Contents

Research papers

Methodological papers

ANNALES MATHEMATICAE ET INFORMATICAE 50. (2019)

ANNALES MATHEMATICAE ET INFORMATICAE

TOMUS 50. (2019)

EGER 1774

ESZTERHÁZY KÁROLY UNIVERSITY

HUNGARIA, EGER

ANNALES MATHEMATICAE ET INFORMATICAE

VOLUME 50. (2019)

INSTITUTE OF MATHEMATICS AND INFORMATICS
ESZTERHÁZY KÁROLY UNIVERSITY
HUNGARY, EGER

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: 2019. december

Research papers

Annales Mathematicae et Informaticae
***50** (2019) pp. 5–17*
DOI: 10.33039/ami.2019.01.001
http://ami.uni-eszterhazy.hu

Surfaces with constant extrinsically Gaussian curvature in the Heisenberg group

Lakehal Belarbi

Department of Mathematics,
Laboratory of Pure and Applied Mathematics,
University of Mostaganem (U.M.A.B.), Mostaganem, Algeria
lakehalbelarbi@gmail.com

Submitted: September 19, 2017
Accepted: January 23, 2019
Published online: February 11, 2019

Abstract

In this work we study constant extrinsically Gaussian curvature translation surfaces in the 3-dimensional Heisenberg group which are invariant under the 1-parameter groups of isometries.

Keywords: Constant extrinsically Gaussian curvature Surfaces, Homogeneous group.

MSC: 49Q20 53C22.

1. Introduction

In 1982, W. P. Thurston formulated a geometric conjecture for three dimensional manifolds, namely every compact orientable 3-manifold admits a canonical decomposition into pieces, each of them having a canonical geometric structure from the following eight maximal and simply connected homogenous Riemannian spaces: $\mathbb{E}^3$, $\mathbb{S}^3$, $\mathbb{H}^3$, $\mathbb{S}^2 \times \mathbb{R}$, $\mathbb{H}^2 \times \mathbb{R}$, $SL(2, \mathbb{R})$, $\mathbb{H}_3$ and Sol_3. See e.g. [34].

During the recent years, there has been a rapidly growing interest in the geometry of surfaces in three homogenous spaces focusing on flat and constant Gaussian curvature surfaces. Many works are studying the geometry of surfaces in homogeneous 3-manifolds. See for example [2–4, 9, 12, 14–16, 21, 22, 24, 36].

The concept of translation surfaces in $\mathbb{R}^3$ can be generalized the surfaces in the three dimensional Lie group, in particular, homogeneous manifolds. In Euclidean

3-space, every cylinder is flat. Conversely, complete flat surfaces in $\mathbb{E}^3$ are cylinders over complete curves. See [20]. López and Munteanu [17] studied invariant surfaces with constant mean curvature and constant Gaussian curvature in Sol_3 space. Yoon and Lee [37] studied translation surfaces in Heisenberg group $\mathbb{H}_3$ whose position vector x satisfies the equation $\Delta x = Ax$, where Δ is the Laplacian operator of the surface and A is a 3×3-real matrix.

Flat G_4-invariant surfaces are nothing but surfaces invariant under $SO(2)$-action, i.e. rotational surfaces. Flat rotational surfaces are classified by Caddeo, Piu and Ratto in [8].

In [14], J. I. Inoguchi give a classification of intrinsically flat G_1-invariant translation surfaces in Heisenberg group $\mathbb{H}_3$. Let M be a surface invariant under G_3, then M is locally expressed as

$$X(u,v) = (0,0,v).(x(u),y(u),0) = (x(u),y(u),v),\ u \in I,\ v \in \mathbb{R}.$$

Here I is an open interval and u is the arclength parameter. Note that $(x,y,0)$ and $(0,0,v)$ commute. Then the sectional curvature $K(X_x \wedge X_y) = \frac{1}{4}$ and the extrinsically Gaussian curvature $K_{ext} = -\frac{1}{4}$. Direct computation show that M is flat. (cf. [12–14, 28]).

The paper is divided according the type of surfaces invariant under 1-parameter subgroups of isometries $\{G_i\}_{i=1,2,3,4}$. So, in section 3 we classify G_1-invariant surfaces of the Heisenberg group $\mathbb{H}_3$ with constant extrinsically Gaussian curvature K_{ext}, including extrinsically flat G_1-invarinant surfaces.

In section 4 we classify G_2-invariant surfaces of the Heisenberg group $\mathbb{H}_3$ with constant extrinsically Gaussian curvature K_{ext}, including extrinsically flat G_2-invariant surfaces.

2. Preliminaries

The 3-dimensional Heisenberg group $\mathbb{H}_3$ is the simply connected and connected 2-step nilpotent Lie group. Which has the following standard representation in $GL(3,\mathbb{R})$

$$\begin{pmatrix} 1 & r & t \\ 0 & 1 & s \\ 0 & 0 & 1 \end{pmatrix}$$

with $r,s,t \in \mathbb{R}$. The Lie algebra $\mathfrak{h}_3$ of $\mathbb{H}_3$ is given by the matrices

$$A = \begin{pmatrix} 0 & x & z \\ 0 & 0 & y \\ 0 & 0 & 0 \end{pmatrix}$$

with $x, y, z \in \mathbb{R}$.The exponential map $exp : \mathfrak{h}_3 \to \mathbb{H}_3$ is a global diffeomorphism, and is given by

$$\exp(A) = I + A + \frac{A^2}{2} = \begin{pmatrix} 1 & x & z + \frac{xy}{2} \\ 0 & 1 & y \\ 0 & 0 & 1 \end{pmatrix}.$$

The Heisenberg group $\mathbb{H}_3$ is represented as the cartesian 3-space $\mathbb{R}^3(x, y, z)$ with group structure:

$$(x_1, y_1, z_1).(x_2, y_2, z_2) := \left(x_1 + x_2, y_1 + y_2, z_1 + z_2 + \frac{1}{2}x_1y_2 - \frac{1}{2}x_2y_1 \right).$$

We equip $\mathbb{H}_3$ with the following left invariant Riemannian metric

$$g := dx^2 + dy^2 + \left(dz + \frac{1}{2}(ydx - xdy) \right)^2.$$

The identity component $I^\circ(\mathbb{H}_3)$ of the full isometry group of $(\mathbb{H}_3, g)$ is the semi-direct product $SO(2) \ltimes \mathbb{H}_3$. The action of $SO(2) \ltimes \mathbb{H}_3$ is given explicitly by

$$\begin{aligned} A &= \left(\begin{bmatrix} \cos\theta & -\sin\theta \\ \sin\theta & \cos\theta \end{bmatrix} . \begin{bmatrix} a \\ b \\ c \end{bmatrix} \right) . \begin{bmatrix} x \\ y \\ z \end{bmatrix} \\ &= \begin{bmatrix} \cos\theta & -\sin\theta & 0 \\ \sin\theta & \cos\theta & 0 \\ \frac{1}{2}(a\sin\theta - b\cos\theta) & \frac{1}{2}(a\cos\theta + b\sin\theta) & 1 \end{bmatrix} . \begin{bmatrix} x \\ y \\ z \end{bmatrix} + \begin{bmatrix} a \\ b \\ c \end{bmatrix}. \end{aligned}$$

In particular, rotational around the z-axis and translations:

$$(x, y, z) \to (x, y, z + a), a \in \mathbb{R}$$

along the z-axis are isometries of $\mathbb{H}_3$.

The Lie algebra $\mathfrak{h}_3$ of $I^\circ(\mathbb{H}_3)$ is generated by the following Killing vector fields:

$$F_1 = \frac{\partial}{\partial x} + \frac{y}{2}\frac{\partial}{\partial z}, \;\; F_2 = \frac{\partial}{\partial y} - \frac{x}{2}\frac{\partial}{\partial z},$$

$$F_3 = \frac{\partial}{\partial z}, \;\; F_4 = -y\frac{\partial}{\partial x} + x\frac{\partial}{\partial y}.$$

One can check that F_1, F_2, F_3 are infinitesimal transformations of the 1-parameter groups of isometries defined by

$$G_1 = \{(t, 0, 0) | t \in \mathbb{R}\}, \;\; G_2 = \{(0, t, 0) | t \in \mathbb{R}\}, \;\; G_3 = \{(0, 0, t) | t \in \mathbb{R}\},$$

respectively. Here this groups acts on $\mathbb{H}_3$ by the left translation. The vector field F_4 generates the group of rotations around the z-axis. Thus G_4 is identified with $SO(2)$.

Definition 2.1. A surface Σ in the Heisenberg space $\mathbb{H}_3$ is said to be invariant surface if it is invariant under the action of the 1-parameter subgroups of isometries $\{G_i\}$, with $i \in \{1,2,3,4\}$.

The Lie algebra $\mathfrak{h}_3$ of $\mathbb{H}_3$ has an orthonormal basis $\{E_1, E_2, E_3\}$ defined by

$$E_1 = \frac{\partial}{\partial x} - \frac{y}{2}\frac{\partial}{\partial z},\ E_2 = \frac{\partial}{\partial y} + \frac{x}{2}\frac{\partial}{\partial z},\ E_3 = \frac{\partial}{\partial z}.$$

The Levi-Civita connection ∇ of g,in terms of the basis $\{E_i\}_{i=1,2,3}$ is explicitly given as follows

$$\begin{cases} \nabla_{E_1}E_1 = 0, \nabla_{E_1}E_2 = \frac{1}{2}E_3, \nabla_{E_1}E_3 = -\frac{1}{2}E_2 \\ \nabla_{E_2}E_1 = -\frac{1}{2}E_3, \nabla_{E_2}E_2 = 0, \nabla_{E_2}E_3 = \frac{1}{2}E_1 \\ \nabla_{E_3}E_1 = -\frac{1}{2}E_2, \nabla_{E_3}E_2 = \frac{1}{2}E_1, \nabla_{E_3}E_3 = 0 \end{cases}$$

The Riemannian curvature tensor R is a tensor field on $\mathbb{H}_3$ defined by

$$R(X,Y)Z = \nabla_X\nabla_Y Z - \nabla_Y\nabla_X Z - \nabla_{[X,Y]}Z.$$

The components $\{R^l_{ijk}\}$ are computed as

$$R^1_{212} = -\frac{3}{4},\ R^1_{313} = \frac{1}{4},\ R^2_{323} = \frac{1}{4}.$$

Let us denote $K_{ij} = K(E_i, E_j)$ the sectional curvature of the plane spanned by E_i and E_j.Then we get easily the following:

$$K_{12} = -\frac{3}{4},\ K_{13} = -\frac{1}{4},\ K_{23} = -\frac{1}{4}.$$

The Ricci curvature Ric is defined by

$$Ric(X,Y) = trace\{Z \to R(Z,X)Y\}.$$

The components $\{R_{ij}\}$ of Ric are defined by

$$Ric(E_i, E_j) = R_{ij} = \sum_{k=1}^{3} \langle R(E_i, E_k)E_k, E_j\rangle .$$

The components $\{R_{ij}\}$ are computed as

$$R_{11} = -\frac{1}{2},\ R_{12} = R_{13} = R_{23} = 0,\ R_{22} = -\frac{1}{2},\ R_{33} = \frac{1}{2}.$$

The scalar curvature S of $\mathbb{H}_3$ is constant and we have

$$S = trRic = \sum_{i=1}^{3} Ric(E_i, E_i) = -\frac{1}{2}.$$

3. Constant extrinsically Gaussian curvature G_1-invariant translation surfaces in Heisenberg group $\mathbb{H}_3$

3.1.

In this subsection we study complete extrinsically flat translation surfaces Σ in Heisenberg group $\mathbb{H}_3$ which are invariant under the one parameter subgroup G_1. Clearly, such a surface is generated by a curve γ in the totally geodesic plane $\{x = 0\}$. Discarding the trivial case of a vertical plane $\{y = y_0\}$. Thus γ is given by $\gamma(y) = (0, y, v(y))$. Therefore the generated surface is parameterized by

$$X(x, y) = (x, 0, 0).(0, y, v(y)) = (x, y, v(y) + \frac{xy}{2}),\ (x, y) \in \mathbb{R}^2.$$

We have an orthogonal pair of vector fields on (Σ), namely,

$$e_1 := X_x = (1, 0, \frac{y}{2}) = E_1 + yE_3.$$

and

$$e_2 := X_y = (0, 1, v' + \frac{x}{2}) = E_2 + v'E_3.$$

The coefficients of the first fundamental form are:

$$E = \langle e_1, e_1\rangle = 1 + y^2,\ F = \langle e_1, e_2\rangle = yv',\ G = \langle e_2, e\rangle = 1 + v'^2.$$

As a unit normal field we can take

$$N = \frac{-y}{\sqrt{1 + y^2 + v'^2}}E_1 - \frac{v'}{\sqrt{1 + y^2 + v'^2}}E_2 + \frac{1}{\sqrt{1 + y^2 + v'^2}}E_3$$

The covariant derivatives are

$$\widetilde{\nabla}_{e_1}e_1 = -yE_2$$

$$\widetilde{\nabla}_{e_1}e_2 = \frac{y}{2}E_1 - \frac{v'}{2}E_2 + \frac{1}{2}E_3$$

$$\widetilde{\nabla}_{e_2}e_2 = v'E_1 + v''E_3.$$

The coefficients of the second fundamental form are

$$l = \langle\widetilde{\nabla}_{e_1}e_1, N\rangle = \frac{yv'}{\sqrt{1 + y^2 + v'^2}}$$

$$m = \langle\widetilde{\nabla}_{e_1}e_2, N\rangle = \frac{-\frac{y^2}{2} + \frac{v'^2}{2} + \frac{1}{2}}{\sqrt{1 + y^2 + v'^2}}$$

$$n = \langle\widetilde{\nabla}_{e_2}e_2, N\rangle = \frac{-yv' + v''}{\sqrt{1 + y^2 + v'^2}}.$$

Let K_{ext} be the extrinsic Gauss curvature of Σ,

$$K_{ext} = \frac{ln - m^2}{EG - F^2} = \frac{-y^2v'^2 + yv'v'' - (-\frac{y^2}{2} + \frac{v'^2}{2} + \frac{1}{2})^2}{(1 + y^2 + v'^2)^2}.$$

Thus Σ is extrinsically flat invariant surface in Heisenberg group $\mathbb{H}_3$ if and only if

$$K_{ext} = 0,$$

that is, if and only if

$$-y^2v'^2 + yv'v'' - \left(-\frac{y^2}{2} + \frac{v'^2}{2} + \frac{1}{2}\right)^2 = 0 \tag{3.1}$$

to classify extrinsically flat invariant surfaces must solve the equation (3.1). We can writes equation (3.1) as

$$y^2 + yv'v'' - \left(\frac{y^2}{2} + \frac{v'^2}{2} + \frac{1}{2}\right)^2 = 0 \tag{3.2}$$

we assume that $z = \frac{y^2}{2} + \frac{v'^2}{2} + \frac{1}{2}$. Then

$$\begin{cases} z' = y + v'v'' \\ v'v'' = z' - y \\ v'^2 = 2z - y^2 - 1. \end{cases} \tag{3.3}$$

Therefore equation (3.2) becomes

$$yz' - z^2 = 0. \tag{3.4}$$

equation (3.4) implies that

$$-\frac{z'}{z^2} = -\frac{1}{y}. \tag{3.5}$$

and equation (3.5) implies that

$$z = \frac{1}{-\ln(y) + \alpha}. \tag{3.6}$$

where $\alpha \in \mathbb{R}$, and if $y \neq e^{\alpha}$.

