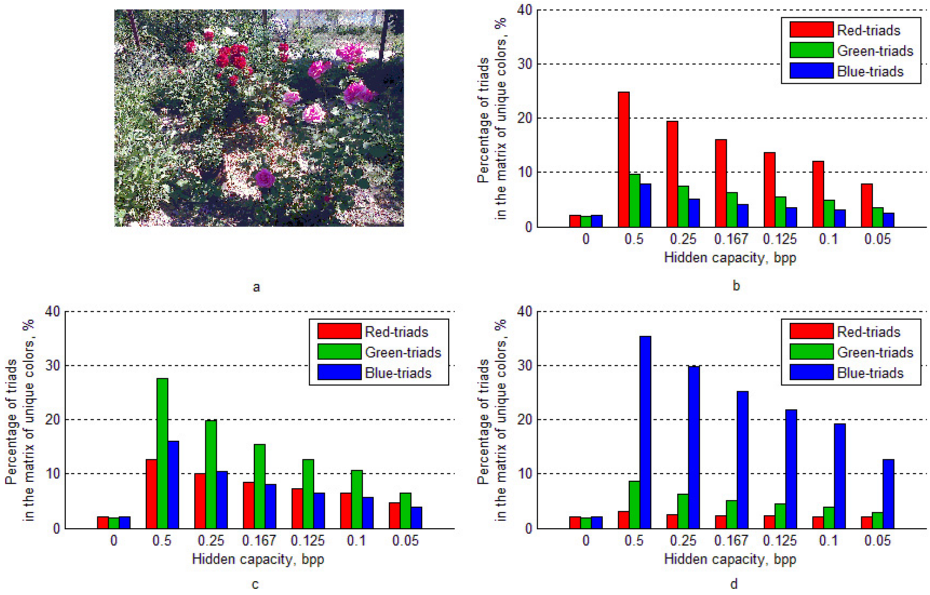


Figure 1: The quantity of sequential color triads in the image stored in losses format: a – original digital image from Set 3; b - percentage of sequential triads in the stego formed by embedding of additional information by LSB Matching into red color component; c - percentage of sequential triads in the stego formed by embedding of additional information by LSB Matching into green color component; d - percentage of sequential triads in the stego formed by embedding of additional information by LSB Matching into blue color component

As can be seen from Figure 1, embedding of additional information into one arbitrary color component causes a significant increase in the quantity of basic triads. At the same time the number of concomitant triads increases too, however less in comparison with the quantity of basic triads in the matrix of unique color of stego. This growth is associated with an increase in the number of unique colors when additional information has been embedded and is random. For example, embedding of additional information into blue color component leads to change of unique triplet $(10, 158, 74)$ on Blue-triad $(10, 158, 73)$, $(10, 158, 74)$ and $(10, 158, 75)$. If there are triplets $(11, 158, 75)$ and $(12, 158, 75)$ in the matrix of unique colors of stego then they with a new triplet $(10, 158, 75)$ forms Red-triad, but embedding of additional information into red color component has not performed.

If as the container to use the image in lossless format then embedding of additional information practically does not influence the percentage of consecutive triads, what is shown in Figure 2, where into image in a lossless format (Figure 2, a) additional information embeds into a red color component (for example), the percentage of Red-, Green- and Blue-triads for which is shown in Figure 2, b. The similar situation is observed when additional information embeds into another color component and is typical for all digital images in a lossless format. Thus, further as containers only digital contents in losses format will consider.

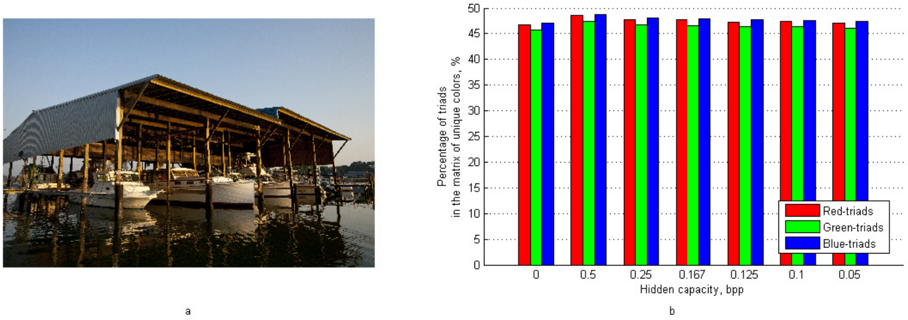


Figure 2: The quantity of sequential color triads in the image stored in lossless format: a – original digital image from Set 4; b - percentage of sequential triads in the stego formed by embedding of additional information by LSB Matching into red color component

Figure 1 shows that the percentage of concomitant triads in the matrix of unique colors of stego increases on average up to 8-12%. Accept preliminary thresholds of the quantity of sequential triads $T_{low} = 2.5$ and $T_{up} = 8$ for detection of the presence/absence of additional information in the digital content. On the basis of digital images from the Set 1, 2, 3 the computational experiment, determining percentage of Red-, Green- and Blue-triads in matrices of unique colors of original containers (hidden capacity 0 bpp) and stegos formed by embedding of additional information only into one arbitrary color component with different values of hidden capacity was carried out. Results of the experiment are shown in Table 1, where pc – the quantity (in %) of all sequential triads (and basic, and concomitant) in matrices of unique colors of digital contents, $max(pc)$ - maximum percentage of sequential triads in matrices of unique colors of digital contents (for hidden capacity 0.05-0.5 bpp $max(pc)$ corresponds to the maximum containing of basic triads in stego sets).

Thus, the computational experiment showed that an average of 90% of digital containers stored in losses format, contain no more than 2.5% Red-, Green- and Blue-triads, their number significantly increases after embedding of additional information. Therefore, the condition for the original container is performance of the relationship:

Set	Threshold	Hidden capacity, pbb						
		0.5	0.25	0.167	0.125	0.1	0.05	0
Set 1	$pc \leq T_{low}$	3.61	6.24	8.21	8.70	9.20	13.30	**94.42**
	$pc > T_{up}$	**89.66**	**80.79**	**72.58**	**64.37**	**56.49**	**35.80**	0.33
	$max(pc)$	56.94	47.44	47.68	45.33	43.22	31.36	9.08
Set 2	$pc \leq T_{low}$	3.15	5.31	6.30	8.62	11.61	18.24	**86.24**
	$pc > T_{up}$	**84.25**	**69.98**	**64.18**	**55.89**	**53.57**	**31.01**	0.83
	$max(pc)$	54.58	51.67	48.76	46.21	41.47	36.02	11.25
Set 3	$pc \leq T_{low}$	1.40	4.65	5.89	6.82	7.60	9.92	**92.25**
	$pc > T_{up}$	**90.70**	**85.74**	**77.36**	**68.22**	**61.24**	**46.98**	0
	$max(pc)$	55.08	51.65	48.28	45.32	43.80	38.53	4.69

Table 1: The percentage of sequential triads in the matrix of unique colors of digital images

$$(pR \leq T_{low}) \text{ AND } (pG \leq T_{low}) \text{ AND } (pB \leq T_{low}),$$

and back again for stego:

$$(pR > T_{low}) \text{ OR } (pG > T_{low}) \text{ OR } (pB > T_{low}),$$

where $T_{low} = 2.5$, pR - the percentage of Red-triads in the matrix of unique colors, pG - the percentage of Green-triads in the matrix of unique colors, pB - the percentage of Blue-triads in the matrix of unique colors.

Threshold $T_{up} = 8$ promotes to correctly detection of unfilled containers, providing the protection against the appearance of "false alarms".

Based on the established features of the changes in the number of color triads in the matrix of unique colors the steganalytic algorithm for detecting embedding of additional information by LSB Matching into spatial domain of digital containers (digital images and digital videos) is proposed. If digital images are as containers, it is necessary only first step of the algorithm for detecting.

2.1. Designations employed in the algorithm

$F_l(m,n), l = \overline{1,K}$ - frame of the analyzed video sequence V consisting of K frames of size $m \times n$.

$resultF$ - a matrix of size $K \times 3$, containing the result of the detection on each frame of the video sequence (sequence of digital images). In the case of a single digital image the matrix $resultF$ is a matrix of size 1×3 and is the end result of the detection. The first column of the matrix corresponds to red color component, the second – to green color component, and the third - to blue color component. Value 1 of the matrix corresponds to the presence of additional information in the corresponding color component, 0 - to its absence.

UCT - a matrix of unique colors of the frame F_l (image I) of size $U \times 3$, containing unique triplets $(r_j, g_j, b_j), j = \overline{1,U}$.

$countR$, $countG$, $countB$ - number of sequential triads in the UCT for red, green and blue color components, respectively.

pR, pG, pB - percentage of sequential triads in the UCT in relation to total number of unique colors for red, green, blue color components, respectively.

kR_{pos}, kG_{pos}, kB_{pos} - number of positive definite frames as containing embedded additional information in red, green and blue color components of the frame, respectively.

kR_{neg}, kG_{neg}, kB_{neg} - number of negative definite frames as not containing embedded additional information in red, green and blue color components of the frame, respectively.

2.2. Steganalytic algorithm

Step 1 (for digital images and frames of video sequence). For each image I (frame F_l, $l = \overline{1, K}$) the detection of the presence of additional information performs.

1. Forming of the matrix of unique colors UCT of the digital image I / frame F_l of video sequence V.

2. Counting of Red-, Green-, Blue-triads.

   (a) If for the current triplet $(r_j, g_j, b_j), j = \overline{1, U}$ in the UCT at the same time there are triplets $(r_j + 1, g_j, b_j)$ and $(r_j - 1, g_j, b_j)$ then $countR = countR + 1$;

   (b) If for the current triplet $(r_j, g_j, b_j), j = \overline{1, U}$ in the UCT at the same time there are triplets $(r_j, g_j + 1, b_j)$ and $(r_j, g_j - 1, b_j)$ then $countG = countG + 1$;

   (c) If for the current triplet $(r_j, g_j, b_j), j = \overline{1, U}$ in the UCT at the same time there are triplets $(r_j, g_j, b_j + 1)$ and $(r_j, g_j, b_j - 1)$ then $countB = countB + 1$.

3. To compute:

$$pR = \frac{countR}{U} \cdot 100,\ pG = \frac{countG}{U} \cdot 100,\ pB = \frac{countB}{U} \cdot 100.$$

4. Detection of the presence/absence of additional information in the digital image I / single frame F_l, $l = \overline{1, K}$, of video sequence V.

   (a) if $(pR = max(pR, pG, pB))$ AND $(pR > T_{up})$
   then $resultF_{l,1} = 1$,
   else if

$$(pR > T_{low} \text{ OR } pG > T_{low} \text{ OR } pB > T_{low}) \text{ AND } (pR > 1.5 \cdot pG \text{ AND } pR > 1.5 \cdot pB)$$

   then $resultF_{l,1} = 1$,
   else $resultF_{l,1} = 0$;

(b) if $(pG = max(pR, pG, pB))$ AND $(pG > T_{up})$
then $resultF_{l,2} = 1$,
else if

$$(pR > T_{low} \text{ OR } pG > T_{low} \text{ OR } pB > T_{low}) \text{ AND } (pG > 1.5 \cdot pR \text{ AND } pG > 1.5 \cdot pB)$$

then $resultF_{l,2} = 1$,
else $resultF_{l,2} = 0$;

(c) if $(pB = max(pR, pG, pB))$ AND $(pB > T_{up})$
then $resultF_{l,3} = 1$,
else if

$$(pR > T_{low} \text{ OR } pG > T_{low} \text{ OR } pB > T_{low}) \text{ AND } (pB > 1.5 \cdot pR \text{ AND } pB > 1.5 \cdot pG)$$

then $resultF_{l,3} = 1$,
else $resultF_{l,3} = 0$.

Step 2 (for video sequences). Counting of positive and negative detection results in frames of a video sequence separately for each color component in the matrix $resultF$:

1. if $resultF_{l,1} = 1$
   then $kR_{pos} = kR_{pos} + 1$,
   else $kR_{neg} = kR_{neg} + 1$;

2. if $resultF_{l,2} = 1$
   then $kG_{pos} = kG_{pos} + 1$,
   else $kG_{neg} = kG_{neg} + 1$;

3. if $resultF_{l,3} = 1$
   then $kB_{pos} = kB_{pos} + 1$,
   else $kB_{neg} = kB_{neg} + 1$.

Step 3 (for video sequences). Detection of the presence/absence of additional information in a digital video:

1. if $kR_{pos} \geq kR_{neg}$
   then additional information contains in a red color component,
   else additional information is absent in a red color component;

2. if $kG_{pos} \geq kG_{neg}$
   then additional information contains in a green color component,
   else additional information is absent in a green color component;

3. if $kB_{pos} \geq kB_{neg}$
   then additional information contains in a blue color component,
   else additional information is absent in a blue color component.

3. Results of the experiment

In the computational experiment aimed at verifying the work of the proposed steganalytic algorithm, digital contents from the Set 1, 2, 3 and 367 video sequences of frame size 320 × 240 obtained by the mobile cameras (Set V) have been used. Each video contains in average 250 frames.

It should be noted that color photos and videos obtained by the cameras of mobile devices (Set 3 and Set V) are the most probable containers as the most widespread due to permanent presence smart phones, IPad or other mobile devices with itself. Original mobile videos are stored in a losses format and has the extension *.3gp or *.mp4. After embedding of additional information video are saved as uncompressed video in *.avi format.

Embedding of additional information was carried out into a randomly selected color component of digital images and videos with different values of hidden capacity: 0.5 bpp, 0.25 bpp, 0.167 bpp, 0.125 bpp, 0.1 bpp, 0.05 bpp. When additional information embeds into video sequence the selected color component is constant for all frames. Such embedding is caused by the fact that, as a rule, in case of steganography data transmission one component is used as the container and its choice is part of the secret key.

By results of experiment type I errors (False Negative FN) – the pass stego in case of its presence and type II errors (False Positive FP) – false detection stego in case of its absence (Table 2) were received.

Set	Errors	Hidden capacity, bpp					
		0.5	0.25	0.167	0.125	0.1	0.05
Set 1	FN	0	0	0	0	0	1.9704
	FP	0	0	0	0	0	0
Set 2	FN	0	0	0.4975	0	2.9851	7.9602
	FP	0	0	0	0	0	0
Set 3	FN	0	0	0	0	0	2.3256
	FP	0	0	0	0	0	0
Set V	FN	0	0	0	0.2725	0.8174	18.2561
	FP	0	0	0	0	0	0

Table 2: The effectiveness of detecting the presence/absence of additional information in digital contents, %

Table 2 shows that errors in detecting the presence of embedding of additional information in a digital content are very small, even for hidden capacity 0.05 bpp, indicating the high effectiveness of the proposed steganalytic algorithm.

For comparing of the effectiveness of the developed steganalytic algorithm with other existing methods for detection of embedding of additional information by LSB Matching in digital images stored in a losses format, the ROC-analysis is used.

ROC-curve analysis method applied to the steganalysis lies in realization of the testing of the group of digital images, that includes both unfilled containers, and stego, and it is known what each image is. Among analyzed digital images in a test group it is necessary to identify the stego (class V_1) and original containers (class V_2) using the developed steganalytic algorithm, as a result of which a positive $\delta = 1$ (stego) or negative $\delta = 0$ (cover) decision is adopted. Test results can be presented as Table 3 (see [3]), where TP is the quantity of correctly identified stego, TN is the quantity of correctly identified unfilled containers.

True state of digital image	Result of detection	
	$\delta = 1$ - stego	$\delta = 0$ - cover
class V_1 - stego	TP	FN
class V_2 - cover	FP	TN

Table 3: Results of detection of the test group of digital images

Obtained detection results are presented in the two-dimensional ROC-space, where the X-axis represents the specificity values that characterize the type II errors:

$$Sp = \frac{TN}{TN + FP},$$

and the Y-axis represents the sensitivity, characterizing the type I errors:

$$Se = \frac{TP}{TP + FN}.$$

In this study values Sp and Se are defined for different values of parameter T_{low}, based on which ROC-curves for values of hidden capacity 0.5, 0.25, 0.167, 0.125, 0.1 and 0.05 bpp shown in Figure 3, 4 are constructed. Figure 3 shows the ROC-space in the range $1 - Sp \subset \overline{0,1}$, $Se \subset \overline{0,1}$, Figure 4 shows a fragment of ROC-space in the range $1 - Sp \subset \overline{0, 0.025}$, $Se \subset \overline{0.8, 1}$.

Based on the constructed ROC-curves an integral parameter ρ characterizing the effectiveness of the studied steganalytic algorithm is obtained, where $\rho = 2A - 1$, A - the area under the ROC-curve. The values of area A and parameter ρ for different values of hidden capacity are given in Table 4.

Comparison of the effectiveness of the steganalytic algorithm based on the analysis of color triads with other modern analogues for digital images (Ker's (see [7]), Liu's (see [8]), HGE, NDH COM, RLH COM, Fused feature, Joint feature set (see [9]), SAVV (see [3])) is carried out by comparing the integral parameters ρ for the corresponding values of hidden capacity. Visual comparison of the above methods is shown in Figure 5 (see [3]).

As can be seen from Figure 5, the results of detecting the presence/absence of embedding of additional information by LSB Matching in digital images are far superior previous solutions (see [3, 7, 8, 9]), especially in the case of a small hidden capacity (0.25 bpp or less). The most revealing is the comparison of the

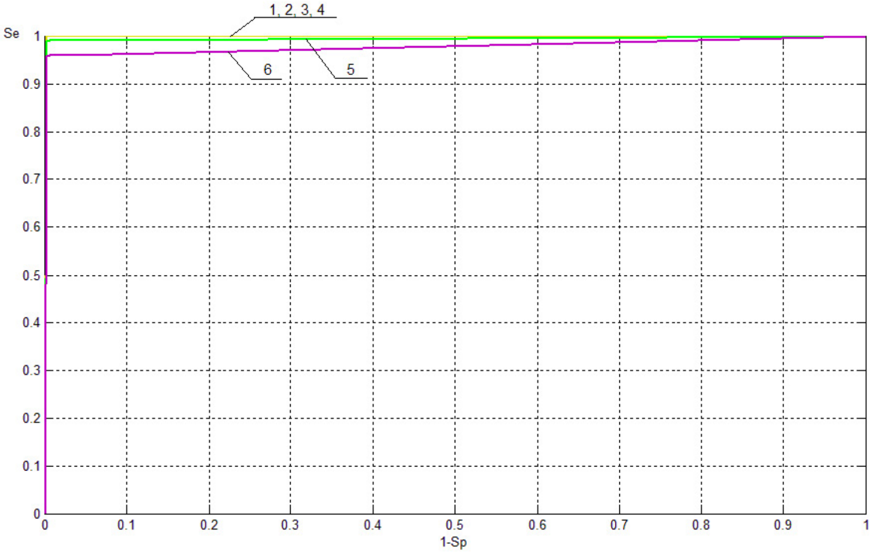


Figure 3: ROC-curves that characterizes the work of the steganalytic algorithm for detection of the presence of additional information embedded by LSB Matching into digital images with different values of hidden capacity: 1 - 0.5 bpp; 2 - 0.25 bpp; 3 - 0.167 bpp; 4 - 0.125 bpp; 5 - 0.1 bpp; 6 - 0.05 bpp

Hidden capacity, bpp	A	ρ
0.5	0.99822294	0.99644588
0.25	0.99822294	0.99644588
0.167	0.997393263	0.994786526
0.125	0.99822294	0.99644588
0.1	0.993238873	0.986477747
0.05	0.977475004	0.954950008

Table 4: Values of the integral parameter ρ for evaluating the effectiveness of the steganalytic algorithm for detection of the presence of additional information embedded by LSB Matching

developed algorithm with method SAVV, analyzing digital images with a small hidden capacity.

In modern papers devoted to steganalysis of digital videos [14-17] in computational experiments the methods of embedding of additional information that are different from LSB Matching, are used, that does not allow correctly to compare the effectiveness of the algorithm based on the accounting of color triads in the matrix of unique colors with other analogues. However, as seen from results of the computational experiment (Table 2, set V), the developed steganalytic algorithm

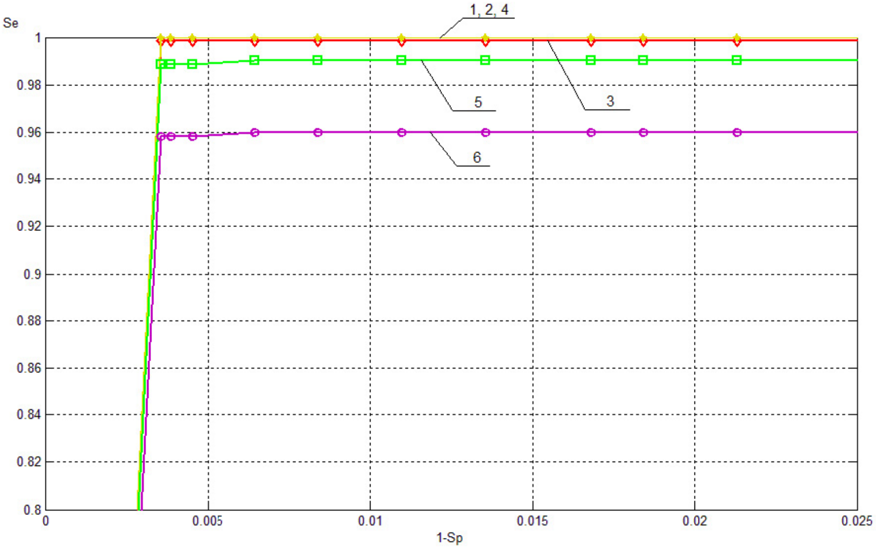


Figure 4: Fragment of ROC-curves in the ROC-space in the range $1 - Sp \subset \overline{0, 0.025}$, $Se \subset \overline{0.8, 1}$ that characterizes the work of the steganalytic algorithm for detection of the presence of additional information embedded by LSB Matching into digital images with different values of hidden capacity: 1 - 0.5 bpp; 2 - 0.25 bpp; 3 - 0.167 bpp; 4 - 0.125 bpp; 5 - 0.1 bpp; 6 - 0.05 bpp

is effective also in case of detection of the presence/absence of additional information in digital video sequences, at the same time both the frame analysis, and the analysis of digital video as a whole are possible.

4. Conclusions

In this paper a new steganalytic algorithm of detection of the presence/absence of additional information embedded by LSB Matching with a small hidden capacity (no more than 0.5 bpp) into one color component of digital images and digital videos stored in losses formats is proposed.

Comparison of the developed steganalytic algorithm with other modern tools of steganalysis showed that the algorithm based on the accounting of color triads in the matrix of unique colors is more effective than analogues, including for very small values of hidden capacity (0.1 and 0.05 bpp). The high effectiveness of the proposed algorithm for small values of hidden capacity is provided by its work in the spatial domain of digital contents, and as a result, there are no additional computational errors.

The proposed algorithm carries out an analysis of digital images and digital

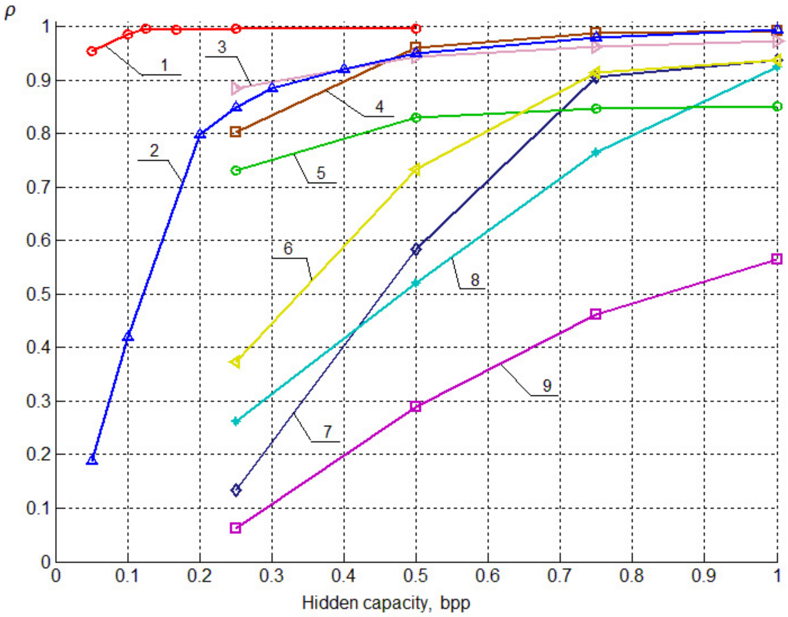


Figure 5: The results of comparing the effectiveness of the work of the steganalytic algorithm Color Triads with existing analogs for the detection of digital images: 1 - Color Triads, 2 - SAVV, 3 - Joint feature set, 4 - Liu's, 5 - RLH COM, 6 - Fused feature, 7 - Ker's, 8 - NDH COM, 9 – HGE

videos, for which its use can be expanded by analysis of single frames in the video sequence if additional information embeds not into all frames, but only into a small part of the total number of frames.

References

[1] Bohme, R., Advanced statistical steganalysis, Springer, 2010.

[2] Bobok, I.I., Steganalytic method for the digital signal-container stored in a losses format, *Modern Information Security*, Vol. 2 (2011), 50–60.

[3] Bobok, I.I., Application of ROC-analysis for integrated assessment of steganalysis method's efficiency, *Informatics and Mathematical Methods in Simulation*, Vol. 2, No. 3 (2012), 221–230.

[4] Alimoradi, D., The effect of correlogram properties on blind steganalysis in JPEG images, *Journal of computing and security*, Vol. 1, No. 1 (2014), 39–46.

[5] Visavalia, S.R., Ganatra A., Improving blind image steganalysis using genetic algorithm and fusion technique, *Journal of computer science*, Vol. 1 (2014), 40–46.

[6] Yamini, B., Sabitha R., Blind steganalysis: to analyse the detection rate of stego images using different steganalytic techniques with support vector machine classifier, *International journal of computer applications*, No. 2 (2014), 22–25.

[7] Ker, A.D., Steganalysis of LSB matching in grayscale images, *IEEE Signal Processing Letters*, Vol. 12, No. 6 (2005), 441–444.

[8] Liu, Q.Z., Sung A.H., Image complexity and feature mining for steganalysis of least significant bit matching steganography, *Information Sciences*, Vol. 178, No. 1 (2008), 21–36.

[9] Zhihua Xia, Lincong Yang, A Learning-Based Steganalytic Method against LSB Matching Steganography, *Radioengineering*, Vol. 20, No. 1 (2011), 102–109.

[10] Jun Zhang, Yuping Hu, Zhibin Yuan, Detection of LSB Matching steganography using the envelope of histogram, *Journal of computers*, Vol. 4, No. 7 (2009), 646–653.

[11] Geetha, S., Sindhu S., Kamaraj N., Close color pair signature ensemble adaptive threshold based steganalysis for LSB embedding in digital images, *Transactions on Data Privacy*, Vol. 1, Iss. 3 (2008), 140–161.

[12] Mitra, S., Roy T., Mazumdar D., Saha A.B., Steganalysis of LSB Encoding in Uncompressed Images by Close Color Pair Analysis, *IIT Kanpur Hackers' Workshop 2004 (IITKHACK04)*, 23-24 Feb 2004, 23–24.

[13] Rudnitskiy, V., Uzun, I., Steganalysis algorithm for images that have been lossy compressed, *Information Security*, Vol. 15, No. 2 (2013), 122–127.

[14] Xikai Xu, Jing Dong, Tieniu Tan, Universal spatial feature set for video steganalysis, *Image Processing (ICIP), 19th IEEE International Conference*, Sept. 30 - Oct. 3 2012, 245–248.

[15] Pankajakshan, V., Ho, A.T.S., Improving video steganalysis using temporal correlation, *3rd International Conference on Intelligent Information Hiding and Multimedia Signal Processing*, 26-28 Nov 2007, Kaohsiung, TAIWAN.

[16] Budhia, U., Kundur, D., Zourntos, T., Digital video Steganalysis exploiting Statistical Visibility in the Temporal domain, *IEEE Transactions on Information Forensics and Security*, Vol. 1, Iss. 4 (2006), 502–516.

[17] Tasdemir, K., Kurugollu, F., Sezer, S., Video steganalysis of LSB based motion vector steganography, *Visual Information Processing (EUVIP), 4th European Workshop*, 10-12 June 2013, 260–264.

[18] Songbin Li, Peng Liu, Qiongxing Dai, Xiuhua Ma, Haojiang Deng, Detection of Information Hiding by Modulating Intra Prediction Modes in H.264/AVC, *Proceedings of the 2nd International Conference on Computer Science and Electronics Engineering (ICCSEE)*, 2013, 590-593.

[19] NRCS Photo Gallery. Online: http://photogallery.nrcs.usda.gov.

[20] WallpapersCraft. Online: http://wallpaperscraft.ru/.

Annales Mathematicae et Informaticae
47 (2017) pp. 19–40
http://ami.uni-eszterhazy.hu

First derivatives of fuzzy surfaces

Jan Caha[a], Jiří Dvorský[b]

[a]Department of Regional Development and Public Administration,
Mendel University in Brno, Zemědělská 1, 613 00, Brno, Czech Republic
jan.caha@mendelu.cz

[b]Department of Computer Science,VŠB - Technical University of Ostrava,
17. listopadu 15, 708 33 Ostrava-Poruba, Czech Republic
jiri.dvorsky@vsb.cz

Submitted October 10, 2016 — Accepted July 25, 2017

Abstract

The presented research shows how the first derivatives (slope and aspect) can be calculated from a fuzzy surface by the means of fuzzy arithmetic within the geographic information system. The proposed method works with fuzzy numbers of arbitrary shape which helps with more precise specification of input values as well as more exact calculation of results. Three most important methods of partial derivatives calculation based on finite elements approximation of a surface are presented and discussed. The presented approach provides an alternative for uncertainty propagation that is commonly performed by the utilization of statistics and the Monte Carlo method in geographic applications. The example calculation shows the differences between the obtained results calculated with the utilization of fuzzy arithmetic and the Monte Carlo method.

Keywords: fuzzy surface, fuzzy arithmetic, surface derivatives, uncertainty propagation

MSC: 03B52, 26E50, 90C70

1. Introduction

The process of modelling surface from a finite set of samples is a common problem in geosciences. Surfaces are often treated as certain and error-free models [41] even though there is a wide set of reasons why they are not. Perhaps the biggest issue arises from incomplete knowledge about the surface under study [32]. A user cannot be sure that the sample of surface values contains values that are representative enough to construct a precise surface. There is also the issue of measurement precision of the individual sample point, some authors point out that every measurement is fuzzy, at least to some extent, because there are no absolutely precise measurements [26, 37]. Another uncertainty can be introduced to the surface by the selection of interpolation technique [32]. Not only there is a range of methods that can be used for interpolation (IDW, spline interpolators, kriging etc.) but some of these methods have parameters (e.q. tension of spline, parameters of variogram in kriging) containing epistemic uncertainty. The values of these parameters are selected by the user and their selection is partially arbitrary [27]. In fact, these parameters are better described as a set of possible values than a single value which may not be correct. Authors [25] argue that much, if not the most, of uncertainty of surfaces in geosciences is interval, fuzzy or possibilistic in its nature. Fisher [13] mentions that the fuzzy set theory should be used if the definition of a class or an individual object is vague. The individual object in the case of a surface, represented commonly in geographic information system (GIS) by a grid, is a cell and its value is definitely vague because it can be based on uncertain data, influenced by epistemic uncertainty in the interpolation method [27] and even the grid model itself is a simplification and idealization of a real surface, which introduces yet another source of uncertainty [11].

Based on these facts a model of surface that would account for its inherent uncertainty is needed [26]. Such model was firstly proposed in [8] and [4]. A fuzzy surface as described in [8] is a result of interpolation with imprecise data, while the model in [4] was based on precise data but imprecise variogram in kriging interpolation. These two studies were the first to introduce fuzzy numbers into spatial modelling and spatial prediction but the applications of fuzzy approaches for predictions and modelling were used in mathematics before [36]. Later, more techniques and approaches for the construction of fuzzy surfaces emerged, including bayesian fuzzy kriging [3], kriging with imprecise variograms was further improved [27], the inverse-distance weighting method [37], and spline interpolators [2, 26, 32].

The definition of a fuzzy surface is only slightly different from an ordinary surface. A fuzzy surface is described by a set of points with known x, y coordinates and a fuzzy number $\tilde{Z}$ that represents the possible values of z at this location. The fuzzy number that describes such set of possible values represents the vague, imprecise or ill-know value. However, this uncertainty of the value does not originate in variability [25]. Similar deduction was done in [30], who mentions that statistical models often require more information about uncertainty than a user actually has. In such situations it might be reasonable and useful to formalize uncertainty in an alternative way which could be amongst others by the usage of the fuzzy set theory. Since methods for the creation of a fuzzy surface are either based on interpolation with imprecise input data, imprecise parameters of the interpolation and rarely on other methods, the outcome naturally contains this uncertainty. Instead of storing only one value of elevation at any point of the surface, the fuzzy surface stores a range of possible values (Fig. 1) that the surface can have at the given point. Fuzzy surfaces are an alternative to probabilistic surfaces that are commonly used in geography to estimate the influence of uncertainty on surface analyses [20, 31, 41]. While both these approaches have a lot in common there are also fundamental differences regarding the way how the resulting uncertainty is calculated and also about semantics of the results [30].

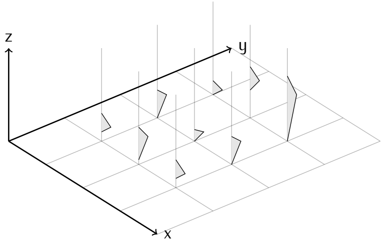


Figure 1: 3D visualization of a small fuzzy surface

The statistical approach to surface uncertainty tries to conceptualize uncertainty that occurs within the whole area of the surface through the specification of its spatial autocorrelation parameters [31]. The most commonly used method for statistical processing of uncertainty within GIS field is the Monte Carlo method [20]. Fuzzy surfaces are focused mainly on modelling uncertainty of a single cell in a grid [1, 25] while not accounting for the spatial autocorrelation explicitly. However, when the spatial autocorrelation of uncertainty is considered, it requires the user to describe it very precisely, by the specification of parameters of the autocorrelation function. This information is very rarely available to the user [30] which leads to providing of expert estimates instead of exact values [31].

As noted in [14], fuzzy mathematics has been rarely employed to actually analyse fuzzy surfaces, even though fuzzy surface exist ín geosciences for a long time. However, some examples of calculations of fuzzy slopes [6, 16, 37] and the visibility analysis on fuzzy surfaces [1] exist. All of them are, however, rather a cases of specific examples that describes only one method with specific type of fuzzy surface. In this article we summarize several methods for the calculation of slope and aspect from the fuzzy surfaces that are working with the arbitrary type of fuzzy numbers. The presented methods should provide an approach for the fuzzy slope and the fuzzy aspect calculation as general as possible. Further the presented method serves as an example of a spatial analysis on a fuzzy surface and the article explains all the necessary steps needed to devise fuzzy equivalent of any subsequent analysis.

2. Fuzzy numbers and fuzzy arithmetic

A fuzzy number is a special case of a fuzzy set [40], that represents a imprecise or uncertain value. Like a fuzzy set a fuzzy number $\tilde{F}$ is also defined by a membership function $\mu_{\tilde{F}}(x)$ that assigns the value from the interval $[0, 1]$ to every x from the universe X. The value of $\mu_{\tilde{F}}(x)$ is denoted as membership value and describes how much likely it is that the given value x belongs to the fuzzy number $\tilde{F}$. Authors [24] explain the semantics of a fuzzy number using the concept of interval of confidence (not to be confused with the confidence interval from statistics) and the level of presumption. The level of presumption is another designation for membership value. The interval of confidence describes the range that the value can take, while the level of presumption determines how likely this interval of confidence is. As the level of presumption increases the interval of confidence never increases [24]. This association corresponds to the mechanism of human thinking about uncertain variables. The less likely values (lower presumption levels) can be found in wider intervals while the values that are more likely (higher presumption levels) are situated in narrower intervals.

There are three main conditions that a fuzzy set has to fulfil in order to be a fuzzy number [19]. The universe on which the set is defined should be real numbers – $\mathbb{R}$. The height of the fuzzy set have to be equal to 1 [40] so that there is at least one value with a full membership to the set. The fuzzy set has to be convex. The convex fuzzy set fulfils:

$$\mu_{\tilde{F}}(\lambda x_1 + (1 - \lambda)x_2) \geq \min(\mu_{\tilde{F}}(x_1), \mu_{\tilde{F}}(x_2)) \tag{2.1}$$

for each x_1 and x_2 from $\mathbb{R}$ and λ from the interval $[0, 1]$ [40].

There are several types of fuzzy numbers, however the piecewise linear fuzzy numbers are of special importance here because they are simple for implementation and calculations [7, 19]. Visualization of different types of fuzzy numbers is shown in Fig. 2.

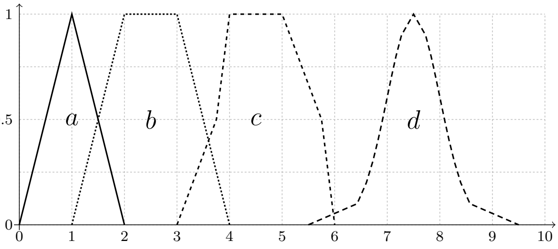


Figure 2: Visualization of fuzzy numbers: *a)* triangular, *b)* trapezoidal, *c)* piecewise linear, *d)* piecewise linear approximating gaussian

There are several important notions describing fuzzy numbers:

- kernel is a set of all x where $\mu_{\tilde{F}}(x) = 1$
- support is a set of all x where $\mu_{\tilde{F}}(x) > 0$
- $\alpha-$cut is a set of all x where $\mu_{\tilde{F}}(x) \geq \alpha$ for all $\alpha \in [0, 1]$, and is denoted as $\tilde{F}_\alpha$

The important property is that all α-cuts (including kernel and support) are crisp sets [9, 40]. Such α-cut can be written as $F_\alpha = [\underline{F_\alpha}, \overline{F_\alpha}]$, where F_α is a closed interval and $\underline{F_\alpha}, \overline{F_\alpha}$ are lower and upper endpoints of this interval. This is useful for further processing of fuzzy sets. The value of α is the level of presumption and the α-cut itself is the interval of confidence in the description provided in [24].

The decomposition theorem states that every fuzzy set can be described by a sequence of α-cuts [19]. The theorem further states that for every $x \in \mathbb{R}$ that belongs to the fuzzy set $\tilde{F}$ applies:

$$\mu_{\tilde{F}}(x) = \sup\{\alpha \in \langle 0, 1\rangle \mid x \in F_\alpha\}. \tag{2.2}$$

This theorem helps with representing the fuzzy set by a finite number of α-cuts. It also allows the transformation from the description by membership function into the description by α-cuts and vice versa.

For practical implementations it may be necessary to describe fuzzy sets by a list of its α-cuts then the number of such cuts has to be specified. After choosing a suitable m as the number of intervals to divide the interval $[0, 1]$ into, $m + 1$ values of α-cuts is given by following equation [19]:

$$\mu_j = \frac{j}{m}, \qquad j = 0, 1, \ldots, m \tag{2.3}$$

Such decomposition of a fuzzy set is very useful for practical applications, especially for the calculation with fuzzy numbers.

Classic crisp numbers can be seen as a special case of fuzzy numbers, where each α-cut is a degenerative interval ($[\underline{x}, \overline{x}]$, where $\underline{x} = \overline{x}$) [19]. This fact allows the combination of crisp values with fuzzy numbers in calculations.

2.1. Fuzzy arithmetic

Fuzzy arithmetic is an extension of a classic arithmetic on fuzzy numbers [24]. It allows complex mathematical operations with vague values. For functional combination fuzzy numbers $\tilde{X}, \tilde{Y}$ the membership function of resulting fuzzy number $\tilde{Z}$ is defined by the extension principle [40]:

$$\mu_{\tilde{Z}}(z) = \sup_{z=F(x,y)} \min(\mu_{\tilde{X}}(x), \mu_{\tilde{Y}}(y)). \tag{2.4}$$

In this equation the $F(x, y)$ is the functional combination of values x and y that belongs to fuzzy sets $\tilde{X}$ and $\tilde{Y}$ respectively. Eq. (2.4) is the most general form of extension principle which can be either simplified for functions of only one fuzzy number or extended to functions of more than two fuzzy numbers [19, 24]. The extension principle provides complete theoretical foundation for the calculation of all possible operations with fuzzy numbers. However, it is not particularly suitable for the practical implementation due to its high computational demand [19]. Three alternatives to this approach exist: the concept of *L*-*R* fuzzy numbers [9], decomposed fuzzy numbers [19] and the usage of interval arithmetic for calculation of α-cuts [19, 24, 29]. This last approach is identified as the most suitable for the practical use in [19].

2.2. Use of interval arithmetic for calculations with fuzzy numbers

The practical implementation of fuzzy arithmetic that uses interval arithmetic is based on the usage of the decomposition theorem (Eq. (2.2)). Fuzzy numbers decomposed on their α-cuts can be combined as ordered sets of intervals according to interval arithmetic defined in [29]. The implementation can be described as a following series of steps. Fuzzy numbers $\tilde{X}$ and $\tilde{Y}$ are divided into $m+1$ α-cuts (Eq. (2.3)). Then for each of those μ_j:

$$[\underline{Z_\alpha}, \overline{Z_\alpha}] = [\underline{X_\alpha}, \overline{X_\alpha}] \diamond [\underline{Y_\alpha}, \overline{Y_\alpha}] = [\min(G), \max(G)] \tag{2.5}$$

$$G = \{\underline{X_\alpha} \diamond \underline{Y_\alpha}, \underline{X_\alpha} \diamond \overline{Y_\alpha}, \overline{X_\alpha} \diamond \underline{Y_\alpha}, \overline{X_\alpha} \diamond \overline{Y_\alpha}\}. \tag{2.6}$$

If the operation $\diamond$ is division, we assume that $0 \notin [\underline{Y_\alpha}, \overline{Y_\alpha}]$, otherwise the operation is not valid.

2.3. Functions of fuzzy numbers

The issue of propagation of fuzzy numbers through functions is much more complex than the basic fuzzy arithmetic. Still some approaches from interval arithmetic are of use and can simplify the process significantly [19]. For functions that are monotonous with respect to all their variables the problem is simple, there is only a need to the propagate combinations of endpoints of α-cuts [29]:

$$f(\tilde{Y_\alpha}) = [\min(f(\underline{X_\alpha}), f(\overline{X_\alpha})), \max(f(\underline{X_\alpha}), f(\overline{X_\alpha}))] \tag{2.7}$$

Other functions like the integer exponentiation of a fuzzy number can be explained by a set of rules [24, 29]. If the exponent n is an odd number, then:

$$\tilde{X}_\alpha^n = \tilde{Z}_\alpha = \begin{cases} [\overline{X_\alpha}^n, \underline{X_\alpha}^n] \text{ if } \overline{X_\alpha} < 0 \\ [\underline{X_\alpha}^n, \overline{X_\alpha}^n] \text{ if } \underline{X_\alpha} > 0 \\ [\min(\underline{X_\alpha}^n, \overline{X_\alpha}^n), \max(\underline{X_\alpha}^n, \overline{X_\alpha}^n)] \text{ otherwise }. \end{cases} \tag{2.8}$$

if the exponent is an even number:

$$\tilde{X}_\alpha^n = \tilde{Z}_\alpha = \begin{cases} [\overline{X_\alpha}^n, \underline{X_\alpha}^n] \text{ if } \overline{X_\alpha} < 0 \\ [\underline{X_\alpha}^n, \overline{X_\alpha}^n] \text{ if } \underline{X_\alpha} > 0 \\ [0, \max(\underline{X_\alpha}^n, \overline{X_\alpha}^n)] \text{ otherwise }. \end{cases} \tag{2.9}$$

In other cases it is usually necessary to use directly the extension principle (Eq. (2.4)) or some technique that simplifies this use – i.e. the transformation method [19]. The extensive description of fuzzy arithmetic and related topics can be found in [19, 24, 29].

3. Comparison of fuzzy arithmetic and monte carlo

Utilization of the Monte Carlo method for the uncertainty propagation is rather common in surface analyses [12, 20, 21, 31], on the other hand fuzzy arithmetic is used rather rarely [14]. The main reason is because the Monte Carlo procedure is very simple to implement [19], as described in [31]. The method can be described by four main steps. The first step is to develop a model of surface and a model of uncertainty. The model of uncertainty is usually based on experts' knowledge and reasonable assumptions about the spatial autocorrelation of uncertainty [31]. The next step is to draw enough random realizations of uncertainty and add it to the surface to produce the uncertain surface. For each realization the calculation of analysis (e.g. slope, aspect, visibility calculation) is performed. The last step is to statistically evaluate the results, usually by calculating mean and standard deviation of the results. The process itself does not require any changes to the calculation of analysis, only its repetition for several times. The method is only demanding on computational power and time, because the number of iterations is generally in hundreds or even thousands.

The utilization of fuzzy arithmetic requires the adjustment to the analysis algorithms, because it has to be done according to the principles of fuzzy arithmetic. So far fuzzy arithmetic is not implemented in any software that would allow calculations of anything else but very simple examples. This is one of the reasons why the use of fuzzy arithmetic is at present limited to the scientific studies [19]. In case of the surface analysis, uncertainty of the surface is directly included in the fuzzy surface, so there is no need to generate random realizations of the surface and the result is calculated in one pass, without the need to iteratively calculate the outcomes.

Fuzzy arithmetic and the Monte Carlo method serve the same purpose – the uncertainty propagation. However, semantics and procedures vary significantly. The biggest difference is in both semantics of the result and its range. Since Monte Carlo is based on probability, it focuses on obtaining the probable outcomes and it cannot guarantee that the results will include all possible outcomes. Fuzzy arithmetic is focused on obtaining all the possible outcomes, so it guarantees that even the extreme combinations of input values will be included in the result. This is the main difference that arises from semantic differences between the Monte Carlo method (probabilistic model of uncertainty) and fuzzy arithmetic (model of uncertainty based on imprecision). The more developed analysis of semantics differences amongst the uncertainty theories is provided in [30].

The Monte Carlo method can be adjusted to produce results that are actually close to the results of fuzzy arithmetic, by use of optimization methods like for example the latin hypercube sampling [28]. But the usage of these optimization techniques in geosciences is not common. Indeed, fuzzy arithmetic provides results that yield larger uncertainties, however, if there is very little knowledge about uncertainty, it is the semantically correct approach [30].

4. Derivatives of surfaces

Derivatives are useful characteristics as they are providing a mathematical description of the surface appearance. The GIS tools for their calculation are based on the approximation of a real surface by a finite number of elements [39]. In the case of a grid structure these elements are cells [37]. This means that the derivative at a specific cell is calculated based on the values of neighbouring cells. There are two first derivatives of the surface: slope and aspect, several second derivatives that describe various types of curvature [39], a complete list of primary and secondary surface parameters and their significance is provided in [38]. All of those are commonly used in geographical and environmental analyses, for example in the fields like hydrology, geomorphology, geology, oceanography, ecology and others [34].

According to [39], two conditions have to be met to allow the calculation of the derivatives of the surface. The cells of the grid have to be aligned to the geographical axes and the distance between the centres of the cell should be the same for the whole grid. If both these conditions are met, the calculation is rather straightforward. Otherwise it is necessary to resample the grid according to those conditions. Other solution would be the modification of the equations which is performed rather rarely due to the complexity of this process [39].

z_7	z_8	z_1
z_6	z_9	z_2
z_5	z_4	z_3

Figure 3: Node numbering convention in the neighbourhood of a central cell z_9 (edited from: [39])

4.1. Methods of derivatives calculation

The basis of the derivatives determination is to calculate partial derivatives of the surface in two directions: North-South (denoted as $\frac{\partial z}{\partial y}$ with respect to the alignment with this axis) and East-West (denoted as $\frac{\partial z}{\partial x}$). There are several methods for calculation of those gradients, their comparison was performed in [23], [42] and also in [34]. The conclusion was that the 4-Cell method [15] provides the most precise results, closely followed by the Horn's method [22]. The third best algorithm was a modified version of Horn's method and as the forth the method of Sharpnack and Akin [33] was evaluated [23], these conclusions are approximately in agreement with the conclusions made in [34]. Study [42] was focused mainly on other elements of calculation. The algorithms were tested with respect to data quality and the resolution of the grid but findings from all these papers [23, 34, 42] suggest that

4-Cell, Horn's, Sharpnack's and Akin's methods are all good estimators of the first derivatives of the surface. Based on these results, these three algorithms for gradient calculation are considered in the article, they happen to be the most commonly implemented in GIS.

In all upcoming formulas the cells are labelled according to Fig. 3. The variable d denotes the size of the grid cell. The arrangement and numbering of the cells vary through the literature and the formulas for calculation of the derivations vary accordingly [39].

The 4-Cell method calculates the values of gradients only from the cells that have direct neighbourhood with the central cell. The method was firstly described in [15]. The equations for this calculation are:

$$\frac{\partial z}{\partial x} = \frac{z_2 - z_6}{2d}, \tag{4.1a}$$

$$\frac{\partial z}{\partial y} = \frac{z_8 - z_4}{2d}. \tag{4.1b}$$

Horn's method considers even the cells in the neighbourhood that have only one point common with the central cell. Cells that have common edge have been assigned higher weight in the calculation. The method was presented in [22] and the equations are:

$$\frac{\partial z}{\partial x} = \frac{(z_1 + 2z_2 + z_3) - (z_7 + 2z_6 + z_5)}{8d}, \tag{4.2a}$$

$$\frac{\partial z}{\partial y} = \frac{(z_7 + 2z_8 + z_1) - (z_5 + 2z_4 + z_3)}{8d}. \tag{4.2b}$$

Sharpnack and Akin's Method is very similar to the Horn's method with the change that all cells have the same weight. This method was proposed in [33] and the equations have the following form:

$$\frac{\partial z}{\partial x} = \frac{(z_1 + z_2 + z_3) - (z_7 + z_6 + z_5)}{6d}, \tag{4.3a}$$

$$\frac{\partial z}{\partial y} = \frac{(z_7 + z_8 + z_1) - (z_5 + z_4 + z_3)}{6d}. \tag{4.3b}$$

4.2. Calculation of slope and aspect

The three methods that were mentioned in the previous section provide three ways for the gradient calculation. These gradients are further used to calculate the slope S and the aspect A. For the slope calculation as a proportional rise the following equation is used:

$$S = \sqrt{\left(\frac{\partial z}{\partial x}\right)^2 + \left(\frac{\partial z}{\partial y}\right)^2}. \tag{4.4}$$

If the result should be in percent, a slightly different variation of the equation is needed:

$$S = 100\sqrt{\left(\frac{\partial z}{\partial x}\right)^2 + \left(\frac{\partial z}{\partial y}\right)^2}. \tag{4.5}$$

If the result should be provided in degrees, a slight modification is necessary. This slope is labelled as geographical:

$$S_g\ (^\circ) = \frac{180}{\pi} \arctan\left(\sqrt{\left(\frac{\partial z}{\partial x}\right)^2 + \left(\frac{\partial z}{\partial y}\right)^2}\right). \tag{4.6}$$

The calculation of aspect is a bit more complicated and requires the use of arctan2 function:

$$A = \text{arctan2}(-\frac{\partial z}{\partial y}, -\frac{\partial z}{\partial x}). \tag{4.7}$$

The mathematical aspect A is different from the geographical aspect A_g, A has a range of $[-\pi, \pi]$ radians, the value of 0 for the East and the values increase in a counter-clockwise direction. A_g has a range of $[0, 2\pi]$ in radians or $[0^\circ, 360^\circ]$ in degrees, the value of 0 for the North and the values increase in a clockwise direction [39]. So there is a need to adjust the values by this formula:

$$A_g\ (^\circ) = \begin{cases} 450^\circ - \frac{180}{\pi}A & \text{if } \frac{180}{\pi}A > 90^\circ \\ 90^\circ - \frac{180}{\pi}A & \text{otherwise.} \end{cases} \tag{4.8}$$

Based on those equations the calculation of approximation of slope and aspect can be calculated from the surface represented by the grid.

5. First derivatives of fuzzy surfaces

In any analysis calculated on a fuzzy surface uncertainty of the surface is propagated through the analysis into the result. Such result then shows uncertainty connected to the input data represented as fuzzy numbers. So far there are three examples of the calculation of a fuzzy slope in the literature provided in [16], [37] and [6]. Unfortunately, in first two cases a fuzzy slope is not the main focus of the research so it is not discussed in detail. [16] use a fuzzy slope to identify the areas having a slope potentially higher than 25 %, but the calculation serves as one of the several examples in the article, so it is discussed very briefly. [37] provided methods for the calculation of partial derivatives using the finite elements method, but the presented method is focused on a fuzzy surface constructed using purely triangular fuzzy numbers. The equations are adjusted to work on such surface, but it does not handle the calculation of a fuzzy slope in general, because triangular fuzzy numbers are only one type of a theoretically infinite set of fuzzy number types. These case specific adjustments of equations are common for presenting methods that utilize fuzzy arithmetic [19]. [6] described the calculation of only fuzzy slope using Horn's algortihm. There has been no attempt (of which the authors are aware) to calculate the aspect of a fuzzy surface.

The basis for the determination of both slope and aspect is the calculation of gradients $\frac{\partial z}{\partial x}$ and $\frac{\partial z}{\partial y}$. Considering that all inputs are uncertain and represented by fuzzy numbers, the results will also contain uncertainty and they will also be represented by fuzzy numbers. The calculation of gradients themselves is based on basic arithmetic operators that have fuzzy equivalents according to Eqs. (2.2, 2.3) and ((2.5, 2.6). This applies for all three methods of the gradient calculation (Eqs. (4.2, 4.1, 4.3)).

5.1. Slope

Calculating slope of a fuzzy surface according to Eqs. (4.5, 4.6) does not need any special approaches. Square of a fuzzy number can be calculated according to Eq. (2.8) and square root is a monotonous function and can be calculated according to Eqs. (2.2) and (2.7). If the slope is to be provided in degrees then the Eq. (4.6) is used. As mentioned previously, there is no problem with using crisp numbers with fuzzy numbers while calculating. Thus obtaining the value of the slope as a fuzzy number is a relatively simple matter.

5.2. Aspect

The aspect calculation is more complicated then the slope calculation. The Eq. (4.7) contains the function arctan2 [18] that has to be calculated for two fuzzy arguments. This in not a trivial operation and the function has to be modified to allow the calculation. The common definition of arctan2 is:

$$\operatorname{arctan2}(y,x) = \begin{cases} \arctan\frac{y}{x} & \text{if } x > 0 \\ \arctan\frac{y}{x} + \pi & \text{if } y \geq 0 \text{ and } x < 0 \\ \arctan\frac{y}{x} - \pi & \text{if } y < 0 \text{ and } x < 0 \\ +\frac{\pi}{2} & \text{if } y > 0 \text{ and } x = 0 \\ -\frac{\pi}{2} & \text{if } y < 0 \text{ and } x = 0 \\ \text{undefined} & \text{if } y = 0 \text{ and } x = 0 \end{cases} \tag{5.1}$$

From the viewpoint of the calculation with fuzzy numbers there are two main problems. The function is not defined for $y = 0$ and $x = 0$ and it is discontinuous for $x > 0$ and $y = 0$ (Fig. 5). This complicates the calculation if $\tilde{Y}$ and $\tilde{X}$ contain 0. Figure 4 shows four examples of areas bounded by supports of fuzzy numbers, labelled FN_1, FN_2, FN_3 and FN_4. As visible from the examples, FN_1 and FN_2 delimit angle intervals with the values of approximately $[215°, 250°]$ and $[30°, 100°]$.[1] Example FN_3 shows a situation in which both intervals contain 0

[1] The presented examples are for the sake of understanding shown in degrees with 0° pointing

and as a result FN_3 has a range of $[0°, 360°]$, which practically means that the aspect cannot be specified. The most interesting situation is FN_4, the smallest value contained in the rectangle is 0° and the highest value is 360° but actually the values between 45° and 290° do not belong to the set. In such situation it is impossible to construct a valid fuzzy number with respect to Eq.(2.1).

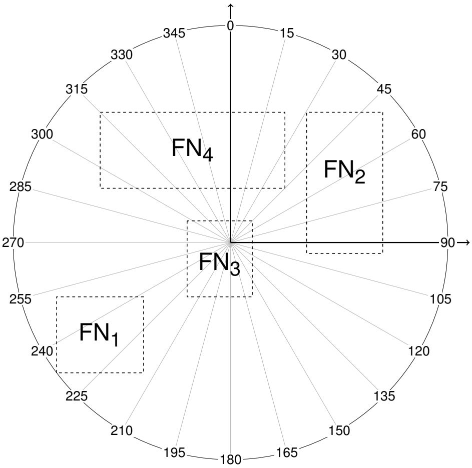


Figure 4: Four examples of bounding boxes on orientation

In order to allow the calculation with fuzzy numbers there is a need for a modified version of the function. As noted in [18] there exists a zero direction problem in directional statistics, the problem that is encountered when arctan2 is calculated for the fuzzy arguments is quite similar. It can be avoided but the obtained results will require a little bit more work in order to be interpreted correctly.

$$\operatorname{arctan2m}(y,x) = \begin{cases} \arctan\frac{y}{x} - 2\pi & \text{if } y \leq 0 \text{ and } x > 0 \\ \arctan\frac{y}{x} & \text{if } y < 0 \text{ and } x > 0 \\ \arctan\frac{y}{x} - \pi & \text{if } y \geq 0 \text{ and } x < 0 \\ \arctan\frac{y}{x} - \pi & \text{if } y < 0 \text{ and } x < 0 \\ +\frac{\pi}{2} - \pi & \text{if } y > 0 \text{ and } x = 0 \\ -\frac{\pi}{2} - \pi & \text{if } y < 0 \text{ and } x = 0 \\ \text{undefined} & \text{if } y = 0 \text{ and } x = 0 \end{cases} \tag{5.2}$$

If the calculation of $\arctan2(\widetilde{Y}, \widetilde{X})$ is to be performed, firstly it needs to be determined if the problem with discontinuity of the function will occur. This will happen if for any α there exist α-cuts such that $0 \in \widetilde{Y}_\alpha$ and $0 > \overline{\widetilde{X}_\alpha}$. If this condition is true, the modified version of arctan2 needs to be used (Eq. (5.2)). The rotated

upward and clockwise rise, even though those angles should be in radians with 0 pointing to the left and counter-clockwise rise of values.

variant of the function has a modified range $[-1.5\pi, 0.5\pi]$ instead of the original range $[-\pi, \pi]$ and it is discontinuous if $x = 0$ and $y < 0$ (see Fig. 5). The problem with undefined value of both functions is solved by setting the result interval to a full range of values if $x = 0$ and also $y = 0$. In either case both functions arctan2 and arctan2m are continuous with respect to both variables and as such they can be propagated by the use of simple approach according to Eq. (2.7).

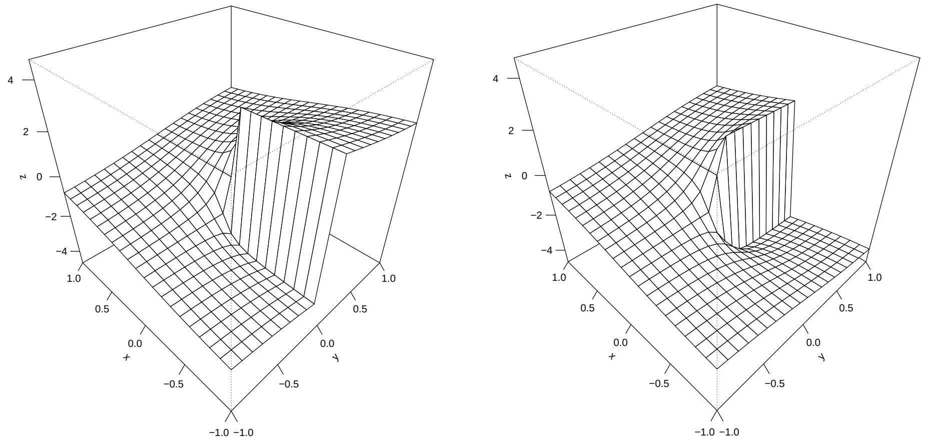


Figure 5: Left figure – values of arctan2(y, x). Right figure – values of arctan2m(y, x). Values of Z are in radians

When recalculating from the mathematical orientation onto the compass orientation (Eq.(4.8)), the value used for the comparison is the maximal value of the kernel – $\overline{A_1}$. After calculating A_g according to Eq. (4.8), the resulting values of A_g then do not fit in the original range of aspect values $[0°, 360°]$, which is the result of the propagation of fuzzy numbers through the calculation. Actually, the resulting values are from the range of $[-90°, 630°]$, meaning that the negative values v have the aspect of $360° + v$ and the positive values v higher than 360° are equal to $v - 360°$. The fuzzy orientation is more complicated for interpretation, but it is necessary to calculate them as such values to allow the correct propagation of the fuzzy numbers through the calculation. For the visualization and interpretation it is necessary to ensure that all those values will be interpreted correctly.

6. Numerical example

In this section an example of the calculation of aspect and slope using Horn's method (Eq. (4.2)) will be shown. The method is chosen because it is the one that is most commonly implemented in GIS. For the sake of readability the calculation will only be done for α-cuts 0 and 1. Each alpha cut of the fuzzy number $\tilde{F}$ will be written as $(\alpha : \underline{\tilde{X}_\alpha}; \overline{\tilde{X}_\alpha})$. The distance between cells has a value of $d = 10$ meters. The input fuzzy numbers or neighbouring cells are triangular and have the following definition:

$$\begin{aligned}
z_1 &= (0.0 : 382.81; 384.01)(1.0 : 383.41; 383.41)\\
z_2 &= (0.0 : 384.34; 385.5)(1.0 : 384.92; 384.92)\\
z_3 &= (0.0 : 385.83; 386.93)(1.0 : 386.38; 386.38)\\
z_4 &= (0.0 : 385.63; 386.63)(1.0 : 386.13; 386.13)\\
z_5 &= (0.0 : 385.46; 386.22)(1.0 : 385.84; 385.84)\\
z_6 &= (0.0 : 384.13; 384.87)(1.0 : 384.5; 384.5)\\
z_7 &= (0.0 : 382.63; 383.53)(1.0 : 383.08; 383.08)\\
z_8 &= (0.0 : 382.74; 383.78)(1.0 : 383.26; 383.26)
\end{aligned}$$

The surface used in this example is visualized in Fig. 6.

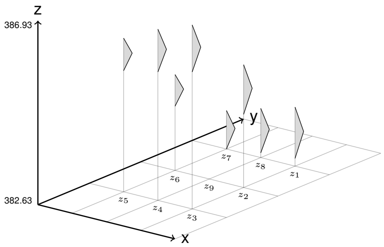


Figure 6: Visualization of a small fuzzy surface used in the example

The first step is to calculate the values of $\frac{\partial z}{\partial x}$ and $\frac{\partial z}{\partial y}$, to do that we firstly extract the necessary α-cuts from the fuzzy numbers according to Eq. (2.3) and then calculate the values according to Eq. (4.2) for each α-cut, applying Eqs. (2.5, 2.6) for each operation. The resulting fuzzy numbers have the values of:

$$\begin{aligned}
\frac{\partial z}{\partial x} &= (0.0 : -0.027; 0.07)(1.0 : 0.021; 0.021)\\
\frac{\partial z}{\partial y} &= (0.0 : -0.19; -0.09)(1.0 : -0.14; -0.14).
\end{aligned}$$

With the knowledge of gradients the calculation of slope is a simple matter (Eq. (4.5)). Eq. (2.8) will be used to calculate the power of fuzzy numbers, the addition of fuzzy numbers is done according to Eqs. (2.5, 2.6) and square root can be calculated as a monotonous function (Eq. (2.7)).

$$\begin{aligned}
\frac{\partial z}{\partial x}^2 &= (0.0 : 0.0; 0.005)(1.0 : 0.00; 0.00)\\
\frac{\partial z}{\partial y}^2 &= (0.0 : 0.009; 0.037)(1.0 : 0.02; 0.02)
\end{aligned}$$

$$S = (0.0 : 0.093; 0.206)(1.0 : 0.145; 0.145)$$

The value of slope S can be further transformed to percent according to Eq. (4.5):

$$S = (0.0 : 9.3; 20.6)(1.0 : 14.5; 14.5).$$

To calculate the aspect the Eq. (4.7) will be used.

$$A = (0.0 : 73.74; 126.92)(1.0 : 98.48; 98.48)$$

The mathematical aspect needs to be turned into the geographical aspect according to Eq. (4.8). For the comparison the value of $\overline{A_1}$ is used and the geographical aspect is obtained.

$$A_g = (0.0 : -36.92; 16.26)(1.0 : -8.48; -8.48)$$

As can be seen this is the case when the range of the resulting fuzzy easpect is outside of the classic aspect range $[0°, 360°]$. That means that this fuzzy aspect needs to be used interpreted as described in section 5. The resulting geographical aspect spans around north, with the modal value being slightly inclined towards west. For the further comparison with Monte Carlo method the range of the result can be described by two intervals $[0, 16.26], [323.08, 360]$ (°) that have the same interpretation as the support of fuzzy aspect.

Through the whole presented example for all variables the kernel value of each fuzzy number is the same as it would be in case of the calculation with crisp values. This fact shows that the propagation was done correctly because if triangular fuzzy numbers, where the kernel value corresponds to what originally was a crisp number, are used, then the kernel value of the result should be equal to the result of crisp calculation [19].

As a comparison the same calculations were performed with the use of the Monte Carlo simulation, using 100, 500, 1 000, 10 000 and 1 000 000 iterations. Triangular probability distributions were used as they are specified by three values [10], which makes them very similar to the triangular fuzzy numbers. The results of the simulations are summarized in Tab. 1. It is obvious that as the number of simulations raises, the ranges get closer to the range identified by fuzzy arithmetic. But it is very unlikely, even with very high number of simulations, for Monte Carlo to identify a complete range of possible outcomes.

The results that Monte Carlo failed to identify have very small probability of occurrence, but that are feasible solutions to the problem, this result is in agreement with the examples provided by [19] and [25]. These solutions can be perceived as the best/worst possible solutions and they can possibly be important for decision making. The complete range of outcomes should be $[9.3, 20.6]$ (%) for the slope and $[0, 16.26], [323.08, 360]$ (°) for the aspect.

The Monte Carlo method did not reach these widths of intervals but it is visible from the Tab. 1 that as the number of simulations increases, the estimates are actually converging towards the results provided by fuzzy arithmetic. However,

Table 1: Monte Carlo experiments

Simulations	Slope range (%)	Aspect range (°)
100	$[12.23, 16.54]$	$[344.24, 359.83]$
500	$[12.08, 17.16]$	$[0.01, 1.28], [339.93, 359.77]$
1 000	$[11.43, 17.33]$	$[0.06, 2.24], [341.30, 359.92]$
10 000	$[11.21, 17.84]$	$[0.06, 2.81], [339.54, 359.54]$
1 000 000	$[10.61, 18.27]$	$[0.00, 5.93], [336.03, 360.00]$

the number of simulations to obtain the true range is likely to be very high since the extension of the intervals is not significant even for a significant increase in the number of simulations. E.g. the change between the fourth and the fifth row of Tab. 1, even though the number of simulations is increased by a factor of 100 the obtained results change relatively insignificantly.

The calculation of slope and aspect for a whole grid instead of just one cell requires simply repeating this step for each cell of a grid. The source code for calculations of fuzzy derivatives with higher number of α-cuts and the other two methods for derivatives calculation are referenced in the Appendix 8.

7. Case study: Analysis of artificial fuzzy surface

For the purpose of practical demonstration analysis of fuzzy surface the artificial dataset is used. The dataset itself as well as code for its creation is described in Appendix 8. The case study demonstrates practical usability of fuzzy surface analyses in geographical applications.

Points from which the fuzzy surface is created are generated as Gaussian random field with gaussian correlation function with sill 200, range 400, nugget 0 and mean value equal to 150. To make the data less dependant on the specific function a random value drawn from normal distribution with 0 mean and standard deviation equal to 4 is added to each point. The dataset consists of 400 measurements randomly distributed in the area of size 4000×4000 meters. The z value (elevation) is interpolated into a grid of 401×401 cells, which makes cell size equal to 10. This datasets simulates real data measured on a surface.

The dataset is interpolated using fuzzy kriging with uncertain variogram proposed originally in [4] and later further developed in [27]. The kriging parameters sill, range and nugget are specified as triangular fuzzy numbers. The specific values are summarized in Tab. 2. The process of calculation of fuzzy kriging as well as source code for fuzzy interpolation are presented in [5].

Based on the previously mentioned fuzzy surface the fuzzy slope and fuzzy aspect can be calculated using procedures shown in section 5. For further use the values of minimum, modal a maximum are the most important as they describe the limits and the most likely value. The visualizations of fuzzy slope and fuzzy

Table 2: The values of semivariogram parameters

Parameter	Minimal value	Modal value	Maximal value
sill	130	138	145
range	390	395	400
nugget	13	15	17

aspect calculated with use of Horn's derivatives equations (Eq. (4.2)) are on Figs. 7 and 8.

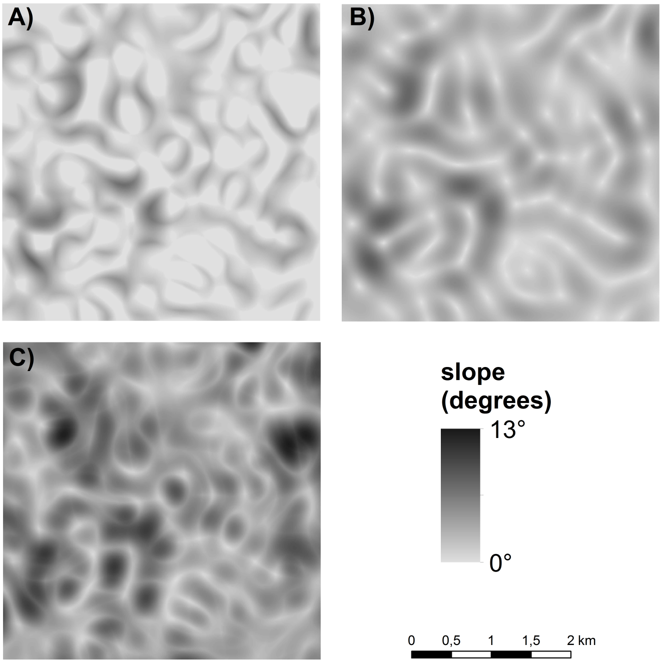


Figure 7: Visualization of minimal (A), modal (B) and maximal (C) slope calculated from the fuzzy surface. The slope unit are degrees

The presented approach is useful in analysing uncertain surfaces, where it would be illogical to present precise outcomes. For example the if the geostatistical estimations based on imprecise information, as presented in [35], should be analysed then using fuzzy arithmetic is the only proper way to do so.

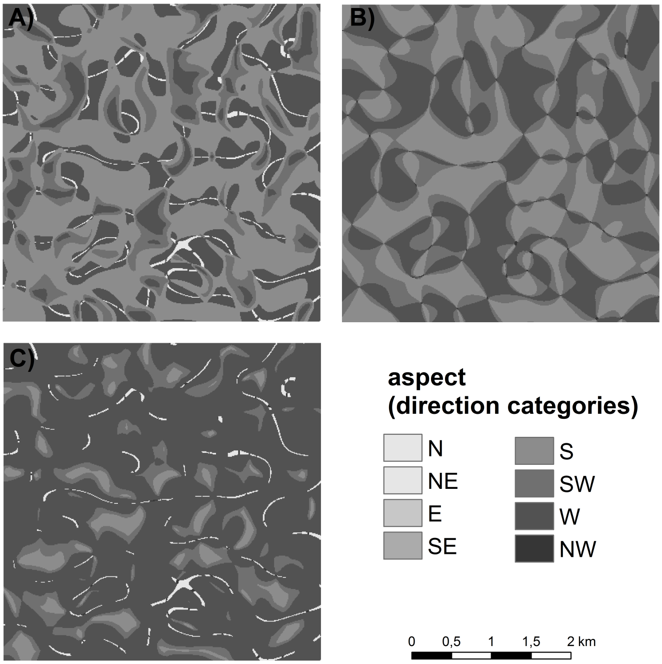


Figure 8: Visualization of minimal (A), modal (B) and maximal (C) aspect calculated from fuzzy surface. The aspect is visualized in directional categories

8. Conclusion

Hanss [19] noted that fuzzy arithmetic has received little attention and that the applications barely exceeded the level of elementary academic examples. The same statement regarding the analyses of fuzzy surfaces is done in [14]. The main reason for this lack of practical utilization is that there is basically no implementation of fuzzy arithmetic in even the mathematical software let alone within GIS. Exceptions are relatively new tools for R project FuzzyNumbers [17] and also FuzzyKrig toolbox [35] for Matlab®. The former allows calculations with fuzzy numbers while the latter is a tool for spatial interpolation with uncertain data and/or uncertain parameters. The secondary reason could be that some operations are not straightforward, like the presented calculation of aspect of a fuzzy surface. The process is more complex when compared to the commonly used Monte Carlo method. But still, such analyses are possible and necessary for further progress in the topic of analyses of fuzzy surfaces.

The presented research is in agreement with the previously performed studies that presented the processes of slope calculation [6, 16, 37], above that the procedure for calculating the aspect of a fuzzy surface is presented as well. The presented algorithms work for fuzzy numbers of arbitrary shape and the precision of the calculation can be adjusted by selecting different amounts of α-cuts. Three types of surface gradients that are the most commonly implemented in software were shown to be compatible with fuzzy arithmetic and can be used to calculate the first derivatives of fuzzy surfaces.

The advantage of the presented approach, when compared to another commonly used technique of the uncertainty propagation – the Monte Carlo method, is that the derivatives of a surface are calculated in one pass and all uncertainty of the fuzzy surface is included in the result. Uncertainty is naturally incorporated in the process by the use of fuzzy numbers and fuzzy arithmetic, so there is no need for iterations in the calculation. Unlike Monte Carlo, fuzzy arithmetic can guarantee inclusion of all possible outcomes (including limit cases) in the result. The Monte Carlo method, on the other hand, focuses only on the most probable results [19]. This is an important difference amongst these two methods that might be important for decision making process based on the result of calculation with uncertainty.

According to the extension principle [40], every operation can be extended to its fuzzy equivalent. That means that every analysis that can be performed on a surface in GIS can be also performed on a fuzzy surface and the result will contain and bound uncertainty of the surface. Such approach to modelling should provide an alternative to the currently used Monte Carlo method and provide GIS users another possibility how to conceptualize and propagate uncertainty through geographic analyses. The need for new approaches and methods is quite significant as the issue of uncertainty propagation within GIS is still relatively undeveloped [21].

Further research should focus on a subsequent surface analysis, which can include but are not limited to second derivatives, optimal path selection, visibility analysis, catchment delimitation and others.

Acknowledgements Authors are thankful for the reviewers' comments that helped in improving this article.

Appendix: Code

The examples in section 6 are performed with the use of FuzzyNumbers package [17]. The source code for the example and its variants, calculated using different method of derivatives determination, can be found in `https://github.com/JanCaha/AMeI-paper`. The data used in section 7 are available as well along with the results calculated by other two methods for gradient calculation.

Procedure of creating the fuzzy surface that was used as input data in the case study can be found in `https://github.com/JanCaha/Hais2015-paper` and in [5].

References

[1] A. M. Anile, P. Furno, G. Gallo, and A. Massolo. A fuzzy approach to visibility maps creation over digital terrains. *Fuzzy Sets and Systems*, 135(1):63–80, 2003.

[2] A. M. Anile and S. Spinella. Modeling Uncertain Sparse Data with Fuzzy B-splines. *Reliable Computing*, 10(5):335–355, 2004.

[3] H. Bandemer and A. Gebhardt. Bayesian fuzzy kriging. *Fuzzy Sets and Systems*, 112(3):405–418, 2000.

[4] A. Bardossy, I. Bogardi, and W. E. Kelly. Kriging with imprecise (fuzzy) variograms. I: Theory. *Mathematical Geology*, 22(1):63–79, 1990.

[5] J. Caha, L. Marek, and J. Dvorský. Predicting PM10 Concentrations Using Fuzzy Kriging. In E. Onieva, I. Santos, E. Osaba, H. Quintián, and E. Corchado, editors, *Hybrid Artificial Intelligent Systems SE - 31*, volume 9121 of *Lecture Notes in Computer Science*, pages 371–381. Springer International Publishing, 2015.

[6] J. Caha, P. Tuček, A. Vondráková, and L. Paclíková. Slope Analysis of Fuzzy Surfaces. *Transactions in GIS*, 16(5):649–661, 2012.

[7] L. Coroianu, M. Gagolewski, and P. Grzegorzewski. Nearest piecewise linear approximation of fuzzy numbers. *Fuzzy Sets and Systems*, 233:26–51, 2013.

[8] P. Diamond. Fuzzy Kriging. *Fuzzy Sets and Systems*, 33(3):315–332, 1989.

[9] D. Dubois and H. Prade. *Fuzzy sets and systems : theory and applications.* Number Nf. Academic Press, New York, 1980.

[10] M. Evans, N. Hastings, and B. Peacock. Triangular Distribution. In *Statistical Distributions*, pages 187–188. Wiley-Interscience, New York, 3rd ed. edition, 2000.

[11] P. Fisher. The pixel : A snare and a delusion. *International Journal of Remote Sensing*, 18(3):679–685, 1997.

[12] P. Fisher. Improved Modeling of Elevation Error with Geostatistics. *GeoInfomatica*, 2(3):215–233, 1998.

[13] P. F. Fisher. Models of uncertainty in spatial data. In P. Longley, M. F. Goodchild, D. J. Maguire, and D. W. Rhind, editors, *Geographical Information Systems: Principles, Techniques, Management, and Applications*, chapter 13, pages 191–205. Wiley, New York, NY, 2nd edition, 2005.

[14] P. F. Fisher and N. J. Tate. Causes and consequences of error in digital elevation models. *Progress in Physical Geography*, 30(4):467–489, 2006.

[15] M. Fleming and R. M. Hoffer. *Machine processing of landsat MSS data and DMA topographic data for forest cover type mapping, Laboratory for Applications of Remote Sensing.* Purdue University, 1979.

[16] C. C. Fonte and W. A. Lodwick. Modelling the Fuzzy Spatial Extent of Geographical Entities. In F. Petry, V. B. Robinson, and M. A. Cobb, editors, *Fuzzy modeling with spatial information for geographic problems*, pages 120–142. Springer, Berlin, 2005.

[17] M. Gagolewski and J. Caha. A Guide to the FuzzyNumbers Package for R. Technical report, 2015.

[18] G. L. Gaile and J. E. Burt. *Directional Statistics.* Geo Abstracts, University of East Anglia, 1980.

[19] M. Hanss. *Applied Fuzzy Arithmetic : An Introduction with Engineering Applications*. Springer-Verlag, Berlin ; New York, 2005.

[20] G. B. M. Heuvelink. *Error Propagation in Environmental Modelling with GIS*. Taylor & Francis Ltd, London, 1998.

[21] G. B. M. Heuvelink. Analysing Uncertainty Propagation in GIS: Why is it not that Simple? In G. M. Foody and P. M. Atkinson, editors, *Uncertainty in remote sensing and GIS*, pages 155–165. Wiley, Chichester, 2002.