From (3.3) and (3.6), we have

$$\begin{aligned} v'^2 &= 2z - y^2 - 1 \\ &= \frac{2}{-\ln(y) + \alpha} - y^2 - 1. \end{aligned}$$

Thus

$$v' = \sqrt{\frac{2}{-\ln(y) + \alpha} - y^2 - 1}.$$

As conclusion, we have

Theorem 3.1. • *The only non-extendable extrinsically flat translation surfaces in the 3-dimensional Heisenberg group* $\mathbb{H}_3$ *invariant under the 1-parameter subgroup* $G_1 = \{(t,0,0) \in \mathbb{H}_3/t \in \mathbb{R}\}$*, are the surfaces whose parametrization is* $X(x,y) = \left(x, y, v(y) + \frac{xy}{2}\right)$ *where* y *and* v *satisfy*

$$v(y) = \int \sqrt{\frac{2}{-\ln(y) + \alpha} - y^2 - 1} dy.$$

where $\alpha \in \mathbb{R}$*, and* $y \neq e^{\alpha}$.

• *There are no complete extrinsically flat translation surfaces in the 3-dimensional Heisenberg group* $\mathbb{H}_3$ *invariant under the 1-parameter subgroup* $G_1 = \{(t,0,0) \in \mathbb{H}_3/t \in \mathbb{R}\}$.

Remark 3.2. Let Σ be a G_1-invariant translation surfaces in the 3-dimensional Heisenberg space. Then Σ is locally expressed as

$$X(x,y) = (0,y,v(y)) \,.\, (x,0,0) = \left(x, y, v(y) - \frac{xy}{2}\right).$$

Then the extrinsically Gaussian curvature K_{ext} of Σ is computed as

$$K_{ext} = \frac{\left((v'-x)^2 - 1\right)^2}{4\left(1 + (v'-x)^2\right)^2}.$$

Thus Σ can not be of constant extrinsically Gaussian curvature.

3.2.

In this subsection we study complete constant extrinsically Gaussian curvature translation surfaces Σ in Heisenberg group $\mathbb{H}_3$ which are invariant under the one parameter subgroup G_1. Clearly, such a surface is generated by a curve γ in the totally geodesic plane $\{x = 0\}$. Discarding the trivial case of a vertical plane $\{y = y_0\}$. Thus γ is given by $\gamma(y) = (0, y, v(y))$. Therefore the generated surface is parameterized by

$$X(x,y) = (x,0,0).(0,y,v(y)) = (x, y, v(y) + \frac{xy}{2}),\ (x,y) \in \mathbb{R}^2.$$

Theorem 3.3. • *The* G_1*-invariant constant extrinsically Gaussian curvature translation surfaces in the 3-dimensional Heisenberg group* $\mathbb{H}_3$*, are:*

1. $K_{ext} = -\frac{1}{4}$.
   The surfaces of equation

$$z = v(y) + \frac{xy}{2} = \frac{xy}{2} + \frac{1}{2} y\sqrt{2\beta - y^2} + \arctan\left(\frac{y}{\sqrt{\beta - y^2}}\right),$$

   where $\beta \in \mathbb{R}$.

2. $K_{ext} \neq -\frac{1}{4}$.
Then y and v satisfy

$$v(y) = \int \sqrt{\frac{1}{-2(K_{ext}+\frac{1}{4})\ln(y)+\gamma} - y^2 - 1}dy.$$

where $\gamma \in \mathbb{R}$, and $y \neq e^{\frac{\gamma}{2(K_{ext}+\frac{1}{4})}}$.

• *There are no complete constant extrinsically Gaussian curvature translation surfaces in the 3-dimensional Heisenberg group $\mathbb{H}_3$ invariant under the 1-parameter subgroup G_1.*

Proof. From (4.1) and (3.2) we have

$$K_{ext} = \frac{ln - m^2}{EG - F^2} = \frac{y^2 + yv'v'' - \frac{1}{4}\left(1 + y^2 + v'^2\right)^2}{(1 + y^2 + v'^2)^2}. \tag{3.7}$$

1. If $K_{ext} = -\frac{1}{4}$. Then equation (3.7) becomes

$$y^2 + yv'v'' = 0 \tag{3.8}$$

We note that y equal zero is solution of the equation(3.8).
If y is different to zero ($y \neq 0$), equation (3.8) becomes

$$v'v'' = -y.$$

Integration gives us

$$v(y) = \frac{1}{2}y\sqrt{2\beta - y^2} + \arctan\left(\frac{y}{\sqrt{\beta - y^2}}\right),$$

where $\beta \in \mathbb{R}$.

2. If $K_{ext} \neq -\frac{1}{4}$. Then equation (3.7) becomes

$$y^2 + yv'v'' = (K_{ext} + \frac{1}{4})(1 + y^2 + v'^2)^2.$$

In fact, put $z = 1 + y^2 + v'^2$. Then z satisfies

$$\frac{1}{2}yz' = (K_{ext} + \frac{1}{4})z^2.$$

Hence we have

$$z = \frac{1}{-2(K_{ext} + \frac{1}{4})y + \gamma},$$

where $\gamma \in \mathbb{R}$, and $y \neq e^{\frac{\gamma}{2(K_{ext}+\frac{1}{4})}}$. Using the equation $z = 1 + y^2 + v'^2$, we get

$$v'^2 = \frac{1}{-2(K_{ext} + \frac{1}{4})y + \gamma} - y^2 - 1. \qquad \square$$

4. Constant extrinsically Gaussian curvature G_2-invariant translation surfaces in Heisenberg group $\mathbb{H}_3$

In this section we study constant complete extrinsically flat translation surfaces Σ in Heisenberg group $\mathbb{H}_3$ which are invariant under the one parameter subgroup G_2. Clearly, such a surface is generated by a curve γ in the totally geodesic plane $\{y = 0\}$. Discarding the trivial case of a vertical plane $\{x = x_0\}$. Thus γ is given by $\gamma(x) = (x, 0, f(x))$. Therefore the generated surface is parameterized by

$$X(x, y) = (0, y, 0).(x, 0, f(x)) = (x, y, f(x) - \frac{xy}{2}),\ (x, y) \in \mathbb{R}^2.$$

We have an orthogonal pair of vector fields on (Σ), namely,

$$e_1 := X_x = (1, 0, f' - \frac{y}{2}) = E_1 + f'E_3.$$

and

$$e_2 := X_y = (0, 1, -\frac{x}{2}) = E_2 - xE_3.$$

The coefficients of the first fundamental form are:

$$E = \langle e_1, e_1\rangle = 1 + f'^2,\ F = \langle e_1, e_2\rangle = -xf',\ G = \langle e_2, e\rangle = 1 + x^2.$$

As a unit normal field we can take

$$N = \frac{-f'}{\sqrt{1 + x^2 + f'^2}}E_1 + \frac{x}{\sqrt{1 + x^2 + f'^2}}E_2 + \frac{1}{\sqrt{1 + x^2 + f'^2}}E_3.$$

The covariant derivatives are

$$\widetilde{\nabla}_{e_1}e_1 = -f'E_2 + f''E_3$$

$$\widetilde{\nabla}_{e_1}e_2 = \frac{f'}{2}E_1 + \frac{x}{2}E_2 - \frac{1}{2}E_3$$

$$\widetilde{\nabla}_{e_2}e_2 = -xE_1.$$

The coefficients of the second fundamental form are

$$l = \langle\widetilde{\nabla}_{e_1}e_1, N\rangle = \frac{-xf' + f''}{\sqrt{1 + x^2 + f'^2}}$$

$$m = \langle\widetilde{\nabla}_{e_1}e_2, N\rangle = \frac{-\frac{f'^2}{2} + \frac{x^2}{2} - \frac{1}{2}}{\sqrt{1 + x^2 + f'^2}}$$

$$n = \langle\widetilde{\nabla}_{e_2}e_2, N\rangle = \frac{-yv' + v''}{\sqrt{1 + y^2 + v'^2}}.$$

Let K_{ext} be the extrinsic Gauss curvature of Σ,

$$K_{ext} = \frac{ln - m^2}{EG - F^2} = \frac{x^2 + xf'f'' - \frac{1}{4}(x^2 + f'^2 + 1)^2}{(1 + x^2 + f'^2)^2}. \tag{4.1}$$

Thus Σ is extrinsically flat invariant surface in Heisenberg group $\mathbb{H}_3$ if and only if

$$K_{ext} = 0,$$

that is, if and only if

$$x^2 + xf'f'' - \frac{1}{4}(x^2 + f'^2 + 1)^2 = 0. \tag{4.2}$$

to classify extrinsically flat invariant surfaces must solve the equation (4.2).

We remark that the equation (4.2) is similarly to the equation (3.1), It is sufficient to change y by x and v by f.

As conclusion, we have

Theorem 4.1. •*The only non-extendable extrinsically flat translation surfaces in the 3-dimensional Heisenberg group $\mathbb{H}_3$ invariant under the 2-parameter subgroup $G_2 = \{(0,t,0) \in \mathbb{H}_3 / t \in \mathbb{R}\}$, are the surfaces whose parametrization is $X(x,y) = \left(x, y, f(x) - \frac{xy}{2}\right)$ where x and f satisfy*

$$f(x) = \int \sqrt{\frac{2}{-\ln(x) + \alpha} - x^2 - 1} dy.$$

where $\alpha \in \mathbb{R}$, and $x \neq e^{\alpha}$.

•*There are no complete extrinsically flat translation surfaces in the 3-dimensional Heisenberg group $\mathbb{H}_3$ invariant under the 1-parameter subgroup $G_2 = \{(0,t,0) \in \mathbb{H}_3 / t \in \mathbb{R}\}$.*

Remark 4.2. Let Σ be a G_2-invariant translation surfaces in the 3-dimensional Heisenberg space. Then Σ is locally expressed as

$$X(x,y) = (x, 0, f(x)) \,.\, (0, y, 0) = \left(x, y, f(x) + \frac{xy}{2}\right).$$

Then the extrinsically Gaussian curvature K_{ext} of Σ is computed as

$$K_{ext} = -\frac{\left((f' + y)^2 - 1\right)^2}{4\left(1 + (v' - x)^2\right)^2}.$$

Thus Σ can not be of constant extrinsically Gaussian curvature.

Theorem 4.3. • *The G_2-invariant constant extrinsically Gaussian curvature translation surfaces in the 3-dimensional Heisenberg group $\mathbb{H}_3$, are:*

1. $K_{ext} = -\frac{1}{4}$.
   The surfaces of equation

$$z = f(x) - \frac{xy}{2} = -\frac{xy}{2} + \frac{1}{2}x\sqrt{2\beta - x^2} + \arctan\left(\frac{x}{\sqrt{\beta - x^2}}\right),$$

   where $\beta \in \mathbb{R}$.

2. $K_{ext} \neq -\frac{1}{4}$.
   Then x *and* f *satisfy*

$$f(x) = \int \sqrt{\frac{1}{-2(K_{ext} + \frac{1}{4})\ln(x) + \gamma} - x^2 - 1}dy.$$

   where $\gamma \in \mathbb{R}$, *and* $x \neq e^{\frac{\gamma}{2(K_{ext}+\frac{1}{4})}}$.

• *There are no complete constant extrinsically Gaussian curvature translation surfaces in the 3-dimensional Heisenberg group* $\mathbb{H}_3$ *invariant under the 1-parameter subgroup* G_2.

Acknowledgements. The author would like to thank the Referees for all helpful comments and suggestions that have improved the quality of our initial manuscript. The author would like also to thank Rabah Souam for their interest and helpful discussions and advice.

References

[1] U. Abresch, H. Rozenberg: *A Hopf differential for constant mean curvature surfaces in* $\mathbb{S}^2 \times \mathbb{R}$ *and* $\mathbb{H}^2 \times \mathbb{R}$, Acta Math 193 (2004), pp. 141–174.

[2] M. Bekkar: *Exemples de surfaces minimales dans l'espace de Heisenberg* $\mathbb{H}_3$, Rend.Sem. Mat. Univ. Cagliari 61 (1991), pp. 123–130.

[3] M. Bekkar, T. Sari: *Surfaces Minimales réglées dans l'espace de Heisenberg* $\mathbb{H}_3$, Rend. Sem. Mat. Univ. Politec. Torino 50.3 (1992), pp. 243–254.

[4] M. Belkhelfa: *Parallel and minimal surfaces in Heisenberg*, in: Proceedings of the Summer School on Differential Geometry, Coimbra (Portugal), September 3–7, 1999, pp. 67–76.

[5] D. Bensikaddour, L. Belarbi: *Minimal Translation Surfaces in Lorentz-Heiesenberg 3-Space*, Nolinear Studies 24.4 (2017), pp. 859–867.

[6] D. Bensikaddour, L. Belarbi: *Minimal Translation Surfaces in Lorentz-Heiesenberg 3-space with Flat Metric*, Differential Geometry-Dynamical Systems 20 (2018), pp. 1–14.

[7] F. Bonahon: *Geometric structures on 3-manifolds*, in: In Handbook of geometric topology, North-Holland, Amsterdam, 2002, pp. 93–164.

[8] R. Cadeo, P. Piu, A. Ratto: $SO(2)$*-invariant minimal and constant mean curvature surfaces in 3-dimensional homogeneous spaces*, Maniscripta Math 87 (1995), pp. 1–12.

[9] B. Daniel: *Isometric immersions into 3-dimensional homogeneous manifolds*, Comment. Math. Helv 82 (2007), pp. 87–131.

[10] B. Daniel, W. H. Meeks, H. Rosenberg: *Half-space theorems for minimal surfaces in Nil and Sol*, J. Differential Geometry 88 (2011), pp. 41–59.

[11] R. S. Erap, E. Toubiana: *Screw motion surfaces in* $\mathbb{H}^2 \times \mathbb{R}$ *and* $\mathbb{S}^2 \times \mathbb{R}$, Illinois J. Math 49 (2005), pp. 1323–1362.

[12] C. B. Figueroa, F. Mercuri, R. H. L. Pedrosa: *Invariant minimal surfaces of the Heisenberg groups*, Ann. Ital 7 (1999), pp. 173–194.

[13] J. Inogochi, T. Kumamoto, N. Ohsugi, Y. Suyama: *Differential geometry of curves and surfaces in 3-dimensional homogeneous speces I*, Fukuoka Univ. Sci. Rep 29 (1999), pp. 155–182.

[14] J. Inoguchi: *Flat translation surfaces in the 3-dimensional Heisenberg group*, J. Geom 82.1-2 (2005), pp. 83–90, doi: 10.1007/s00022-005-1730-1.

[15] J. Inoguchi, R. López, M. I. Munteanu: *Minimal translation surfaces in the Heisenberg group* Nil_3, Geom. Dedicata 161 (2012), pp. 221–231, doi: 10.1007/s10711-012-9702-8.

[16] R. López: *Constant mean curvature surfaces in Sol with non-empty boundary*, Houston J. Math 38.4 (2012), pp. 1091–1105.

[17] R. López, M. I. Munteanu: *Invariant surfaces in homogeneous space Sol with constant curvature*, Math. Nach 287.8 (2014), pp. 1013–1024, doi: 10.1002/mana.201010083.

[18] R. López, M. I. Munteanu: *Minimal translation surfaces in* Sol_3, J. Math. Soc. Japan 64.3 (2012), pp. 985–1003, doi: 10.2969/jmsj/06430985.

[19] J. M. Manzano, R. Souam: *The classification of totally ombilical surfaces in homogeneous 3-manifolds*, Math. Z 279 (2015), pp. 557–576.

[20] W. S. Massey: *Surfaces of Gaussian curvature zero in euclidean 3-space*, Tohoku Math. J. 14.1 (1962), pp. 73–79.

[21] W. H. Meeks: *Constant mean curvature spheres in* Sol_3, Amer. J. Math 135 (2013), pp. 1–13.

[22] W. H. Meeks, J. Pérez: *Constant mean curvature in metric Lie groups*, Contemp. Math 570 (2012), pp. 25–110.

[23] W. H. Meeks, H. Rosenberg: *The theory of minimal surfaces in* $\mathbb{M} \times \mathbb{R}$, Comment. Math. Helv 80 (2005), pp. 811–885.

[24] E. Molnár, J. Szirmai: *Symmetries in the 8 homogeneous 3-geometries*, Symmetry: Culture and Science 21.1-3 (2010), pp. 87–117.

[25] B. Nelli, H. Rozenberg: *Minimal surfaces in* $\mathbb{H}^2 \times \mathbb{R}$, Bull. Braz. Math. Soc 33 (2002), pp. 263–292.

[26] J. Pallagi, B. Schultz, J. Szirmai: *Equidistant surfaces in Nil space*, Stud. Univ. Zilina. Math. Ser 25 (2011), pp. 31–40.

[27] H. Rosenberg: *Minimal surfaces in* $\mathbb{M}^2 \times \mathbb{R}$, Illinois J. Math 46 (2002), pp. 1177–1195.

[28] A. Sanini: *Gauss map of surfaces of Heisenberg group*, Boll. Un. Math. Ital 7.11-B, Supl. Facs. 2 (1997), pp. 79–93.

[29] P. Scott: *The geometries of 3-manifolds*, Bull. London Math. Soc 15 (1983), pp. 401–487, doi: 10.1112/blms/15.5.401.

[30] R. Souam: *On stable constant mean curvature surfaces in* $\mathbb{S}^2 \times \mathbb{R}$ *and* $\mathbb{H}^2 \times \mathbb{R}$, Trans. Amer. Math. Soc 362.6 (2010), pp. 2845–2857.

[31] R. Souam, E. Toubiana: *On the classification and regularity of umbilic surfaces in homogeneous 3-manifolds*, Mat. Contemp 30 (2006), pp. 201–215.

[32] R. Souam, E. Toubiana: *Totally umbilic surfaces in homogeneous 3-manifolds*, Comm. Math. Helv 84 (2009), pp. 673–704.

[33] J. Szirmai: *Lattice-like translation ball packings in Nil space*, Publ. Math. Debrecen 80.3-4 (2012), pp. 427–440.

[34] W. M. Thurston: *Three-dimensional Geometry and Topology I*, ed. by S. Levi, Princeton Math.Series, 35, 1997.

[35] F. Torralbo, F. Urbano: *On the Gauss curvature of compact surfaces in homogeneous 3-manifolds*, Proc. Amer. Math. Soc 138.7 (2010), pp. 2561–2567.

[36] D. W. Yoon, C. W. Lee, M. K. Karacan: *Some Translation in the 3-dimensional Heisenberg group*, Bull. Korean Math. Soc 50 (2013), pp. 1329–1343, doi: 10.4134/BKMS.2013.50.4.1329.

[37] D. W. Yoon, J. W. Lee: *Translation invariant surfaces in the 3-dimensional Heisenberg group*, Bull. Iranian Math. Soc. 40 (2014), pp. 1373–1385.

Annales Mathematicae et Informaticae
***50** (2019) pp. 19–37*
DOI: 10.33039/ami.2018.09.002
http://ami.uni-eszterhazy.hu

Some k-hop based graph metrics and node ranking in wireless sensor networks*

Csaba Biró, Gábor Kusper

Eszterházy Károly University, Eger, Hungary
biro.csaba@uni-eszterhazy.hu
kusper.gabor@uni-eszterhazy.hu

Submitted: December 14, 2017
Accepted: September 13, 2018
Published online: February 27, 2019

Abstract

Node localization and ranking is an essential issue in wireless sensor networks (WSNs). We model WSNs by communication graphs. In our interpretation a communication graph can be directed, in case of heterogeneous sensor nodes, or undirected, in case of homogeneous sensor nodes, and must be strongly connected. There are many metrics to characterize networks, most of them are either global ones or local ones. The local ones consider only the immediate neighbors of the observed nodes. We are not aware of a metric which considers a subgraph, i.e., which is between global and local ones. So our main goal was to construct metrics that interpret the local properties of the nodes in a wider environment. For example, how dense the environment of the given node, or in which extent it can be relieved within its environment. In this article we introduce several novel k-hop based density and redundancy metrics: Weighted Communication Graph Density ($\mathcal{WCGD}$), Relative Communication Graph Density ($\mathcal{RCGD}$), Weighted Relative Communication Graph Density ($\mathcal{WRCGD}$), Communication Graph Redundancy ($\mathcal{CGR}$), Weighted Communication Graph Redundancy ($\mathcal{WCGR}$). We compare them to known graph metrics, and show that they can be used for node ranking.

Keywords: k-hop based graph metrics, communication graph density, com-

*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"

munication graph redundancy, node ranking

MSC: 68M07 68M17 68M12 68M15 68R10 05C05 05C12 05C40 05C82 90B10

1. Introduction

The modeling and analysis of complex networks is an important interdisciplinary field of science. The networks are mathematically related to graph theory. It is known that topology represents the properties of the whole network structure. A topology describes a real network *(with constraints)* and it can be converted to an undirected or directed graph. The common property of topological models is that they are usually calculated based on probabilities [2–4, 11]. The objects of the model can be matched by the vertices of the graph. Edges can be used to describe the relations between the objects. Graph-based modeling can be of two types: ad-hoc or measurement-based. On large wireless networks the traditional measurements based procedures can not be applied efficiently, but k-hop based approaches can be computed effectively also for large networks.

There are many graph-based metrics for modeling complex networks [9]. Topological metrics commonly used on networks: number of nodes and edges, average degree, degree distribution, connectedness, diameter, number of independent paths. Parameters for measuring the effectiveness of wireless networks: scope and coverage, scalability, expected transmission number, hop count *(number of hops)*, power consumption / lifetime.

In graph theory, the density of a graph $(\mathcal{V};\mathcal{E})$ can be calculated as $\frac{|\mathcal{E}|}{|\mathcal{V}|(|\mathcal{V}|-1)}$ [6]. Since the number of edges for a complete directed graph is $|\mathcal{V}|\,(|\mathcal{V}|-1)$, the maximum density is 1. Clearly, the minimum density is 0 (for empty graphs). There are two different approaches [13, 16], but there is no strict distinction between sparse and dense graphs.

Distance-based metrics

The eccentricity of a node u is defined as the longest hop count between the node u and any other node in the graph.

Centralization [8] is a general method for calculating a graph-level centrality score based on some node-level centrality metric. Centrality based metrics are the following ones: degree centrality *(based on degree)*, closeness centrality *(based on average distances)*, betweenness centrality (*based on geodesics)*, eigenvector centrality *(recursive: similar to page rank methods)*, eccentricity centrality. In the case of eccentricity centrality, we not use the reciprocal to assure that more central nodes have a higher value of eccentricity.

Connection-based metrics

The most basic connection-based metrics are the degree of a node, which is the number of edges to other nodes, and the degree distribution. The degree distribu-

tion $P(k)$ of a graph is then defined to be the fraction of nodes in the network with degree k. Thus if there are n nodes in total in a graph and n_k of them have degree k, we have $P(k) = n_k/n$.

Clustering is a fundamental and important property of networks, just like degree and degree distribution. Clustering coefficient is the measurement that shows the opportunity of a graph to be divided into clusters. Clusters are disjoint subgraphs of the graph. A cluster usually should be a complete subgraph, so in this way it is similar to a clique, but a cluster may consists of one node, on the other hand a clique is a complete subgraph which contains always at least two nodes in case of a communication graph. The clustering coefficient can globally [12, 18] or locally [19] characterize a graph. The global clustering coefficient is based on triplets of nodes. The global clustering coefficient of a network, also known as transitivity T, which is the ratio of the number of loops of length three and the number of paths of length two.

Let u be a vertex with k degree and given by the proportion of e edges between the v within it is neighborhood G, then the Local clustering coefficient of u in G is given by $C_u = \frac{2e}{k(k-1)}$. Thus, C_u measures the ratio of the number of edges between the neighbors of u to the total possible number of such edges. The average clustering coefficient is the average of local clustering coefficients.

Wireless sensor networks

The ad hoc wireless sensor networks (WSN) are used widely *(for example in military to observe environment).* They have the advantage that they consist of sensors with low energy consumption, which can be deployed easily in a cheap way on such areas which are out-of-the-way. These sensors are the nodes of WSN. They are capable to process some limited information and to use wireless communication. A big effort is used to research how to deploy them in an optimal way to keep efficient energy consumption and communication. Although there are many WSN solutions, the deployment of a WSN is still an active research field [1].

One of the important property of an ad hoc wireless network is node density. The dense layout makes the following properties available: high fault tolerance, high-coverage characteristics, but also cause some problems. The interference is high near to dense node areas, and there are a lot of collisions in case of message passing, which requires complicated operations for routing protocols, because of too many possible routes, routing needs lots of resources [5].

The aim of topology control techniques is to reduce the cost of the distributed algorithms interpreted on the network. The graph, which represents a network, has to be thinned because of cost-reduction by techniques like disconnection of nodes, removing links, changing scopes, etc., but the network-quality characteristics *(like scalability, coverage, fault tolerance, etc.)* must not fall below a required level. The overall aim is to create a scalable, fault-tolerant sparse topology, where the degree of the nodes are low, the maximum load is low, energy consumption is low and the paths are short. The following techniques are used to create an optimal topology: reducing the scope of nodes, removing some nodes, introducing a dominating set

of nodes, clustering, and add some new nodes to gain all-all communication [15, 17].

In multi-hop networks one hop is the unit of the path between source and destination. The hop count refers to the number of intermediate nodes through which data must pass between source and destination. Networks can be classified by the number of hops between source nodes, which measures their environment, and a sink node, which collects data. In a single-hop network there is only one *(single)* hop between the source nodes and the sink node. In a multi-hop network a sensor can also transmit data from the source to the sink because there are more than one hop from the source to the sink.

The rest of this article is organized as follows. In Section 2 we give some preliminary definitions, like communication graph. In Section 3 we introduce the new metrics, each of them are k-hop based. In Section 4 we compare existing metrics and the new ones. In section 5 we show how to use this metrics to rank nodes and Section 6 contains our conclusions.

2. Preliminaries

Given a randomly-deployed sensor network with homogeneous or heterogeneous nodes. Also given a mapping which sensor is able to communicate with which sensors directly. Accordingly, by communication graph we mean a weighted directed graph $\mathcal{D} = (\mathcal{S}; \mathcal{E}_\mathcal{C}, \mathcal{W})$, where $\mathcal{S}$ is the set of nodes, which represents the sensors, $\mathcal{E}_\mathcal{C} \subseteq \mathcal{S} \times \mathcal{S}$ is the set of edges, and $\mathcal{W}$ is the set of weights. An edge $(x_i, x_j) \in \mathcal{E}_\mathcal{C}$ represents the possibility of messaging from node x_i to x_j in $\mathcal{D}$, i.e., the sensor represented by x_j is in the transmission range of x_i. The $\mathcal{W}_{ij}$ denotes the communication cost of the (x_i, x_j) message. In the case of homogeneous sensors the $\mathcal{D}$ graph is symmetric, accordingly $\mathcal{D} = (\mathcal{S}; \mathcal{E}_\mathcal{C}, \mathcal{W})$ is equivalent to a simple weighted undirected graph $\mathcal{G} = (\mathcal{S}; \mathcal{E}_\mathcal{C}, \mathcal{W})$.

In case of an weighted undirected graph $\mathcal{G} = (\mathcal{S}; \mathcal{E}_\mathcal{C}, \mathcal{W})$ we define a clique as a subset of the node set $\mathcal{C}l \subseteq \mathcal{S}$, such that for every two nodes in $\mathcal{C}l$, there exists an edge connecting the two. The weight of a $\mathcal{C}l$ is the sum of the weight of their edges.

If our communication graph is directed, we define a clique as a subset of the node set $\mathcal{C}l \subseteq \mathcal{S}$, such that for every two nodes in $\mathcal{C}l$, there exists an edge from the first one to the second one, and from the second one back to the fists one. The weight of the $\mathcal{C}l$ is defined as above, considering that $\mathcal{W}_{ij}$ and $\mathcal{W}_{ji}$ are not necessarily equal. A maximal clique is a clique which is not a proper subset of any other clique. A n-clique is a clique which contains exactly n vertices.

In this paper we assume that the communication graph is strongly connected, the cost of communication between each node is constant *(we do not use weights)*, and the network consists of homogeneous nodes. To test the new metrics we used our own representation [7] and SAT solver [10].

3. *k*-hop based graph density and redundancy metrics

In this section we present some spanning tree and clique-based graph density metrics. With spanning tree-based metrics, we define graph density, whereas clique-based redundancy metrics mean the degree of relieving in our interpretation. We use the notion of k-hop environment of a node u, denoted by $\mathcal{G}^{[n]}(u)$, which is a subgraph of graph $\mathcal{G}$, which consists u and the nodes which can be reached from u from an path, which length is smaller or equal than k, and which contains edges between these nodes from $\mathcal{G}$. We compute local metrics for u by computing a graph metrics for $\mathcal{G}^{[n]}(u)$. The parameter k should be a relatively small number because otherwise $\mathcal{G}^{[n]}(u)$ could be the whole graph. The metrics over the k-hop environment of a node can characterize the node more properly then considering merely the node itself. On the other hand these metrics characterize not only the node but its environment.

Taking into account the constraints mentioned in Section 2, the basic notations are:

- u: the candidate node;
- k: the number of hops;
- $\mathcal{N}$, $\mathcal{V}$: the number of nodes and edges of graph $\mathcal{G}$;
- $\mathcal{N}^{[k]}(u)$,$\mathcal{V}^{[k]}(u)$: the number of nodes and edges of graph $\mathcal{G}^{[k]}(u)$;
- $\mathcal{C}l$, $\mathcal{M}$: the set of maximum cliques of graph $\mathcal{G}$ and the cardinality of this set;
- $\mathcal{C}l^{[k]}(u)$, $\mathcal{M}^{[k]}(u)$: the set of maximal cliques of graph $\mathcal{G}^{[n]}(u)$ and the cardinality of this set;
- $\mathcal{T}^{[k]}(u)$, $\mathcal{T}$: the number of edges of the minimum cost spanning tree of graph $\mathcal{G}^{[k]}(u)$ and $\mathcal{G}$. Note, that in case of a communication graph we have that $\mathcal{T} = \mathcal{N} - 1$, regardless whether the graph is directed or undirected;
- s: the spreading factor, which is rather a technical value to enlarge small differences in the metrics, in this article we set $s = 2.71$;
- cs: the clique size, minimum value is 2.

3.1. Spanning tree-based metrics

Sanning tree-based approaches can be found in the wide area of network protocols. For example, a known technique is Time-To-Live *(TTL)*. It works as follows, routing methods try to find the best path for forwarding the collected data, the TTL mechanism is used to limit the number of hops to avoid over-overlapping of paths

and to balance the data load on the nodes and the energy consumption [14]. They use also small k values.