[22] B. Horn. Hill shading and the reflectance map. *Proceedings of the IEEE*, 69(1):14–47, 1981.

[23] K. H. Jones. A comparison of algorithms used to compute hill slope as a property of the dem. *Computers & Geosciences*, 24:315–323, 1998.

[24] A. Kaufmann and M. M. Gupta. *Introduction to Fuzzy Arithmetic*. Van Nostrand Reinhold Company, New York, 1985.

[25] W. Lodwick, A. M. Anile, and S. Spinella. Introduction. In W. Lodwick, editor, *Fuzzy surfaces in GIS and geographical analysis : theory, analytical methods, algorithms, and applications*, pages 1–46. CRC Press, Boca Raton, 2008.

[26] W. A. Lodwick and J. Santos. Constructing consistent fuzzy surfaces from fuzzy data. *Fuzzy Sets and Systems*, 135(2):259–277, 2003.

[27] K. Loquin and D. Dubois. Kriging with Ill-Known Variogram and Data. In A. Deshpande and A. Hunter, editors, *Scalable Uncertainty Management SE - 5*, volume 6379 of *Lecture Notes in Computer Science*, pages 219–235. Springer Berlin / Heidelberg, 2010.

[28] M. McKay, R. Beckman, and W. Conover. A comparison of three methods for selecting values of input variables in the analysis of output from a computer code. *Technometrics*, 21(2):239–245, 1979.

[29] R. E. Moore, R. B. Kearfott, and M. J. Cloud. *Introduction to interval analysis*. Society for Industrial and Applied Mathematics, Philadelphia, PA, 2009.

[30] M. Oberguggenberger. The mathematics of uncertainty: models, methods and interpretations. In W. Fellin, H. Lessmann, M. Oberguggenberger, and R. Vieider, editors, *Analyzing Uncertainty in Civil Engineering*, page 252. Springer, Berlin, 2005.

[31] J. Oksanen and T. Sarjakoski. Error propagation analysis of DEM-based drainage basin delineation. *International Journal of Remote Sensing*, 26(14):3085–3102, 2005.

[32] J. Santos, W. A. Lodwick, and A. Neumaier. A New Approach to Incorporate Uncertainty in Terrain Modeling. In M. Egenhofer and D. Mark, editors, *Geographic Information Science*, volume 2478 of *Lecture Notes in Computer Science*, pages 291–299. Springer Berlin Heidelberg, 2002.

[33] D. A. Sharpnack and G. Akin. An algorithm for computing slope and aspect from elevations. *Photogrammetric Engineering*, 35:247–248, 1969.

[34] A. K. Skidmore. A comparison of techniques for calculating gradient and aspect from a gridded digital elevation model. *International journal of geographical information systems*, 3(4):323–334, 1989.

[35] S. Soltani-Mohammadi. FuzzyKrig: a comprehensive matlab toolbox for geostatistical estimation of imprecise information. *Earth Science Informatics*, 2015.

[36] H. Tanaka, S. Uejima, and K. Asai. Linear Regression Analysis with Fuzzy Model. *IEEE Transactions on Systems, Man and Cybernetics*, 12(6):903–907, 1982.

[37] O. Waelder. An application of the fuzzy theory in surface interpolation and surface deformation analysis. *Fuzzy Sets and Systems*, 158(14):1535–1545, 2007.

[38] J. P. Wilson. Digital terrain modeling. *Geomorphology*, 137(1):107–121, 2012.

[39] J. P. Wilson and J. C. Gallant. *Terrain analysis : principles and applications*. Wiley, New York, 2000.

[40] L. A. Zadeh. Fuzzy Sets. *Information and Control*, 8(3):338–353, 1965.

[41] J. Zhang and M. F. Goodchild. *Uncertainty in geographical information.* Taylor & Francis, London, 2002.

[42] Q. M. Zhou and X. J. Liu. Analysis of errors of derived slope and aspect related to DEM data properties. *Computers & Geosciences*, 30(4):369–378, 2004.

Annales Mathematicae et Informaticae
47 (2017) pp. 41–74
http://ami.uni-eszterhazy.hu

Twelve subsets of permutations enumerated as maximally clustered permutations

David Callan[a], Toufik Mansour[b], Mark Shattuck[c*]

[a]Department of Statistics, University of Wisconsin, Madison, WI 53706
callan@stat.wisc.edu

[b]Department of Mathematics, University of Haifa, 3498838 Haifa, Israel
tmansour@univ.haifa.ac.il

[c]Institute for Computational Science & Faculty of Mathematics and Statistics
Ton Duc Thang University, Ho Chi Minh City, Vietnam
mark.shattuck@tdt.edu.vn

Submitted February 9, 2017 — Accepted June 18, 2017

Abstract

The problem of avoiding a single pattern or a pair of patterns of length four by permutations has been well studied. Less is known about the avoidance of three 4-letter patterns. In this paper, we show that the number of members of S_n avoiding any one of twelve triples of 4-letter patterns is given by sequence A129775 in OEIS, which is known to count maximally clustered permutations. Numerical evidence confirms that there are no other (non-trivial) triples of 4-letter patterns giving rise to this sequence and hence one obtains the largest $(4,4,4)$-Wilf-equivalence class for permutations. We make use of a variety of methods in proving our result, including recurrences, the kernel method, direct counting, and bijections.

Keywords: pattern avoidance; Wilf-equivalence; kernel method; maximally clustered permutations

MSC: 05A15, 05A05

[*] Corresponding author.

1. Introduction

1.1. Background

The pattern avoidance question is an extensively studied problem in enumerative and algebraic combinatorics. It has its origins with Knuth [5] and Simion and Schmidt [8] who considered the problem on permutations and enumerated the number of members of S_n avoiding a particular element or subset, respectively, of S_3. Since then the problem has been addressed on several other discrete structures, such as compositions, k-ary words, and set partitions; see, e.g., the texts [3, 7] and references contained therein. Here, we provide further enumerative results concerning the classical avoidance problem on permutations.

Members of S_n avoiding a single 4-letter pattern have been well studied (see, e.g., [9, 10, 11]). There are 56 symmetry classes of pairs of 4-letter patterns, for all but 8 of which the avoiders have been enumerated [2]. Less is known about the 317 symmetry classes of triples of 4-letter patterns. In this paper, we show that precisely 12 of them have the counting sequence of maximally clustered permutations (sequence A129775 in OEIS), which has generating function

$$\frac{2(1-4x)}{2-9x+4x^2-x\sqrt{1-4x}}=1+\frac{x}{2-x-C(x)},$$

where $C(x)=\frac{1-\sqrt{1-4x}}{2x}$ is the generating function for the Catalan numbers. Based on numerical evidence, this corresponds to the largest $(4,4,4)$-Wilf-equivalence class for permutations.

A computer check of initial terms eliminates all but 12 candidate classes for this counting sequence. We next recall basic terminology, review some standard results, list a representative triple π_i, $i=1,2,\ldots,12$, for each class, and state the main result. Then, in Section 2, we treat each π_i in turn. Our methods include recurrences, the kernel method for solving them, direct counting, and bijections.

1.2. Notation, terminology and main result

Let $\pi=\pi_1\pi_2\cdots\pi_n\in S_n$ and $\tau\in S_k$ be two permutations. We say that π *contains* τ if there exists a subsequence $1\le i_1<i_2<\cdots<i_k\le n$ such that $\pi_{i_1}\pi_{i_2}\cdots\pi_{i_k}$ is *order-isomorphic* to τ; in this context τ is usually called a *pattern*. We say that π *avoids* τ, or is τ-*avoiding*, if such a subsequence fails to exist. The set of all τ-avoiding permutations in S_n is denoted $S_n(\tau)$. For an arbitrary finite collection of patterns T, we say that π *avoids* T if π avoids every $\tau\in T$; the corresponding subset of S_n is denoted $S_n(T)$. Two sets of patterns T and T' are said to be *Wilf-equivalent* if $|S_n(T)|=|S_n(T')|$ for all $n\ge 0$.

The maximally clustered permutations are those that avoid 3421, 4312 and 4321, and this triple is in the same symmetry class as π_3 in Theorem 1.1 below. (See [1], where a different proof is given in this particular case.) Here, symmetry refers to the action of the dihedral group of order 8 generated by the operations reverse, complement, and inverse on permutation patterns. Two pattern sets so related obviously have equinumerous avoiders, in short, are trivially Wilf-equivalent.

For a permutation p on a set of positive integers, the standardization of p, denoted $\mathrm{St}(p)$, is obtained by replacing the smallest entry of p by 1, the next smallest by 2, and so on. Thus π avoids τ iff no subsequence of π has standardization equal to τ. It is well known [8] that, for each 3-letter pattern τ, $|S_n(\tau)|$ is the Catalan number $C_n = \frac{1}{n+1}\binom{2n}{n}$. Throughout, we use $C(x) = \frac{1-\sqrt{1-4x}}{2x}$ to denote the generating function $\sum_{n\geq 0} C_n x^n$.

Theorem 1.1 (Main Theorem). *Define*

$$\begin{aligned}
\pi_1 &= \{1324, 2134, 2143\}, & \pi_2 &= \{1243, 1324, 2134\}, & \pi_3 &= \{1234, 1243, 2134\},\\
\pi_4 &= \{2314, 2341, 2413\}, & \pi_5 &= \{2314, 2413, 2431\}, & \pi_6 &= \{1423, 3142, 4132\},\\
\pi_7 &= \{1324, 1342, 3142\}, & \pi_8 &= \{1324, 1342, 3124\}, & \pi_9 &= \{1324, 1342, 2314\},\\
\pi_{10} &= \{1324, 1432, 2431\}, & \pi_{11} &= \{1423, 1432, 4132\}, & \pi_{12} &= \{1342, 1423, 4123\}.
\end{aligned}$$

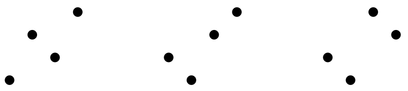
Then, for all $j = 1, 2, \ldots, 12$,

$$\sum_{n\geq 0} \#S_n(\pi_j)x^n = \frac{2(1-4x)}{2-9x+4x^2-x\sqrt{1-4x}}.$$

2. Proof of main theorem

2.1. Class 1

$\pi_1 = \{1324, 2134, 2143\}$, with graphical representation



Let $A_n = S_n(\pi_1)$. Define $a_n = \#A_n$ and $a_n(i_1, \ldots, i_s)$ to be the number of permutations $\sigma_1\sigma_2\cdots\sigma_n \in A_n$ such that $\sigma_1\sigma_2\cdots\sigma_s = i_1 i_2 \cdots i_s$. Then we have the following recurrence.

Lemma 2.1. *For all* $1 \leq i \leq n-2$,

$$a_n(i) = 2a_{n-1}(i) + a_{n-2}(i)\delta_{i\leq n-3} + \sum_{j=i+2}^{n-2} C_{n-j}a_{j-1}(i),$$

with $a_n(n-1) = a_n(n) = a_{n-1}$.

Proof. By the definitions, $a_n(n) = a_n(n-1) = a_{n-1}$. If $1 \leq i \leq n-2$, then

$$a_n(i) = a_n(i,i+1) + a_n(i,n) + a_n(i,n-1)\delta_{i\leq n-3} + \sum_{j=i+2}^{n-2} a_n(i,j)$$

$$
\begin{aligned}
&= 2a_{n-1}(i) + a_n(i, n-1, n)\delta_{i\le n-3} + \sum_{j=i+2}^{n-2} a_n(i,j)\\
&= 2a_{n-1}(i) + a_{n-2}(i)\delta_{i\le n-3} + \sum_{j=i+2}^{n-2} a_n(i,j).
\end{aligned}
$$

Note that any permutation $\pi = ij\pi' \in A_n$ with $i+2 \le j \le n-2$ can be decomposed as $\pi = ij\alpha\beta$, where each letter of α is greater than each letter of β and α avoids 213 and $i\beta \in A_{j-1}$. Thus, by the fact that the number of permutations of length d that avoid 213 is given by the d-th Catalan number (see [5]), we obtain that $a_n(i,j) = C_{n-j}a_{j-1}(i)$, which completes the proof. □

Define $A_n(v)$ to be the polynomial $\sum_{i=1}^n a_n(i)v^{i-1}$. Then Lemma 2.1 can be translated in terms of $A_n(v)$ as

$$
\begin{aligned}
&A_n(v) - A_{n-1}(1)(v^{n-2} + v^{n-1})\\
&\quad = 2A_{n-1}(v) + A_{n-2}(v) - 2A_{n-2}(1)v^{n-2} - A_{n-3}(1)v^{n-3}\\
&\quad + \sum_{j=3}^{n-2} C_{n-j}(A_{j-1}(v) - A_{j-2}(1)v^{j-2}).
\end{aligned}
$$

Note that $A_0(v) = A_1(v) = 1$ and $A_2(v) = 1 + v$. Define $A(x,v) = \sum_{n\ge0} A_n(v)x^n$. Multiplying the last recurrence by x^n, and summing over $n \ge 3$, yields

$$
\begin{aligned}
&A(x,v) - \frac{x}{v}(A(xv,1) - 1) - xA(xv,1) - 1\\
&\quad = x(2+x)(A(x,v) - 1) - x^2(2+x)A(xv,1)\\
&\quad + x(C(x) - 1 - x)(A(x,v) - 1 - x) - x^2(C(x) - 1 - x)(A(xv,1) - 1),
\end{aligned}
$$

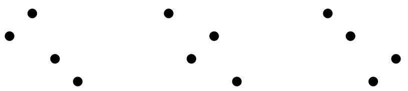
which, upon setting $v = 1$, gives the following result.

Theorem 2.2. *The generating function for the number of permutations of length n that avoid π_1 is given by*

$$\frac{2(1-4x)}{2 - 9x + 4x^2 - x\sqrt{1-4x}}.$$

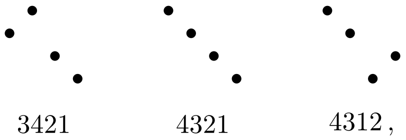
2.2. Class 2

We use the representative triple $\pi_2 := \{X, Y, Z\}$, as illustrated,



$X = 3421$ $Y = 4231$ $Z = 4312$,

compared with



the pattern set π_3 considered in Class 3 below. Note that they differ only in the middle of the middle pattern. Clearly, a permutations avoids π_2 if and only if each of its components does so and the same is true of π_3. So the following result shows that $|S_n(\pi_2)| = |S_n(\pi_3)|$.

Theorem 2.3. *The map "locate the maximal runs of consecutive fixed points and reverse each run" is a bijection from the indecomposable permutations in $S_n(\pi_3)$ to the indecomposable permutations in $S_n(\pi_2)$.*

Proof. As an example,

$$\begin{pmatrix} 1 & 2 & 3 & 4 & 5 & 6 & 7 & 8 & 9 & 10 & 11 \\ 3 & 11 & 1 & 4 & 5 & 6 & 2 & 8 & 9 & 7 & 10 \end{pmatrix}$$
$$\mapsto \begin{pmatrix} 1 & 2 & 3 & 4 & 5 & 6 & 7 & 8 & 9 & 10 & 11 \\ 3 & 11 & 1 & 6 & 5 & 4 & 2 & 9 & 8 & 7 & 10 \end{pmatrix}.$$

From the characterization of indecomposable π_3-avoiders in Class 3 below, it is clear that the map is one-to-one and into; the only issue is whether it is onto. To show that it is, we investigate the structure of π_2-avoiders.

Lemma 2.4. *Suppose $c > b_1 > b_2 > \cdots > b_r > a$, $r \geq 1$, is a maximal decreasing subsequence of length ≥ 3 in a π_2-avoider p. Then, in the matrix diagram of p, the entries $b_1, b_2, \ldots, b_r$ form the reverse (NW to SE) diagonal of a square bisected by the main diagonal and c is the only entry lying NW of this square and a is the only entry lying SE of it.*

Proof. Consider the rectangles in the matrix determined by the subsequence as shown in Figure 1 for $r \geq 2$ (collapsing some regions covers the case $r = 1$). The gray rectangles are all empty for the indicated reason where M refers to the maximal condition in the hypothesis, and X, Z refer to offending patterns. The entries in the rectangle B are decreasing (else a Y offender is present). Furthermore, since the rest of the row and column containing B is empty, the entries in $b_1 B\, b_r$ must be consecutive and B must be a square of side length $r-2$. Also, the entries in rectangle A consist of $[b_r - 1] \setminus \{a\}$. This means that A is a square of side length $b_r - 1$, and so B is bisected by the main diagonal. Thus, all parts of the lemma have been established. □

It follows from Lemma 2.4 that the mapping is onto and, hence, a bijection. □

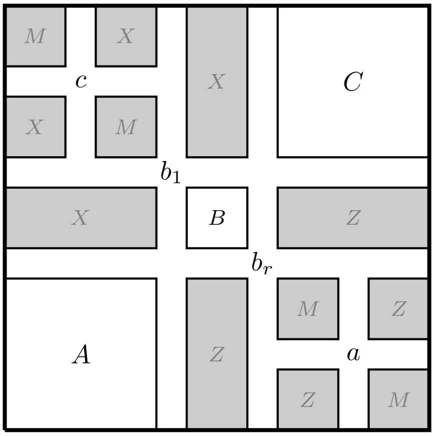


Figure 1: A decomposition

2.3. Class 3

We use the representative triple $\pi_3 := \{3421, 4321, 4312\}$.

Losonczy [6] introduced the notion of maximally clustered elements in a Coxeter group and showed that for Type A (symmetric) groups, they are characterized precisely by avoiding the 3 patterns in π_3. Soon after, Denoncourt and Jones [1] considered heaps in Coxeter groups and found an expression for the generating function for permutations that avoid both π_3 and a heap H as a rational function of the generating function for permutations that avoid 321 and H. The enumeration of π_3-avoiders follows by setting $H = \emptyset$.

For our bijective enumeration, we note that a permutation p avoids π_3 if and only if each of its components does so. So it suffices to determine u_n, the number of indecomposable π_3-avoiders of length n, for then the Invert transform of $(u_n)_{n\geq 1}$ gives the unrestricted counting sequence. Clearly, $u_1 = 1$ and we will show that $u_n = \frac{1}{2}\binom{2(n-1)}{n-1}$ for $n \geq 2$.

The left to right maxima (LR maxima) of a permutation determine a (rotated) Dyck path P with the LR maxima at the NE corners (N = North, E = East), as in Figure 2. The returns to the diagonal split P into its *components*, and P is *indecomposable* if it has exactly one return (necessarily at its endpoint). Components of the permutation p correspond to components of the Dyck path P and so p is indecomposable iff P is.

We begin with an obvious connection between fixed points and 321 patterns.

Lemma 2.5. *For any permutation p and a fixed point b of p, either b is a component of p or b is the "2" of a 321 pattern in p.*

Now we look at the structure of indecomposable π_3-avoiders.

Lemma 2.6. *Let p be an indecomposable π_3-avoider.*

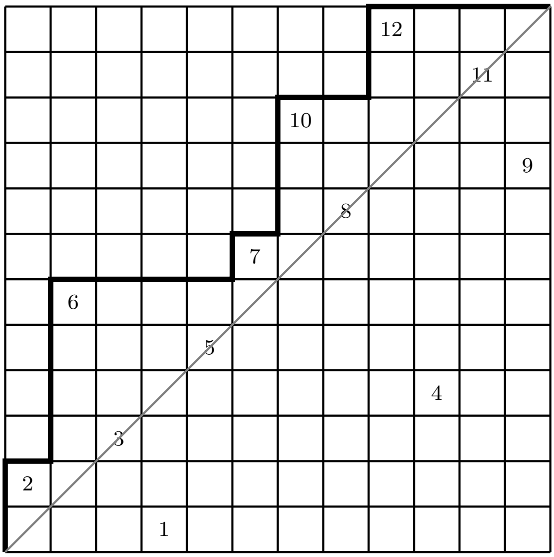


Figure 2: A permutation with LR maxima $2, 6, 7, 10, 12$ and its Dyck path. This permutation is indecomposable

(i) An entry b of p can be the "2" of at most one 321 *pattern.*

(ii) If cba is a 321 *pattern in p, then b is a fixed point of p.*

(iii) A fixed point b is the "2" of exactly one 321 *pattern in p.*

Proof. (i) If b was the "2" of more than one 321 pattern, a forbidden pattern would be present.

(ii) By (i), the entries preceding b are precisely $\{c\} \cup [b-1] \backslash \{a\}$ and so b is the b-th entry.

(iii) This follows from part (i) and Lemma 2.5. □

Corollary 2.7. *An indecomposable permutation is a π_3-avoider if and only if it either avoids* 321 *or the "2"s of its* 321 *patterns are all distinct and all fixed points.*

Proof. The "only if" part follows from Lemmas 2.5 and 2.6, and the presence of any one of the offending patterns would imply two 321 patterns with the same 2. □

Lemma 2.8. *An indecomposable π_3-avoider is determined by the locations in the matrix diagram of its LR maxima and its fixed points.*

Proof. All other entries must be increasing. Suppose not and that $b > a$ were two other entries, with b to the left of a. Then a LR maximum would precede b, so b would be the "2" of a 321 and hence a fixed point, which it is not. □

Arbitrary indecomposable Dyck paths are possible for an indecomposable π_3-avoider, but what about the fixed points? For b to be a fixed point, b cannot be either the value or position of a LR maximum and there must be exactly one LR maximum preceding b and $> b$. In terms of the Dyck path in a matrix diagram, b cannot be in the row or column of a NE corner, and the b-th E step (among the E steps) and its bounce N step must be the end steps of a subpath with just one peak ($= NE$ corner). Any b meeting these conditions can be a fixed point. More precisely, given an indecomposable Dyck path (determining the LR maxima and their positions) and a subset B of the b's meeting the above conditions, there is exactly one indecomposable π_3-avoider with this Dyck path and fixed point set B, namely, the permutation in which all other entries are increasing.

It is convenient to focus on the vertices of the Dyck path, and call a vertex *good* if it is the left endpoint of the E step directly above a possible fixed point b. Since the Dyck path is indecomposable, we may delete the first and last step to get a new (unrestricted) Dyck path of semilength $n-1$ with a new diagonal line joining its endpoints. In this formulation, a vertex is good if (i) it joins 2 E steps, (ii) its *bounce* vertex (down to the diagonal, left to the path) joins 2 N steps, and (iii) the subpath bounded by the vertex and its bounce contains only one peak. Some examples are shown in Figure 3.

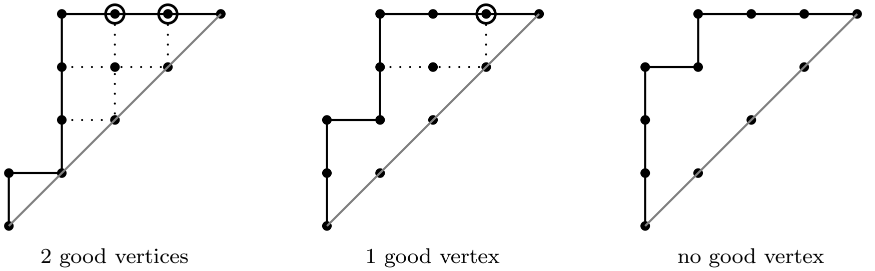


Figure 3: Good vertices

Thus we have shown that indecomposable π_3-avoiders of length n correspond to Dyck paths of semilength $n-1$ in which some (maybe all or none) of the good vertices are marked (with marked vertices corresponding to the fixed points). We now give a bijection from these marked Dyck paths to the set of all balanced paths of $n-1$ N steps and $n-1$ E steps that end with an E step, counted by $\frac{1}{2}\binom{2(n-1)}{n-1}$. For each marked vertex v, draw a line from v down to the diagonal and then, in gray, left to the bounce vertex of v, so the new E steps are colored gray. Erase all lines that can't be "seen" from the diagonal, leaving a new Dyck path with (possibly) some gray steps. Lastly, take each component that ends with a gray step and flip it over the diagonal, and then "forget" the coloring. The result is the desired balanced path. The terminal E step of the Dyck path remains undisturbed and so the balanced path always ends with an E. For example, the permutation in Figure 2 is an indecomposable π_3-avoider of length $n = 12$ with 4 fixed points and it produces the Dyck path of semilength $n - 1 = 11$ with 4 marked vertices in Figure 4a corresponding bijectively to the balanced path in Figure 4b. To reverse the map, record the points p on the diagonal that terminate an N step lying below the diagonal. Flip over the diagonal each component that lies below the diagonal. Then, for each p, there is a new E segment (= maximal sequence of contiguous E steps) into p and a N segment out of p that may be new or original. In any case, interchange these E and N segments in the path. Lastly, mark the vertex directly above each p.

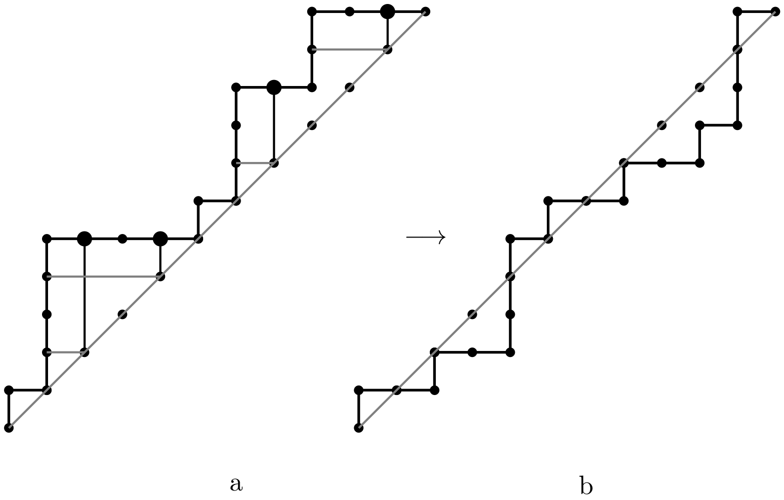


Figure 4: A marked Dyck path (a) and its corresponding balanced path (b)

2.3.1. Class 3, alternative count

Let a_n be the number of permutations of length n that avoid π_3. In order to study the sequence a_n, we extend our notation by defining $a_n(i_1, i_2, \ldots, i_s)$ to be the number of permutations $\sigma_1\sigma_2\cdots\sigma_n$ of length n that avoid π_3 such that $\sigma_1\sigma_2\cdots\sigma_s = i_1 i_2 \cdots i_s$.

Lemma 2.9. *We have*

$$a_n(i) = 2a_{n-1}(i) + \sum_{j=1}^{i} a_{n-1}(j) - 2\sum_{j=1}^{i} a_{n-2}(j), \qquad 1 \le i \le n-3,$$

$$a_n(n-2) = 2a_{n-1}(n-2) + \sum_{j=1}^{n-3} a_{n-1}(j) - 2\sum_{j=1}^{n-3} a_{n-2}(j) + a_{n-2} - a_{n-3},$$

with the initial conditions $a_n(n) = a_n(n-1) = a_{n-1}$.

Proof. By the definitions the initial conditions hold, and for $1 \le i \le n-2$,

$$a_n(i) = \sum_{j=1}^{i-1} a_n(i,j) + \sum_{j=i+1}^{n-2} a_n(i,j) + a_n(i,n-1) + a_n(i,n).$$

Clearly, $a_n(i,j) = 0$ for all $1 \le i < j \le n-2$ and $a_n(i,n-1) = a_n(i,n) = a_{n-1}(i)$ for all $1 \le i \le n-2$. Thus,

$$a_n(i) = 2a_{n-1}(i) + \sum_{j=1}^{i-1} a_n(i,j). \tag{2.1}$$

Also, for $1 \le j < i \le n-3$,

$$a_n(i,j) = \sum_{\ell=1}^{j-1} a_n(i,j,\ell) + \sum_{\ell=j+1}^{n-1} a_n(i,j,\ell) + a_n(i,j,n) = \sum_{\ell=1}^{j-1} a_{n-1}(j,\ell) + a_{n-1}(i,j),$$

which, by (2.1), implies $a_n(i,j) = a_{n-1}(j) - 2a_{n-2}(j) + a_{n-1}(i,j)$. Hence, (2.1) gives

$$a_n(i) = 2a_{n-1}(i) + \sum_{j=1}^{i} a_{n-1}(j) - 2\sum_{j=1}^{i} a_{n-2}(j), \qquad 1 \le i \le n-3,$$

$$a_n(n-2) = 2a_{n-1}(n-2) + \sum_{j=1}^{n-3} a_{n-1}(j) - 2\sum_{j=1}^{n-3} a_{n-2}(j) + a_{n-2} - a_{n-3}.$$

□

In order to solve the recurrence in Lemma 2.9, we define $A_n(v)$ to be the polynomial $\sum_{i=1}^{n} a_n(i)v^{i-1}$. Then, by translating these recurrences in terms of $A_n(v)$, we obtain

$$\begin{aligned} &A_n(v) - a_{n-3}v^{n-3} - a_{n-1}v^{n-1} - a_{n-1}v^{n-2} \\ &= 2(A_{n-1}(v) - a_{n-2}v^{n-2}) + \frac{A_{n-1}(v) - v^{n-2}A_{n-1}(1)}{1-v} \\ &- \frac{2(A_{n-2}(v) - v^{n-2}A_{n-2}(1))}{1-v}, \end{aligned}$$

which is equivalent to

$$\begin{aligned} A_n(v) = {} & A_{n-3}(1)v^{n-3} + 2A_{n-1}(v) \\ & + \frac{A_{n-1}(v) - v^n A_{n-1}(1)}{1-v} - \frac{2(A_{n-2}(v) - v^{n-1}A_{n-2}(1))}{1-v}, \end{aligned} \tag{2.2}$$

for all $n \geq 3$. By the definitions, we have $A_0(v) = A_1(v) = 1$ and $A_2(v) = 1 + v$.

Now let $A(x, v) = \sum_{n\geq 0} A_n(v)x^n$ be the generating function for the sequence $A_n(v)$. Multiplying (2.2) by x^n, and summing over all $n \geq 3$, yields

$$\begin{aligned} A(x,v) - (1+v)x^2 - x - 1 = {}& x^3A(xv,1) + 2x(A(x,v) - x - 1) \\ &+ \frac{x}{1-v}(A(x,v) - 1 - x - v(A(xv,1) - 1 - xv)) \\ &- \frac{2x^2}{1-v}(A(x,v) - 1 - v(A(xv,1) - 1)), \end{aligned}$$

which may be written as

$$\begin{aligned} &\left(1 - 2\frac{x}{v} - \frac{x}{v(1-v)} + \frac{2x^2}{v^2(1-v)}\right) A(x/v, v) \\ &\qquad = \frac{x^2}{v^2} + \frac{x^2}{v} - x - \frac{3x}{v} + 1 + \left(\frac{x^3}{v^3} - \frac{x}{1-v} + \frac{2x^2}{v(1-v)}\right) A(x,1). \end{aligned}$$

To solve the preceding functional equation, we make use of the *kernel method* (see, e.g., [4]). Let $v = v_0(x) = \frac{1+\sqrt{1-4x}}{2}$. Then, the kernel $1 - 2\frac{x}{v} - \frac{x}{v(1-v)} + \frac{2x^2}{v^2(1-v)}$ at $v = v_0(x)$ equals zero, which implies

$$\left(\frac{x}{1-v_0(x)} - \frac{x^3}{v_0^3(x)} - \frac{2x^2}{v_0(x)(1-v_0(x))}\right) A(x,1) = \frac{x^2}{v_0^2(x)} + \frac{x^2}{v_0(x)} - x - \frac{3x}{v_0(x)} + 1.$$

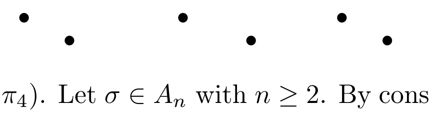
Simplifying the formula found for $A(x, 1)$ yields, after several algebraic steps, the following result.

Theorem 2.10. *The generating function for the number of permutations of length n that avoid π_3 is given by*

$$\frac{2(1-4x)}{2 - 9x + 4x^2 - x\sqrt{1-4x}}.$$

2.4. Class 4

$\pi_4 = \{2314, 2341, 2413\}$



Let $A_n = S_n(\pi_4)$. Let $\sigma \in A_n$ with $n \geq 2$. By considering the positions of $n-1$ and n within σ, one can show the following block decomposition result.

Lemma 2.11. *Let $n \geq 2$. A permutation σ of length n avoids π_4 if and only if either*

- $\sigma=\sigma'(n-1)\sigma''n\sigma'''$ *such that* $\sigma'>\sigma''\sigma'''$ *(that is, each letter of* σ' *is greater than each letter of* σ'' *or* σ'''*),* σ' *is a permutation of* $[n-j+1,n-2]$ *that avoids* 231*, and* $\sigma''n\sigma'''$ *is a permutation of* $\{1,2,\dots,n-j,n\}$ *that avoids* π_4*; or*

- $\sigma=\sigma'n\sigma''n-1\sigma'''$*: If* $\sigma'=\emptyset$*, then* $\sigma\in A_n$ *if and only if* $\sigma''(n-1)\sigma'''\in A_{n-1}$*. If* $\sigma'\neq\emptyset$ *and* $\sigma''=\emptyset$*, then* $\sigma\in A_n$ *if and only if* $\sigma'(n-1)\sigma'''\in A_{n-1}$*. If* $\sigma',\sigma''\neq\emptyset$*, then* $\sigma'>\sigma''\sigma'''$*,* σ' *avoids* 231*, and* $\sigma''(n-1)\sigma'''$ *avoids* π_4*.*

Define $A(x)=\sum_{n\geq0}\#A_nx^n$. Since 231-avoiders are counted by Catalan numbers, we have by Lemma 2.11,

$$\begin{aligned}A(x)&=1+x+xC(x)(A(x)-1)\\&\quad+x(A(x)-1)+x(A(x)-1-xA(x))+x(C(x)-1)(A(x)-1-xA(x)),\end{aligned}$$

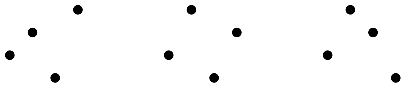
where $A(x)-1-xA(x)$ is the generating function for the number of permutations $\sigma_1\cdots\sigma_n$ in A_n, $n\geq2$, with $\sigma_1\neq n$. Thus, we can state the following result.

Theorem 2.12. *The generating function for the number of permutations of length* n *that avoid* π_4 *is given by*

$$\frac{2(1-4x)}{2-9x+4x^2-x\sqrt{1-4x}}.$$

2.5. Class 5

$\pi_5=\{2314,2413,2431\}$



Let $A_n=S_n(\pi_5)$. Let $\sigma\in A_n$ with $n\geq2$. Again, we have a block decomposition of σ.

Lemma 2.13. *Let* $n\geq2$*. A permutation* σ *of length* n *avoids* π_5 *if and only if either*

- $\sigma=\sigma'n\sigma''(n-1)\sigma'''$ *such that* $\sigma''(n-1)\sigma'''>\sigma'$*,* σ' *is a permutation of* $[j]$ *that avoids* 231*, and* $\sigma''(n-1)\sigma'''$ *is a permutation of* $[j+1,n-1]$ *that avoids* π_5*; or*

- $\sigma=\sigma'(n-1)\sigma''n\sigma'''$*: If* $\sigma'=\emptyset$*, then* $\sigma\in A_n$ *if and only if* $\sigma''(n-1)\sigma'''\in A_{n-1}$*. If* $\sigma'\neq\emptyset$ *and* $\sigma''=\emptyset$*, then* $\sigma\in A_n$ *if and only if* $\sigma'(n-1)\sigma'''\in A_{n-1}$*. If* $\sigma',\sigma''\neq\emptyset$*, then* $\sigma''n\sigma'''>\sigma'$*,* σ' *is a permutation of* $[j]$ *that avoids* 231*, and* $\sigma''(n-1)\sigma'''$ *is a permutation of* $\{j+1,j+2,\dots,n-2,n\}$ *that avoids* π_5*.*

Define $A(x) = \sum_{n \geq 0} \#A_n x^n$. By Lemma 2.13, we have

$$\begin{aligned} A(x) &= 1 + x + xC(x)(A(x) - 1) \\ &\quad + x(A(x) - 1) + x(A(x) - 1 - xA(x)) + x(C(x) - 1)(A(x) - 1 - xA(x)), \end{aligned}$$

where $A(x) - 1 - xA(x)$ is the generating function for the number of permutations $\sigma_1 \cdots \sigma_n$ in A_n, $n \geq 2$, with $\sigma_1 \neq n$. Thus, we can state the following result.

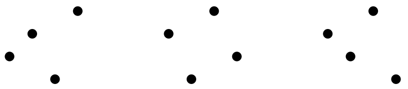
Theorem 2.14. *The generating function for the number of permutations of length n that avoid π_5 is given by*

$$\frac{2(1-4x)}{2 - 9x + 4x^2 - x\sqrt{1-4x}}.$$

Note that Lemmas 2.11 and 2.13 yield a recursive bijection between $S_n(\pi_4)$ and $S_n(\pi_5)$.

2.6. Class 6

We use the representative triple $\pi_6 = \{2314, 3142, 3241\}$



In order to determine the number of π_6-avoiders of length n, we refine the set by considering a couple of auxiliary statistics as follows. Given $n \geq 2$, $\ell \in [n-1]$, and $1 \leq i \leq \ell$, let $u(n; \ell, i)$ denote the number of permutations of length n avoiding the patterns in π_6 in which the largest letter (if it exists) to the left of n is ℓ wherein there are exactly $i - 1$ positions separating ℓ and n. Let $u(n; \ell) := \sum_{i=1}^{\ell} u(n; \ell, i)$. Denote by $u(n)$ the number of permutations of length n avoiding the patterns in π_6, the set of which we will denote by $\mathcal{U}_n$. Since members of $\mathcal{U}_n$ starting with n are synonymous with members of $\mathcal{U}_{n-1}$, we have the relation

$$u(n) = u(n-1) + \sum_{\ell=1}^{n-1} u(n; \ell), \qquad n \geq 2, \tag{2.3}$$

with $u(1) = u(0) = 1$. The following lemma provides a recurrence for the array $u(n; \ell, i)$ which we will use to determine $u(n)$.

Lemma 2.15. *If $2 \leq i \leq \ell < n$, then*

$$u(n; \ell, i) = C_{\ell-i} C_{i-1} u(n - \ell - 1), \qquad i \geq 2, \tag{2.4}$$

with

$$u(n; \ell, 1) = C_{n-\ell-1} u(\ell) + C_{\ell-1} u(n - \ell - 1) - C_{\ell-1} C_{n-\ell-1}$$

$$+ (C_\ell - C_{\ell-1}) \sum_{r=0}^{n-\ell-2} C_r u(n-\ell-r-1), \tag{2.5}$$

for $\ell \geq 2$, *and* $u(n;1,1) = u(n-2)$ *for* $n \geq 2$.

Proof. That $u(n;1,1) = u(n-2)$ for $n \geq 2$ follows from the definitions. We give combinatorial proofs of (2.4) and (2.5). Let $\mathcal{U}_{n,\ell,i}$ denote the subset of $\mathcal{U}_n$ enumerated by $u(n;\ell,i)$. To show (2.4), note first that members of $\mathcal{U}_{n,\ell,i}$, where $2 \leq i \leq \ell$, must be of the form $\alpha = \alpha_1 \ell \alpha_2 n \delta$, where α_2 has length $i-1$ and δ comprises the elements of $[\ell+1, n-1]$. (Note α_2 non-empty implies that there can be no members of $[\ell-1]$ to the right of n, for otherwise there would be an occurrence of 3241 or 3142 in which the roles of the “3” and “4” are played by the ℓ and n, respectively.) Furthermore, any letter in α_1 must be smaller than any letter in α_2 in order to avoid 2314. Finally, the subwords α_1 and α_2 must both avoid 231 (since n lies to their right), while there is no further restriction on δ (i.e., it must only avoid the original patterns in π_6). Conversely, any permutation α of $[n]$ of the form described above in which α_1 and α_2 both avoid 231, each letter of α_2 is greater than each letter of α_1, and δ avoids the patterns in π_6 is seen to be a member of $\mathcal{U}_{n,\ell,i}$. This implies $u(n;\ell,i) = C_{\ell-i}C_{i-1}u(n-\ell-1)$ for $2 \leq i \leq \ell$, as desired.

To show (2.5), let $X = \mathcal{U}_{n,\ell,1}$ and first consider the case in which there are no elements of $[\ell-1]$ occurring to the right of the letter n within a member of X. Then such members of X may be decomposed as $\alpha\ell n\beta$, where α is a permutation of $[\ell-1]$ avoiding the pattern 231 and $\beta \in \mathcal{U}_{n-\ell-1}$ (on the letters in $[\ell+1, n-1]$). Furthermore, permutations of this form are seen to avoid the patterns in π_6. Thus, there are $C_{\ell-1}u(n-\ell-1)$ possibilities in this case.

Now suppose that all elements of $[\ell-1]$ occur to the right of n within $\rho \in X$. We consider subcases as follows. First assume ρ is of the form $\rho = \ell n \rho_1 \rho_2$, where ρ_1 and ρ_2 are permutations of $[\ell+1, n-1]$ and $[\ell-1]$, respectively. Then ρ_1 must avoid the pattern 213 since $\rho_2 \neq \emptyset$, while ρ_2 has no restrictions other than those imposed by π_6. This implies that there are $C_{n-\ell-1}u(\ell-1)$ possibilities in this case. Now assume that at least one letter of $[\ell-1]$ lies to the left of some letter of $[\ell+1, n-1]$ within ρ. Then ρ must be of the form $\rho = \ell n \delta_1 \gamma \delta_2$ in this case, where γ consists of all the letters in $[\ell-1]$ and δ_1 and δ_2 together comprise all of the letters in $[\ell+1, n-1]$, with δ_2 non-empty. (For otherwise, there would be a guaranteed occurrence of 3241 or 3142, with the ℓ playing the role of the “3”.) Furthermore, since $\ell \geq 2$ implies γ is non-empty, it must be the case that all letters of δ_1 are larger than all letters of δ_2 in order to avoid 2314. In addition, γ non-empty implies δ_1 must avoid 213 and δ_2 non-empty implies γ must avoid 231. Finally, the subword δ_2 is seen to have no restrictions other than those imposed by π_6 since all letters of δ_1 and γ are larger or smaller, respectively, than all letters of δ_2. Since the preceding conditions on γ, δ_1 and δ_2 are seen also to be sufficient for membership of ρ within X, it follows that there are $C_{\ell-1}\sum_{r=0}^{n-\ell-2} C_r u(n-\ell-r-1)$ possibilities in this case, where r denotes the length of δ_1.

Now suppose that there is at least one element of $[\ell-1]$ to the left and to the right of n within $\beta \in X$, whence $\ell \geq 3$ in this case. Then β can be expressed in the form $\beta = \beta_1 \ell n \delta_1 \beta_2 \delta_2$, where β_1, β_2 are non-empty words in $[\ell - 1]$ and δ_1, δ_2 are words in $[\ell + 1, n - 1]$. First assume δ_2 is non-empty. Note that all elements of β_1 must be less than all of those in β_2 in this case in order to avoid 2314 (for otherwise, there would be an occurrence of 2314 in which the ℓ plays the role of the "3" and any member of δ_2 plays the role of the "4"). Let p be the smallest element of β_2. Then $2 \leq p \leq \ell - 1$ since both β_1 and β_2 are non-empty. Furthermore, δ_2 non-empty implies both β_1 and β_2 avoid 231, which implies that there are $\sum_{p=2}^{\ell-1} C_{p-1}C_{\ell-p} = C_\ell - 2C_{\ell-1}$ possibilities for β_1 and β_2. Once the positions of the letters in β_1 and β_2 have been determined, there are $\sum_{r=0}^{n-\ell-2} C_r u(n-\ell-r-1)$ possibilities for the letters in δ_1 and δ_2, upon considering the length r of δ_1 (note that all letters in δ_2 must be smaller than all letters in δ_1 in order to avoid 2314). Thus, there are $(C_\ell - 2C_{\ell-1}) \sum_{r=0}^{n-\ell-2} C_r u(n - \ell - r - 1)$ members β of the form above in which $\delta_2 \neq \emptyset$.

Finally, suppose $\delta_2 = \emptyset$ in the decomposition of β above. In this case, the subsequence $\beta_1 \ell \beta_2$ constitutes a permutation of $[\ell]$ avoiding the patterns in π_6 which does not start or end with the letter ℓ. By subtraction, there are $u(\ell) - u(\ell - 1) - C_{\ell-1}$ possibilities for this subsequence. The letters of δ_1 must avoid 213, with no other restrictions on δ_1. Furthermore, any permutation β of the form above satisfying the stated conditions on its constituent parts is seen to avoid the patterns in π_6. Since there are $C_{n-\ell-1}$ possibilities for δ_1, it follows that there are $(u(\ell) - u(\ell-1) - C_{\ell-1})C_{n-\ell-1}$ permutations β of the form above in which $\delta_2 = \emptyset$. Combining all of the previous cases implies that the cardinality of X is given for $\ell \geq 2$ by

$$\begin{aligned}
&C_{\ell-1}u(n-\ell-1) + C_{n-\ell-1}u(\ell-1) + C_{\ell-1}\sum_{r=0}^{n-\ell-2} C_r u(n-\ell-r-1)\\
&+ (C_\ell - 2C_{\ell-1})\sum_{r=0}^{n-\ell-2} C_r u(n-\ell-r-1) + (u(\ell) - u(\ell-1) - C_{\ell-1})C_{n-\ell-1}\\
&= C_{n-\ell-1}u(\ell) + C_{\ell-1}u(n-\ell-1) - C_{\ell-1}C_{n-\ell-1}\\
&+ (C_\ell - C_{\ell-1})\sum_{r=0}^{n-\ell-2} C_r u(n-\ell-r-1),
\end{aligned}$$

which gives (2.5). □

Summing (2.4) over $2 \leq i \leq \ell$, and using the recurrence for Catalan numbers, implies

$$u(n;\ell) = u(n;\ell,1) + (C_\ell - C_{\ell-1})u(n-\ell-1), \qquad \ell \geq 1. \tag{2.6}$$

Summing (2.6) over $1 \leq \ell \leq n-1$, and using (2.5), implies

$$\sum_{\ell=1}^{n-1} u(n;\ell) = \sum_{\ell=1}^{n-1} u(n;\ell,1) + \sum_{\ell=1}^{n-1}(C_\ell - C_{\ell-1})u(n-\ell-1)$$

$$
\begin{aligned}
&= u(n-2) + \sum_{\ell=2}^{n-1} C_{n-\ell-1}u(\ell) + \sum_{\ell=2}^{n-1} C_{\ell-1}u(n-\ell-1) - \sum_{\ell=2}^{n-1} C_{\ell-1}C_{n-\ell-1} \\
&+ \sum_{\ell=2}^{n-1}(C_\ell - C_{\ell-1}) \sum_{r=0}^{n-\ell-2} C_r u(n-\ell-r-1) + \sum_{\ell=1}^{n-1}(C_\ell - C_{\ell-1})u(n-\ell-1).
\end{aligned}
$$

Thus, we have by (2.3),

$$
\begin{aligned}
u(n) &= \sum_{\ell=0}^{n-1} C_{n-\ell-1}u(\ell) - C_{n-1} + \sum_{\ell=1}^{n-1} C_{\ell-1}u(n-\ell-1) - \sum_{\ell=1}^{n-1} C_{\ell-1}C_{n-\ell-1} \\
&+ \sum_{\ell=1}^{n-1}(C_\ell - C_{\ell-1}) \sum_{r=0}^{n-\ell-2} C_r u(n-\ell-r-1) + \sum_{\ell=0}^{n-1}(C_\ell - C_{\ell-1})u(n-\ell-1) \\
&= 2\sum_{\ell=1}^{n-1} C_{n-\ell-1}u(\ell) + C_{n-1} - \sum_{\ell=1}^{n-1} C_{\ell-1}C_{n-\ell-1} \\
&+ \sum_{\ell=1}^{n-1}(C_\ell - C_{\ell-1}) \sum_{r=0}^{n-\ell-2} C_r u(n-\ell-r-1) \\
&= 2\sum_{\ell=1}^{n-1} C_{n-\ell-1}u(\ell) + \sum_{\ell=1}^{n-1}(C_\ell - C_{\ell-1}) \sum_{r=0}^{n-\ell-2} C_r u(n-\ell-r-1), \qquad n \geq 2.
\end{aligned}
\tag{2.7}
$$

Let $f(x) = \sum_{n\geq 1} u(n)x^n$. Multiplying both sides of (2.7) by x^n, and summing over $n \geq 2$, yields

$$
\begin{aligned}
f(x) &= x + 2xC(x)f(x) + \sum_{\ell\geq 1}(C_\ell - C_{\ell-1}) \sum_{r\geq 0} C_r \sum_{n\geq \ell+r+2} u(n-\ell-r-1)x^n \\
&= x + 2xC(x)f(x) + x\sum_{\ell\geq 1}(C_\ell - C_{\ell-1})x^\ell \sum_{r\geq 0} C_r x^r \sum_{n\geq 1} u(n)x^n \\
&= x + 2xC(x)f(x) + x((1-x)C(x)-1)C(x)f(x) \\
&= x + xC(x)f(x) + (1-x)(C(x)-1)f(x) \\
&= x + (C(x)+x-1)f(x),
\end{aligned}
$$

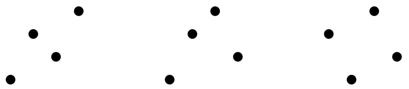
where we have used the fact $xC^2(x) = C(x) - 1$. Thus, we have

$$
\begin{aligned}
\sum_{n\geq 0} u(n)x^n = 1 + f(x) = \frac{2-C(x)}{2-x-C(x)} &= \frac{1-4x-\sqrt{1-4x}}{1-4x+2x^2-\sqrt{1-4x}} \\
&= \frac{2(1-4x)}{2-9x+4x^2-x\sqrt{1-4x}},
\end{aligned}
$$

as desired.

2.7. Class 7

$\pi_7 = \{1324, 1342, 3142\}$



To count π_7-avoiders by first entry m, set $u(n) = |S_n(\pi_7)|$ and $u(n,m) = |\{p \in S_n(\pi_7) : p_1 = m\}$.

Clearly, $u(n,m) = u(n-1)$ for $m = n$. For $1 \le m < n$, use the left to right maxima $(m_i)_{i=1}^{k+1}$, where $k \ge 1$, $m_1 = m$, and $m_{k+1} = n$, to decompose p as

$$p = m_1 A_1 m_2 A_2 \cdots m_k A_k m_{k+1} A_{k+1}. \tag{2.8}$$

Proposition 2.16.
(i) $m_1, \ldots, m_k$ *are consecutive integers.*
(ii) $A_1 > A_2 > \cdots > A_k > [m-1] \cap A_{k+1}$*, where* $A_i > A_j$ *means* $min(A_i) > max(A_j)$*.*
(iii) For $1 \le i \le k$*,* A_i *avoids* 132*.*

Proof. (i) Say $m_i = a$ and $m_{i+1} = c \ge a+2$. Then $b := a+1$ occurs after m_{i+1} and $\{a,c,b,n\}$ occur either in the order $acbn$ (1324) or $acnb$ (1342), both forbidden.

(ii) If $a_i < a_j$ with $1 \le i < j \le k+1$, $a_i \in A_i$, $a_j \in A_j$, then $m_i a_i m_j a_j$ is a 3142.

(iii) If not, then $n = m_{k+1}$ would be the "4" of a 1324. □

Thus p is captured by the list (recall St refers to standardizing a list)

$$\mathrm{St}(A_1), \ldots, \mathrm{St}(A_k), \mathrm{St}(m_1 A_{k+1})\,.$$

Conversely, if these conditions hold and $\mathrm{St}(m_1 A_{k+1})$ is a π_7-avoider, then so is p.

Since 132-avoiders of length n are equinumerous with Dyck paths of size (semilength) n, and $(k+1)$-lists of Dyck paths of total size n are counted by the generalized Catalan number $C(n,k) := (k+1)\binom{2n+k+1}{n}/(2n+k+1)$, the decomposition (2.8) leads to the recurrence

$$u(n,m) = \sum_{k=1}^{n-m} \sum_{h=0}^{m-1} C(m-h-1, k-1)u(n-m+h-k+1, h+1),$$

where the index h refers to the number of entries of A_{k+1} that are $< m_1$. Recall that the generating function $C(x,y) := \sum_{n,k\ge 0} C(n,k)x^n y^k$ is given by $C(x,y) = C(x)/(1-yC(x))$ where $C(x)$ is the generating function for the Catalan numbers.

Now define generating functions $F(x) = \sum_{n\ge 1} u(n)x^n$ and

$$F(x,y) = \sum_{n\ge 1} \sum_{m=1}^{n} u(n,m)x^n y^m.$$

Note that $F(x) = F(x,1)$.

Split $F(x,y)$ into $F_1 + F_2$, where F_1 is the sum over $m < n$ and F_2 is the sum over $m = n$. Using the recurrence, we have

$$F_1 = \sum_{n\geq 2}\sum_{m=1}^{n-1}\sum_{k=1}^{n-m}\sum_{h=0}^{m-1} C(m-h-1,k-1)u(n-m+h-k+1,h+1)x^n y^m.$$

Introduce new summation indices $r = m-h-1, s = k-1, t = n-m+h-k+1, j = h+1$ to get

$$F_1 = \sum_{r,s\geq 0,t\geq 1}\sum_{j=1}^{t} C(r,s)u(t,j)x^{r+s+t+1}y^{j+r} = xC(xy,x)F(x,y)\,.$$

Also, we have

$$F_2 = \sum_{n\geq 1} u(n-1)(xy)^n = xy\big(1+F(xy,1)\big)\,.$$

So $F(x,y)$ satisfies

$$F(x,y) = xC(xy,x)F(x,y) + xy + xyF(xy,1)\,. \tag{2.9}$$

Set $y = 1$ in (2.9) to get $F(x,1) = x/(1-x-xC(x,x))$, leading to

$$F(x) = \frac{x}{1-x-\frac{xC(x)}{1-xC(x)}},$$

and, after expansion,

$$F(x,y) = \frac{xy\big(1+F(xy)\big)}{1-xC(xy,x)}\,,$$

and $1+F(x) = \frac{2(1-4x)}{2-9x+4x^2-x\sqrt{1-4x}}$.

As an aside, the decomposition (2.8) readily yields a bijection from $S_n(\pi_7)$ to a certain subset of the Schroder paths of size $n-1$. We represent a Schroder path as a Motzkin path consisting of upsteps $U = (1,1)$, flatsteps $F = (1,0)$ and downsteps $D = (1,-1)$, but with size measured by # U's + # F's rather than length. Let $\mathcal{A}_n$ denote the set of Schroder paths of size n with all flatsteps at ground level, ending with an F, and decorated so that, for each descent (maximal sequence of contiguous downsteps) that ends at ground level, one of its downsteps is marked. Let $\mathcal{B}_n$ denote the set of Schroder paths of size n such that, for each flatstep not at ground level, the portion of the path between the flatstep and the next vertex at ground level consists of a Dyck path (possibly empty) followed only by downsteps. There is a simple bijection from $\mathcal{A}_n$ to $\mathcal{B}_{n-1}$: delete the last step (necessarily F) and, for each marked downstep, if it is the last downstep of a descent, just erase the mark, otherwise delete the marked step and turn its matching upstep into a flatstep. For example, here is a bijection from $S_n(\pi_7)$ to the paths in $\mathcal{A}_n$. Let ϕ be your favorite bijection from 312-avoiders to Dyck paths. Given $p \in S_n(\pi_7)$, if the first

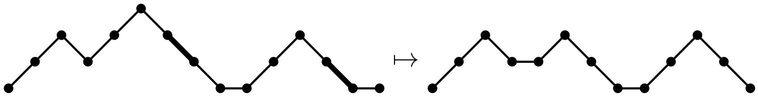


Figure 5: The bijection $\mathcal{A}_n \longrightarrow \mathcal{B}_{n-1}$

entry of p is n, begin the path with a flatstep, delete n, and start over. Otherwise, consider the decomposition (2.8). Replace each $m_i, 1 \leq i \leq k$, by an upstep, each $A_i, 1 \leq i \leq k$, by the Dyck path $\phi\big(\mathrm{St}(A_i)\big)$, append k downsteps and mark the first one. These replacements and the appendage produce a primitive Dyck path with one marked downstep on the last descent. Next, ignore the entry $m_{k+1} = n$ and start over with $\mathrm{St}(m_1 A_{k+1})$. The process will end when $\mathrm{St}(m_1 A_{k+1}) = 1$, which will terminate the path with a flatstep.

2.7.1. Class 7, alternative count

Let $b(n;i,j)$ denote the number of permutations of length n avoiding the patterns in π_7 in which the first letter is i and the second is j. If $n \geq 2$, then define $b(n;i) = \sum_{j=1, j\neq i}^{n} b(n;i,j)$ and $b(n) = \sum_{i=1}^{n} b(n;i)$, with $b(1) = b(1;1) = 1$. Put $b(n;i,j) = 0$ if $i = 0$ or $j = 0$.

We have the following obvious initial values. If $n = 2$, then $b(2) = 2$, with $b(2;1) = b(2;1,2) = 1$ and $b(2;2) = b(2;2,1) = 1$. If $n = 3$, then $b(3) = 6$, with $b(3;1) = b(3;2) = b(3;3) = 2$ and $b(3;1,2) = b(3;1,3) = b(3;2,1) = b(3;2,3) = b(3;3,1) = b(3;3,2) = 1$.

If $n \geq 4$, then the array $b(n;i,j)$ is determined as follows.

Lemma 2.17. *If $1 \leq i \leq n-1$, then $b(n;i,i+1) = b(n;i,n) = b(n;n,i) = b(n-1;i)$, with $b(n;i,i-1) = b(n-1;i-1)$ for $1 < i \leq n$. If $1 \leq i < j-1 < n-1$, then $b(n;i,j) = 0$. If $1 \leq j < i-1 < n-1$, then*

$$b(n;i,j) = b(n-1;i-1,j) + \sum_{k=1}^{j-1} b(n-1;i-1,k). \tag{2.10}$$

Proof. Let $\mathcal{B}_n$ denote the subset of the permutations of length n avoiding the patterns in π_7 and $\mathcal{B}_{n,i,j}$ the subset of $\mathcal{B}_n$ enumerated by $b(n;i,j)$. The first statement is clear since a letter n in either the first or second position is seen to be extraneous concerning the avoidance of the patterns in π_7, as is the letter $i+1$ within members of $\mathcal{B}_{n,i,i+1}$ and the letter $i-1$ within members of $\mathcal{B}_{n,i,i-1}$. Permutations of length at least four starting with the letters i, j where $1 \leq i < j-1 < n-1$ always contain an occurrence of either 1324 or 1342, which implies $b(n;i,j) = 0$ in these cases.

To show (2.10), we consider the third letter k within a member of $\mathcal{B}_{n,i,j}$ where $1 \leq j < i-1 < n-1$. Note that k cannot belong to $[i+1, n]$, for if it did, then there would be an occurrence of 3142, as witnessed by any subsequence $ijkx$, where $x \in [j+1, i-1]$. It also cannot be the case that k belongs to $[j+2, i-1]$, for otherwise there would be an occurrence of 1342 or 1324 with either $jkn(j+1)$ or $jk(j+1)n$. Thus, it must be the case that $k=j+1$ or $k \in [j-1]$. The first term on the right-hand side of (2.10) accounts for when $k=j+1$ since the letter k is seen to be extraneous in this case concerning the avoidance of the patterns in π_7 and thus may be deleted.

So assume $k \le j-1$, and we will show that the letter j may be deleted from members of $\mathcal{B}_{n,i,j}$ in this case. Given $\lambda \in \mathcal{B}_{n-1,i-1,k}$, let λ' be obtained from λ by inserting j between the $i-1$ and k and increasing all letters of λ in $[j, n-1]$ by one. We will show that if λ avoids the patterns in π_7, then so must λ'. Suppose, to the contrary, that λ' contains an occurrence of some pattern $\rho \in \pi_7$. Then ρ cannot be either 1342 or 1324, for otherwise the letter j would play the role of the "1" in an occurrence of either pattern within λ', and replacing j with $k<j$ would imply λ contains one of these patterns, a contradiction. Thus ρ must be 3142. Note that the role of the "3" must be played by the letter j, for otherwise λ would contain an occurrence of 3142 with the "3" and "1" played by $i-1$ and k, respectively.

Thus, the occurrence of 3142 in λ' is realized by a subsequence $j\ell rs$. Note that $r<i$, for otherwise λ would contain an occurrence of 3142 with $(i-1)\ell(r-1)s$, which is impossible. We now consider the position of the element n within λ'. If n lies to the left of r within λ', then $(i-1)k(n-1)(r-1)$ would form an occurrence of 3142 in λ, a contradiction. On the other hand, if n lies to the right of r within λ', then there would be an occurrence of 1324 or 1342 within λ' as witnessed by either ℓrsn or ℓrns, a contradiction. Thus, λ' must avoid the patterns in π_7 if λ does, which completes the proof. □

Define $b(n;i|w) = \sum_{j=1}^{n} b(n;i,j)w^{j-1}$ and

$$B_n(v,w) = \sum_{i=1}^{n}\sum_{j=1}^{n} b(n;i,j)v^{i-1}w^{j-1}.$$

Then the recurrence (2.10) implies

$$\begin{aligned}
&b(n;i|w) - b(n-1;i-1)w^{i-2} - b(n-1;i)w^{i}\delta_{i<n-1} - b(n-1;i)w^{n-1}\\
&= \sum_{j=1}^{i-2} w^{j-1} \sum_{k=1}^{j} b(n-1;i-1,k)\\
&= \frac{1}{1-w}\Big(b(n-1;i-1|w) - w^{i-2}b(n-1;i-1|1)\\
&\qquad + b(n-2;i-1)((1+\delta_{i<n-1})w^{i-2} - w^{i-1} - w^{n-2}\delta_{i<n-1})\Big),
\end{aligned}$$

which implies

$$\begin{aligned}
b(n;i|w) &= b(n-1;i-1)w^{i-2} + b(n-1;i)w^{i}\delta_{i<n-1} + b(n-1;i)w^{n-1}\\
&\quad + b(n-2;i-1)w^{i-2} + \frac{1}{1-w}(b(n-1;i-1|w) - w^{i-2}b(n-1;i-1|1)
\end{aligned}$$

$$+ b(n-2; i-1)(w^{i-2} - w^{n-2})\delta_{i<n-1}). \tag{2.11}$$

Note that $b(n;1|w) = 2^{n-3}(w + w^{n-1})$ and $b(n;n|w) = \sum_{j=1}^{n-1} b(n-1;j)w^{j-1} = B_{n-1}(w,1)$. Also, $b(n;2,j)$ equals 2^{n-3}, 0, or $b(n-1;2)$ when $j = 1$, $4 \le j \le n-1$, or $j = 3, n$, respectively. Thus, $b(n;2|w) = 2^{n-3} + b(n-1;2|1)(w^2 + w^{n-1})$, which, by induction, implies $b(n;2|w) = 2^{n-3} + (n-2)2^{n-4}(w^2 + w^{n-1})$.

Multiplying (2.11) by v^{i-1}, and summing over $i = 3, 4, \ldots, n-1$, implies

$$\begin{aligned}B_n(v,w) &= B_{n-1}(w,1)v^{n-1} + (v+w)B_{n-1}(vw,1) - (vw)^{n-2}(v+w)B_{n-2}(1,1)\\ &\quad + w^{n-1}B_{n-1}(v,1) + vB_{n-2}(vw,1) - v^{n-2}w^{n-3}B_{n-3}(1,1)\\ &\quad + \frac{v}{1-w}\Big(B_{n-1}(v,w) - v^{n-2}B_{n-2}(w,1) - B_{n-1}(vw,1)\\ &\qquad + (vw)^{n-2}B_{n-2}(1,1) + B_{n-2}(vw,1) - w^{n-2}B_{n-2}(v,1)\Big),\end{aligned}$$

with $B_0(v,w) = B_1(v,w) = 1$, $B_2(v,w) = v + w$ and $B_3(v,w) = v + v^2 + w + w^2 + vw^2 + wv^2$.

Define $B(x;v,w) = \sum_{n\ge0} B_n(v,w)x^n$. Multiplying the last recurrence by x^n and summing over $n \ge 4$, we obtain after several algebraic steps

$$\begin{aligned}\frac{1-vx-w}{1-w}B(x;v,w) &= 1 - (v+w+1)x - vx^2\\ &\quad - \frac{x(vwx + vw - 2vx - w + w^2)}{1-w}B(x;vw,1)\\ &\quad + \frac{x(1-vx-w)}{1-w}(B(vx,w,1) + B(wx;v,1))\\ &\quad + \frac{x^2(vw + wvx + w^2 - w - vx)}{1-w}B(vwx;1,1).\end{aligned}$$

Substituting $w = 1 - vx$ into the preceding functional equation yields

$$1 = (2+v)x + (1 - vx - 2x)B(x;v(1-vx),1) - x(1 - vx - x)B(vx(1-vx);1,1).$$

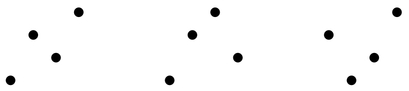
Let v be a solution of the equality $v(1-vx) = 1$, namely, $v = C(x) = \frac{1-\sqrt{1-4x}}{2x}$. Replacing v by $C(x)$ in the last functional equation then gives

$$B(x;1,1) = \frac{2(1-4x)}{2 - 9x + 4x^2 - x\sqrt{1-4x}},$$

as desired.

2.8. Class 8

$\pi_8 = \{1324, 1342, 3124\}$



Let $a(n;i,j)$ denote the number of permutations of length n avoiding the patterns in π_8 in which the first letter is i and the second is j. If $n \geq 2$, then define $a(n;i) = \sum_{j=1,j\neq i}^{n} a(n;i,j)$ and $a(n) = \sum_{i=1}^{n} a(n;i)$, with $a(1) = a(1;1) = 1$. Put $a(n;i,j) = 0$ if $i = 0$ or $j = 0$.

We have the following obvious initial values. If $n = 2$, then $a(2) = 2$, with $a(2;1) = a(2;1,2) = 1$ and $a(2;2) = a(2;2,1) = 1$. If $n = 3$, then $a(3) = 6$, with $a(3;1) = a(3;2) = a(3;3) = 2$ and $a(3;1,2) = a(3;1,3) = a(3;2,1) = a(3;2,3) = a(3;3,1) = a(3;3,2) = 1$.

If $n \geq 4$, then the array $a(n;i,j)$ is determined as follows.

Lemma 2.18. *If $1 \leq i \leq n-1$, then $a(n;i,i+1) = a(n;i,n) = a(n;n,i) = a(n-1;i)$, with $a(n;i,i-1) = a(n-1;i-1)$ for $1 < i \leq n$. If $1 \leq i < j-1 < n-1$, then $a(n;i,j) = 0$. If $1 \leq j < i-1 < n-1$, then*

$$a(n;i,j) = a(n-1;i,j) + a(n-1;i-1,j-1) + \sum_{k=1}^{j-2} a(n-1;j,k). \tag{2.12}$$

Proof. Let $\mathcal{A}_n = S_n(\pi_8)$ and $\mathcal{A}_{n,i,j}$ be the subset of $\mathcal{A}_n$ enumerated by $a(n;i,j)$. The first statement is clear since a letter n in either the first or second position is seen to be extraneous concerning the avoidance of the patterns in π_8, as is the letter $i+1$ within members of $\mathcal{A}_{n,i,i+1}$ and the letter $i-1$ within members of $\mathcal{A}_{n,i,i-1}$. Permutations of length at least four starting with the letters i, j where $1 \leq i < j-1 < n-1$ must contain an occurrence of either 1324 or 1342, whence $a(n;i,j) = 0$ in these cases.

We now show (2.12). To do so, we consider the third letter k within a member of $\mathcal{A}_{n,i,j}$ where $1 \leq j < i-1 < n-1$. Note that k cannot belong to $[i+1, n-1]$, for if it did, then there would be an occurrence of 1342 or 1324, as witnessed by either $jkn(i-1)$ or $jk(i-1)n$. It also cannot belong to $[j+1, i-1]$, for if it did, then there would be an occurrence of 3124, as witnessed by $ijkn$. Thus, it must be the case that $k = n$ or $k \in [j-1]$. It is seen that the first two terms on the right-hand side of (2.12) account for the cases in which $k = n$ or $k = j-1$, respectively. Now assume $k \in [j-2]$. In this case, we will argue that the letter i is superfluous when considering the avoidance of patterns in π_8, whence it may be deleted. This will give the sum on the right-hand side of (2.12) and complete the proof. Given $\lambda \in \mathcal{A}_{n-1,j,k}$, let λ' be obtained from λ by writing the letter i before λ and increasing all elements of $[i, n-1]$ within λ by one. We will show that if λ avoids the patterns in π_8, then so does λ'. Suppose, to the contrary, that λ' contains an occurrence of some pattern ρ of π_8. Since λ avoids the patterns in π_8, we must have $\rho = 3124$, with the letter i playing the role of the "3".

Suppose that the 3124 subsequence within λ' is witnessed by $i\ell rs$. Note that $r > j$, for otherwise λ would contain an occurrence of 3124 with the subsequence

$j\ell r(s-1)$. We consider several cases on ℓ. First assume $\ell \in [j+1, i-1]$. Then all elements of $[k+1, j-1]$ within λ' must occur to the left of r in order to avoid 1342, and thus to the left of ℓ as well in order to avoid 1324. But then λ would contain 3124 as witnessed by $jkx\ell$, where x is any element of $[k+1, j-1]$, a contradiction. On the other hand, if $\ell \in [k+1, j-1]$, then λ would contain 3124 with the subsequence $jk\ell r$, which is again not possible. Finally, let us assume $\ell \in [k]$; note that $\ell = j$ is included in this case, for if the second letter in an occurrence of 3124 starting with i is j, then one may replace j with k since $k < j$. Note that then any $x \in [k+1, j-1]$ must lie to the left of s within λ', for if x was to the right of s, then $kr(s-1)x$ would be an occurrence of 1342 within λ, which is impossible. But x lying to the left of s within λ' would cause λ to contain an occurrence of 3124 as witnessed by $jkx(s-1)$. Thus, it must be the case that λ' avoids the patterns in π_8 if λ does, as desired. □

Summing (2.12) over $1 \le j \le i-2$ yields the recurrence

$$\begin{aligned} a(n;i) &= a(n-1;i-1) + 2a(n-1;i) + a(n-3)\delta_{i=n-2} \\ &\quad + \sum_{j=1}^{\min(i,n-2)} (a(n-1;j) - a(n-2;j-1) - 2a(n-2;j)), \qquad 3 \le i \le n-1. \end{aligned} \tag{2.13}$$

Since $a(n;2) = a(n-1;1) + 2a(n-1;2)$, recurrence (2.13) is seen to hold for $i = 2$ and $n \ge 3$ as well, with $a(n;1) = \#S_{n-1}(231, 213) = 2^{n-2}$ and $a(n;n) = a(n-1)$.

Define the generating functions

$$A(x,y) = \sum_{n\ge1}\sum_{i=1}^{n} a(n;i)x^n y^i$$

and

$$A(x) = \sum_{n\ge1} a(n)x^n.$$

Note that $A(x) = A(x,1)$. The following lemma, valid for arbitrary $a(n;i)$, will be useful. Its proof is routine.

Lemma 2.19.

$$\sum_{n\ge1}\sum_{i=1}^{n}\sum_{j=1}^{i} a(n;j)x^n y^i = \frac{A(x,y) - yA(xy,1)}{1-y}.$$

Using (2.13) for $n \ge 3$ and Lemma 2.19 yields after several algebraic steps the functional equation

$$A(x,y) = xy(1-x)(1-x-xy) - \frac{x\left(x(y+2) + y^2 + y - 3\right)}{1-y}A(x,y)$$

$$+\frac{xy\left(x^2\left(1-y^2\right)+x\left(y^2+3y-1\right)-y\right)}{1-y}A(xy,1). \tag{2.14}$$

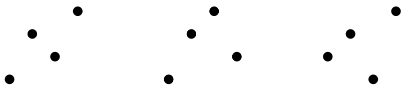
Taking $y=1-x$ in (2.14) implies

$$A(x-x^2,1)=\frac{x(1-x)^2}{1-3x+2x^2-x^3}\,,$$

which gives the generating function $\frac{2(1-4x)}{2-9x+4x^2-x\sqrt{1-4x}}$ for $1+A(x)$. Since $A(xy,1)=A(xy)$, substituting in (2.14) gives the bivariate generating function $A(x,y)$.

2.9. Class 9

$\pi_9=\{1324,1342,2314\}$



Let $d(n;i)$ denote the number of permutations of length n avoiding the three patterns in question and starting with the letter i and let $d(n)=\sum_{i=1}^{n}d(n;i)$. We have the following recurrence formula for the $d(n;i)$.

Lemma 2.20. *If $n\geq 2$, then $d(n;1)=2^{n-2}$ and $d(n;n)=d(n;n-1)=d(n-1)$, with $d(1)=d(1;1)=1$. If $n\geq 4$, then*

$$d(n;i)=2^{n-i-1}d(i-1)+\sum_{\ell=i+1}^{n}\sum_{j=1}^{i-1}d(\ell-1;j), \qquad 2\leq i\leq n-2. \tag{2.15}$$

Proof. That $d(n;n)=d(n;n-1)=d(n-1)$ is clear since neither n nor $n-1$ can start an occurrence of any pattern in π_9. Let $\mathcal{D}_{n,i}$ denote the subset of the permutations of length n enumerated by $d(n;i)$ and let $\mathcal{D}_n=\cup_{i=1}^{n}\mathcal{D}_{n,i}$. That $d(n;1)=2^{n-2}$ follows from the fact that members of $\mathcal{D}_{n,1}$ are synonymous with permutations of length $n-1$ avoiding both 213 and 231 (which are seen to number 2^{n-2}). We now assume $2\leq i\leq n-2$ and show (2.15). We first count members $\alpha\in\mathcal{D}_{n,i}$ in which all elements of $[i+1,n]$ occur to the left of all elements of $[i-1]$, i.e., α that may be decomposed as $\alpha=i\alpha_1\alpha_2$ where α_1 and α_2 are permutations of $[i+1,n]$ and $[i-1]$, respectively. Note that α_1 must avoid both 213 and 231, while α_2 need only avoid the original patterns in π_9. Thus, there are $2^{n-i-1}d(i-1)$ possibilities in this case.

Now assume that the leftmost element j of $[i-1]$ occurs earlier than some element of $[i+1,n]$ within $\alpha\in\mathcal{D}_{n,i}$. Then α must have the form $\alpha=i\alpha_1 j\alpha_2$, where $\alpha_1=n(n-1)\cdots(\ell+1)$ for some $i+1\leq\ell\leq n$. To see this, note first that $i+1$ must occur somewhere to the right of j, for if it occurred to the left of j, then some element x of $[i+1,n]$ occurring to the right of j within α implies that there would be an occurrence of 2314 as witnessed by the subsequence $i(i+1)jx$. Then $i+1$ occurring to the right of j implies any elements of $[i+2,n]$ to the left of j must be in descending order so as to avoid 1342. Finally, if $i<y<n$ lies to the left of j, then so must $y+1$, for otherwise there would be an occurrence of 2314 witnessed by the subsequence $iyj(y+1)$. Thus, α_1 has the stated form. Furthermore, it is seen that the letters in α_2 constitute a member of $\mathcal{D}_{\ell-1,j}$, upon arguing that $j\alpha_2$ avoids the patterns in π_9 if and only if $ij\alpha_2$ does. Conversely, any permutation of the form α above with the stated restrictions on its constituent parts is seen to avoid the patterns in π_9. Considering all possible ℓ and j, it follows that there are $\sum_{\ell=i+1}^{n}\sum_{j=1}^{i-1} d(\ell-1;j)$ members of $\mathcal{D}_{n,i}$ in which some element of $[i+1,n]$ occurs to the right of some element of $[i-1]$. Combining this case with the previous one yields (2.15). □

Let $v(n;y)=\sum_{i=1}^{n} d(n;i)y^i$. Multiplying both sides of (2.15) by y^n, and summing over $2\le i\le n-2$, implies

$$\begin{aligned}
v(n;y) &= 2^{n-2}y+(1+y)d(n-1)y^{n-1}+2^{n-1}\sum_{i=2}^{n-2} d(i-1)\left(\frac{y}{2}\right)^i\\
&+\sum_{i=2}^{n-1} y^i\sum_{\ell=i+1}^{n}\sum_{j=1}^{i-1} d(\ell-1;j)-y^{n-1}\sum_{j=1}^{n-2} d(n-1;j)\\
&=2^{n-2}y+(1+y)d(n-1)y^{n-1}+2^{n-1}\sum_{i=2}^{n-2} d(i-1)\left(\frac{y}{2}\right)^i\\
&+\frac{y}{1-y}\sum_{\ell=2}^{n-1}(v(\ell;y)-y^{\ell}v(\ell;1))-y^{n-1}(v(n-1;1)-v(n-2;1)),\ n\ge 3. \quad (2.16)
\end{aligned}$$

Let $v(x,y)=\sum_{n\ge1} v(n;y)x^n$. Then recurrence (2.16) implies

$$\begin{aligned}
v(x,y)-v(1;y)x-v(2;y)x^2 &= \frac{2x^3y}{1-2x}(1+v(xy,1))+x(1+y)(v(xy,1)-xy)\\
&+\frac{xy}{(1-x)(1-y)}(v(x,y)-v(xy,1))-x(v(xy,1)-xy)+x^2yv(xy,1),
\end{aligned}$$

which may be rewritten as

$$\begin{aligned}
&\left(1-\frac{xy}{(1-x)(1-y)}\right)v(x,y)\\
&=\frac{xy(1-x)}{1-2x}+\left(\frac{xy(1-x)}{1-2x}-\frac{xy}{(1-x)(1-y)}\right)v(xy,1). \quad (2.17)
\end{aligned}$$

To solve functional equation (2.17), we use the kernel method and let $y=1-x$ to obtain

$$v(x(1-x),1)=\frac{x(1-x)^2}{1-2x-x(1-x)^2}.$$

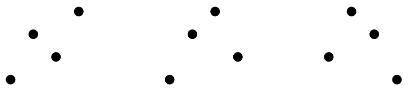
Replacing x with $\frac{1-\sqrt{1-4x}}{2}$ then implies

$$\begin{aligned}1+v(x,1) &= \frac{\sqrt{1-4x}}{\sqrt{1-4x}-\left(\frac{1-\sqrt{1-4x}}{2}\right)\left(\frac{1+\sqrt{1-4x}}{2}-x\right)} = \frac{2\sqrt{1-4x}}{(2-x)\sqrt{1-4x}-x} \\ &= \frac{2(1-4x)}{2-9x+4x^2-x\sqrt{1-4x}},\end{aligned}$$

as desired. (Note that replacing x with $\frac{1+\sqrt{1-4x}}{2}$ leads to a power series whose coefficients are not all positive integers.)