We define graph density of the graphs $\mathcal{G}$ and $\mathcal{G}^{[k]}(u)$ as follows:

$$\mathcal{GD} = \frac{\mathcal{V}}{\mathcal{T}}$$

$$\mathcal{GD}^{[k]}(u) = \frac{\mathcal{V}^{[k]}(u)}{\mathcal{T}^{[k]}(u)}.$$

The graph density takes its maximum if the graph is complete. In case of undirected graphs the maximum is: $\frac{\mathcal{N}(\mathcal{N}-1)}{2(\mathcal{N}-1)} = \frac{\mathcal{N}}{2}$. In case of directed graphs the maximum is: $\frac{\mathcal{N}(\mathcal{N}-1)}{\mathcal{N}-1} = \mathcal{N}$. The graph density takes its maximum if the graph is a tree. In case of undirected graphs the minimum is: $\frac{\mathcal{N}-1}{\mathcal{N}-1} = 1$, since the graph is a communication graph, i.e., it is strongly connected. If the graph is directed, then the minimum is: $\frac{2(\mathcal{N}-1)}{\mathcal{N}-1} = 2$, because of the same reason.

Communication and Weighted Communication Graph Density

We define the communication graph density of node u in its k-hop environment as follows:

$$\mathcal{CGD}^{[k]}(u) = s^{\frac{\mathcal{V}^{[k]}(u)}{\mathcal{T}^{[k]}(u)}}.$$

The $\mathcal{CGD}^{[k]}(u)$ can be used also as a local metric for a node, and computed quickly for all nodes and use to rank them.

We define the weighted communication graph density of node u in its k-hop environment as follows:

$$\mathcal{WCGD}^{[k]}(u) = s^{\frac{\mathcal{V}^{[k]}(u)}{\mathcal{T}^{[k]}(u)}} \frac{\mathcal{N}^{[k]}(u)}{\mathcal{N}}.$$

The $\mathcal{WCGD}^{[k]}(u)$ is no longer a purely local metric, but takes into account the number of nodes in the k-hop environment.

Relative Communication Graph Density

We define the relative communication graph density of node u in its k-hop environment as follows:

$$\mathcal{RCGD}^{[k]}(u) = s^{\frac{\mathcal{CGD}^{[k]}(u)}{\mathcal{CGD}}} = s^{\frac{\mathcal{V}^{[k]}(u)\mathcal{T}}{\mathcal{T}^{[k]}(u)\mathcal{V}}}.$$

It maximizes its value when the k-hop environment of u, i.e., $\mathcal{G}^{[k]}(u)$ is a complete graph and the rest of the graph is a tree, or consists of several trees.

The minimum is - vice versa - assumes that the k-hop environment of u is a tree and the rest of the graph is a complete graph.

If we consider the two extremes, i.e., if the communication graph is a complete graph or if it is a tree, interestingly enough, we get the same relative communication

graph density, which is s. If the communication graph is a complete graph, then for any $k >= 1$ and for any node u we have that $\mathcal{G}^{[k]}(u)$ is equal to $\mathcal{G}$, so, $\frac{\mathcal{V}^{[k]}(u)}{\mathcal{T}^{[k]}(u)} = \frac{\mathcal{V}}{\mathcal{T}}$, i.e., $\frac{\mathcal{V}^{[k]}(u)\mathcal{T}}{\mathcal{T}^{[k]}(u)\mathcal{V}} = 1$. On the other hand, if the communication graph is a tree, then its communication graph density is a constant (1 if the graph is undirected, 2 if it is directed) for any n and u, so again $\frac{\mathcal{V}^{[k]}(u)\mathcal{T}}{\mathcal{T}^{[k]}(u)\mathcal{V}} = 1$.

We get the same result for the two extreme cases, because this metric shows the relative density of subgraph related to the whole graph. A tree has a very small density, and a complete graph has a very high density, but if we take a subgraph of a tree then it has the same density as the whole, and the same is true for a complete graph. So they have the same relative density.

This metric shows whether the k-hop environment of a node is more dense as the whole graph, or has the same density, or it is less dense. This means that if

- $\mathcal{RCGD}^{[k]}(u) = s$, then $\mathcal{G}^{[k]}(u)$ has the same cgd as $\mathcal{G}$;
- $\mathcal{RCGD}^{[k]}(u) < s$, then $\mathcal{G}^{[k]}(u)$ has smaller cgd than $\mathcal{G}$;
- $\mathcal{RCGD}^{[k]}(u) > s$, then $\mathcal{G}^{[k]}(u)$ has bigger cgd than $\mathcal{G}$;

where cgd means communication graph density.

Note, that this metric is computed by dividing a local property by a global one.

Weighted Relative Communication Graph Density

We define the weighted relative communication graph density of node u in its k-hop environment as follows:

$$\mathcal{WRCGD}^{[k]}(u) = \mathcal{RCGD}^{[k]}(u)\frac{\mathcal{N}^{[k]}(u)}{\mathcal{N}} = s\frac{\mathcal{V}^{[k]}(u)\mathcal{T}}{\mathcal{T}^{[k]}(u)\mathcal{V}}\frac{\mathcal{N}^{[k]}(u)}{\mathcal{N}}.$$

Note, that this metric is computed as a multiplication of two numbers, which are both computed by dividing a local property by a global one, so we have $(local'/global') * (local''/global'')$.

This metric takes in consideration also how many nodes are in the n-hop environment of the node u. A node is more valuable if its k-hope environment is bigger.

3.2. Clique-based metrics

During the work of a WSN the topology of the network may change because some sensors may go wrong, or the transmission range can be less. If a node can be found in a dense (redundant) environment then it may happen more often that communication interference occurs and routing is more resource consuming; on the other hand, the environment itself is more fault tolerant. In a sparse environment routing is easier, communication interference is less frequent, but the environment is less fault tolerant. The aim of topology control techniques is to reduce the cost

of the distributed algorithms interpreted on the network. But the network-quality characteristics *(like scalability, coverage, fault tolerance, etc.)* must not fall below a required level. A clique is a complete subgraph, so they have high communication redundancy, on the other hand they allow high fault tolerance, results in high coverage, etc.

First of all we define the average clique size as follows:

$$\overline{\mathcal{CL}} = \frac{1}{\mathcal{M}} \sum_{i=1}^{\mathcal{M}} |\mathcal{C}l_i|_{>=cs} \,.$$

The average clique size is maximal, if the graph is complete. Its minimum is cs if all maximal cliques have the size cs. It is not defined if there is no clique with size at least cs. Its maximum is $\mathcal{N}$ if the communication graph is complete, because then we have only one maximal clique, the graph itself. The clique problem, the problem of finding all maximal size cliques, is a well-known NP-complete problem. It meas that is not feasible to find all maximal cliques in a large graph. So one can not use clique based metrics to guide topology control techniques, except if we work with relatively small graphs, like in the k-hop environment of a node.

Clique size-based metrics

So we define the clique size-based communication graph redundancy of node u within k-hop environment as follows:

$$\mathcal{CGR}_{sb}{}^{[k]}(u) = \frac{1}{\mathcal{M}^{[k]}(u)} \sum_{i=1}^{\mathcal{M}^{[k]}(u)} \left|\mathcal{C}l^{[k]}(u)_i\right|_{>=cs}$$

It only shows the average clique size within k-hop environment of node u, but it ignores the number of nodes within the k-hop environment.

We define weighted communication graph redundancy of node u within k-hop environment as follows:

$$\mathcal{WCGR}_{sb}{}^{[k]}(u) = \mathcal{CGR}_{sb}{}^{[k]}(u) \frac{\mathcal{N}^{[k]}(u)}{\mathcal{N}}.$$

This metric uses also the number of nodesThis can be considered to be a local metric, because the computationally intensive tasks *(find cliques)* typically occur in a k-hop environment.

Clique value-based metrics

Since a clique of size 4 is more valuable in a graph than 6 in a graph with 100 nodes, we shall take into consideration the number of nodes in the graph, which is denoted by $\mathcal{N}$, to compute the value of a clique. We also use the average clique size to normalize this value.

So we define the value of a clique as follows:

$$\mathcal{CL}_V = \frac{|\mathcal{C}l|_{>=cs}}{\mathcal{N}} s^{\frac{|\mathcal{C}l|_{>=cs}}{\overline{\mathcal{CL}}}}.$$

We define also the average value of cliques as follows:

$$\overline{\mathcal{CL}_V} = \frac{1}{\mathcal{M}} \sum_{i=1}^{\mathcal{M}} \frac{|\mathcal{C}l_i|_{>=cs}}{\mathcal{N}} s^{\frac{|\mathcal{C}l_i|_{>=cs}}{\overline{\mathcal{CL}}}}$$

We define also the average value of cliques within the k-hop environment, also called clique value-based communication graph redundancy as follows:

$$\mathcal{CGR}_{vb}{}^{[k]}(u) = \frac{1}{\frac{1}{\mathcal{M}^{[k]}(u)} \sum_{i=1}^{\mathcal{M}^{[k]}(u)} \frac{|\mathcal{C}l^{[k]}(u)_i|_{>=cs}}{\mathcal{N}^{[k]}(u)} s^{\frac{|\mathcal{C}l^{[k]}(u)_i|_{>=cs}}{\overline{\mathcal{CL}}^{[k]}(u)}}}.$$

This metric is the pair of $\mathcal{CL}_V$ in case of $\mathcal{G}^{[k]}(u)$. This is a local metric, but this notion does not takes into consideration the number of nodes in the k-hop environment of u. Without reciprocal, the peripheral but relievable nodes are ranked in advance.

After considerating the number of nodes in the k-hop and conversion we define weighted clique value-based communication graph redundancy as follows:

$$\mathcal{WCGR}_{vb}{}^{[k]}(u) = \frac{1}{\mathcal{CGR}_{vb}{}^{[k]}(u)} \frac{\mathcal{N}^{[k]}(u)}{\mathcal{N}} =$$

$$= \frac{1}{\mathcal{M}^{[k]}(u)} \sum_{i=1}^{\mathcal{M}^{[k]}(u)} \frac{\mathcal{N}^{[k]}(u) |\mathcal{C}l^{[k]}(u)_i|_{>=cs}}{\mathcal{N}^2} s^{\frac{|\mathcal{C}l^{[k]}(u)_i|_{>=cs}}{\overline{\mathcal{CL}}^{[k]}(u)}}$$

This metric is the pair of $\overline{\mathcal{CL}_V}$ in case of $\mathcal{G}^{[k]}(u)$.

4. Comparisons with other metrics

In this article we considered networks with 200-500 nodes at 15-40% densities. The k value in each case is less than 3. An important constraint was that the largest k-hop environment must be smaller than the quarter of a complete graph.

For simulating and analyzing networks we used a self-developed *Python* tool *based on NetworkX* [1]. For computing pairwise correlation of metrics we used *pandas*[2]. Many metrics are only implemented for undirected graphs in *NetworkX*, therefore, the comparisons were done only on undirected graphs.

The results of the correlation analysis are presented in Table 1–3 *(average values of 1000 runs)*, shows some interesting phenomena and experience. The abbrevation c. means centrality.

[1] `https://networkx.github.io`
[2] `https://pandas.pydata.org`

4.1. 1-hop based environment

	k-hop based metrics					
	$\mathcal{WCGD}^{[k]}$	$\mathcal{RCGD}^{[k]}$	$\mathcal{WRCGD}^{[k]}$	$\mathcal{WCGR}_{sb}{}^{[k]}$	$\mathcal{CGR}_{vb}{}^{[k]}$	$\mathcal{WCGR}_{vb}{}^{[k]}$
Clustering coeff.	0,06	0,19	0,00	0,37	-0,17	0,14
Eccentricity	-0,07	-0,21	-0,24	-0,15	-0,31	-0,21
Betweenness c.	-0,04	-0,05	0,06	-0,12	0,22	-0,01
Degree c.	**0,54**	**0,87**	**0,94**	**0,76**	**0,84**	**0,83**
Closeness c.	0,05	0,23	0,28	0,17	0,37	0,23
Eigenvector c.	**0,67**	**0,61**	**0,64**	0,49	0,28	**0,68**

Table 1: Correlations with other metrics, where k is 1

It can be seen from the Table 1 that within 1-hop environment the defined metrics show their most significant correlation with degree centrality. The correlation is the strongest between $\mathcal{WRCGD}$ and degree centrality, the correlation is over 90%. $\mathcal{WCGD}$, $\mathcal{RCGD}$, $\mathcal{WRCGD}$ and $\mathcal{WCGR}_{vb}$ metrics are also strongly correlated with the eigenvector centrality.

4.2. 2-hop based environment

	k-hop based metrics					
	$\mathcal{WCGD}^{[k]}$	$\mathcal{RCGD}^{[k]}$	$\mathcal{WRCGD}^{[k]}$	$\mathcal{WCGR}_{sb}{}^{[k]}$	$\mathcal{CGR}_{vb}{}^{[k]}$	$\mathcal{WCGR}_{vb}{}^{[k]}$
Clustering coeff.	-0,07	0,04	-0,15	0,08	-0,26	-0,06
Eccentricity	-0,15	-0,24	-0,38	-0,22	-0,45	-0,33
Betweenness c.	0,06	0,03	0,22	0,05	0,32	0,17
Degree c.	**0,54**	**0,78**	**0,83**	**0,72**	**0,67**	**0,72**
Closeness c.	0,08	0,26	0,44	0,24	0,53	0,36
Eigenvector c.	**0,87**	**0,68**	**0,67**	**0,57**	0,26	**0,71**

Table 2: Correlations with other metrics, where k is 2

It can be seen from the Table 2 that the defined metrics within 2-hop environment showed a weaker correlation with the degree centrality and stronger with the eigenvector centrality, since the degree of neighbors of the examined node also affects the density and redundancy of the environment. The strongest correlation with the degree centrality is still shown with $\mathcal{WRCGD}$, while with the eigenvector centrality correlats best with$\mathcal{WCGD}$.

4.3. 3-hop based environment

The correlations in the 3-hop environment are shown in the Table 3. In general, the correlations with the degree centrality and the eigenvector centrality are no longer significant, the eccentricity and the closeness centraliy correlations are reinforced.

In the following, we will analyze in detail the relationship between the new and already known metrics.

	k-hop based metrics					
	$\mathcal{WCGD}^{[k]}$	$\mathcal{RCGD}^{[k]}$	$\mathcal{WRCGD}^{[k]}$	$\mathcal{WCGR}_{sb}{}^{[k]}$	$\mathcal{CGR}_{vb}{}^{[k]}$	$\mathcal{WCGR}_{vb}{}^{[k]}$
Clustering coeff.	0,12	0,03	-0,20	0,03	-0,24	-0,19
Eccentricity	-0,27	-0,18	**-0,55**	-0,30	**-0,56**	**-0,51**
Betweenness c.	0,18	0,07	0,36	0,14	0,38	0,29
Degree c.	0,46	**0,55**	**0,63**	**0,56**	**0,54**	**0,62**
Closeness c.	0,38	0,25	**0,63**	0,34	**0,67**	**0,54**
Eigenvector c.	**0,72**	**0,55**	**0,52**	**0,52**	-0,26	**0,59**

Table 3: Correlations with other metrics, where k is 3

Weighted Communication Graph Density

The metric $\mathcal{WCGD}^{[k]}(u)$ correlats strongly with eigenvector centrality. There is a not too strong but significant correlation with degree centrality also, and there is no relevant correlation with other metrics. If we want to characterize this metric on the basis of the above, then a high $\mathcal{WCGD}^{[k]}(u)$ value node has the following properties (in k-hop environment, if $k = 3$):

- average probability of high number of direct connections,
- high probability of high degree of neighbors,
- weak probability of central location.

Relative Communication Graph Density

The metric $\mathcal{RCGD}^{[k]}(u)$ has average correlation with degree centrality and eigenvector centrality, weak but significant contact with closeness centrality, and there is no relevant correlation with other metrics. So a high $\mathcal{RCGD}^{[k]}(u)$ value node has the following properties (in k-hop environment, if $k = 3$):

- average probability of high number of direct connections,
- average probability of high degree of neighbors.

Weighted Relative Communication Graph Density

The metric $\mathcal{WRCGD}^{[k]}(u)$ has an average linear correlation with degree centrality, closeness centrality, and eigenvector centrality, and suggests a weak correlation with betweenness centrality, but with eccentricity shows an average but inverse correlation. So a high $\mathcal{WRCGD}^{[k]}(u)$ value node has the following properties (in k-hop environment, if $k = 3$):

- average probability of high number of direct connections,
- average probability of central location,
- average probability of high degree of neighbors,
- weak probability of high geodesic distance from any other node.

Weighted Communication Graph Redundancy *(size-based)*

The metric ${\mathcal{WCGR}_{sb}}^{[k]}(u)$ has average correlation with degree centrality and eigenvector centrality. It suggests a weak but significant correlation with closeness centrality and inverse correlation with eccenticity. There is no relevant correlation with other metrics. So a high $\mathcal{WCGR}_{sb}^{[k]}(u)$ value node has the following properties (in k-hop environment, if $k = 3$):

- average probability of high number of direct connections,
- average probability of high degree of neighbors,
- weak probability of central location.

Communication Graph Redundancy *(value-based)*

The metric $\mathcal{CGR}_{vb}^{[k]}(u)$ has an average linear correlation with degree centrality and closeness centrality. It suggests a weak correlation with betweenness centrality. It shows shows an average but inverse correlation with eccentricity. So a high $\mathcal{CGR}_{vb}^{[k]}(u)$ value node has the following properties (in k-hop environment, if $k = 3$):

- average probability of high number of direct connections,
- average probability of central location,
- weak probability of great geodesic distance from any other node.

Weighted Communication Graph Redundancy *(value-based)*

The metric $\mathcal{WCGR}_{vb}^{[k]}(u)$ has an average linear correlation with degree centrality, closeness centrality, and eigenvector centrality. It suggests a weak correlation with betweenness centrality, but with eccentricity shows an average but inverse correlation. So a high $\mathcal{WCGR}_{vb}^{[k]}(u)$ value node has the following properties (in k-hop environment, if $k = 3$):

- average probability of high number of direct connections,
- average probability of high degree of neighbors,
- average probability of central location,
- weak probability of great geodesic distance from any other node.

5. Node ranking

In this section we show how to use the different metrics to make node ranking *(top 30 selection)*. The generated network (shown in Fig. 1) contains 100 randomly deployed and homogeneous sensor nodes *(vertices)* with 926 connections *(edges)*. The density is 18.7%, the transmission range is 55 m, the area is $300m \times 300m$ and the k-hop number is 3. The communication graph of the exemplary network are shown Fig. 2.

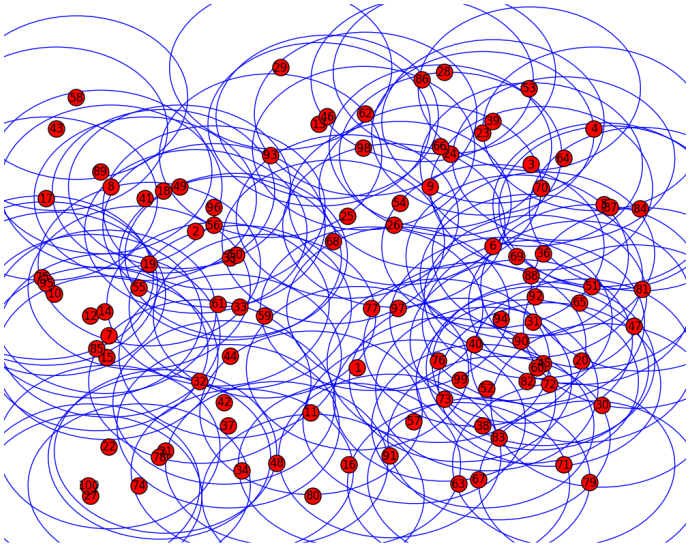


Figure 1: Randomly Deployed Sensor Network

Both the spanning tree and the clique based metrics show the denser environments of the network. Since this network consists of only 100 nodes it does not give a real picture of the metrics, but we can still see the tendencies.

Spanning tree-based metrics (density)

Figures 3–5 show how to use the spanning tree based metrics to make ranking nodes. The task of all three metrics is to designate the densest environments. Based on the comparisons, the most significant feature of $\mathcal{WCGD}$ is that the neighbors and the neighbors' neighbors have a high degree. Figure 3 shows that the top 3 node and the at least 40% of the selected nodes are centrally located. The metric $\mathcal{RCGD}$ showed no significant correlation with the closeness centrality and eccentricity. In Figure 4 we can see that among the selected nodes there are only few nodes in central position. The metric $\mathcal{RCGD}$ primarily marks the nodes within the densest areas. The metric $\mathcal{WCGD}$ selects those nodes whose density is high within their

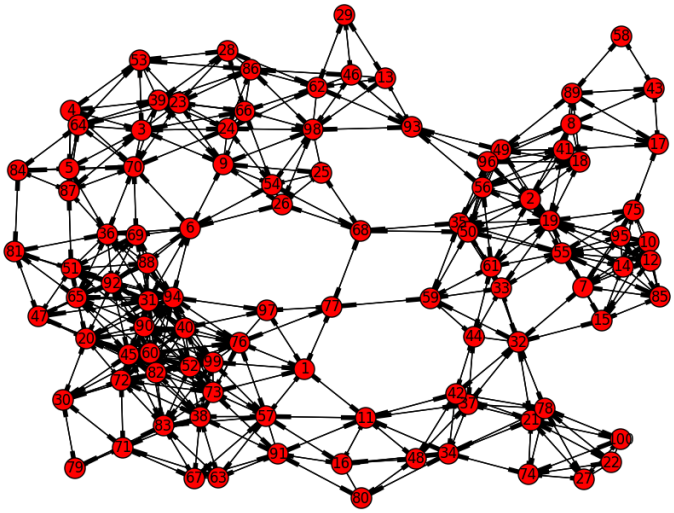


Figure 2: Communication Graph

k-hop environment and centrally located. Figure 5 shows that the top 3 node and the at least 80% of the selected nodes are centrally located.

In these figures we use the following colour codes: top 1 rank node is red, top 2 is green, top 3 is yellow, top 4–10 are blue, top 11–20 are pink, top 21–30 are orange, the rest is cyan.

Figure 3 shows the top 30 ranked nodes based on the *Weighted Communication Graph Density* metric. In 3-hop environments, the highest weighted communication graph density has nodes 77, 68, and 26.

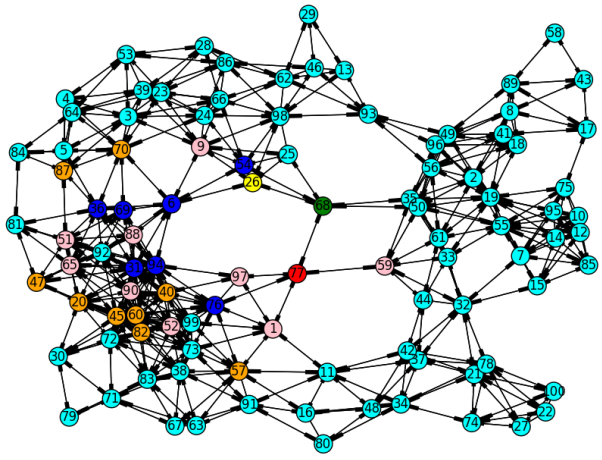


Figure 3: Weighted Communication Graph Density

Figure 4 shows the top 30 ranked nodes based on the *Relative Communication Graph Density* metric. In 3-hop environments, the highest relative communication graph density has nodes 30, 79, and 71.

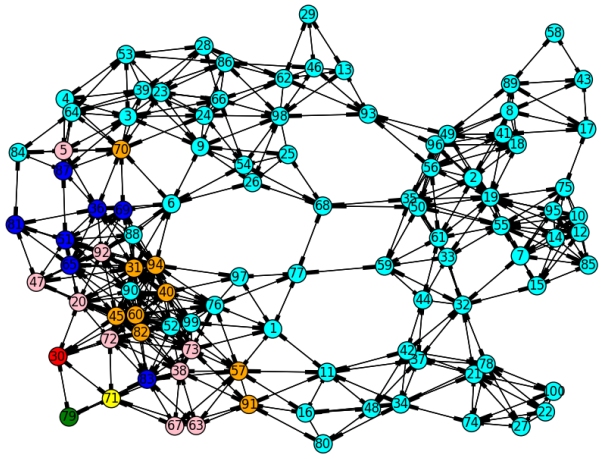


Figure 4: Relative Communication Graph Density

Figure 5 shows the top 30 ranked nodes based on the *Weighted Relative Communication Graph Density* metric. In 3-hop environments, the highest weighted relative communication graph density has nodes 68, 77, and 26.

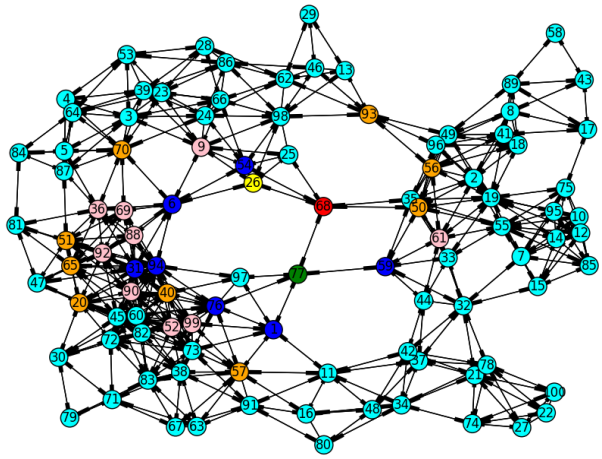


Figure 5: Weighted Relative Communication Graph Density

Clique-based metrics (degree of relieving)

Figures 6–8 show how can we use the clique based metrics to make ranking nodes. The task of all three metrics is to designate the degree of relieving nodes within their k-hop environment. Figure 6 shows node ranking created by $\mathcal{WCGR}_{sb}$. In case of $\mathcal{WCGR}_{sb}$ only the size of cliques in the k-hop environment of the examined node is relevant. $\mathcal{WCGR}_{sb}$ marks primarily the nodes within the densest areas, just lik $\mathcal{RCGD}$. Figure 7 shows node ranking created by $\mathcal{WCGR}_{vb}$. The significant difference between $\mathcal{WCGR}_{vb}$ and $\mathcal{WCGR}_{sb}$ is that $\mathcal{WCGR}_{vb}$ takes into consideration also the degree of neighbors. Figure 8 clearly shows that $\mathcal{CGR}_{vb}$ primarily focuses on centrally located nodes so the top 3 node and the at least 80% of the selected nodes are centrally located.

Figure 6 shows the top 30 ranked nodes based on the *Clique size-based Weighted Communication Graph Redundancy* metric. In 3-hop environments, the most relieved nodes has nodes 79, 30, and 81.

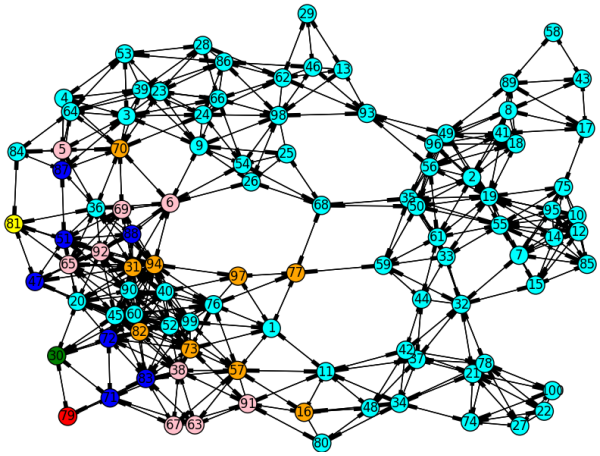


Figure 6: Clique size-based Weighted Communication Graph Redundancy

Figure 7 shows the top 30 ranked nodes based on the *Clique value-based Communication Graph Redundancy* metric. In 3-hop environments, the most relieved nodes has nodes 68, 77, and 26.

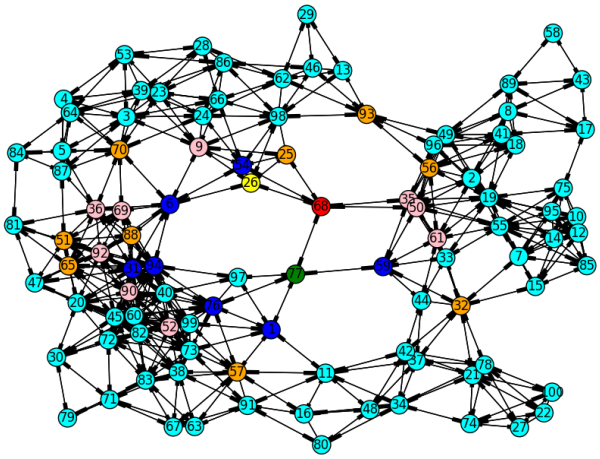


Figure 7: Clique value-based Communication Graph Redundancy

Figure 8 shows the top 30 ranked nodes based on the *Clique value-based Weighted Communication Graph Redundancy* metric. In 3-hop environments, the most relieved nodes has nodes 77, 68, and 26.

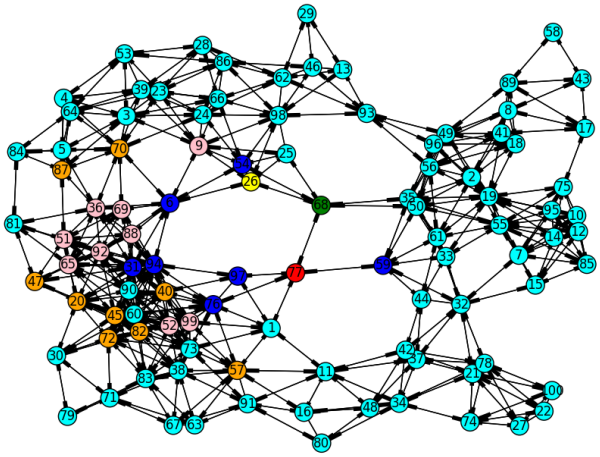


Figure 8: Clique value-based Weighted Communication Graph Redundancy

These figures show that 1-1 densities and redundancy-based metrics similarly rank the nodes. Why? The first reason is that the two concepts are closely related, if the density is high, then the redundancy is high, too. However, if the test is performed with directed graphs, there will be significant differences, because if node u can send a message to node v, then v may not be able to send a message to u, which means that high density does not mean necessarily also high redundancy.