2.10. Class 10

$\pi_{10} = \{1324,\ 1432,\ 2431\}$

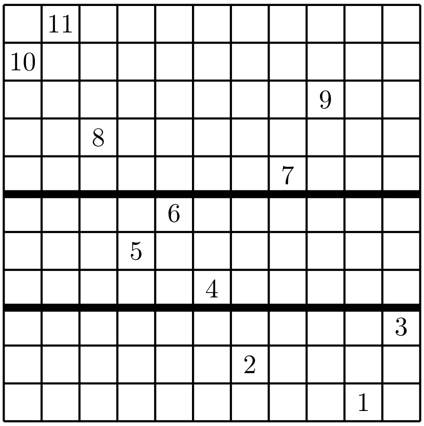


We will count the number $u(n)$ of length-n π_{10}-avoiders directly. The first 3 letters of each pattern in π_{10} form a 132 pattern. So, not surprisingly, 132-avoiders, counted by the Catalan numbers $C(n)$, will figure prominently. Every 132-avoider is a π_{10}-avoider. Let $\mathcal{V}(n)$ denote the set of length-n π_{10}-avoiders that do contain a 132, and set $v(n) = |\mathcal{V}(n)|$. Thus $u(n) = C(n)[\,\text{avoids } 132\,] + v(n)[\,\text{contains } 132\,]$.

Now suppose acb is a 132 pattern in $p \in \mathcal{V}(n)$. Then every entry of p after b is $< c$ (else a 1324 is present) and $> b$ (else a 1432 or 2431 is present), and the entries after b are increasing (else a 1432 is present). This stringent restriction implies that only one entry, say $b = b(p)$, is the "2" of a 132 in p. Note that if all entries after b in a permutation $p \in \mathcal{V}(n)$ are deleted, the resulting permutation, when standardized, is a *132-ender*, defined to be a π_{10}-avoider that contains a 132 and such that all its 132's end at its last entry.

Our strategy will be to start with a length-k 132-ender p and, viewing it as a permutation matrix, determine how many ways to append $n-k$ increasing entries all lying between the appropriate bounds without introducing a 1324 (we need not worry about introducing a 1432 or 2431 since these new entries are increasing). Then we sum over all k and p.

For a length-k 132-ender p, let b denote its last entry and c the smallest entry that serves as the "3" of a 132. Draw heavy lines above b and below c as in Figure 6. These heavy lines determine the *inner* and *outer* permutations of p, denoted $\text{Inn}(p)$ and $\text{Out}(p)$ respectively: standardize the subpermutation consisting of the entries between the 2 heavy lines to get $\text{Inn}(p)$ and standardize the entries outside the heavy lines to get $\text{Out}(p)$. The original permutation p can be recovered from $\text{Inn}(p)$ and $\text{Out}(p)$ because, as is easily seen, the entries between the heavy lines necessarily form a contiguous block (factor) of p that lies immediately to the left of the leftmost entry $< b$.



$\text{Out}(p) =$
$\text{St}(10\,11\,8\,2\,7\,9\,1\,3) =$
$7\,8\,5\,2\,4\,6\,1\,3$

$\text{Inn}(p) =$
$\text{St}(5\,6\,4) = 2\,3\,1$

Figure 6: A 132-ender, $10\,11\,8\,5\,6\,4\,2\,7\,9\,1\,3$, with $b = 3$ and $c = 7$

Any 132-avoider can be an inner permutation, and outer permutations are characterized by the properties (i) is a 132-ender, (ii) the smallest c that serves as the "3" of a 132 is $b + 1$ where b is the last entry. Let $\mathcal{A}_m$ denote the set of length-m permutations meeting these two conditions and set $w_0(m) = |\mathcal{A}_m|$.

The number of ways to append $n-k$ increasing entries as specified to a length-k 132-ender p depends only on the 132-avoider q:=Inn(p) and $t := n - k$. Let $w_1(q, t)$ denote this number. Then, refining the count by the length m of Inn(p), we have

$$v(n) = \sum_{k=3}^{n} \sum_{m=0}^{k-3} w_0(k - m) \sum_{q \in S_m(132)} w_1(q, n - k). \tag{2.18}$$

To evaluate the inner sum, we use a bijection from $S_m(132)$ to certain restricted growth sequences. Set $RG_m = \{a_1 a_2 \cdots a_{m+1} : a_1 = 1,\ 2 \le a_i \le a_{i-1} + 1 \text{ for } 2 \le i \le m + 1\}$. Thus $RG_0 = \{1\}$, $RG_1 = \{12\}$, $RG_2 = \{122, 123\}$, $RG_3 = \{1222, 1223, 1232, 1233, 1234\}$. There is an obvious correspondence between RG_m and primitive Dyck paths of semilength $m + 1$ via upstep heights; thus

$$UUDUUDDD \mapsto 1223.$$

The bijection $S_m(132) \to RG_m$ is illustrated in Figure 7 below. Given $q \in S_m(132)$, append $0\ m + 1$, and in the matrix diagram, draw a line segment from each nonterminal entry to the next larger entry. Set a_i = number of segments crossing the i-th interior horizontal line. To reverse the map, discard a_1 and set $b_i = a_{i+1} - 1,\ 1 \le i \le m$. Start with 1 and then, for $2 \le i \le m$, build the permutation by successively inserting i in the b_i-th currently available slot (right to left), where

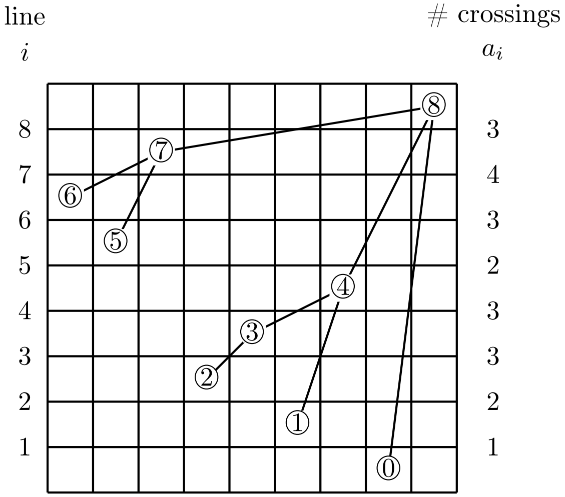


Figure 7: The bijection $S_m(132) \to RG_m$ with $m = 7$: $q = 6572314 \mapsto a = 12332343$

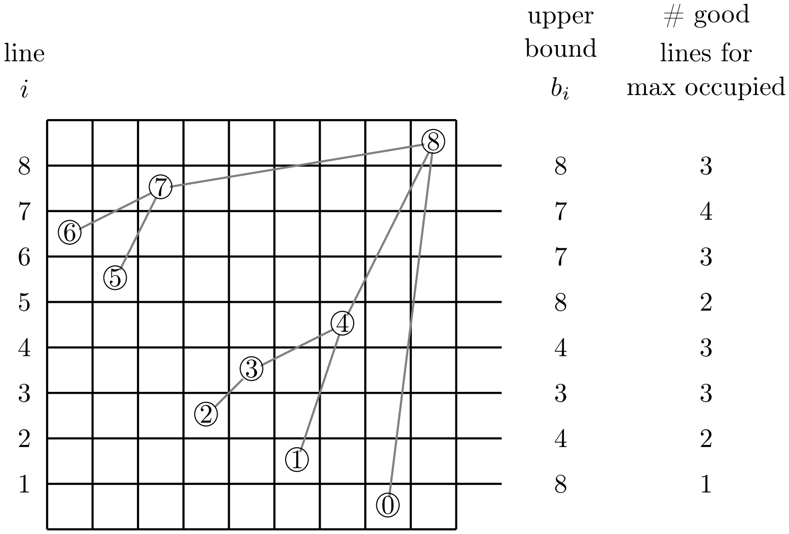


Figure 8: Counting the ways to append entries

available means "won't introduce a 132". Now let us count the number of ways to suitably append t increasing entries to a 132-avoider, using the permutation q of Figure 7 as an example. Since the new entries are increasing, this amounts to inserting t balls into $m+1$ boxes, the boxes being the protruding horizontal lines in Figure 8 above. But there are restrictions. The presence of a ball on line i means all balls lie on or below line b_i, where b_i is the next larger entry after $i-1$ (else a 1324 is present). This implies the upper bounds b_i listed in Figure 8.

Consequently, if i is the largest numbered line containing a ball, then the other $t-1$ balls are constrained to lie on a line j satisfying $j \le i$ and $b_j \ge b_i$. The number of such lines is given in the last column and this column *coincides with* the image $a \in RG_n$ of q under the preceding bijection. So the total number of ways to extend q is $\sum_{i=1}^{m+1} \binom{a_i+(t-1)-1}{t-1}$ using the familiar balls-in-boxes formula.

Hence, with $t := n-k$, the inner sum in (2.18) becomes

$$\begin{aligned}\sum_{q\in S_m(132)} w_1(q,t) &= \sum_{j=1}^{m+1} (\text{total number of } j\text{'s in } RG_m) \times \binom{j+(t-1)-1}{t-1} \\ &= \sum_{j=1}^{m+1} C(m+1-j, 2j-2)\binom{j+t-2}{t-1},\end{aligned} \tag{2.19}$$

where $C(n,k) = \frac{k+1}{2n+k+1}\binom{2n+k+1}{n} = \binom{2n+k}{n} - \binom{2n+k}{n-1}$ is the generalized Catalan number that counts nonnegative lattice paths of $n+k$ upsteps and n downsteps. The second equality in (2.19) is left as an exercise for the reader.

Next, we compute $w_0(m) = |\mathcal{A}_m|$. A 132-ender with consecutive bc arises by suitably appending an entry to a 132-avoider. As Figure 9 illustrates, you can append an entry on any non-top line except just below a LR min. There are

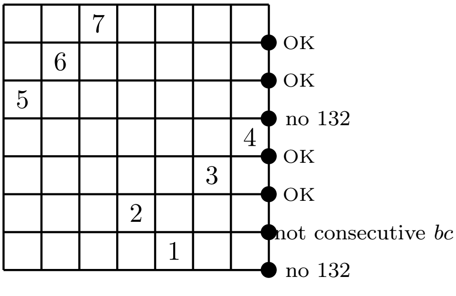


Figure 9: Constructing 132 enders with consecutive *bc*

$N(m-1,k)$ (Narayana number, $N(n,k) = \frac{1}{k}\binom{n-1}{k-1}\binom{n}{k-1}$) 132-avoiders of length $m-1$ with k LR minima, each of which contributes $m-1-k$ elements to $\mathcal{A}_m$. Hence $w_0(m) = \sum_{k=1}^{m-1}(m-1-k)N(m-1,k) = \binom{2m-3}{m-3}$.

So (2.18) becomes

$$\begin{aligned}
v(n) &= \sum_{k=3}^{n}\sum_{m=0}^{k-3}\binom{2(k-m)-3}{k-m-3}\sum_{j=1}^{m+1} C(m+1-j,2j-2)\binom{j+n-k-2}{n-k-1}\\
&= \sum_{k=3}^{n}\sum_{j=1}^{k-2}\binom{j+n-k-2}{n-k-1}\sum_{m=j-1}^{k-3}\binom{2(k-m)-3}{k-m-3}C(m+1-j,2j-2)\\
&= \sum_{k=3}^{n}\sum_{j=1}^{k-2}\binom{j+n-k-2}{n-k-1}\binom{2k-2}{k-j-2}.
\end{aligned} \tag{2.20}$$

The last equality follows from the identity

$$\sum_{i=1}^{n-k}\binom{2i+1}{i-1}C(n-k-i,2k) = \binom{2n+2}{n-k-1},$$

which has a simple combinatorial proof: it counts lattice paths of $n+k+3$ upsteps and $n-k-1$ downsteps, starting at the origin, by the x-coordinate, $2i+1$, of the last vertex at height 3. This vertex is the left endpoint of an upstep whose removal splits the path into a pair of paths counted by the summand on the left.

Now let us find the generating function for the sequence $u(n)$, that is, $U(x) = \sum_{n\geq 0} u(n)x^n$. By the above, we have

$$\begin{aligned}
U(x) &= \sum_{n\geq 3} v(n)x^n + \sum_{n\geq 0}\frac{1}{n+1}\binom{2n}{n}x^n\\
&= \sum_{n\geq 4}\left(\sum_{k=3}^{n-1}\sum_{j=1}^{k-2}\binom{j+n-k-2}{n-k-1}\binom{2k-2}{k-2-j}x^n\right)\\
&+ \sum_{n\geq 3}\binom{2n-2}{n-3}x^n + \frac{2}{1+\sqrt{1-4x}}\\
&= \sum_{k\geq 3}\sum_{j=1}^{k-2}\binom{2k-2}{k-2-j}\left(\sum_{n\geq k+1}\binom{j+n-k-2}{n-k-1}x^n\right)\\
&+ \frac{16x^3}{\sqrt{1-4x}(1+\sqrt{1-4x})^4} + \frac{2}{1+\sqrt{1-4x}}\\
&= \sum_{j\geq 1}\sum_{k\geq j+2}\binom{2k-2}{k-2-j}\frac{x^{k+1}}{(1-x)^j}\\
&+ \frac{16x^3}{\sqrt{1-4x}(1+\sqrt{1-4x})^4} + \frac{2}{1+\sqrt{1-4x}}\\
&= \sum_{j\geq 1}\frac{4^{j+1}x^{j+3}}{(1-x)^j\sqrt{1-4x}(1+\sqrt{1-4x})^{2j+2}}
\end{aligned}$$

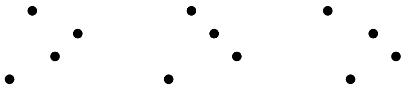
$$\begin{aligned}
&+\frac{16x^3}{\sqrt{1-4x}(1+\sqrt{1-4x})^4}+\frac{2}{1+\sqrt{1-4x}}\\
&=\frac{16x^4(1-2x+\sqrt{1-4x})}{\sqrt{1-4x}(1+\sqrt{1-4x})^4((1-x)\sqrt{1-4x}+1-5x+2x^2)}\\
&+\frac{16x^3}{\sqrt{1-4x}(1+\sqrt{1-4x})^4}+\frac{2}{1+\sqrt{1-4x}},
\end{aligned}$$

which implies

$$\sum_{n\geq 0} U(n)x^n=\frac{2(1-4x)}{2-9x+4x^2-x\sqrt{1-4x}}.$$

2.11. Class 11

We use the representative triple $\pi_{11}=\{1423, 1432, 4132\}$



Let $A_n=S_n(\pi_{11})$. Let $\sigma\in A_n$ with $n\geq 1$. Then σ can be decomposed as either which can be described as follows.

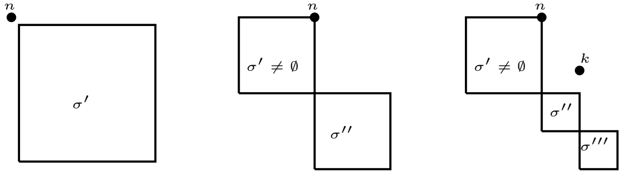


Figure 10: Decompositions

Lemma 2.21. *Let $n\geq 2$. A permutation π of $[n]$ avoids π_{11} if and only if either*

- *$\pi=n\pi'$ such that π' is a permutation of $[n-1]$ that avoids 132; or*
- *$\pi=\pi'n\pi''$ such that $\pi'>\pi''$, where π' is a non-empty permutation of $[n-j+1, n-1]$ that avoids π_{11} and π'' is a permutation of $[n-j]$ that avoids 132; or*
- *$\pi=\pi'n\pi''k\pi'''$ such that $\pi'>\pi''>\pi'''$, where $\pi'k$ is a permutation of $[n-j+1, n-1]$ avoiding π_{11} of length at least two such that $k\neq n-j+1$, π'' is a permutation of $[d+1, n-j]$ that avoids 132, and π''' is a permutation of $[d]$ that avoids 132.*

Let $A(x)=\sum_{n\geq 0}\#A_nx^n$. Using Lemma 2.21, we obtain

$$A(x)=1+xC(x)+x(A(x)-1)C(x)+x(A(x)-1-xA(x))C(x),$$

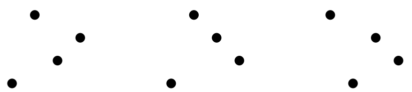
where $A(x)-1-xA(x)$ is the generating function for the number of permutations $\sigma=\sigma_1\cdots\sigma_n$ of A_n, $n\geq 2$, such that $\sigma_n\neq 1$. Thus, we can state the following result.

Theorem 2.22. *The generating function for the number of permutations of length n that avoid π_{11} is given by*

$$\frac{2(1-4x)}{2-9x+4x^2-x\sqrt{1-4x}}.$$

2.12. Class 12

We use the representative triple $\pi_{12}=\{2314, 2341, 3124\}$



Let $c(n;i,j)$ denote the number of permutations of length n avoiding the patterns in π_{12} in which the first letter is i and the second is j. For $n\geq 2$, define $c(n;i)=\sum_{j=1,j\neq i}^{n}c(n;i,j)$ and $c(n)=\sum_{i=1}^{n}c(n;i)$, with $c(1)=c(1;1)=1$. The values of the array $c(n;i,j)$ for $n\leq 3$ clearly are the same as those given above for $a(n;i,j)$.

If $n\geq 4$, then the array $c(n;i,j)$ satisfies the following relations.

Lemma 2.23. *If $1\leq i\leq n-1$, then $c(n;i,n)=c(n;n,i)=c(n-1;i)$, with $c(n;1,i)=c(n;i,i-1)=c(n-1;i-1)$ for $1<i\leq n$. If $2\leq i<j<n$, then $c(n;i,j)=0$. If $1\leq j<i-1<n-1$, then*

$$c(n;i,j)=c(n-1;i,j)+c(n-1;i-1,j-1)+\sum_{k=1}^{j-2}c(n-1;j,k). \tag{2.21}$$

Proof. Let $\mathcal{C}_n$ denote the subset of the permutations of length n avoiding the patterns in π_{12} and $\mathcal{C}_{n,i,j}$ the subset of $\mathcal{C}_n$ enumerated by $c(n;i,j)$. The first statement is clear since a letter n in either the first or second position is seen to be extraneous concerning the avoidance of the patterns in π_{12}, as is the letter 1 within members of $\mathcal{C}_{n,1,i}$ and the letter $i-1$ within members of $\mathcal{C}_{n,i,i-1}$. Permutations of length at least four starting with the letters i, j where $2\leq i<j<n$ must contain an occurrence of either 2314 or 2341, whence $c(n;i,j)=0$ in these cases. We now show (2.21). To do so, consider the third letter k within a member of $\mathcal{C}_{n,i,j}$ where $1\leq j<i-1<n-1$. The letter k cannot belong to $[i+1,n-1]$, for if it did, then there would be an occurrence of 2314 or 2341, and it cannot belong to $[j+1,i-1]$, for if it did, then 3124 would occur. Thus, we must have $k=n$ or $k\in[j-1]$, and the first two terms on the right-hand side of (2.21) are seen to account for the cases in which $k=n$ or $k=j-1$, respectively.

So let us assume $k \le j-2$. Given $\lambda \in \mathcal{C}_{n-1,j,k}$, let λ' be the permutation obtained from λ by writing the letter i before λ and increasing all elements of $[i, n-1]$ within λ by one. We will show that λ avoiding the patterns in π_{12} implies λ' does. Suppose, to the contrary, that λ' contains an occurrence of some pattern $\rho \in \pi_{12}$ and that ρ is realized within λ' by the subsequence $i\ell rs$. First assume $\rho = 3124$. Note that one may take $\ell \le k$ within an occurrence of ρ in this case, for if $\ell > k$, one may replace ℓ with k. Furthermore, observe that we must have $r > j$, for if not, then λ would contain ρ with the subsequence $j\ell r(s-1)$, which is impossible. Now consider the position of any $y \in [k+1, j-1]$. If y lies (i) to the right of s, (ii) between r and s, or (iii) to the left of r, then there would be an occurrence within λ of 2341, 2314, or 3124, respectively, as witnessed by the subsequences $jr(s-1)y$, $jry(s-1)$, or $jkyr$, with each scenario being impossible. This implies $\rho = 3124$ is not possible.

Now assume $\rho = 2314$. Note that $r > j$, for otherwise λ would contain 2314 with $j(\ell-1)r(s-1)$. But then $r > j$ implies λ' contains an occurrence of 3124 with $ijrs$, which is impossible by the preceding case. Finally, assume $\rho = 2341$. If $y \in [k+1, j-1]$, then λ would contain an occurrence of 2341, 2314, or 3124, respectively, as witnessed by the subsequences $j(\ell-1)(r-1)y$, $j(\ell-1)y(r-1)$, or $jky(\ell-1)$, depending on whether y lies (i) to the right of r, (ii) between ℓ and r, or (iii) to the left of ℓ. Thus, $\rho = 2341$ is also not possible, which implies λ' avoids the patterns in π_{12} if λ does, as desired. □

Note that (2.21) implies for $2 \le i \le n-1$,

$$\begin{aligned} c(n;i) &= c(n-1;i-1) + c(n-1;i) \\ &\quad + \sum_{j=3}^{i} (c(n-1;j) - c(n-2;j-1) - c(n-2;j)), \end{aligned} \tag{2.22}$$

with $c(n;n) = c(n;1) = c(n-1)$.

Define $C_n(v) = \sum_{i=1}^{n} c(n;i)v^{i-1}$. Multiplying both sides of (2.22) by v^{i-1}, and summing over $2 \le i \le n-1$, yields

$$\begin{aligned} C_n(v) &= (1+v^{n-1})C_{n-1}(1) + (1+v)C_{n-1}(v) - (1+v^{n-1})C_{n-2}(1) \\ &\quad + \frac{1}{1-v}(C_{n-1}(v) - C_{n-2}(1) - v^{n-1}C_{n-1}(1) + v^{n-1}C_{n-2}(1)) \\ &\quad - \frac{1+v}{1-v}(C_{n-2}(v) - C_{n-3}(1) - v^{n-2}C_{n-2}(1) + v^{n-2}C_{n-3}(1)) \\ &\quad - \frac{v-v^{n-1}}{1-v}C_{n-3}(1) - v^{n-2}(C_{n-2}(1) - C_{n-3}(1)), \qquad n \ge 3, \end{aligned}$$

with $C_0(v) = C_1(v) = 1$ and $C_2(v) = 1+v$.

Define $C(x,v) = \sum_{n \ge 0} C_n(v)x^n$. Multiplying both sides of the last recurrence by x^n, and summing over $n \ge 3$, we obtain

$$\frac{(1-x-xv)(1-x-v)}{1-v}C(x,v) = (1-x)^2 - vx + \frac{x(1-x)(1-x-v)}{1-v}C(x,1)$$

$$-\frac{vx(1-vx-2x+vx^2)}{1-v}C(xv,1).$$

Substituting $v=1-x$ in the preceding functional equation yields $C(x(1-x),1)=\frac{1-2x}{(1-x)^3-x^2}$, which implies

$$C(x,1)=\frac{2(1-4x)}{2-9x+4x^2-x\sqrt{1-4x}}.$$

References

[1] Denoncourt, H., Jones, B.C., The enumeration of maximally clustered permutations, *Ann. Comb.* 14 (2010), 65–84.

[2] Enumerations of specific permutation classes, Wikipedia.

[3] Heubach, S., Mansour, T., *Combinatorics of Compositions and Words*, CRC Press, Boca Raton, FL, 2009.

[4] Hou, Q.H., Mansour, T., Kernel method and systems of functional equations with several conditions, *J. Comput. Appl. Math.* 235:5 (2011), 1205–1212.

[5] Knuth, D.E., *The Art of Computer Programming*, 2nd edition, Addison Wesley, Reading, MA, 1973.

[6] Losonczy, J., Maximally clustered elements and Schubert varieties, *Ann. Comb.* 11 (2007), 195–212.

[7] Mansour, T., *Combinatorics of Set Partitions*, CRC Press, Boca Raton, FL, 2012.

[8] Simion, R., Schmidt, F.W., Restricted permutations, *European J. Combin.* 6 (1985), 383–406.

[9] Stankova, Z.E., Forbidden subsequences, *Discrete Math.* 132 (1994), 291–316.

[10] Stankova, Z.E., Classification of forbidden subsequences of length four, *European J. Combin.* 17 (1996), 501–517.

[11] West, J., Generating trees and the Catalan and Schröder numbers, *Discrete Math.* 146 (1995), 247–262.

Annales Mathematicae et Informaticae
47 (2017) pp. 75–83
http://ami.uni-eszterhazy.hu

Determining special roots of quaternion polynomials

Petroula Dospra[a], Dimitrios Poulakis[b]

[a]petroula.dospra@gmail.com
[b]Department of Mathematics
Aristotle University of Thessaloniki
Thessaloniki, Greece
poulakis@math.auth.gr

Submitted June 19, 2017 — Accepted August 8, 2017

Abstract

In this paper we determine the sets of spherical roots, real roots, isolated complex roots, pure imaginary quaternion roots and roots in $\mathbb{R} + \mathbb{R}\mathbf{j}$ and $\mathbb{R} + \mathbb{R}\mathbf{k}$ of a quaternion polynomial $Q(t)$ by corresponding these sets to the sets of real or complex roots of some real or complex polynomials determined by $Q(t)$. Thus, the counting and classifying methods for such polynomials can be used for the counting and classifying of the aforementioned roots of quaternion polynomials.

Keywords: Quaternion polynomial; Spherical Root; Isolated Root; Complex Root.

MSC: 12E15; 11R52; 16H05.

1. Introduction

The problem of counting and classifying the real/imaginary roots of a given real polynomial has been extensively studied. The classical Sturm's algorithm is an efficient method for determining the numbers of real roots of constant coefficient polynomials, but very inconvenient for those with symbolic coefficients. On the other hand, since the complete root classification of a parametric polynomial has been applied in studies of ordinary differential equations, of integral equations, of mechanics problems, and to real quantifier elimination (see [16]), several methods have been developed to fulfil this task [9, 17, 26, 27]. As far as we know, similar results do not exist for the roots of polynomials

$$Q(t) = a_0 t^n + a_1 t^{n-l} + \cdots + a_n$$

with coefficients $a_0, \ldots, a_n$ lying in the skew field of quaternions $\mathbb{H}$.

The fundamental theorem of algebra holds for quaternion polynomials [18, 5, 8, 14]. Algorithms for the computation of roots of a quaternion polynomial and its expression as a product of linear factors have been investigated in several papers [6, 7, 10, 11, 12, 13, 14, 19, 20, 21, 22, 24, 25]. Recently, a method for finding the pure imaginary quaternion roots of a quaternion polynomial was given [1]. Furthermore, necessary and sufficient conditions for a quaternion polynomial to have a special kind of root were obtained [4]. Note that a special class of space curves, the so-called Pythagorean hodograph curves, may be generated by quaternion polynomials [2, 3], and as it is has been studied in [3, Chapter 6], such a curve is generated by another curve of lower degree if and only if its associated quaternion polynomial has a complex root.

In this paper we determine the sets of spherical roots, real roots, isolated complex roots, pure imaginary quaternion roots and roots in $\mathbb{R} + \mathbb{R}\mathbf{j}$ and $\mathbb{R} + \mathbb{R}\mathbf{k}$ of a quaternion polynomial $Q(t)$ by defining a bijection from each of these sets onto the sets of real or complex roots of some real or complex polynomials which are determined by $Q(t)$. So the counting and classifying methods for real and complex polynomials can be used for counting and classifying the aforementioned roots of quaternion polynomials. Our method for the study of pure imaginary quaternion roots has the same initial point as Chapman's approach [1, Chapter 4, Section 1.3], but is quite different in the sequel, since we use simpler equations and we determine exactly the sets of spherical and isolated pure imaginary quaternion roots.

The paper is organized as follows. In Section 2, we recall basic facts about quaternions and quaternion polynomials. Section 3 is devoted to the study of pure imaginary quaternion roots of quaternion polynomials. The spherical roots, real roots, complex isolated roots, and roots in $\mathbb{R}+\mathbb{R}\mathbf{j}$ and $\mathbb{R}+\mathbb{R}\mathbf{k}$ are studied in Section 4. Finally, in Section 5, we illustrate our results with three examples.

2. Quaternions

Let $\mathbb{R}$ and $\mathbb{C}$ be the fields of real and complex numbers, respectively. We denote by $\mathbb{H}$ the skew field of real quaternions. Its elements are of the form $q = x_0 + x_1\mathbf{i} + x_2\mathbf{j} + x_3\mathbf{k}$, where $x_1, x_2, x_3, x_4 \in \mathbb{R}$, and $\mathbf{i}$, $\mathbf{j}$, $\mathbf{k}$ satisfy the following multiplication rules:

$$\mathbf{i}^2 = \mathbf{j}^2 = \mathbf{k}^2 = -1, \quad \mathbf{ij} = -\mathbf{ji} = \mathbf{k}, \quad \mathbf{jk} = -\mathbf{kj} = \mathbf{i}, \quad \mathbf{ki} = -\mathbf{ik} = \mathbf{j}.$$

The *conjugate* of q is defined as $\bar{q} = x_0 - x_1\mathbf{i} - x_2\mathbf{j} - x_3\mathbf{k}$. The *real* and the *imaginary part* of q are $\operatorname{Re} q = x_0$ and $\operatorname{Im} q = x_1\mathbf{i} + x_2\mathbf{j} + x_3\mathbf{k}$, respectively. If $\operatorname{Re} q = 0$, then q

is called *pure imaginary quaternion.* The *norm* $|q|$ of q is defined to be the quantity

$$|q| = \sqrt{q\bar{q}} = \sqrt{x_0^2 + x_1^2 + x_2^2 + x_3^2}.$$

Two quaternions q and q' are said to be *congruent* or *equivalent,* written $q \sim q'$, if there is $w \in \mathbb{H} \setminus \{0\}$ such that $q' = wqw^{-1}$. By [28], we have $q \sim q'$ if and only if $\operatorname{Re} q = \operatorname{Re} q'$ and $|q| = |q'|$. The *congruence class* of q is the set

$$[q] = \{q' \in \mathbb{H}/\ q' \sim q\} = \{q' \in \mathbb{H}/\ \operatorname{Re} q = \operatorname{Re} q',\ |q| = |q'|\}.$$

Note that every class $[q]$ contains exactly one complex number z and its conjugate $\bar{z}$, which are $x_0 \pm \mathbf{i}\sqrt{x_1^2 + x_2^2 + x_3^2}$.

Let $\mathbb{H}[t]$ be the polynomial ring in the variable t over $\mathbb{H}$. Every polynomial $f(t) \in \mathbb{H}[t]$ is written as $a_0t^n + a_1t^{n-l} + \cdots + a_n$ where n is an integer ≥ 0 and $a_0, \ldots, a_n \in \mathbb{H}$ with $a_0 \neq 0$. The addition and the multiplication of polynomials are defined in the same way as the commutative case, where the variable t is assumed to commute with quaternion coefficients [15, Chapter 5, Section 16]. For every $q \in \mathbb{H}$ we define the evaluation of $f(t)$ at q to be the element

$$f(q) = a_0q^n + a_1q^{n-l} + \cdots + a_n.$$

Note that it is not in general a ring homomorphism from $\mathbb{H}[t]$ to $\mathbb{H}$.

We say that a quaternion q is a zero or a root of $f(t)$ if $f(q) = 0$. The polynomial $B(t) \in \mathbb{H}[t]$ is called a *right factor* of $Q(t)$ if there exists $C(t) \in \mathbb{H}[t]$ such that $Q(t) = C(t)B(t)$. Note that q is a root of $f(t)$ if and only if $t - q$ is a right factor of $Q(t)$, i.e. there exists $g(t) \in \mathbb{H}[t]$ such that $f(t) = g(t)(t - q)$ [15, Proposition 16.2].

Let q be a root of $f(t)$. If q is not real and has the property that $f(z) = 0$ for all $z \in [q]$, then we will say that q generates a *spherical root.* For short, we will also say that q is, rather than generates, a spherical root. If q is real or does not generate a spherical zero, it is called an isolated root. If two elements of a class are roots of $f(t)$, then [10, Theorem 4] implies that all elements of this class are zeros of $f(t)$. Therefore, since every congruence class contains exactly one complex number z and its conjugate $\bar{z}$, the pairs of complex numbers $\{z, \bar{z}\}$ which are roots of $f(t)$ determine all spherical roots of $f(t)$.

3. Pure imaginary quaternion roots

In this section we determine the set of pure imaginary quaternion roots of a quaternion polynomial. Let $Q(t) = a_nt^n + a_{n-1}t^{n-1} + \cdots + a_0$. We set

$$g(t) = \sum_{m=0}^{\lfloor (n-1)/2 \rfloor} a_{2m+1}(-1)^m t^m \quad \text{and} \quad h(t) = \sum_{m=0}^{\lfloor n/2 \rfloor} a_{2m}(-1)^m t^m.$$

Let $g_i(t), h_i(t) \in \mathbb{R}[t]$ $(i = 1, 2, 3, 4)$ such that

$$g(t) = g_1(t) + g_2(t)\mathbf{i} + g_3(t)\mathbf{j} + g_4(t)\mathbf{k},$$

$$h(t) = h_1(t) + h_2(t)\mathbf{i} + h_3(t)\mathbf{j} + h_4(t)\mathbf{k}.$$

We denote by $E(t)$ the greatest common divisor of polynomials $g_1(t), \dots, g_4(t)$, $h_1(t), \dots, h_4(t)$. Further, we consider the polynomials

$$F(t) = (g_1(t)^2 + g_2(t)^2 + g_3(t)^2 + g_4(t)^2)t - (h_1(t)^2 + h_2(t)^2 + h_3(t)^2 + h_4(t)^2)$$

and

$$G(t) = g_1(t)h_1(t) + g_2(t)h_2(t) + g_3(t)h_3(t) + g_4(t)h_4(t).$$

Let $L(t) = \gcd(F(t), G(t))$. Note, that $E(t)$ divides $L(t)$.

Theorem 3.1. *Let $\mathcal{E}$ be the set of the positive roots of $E(t)$ and $\mathcal{L}$ the set of the positive roots of $L(t)$ which are not roots of $E(t)$. We denote by S and I the sets of distinct spherical and isolated pure imaginary quaternion roots of $Q(t)$, respectively. Then the maps*

$$\begin{aligned} \sigma\colon \mathcal{E} &\longrightarrow S, \\ N &\longmapsto \{x_1\mathbf{i} + x_2\mathbf{j} + x_3\mathbf{k}/\ x_1, x_2, x_3 \in \mathbb{R},\ x_1^2 + x_2^2 + x_3^2 = N\} \end{aligned}$$

and

$$\begin{aligned} \tau\colon \mathcal{L} &\longrightarrow I, \\ N &\longmapsto -g(N)^{-1}h(N) \end{aligned}$$

are bijective.

Proof. Let $x = x_1\mathbf{i} + x_2\mathbf{j} + x_3\mathbf{k}$, with $x \neq 0$, be a pure imaginary quaternion which is a root of $Q(t)$. Then we have $x^2 = -(x_1^2 + x_2^2 + x_3^2) = -|x|^2$ and setting $N = x_1^2 + x_2^2 + x_3^2$, we get $x^2 = -N$. Thus, the equality $Q(x) = 0$ implies $g(N)x + h(N) = 0$. Further, since $x \neq 0$, we have $g(N) = 0$ if and only if $h(N) = 0$.

Suppose that x defines a spherical root of $Q(t)$. It follows that every $y \in [x]$ is a root of $Q(t)$. Further, y is a pure imaginary quaternion with $|x| = |y|$, and so $y^2 = -|y|^2 = -|x|^2 = x^2$, whence $g(N)y = -h(N)$. Thus, for every $y \in [x]$ we have $g(N)y = -h(N)$, whence we deduce $g(N) = h(N) = 0$. Since N is a real number, we have $E(N) = 0$, and so $N \in \mathcal{E}$. Conversely, if $N \in \mathcal{E}$, then $g(N) = h(N) = 0$. Thus, for every pure imaginary quaternion x with $|x| = \sqrt{N}$, we have $Q(x) = g(N)x + h(N) = 0$. Hence, $[x]$ is a spherical root of $Q(t)$. Therefore, σ is a bijection.

We have that $E(t)$ divides $F(t)$ and $G(t)$, whence we have that $E(t)$ divides $L(t)$. Suppose now that $L(t) \neq E(t)$. Let x be an isolated pure imaginary quaternion root of $Q(t)$ with $g(N)h(N) \neq 0$. Then, we get $g(N)x + h(N) = 0$, whence we get $|g(N)|^2|x|^2 = |h(N)|^2$, and so, N is a root of $F(t)$. Furthermore, we have

$$x = -g(N)^{-1}h(N) = -\overline{g(N)}h(N)/|g(N)|^2$$

and, since $\operatorname{Re} x = 0$, we get $G(N) = 0$. Hence N is a real positive root of $L(t)$, and so $N \in \mathcal{L}$. Conversely, suppose that $N \in \mathcal{L}$. Set $x = -\overline{g(N)}h(N)/|g(N)|^2$.

Since $G(N) = 0$, we have $\operatorname{Re} x = 0$. Furthermore, N is a root of $F(t)$, and so, we get $N = |h(N)|^2/|g(N)|^2 = |x|^2$. Thus, x is a purely imaginary quaternion with $x^2 = -|x|^2 = -N$, and hence $Q(x) = g(N)x + h(N) = 0$. Therefore, τ is a bijection. □

Corollary 3.2. *The numbers of spherical roots of $Q(t)$ is equal to the number of positive roots of $E(t)$, and the number of isolated pure imaginary quaternion roots of $Q(t)$ is equal to the number of positive roots of $L(t)$ which are not roots of $E(t)$.*

Corollary 3.3. *Let $l_1, \dots, l_\nu$ be the positive real roots of $L(t)$ satisfying $g(l_i)h(l_i) \neq 0$ $(i = 1, \dots, \nu)$. Then, the quaternions $q_i = -g(l_i)^{-1}h(l_i)$ $(i = 1, \dots, \nu)$ are all the isolated pure quaternion roots of $Q(t)$.*

Proof. The map τ of Theorem 3.1 is a bijection. Further, the roots of $L(t)$ which are also roots of $E(t)$ are the roots ρ such that $g(\rho) = h(\rho) = 0$. The other roots satisfy $g(\rho) \neq 0$ and $h(\rho) \neq 0$. Thus, these roots ρ yield the isolated pure imaginary quaternion roots of $Q(t)$ which are the quaternions $-g(\rho)^{-1}h(\rho)$. □

4. Spherical roots and roots in $\mathbb{C}$, $\mathbb{R} + \mathbb{R}\mathbf{j}$ and $\mathbb{R} + \mathbb{R}\mathbf{k}$

In this section we study first the sets of spherical roots, complex isolated roots and real roots of a quaternion polynomial. Let $Q(t) \in \mathbb{H}[t] \setminus \mathbb{C}[t]$ be a monic polynomial of degree ≥ 1. Write $Q(t) = f_1(t) + f_2(t)\mathbf{i} + g_1(t)\mathbf{j} + g_2\mathbf{k}$, where $f_1(t), f_2(t), g_1(t), g_2(t) \in \mathbb{R}[t]$, and let $\Delta(t) = \gcd(f_1(t), f_2(t), g_1(t), g_2(t))$. Set $f(t) = f_1(t) + f_2(t)\mathbf{i}$, $g(t) = g_2(t) + g_1(t)\mathbf{i}$, $E(t) = \gcd(f(t), g(t))$ and $\Lambda(t) = E(t)/\Delta(t)$.

Theorem 4.1. *a) The set of real roots of $Q(t)$ coincides with the set of real roots of $\Delta(t)$.*
b) The spherical roots of $Q(t)$ are represented by the pairs of complex conjugate roots of $\Delta(t)$.
c) The set of isolated complex roots of $Q(t)$ is equal to the set of roots of $\Lambda(t)$.

Proof. a) Let $x \in \mathbb{R}$. Then we have $Q(x) = 0$ if and only if

$$f_1(x) = f_2(x) = g_1(x) = g_2(x) = 0$$

which is equivalent to $\Delta(x) = 0$. It follows that the set of real roots of $Q(t)$ is the same with the set of real roots of $\Delta(t)$.

b) We have $Q(t) = f(t) + \mathbf{k}g(t)$. Let $z \in \mathbb{C}$. We have $Q(z) = 0$ if and only if $f(z) + \mathbf{k}g(z) = 0$ which is equivalent to $f(z) = g(z) = 0$. Suppose now that $Q(t)$ has a spherical root q. Let z and $\bar{z}$ be the only complex numbers of the class of q. Then we have $Q(z) = Q(\bar{z}) = 0$, whence we get $f(z) = f(\bar{z}) = 0$ and $g(z) = g(\bar{z}) = 0$. It follows that the real polynomial $(t - z)(t - \bar{z})$ divides $f(z)$ and $g(z)$ and hence $(t - z)(t - \bar{z})$ divides the polynomials $f_1(t), f_2(t), g_1(t), g_2(t)$. Therefore z and $\bar{z}$ is a pair of conjugate complex root of $\Delta(t)$. Conversely, suppose z and $\bar{z}$ is a pair of conjugate complex root of $\Delta(t)$. It follows that z and $\bar{z}$ are roots of $f(t)$ and $g(t)$ and hence of $Q(t)$. Therefore, the class of z is a spherical root of $Q(t)$. So, there is a bijection between the spherical roots of $Q(t)$ and the pairs of complex conjugate roots of $\Delta(t)$.

c) Suppose that $C(t) \in \mathbb{C}[t]$. The polynomial $C(t)$ is a right factor of $Q(t)$ if and only if there is $A(t) \in \mathbb{H}[t] \setminus \mathbb{C}[t]$ such that $Q(t) = A(t)C(t)$. This happens if and only if $C(t)$ divides $f(t)$ and $g(t)$ which is equivalent to the fact that $C(t)$ divides $E(t)$. Thus, $C(t)$ is a right factor of $Q(t)$ if and only if $C(t)$ divides $E(t)$. In case where $C(t) \in \mathbb{R}[t]$, we similarly deduce that $C(t)$ is a right factor of $Q(t)$ if and only if $C(t)$ divides $\Delta(t)$. Thus, we have that $\Delta(t)$ divides $E(t)$ and the polynomial $\Lambda(t) = E(t)/\Delta(t)$ has no real factor. Suppose that z is a complex no real isolated root of $Q(t)$. Then its conjugate $\bar{z}$ is not a root of $Q(t)$ and so, $\bar{z}$ is not a root of $E(t)$. It follows that z is a root of $\Lambda(t)$. Conversely, suppose that z is a root of $\Lambda(t)$. If its conjugate $\bar{z}$ is also a root of $\Lambda(t)$, then $(t-z)(t-\bar{z})$ is a real factor of $\Lambda(t)$ which is a contradiction. Then, $\bar{z}$ is not a root of $E(t)$, and so, it is not a root of $Q(t)$. Hence, z is an isolated complex no real root of $Q(t)$. Thus, the complex no real isolated roots of $Q(t)$ are precisely the roots of $\Lambda(t)$. □

Next, we deal with the roots of $Q(t)$ in $\mathbb{R} + \mathbb{R}\mathbf{j}$ and $\mathbb{R} + \mathbb{R}\mathbf{k}$. Set $\bar{f}(t) = f_1(t) + g_1(t)\mathbf{j}$, $\bar{g}(t) = g_2(t) - f_2(t)\mathbf{j}$ and $\bar{E}(t) = \gcd(\bar{f}(t), \bar{g}(t))$. Further, we put $\tilde{f}(t) = f_1(t) + g_2(t)\mathbf{k}$, $\tilde{g}(t) = g_1(t) - f_2(t)\mathbf{k}$ and $\tilde{E}(t) = \gcd(\tilde{f}(t), \tilde{g}(t))$.

Theorem 4.2. *a) The set of roots of $Q(t)$ in $\mathbb{R} + \mathbb{R}\mathbf{j}$ is equal to the set of roots of $\bar{E}(t)$.*
b) The set of roots of $Q(t)$ in $\mathbb{R} + \mathbb{R}\mathbf{k}$ is equal to the set of roots of $\tilde{E}(t)$.

Proof. For (a), we write $Q(t) = \bar{f}(t) + \mathbf{k}\bar{g}(t)$. Let $x \in \mathbb{R} + \mathbb{R}\mathbf{j}$. Then $Q(x) = 0$ if and only if $\bar{f}(x) = \bar{g}(x) = 0$ which is equivalent to $\bar{E}(x) = 0$. For (b), we write $Q(t) = \tilde{f}(t) + \mathbf{i}\tilde{g}(t)$, and similarly we deduce the result. □

5. Examples

In this section we give three examples using the results of previous sections.

Example 5.1. By [1, Chapter 4, Example 1.4.1], the roots of the polynomial

$$P(t) = t^3 + (2 + \mathbf{k})t + \mathbf{i} - \mathbf{j}.$$

are the pure imaginary quaternions $\mathbf{j}$ and $\mathbf{i} + \mathbf{j}$.

We shall compute the pure quaternion roots of $P(t)$ using Corollary 3.3. We follow the notations of Section 3. We have $g(t) = -t + 2 + \mathbf{k}$ and $h(t) = \mathbf{i} - \mathbf{j}$. Thus, we get $g_1(t) = -t + 2$, $g_2(t) = g_3(t) = 0$, $g_4(t) = 1$ and $h_1(t) = h_4(t) = 0$, $h_2(t) = 1$, $h_3(t) = -1$. Hence, the greatest common divisor $E(t)$ of these polynomial is 1. It follows that $P(t)$ has not a spherical pure imaginary quaternion root. Next, we obtain the polynomials

$$F(t) = ((-t + 2)^2 + 1)t - (1 + 1) = t^3 - 4t^2 + 5t - 2 \quad \text{and} \quad G(t) = 0.$$

Then $L(t) = \gcd(F(t), G(t)) = t^3 - 4t^2 + 5t - 2$. The roots of $L(t)$ are 1 and 2. Next, we compute:

$$-g(1)^{-1}h(1) = -(1+\mathbf{k})^{-1}(\mathbf{i}-\mathbf{j}) = \mathbf{j}, \quad -g(2)^{-1}h(2) = -\mathbf{k}^{-1}(\mathbf{i}-\mathbf{j}) = \mathbf{j}+\mathbf{i}.$$

Hence, the isolated pure imaginary quaternion roots of $P(t)$ are $\mathbf{j}$ and $\mathbf{i}+\mathbf{j}$.

Example 5.2. According to [12], the polynomial

$$Q(t) = t^6 + \mathbf{j}\,t^5 + \mathbf{i}\,t^4 - t^2 - \mathbf{j}\,t - \mathbf{i}$$

has the four isolated roots

$$t_1 = 1, \quad t_2 = -1, \quad t_3 = \frac{1}{2}(1-\mathbf{i}-\mathbf{j}-\mathbf{k}), \quad t_4 = \frac{1}{2}(-1+\mathbf{i}-\mathbf{j}-\mathbf{k})$$

and the spherical root generated by $t_5 = \mathbf{i}$.

Following the notations of Section 4, we have:

$$f_1(t) = t^6 - t^2, \quad f_2(t) = t^4 - 1, \quad g_1(t) = t^5 - t, \quad g_2(t) = 0.$$

Then

$$\Delta(t) = \gcd(f_1(t), f_2(t), g_1(t), g_2(t)) = t^4 - 1.$$

By Theorem 4.1, we have that $Q(t)$ has the real roots ± 1 and one spherical root defined by $\mathbf{i}$.

Example 5.3. We shall compute the roots of the polynomial

$$R(t) = t^4 - (2+\mathbf{k})t^3 + (3+\mathbf{j}+2\mathbf{k})t^2 - 2(1+\mathbf{j}+\mathbf{k})t + 2(1+\mathbf{j}).$$

Following the notations of Section 4, we write $R(t) = f(t) + \mathbf{k}g(t)$, where

$$f(t) = t^4 - 2t^3 + 3t^2 - 2t + 2, \quad g(t) = -t^3 + (2+\mathbf{i})t^2 - 2(1+\mathbf{i})t + 2\mathbf{i}.$$

We have:

$$E(t) = \gcd(f(t), g(t)) = t^3 - (2+\mathbf{i})t^2 + 2(1+\mathbf{i})t - 2\mathbf{i}.$$

Next, we write $R(t) = f_1(t) + f_2(t)\mathbf{i} + g_1(t)\mathbf{j} + g_2(t)\mathbf{k}$, where

$$f_1(t) = t^4 - 2t^3 + 3t^2 - 2t + 2, \; f_2(t) = 0, \; g_1(t) = t^2 - 2t + 2, \; g_2(t) = -t^3 + 2t^2 - 2t.$$

Then, we have:

$$\Delta(t) = \gcd(f_1(t), f_2(t), g_1(t), g_2(t)) = t^2 - 2t + 2.$$

The roots of $\Delta(t)$ are the complex numbers $1 \pm \mathbf{i}$ which define a spherical root of $R(t)$. Further, we get:

$$\Lambda(t) = \frac{E(t)}{\Delta(t)} = t - \mathbf{i},$$

and so $\mathbf{i}$ is a complex isolated root of $R(t)$.

Next, we shall compute the pure quaternion roots of $R(t)$. Following the notations of Section 3, we compute:

$$g(t) = -2 + 2t - 2\mathbf{j} + (-2 + t)\mathbf{k}, \qquad h(t) = 2 - 3t + t^2 + (2 - t)\mathbf{j} - 2t\mathbf{k}.$$

It follows:

$$F(t) = 28t - 30t^2 + 11t^3 - 8 - t^4 \quad \text{and} \quad G(t) = -4t + 6t^2 - 2t^3.$$

We have:

$$L(t) = \gcd(F(t), G(t)) = t^2 - 3t + 2 = (t - 1)(t - 2).$$

We obtain $-g(1)^{-1}h(1) = \mathbf{i}$ and $-g(2)^{-1}h(2) = \mathbf{k} + \mathbf{i}$. Accordingly to [6], the polynomial $R(t)$ has not other roots. Thus, $R(t)$ has the isolated roots $\mathbf{i}$ and $\mathbf{i} + \mathbf{k}$ and the spherical root defined by $1 + \mathbf{i}$.

References

[1] Chapman, A., p-Central Subspaces of Central Simple Algebras Thesis, Bar-Ilan University (Israel), August 2013, arXiv:1406.0069v1.

[2] Choi, H. I., Lee, D. S., Moon, H. P., Clifford algebra, spin representation, and rational parametrization of curves and surfaces, *Advances in Computational Mathematics*, Vol. 17 (2002), 5–48.

[3] Dospra, P., Quaternion polynomials and rational rotation minimizing frame curves, PhD Thesis, Agricultural University of Athens, 2015.

[4] Dospra, P., Poulakis, D., Complex Roots of Quaternion Polynomials, Proceedings of the conference: "Applications of Computer Algebra (ACA 2015)", July 20-23, 2015, Kalamata, Greece, *Springer Proceedings in Mathematics and Statistics* (2017), 45–58.

[5] Eilenberg, S., Niven, I., The "fundamental theorem of algebra" for quaternions, *Bulletin of the American Mathematical Society*, Vol. 50 (1944), 246–248.

[6] Gentili, G., Struppa, D. C., On the multiplicity of zeroes of polynomials with quaternionic coefficients, *Milan Journal of Mathematics,* Vol. 76 (2008), 15–25.

[7] Gentili, G., Stoppato, C., Zeros of regular functions and polynomials of a quaternionic variable, *Michigan Mathematical Journal*, Vol. 56 (2008), 655–667.

[8] Gentili, G., Struppa, D. C., Vlacc, F., The fundamental theorem of algebra for Hamilton and Cayley numbers, *Math. Z.*, Vol. 259 (2008), 895–902.

[9] Gonzalez-Vega, L., Lombardi, H., Recio, T., Roy, M. F., Sturm - Habicht sequences, determinants and real roots of univariate polynomials, in: B.F. Caviness, J.R. Johnson (Eds.), Quantifier Elimination and Cylindrical Algebraic Decomposition, *Springer-Verlag*, 1998, 300–316.

[10] Gordon, B., Motzkin, T. S., On the zeros of polynomials over division rings, *Transactions of the American Mathematical Society*, Vol. 116 (1965), 218–226.

[11] Huang, L., So, W., Quadratic formulas for quaternions, *Applied Mathematics Letters*, Vol. 15, (2002), 533–540.

[12] Janovská, D., Opfer, G., A note on the computation of all zeros of simple quaternionic polynomials *SIAM Journal on Numerical Analysis*, Vol. 48 (2010), 244–256.

[13] Jia, Z., Cheng, X., Zhao, M., A new method for roots of monic quaternionic quadratic polynomial, *Computers and Mathematics with Applications*, Vol. 58 (2009), 1852–1858.

[14] Kalantari, B., Algorithms for quaternion polynomial root-finding, *Journal of Complexity*, Vol. 29 (2013), 302-322.

[15] Lam, T. Y., A First Course in Noncommutative Rings, 2nd edition, Springer, New York 2001.

[16] Liang, S., Jeffrey, D. J., An algorithm for computing the complete root classification of a parametric polynomial. Proceedings of 8th International Conference, AISC 2006 Beijing, China, September 20-22, 2006. LNCS 4120, (2006), 116–130.

[17] Liang, S., Jeffrey, D. J., *Automatic Computation of the Complete Root Classification for a Parametric Polynomial, Journal of Symbolic Computation*, Vol. 44, 10 (2009), 1487–1501.

[18] Niven, I., Equations in quaternions, *American Mathematical Monthly*, Vol. 48 (1941), 654–661.

[19] Opfer, G., Polynomials and Vandermonde Matrices over the field of Quaternions, *Electronic Transactions on Numerical Analysis*, Vol. 36 (2009), 9–16.

[20] Pogorui, A., Shapiro, M. V., On the structure of the set of zeros of quaternionic polynomials, *Complex Variables*, Vol. 49, 6 (2004), 379–389.

[21] Serôdio, R., Pereira, E., Vitória, J., Computing the zeros of quaternion polynomials, *Comput. Math. Appl.* Vol. 42 (2001), 1229–1237.

[22] Serôdio, R., Siu, L., Zeros of quaternion polynomials, *Appl. Math. Lett.* Vol. 14 (2001), 237–239.

[23] Sagraloff M., Mehlhorn, K., Computing real roots of real polynomials, *J. Symbolic Computation*, Vol. 73 (2016), 46–86.

[24] Topuridze, N., On Roots of Quaternion Polynomials, *Journal of Mathematical Sciences*, Vol. 160, 6 (2009), 843–855.

[25] Wedderburn, J. H. M., On division algebras, *Transactions of the American Mathematical Society*, Vol. 22 (1921), 129–135.

[26] Yang, L., Hou, X. R., Zeng, Z. B., A complete discrimination system for polynomials, *Sci. China* E, Vol. 39 (1996), 628–646.

[27] Yang, L., Recent Advances on Determining the Number of Real Roots of Parametric Polynomials, *J. Symbolic Computation*, Vol. 28 (1999), 225–242.

[28] Zhang, F., Quaternions and matrices of quaternions, *Linear Algebra Appl.*, Vol. 251, (1997), 21–57.

Annales Mathematicae et Informaticae
47 (2017) pp. 85–107
http://ami.uni-eszterhazy.hu

Generic lightlike submanifolds of an indefinite trans-Sasakian manifold with a non-metric ϕ-symmetric connection

Dae Ho Jin[a], **Jae Won Lee**[b]*

[a]Department of Mathematics Education
Dongguk University, Gyeongju 38066, Republic of Korea
jindh@dongguk.ac.kr

[b]Department of Mathematics Education and RINS
Gyeongsang National University, Jinju 52828, Republic of Korea
leejaew@gnu.ac.kr

Submitted September 20, 2016 — Accepted August 7, 2017

Abstract

Jin [13] introduced the notion of non-metric ϕ-symmetric connection on semi-Riemannian manifolds and studied lightlike hypersurfaces of an indefinite trans-Sasakian manifold with a non-metric ϕ-symmetric connection [12]. We study further the geometry of this subject. In this paper, we study generic lightlike submanifolds of an indefinite trans-Sasakian manifold with a non-metric ϕ-symmetric connection.

Keywords: non-metric ϕ-symmetric connection, generic lightlike submanifold, indefinite trans-Sasakian structure

MSC: 53C25, 53C40, 53C50

1. Introduction

The notion of non-metric ϕ-symmetric connection on indefinite almost contact manifolds or indefinite almost complex manifolds was introduced by Jin [12, 13]. Here we quote Jin's definition in itself as follows:

*Corresponding author

A linear connection $\bar{\nabla}$ on a semi-Riemannian manifold $(\bar{M}, \bar{g})$ is called a *non-metric ϕ-symmetric connection* if it and its torsion tensor $\bar{T}$ satisfy

$$(\bar{\nabla}_{\bar{X}}\bar{g})(\bar{Y}, \bar{Z}) = -\,\theta(\bar{Y})\phi(\bar{X}, \bar{Z}) - \theta(\bar{Z})\phi(\bar{X}, \bar{Y}), \tag{1.1}$$

$$\bar{T}(\bar{X}, \bar{Y}) = \theta(\bar{Y})J\bar{X} - \theta(\bar{X})J\bar{Y}, \tag{1.2}$$

where ϕ and J are tensor fields of types $(0, 2)$ and $(1, 1)$ respectively, and θ is an 1-form associated with a smooth vector field ζ by $\theta(\bar{X}) = \bar{g}(\bar{X}, \zeta)$. Throughout this paper, we denote by $\bar{X}$, $\bar{Y}$ and $\bar{Z}$ the smooth vector fields on $\bar{M}$.

In case $\phi = \bar{g}$ in (1.1), the above non-metric ϕ-symmetric connection reduces to so-called the quarter-symmetric non-metric connection. Quarter-symmetric non-metric connection was intorduced by S. Golad [7], and then, studied by many authors [2, 4, 19, 20]. In case $\phi = \bar{g}$ in (1.1) and $J = I$ in (1.2), the above non-metric ϕ-symmetric connection reduces to so-called the semi-symmetric non-metric connection. Semi-symmetric non-metric connection was intorduced by Ageshe and Chafle [1] and later studied by many geometers.

The notion of generic lightlike submanifolds on indefinite almost contact manifolds or indefinite almost complext manifolds was introduced by Jin-Lee [14] and later, studied by Duggal-Jin [6], Jin [9, 10] and Jin-Lee [16] and several geometers. We cite Jin-Lee's definition in itself as follows:

A lightlike submanifold M of an indefinite almost contact manifold $\bar{M}$ is said to be *generic* if there exists a screen distribution $S(TM)$ on M such that

$$J(S(TM)^{\perp}) \subset S(TM), \tag{1.3}$$

where $S(TM)^{\perp}$ is the orthogonal complement of $S(TM)$ in the tangent bundle $T\bar{M}$ on $\bar{M}$, *i.e.*, $T\bar{M} = S(TM) \oplus_{orth} S(TM)^{\perp}$. The geometry of generic lightlike submanifolds is an extension of that of lightlike hypersurfaces and half lightlike submanifolds of codimension 2. Much of its theory will be immediately generalized in a formal way to general lightlike submanifolds.

The notion of trans-Sasakian manifold, of type (α, β), was introduced by Oubina [18]. If $\bar{M}$ is a semi-Riemannian manifold with a trans-Sasakian structure of type (α, β), then $\bar{M}$ is called an *indefinite trans-Sasakian manifold of type* (α, β). Indefinite Sasakian, Kenmotsu and cosymplectic manifolds are important kinds of indefinite trans-Sasakian manifolds such that

$$\alpha = 1, \ \ \beta = 0; \quad \alpha = 0, \ \ \beta = 1; \quad \alpha = \beta = 0, \quad \text{respectively.}$$

In this paper, we study generic lightlike submanifolds M of an indefinite trans-Sasakian manifold $\bar{M} = (\bar{M}, J, \zeta, \theta, \bar{g})$ with a non-metric ϕ-symmetric connection, in which the tensor field J in (1.2) is identical with the indefinite almost contact structure tensor field J of $\bar{M}$, the tensor field ϕ in (1.1) is identical with the fundamental 2-form associated with J, that is,

$$\phi(\bar{X}, \bar{Y}) = \bar{g}(J\bar{X}, \bar{Y}), \tag{1.4}$$

and the 1-form θ, defined by (1.1) and (1.2), is identical with the structure 1-form θ of the indefinite almost contact metric structure $(J,\, \zeta,\, \theta,\, \bar{g})$ of $\bar{M}$.

Remark 1.1. Denote $\widetilde{\nabla}$ by the unique Levi-Civita connection of $(\bar{M}, \bar{g})$ with respect to the metric $\bar{g}$. It is known [13] that *a linear connection $\bar{\nabla}$ on $\bar{M}$ is non-metric ϕ-symmetric connection if and only if it satisfies*

$$\bar{\nabla}_{\bar{X}}\bar{Y} = \widetilde{\nabla}_{\bar{X}}\bar{Y} + \theta(\bar{Y})J\bar{X}. \tag{1.5}$$

For the rest of this paper, by the *non-metric ϕ-symmetric connection* we shall mean the *non-metric ϕ-symmetric connection defined by* (1.5).

2. Non-metric ϕ-symmetric connections

An odd-dimensional semi-Riemannian manifold $(\bar{M},\, \bar{g})$ is called an *indefinite trans-Sasakian manifold* if there exist (1) a structure set $\{J,\, \zeta,\, \theta,\, \bar{g}\}$, where J is a $(1,1)$-type tensor field, ζ is a vector field and θ is a 1-form such that

$$\begin{aligned} J^2\bar{X} = -\bar{X} + \theta(\bar{X})\zeta, \quad \theta(\zeta) = 1, \quad \theta(\bar{X}) = \epsilon\, \bar{g}(\bar{X}, \zeta),\\ \theta \circ J = 0, \qquad \bar{g}(J\bar{X},\, J\bar{Y}) = \bar{g}(\bar{X},\, \bar{Y}) - \epsilon\, \theta(\bar{X})\theta(\bar{Y}), \end{aligned} \tag{2.1}$$

(2) two smooth functions α and β, and a Levi-Civita connection $\widetilde{\nabla}$ such that

$$(\widetilde{\nabla}_{\bar{X}}J)\bar{Y} = \alpha\{\bar{g}(\bar{X}, \bar{Y})\zeta - \epsilon\, \theta(\bar{Y})\bar{X}\} + \beta\{\bar{g}(J\bar{X}, \bar{Y})\zeta - \epsilon\, \theta(\bar{Y})J\bar{X}\},$$

where ϵ denotes $\epsilon = 1$ or -1 according as ζ is spacelike or timelike respectively. $\{J,\, \zeta,\, \theta,\, \bar{g}\}$ is called an *indefinite trans-Sasakian structure of type* (α, β).

In the entire discussion of this article, we shall assume that the vector field ζ is a spacelike one, *i.e.*, $\epsilon = 1$, without loss of generality.

Let $\bar{\nabla}$ be a non-metric ϕ-symmetric connection on $(\bar{M}, \bar{g})$. Using (1.5) and the fact that $\theta \circ J = 0$, the equation in the item (2) is reduced to

$$\begin{aligned} (\bar{\nabla}_{\bar{X}}J)\bar{Y} \;=\;& \alpha\{\bar{g}(\bar{X}, \bar{Y})\zeta - \theta(\bar{Y})\bar{X}\}\\ &+\; \beta\{\bar{g}(J\bar{X}, \bar{Y})\zeta - \theta(\bar{Y})J\bar{X}\} + \theta(\bar{Y})\{\bar{X} - \theta(\bar{X})\zeta\}. \end{aligned} \tag{2.2}$$

Replacing $\bar{Y}$ by ζ to (2.2) and using $J\zeta = 0$ and $\theta(\bar{\nabla}_{\bar{X}}\zeta) = 0$, we obtain

$$\bar{\nabla}_{\bar{X}}\zeta \;= -(\alpha - 1)J\bar{X} + \beta\{\bar{X} - \theta(\bar{X})\zeta\}. \tag{2.3}$$

Let (M, g) be an m-dimensional lightlike submanifold of an indefinite trans-Sasakian manifold $(\bar{M}, \bar{g})$ of dimension $(m + n)$. Then the radical distribution $Rad(TM) = TM \cap TM^{\perp}$ on M is a subbundle of the tangent bundle TM and the normal bundle $TM^{\perp}$, of rank $r\,(1 \leq r \leq \min\{m,\, n\})$. In general, there exist two complementary non-degenerate distributions $S(TM)$ and $S(TM^{\perp})$ of $Rad(TM)$

in TM and $TM^{\perp}$ respectively, which are called the *screen distribution* and the *co-screen distribution* of M, such that

$$TM = Rad(TM) \oplus_{orth} S(TM), \quad TM^{\perp} = Rad(TM) \oplus_{orth} S(TM^{\perp}),$$

where $\oplus_{orth}$ denotes the orthogonal direct sum. Denote by $F(M)$ the algebra of smooth functions on M and by $\Gamma(E)$ the $F(M)$ module of smooth sections of a vector bundle E over M. Also denote by $(2.1)_i$ the i-th equation of (2.1). We use the same notations for any others. Let X, Y, Z and W be the vector fields on M, unless otherwise specified. We use the following range of indices:

$$i,\, j,\, k,\, ... \quad \in \{1, \,...\,, r\}, \qquad a,\, b,\, c,\, ... \quad \in \{r\, +\, 1, \,...\,, n\}.$$

Let $tr(TM)$ and $ltr(TM)$ be complementary vector bundles to TM in $T\bar{M}_{|M}$ and $TM^{\perp}$ in $S(TM)^{\perp}$ respectively and let $\{N_1, \cdots, N_r\}$ be a lightlike basis of $ltr(TM)_{|\mathcal{U}}$, where $\mathcal{U}$ is a coordinate neighborhood of M, such that

$$\bar{g}(N_i, \xi_j) = \delta_{ij}, \quad \bar{g}(N_i, N_j) = 0,$$

where $\{\xi_1, \cdots, \xi_r\}$ is a lightlike basis of $Rad(TM)_{|\mathcal{U}}$. Then we have

$$\begin{aligned} T\bar{M} &= TM \oplus tr(TM) = \{Rad(TM) \oplus tr(TM)\} \oplus_{orth} S(TM) \\ &= \{Rad(TM) \oplus ltr(TM)\} \oplus_{orth} S(TM) \oplus_{orth} S(TM^{\perp}). \end{aligned}$$

We say that a lightlike submanifold $M = (M, g, S(TM), S(TM^{\perp}))$ of $\bar{M}$ is

(1) *r-lightlike submanifold* if $1 \le r < \min\{m,\, n\}$;

(2) *co-isotropic submanifold* if $1 \le r = n < m$;

(3) *isotropic submanifold* if $1 \le r = m < n$;

(4) *totally lightlike submanifold* if $1 \le r = m = n$.

The above three classes (2)$\sim$(4) are particular cases of the class (1) as follows:

$$S(TM^{\perp}) = \{0\}, \qquad S(TM) = \{0\}, \qquad S(TM) = S(TM^{\perp}) = \{0\}$$

respectively. The geometry of r-lightlike submanifolds is more general than that of the other three types. For this reason, we consider only r-lightlike submanifolds M, with following local quasi-orthonormal field of frames of $\bar{M}$:

$$\{\xi_1, \cdots, \xi_r\,,\; N_1, \cdots, N_r\,,\; F_{r+1}, \cdots, F_m\,,\; E_{r+1}, \cdots, E_n\},$$

where $\{F_{r+1}, \cdots, F_m\}$ and $\{E_{r+1}, \cdots, E_n\}$ are orthonormal bases of $S(TM)$ and $S(TM^{\perp})$, respectively. Denote $\epsilon_a = \bar{g}(E_a, E_a)$. Then $\epsilon_a \delta_{ab} = \bar{g}(E_a, E_b)$.

In the sequel, we shall assume that ζ is tangent to M. Călin [5] proved that *if ζ is tangent to M, then it belongs to $S(TM)$* which we assumed in this paper. Let P

be the projection morphism of TM on $S(TM)$. Then the local Gauss-Weingarten formulae of M and $S(TM)$ are given respectively by

$$\begin{aligned}
\bar{\nabla}_X Y &= \nabla_X Y + \sum_{i=1}^{r} h_i^{\ell}(X,Y)N_i + \sum_{a=r+1}^{n} h_a^s(X,Y)E_a, && (2.4)\\
\bar{\nabla}_X N_i &= -A_{N_i}X + \sum_{j=1}^{r} \tau_{ij}(X)N_j + \sum_{a=r+1}^{n} \rho_{ia}(X)E_a, && (2.5)\\
\bar{\nabla}_X E_a &= -A_{E_a}X + \sum_{i=1}^{r} \lambda_{ai}(X)N_i + \sum_{b=r+1}^{n} \sigma_{ab}(X)E_b; && (2.6)\\
\nabla_X PY &= \nabla^*_X PY + \sum_{i=1}^{r} h_i^*(X,PY)\xi_i, && (2.7)\\
\nabla_X \xi_i &= -A^*_{\xi_i}X - \sum_{j=1}^{r} \tau_{ji}(X)\xi_j, && (2.8)
\end{aligned}$$

where ∇ and ∇^* are induced linear connections on M and $S(TM)$ respectively, h_i^{ℓ} and h_a^s are called the *local second fundamental forms* on M, h_i^* are called the *local second fundamental forms* on $S(TM)$. A_{N_i}, A_{E_a} and $A^*_{\xi_i}$ are called the *shape operators*, and τ_{ij}, ρ_{ia}, λ_{ai} and σ_{ab} are 1-forms.

Let M be a generic lightlike submanifold of $\bar{M}$. From (1.3) we show that $J(Rad(TM))$, $J(ltr(TM))$ and $J(S(TM^{\perp}))$ are subbundles of $S(TM)$. Thus there exist two non-degenerate almost complex distributions H_o and H with respect to J, *i.e.*, $J(H_o) = H_o$ and $J(H) = H$, such that

$$\begin{aligned}
S(TM) = \{J(Rad(TM)) \oplus J(ltr(TM))\}& \\
\oplus_{orth} J(S(TM^{\perp})) \oplus_{orth} H_o&,\\
H = Rad(TM) \oplus_{orth} J(Rad(TM)) \oplus_{orth} H_o&.
\end{aligned}$$

In this case, the tangent bundle TM on M is decomposed as follows:

$$TM = H \oplus J(ltr(TM)) \oplus_{orth} J(S(TM^{\perp})). \tag{2.9}$$

Consider local null vector fields U_i and V_i for each i, local non-null unit vector fields W_a for each a, and their 1-forms u_i, v_i and w_a defined by

$$\begin{aligned}
U_i &= -JN_i, & V_i &= -J\xi_i, & W_a &= -JE_a, && (2.10)\\
u_i(X) &= g(X,V_i), & v_i(X) &= g(X,U_i), & w_a(X) &= \epsilon_a g(X,W_a). && (2.11)
\end{aligned}$$

Denote by S the projection morphism of TM on H and by F the tensor field of type $(1,1)$ globally defined on M by $F = J \circ S$. Then JX is expressed as

$$JX = FX + \sum_{i=1}^{r} u_i(X)N_i + \sum_{a=r+1}^{n} w_a(X)E_a. \tag{2.12}$$

Applying J to (2.12) and using $(2.1)_1$ and (2.10), we have

$$F^2X = -X + \theta(X)\zeta + \sum_{i=1}^{r} u_i(X)U_i + \sum_{a=r+1}^{n} w_a(X)W_a. \tag{2.13}$$

In the following, we say that F is the *structure tensor field* on M.

3. Structure equations

Let $\bar{M}$ be an indefinite trans-Sasakian manifold with a non-metric ϕ-symmetric connection $\bar{\nabla}$. In the following, we shall assume that ζ is tangent to M. Călin [5] proved that if ζ is tangent to M, then it belongs to $S(TM)$ which we assumed in this paper. Using (1.1), (1.2), (1.4), (2.4) and (2.12), we see that

$$\begin{aligned}(\nabla_X g)(Y,Z) = \sum_{i=1}^{r}\{h_i^\ell(X,Y)\eta_i(Z) + h_i^\ell(X,Z)\eta_i(Y)\} \qquad (3.1)\\ - \theta(Y)\phi(X,Z) - \theta(Z)\phi(X,Y),\end{aligned}$$

$$T(X,Y) = \theta(Y)FX - \theta(X)FY, \tag{3.2}$$

$$h_i^\ell(X,Y) - h_i^\ell(Y,X) = \theta(Y)u_i(X) - \theta(X)u_i(Y), \tag{3.3}$$

$$h_a^s(X,Y) - h_a^s(Y,X) = \theta(Y)w_a(X) - \theta(X)w_a(Y), \tag{3.4}$$

$$\begin{aligned}&\phi(X,\xi_i) = u_i(X), \quad \phi(X,N_i) = v_i(X), \quad \phi(X,E_a) = w_a(X), \qquad (3.5)\\ &\phi(X,V_i) = 0, \quad \phi(X,U_i) = -\eta_i(X), \quad \phi(X,W_a) = 0,\end{aligned}$$

for all i and a, where η_i's are 1-forms such that $\eta_i(X) = \bar{g}(X,N_i)$.

From the facts that $h_i^\ell(X,Y) = \bar{g}(\bar{\nabla}_XY,\xi_i)$ and $\epsilon_a h_a^s(X,Y) = \bar{g}(\bar{\nabla}_XY,E_a)$, we know that h_i^ℓ and h_a^s are independent of the choice of $S(TM)$. Applying $\bar{\nabla}_X$ to $g(\xi_i,\xi_j) = 0$, $\bar{g}(\xi_i,E_a) = 0$, $\bar{g}(N_i,N_j) = 0$, $\bar{g}(N_i,E_a) = 0$ and $\bar{g}(E_a,E_b) = \epsilon\delta_{ab}$ by turns and using (1.1) and (2.4) ∼ (2.6), we obtain

$$\begin{aligned}&h_i^\ell(X,\xi_j) + h_j^\ell(X,\xi_i) = 0, \qquad h_a^s(X,\xi_i) = -\epsilon_a\lambda_{ai}(X),\\ &\eta_j(A_{N_i}X) + \eta_i(A_{N_j}X) = 0, \quad \eta_i(A_{E_a}X) = \epsilon_a\rho_{ia}(X), \qquad (3.6)\\ &\epsilon_b\sigma_{ab} + \epsilon_a\sigma_{ba} = 0; \quad h_i^\ell(X,\xi_i) = 0, \quad h_i^\ell(\xi_j,\xi_k) = 0, \quad A^*_{\xi_i}\xi_i = 0.\end{aligned}$$

Definition 3.1. We say that a lightlike submanifold M of $\bar{M}$ is

(1) *irrotational*[17] if $\bar{\nabla}_X\xi_i \in \Gamma(TM)$ for all $i \in \{1, \cdots, r\}$,

(2) *solenoidal* [15] if A_{W_a} and A_{N_i} are $S(TM)$-valued for all α and i.

From (2.4) and $(3.1)_2$, the item (1) is equivalent to

$$h_j^\ell(X,\xi_i) = 0, \qquad h_a^s(X,\xi_i) = \lambda_{ai}(X) = 0.$$

By using $(3.1)_4$, the item (2) is equivalent to

$$\eta_j(A_{N_i}X)=0, \qquad \rho_{ia}(X)=\eta_i(A_{E_a}X)=0.$$

The local second fundamental forms are related to their shape operators by

$$h_i^\ell(X,Y)=g(A^*_{\xi_i}X,Y)+\theta(Y)u_i(X)-\sum_{k=1}^r h_k^\ell(X,\xi_i)\eta_k(Y), \tag{3.7}$$

$$\epsilon_a h_a^s(X,Y)=g(A_{E_a}X,Y)+\theta(Y)w_a(X)-\sum_{k=1}^r \lambda_{ak}(X)\eta_k(Y), \tag{3.8}$$

$$h_i^*(X,PY)=g(A_{N_i}X,PY)+\theta(PY)v_i(X). \tag{3.9}$$

Replacing Y by ζ to (2.4) and using (2.3), (2.12), (3.7) and (3.8), we have

$$\nabla_X\zeta=-(\alpha-1)FX+\beta(X-\theta(X)\zeta), \tag{3.10}$$

$$\theta(A^*_{\xi_i}X)=-\alpha u_i(X), \qquad h_i^\ell(X,\zeta)=-(\alpha-1)u_i(X), \tag{3.11}$$

$$\theta(A_{E_a}X)=-\{\epsilon_a(\alpha-1)+1\}w_a(X), \tag{3.12}$$
$$h_a^s(X,\zeta)=-(\alpha-1)w_a(X).$$

Applying $\bar\nabla_X$ to $\bar g(\zeta,N_i)$ and using (2.3), (2.5) and (3.9), we have

$$\theta(A_{N_i}X)=-\alpha v_i(X)+\beta\eta_i(X), \tag{3.13}$$
$$h_i^*(X,\zeta)=-(\alpha-1)v_i(X)+\beta\eta_i(X).$$

Applying $\bar\nabla_X$ to $(2.10)_{1,2,3}$ and (2.12) by turns and using (2.2), (2.4) $\sim$ (2.8), (2.10) $\sim$ (2.12) and (3.7) $\sim$ (3.9), we have

$$\begin{aligned}
&h_j^\ell(X,U_i)=h_i^*(X,V_j), \qquad \epsilon_a h_i^*(X,W_a)=h_a^s(X,U_i),\\
&h_j^\ell(X,V_i)=h_i^\ell(X,V_j), \qquad \epsilon_a h_i^\ell(X,W_a)=h_a^s(X,V_i),\\
&\epsilon_b h_b^s(X,W_a)=\epsilon_a h_a^s(X,W_b),
\end{aligned} \tag{3.14}$$

$$\begin{aligned}
\nabla_XU_i \;=\; & F(A_{N_i}X)+\sum_{j=1}^r\tau_{ij}(X)U_j+\sum_{a=r+1}^n\rho_{ia}(X)W_a\\
&-\{\alpha\eta_i(X)+\beta v_i(X)\}\zeta,
\end{aligned} \tag{3.15}$$

$$\begin{aligned}
\nabla_XV_i \;=\; & F(A^*_{\xi_i}X)-\sum_{j=1}^r\tau_{ji}(X)V_j+\sum_{j=1}^r h_j^\ell(X,\xi_i)U_j\\
&-\sum_{a=r+1}^n\epsilon_a\lambda_{ai}(X)W_a-\beta u_i(X)\zeta,
\end{aligned} \tag{3.16}$$

$$\nabla_XW_a \;=\; F(A_{E_a}X)+\sum_{i=1}^r\lambda_{ai}(X)U_i+\sum_{b=r+1}^n\sigma_{ab}(X)W_b \tag{3.17}$$

$$\begin{aligned}
&\qquad - \beta w_a(X)\zeta,\\
(\nabla_X F)(Y) &= \sum_{i=1}^{r} u_i(Y)A_{N_i}X + \sum_{a=r+1}^{n} w_a(Y)A_{E_a}X \qquad (3.18)\\
&- \sum_{i=1}^{r} h_i^{\ell}(X,Y)U_i - \sum_{a=r+1}^{n} h_a^s(X,Y)W_a\\
&+ \{\alpha g(X,Y) + \beta\bar{g}(JX,Y) - \theta(X)\theta(Y)\}\zeta\\
&- (\alpha-1)\theta(Y)X - \beta\theta(Y)FX,\\
(\nabla_X u_i)(Y) &= -\sum_{j=1}^{r} u_j(Y)\tau_{ji}(X) - \sum_{a=r+1}^{n} w_a(Y)\lambda_{ai}(X) \qquad (3.19)\\
&- h_i^{\ell}(X,FY) - \beta\theta(Y)u_i(X),\\
(\nabla_X v_i)(Y) &= \sum_{j=1}^{r} v_j(Y)\tau_{ij}(X) + \sum_{a=r+1}^{n} \epsilon_a w_a(Y)\rho_{ia}(X) \qquad (3.20)\\
&+ \sum_{j=r+1}^{r} u_j(Y)\eta_i(A_{N_j}X) - g(A_{N_i}X, FY)\\
&- (\alpha-1)\theta(Y)\eta_i(X) - \beta\theta(Y)v_i(X).
\end{aligned}$$

Theorem 3.2. *There exist no generic lightlike submanifolds of an indefinite trans-Sasakian manifold with a non-metric ϕ-symmetric connection such that ζ is tangent to M and F satisfies the following equation*:

$$(\nabla_X F)Y = (\nabla_Y F)X, \qquad \forall\, X,\, Y \in \Gamma(TM).$$

Proof. Assume that $(\nabla_X F)Y - (\nabla_Y F)X = 0$. From (3.18) we obtain

$$\begin{aligned}
&\sum_{i=1}^{r}\{u_i(Y)A_{N_i}X - u_i(X)A_{N_i}Y\} \qquad (3.21)\\
&+ \sum_{a=r+1}^{n}\{w_a(Y)A_{E_a}X - w_a(X)A_{E_a}Y\} - 2\beta\bar{g}(X,JY)\zeta\\
&+ \{\theta(X)u_i(Y) - \theta(Y)u_i(X)\}U_i + \{\theta(X)w_a(Y) - \theta(Y)w_a(X)\}W_a\\
&+ (\alpha-1)\{\theta(X)Y - \theta(Y)X\} + \beta\{\theta(X)FY - \theta(Y)FX\} = 0.
\end{aligned}$$

Taking the scalar product with ζ and using $(3.12)_1$ and $(3.13)_1$, we have

$$\begin{aligned}
&\alpha\sum_{i=1}^{r}\{u_i(Y)v_i(X) - u_i(X)v_i(Y)\}\\
&= \beta\sum_{i=1}^{r}\{u_i(Y)\eta_i(X) - u_i(X)\eta_i(Y)\} - 2\beta\bar{g}(X,JY).
\end{aligned}$$

Taking $X = V_j$, $Y = U_j$ and $X = \xi_j$, $Y = U_j$ to this equation by turns, we obtain $\alpha = 0$ and $\beta = 0$, respectively. Taking $X = \xi_i$ to (3.21), we have

$$\theta(X)\xi_i + \sum_{j=1}^{r} u_j(X)A_{N_j}\xi_i + \sum_{a=r+1}^{n} w_a(X)A_{E_a}\xi_i = 0.$$

Taking $X = U_k$ and $X = W_b$ to this equation, we have

$$A_{N_k}\xi_i = 0, \qquad A_{E_b}\xi_i = 0.$$

Therefore, we get $\theta(X)\xi_i = 0$. It follows that $\theta(X) = 0$ for all $X \in \Gamma(TM)$. It is a contradiction to $\theta(\zeta) = 1$. Thus we have our theorem. □

Corollary 3.3. *There exist no generic lightlike submanifolds of an indefinite trans-Sasakian manifold with a non-metric ϕ-symmetric connection such that ζ is tangent to M and F is parallel with respect to the connection ∇.*

Theorem 3.4. *Let M be a generic lightlike submanifold of an indefinite trans-Sasakian manifold $\bar{M}$ with a non-metric ϕ-symmetric connection such that ζ is tangent to M. If U_is or V_is are parallel with respect to ∇, then $\alpha = \beta = 0$, i.e., $\bar{M}$ is an indefinite cosymplectic manifold. Furthermore, if U_i is parellel, M is solenoidal and $\tau_{ij} = 0$, if V_i is parallel, M is irrotational and $\tau_{ij} = 0$.*

Proof. (1) If U_i is parallel with respect to ∇, then, taking the scalar product with ζ, V_j, W_a, U_j and N_j to (3.15) such that $\nabla_X U_i = 0$ respectively, we get

$$\alpha = \beta = 0, \quad \tau_{ij} = 0, \quad \rho_{ia} = 0, \quad \eta_j(A_{N_i}X) = 0, \quad h_i^*(X, U_j) = 0. \tag{3.22}$$

As $\alpha = \beta = 0$, $\bar{M}$ is an indefinite cosymplectic manifold. As $\rho_{ia} = 0$ and $\eta_j(A_{N_i}X) = 0$, M is solenoidal.

(2) If V_i is parallel with respect to ∇, then, taking the scalar product with ζ, U_j, V_j, W_a and N_j to (3.16) with $\nabla_X V_i = 0$ respectively, we get

$$\beta = 0, \quad \tau_{ji} = 0, \quad h_j^\ell(X, \xi_i) = 0, \quad \lambda_{ai} = 0, \quad h_i^\ell(X, U_j) = 0. \tag{3.23}$$

As $h_j^\ell(X, \xi_i) = 0$ and $\lambda_{ai} = 0$, M is irrotational.

As $h_i^\ell(X, U_j) = 0$, we get $h_i^\ell(\zeta, U_j) = 0$. Taking $X = U_j$ and $Y = \zeta$ to (3.3), we get $h_i^\ell(U_j, \zeta) = \delta_{ij}$. On the other hand, replacing X by U to $(3.12)_1$, we have $h_i^\ell(U_j, \zeta) = -(\alpha - 1)\delta_{ij}$. It follows that $\alpha = 0$. Since $\alpha = \beta = 0$, $\bar{M}$ is an indefinite cosymplectic manifold. □

4. Recurrent and Lie recurrent structure tensors

Definition 4.1. The structure tensor field F of M is said to be

(1) *recurrent* [11] if there exists a 1-form ϖ on M such that

$$(\nabla_X F)Y = \varpi(X)FY,$$

(2) *Lie recurrent* [11] if there exists a 1-form ϑ on M such that

$$(\mathcal{L}_X F)Y = \vartheta(X)FY,$$

where $\mathcal{L}_X$ denotes the Lie derivative on M with respect to X, that is,

$$(\mathcal{L}_X F)Y = [X, FY] - F[X, Y]. \tag{4.1}$$

In case $\vartheta = 0$, *i.e.*, $\mathcal{L}_X F = 0$, we say that F is *Lie parallel.*

Theorem 4.2. *There exist no generic lightlike submanifolds of an indefinite trans-Sasakian manifold with a non-metric ϕ-symmetric connection such that ζ is tangent to M and the structure tensor field F is recurrent.*

Proof. Assume that F is recurrent. From (3.18), we obtain

$$\begin{aligned}\varpi(X)FY \;=\; & \sum_{i=1}^{r} u_i(Y)A_{N_i}X + \sum_{a=r+1}^{n} w_a(Y)A_{E_a}X \\ & - \sum_{i=1}^{r} h_i^\ell(X,Y)U_i - \sum_{a=r+1}^{n} h_a^s(X,Y)W_a \\ & + \{\alpha g(X,Y) + \beta\bar{g}(JX,Y) - \theta(X)\theta(Y)\}\zeta \\ & - (\alpha - 1)\theta(Y)X - \beta\theta(Y)FX.\end{aligned}$$

Replacing Y by ξ_j to this and using the fact that $F\xi_j = -V_j$, we get

$$\varpi(X)V_j = \sum_{k=1}^{r} h_k^\ell(X,\xi_j)U_k + \sum_{b=r+1}^{n} h_b^s(X,\xi_j)W_b - \beta u_j(X)\zeta.$$

Taking the scalar product with U_j, we get $\varpi = 0$. It follows that F is parallel with respect to ∇. By Corollary 3.2, we have our theorem. □

Theorem 4.3. *Let M be a generic lightlike submanifold of an indefinite trans-Sasakian manifold $\bar{M}$ with a non-metric ϕ-symmetric connection such that ζ is tangent to M and F is Lie recurrent. Then we have the following results:*

(1) *F is Lie parallel,*

(2) *the function α satisfies $\alpha = 0$,*

(3) *τ_{ij} and ρ_{ia} satisfy $\tau_{ij} \circ F = 0$ and $\rho_{ia} \circ F = 0$. Moreover,*

$$\tau_{ij}(X) = \sum_{k=1}^{r} u_k(X)g(A_{N_k}V_j, N_i) - \beta\theta(X)\delta_{ij}.$$

Proof. (1) Using (2.13), (3.2), (3.18), (4.1) and the fact that $\theta \circ F = 0$, we get

$$\begin{aligned} \vartheta(X)FY &= -\nabla_{FY}X + F\nabla_Y X \\ &+ \sum_{i=1}^{r} u_i(Y)A_{N_i}X + \sum_{a=r+1}^{n} w_a(Y)A_{E_a}X \\ &- \sum_{i=1}^{r}\{h_i^\ell(X,Y) - \theta(Y)u_i(X)\}U_i \\ &- \sum_{a=r+1}^{n}\{h_a^s(X,Y) - \theta(Y)w_a(X)\}W_a \\ &+ \alpha\{g(X,Y)\zeta - \theta(Y)X\} - \beta\theta(Y)FX. \end{aligned} \tag{4.2}$$

Replacing Y by ξ_j and then, Y by V_j to (4.2), respectively, we have

$$\begin{aligned} -\vartheta(X)V_j &= \nabla_{V_j}X + F\nabla_{\xi_j}X \\ &- \sum_{i=1}^{r} h_i^\ell(X,\xi_j)U_i - \sum_{a=r+1}^{n} h_a^s(X,\xi_j)W_a, \end{aligned} \tag{4.3}$$

$$\begin{aligned} \vartheta(X)\xi_j &= -\nabla_{\xi_j}X + F\nabla_{V_j}X + \alpha u_j(X)\zeta \\ &- \sum_{i=1}^{r} h_i^\ell(X,V_j)U_i - \sum_{a=r+1}^{n} h_a^s(X,V_j)W_a. \end{aligned} \tag{4.4}$$

Taking the scalar product with U_i to (4.3) and N_i to (4.4) respectively, we get

$$\begin{aligned} -\delta_{ij}\vartheta(X) &= g(\nabla_{V_j}X, U_i) - \bar{g}(\nabla_{\xi_j}X, N_i), \\ \delta_{ij}\vartheta(X) &= g(\nabla_{V_j}X, U_i) - \bar{g}(\nabla_{\xi_j}X, N_i). \end{aligned}$$

Comparing these two equations, we get $\vartheta = 0$. Thus F is Lie parallel.

(2) Taking the scalar product with ζ to (4.4), we get $g(\nabla_{\xi_j}X, \zeta) = \alpha u_j(X)$. Taking $X = U_i$ to this result and using (3.15), we obtain $\alpha = 0$.

(3) Taking the scalar product with N_i to (4.3) such that $X = W_a$ and using (3.4), $(3.6)_4$, (3.8) and (3.17), we get $h_a^s(U_i, V_j) = \rho_{ia}(\xi_j)$. On the other hand, taking the scalar product with W_a to (4.4) such that $X = U_i$ and using (3.15), we have $h_a^s(U_i, V_j) = -\rho_{ia}(\xi_j)$. Thus $\rho_{ia}(\xi_j) = 0$ and $h_a^s(U_i, V_j) = 0$.

Taking the scalar product with U_i to (4.3) such that $X = W_a$ and using (3.4), $(3.6)_{2,4}$, (3.8) and (3.17), we get $\epsilon_a\rho_{ia}(V_j) = \lambda_{aj}(U_i)$. On the other hand, taking the scalar product with W_a to (4.3) such that $X = U_i$ and using $(3.1)_2$ and (3.15), we get $\epsilon_a\rho_{ia}(V_j) = -\lambda_{aj}(U_i)$. Thus $\rho_{ia}(V_j) = \lambda_{aj}(U_i) = 0$.

Taking the scalar product with V_i to (4.3) such that $X = W_a$ and using (3.4), $(3.6)_2$, $(3.14)_4$ and (3.17), we obtain $\lambda_{ai}(V_j) = -\lambda_{aj}(V_i)$. On the other hand, taking the scalar product with W_a to (4.3) such that $X = V_i$ and using $(3.6)_2$ and (3.16), we have $\lambda_{ai}(V_j) = \lambda_{aj}(V_i)$. Thus we obtain $\lambda_{ai}(V_j) = 0$.

Taking the scalar product with W_a to (4.3) such that $X = \xi_i$ and using (2.8), (3.3), $(3.6)_2$ and (3.7), we get $h^\ell_i(V_j, W_a) = \lambda_{ai}(\xi_j)$. On the other hand, taking the scalar product with V_i to (4.4) such that $X = W_a$ and using (3.3) and (3.17), we get $h^\ell_i(V_j, W_a) = -\lambda_{ai}(\xi_j)$. Thus $\lambda_{ai}(\xi_j) = 0$ and $h^\ell_i(V_j, W_a) = 0$.

Summarizing the above results, we obtain

$$\begin{aligned}&\rho_{ia}(\xi_j) = 0, \quad \rho_{ia}(V_j) = 0, \quad \lambda_{ai}(U_j) = 0, \quad \lambda_{ai}(V_j) = 0, \quad \lambda_{ai}(\xi_j) = 0, \\ &h^s_a(U_i, V_j) = h^\ell_j(U_i, W_a) = 0, \qquad h^\ell_i(V_j, W_a) = h^s_a(V_j, V_i) = 0.\end{aligned} \tag{4.5}$$

Taking the scalar product with N_i to (4.2) and using $(3.1)_4$, we have

$$\begin{aligned}&-\bar{g}(\nabla_{FY}X, N_i) + g(\nabla_Y X, U_i) - \beta\theta(Y)v_i(X) \\ &+\sum_{k=1}^{r} u_k(Y)\bar{g}(A_{N_k}X, N_i) + \sum_{a=r+1}^{n} \epsilon_a w_a(Y)\rho_{ia}(X) = 0.\end{aligned} \tag{4.6}$$

Replacing X by V_j to (4.6) and using (3.7), (3.16) and $(4.5)_2$, we have

$$h^\ell_j(FX, U_i) + \tau_{ij}(X) + \beta\theta(X)\delta_{ij} = \sum_{k=1}^{r} u_k(X)\bar{g}(A_{N_k}V_j, N_i). \tag{4.7}$$

Replacing X by ξ_j to (4.6) and using (2.8), (3.7) and $(4.5)_1$, we have

$$h^\ell_j(X, U_i) = \sum_{k=1}^{r} u_k(X)\bar{g}(A_{N_k}\xi_j, N_i) + \tau_{ij}(FX). \tag{4.8}$$

Taking $X = U_k$ to this equation and using $(3.14)_1$, we have

$$h^*_i(U_k, V_j) = \bar{g}(A_{N_k}\xi_j, N_i). \tag{4.9}$$

Taking $X = U_i$ to (4.2) and using (2.13), (3.3), (3.4) and (3.15), we get

$$\begin{aligned}&\sum_{k=1}^{r} u_k(Y)A_{N_k}U_i + \sum_{a=r+1}^{n} w_a(Y)A_{E_a}U_i - A_{N_i}Y \\ &- F(A_{N_i}FY) - \sum_{j=1}^{r}\tau_{ij}(FY)U_j - \sum_{a=r+1}^{n}\rho_{ia}(FY)W_a = 0.\end{aligned} \tag{4.10}$$

Taking the scalar product with V_j to (4.10) and using (3.8), (3.9), $(3.14)_1$, $(4.5)_6$ and (4.9), we get

$$h^\ell_j(X, U_i) = -\sum_{k=1}^{r} u_k(X)\bar{g}(A_{N_k}\xi_j, N_i) - \tau_{ij}(FX).$$

Comparing this equation with (4.8), we obtain

$$\tau_{ij}(FX) + \sum_{k=1}^{r} u_k(X)\bar{g}(A_{N_k}\xi_j, N_i) = 0.$$

Replacing X by U_h to this equation, we have $\bar{g}(A_{N_k}\xi_j, N_i) = 0$. Therefore,

$$\tau_{ij}(FX) = 0, \qquad h_j^\ell(X, U_i) = 0. \tag{4.11}$$

Taking $X = FY$ to $(4.11)_2$, we get $h_j^\ell(FX, U_i) = 0$. Thus (4.7) is reduced to

$$\tau_{ij}(X) = \sum_{k=1}^{r} u_k(X)\bar{g}(A_{N_k}V_j, N_i) - \beta\theta(X)\delta_{ij}.$$

Taking the scalar product with U_j to (4.10) such that $Y = W_a$ and using (3.4), (3.8), (3.9) and $(3.14)_2$, we have

$$h_i^*(W_a, U_j) = \epsilon_a h_a^s(U_i, U_j) = \epsilon_a h_a^s(U_j, U_i) = h_i^*(U_j, W_a). \tag{4.12}$$

Taking the scalar product with W_a to (4.10), we have

$$\begin{aligned}\epsilon_a\rho_{ia}(FY) &= -h_i^*(Y, W_a)\\ &\quad + \sum_{k=1}^{r} u_k(Y)h_k^*(U_i, W_a) + \sum_{b=r+1}^{n} \epsilon_b w_b(Y)h_b^s(U_i, W_a).\end{aligned}$$

Taking the scalar product with U_i to (4.2) and then, taking $X = W_a$ and using (3.4), $(3.6)_4$, (3.8), (3.9), $(3.14)_2$, (3.17) and (4.12), we obtain

$$\begin{aligned}\epsilon_a\rho_{ia}(FY) &= h_i^*(Y, W_a)\\ &\quad - \sum_{k=1}^{r} u_k(Y)h_k^*(U_i, W_a) - \sum_{b=r+1}^{n} \epsilon_b w_b(Y)h_b^s(U_i, W_a).\end{aligned}$$

Comparing the last two equations, we obtain $\rho_{ia}(FY) = 0$. □

5. Indefinite generalized Sasakian space forms

Definition 5.1. An indefinite trans-Sasakian manifold $\bar{M}$ is said to be a *indefinite generalized Sasakian space form* and denote it by $\bar{M}(f_1, f_2, f_3)$ if there exist three smooth functions f_1, f_2 and f_3 on $\bar{M}$ such that

$$\begin{aligned}\widetilde{R}(\bar{X}, \bar{Y})\bar{Z} &= f_1\{\bar{g}(\bar{Y}, \bar{Z})\bar{X} - \bar{g}(\bar{X}, \bar{Z})\bar{Y}\}\\ &\quad + f_2\{\bar{g}(\bar{X}, J\bar{Z})J\bar{Y} - \bar{g}(\bar{Y}, J\bar{Z})J\bar{X} + 2\bar{g}(\bar{X}, J\bar{Y})J\bar{Z}\}\\ &\quad + f_3\{\theta(\bar{X})\theta(\bar{Z})\bar{Y} - \theta(\bar{Y})\theta(\bar{Z})\bar{X}\\ &\qquad + \bar{g}(\bar{X}, \bar{Z})\theta(\bar{Y})\zeta - \bar{g}(\bar{Y}, \bar{Z})\theta(\bar{X})\zeta\},\end{aligned} \tag{5.1}$$

where $\widetilde{R}$ is the curvature tensor of the Levi-Civita connection $\bar{\nabla}$.

The notion of generalized Sasakian space form was introduced by Alegre *et. al.* [3], while the indefinite generalized Sasakian space forms were introduced by Jin [8]. Sasakian space form, Kenmotsu space form and cosymplectic space form are important kinds of generalized Sasakian space forms such that

$$f_1 = \tfrac{c+3}{4}, f_2 = f_3 = \tfrac{c-1}{4}; \quad f_1 = \tfrac{c-3}{4}, f_2 = f_3 = \tfrac{c+1}{4}; \quad f_1 = f_2 = f_3 = \tfrac{c}{4}$$

respectively, where c is a constant J-sectional curvature of each space forms.

Denote by $\bar{R}$ the curvature tensors of the non-metric ϕ-symmetric connection $\bar{\nabla}$ on $\bar{M}$. By directed calculations from (1.2), (1.5) and (2.1), we see that

$$\begin{aligned}\bar{R}(\bar{X},\bar{Y})\bar{Z} = {} & \widetilde{R}(\bar{X},\bar{Y})\bar{Z} + (\bar{\nabla}_{\bar{X}}\theta)(\bar{Z})J\bar{Y} - (\bar{\nabla}_{\bar{Y}}\theta)(\bar{Z})J\bar{X} \\ & - \theta(\bar{Z})\{\alpha[\theta(\bar{Y})\bar{X} - \theta(\bar{X})\bar{Y}] + \beta[\theta(\bar{Y})J\bar{X} - \theta(\bar{X})J\bar{Y}] \\ & + 2\beta\bar{g}(X, JY)\zeta\}.\end{aligned} \tag{5.2}$$

Denote by R and R^* the curvature tensors of the induced linear connections ∇ and ∇^* on M and $S(TM)$ respectively. Using the Gauss-Weingarten formulae, we obtain Gauss-Codazzi equations for M and $S(TM)$ respectively:

$$\begin{aligned}\bar{R}(X,Y)Z = {} & R(X,Y)Z \\ & + \sum_{i=1}^{r}\{h_i^{\ell}(X,Z)A_{N_i}Y - h_i^{\ell}(Y,Z)A_{N_i}X\} \\ & + \sum_{a=r+1}^{n}\{h_a^{s}(X,Z)A_{E_a}Y - h_a^{s}(Y,Z)A_{E_a}X\} \\ & + \sum_{i=1}^{r}\{(\nabla_X h_i^{\ell})(Y,Z) - (\nabla_Y h_i^{\ell})(X,Z) \\ & + \sum_{j=1}^{r}[\tau_{ji}(X)h_j^{\ell}(Y,Z) - \tau_{ji}(Y)h_j^{\ell}(X,Z)] \\ & + \sum_{a=r+1}^{n}[\lambda_{ai}(X)h_a^{s}(Y,Z) - \lambda_{ai}(Y)h_a^{s}(X,Z)] \\ & - \theta(X)h_i^{\ell}(FY,Z) + \theta(Y)h_i^{\ell}(FX,Z)\}N_i \\ & + \sum_{a=r+1}^{n}\{(\nabla_X h_a^{s})(Y,Z) - (\nabla_Y h_a^{s})(X,Z) \\ & + \sum_{i=1}^{r}[\rho_{ia}(X)h_i^{\ell}(Y,Z) - \rho_{ia}(Y)h_i^{\ell}(X,Z)] \\ & + \sum_{b=r+1}^{n}[\sigma_{ba}(X)h_b^{s}(Y,Z) - \sigma_{ba}(Y)h_b^{s}(X,Z)] \\ & - \theta(X)h_a^{s}(FY,Z) + \theta(Y)h_a^{s}(FX,Z)\}E_a,\end{aligned} \tag{5.3}$$

$$\begin{aligned}R(X,Y)PZ = {} & R^*(X,Y)PZ \\ & + \sum_{i=1}^{r}\{h_i^*(X,PZ)A^*_{\xi_i}Y - h_i^*(Y,PZ)A_{\xi_i}X\}\end{aligned} \tag{5.4}$$

$$
\begin{aligned}
&+\sum_{i=1}^{r}\{(\nabla_X h_i^*)(Y,PZ)-(\nabla_Y h_i^*)(X,PZ)\\
&\qquad+\sum_{k=1}^{r}[\tau_{ik}(Y)h_k^*(X,PZ)-\tau_{ik}(X)h_k^*(Y,PZ)]\\
&\qquad-\theta(X)h_i^*(FY,PZ)+\theta(Y)h_i^*(FX,PZ)\}\xi_i,
\end{aligned}
$$

Taking the scalar product with ξ_i and N_i to (5.2) by turns and then, substituting (5.3) and (5.1) and using $(3.6)_4$ and (5.4), we get

$$
\begin{aligned}
&(\nabla_X h_i^\ell)(Y,Z)-(\nabla_Y h_i^\ell)(X,Z)\\
&+\sum_{j=1}^{r}\{\tau_{ji}(X)h_j^\ell(Y,Z)-\tau_{ji}(Y)h_j^\ell(X,Z)\}\\
&+\sum_{a=r+1}^{n}\{\lambda_{ai}(X)h_a^s(Y,Z)-\lambda_{ai}(Y)h_a^s(X,Z)\}\\
&-\theta(X)h_i^\ell(FY,Z)+\theta(Y)h_i^\ell(FX,Z)\\
&-(\bar\nabla_X\theta)(Z)u_i(Y)+(\bar\nabla_Y\theta)(Z)u_i(X)\\
&+\beta\theta(Z)\{\theta(Y)u_i(X)-\theta(X)u_i(Y)\}\\
&=\ f_2\{u_i(Y)\bar g(X,JZ)-u_i(X)\bar g(Y,JZ)+2u_i(Z)\bar g(X,JY)\},
\end{aligned}
\tag{5.5}
$$

$$
\begin{aligned}
&(\nabla_X h_i^*)(Y,\,PZ)-(\nabla_Y h_i^*)(X,\,PZ)\\
&-\sum_{j=1}^{r}\{\tau_{ij}(X)h_j^*(Y,\,PZ)-\tau_{ij}(Y)h_j^*(X,\,PZ)\}\\
&-\sum_{a=r+1}^{n}\epsilon_a\{\rho_{ia}(X)h_a^s(Y,\,PZ)-\rho_{ia}(Y)h_a^s(X,\,PZ)\}\\
&+\sum_{j=1}^{r}\{h_j^\ell(X,\,PZ)\eta_i(A_{N_j}Y)-h_j^\ell(Y,\,PZ)\eta_i(A_{N_j}X)\}\\
&-\theta(X)h_i^*(FY,PZ)+\theta(Y)h_i^*(FX,PZ)\\
&-(\bar\nabla_X\theta)(PZ)v_i(Y)+(\bar\nabla_Y\theta)(PZ)v_i(X)\\
&+\alpha\theta(PZ)\{\theta(Y)\eta_i(X)-\theta(X)\eta_i(Y)\}\\
&+\beta\theta(PZ)\{\theta(Y)v_i(X)-\theta(X)v_i(Y)\}\\
&=\ f_1\{g(Y,PZ)\eta_i(X)-g(X,PZ)\eta_i(Y)\}\\
&+f_2\{v_i(Y)\bar g(X,JPZ)-v_i(X)\bar g(Y,JPZ)+2v_i(PZ)\bar g(X,JY)\}\\
&+f_3\{\theta(X)\eta_i(Y)-\theta(Y)\eta_i(X)\}\theta(PZ).
\end{aligned}
\tag{5.6}
$$

Theorem 5.2. *Let M be a generic lightlike submanifold of an indefinite generalized Sasakian space form $\bar M(f_1,f_2,f_3)$ with a non-metric ϕ-symmetric connection such that ζ is tangent to M. Then α, β, f_1, f_2 and f_3 satisfy*

(1) α *is a constant on* M*,*

(2) $\alpha\beta = 0$*, and*

(3) $f_1 - f_2 = \alpha^2 - \beta^2$ *and* $f_1 - f_3 = \alpha^2 - \beta^2 - \zeta\beta$*.*

Proof. Applying $\bar{\nabla}_X$ to $\theta(U_i) = 0$ and $\theta(V_i) = 0$ by turns and using (2.4), (3.15), (3.16) and the facts that $F\zeta = 0$ and ζ belongs to $S(TM)$, we get

$$(\bar{\nabla}_X\theta)(U_i) = \alpha\eta_i(X) + \beta v_i(X), \qquad (\bar{\nabla}_X\theta)(V_i) = \beta u_i(X). \tag{5.7}$$

Applying ∇_X to $(3.14)_1$: $h^\ell_j(Y, U_i) = h^*_i(Y, V_j)$ and using (2.1), (2.12), (3.7), (3.9), (3.11), (3.12), $(3.14)_{1,2,4}$, (3.15) and (3.16), we obtain

$$\begin{aligned}(\nabla_X h^\ell_j)(Y, U_i) &= (\nabla_X h^*_i)(Y, V_j) \\ &- \sum_{k=1}^{r}\{\tau_{kj}(X)h^\ell_k(Y, U_i) + \tau_{ik}(X)h^*_k(Y, V_j)\} \\ &- \sum_{a=r+1}^{n}\{\lambda_{aj}(X)h^s_a(Y, U_i) + \epsilon_a\rho_{ia}(X)h^s_a(Y, V_j)\} \\ &+ \sum_{k=1}^{r}\{h^*_i(Y, U_k)h^\ell_k(X, \xi_j) + h^*_i(X, U_k)h^\ell_k(Y, \xi_j)\} \\ &- g(A^*_{\xi_j}X, F(A_{N_i}Y)) - g(A^*_{\xi_j}Y, F(A_{N_i}X)) \\ &- \sum_{k=1}^{r} h^\ell_j(X, V_k)\eta_k(A_{N_i}Y) \\ &- \beta(\alpha - 1)\{u_j(Y)v_i(X) - u_j(X)v_i(Y)\} \\ &- \alpha(\alpha - 1)u_j(Y)\eta_i(X) - \beta^2 u_j(X)\eta_i(Y).\end{aligned}$$

Substituting this equation into the modification equation, which is change i into j and Z into U_i from (5.5), and using $(3.6)_3$ and $(3.14)_3$, we have

$$\begin{aligned}&(\nabla_X h^*_i)(Y, V_j) - (\nabla_Y h^*_i)(X, V_j) \\ &- \sum_{k=1}^{r}\{\tau_{ik}(X)h^*_k(Y, V_j) - \tau_{ik}(Y)h^*_k(X, V_j)\} \\ &- \sum_{a=r+1}^{n}\epsilon_a\{\rho_{ia}(X)h^s_a(Y, V_j) - \rho_{ia}(Y)h^s_a(X, V_j)\} \\ &+ \sum_{k=1}^{r}\{h^\ell_k(X, V_j)\eta_i(A_{N_k}Y) - h^\ell_k(Y, V_j)\eta_i(A_{N_k}X)\} \\ &- \theta(X)h^*_i(FY, V_j) + \theta(Y)h^*_i(FX, V_j) \\ &- \beta(2\alpha - 1)\{u_j(Y)v_i(X) - u_j(X)v_i(Y)\}\end{aligned}$$

$$\begin{aligned}&- (\alpha^2 - \beta^2)\{u_j(Y)\eta_i(X) - u_j(X)\eta_i(Y)\}\\ &= \ f_2\{u_j(Y)\eta_i(X) - u_j(X)\eta_i(Y) + 2\delta_{ij}\bar{g}(X, JY)\}.\end{aligned}$$

Comparing this equation with (5.6) such that $PZ = V_j$, we obtain

$$\begin{aligned}&\{f_1 - f_2 - \alpha^2 + \beta^2\}\{u_j(Y)\eta_i(X) - u_j(X)\eta_i(Y)\}\\ &= \ 2\alpha\beta\{u_j(Y)v_i(X) - u(_jX)v_i(Y)\}.\end{aligned}$$

Taking $Y = U_j$, $X = \xi_i$ and $Y = U_j$, $X = V_i$ to this by turns, we have

$$f_1 - f_2 = \alpha^2 - \beta^2, \qquad\qquad \alpha\beta = 0.$$

Applying $\bar{\nabla}_X$ to $\theta(\zeta) = 1$ and using (2.3) and the fact: $\theta \circ J = 0$, we get

$$(\bar{\nabla}_X\theta)(\zeta) = 0. \tag{5.8}$$

Applying $\bar{\nabla}_X$ to $\eta_i(Y) = \bar{g}(Y, N_i)$ and using (1.1) and (2.5), we have

$$(\nabla_X\eta)(Y) = -g(A_{N_i}X, Y) + \sum_{j=1}^{r}\tau_{ij}(X)\eta_j(Y) - \theta(Y)v_i(X). \tag{5.9}$$

Applying ∇_X to $h_i^*(Y, \zeta) = -(\alpha - 1)v_i(Y) + \beta\eta_i(Y)$ and using (3.9), (3.10), (3.20), (5.9) and the fact that $\alpha\beta = 0$, we get

$$\begin{aligned}(\nabla_X h_i^*)(Y, \zeta) \ &= -(X\alpha)v_i(Y) + (X\beta)\eta_i(Y)\\ &+ (\alpha - 1)\{g(A_{N_i}X, FY) + g(A_{N_i}Y, FX)\\ &\qquad - \sum_{j=1}^{r} v_j(Y)\tau_{ij}(X) - \sum_{a=r+1}^{n}\epsilon_a w_a(Y)\rho_{ia}(X)\\ &\qquad - \sum_{j=1}^{r} u_j(Y)\eta_i(A_{N_j}X) - (\alpha - 1)\theta(Y)\eta_i(X)\}\\ &- \beta\{g(A_{N_i}X, Y) + g(A_{N_i}Y, X) - \sum_{j=1}^{r}\tau_{ij}(X)\eta_j(Y)\\ &\qquad - \beta\theta(X)\eta_i(Y)\}.\end{aligned}$$

Substituting this and $(3.13)_2$ into (5.6) with $PZ = \zeta$ and using (5.8), we get

$$\begin{aligned}&\{X\beta + (f_1 - f_3 - \alpha^2 + \beta^2)\theta(X)\}\eta_i(Y)\\ &- \{Y\beta + (f_1 - f_3 - \alpha^2 + \beta^2)\theta(Y)\}\eta_i(X)\\ &= \ (X\alpha)v_i(Y) - (Y\alpha)v_i(X).\end{aligned}$$

Taking $X = \zeta$, $Y = \xi_i$ and $X = U_j$, $Y = V_i$ to this by turns, we have

$$f_1 - f_3 = \alpha^2 - \beta^2 - \zeta\beta, \qquad\qquad U_j\alpha = 0.$$

Applying ∇_Y to $(3.11)_2$ and using (3.10) and (3.19), we get

$$\begin{aligned}(\nabla_X h_i^\ell)(Y,\zeta) = &-(X\alpha)u_i(Y)\\ &+(\alpha-1)\{\sum_{j=1}^{r} u_j(Y)\tau_{ij}(X)+\sum_{a=r+1}^{n}\epsilon_a w_a(Y)\lambda_{ai}(X)\\ &\qquad + h_i^\ell(X,FY)+h_i^\ell(Y,FX)\}\\ &-\beta\{h_i^\ell(Y,X)+\theta(Y)u_i(X)-\theta(X)u_i(Y)\}.\end{aligned}$$

Substituting this into (5.5) such that $Z=\zeta$ and using (3.3) and (5.8), we have

$$(X\alpha)u_i(Y)=(Y\alpha)u_i(X).$$

Taking $Y=U_i$ to this result and using the fact that $U_i\alpha=0$, we have $X\alpha=0$. Therefore α is a constant. This completes the proof of the theorem. □

Theorem 5.3. *Let M be a generic lightlike submanifold of an indefinite generalized Sasakian space form $\bar M(f_1,f_2,f_3)$ with a non-metric ϕ-symmetric connection such that ζ is tangent to M. If F is Lie recurrent, then*

$$\alpha=0,\qquad f_1=-\beta^2,\qquad f_2=0,\qquad f_3=-\zeta\beta.$$

Proof. By Theorem 4.2, we shown that $\alpha=0$ and we have $(4.11)_2$. Applying ∇_X to $(4.11)_2$: $h_i^\ell(Y,U_j)=0$ and using $(3.11)_2$, (3.15) and $(4.11)_2$, we have

$$\begin{aligned}(\nabla_X h_i^\ell)(Y,U_j) = &-h_i^\ell(Y,F(A_{N_j}X))-\sum_{a=r+1}^{n}\rho_{ja}(X)h_i^\ell(Y,W_a)\\ &+\beta u_i(Y)v_j(X).\end{aligned}$$

Substituting this into (5.5) with $Z=U_j$ and using $(5.7)_1$, we obtain

$$\begin{aligned}&h_i^\ell(X,F(A_{N_j}Y))-h_i^\ell(Y,F(A_{N_j}X))\\ &+\sum_{a=r+1}^{n}\{\rho_{ja}(Y)h_i^\ell(X,W_a)-\rho_{ja}(X)h_i^\ell(Y,W_a)\}\\ &+\sum_{a=r+1}^{n}\{\lambda_{ai}(X)h_a^s(Y,U_j)-\lambda_{ai}(Y)h_a^s(X,U_j)\}\\ =\;& f_2\{u_i(Y)\eta_j(X)-u_i(X)\eta_j(Y)+2\delta_{ij}\bar g(X,JY)\}.\end{aligned}$$

Taking $Y=U_i$ and $X=\xi_j$ to this and using (3.3) and $(4.5)_{1,3,5}$, we have

$$3f_2=h_i^\ell(\xi_j,F(A_{N_j}U_i))+\sum_{a=r+1}^{n}\rho_{ja}(U_i)h_i^\ell(\xi_j,W_a). \tag{5.10}$$

In general, replacing X by ξ_j to (3.7) and using (3.3) and $(3.6)_7$, we get $h_i^\ell(X,\xi_j)=g(A^*_{\xi_i}\xi_j,X)$. From this and $(3.6)_1$, we obtain $A^*_{\xi_i}\xi_j=-A^*_{\xi_j}\xi_i$. Thus

$A^*_{\xi_i}\xi_j$ are skew-symmetric with respect to i and j. On the other hand, in case M is Lie recurrent, taking $Y = U_j$ to (4.10), we have $A_{N_i}U_j = A_{N_j}U_i$. Thus $A_{N_i}U_j$ are symmetric with respect to i and j. Therefore, we get

$$h^\ell_i(\xi_j,\, F(A_{N_j}U_i)) = g(A^*_{\xi_i}\xi_j, F(A_{N_j}U_i)) = 0.$$

Also, by using (3.4), $(3.6)_2$, $(3.14)_4$ and $(4.5)_4$, we have

$$h^\ell_i(\xi_j, W_a) = \epsilon_a h^s_a(\xi_j, V_i) = \epsilon_a h^s_a(V_i, \xi_j) = -\lambda_{ja}(V_i) = 0.$$

Thus we get $f_2 = 0$ by (5.10). Therefore, $f_1 = -\beta^2$ and $f_3 = -\zeta\beta$. □

Theorem 5.4. *Let M be a generic lightlike submanifold of an indefinite generalized Sasakian space form $\bar{M}(f_1, f_2, f_3)$ with a non-metric ϕ-symmetric connection such that ζ is tangent to M. If U_is or V_is are parallel with respect to ∇, then $\bar{M}(f_1, f_2, f_3)$ is a flat manifold with an indefinite cosymplectic structure;*

$$\alpha = \beta = 0, \qquad f_1 = f_2 = f_3 = 0.$$

Proof. (1) If U_is are parallel with respect to ∇, then we have (3.22). As $\alpha = 0$, we get $f_1 = f_2 = f_3$ by Theorem 5.2. Applying ∇_Y to $(3.22)_5$, we obtain

$$(\nabla_X h^*_i)(Y, U_j) = 0.$$

Substituting this equation and (3.22) into (5.6) with $PZ = U_j$, we have

$$f_1\{v_j(Y)\eta_i(X) - v_j(X)\eta_i(Y)\} + f_2\{v_i(Y)\eta_j(X) - v_i(X)\eta_j(Y)\} = 0.$$

Taking $X = \xi_i$ and $Y = V_j$ to this equation, we get $f_1 + f_2 = 0$. Thus we see that $f_1 = f_2 = f_3 = 0$ and $\bar{M}$ is flat.

(2) If V_is are parallel with respect to ∇, then we have (3.23) and $\alpha = 0$. As $\alpha = 0$, we get $f_1 = f_2 = f_3$ by Theorem 5.2. From $(3.14)_1$ and $(3.23)_5$, we have

$$h^*_i(Y, V_j) = 0.$$

Applying ∇_X to this equation and using the fact that $\nabla_X V_j = 0$, we have

$$(\nabla_X h^*_i)(Y, V_j) = 0.$$

Substituting these two equations into (5.6) such that $PZ = V_j$, we obtain

$$\begin{aligned}
&\sum_{a=r+1}^{n} \epsilon_a\{\rho_{ia}(Y)h^s_a(X,\, V_j) - \rho_{ia}(X)h^s_a(Y,\, V_j)\}\\
&+ \sum_{k=1}^{r} \{h^\ell_k(X,\, V_j)\eta_i(A_{N_k}Y) - h^\ell_k(Y,\, V_j)\eta_i(A_{N_k}X)\}\\
=\;& f_1\{u_j(Y)\eta_i(X) - u_j(X)\eta_i(Y)\} + 2f_2\delta_{ij}\bar{g}(X, JY).
\end{aligned}$$

Taking $X = \xi_i$ and $Y = U_j$ to this equation and using (3.3), $(3.23)_{3,4,5}$ and the fact that $h^s_a(U_j, V_j) = \epsilon_a h^\ell_i(U_j, W_a) = 0$ due to (3.3), $(3.14)_4$ and $(3.23)_5$, we obtain $f_1 + 2f_2 = 0$. It follows that $f_1 = f_2 = f_3 = 0$ and $\bar{M}$ is flat. □

Definition 5.5. An r-lightlike submanifold M is called *totally umbilical* [6] if there exist smooth functions $\mathcal{A}_i$ and $\mathcal{B}_a$ on a neighborhood $\mathcal{U}$ such that

$$h_i^\ell(X,Y) = \mathcal{A}_i g(X,Y), \qquad h_a^s(X,Y) = \mathcal{B}_a g(X,Y). \tag{5.11}$$

In case $\mathcal{A}_i = \mathcal{B}_a = 0$, we say that M is *totally geodesic.*

Theorem 5.6. *Let M be a generic lightlike submanifold of an indefinite generalized Sasakian space form $\bar{M}(f_1, f_2, f_3)$ with a non-metric ϕ-symmetric connection such that ζ is tangent to M. If M is totally umbilical, then $\bar{M}(f_1, f_2, f_3)$ is an indefinite Sasakian space form such that*

$$\alpha = 1, \quad \beta = 0; \qquad f_1 = \frac{2}{3}, \quad f_2 = f_3 = -\frac{1}{3}.$$

Proof. Taking $Y = \zeta$ to $(5.11)_{1,\,2}$ by turns and using $(3.12)_{1,\,2}$, we have

$$\mathcal{A}_i\theta(X) = -(\alpha - 1)u_i(X), \qquad \mathcal{B}_a\theta(X) = -(\alpha - 1)w_a(X),$$

respectively. Taking $X = \zeta$ and $X = U_i$ to the first equation by turns, we have $\mathcal{A}_i = 0$ and $\alpha = 1$ respectively. Taking $X = \zeta$ to the second equation, we have $\mathcal{B}_a = 0$. As $\mathcal{A}_i = \mathcal{B}_a = 0$, M is totally geodesic. As $\alpha = 1$ and $\beta = 0$, $\bar{M}$ is an indefinite Sasakian manifold and $f_1 - 1 = f_2 = f_3$ by Theorem 5.2.

Taking $Z = U_j$ to (5.5) and using $(5.7)_1$ and $h_i^\ell = h_a^s = 0$, we get

$$(f_2 + 1)\{u_i(Y)\eta_j(X) - u_i(X)\eta_j(Y)\} + 2\delta_{ij} f_2 \bar{g}(X, JY) = 0.$$

Taking $X = \xi_j$ and $Y = U_i$, we have $f_2 = -\frac{1}{3}$. Thus $f_1 = \frac{2}{3}$ and $f_3 = -\frac{1}{3}$. □

Definition 5.7. (1) A screen distribution $S(TM)$ is said to be *totally umbilical* [6] in M if there exist smooth functions γ_i on a neighborhood $\mathcal{U}$ such that

$$h_i^*(X, PY) = \gamma_i g(X, PY).$$

In case $\gamma_i = 0$, we say that $S(TM)$ is *totally geodesic* in M.

(2) An r-lightlike submanifold M is said to be *screen conformal* [8] if there exist non-vanishing smooth functions φ_i on $\mathcal{U}$ such that

$$h_i^*(X, PY) = \varphi_i h_i^\ell(X, PY). \tag{5.12}$$

Theorem 5.8. *Let M be a generic lightlike submanifold of an indefinite generalized Sasakian space form $\bar{M}(f_1, f_2, f_3)$ with a non-metric ϕ-symmetric connection such that ζ is tangent to M. If $S(TM)$ is totally umbilical or M is screen conformal, then $\bar{M}(f_1, f_2, f_3)$ is an indefinite Sasakian space form;*

$$\alpha = 1, \quad \beta = 0; \qquad f_1 = 0, \quad f_2 = f_3 = -1.$$

Proof. (1) If $S(TM)$ is totally umbilical, then $(3.13)_2$ is reduced to

$$\gamma_i\theta(X) = -(\alpha - 1)v_i(X) + \beta\eta_i(X).$$

Replacing X by V_i, ξ_i and ζ respectively, we have $\alpha = 1$, $\beta = 0$ and $\gamma_i = 0$. As $\gamma_i = 0$, $S(TM)$ is totally geodesic, and $h^s_a(X, U_k) = 0$ and $h^\ell_j(X, U_k) = 0$. As $\alpha = 1$ and $\beta = 0$, $\bar{M}$ is an indefinite Sasakian manifold and $f_1 - 1 = f_2 = f_3$ by Theorem 5.1. Taking $PZ = U_k$ to (5.6) with $h^*_i = 0$, we get

$$f_1[\{v_k(Y)\eta_i(X) - v_k(X)\eta_i(Y)\} + \{v_i(Y)\eta_k(X) - v_i(X)\eta_k(Y)\}] = 0.$$

Taking $X = \xi_i$ and $Y = V_k$, we have $f_1 = 0$. Thus $f_2 = f_3 = -1$.

(2) If M is screen conformal, then, from $(3.12)_2$, $(3.13)_2$ and (5.12), we have

$$(\alpha - 1)\{v_i(X) - \beta\eta_i(X) = \varphi_i(\alpha - 1)u_i(X)\}.$$

Taking $X = V_i$ and $X = \xi_i$ to this equation by turns, we have $\alpha = 1$ and $beta = 0$. As $\alpha = 1$ and $\beta = 0$, $\bar{M}$ is an indefinite Sasakian manifold and $f_1 - 1 = f_2 = f_3$ by Theorem 5.1.

Denote by μ_i the r-th vector fields on $S(TM)$ such that $\mu_i = U_i - \varphi_i V_i$. Then $J\mu_i = N_i - \varphi_i\xi_i$. Using $(3.14)_{1,2,3,4}$ and (5.12), we get

$$h^\ell_j(X, \mu_i) = 0, \qquad h^s_a(X, \mu_i) = 0. \tag{5.13}$$

Applying ∇_Y to (5.12), we have

$$(\nabla_X h^*_i)(Y, PZ) = (X\varphi_i)h^\ell_i(Y, PZ) + \varphi_i(\nabla_X h^\ell_i)(Y, PZ).$$

Substituting this equation and (5.12) into (5.6) and using (5.5), we have

$$\begin{aligned}
&\sum_{j=1}^{r}\{(X\varphi_i)\delta_{ij} - \varphi_i\tau_{ji}(X) - \varphi_j\tau_{ij}(X) - \eta_i(A_{N_j}X)\}h^\ell_j(Y, PZ)\\
&- \sum_{j=1}^{r}\{(Y\varphi_i)\delta_{ij} - \varphi_i\tau_{ji}(Y) - \varphi_j\tau_{ij}(Y) - \eta_i(A_{N_j}Y)\}h^\ell_j(X, PZ)\\
&- \sum_{a=r+1}^{n}\{\epsilon_a\rho_{ia}(X) + \varphi_i\lambda_{ai}(X)\}h^s_a(Y, PZ)\\
&+ \sum_{a=r+1}^{n}\{\epsilon_a\rho_{ia}(Y) + \varphi_i\lambda_{ai}(Y)\}h^s_a(X, PZ)\\
&- (\bar{\nabla}_X\theta)(PZ)\{v_i(Y) - \varphi u_i(Y)\} + (\bar{\nabla}_Y\theta)(PZ)\{v_i(X) - \varphi u_i(X)\}\\
&- \alpha\{\theta(X)\eta_i(Y) - \theta(Y)\eta_i(X)\}\theta(PZ)\\
&= \ f_1\{g(Y, PZ)\eta_i(X) - g(X, PZ)\eta_i(Y)\}\\
&+ f_2\{[v_i(Y) - \varphi_i u_i(Y)]\bar{g}(X, JPZ) - [v_i(X) - \varphi_i u_i(X)]\bar{g}(Y, JPZ)
\end{aligned}$$

$$\begin{aligned}&+2[v_i(PZ)-\varphi_i u_i(PZ)]\bar{g}(X,JY)\}\\ &+f_3\{\theta(X)\eta_i(Y)-\theta(Y)\eta_i(X)\}\theta(PZ).\end{aligned}$$

Replacing PZ by μ_j to this and using (5.7) and (5.13), we obtain

$$\begin{aligned}&f_1\{[v_j(Y)\eta_i(X)-v_j(X)\eta_i(Y)]-\varphi_j[u_j(Y)\eta_i(X)-u_j(X)\eta_i(Y)]\}\\ &+f_1\{[v_i(Y)\eta_j(X)-v_i(X)\eta_j(Y)]-\varphi_i[u_i(Y)\eta_j(X)-u_i(X)\eta_j(Y)]\}\\ &-2f_2(\varphi_j+\varphi_i)\delta_{ij}\bar{g}(X,JY)=0.\end{aligned}$$

Taking $X=\xi_i$ and $Y=V_j$, we get $f_1=0$. Thus $f_2=f_3=-1$. □

Acknowledgements. We would like to express our gratitude to the referees for valuable comments and suggestions.

References

[1] Ageshe,N. S., Chafle, M. R., A semi-symmetric non-metric connection on a Riemannian manifold, *Indian J. Pure Appl. Math.*, Vol. 23 (1992), No. 6, 399–409.

[2] Ahmad, M., Haseeb, A., Hypersurfaces of an almost r-paracontact Riemannian manifold endowed with a quarter-symmetric non-metric connection, *Kyungpook Math. J.*, Vol 49 (2009), 533–543.

[3] Alegre, P, Blair, D. E., Carriazo, A., Generalized Sasakian space form, *Israel J. Math.*, Vol 141 (2004), 157–183.

[4] Barman, A., On a type of quarter-symmetric non-metric ϕ-connection on a Kenmotsu manifold, *Bull. Math. Analy. and Appl.*, Vol 4 (2012), No.3, 1–11.

[5] Călin, C., Contributions to geometry of CR-submanifold, *Thesis, University of Iasi (Romania)* (1998).

[6] Duggal, K. L., Jin, D. H., Generic lightlike submanifolds of an indefinite Sasakian manifold, *Int. Elec. J. Geo.*, Vol 5 (2012), No.1, 108–119.

[7] Golab, S. On semi-symmetric and quarter-symmetric connections, *Tensor, N.S.*, Vol 29 (1975), 249–254.

[8] Jin, D. H., Half lightlike submanifolds of an indefinite trans-Sasakian manifold, *Bull. Korean Math. Soc.*, Vol 51 (2014), No.4, 979–994.

[9] Jin, D. H., Indefinite generalized Sasakian space form admitting a generic lightlike submanifold, *Bull. Korean Math. Soc.*, Vol 51 (2014), No.6, 1711–1726.

[10] Jin, D. H., Generic lightlike submanifolds of an indefinite trans-Sasakian manifold of a quasi-constant curvature, *Appl. Math. Sci.*, Vol 9 (2015), No.60, 2985–2997.

[11] Jin, D. H., Special lightlike hypersurfaces of indefinite Kaehler manifolds, *Filomat*, Vol 30 (2016), No.7, 1919–1930.

[12] Jin, D. H., Lightlike hypersurfaces of an indefinite trans-Sasakian manifold with a non-metric ϕ-symmetric connection, *Bull. Korean Math. Soc.*, Vol 53 (2016), No.6, 1771–1783.

[13] Jin, D. H., Lightlike hypersurfaces of an indefinite Kaehler manifold with a non-metric ϕ-symmetric connection, *Bull. Korean Math. Soc.*, Vol 54 (2017), No.2, 619–632.

[14] Jin, D. H., Lee, J. W., Generic lightlike submanifolds of an indefinite cosymplectic manifold, *Math. Probl. in Engin.*, Vol 2011, Art ID 610986, 1–16.

[15] Jin, D. H., Lee, J. W., A semi-Riemannian manifold of quasi-constant curvature admits lightlike submanifolds, *Inter. J. of Math. Analysis*, Vol 9 (2015), No.25, 1215–1229.

[16] Jin, D. H., Lee, J. W., Generic lightlike submanifolds of an indefinite Kaehler manifold, *Inter. J. Pure and Appl. Math.*, Vol 101 (2015), No.4, 543–560.

[17] Kupeli, D. N., Singular Semi-Riemannian Geometry, *Kluwer Academic*, Vol 366 (1996).

[18] Oubina, J. A., New classes of almost contact metric structures, *Publ. Math. Debrecen*, Vol 32 (1985), 187–193.

[19] Shaikh, A. A., Sanjib Kumar Jana, On quarter symmetric metric connections on a (k, m)-contact metric manifolds, *Commun. Fac. Sci. Univ. Ank. Series A1*, Vol 55 (2006), No.1, 33–45.

[20] Yadav, S., Suthar, D. L., A quart symmetric non-metric connection in a generalized co-symmpletic manifolds, *Glob. J. Sci. Fron. Res.*, Vol 11 (2011), No.9, 1–7.

Annales Mathematicae et Informaticae
47 (2017) pp. 109–118
http://ami.uni-eszterhazy.hu

Gardener's spline curve

Imre Juhász

Department of Descriptive Geometry, University of Miskolc
agtji@uni-miskolc.hu

Submitted August 24, 2017 — Accepted November 29, 2017

Abstract

In the well-known gardener's construction of the ellipse we replace the two foci by a finite set of points in the plane, that results in a G^1 spline curve that consists of elliptic arcs, if the set contains at least three non-collinear points. An algorithm is provided for the specification of these elliptic arcs, along with their control point based representation.

Keywords: Gardener's construction, spline curve, rational Bézier representation

MSC: 65D17, 68U07

1. Introduction

A well-known way of drawing an ellipse is as follows. Place a piece of paper on the board, stick in two pins, loop a thread around the pins, pull taut with the tip of a pen and move the pen around, always keeping the loop of thread taut. As the pen moves around the two pins it will trace out an ellipse. This makes use of the fact that an ellipse is the locus of points, whose sum of distances from two fixed points is a constant. (Certainly, the usage of a thread is not a construction in the Euclidean sense.)

Replacing the paper by the ground, the pins by pegs, the thread with a string (or a rope) and the pen with a peg (or a spade), this procedure is used by gardeners to outline an elliptical flower bed. Therefore, this method is called the gardener's construction (or the string method). It is not known when, where and by whom it has been invented but it is doubtless that this has been used by gardeners for quite a while.

A generalization of the gardener's construction is attributed to Charles Graves, cf. [1], who replaced the two pins (the foci) by an ellipse and proved that the gardener's construction results in another ellipse, confocal to the original one.

In what follows, we replace the two foci with a finite set of points in the plane, in which case the gardener's construction produces a closed G^1 spline curve, which is composed of elliptic arcs. We provide an algorithm for the specification of the arcs, that enables us to draw this closed curve by means of a computer. This work was motivated by an incomplete, erroneous text (cf. [2]) found on the internet on a similar topic (its title is misleading).

2. Problem statement

If we loop a finite set of points in the plane (pins stuck in a board) with a string, pull taut with the tip of a pen and move the pen around, the string will always tighten on a part of the perimeter of the convex hull of the set. The convex hull of a finite set of points in the plane is always bounded by a closed convex polygon, the vertices of which are elements of the given set. There are several methods to compute the convex hull of a finite set of points, one can find them, e.g., in [3] or [4].

If the set contains just a single point or all the points are collinear, the convex hull degenerates to a single point or to a straight line segment, respectively. In these degenerate cases the gardener's construction produces a circle or an ellipse. We exclude these trivial cases in our further study, i.e., we assume that the set contains at least three non-collinear points. From now on, we will examine closed convex planar polygons instead of a finite set of points.

Let $\mathbf{p}_1, \mathbf{p}_2, \ldots, \mathbf{p}_n$ be the vertices of a closed convex planar polygon $\mathcal{P}$. Throughout the paper we make use of the convention

$$\mathbf{p}_i \equiv \mathbf{p}_{i \bmod n}. \tag{2.1}$$

This convention is applied for the half-lines and support lines as well, which will be introduced later. We assume that the orientation of the polygon is counterclockwise. In what follows, if we list certain entities, like foci or delimiting lines, we always do it in the counterclockwise direction. The perimeter of the polygon is denoted by L_p, i.e.,

$$L_p = \sum_{i=0}^{n} \left\| \mathbf{p}_{i+1} - \mathbf{p}_i \right\| .$$

Consider the distance $L = L_p + \delta,\ 0 < \delta \in \mathbb{R}$.

The support lines

$$\mathbf{l}_i\left(t\right) = \left(1-t\right)\mathbf{p}_i + t\mathbf{p}_{i+1},\ t \in \mathbb{R},\ i = 1, 2, \ldots, n$$

of sides of the polygon $\mathcal{P}$ divide the plane into $2n$ external parts (outside of the polygon), each part containing an arc of the closed curve. Each side $\mathbf{s}_i$ – bounded by vertices $\mathbf{p}_i$ and $\mathbf{p}_{i+1}$ – and each vertex $\mathbf{p}_i$ has its own region that will contain an elliptic arc, that may degenerate to a single point in special circumstances. Side and vertex arcs alternately follow each other, the sequence is $\mathbf{s}_i, \mathbf{v}_{i+1}, (i = 1, 2, \dots, n)$. Each arc is a part of an ellipse, since the construction inherently ensures that the sum of the distances of any point of the arc from two certain vertices of the polygon is constant. Certainly, these two points (the foci) and the constant (the length of the major axis) vary arc by arc.

3. Triangle

At first we consider the simplest case, the triangle. In this case we have the sequence of arcs $\mathbf{s}_i, \mathbf{v}_{i+1}, (i = 1, 2, 3)$. Foci of the side arc $\mathbf{s}_i$ are $\mathbf{p}_i$,$\mathbf{p}_{i+1}$ and its endpoints are on support lines $\mathbf{l}_{i-1}, \mathbf{l}_{i+1}$ that we will refer to as delimiters of the arc. The vertex arc $\mathbf{v}_i$ has the foci $\mathbf{p}_{i-1}$,$\mathbf{p}_{i+2}$, and its delimiters are $\mathbf{l}_i, \mathbf{l}_{i-1}$. Consecutive arcs share a focus and a delimiter, moreover, at the common point of the two arcs the tangent lines are also coinciding, cf. Fig. 1. Thus, the result is a G^1 spline curve that consists of elliptic arcs.

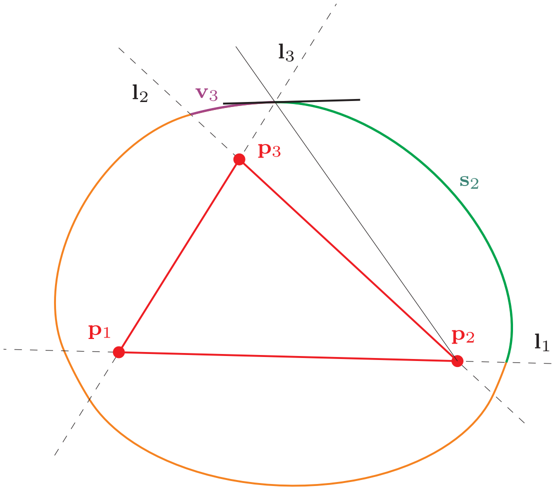


Figure 1: The case of the triangle: side arc $\mathbf{s}_2$ has the foci $\mathbf{p}_2, \mathbf{p}_3$ and delimiters $\mathbf{l}_1, \mathbf{l}_3$; the consecutive vertex arc $\mathbf{v}_3$ has foci $\mathbf{p}_2, \mathbf{p}_1$ and delimiters $\mathbf{l}_3, \mathbf{l}_2$. The two arcs have a common tangent line at their joint.

In the case of the triangle and the parallelogram, i.e., when support lines meet only at vertices of the polygon, this simple structure always works. Otherwise, if δ is big enough, the determination of foci and delimiters of arcs may become much more complicated. Therefore, we have to construct an algorithm which can cope with any convex polygon.

4. Generic case

Our objective is to determine the defining data of consecutive elliptic arcs, i.e., the two foci, the delimiters of the arc and the length of the major axis, for any configuration of closed convex polygons. To this end, we introduce half-lines

$$\mathbf{h}_i(t) = (1-t)\,\mathbf{p}_i + t\mathbf{p}_{i+1},\ 0 \le t \in \mathbb{R},\ (1, 2, \dots, n)$$

and build an intersection matrix M of size $n \times n$. Rows of this matrix correspond to half-lines $\mathbf{h}_i$ and its columns to support lines $\mathbf{l}_j$. We examine only those intersection points of half-lines and support lines which differ from the vertices of the polygon. Entry $m_{i,j}$ of matrix M equals 1 if the support line $\mathbf{l}_j, (j > i+1)$ intersects the half-line $\mathbf{h}_i$ and for the intersection point $\mathbf{q}_{ij}$ inequality

$$L_p - \sum_{k=i+1}^{j-1} \|\mathbf{p}_{k+1} - \mathbf{p}_k\| + \|\mathbf{q}_{ij} - \mathbf{p}_{i+1}\| + \|\mathbf{q}_{ij} - \mathbf{p}_j\| < L$$

holds, otherwise $m_{i,j} = 0$. Thus, the intersection matrix depends not only on the location of the vertices but on δ as well. If in the above expression we have equality, it means that the arc is degenerated to a single point, which does not need any further study. The first support line (if there is any) that intersects the half-line $\mathbf{h}_i$ has to be $\mathbf{l}_{i+2}$ and if there are more than one such support lines, they must be consecutive ones, since the polygon is convex. (Note, that vertices, half-lines and support lines are cyclically arranged, convention (2.1) is used.) The intersection matrix of the configuration in Fig. 2 is

$$M = \begin{bmatrix} 0 & 0 & 1 & 0 & 0 & 0 \\ 0 & 0 & 0 & 1 & 0 & 0 \\ 0 & 0 & 0 & 0 & 1 & 1 \\ 1 & 0 & 0 & 0 & 0 & 1 \\ 1 & 1 & 0 & 0 & 0 & 0 \\ 0 & 1 & 0 & 0 & 0 & 0 \end{bmatrix}.$$

Subsequently we provide algorithms for the specification of the list of foci and delimiters of the elliptic arcs, by processing the intersection matrix.

4.1. Specification of foci

In this subsection we produce the pair of foci of consecutive elliptic arcs by processing the intersection matrix row by row. Rows and columns of the intersection matrix can be considered as cycles, that is any element has a predecessor and a subsequent element according to the convention (2.1). We always process the vertices in counterclockwise direction. If in a list or in a sum the lower limit r_ℓ happens to be greater than the upper limit r_u, then the sequence $r_\ell, r_\ell + 1, \dots, n, n+1, r_u$ is meant.

We process the intersection matrix row by row. Processing the ith row of the intersection matrix $(i = 1, 2, \dots, n)$ is as follows.

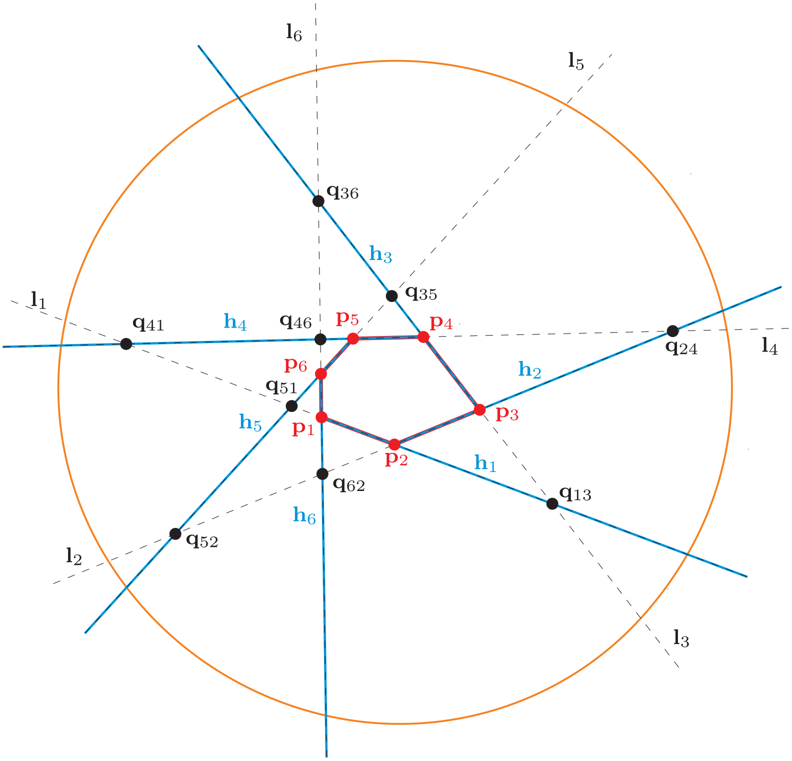


Figure 2: Support lines $\mathbf{l}_i$, half-lines $\mathbf{h}_i$ $(i = 1, 2, \ldots, 6)$ and the intersection $\mathbf{q}_{ij}$ of halflines and support lines of a hexagon.

1. If $m_{i,j} = 0, (j = 1, 2, \ldots, n)$ then the pairs of foci are

   1.1. if $m_{i-1,i+1} = 0$ then insert $\mathbf{p}_i, \mathbf{p}_{i+1}$

   1.2. if $m_{i-1,i+2} = 0$ then insert $\mathbf{p}_i, \mathbf{p}_{i+2}$

2. Otherwise, find the first element k_1 of the row which equals 1 and the corresponding entry in the previous row is 0, i.e., $m_{i,k_1} = 1$ and $m_{i-1,k_1} = 0$, moreover find k_2 which is the last element of the ith row of this property. (The search always starts at column $j = i + 1$ and goes around.) If there is no column that fulfills both requirements, then set $k_1 = 0$ and find k_ℓ which is the last column containing 1 (regardless of the previous row). The pairs of foci are as follows.

   2.1. If $k_1 > 0$ then

   2.1.1. if the previous row is the zero vector, i.e., $m_{i-1,j} = 0$, $(i = 1, 2, \ldots, n)$ then insert $\mathbf{p}_i, \mathbf{p}_{k_1-1}$

   2.1.2. insert $\mathbf{p}_i, \mathbf{p}_{k_1}$; $\mathbf{p}_i, \mathbf{p}_{k_1+1}$; $\ldots, \mathbf{p}_i, \mathbf{p}_{k_2+1}$

2.2. else insert $\mathbf{p}_i, \mathbf{p}_{k_\ell+1}$

The pairs of foci of the configuration in Fig. 2 are

$$\begin{bmatrix} 1 & 1 & 2 & 2 & 3 & 3 & 3 & 4 & 4 & 5 & 5 & 6 \\ 3 & 4 & 4 & 5 & 5 & 6 & 1 & 1 & 2 & 2 & 3 & 3 \end{bmatrix}.$$

4.2. Specification of delimiters

We specify the delimiters of consecutive elliptic arcs based on the intersection matrix. Consecutive pairs of the delimiters share an element, namely the second element of the current pair is the first element of the next one.

Processing of the ith row of the intersection matrix is as follows.

1. If $m_{i,j} = 0, (j = 1, 2, \dots, n)$, i.e., the row is the zero vector, then

   1.1. if $m_{i-1,i+1} = 0$ then insert $\mathbf{l}_{i-1}, \mathbf{l}_{i+1};\ \mathbf{l}_{i+1}, \mathbf{l}_i$

   1.2. else insert $\mathbf{l}_{i-1}, \mathbf{l}_i$

2. Otherwise find the first element k_1 of the row which equals 1 and the corresponding entry in the previous row is 0, i.e., $m_{i,k_1} = 1$ and $m_{i-1,k_1} = 0$, moreover find k_2 which is the last column of the ith row of this property. (The search always starts at column $j = i + 1$ and goes round.) If there is no column that fulfills both requirements, then set $k_1 = 0$. Pairs of delimiters are:

3. If $k_1 > 0$ then

   3.1. if the previous row is the zero vector, then insert $\mathbf{l}_{i-1}, \mathbf{l}_{k_1-1}; \mathbf{l}_{k_1-1}, \mathbf{l}_{k_1}; \mathbf{l}_{k_1}, \mathbf{l}_{k_1+1}; \mathbf{l}_{k_1+1}, \mathbf{l}_{k_1+2}; \dots; \mathbf{l}_{k_2-1}, \mathbf{l}_{k_2}$

   3.2. else insert $\mathbf{l}_{i-1}, \mathbf{l}_{k_1};\quad \mathbf{l}_{k_1}, \mathbf{l}_{k_1+1}; \mathbf{l}_{k_1+1}, \mathbf{l}_{k_1+2}; \dots; \mathbf{l}_{k_2-1}, \mathbf{l}_{k_2}$

   3.3. insert $\mathbf{l}_{k_2}, \mathbf{l}_i$

4. else insert $\mathbf{l}_{i-1}, \mathbf{l}_i$

The delimiters of elliptic arcs of the configuration in Fig. 2 are

$$\begin{bmatrix} 6 & 3 & 1 & 4 & 2 & 5 & 6 & 3 & 1 & 4 & 2 & 5 \\ 3 & 1 & 4 & 2 & 5 & 6 & 3 & 1 & 4 & 2 & 5 & 6 \end{bmatrix}.$$

4.3. The length of the major axis

The length of the major axis of the ellipse defined by foci $\mathbf{p}_i, \mathbf{p}_j$ (the order does matter) is

$$\delta + \sum_{k=i}^{j-1} \|\mathbf{p}_{k+1} - \mathbf{p}_k\| .$$

5. Control point based representation

The best way for the description of elliptic arcs seems to be the quadratic rational Bézier form

$$\sum_{i=0}^{2} \mathbf{b}_i \frac{w_i B_i^2 (t)}{\sum_{j=0}^{2} w_j B_j^2 (t)}, \; t \in [0,1], \tag{5.1}$$

where $B_i^2 (t)$, $(i = 0, 1, 2)$ denote quadratic Bernstein polynomials, and w_i are non-negative weights, such that $w_0 + w_1 + w_2 > 0$. The three control points can easily be computed, since we know the two endpoints and the tangent lines there. If we use the standard form for the specification of the weights $(w_0 = w_2 = 1)$ just the weight of the middle control point, i.e., w_1 has to be computed which is also a routine exercise.

An alternative representation could be the trigonometric one (cf. [5])

$$\sum_{i=0}^{2} A_{2,i}^{\alpha} (t) \, \mathbf{b}_i, \; t \in [0, \alpha], \tag{5.2}$$

where

$$\begin{aligned} A_{2,0}^{\alpha} (t) &= \frac{1}{\sin^2 \left(\frac{\alpha}{2}\right)} \sin^2 \left(\frac{\alpha - t}{2}\right), \\ A_{2,1}^{\alpha} (t) &= \frac{2 \cos \left(\frac{\alpha}{2}\right)}{\sin^2 \left(\frac{\alpha}{2}\right)} \sin \left(\frac{\alpha - t}{2}\right) \sin \left(\frac{t}{2}\right), \\ A_{2,2}^{\alpha} (t) &= \frac{\sin^2 \left(\frac{t}{2}\right)}{\sin^2 \left(\frac{\alpha}{2}\right)}. \end{aligned}$$

Its control points $\mathbf{b}_i$. $(i = 0, 1, 2)$ coincide with that of (5.1), only shape parameter α has to be calculated. Actually, (5.2) is just a reparametrization of (5.1). We prefer the rational Bézier representation. Fig. 3 shows the rational Bézier representation of the gardener's spline curve for a quadrilateral.

If we use the standard form of weights for all elliptic arcs $\mathbf{e}_i, (i = 1, 2, \ldots, 2n)$ then we obtain a G^1 description of the gardener's spline curve. In what follows we show a method for the C^1 description of the curve that will be achieved by the transformation of weights.

Let us consider two consecutive arcs

$$\mathbf{e}_i (t) = \sum_{j=0}^{2} \mathbf{b}_{i,j} \frac{w_{i,j} B_j^2 (t)}{\sum_{k=0}^{2} w_{i,k} B_k^2 (t)}, \; t \in [0,1]$$

and

$$\mathbf{e}_{i+1} (t) = \sum_{j=0}^{2} \mathbf{b}_{i+1,j} \frac{w_{i+1,j} B_j^2 (t)}{\sum_{k=0}^{2} w_{i+1,k} B_k^2 (t)}, \; t \in [0,1].$$

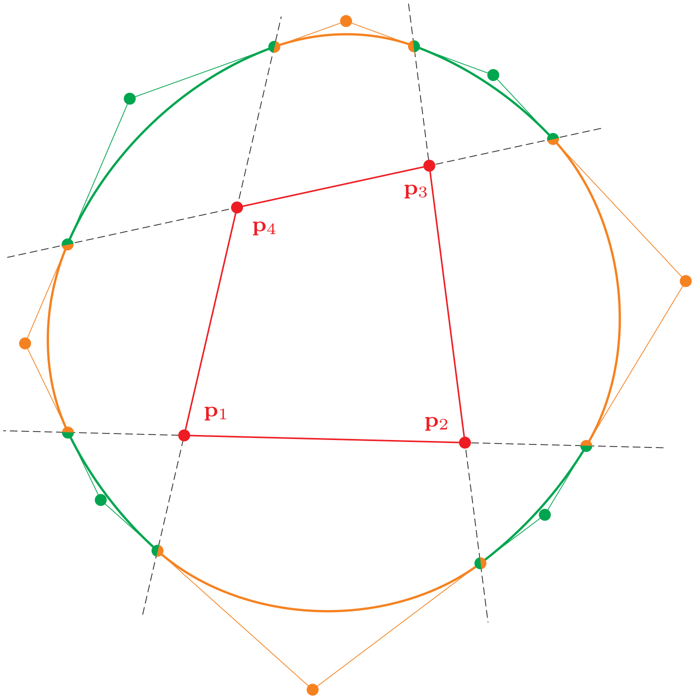


Figure 3: The gardener's spline curve of a quadrilateral, along with the control polygon of each quadratic Bézier curve that describe the elliptic arcs, from which the curve is composed of

For C^1 continuity conditions

$$\begin{aligned} \mathbf{e}_i\left(t\right)\big|_{t=1} &= \mathbf{e}_{i+1}\left(t\right)\big|_{t=0}\,, \\ \frac{\mathrm{d}}{\mathrm{d}\,t}\mathbf{e}_i\left(t\right)\bigg|_{t=1} &= \frac{\mathrm{d}}{\mathrm{d}\,t}\mathbf{e}_{i+1}\left(t\right)\bigg|_{t=0} \end{aligned}$$

have to be fulfilled. The first equality is guaranteed by the construction, only the second condition needs further study. Since

$$\begin{aligned} \frac{\mathrm{d}}{\mathrm{d}\,t}\mathbf{e}_i\left(t\right)\bigg|_{t=1} &= 2\frac{w_{i,1}}{w_{i,2}}\left(\mathbf{b}_{i,2}-\mathbf{b}_{i,1}\right), \\ \frac{\mathrm{d}}{\mathrm{d}\,t}\mathbf{e}_{i+1}\left(t\right)\bigg|_{t=0} &= 2\frac{w_{i+1,1}}{w_{i+1,0}}\left(\mathbf{b}_{i+1,1}-\mathbf{b}_{i+1,0}\right) \end{aligned}$$

the equality

$$\frac{w_{i,1}}{w_{i,2}}\left(\mathbf{b}_{i,2}-\mathbf{b}_{i,1}\right) = \frac{w_{i+1,1}}{w_{i+1,0}}\left(\mathbf{b}_{i+1,1}-\mathbf{b}_{i+1,0}\right) \tag{5.3}$$

has to be fulfilled for $i = 1, 2, \ldots, 2n$. If we have the standard form of weights, $w_{i,2} = w_{i+1,0} = 1$, $(i = 1, 2, \ldots, 2n)$ this equality will not hold in general.

We will apply a suitable projective transformation of these standard weights (cf. [6]), which is equivalent to a linear fractional transformation of the parameter. Thus we perform substitutions

$$w_{i,0} \to \alpha_i^2, w_{i,1} \to \alpha_i w_{i,1}, w_{i,2} \to 1,$$

where α_i, $(i = 1, 2, \ldots, 2n)$ are positive real values to be determined. Applying these substitutions in Eq. (5.3), we obtain equalities

$$\alpha_i w_{i,1} \left(\mathbf{b}_{i,2} - \mathbf{b}_{i,1}\right) = \frac{w_{i+1,1}}{\alpha_{i+1}} \left(\mathbf{b}_{i+1,1} - \mathbf{b}_{i+1,0}\right), \ (i = 1, 2, \ldots, 2n)$$

which results in the system of equations

$$\alpha_i \alpha_{i+1} = \beta_{i,i+1}, \ (i = 1, 2, \ldots, 2n) \tag{5.4}$$

for the unknowns $\alpha_i, (i = 1, 2, \ldots, 2n)$, where the positive constants $\beta_{i,i+1}$ are of the form

$$\beta_{i,i+1} = \frac{w_{i+1,1} \left\|\mathbf{b}_{i+1,1} - \mathbf{b}_{i+1,0}\right\|}{w_{i,1} \left\|\mathbf{b}_{i,2} - \mathbf{b}_{i,1}\right\|}, \ (i = 1, 2, \ldots, 2n).$$

The solution is

$$\begin{aligned}
\alpha_2 &= \frac{1}{\alpha_1} \beta_{1,2}, \\
\alpha_3 &= \alpha_1 \frac{\beta_{2,3}}{\beta_{1,2}}, \\
&\vdots \\
\alpha_{2n} &= \frac{1}{\alpha_1} \frac{\beta_{1,2}\beta_{3,4} \cdots \beta_{2n-1,2n}}{\beta_{2,3}\beta_{4,5} \cdots \beta_{2n-2,2n-1}}, \\
\alpha_1 &= \alpha_1 \frac{\beta_{2,3}\beta_{4,5} \cdots \beta_{2n-2,2n-1}\beta_{2n,1}}{\beta_{1,2}\beta_{3,4} \cdots \beta_{2n-1,2n}},
\end{aligned}$$

that is for the closed curve we have a solution if

$$\frac{\beta_{2,3}\beta_{4,5} \cdots \beta_{2n-2,2n-1}\beta_{2n,1}}{\beta_{1,2}\beta_{3,4} \cdots \beta_{2n-1,2n}} = 1$$

which is a very serious restriction. However, if we do not require C^1 joint of arcs $\mathbf{e}_1$ and $\mathbf{e}_{2n}$, i.e., if we discard equation

$$\alpha_{2n}\alpha_1 = \beta_{2n,1},$$

we always have solutions, where α_1 is a free parameter.

References

[1] Glaeser, G., Stachel, H., Odehnal, B., *The Universe of Conics: From the ancient Greeks to 21st century developments*, Springer, 2016.

[2] Khilji, M. J., Thatipur, D. G., Multi foci closed curves, *Journal of Theoretics* 6.

[3] O'Rourke, J., *Computational Geometry in C*, 2nd Edition, Cambridge University Press, 1998.

[4] de Berg, M., Cheong, O., van Kreveld, M., Overmars, M., *Computational Geometry: Algorithms and Applications*, 3rd Edition, Springer, 2008.

[5] Sánchez-Reyes, J., Harmonic rational Bézier curves, p-Bézier curves and trigonometric polynomials, *Computer Aided Geometric Design*, Vol. 15(1998) (9), 909–923.

[6] Patterson, R. R., Projective transformations of the parameter of a Bernstein-Bézier curve, *ACM Transactions on Graphics*, Vol. 4(1985) (4), 276–290.

Annales Mathematicae et Informaticae
47 (2017) pp. 119–128
http://ami.uni-eszterhazy.hu

The $h(x)$-Lucas quaternion polynomials

Nayil Kilic

Department of Mathematics
Sinop University, Sinop, Turkey
nayilkilic@gmail.com

Submitted May 16, 2017 — Accepted November 30, 2017

Abstract

In this paper, we study $h(x)$-Lucas quaternion polynomials considering several properties involving these polynomials and we present the exponential generating functions and the Poisson generating functions of the $h(x)$-Lucas quaternion polynomials. Also, by using Binet's formula we give the Cassini's identity, Catalan's identity and d'Ocagne's identity of the $h(x)$-Lucas quaternion polynomials.