6. Conclusions and Future work

In this paper we introduced several novel k-hop based density and redundancy metrics. We compared them to well-known graph metrics and we showed how can them be used for node ranking. Our primary goal was to define metrics that are able to rank nodes depending on their immediate environment within the whole network. Based on the results, we think that more sophisticated node ranking can be given using the new metrics. We primarily focused on modelling small heterogeneous networks. Metrics are defined so that they can be used also on networks where communication costs are different *(weighted directed graphs)*. Our further goal is to investigate also such networks. An interesting questions is how to use these metrics to increase the efficiency of different *(Tx range-based, hierarchical)* topology control methods and how to use them in different hierarchical topological models *(e.g. clustering, cluster head selection)*.

References

[1] I. F. Akyildiz, W. Su, Y. Sankarasubramaniam, E. Cayirci: *A survey on sensor networks*, IEEE Communications magazine 40.8 (2002), pp. 102–114, doi: 10.1109/MCOM.2002.1024422.

[2] R. Albert, A.-L. Barabási: *Statistical mechanics of complex networks*, Reviews of modern physics 74.1 (2002), p. 47.

[3] A.-L. Barabási, R. Albert, H. Jeong: *Scale-free characteristics of random networks: the topology of the world-wide web*, Physica A: statistical mechanics and its applications 281.1-4 (2000), pp. 69–77.

[4] S. Boccaletti, V. Latora, Y. Moreno, M. Chavez, D.-U. Hwang: *Complex networks: Structure and dynamics*, Physics Reports 424.4 (2006), pp. 175–308, issn: 0370-1573, doi: 10.1016/j.physrep.2005.10.009.

[5] Bolic, Miodrag, Simplot-Ryl, David Stojmenovic, Ivan (eds.): *RFID Systems - Research Trends and Challenges*, New York: John Wiley & Sons, 2010, p. 576, isbn: 978-0-470-74602-8, doi: 10.1002/9780470665251.

[6] T. F. Coleman, J. J. Moré: *Estimation of sparse Jacobian matrices and graph coloring blems*, SIAM journal on Numerical Analysis 20.1 (1983), pp. 187–209.

[7] G. K. Csaba Biró: *Equivalence of Strongly Connected Graphs and Black-and-White 2-SAT Problems*, Miskolc Mathematical Notes 19.2 (2018), pp. 755–768, doi: 10.18514/MMN.2018.2140.

[8] L. C. Freeman: *Centrality in social networks conceptual clarification*, Social Networks (1978), p. 215, doi: 10.1.1.227.9549.

[9] J. M. Hernández, P. Van Mieghem: *Classification of graph metrics*, Delft University of Technology: Mekelweg, The Netherlands (2011), pp. 1–20.

[10] G. Kusper, C. Biró, G. B. Iszály: *SAT solving by CSFLOC, the next generation of full-length clause counting algorithms*, in: 2018 IEEE International Conference on Future IoT Technologies (Future IoT), IEEE, 2018, pp. 1–9, doi: 10.1109/FIOT.2018.8325589.

[11] A. Lesne: *Complex networks: from graph theory to biology*, Letters in Mathematical Physics 78.3 (2006), pp. 235–262, doi: 10.1007/s11005-006-0123-1.

[12] R. D. Luce, A. D. Perry: *A method of matrix analysis of group structure*, Psychometrika 14.2 (June 1949), pp. 95–116, issn: 1860-0980, doi: 10.1007/BF02289146, url: 10.1007/BF02289146.

[13] J. Nešetřil, P. O. De Mendez: *First order properties on nowhere dense structures*, The Journal of Symbolic Logic 75.3 (2010), pp. 868–887, doi: 10.2178/jsl/1278682204.

[14] D. Rewadkar, M. P. Madhukar: *An adaptive routing algorithm using dynamic TTL for data aggregation in Wireless Sensor Network*, in: Second International Conference on Current Trends In Engineering and Technology-ICCTET 2014, IEEE, 2014, pp. 192–197, doi: 10.1109/ICCTET.2014.6966286.

[15] P. Santi: *Topology Control in Wireless Ad Hoc and Sensor Networks*, ACM Computing Surveys 37.2 (2005), pp. 164–194, issn: 0360-0300, doi: 10.1145/1089733.1089736.

[16] I. Streinu, L. Theran: *Sparse hypergraphs and pebble game algorithms*, European Journal of Combinatorics 30.8 (2009), Combinatorial Geometries and Applications: Oriented Matroids and Matroids, pp. 1944–1964, issn: 0195-6698, doi: 10.1016/j.ejc.2008.12.018.

[17] Y. Wang: *Topology Control for Wireless Sensor Networks*, Springer - Wireless Sensor Networks and Applications 148.2 (2008), pp. 113–147, issn: 1860-4862, doi: 10.1007/978-0-387-49592-7.

[18] S. Wasserman, K. Faust: *Social network analysis: Methods and applications*, vol. 8, Cambridge university press, 1994.

[19] D. J. Watts, S. H. Strogatz: *Collective dynamics of 'small-world' networks*, Nature 393.6684 (1998), pp. 440–442, issn: 00280836, doi: 10.1038/30918.

Annales Mathematicae et Informaticae
***50** (2019) pp. 39–56*
DOI: 10.33039/ami.2019.03.004
http://ami.uni-eszterhazy.hu

Capacity of permutations*

Aubrey Blecher[a], Charlotte Brennan[a], Arnold Knopfmacher[a], Mark Shattuck[b]

[a]The John Knopfmacher Centre for Applicable Analysis and Number Theory
School of Mathematics, University of the Witwatersrand
Johannesburg, South Africa
Aubrey.Blecher@wits.ac.za
Charlotte.Brennan@wits.ac.za
Arnold.Knopfmacher@wits.ac.za

[b]Department of Mathematics, University of Tennessee
Knoxville, Tennessee, USA
shattuck@math.utk.edu

Submitted: July 27, 2018
Accepted: March 25, 2019
Published online: May 21, 2019

Abstract

Permutations of $[n] = \{1, 2, \ldots, n\}$ may be represented geometrically as bargraphs with column heights in $[n]$. We define the notion of capacity of a permutation to be the amount of water that the corresponding bargraph would hold if the region above it could retain water assuming the usual rules of fluid flow. Let $C(n)$ be the sum of the capacities of all permutations of $[n]$. We obtain, in a unique manner, all permutations of length $n+1$ from those of length n, which yields a recursion for $C(n+1)$ in terms of $C(n)$ that we can subsequently solve. Finally, we consider permutations that have a single dam (i.e., a single area of water containment) and compute the total number and capacity of all such permutations of a given length. We also provide bijective proofs of these formulas and an asymptotic estimate is found for the average capacity as n increases without bound.

Keywords: permutation statistic, generating functions, asymptotics

MSC: 05A05, 05A15, 05A16

*The second and third authors are supported by the National Research Foundation under grant numbers 86329 and 81021, respectively.

1. Introduction

A permutation of $[n]$ is an ordering of the elements of $[n]$. In recent years, a variety of different statistics on permutations have been studied in the literature; see, for example, [1–3, 6–12, 14, 15, 17, 18]. In order to describe our new statistic, we represent a permutation of $[n]$ as a bargraph with column heights in $[n]$. The height of the i-th column of the bargraph equals the size of the i-th letter of the permutation. We define the *capacity* of a permutation to be the amount of water the representing bargraph would retain if water is poured onto it from above and allowed to escape in any direction (if needed) subject to the usual rules of fluid flow. It is thus a measure of the area in the plane where the water would be retained. See [16] where the capacity statistic is considered on compositions and finite set partitions, represented geometrically as bargraphs, and also [4, 5] for further related results.

The organization of this paper is as follows. In the next section, we find an explicit formula for the sum of the capacities of all permutations of length n. In the third section, we consider the situation in which the retained water is restricted to a single area, i.e., to a single subsequence of consecutive entries, and refer to such permutations as having one dam. We then prove an analogous formula for the total capacity taken over all one-dam permutations of length n as well as an explicit formula for the total number of such permutations by considering a refinement according to the width of the dam. Some asymptotic estimates as n approaches infinity are also found for these quantities, and in the final section, bijective proofs are provided.

Illustrated below in Figure 1 is the capacity of the permutation 526134 of [6].

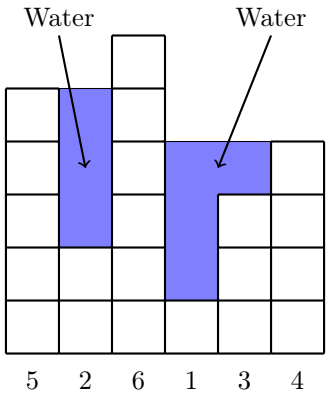


Figure 1: Permutation 526134 of [6] with capacity 7

2. Total capacity of permutations

Let $C(n)$ be the total capacity of all permutations of $[n]$. We employ a direct counting approach in order to obtain a recurrence for $C(n+1)$. This involves the following procedure. Consider an arbitrary permutation of $[n]$; from this, we obtain a unique permutation of $[n+1]$ via a simple two-step process:

- We raise the permutation of $[n]$ by adding one to each element in the original permutation. This produces a permutation of the elements of $[n+1] \setminus \{1\}$ as illustrated below in Figure 2.

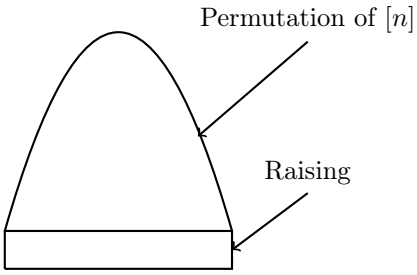


Figure 2: Raising a permutation of $[n]$ by one

- To convert this to an arbitrary permutation of $[n+1]$, we insert the element 1 within the raised permutation in any one of $n+1$ possible positions as shown in Figure 3.

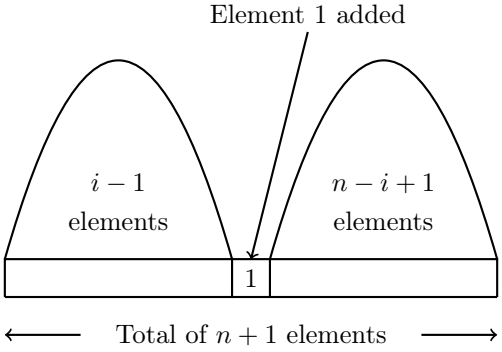


Figure 3: Element 1 added in the i-th position, $1 \leq i \leq n+1$

We denote the set of all permutations of $[n]$ by $\mathcal{S}_n$. Note that each member of $\mathcal{S}_{n+1}$ arises uniquely upon applying the above procedure to $\mathcal{S}_n$.

If the element 1 is added in either the first or the last position, there is no change to the capacity of the original permutation. In general, we will consider

adding the 1 in all other positions i, where $2 \leq i \leq n$, and determine what addition this makes to the capacity of the member of $\mathcal{S}_n$ from which it arose. Note that the two-step procedure above is seen to leave the capacity of the precursor permutation unchanged, except for the additional capacity above the added element 1.

So our method will consist of calculating (see Figure 3) how many times the unchanged original capacity is to be counted, and secondly what is the total additional contribution above the 1 over all the possible original permutations of $[n]$.

So let us consider our general case where the 1 is added in the i-th position. Let r denote the maximum element to the left of 1, where $i \leq r \leq n+1$. First, consider the case $i \leq r \leq n$, which is illustrated in Figure 4. Then $n+1$ must occur to the right of the 1 and hence the additional capacity above the 1 is $r-1$. For each maximum r, the set of numbers to the left of 1 can be chosen, and then permuted, in $\binom{r-2}{i-2}(i-1)!$ ways, while the remaining numbers to the right of 1 can be permuted in $(n-i+1)!$ ways.

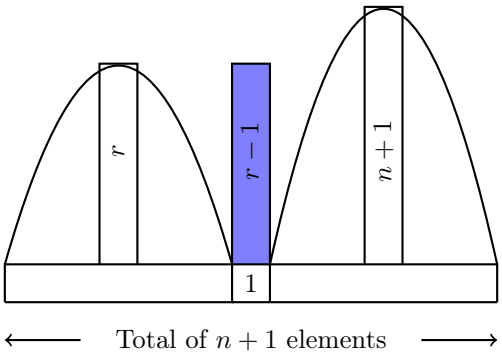


Figure 4: Additional capacity above the element 1, $i \leq r \leq n$

Thus, the total additional capacity is

$$\sum_{i=2}^{n} \sum_{r=i}^{n} \binom{r-2}{i-2} (i-1)!(n-i+1)!(r-1). \tag{2.1}$$

Now let us consider the case $r = n+1$. The sketch for this case is in Figure 5.

Here, by the pigeonhole principle, we have $n-i+2 \leq s \leq n$, and by a similar argument as for equation (2.1), the total additional capacity in this case is

$$\sum_{i=2}^{n} \sum_{s=n-i+2}^{n} \binom{s-2}{n-i} (n-i+1)!(i-1)!(s-1). \tag{2.2}$$

Expression (2.2) is equivalent to (2.1), which can also be realized by applying the reversal operation.

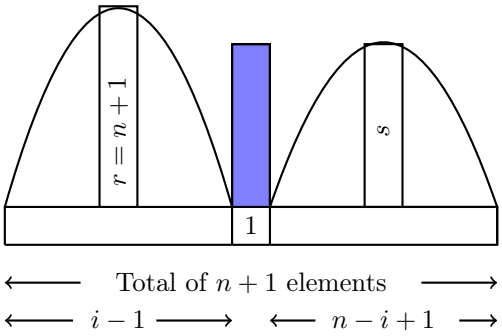


Figure 5: Additional capacity above the element 1

Thus, the total additional capacity over all permutations is

$$\begin{aligned}
&2\sum_{i=2}^{n}(i-1)!(n-i+1)!\sum_{r=i}^{n}\binom{r-2}{i-2}(r-1)\\
&=2\sum_{i=2}^{n}(i-1)(i-1)!(n-i+1)!\binom{n}{i}\\
&=2n!\sum_{i=2}^{n}\frac{(i-1)(n-i+1)}{i}\\
&=2n!\sum_{i=1}^{n}\left(-i+(n+2)-\frac{n+1}{i}\right)\\
&=2n!\left(\binom{n}{2}+2n-(n+1)H_n\right),
\end{aligned}$$

where H_n is the n-th Harmonic number $\sum_{i=1}^{n}\frac{1}{i}$.

So the recursion is

$$C(n+1)=(n+1)C(n)+2n!\left(\binom{n}{2}+2n-(n+1)H_n\right),\quad n\geq 1,$$

with $C(1)=0$.

We solve this first order linear recursion and obtain the following result.

Theorem 2.1. *The total capacity $C(n)$ over all permutations of $[n]$ is*

$$C(n)=\frac{n!}{2}\left(n(n+7)-4(n+1)H_n\right).$$

The values of $C(n)$ for $1 \leq n \leq 12$ are

$$0, 0, 2, \mathbf{28}, 312, 3384, 37872, 446688, 5595840, 74617920, 1058711040, 15958667520.$$

To illustrate, we list all the permutations of length 4 and their respective capacities in the table below. Note that the total is indeed 28, shown in bold in the list above.

Permutation	1234	1243	1324	1342	1423	1432
Capacity	0	0	1	0	1	0
Permutation	2134	2143	2314	2341	2413	2431
Capacity	1	1	2	0	2	0
Permutation	3124	3142	3214	3241	3412	3421
Capacity	3	2	3	1	1	0
Permutation	4123	4132	4213	4231	4312	4321
Capacity	3	2	3	1	1	0

Using the asymptotic expansion of H_n, we obtain the following estimate.