Keywords: Lucas polynomials, recurrences, quaternion.

MSC: 11B39, 11B37, 11R52

1. Introduction

The Lucas sequence, $\{L_n\}$, is defined by the recurrence relation, for $n > 1$

$$L_{n+1} = L_n + L_{n-1}$$

where $L_0 = 2$, $L_1 = 1$.

In [13], Nalli and Haukkanen introduced the $h(x)$-Lucas polynomials.

Definition 1.1 ([13])**.** Let $h(x)$ be a polynomial with real coefficients. The $h(x)$-Lucas polynomials $\{L_{h,n}(x)\}_{n=0}^{\infty}$ are defined by the recurrence relation

$$L_{h,n+1}(x) = h(x)L_{h,n}(x) + L_{h,n-1}(x), n \geq 1, \tag{1.1}$$

with initial conditions $L_{h,0}(x) = 2$, $L_{h,1}(x) = h(x)$.

The quaternions are such numbers which extend the complex numbers. They are members of noncommutative algebra. A quaternion p is defined in the form

$$p = a_0 + a_1 i + a_2 j + a_3 k$$

where a_0, a_1, a_2 and a_3 are real numbers and i, j, k are standart orthonormal basis in $\mathbb{R}^3$ which satisfy the quaternion multiplication rules as

$$i^2 = j^2 = k^2 = -1, \quad ij = -ji = k, \quad jk = -kj = i \quad ki = -ik = j.$$

The conjugate of the quaternion p is denoted by $\overline{p}$ and $\overline{p} = a_0 - a_1 i - a_2 j - a_3 k$.

We start by recalling some basic results concerning quaternion algebra $\mathbb{H}$, it is well known that the algebra $\mathbb{H} = \{a = a_0 e_0 + a_1 e_1 + a_2 e_2 + a_3 e_3 | a_i \in \mathbb{R}, i = \{0, 1, 2, 3\}\}$ of real quaternions define a $four-$ dimensional vector space over $\mathbb{R}$ having basis $e_0 \cong 1$, $e_1 \cong i$, $e_2 \cong j$ and $e_3 \cong k$ which satisfies the following multiplication rules.

$$\begin{aligned} {e_s}^2 = -1, s \in \{1, 2, 3\}, \quad e_1 e_2 = -e_2 e_1 = e_3, \quad & e_2 e_3 = -e_3 e_2 = e_1, \\ & e_3 e_1 = -e_1 e_3 = e_2. \end{aligned} \tag{1.2}$$

In [8], Horodam defined the nth Lucas quaternions as follows.

Definition 1.2 ([8])**.** The Lucas quaternion numbers that are given for the nth classic Lucas L_n number are defined by the following recurrence relations:

$$T_n = L_n + iL_{n+1} + jL_{n+2} + kL_{n+3}$$

where $n = 0, \mp 1, \mp 2, \dots$.

The Lucas quaternions have been studied in several papers (see, for example [1, 2, 7, 10, 15]). Recently, in [2], Ari considered the $h(x)$-Lucas quaternion polynomials, he derived the Binet formula and generating function of $h(x)$-Lucas quaternion polynomial sequence.

In this paper, we study $h(x)$-Lucas quaternion polynomials considering several properties involving these polynomials and we present the exponential generating functions and the Poisson generating functions of the $h(x)$-Lucas quaternion polynomials. Also, by using Binet's formula we give the Cassini's identity, the Catalan's identity and the d'Ocagne's identity of the $h(x)$-Lucas quaternion polynomials.

2. The $h(x)$-Lucas quaternion polynomials and some properties

Let e_i $(i = 0, 1, 2, 3)$ be a basis of $\mathbb{H}$, which satisfy the multiplication rules (1.2). Let $h(x)$ be a polynomial with real coefficients. In [2], Ari introduced the $h(x)$-Lucas quaternion polynomials as follows:

Definition 2.1 ([2])**.** Let $h(x)$ be a polynomial with real coefficients. The $h(x)$-Lucas quaternion polynomials $\{T_{h,n}(x)\}_{n=0}^{\infty}$ are defined by the recurrence relation

$$T_{h,n}(x) = \sum_{s=0}^{3} L_{h,n+s}(x)e_s \tag{2.1}$$

where $L_{h,n}(x)$ is the nth $h(x)$-Lucas polynomial.

The conjugate of $T_{h,n}(x)$ is given by

$$\overline{T_{h,n}(x)} = L_{h,n}(x)e_0 - L_{h,n+1}(x)e_1 - L_{h,n+2}(x)e_2 - L_{h,n+3}(x)e_3.$$

For $n = 0$,

$$\begin{aligned} T_{h,0}(x) &= \sum_{s=0}^{3} L_{h,s}(x)e_s \\ &= L_{h,0}(x)e_0 + L_{h,1}(x)e_1 + L_{h,2}(x)e_2 + L_{h,3}(x)e_3 \\ &= 2e_0 + h(x)e_1 + (h^2(x) + 2)e_2 + (h^3(x) + 3h(x))e_3. \end{aligned}$$

For $n = 1$,

$$\begin{aligned} T_{h,1}(x) &= \sum_{s=0}^{3} L_{h,s+1}(x)e_s \\ &= L_{h,1}(x)e_0 + L_{h,2}(x)e_1 + L_{h,3}(x)e_2 + L_{h,4}(x)e_3 \\ &= h(x)e_0 + (h^2(x) + 2)e_1 + (h^3(x) + 3h(x))e_2 \\ &\quad + (h^4(x) + 4h^2(x) + 2)e_3. \end{aligned}$$

From the recurrence relation (2.1), using the recurrence relation (1.1) and some properties of summation formulas, we obtain that

$$\begin{aligned} T_{h,n+1}(x) &= \sum_{s=0}^{3} L_{h,s+1+n}(x)e_s \\ &= \sum_{s=0}^{3} \Big(h(x)L_{h,s+n}(x) + L_{h,s+n-1}(x)\Big)e_s \\ &= h(x)\sum_{s=0}^{3} L_{h,s+n}(x)e_s + \sum_{s=0}^{3} L_{h,s+n-1}(x)e_s \\ &= h(x)T_{h,n}(x) + T_{h,n-1}(x) \end{aligned}$$

and so

$$T_{h,n+1}(x) = h(x)T_{h,n}(x) + T_{h,n-1}(x).$$

In [13], authors studied some combinatorial properties of $h(x)$-Fibonacci and $h(x)$-Lucas polynomials and present properties of these polynomials. They obtained the following Binet's formula for $L_{h,n}(x)$

$$L_{h,n}(x) = \alpha^n(x) + \beta^n(x) \tag{2.2}$$

where

$$\alpha(x) = \frac{h(x) + \sqrt{h^2(x)+4}}{2}, \qquad \beta(x) = \frac{h(x) - \sqrt{h^2(x)+4}}{2} \tag{2.3}$$

are roots of the characteristic equation $y^2 - h(x)y - 1 = 0$ of the recurrence relation (1.1).

Ari in [2] calculated the Binet-style formula for $T_{h,n}(x)$,

$$T_{h,n}(x) = \alpha^*(x)\alpha^n(x) + \beta^*(x)\beta^n(x) \tag{2.4}$$

where $\alpha(x)$ and $\beta(x)$ as in (2.3) and $\alpha^*(x) = \sum\limits_{s=0}^{3} \alpha^s(x)e_s$, $\beta^*(x) = \sum\limits_{s=0}^{3} \beta^s(x)e_s$. The following basic identities are needed for our purpose in proving.

$$\alpha(x) + \beta(x) = h(x), \quad \alpha(x)\beta(x) = -1, \quad \alpha(x) - \beta(x) = \sqrt{h^2(x)+4} \tag{2.5}$$

and

$$\frac{\alpha(x)}{\beta(x)} = -\alpha^2(x), \qquad \frac{\beta(x)}{\alpha(x)} = -\beta^2(x).$$

Also,

$$1 + h(x)\alpha(x) = \alpha^2(x), \qquad 1 + h(x)\beta(x) = \beta^2(x), \tag{2.6}$$

and

$$1 + \alpha^2(x) = \alpha(x)\sqrt{h^2(x)+4}, \qquad 1 + \beta^2(x) = -\beta(x)\sqrt{h^2(x)+4}. \tag{2.7}$$

The following Lemma, related with the $h(x)$-Lucas polynomials and it will be useful in the proof of one property of the $h(x)$-Lucas quaternion polynomials in the next Theorem.

Lemma 2.2. *For $n \geq 0$,*

$$L^2{}_{h,n}(x) + L^2{}_{h,n+1}(x) = L_{h,2n}(x) + L_{h,2n+2}(x).$$

Proof. Using (2.2) and (2.5), we get

$$\begin{aligned} L^2{}_{h,n}(x) + L^2{}_{h,n+1}(x) &= (\alpha^n(x) + \beta^n(x))^2 + (\alpha^{n+1}(x) + \beta^{n+1}(x))^2 \\ &= \alpha^{2n}(x) + 2\alpha^n(x)\beta^n(x) + \beta^{2n}(x) \\ &\quad + \alpha^{2n+2}(x) + 2\alpha^{n+1}(x)\beta^{n+1}(x) + \beta^{2n+2}(x) \\ &= \alpha^{2n}(x) + \beta^{2n}(x) + \alpha^{2n+2}(x) + \beta^{2n+2}(x) \\ &= L_{h,2n}(x) + L_{h,2n+2}(x). \end{aligned}$$

So the proof is complete. □

Theorem 2.3. *For $n \geq 0$, the following statements hold:*

(*i*) $(T_{h,n}(x))^2+(T_{h,n+1}(x))^2 = (\alpha^{2*}(x)\alpha^{2n+1}(x)-\beta^{2*}(x)\beta^{2n+1}(x))(\alpha(x)-\beta(x))$.

(*ii*) $\frac{(T_{h,n}(x))^2+(T_{h,n+1}(x))^2}{(\alpha(x)-\beta(x))} = (\alpha^{2*}(x)\alpha^{2n+1}(x) - \beta^{2*}(x)\beta^{2n+1}(x))$.

(*iii*) $\overline{T_{h,n}(x)} + T_{h,n}(x) = 2L_{h,n}(x)e_0$.

(*iv*) $(T_{h,n}(x))^2 = 2L_{h,n}(x)e_0T_{h,n}(x) - T_{h,n}(x)\overline{T_{h,n}(x)} = T_{h,n}(x)(2L_{h,n}(x)e_0 - \overline{T_{h,n}(x)})$.

(*v*) $T_{h,n}(x)\overline{T_{h,n}(x)} = ((h(x))^2 + 2)(L_{h,2n+4}(x) + L_{h,2n+2}(x))$.

(*vi*) $T_{h,1}(x) - \alpha(x)T_{h,0}(x) = -\beta^*(x)\sqrt{h^2(x)+4}$.
In particular $\frac{T_{h,1}(x)-\alpha(x)T_{h,0}(x)}{\alpha(x)-\beta(x)} = -\beta^*(x)$.

(*vii*) $T_{h,1}(x) - \beta(x)T_{h,0}(x) = \alpha^*(x)\sqrt{h^2(x)+4}$.
In particular $\frac{T_{h,1}(x)-\beta(x)T_{h,0}(x)}{\alpha(x)-\beta(x)} = \alpha^*(x)$.

Proof. (*i*) From (2.4), (2.5) and (2.7), we obtain

$$\begin{aligned}
&(T_{h,n}(x))^2 + (T_{h,n+1}(x))^2\\
&= (\alpha^*(x)\alpha^n(x) + \beta^*(x)\beta^n(x))^2 + (\alpha^*(x)\alpha^{n+1}(x) + \beta^*(x)\beta^{n+1}(x))^2\\
&= \alpha^{2*}(x)\alpha^{2n}(x)(1+\alpha^2(x)) + \beta^{2*}(x)\beta^{2n}(x)(1+\beta^2(x))\\
&= \alpha^{2*}(x)\alpha^{2n+1}(x)\sqrt{h^2(x)+4} - \beta^{2*}(x)\beta^{2n+1}(x)\sqrt{h^2(x)+4}\\
&= (\alpha^{2*}(x)\alpha^{2n+1}(x) - \beta^{2*}(x)\beta^{2n+1}(x))(\alpha(x)-\beta(x)).
\end{aligned}$$

(*ii*) The proof of (*ii*) follows immediately from (*i*).

(*iii*) Using the definition of $\overline{T_{h,n}(x)}$ and some computations, we have

$$\begin{aligned}
\overline{T_{h,n}(x)} &= L_{h,n}(x)e_0 - L_{h,n+1}(x)e_1 - L_{h,n+2}(x)e_2 - L_{h,n+3}(x)e_3\\
&= 2L_{h,n}(x)e_0 - \sum_{s=0}^{3} L_{h,n+s}(x)e_s\\
&= 2L_{h,n}(x)e_0 - T_{h,n}(x),
\end{aligned}$$

and the result follows.

(*iv*) By (*iii*), (*iv*) holds.

(*v*) Using Definition 2.1, the definition of $\overline{T_{h,n}(x)}$, Lemma 2.2 and (1.1) we obtain

$$\begin{aligned}
T_{h,n}(x)\overline{T_{h,n}(x)} &= \sum_{s=0}^{3} L_{h,n+s}(x)e_s\overline{T_{h,n}(x)}\\
&= \big(L_{h,n}(x)e_0 + L_{h,n+1}(x)e_1 + L_{h,n+2}(x)e_2 + L_{h,n+3}(x)e_3\big)
\end{aligned}$$

$$\begin{aligned}&\times \big(L_{h,n}(x)e_0 - L_{h,n+1}(x)e_1 - L_{h,n+2}(x)e_2 - L_{h,n+3}(x)e_3\big)\\
&= {L^2}_{h,n}(x) + {L^2}_{h,n+1}(x) + {L^2}_{h,n+2}(x) + {L^2}_{h,n+3}(x)\\
&= L_{h,2n}(x) + L_{h,2n+2}(x) + L_{h,2n+4}(x) + L_{h,2n+6}(x)\\
&= L_{h,2n}(x) + L_{h,2n+2}(x) + L_{h,2n+4}(x) + h(x)L_{h,2n+5}(x)\\
&\quad + L_{h,2n+4}(x)\\
&= 2L_{h,2n+2}(x) + h^2(x)L_{h,2n+2}(x) + 2L_{h,2n+4}(x)\\
&\quad + h^2(x)L_{h,2n+4}(x)\\
&= \big(2 + (h(x))^2\big)\big(L_{h,2n+2}(x) + L_{h,2n+4}(x)\big).\end{aligned}$$

(vi) Since

$$L_{h,s+1}(x) - \beta(x)L_{h,s}(x) = \alpha^s(x)\big(\alpha(x) - \beta(x)\big)$$

and

$$L_{h,s+1}(x) - \alpha(x)L_{h,s}(x) = \beta^s(x)\big(\alpha(x) - \beta(x)\big),$$

using the definition of $\beta^*(x)$, Definition2.1 and Eq.(2.5), we have

$$\begin{aligned}&T_{h,1}(x) - \alpha(x)T_{h,0}(x)\\
&= L_{h,1}(x)e_0 + L_{h,2}(x)e_1 + L_{h,3}(x)e_2 + L_{h,4}(x)e_3\\
&\quad - \alpha(x)\big(L_{h,0}(x)e_0 + L_{h,1}(x)e_1 + L_{h,2}(x)e_2 + L_{h,3}(x)e_3\big)\\
&= \big(L_{h,1}(x) - \alpha(x)L_{h,0}(x)\big)e_0 + \big(L_{h,2}(x) - \alpha(x)L_{h,1}(x)\big)e_1\\
&\quad + \big(L_{h,3}(x) - \alpha(x)L_{h,2}(x)\big)e_2 + \big(L_{h,4}(x) - \alpha(x)L_{h,3}(x)\big)e_3\\
&= -\beta^0(x)(\alpha(x) - \beta(x))e_0 - \beta^1(x)(\alpha(x) - \beta(x))e_1\\
&\quad - \beta^2(x)(\alpha(x) - \beta(x))e_2 - \beta^3(x)(\alpha(x) - \beta(x))e_3\\
&= -\sqrt{h^2(x)+4}(e_0 + \beta^1(x)e_1 + \beta^2(x)e_2 + \beta^3(x)e_3)\\
&= -\sqrt{h^2(x)+4}\sum_{s=0}^{3}\beta^s(x)e_s\\
&= -\sqrt{h^2(x)+4}\beta^*(x).\end{aligned}$$

which completes the first part of the proof of (vi). The proof of the remaining part can be obtained from previous result.

(vii) The proof is similar to part (vi) and thus, omitted. □

Theorem 2.4. *For $n \geq 0$, $\sum\limits_{k=0}^{n}\binom{n}{k}(h(x))^k T_{h,k}(x) = T_{h,2n}(x)$.*

Proof. Using (2.4) and (2.6), we obtain

$$\sum_{k=0}^{n}\binom{n}{k}(h(x))^k T_{h,k}(x) = \sum_{k=0}^{n}\binom{n}{k}(h(x))^k[\alpha^*(x)\alpha^k(x) + \beta^*(x)\beta^k(x)]$$

$$\begin{aligned}
&= \alpha^*(x)\sum_{k=0}^{n}\binom{n}{k}(h(x))^k\alpha^k(x)\\
&\quad + \beta^*(x)\sum_{k=0}^{n}\binom{n}{k}(h(x))^k\beta^k(x)\\
&= \alpha^*(x)(1+h(x)\alpha(x))^n + \beta^*(x)(1+h(x)\beta(x))^n\\
&= \alpha^*(x)\alpha^{2n}(x) + \beta^*(x)\beta^{2n}(x)\\
&= T_{h,2n}(x).
\end{aligned}$$

□

Theorem 2.5. *The sum of the first m terms of the sequence $\{T_{h,m}(x)\}_{m=0}^{\infty}$ is given by*

$$\sum_{k=0}^{m} T_{h,k}(x) = \frac{T_{h,0}(x) - T_{h,m}(x) - T_{h,m+1}(x) - \alpha^*(x)\beta(x) - \beta^*(x)\alpha(x)}{(1-\alpha(x))(1-\beta(x))}.$$

Proof. From (2.4), (2.5) and some calculations, we get

$$\begin{aligned}
\sum_{k=0}^{m} T_{h,k}(x) &= \sum_{k=0}^{m}(\alpha^*(x)\alpha^k(x) + \beta^*(x)\beta^k(x))\\
&= \alpha^*(x)\sum_{k=0}^{m}\alpha^k(x) + \beta^*(x)\sum_{k=0}^{m}\beta^k(x)\\
&= \alpha^*(x)\Big(\frac{1-\alpha^{m+1}(x)}{1-\alpha(x)}\Big) + \beta^*(x)\Big(\frac{1-\beta^{m+1}(x)}{1-\beta(x)}\Big)\\
&= \frac{\alpha^*(x) - \alpha^*(x)\beta(x) - \alpha^*(x)\alpha^{m+1}(x) + \alpha^*(x)\alpha^m(x)\alpha(x)\beta(x)}{(1-\beta(x))(1-\alpha(x))}\\
&\quad + \frac{\beta^*(x) - \beta^*(x)\alpha(x) - \beta^*(x)\beta^{m+1}(x) + \beta^*(x)\alpha(x)\beta(x)\beta^m(x)}{(1-\beta(x))(1-\alpha(x))}\\
&= \frac{T_{h,0}(x) - T_{h,m}(x) - T_{h,m+1}(x) - \alpha^*(x)\beta(x) - \beta^*(x)\alpha(x)}{(1-\alpha(x))(1-\beta(x))}.
\end{aligned}$$

So the proof is complete. □

3. Exponential generating functions for the $h(x)$-Lucas quaternion polynomials

In this section, we give the exponential generating functions for the sequence of the $h(x)$-Lucas quaternion polynomials. The exponential generating function of a sequence $\{b_k\}_{k=0}^{\infty}$ is given by

$$EG(b_k, l) = \sum_{k=0}^{\infty} b_k \frac{l^k}{k!}.$$

Theorem 3.1. *The exponential generating function for the $h(x)$-Lucas quaternion polynomials are*

$$\sum_{k=0}^{\infty} \frac{T_{h,k}(x)}{k!} l^k = \alpha^*(x) e^{\alpha(x) l} + \beta^*(x) e^{\beta(x) l}. \tag{3.1}$$

Proof. From the Binet-style formula for the $h(x)$-Lucas quaternion polynomials, we have

$$\begin{aligned}\sum_{k=0}^{\infty} \frac{T_{h,k}(x)}{k!} l^k &= \sum_{k=0}^{\infty} \Big(\alpha^*(x)\alpha^k(x) + \beta^*(x)\beta^k(x)\Big)\frac{l^k}{k!} \\ &= \alpha^*(x) \sum_{k=0}^{\infty} \frac{(\alpha(x) l)^k}{k!} + \beta^*(x) \sum_{k=0}^{\infty} \frac{(\beta(x) l)^k}{k!} \\ &= \alpha^*(x) e^{\alpha(x) l} + \beta^*(x) e^{\beta(x) l}.\end{aligned}$$

□

4. Poisson generating functions for the $h(x)$-Lucas quaternion polynomials

In this section, we present Poisson generating functions for the sequence of the $h(x)$-Lucas quaternion polynomials.

Lemma 4.1. *The Poisson generating functions for the $h(x)$-Lucas quaternion polynomials are*

$$\sum_{k=0}^{\infty} \frac{T_{h,k}(x)}{k!} l^k e^{-l} = \frac{\alpha^*(x) e^{\alpha(x) l} + \beta^*(x) e^{\beta(x) l}}{e^l}. \tag{4.1}$$

Proof. Since $PG(b_n, x) = e^{-l} EG(b_n, x)$, we have the result by Theorem 3.1. □

5. Catalan's, Cassini's and d'Ocagne's identity for the $h(x)$-Lucas quaternion polynomials

In this section, we compute Catalan's identity, Cassini's identity and d'Ocagne's identity for the $h(x)$-Lucas quaternion polynomials, we start with Catalan's identity.

Theorem 5.1. *For $n \geq m \geq 1$, Catalan identity for the $h(x)$-Lucas quaternion polynomials is*

$$\begin{aligned}T_{h,n+m}(x) T_{h,n-m}(x) - T^2{}_{h,n}(x) &= (-1)^{n-m}(\alpha^m(x) - \beta^m(x)) \\ &\quad \times 1\Big(\alpha^*(x)\beta^*(x)\alpha^m(x) - \beta^*(x)\alpha^*(x)\beta^m(x)\Big).\end{aligned}$$

Proof. Using (2.4) and (2.5), we obtain

$$\begin{aligned}
&T_{h,n+m}(x)T_{h,n-m}(x) - T^2{}_{h,n}(x)\\
&= \Big(\alpha^*(x)\alpha^{n+m}(x) + \beta^*(x)\beta^{n+m}(x)\Big)\Big(\alpha^*(x)\alpha^{n-m}(x) + \beta^*(x)\beta^{n-m}(x)\Big)\\
&\quad - \Big(\alpha^*(x)\alpha^n(x) + \beta^*(x)\beta^n(x)\Big)^2\\
&= \alpha^*(x)\beta^*(x)\alpha^{n+m}(x)\beta^{n-m}(x) + \beta^*(x)\alpha^*(x)\beta^{n+m}(x)\alpha^{n-m}(x)\\
&\quad - \alpha^*(x)\beta^*(x)\alpha^n(x)\beta^n(x) - \beta^*(x)\alpha^*(x)\alpha^n(x)\beta^n(x)\\
&= \alpha^*(x)\beta^*(x)\big(\alpha(x)\beta(x)\big)^n\big(\frac{\alpha^m(x)}{\beta^m(x)} - 1\big)\\
&\quad + \beta^*(x)\alpha^*(x)\big(\alpha(x)\beta(x)\big)^n\big(\frac{\beta^m(x)}{\alpha^m(x)} - 1\big)\\
&= \alpha^*(x)\beta^*(x)(-1)^n\alpha^m(x)\big(\frac{\alpha^m(x) - \beta^m(x)}{(\alpha(x)\beta(x))^m}\big)\\
&\quad + \beta^*(x)\alpha^*(x)(-1)^n\beta^m(x)\big(\frac{\beta^m(x) - \alpha^m(x)}{(\alpha(x)\beta(x))^m}\big)\\
&= (-1)^{n-m}(\alpha^m(x) - \beta^m(x))\Big(\alpha^*(x)\beta^*(x)\alpha^m(x) - \beta^*(x)\alpha^*(x)\beta^m(x)\Big).
\end{aligned}$$

So Theorem 5.1 is proved. □

Theorem 5.2. *For any natural number n, Cassini identity for the $h(x)$-Lucas quaternion polynomials is*

$$\begin{aligned}
T_{h,n+1}(x)T_{h,n-1}(x) - T^2{}_{h,n}(x) &= (-1)^{n-1}(\alpha(x) - \beta(x))\\
&\quad \times \Big(\alpha^*(x)\beta^*(x)\alpha(x) - \beta^*(x)\alpha^*(x)\beta(x)\Big).
\end{aligned}$$

Proof. Taking $m = 1$ in Catalan's identity, the proof is completed. □

Theorem 5.3 (d'Ocagne's identity)**.** *Suppose that n is a nonnegative integer number and m any natural number. If $m > n$, then*

$$\begin{aligned}
&T_{h,m}(x)T_{h,n+1}(x) - T_{h,m+1}(x)T_{h,n}(x)\\
&= (-1)^n(\alpha(x) - \beta(x))\Big(\beta^*(x)\alpha^*(x)\beta^{m-n}(x) - \alpha^*(x)\beta^*(x)\alpha^{m-n}(x)\Big).
\end{aligned}$$

Proof. From (2.4) and (2.5), we obtain

$$\begin{aligned}
&T_{h,m}(x)T_{h,n+1}(x) - T_{h,m+1}(x)T_{h,n}(x)\\
&= \Big(\alpha^*(x)\alpha^m(x) + \beta^*(x)\beta^m(x)\Big)\Big(\alpha^*(x)\alpha^{n+1}(x) + \beta^*(x)\beta^{n+1}(x)\Big)\\
&\quad - \Big(\alpha^*(x)\alpha^{m+1}(x) + \beta^*(x)\beta^{m+1}(x)\Big)\Big(\alpha^*(x)\alpha^n(x) + \beta^*(x)\beta^n(x)\Big)\\
&= \alpha^*(x)\beta^*(x)\alpha^m(x)\beta^n(x)\Big(\beta(x) - \alpha(x)\Big) + \beta^*(x)\alpha^*(x)\beta^m(x)\alpha^n(x)
\end{aligned}$$

$$\begin{aligned}
&\times \Big(\alpha(x)-\beta(x)\Big)\\
&=\alpha^*(x)\beta^*(x)\alpha^{m-n}(x)\Big(\alpha(x)\beta(x)\Big)^n\Big(\beta(x)-\alpha(x)\Big)+\beta^*(x)\alpha^*(x)\beta^{m-n}(x)\\
&\times \Big(\alpha(x)\beta(x)\Big)^n\Big(\alpha(x)-\beta(x)\Big)\\
&=(-1)^n(\alpha(x)-\beta(x))\Big(\beta^*(x)\alpha^*(x)\beta^{m-n}(x)-\alpha^*(x)\beta^*(x)\alpha^{m-n}(x)\Big).
\end{aligned}$$

So, the proof is complete. □

References

[1] Akyigit, M., Kosal, H. H., Tosun, M., Fibonacci Generalized Quaternions, *Advances in Applied Clifford Algebras*, Vol. 24(3) (2014), 631–641.

[2] Ari, K., On $h(x)$-Lucas quaternion polynomials, *Ars Combinatoria*, Vol. 121 (2015), 291–303.

[3] Catarino, P., A note on $h(x)-$ Fibonacci quaternion polynomials, *Chaos, Solitons Fractals*, Vol. 77 (2015), 1–5.

[4] Catarino, P., A note on certain matrices with $h(x)-$ Fibonacci quaternion polynomials, *Journal of Difference Equations and Applications*, Vol. 22(2) (2016), 343–351.

[5] Catarino, P., The $h(x)-$ Fibonacci Quaternion Polynomials: Some Combinatorial Properties, *Advances in Applied Clifford Algebras*, Vol. 26 (2016), 71–79.

[6] Falcón, S., On the generating matrices of the $k-$ Fibonacci numbers, *Proyecciones J. Math.* Vol. 32(4) (2013), 347–357.

[7] Halici, S., On Fibonacci quaternions, *Advances in Applied Clifford Algebras*, Vol. 22(2) (2012), 321–327.

[8] Horodam, A. F., Complex Fibonacci numbers and Fibonacci quaternions, *Am. Math. Mon.* Vol. 70 (1963), 289–291.

[9] Horodam, A. F., Recurrence relations, *Ulam Q.* Vol. 2(2) (1993), 23–33.

[10] Iyer, M. R., Some results on Fibonacci quaternions, *Fibonacci Q.* Vol. 7(2) (1969), 201–210.

[11] Iyer, M. R., Note on Fibonacci quaternions, *Fibonacci Q.* **7**(3) (1969), 225–229.

[12] Kilic, E., Tasci, D., Haukkanen, P., On the generalized Lucas sequences by Hessenberg matrices, *Ars Combinatoria*, Vol. 95 (2010), 383–395.

[13] Nalli, A., Haukkanen, P., On generalized Fibonacci and Lucas polynomials, *Chaos, Solitons Fractals*, Vol. 42 (2009), 3179–3186.

[14] Polatli, E., Kesim, S., A Note on Catalan's identity for the k-Fibonacci quaternions, *Journal of Integer Sequences*, Vol. 18 (2015), 1–4.

[15] Tan, E., Yilmaz, S., Sahin, M., On a new generalization of Fibonacci Quaternions, *Chaos, Solitons Fractals*, Vol. 82 (2016), 1–4.

Annales Mathematicae et Informaticae
47 (2017) pp. 129–139
http://ami.uni-eszterhazy.hu

Hyperbolic distance between hyperbolic lines

Riku Klén

Department of Mathematics and Statistics, University of Turku, Finland
The Institute of Natural and Mathematical Sciences, Massey University, New Zealand
riku.klen@utu.fi

Submitted June 2, 2016 — Accepted November 20, 2017

Abstract

We derive formulas for the hyperbolic distance between hyperbolic lines in the unit disk and in the upper half plane. We also build an algorithm in MATLAB/Octave to compute the hyperbolic distance.

Keywords: algorithms, hyperbolic geometry, hyperbolic distance, Poincaré model

MSC: 51M10

1. Introduction

The hyperbolic geometry was founded in the 19th century as an answer to the two millenniums old question about the parallel postulate. The hyperbolic geometry shows that the parallel postulate cannot be derived from the other four Euclid's postulates. The hyperbolic geometry has turned out to be a very useful tool in geometric function theory [9] and many applications including cosmology [1], Einstein's theory of general relativity [4, 8] and celestial mechanics [5].

The basic models of the hyperbolic geometry are the unit ball and the upper half space models. These models can be used to obtain geometry on any plane domain with at least 2 boundary points via the Riemann mapping theorem. Despite the hyperbolic geometry has many applications, some of the elementary properties has not been implemented to algorithms. In this article we consider one of these, namely the hyperbolic distance between two lines. We introduce an algorithm for the hyperbolic distance between two hyperbolic lines in the unit disk (Algorithm 1) and in the upper half plane (Algorithm 2).

2. Preliminary results

In this section we introduce notation and preliminary results. For basics of the hyperbolic geometry we refer reader to [2] and [3]. We denote the Euclidean n-space by $\mathbb{R}^n$, $n \geq 2$, and identify $\mathbb{R}^2$ with the complex plane $\mathbb{C}$.

For $x \in \mathbb{R}^n$ and $r > 0$ we denote Euclidean sphere with center x and radius r by $S^{n-1}(x, r) = \{y \in \mathbb{R}^n \colon |x - y| = r\}$.

When $a, b \in \mathbb{R}$, $a < b$, we denote open and closed intervals by $(a, b) = \{z \in \mathbb{R} \colon a < z < b\}$ and $[a, b] = \{z \in \mathbb{R} \colon a \leq z \leq b\}$. For half-open intervals we use notation $(a, b]$ and $[a, b)$. If $x, y \in \mathbb{R}^n$, $x \neq y$ and $n \geq 2$, we denote the closed Euclidean line segment by $[x, y] = \{z \in \mathbb{R}^n \colon z = x + t(y - x),\ t \in [0, 1]\}$. If one or both of the end points are not included in the line segment, we use notation $(x, y]$, $[x, y)$ or (x, y).

We define the upper half space by

$$\mathbb{H}^n = \{x = (x_1, x_2, \ldots, x_n) \in \mathbb{R}^n \colon x_n > 0\}$$

and the unit ball by

$$\mathbb{B}^n = \{x \in \mathbb{R}^n \colon |x| < 1\}.$$

Next we define the hyperbolic distance in these two domains. For $x, y \in \mathbb{H}^n$

$$\rho_{\mathbb{H}^n}(x, y) = \operatorname{arc\,cosh}\left(1 + \frac{|x - y|^2}{2x_n y_n}\right). \tag{2.1}$$

For $x, y \in \mathbb{B}^n$

$$\rho_{\mathbb{B}^n}(x, y) = 2 \operatorname{arc\,sinh} \frac{|x - y|}{\sqrt{1 - |x|^2}\sqrt{1 - |y|^2}}. \tag{2.2}$$

For noncollinear $a, b, c \in \mathbb{R}^n$ there exist a unique circle $C \subset \mathbb{R}^n$ containing the three points. We denote the center of the circle C by $\operatorname{center}(a, b, c)$. If $a, b, c \in \mathbb{C}$ then $\operatorname{center}(a, b, c)$ can be found by the following formula.

Lemma 2.1 ([6, Proposition 2.2]). *Let $C \subset \mathbb{C}$ be a circle and $a, b, c \in C$ be distinct. Then the center of C is*

$$\operatorname{center}(a, b, c) = \frac{(b - c)|a|^2 + (c - a)|b|^2 + (a - b)|c|^2}{(b - c)\bar{a} + (c - a)\bar{b} + (a - b)\bar{c}}.$$

Let $S \subset \mathbb{C}$ be a circular arc and C be the circle that contains S. We denote the center of C by $\operatorname{center}(S)$.

In the unit ball $\mathbb{B}^n$ for $z \in \mathbb{B}^n$ we can define a useful Möbius mapping T_z as in [10, 1.34]. For all $x \in \mathbb{B}^n$ define

$$T_z(x) = (p_z \circ q_z)(x), \tag{2.3}$$

where

$$q_z(x) = \frac{z}{|z|^2} + \left(\frac{1}{|z|^2} - 1\right)\left(x - \frac{z}{|z|^2}\right) \Big/ \left|x - \frac{z}{|z|^2}\right|^2$$

and

$$p_z(x) = x - 2x\frac{z^2}{|z|^2}.$$

Geometrically q_z is the inversion in sphere $S^{n-1}(z/|z^2|, 1/|z|^2 - 1)$ and p_z is the reflection in the $n-1$-dimensional hyperplane through 0, which is perpendicular to the line that contains 0 and z.

A useful property of the mapping T_z is the fact that it is a Möbius mapping and thus it preserves the hyperbolic distance in $\mathbb{B}^n$: For all $x, y \in \mathbb{B}^n$

$$\rho_{\mathbb{B}^n}(x, y) = \rho_{\mathbb{B}^n}(T_z(x), T_z(y)).$$

Since hyperbolic lines in the upper half space and the unit ball are arcs of Euclidean circles, we need repeatedly to find intersection points of two circles. Mathematically this is very straightforward and a solution is obtained by solving a pair of equations. Algorithmically this is also very simple, for example there are functions in MATLAB (function `circcirc`) and Octave (function `intersectCircles`). Our algorithms are independent of programming language and thus we introduce the formula for finding the intersection of two circles in the complex plane.

Let $C_1 = S^1(x, r)$ be circle with center $x \in \mathbb{C}$ and radius $r > 0$, and $C_2 = S^1(y, s)$ be circle with center $y \in \mathbb{C}$ and radius $s > 0$. If $r + s < |x - y|$ or $|x - y| + \min\{r, s\} < \max\{r, s\}$, then $C_1 \cap C_2 = \emptyset$.

We assume that $r + s < |x - y| < \max\{r, s\} - \min\{r, s\}$. Now $C_1 \cap C_2 \neq \emptyset$ and we derive a formula for the intersection points v. Let $v \in C_1 \cap C_2$ and choose a point z from the Euclidean line through x and y such that (x, v, z) and (y, z, v) form two right-angled triangle with the right-angle at z. If we denote $|v - z| = h$ and $|x - z| = t$, then $|y - z| = |x - y| - t$ and by the Pythagorean theorem

$$r^2 = h^2 + t^2 \quad \text{and} \quad s^2 = (|x - y| - t)^2 + h^2.$$

Now $h = \sqrt{r^2 - t^2}$ and

$$h^2 = r^2 - t^2 = s^2 - (|x - y| - t)^2,$$

which is equivalent to

$$t = \frac{r^2 - s^2 + |x - y|^2}{2|x - y|}.$$

We obtain

$$z = x + (y - x)\frac{t}{|x - y|} = x + (y - x)\frac{r^2 - s^2 + |x - y|^2}{2|x - y|^2}$$

and

$$v = z \pm i(x - y)\frac{h}{|x - y|} = z \pm i(x - y)\frac{\sqrt{4r^2|x - y|^2 - (r^2 - s^2 + |x - y|^2)^2}}{2|x - y|^2}.$$

Finally, we introduce an elementary lemma, which we need for our algorithm in the unit ball.

Lemma 2.2. *The function*

$$f(\alpha)=\frac{\frac{a}{\cos(\pi-\alpha)}-a}{1-(1-a\tan(\pi-\alpha))^2}$$

is decreasing on $(0,\pi)$ *and* $f(\alpha)\to 0$ *as* $\alpha\to\pi$*.*

Proof. By differentiation we obtain

$$f'(\alpha)=-\frac{2a\sin\alpha+a\sin(2\alpha)+2\cos(2\alpha)+2}{2(1-\cos\alpha)(a\sin(\alpha)+2\cos\alpha)^2}$$

for $\alpha\in(0,\pi)$. We observe that $f'(\alpha)<0$ implying $f(\alpha)$ is decreasing, because $a\sin(2\alpha)=2a\sin\alpha\cos\alpha$ and thus

$$2a\sin\alpha+a\sin(2\alpha)+2\cos(2\alpha)+2=2a\sin\alpha(1+\cos\alpha)+2(1+\cos(2\alpha))\geq 0$$

for $\alpha\in(0,\pi)$.

We denote $\beta=\pi-\alpha$ and calculate using l'Hospital's rule

$$\begin{aligned}\lim_{\alpha\to\pi}f(\alpha)=\lim_{\beta\to 0}f(\beta)&=\lim_{\beta\to 0}\frac{\dfrac{2a\sin\beta}{1+\cos(2\beta)}}{\dfrac{2a(1-a\tan\beta)}{(\cos\beta)^2}}\\&=\lim_{\beta\to 0}\frac{2a\sin\beta(\cos\beta)^2}{(1+\cos(2\beta))(2a(1-a\tan\beta))}=0\end{aligned}$$

and the assertion follows. □

3. The upper half plane

Let $a,b\in\mathbb{H}^2$ be two distinct points. If a, b and $\overline{a}$ are collinear, then $\mathrm{Re}(a)=\mathrm{Re}(b)$ and the hyperbolic line through a and b is the Euclidean ray

$$\{z\in\mathbb{H}^2\colon z=(\mathrm{Re}(a),t),\ t>0\}. \tag{3.1}$$

If a, b and $\overline{a}$ are not collinear the hyperbolic line through a and b is the Euclidean semicircle

$$S^1(c,|a-c|)\cap\mathbb{H}^2,\quad c=\mathrm{center}(a,b,\overline{a}), \tag{3.2}$$

where the function center is defined in Lemma 2.1 and $c\in\partial\mathbb{H}^2$.

We derive next a formula for the hyperbolic distance between two hyperbolic lines.

Let $l_1, l_2 \subset \mathbb{H}^2$ be two distinct hyperbolic lines. If $l_1 \cap l_2 \neq \emptyset$ or both l_1 and l_2 are Euclidean rays as in (3.1), then $\rho_{\mathbb{H}^2}(l_1, l_2) = 0$. The latter one can be seen by selecting $x \in l_1$ and $y \in l_2$ with $\mathrm{Im}(x) = \mathrm{Im}(y) = t > 0$. Now (see Figure 1)

$$\rho_{\mathbb{H}^2}(l_1, l_2) \leq \rho_{\mathbb{H}^2}(x, y) = \operatorname{arc\,cosh}\left(1 + \frac{\mathrm{Re}(x-y)^2}{2t^2}\right) \to 0 \tag{3.3}$$

as $t \to 0$.

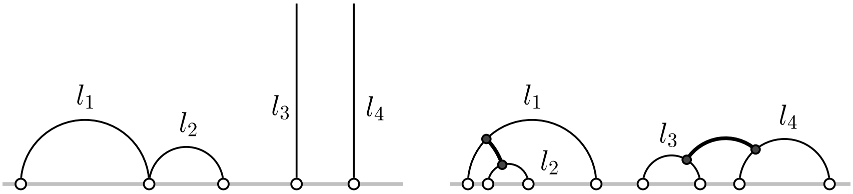


Figure 1: Left: Hyperbolic lines l_1, l_2, l_3 and l_4 with $\rho_{\mathbb{H}^2}(l_1, l_2) = 0 = \rho_{\mathbb{H}^2}(l_3, l_4)$ as in Proposition 3.1. Here $l_1 \cap l_2 = \emptyset$ but $\overline{l_1} \cap \overline{l_2} \neq \emptyset$ and l_3 and l_4 are as in (3.3). Right: Proposition 3.2 where the black points indicate the points p_i that give $\rho_{\mathbb{H}^2}(l_1, l_2)$ and $\rho_{\mathbb{H}^2}(l_3, l_4)$.

We assume that at least one of l_1 and l_2 is a semicircle of type (3.2). Let us now assume that $\overline{l_1} \cap \overline{l_2} = \{d\} \subset \partial\mathbb{H}^2$ and $l_1 \cap l_2 = \emptyset$. To simplify notation, we may assume that $d = 0$. We first assume that both l_1 and l_2 are Euclidean semicircles of type (3.2). We denote $l_1 \subset S^1((-r, 0), r)$ for some $r > 0$ and $l_2 \subset S^1((\pm s, 0), s)$ for some $s > r$. We choose $x \in l_1$ to be $x = r(e^{\alpha i} - 1)$ and $y \in l_2$ to be $y = s(e^{\alpha i} - 1)$ or $y = s(e^{(\pi-\alpha)i} + 1)$ for some $\alpha \in (0, \pi/2)$. Now

$$|x - y| \leq r - r\cos\alpha + s - s\cos\alpha, \quad x_2 = r\sin\alpha, \quad y_2 = s\sin\alpha$$

and thus by (2.1) we obtain

$$\rho_{\mathbb{H}^2}(l_1, l_2) \leq \rho_{\mathbb{H}^2}(x, y) \leq \operatorname{arc\,cosh}\left(1 + \frac{(r+s)^2\cos^2\alpha}{2rs\sin^2\alpha}\right) \to 0$$

as $\alpha \to 0$.

At least one of l_1 and l_2 has to be a Euclidean semicircle and the case that the other one is a Euclidean ray can be considered similarly as above. Let l_1 and l_2 be as above with $\mathrm{center}(l_2) = (s, 0)$. Denote the hyperbolic line that is the Euclidean ray by $l_2' = \{z \in \mathbb{H}^2 \colon \mathrm{Re}(z) = 0\}$. Then it is clear that

$$\rho_{\mathbb{H}^2}(l_1, l_2') \leq \rho_{\mathbb{H}^2}(l_1, l_2) \to 0, \quad \text{as } \alpha \to 0$$

and we conclude that for any two hyperbolic lines l_1 and l_2 in the case $\overline{l_1} \cap \overline{l_2} \neq \emptyset$ we have $\rho_{\mathbb{H}^2}(l_1, l_2) = 0$.

Proposition 3.1. *If l_1 and l_2 are hyperbolic lines in $\mathbb{H}^2$ with $\overline{l_1} \cap \overline{l_2} \neq \emptyset$, then $\rho_{\mathbb{H}^2}(l_1, l_2) = 0$.*

Next we assume, that $\overline{l_1} \cap \overline{l_2} = \emptyset$. Now by (3.3) at least one of the hyperbolic lines has to be a Euclidean semicircle.

We first assume that both l_1 and l_2 are Euclidean semicircles. We denote $l_1 \subset S^1((x,0),r)$ and $l_2 \subset S^1((y,0),s)$ for $x, y \in \mathbb{R}$, $x \neq y$, and $r, s > 0$ with $r > |x-y|$ and $s < r-|x-y|$. Let u be the radius and z the center of the Euclidean semicircle that is perpendicular to l_1 and l_2. By the Pythagorean theorem

$$u^2 = |x-z|^2 - r^2 = (|x-z| - |x-y|)^2 - s^2$$

and thus

$$|x-z| = \frac{r^2 - s^2 + |x-y|^2}{2|x-y|}, \quad u = \sqrt{|x-z|^2 - r^2}. \tag{3.4}$$

Now the circle

$$C_2 = \begin{cases} S^1((x-|x-z|,0),u), & \text{if } y < x, \\ S^1((x+|x-z|,0),u), & \text{if } x < y, \end{cases} \tag{3.5}$$

is perpendicular to both l_1 and l_2.

Proposition 3.2. *If l_1 and l_2 are hyperbolic lines of type* (3.2) *in* $\mathbb{H}^2$ *with* $\overline{l_1} \cap \overline{l_2} = \emptyset$ *and* $\text{center}(l_1) \neq \text{center}(l_2)$. *Then* $\rho_{\mathbb{H}^2}(l_1,l_2) = \rho_{\mathbb{H}^2}(p_1,p_2)$, *where* $\{p_1\} = l_1 \cap C_2$ *and* $\{p_2\} = l_2 \cap C_2$. *Here* C_2 *is as in* (3.5) *and* (3.4).

Next we deal with the case $x = y$. Let l_1 and l_2 be hyperbolic lines of type (3.2) in $\mathbb{H}^2$ with $\overline{l_1} \cap \overline{l_2} = \emptyset$ and $l_1 \subset S^1((x,0),r)$ and $l_2 \subset S^1((x,0),s)$ for $x \in \mathbb{R}$ and $r, s > 0$. Now (see Figure 2)

$$\rho_{\mathbb{H}^2}(l_1,l_2) = \rho_{\mathbb{H}^2}((x,r),(x,s)). \tag{3.6}$$

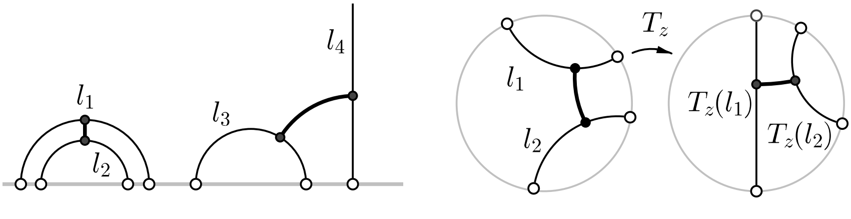


Figure 2: Left: Formula (3.6) for l_1 and l_2. Proposition 3.3 for l_3 and l_4. The black points indicate the points p_i that give $\rho_{\mathbb{H}^2}(l_1,l_2)$ and $\rho_{\mathbb{H}^2}(l_3,l_4)$. Right: Function T_z can be used to map l_1 to a Euclidean line segment. Proposition 4.1 gives the hyperbolic distance $\rho_{\mathbb{B}^2}(T_z(l_1),T_z(l_2)) = \rho_{\mathbb{B}^2}(l_1,l_2)$. The black points indicate the points p_i that give $\rho_{\mathbb{B}^2}(l_1,l_2)$ and $\rho_{\mathbb{B}^2}(T_z(l_1),T_z(l_2))$.

We then assume that l_1 is a Euclidean ray and l_2 is a Euclidean semicircle. We denote $l_1 = \{z \in \mathbb{H}^2 \colon z = (x,t),\, t > 0\}$ and $l_2 \subset S^1((y,0),r)$ for $x, y \in \mathbb{R}$ and $0 < r < |x-y|$. By the Pythagorean theorem we obtain that the circle

$$C_3 = S^1((x,0), \sqrt{|x-y|^2 - r^2}) \tag{3.7}$$

is perpendicular to l_1 and l_2.

Proposition 3.3. *Let l_1 and l_2 be hyperbolic lines in $\mathbb{H}^2$ with $\overline{l_1} \cap \overline{l_2} = \emptyset$. If l_1 is of type* (3.1) *and l_2 is of type of type* (3.2)*, then $\rho_{\mathbb{H}^2}(l_1, l_2) = \rho_{\mathbb{H}^2}(p_1, p_2)$, where $\{p_1\} = l_1 \cap C_3$ and $\{p_2\} = l_2 \cap C_3$. Here C_3 is as in* (3.7)*.*

Putting (3.6) and the results of Propositions 3.1, 3.2 and 3.3 together gives us Algortihm 1, the algorithm for the hyperbolic distance between hyperbolic lines in $\mathbb{H}^2$.

```
Data: points $a, b, c, d \in \mathbb{H}^2$ with $a \neq b$ and $c \neq d$
Result: $\rho_{\mathbb{H}^2}(l_1, l_2)$ for the hyperbolic line $l_1$ through $a$ and $b$, and the
        hyperbolic line $l_2$ through $c$ and $d$
/* Case A                                                              */
if $l_1$ and $l_2$ are Euclidean rays then
    return 0
/* Case B                                                              */
else if $l_1$ and $l_2$ are a Euclidean ray and a semicircle then
    if $\overline{l_1} \cap \overline{l_1} \neq \emptyset$ then
        return 0
    else
        calculate $\rho_{\mathbb{H}^2}(l_1, l_2)$ using Proposition 3.3
    end
/* Case C: $l_1$ and $l_2$ are semicircles                               */
else
    if $\overline{l_1} \cap \overline{l_1} \neq \emptyset$ then
        return 0
    else if center($l_1$) == center($l_2$) then
        calculate $\rho_{\mathbb{H}^2}(l_1, l_2)$ using (3.6)
    else
        calculate $\rho_{\mathbb{H}^2}(l_1, l_2)$ using Proposition 3.2
    end
end
```

Algorithm 1: Algorithm for hyperbolic distance between hyperbolic lines in $\mathbb{H}^2$.

4. The unit disk

In this section we find the hyperbolic distance between two hyperbolic lines in the unit disk. For all $x \in \mathbb{B}^2$ we denote $x^* = x/|x|^2$.

Let $a, b \in \mathbb{B}^2$ be two distinct points, $a \neq 0$. If points a, b and a^* are collinear, then the hyperbolic line through a and b is the Euclidean line segment

$$\{z \in \mathbb{B}^2 \colon z = ta/|a|,\, t \in (-1, 1)\}. \tag{4.1}$$

If points a, b and a^* are not collinear, then the hyperbolic line through a and b is the circular arc

$$S^1(c,|a-c|)\cap\mathbb{B}^2, \quad c=\text{center}(a,b,a^*). \tag{4.2}$$

By mapping T_z defined in (2.3) we can map any hyperbolic line of type (4.2) to type (4.1) and preserve hyperbolic distances, see Figure 2. The selection of z that does the trick is

$$z=c\left(1-\frac{|a-c|}{|c|}\right), \quad c=\text{center}(a,b,a^*). \tag{4.3}$$

Let $l_1,l_2\subset\mathbb{B}^2$ by hyperbolic lines. By the discussion above, we may assume that l_1 is of type (4.1) and after rotation about the origin we may choose $l_1=(-i,i)$.

All the hyperbolic lines l_3 perpendicular to l_1 have $\text{Re}(\text{center}(l_3))=0$. Let us denote the end points of l_2 by a_2 and b_2, and the Euclidean line through points a_2 and b_2 by L. The hyperbolic lines perpendicular to l_2 satisfy $\text{center}(l_2)\in L$. Since the shortest hyperbolic segment joining l_1 and l_2 is perpendicular to both l_1 and l_2, we want hyperbolic line l_3 with $\text{center}(l_3)\in L\cap\{z\in\mathbb{B}^2\colon\ \text{Re}(z)=0\}$.

The last thing we need to do, is to find the radius of the circle C_3 that contains l_3. Since C_3 is perpendicular to the unit circle we obtain by the Pythagorean theorem

$$C_3=S^1(\text{center}(l_3),r_3)), \quad r_3=\sqrt{\text{center}(l_3)^2-1}. \tag{4.4}$$

We have obtained the following proposition.

Proposition 4.1. *Let l_1 and l_2 be hyperbolic lines in $\mathbb{B}^2$ with $\overline{l_1}\cap\overline{l_2}=\emptyset$. If l_1 is of type* (4.1) *and l_2 is of type of type* (4.2)*, then $\rho_{\mathbb{B}^2}(l_1,l_2)=\rho_{\mathbb{B}^2}(p_1,p_2)$, where $\{p_1\}=l_1\cap C_3$ and $\{p_2\}=l_2\cap C_3$. Here C_3 is as in* (4.4).

Finally, we need to consider the case $\overline{l_1}\cap\overline{l_2}\neq\emptyset$. If $l_1\cap l_2\neq\emptyset$, then clearly $\rho_{\mathbb{B}^2}(l_1,l_2)=0$. We assume that $l_1\cap l_2=\emptyset$. As above, we may assume that $l_1=(-i,i)$. Now we can consider

$$l_2=\{z\in\mathbb{B}^2\colon\ z=a-i+ae^{i\alpha}\}$$

for $a>0$. We choose $y=a+i+ae^{i\alpha}$ for small enough α and $x\in l_1$ such that $x\in L'$, where L' is a Euclidean line through y and $a-i$. Now $y\to -i$ and $x\to -i$ as $\alpha\to\pi$. Since $|x|=1-a\tan(\pi-\alpha)$ and $|x-(a-i)|=a/\cos(\pi-\alpha)$ we can estimate

$$\begin{aligned}\rho_{\mathbb{B}^2}(l_1,l_2)\le\rho_{\mathbb{B}^2}(x,y)&=2\,\text{arc}\sinh\frac{|x-y|}{1-|x|^2}\\&=2\,\text{arc}\sinh\frac{\frac{a}{\cos(\pi-\alpha)}-a}{1-(1-a\tan(\pi-\alpha))^2}\to 0\end{aligned}$$

as $\alpha\to 0$, where the limit follows from Lemma 2.2. We conclude that $\rho_{\mathbb{B}^2}(l_1,l_2)=0$ whenever $\overline{l_1}\cap\overline{l_2}\neq\emptyset$.

Combining Proposition 4.1 with the above discussion we obtain the following algorithm.

```
Data: points $a, b, c, d \in \mathbb{B}^2$ with $a \neq b$ and $c \neq d$
Result: $\rho_{\mathbb{B}^2}(l_1, l_2)$ for the hyperbolic line $l_1$ through $a$ and $b$, and the
        hyperbolic line $l_2$ through $c$ and $d$
if $l_1$ and $l_2$ are circular arcs then
    use function $T_z$ defined in (2.3) for $z$ as in (4.3) to transform $l_1$ into a
      Euclidean line segment
end
if $\overline{l_1} \cap \overline{l_2} \neq \emptyset$ then
    return 0
else
    calculate $\rho_{\mathbb{B}^2}(l_1, l_2)$ Proposition 4.1
end
```

Algorithm 2: Algorithm for hyperbolic distance between hyperbolic lines in $\mathbb{B}^2$.

5. Testing the algorithms

We compared Algorithms 1 and 2 with other solutions to the problem using random points. We implemented the algorithms in MATLAB/Octave and tested the performance. We made additionally visual testing for strategically chosen points and random points for Algorithms 1 and 2. Next we shortly introduce other methods.

The easiest way to find the minimum distance between two hyperbolic lines is to generate m points for each line and find the shortest hyperbolic distance between the points pairwise. We call this method the linear search (LS). In the LS algorithm the points on the hyperbolic line are equally spaced in the Euclidean distance.

An other way is to represent the each hyperbolic line with a real variable and minimise the hyperbolic distance with respect to the variables. For example, if a hyperbolic line is an arc of a circle we can write it as $x + re^{it}$ for t in a suitable interval. If both hyperbolic lines are circular arcs we can minimise

$$\rho(x + re^{it}, x' + r'e^{is})$$

with respect to variables t and s. In MATLAB/Octave we may use function `fminsearch`. We call this algorithm the minimum search (MS). The starting point for minimisation was selected to be the midpoints of the domains for t and s.

We tested the algorithms with 1000 random quadruples of points and compared the running times. For the linear search we also varied the number of points m with values 50, 250 and 500. Additionally we checked which of the three algorithms gave the lowest value. For the LM algorithm we estimated the error by comparing the value to Algorithm 1 or Algorithm 2 depending on the domain. The MS algorithm uses minimisation, which gives the points that are used to compute the hyperbolic distance in the domain. If the minimisation points were not in the original domain, then the minimisation did not work and we did not include the result to our test. We kept track how often this happened and reported the success rate.

For every set of random point Algorithm 1 or Algorithm 2 gave the minimum value of the 3 algorithms. However, in some cases also the MS algorithm gave the same value.

	LS, $m = 50$	LS, $m = 250$	LS, $m = 500$	MS	Alg. 1 / 2
$\mathbb{H}^2$	1.1 (0.08)	12.3 (0.01)	58.0 (0.009)	49.1 (53%)	1.4
$\mathbb{B}^2$	1.6 (0.02)	16.3 (0.004)	77.5 (0.002)	56.4 (22%)	2.6

Table 1: Average evaluation time (in ms) for LS, MS and Algorithms 1 and 2. For LS algorithm error is given in parentheses and MS algorithm success rate is given in parentheses.

From Table 1 we can see that the success rate for MS algorithm is poor. The algorithm gives good results when it works, but it is much slower compared to Algorithms 1 and 2. Table 1 also shows that LS algorithm works, but the quality is poor (10^{-2}) with $m = 50$. Choosing $m = 500$ gives better quality, but the evaluation time becomes longer than for the other algorithms. We may conclude that Algorithms 1 and 2 outperform LS and MS algorithms.

Finally, we note that Algorithms 1 and 2 do not work in higher dimensions ($n \geq 3$) in general and it remains an open problem how the generalisation should be implemented.

References

[1] A. Aigon-Dupuy, P. Buser, K.-D. Semmler: Hyperbolic geometry. Hyperbolic geometry and applications in quantum chaos and cosmology, 1–81, London Math. Soc. Lecture Note Ser., 397, Cambridge Univ. Press, Cambridge, 2012.

[2] J.W. Anderson: Hyperbolic geometry. Second edition. Springer Undergraduate Mathematics Series. Springer-Verlag London, Ltd., London, 2005.

[3] A.F. Beardon: The geometry of discrete groups. Corrected reprint of the 1983 original. Graduate Texts in Mathematics, 91. Springer-Verlag, New York, 1995.

[4] A. Einstein: Zur Elektrodynamik bewegter Körper. Annalen der Physik 17 (1905), 891–921.

[5] H. Geiges: The Geometry of Celestial Mechanics. To appear in London Mathematical Society Student Texts.

[6] E. Harmaala, R. Klén: Ptolemy's constant and uniformity. Manuscript, 2016, arXiv:1604.05367.

[7] R. Klén: Local convexity properties of quasihyperbolic balls in punctured space. J. Math. Anal. Appl. 342 (2008), no. 1, 192–201.

[8] H. Minkowski: Das Relativitätsprinzip. Ann. Phys. 352 (1915), 927–938.

[9] R. Nevanlinna: Eindeutige analytische Funktionen. - 2te Aufl. Die Grundlehren der math- ematischen Wissenschaften in Einzeldarstellungen mit besonderer Berücksichtigung der An- wendungsgebiete, Bd XLVI. Springer-Verlag, Berlin-Göttingen-Heidelberg, 1953.

[10] M. Vuorinen: Conformal geometry and quasiregular mappings. Lecture Notes in Mathematics, 1319. Springer-Verlag, Berlin, 1988.

Annales Mathematicae et Informaticae
47 (2017) pp. 141–148
http://ami.uni-eszterhazy.hu

A combinatorial generalization of the gcd-sum function using a generalized Möbius function

Vichian Laohakosol[a], **Pinthira Tangsupphathawat**[b]

[a]Department of Mathematics, Faculty of Science, Kasetsart University,
Bangkok 10900, Thailand
fscivil@ku.ac.th

[b]Department of Mathematics, Faculty of Science and Technology,
Phranakorn Rajabhat University, Bangkok 10220, Thailand
t.pinthira@hotmail.com

Submitted September 4, 2017 — Accepted November 30, 2017

Abstract

Using the Souriau-Hsu-Möbius function with a natural parameter, a generalized Cesáro formula which is an extension of the classical gcd-sum formula is derived. The formula connects a combinatorial aspect of the generalized Möbius function with the number of integers whose prime factors have sufficiently high powers.

Keywords: gcd-sum function, Souriau-Hsu-Möbius function

MSC: 11A25

1. Introduction

Let $\mathcal{A} := \{F : \mathbb{N} \to \mathbb{C}\}$ be the set of complex-valued arithmetic functions. For $F, G \in \mathcal{A}$, their addition and Dirichlet product (or convolution) are defined, respectively, by

$$(F+G)(n) = F(n) + G(n), \quad (F * G)(n) = \sum_{d \mid n} F(d)G(n/d).$$

It is well known, [7, Chapter 7] that $(\mathcal{A}, +, *)$ is commutative ring with identity I, where $I(n) = 1$ if $n = 1$ and $I(n) = 0$ when $n > 1$. The Souriau-Hsu-Möbius function ([9], [2]) is defined, for $\alpha \in \mathbb{C}$, by

$$\mu_\alpha(n) = \prod_{p|n} \binom{\alpha}{\nu_p(n)} (-1)^{\nu_p(n)},$$

where $n = \prod p^{\nu_p(n)}$ denotes the unique prime factorization of n. For some particular values of α, the corresponding Souriau-Hsu-Möbius functions represent certain well-known arithmetic functions, namely,

(i) when $\alpha = 0$, this corresponds to the convolution identity $\mu_0 = I$;

(ii) when $\alpha = 1$, this is the classical Möbius function $\mu_1 = \mu$;

(iii) when $\alpha = -1$, this is the inverse of the Möbius function $\mu_{-1} = \mu^{-1} =: u$, where $u(n) = 1$ $(n \in \mathbb{N})$ is the constant 1 function;

(iv) when $\alpha = -2$, this is the number of divisors function, $\mu_{-2} = d$, [2, p. 75]. Following [11], see also [6], for $\alpha \in \mathbb{C}$, $k \in \mathbb{Z}$, the (k, α)-Euler's totient is defined as an arithmetic function of the form

$$\varphi_{k,\alpha} := \zeta_k * \mu_\alpha \qquad \zeta_k(n) := n^k.$$

When $k = \alpha = 1$, this function is the classical Euler's totient

$$\varphi_{1,1}(n) := \varphi(n) = \zeta_1 * \mu_1(n) = \sum_{d|n} d\mu\left(\frac{n}{d}\right) = n \prod_{p|n} \left(1 - \frac{1}{p}\right),$$

which counts the number of integers in $\{1, 2, \ldots, n\}$ that are relatively prime to n. Fixing $k = 1, \alpha = r \in \mathbb{N}$, the corresponding Euler's totient, referred to as *r-Euler totient*, is

$$\varphi_r(n) := \varphi_{1,r}(n) := \zeta_1 * \mu_r(n) = \sum_{d|n} d\mu_r\left(\frac{n}{d}\right).$$

As mentioned in [5, Example 6], the r-Euler totient has the following combinatorial meaning: an integer a is said to be *r^{th}-degree prime* to n (≥ 2), briefly written as $(a, n)_r = 1$, if for each prime divisor p of n, there are integers $a_0, a_1, \ldots, a_{r-1}$ with $0 < a_i < p$ such that

$$a \equiv a_0 + a_1 p + \cdots + a_{r-1} p^{r-1} \pmod{p^r}.$$

As a convention, we define

$$(a, 1)_r := 1 \quad \text{for any } a \in \mathbb{N}.$$

When $r = 1$, the concept of being r^{th}-degree prime is merely that of being relatively prime.

In order to connect the concept of being r^{th}-degree prime with the r-Euler totient, we introduce another notion. A positive integer n is said to be *r-powerful* if each of its prime factor appears with multiplicity at least r, i.e., $\nu_p(n) \geq r$ for each prime divisor p of n; as a convention, the integer 1 is adopted to be r-powerful for any $r \in \mathbb{N}$. Note that if n is r-powerful, then it is also s-powerful for all $s \in \mathbb{N}$ with $s \leq r$. The following lemma shows that when n is r-powerful, the function $\varphi_r(n)$ counts the number of a's in the set $\{1, 2, \dots, n\}$ such that $(a, n)_r = 1$; the proof given here is extracted from [11].

Lemma 1.1. *Let $n, r \in \mathbb{N}$, and let $N_r(n)$ denote the number of integers $a \in \{1, 2, \dots, n\}$ such that $(a, n)_r = 1$. We have :*

1) *The function $F(n) = N_r(n)$ if n is r-powerful and zero otherwise, is multiplicative.*

2) *If n is r-powerful, then $N_r(n) = \varphi_r(n) = n \prod_{p|n} (1 - 1/p)^r$,*
$N_r(1) = \varphi_r(1) := 1$.

Proof. 1) Let n be an r-powerful positive integer whose prime factorization is $n = p_1^{e_1} \cdots p_s^{e_s}$. By the Chinese remainder theorem, for any integers $\alpha_1, \dots, \alpha_s$, there is a unique $a \pmod n$ such that

$$a \equiv \alpha_1 \pmod{p_1^{e_1}}, \; \dots \;, a \equiv \alpha_s \pmod{p_s^{e_s}}.$$

Conversely, for any $a \pmod n$, there uniquely exist $\alpha_i \pmod{p_i^{e_i}}$ $(i = 1, \dots, s)$ satisfying the above system of congruences. Thus,

$$(a, n)_r = 1 \iff (a, p_i^{e_i})_r = 1 \text{ holds for every } i \in \{1, \dots, s\},$$

which shows at once that $N_r(n)$ is a multiplicative function of n.

2) Using part 1), it suffices to check that N_r and φ_r are equal on any prime power p^e with $e \geq r$. Recall that $N_r(p^e)$ is the number of $a \in \{1, 2, \dots, p^e\}$ such that $(a, p^e)_r = 1$, i.e., such that there are integers $a_0, a_1, \dots, a_{r-1}$ with $0 < a_i < p$ satisfying

$$a \equiv a_0 + a_1 p + \cdots + a_{r-1} p^{r-1} \pmod{p^r}.$$

Thus, the number of such $a \pmod{p^r}$ is $(p-1)^r$, and so the total number of such $a \pmod{p^e}$ is $N_r(p^e) = p^{e-r}(p-1)^r$. On the other hand using $e \geq r$, we have

$$\begin{aligned}
\varphi_r(p^e) &= \sum_{d|p^e} d\mu_r(p^e/d) = p^0 \binom{r}{e} (-1)^e + p \binom{r}{e-1} (-1)^{e-1} + \cdots + p^e \binom{r}{0} (-1)^0 \\
&= p^{e-r} \binom{r}{r} (-1)^r + p^{e-r+1} \binom{r}{r-1} (-1)^{r-1} + \cdots + p^e \binom{r}{0} (-1)^0 \\
&= p^{e-r}(p-1)^r = N_r(p^e).
\end{aligned}$$

□

The classical *gcd-sum function* is an arithmetical function defined by

$$g(n) := \sum_{j=1}^{n} \gcd(j, n), \tag{1.1}$$

and the classical *gcd-sum formula* states that

$$g(n)=\sum_{j=1}^{n}\gcd(j,n)=\sum_{d|n}d\varphi\left(\frac{n}{d}\right). \tag{1.2}$$

There have recently appeared quite a number of works related to the gcd-sum function and the gcd-sum formula. In [3], the gcd-sum function (1.1) is shown to be multiplicative, has a polynomial growth, and arises in the context of a lattice point counting problem, while the paper [1] studies the function $\sum_{j=1,\ \gcd(j,n)|d}^{n}\gcd(j,n)$, which is a generalization of the gcd-sum function (1.1).

In [8], the function $\sum_{j=1}^{n}\gcd(j,n)^{-1}$, which counts the orders of a generator of a cyclic group, is studied. In [4], an extended Cesáro formula

$$\sum_{j=1}^{n}f(\gcd(j,n))=\sum_{d|n}f(d)\varphi\left(\frac{n}{d}\right)\qquad(f\in\mathcal{A}), \tag{1.3}$$

which is another extension of (1.2), is investigated. Various properties of the gcd-sum function (1.1) and its analogues are surveyed in [10]. Our objective here is to establish yet another generalization of the gcd-sum formula (1.2) by relating the r-Euler totient with the counting of r-powerful integers that are r^{th}-degree prime.

2. Generalized gcd-sum formula

Our generalized gcd-sum formula arises from replacing the usual Euler's totient on the right-hand side of (1.3) by the r-Euler totient, and using its combinatorial meaning to derive its corresponding generalized form. To do so, we need to extend the notion of r^{th}-degree primeness to that of r-gcd.

Definition. Let $r\in\mathbb{N}$, and let $n\in\mathbb{N}$ be r-powerful. For $j\in\mathbb{N}$, the integer g is the *r-gcd* of j and n, denoted by $g:=(j,n)_r$, if $g=\gcd(j,n)$ satisfies two additional requirements

1. $\left(\frac{j}{g},\frac{n}{g}\right)_r=1$, and
2. n/g is r-powerful.

When $r=1$, the above definition of r-gcd is identical with the usual greatest common divisor. Let us look at some examples.

Example 1. Let $n=2^3\cdot3^2=72$. The divisors of n are $1,2,3,4,6,8,9,12,18,24,36,72$. Consider $j\in\{1,2,\dots,72\}$. For $r=1$, we have
$(j,72)_1=1$ when $j\in\{1,5,7,11,13,17,19,23,25,29,31,35,37,41,43,47,49,53,$
$55,59,61,65,67,71\}$
$(j,72)_1=2$ when $j\in\{2,10,14,22,26,34,38,46,50,58,62,70\}$
$(j,72)_1=3$ when $j\in\{3,15,21,33,39,51,57,69\}$

$(j,72)_1 = 4$ when $j \in \{4,20,28,44,52,68\}$
$(j,72)_1 = 6$ when $j \in \{6,30,42,66\}$
$(j,72)_1 = 8$ when $j \in \{8,16,32,40,56,64\}$
$(j,72)_1 = 9$ when $j \in \{9,27,45,63\}$
$(j,72)_1 = 12$ when $j \in \{12,60\}$
$(j,72)_1 = 18$ when $j \in \{18,54\}$
$(j,72)_1 = 24$ when $j \in \{24,48\}$
$(j,72)_1 = 36$ when $j \in \{36\}$
$(j,72)_1 = 72$ when $j \in \{72\}$.

For $r = 2$, we have
$(j,72)_2 = 1$ when $j \in \{7,23,31,35,43,59,67,71\}$
$(j,72)_2 = 2$ when $j \in \{14,46,62,70\}$
$(j,72)_2 = 8$ when $j \in \{32,40,56,64\}$
$(j,72)_2 = 9$ when $j \in \{27,63\}$
$(j,72)_2 = 18$ when $j \in \{54\}$
$(j,72)_2 = 72$ when $j \in \{72\}$,
and for the other values of j, the 2-gcd $(j,72)_2$ are not defined.

For $r = 3$, we have $(j,72)_3 = 9$ when $j \in \{63\}$, $(j,72)_3 = 72$ when $j \in \{72\}$, and for the other values of j, the 3-gcd $(j,72)_3$ are not defined.

For $r \geq 4$, we have $(j,72)_3 = 72$ when $j \in \{72\}$, and the r-gcd $(j,72)_r$ are not defined for the remaining j's.

Our next lemma connects the r-gcd with the r-Euler totient.

Lemma 2.1. *Let $r, n \in \mathbb{N}$ with n being r-powerful, and let*

$$A_{r,d}(n) := \{a \in \{1,2,\ldots,n\}\ ;\ (a,n)_r = d\}\,.$$

Then

$$|A_{r,d}(n)| = \varphi_r\,(n/d)\,.$$

Proof. The result follows at once from the observation that

$$a \in \{1,2,\ldots,n\} \quad \text{and} \quad (a,n)_r = d,$$

if and only if $\frac{a}{d} \in \left\{1,2,\ldots,\frac{n}{d}\right\}$, $\left(\frac{a}{d},\frac{n}{d}\right)_r = 1$, $\frac{n}{d}$ is r-powerful, and the set of elements on the right-hand side has cardinality $\varphi_r\,(n/d)$. □

We now state and prove our generalized Cesáro formula.

Theorem 2.2. *Let $r, n \in \mathbb{N}$ with n being r-powerful. For an arithmetic function f, we have the following generalized Cesáro formula*

$$\sum_{j=1}^{n} f((j,n)_r) = \sum_{d\,|_r\,n} f(d)\varphi_r\left(\frac{n}{d}\right). \tag{2.1}$$

where the symbol $d\,|_r\,n$ in the summation on the right-hand side indicates that the sum extends over all the divisors d for which n/d is r-powerful.

In particular, taking $f(n) = n$, the generalized Cesáro formula becomes the generalized gcd-sum formula

$$\sum_{j=1}^{n}(j,n)_r = \sum_{d\,|_r\,n} d\varphi_r\left(\frac{n}{d}\right). \tag{2.2}$$

Proof. Writing $(j,n)_r = d$, using the above notation and Lemma 2.1, we get

$$\sum_{j=1}^{n} f\left((j,n)_r\right) = \sum_{j=1}^{n}\sum_{(j,n)_r=d} f(d) = \sum_{d\,|_r\,n} f(d)\left|A_{r,d}(n)\right| = \sum_{d\,|_r\,n} f(d)\varphi_r\left(\frac{n}{d}\right). \quad \square$$

When $r = 1$, the generalized Cesáro formula is simply the classical Cesáro formula, and its representation via generalized Möbius function becomes a Dirichlet product of two arithmetic functions, viz.,

$$\sum_{j=1}^{n} f((j,n)_1) = \sum_{d\,|_1\,n} f(d)\varphi_1\left(\frac{n}{d}\right) = (f*\varphi)(n).$$

Example 2. Continuing from Example 1, let $n = 2^3 \cdot 3^2 = 72$.

For $r = 1$, we have

$$\{d \in \{1,2,\ldots,72\}\,;\, d\,|_1\,72\} = \{1,2,3,4,6,8,9,12,18,24,36,72\},$$

and the values of $\varphi_1\,(72/d)$ with $d\,|_1\,72$ are

$$\begin{aligned}&\varphi_1(72) = 24, \varphi_1(36) = 12, \varphi_1(24) = 8, \varphi_1(18) = 6, \varphi_1(12) = 4, \varphi_1(9) = 6,\\ &\varphi_1(8) = 4, \varphi_1(6) = 2, \varphi_1(4) = 2, \varphi_1(3) = 2, \varphi_1(2) = 1, \varphi_1(1) = 1.\end{aligned}$$

Using Example 1, the left-hand side of (2.1) is

$$\begin{aligned}\sum_{j=1}^{n} f((j,n)_1) &= f(1)\times 24 + f(2)\times 12 + f(3)\times 8 + f(4)\times 6 + f(6)\times 4\\ &\quad + f(8)\times 6 + f(9)\times 4 + f(12)\times 2 + f(18)\times 2 + f(24)\times 2\\ &\quad + f(36)\times 1 + f(72)\times 1\\ &= \sum_{d\,|_1\,n} f(d)\varphi_1\left(\frac{n}{d}\right)\end{aligned} \tag{2.3}$$

which agrees with the theorem.

For $r = 2$, we have $\{d \in \{1,2.\ldots,72\}\,;\, d\,|_2\,72\} = \{1,2,8,9,18,72\}$, and the values of $\varphi_2\,(72/d)$ with $d\,|_2\,72$ are

$$\varphi_2(72) = 8, \varphi_2(36) = 4, \varphi_2(9) = 4, \varphi_2(8) = 2, \varphi_2(4) = 1, \varphi_2(1) = 1.$$

Using Example 1, the left-hand side of (2.1) is

$$\sum_{j=1}^{n} f((j,n)_2) = f(1)\times 8 + f(2)\times 4 + f(8)\times 4 + f(9)\times 2 + f(18)\times 1 + f(72)\times 1$$
$$= \sum_{d\,|_2\,n} f(d)\varphi_2\left(\frac{n}{d}\right). \tag{2.4}$$

For $r = 3$, we have $\{d \in \{1, 2. \dots, 72\}\,;\, d\,|_3\,72\} = \{9, 72\}$, and the values of $\varphi_3\,(72/d)$ with $d\,|_3\,72$ are $\varphi_3(8) = 1, \varphi_3(1) = 1$. Using Example 1, the left-hand side of (2.1) is

$$\sum_{j=1}^{n} f((j,n)_3) = f(9)\times 1 + f(72)\times 1 = \sum_{d\,|_3\,n} f(d)\varphi_3\left(\frac{n}{d}\right). \tag{2.5}$$

For $r \geq 4$, we have $\{d \in \{1, 2. \dots, 72\}\,;\, d\,|_r\,72\} = \{72\}$, and the values of $\varphi_3\,(72/d)$ with $d\,|_r\,72$ is $\varphi_r(1) = 1$. The left-hand side of (2.1) is

$$\sum_{j=1}^{n} f((j,n)_r) = f(72)\times 1 = \sum_{d\,|_r\,n} f(d)\varphi_r\left(\frac{n}{d}\right). \tag{2.6}$$

The formulae such as (2.1)–(2.6) deal with a single r. We end this paper by remarking that such formulae can be absorbed into one single formula. For a positive integer n whose prime representation is $n = p_1^{\nu_1(n)} p_2^{\nu_2(n)} \cdots p_t^{\nu_t(n)}$, let

$$\nu(n) := \max\{\nu_i(n); 1 \leq i \leq t\}.$$

Corollary 2.3. *For an arithmetic function f, we have the following generalized Cesáro formula*

$$\sum_{r=1}^{\nu(n)} \sum_{j=1}^{n} f((j,n)_r) = \sum_{r=1}^{\nu(n)} \sum_{d\,|_r\,n} f(d)\varphi_r\left(\frac{n}{d}\right).$$

Acknowledgements. The first author is supported by Center for Advanced Studies in Industrial Technology and Faculty of Science, Kasetsart University. The second author is supported by Phranakhon Rajabhat University.

References

[1] Abel, U., Awan W., Kushnirevych, V., A generalization of the gcd-sum function, *J. Integer Sequences* Vol. 16 (2013), Article 13.6.7.

[2] Brown, T. C., Hsu, L. C., Wang, J., Shiue, P. J. S., On a certain kind of generalized number-theoretical Möbius function, *Math. Sci.* Vol. 25 (2000), 72–77.

[3] Broughan, K. A., The gcd-sum function, *J. Integer Sequences* Vol. 4 (2001), Article 01.2.2.

[4] Haukkanen, P., On a gcd-sum function, *Aequationes Math.* Vol. 76 (2008), 168–178.

[5] Hsu, L. C., Wang, J., Some Möbius-type functions and inversions constructed via difference operators, *Tamkang J. Math.* Vol. 29 (1998), 89–99.

[6] Pabhapote, N., Laohakosol V., Combinatorial aspects of the generalized Euler's totient, *Intern. J. Math. Math. Sci.* (2010), Article ID 648165, 15 pages.

[7] Redmond, D., Number Theory, An Introduction, Marcel Dekker, New York, 1996.

[8] Sándor, J., Kramer, A. V., Über eine zahlentheoretische Funktion, *Math. Moravica* Vol. 3 (1999), 53–62.

[9] Souriau, Jean-Marie, Généralisation de certaines formules arithmétiques d' inversion. Applications, *Revue Scientific (Rev. Rose Illus.)* Vol. 82 (1944), 204–211.

[10] Tóth, L., A survey of gcd-sum functions, *J. Integer Sequences* Vol. 13 (2010), Article 10.8.1.

[11] Wang, J., Hsu, L. C., On certain generalized Euler-type totients and Möbius-type functions, Dalian University of Technology, China, preprint.

Annales Mathematicae et Informaticae
47 (2017) pp. 149–176
http://ami.uni-eszterhazy.hu

Decision structure based object-oriented design principles

Szabolcs Márien

Eszterházy Károly University of Applied Sciences
Institute of Mathematics and Informatics
Eger, Hungary
szabolcs.marien@innovitech.hu

Submitted August 24, 2016 — Accepted January 3, 2017

Abstract

The major part of program complexity is based on the logic of conditions, but the existing refactoring methods do not detail the options of decision merging according to the cases of decision redundancies, which are the main options of optimizing the decision structures by refactoring. To extinguish decision redundancies in the source code, we have an option to merge decisions, which can be interpreted as refactoring tools, by which the quality of code structures can be optimized. I intend to complete the definitions of decision, decision raising, and introduce a novel concept, decision merging, based on the concept of behavioural contract. According to the decision merging cases, new design principles can be created. The principle "Using inheritance to dissolve decision redundancy" identifies the cases, when the usage of inheritance as an object-oriented tool is more reasonable than object composition. The other new principle is "Avoid decision redundancy", by which decision redundancies can be eliminated based on the decision merging rules. I initiate new object-oriented metrics as well, giving the opportunity to determine the degree of decision redundancies in the software. The properties of these metrics are analysed empirically.

Keywords: Design principles, metrics, inheritance, decision raising, decision merging, decision redundancy.

1. Introduction

1.1. Optimizing decision structures by refactoring

The refactoring of conditional statements by polymorphic methods ("Replace Conditional with Polymorphism") is an interesting, existing refactoring method [11, 16], where the branches of conditional ("if-then-else") statements can be realized as a class with an abstract polymorphic method, which is overridden by the subclasses. The interface of a decision will be an abstract polymorphic method in the parent class [11]. The advantage of replacing a conditional statement with a polymorph method is prevailed if the conditional statement has equal occurrences in the program. In this case the subclasses are not necessary to be known, which reduces dependencies significantly [11]. Consequently, the introduction of new decision options does not result in the change of places where they are used, only the introduction of a new subclass is necessary [11].

The "Replace Type Code with Subclasses" and the "Replace Type Code with State/Strategy" – as conditional statement specific refactoring methods – are based on the previously described "Replace Conditional with Polymorphism" refactoring method [11]. Furthermore there are the "Move Embellishment to Decorator" and the "Replace Conditional Dispatcher with Command" conditional statement optimizing refactoring methods, which are also based on design patterns [16].

According to my concept, class hierarchies can be viewed as abstract decisions [24, 25], based on which I define decision raising and decision merging as the extended interpretations of decision structure optimization methods. When we define a decision, we give the functionality and/or the data structure (state description) of decision options. Decision predicate decides which decision option will set off at a given decision location [24, 25]. In order to simplify the problem, every decision consists of two decision options so every decision tree is a binary tree. As every tree can be transformed into a binary tree, we do not lose generality (see Section 4, where behavioral contract based definitions of decisions and decision raisings are introduced).

The following rules for avoiding decision redundancy are defined in Section 5:

- Decisions should not Recur Rule 1 (DnR Rule 1): "decisions with equivalent decision predicates and equivalent or partly equivalent data structures and behaviors should not recur"
- Decisions should not Recur Rule 2 (DnR Rule 2): "decisions with equivalent decision predicates that define diverse data structures and behaviors should not recur".

I will introduce the decision merging cases as new refactoring tools (see Section 6), the help of which the defined decision redundancies (see Section 5) can be eliminated. The cases of decision merging are determined based on the cases of decision redundancies. The merging of nonraised decision can be realized after raising decisions. The decision raising is a transformation method, by which the decisions can be defined by class hierarchies without conditional statements. The subclasses of hierarchies define the decision options, where the interfaces of decisions are the polymorph methods of parent class. There are the following cases of decision merging:

- The merging/partial merging of fully or partially equivalent decisions: Decision merging/partial merging is necessary if the decision predicates of decisions are equivalent, and decision option declared data structures and behaviors are equivalent or partially equivalent.
- The merging of decisions with equivalent decision predicates and nonequivalent behavioral contracts: If there are two nonraised or raised decisions, which have equivalent decision predicates, then these decisions can be merged.

1.2. Behavioral contracts

In order to examine decision structures and optimization transformations based on the optimization cases (when the transformations are justified) the introduction of the behavioral contracts of decision structures is necessary. There is a contract ("Design by contract" – DBC) between a class and its client, according to which the client has to realize the declared conditions in the course of the calling method of the class. In addition, the class guarantees specified conditions in the course of the returning of the method, which specifies the required behavior of the method [18]. The behavior of a program/object can be specified by a behavioral contract [18, 27], where the contract declares a set of possible behaviors [27] ("Concept of a behavioral contract" – "Design by Contract"). These contracts can be specified by the pre- and post-conditions of methods. Method behavior independent conditions, which are always satisfied, can be specified as class invariants, which describe the general aspects of the behavior contracts of classes [22]. Accordingly invariants specify the general constraints of the values of class variables. General state transition constraints can be specified by history constraints [27]. The parts of state describing data structures which are changed by methods as state transitions are declared by frame conditions [27].

1.3. Object-oriented design principles

Cohesion, coupling and complexity are highlighted as the main targets of quality ensuring metrics [1]. Cohesion examines the inner-consistencies of parts [8]. Accordingly cohesion determines the collaboration levels of elements inside modules. In case of high cohesion, the major part of components realizes the same functionality [1, 5]. In conformity with this, the functional cohesion of a component is high if it serves a properly encompassed behaviour [4]. In case of good program design the cohesion of program structure is high and its coupling is low [8]. According to Wand and Weber's coupling definition [30], there is coupling between two "things" if there is at least one connection between their state histories. The strongest type of coupling is inheritance. Based on loose coupling, the realization of independent system components can be facilitated, meaning that the changing and the maintenance of programs become easier. The aim of object-oriented design principles is to eliminate of dependencies and couplings, to increase cohesion and to decrease complexity. Using object-oriented design principles as abstract concepts, the mentioned designing failures can be avoided.

Liskov substitution principle – LSP [12, 19, 27, 28]: The LSP was extended by Schreiner, who added that subtypes could specialize and refine the parent-type declared contracts, which does not violate the definitions of subtype and substitution. With respect to behavior contracts, it means the following: the preconditions of the subtype are not stronger than in the parent type, the post conditions of the subtype are not weaker than in the parent type, the invariants of the parent-type-based member variables of the subtype are not weaker than in the parent type, the subtype realizes the history constraints of the parent type [27].

Favor object composition over class inheritance [12]: In consideration of reusability and maintainability the appropriate usage of composition vs. inheritance is a critical issue. The aim is a harmonic class and object structure. Inheritance is the tightest couple between classes, which can be used only in well-defined cases.

Single Responsibility Principle [20]: In the course of determining class behavior, we have to take separated task responsibility into consideration. The determined class behavior should be responsible for "one task", other classes need to be introduced in order to supply other tasks. If we don't take it into consideration, and a single class realizes more tasks, then if one of the task behaviors needs to be changed, it may result in the change of the behavior of the other tasks as a side effect, generating more risk and cost.

The issues of the previously mentioned object-oriented design principles are concerned by the concept of decision merging. Accordingly decision merging (discussed in section 6) can support the issue of the separation of class behaviors into subclasses ("Liskov substitution principle" [12, 19, 27, 28], "Single Responsibility Principle" [20]). Furthermore, decision redundancies and decision merging cases (which eliminate decision redundancies) support the appropriate usage of inheritance and object composition ("Favor object composition over class inheritance" [12]). According to the cases of decision redundancies and the decision merging rules, two object-oriented design principles are created (see Section 3):

- "Using inheritance to dissolve decision redundancy": If one case of decision redundancies induces the usage of decision merging and decision raisings transformations, then the usage of inheritance is confirmed.

- "Avoid decision redundancy": If there are decision redundancies in a source code, then based on the decision merging rules the decision redundancies need to be eliminated.

These new design principles are useful when deciding whether the usage of inheritance or object composition can be confirmed, which is one of the subjects of this paper.

1.4. Object-oriented design metrics

The main aspects of the quality insurance of program developing are maintainability, extendibility, intelligibility, reusability [10] and changeability [15].

Six object-oriented design metrics were specified by Chidamber and Kemerer [8]. These metrics are the following: WMC – "Weighted methods per class", DIT – "Depth of inheritance tree", NOC – "Number of children", CBO – "Coupling between objects", RFC – "Response for a class", LCOM – "Lack of Cohesion in Methods".

One of the first metrics, and at the same time probably the most determinative cohesion metric is the LCOM – "Lack of Cohesion in Methods" metric [8]. The interpretation of this metric is based on dependencies between methods, which can be determined by the sets of the method-used member variables of classes.

Eder et al. introduced the concepts of method cohesion, class cohesion and inheritance cohesion [10]. However, the specified cohesion measuring methods require semantic analysis, which obstructs the industrial usage of them.

Badri et al. analysed the correlation between coupling and cohesion [1], which has not been justified previously. However the correlation between them was mentioned several times. According to these, the high (low) cohesion is associated with the low (high) coupling values [17]. They measured the correlation between their new cohesion metric and coupling metric by empirical analysis. Meeting the requirements they revealed a significant negative correlation between their new cohesion metric and the CBO [7, 8] coupling metric [1].

Several measurements have tried to confirm the correlation between the different metrics and the changeability aspects of programs, in many cases without successfully detecting the correlations between them [15]. Chae and Kwon stated that the existing cohesion metrics will not be good measuring indicators of changeability until such cohesion metrics that can measure the cohesion properties appropriately are realized [6].

Complexity metrics facilitate the detection of complex objects and classes. Implementation, testing and verification efforts are higher in case of classes with high complexity [31]. Some examples of complexity metrics are listed in [31]: "Cyclomatic Complexity" [21], "Depth of Inheritance Tree" (DIT) [8], "Number of Children" (NOC) [8], "Weighted Methods Complexity" (WMC) [8]. These complexity metrics are based on the static aspects of systems (for example class diagrams, source codes), so they are static metrics [31].

Munson and Khoshgoftaar introduced dynamic complexity metrics [23]. They separated the concepts of "Operational complexity" and "Functional complexity" [23]. The "Operational complexity" of objects uses the "Cyclomatic Complexity" metric [21], which is based on the "Control Flow Graph" [31].

The Qi ("Quality Indicator") [2] metric – similarly to the previously mentioned metric – is based on the branches of programs, and with its help, various software attributes such as complexity, cohesion and coupling can be examined together. It uses controlling paths and their probabilities. One curiosity is the appearance of the representations of polymorphic callings with graphs, where the opportunities of dynamic couplings – or method callings – are represented by graph edges. Branches show polymorphic possibilities.

2. Motivation

The appropriate usage of inheritance is the key point of object-oriented programming, the clarifying of this question (inheritance vs. object composition/aggregation) is the aim of numerous design principles [12, 19, 27, 28] and design patterns [12, 16]. At the same time, in spite of these clarifying attempts, it is not obvious, which tool of the object-oriented technology (inheritance, object composition, aggregation) should be used in different cases.

Fowler et al. specified the "Replace Conditional with Polymorphism" refactoring method [11], by which the interpretation of inheritance was extended. It states that replacing is necessary if there are equal conditional statements in a program [11]. Additional details about the necessity of using this refactoring method are not elaborated. The merging method of the concept of "Parallel Inheritance Hierarchies" [11] eliminates the redundancies of the declaration and/or the usage of class hierarchies by defining them as raised decisions. It describes the cases where merging is necessary as follows: The merging of class hierarchies is necessary if the changing of one hierarchy results in the changing of another one [11]. The details of the "Move Embellishment to Decorator" and the "Replace Conditional Dispatcher with Command" refactoring methods [16] don't describe the decision structures that would help to recognize the necessity of the using of the "Decorator" and the "Command" Design Patterns [12].

We must see that the description of the cases of decision redundancies – which realize the necessity of decision raising and/or decision merging – is incomplete. In order to complement this, I specified the cases of decision redundancies [24, 25] (see Section 5), and I clarify it, where the raising and/or the merging of decisions are justified (see Section 6).

In order to achieve better program quality and structure, many object-oriented design principles were formulated that provide quality improvement by increasing cohesion between program units, decreasing coupling and eliminating dependencies. The LSP [12, 19, 27, 28] analyses inheritance quality. Its definition is based on behavior contracts [18, 27], which are the bases of the definitions of decision, decision raising and decision merging. The LSP is the most elaborated design principle, by the help of which the cases when inheritance relation can be used between two types can be determined. At the same time it doesn't help the detection of cases when the introduction of inheritances is confirmed by program structures. The "Favor object composition over class inheritance" [12] design principle is not accurately defined, its theoretical background is not elaborated. The principle tries to give intuitive, practical suggestions in connection with the question of using inheritance and object composition. The "Single Responsibility Principle" [20] is related to cohesion [9, 20]. In the course of defining this principle the use of behavior contracts is suggested, based on which the LSP principle was extended, and furthermore, it is used in this paper as well.

The known design principles are concerned with the critical issue of which object-oriented construction's usage is justified in different programming cases. However, in my opinion, the clarification of this question is necessary, therefore, based on my previously mentioned concept, I define new design principles to answer this issue (see Section 3).

In conformity with an examination, which is based on two metrics Tight Class Cohesion (TCC) and Loose Class Cohesion (LCC) [3], the conclusion is the following: Those classes have lower cohesion, which frequently use inheritances [3]. The experienced inverse dependencies between inheritance based code reusability and cohesion [3] indicate the unclarified status of the usability of appropriate inheritance. Accordingly, based on the Lack of Cohesion in Methods (LCOM) [8], the cohesion specific influences of inheritances are not taken into account [3, 5, 6, 13, 14, 15]. We must note that the aim of inheritances is not reusing, but the extension of the functionality of the classes with specific behaviour. Accordingly the reusing specific application of inheritance can result in the decrease of the optimal structure of the code. In order to promote the appropriate usage of inheritance, there are numerous concepts as I described above. In order to clarify this question I introduce a new concept (decision merging) about the use of inheritance in this paper. Intuitively, we must see that the class-subclass inheritance structures with optimized decision structures – which are resulted based on the eliminations of decision redundancies using decision merging cases – contribute to the decrease of the divisibility of classes, by which the cohesion of classes can be increased (see the empirical validations in Section 8).

Beside the supposed similarity between cohesion metrics and the new measuring method introduced in this paper, I also find the comparison of these new metrics and complexity metrics necessary. The reason for this is there are more complexity metrics which examine the branches of the conditional statements of object-oriented classes. Some of these metrics are the following: "Cyclomatic Complexity" – CC [21, 31], "Weighted Methods Complexity" – WMC [8], "Operational Complexity", "Functional Complexity" [23, 31], "Quality Indicator" – Qi [2]. Furthermore, there are additional C&K metrics [8] which describe the inheritance structure of programs. They are the following: "Depth of Inheritance Tree" (DIT) [8], "Number of Children" (NOC) [8]. These metrics are interesting in the consideration of the appropriate usage of inheritances according to the concepts specified in this paper.

Note that there is no complexity metric which would consider the structurally critical question of whether complexity growing conditional statements and inheritance structures are used appropriately, or whether the structures could be optimized. The method complexity which is measured by the CC metric shows the complexity of tasks, which is realized by the method. The high value of method complexity is not necessarily the sign of wrong code structures, it only shows the complexity of tasks. There is a similar conclusion according to the WMC. The "Operational Complexity" (OCPX) [23, 31] metric is based on dynamic complexity, which takes the CC of running paths into consideration. Therefore, according to the previously mentioned metrics, it can't be used to measure the quality of the decision structures.

I initiate such new object-oriented metrics that give opportunity to determine the rate of decision redundancies in the source code of a program (see Section 7). In order to determine the relationship between the "Metric of decision abstraction" (MDA), the "Ratio of inheritances coming into existence by the elimination of decision redundancies" (RIEDR) metrics and the level of code integrity, I analysed 10-10 states of several open source projects empirically.

To summarize, the following questions have to been answered:

In which cases can we talk about decision redundancies?

This is the most important question from the point of view of my topic. We need to clarify the cases where the use of decision merging is justified. In order to clarify this question the following metrics are introduced: "Metric of decision abstraction" (MDA), "Ratio of equivalent decision cases" (REDC), "Ratio of decision cases with equivalent decision predicates" (RDCEDP). Furthermore, the "Avoid decision redundancy" design principle is defined, by which the elimination of decision redundancies is targeted based on decision merging cases.

In which cases can we use inheritance?

Beyond the previously defined general aspect the aim of the inheritance-specific question is to clarify whether the using cases of inheritances are justified in the source code. We must see that there are overlaps between this and the previously mentioned questions, but in consideration of the prominent role of inheritances it must be specified separately. This question is answered by one of the introduced object-oriented design principles, namely it is the "Using inheritance to dissolve decision redundancy". This principle clarifies the appropriate usage of inheritances based on the decision redundancy cases. In order to determine the scale of appropriate inheritance usage, the "Ratio of inheritances coming into existence by the elimination of decision redundancies" (RIEDR) metric is introduced, by which the polymorph methods are analysed in the inheritances.

In order to clarify these questions, the cases of decision redundancies and decisions merging are defined. In the course of the evaluation of the new metrics I analyse the correlations between decision redundancies and cohesion, complexity and coupling, by which we can notice their relations with the general code quality aspects.

3. Extending the object-oriented design principles

In the following I suggest two new object-oriented design principles, one of which unequivocally tries to highlight those cases, where the use of inheritance is justified ("Using inheritance to dissolve decision redundancy") complementing the "Favor object composition over class inheritance" design principle [12]. Furthermore, I try to determine a general program structure organizing principle, which – beyond the subject of the appropriate usage of inheritance – helps to find optimal structures ("Avoid decision redundancy"). The new design principles contain – as a part of their definitions – the rules of decision merging (see Section 6), by the help of which decision redundancies can be avoided (see Section 5). Based on the cases of decision redundancies according to the decision merging rules, the new object-oriented design principles are the following.

3.1. Using inheritance to dissolve decision redundancy

According to the definitions of decision, decision raising and decision merging, the aim of inheritances is to define decisions in an abstract form, based on which the facility of decision merging can be realized. The use of inheritance is justified if the decision structure based dependencies confirm its usage, that is if one case of decision redundancies which justifies the usage of decision merging is fulfilled. In these cases, the elimination of decision redundancies can be realized by one of the decision merging rules.

3.2. Avoid decision redundancy

If the use of decision merging is confirmed by decision redundancies, then the decision redundancies have to be eliminated based on the decision merging rules. Using this principle, according to the rules of avoiding decision redundancies, a more optimal decision structure can be achieved. This principle determines generally the optimization facilities of decision structures based on decision redundancies and decision merging rules, accordingly it helps determine the using facilities of inheritance. It complement the "Using inheritance to dissolve decision redundancy" principle, which approaches this issue from the appropriate usage of inheritance.

3.3. Comparing the new principles with other ones

The "Using inheritance to dissolve decision redundancy" and the "Avoid decision redundancy" principles specify the cases accurately based on decision redundancies and the decision merging rules, where the use of inheritances is necessary. It complements the "Favor object composition over class inheritance" design principle [12], where the using facility can be decided based on some intuitive concepts. The LSP specifies the relationships between the type and the subtype, but it doesn't mention anything about the initial structures, where the introduction of inheritance is necessary. Based on the new design principles we can detect those structural surroundings, where the inheritances can resolve the decision redundancies. The "Single Responsibility Principle" [20] is the principle of cohesion [9]. Furthermore the eliminations of decision redundancies increase the cohesions (see the empirical evaluation of new metrics in Section 8). Therefore the fulfilment of "Single Responsibility Principle" can be facilitated by using the new principles to reduce decision redundancies.

4. The definitions of decision and decision raising according to behavioral contracts

The decision structure of programs is defined irrespectively from the program implementation. The realization of this structure strongly determines the optimization level of programs. Decision structures can be optimized by different transformations, by which the behavioral aspects of programs are not changed. In order to examine decision structures, transformation methods and optimization cases (when transformations are required) the introduction of the following concepts is necessary.

4.1. Behavior of decision

The behavior of the decision options D_{O_1}, D_{O_2} of decision D can be declared by behavioral contracts $C_{D_{O_1}}$, $C_{D_{O_2}}$ [18, 27]. (The behavior of a decision option is declared by one behavioral contract.) The D_{O_1}, D_{O_2} are the implementations of decision options, which have to realize the declared decision requirements $(C_{D_{O_1}}, C_{D_{O_2}})$. The changing of decision structure implementations does not always result in the altering of behavioral contracts.

The behavioral contracts of decision options declare the pre-conditions of decision options as their decision predicates and the post-conditions of decision options as state transitions. Behavioral contracts define the data structures, on which the state transitions of behavioral contracts are interpreted. The invariants [19] – which narrow the state-space of behavioral contracts – and the history constraints [19] – which define general state-transitions – are handled as parts of the pre- and post-conditions of decision options. The interpretations of these as invariants and history constraints are not important in consideration of the behavioral contracts of decision options.

A decision case is one case of a decision, in the course of which an appropriate decision option is selected based on the evaluation of its decision predicate. According to a selected decision option, a decision option specified state transition is executed, by which the modifications (the modification or/and the extension of the state) are realized based on the data structure of the decision option (the concerning part of the state description).

4.2. Decision raising

It is a transformation, by which decision dependencies can be eliminated. After using this transformation, the behavior and the data structures of decision options are defined by class hierarchies without using “if-then-else” statements. The subclasses of class hierarchies define the decision options, which are integrated by parent classes. The “interface” of a decision is a polymorph method of a parent class, which has to be overridden by its subclasses [24, 25]. After decision raising decisions can be implemented – without “if-then-else” statements – with references which have the same type as the parent class of decision declaration class hierarchies. They refer to the instances of the subclasses of class hierarchies according to the appropriate decision options [24, 25].

Let D_{NR} be a nonraised decision, where it's decision options $D_{NR_{O_1}}$, $D_{NR_{O_2}}$ implement the behavioral contracts $C_{D_{NR_{O_1}}}$, $C_{D_{NR_{O_2}}}$. The decision D_R is the raised decision of the decision D_{NR} if the behavioral contracts $C_{D_{R_{O_1}}}$, $C_{D_{R_{O_2}}}$ of the decision options $D_{R_{O_1}}$, $D_{R_{O_2}}$ of the decision D_R are defined according to the following: $C_{D_{R_{O_1}}} = C_{D_{NR_{O_1}}}$, $C_{D_{R_{O_2}}} = C_{D_{NR_{O_2}}}$. It means that the behavioral contracts of the decision options of nonraised and raised decisions are equivalent.

If the decisions of decision cases have already been raised, there are two types of decision cases: initial decision cases and reusing decision cases. In the course of initial decision cases, a decision option is archived by using a reference. The type of this reference is the parent class of the class hierarchy of a raised decision. This reference points to an instance of the subclass of the selected decision option. In the course of the following decision cases (reusing decision cases), the result of initial decision case is reused based on the calling polymorph method of archiving polymorph reference without the need to re-evaluate the decision. In case of nonraised decisions, the reevaluation of the decisions is necessary, but in case of raised decisions, the archived decisions can be reused (reusing decision cases), so the re-evaluation of the decisions is not necessary.

5. Avoiding decision redundancies

We must see that decision raisings are reasonable if existing or expected decision redundancies can be eliminated. These redundancies result in implementation dependencies that reduce the maintainability and transparency of codes. The conditions of avoiding decision redundancies are the following:

- Decisions should not Recur Rule 1 (DnR Rule 1): Decisions with equivalent decision predicates and equivalent or partly equivalent data structures and behaviors should not recur, so the equivalent or partly equivalent decision should not be realized again in the course of the same running. Therefore, the declarations of the behavioral and the data structure aspects of such decisions should be defined at one place.

- Decisions should not Recur Rule 2 (DnR Rule 2): Decisions with equivalent decision predicates that define diverse data structures and behaviors should not recur. Accordingly such decisions have to be defined in merged forms at one place.

The aim of avoiding decision redundancies is the reduction of decision dependency. Decision dependency can be interpreted as a type of implementation dependency, which is based on the decision structure of programs. If the change of the behavioral contracts of decision options or the introduction of new decision options forces changes in several decision cases, then there is a decision dependency. Using raised and merged decisions (see Section 6) only the initial decision case needs to be changed if the behavior of a decision option is changed or a new decision option is introduced, the changing of other decision cases is not necessary. Inheritance is used rightfully if decision structure dependencies make it reasonable.

6. Decision merging

Decision merging is the tool of eliminating decision redundancies, which can be interpreted as a new refactoring tool. The cases of decision merging are based on the cases of decision redundancies. I use Liskov's subtype-parent type substitution principle [19] based on behavioral contracts [27].

The behavioral contract C' is the strengthening – in my interpretation, the *real refinement* – of the behavioral contract C: $C' \supset C$ (pronounced as: the behavioral contract C is implicated from the behavioral contract C') if the behavioral contract C' realizes the requirements of the behavioral contract C, but it specifies additional statements as well. *Real refinement* means "strengthening" for post-conditions, but it means "weakening" for pre-conditions. In case of the behavioral contracts of decisions pre-conditions as decision predicates cannot be changed.

The behavioral contract C' is the *refinement* of the behavioral contract C: $C' \supseteq C$ (pronounced as: the behavioral contract C is implicated from or is equal to the behavioral contract C') if the behavioral contract C' realizes the requirements of the behavioral contract C, but it specifies additional statements as well, or their behavioral contracts are equal. Regarding the post-conditions, refinement means equivalence (keeping conditions), or "strengthening" (realizing additional conditions). Regarding the pre-conditions, it means equivalence or "weakening". Pre-conditions as decision predicates cannot be changed.

We must see that there may be partial or total equivalence in the behavioral contracts of nonraised and raised decisions, and in such cases in order to eliminate decision redundancies the using of partial or total decision merging is justified. It is important to note that the merging of nonraised decisions can be realized after raising decisions into class-subclass structures. Decision merging may also be necessary in case of raised decisions, which means that there are decisions merging cases where raised decisions need to be merged. In order to determine whether two decisions can be merged, the behavioral contracts of decisions need to be examined, based on which the fulfillment of one case of decision merging rules can be determined.

According to the previously mentioned rules of avoiding decision redundancies, I describe the conditions where the use of decision merging or partial decision merging is justified below.

6.1. The merging/partial merging of fully or partially equivalent decisions

Decision merging/partial merging is necessary if the decision predicates of decisions are equivalent, and decision option declared data structures and behaviors are equivalent or partially equivalent, which means that one of them extends the other or both of them extend a common part. Evidently if there are raised decisions, which complete the conditions of decision merging, the merging must be executed.

6.1.1. Merging of equivalent or extending decisions ("Decision merging")

Two decisions can be merged in the following cases: If the decision options of two decisions realize equivalent behavioral contracts. If the behavioral contract of one decision is refined, strengthened by the behavioral contract of the other decision.

Let there be decisions D_1, D_2 and decision options $D_{1_{O_1}}, D_{1_{O_2}}, D_{2_{O_1}}, D_{2_{O_2}}$, which realize the behavioral contracts $C_{D_{1_{O_1}}}, C_{D_{1_{O_2}}}, C_{D_{2_{O_1}}}, C_{D_{2_{O_2}}}$. The decisions D_1, D_2 can be merged if there are such behavioral contracts $C_{D_{O_1}}, C_{D_{O_2}}$ for which the following are true:

$$C_{D_{O_1}} = C_{D_{1_{O_1}}}, \quad C_{D_{O_1}} \subseteq C_{D_{2_{O_1}}}, \quad C_{D_{O_2}} = C_{D_{1_{O_2}}}, \quad C_{D_{O_2}} \subseteq C_{D_{2_{O_2}}}.$$

Accordingly if the behavioral contracts $(C_{D_{2_{O_1}}}, C_{D_{2_{O_2}}})$ of one of the decisions that are merged match or refine/strengthen the behavioral contracts $(C_{D_{1_{O_1}}}, C_{D_{1_{O_2}}})$ of the other decision, then the behavioral contracts $(C_{D_{O_1}}, C_{D_{O_2}})$ of the merged decision are equivalent with the behavioral contracts of one of the merging decisions, and the behavioral contracts of the other decision are the refinements of the behavioral contracts of the merged decision.

Therefore, we have to examine the equivalence of the data structures and behaviors of decisions, paying attention to the partial equivalence if one is the refinement of the other one. For these reasons decisions can be merged if their data structures and behaviors perform the following:

- Concerning data structure: The sets of available states based on the variables of decisions are equal, or the states based on the data structure of one of the decisions are a subset of the data structure based states of the other decision.

- Concerning behavior: If the pre- and post-conditions as the behaviors of the decision options of decisions are equal, or one of the decisions declares additional post-conditions while pre-conditions are unchanged.

6.1.2. Merging of partially equivalent decisions ("Decision partial merging")

The partial merging of two decisions is possible if the behavioral contracts of decisions have an equal common part, which is extended with additional conditions by both of the merging decisions. These additional conditions are not part of the merging.

Let there be decisions D_1, D_2 and decision options $D_{1_{O_1}}, D_{1_{O_2}}, D_{2_{O_1}}, D_{2_{O_2}}$, which perform the behavioral contracts $C_{D_{1_{O_1}}}, C_{D_{1_{O_2}}}, C_{D_{2_{O_1}}}, C_{D_{2_{O_2}}}$. The decisions D_1, D_2 can be merged partially if the behavioral contracts $C_{D_{O_1}}, C_{D_{O_2}}$ can be described according to the following:

$$C_{D_{O_1}} \subset C_{D_{1_{O_1}}}, \quad C_{D_{O_1}} \subset C_{D_{2_{O_1}}}, \quad C_{D_{O_2}} \subseteq C_{D_{1_{O_2}}}, \quad C_{D_{O_2}} \subseteq C_{D_{2_{O_2}}}.$$

That is, take the separated common parts $(C_{D_{O_1}}, C_{D_{O_2}})$ of the behavioral contracts of the decision options of the decisions D_1 and D_2, which are to be merged. The behavioral contracts of the decision options of decisions D_1, D_2 are the real-refinements or refinements of the behavioral contracts $C_{D_{O_1}}, C_{D_{O_2}}$ of decision D which is the common part of the decisions D_1 and D_2. It means that at least one of the behavioral contracts of the decision options of every merged decision has a real-refinement connection. (If refinement relations were allowed for every merging behavioral contract, then those cases would be interpreted as partial merging where the behavioral contracts of merging decision options are equivalent, or where one of the decisions extends the other decision. However, these are the cases of the "Merging of equivalent or extending decisions".)

It can be stated that a common part is an intersection of the behavioral contracts of merging decisions, so the following must be met:

$$\begin{aligned} C_{D_{1_{O_1}}} &= C_{D_{O_1}} \wedge C_{D_{1_{O_1}}\text{extend}}, \quad & C_{D_{1_{O_2}}} &= C_{D_{O_2}} \wedge C_{D_{1_{O_2}}\text{extend}} \\ C_{D_{2_{O_1}}} &= C_{D_{O_1}} \wedge C_{D_{2_{O_1}}\text{extend}}, \quad & C_{D_{2_{O_2}}} &= C_{D_{O_2}} \wedge C_{D_{2_{O_2}}\text{extend}} \end{aligned}$$

where the behavioral contracts $C_{D_{1_{O_1}}\text{extend}}, C_{D_{1_{O_2}}\text{extend}}, C_{D_{2_{O_1}}\text{extend}}, C_{D_{2_{O_2}}\text{extend}}$ determine the decision specific aspects of merging decisions.

Accordingly the data structures and the behaviors of decisions must be examined in order to determine whether there is a common part. So decisions can be merged if their data structures and behaviors perform the following:

- Concerning data structure: The state sets which are realized based on the data variables of decisions have an intersection. It means that there is a common part, which is extended by the state sets of the examined decisions.
- Concerning behavior: The post-conditions of decision options specify additional conditions in relation to the post-conditions of a common behavior with equivalent pre-conditions.

6.1.3. Demonstration of merging equivalent or extending decisions

Decisions, decision options, decision predicates are indicated according to the following: Decisions: D1, D2. Merged decision: D. The decision predicates of the decisions D1 and D2: D1P, D2P. The decision options of the decisions D1 and D2: D1.D1_O1, D1.D1_O2, D2.D2_O1, D2.D2_O2. The decision options of the merged decision D: D_O1.D_O, D_O2.D_O, D_O1.D1_O, D_O1.D2_O, D_O2.D1_O, D_O2.D2_O.

I show the facilities of the merging of equivalent decisions based on Activity, Class and Sequence UML diagrams [26]. In the course of demonstrating, the "Merging of partially equivalent decisions" case is avoided. The decision structure of equivalent decisions can be represented with an Activity diagram [26] (Figure 1).

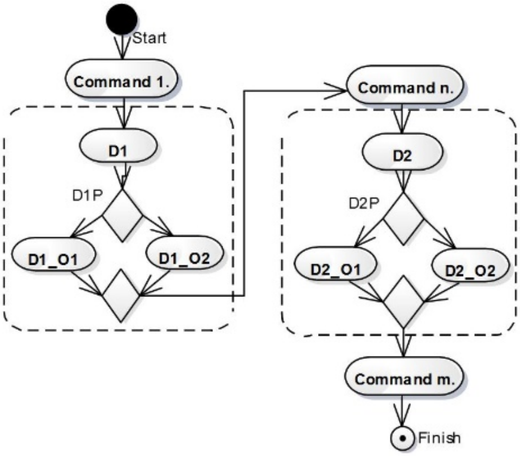


Figure 1: The decision structure of decisions

According to the decision structure, the equivalencies are the following: D1P = D2P, the decision predicates are equivalent. The behavior contracts of the decision options D1_O1 and D2_O1 are equivalent. The behavior contracts of the decision options D1_O2 and D2_O2 are equivalent.

The case of nonmerged decisions can be modeled as follows: One of the possible implementation cases is when separated methods implement the behavior of the decision options of decisions (see Figure 2).

The two cases which are implied from the equivalent decision predicates can be demonstrated with the sequence diagrams in Figure 2. According to these, the similar decision options have to be selected and executed in the course of the same running.

The case of merged decisions can be represented by a class and a sequence diagrams (Figure 3). So the decision option specific operation can be specified by one sequence diagram, on which decision specific behavior is not shown, because it is obscured by polymorph functioning.

The classes and subclasses which represent the merged decisions fulfill the following: The behavior contract of the method D_O1.D_O is equivalent with the behavior contracts of the methods D1.D1_O1, D2.D2_O1, which represent the decision options. The behavior contract of the method D_O2.D_O is equivalent with the behavior contracts of the methods D1.D1_O2, D2.D2_O2, which represent the decision options.

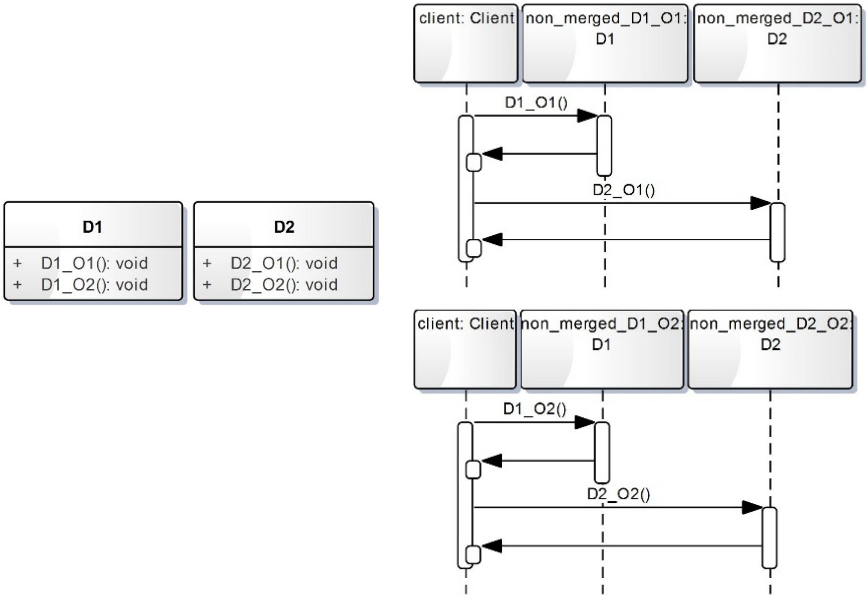


Figure 2: The class diagram of the implementation of the decisions D1 and D2 before decision merging and the sequence diagrams of the decision cases of the decisions D1, D2

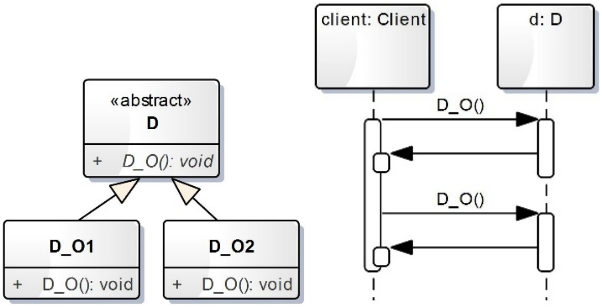


Figure 3: The class diagram representation of the parent class – subclass relationship of merged decisions and the decision cases of merged decisions represented by a sequence diagram

6.2. The merging of decisions with equivalent decision predicates and non-equivalent behavioral contracts

If there are two nonraised or raised decisions, which have equivalent decision predicates, then these decisions can be merged. Accordingly decisions can be merged if their data structures and post conditions are not equivalent from the behavioral aspect, only their decision predicates as preconditions are equivalent.

The decision predicates of the decisions are equivalent: $P_{D_1} = P_{D_2}$ if the decision predicates are equivalent based on the program behavior for every valued-states of state rows/executions. The state of decision predicate is valued if the expression of the decision predicate is evaluated.

Let there be decisions D_1, D_2 and their decision options: $D_{1_{O_1}}$, $D_{1_{O_2}}$, $D_{2_{O_1}}$, $D_{2_{O_2}}$ realize the behavioral contracts $C_{D_{1_{O_1}}}, C_{D_{1_{O_2}}}, C_{D_{2_{O_1}}}, C_{D_{2_{O_2}}}$. The decisions D_1, D_2 can be merged by a decision D with its decision options D_{O_1}, D_{O_2} if there are behavioral contracts $C_{D_{O_1}}, C_{D_{O_2}}$ according to the decision options D_{O_1}, D_{O_2}, which are the disjunctions of the behavior contracts of merged decisions:

$$C_{D_{O_1}} = C_{D_{1_{O_1}}} \vee C_{D_{2_{O_1}}}, \quad C_{D_{O_2}} = C_{D_{1_{O_2}}} \vee C_{D_{2_{O_2}}}.$$

The decision predicates – as the parts of behavior contracts – fulfill the following:

$$P_D = P_{D_1} = P_{D_2},$$

where P_D is the decision predicate of the decision D, furthermore, P_{D_1}, P_{D_2} indicate the decision predicates of the decisions D_1, D_2.

In the following, I show the facilities of decisions with equivalent decision predicates, but different behavior contracts. It is based on Activity, Class and Sequence UML diagrams [26]. The Activity diagram of decisions with equivalent decision predicates equals to the Activity diagram of the decision structure of equivalent decisions. Furthermore, the Class and Sequence diagrams of nonmerged decisions with equivalent decision predicates are equal to the Class and Sequence diagrams of nonmerged equivalent decisions.

According to the decision structure (see Figure 1), the equivalencies are the following: D1P = D2P, the decision predicates are equivalent. The behavior contracts of the decision options D1O1 and D2O1 are NOT equivalent. The behavior contracts of the decision options D1O2 and D2O2 are NOT equivalent.

The case of nonmerged decisions can be modeled according to the following: One of the possible implementation cases is when separated methods implement the behavior of the decision options of decisions. The behavior of methods which represent the decision options is not equivalent (see Figure 2).

The two cases which are implied from the equivalent decision predicates can be demonstrated with the sequence diagrams in Figure 2. According to these – contrary to the previously mentioned decision merging case – the behavior of the executed decision options is not equivalent in the course of the same running. The case of merged decisions can be represented by the class and sequence diagrams of Figure 4.

The class diagram shows how the behavior of two decisions can be defined parallelly in the same subclass, and how the abstract methods represent them in the parent class. The sequence diagram demonstrates the cases, where the different decision options of merged decisions are executed according to the equivalent decision predicates.

The classes and subclasses which represent the merged decisions fulfill the following: The behavior contract of the method D_O1.D1_O / D_O1.D2_O is

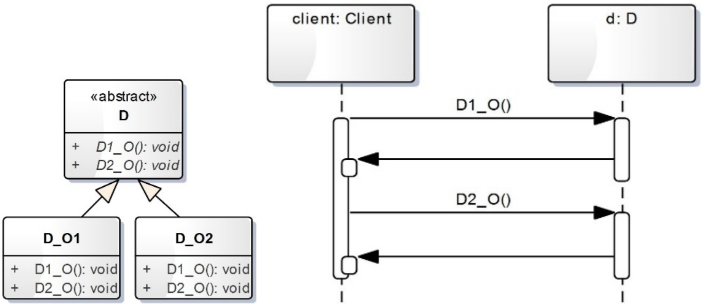


Figure 4: The class diagram representation of the parent class – subclass relationship of merged decisions (with equivalent decision predicates, but different behavior) and the decision cases of merged decisions represented by a sequence diagram

equivalent with the behavior contract of the method D1.D1_O1 / D2.D2_O1. The behavior contract of the method D_O2.D1_O / D_O2.D2_O is equivalent with the behavior contract of the method D1.D1_O2 / D2.D2_O2.

7. The introduction of metrics

The cases of avoiding decision redundancies and the definitions of decision merging specify designing viewpoints that need to be measured and for which measuring methods need to be defined. Accordingly the new metrics which are specified in the following measure the performance of the introduced, new design principles and the decision merging cases which are the theoretical backgrounds of them.

"Metric of decision abstraction": (MDA)

The new metric represents the ratio of polymorph decision cases and the total number of decision cases.

$$\frac{N_{PDC}}{N_{DC}}$$

N_{PDC} – The number of "Polymorph decision cases".
N_{DC} – The total number of "Decision cases".

Under the "Polymorph decision cases" I mean the following: After a decision raising, a decision is realized in a class hierarchy with its classes and subclasses. At the places of use, the callings of polymorph methods represent the decision cases which are called through parent class typed references. Accordingly every polymorph method calling is a decision case. Under the "Decision cases" I mean the conditional statements and polymorph method callings of programs. In conformity with this metric, the decisions represented by conditional statements are not analysed from the point of view of whether the use of decision raising and merging is confirmed or not. It could be fulfilled by analysing the behaviour contracts of decisions. Its value range is 0-1, where the higher value indicates the good structure of systems. According to my presumption the increasing rate of polymorph decision cases decreases complexity, and it promotes the increase of maintainability.

"Ratio of inheritances coming into existence by the elimination of decision redundancies": (RIEDR)

In order to measure the fulfilment of the new "Using inheritance to dissolve decision redundancy" design principle, we have to examine whether inheritances are used for the elimination of decision redundancies. I suppose if a parent-subclass inheritance structure contains polymorph methods (the interfaces of decisions are represented by polymorph methods), then the introduction of that inheritance resulted in the elimination of decision redundancies. Accordingly the fulfilment of this principle can be measured based on the analysis of inheritances, whether they contain polymorph methods which are the interfaces of raised decisions. The fulfilment rate of this principle is better when several inheritances of class structure contain polymorph methods. Determining the ratio:

$$\frac{N_{PI}}{N_I}$$

N_{PI} – The number of inheritances containing polymorph methods.
N_I – The number of inheritances.

Its value range is 0-1, where the higher value indicates the good structure of systems. According to my presumption, a higher metric rate value indicates better code integrity and organizing level. In conformity with this, the increase of this ratio decreases the complexity of codes.

The introduction of additional metrics is suggested according to the previously specified decision redundancies and decision merging cases. These metrics can show the rates of the decision redundancies of programs more sophisticatedly. According to the cases of avoiding decision redundancies, the following new metrics are introduced, which can express the fulfilment rate of the "Avoid decision redundancy" design principle at the same time:

The *"Ratio of equivalent decision cases" (REDC)* metric specifies the ratio of the number of equivalent decision cases (the behaviour contracts are fully or partly equivalent according to the decision cases) and the total number of decision cases. The archived decision cases of raised decisions are not considered as equivalent decision cases.

$$\frac{N_{EDC}}{N_{DC}}$$

N_{EDC} – The number of "Equivalent decision cases".
N_{DC} – The total number of "Decision cases".

Its value range is 0-1, where the higher value indicates the wrong structure of systems.

The *"Ratio of decision cases with equivalent decision predicates" (RDCEDP)* metric specifies the ratio of the number of decision cases the decisions of which have equivalent decision predicates and define diverse behaviours and the total number of decision cases. The archived decision cases of raised decisions are not considered as decision cases with equivalent decision predicates.

$$\frac{N_{DCEP}}{N_{DC}}$$

N_{DCEP} – The number of decision cases with equivalent decision predicates and diverse behaviours.
N_{DC} – The total number of "Decision cases".

Its value range is 0-1, where the higher value indicates the wrong structure of systems.

8. The empirical validations of new metrics

We must see that in order to assess the new metrics REDC and the RDCEDP, the analysis of the behaviour contracts of decision options must be taken into consideration. Using the JML – Java behaviour specification language [18], the behavioural contract based examination of decision options and decision structures is possible [24]. In the future, I intend to analyse the behavioural contract based aspects of decision raising and decisions merging by using the JML specifications of decision structures. According to this I intend to realize the empirical validation of the REDC and RDCEDP metrics based on the JML specific examinations of decision structures.

At the same time, the measurement facilities of the MDA and the RIEDR metrics can be automated easier, therefore the empirical validation of them is easier as well. I analysed the sources of several open source projects from "sourceforge.net"[1] empirically in order to justify the relationship between the decision structure based metrics and code integrity. In the course of these measurements the MDA and the RIEDR metrics were evaluated. The scopes and the sizes of the analysed systems were different, which provide good measurement basis. In the following I described the analysed systems shortly:

ProGuard[2]: It is a free class compressing, optimizing, obfuscator and pre-analyser tool, which can search and eliminate non-used classes, member variables, methods and attributes. The range of 10 analysed versions is 3.0–3.9. The number of examined classes is between 317–391, the number of "useful" lines is between 30,573–39,669.

LWJGL[3]: It supports the development of commercial Java-based games. The

[1] `http://sourceforge.net`
[2] `http://proguard.sourceforge.net`
[3] `http://sourceforge.net/projects/java-game-lib`

range of 10 analysed versions is 2.4.2–2.8.5. The number of examined classes is between 254–416, the number of "useful" lines is between 29,292–42,681.

LaTeXDraw[4]: It is a free, Java-based PSTricks code generator and editing tool. The range of 10 analysed versions is 1.5.0–2.0.6. The number of examined classes is between 69–225, the number of "useful" lines is between 28,368–58,483.

Neuroph[5]: It is a freeware, open source neuron network framework, by which neuron network architectures can be developed. The range of 10 analysed versions is 2.1.0–2.8.0. The number of examined classes is between 69–156, the number of "useful" lines is between 2640–6769.

Finding a properly used outer property as a quality indicator is difficult, furthermore, numerous realized measurements confirmed the correlation between the previously specified metrics (complexity, cohesion and coupling metrics) and outer properties (which are based on maintainability and error-proneness) (Subsection 1.4 and Section 2). In conformity with this, I analysed the relationship between the previously specified complexity, cohesion and coupling metrics and the decision structure based metrics which are specified in this paper. Correlations between them were analysed by the Pearson correlation method, by which linear relationships between independent variables can be detected.

In the course of examinations, I took into consideration the complexity, cohesion and coupling metrics, which were introduced by Chidamber and Kemerer (C&K metrics) [8]. Namely, these metrics are the following: "Weighted methods per class" – WMC, "Coupling between objects" – CBO, "Response for a class" – RFC, "Lack of cohesion in methods" – LCOM. The DIT and the NOC metrics (C&K metrics) were not considered according to the arguments which are listed in the following section. I used the CKJM measurement tool [29] in the course of examinations.

In order to measure the MDA and the RIEDR metrics, which were introduced by this paper, a self-made static code analyzer was used, by which the following parameters of programs can be collected:

- The number of inheritances: The number of inheritance relationships between parent and child classes, including interface implementations as well.
- The number of inheritances, where there is at least one polymorph method.
- The number of branches: Branches are the following conditional statements: "if-then-else", "switch", "while", "for".
- The number of polymorph method callings.

Parameters which can be measured by this tool allow the measuring of the MDA and the RIEDR metrics.

[4] `http://latexdraw.sourceforge.net`

[5] `http://neuroph.sourceforge.net`

8.1. C&K metrics descriptions

The metrics which were specified by Chidember and Kemerer in [8] are described as follows, extended with some intuitive reflections:

"Weighted methods per class" – WMC: It measures the complexity of classes. It has two types. According to the first of them, the weight of methods is 1, therefore the number of methods clearly determines the complexity of classes. According to the second case of this metric, the methods are weighted based on their inner complexity [8]. If the inner complexity of methods is not taken into consideration in the course of their evaluation, then this metric may not work properly, because the change of inner method complexity could compensate for the increasing number of methods.

"Depth of Inheritance Tree" – DIT: It is the maximum depth (the case of multiple inheritances) of a class hierarchy, from the examined class to the root parent class [8]. In case of appropriate inheritance usage, a higher DIT value means more complex decision structures, the decisions of which include each other. It also describes problem complexity, the optimizing of which cannot be realized based on the reduction of the levels of class hierarchies, because the complexity of programs is not changeable. But if the introduction of inheritances is not based on the rules of decision merging, then the elimination of non-properly used inheritances may result in the decrease of DIT metric values, according to the appropriate code structure realizations.

"Number of children" – NOC: It is the number of the subclasses of a class. The high number of subclasses increases the probability of non-proper abstractions. Accordingly if a class has lots of subclasses, then it may be the result of non-proper inheritance usage [8]. The metric is not capable of measuring the number of rightly used inheritances. In several cases, the decrease in the number of child classes by introducing new inheritance levels is not confirmed. Based on decision redundancies, it can be found out that there are decisions that can be "linearised" to a level, namely, their merging can be used to lower the number of subclasses.

"Coupling between objects" – CBO: It determines the number of connections between classes. The exaggerated usage of coupling is detrimental to modularity and it decreases re-usability. So the independency of a class increases re-using capability [8]. In the course of the measure of coupling, inheritances are taken into consideration as one type of coupling, which disfigures the measure of dependencies between coupling and re-usage capability. The unsuitable consideration of inheritances as coupling leads to the incorrect conclusion that NOC metric values are high if classes have high CBO metric values [8]. This conclusion is not good, because inheritances do not necessarily spoil the structural quality of coupling. The aim of inheritance is not class reusing, but the extension of classes with a specific behaviour. This approach was introduced by the "Liskov Substitution Principle" (LSP) [12, 19, 27, 28] and by the inheritance cohesion [10]. Inheritance cohesion is strong if inheritances are used to introduce specialized child classes. Respectively it is weak if the main aim of inheritances is reusing.

"Response for a class" – RFC: The response set of a class consists of those methods that can be executed as an effect of the messages sent by the instances of a given class [8].

"Lack of Cohesion in Methods" – LCOM: The interpretation of this metric is based on the dependencies between the methods, which can be determined by the sets of the member variables of classes used by the method. The lack of cohesion may mean that classes should be split into subclasses [8]. The MDA indicates the increase of the number of polymorph method invocations, accordingly the number of raised and merged decisions is growing as well. This results in the decrease of decision separation based behaviour, which causes low cohesion within a class. At the same time, the LCOM metric has more similarities with the REDC and the RDCEDP metrics, which are based on the similarity and the overlapping of behaviour contracts. The concept of these metrics is more similar to the cohesion theoretical basis, by which the functional separation of classes can be expressed. Based on this idea, my future plan is to investigate whether cohesion can be determined by the examination of the similarities between the behaviour contracts of methods.

From the mentioned C&K metrics DIT and NOC metrics are strongly related to the complexity of inheritance hierarchies. At the same time, these two metrics do not clarify the cases where inheritances can be used rightfully. So it is possible that complex structures signed by DIT and NOC only indicate the complexity of the realized problem, which can be optimal from the point of view of the code structure. Therefore I do not analyse the aspects of these metrics.

8.2. The measuring results of the MDA

For the measurement of the MDA the following ratio must be determined:

$$\frac{N_{PMI}}{N_{CS} + N_{PMI}}$$

N_{PMI} – The number of polymorph method invocations.
N_{CS} – The number of conditional statements.

This ratio can be determined using the results of the developed measuring tool. According to my supposition, N_{PMI} approximately determines N_{PDC}, and N_{DC} can be determined by summing up N_{CS} and N_{PMI}.

I analysed the versions of the previously described ProGuard, LWJGL, LaTeX-Draw, Neuroph projects. The correlations between the MDA and the WMC, CBO,

	WMC	**CBO**	**RFC**	**LCOM**
ProGuard	**−0.893**	**−0.881**	**−0.905**	**−0.773**
LWJGL	**−0.907**	0.821	**−0.678**	**−0.928**
LaTeXDraw	**−0.648**	0.508	**−0.623**	**−0.684**
Neuroph	−0.433	0.067	−0.130	**−0.685**

Table 1: The Pearson product-moment correlation coefficients between the MDA and the WMC, CBO, RFC, LCOM metrics

RFC, LCOM metrics [8] were examined based on the Pearson product-moment correlation coefficient (see Table 1). According to the measurement of the three systems, the correlation between the MDA and the WMC metric is significant, however, there is one system (Neuroph), where the correlation is low, but exists. The correlation measurements between the MDA and the CBO metric are ambiguous, therefore there is no correlation between them. Based on the measurements of three systems, the correlation between the MDA and the RFC metric is significant, however, there is one correlation measurement which indicates no correlation between them (Neuroph system). The measurements confirmed the significant relationship between the MDA and the LCOM metric. In conformity with the measurements, the correlation between these metrics is the most significant.

8.3. The measuring results of the RIEDR metric

For the measurement of the RIEDR metric the following ratio must be determined:

$$\frac{N_{PI}}{N_I}$$

N_{PI} – The number of inheritances containing polymorph methods.
N_I – The number of inheritances.

I analysed the versions of the previously described ProGuard, LWJGL, LaTeXDraw, Neuroph projects. The correlations between the RIEDR metric and the WMC, CBO, RFC, LCOM metrics [8] were examined based on the Pearson product-moment correlation coefficient (see Table 2).

	WMC	**CBO**	**RFC**	**LCOM**
ProGuard	**−0.830**	**−0.863**	**−0.798**	**−0.925**
LWJGL	**−0.866**	0.739	−0.255	**−0.772**
LaTeXDraw	0.737	0.446	0.745	0.681
Neuroph	−0.278	0.481	0.330	**−0.861**

Table 2: The Pearson product-moment correlation coefficients between the RIEDR and the WMC, CBO, RFC, LCOM metrics

The correlation measurements between the RIEDR metric and the WMC, CBO, RFC metrics are ambiguous. Accordingly, there is no correlation between these metrics. In conformity with the measurements of three systems, the correlation between the RIEDR metric and the LCOM is significant, but there is one system (LaTeXDraw), where the measurement indicates inverse correlation. In order to determine the reason of the inverse correlation in case of the LaTeXDraw system further examinations are needed.

8.4. The summary of empirical validations

In case of the WMC metric [8] the optionally considered inner complexity of methods promote the correlation with the MDA, because inner complexity is related to the quality of decision structures which can be measured by the new metrics.

The relationship between the LCOM metric [8] and the new metrics can be perceived based on behaviour contracts [18, 27], which should be considered in the course of the determination of cohesion. These behaviour contracts specify the basic concepts of decision merging examinations and the introduction of new decision structure quality specific metrics. These supposed relationships were confirmed by empirical validations.

The empirically perceived relationship between the RFC metric [8] and the MDA was not supposed intuitively. To find the cause of the empirical connection between the two metrics requires further examinations.

9. Conclusions

In the course of the paper the definitions of decision, decision raising [24, 25] and the newly introduced decision merging are extended based on the concept of behavioural contract [18, 27].

Using the behaviour contract-based definitions, the behavioural contract specific aspects of the transformations of decision raisings and decision merging can be showed. Using the JML – Java behaviour specification language [18], the behavioural contract-based examination of decision structures is possible [24]. In the future, I intend to analyse the behavioural contract-based aspects of decision raising and decisions merging by using the JML specifications of decision structures.

Based on the described concepts of decision redundancies and the rules of decision merging, I introduced new object-oriented design principles ("Using inheritance to dissolve decision redundancy", "Avoid decision redundancy"). These principles determine the cases, where the use of inheritance as an object-oriented tool is justified. Several existing object-oriented design principles are engaged in detecting the cases where the use of inheritance vs. object composition is confirmed. I intend to examine the relationship between the existing design principles and the quality of decision structures.

I will deal with the examination of the designing circumstances of design patterns. In the course of the examination of the decision structures of design patterns, I plan to examine high-level optimizing facilities and low-level refactoring methods. One of the new directions could be the examination of the suspected relationship between decision structures and design patterns. According to this relationship the decision structure circumstances of design patterns appear in Use Case models [26]. In conformity with this, a new research direction is to find out how the design patterns appear in Use Case models, or rather, how the decision structures of design patterns reflect on the level of Use Cases.

I initiated new object-oriented metrics that give the opportunity to examine the quality of decision structures. The introduced MDA and RIEDR metrics are examined empirically compared to the previously specified complexity, cohesion and coupling metrics. The correlations between them are analysed by the Pearson correlation method, by which the linear relationship between independent variables can be analysed. According to the measurements, the correlations between the MDA and the WMC, RFC, LCOM [8] are significant, furthermore, there is a significant correlation between the RIEDR metric and the LCOM [8] metric as well. In the cases of the WMC and LCOM metrics [8], the detected relationship can be perceived intuitively, but the empirically confirmed relationship between the RFC metric [8] and the MDA requires additional examinations. The relationship between the LCOM [8] and the decision structure based metrics is based on the dependencies between the functional separation signing capability of cohesion and decision structure anomalies.

References

[1] L. Badri, M. Badri, and B. Gueye. Revisiting class cohesion: An empirical investigation on several system. *Journal of Object Technology*, 7(6):55–75, 2008.

[2] M. Badri, L. Badri, and F. Touré. Empirical analysis of object-oriented design metrics: Towards a new metric using control flow paths and probabilities. *Journal of Object Technology*, 8(6):123–142, 2009.

[3] J.M. Bieman and B.K. Kang. Cohesion and reuse in an object-oriented system. *In Proceedings of the ACM Symposium on Software Reusability (SSR'95)*, pages 259–262, 1995.

[4] G. Booch, R.A. Maksimchuk, M.W. Engel, B.J. Young, J. Conallen, and K.A. Houston. *Object-Oriented Analysis and Design with Applications.* Addison Wesley Longman Publishing Co., Inc., 3rd edition, 2007.

[5] L.C. Briand, J.W. Daly, and J. Wüst. A unified framework for cohesion measurement in object-oriented systems. *Empirical Software Engineering*, 3(1):65–117, 1998.

[6] H.S. Chae and Y.R. Kwon. A cohesion measure for classes in object-oriented systems. *In Proceedings of the 5th. International Software Metrics Symposium. Bethesda, MD*, pages 158–166, 1998.

[7] S.R. Chidamber and C.F. Kemerer. Towards a metrics suite for object oriented design. *In Proceedings of Conference on Object-Oriented Programming Systems, Languages, and Applications (OOPSLA '91)*, pages 197–211, 1991.

[8] S.R. Chidamber and C.F. Kemerer. A metrics suite for object oriented design. *IEEE Transactions on Software Engineering*, 20(6):476–493, 1994.

[9] T. DeMarco. *Structured analysis and system specification.* Yourdon Press, Prentice Hall, Inc., 1979.

[10] J. Eder, G. Kappel, and M. Schrefl. Coupling and cohesion in object-oriented systems. *Technical Report, University of Klagenfurt, Austria*, pages 1–34, 1994.

[11] M. Fowler, K. Beck, J. Brant, W. Opdyke, and D. Roberts. *Refactoring: Improving the design of existing code.* Addison Wesley Longman Publishing Co., Inc., 1999.

[12] E. Gamma, R. Helm, R. Johnson, and J. Vlissides. *Design Patterns: Elements of Reusable Object-Oriented Software.* Addison-Wesley Professional Computing Series, 1995.

[13] B. Henderson-Sellers. *Object-Oriented Metrics. Measures of Complexity.* Prentice Hall, Inc., 1996.

[14] M. Hitz and B. Montazeri. Measuring coupling and cohesion in object-oriented systems. *In Proceedings of the International Symposium on Applied Corporate Computing, Monterrey, Mexico*, 50:75–76, 1995.

[15] H. Kabaili, R.K. Keller, and F. Lustman. Cohesion as changeability indicator in object-oriented systems. *In Proceedings of the 5th European Conference on Software Maintenance and Reengineering (CSMR 2001), IEEE, Lisbon, Portugal*, pages 39–46, 2001.

[16] J. Kerievsky. *Refactoring to patterns.* Addison Wesley Longman Publishing Co., Inc., 2004.

[17] C. Larman. *Applying UML and Patterns: An Introduction to Object-Oriented Analysis and Design and the Unified Process.* Prentice Hall, Inc., 3rd edition, 2005.

[18] G.T. Leavens and Y. Cheon. Design by contract with JML. *Dept. of Computer Science, Iowa State University, Dept. of Computer Science, University of Texas at El Paso*, pages 1–13, 2006.

[19] B.H. Liskov and J.M. Wing. A behavioral notion of subtyping. *ACM Transactions on Programming Languages and Systems*, 16(6):1811–1841, 1994.

[20] R.C. Martin and M. Micah. *Agile principles, patterns, and practices in C#.* Prentice Hall, Inc., 2006.

[21] T.J. McCabe. A complexity measure. *IEEE Transactions on Software Engineering*, 2(4):308–320, 1976.

[22] B. Meyer. Applying “design by contract”. *IEEE Computer*, 25(10):40–51, 1992.

[23] J.C. Munson and T.M. Khoshgoftaar. *Handbook of Software Reliability Engineering. Chapter 12.: Software Metrics for Reliability Assessment.* IEEE Computer Society Press, McGraw-Hill, 1996.

[24] Sz. Márien. Decision based examination of object-oriented methodology using JML. *Annales Mathematicae et Informaticae*, 35:95–121, 2008.

[25] Sz. Márien. Decision based examination of object-oriented programming and design patterns. *Teaching Mathematics and Computer Science*, 6(1):83–109, 2008.

[26] J. Rumbaugh, I. Jacobson, and G. Booch. *The Unified Modeling Language reference manual.* Addison Wesley Longman Publishing Co., Inc., 2nd edition, 2004.

[27] W. Schreiner. From types to contracts: Supporting by light-weight specifications the liskov substitution principle. *Technical Report no. 10-22 in RISC Report Series. Research Institute for Symbolic Computation (RISC), Johannes Kepler University Linz, Austria*, 2010.

[28] R.W. Sebesta. *Concepts of Programming Languages.* Addison Wesley Longman Publishing Co., Inc., 7th edition, 2006.

[29] D. Spinellis. Tool writing: A forgotten art? *IEEE Software*, 22(4):9–11, 2005.

[30] Y. Wand and R. Weber. An ontological model of an information system. *IEEE Transactions on Software Engineering*, 16(11):1282–1292, 1990.

[31] S. Yacoub, T. Robinson, and H.H. Ammar. Dynamic metrics for object oriented designs. *In Proceedings of the 6th International Software Metrics Symposium*, pages 50–61, 1999.

Annales Mathematicae et Informaticae
47 (2017) pp. 177–183
http://ami.uni-eszterhazy.hu

Period of balancing sequence modulo powers of balancing and Pell numbers

Bijan Kumar Patel[a], **Utkal Keshari Dutta**[b]
Prasanta Kumar Ray[c*]

[a]International Institute of Information Technology, Bhubaneswar, India
iiit.bijan@gmail.com

[b]Veer Surendra Sai University of Technology, Burla, India
utkaldutta@gmail.com

[c]Sambalpur University, Sambalpur, India
prasantamath@suniv.ac.in

Submitted January 10, 2017 — Accepted June 23, 2017

Abstract

The period of balancing numbers modulo m, denoted by $\pi(m)$, is the least positive integer n such that $(B_n, B_{n+1}) \equiv (0, 1) \pmod{m}$, where B_n is the n-th balancing number. While studying periodicity of balancing numbers, Panda and Rout found the results for $\pi(B_n)$ and $\pi(P_n)$, where P_n denotes the n-th Pell number. In this article we obtain the formulas of $\pi(B_n^{k+1})$ and $\pi(P_n^{k+1})$ for all $k \geq 1$.

Keywords: Balancing numbers; Lucas-balancing numbers; Periodicity; p-Adic order.

MSC: 11A05, 11B39, 11B50

1. Introduction

Let the sequence of balancing and Pell numbers be $\{B_n\}_{n\geq 1}$ and $\{P_n\}_{n\geq 1}$ respectively. These two sequences satisfy the recurrence relations $B_{n+1} = 6B_n - B_{n-1}$

[*] Corresponding author. E-mail: prasantamath@suniv.ac.in

and $P_{n+1} = 2P_n + P_{n-1}$ with initial values $(B_0, B_1) = (0, 1) = (P_0, P_1)$ [1, 7]. Balancing numbers B_n and their associate Lucas-balancing numbers C_n are obtained from the Pell equation $C_n^2 - 8B_n^2 = 1$ [6]. The Lucas-balancing numbers satisfy the same recurrence relation as that of balancing numbers but with different initials $(C_0, C_1) = (1, 3)$ [6]. Some developments on balancing numbers and their related sequences can be found in [2, 3, 4, 12, 13].

Panda and Rout, in [8], defined the period of balancing numbers modulo m, denoted by $\pi(m)$, the least positive integer n satifying $(B_n, B_{n+1}) \equiv (0, 1) \pmod{m}$. They have derived the formulas $\pi(B_n) = 2n$ and $\pi(P_n) = n$ or $2n$ for the parity of n [8]. Later, Patel and Ray [9] studied the rank $r(m)$ and order $o(m)$ of the balancing sequence and established some connections between the period, rank and order. Among other relations, one important connection between them is that the product of rank and order is equal to period.

In [5], Marques obtained the order of appearance (rank) of Fibonacci numbers modulo powers of Fibonacci and Lucas numbers and derived the formula of $r(L_n^k)$ in some cases. Later, Pongsriiam extended the work of Marques and obtained the complete formula of $r(L_n^k)$ for all $n, k \geq 1$ [10]. Recently, Sanna [14] studied the p-adic valuation of Lucas sequences $u_n = au_{n-1} + bu_{n-2}$ with $u_0 = 0$ and $u_1 = 1$ for all $n \geq 2$. Among other results he has also established the following identity.

Theorem 1.1. *If p is a prime number such that $p \nmid b$, then*

$$\nu_p(u_n) = \begin{cases} \nu_p(n) + \nu_p(u_p) - 1, & \text{if } p|\Delta, p|n; \\ 0, & \text{if } p|\Delta, p \nmid n; \\ \nu_p(n) + \nu_p(u_{pr(p)}) - 1, & \text{if } p \nmid \Delta, r(p)|n, p|n; \\ \nu_p(u_{pr(p)}), & \text{if } p \nmid \Delta, r(p)|n, p \nmid n; \\ 0, & \text{if } p \nmid \Delta, r(p) \nmid n. \end{cases}$$

for each positive integer n, where $\Delta = a^2 + 4b$.

In the present study, we obtain the formula of $\pi(B_n^{k+1})$ and $\pi(P_n^{k+1})$ for every $k \geq 1$.

2. Preliminaries

The following results concerning about the periodicity of balancing numbers are found in [8].

Lemma 2.1. *If m divides n, then $r(m)$ divides $r(n)$.*

Lemma 2.2. *If m divides B_n if and only if $r(m)$ divides n.*

The following result is found in [6].

Lemma 2.3. *For every positive integers m and n, $B_{m+n} = B_mC_n + C_mB_n$.*

The following result is found in [7].

Lemma 2.4. *For $n \geq 0$, $C_n = 2Q_n^2 - (-1)^n$.*

The following two corollaries are direct consequences of theorem 1.1. The first one obtained for $(a, b) = (6, -1)$ and the second one for $(a, b) = (2, 1)$.

Corollary 2.5. *For all prime p,*

$$\nu_p(B_n) = \begin{cases} \nu_p(n), & \text{if } p|\Delta, p|n, \\ \nu_p(n) + \nu_p(B_{r(p)}), & \text{if } p \nmid \Delta, r(p)|n, \\ 0, & \text{otherwise.} \end{cases}$$

Corollary 2.6. *For all prime p,*

$$\nu_p(P_n) = \begin{cases} \nu_p(n), & \text{if } p|\Delta, p|n, \\ \nu_p(n) + \nu_p(P_{r(p)}), & \text{if } p \nmid \Delta, r(p)|n, \\ 0, & \text{otherwise.} \end{cases}$$

3. Main results

In order to prove the main results we need the following lemma.

Lemma 3.1. *For $n \geq 2, k \geq 1$, $r(B_n^{k+1}) = nB_n^k$ and $r(P_n^{k+1}) = nP_n^k$.*

Proof. Since B_n divides B_n^{k+1}, $r(B_n) = n$ divides $r(B_n^{k+1})$ by Lemma 2.1. Further to prove $r(B_n^{k+1})$ divides nB_n^k analogously B_n^{k+1} divides $B_{nB_n^k}$ [Lemma 2.2], it is enough to exhibit $\nu_p(B_n^{k+1}) \leq \nu_p(B_{nB_n^k})$, for all prime p. Now for $p = 2$, using the Corollary 2.5, we have

$$\nu_2(B_{nB_n^k}) = \nu_2(nB_n^k) = \nu_2(n) + \nu_2(B_n^k) = \nu_2(B_n) + \nu_2(B_n^k) = \nu_2(B_n^{k+1}).$$

On the other hand for all odd primes, we have

$$\begin{aligned} \nu_p(B_{nB_n^k}) &= \nu_p(nB_n^k) + \nu_p(B_{r(p)}) \\ &= \nu_p(n) + \nu_p(B_{r(p)}) + \nu_p(B_n^k) \\ &= \nu_p(B_n) + \nu_p(B_n^k) \\ &= \nu_p(B_n^{k+1}). \end{aligned}$$

Now to see that $r(B_n^{k+1}) = nB_n^k/p^t$, for some $t \geq 0$. It suffices to show that B_n^{k+1} does not divides $B_{\frac{n}{2}B_n^k}$ for $p = 2$. That is $\nu_2\left(B_{\frac{nB_n^k}{2}}\right) < \nu_2(B_n^{k+1})$. So,

$$\begin{aligned} \nu_2\left(B_{nB_n^k/2}\right) &= \nu_2(nB_n^k/2) \\ &= \nu_2(n) + \nu_2(B_n^k) - \nu_2(2) \\ &= \nu_2(B_n) + \nu_2(B_n^k) - 1 \\ &< \nu_2(B_n^{k+1}). \end{aligned}$$

Again for all odd primes

$$\begin{aligned}\nu_p(B_{nB_n^k/p}) &= \nu_p(nB_n^k/p) + \nu_p(B_{r(p)})\\ &= \nu_p(B_n) + \nu_p(B_n^k) - \nu_p(p)\\ &< \nu_p(B_n^{k+1}),\end{aligned}$$

which completes the first part of the lemma. The second part can be proved analogously. □

Theorem 3.2. *For $n \geq 2$ and $k \geq 1$,* $\pi(B_n^{k+1}) = \begin{cases} nB_n^k, & \textit{if } n \equiv 0 \pmod 2;\\ 2nB_n^k, & \textit{if } n \equiv 1 \pmod 2.\end{cases}$

Proof. By virtue of Lemma 3.1, $B_{nB_n^k} \equiv 0 \pmod{B_n^{k+1}}$. In order to prove the theorem, it suffices to show the following cases, $B_{nB_n^k+1} \equiv 1 \pmod{B_n^{k+1}}$ for n is even only and for odd $B_{2nB_n^k+1} \equiv 1 \pmod{B_n^{k+1}}$. Using Lemma 2.3, we have

$$B_{2nB_n^k+1} = B_{2nB_n^k}C_1 + C_{2nB_n^k}B_1 \equiv C_{2nB_n^k} \pmod{B_n^{k+1}}. \tag{3.1}$$

Therefore using the identity $C_n^2 = 8B_n^2 + 1$ [6], we have

$$C_{2nB_n^k}^2 = 8B_{2nB_n^k}^2 + 1 \equiv 1 \pmod{B_n^{k+1}}.$$

Consequently, $C_{nB_n^k}^2 \equiv 1 \pmod{B_n^{k+1}}$, which implies

$$Q_{2nB_n^k}^2 \equiv 1 \pmod{B_n^{k+1}}. \tag{3.2}$$

By Lemma 2.4, we get

$$C_{2nB_n^k} = 2Q_{2nB_n^k}^2 - (-1)^{2nB_n^k} \equiv 1 \pmod{B_n^{k+1}},$$

which completes the second case. Again the use of Lemma 2.3 gives the identity $B_{nB_n^k+1} \equiv C_{nB_n^k} \pmod{B_n^{k+1}}$. In order to prove the first case, it suffices to show $C_{nB_n^k} \equiv 1 \pmod{B_n^{k+1}}$ for n is even only. Let $n = 2m$. As $C_{2n} = 16B_n^2 + 1$ [6], $C_{2mB_{2m}^k} = 16B_{mB_{2m}^k}^2 + 1$. We will now show that B_{2m}^{k+1} divides $B_{mB_{2m}^k}^2$, that is, $\nu_p(B_{2m}^{k+1}) \leq \nu_p(B_{mB_{2m}^k}^2)$ for all prime p.

When $p = 2$, by virtue of Corollary 2.5, we have

$$\begin{aligned}\nu_2(B_{mB_{2m}^k}^2) &= 2\cdot\nu_2(B_{mB_{2m}^k}) = 2\cdot\nu_2(mB_{2m}^k)\\ &= 2\cdot\nu_2(m) + 2\cdot\nu_2(B_{2m}^k)\\ &\geq \nu_2(B_{2m}^{2k})\\ &\geq \nu_2(B_{2m}^{k+1}).\end{aligned}$$

On the other hand, for any odd prime p, we obtain

$$\begin{aligned}\nu_p(B^2_{mB^k_{2m}}) &= 2\cdot\nu_p(B_{mB^k_{2m}})\\ &= 2[\nu_p(mB^k_{2m})+\nu_p(B_{r(p)})]\\ &= \nu_p(B^{2k}_{2m})+2\cdot\nu_p(m)+2\cdot\nu_p(B_{r(p)})\\ &\geq \nu_p(B^{k+1}_{2m}).\end{aligned}$$

This completes for n even. Now for n is odd, that is, for $n=2m+1$, we need to show that

$$C_{(2m+1)B^k_{2m+1}} \not\equiv 1 \pmod{B^{k+1}_{2m+1}}.$$

From Eq. (3.2), we have

$$Q^2_{2(2m+1)B^k_{2m+1}} \equiv 1 \pmod{B^{k+1}_{2m+1}}.$$

Since $C_{(2m+1)B^k_{2m+1}} = 2Q^2_{(2m+1)B^k_{2m+1}}+1$, it follows that

$$C_{(2m+1)B^k_{2m+1}} \not\equiv 1 \pmod{B^{k+1}_{2m+1}}.$$

This ends the proof. □

Theorem 3.3. *For $n\geq 2$ and $k\geq 1$,*

$$\pi(P^{k+1}_n) = \begin{cases} nP^k_n, & \text{if } n\equiv 0 \pmod 2;\\ 2nP^k_n, & \text{if } n\equiv 1 \pmod 2.\end{cases}$$

Proof. In order to derive the above result, we need to prove the following two cases, $B_{nP^k_n+1}\equiv 1 \pmod{P^{k+1}_n}$ for n is even only and $B_{2nP^k_n+1}\equiv 1 \pmod{P^{k+1}_n}$ for n is odd. For the proof of the first case, since $B_{nP^k_n+1}\equiv C_{nP^k_n} \pmod{P^{k+1}_n}$ by Lemma 2.3, it suffices to show that $C_{nP^k_n}\equiv 1 \pmod{P^{k+1}_n}$ for n even only.

Let $n=2m$, we need to show that $C_{2mP^k_{2m}}\equiv 1 \pmod{P^{k+1}_{2m}}$. As $C_{2mP^k_{2m}} = 16B^2_{mP^k_{2m}}+1$, it is enough to prove that P^{k+1}_{2m} divides $B^2_{mP^k_{2m}}$, that is, $\nu_p(P^{k+1}_{2m})\leq \nu_p(B^2_{mP^k_{2m}})$ for all prime p. By virtue of Corollary 2.6, for even prime p,

$$\begin{aligned}\nu_2(B^2_{mP^k_{2m}}) &= 2\cdot\nu_2(B_{mP^k_{2m}}) = 2\cdot\nu_2(mP^k_{2m})\\ &= 2\cdot\nu_2(m)+2\cdot\nu_2(P^k_{2m})\\ &\geq \nu_2(P^{k+1}_{2m}).\end{aligned}$$

Further for any odd prime p, we have

$$\begin{aligned}\nu_p(B^2_{mP^k_{2m}}) &= 2\cdot\nu_p(B_{mP^k_{2m}})\\ &= 2[\nu_p(mP^k_{2m})+\nu_p(P_{r(p)})]\\ &= \nu_p(P^{2k}_{2m})+2\cdot\nu_p(m)+2\cdot\nu_p(P_{r(p)})\end{aligned}$$

$$\geq \nu_p(P_{2m}^{k+1}).$$

This completes for n even.

Now for n is odd, we have to claim

$$C_{(2m+1)P_{2m+1}^k} \not\equiv 1 \pmod{P_{2m+1}^{k+1}}.$$

Since $C_{nP_n^k}^2 \equiv 1 \pmod{P_n^{k+1}}$, which implies

$$Q_{2nP_n^k}^2 \equiv 1 \pmod{P_n^{k+1}}. \tag{3.3}$$

From (3.3), we have

$$Q_{2(2m+1)P_{2m+1}^k}^2 \equiv 1 \pmod{P_{2m+1}^{k+1}}.$$

Further, using Lemma 2.4, we have $C_{(2m+1)P_{2m+1}^k} = 2Q_{(2m+1)P_{2m+1}^k}^2 + 1$ and the result follows. This ends the proof of first case. Furthermore,

$$B_{2nP_n^k+1} \equiv C_{2nP_n^k} = 16B_{nP_n^k}^2 + 1 \equiv 1 \pmod{P_n^{k+1}}.$$

This completes the proof of the second case. □

Acknowledgement. The authors are grateful to the anonymous referee for his useful suggestions and comments which improved the quality of this manuscript.