Corollary 2.2. *The average capacity for permutations of* $[n]$ *is*

$$\frac{1}{2}\left(n(n+7) - 4(n+1)H_n\right) = \frac{n^2}{2} - 2n\ln n + \left(\frac{7}{2} - 2\gamma\right)n - 2\ln n + O(1)$$

as $n \to \infty$*, where* γ *is Euler's constant.*

3. Total capacity in the one-dam situation

For permutations of $[n]$, we have computed the total capacity $C(n)$. We now determine the total capacity of permutations having exactly one dam defined as follows.

A permutation $\sigma = \sigma_1\sigma_2\cdots\sigma_n$ of $[n]$ is said to have exactly *one dam* if there exists only a single connected area of water containment. More precisely, we define the one-dam situation as that in which all of the water retained by a permutation σ is contained within a subsequence of σ of the form $r\sigma_i\sigma_{i+1}\cdots\sigma_j s$, where $2 \leq r, s \leq n$ and $\sigma_i, \sigma_{i+1}, \ldots, \sigma_j < \min\{r, s\}$. Moreover, the contribution of each σ_ℓ for $1 \leq \ell < i$ or $j < \ell \leq n$ towards the capacity is zero.

For example, the permutation $\sigma = 463152$ of [6] has only one dam, with $r = 6$ and $s = 5$, whereas the permutation in Figure 1 above has two. We give, in Figure 6 below, a symbolic sketch of a generic permutation having a single dam.

Let us define the *dam width* p of a one-dam permutation as the number of letters p that actually contribute to the capacity, i.e., the aforementioned

$$\sigma_i\sigma_{i+1}\cdots\sigma_j \text{ has } j - i + 1 = p.$$

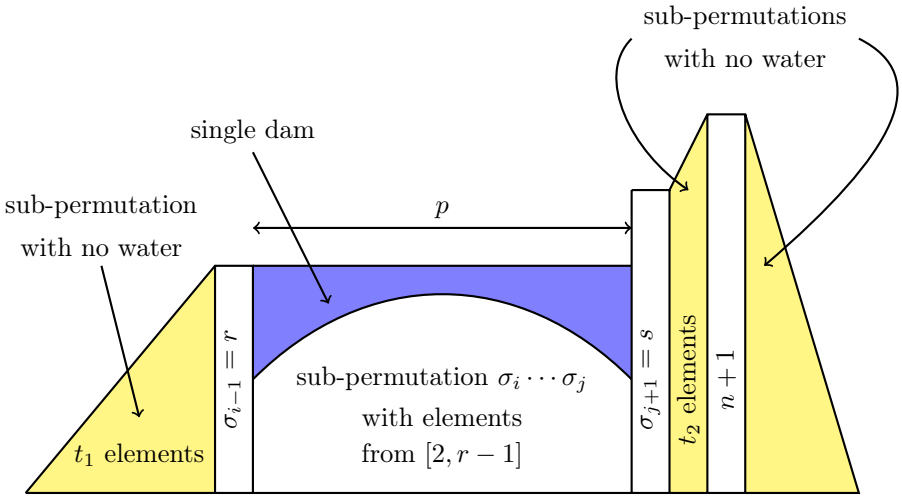


Figure 6: Permutation with one dam only, after raising but before adding 1

Let $C_1(n,p)$ be the *total* capacity taken over all permutations of $[n]$ with one dam of width p. Now let us obtain *all* one-dam permutations of $[n+1]$ of width $p+1$ from all possible precursors in $\mathcal{S}_n$. Each one-dam member of $\mathcal{S}_{n+1}$ of width $p+1$ can be obtained *in a unique way* from a certain subset of $\mathcal{S}_n$ by the following modified two-step procedure:

- Raising such permutations by one,
- Adding 1 to these permutations in every possible way that results in a one-dam permutation of $[n+1]$.

Let us first write a recursion for $C_1(n+1,1)$. We consider the following cases:
First case, where we add the element 1 to any raised unimodal permutation at all points other than the ends.
Second case, where there is a single dam of width one both before and after adding the element 1 to either end of a raised permutation.

So for the first case, we fix a raised unimodal permutation. The total contribution of adding the element 1 in any of the specified positions is

$$1+2+3+\cdots+(n-1) = \binom{n}{2}.$$

There are precisely

$$\binom{n-1}{0} + \binom{n-1}{1} + \cdots \binom{n-1}{n-1} = 2^{n-1}$$

unimodal permutations of length n. Hence, the contribution towards $C_1(n+1,1)$ is $2^{n-1}\binom{n}{2}$.

For the second case, the contribution is seen to be $2C_1(n,1)$. Combining the prior two cases, we have the recurrence

$$C_1(n+1,1) = 2C_1(n,1) + \binom{n}{2}2^{n-1}, \quad n \geq 1,$$

with the initial condition $C_1(1,1) = 0$, which yields the following result.

Proposition 3.1. *The total capacity of all one-dam permutations of $[n]$ with dam width 1 is*

$$C_1(n,1) = \frac{2^n\, n}{24}(n-1)(n-2).$$

We now write a recurrence for $C_1(n+1,p+1)$ where $p \geq 1$. For this, note that obtaining all one-dam permutations of length $n+1$ having width $p+1$ entails either

i) Adding 1 to any of the permutations counted in $C_1(n,p)$ (after first raising them) in any of the $p+1$ positions available inside the dam, or

ii) Adding 1 to either end of a permutation counted by $C_1(n,p+1)$ (after raising).

Now for case i) above, let r be the left bound of the dam in a one-dam permutation and s be the right bound. Assume for now that $r < s$ where $s \leq n$. (The case $s = n+1$ must be considered separately.) The width of the dam is p. Let there be t_1 increasing parts to the left of r and t_2+1 increasing parts to the right of s of which the last part must be $n+1$.

We note the following restrictions:

$$1 \leq p \leq n-3,$$

$$0 \leq t_1 \leq (r-1)-(p+1) = r-2-p,$$

$$0 \leq t_2 \leq n-s.$$

After raising and inserting the 1, we see that

$$p+1 < r \leq n-1$$

(because all $p+1$ elements of the new wider dam must be $< r$).

When we add 1 to the dam (in any of the $p+1$ possible positions), the additional capacity above the 1 is $r-1$. There are $\binom{r-2-p}{t_1}$ and $\binom{n-s}{t_2}$ ways to choose t_1 and t_2 elements, respectively, to form the increasing sequences. There are $\binom{r-2}{p}p!$ ways to choose and order the p elements in the dam prior to inserting 1. Thus, the additional contribution for permutations enumerated by $C_1(n,p)$ with given parameters r and s as stated is

$$\sum_{t_1=0}^{r-2-p}\sum_{t_2=0}^{n-s}\binom{r-2}{p}p!(r-1)(p+1)\binom{r-2-p}{t_1}\binom{n-s}{t_2}$$

$$= \binom{r-2}{p}(r-1)(p+1)!2^{n-s}\sum_{t_1=0}^{r-2-p}\binom{r-2-p}{t_1}$$
$$= \binom{r-2}{p}(r-1)(p+1)!2^{n-s}\,2^{r-2-p}. \tag{3.1}$$

Summing (3.1) over all possible values of s yields

$$\sum_{s=r+1}^{n}\binom{r-2}{p}(r-1)(p+1)!2^{n-s+r-2-p}$$
$$= (p+1)!\binom{r-2}{p}(r-1)2^{n-2-p}\left(1-2^{r-n}\right). \tag{3.2}$$

Finally, summing (3.2) over all possible values of r, the total additional capacity is

$$\sum_{r=p+2}^{n-1}(p+1)!\binom{r-2}{p}(r-1)2^{n-2-p}\left(1-2^{r-n}\right)$$
$$= (p+1)!2^{n-2-p}\sum_{r=p+2}^{n-1}\binom{r-2}{p}(r-1)\left(1-2^{r-n}\right). \tag{3.3}$$

Now for the case $s = n+1$, the restrictions are

$$1 \le p \le n-2,$$

$$0 \le t_1 \le (r-1)-(p+1) = r-2-p.$$

Considering all possible values of r and t_1, the additional contribution for permutations in the case $s = n+1$ is

$$\sum_{r=p+2}^{n}\sum_{t_1=0}^{r-2-p}\binom{r-2}{p}p!(r-1)(p+1)\binom{r-2-p}{t_1}$$
$$= \sum_{r=p+2}^{n}\binom{r-2}{p}(r-1)(p+1)!2^{r-2-p}. \tag{3.4}$$

Finding the total capacity requires taking into account the cases when $r > s$ and exploiting the obvious symmetry (i.e., multiplying by 2). Thus, by (3.3) and (3.4), the total additional capacity in case i) above is

$$(p+1)!2^{n-1-p}\sum_{r=p+2}^{n-1}\binom{r-2}{p}(r-1)\left(1-2^{r-n}\right)$$
$$+\sum_{r=p+2}^{n}\binom{r-2}{p}(r-1)(p+1)!2^{r-1-p}$$

$$= (p+1)!2^{n-1-p}\sum_{r=p+2}^{n}\binom{r-2}{p}(r-1) = (p+1)(p+1)!2^{n-1-p}\binom{n}{p+2}, \quad (3.5)$$

where we have made use of [13, p. 174] to obtain the last equality.

The original total capacity from i) is

$$(p+1)C_1(n,p). \quad (3.6)$$

Case ii) leads to a contribution towards $C_1(n+1,p+1)$ of

$$2C_1(n,p+1). \quad (3.7)$$

So adding (3.5), (3.6) and (3.7), we have the recurrence:

$$\begin{aligned} C_1(n+1,p+1) &= 2C_1(n,p+1) + (p+1)C_1(n,p) \\ &\quad + (p+1)(p+1)!2^{n-1-p}\binom{n}{p+2}. \end{aligned} \quad (3.8)$$

We have the following explicit formula for $C_1(n,p)$.

Theorem 3.2. *The total capacity of all one-dam permutations of $[n]$ with dam width p is*

$$C_1(n,p) = \frac{p}{p+2}2^{n-2-p}\frac{n!}{(n-2-p)!},$$

for $1 \le p \le n-2$.

Proof. We prove the result for a given $n \ge 3$ and all $p \in [n-2]$ by induction on n. The $n=3$ case is clear since $C_1(3,1) = 2$. If $n \ge 3$ and $p \ge 1$, then the formula for $C_1(n+1,p+1)$ follows from (3.8) and the induction hypothesis, upon considering separately the cases when $p \le n-3$ and $p = n-2$. By Proposition 3.1, the formula holds for $p = 1$ and all $n \ge 3$, which fully establishes the $n+1$ case and completes the induction. □

Remark 3.3. From Theorem 3.2, we obtain the generating function

$$\sum_{n \ge p+2} C_1(n,p)\,x^n = \frac{p(p+1)!x^{p+2}}{(1-2x)^{p+3}}, \quad p \ge 1.$$

Below is an array of values for $C_1(n,p)$ for small n and p:

$$[C_1(n,p)]_{n\ge 3, p\ge 1} = \begin{pmatrix} 2 & 0 & 0 & 0 & 0 & 0 \\ 16 & 12 & 0 & 0 & 0 & 0 \\ 80 & 120 & 72 & 0 & 0 & 0 \\ 320 & 720 & 864 & 480 & 0 & 0 \\ 1120 & 3360 & 6048 & 6720 & 3600 & 0 \\ 3584 & 13440 & 32256 & 53760 & 57600 & 30240 \end{pmatrix}.$$

Corollary 3.4. *The total capacity of one-dam permutations of* $[n]$ *is*

$$C_1(n) = \sum_{p=1}^{n-2} \frac{p}{p+2} 2^{n-2-p} \frac{n!}{(n-2-p)!}.$$

The values of $C_1(n)$ for $1 \le n \le 12$ are

$$0, 0, 2, 28, 272, 2384, 20848, 190880, 1871808, 19832448, 227360256, 2814303232.$$

4. Total number of one-dam permutations

In this section, we find the number of permutations of $[n]$ that have exactly one dam. Let $N(n,p)$ be the number of one-dam permutations of size n with width p. In order to obtain a recursion for $N(n+1,p+1)$ in terms of $N(n,p)$, we apply the same two-step procedure as before. We again consider separately the cases $p = 1$ and $p > 1$.

4.1. Case where $p = 1$

First, we add 1 at all points other than the ends to a raised unimodal permutation; then the contribution to the number of permutations is $(n-1)2^{n-1}$.

Next, we add 1 to the ends of a one-dam permutation, which yields a contribution of $2N(n,1)$. Combining the prior cases gives

$$N(n+1,1) = (n-1)2^{n-1} + 2N(n,1), \quad n \ge 1,$$

with initial condition $N(1,1) = 0$.

Solving this first order linear recursion gives the following result.

Proposition 4.1. *The number of one-dam permutations of* $[n]$ *with dam width* 1 *is*

$$N(n,1) = 2^{n-3}(n-1)(n-2).$$

4.2. Case where $p > 1$

First, we add 1 to a permutation counted in $N(n,p)$ in any of the $p+1$ positions within the dam, which gives a contribution of $(p+1)N(n,p)$. Otherwise, add the 1 to either end of a permutation counted by $N(n,p+1)$.

Thus, the recursion (3.8) is replaced by

$$N(n+1,p+1) = (p+1)N(n,p) + 2N(n,p+1). \tag{4.1}$$

One then has the following explicit formula for $N(n,p)$.

Theorem 4.2. *The number of one-dam permutations of* $[n]$ *with dam width* p *is*

$$N(n,p) = \frac{1}{p+1} 2^{n-1-p} \frac{(n-1)!}{(n-2-p)!},$$

for $1 \le p \le n-2$.

Proof. This is shown by induction on n as before using (4.1) and Proposition 4.1. □

Remark 4.3. From Theorem 4.2, we obtain the generating function

$$\sum_{n \ge p+2} N(n,p)\, x^n = \frac{2p! x^{p+2}}{(1-2x)^{p+2}}, \quad p \ge 1.$$

Below are the values for $N(n,p)$ for small n and p:

$$[N(n,p)]_{n\ge 3, p\ge 1} = \begin{pmatrix} 2 & 0 & 0 & 0 & 0 & 0 \\ 12 & 4 & 0 & 0 & 0 & 0 \\ 48 & 32 & 12 & 0 & 0 & 0 \\ 160 & 160 & 120 & 48 & 0 & 0 \\ 480 & 640 & 720 & 576 & 240 & 0 \\ 1344 & 2240 & 3360 & 4032 & 3360 & 1440 \end{pmatrix}.$$

Corollary 4.4. *The number of permutations of* $[n]$ *with one dam is*

$$N(n) = \sum_{p=1}^{n-2} \frac{1}{p+1} 2^{n-1-p} \frac{(n-1)!}{(n-2-p)!}.$$

The values of $N(n)$ for $1 \le n \le 12$ are

$$0, 0, 2, 16, 92, 488, 2656, 15776, 105696, 806592, 6974592, 67573504.$$

5. Asymptotics for $C_1(n)$ and $N(n)$

5.1. Asymptotics for $C_1(n)$

In order to find the asymptotic average capacity for one-dam permutations of $[n]$, we need asymptotic estimates of the quantities $C_1(n)$ and $N(n)$ in Corollaries 3.4 and 4.4.

We first find the maximum value of $C_1(n,p)$ over p for a fixed n. For this, we compute the ratio $C_1(n,p+1)/C_1(n,p)$ and determine where it is greater than or less than one.