References

[1] Behera, A., Panda, G. K., On the square roots of triangular numbers, *Fibonacci Quart.*, Vol. 37 (1999), 98–105.

[2] Berczes, A., Liptai, K., Pink, I., On generalized balancing numbers, *Fibonacci Quart.*, Vol. 48 (2010), 121–128.

[3] Kovacs, T., Liptai, K., Olajos, P., On (a, b)-balancing numbers, *Publ. Math. Debrecen*, Vol. 77 (2010), 485–498.

[4] Liptai, K., Luca, F., Pinter, A., Szalay, L., Generalized balancing numbers, *Indag. Math. (N.S.)*, Vol. 20 (2009), 87–100.

[5] Marques, D., The order of appearance of powers of Fibonacci and Lucas numbers, *Fibonacci Quart.*, Vol. 50 (2012), 239–245.

[6] Panda, G. K., Some fascinating properties of balancing numbers, *Congr. Numer.*, Vol. 194 (2009), 185–189.

[7] Panda, G. K., Ray, P. K., Some links of balancing and cobalancing numbers with Pell and associated Pell numbers, *Bull. Inst. Math. Acad. Sin. (N.S.)*, Vol. 6 (2011), 41–72.

[8] Panda, G. K., Rout, S. S., Periodicity of balancing numbers, *Acta Math. Hungar.*, Vol. 143 (2014), 274–286.

[9] Patel, B. K., Ray, P. K., The period, rank and order of the sequence of balancing numbers modulo m, *Math. Rep. (Bucur.)*, Vol. 18 (2016), Article No.9.

[10] Pongsriiam, P., A complete formula for the order of appearance of the powers of Lucas numbers, *Commun. Korean Math. Soc.*, Vol. 31 (2016), 447–450.

[11] Ray, P. K., Curious congruences for balancing numbers, *Int. J. Contemp. Math. Sci.*, Vol. 7 (2012), 881–889.

[12] Ray, P. K., Some congruences for balancing and Lucas-balancing numbers and their applications, *Integers*, Vol. 14 (2014), #A8.

[13] Rout, S. S., Balancing non-Wieferich primes in arithmetic-progression and *abc* conjecture, *Proc. Japan Acad.*, Vol. 92 (2016), 112–116.

[14] Sanna, C., The p-Adic valuation of Lucas sequences, *Fibonacci Quart.*, Vol. 54 (2016), 118–124.

Annales Mathematicae et Informaticae
47 (2017) pp. 185–200
http://ami.uni-eszterhazy.hu

Comparison and affine combination of generalized barycentric coordinates for convex polygons

Ákos Tóth

Department of Computer Graphics and Image Processing
Faculty of Informatics, University of Debrecen
toth.akos@inf.unideb.hu

Submitted April 10, 2017 — Accepted June 18, 2017

Abstract

In this paper, we study and compare different types of generalized barycentric coordinates in detail, including Wachspress, discrete harmonic and mean value coordinates for convex, n-sided polygons. Contour lines are computed in each barycentric coordinate method, and curvature plots of these contour line curves are visualized. Moreover, different distortions of uniform patterns are also shown, providing exact visual method to compare these methods. To overcome the shortcomings of different generalized barycentric computations, affine combination of methods is provided.

Keywords: barycentric coordinates, Wachspress coordinates, discrete harmonic coordinates, mean value coordinates, affine combination

MSC: 52B55, 52A38, 65D05

1. Introduction

Barycentric coordinates were first introduced by Möbius [1] in 1827. Any point v inside a triangle v_1, v_2, v_3 can be obtained by weighted sum of these vertices, if corresponding weights w_1, w_2, w_3 are placed at the vertices of triangle. These weights w_1, w_2, w_3 are the barycentric coordinates of point v. This can be generalized for arbitrary n-sided polygons in the plane where an inner point v can be defined as the weighted sum of vertices $v_1, \dots, v_n$ as

$$v = \frac{w_1(v)v_1 + \ldots + w_n(v)v_n}{w_1(v) + \ldots + w_n(v)}.$$

These barycentric coordinates can be normalized that values sum to one, thus they vary linearly inside the polygon. Therefore many applications often use them to interpolate different values which are placed at the vertices of the polygon. In computer graphics, this interpolation is used e.g. for shading or geometry deformation.

In the last couple of years, many approaches have been released which tried to generalize the barycentric coordinates. The first generalization appeared in Wachspress's pioneering work [2] in 1975. The computation of Wachspress coordinates is simple for any convex polygons, because they are rational functions and their derivatives can also be easily evaluated. They have many nice properties [3] such as affine invariance or smoothness, but they are not well-defined for star-shaped polygons and for arbitrary concave polygons.

Later, new generalizations of barycentric coordinates are published, such as discrete harmonic [4] and mean value coordinates [5]. The discrete harmonic coordinates have the same requirements as Wachspress coordinates. They are well defined for convex polygons and they are based on the minimization of an energy function, but unlike Wachspress or mean value functions these coordinates are not necessarily positive over the interior of any convex polygon. The mean value coordinates are probably the most popular type of generalized barycentric coordinates. They are well defined everywhere in the plane for any simple, star-shaped or arbitrary polygon. If the polygon is not star-shaped or convex, these coordinates are not necessarily positive, but in case of complex geometric shapes they are very robust. Owing to the above-mentioned properties and advantages, the generalized barycentric coordinates are commonly used in computer graphics and image processing for parameterization of meshes [9, 10, 12], mesh deformation [7, 6, 8], transfinite interpolation [18], image warping [17], cloning [13] or symmetrization [11].

Floater et al. [14] had already provided an overall picture of barycentric coordinates and visualized these coordinates using the contour lines of the coordinate functions. These contour lines mean those points of the polygon where one of the barycentric coordinates is constant. In this paper, we provide a much detailed comparison of the three different methods, we examine the Wachspress, the harmonic and the mean value coordinates and compare these functions for convex, n-sided polygons in an exact way. Moreover, to overcome the drawbacks of each method, we investigate an affine combination of these generalized barycentric coordinates.

In the next section, we give a short overview of some important definitions and properties of these barycentric coordinate methods. In Section 3, we discuss how the different coordinate functions can be comparable. Then, in Subsection 3.3, we present our results and we highlight the advantages and disadvantages of the distinct barycentric coordinates. The affine combination of the methods is described in Section 4.

2. Barycentric coordinates on n-sided polygons

Definition 2.1. Let P be a convex polygon in the plane, with vertices $v_1, v_2, \ldots, v_n$ and $n \geq 3$. We call any functions $b_i : P \to \mathbb{R}, i = 1 \ldots n$, barycentric coordinates, if they satisfy the following properties for all $v \in P$:

$$b_i(v) \geq 0, \quad i = 1, \ldots, n,$$
$$\sum_{i=1}^{n} b_i(v) = 1,$$
$$\sum_{i=1}^{n} b_i(v) v_i = v.$$

2.1. Wachspress coordinates

Wachspress coordinates are the simplest and earliest generalized barycentric coordinate functions and they have some important properties, for example, they are affine invariant, smooth (C^∞) and can be comuted by rational polynomials. These coordinate functions were published by Wachspress [2] and Warren [15], then Meyer et al. [3] simplified the formula and defined the coordinates in the following way:

$$b_i(v) = \frac{w_i(v)}{\sum_{j=1}^{n} w_j(v)},$$

with

$$w_i(v) = \frac{C_i(v)}{A_{i-1}(v) A_i(v)},$$

where $C_i(v), A_{i-1}(v), A_i(v)$ are areas of triangles shown in Figure 1.

2.2. Discrete harmonic coordinates

Discrete harmonic coordinates were first appeared in the work of Pinkall et al. [16], and the method is based on the minimization of discrete Dirichlet energy. It is an interesting fact, that the discrete harmonic coordinates and the Wachspress coordinates are identical in case of cyclic polygons, i.e. if the vertices of the polygon lie on a circle [14]. The discrete harmonic coordinates on convex polygons are defined by the following equation (following the notation of Figure 1):

$$b_i(v) = \frac{w_i(v)}{\sum_{j=1}^{n} w_j(v)},$$

where

$$w_i(v) = \frac{r_{i+1}^2(v) A_{i-1}(v) - r_i^2(v) B_i(v) + r_{i-1}^2(v) A_i(v)}{A_{i-1}(v) A_i(v)}$$

and

$$B_i(v) = r_{i-1}(v)r_{i+1}(v)sin(\alpha_{i-1}(v) + \alpha_i(v))/2$$

are signed triangle areas while

$$r_i(v) = ||v_i - v||$$

is the Euclidean distance of points v_i and v.

2.3. Mean value coordinates

Another popular type of barycentric coordinates is the mean value method [5] which can be generalized to non-convex polygons, unlike the previously mentioned coordinate functions. If the polygon P is convex, the mean value coordinates are defined by

$$b_i(v) = \frac{w_i(v)}{\sum_{j=1}^{n} w_j(v)},$$

where

$$w_i(v) = \frac{\tan(\alpha_i(v)/2) + \tan(\alpha_{i-1}(v)/2)}{||v_i - v||}$$

following again the notations of Figure 1.

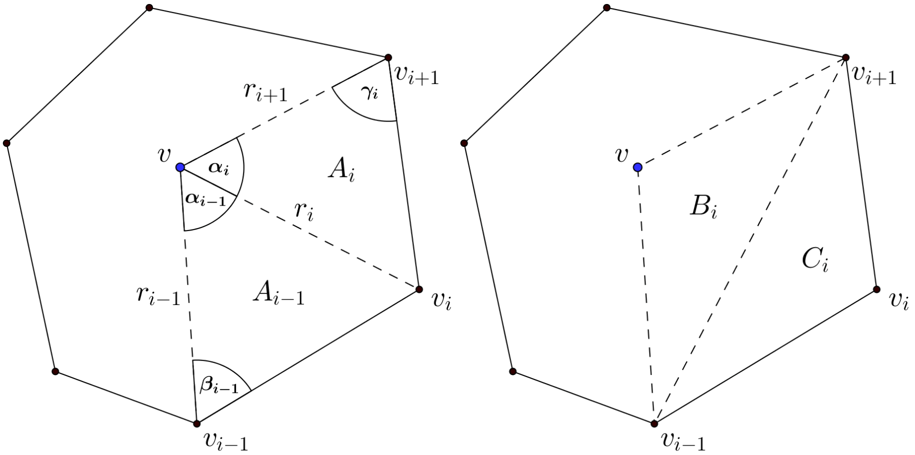


Figure 1: Notations for various barycentric coordinate functions

3. Comparison of different barycentric coordinate methods

As we have mentioned previously, in the work of Floater et al. [5], the authors had provided an overall picture of the so-called contour lines of the coordinate functions, but without any analysis. By contour lines we mean those points of the polygon P where one of the barycentric coordinates, say, the one assigned with vertex v_i, is constant. Evidently, the plot of these contour lines gives only a superficial image of the (dis)similarity of the functions. With all this in mind, our aim here is to examine the three different barycentric coordinates defined above and compare the behavior of these functions in similar conditions by contour lines and their patterns. For the comparison of the barycentric coordinate functions, first we use the curvature plot of the contour lines, thus we get more precise results than the aforementioned work.

3.1. The extraction of contour lines

In order to compare the curvature functions of the contour lines of the different barycentric coordinate functions, we need to compute these contour lines. Since it cannot be expressed and computed in a closed form, we consider a coordinate value $b_i \in [0,1]$ assigned with vertex v_i of polygon P with fixed i, and find those interior points of P, where the chosen barycentric coordinate b_i is equal or within a predefined limit (in our case it is 0.01) to the given value with respect to the vertex v_i of P. This way we get a set of points (see Figure 2) in P. Now if we fit a curve to these points, we get the contour line of the chosen barycentric coordinate value b_i with respect to the vertex v_i.

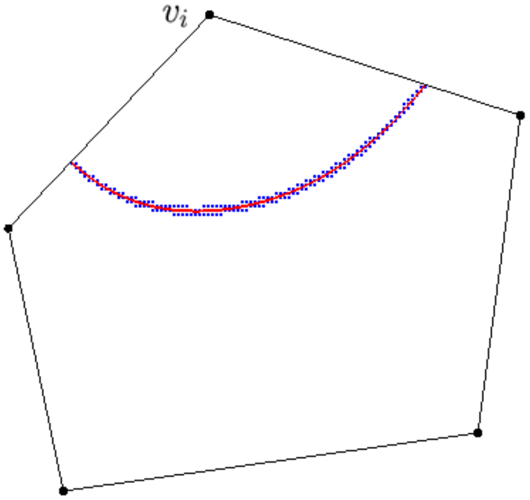


Figure 2: Blue pixels mark those points where barycentric coordinate b_i is in $[0.3, 0.32]$ with respect to the vertex v_i, while red line shows the fitted contour line curve

To fit the contour line curve to this set of points (see Figure 2), we use a polynomial fitting algorithm. At first, we compute the coefficients of the fitted polynomial $p(x)$ of degree n which best fits for the given point set (x_i, y_i). With these points, we can construct the following system of linear equations:

$$\begin{pmatrix} x_1^{n+1} & x_1^n & \cdots & 1 \\ x_2^{n+1} & x_2^n & \cdots & 1 \\ \vdots & \vdots & \ddots & \vdots \\ x_n^{n+1} & x_n^n & \cdots & 1 \end{pmatrix} \begin{pmatrix} p_1 \\ p_2 \\ \vdots \\ p_n \end{pmatrix} = \begin{pmatrix} y_1 \\ y_2 \\ \vdots \\ y_n \end{pmatrix},$$

where the matrix on the left is a *Vandermonde matrix.*

We have to solve this system to get the coefficients of $p(x)$. With the resulted coefficients, we can compute the contour line by the following explicit equation:

$$y = p_1 x^n + p_2 x^{n-1} + \cdots + p_n x + p_{n+1}.$$

3.2. The curvature function of the contour lines

As we have already stated, we use curvature functions for the comparison of the distinct coordinate functions, because it provides a precise image of the behavior of the contour line curves. After we have computed a contour line of the different barycentric coordinate functions as $f(x, y) = 0$ by simply converting the above equation to implicit form, the curvature of this curve at a regular point x_0 can be calculated easily by the well-known equation below:

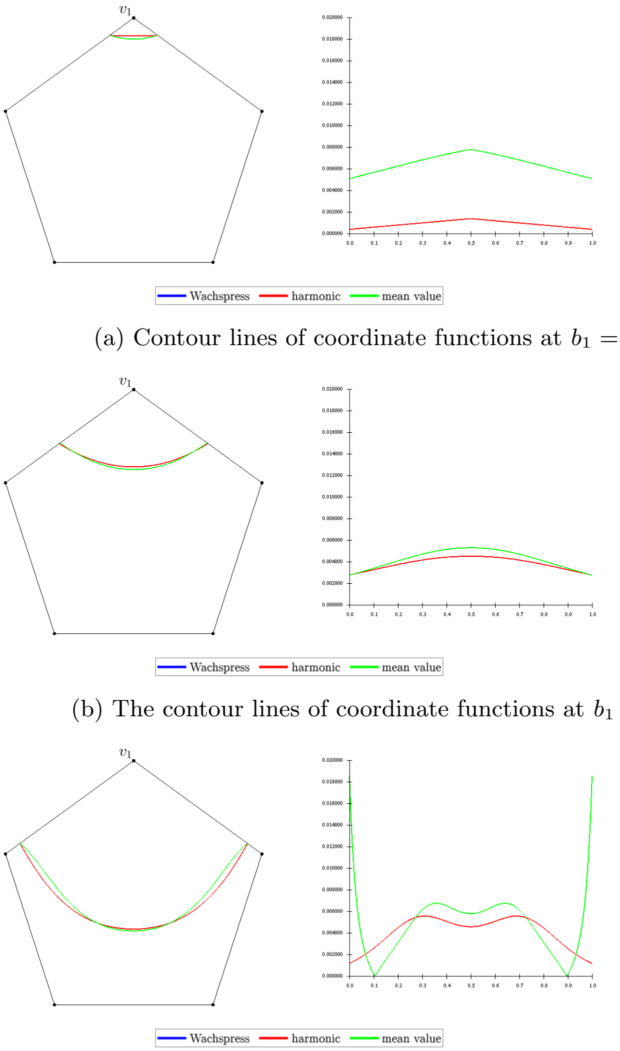
$$\kappa(x_o) = \frac{f''(x_0)}{(1 + f'(x_0)^2)^{3/2}}.$$

For each contour line, we calculate the curvature at every point of the curve, then we visualize these values on a curvature plot (see the right side of the figures below). The characteristic of the contour line curves can be assessed properly with these functions and they provide a good basis for the comparison, especially in terms of inflection points.

3.3. Curvature plot and inflection points

In this subsection, we present our results of the comparison and we also discuss the advantages and disadvantages of the studied methods. We compare the Wachspress, the discrete harmonic and the mean value coordinates for convex, n-sided polygons. In all cases, we compute the contour lines of each barycentric coordinate functions for various values b_i with respect to the vertex v_i, then visualize the curvature plot of the contour curves.

In the first example (see Figures 3a to 3c) we displayed contour lines for three different values of the barycentric coordinate assigned to the corresponding vertex v_1, that is for the value b_1. The Wachspress and the harmonic coordinate functions are identical in every figures, because regular polygons are all cyclic polygons (see Figure 5) [14]. We can also observe, that close to the corresponding vertex v_1, the three different coordinate functions almost behave like a circular arc (shown in



(a) Contour lines of coordinate functions at $b_1 = 0.8$

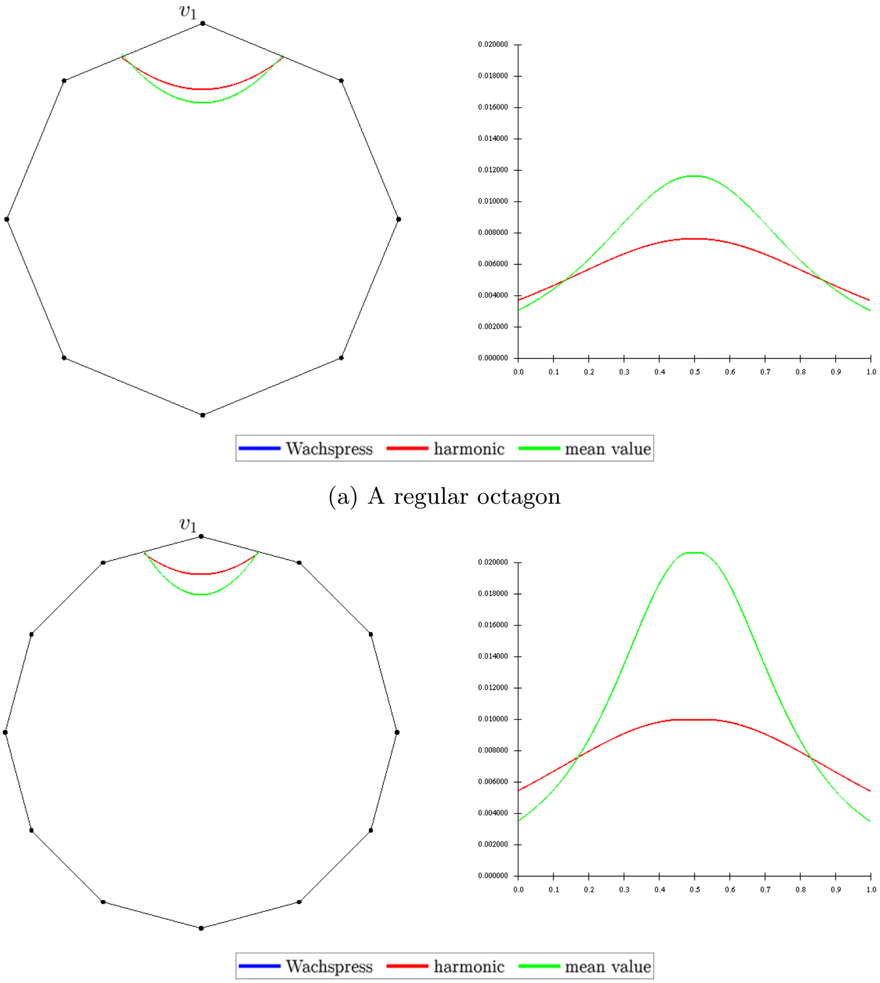
(b) The contour lines of coordinate functions at $b_1 = 0.4$

(c) The contour lines of coordinate functions at $b_1 = 0.1$

Figure 3: Left: Barycentric coordinate functions for regular polygons with respect to vertex v_1. Right: Curvature plots of the contour line curves

Figure 3b), with near constant curvature plot, but in lower regions the mean value coordinates produce unwanted inflection points (see Figure 3c).

As shown in Figure 4, the larger the angle of the regular polygon at the corresponding vertex v_1, the larger the difference between the mean value and the other two coinciding coordinate functions. Again, Wachspress and discrete harmonic coordinates produce contour lines much closer to circular arcs with constant curvature.

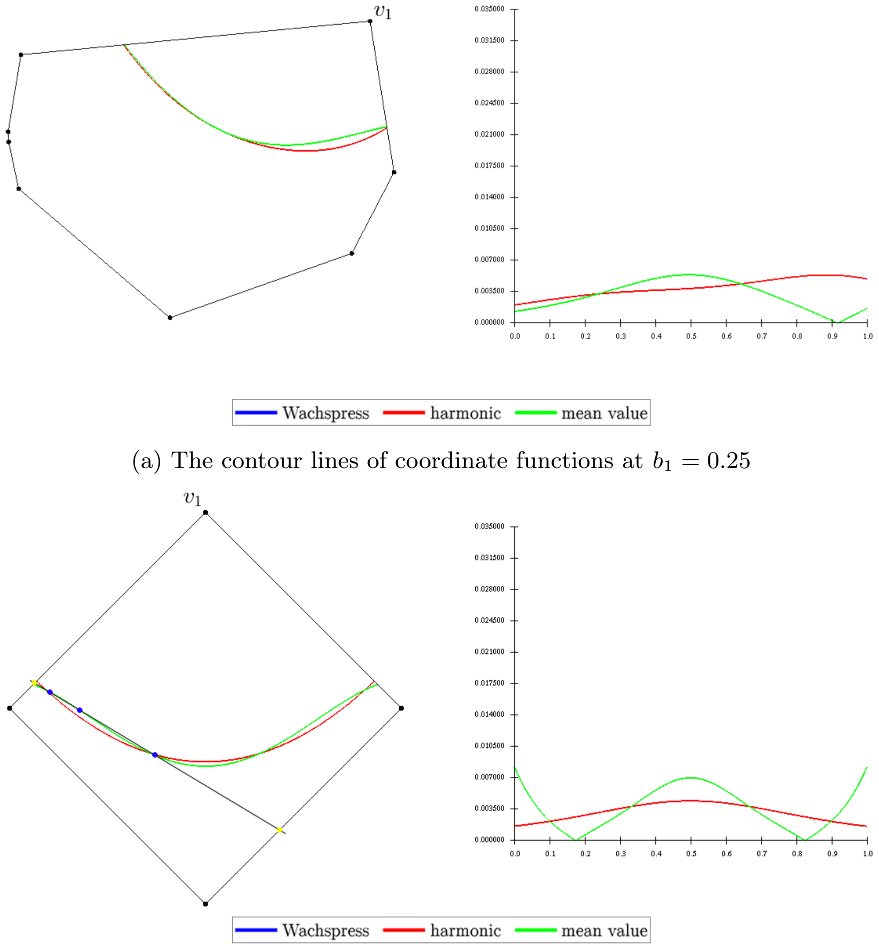


(a) A regular octagon

(b) A regular dodecagon

Figure 4: Left: Barycentric coordinate functions for regular polygons with respect to the vertex v_1. Right: Curvature plots of the contour line curves at $b_1 = 0.25$

As we have mentioned above, the Wachspress and the harmonic coordinate functions are identical in the case of cyclic polygons (shown in Figure 5). Moreover, it can be stated generally that the Wachspress and the harmonic coordinates do not generate inflection points for regular and cyclic polygons, while it can happen in case of mean value coordinate functions, as it can be seen e.g. in the curvature plots of Figure 5.



(a) The contour lines of coordinate functions at $b_1 = 0.25$

(b) The contour lines of coordinate functions at $b_1 = 0.15$. The blue dots mark the intersections of the line with the curve of the mean value function, while the yellow ones mark the intersections of the line with the polygon

Figure 5: Left: Barycentric coordinate functions for cyclic polygons with respect to the vertex v_1. Right: Curvature plots of the contour line curves

Therefore, we can say that the Wachspress and the harmonic coordinate functions satisfy the *variation diminishing* property in the case of cyclic polygons, while the mean value method does not fulfill this requirement. This property is originally required to be fulfilled by free-form curves: the number of intersections of a straight line with the curve is less than or equal to the number of intersections of the line with the control polygon. In our case the basis polygon P plays the role of the control polygon (see Figure 5b).

Now let us consider a polygon of irregular shape, where the different barycentric methods have significantly different types of contour lines. In the case of Wachspress coordinate functions, the contour line curve is closer to those vertices v_j, where the area of the triangle C_j (see Figure 1), which is specified by the given vertex v_j and its neighbors v_{j-1}, v_{j+1}, is larger. This behavior is clearly visible in Figure 6, where we display some non-cyclic polygons with vertices where the corresponding area of the triangle is much smaller than neighbouring triangle areas. Furthermore, it is worth examining the curvature plots of the Wachspress coordinate contour lines in these examples, because it shows the behavior of the curve perfectly. In Figure 6a, the area of the triangle C_4 at the vertex v_4 is evidently larger than the others, thus the curve is predominantly closer to this vertex and the curvature function is increasing on the interval $[0.5, 0.8]$ significantly. In the same way, we can observe in Figure 6b that the areas of the triangles C_3 and C_5 are the same, thus the curve is almost equally close to vertices v_3 and v_5, but because the area of the triangle C_4 is very small, the middle part of the curve flattens and it behaves like a straight line. The curvature plot also displays this behavior, because it is almost identical at interval $[0.1, 0.4]$ and $[0.6, 0.9]$, and the value of it at $x = 0.5$ is close to zero.

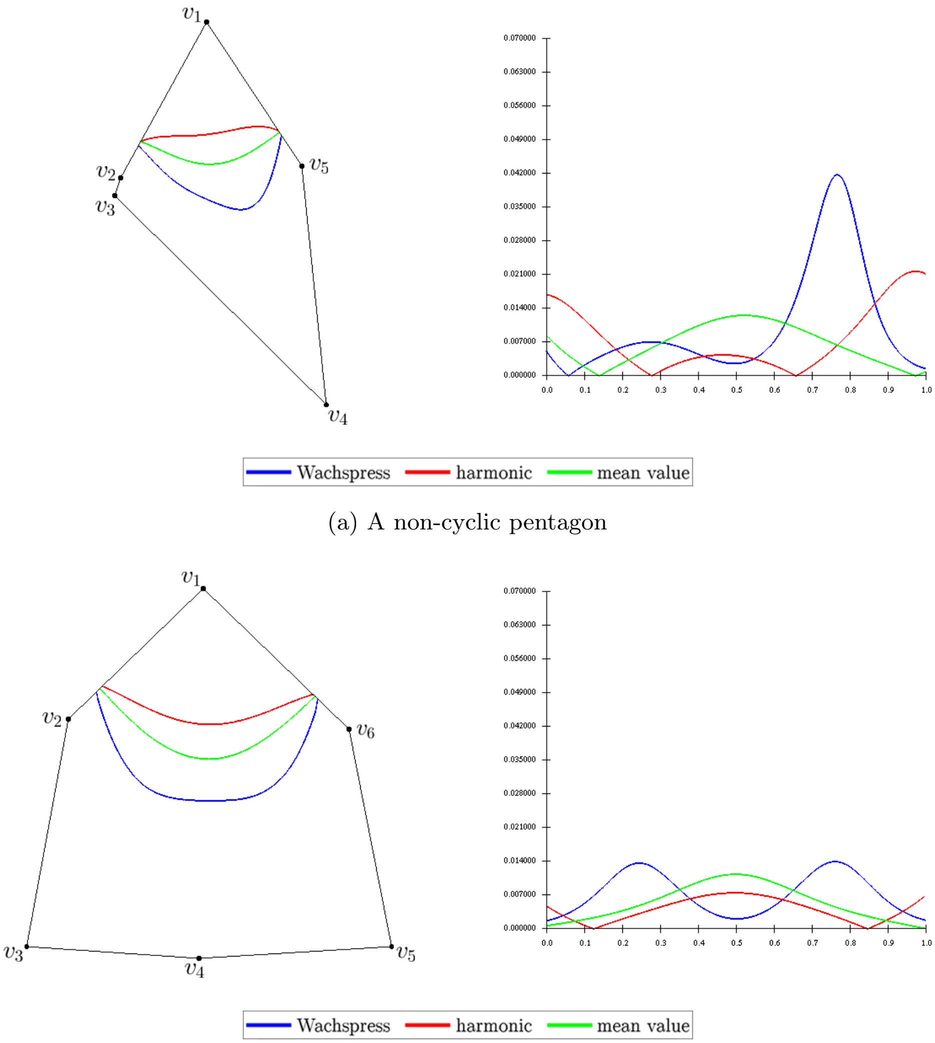
The contour line of the harmonic coordinate function is close to the corresponding vertex v_i, if the angles β_{i-1} and γ_i (see Figure 1) are obtuse. Therefore, in Figure 6a, it behaves contrary to the Wachspress coordinates, because the angles of the polygon at vertices v_2 and v_5 are obtuse. Furthermore, the harmonic coordinate method often produces inflection points on the contour line curve in these cases (shown in Figure 6a and 6b).

The mean value coordinate function is more robust than the other methods, as it is clearly visible in all examples of Figure 6.

3.4. Patterns of the barycentric coordinates

In the following examples, we displayed the contour line patterns of the three different barycentric coordinate functions with respect to the vertex v_i, from zero to one by step 0.05. These patterns clearly show the differences between the coordinate functions.

In the case of heavily non-cyclic polygons (see Figure 7a), the three methods produce significantly different patterns. The Wachspress coordinate function provides the most uniform shape of contour line pattern, which we may naively expect from this kind of shapes. We can also see, that there are significantly thicker stripes in the last intervals of the mean value coordinates than in the others, while the stripes of the harmonic coordinate function pattern behave unconventionally, because the upper ones approximate the corresponding vertex v_i. This behavior occurs, when the angles of the polygon at the neighbor vertices of the correspond-



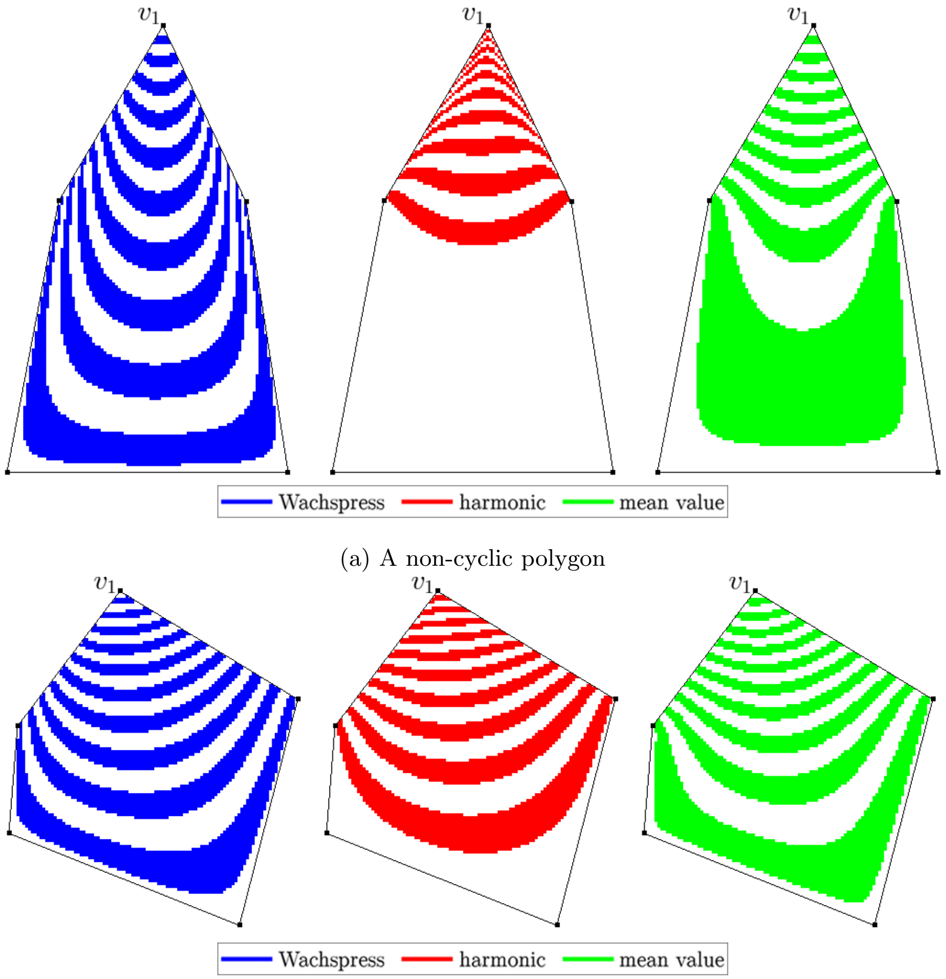
(a) A non-cyclic pentagon

(b) A non-cyclic hexagon

Figure 6: Left: Barycentric coordinate functions for non-cyclic polygons with respect to the vertex v_1. Right: Curvature plots of the contour line curves at $b_1 = 0.25$

ing vertex v_i are obtuse, because if the angles $\beta_{i-1} + \gamma_i > \pi$ (see Figure 1), there may be some interior vertices of the polygon the barycentric coordinate of which is negative with respect to v_i [14].

Furthermore, if we consider a nearly regular polygon (shown in Figure 7b), we can notice that the Wachspress coordinate method almost behaves like in the



(a) A non-cyclic polygon

(b) A nearly regular polygon

Figure 7: Iso-barycentric contour line patterns of the three different functions with respect to the vertex v_1 from zero to one by step 0.05

previous figure (see Figure 7a), while the other two work differently, providing more uniform shapes in this case.

4. Affine combination of barycentric functions

As we have seen in the previous sections, each generalized barycentric method has its advantages and drawbacks. For example, discrete harmonic coordinate functions are based on the minimization of the Dirichlet energy, but provide unusual and irregular shape of contour line patterns. In every application, one has to decide which method fits better to the problem. To overcome this restriction, in this section we introduce the affine combination of barycentric functions. Suppose a polygon P with n vertices is given and the Wachspress coordinates b_i^W and discrete harmonic coordinates b_i^H are computed, respectively. Now consider the affine combination b_i^A of these coordinates

$$b_i^A(\lambda) = (1-\lambda)b_i^W + \lambda b_i^H,$$

where $\lambda \in [0,1]$ is a free parameter. The new coordinates evidently satisfy the requirements formulated in Section 2 to be barycentric coordinates: $b_i^A(\lambda) \geq 0$ for each $i = 1, ..., n$, and the sum of these coordinates is as follows

$$\sum_{i=1}^{n} b_i^A(v) = \sum_{i=1}^{n} \big((1-\lambda)b_i^W(v) + \lambda b_i^H(v)\big) = (1-\lambda)\sum_{i=1}^{n} b_i^W(v) + \lambda \sum_{i=1}^{n} b_i^H(v) = 1.$$

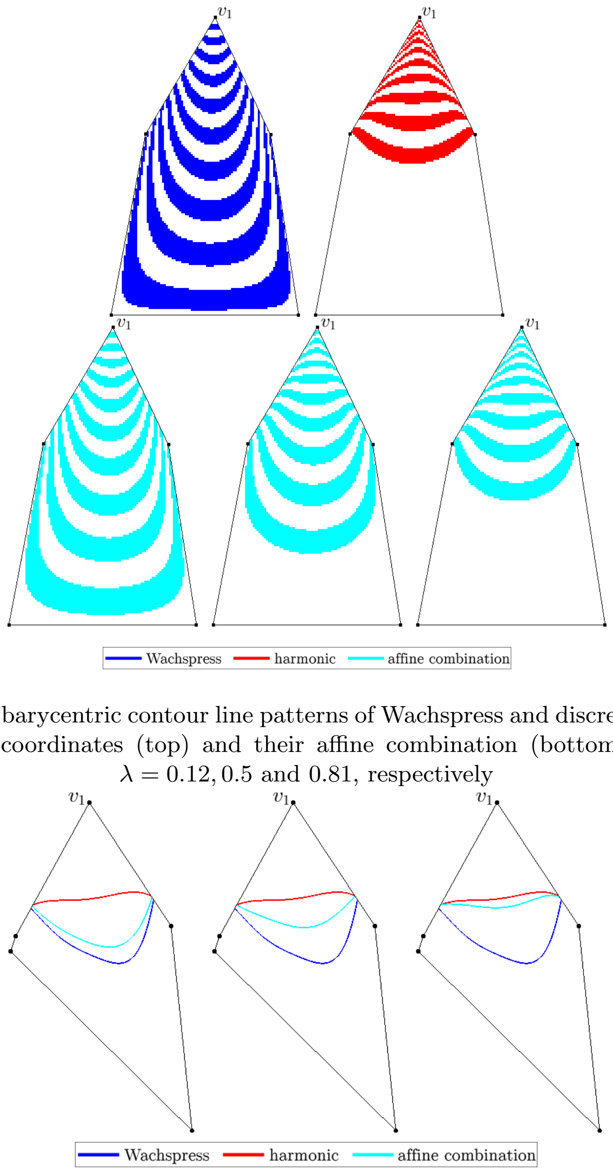
This affine combination is a tool to merge advantages of two barycentric methods. The choice of the free parameter λ gives us an extra flexibility in order to weight the two methods, which can yield different versions of affine combinations, various trade-offs between uniform patterns and energy minimization, as we can observe in Figure 8.

5. Conclusions

In this paper, we studied the different types of the generalized barycentric coordinates (Wachspress, discrete harmonic, mean value) and we compared these functions for convex, n-sided polygons in detail. For the comparison, we used the curvature functions of the contour line curves which are defined by a polynomial fitting algorithm. The curvature functions provided descriptions of the behavior of the contour line curves.

In the case of regular and cyclic polygons, the behavior of the Wachspress and the harmonic coordinate functions is equivalent, and they satisfy the property of *variation diminishing*, while the mean value coordinates may generate unnecessary inflection points. It is also shown that angle of the polygon at the corresponding vertex heavily affects the different barycentric coordinate methods. The larger the angle, the larger the divergence of the mean value from the other coordinate functions.

With regard to non-cyclic polygons, the three different coordinate methods provide significantly different patterns of contour lines. It can be said generally that the Wachspress coordinates follow the sides of the polygon better, than the others, and provide more uniform patterns, while the harmonic coordinate method behaves unconventionally in some cases, because it produces negative coordinates. Moreover, we can say that the mean value coordinate function is the most robust for non-cyclic polygons.



(a) Iso-barycentric contour line patterns of Wachspress and discrete harmonic coordinates (top) and their affine combination (bottom) with $\lambda = 0.12, 0.5$ and 0.81, respectively

(b) Contour line curves of Wachspress and harmonic coordinates and their affine combination with $\lambda = 0.22, 0.5$ and 0.83, respectively

Figure 8: The affine combination of Wachspress and discrete harmonic coordinate functions with respect to the vertex v_1

In order to overcome the shortcomings of these methods, we introduced the affine combination of two barycentric coordinate functions which method also gives us an extra flexibility by the free parameter λ. This way one can provide a trade-off between various advantageous properties and drawbacks of the methods.

References

[1] Möbius, A. F., Der barycentrische calcul, Leipzig, 1827.

[2] Wachspress, E. L., A rational finite element basis, Elsevier, 1975.

[3] Meyer, M., Barr, A., Lee, H., Desbrun, M., Generalized barycentric coordinates on irregular polygons, *Journal of Graphics Tools*, Vol. 7 (2002), 13–22.

[4] Eck, M., DeRose, T., Duchamp, T., Hoppe, H., Lounsbery, M., Stuetzle, W., Multiresolution analysis of arbitrary meshes, *Proceedings of SIGGRAPH '95*, (1995), 173–182.

[5] Floater, M. S., Mean value coordinates, *Computer Aided Geometric Design*, Vol. 20 (2003), 19–27.

[6] Ju, T., Schaefer, S., Warren, J. , Mean value coordinates for closed triangular meshes, *ACM Transactions on Graphics (TOG)*, Vol. 24 (2005), 561–566.

[7] Joshi, P., Meyer, M., DeRose, T., Green, B., Sanocki, T., Harmonic coordinates for character articulation, *ACM Transactions on Graphics (TOG)*, Vol. 26 (2007).

[8] Lipman, Y., Levin, D., Cohen-Or, D., Green coordinates, *ACM Transactions on Graphics (TOG)*, Vol. 27 (2008), 78:1–78:10.

[9] Liu, C., Luo, Z., Shi, X., Liu, F., Luo, X., A fast mesh parameterization algorithm based on 4-point interpolatory subdivision, *Applied Mathematics and Computation*, Vol. 219 (2013), 5339–5344.

[10] Morigi, S., Feature-sensitive parameterization of polygonal meshes, *Applied Mathematics and Computation*, Vol. 215 (2009), 1561–1572.

[11] Guessab, A., Guessab, F., Symmetrization, convexity and applications, *Applied Mathematics and Computation*, Vol. 240 (2014), 149–160.

[12] Weber, O., Ben-Chen, M., Gotsman, C., Hormann, K., A complex view of barycentric mappings, *Computer Graphics Forum*, Vol. 30 (2011), 1533–1542.

[13] Farbman, Z., Hoffer, G., Lipman, Y., Cohen-Or, D., Lischinski, D., Coordinates for instant image cloning, *ACM Transactions on Graphics (TOG)*, Vol. 28 (2009), 67:1–67:9.

[14] Floater, M. S., Hormann, K., Kós, G., A general construction of barycentric coordinates over convex polygons, *Advances in Computational Mathematics*, Vol. 24 (2006), 311–331.

[15] Warren, J., Barycentric coordinates for convex polytopes, *Advances in Computational Mathematics*, Vol. 6 (1996), 97–108.

[16] Pinkall, U., Polthier, K., Computing discrete minimal surfaces and their conjugates, *Experimental Mathematics*, Vol. 2 (1993), 15–36.

[17] Hormann, K., Floater, M. S., Mean value coordinates for arbitrary planar polygons, *ACM Transactions on Graphics (TOG)*, Vol. 25 (2006), 1424–1441.

[18] Dyken, C., Floater, M. S., Transfinite mean value interpolation, *Computer Aided Geometric Design*, Vol. 26 (2009), 117–134.

Annales Mathematicae et Informaticae
47 (2017) pp. 201–209
http://ami.uni-eszterhazy.hu

Caustics of spline curves[*]

Ede Troll, Miklós Hoffmann

Institute of Mathematics and Computer Science
Eszterházy Károly University, Eger, Hungary
[troll.ede,hoffmann.miklos]@uni-eszterhazy.hu

Submitted September 26, 2017 — Accepted December 18, 2017

Abstract

If we consider a curve as a mirror, then parallel light rays reflected by this mirror curve form a family of lines, which generally has an envelope. This envelope is called caustic curve of the given mirror curve. The aim of this paper is to describe the caustic curve of a free-form curve and study its alteration by the modification of a control point. We will provide results of general form, and case studies in terms of Bézier and B-spline curves.

Keywords: caustic curve, Bézier curve, B-spline curve

1. Introduction

The classical definition of caustic curves can be found in several books (e.g. in [1, 2]), which can be adapted for parameterical curves as follows.

Definition 1.1. Given a direction vector $\mathbf{v}(v_x, v_y)$ of light rays and the mirror curve $\mathbf{r}(x(t), y(t)), t \in [0,1]$, the envelope of rays (if exists) reflected by the curve is called caustic curve (or simply caustic) of the given curve.

More precisely the curve defined above is called *catacaustic* curve, making distinction between reflected and refracted rays – in this latter case the envelope curve is called *diacaustic* curve. Throughout the paper we will consider only reflected rays, therefore the curve will simply be called caustic.

Although the study of caustic curves has a long history in mathematics and optics, the shape of the mirror is generally a classical curve, such as line, circle

[*]This research was supported by the grant EFOP-3.6.1-16-2016-00001 (Complex improvement of research capacities and services at Eszterházy Károly University).

or parabola. In the case of reflection by a circle, the caustic is an epicycloid, the locus of a fixed point on the circumference of a circle that rolls on the exterior of another circle, whilst in the case of reflection by the parabola, the caustic is a cubic algebraic curve called Tschirnhausen's cubic (see e.g. [3]).

But there are limited results in geometric modeling considering caustic curves of other curves, especially caustic curve of free-form curves. While it is relatively easy to visualize the family of rays for any type of mirror curve, and the human perception can 'realize' the envelope if the family is dense enough, which can also been approximated numerically providing predefined shapes ([4] and references therein), the exact computation of the caustic curve yielding closed form of these curves is a challenging problem. Exact computation of the evolute of Bézier curves using computer algebra software can be found in [5]. The aim of this paper is to provide exact mathematical description of the caustic curve of free-form mirror curves with arbitrary basis functions in a closed form.

2. The caustic of a free-form curve

In this section we will discuss the method of calculating the caustic of a control point based free-form curve. Let the curve be given in the form

$$\mathbf{r}(t) = \sum_{i=0}^{n} A_i(t)\mathbf{p}_i,$$

where $t \in [0,1]$, $A(t)_i$ are the basis functions and $\mathbf{p}_i$ are the control points.

When we emit parallel light rays onto the curve with the direction vector $\mathbf{v}(v_x, v_y)$ then these rays will form the family of straight lines

$$\mathbf{l}(u,t) = \mathbf{r}(t) + u\mathbf{v},$$

where $u \in \mathbb{R}$ is the running parameter of the lines, while t is the family parameter, coinciding the running parameter of the mirror curve.

To obtain the reflected rays, we have to mirror these incoming rays onto the corresponding tangent lines

$$\mathbf{t}(w,t) = \mathbf{r}(t) + w\mathbf{r}'(t),$$

where $w \in \mathbb{R}$.

To define the reflected rays we translated the points of the curve at every t with the vector $\mathbf{v}$ and mirrored them onto the tangent lines. The mirrored points are $\mathbf{m}_r(t)$ $(x_m(t), y_m(t))$, where

$$\begin{aligned} x_m(t) &= x(t) - 2\langle \mathbf{v}, \mathbf{n}_t(t)\rangle x_n(t), \\ y_m(t) &= y(t) - 2\langle \mathbf{v}, \mathbf{n}_t(t)\rangle y_n(t), \end{aligned}$$

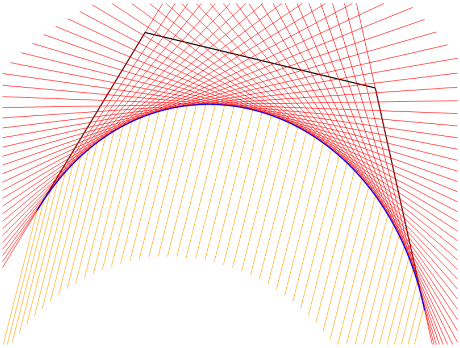


Figure 1: The family of incoming light rays and the corresponding tangents

where $\mathbf{n}_t(t)\,(x_n(t), y_n(t))$ are the unit normal vectors of the tangents and

$$x_n(t) = \frac{y'(t)}{\sqrt{x'^2(t) + y'^2(t)}},$$
$$y_n(t) = \frac{-x'(t)}{\sqrt{x'^2(t) + y'^2(t)}}.$$

Now the direction vector of the reflected rays are $\mathbf{v}_r(t) = \mathbf{m}_r(t) - \mathbf{r}(t)$ and the family of the reflected rays can be described as

$$\mathbf{l}_r(u,t) = \mathbf{r}(t) + u\mathbf{v}_r(t),$$

where $u \in \mathbb{R}$ and the coordinate functions are

$$x_l(u,t) = x(t) + ux_{v_\mathbf{r}}(t)$$
$$y_l(u,t) = y(t) + uy_{v_\mathbf{r}}(t).$$

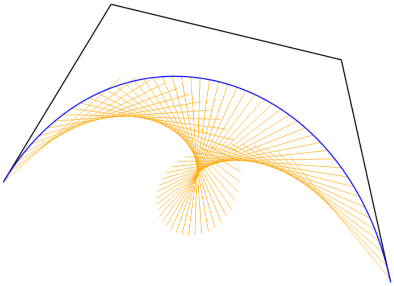


Figure 2: The family of reflected light rays

To compute the envelope, that is the caustic curve, we have to solve the following equation for u

$$\begin{vmatrix} \frac{\partial x_l}{\partial t} & \frac{\partial x_l}{\partial u} \\ \frac{\partial y_l}{\partial t} & \frac{\partial y_l}{\partial u} \end{vmatrix} = 0,$$

where

$$\begin{aligned}
\frac{\partial x_l}{\partial t} &= 2u \left(\frac{2\left(v_x {y'}^2(t) - v_y x'(t)y'(t)\right)\left(x'(t)x''(t) + y'(t)y''(t)\right)}{\left({x'}^2(t) + {y'}^2(t)\right)^2} \right. \\
&\qquad \left. - \frac{2v_x y'(t)y''(t) - v_y\left(x'(t)y''(t) + y'(t)x''(t)\right)}{{x'}^2(t) + {y'}^2(t)} \right) + x'(t), \\
\frac{\partial x_l}{\partial u} &= v_x - \frac{2y'(t)\left(v_x y'(t) - v_y x'(t)\right)}{{x'}^2(t) + {y'}^2(t)}, \\
\frac{\partial y_l}{\partial t} &= 2u \left(\frac{2\left(v_y {x'}^2(t) - v_x x'(t)y'(t)\right)\left(x'(t)x''(t) + y'(t)y''(t)\right)}{\left({x'}^2(t) + {y'}^2(t)\right)^2} \right. \\
&\qquad \left. - \frac{2v_y x'(t)x''(t) - v_x\left(x'(t)y''(t) + y'(t)x''(t)\right)}{{x'}^2(t) + {y'}^2(t)} \right) + y'(t), \\
\frac{\partial y_l}{\partial u} &= v_y - \frac{2x'(t)\left(v_y x'(t) - v_x y'(t)\right)}{{x'}^2(t) + {y'}^2(t)}.
\end{aligned}$$

Substituting these forms into the original equation

$$\frac{\partial x_l}{\partial u}\frac{\partial y_l}{\partial t} - \frac{\partial y_l}{\partial u}\frac{\partial x_l}{\partial t} = 0$$

after some calculation the equation can be simplified in the following form

$$\begin{aligned}
&\frac{2u\left(v_x^2\left(x'(t)y''(t) - y'(t)x''(t)\right) + v_y^2\left(x'(t)y''(t) - y'(t)x''(t)\right)\right)}{{x'}^2(t) + {y'}^2(t)} \\
&\qquad + \frac{v_y\left({x'}^3(t) + x'(t){y'}^2(t)\right) - v_x\left({y'}^3(t) + y'(t){x'}^2(t)\right)}{{x'}^2(t) + {y'}^2(t)} = 0
\end{aligned}$$

and the determinant equation gives us $u(t)$ what is the parameter u of the rays expressed by the family parameter t as follows

$$u(t) = -\frac{1}{2}\frac{v_y\left({x'}^3(t) + x'(t){y'}^2(t)\right) - v_x\left({y'}^3(t) + {x'}^2(t)y'(t)\right)}{\left(v_x^2 + v_y^2\right)\left(x'(t)y''(t) - x''(t)y'(t)\right)}.$$

Now we can express the envelope of the reflected rays $\mathbf{c}(t)\left(x_c(t), y_c(t)\right)$ as the function of the parameter t of the original free-formed curve where

$$x_c(t) = x_l\left(u(t), t\right)$$

$$= -\frac{1}{2}\frac{\left(v_y x'(t) - v_x y'(t)\right)\left(v_x {x'}^2(t) + 2v_y x'(t)y'(t) - v_x {y'}^2(t)\right)}{\left(v_x^2 + v_y^2\right)\left(x'(t)y''(t) - x''(t)y'(t)\right)} + x(t),$$

$$y_c(t) = y_l\left(u(t), t\right)$$

$$= \frac{1}{2}\frac{\left(v_y x'(t) - v_x y'(t)\right)\left(v_y {x'}^2(t) - 2v_x x'(t)y'(t) - v_y {y'}^2(t)\right)}{\left(v_x^2 + v_y^2\right)\left(x'(t)y''(t) - x''(t)y'(t)\right)} + y(t).$$

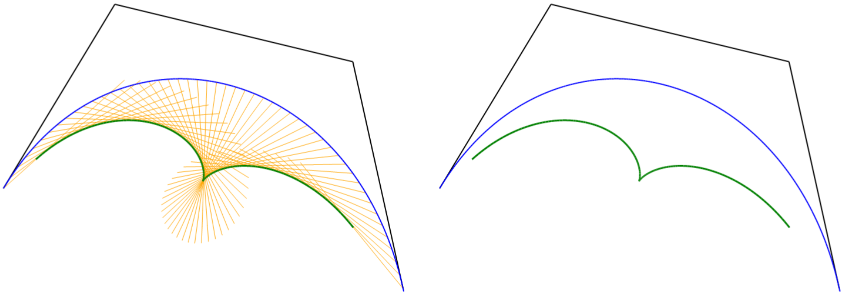


Figure 3: The caustic curve of a control point based free-formed curve

In these computations we extensively used computer algebra softwares. As we can observe from the equation above, the caustic curve of a control point based free formed curve is rational, and its degree depends on the degree of the original basis functions.

Example 2.1. The construction of the caustic curve described in the preceding section is general, that is valid for every control point based free-form curve, and it is independent of the degree and type of the basis functions. Here we provide some case studies of the results for the most popular free-form curves.

In Figure 4 a Bézier curve of degree 4 can be seen in the same position but with different incoming direction of light rays.

Since the caustic has the same parameter interval $t \in [0, 1]$ as the original curve, the piecewisely computed caustics of connect B-Spline arcs will also be smoothly connected automatically. In Figure 5 a uniform cubic B-Spline and its piecewisely connected caustic curve can be seen. To show the arcs of the caustics we drew them with different colours.

The last example shows a closed uniform cubic B-Spline and its caustics with various light ray directions, where the control points are vertices of a square (Fig 6).

In the leftmost image the incoming rays are parallel to the y axis, while in the right image the rays are parallel to the x axis. These caustics look very similar to the nephroid what is the caustic of the circle. In the middle image the direction vector of the incoming rays does not parallel to any of the axes and because the cubic B-Spline can not describe an exact circle, the caustic is slightly distorted.

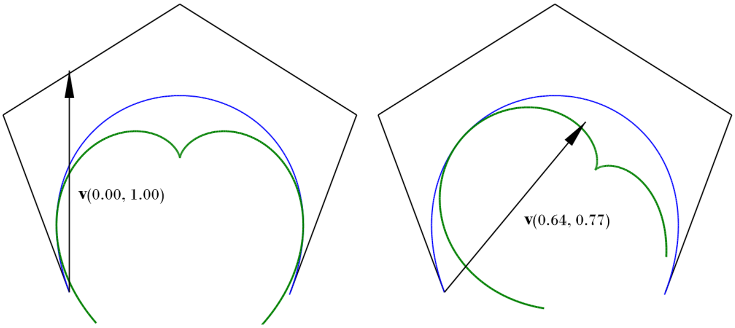


Figure 4: The caustic of a Bézier curve of degree 4 with different light directions

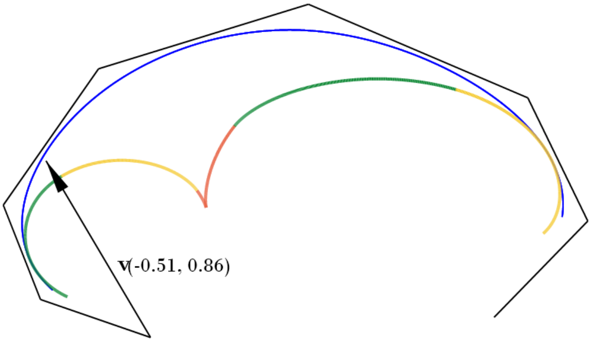


Figure 5: The caustic curve of the uniform cubic B-spline curve (different colours means different arcs of the caustic)

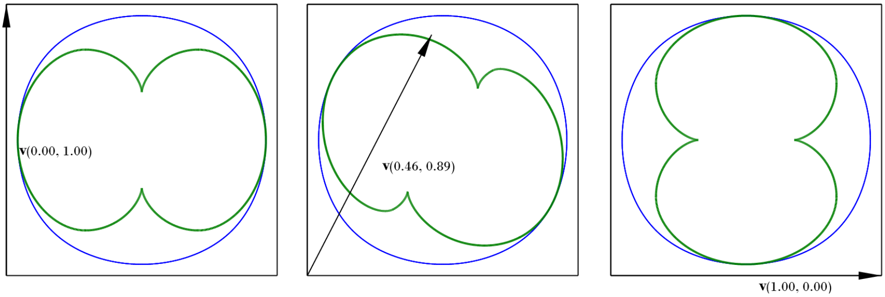


Figure 6: The caustic of the closed uniform cubic B-Spline curve with different light ray directions

2.1. Geometric effect of the control points

In this section, our aim is to describe the alteration of the caustic curve by the modification of a control point. This means that we would like to compute the path of a point $\mathbf{c}(t_0), t_0 \in [0,1]$ of the caustic curve when a control point $\mathbf{p}_i, i \in 0, 1, \ldots, n$ moves along a straight line. To describe the effect of the modification we have to express the caustic curve in a way that the running parameter t is considered to be constant ($t = t_0$), and the control point $\mathbf{p}_i(x_{\mathbf{p}_i}, y_{\mathbf{p}_i})$ is the parameter (more precisely, the coordinates of the control point are the parameters) of the path. To emphasize this fact the path of the point $\mathbf{c}(t_0)$ is denoted by $\mathbf{s}(t_0, \mathbf{p}_i)$. Let us express the coordinate functions of the free-form curve and their derivatives as a constant part plus the part depending on the coordinates of the moving control points:

$$\begin{aligned}
x(t_0, x_{\mathbf{p}_i}) &= \sum_{j \neq i} A_j(t_0) x_{\mathbf{p}_j} + A_i(t_0) x_{\mathbf{p}_i}, & y(t_0, y_{\mathbf{p}_i}) &= \sum_{j \neq i} A_j(t_0) y_{\mathbf{p}_j} + A_i(t_0) y_{\mathbf{p}_i}, \\
x'(t_0, x_{\mathbf{p}_i}) &= \sum_{j \neq i} A'_j(t_0) x_{\mathbf{p}_j} + A'_i(t_0) x_{\mathbf{p}_i}, & y'(t_0, y_{\mathbf{p}_i}) &= \sum_{j \neq i} A'_j(t_0) y_{\mathbf{p}_j} + A'_i(t_0) y_{\mathbf{p}_i}, \\
x''(t_0, x_{\mathbf{p}_i}) &= \sum_{j \neq i} A''_j(t_0) x_{\mathbf{p}_j} + A''_i(t_0) x_{\mathbf{p}_i}, & y''(t_0, y_{\mathbf{p}_i}) &= \sum_{j \neq i} A''_j(t_0) y_{\mathbf{p}_j} + A''_i(t_0) y_{\mathbf{p}_i},
\end{aligned}$$

where j runs from 0 to n.

Using the expressions above we can expand the coordinate functions of the caustics in the form

$$\begin{aligned}
x_c(t_0, x_{\mathbf{p}_i}) = {} & -\frac{1}{2} \frac{(v_y x'(t_0, x_{\mathbf{p}_i}) - v_x y'(t_0, x_{\mathbf{p}_i}))}{\left(v_x^2 + v_y^2\right)\left(x'(t_0, x_{\mathbf{p}_i}) y''(t_0, x_{\mathbf{p}_i}) - x''(t_0, x_{\mathbf{p}_i}) y'(t_0, x_{\mathbf{p}_i})\right)} \\
& \times \frac{\left(v_x x'^2(t_0, x_{\mathbf{p}_i}) + 2 v_y x'(t_0, x_{\mathbf{p}_i}) y'(t_0, x_{\mathbf{p}_i}) - v_x y'^2(t_0, x_{\mathbf{p}_i})\right)}{\left(v_x^2 + v_y^2\right)\left(x'(t_0, x_{\mathbf{p}_i}) y''(t_0, x_{\mathbf{p}_i}) - x''(t_0, x_{\mathbf{p}_i}) y'(t_0, x_{\mathbf{p}_i})\right)} \\
& + x(t_0, x_{\mathbf{p}_i}), \\
y_c(t_0, y_{\mathbf{p}_i}) = {} & \frac{1}{2} \frac{(v_y x'(t_0, y_{\mathbf{p}_i}) - v_x y'(t_0, y_{\mathbf{p}_i}))}{\left(v_x^2 + v_y^2\right)\left(x'(t_0, y_{\mathbf{p}_i}) y''(t_0, y_{\mathbf{p}_i}) - x''(t_0, y_{\mathbf{p}_i}) y'(t_0, y_{\mathbf{p}_i})\right)} \\
& \times \frac{\left(v_y x'^2(t_0, y_{\mathbf{p}_i}) - 2 v_x x'(t_0, y_{\mathbf{p}_i}) y'(t_0, y_{\mathbf{p}_i}) - v_y y'^2(t_0, y_{\mathbf{p}_i})\right)}{\left(v_x^2 + v_y^2\right)\left(x'(t_0, y_{\mathbf{p}_i}) y''(t_0, y_{\mathbf{p}_i}) - x''(t_0, y_{\mathbf{p}_i}) y'(t_0, y_{\mathbf{p}_i})\right)} \\
& + y(t_0, y_{\mathbf{p}_i}).
\end{aligned}$$

By these expressions we can observe that the paths $\mathbf{s}(t_0, \mathbf{p}_i)$ are rational curves (see Fig. 7) and their coordinate functions s_x and s_y can be expressed as

$$\begin{aligned}
s_x(t_0, x_{\mathbf{p}_i}) &= \frac{a_1 x_{\mathbf{p}_i}^3 + a_2 x_{\mathbf{p}_i}^2 + a_3}{a_4 x_{\mathbf{p}_i} + a_5}, \\
s_y(t_0, y_{\mathbf{p}_i}) &= \frac{b_1 y_{\mathbf{p}_i}^3 + b_2 y_{\mathbf{p}_i}^2 + b_3}{b_4 y_{\mathbf{p}_i} + b_5},
\end{aligned}$$

where a_k and b_k, ($k \in \{1, 2, \ldots, 5\}$) are constants depending on the direction vector of the incoming rays, the fixed value of the running parameter t_0 and the position of the other control points.

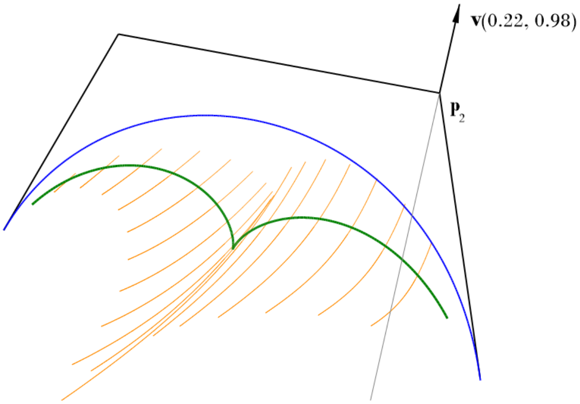


Figure 7: The geometric effect of the alteration of the control point $\mathbf{p}_2$ on the caustic curve and paths of points of a cubic Bézier curve when the control point moves along the straight line defined by the direction vector $\mathbf{d}\,(0.22, 0.98)$

3. Conclusion and future research

Caustic curves of control point based free-form curves have been computed and provided in closed form in this paper. The theoretical results hold for any free-form curve, independently of its specific basis functions. Beside these results, some examples of the most popular free-form curves are also considered. As an important aspect of the design of free-form mirrors and their caustics we have seen, that the alteration of a control point will evidently affect the shape of the caustic, and fixed points of the caustic curve move along rational curves as one of the control points is altered along a straight line.

These results may form the basis for the spatial extension of the computation of caustics of free-form surfaces, which is an important aspect of geometric modeling and free-form architecture as well. It can be the direction of our future research.

References

[1] Lawrence, J. D., A Catalog of Special Plane Curves, New York: Dover, 1972.

[2] Lockwood, E. H., A Book of Curves, Cambridge University Press, 1967, pp. 182–185.

[3] Farouki, R. T., Pythagorean-Hodograph Curves: Algebra and Geometry Inseparable, Springer, 2008.

[4] Schwartzburg, Y., Testuz, R., Tagliasacchi, A., Pauly, M., High-contrast computational caustic design, *ACM Transactions on Graphics (TOG)*, 33(4) (2014), 1–11.

[5] Tsukada, T., Caustics on Spline Curves, Wolfram Demonstration Project `http://demonstrations.wolfram.com/CausticsOnSplineCurves/` (visited on: December 13, 2017).

[6] Fowler, B., Bartels, R., Constraint-based curve manipulation, *IEEE Comp. Graph. and Appl.*, Vol. 13 (1993), 43–49.

Annales Mathematicae et Informaticae
47 (2017) pp. 211–218
http://ami.uni-eszterhazy.hu

Infinitary superperfect numbers

Tomohiro Yamada

Center for Japanese language and culture, Osaka University,
562-8558, 8-1-1, Aomatanihigashi, Minoo, Osaka, Japan
tyamada1093@gmail.com

Submitted June 13, 2017 — Accepted June 30, 2017

Abstract

We shall show that 9 is the only odd infinitary superperfect number.

Keywords: Odd perfect numbers, infinitary superperfect numbers, unitary divisors, infinitary divisors, the sum of divisors

MSC: 11A05, 11A25

1. Introduction

As usual, $\sigma(N)$ denotes the sum of divisors of a positive integer N. N is called perfect if $\sigma(N) = 2N$. It is a well-known unsolved problem to decide whether or not an odd perfect number exists. Interest to this problem has produced many analogous notions and problems concerning divisors of an integer. For example, Suryanarayana [15] called N to be superperfect if $\sigma(\sigma(N)) = 2N$. It is asked in this paper and still unsolved whether there are any odd superperfect numbers.

Some special classes of divisors have also been studied in several papers. One of them is the class of unitary divisors defined by Eckford Cohen [2]. A divisor d of n is called a unitary divisor if $\gcd(d, n/d) = 1$. Wall [16] introduced the notion of biunitary divisors. Letting $\gcd_1(a, b)$ denote the greatest common unitary divisor of a and b, a divisor d of a positive integer n is called a biunitary divisor if $\gcd_1(d, n/d) = 1$.

Graeme L. Cohen [3] generalized these notions and introduced the notion of k-ary divisors for any nonnegative integer k recursively. Any divisor of a positive integer n is called a 0-ary divisor of n and, for each nonnegative integer k, a divisor d of a positive integer n is called a $(k+1)$-ary divisor if d and n/d does not have a common k-ary divisor other than 1. Clearly, a 1-ary divisor is a unitary divisor and a 2-ary divisor is a biunitary divisor. We note that a positive integer $d = \prod_i p_i^{f_i}$ with p_i distinct primes and $f_i \geq 0$ is a k-ary divisor of $n = \prod_i p_i^{e_i}$ if and only if $p_i^{f_i}$ is a k-ary divisor of $p_i^{e_i}$ for each i. G. L. Cohen [3, Theorem 1] showed that, if p^f is an $(e-1)$-ary divisor of p^e, then p^f is a k-ary divisor of p^e for any $k \geq e-1$ and called such a divisor to be an infinitary divisor. For any positive integer n, a divisor $d = \prod_i p_i^{f_i}$ of $n = \prod_i p_i^{e_i}$ is called an infinitary divisor if $p_i^{f_i}$ is an infinitary divisor of $p_i^{e_i}$ for each i, which is written as $d \mid_\infty n$.

According to E. Cohen [2], Wall [16] and G. L. Cohen [3] respectively, henceforth $\sigma^*(N), \sigma^{**}(N)$ and $\sigma_\infty(n)$ denote the sum of unitary, biunitary and infinitary divisors of N, respectively.

Replacing σ by σ^*, Subbarao and Warren [14] introduced the notion of a unitary perfect number. N is called unitary perfect if $\sigma^*(N) = 2N$. They proved that there are no odd unitary perfect numbers and $6, 60, 90, 87360$ are the first four unitary perfect numbers. Later the fifth unitary perfect number has been found by Wall [17], but no further instance has been found. Subbarao [13] conjectured that there are only finitely many unitary perfect numbers. Similarly, a positive integers N is called biunitary perfect if $\sigma^{**}(N) = 2N$. Wall [16] showed that $6, 60$ and 90, the first three unitary perfect numbers, are the only biunitary perfect numbers.

G. L. Cohen [3] introduced the notion of infinitary perfect numbers; a positive integer n is called infinitary perfect if $\sigma_\infty(n) = 2n$. Cohen [3, Theorem 16] showed that $6, 60$ and 90, exactly all of the biunitary perfect numbers, are also all of the infinitary perfect numbers not divisible by 8. Cohen gave 14 infinitary perfect numbers and Pedersen's database, which is now available at [8], contains 190 infinitary perfect numbers.

Combining the notion of superperfect numbers and the notion of unitary divisors, Sitaramaiah and Subbarao [10] studied unitary superperfect numbers, integers N satisfying $\sigma^*(\sigma^*(N)) = 2N$. They found all unitary superperfect numbers below 10^8. The first ones are $2, 9, 165, 238$. Thus, there are both even and odd ones. The author [18] showed that $9, 165$ are all the odd ones.

Now we can call an integer N satisfying $\sigma_\infty(\sigma_\infty(N)) = 2N$ to be infinitary superperfect. We can see that 2 and 9 are infinitary superperfect, while 2 is also superperfect (in the ordinary sense) and 9 is also unitary superperfect. Below 2^{29}, we can find some integers n dividing $\sigma_\infty(\sigma_\infty(n))$ but we cannot find any other infinitary superperfect numbers.

Analogous to [18], we can show that following result.

Theorem 1.1. 9 *is the only odd infinitary superperfect number.*

We can see that this immediately follows from the following result.

Theorem 1.2. *If N is an infinitary superperfect number with $\omega(\sigma_\infty(N)) \leq 2$, then $N = 2$ or $N = 9$.*

Indeed, if N is odd and $\sigma_\infty(\sigma_\infty(N)) = 2N$, then $\sigma_\infty(N)$ is a prime power or of the form $2^f q^{2^l}$ with $f, l \geq 0$ as shown in Section 3.

Table 1: All integers $N \leq 2^{29}$ for which $\sigma_\infty(\sigma_\infty(N)) = kN$

N	k
1	1
2	2
$8 = 2^3$	3
$9 = 3^2$	2
$10 = 2 \cdot 5$	3
$15 = 3 \cdot 5$	4
$18 = 2 \cdot 3^2$	4
$24 = 2^3 \cdot 3$	5
$30 = 2 \cdot 3 \cdot 5$	5
$60 = 2^2 \cdot 3 \cdot 5$	6
$720 = 2^4 \cdot 3^2 \cdot 5$	3
$1020 = 2^2 \cdot 3 \cdot 5 \cdot 17$	4
$4080 = 2^4 \cdot 3 \cdot 5 \cdot 17$	3
$8925 = 3 \cdot 5^2 \cdot 7 \cdot 17$	4
$14688 = 2^5 \cdot 3^3 \cdot 17$	5
$14976 = 2^7 \cdot 3^2 \cdot 13$	5
$16728 = 2^3 \cdot 3 \cdot 17 \cdot 41$	4
$17850 = 2 \cdot 3 \cdot 5^2 \cdot 7 \cdot 17$	8
$35700 = 2^2 \cdot 3 \cdot 5^2 \cdot 7 \cdot 17$	6
$36720 = 2^4 \cdot 3^3 \cdot 5 \cdot 17$	6
$37440 = 2^6 \cdot 3^2 \cdot 5 \cdot 13$	6
$66912 = 2^5 \cdot 3 \cdot 17 \cdot 41$	3
$71400 = 2^3 \cdot 3 \cdot 5^2 \cdot 7 \cdot 17$	12
$285600 = 2^5 \cdot 3 \cdot 5^2 \cdot 7 \cdot 17$	9
$308448 = 2^5 \cdot 3^4 \cdot 7 \cdot 17$	5
$381888 = 2^6 \cdot 3^3 \cdot 13 \cdot 17$	5
$428400 = 2^4 \cdot 3^2 \cdot 5^2 \cdot 7 \cdot 17$	3
$602208 = 2^5 \cdot 3^3 \cdot 17 \cdot 41$	6
$636480 = 2^6 \cdot 3^2 \cdot 5 \cdot 13 \cdot 17$	4
$763776 = 2^7 \cdot 3^3 \cdot 13 \cdot 17$	10
$856800 = 2^5 \cdot 3^2 \cdot 5^2 \cdot 7$	6
$1321920 = 2^6 \cdot 5^5 \cdot 5 \cdot 17$	7
$1505520 = 2^4 \cdot 3^3 \cdot 5 \cdot 17 \cdot 41$	4
$3011040 = 2^5 \cdot 3^3 \cdot 5 \cdot 17 \cdot 41$	8
$3084480 = 2^6 \cdot 3^4 \cdot 5 \cdot 7 \cdot 17$	5
$21679488 = 2^7 \cdot 3^5 \cdot 17 \cdot 41$	7
$22276800 = 2^6 \cdot 3^2 \cdot 5^2 \cdot 7 \cdot 13 \cdot 17$	6
$30844800 = 2^{10} \cdot 3^4 \cdot 5^3 \cdot 7 \cdot 17$	7
$31615920 = 2^4 \cdot 3^4 \cdot 5 \cdot 7 \cdot 17 \cdot 41$	4
$44553600 = 2^7 \cdot 3^2 \cdot 5^2 \cdot 7 \cdot 13 \cdot 17$	12
$50585472 = 2^7 \cdot 3^4 \cdot 7 \cdot 17 \cdot 41$	5
$63231840 = 2^5 \cdot 3^4 \cdot 5 \cdot 7 \cdot 17 \cdot 41$	8
$126463680 = 2^6 \cdot 3^4 \cdot 5 \cdot 7 \cdot 17 \cdot 41$	6
$213721200 = 2^4 \cdot 3^3 \cdot 5^2 \cdot 7 \cdot 11 \cdot 257$	4
$230177280 = 2^9 \cdot 3 \cdot 5 \cdot 17 \cdot 41 \cdot 43$	9
$252927360 = 2^7 \cdot 3^4 \cdot 5 \cdot 7 \cdot 17 \cdot 41$	12
$307758528 = 2^6 \cdot 3^5 \cdot 7 \cdot 11 \cdot 257$	5
$345265920 = 2^8 \cdot 3^2 \cdot 5 \cdot 17 \cdot 41 \cdot 43$	3
$427442400 = 2^5 \cdot 3^3 \cdot 5^2 \cdot 7 \cdot 11 \cdot 257$	8
$437898240 = 2^{10} \cdot 3^2 \cdot 5 \cdot 13 \cdot 17 \cdot 43$	5
$466794240 = 2^8 \cdot 3 \cdot 5 \cdot 11 \cdot 43 \cdot 257$	3
$512930880 = 2^6 \cdot 3^4 \cdot 5 \cdot 7 \cdot 11 \cdot 257$	4

Our method does not seem to work to find all odd super perfect numbers since $\sigma(\sigma(N)) = 2N$ does not seem to imply that $\omega(\sigma(N)) \leq 2$. Even assuming that $\omega(\sigma(N)) \leq 2$, the property of σ that $\sigma(p^e)/p^e > 1 + 1/p$ prevents us from showing that $\sigma(\sigma(N)) < 2$. All that we know is the author's result in [19] that there are only finitely many odd superperfect numbers N with $\omega(\sigma(N)) \leq k$ for each k. For the biunitary analogues, the author [20] showed that 2 and 9 are the only integers N (even or odd!) such that $\sigma^{**}(\sigma^{**}(N)) = 2N$.

In Table 1, we give all integers $N \leq 2^{29}$ dividing $\sigma_\infty(\sigma_\infty(N))$. We found no other infinitary superperfect numbers other than 2 and 9, while we found several integers N dividing $\sigma_\infty(\sigma_\infty(N))$. From this table, we are led to conjecture that 2 is the only even infinitary superperfect number. On the other hand, it seems that for any integer $k \geq 3$, there exist infinitely many integers N for which $\sigma_\infty(\sigma_\infty(N)) = kN$.

2. Preliminary lemmas

In this section, we shall give several preliminary lemmas concerning the sum of infinitary divisors used to prove our main theorems.

We begin by introducing Theorem 8 of [3]: writing binary expansions of e, f as $e = \sum_{i \in I} 2^i$ and $f = \sum_{j \in J} 2^j$, p^f is an infinitary divisor of p^e if and only if J is a subset of I.

Hence, factoring $n = \prod_{i=1}^{r} p_i^{e_i}$ and writing a binary expansion of each e_i as $e_i = \sum_j y_{ij} 2^j$ with $y_{ij} \in \{0, 1\}$, we observe that, as is shown in [3][Theorem 13],

$$\sigma_\infty(n) = \prod_{i=1}^{r} \prod_{y_{ij}=1} \left(1 + p_i^{2^j}\right). \tag{2.1}$$

From this, we can easily deduce the following lemma.

Lemma 2.1. *Let $v_p(n)$ denote the exponent of a prime p in the factorization of the integer n and let $l(e)$ denote the number of 1's in the binary expansion of e. Then we have $v_2(\sigma_\infty(n)) \geq \sum_{p>2} l(v_p(n)) \geq \omega(n) - 1$. In particular, $\sigma_\infty(n)$ is odd if and only if n is a power of 2.*

Proof. For each prime factor p_i, write a binary expansion of each e_i as $e_i = \sum_j y_{ij} 2^j$ with $y_{ij} \in \{0, 1\}$. Hence, $l(e_i) = \sum_j y_{ij}$ holds for each i. Unless $p_i = 2$, $p_i^{2^j} + 1$ is even for any $j \geq 0$. By (2.1), each product $\sigma_\infty(p_i^{e_i}) = \prod_{y_{ij}=1} \left(1 + p_i^{2^j}\right)$ except $p_i = 2$ is divisible by 2 at least $l(e_i)$ times and $\sigma_\infty(n)$ at least $\sum_{p_i \neq 2} l(e_i)$ times. We can easily see that $\sum_{p_i \neq 2} l(e_i) \geq \omega(n) - 1$ since $l(m) > 0$ for any nonzero integer m. □

The following two lemmas follow almost immediately from Bang's result [1]. But we shall include direct proofs.

Lemma 2.2. *If p is a prime and $\sigma_\infty(p^e)$ is a prime power, then p is a Mersenne prime and $e = 1$ or $p = 2, e = 2^l$ and $\sigma_\infty(p^e)$ is a Fermat prime.*

Proof. If $e = 1$, then $p + 1$ must be a prime power. If p is odd, then $p + 1$ must be even and therefore a power of two. Hence, $p = 2$ or p is a Mersenne prime.

If $e = 2^l \geq 2$ is a power of two, then $\sigma_\infty(p^e) = p^{2^l} + 1$ must be a prime power, which is shown to be impossible by Lebesgue [6]. Hence, $p^{2^l} + 1$ must be prime. If $p > 2$, then $p^{2^l} + 1 > 2$ is even and therefore cannot be prime. If $p = 2$, then $\sigma_\infty(2^e) = 2^{2^l} + 1$ must be a Fermat prime.

If $l(e) > 0$, then $\sigma_\infty(p^e)$ has at least two factors $p^{2^k} + 1$ and $p^{2^l} + 1$ with $l > k$. If p is odd, then $p^{2^l} + 1$ cannot be prime power as above. If $p = 2$, then these two factors must give distinct Fermat primes. Hence, in both cases, $(p^{2^k} + 1)(p^{2^l} + 1)$ cannot be a prime power and neither can $\sigma_\infty(p^e)$. □

Lemma 2.3. *$\sigma_\infty(2^e)$ has at least $l(e)$ distinct prime factors. If p is an odd prime, then $\sigma_\infty(p^e)$ has at least $l(e) + 1$ distinct prime factors.*

Proof. Whether p is odd or two, $\sigma_\infty(p^e)$ is the product of $l(e)$ distinct numbers of the form $p^{2^l}+1$. If $k>l$, then $p^{2^k}+1\equiv 2 \pmod{p^{2^l}+1}$ and therefore $p^{2^k}+1$ has a odd prime factor not dividing $p^{2^l}+1$. □

Finally, we shall introduce two technical lemmas needed in the proof.

Lemma 2.4. *If $p^2+1=2q^m$ with $m\geq 2$, then m must be a power of 2 and, for any given prime q, there exists at most one such m. If $p^{2^k}+1=2q^m$ with $k>1$, then $m=1$.*

Proof. Cohn [4] showed that $x^2+1=2y^n$ has no solution in positive integers x,y,n with $xy>1$ and $n>2$ other than $(x,y,n)=(239,13,4)$, quoting the result of Ljunggren [7] and the simpler proof of Steiner and Tzanakis [11] for $n=4$ and rediscovering the result of Størmer [12, Théorème 8] for odd n.

Hence, if $p^2+1=2q^m$ with $m\geq 2$, then we must have $m=2$ for any prime $q\neq 239$ and $m=4$ for $q=239$. This implies the former statement.

If $p^{2^k}+1=2q^m$ with $k>1$, then $m=2^l$ for some integer $l\geq 0$. Now the latter statement follows observing that $x^4+1=2y^2$, equivalent to $y^4-x^4=(y^2-1)^2$, has no solution other than $(1,1)$ by Fermat's well-known right triangle theorem (see for example Theorem 2 in Chapter 4 of Mordell [9]). □

Lemma 2.5. *If p,q are odd primes satisfying $p^{2^k}+1=2q$ and $2^{2^{k+1}}\equiv 1 \pmod q$ with $k>0$, then $(p,q)=(3,5)$ and $k=1$.*

Proof. Since q divides $2^{2^{k+1}}-1=(2^{2^k}+1)(2^{2^k}-1)$, q must divide either of $2^{2^k}+1$ or $2^{2^k}-1$. In both cases, $q\leq 2^{2^k}+1$ and therefore, noting that $k>0$,

$$2^{(2^k+1)(\log p/\log 2)}=p^{2^k+1}<2q\leq 2(2^{2^k}+1)=2^{2^k+1}+2<2^{2^k+2}. \tag{2.2}$$

Hence, we have $(2^k+1)(\log p/\log 2)<2^k+2$ and $\log p/\log 2<1+1/(2^k+1)$, which leads to $k=1, p=3$ and $q=(3^2+1)/2=5$. □

3. Proofs of Theorems 1.1 and 1.2

We begin by noting that Theorem 1.1 follows from Theorem 1.2. Indeed, if N is odd and $\sigma_\infty(\sigma_\infty(N))=2N$, then Lemma 2.1 gives that $\omega(\sigma_\infty(N))\leq 2$ and therefore Theorem 1.2 would yield Theorem 1.1.

In order to prove Theorem 1.2, we shall first show that if $\sigma_\infty(N)$ is odd or a prime power, then N must be 2. If $\sigma_\infty(N)$ is a prime power, then Lemma 2.2 immediately yields that $N=2^e$ or $\sigma_\infty(N)$ is a power of 2, where the latter case cannot occur since $\sigma_\infty(\sigma_\infty(N))$ must be odd in the latter case while we must have $\sigma_\infty(\sigma_\infty(N))=2N$. If $\sigma_\infty(N)$ is odd, then N must be a power of 2 by Lemma 2.1.

Thus, we see that if $\sigma_\infty(N)$ is odd or a prime power, then $N=2^e$ must be a power of 2. Now we can easily see that $\sigma_\infty(\sigma_\infty(N))=2N=2^{e+1}$ must also be a power of 2. Hence, for each prime-power factor $q_i^{f_i}$ of $\sigma_\infty(N)$, $\sigma_\infty(q_i^{f_i})$ is also a power of 2. By Lemma 2.2, each $f_i = 1$ and q_i is a Mersenne prime. Hence, we see that $\sigma_\infty(N) = \sigma_\infty(2^e)$ must be a product of Mersenne primes. Let r be an integer such that $2^{2^r} \mid_\infty N$. Then $2^{2^r} + 1$ must also be a product of Mersenne primes. By the first supplementary law, only $r = 0$ is appropriate and therefore $e = 0$. Thus, we conclude that if $\sigma_\infty(N)$ is odd or a prime power, then $N = 2$.

Henceforth, we are interested in the case $\sigma_\infty(N) = 2^f q^{2^l}$ with $f > 0$ and $l \geq 0$. Factor $N = \prod_i p_i^{e_i}$. Our proof proceeds as follows: (I) if $l = 0$, then there exists exactly one prime factor p_i of N such that q divides $\sigma_\infty(p_i^{e_i})$, (IA) if $l = 0$ and $f = 1$, then $N = 9$, (IBa) it is impossible that $l = 0, f > 1$ and $p_i \mid q + 1$, (IBb) it is impossible that $l = 0, f > 1$ and p_i does not divide $q + 1$, (II) if $l > 0$, then there exists at most one prime factor p_i of $q^{2^l} + 1$ such that $p_i^{2^k} + 1 = 2q$, (IIa) it is impossible that $q^{2^l} + 1$ has no such prime factor, (IIb) it is impossible that $q^{2^l} + 1$ has one such prime factor p_i.

First we shall settle the case $l = 0$, that is, $\sigma_\infty(N) = 2^f q$. Since q divides N exactly once, there exists exactly one index i such that q divides $\sigma_\infty(p_i^{e_i})$.

For any index j other than i, we must have $\sigma_\infty(p_j^{e_j}) = 2^{k_j}$ and therefore, by Lemma 2.2, we have $e_j = 1$ and $p_j = 2^{k_j} - 1$ for some intger k_j. Clearly, p_j must divide $2N = \sigma_\infty(\sigma_\infty(N)) = \sigma_\infty(2^f)(q + 1)$ and the first supplementary law yields that $p_j \mid (q + 1)$ unless $p_j = 3$.

If $f = 1$, then $N = 2^k p^e$ for an odd prime p by Lemma 2.1 and $2N = \sigma_\infty(2q) = 3(q+1)$. Hence, $p = 3$ and $\sigma_\infty(2^k 3^e) = 2q$. But, we observe that $k = 0$ and $N = 3^e$ since $\sigma_\infty(3^e) > 2$ is even. By Lemma 2.1, we have $e = 2^u$ and $3^e + 1 = 2q = 2(2 \times 3^{e-1} - 1) = 4 \times 3^{e-1} - 2$. Hence, $3^{e-1} = 3$, that is, $N = 9$ and $q = 5$. This gives an infinitary superperfect number 9.

Now we consider the case $f > 1$. If $2^{2^m} \mid_\infty 2^f$ with $m > 0$ and p divides $2^{2^m} + 1$, then p must be congruent to 1 (mod 4) and therefore must be p_i. By Lemma 2.2, we must have $2^{2^m} + 1 = p_i$. Hence, $f = 2^m$ and $\sigma_\infty(2^f) = p_i$ or $f = 2^m + 1$ and $\sigma_\infty(2^f) = 3p_i$.

If p_i divides $q + 1$, then $e_i \geq 2$. By Lemma 2.3, we must have $e_i = 2^v$ and $p_i^{e_i} + 1 = 2q$. Since $p_i = \sigma_\infty(2^f)$, $p_i^{e_i - 1}$ divides $q + 1$ and therefore $2(q+1) = p_i^{e_i} + 3$, Hence, $p_i^{e_i} \equiv -3 \pmod{p_i^{e_i - 1}}$, which is impossible since $p_i > 3$ now.

If p_i does not divide $q + 1$, then $e_i = 1$ and $2^{k_i} q = p_i + 1 = 2^{2^m} + 2$. Hence, $k_i = 1$ and $q = 2^{2^m - 1} + 1$. Now $m = 1$ with $q = 3$ is the only m such that q is prime. Hence, we have $p_i = 2q - 1 = 5$, $\sigma_\infty(2^f) = 5$ or 15 and $N = \sigma_\infty(2^f)(q + 1)/2 = 10$ or 30, neither of which is infinitary superperfect. Thus, the case $\sigma_\infty(N) = 2^f q$ with $f > 1$ has turned out to be impossible and $N = 3^2$ is the only infinitary superperfect number with $\sigma_\infty(N) = 2q$.

Now the remaining case is $\sigma_\infty(N) = 2^f q^g$ with $g > 1$. We can take a positive integer l such that $q^{2^l} \mid_\infty q^g$. If p is odd and divides $\sigma_\infty(q^{2^l}) = q^{2^l} + 1$, then p divides $\sigma_\infty(\sigma_\infty(N)) = 2N$ and therefore p divides N. If $p^{2^k} \mid_\infty N$, then $p^{2^k} + 1$ divides $\sigma_\infty(N) = 2^f q^{2^l}$ and therefore we can write $p^{2^k} + 1 = 2q^t$. We note that $p \equiv 1 \pmod 4$ since p is odd and divides $q^{2^l} + 1$ with $l > 0$. Hence, we see that a) if $k = 0$, then $p + 1 = 2q^t$, b) if $k = 1$, then $p^2 + 1 = 2q$ or $2q^{2^u}$ by Lemma 2.4 and

c) if $k > 1$, then $p^{2^k} + 1 = 2q$ by Lemma 2.4.

Clearly, there exists at most one prime factor p_i of N such that $p_i^{2^k} + 1 = 2q$ for some integer $k > 0$. Moreover, by Lemma 2.4, there exists at most one prime factor p_j of N such that $p_j^2 + 1 = 2q^{2^u}$ for some integer $u > 0$. Letting i and j denote the indices of such primes respectively if these exist, $q^{2^l} + 1$ can be written in the form

$$q^{2^l} + 1 = 2p_i^{s_i} p_j^{s_j} (2q^{t_1} - 1)(2q^{t_2} - 1)..., \tag{3.1}$$

where $s_i, s_j \geq 0$ may be zero.

If $s_i \neq 0$, then we have $2p_i^{s_i} p_j^{s_j} \equiv \pm 1 \pmod{q}$ and therefore, observing that $p_i^{2^{k+1}} \equiv p_j^4 \equiv 1 \pmod{q}$, we have $2^{2^{k+1}} \equiv 1 \pmod{q}$. By Lemma 2.5, we must have $p_i = 3, e_i = 2$ and $q = 5$ and p_j cannot exist. Since $p_i = 3$ divides $q^{2^l} + 1$, we must have $l = 0$, contrary to the assumption $l > 0$.

If $s_i = 0$, then we must have $2p_j^{s_j} \equiv \pm 1 \pmod{q}$. If s_j is even, then $2p_j^{s_j} \equiv 2(-1)^{s_j/2} \equiv \pm 2 \pmod{q}$ cannot be $\pm 1 \pmod{q}$. Hence, s_j must be odd and $2p_j \equiv \pm 1 \pmod{q}$. Since $p_j^4 \equiv 1 \pmod{q}$, we have $2^4 \equiv 1 \pmod{q}$ and $q \equiv 1 \pmod{4}$. Equivalently, we have $q = 5$ and therefore $p_j^2 + 1 = 2 \times 5^{2^k}$ with $k > 0$. By Lemma 2.5, we must have $p_j = 7$ and $k = 1$. However, this is impossible since 7 divides neither $\sigma_\infty(5^2) = 2 \times 13$ nor $\sigma_\infty(2^f)$ by the first supplementary law. Now our proof is complete.

References

[1] A. S. Bang, Taltheoretiske Undersøgelser, *Tidsskrift Math.* Vol. 5 IV (1886), 70–80 and 130–137.

[2] Eckford Cohen, Arithmetical functions associated with the unitary divisors of an integer, *Math. Z.* Vol. 74 (1960), 66–80.

[3] Graeme L. Cohen, On an integer's infinitary divisors, *Math. Comp.* Vol. 54(189) (1990), 395–411.

[4] J. H. E. Cohn, Perfect Pell powers, *Glasgow Math. J.* Vol. 38 (1996), 19–20.

[5] G. H. Hardy and E. M. Wright, revised by D. R. Heath-Brown and J. H. Silverman, *An Introduction to the Theory of Numbers*, Sixth edition, Oxford University Press, Oxford, 2008.

[6] M. Lebesgue, Sur l'impossibilité, en nombres entiers, de l'équation $x^m = y^2 + 1$, *Nouv. Ann. Math. sér. 1*, Vol. 9 (1850), 178–181.

[7] W. Ljunggren, Zur theorie der Gleichung $X^2 + 1 = DY^4$, *Avh. Norske, Vid. Akad. Oslo* Vol. 1, No. 5 (1942).

[8] David Moews, A database of aliquot cycles, `http://djm.cc/aliquot-database/aliquot-database.uhtml`.

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

[10] V. Sitaramaiah and M. V. Subbarao, On the equation $\sigma^*(\sigma^*(N)) = 2N$, *Util. Math.* Vol. 53 (1998), 101–124.

[11] Ray Steiner and Nikos Tzanakis, Simplifying the solution of Ljunggren's equation $X^2 + 1 = 2Y^4$, *J. Number Theory* Vol. 37(2) (1991), 123–132.

[12] Carl Størmer, Quelques théorèmes sur l'équation de Pell $x^2 - Dy^2 = \pm 1$ et leurs applications, *Skrift. Vidensk. Christiania I. Math. -naturv. Klasse* (1897), Nr. 2, 48 pages.

[13] M. V. Subbarao, Are there an infinity of unitary perfect numbers?, *Amer. Math. Monthly* Vol. 77(4) (1970), 389–390.

[14] M. V. Subbarao and L. J. Warren, Unitary perfect numbers, *Canad. Math. Bull.* Vol. 9 (1966), 147–153.

[15] D. Suryanarayana, Super perfect numbers, *Elem. Math.* Vol. 24 (1969), 16–17.

[16] Charles R. Wall, Bi-unitary perfect numbers, *Proc. Amer. Math. Soc.* Vol. 33 (1972), 39–42.

[17] Charles R. Wall, The fifth unitary perfect number, *Canad. Math. Bull.* Vol. 18 (1975), 115–122.

[18] T. Yamada, Unitary super perfect numbers, *Math. Pannon.* Vol. 19(1) (2008), 37–47.

[19] T. Yamada, On finiteness of odd superperfect numbers, `https://arxiv.org/abs/0803.0437`.

[20] T. Yamada, 2 and 9 are the only biunitary superperfect numbetrs, `https://arxiv.org/abs/1705.00189`.

Annales Mathematicae et Informaticae
47 (2017) pp. 219–225
http://ami.uni-eszterhazy.hu

On bounded and unbounded curves determined by their curvature and torsion

Oleg Zubelevich[*]

Department of Theoretical mechanics, Mechanics and Mathematics Faculty
M. V. Lomonosov moscow State University, Russia, Moscow
ozubel@yandex.ru

Submitted December 7, 2016 — Accepted January 23, 2017

Abstract

We consider a curve in $\mathbb{R}^3$ and provide sufficient conditions for the curve to be unbounded in terms of its curvature and torsion. We also present sufficient conditions on the curvatures for the curve to be bounded in $\mathbb{R}^4$.

Keywords: Curves, intrinsic equation, curvature, torsion, Frenet-Serret formulas.

MSC: 53A04

1. Introduction

This short note concerns a smooth curve γ in the standard three-dimensional Euclidean space $\mathbb{R}^3$. It is well known that the curve is uniquely defined (up to translations and rotations of $\mathbb{R}^3$) by its curvature $\kappa(s)$ and its torsion $\tau(s)$, the argument s is the arc-length parameter. The pair $(\kappa(s), \tau(s))$ is called the intrinsic equation of the curve.

In the sequel we assume that $\kappa, \tau \in C[0, +\infty)$.

To obtain the radius-vector of the curve γ one must solve the system of Frenet-Serret equations:

$$\begin{aligned} \boldsymbol{v}'(s) &= \kappa(s)\boldsymbol{n}(s), \\ \boldsymbol{n}'(s) &= -\kappa(s)\boldsymbol{v}(s) + \tau(s)\boldsymbol{b}(s), \end{aligned} \tag{1.1}$$

[*] Partially supported by grant RFBR 15-01-03747.

$$\boldsymbol{b}'(s) = -\tau(s)\boldsymbol{n}(s).$$

The vectors $\boldsymbol{v}(s), \boldsymbol{n}(s), \boldsymbol{b}(s)$ stand for the Frenet-Serret frame at the curve's point with parameter s. Then the radius-vector of the curve is computed as follows $\boldsymbol{r}(s) = \int_0^s \boldsymbol{v}(\xi)d\xi + \boldsymbol{r}(0)$.

If the curve γ is flat (it is so iff $\tau(s) = 0$) then the system (1.1) is integrated explicitly. In three dimensional case nobody can integrate this system with arbitrary sooth functions τ, κ.

So we obtain a very natural and pretty problem: to restore the properties of the curve γ having the curvature $\kappa(s)$ and the torsion $\tau(s)$.

For example, under which conditions on the functions κ, τ a curve γ is closed? This is a hard open problem. There may by another question: Which are sufficient conditions for the whole curve to be contained in a sphere? This question is much simpler. Such a type questions have been discussed in [4, 3, 5].

There is a sufficient condition for the curve to be unbounded [1]. In this article the condition is formulated in terms of curvature only and this condition is valid in a big class of spaces, including Hilbert spaces and Riemannian manifolds of non-positive curvature.

In general case, (1.1) is a linear system of ninth order with matrix depending on s. To describe the properties of γ one must study this system.

In this note we formulate and prove some sufficient conditions for unboundedness of the curve γ.

We also present sufficient conditions for the curve to be bounded in the four dimensional Euclidean space.

It is interesting that in $\mathbb{R}^m$ of odd m the curves are in generic case unbounded but for the even m they are generically bounded. Some justification of this very informal observation is given below.

2. Main theorem

We shall say that γ is unbounded iff $\sup_{s\ge 0} |\boldsymbol{r}(s)| = \infty$.

Theorem 2.1. *Suppose there exists a function $\lambda(s)$ such that functions*

$$k(s) = \lambda(s)\kappa(s), \quad t(s) = \lambda(s)\tau(s)$$

are monotone[1] *and belong to $C[0,\infty)$.*

Introduce a function $T(s) = \int_0^s t(\xi)d\xi$.

Suppose also that the following equalities hold

$$\lim_{s\to\infty} T(s) = \infty, \quad \lim_{s\to\infty} \frac{k(s)}{T(s)} = \lim_{s\to\infty} \frac{t(s)}{T(s)} = 0. \tag{2.1}$$

[1] E.g. one of these functions, $k(s)$ is monotonically increasing: $s' < s'' \Rightarrow k(s') \le k(s'')$, $s', s'' \in [0,\infty)$ while other one $t(s)$ is monotonically decreasing: $s' < s'' \Rightarrow t(s') \ge t(s'')$, $s', s'' \in [0,\infty)$. The inverse situation is also allowed, or the both functions can be increasing or decreasing simultaneously.

Then the curve γ is unbounded.

The proof of this theorem is contained in Section 4.1.

Putting $\lambda = 1/\tau$ in this Theorem , we deduce the following corollary.

Corollary 2.2. *Suppose that the function $\kappa(s)/\tau(s)$ is monotone and*

$$\lim_{s\to\infty} \frac{\kappa(s)}{s\cdot\tau(s)} = 0. \tag{2.2}$$

Then the curve γ is unbounded.

Note that the geodesic curvature of the tantrix[2] $\kappa_T(s)$ is equal to $\tau(s)/\kappa(s)$ [3]. So that formula (2.2) can be rewritten as follows

$$\lim_{s\to\infty} \kappa_T(s)s = \infty.$$

Theorem 2.1 is not reduced to Corollary 2.2. Consider an example. Let the curve γ be given by

$$\kappa(s) = 1, \quad \tau(s) = \frac{1}{1+s}.$$

Since $\tau(s) \to 0$ as $s \to \infty$ it may seem that this curve is about a circle with $\kappa(s) = 1$. Nevertheless applying Theorem 2.1 with $\lambda = 1$ we see that the curve γ is unbounded.

Consider a system which consists of (1.1) together with the equation $\boldsymbol{r}'(s) = \boldsymbol{v}(s)$. From viewpoint of stability theory, Theorem 2.1 states that under certain conditions this system is unstable.

Since $|\boldsymbol{r}(s)| = O(s)$ as $s \to \infty$, this instability is too weak to study it by standard methods such as the Lyapunov exponents method.

3. Supplementary remarks: Bounded curves in $\mathbb{R}^4$

Actually the above developed technique can be generalized to the curves in any multidimensional Euclidean space $\mathbb{R}^m$. For the case of the odd m we can prove a theorem similar to Theorem 2.1. But for the case when m is even our method allows to obtain sufficient conditions for the curve to be bounded.

In this section we illustrate such an effect. To avoid of big formulas we consider only the case $m = 4$.

So let a curve $\gamma \subset \mathbb{R}^4$ be given by its curvatures

$$\kappa_i(s) \in C[0,\infty), \quad i = 1, 2, 3.$$

And let $\boldsymbol{v}_j(s)$, $j = 1, 2, 3, 4$ be the Frenet-Serret frame.

[2] The tangential spherical image of the curve γ is the curve on the unit sphere. This curve has the radius-vector $\boldsymbol{r}'(s)$.

Then the Frenet-Serret equations are

$$\frac{d}{ds}\begin{pmatrix}\boldsymbol{v_1}\\ \boldsymbol{v_2}\\ \boldsymbol{v_3}\\ \boldsymbol{v_4}\end{pmatrix}(s)=A(s)\begin{pmatrix}\boldsymbol{v_1}\\ \boldsymbol{v_2}\\ \boldsymbol{v_3}\\ \boldsymbol{v_4}\end{pmatrix}(s),$$

$$A(s)=\begin{pmatrix}0&\kappa_1(s)&0&0\\ -\kappa_1(s)&0&\kappa_2(s)&0\\ 0&-\kappa_2(s)&0&\kappa_3(s)\\ 0&0&-\kappa_3(s)&0\end{pmatrix}$$

Theorem 3.1. *Suppose that the function $\kappa_1(s)\kappa_3(s)$ does not take the value zero. The functions*

$$f_1(s)=\frac{1}{\kappa_1(s)},\quad f_2(s)=\frac{\kappa_2(s)}{\kappa_1(s)\kappa_3(s)}$$

are monotone and

$$\sup_{s\geq 0}|f_i(s)|<\infty,\quad i=1,2.$$

Then the curve γ is bounded.

The proof of this theorem is contained in Section 4.2.

4. Proofs

4.1. Proof of Theorem 2.1

Let us expand the radius-vector by the Frenet-Serret frame

$$\boldsymbol{r}(s)=r_1(s)\boldsymbol{v}(s)+r_2(s)\boldsymbol{n}(s)+r_3(s)\boldsymbol{b}(s).$$

Differentiating this formula we obtain

$$\begin{aligned}\boldsymbol{v}(s)&=r_1'(s)\boldsymbol{v}(s)+r_2'(s)\boldsymbol{n}(s)+r_3'(s)\boldsymbol{b}(s)\\ &\quad+r_1(s)\boldsymbol{v}'(s)+r_2(s)\boldsymbol{n}'(s)+r_3(s)\boldsymbol{b}'(s).\end{aligned}$$

Using the Frenet-Serret equations, one yields

$$r'(s)=\begin{pmatrix}0&\kappa(s)&0\\ -\kappa(s)&0&\tau(s)\\ 0&-\tau(s)&0\end{pmatrix}r(s)+\begin{pmatrix}1\\0\\0\end{pmatrix},\qquad r=\begin{pmatrix}r_1\\r_2\\r_3\end{pmatrix}.\tag{4.1}$$

The author was informed about system (4.1) by Professor Ya. V. Tatarinov.

Let us multiply both sides of system (4.1) by the row-vector

$$\lambda(s)\big(\tau(s),0,\kappa(s)\big)$$

from the left:

$$t(s)r_1'(s) + k(s)r_3'(s) = t(s).$$

Then we integrate this equation:

$$\int\limits_0^s t(a)r_1'(a)da + \int\limits_0^s k(a)r_3'(a)da = T(s). \tag{4.2}$$

From the Second Mean Value Theorem [2], we know that there is a parameter $\xi \in [0, s]$ such that

$$\begin{aligned}\int\limits_0^s t(a)r_1'(a)da &= t(0)\int\limits_0^\xi r_1'(a)da + t(s)\int\limits_\xi^s r_1'(a)da \\ &= t(0)\big(r_1(\xi) - r_1(0)\big) + t(s)\big(r_1(s) - r_1(\xi)\big)\end{aligned}$$

By the same argument for some $\eta \in [0, s]$ we have

$$\int\limits_0^s k(a)r_3'(a)da = k(0)\big(r_3(\eta) - r_3(0)\big) + k(s)\big(r_3(s) - r_3(\eta)\big).$$

Thus formula (4.2) takes the form

$$\begin{aligned}t(0)\big(r_1(\xi) - r_1(0)\big) + t(s)\big(r_1(s) - r_1(\xi)\big)& \\ + k(0)\big(r_3(\eta) - r_3(0)\big) + k(s)\big(r_3(s) - r_3(\eta)\big) &= T(s).\end{aligned} \tag{4.3}$$

Since the Frenet-Serret frame is orthonormal we have

$$|\boldsymbol{r}(s)|^2 = r_1^2(s) + r_2^2(s) + r_3^2(s) = |r(s)|^2.$$

Assume the Theorem is not true: the curve γ is bounded, i.e. $\sup_{s\geq 0} |\boldsymbol{r}(s)| < \infty$. Then due to conditions (2.1) the left side of formula (4.3) is $o(T(s))$ as $s \to \infty$. This contradiction proves the theorem.

The Theorem is proved.

4.2. Proof of Theorem 3.1

Let $\boldsymbol{r}(s)$ be a radius-vector of the curve γ. Then one can write

$$\boldsymbol{r}(s) = \sum_{i=1}^4 r_i \boldsymbol{v}_i(s), \quad \boldsymbol{r}'(s) = \boldsymbol{v}_1(s).$$

Similarly as in the previous section, due to the Frenet-Serret equations this gives

$$r'(s) = A(s)r(s) + \begin{pmatrix}1\\0\\0\\0\end{pmatrix}, \qquad r = \begin{pmatrix}r_1\\r_2\\r_3\\r_4\end{pmatrix}.$$

First we multiply this equation by $r'^T(s)A^{-1}(s)$, $\quad(\det A=(\varkappa_1\varkappa_3)^2)$:

$$r'^T(s)A^{-1}(s)r'(s)=r'^T(s)r(s)+r'^T(s)A^{-1}(s)\begin{pmatrix}1\\0\\0\\0\end{pmatrix}. \tag{4.4}$$

Since A^{-1} is a skew-symmetric matrix we have $r'^T(s)A^{-1}(s)r'(s)=0$, and some calculation yields

$$r'^T(s)A^{-1}(s)\begin{pmatrix}1\\0\\0\\0\end{pmatrix}=r_2'(s)f_1(s)+r_4'(s)f_2(s).$$

Then formula (4.4) takes the form

$$-\frac{1}{2}\Big(|r(s)|^2\Big)'=r_2'(s)f_1(s)+r_4'(s)f_2(s).$$

Integrating this formula we obtain

$$-\frac{1}{2}\Big(|r(s)|^2-|r(0)|^2\Big)=\int\limits_0^s r_2'(a)f_1(a)+r_4'(a)f_2(a)da.$$

By the same argument which was employed to obtain formula (4.3), it follows that

$$\begin{aligned}&-\frac{1}{2}\Big(|r(s)|^2-|r(0)|^2\Big)=\\&f_1(0)\big(r_2(\xi)-r_2(0)\big)+f_1(s)\big(r_2(s)-r_2(\xi)\big)+\\&f_2(0)\big(r_4(\eta)-r_4(0)\big)+f_2(s)\big(r_4(s)-r_4(\eta)\big),\end{aligned} \tag{4.5}$$

here $\xi,\eta\in[0,s]$.

To proceed with the proof assume that the curve γ be unbounded:

$$\sup_{s\ge 0}|r(s)|=\infty.$$

Take a sequence s_k such that

$$|r(s_k)|=\max_{s\in[0,k]}|r(s)|,\quad k\in\mathbb{N},\quad s_k\in[0,k].$$

It is easy to see that

$$s_k\to\infty,\quad |r(s)|\le|r(s_k)|,\quad s\in[0,s_k]$$

and $|r(s_k)|\to\infty$ as $k\to\infty$.

Substitute this sequence to formula (4.5):

$$-\frac{1}{2}\Big(|r(s_k)|^2 - |r(0)|^2\Big) =$$
$$f_1(0)\big(r_2(\xi_k) - r_2(0)\big) + f_1(s_k)\big(r_2(s_k) - r_2(\xi_k)\big)+$$
$$f_2(0)\big(r_4(\eta_k) - r_4(0)\big) + f_2(s_k)\big(r_4(s_k) - r_4(\eta_k)\big), \qquad (4.6)$$

here $\xi_k, \eta_k \in [0, s_k]$ and thus $|r_2(\xi_k)| \leq |r(s_k)|, \quad |r_4(\eta_k)| \leq |r(s_k)|$.

Due to conditions of the Theorem and the choice of the sequence s_k the right-hand side of formula (4.6) is $O(|r(s_k)|)$ as $k \to \infty$. But the left-hand one is of order $-|r(s_k)|^2/2$. This gives a contradiction.

The Theorem is proved.

References

[1] S. Alexander, R. Bishop, R. Ghrist, Total curvature and simple pursuit on domains of curvature bounded above, Geometriae Dedicata, 147(2010), 275–290.

[2] R. Courant, Differential and Integral Calculus, vol. 1, John Wiley and Sons, 1988.

[3] W. Frenchel, On the differential geometry of closed space curves, Bull. Amer. Math. Soc., 57(1951), 44–54.

[4] P. W. Gifford, Some refinements in theory of specialized space curves, Amer. Math. Monthly, 60(1953), 384–393.

[5] Yung-Chow Wong, Hon-Fei Lai, A Critical Examination of the Theory of Curves in Three Dimensional Differential Geometry, Tohoku Math. Journ. Vol. 19, No. 1, 1967.

Methodological papers

Annales Mathematicae et Informaticae
47 (2017) pp. 229–242
http://ami.uni-eszterhazy.hu

Teaching digital image processing – eyes and eyesight

István Gerják

Óbuda University
gerjakist@gmail.com

Submitted April 30, 2016 — Accepted November 12, 2017

Abstract

The necessity of digital image processing and its teaching have risen dramatically in the past few years. Besides the appearance of digital cameras, the general use of smart phones suitable for making photos made it possible for anybody any time to take snapshots of the important events of their lives. Through social community networks (for example Twitter or Facebook) we can share our photos with our friends and acquaintances.

The Internet craze makes computers attractive for youngsters and we have to reconsider some different forms of education in the world of emails and chatting radically. It refers to the fact that better and better digital teaching materials and their e-Book counterparts appear in increasing numbers. They try to surpass the traditional curriculum and get young people to study in this way as well.

In the present paper, which is part of a forthcoming series of articles we are looking for novel ways and opportunities of teaching digital image processing in secondary schools in this spirit and presenting the first chapter of the digital learning material developed by ourselves.

Keywords: Teaching digital image processing, seeing, functioning of human eyes, visual defects and their correction

MSC: 68U10

1. Introduction

Infocommunication technologies (ICT) and digital competency play a key role in teaching in secondary schools [17]. Teachers who graduated more than 10 years ago at universities did not learn, moreover, were not able to learn the latest modern Internet or mobile technologies. The knowledge they should obtain could be accessed by taking part in courses or asking their IT teacher colleagues if they had enough time for this.

In our articles, which review the chapters of our coursebook separately we want to shed light on how important the computer, information technologies and the multimedia in education are and would like to publish contents and methods that illustrate the potential possibilities of this novel type of educational environment.

Also, we want to represent the connection between our curriculum and NAT objectives in which they have declared that the acquired knowledge must be durable, and it must meet the demands of our age. Our curriculum also fits into the idea that a common educational material must be available for everyone.[18]

Hereinafter we are also dealing with what and how we should teach about digital image processing at secondary schools and how our digital material supports teaching original thinking [19]. Our material assists teaching how to study as students while applying it can organise their own studies either individually or in groups, they learn to be able to manage time and information [20][21] When applying our material students can get to know how the media work and make their influence felt [20].

We are using a Moodle framework, which helps students put forth their individual abilities and talents so assists their applying controlling and assessing procedures [22].

According to NAT experimentation, observation, and the differentiated development of natural scientific thinking are particularly important. NAT also considers important that the teachers of different subjects should help each other to make students' knowledge complete and interesting.

In our material, the first chapter deals with the eye and eyesight, which is the main topic our present paper, we kept these two objectives in view[23].

No books are available for teaching Digital Image Processing in secondary schools they would be necessary though. This is the reason why we have developed our website (`www.gerjak.hu`). The first chapter in this material is about eyes and sight which is the main theme of this article. Since our paper mainly focuses on secondary school students, our references come mostly from popular books instead of scientific papers, which are difficult for them to understand.

Students could learn about eyes and eyesight in four subjects. Under the study of biology course there is only one page about the structure and the diseases of eyes (strabismus, cataracts, nearsightedness, and farsightedness [6]. Physics deals with the eyepiece, the myopic and hyperopic eyes on half a page [9].

Drawing and visual culture do not teach about the eye [1]. Besides, the creation of multimedia contents and web pages also require getting to know how the eyes function and it is necessary to know where and in what colour the contents, which are important for us should be placed. Therefore, the main questions about eyes and eyesight could be taught in informatics lessons as well.

As these days this topic has not been developed properly in textbooks, it leads to the task: let's work it out. In Chapter 2 of this paper after the introduction, we discuss the material that can be found on our website and which includes the knowledge students need to learn to perform the practices described in Chapter 2. Chapter 4 is about the development and the protection of the eyes and repairing visual errors.

1.1. Teaching digital image processing

Why is our curriculum necessary and what should it be about? What we need to learn and what we need largely depends on our profession and the life we choose. Earlier books were published by printers and they learned the craft of printing, the colours, the fonts, the properties of light, and their 'treatment'.

Today, everyone documents his own life (portfolio, website) and everyone is a printer as well. They need to learn the rules of visualisation so they also need to know the rules of 'eyesight'. Doctors study different things about eyes, painters need to see other things than turners, physicist, and IT specialists. We should teach complex skills together with other scientific fields.

Our curriculum enhances the teaching of Digital Image Processing. Our choice is supported not only by the complexity and timeliness of the topic and the need for practical skills on the subject but it is fully in line with the following statements of Comenius [8]: "Everything must be placed before the senses. Everything visible should be brought before the organ of sight, everything audible before that of hearing. Odours should be placed before the sense of smell, and things that are testable and tangible before the sense of taste and of touch respectively. If something can be sensed with more than one organ, it should be placed before each organ."

When and where should knowledge on Digital Image Processing be taught? According to the international and domestic trends word processing and spreadsheets should not be taught in IT lessons for students, but in the linguistic and mathematical training courses [15].

Instead of them, extensive use of ICT tools for teaching digital image processing techniques should be taught. The topics in relation to this could be the following: the proper use of human eyes ('proper' eyesight, usage of colours, prevention of eye diseases and eye fatigue, photography, scanning techniques, compression, filtering, improvement of photos, animation- and film-making (e.g. cutting details from existing films), converting audio and image files and presentation of web-based materials.

Accordingly, the main chapters of our curriculum are the following:

- Human vision
- Human vision / Interpretation
- Digitalization
- Photography
- Light / colours
- Image processing skills
- Creation of multimedia files / Conversion
- Creation of movies / Uploading them on websites

In our opinion, the number of lessons for teaching them is enough but there is neither curriculum nor textbook for teaching them. Skimming through the websites of secondary schools and their curriculum of informatics we can see that mostly only the creation of presentation is taught except for schools where students are taught image processing by using GIMP and creating websites only as much as it is necessary for the final exam requirements.

Most of our colleagues believe and surveys also show that the lessons given are not enough to practice and solve problems. The new tools could ease this difficulty. Using digital and e-learning materials through the Internet students could practice enough and these materials could arouse their interest more than traditional exercise books.

Besides, ICT devices maintain the students' curiosity and ongoing activity and the use of constructive teaching methods. During the creation of our curriculum we paid attention to the results of modern brain research of Josef Kraus who defined the following principles for teaching [16]:

- teaching must be activating
- The teaching-learning process must be multi-channel
- More relaxed and restful phases must be ensured
- There is a need to use surprising and unexpected situations to provoke the students' attention

But why should we study the human eye and eyesight first?

Man has always grabbed beautiful, colourful or spectacular things. Design elements are always better stuck in his memory. Although there has been no scientific research about it, almost all books, which deal with this topic mention that the efficiency of getting information by reading, hearing and seeing parallel is much better and the efficiency of seeing is two or three times better than the efficiency of the other senses [12, 5].

Why do pictures, moving figures, animations, flashes and films have such great influence on us? We will give an answer to this question at the end of our articles as well.

In earlier years teachers did their experiments with real appliances in their laboratories of physics or chemistry. All our senses played an important part in receiving and understanding information during those experiments. We touched, smelled, tasted and looked at the objects. Nowadays, in a computerized, virtual world teaching is done almost exclusively through the monitor or the projector, so the role of sight has become much more important in receiving information. That is why this paper gives such a detailed description of the parts of the eye and shows their interesting and humorous presentation during lessons.

Nowadays, besides learning everything appeals to the eye-marketing, commercials and food. So it is essential that both as a creator and a recipient we should be aware of all the impulses that affect the eye. Unfortunately this new 'seeing' and the overstrain of the eye have led to more and more eye diseases. That is why at the end of this paper we are drawing attention to the health hazards of work on the computer.

We should start teaching the causes of eye fatigue and lachrymation and also finding a solution to them at secondary schools.

It is difficult to write about everything on this topic in one paper, so those interested can read further useful tips that have been left out in our next articles.

In this paper we review the first chapter (Human sight) from our developed curriculum. In our forthcoming paper we will deal with the second chapter (Human sight/interpretation) and the Moodle-material and we are planning to publish further chapters as well.

2. The eye and eyesight

Teaching is not about 'frontal' lessons any more. Plenty of curriculum and video can be found on the Internet. Our task is to guide our students along the 'right' route. This is the route we follow while getting to know the eye. We intend three lessons for this topic. During the first lesson we talk about the eye as 'technical' means.

We throw light on new notions connected to the functioning of the eye then the teacher calls the students' attention to websites on the topic and finally we put it all to practice in the course of project work. At the end of the lesson we talk about the experience we gained during the lesson, then in the final five minutes we do exercises related to the eye. The necessity and types of these exercises are dealt with in Chapter 4.

In the next lesson the students give account of the results of their search on the Internet, which will be discussed together, then we will deal with signals coming from the eye to the brain, that is sight. Having talked over the related notions, the course of this lesson is the same as that of the previous lesson. We finish the topic in the third lesson.

Why should we start our curriculum with the eye?

The knowledge of multiplication table is indispensable in mathematics as well. To know the eye is also important, as 80% of our knowledge comes to us through it. It is important to know the size, and the colour of the eye and also what colour-environment information it sees 'well' and why. What calls our attention and what disturbs us? It is essential to be aware of central and peripheral sight as they are the basis of commercials put on the sides of websites and flashing notices.

2.1. The structure and funcioning of the eye

In the course of the lesson (we are following it in this paper) we give short pieces of knowledge, then we discuss the notions, which we want to find (key words), and finally we do exercises on the given topic. We can find some videos on the functions of the eye on YouTube, e.g.: `http://youtube.com/watch?v=Sqr6LKIR2b8`

The most perfect known organ of sight is the complex eye, which can recognise formations already. Homework: searching on the Internet – complex eye, types of eyes.

As teachers we should call our students' attention to some websites, which were created for educational purposes.

Prompt: `https://hu.wikipedia.org/wiki/Szem,www.gerjak.hu`

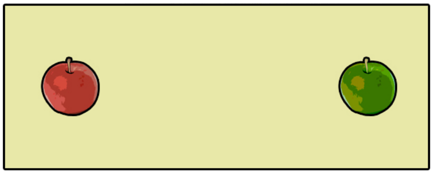


Figure 1: Experiment to prove the existence of the macula cocea

The eyes of vertebrates contain a blind spot, which can have advantages and disadvantages as well.

Homework. Searching on the internet – blind spot, left and right cerebral hemispheres, 3Dsight

Exercise. When we move closer to the red apple (left), the green apple disappears. When we move even closer, the green apple appears again. We can do this experiment with the right eye as well (see Figure 1)

Our eyes, which have more or less spherical forms, are located in the eye-sockets and are protected by seven different bones. The eyelids protect our eyes from the outside world and light.

Homework. Search on the Internet-bones, eyelids

Exercise. We should press our closed eyelids with our nails

Effect. We can see light coming from the direction where light falls on the pressed place. This proves that senses are created in the brain and not in the sense organs.

We move our eyes up and down and sideways so that we can see our environment as widely as possible. These movements are made by three pairs of ocular muscles.

According to latest studies, visual defects can be cured with the help of ocular muscles [2, 3, 4, 25]. The deformation of the eyeball caused by old age can also be improved by ocular exercises.

Exercise. Ocular exercise –looking near and far with our eyes We can experience the sizzling noise of the muscles of the eye.

Exercise. When we keep closing our eyes quickly and strongly, we can hear some kind of sizzling noise Prrrrrrrr when compressing our eyelids. This noise can be heard for some seconds then it stops. The reason is: the palpebral muscle and the stirrup muscle belong to the facial nerve (nervus facialis).

The following movements of the eye are performed motorized by the brain when we focus on an object with both eyes.

Exercise. We close one of our eyes and look at the object with the open eye. We can feel the movement of the closed eye to the direction of the target object.

Another important movement of the eye is the saccadic following movement. In this case, we change the peripheral vision (where there is poor vision) for the fovea (where there is sharp vision).

Exercise. Watching a video on central and peripheral sight Peripheral vision is quicker and it helps us recognise our environment. Central vision is used for recognising objects [6].

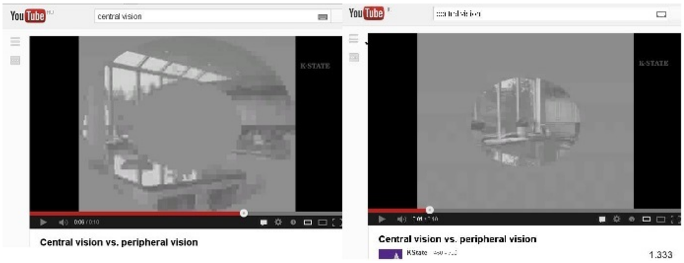

YouTube
central vision
K-STATE
0:06 / 0:10
Central vision vs. peripheral vision
YouTube
K-STATE
Central vision vs. peripheral vision
KState
1.333

Figure 2: Comparing central and peripheral vision

Getting to know the eyeball, the iris, the cornea and the crystelline lens is also given as homework.

Exercise. Students can see the clearness of the vitreous humour individually when they lie on the grass in sunny weather and look at the blue sky, taking care that they do not look into the sun. They can see small, fluffy things floating, of which they might think that it is their aura, but it is only the clarification process of the vitreous humour.

The slit in the middle of the iris, the diameter of the iris is moved reflex-like, that is unintentionally by the ocular muscles according to the strength of light coming into the eye [10, 29]. It tightens in strong light and widens in dim light, thus regulating the quantity of light falling on the retina.

Exercise. The narrowing of the pupil can be examined carefully with the help of a torch. Approaching the eye with a torch and removing the torch result in the narrowing and the widening of the pupil. We have shown the movement of the pupil on our website (`www.gerjak.hu`).

We can perceive other people's feelings and thoughts if we watch the widening of their pupils. Under the influence of joy, excitement or desire the pupil can widen even quadruple of its original state [11]. When we are angry or nervous, our eyes become piercing and our pupils become narrow.

Men and women can realise whether they are attractive to each other or not by watching each other's pupils. Children also look at grown-ups with their eyes wide open, thus expressing their admiration. Making use of it, models of beauty products and fashion design clothes are taken photos with wider pupils so that the products they advertise could be sold more easily [26]. That is also the reason why card-gamers wear glasses. They do not want their wide pupils to reveal their lucky run of the cards.

The eye is the most significant means of communication because of its central place and also because it functions beyond our control, this way providing us with accurate information about the other person. The brain performs the widening of the pupil automatically and interprets it. We can examine it with Eckhard Hess' pupil test [27].

When we have our students look at Figure 3 their pupils get constricted because their brains perceive it as if a pair of piercing eyes was looking at them.

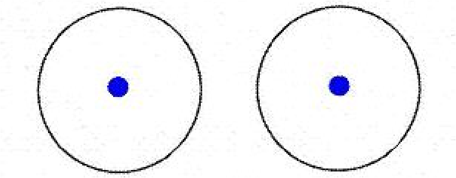


Figure 3: Eye illusion

Then when looking at Figure 4, their pupils widen because they perceive it as if a pair of attractive eyes was looking at them.

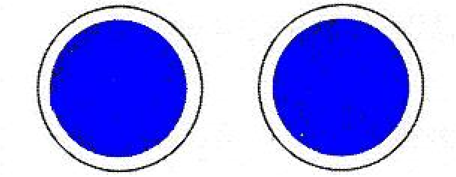


Figure 4: Eye illusion (2)

Light stimulating nerve nuclei in the brainstem affect pupil constricting muscles.

Exercise. As nerve fibres cross each other, when we shine light into the right eye, the pupil of the left eye gets constricted as well.

The shape of the lens changes according to distance [5]. When we look at a shorter distance, the ciliary muscle contracts, so the suspensory ligaments of the lens loosen and the lens, as it is flexible, becomes more convex. When we look at a greater distance, the ciliary muscle loosens and the lens becomes flat due to the tightening of the ligaments.

Here is a good exercise to move and accommodate the crystalline lens. Hold your fingers at two cms from each other and look at your fingers first and then look through them and focus a bit farther. (About two-three cms) The so-called sausage-effect occurs when we see a sausage between our two fingertips [28]. The reason is: we can see with both of our eyes and as we focus at a greater distance, our brain cannot decide whether it can really see two fingers or whether it should erase one of them. The topic of visual illusions is so rich that it deserves a separate study.

The eye can be seen as an optical low pass. We can perceive changes only to some extent [30].

Exercise. This proves that cut-off frequency decreases with nictating and closing the eyes quickly. Then only dark and light spots remain identifiable (see Figure 5).

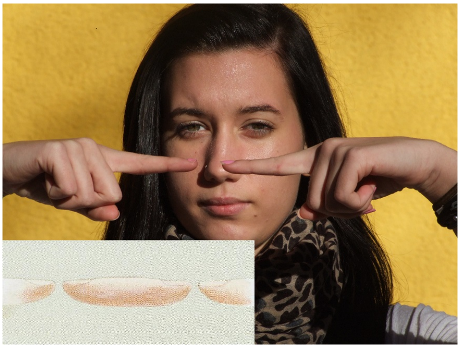


Figure 5: Sausage effect of the eye

3. New approach required in the teaching of sight

We perceive the world around us with our eyes and evaluate it with our brains. Out of the sense organs the eye is the quickest provider of data transfer. It has the longest perceiving distance and the best adaptability. It is capable of adapting to different light conditions, cold and heat, dryness and wetness.

We have to be 'wonderful' and 'interesting' in teaching as well if we want to become acquainted with this wonder of our body in its true aspect. Fortunately there are all kinds of IT and communication devices and we can also use other people's ideas, films and photos to expand our knowledge.

Students are susceptible to interesting things but then they are unable to listen attentively for a long time, so you have to arouse their interest again.

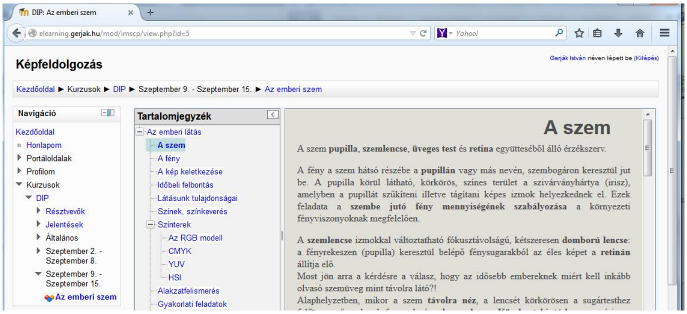
In secondary education it is very important that knowledge should be conveyed in a funny, entertaining way and possibly be accompanied by short animations and film projection.

After teaching our students the theory about the different parts of the eye during a lesson, we tried to attach some interesting stories or exercises to each part of our work on human eye to fix what we have learnt in the students' memories. Human brain can remember interesting and especially cathartic experiences much more easily.

Although in our next paper we will write in details about our tools, which assist teaching, we are presenting a Moodle screen picture from our material on the eye (see Figure 6).

4. Possible ways of the development of the eye

The development of our eyes has always been directed by evolution. It has developed in a certain direction if was advantageous for the preservation of our race or selection. There are different phases of development to prove this.



A szem

A szem **pupilla**, **szemlencse**, **üveges test** és **retina** együtteséből álló érzékszerv.

A fény a szem hátsó részébe a **pupillán** vagy más nevén, szembogáron keresztül jut be. A pupilla körül látható, körkörös, színes terület a szivárványhártya (irisz), amelyben a pupillát szűkíteni illetve tágítani képes izmok helyezkednek el. Ezek feladata a **szembe jutó fény mennyiségének szabályozása** a környezeti fényviszonyoknak megfelelően.

A **szemlencse** izmokkal változtatható fókusztávolságú, kétszeresen **domború lencse**: a fényrekeszen (pupilla) keresztül belépő fénysugarakból az éles képet a **retinán** állítja elő.
Most jön arra a kérdésre a válasz, hogy az idősebb embereknek miért kell inkább olvasó szemüveg mint távolra látó?!
Alaphelyzetben, mikor a szem **távolra néz**, a lencsét körkörösen a sugártesthez

Figure 6: Moodle website

Predatory animals and man have their eyes on the front because stereo sight helps to estimate distance and grab food. 'Prey animals' have their eyes on the sides of their heads, which enables them to see the world in 360 degrees, thus they can escape attacks coming from any direction [7]. Our eyes have only a 180 degree visual field but we can improve this by hunting in 'hordes'. This is what most predatory animals also do.

Seeing colours also serves this development. First we had to distinguish only between blue and green colours so that we could perceive the sky and water and also because we had leaf-eating habits similarly to monkeys. Later on we had to be able to perceive red to find ripe fruit.

The camouflage of insects is impossible in three-dimensional space, so they have developed depth perception [14]. So we can say when necessity arises, our body satisfies all our needs. This has been mostly the case for a couple of millions of years. What about our eyes? Will working a lot on the computer turn them from 3D eyes into 2D ones? Will we have to look at the world around us through special glasses as in 3D movies and also with some equipment, which can see in the dark?

Will young people become more passive in their community contacts and way of thinking because they start to use the computer too early? When we read we can imagine what we have read about while TV and computers provide us with everything. Future will tell but we should take care of the wonderful abilities of our eyes, which also contribute to the fact that man could become the masterpiece of creation or from another point of view the 'top predatory animal' in the world.

4.1. Our actions, which endanger our sight, protection of the eye, eye-defects and the ways of curing them

In a German survey 70% of the people working on the computer complained that their eyes got tired very soon and 65% of them complained of light sensitivity. 55% of the people asked had a stinging pain in their eyes. 32% of them complained about dry eyes and diminishing visual abilities. Almost 30% of them had moist eyes and a growing pressure in their eyes. 25% of them experienced red eyes [24].

We have no information about a similar survey in Hungary but teachers and students have the same complaints. What are these complaints caused by? Our eyes were made to see in 3D. Following movements, focusing, and accommodating, saccadic movement – all use certain parts of the eye and the body has prepared for them gradually of computers. Looking at the monitor fixedly, less nictating, steady focusing at 30-40 cms cause the tiredness of the ciliary muscles and the dryness of the eye.

Our eyes keep moving in their natural circumstances. During evolution hunting, searching for food and orientation required of our eyes to focus frequently at distant and close objects. We call it accommodation.

During work on the computer we look at a short distance permanently, thus our ciliary muscles are kept tensely for a long time, so they get tired. After work the lens often needs an hour to get back to normal. (When looking at short distances the ciliary muscles are tense, the zonular fibres are loose. When looking at long distances it is the other way round (see Figure 7).

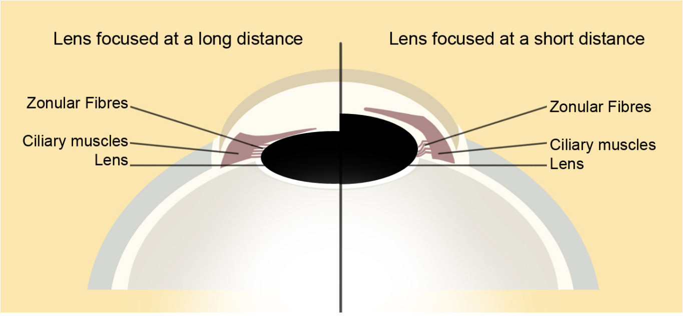


Figure 7: Accommodation of the lens

Our eyes have to adapt to different light conditions, as in the open air sunshine and shade, dusk and darkness, open and close spaces alternate with each other. Our eyes need to adapt to changes from dark to light and vice versa.

While working on the computer the amount of light, which falls into the eye changes constantly, light and dark surfaces alternate, so our eyes need constant adaptation, which is very tiring. Our eyes can be irritated by looking directly at the source of light or at a luminous surface.

We have to nictate 25-30 times a minute to clean and nourish our eyes. Thus our eyelids spread a layer of tears on our eyes. When we stare at the monitor permanently, we forget about nictating, so we do it only once or twice a minute. This way our eyes get dry and it causes a burning, itchy feeling. So Japanese scientists produced a pair of glasses, which becomes dark after not nictating for a long time and it becomes clear after nictating, thus getting computer users to nictate [13].

During work on the computer we overburden our spine, so burdening the second cervical vertebra (which is related to the functioning of the eyes) does harm to our sight as well. Sitting while working, pressing our stomach and staying almost motionless hinders the provision of our brains and sight with oxygen and blood.

As we have seen, working on the computer has an effect on the whole body. According to Dr.Arnold Gesell, 'The whole human body takes part in the process of sight', so we have to cure the whole body to improve our sight [24].

We can find health sanitary courses in foreign countries like Japan and Germany, which improve the sight of those people who do computer work. There are no similar initiatives known in Hungary. We should teach secondary school children how to protect their eyes and bodies from the dangers of computer work.

In the last five minutes of our information technology lessons at school we have exercises, which improve the students' eyesight and health.

There are three main areas of these exercises: practising eye movements (nictating), loosening muscles and soothing the nerval system.

Exercise 1. We look at a distant point with our eyes, then at out nose 10 times alternately

Exercise 2. neck-circle, then shoulder-circle 10 times

Exercise 3. In pairs we bend our backbones, one spondyle after the other forward, then back, while our pair carefully follows the movement on our spine with his/her hand.

5. Summary

In secondary school education we should make an effort to achieve that the material taught at different lessons should complement each other.

Our eyes are used in every lesson but their functioning is mentioned only in biology lessons, though very insufficiently and leaving out the functioning of the eye.

The learning process takes place more and more with the help of the Internet, so we are supposed to adapt to this tendency. During lessons it is recommended to follow the order presented in this work. First we should tell our students to find information on the functioning of the eye and use the proper websites, including ours (see Chapter 2). Then as it is shown in Chapter 2, we should commit what we have taught in their memories through experiments.

In the next lesson we should revise everything that we have taught about the eye and ask our students if they can find a connection between work on the computer and the way we use our eyes (with special regard to what we have written in Chapter 4). We should try hard to get our students to notice any possible eye problems and the ways of prevention and cure.

This article has presented only tutorial topics and we have touched upon the methodology of teaching. We are planning to write about the results in our forthcoming articles.

We cannot talk about everything within the framework of one article. Our aim was to demonstrate that within the scope of digital image processing the topic of human eye and sight has provided us with means, which teachers would never have dreamt of earlier. It is up to us to what extent and how much we can get our students involved and how to make use of it.

References

[1] A képzelet világa 1-2. Imrehné Sebestyén Margit. OM Kerettanterv 17/2004.(V.20.) Apáczai Kiadó., 2011.

[2] Balázs Rozália: Természetes nézés (Natural looking). 2015. Sites.google.com http://sites.google.com/site/balazsrozalia/home/gygymdok-terpik/termszetes-nzs (letöltve:2015.05.25)

[3] Bates W. H.: Tökéletes látás szemüveg nélkül (Perfect vision without glasses). http://www.bates-osszes.hupont.hu/2/tokeletes-latas-szemuveg-nelkul (letöltés: 2015.05.25.)

[4] Benjamin Harry: Tökéletes látás szemüveg nélkül (Perfect vision without glasses). Bioenergetic Kiadó Kft. 2012. ISBN 9789632911410

[5] Berke József, Hegedűs Gy.Gyula, Kelemen Dezső, Szabó József: Digitális képfeldolgozás és alkalmazásai (Digital image processing and applications). LSI, 14–16.

[6] Biológia 10., Berger Józsefné. Prizma Könyvek. Nemzeti Tankönyvkiadó Zrt., 2006. 79–80.

[7] Blakemore, Gombrich, Gregory: Illúzió a természetben és a művészetben (Illusion in nature and art). Budapest. 11., 23., 24., 32., 45–46., 93. oldal

[8] Comenius, J. A.: A látható világ (Orbis pictus). Magyar Helikon Kk., 1959/ Az 1669-ben megjelent kiadás alapján.

[9] Fizika 10., Medgyes Sándorné. Prizma Könyvek. Nemzeti Tankönyvkiadó Zrt., 2006. 16. oldal

[10] Greenfield, S. A.: Utazás az agy körül (The Human Brain). VILÁG-EGYETEM, Kulturtrade Kiadó Kft., 1998. ISBN 963 9069 65 5. 104. oldal

[11] Greenfield, S. A.: Utazás az agy körül (The Human Brain). VILÁG-EGYETEM, Kulturtrade Kiadó Kft., 1998. ISBN 963 9069 65 5. 105. oldal

[12] Holzinger, Andreas: A multimédia alapjai (The basics of multimedia). Kiskapu Kft., 2004. ISBN 963 9301 71. 7. oldal

[13] hvg.hu/tudomany/20091027

[14] Julesz Béla: Dialógusok az észlelésről (Dialogues about perception). Typotex Kiadó, Budapest, 2000. 88. oldal

[15] Kiss Gábor: A Magyar és a nemzetközi informatikaoktatás összehasonlítása (Comparison of Hungarian and International Informatics Education). PhD értekezés, Debreceni Egyetem, 2012. 15. oldal

[16] Kraus J. (Gordana Stankov: Konkrét és kepi reprezentációk használata a hetedik osztályos algebratanításban (CONCRETE AND IMAGE REPRESENTATIONS USE THE SECOND CLASS ALGEBRA IN TEACHING). PhD értekezés, Debreceni Egyetem, 2008. 6. oldal.

[17] NAT 2012.In: A Kormány 110/2012. (VI.4.) Korm. Rendelete a Nemzeti alaptanterv kiadásáról, bevezetéséről és alkalmazásáról,*Magyar Közlöny 2012.* 66. sz. Page:10654-10655.

[18] NAT 2012.In: A Kormány 110/2012. (VI.4.) Korm. Rendelete a Nemzeti alaptanterv kiadásáról, bevezetéséről és alkalmazásáról,*Magyar Közlöny 2012.* 66. sz. Page:10640.

[19] NAT 2012.In: A Kormány 110/2012. (VI.4.) Korm. Rendelete a Nemzeti alaptanterv kiadásáról, bevezetéséről és alkalmazásáról,*Magyar Közlöny 2012.* 66. sz. Page:10641.

[20] NAT 2012.In: A Kormány 110/2012. (VI.4.) Korm. Rendelete a Nemzeti alaptanterv kiadásáról, bevezetéséről és alkalmazásáról,*Magyar Közlöny 2012.* 66. sz. Page:10644.

[21] NAT 2012.In: A Kormány 110/2012. (VI.4.) Korm. Rendelete a Nemzeti alaptanterv kiadásáról, bevezetéséről és alkalmazásáról,*Magyar Közlöny 2012.* 66. sz. Page:10657.

[22] NAT 2012.In: A Kormány 110/2012. (VI.4.) Korm. Rendelete a Nemzeti alaptanterv kiadásáról, bevezetéséről és alkalmazásáról,*Magyar Közlöny 2012.* 66. sz. Page:10645.

[23] NAT 2012.In: A Kormány 110/2012. (VI.4.) Korm. Rendelete a Nemzeti alaptanterv kiadásáról, bevezetéséről és alkalmazásáról,*Magyar Közlöny 2012.* 66. sz. Page:10648-10649.

[24] Ostermeier, Uschi-Sitkowski:Szemtorna a számítógépnél (Eyes exercise at the computer). M-érték Könyvkiadó Kft., Budapest, 2003.

[25] Ozsváth Mária dr.:Szemtréning 2015. `http://www.alternativszemdr.hu/indexphp?modul=szemtrening-iriszdiagnosztika` (letöltés:2015.05.25)

[26] Pease, B. és A.: A testbeszéd enciklopédiája (The Definitive Book of Body Language). Park Könyvkiadó, Budapest 2008. 153. oldal

[27] Pease, B. és A.: A testbeszéd enciklopédiája (The Definitive Book of Body Language). Park Könyvkiadó, Budapest 2008. 155. oldal

[28] Reader's Digest: A lélek és az értelem ABC-je (ABC of the soul and the mind). London 1992 ISBN 963 8475 13 7.

[29] Smith, M.: Az emberi test (The Human Body). Szalay Könyvek. 33. oldal

[30] Steinmetz, R.: Multimédia (Multimedia). Springer Hungarica Kiadó Kft. Budapest, 1995. ISBN 963 8455 79 9. 55. oldal

Annales Mathematicae et Informaticae
47 (2017) pp. 243–251
http://ami.uni-eszterhazy.hu

The development of mathematical competences in Hungarian teacher training education*

Szilvia Petz[a], Miklós Hoffmann[b]

[a]Széchenyi István University, Apáczai Csere János Faculty
Győr, Hungary
petztiborne@gmail.com,

[b]Institute of Mathematics and Computer Science, Eszterházy Károly University
Eger, Hungary
hoffmann.miklos@uni-eszterhazy.hu

Submitted May 18, 2017 — Accepted November 30, 2017

Abstract

In this paper we present research results on the assessment of K1-4 (pupils from age 7 to age 10) teacher training student's mathematical knowledge and competences as one of the most important parameters of school teaching quality. Teachers' abilities in grades K1-4 are among the most important in-school factors influencing the quality of pupils' learning. A large-scale longitudinal study was conducted in which the elementary mathematical knowledge and skills of a group of teacher training students from 5 different institutes was assessed by means of a paper and pencil test that was administered both at the beginning and at the end of their second year mathematical course in the 2016/2017 academic year. This course is a methodical course in some institutions, which has an essential influence on the final results of this research. The 27-item-test covered the new standards for mathematics in the K1-4 elementary school curriculum. We have observed that those students coming from institutions providing separate methodical courses can gain better knowledge in explaining simple mathematical relations and notions than

*This research was supported by the grant EFOP-3.6.1-16-2016-00001 (Complex improvement of research capacities and services at Eszterházy Károly University).

those students whose institutions do not provide methodical background in specific courses.

Keywords: mathematical competences, teacher training

MSC: 97D40, 00A35, 97D60

1. Introduction and problem statement

Over the past three decades, higher education in OECD countries has changed profoundly. Not only has participation soared but student populations have become much more diverse. In response, systems have expanded and new providers with new offerings have emerged. This long period of expansion has distracted attention from the actual outcomes of higher education, but OECD countries are now looking more closely at how to ensure quality in education. As pointed out by the OECD teachers' review in 2005 (see [4]), education systems need to invest in intensive teacher education and training if teachers are expected to deliver high-quality outcomes. This also refers to the ECEC sector [5]: specific knowledge, skills and competencies are expected of ECEC practitioners. What does it mean to be a competent mathematics teacher? The answer is not simple: competent mathematics teacher has been trained in mathematics and receives some additional pedagogical and didactical training. Furthermore, teachers have to learn mathematics in ways that are specifically focused on teaching at a certain level, which is called "pedagogical content knowledge" by Shulman [3]. To possess mathematical competences means having knowledge of understanding, doing, using, and having a well-founded opinion about mathematics in variety of situation and contexts where mathematics plays or can play a role. Niss identified eight main constituents in that competence in [1], each of which is called a mathematical competence:

- The ability to ask and answer questions in and with mathematics:
  - Mathematical thinking competence: mastering mathematical modes of thought
  - Problem handling competence: formulating and solving mathematical problems
  - Modelling competence: being able to analyse and build mathematical models
  - Reasoning competence: being able to reason mathematically
- The ability to deal with mathematical language and tools:
  - Representation competence: being able to handle different representations of mathematical entities
  - Symbol and formalism competence: being able to handle symbol language and mathematical formalism

- Communication competence: being able to communicate in, with and about mathematics
- Aids and tools competence: being able to make, use of and relate to aids and tools of mathematics.

A competent mathematics teacher is someone who is able to help his or her students in an effective and efficient way to build and develop their mathematical competencies. A competent mathematics teacher must be mathematically and methodically competent as well in the outlined sense. This study tries to explore to what extent the teacher training students possess the mathematical competences mentioned above.

The following three domains received a primary focus in our research programme: knowledge of subject matter, pedagogical content knowledge, and knowledge of learners' cognition, based on the partition published by Shulman in [3] (see also: [2]). Applied to the domain of mathematics education, these three domains of knowledge can be described as follows.

First, subject-matter knowledge includes mastery of the key facts, concepts, principles and explanatory frameworks, procedures and problem solving techniques and strategies within the given domain of instruction. Crucial in this respect is also the level of teachers' understanding of the domain. The second category of teachers' knowledge can be defined as "knowledge of subject matter for teaching" ([3, p. 9]). It consists of an understanding of how to represent specific subject matter appropriately to the diverse abilities and interests of learners. It includes several issues, such as knowledge of mathematics lesson scripts and mathematics teaching routines, knowledge about various problem types, graphical representations, etc. that are best suited to introduce particular mathematical notions and skills to pupils. Furthermore, knowledge of instructional materials (textbooks, manipulatives, software, tests, etc.) available for teaching various mathematical topics is also essential. Third, there is teachers' knowledge of how students think and learn with respect to mathematics. This third component consists of the teachers' knowledge of the mathematical concepts and procedures that students bring to the learning of a topic, the misconceptions and buggy procedures that they may have developed, and the stages of understanding and skill that they are likely to pass through in the course of gaining mastery of it. In the Hungarian universities of teacher training the first years of training is primarily theoretical. The proportion of hours spent on theory decreases during the 4 years of the training, while gradually more time becomes available for practice. As far as mathematics education is concerned, the three major components of professional domain-specific knowledge discussed above (i.e., mathematical competence, pedagogical content knowledge, and knowledge of students) are typically addressed in one course. The undergraduate courses are followed by the Mathematics Methodology course, where students can learn how to teach each topic. However, there are institutions where there is no dedicated methodical course, but every course has a methodical aspect as well.

Overall, there are substantial differences between the institutes in terms of the relative proportion of instruction time that is devoted to each of these three components, the level of integration of these components, and, what exactly is being taught and learnt during this course. In Hungary there is no entrance exam or any other form of selection at the beginning of programs for higher education, including the training of elementary school teachers. Anyone who finished secondary school successfully and received his or her matriculation, can enter this teacher-training program. As a consequence, many students drop out during the first year of training or do not succeed in their exams. The low level of mathematical content knowledge and skills of students who want to become an elementary school teacher is increasingly being considered a major issue of concern among policy makers, curriculum developers, and teacher trainers involved in the training of future elementary school teachers. This growing concern was the major reason to set up this study.

2. The survey

A survey of mathematical competencies among K1-4 primary school teacher students of Hungarian teacher training institutions was held during the 2016/2017 academic year. Participants were 177 teacher training students who started mathematical methodological course (in those institutions where this course exists). These teacher students belonged to 5 institutions for teacher training. A paper-and-pencil mathematics test was administered to these 177 students during the first week of the academic year. At the end of the course a parallel version of this pretest was administered. The test was divided in six subsets differing in terms of the curricular subdomains and of the cognitive operations being addressed by the item.

2.1. Survey materials

The starting point for the construction of the mathematical competence pretest and posttest were the new standards for elementary education that have become operational in the Hungary since 2012. These standards cover all domains of the curriculum, including mathematics, and state the competencies that children should possess at the end of elementary school. The Hungarian curriculum standards for mathematics education are officially classified into different categories. Starting from this classification, we decided to divide the standards into six subdomains that were formed by combining a content and a cognitive dimension. The content dimension divided the mathematical content into two categories: Numbers and Arithmetics; Measurement and Geometry. Because more than half of the standards refers to the content domain of number and Arithmetics, we decided to combine the two other content domains (measurement and geometry) into one single domain. The cognitive dimension distinguished among three categories: Declarative knowledge, Procedural knowledge, and Strategic and problem solving skills. This resulted in a classification scheme consisting of six subdomains. It is worth noting, that none of the items required mathematical knowledge or skills beyond the content of the mathematics curriculum of the elementary school in Hungary. Nevertheless, the test contained several items that required good understanding of these elementary school mathematical notions and/or the application of problem solving strategies for using these mathematical notions in contextual problems. In a further stage of the project, we also constructed a parallel version of the first mathematics test. This parallel test—to be used as posttest—contained problems that were isomorphic to the problems from the pretest, but that were different in terms of superficial task characteristics (i.e., the concrete numbers used, the names of the persons and objects in the word problems, etc.).

2.2. Survey procedure

Shortly before the start of the academic year 2016/2017, copies of the pretest were sent to the 5 participating institutes, together with specific instructions on how the test had to be administered to the student teachers and how the completed forms had to be returned to the researchers. The pretest was administered in all institutes during the first week of the academic year 2016/2017. The administration of the test took 90 min. At the beginning of the pretest session the teacher trainer introduced the test and motivated the student teachers to do their best. At the same time, it was emphasized that the results would not be used for evaluative purposes within the context of their teacher training. At the end of the pretest session all copies were returned to the researchers who scored all test sheets according to strict criteria, leading to either 0-5 points for each of the 27 items. Most items were scored dichotomously on correctness of the answer. For the other items credit was given to partially correct answers. These could either refer to a correct response on a subset of questions or problems that were framed within one item, or to a partially correct response to a simple item (like when solving correctly the first step of a multi-step word problem). The organization and administration of the posttest was done in the same way at the same institutions as it happend in the case of the pretest.

2.3. Typical survey questions

Here we provide some of the examples that were included in the study tests.

- Numbers and Arithmetics – Declarative knowledge: What digit represents the tens, and what digit represents the thousands in the number 654,372?
- Numbers and Arithmetics – Procedural knowledge: Solve the operation! $3717 + 8635$
- Numbers and Arithmetics – Strategic and problem solving skills: In a container, there is 6845 l of oil, 5947 l more than in the barrel. How many liters of oil are in the barrel?
- Measurement and Geometry – Declarative knowledge: Are the following statements right or wrong? Explain your answer!

  - Every deltoid is square
  - Every square is a deltoid

- Measurement and Geometry – Procedural knowledge: What is the volume in cm^3 of a bottle of 50 dl?

- Measurement and Geometry – Strategic and problem solving skills: New carpeting of a room is planned. How long is the circumference when a room is 2 m 75 cm wide and 4 m 30 cm long? Make a drawing!

3. Results

Results of the pretest and the posttest are shown in Table 1 and 2 (results are given in percentage, because scoring of different parts may vary from part to part):

With respect to the results, although the tasks of the test cover the lower-level curriculum, there are still some items that do not reach 50%. The general comparison of the mean scores during the pretest and posttest reported in the two previous sections suggests that the mathematical method course had a significant and beneficial impact on the student teachers' competence in elementary mathematics. To summarize the results of the test it can be said that the knowledge of the teacher training's students is limited and uncertain. The results confirmed the frequently heard concern that at the beginning of their course students have rather weak mathematical competencies. At the end of their mathematical (methodological) course, the overall test performance had become substantially better, although there were still reasons to be seriously concerned about the readiness of some student teachers to teach mathematics to elementary school children.

Although not presented in terms of percentages, it is clear from the study, that those students who attended in the methodical course perform better in the posttest comparing with those ones whose institution has no specific methodical course. This very fact underlines the absolute necessity of methodical studies, not only from the pedagogical viewpoint, but also from scientific point of view - it seems to be an essential part of the curriculum to improve the knowledge of teacher training students in terms of basic mathematical notions and elementary strategic thinking. The authors think that every teacher training institutions must include methodical courses in their curriculum.

Unfortunately, the increasing of the number of students in higher education yields the consequence that more and more young people can be admitted from those ones who can not comply with the minimum requirements or can do it only in a very difficult way. This will effect on their work in the future where they will be uncertain and in worse case they will teach the next generation badly and faulty. As a result, the work in the lessons will be also irregularly which will be noticed by the students, too. The other source of the problem can probably be found in public education: students must learn a lot of material, but the world is changing and the attitude "I can get everything easily and only with little effort" does not

		Declarative	Procedural	Strategic
Numbers and Arithmetics	mean	61.0	63.3	47.8
	st.dev.	21.5	19.8	23.5
Measure and Geometry	mean	46.5	56.8	31.7
	st.dev.	24.1	29.0	28.2

Table 1: Pretest results (percentages) in terms of Declarative knowledge, Procedural knowledge and Strategic and problem solving skills in the two fields of mathematics

		Declarative	Procedural	Strategic
Numbers and Arithmetics	mean	70.1	68.8	60.2
	st.dev.	20.1	18.1	20.8
Measure and Geometry	mean	54.8	64.7	41.0
	st.dev.	20.9	25.3	26.4

Table 2: Posttest results (percentages) in terms of Declarative knowledge, Procedural knowledge and Strategic and problem solving skills in the two fields of mathematics

help the education of mathematics where you need precise knowledge of notions, strategies and a lot of practicing. Competence of mathematics is very important for the following generations. Problems from the real life can be used also in other sciences: the purpose and task of mathematics' teaching acquaints students with the concrete environment relations of quantitative and spatial circumstances that are surrounding them. Establishing their modern mathematical literacy therefore is of utmost importance, which makes them able to apply and develop mathematical thinking. Particular attention should be paid to the development and improvement of primary concepts, which should include various activities. Mathematics as a profession is to develop self-awareness of a starting experience, to improve independent thinking needs, to describe the joy of problem-solving and to develop positive personality traits. Some of the mathematical knowledge is abstract and a significant part is still connected to a specific knowledge. But emphasis should be placed on the diversity of activities to raise awareness of the experience, to record different ways, interpretation and systematization to search of correlations.

4. Discussion and concluding remarks

Starting from the state-of-the-art in the international research literature on preservice and in-service teachers' insufficient mastery in one of the major components of their domain-specific professional competence, namely their mastery of the content to be taught to their students, and its relationship with classroom practice, a longitudinal study was set up in which we assessed the elementary mathematical content knowledge and skills of a large group of Hungarian teacher training's students at the beginning and at the end of their methodical studies. Taking into account the Hungarian standards for elementary school mathematics, a pretest and a parallel posttest were constructed consisting of 27 items divided in six subtests, representing the major categories of these standards. Although none of the items required mathematical knowledge or skills beyond the content of the mathematics curriculum for the elementary school in Hungary, the test contained several items that demanded a thorough understanding of certain mathematical notions and/or the application of problem-solving strategies for using these mathematical notions in context problems. The results of the pretest confirmed the frequently heard concerns about the problematic level of mathematical competence of students who want to become an elementary school teacher given the low overall mean score as well as the detailed results for some very difficult items and for some very low performing subjects. The comparison of the actual mean pretest score and the score predicted by the teacher trainers indicated that Hungarian teacher trainers certainly do not underestimate the weakness of the mathematical content knowledge of their incoming students. Although the posttest results were considerably better than those for the pretest, the overall mean score was still very low. The design of the present study does not allow a more fine tuned analysis of the relative contribution of the instruction factor, and even leaves open the possibility that other factors besides this contributed to the observed gain in test scores between pretest and posttest. But we can definitely say that the methodological course help with the development of the students' mathematical knowledge and the correction of the wrong rooted concepts. There is an opportunity to the substitution of the missing knowledge, too. Meanwhile the methodological culture of the students is improving as well. The substantial differences in test score gain from pretest to posttest between the 5 institutes for teaching training that participated in the study suggest that these institutes were almost equally successful in developing the elementary mathematical competencies of their student teachers. But it is evident from the details of test results and answers, that students from those institutions, who provide separate methodical courses, can gain more well-established knowledge in explaining simple mathematical relationships and notions, than those students coming from institutions without separate methodical courses.

Additional research is needed to further unravel what characteristics of the teacher-training program in general, and of the specific mathematics (education) courses in particular are decisive for the development of the elementary mathematical competence of preservice teachers. Beside documenting the development of mathematical content knowledge and skills of preservice elementary school teachers in Hungary, the present study also resulted in two parallel versions of an instrument that is useful for the (self-) assessment of student teachers' mastery of the mathematical content they will have to teach after their graduation. The test as a whole proved to be a valuable instrument to assess the entrance level and the progress of mathematical content knowledge of our students, or to assist our student teachers in the self-assessment of (the development of) that level.

Currently, we are planning a follow-up study aimed at the development of a computer based instrument for continuous (self-)assessment of the mathematical content knowledge and skills of preservice teachers. Research evidence suggests that effective mathematics instruction involves the use of a variety of teaching methods. At the same time, there is general agreement that certain methods such as problem-based learning, investigation and contextualisation are particularly effective for raising achievement and improving students' attitudes toward mathematics.

Acknowledgement. The authors wish to thank the teacher trainers and students of all teacher training institutions involved in the study for their help in data providing and collection.

References

[1] Niss, M., Mathematical Competencies and the Learning of Mathematics: The Danish KOM Poject. In Gagatsis, A. and Papastavrides, S (eds): 3rd Mediteranean Conference on Mathematical Education Athen, Hellas 3-5 January 2003. Athens: Hellenic Mathematical Society, 2003, 115—124.

[2] De Corte, E., Greer, B. and Verschaffel, L., Mathematics teaching and learning. In D.C. Berliner and R. Calfee (Eds.), The handbook of educational psychology. New York, Macmillan, 1996, 490—549.

[3] Shulman, L., Those Who Understand: Knowledge Growth in Teaching, *Educational Researcher*, Vol. 15 (1986), 4-–14.

[4] `http://www.oecd.org/newsroom/34711139.pdf`, 21.04.2017

[5] `https://www.oecd.org/education/skills-beyond-school/37376068.pdf`, 21.04.2017