By the formula $C_1(n,p) = \frac{p}{p+2} 2^{n-2-p} \frac{n!}{(n-2-p)!}$ from Theorem 3.2, we have

$$\frac{C_1(n,p+1)}{C_1(n,p)} = \frac{(n-2-p)(p+1)(p+2)}{2p(p+3)}.$$

Since $\frac{(p+1)(p+2)}{p(p+3)} > 1$, the ratio $\frac{C_1(n,p+1)}{C_1(n,p)}$ exceeds 1 if $p \leq n-4$. Comparing directly $C_1(n,n-3) = \frac{2(n-3)n!}{n-1}$ and $C_1(n,n-2) = (n-2)(n-1)!$, we have $C_1(n,n-3) > C_1(n,n-2)$ if $n \geq 4$, which we will assume. Thus, the size of the largest term is given by $C_1(n,n-3)$.

We represent the general term $C_1(n,p)$ for $p \leq n-3$ by $C(n,n-3-j)$, where j runs from 0 to $n-4$. Thus, the ratio of the general term $C_1(n,p)$ to the maximum term $C_1(n,n-3)$ is

$$\frac{C_1(n,n-3-j)}{C_1(n,n-3)} = \frac{2^j(3+j-n)(n-1)}{(1+j-n)(j+1)!(n-3)}.$$

At this stage, the final term where $p = n-2$ is omitted and will be reintroduced later.

Summing over all possible values of j yields

$$\sum_{j=0}^{n-4} \frac{2^j(3+j-n)(n-1)}{(1+j-n)(j+1)!(n-3)} = \frac{n-1}{n-3} \sum_{j=0}^{n-4} \frac{2^j(3+j-n)}{(1+j-n)(j+1)!}. \tag{5.1}$$

To estimate this sum, we perform a series expansion on the summand

$$\frac{2^j(3+j-n)}{(1+j-n)(j+1)!} = \frac{2^j}{(1+j)!} - \frac{2^{1+j}}{(1+j)!n} + O\left(\frac{1}{n^2}\right).$$

We shall replace the original summand by $\frac{2^j}{(1+j)!} - \frac{2^{1+j}}{(1+j)!n}$.

Thus, consider the sum

$$\sum_{j=0}^{n-4} \left(\frac{2^j}{(1+j)!} - \frac{2^{1+j}}{(1+j)!n} \right). \tag{5.2}$$

The terms

$$\sum_{j=0}^{n-4} \frac{2^{1+j}}{(1+j)!n}$$

may be ignored as they only make a small contribution for large n since

$$\frac{1}{n} \sum_{j=0}^{n-4} \frac{2^{1+j}}{(1+j)!} < \frac{e^2}{n}.$$

Therefore, the sum in (5.2) can be approximated by an infinite sum

$$\sum_{j=0}^{\infty} \frac{2^j}{(1+j)!},$$

since the terms for $j \geq n-3$ are exponentially small. Thus, the sum in (5.2) equals

$$\sum_{j=0}^{\infty} \frac{2^j}{(1+j)!} + O\left(\frac{1}{n}\right) = \frac{e^2-1}{2} + O\left(\frac{1}{n}\right).$$

Finally, we include the factor $\frac{n-1}{n-3}$ from equation (5.1) above that was left out, multiply by the largest term $C(n, n-3) = \frac{2(n-3)n!}{n-1}$ and then add the missing last term when $p = n-2$ to obtain

$$(e^2-1)n!\left(1+O\left(\frac{1}{n}\right)\right)+n!\left(1+O\left(\frac{1}{n}\right)\right),$$

which yields the following result.

Theorem 5.1. *As $n \to \infty$, the asymptotic expression for $C_1(n)$, the total capacity of all one-dam permutations of $[n]$, is given by*

$$C_1(n) = e^2\, n!\left(1+O\left(\frac{1}{n}\right)\right).$$

5.2. Asymptotics for $N(n)$

One can also find an asymptotic expression for the number of permutations of $[n]$ with one dam, following the method used for $C_1(n)$. By Theorem 4.2, the ratio of $N(n, p+1)$ to $N(n,p)$ simplifies to $\frac{(n-2-p)(1+p)}{2(2+p)}$. Since $\frac{2}{3} \le \frac{1+p}{2+p} < 1$, we have $\frac{N(n,p+1)}{N(n,p)} \ge 1$ if $1 \le p \le n-5$ and $\frac{N(n,p+1)}{N(n,p)} < 1$ if $p = n-3$ or $n-4$. (Note that there is equality in the inequality $\frac{N(n,p+1)}{N(n,p)} \ge 1$ if and only if $n = 6$ and $p = 1$.) Thus, the maximum value of $N(n,p)$ for $1 \le p \le n-2$ where $n \ge 5$ is given by

$$N(n, n-4) = \frac{4(n-1)!}{n-3}.$$

This time however there are two cases to add at the end, namely, when $p = n-2$ and $p = n-3$. We consider the ratio $\frac{N(n,n-4-j)}{N(n,n-4)}$ of the general term to the largest term for $0 \le j \le n-5$ and sum over j to get

$$\sum_{j=0}^{n-5} \frac{N(n, n-4-j)}{N(n, n-4)} = \sum_{j=0}^{n-5} \frac{2^{1+j}(n-3)}{(n-3-j)(j+2)!}.$$

Similar to before, we have

$$\sum_{j=0}^{n-5} \frac{2^{1+j}(n-3)}{(n-3-j)(j+2)!} = \sum_{j=0}^{n-5} \left(\frac{2^{1+j}}{(2+j)!} + \frac{2^{1+j} j}{(2+j)!n}\right) + O\left(\frac{1}{n}\right).$$

We approximate this last sum, ignoring the second part, by the infinite sum

$$\sum_{j=0}^{\infty} \frac{2^{1+j}}{(2+j)!} = \frac{e^2-3}{2}.$$

Multiplying by the largest term and adding the two missing terms for $p = n-2$ and $p = n-3$, we have

$$\left(\frac{e^2-3}{2}\,\frac{4(n-1)!}{n} + \frac{4(n-1)!}{n} + \frac{2(n-1)!}{n}\right)\left(1+O\left(\frac{1}{n}\right)\right),$$

which yields the following result.

Theorem 5.2. *As $n \to \infty$, the asymptotic expression for $N(n)$, the number of permutations of $[n]$ with one dam, is given by*

$$N(n) = \frac{2e^2(n-1)!}{n}\left(1+O\left(\frac{1}{n}\right)\right).$$

Finally, dividing the result of Theorem 5.1 by that of Theorem 5.2 yields the following estimate.

Theorem 5.3. *As $n \to \infty$, the average capacity for the permutations of $[n]$ with one dam is*

$$\frac{C_1(n)}{N(n)} = \frac{n^2}{2}\left(1+O\left(\frac{1}{n}\right)\right).$$

6. Combinatorial proofs

In this section, we provide bijective proofs of Theorems 3.2 and 4.2 above. Since our combinatorial proof of the former makes use of ideas from the latter, we first argue the latter.

6.1. Combinatorial proof of Theorem 4.2.

Equivalently, we show $N(n,p) = 2^{n-1-p}p!\binom{n-1}{p+1}$. To do so, first let $S = \{s_1 < s_2 < \cdots < s_{p+1}\}$ be an arbitrary subset of $[n-1]$ of size $p+1$. We reorder the elements $s_1, s_2, \ldots, s_p$ according to an arbitrary permutation α of $[p]$ as $s_{\alpha(1)}, s_{\alpha(2)}, \ldots, s_{\alpha(p)}$, which we will denote by α^*. Next, we assign to each member of $[n]-S$ either a or b. From this configuration enumerated by $2^{n-1-p}p!\binom{n-1}{p+1}$, we create a permutation $\pi = \pi_1\pi_2\cdots\pi_n$ of $[n]$ having a single dam $\pi_{i-1}\pi_i\cdots\pi_j\pi_{j+1}$ of width p such that the section $\pi_i\cdots\pi_j$ is a permutation of $\{s_1,\ldots,s_p\}$ and $s_{p+1} = \min\{\pi_{i-1}, \pi_{j+1}\}$. In creating π, we will first form the subsequence Q of π comprising the elements of $S \cup [s_{p+1}+1, n]$; note that Q must consist of consecutive letters of π.

Consider the sequence $c = c_1c_2\cdots c_\ell$ of letters in $\{a,b\}$ assigned to the elements $s_{p+1}+1, s_{p+1}+2, \ldots, n$, where $\ell = n - s_{p+1}$. If $c = a^\ell$ or $c = b^\ell$, then let Q be given by $Q = n(n-1)\cdots(s_{p+1}+1)\alpha^* s_{p+1}$ or $Q = s_{p+1}\alpha^*(s_{p+1}+1)\cdots(n-1)n$, respectively. If $c = b^{\ell-1}a$ or $c = a^{\ell-1}b$, then let $Q = (s_{p+1}+1)\cdots(n-1)n\alpha^* s_{p+1}$ or $Q = s_{p+1}\alpha^* n(n-1)\cdots(s_{p+1}+1)$. So assume c starts with $b^t a$ or $a^t b$, where

$1 \le t \le \ell - 2$. We consider cases based on the final letter c_ℓ to define Q. First assume $c_\ell = a$. If c starts with $b^t a$ for some $1 \le t \le \ell - 2$, then let

$$Q = (s_{p+1} + t + 1)\beta' n \beta''(s_{p+1} + t) \cdots (s_{p+1} + 1)\alpha^* s_{p+1},$$

where β' is increasing, β'' is decreasing and $\beta' \cup \beta'' = [s_{p+1} + t + 2, n - 1]$, with membership in the string β' or β'' dependent on whether a or b is assigned to the element in question. If c starts with $a^t b$, then let

$$Q = (s_{p+1} + 1) \cdots (s_{p+1} + t)\beta' n \beta''(s_{p+1} + t + 1)\alpha^* s_{p+1},$$

where β' and β'' are as before. Now assume $c_\ell = b$. If c starts with $b^t a$, then let Q be obtained by reversing the Q from the corresponding case above when $c_\ell = a$. Likewise, if c starts with $a^t b$, then reverse Q from the corresponding case when $c_\ell = a$.

Finally, if $x \in [s_{p+1}] - S$, then either place x before Q if x is assigned a or after Q if x is assigned b such that any elements of $[s_{p+1}] - S$ before (after) Q occur in increasing (decreasing) order. Let π be the permutation of $[n]$ obtained by applying the operations described above. One may verify that π contains a single dam of width p and that the procedure above is reversible.

6.2. Proof of Theorem 3.2.

Let $\mathcal{N}(n, p)$ denote the set of permutations enumerated by $N(n, p)$. To compute the sum of the capacities of all members of $\mathcal{N}(n, p)$, it is enough to consider the contribution from the first letter of each dam, by symmetry, and multiply the result by p. Let $\lambda \in \mathcal{N}(n, p)$ be formed in the manner described above from an ordered triple (S, α, d), where S and α are as before with $S = \{s_1 < s_2 < \cdots < s_{p+1}\}$ and d is a binary sequence in $\{a, b\}$ of length $n - 1 - p$. Let λ' be the member of $\mathcal{N}(n, p)$ obtained from the triple $(S', \gamma\alpha, d)$, where γ denotes the complement operation (i.e., $\gamma(i) = p + 1 - i$ for all $i \in [p]$) and $S' = \{s_{p+1}\} \cup \{s_{p+1} - s_i : 1 \le i \le p\}$. Note that $\lambda = \lambda'$ if and only if $p = 1$, s_2 is even and $s_1 = \frac{s_2}{2}$, which is permitted. Taken together, λ and λ' contribute s_{p+1} towards the total capacity of all members of $\mathcal{N}(n, p)$ for all λ (considering only the contribution of the first position within a dam). So we must replace $\binom{n-1}{p+1}$ as the enumerator of S with the sum $\sum_{r=p+1}^{n-1} \binom{r-1}{p} r = \binom{n}{p+2}(p + 1)$, where r denotes s_{p+1}; this identity is shown below bijectively. Upon considering separately the cases when $\lambda = \lambda'$ and $\lambda \ne \lambda'$, it is seen that the contribution of each λ is counted twice (note that if $p > 1$, then $\lambda \ne \lambda'$ for all λ with the mapping $\lambda \mapsto \lambda'$ an involution for all p). Thus, multiplying by p, the total capacity of all members of $\mathcal{N}(n, p)$ is given by

$$\frac{1}{2}\left(2^{n-1-p} p \cdot p! \binom{n}{p+2}(p + 1)\right) = \frac{p}{p+2} 2^{n-2-p} \frac{n!}{(n - 2 - p)!},$$

as desired.

For completeness, we provide a bijective proof of the identity

$$\sum_{r=p+1}^{n-1} \binom{r-1}{p} r = \binom{n}{p+2}(p+1), \quad 1 \le p \le n-2, \tag{6.1}$$

used above, since the authors were unable to find such a proof in the literature. Note that the right side of (6.1) clearly counts members of the set $\mathcal{A}$ consisting of "marked" subsets of $[n]$ of size $p+2$ wherein one of the elements, not the largest, is marked. To complete the proof, we construct another set $\mathcal{B}$ enumerated by the left side of (6.1) as well as a bijection between the sets $\mathcal{B}$ and $\mathcal{A}$. Given $p+1 \le r \le n-1$, let $\mathcal{B}_r$ denote the set of configurations wherein the members of $[r]$ are written in a row, exactly $p+1$ numbers are circled, among them r itself, and a dot is placed directly prior to some member of $[r]$. Let $\mathcal{B} = \bigcup_{r=p+1}^{n-1} \mathcal{B}_r$. To define a bijection from $\mathcal{B}$ to $\mathcal{A}$, renumber the elements to the right of the dot where the dot now receives a number (the number assigned the position of the dot will become the marked element of $A \in \mathcal{A}$). Note that the element r becomes $r+1$ and thus the largest element of A.

References

[1] J.-L. Baril, T. Mansour, A. Petrossian: *Equivalence classes of permutations modulo excedances*, J. Comb. 5.4 (2014), pp. 453–469, doi: 10.4310/joc.2014.v5.n4.a4.

[2] A. Baxter: *Refining enumeration schemes to count according to permutation statistics*, Electron. J. Combin. 21.2 (2014), Art. Num. 2.50.

[3] R. Biagioli: *Major and descent statistics for the even-signed permutation group*, Adv. in Appl. Math. 31.1 (2003), pp. 163–179, doi: 10.1016/s0196-8858(02)00561-4.

[4] A. Blecher, C. Brennan, A. Knopfmacher: *Capacity of words*, J. Combin. Math. Combin. Comput. 107 (2018), pp. 245–258.

[5] A. Blecher, C. Brennan, A. Knopfmacher: *The water capacity of integer compositions*, Online J. Anal. Comb. 13 (2018), Art. Num. 6.

[6] M. Bóna: *Combinatorics of Permutations*, 2nd ed., London: CRC Press, Taylor and Francis Group, 2012, doi: 10.1201/b12210.

[7] C.-O. Chow, S.-M. Ma, T. Mansour, M. Shattuck: *Counting permutations by cyclic peaks and valleys*, Ann. Math. Inform. 43 (2014), pp. 43–54.

[8] C.-O. Chow, T. Mansour: *Asymptotic probability distributions of some permutation statistics for the wreath product $C_r \sim S_n$*, Online J. Anal. Comb. 7 (2012), Art. Num. 2.

[9] S. Corteel, I. Gessel, C. Savage, H. Wilf: *The joint distribution of descent and major index over restricted sets of permutations*, Ann. Comb. 11.3-4 (2007), pp. 375–386, doi: 10.1007/s00026-007-0325-y.

[10] E. Deutsch, W. P. Johnson: *Create your own permutation statistics*, Math. Mag. 77.2 (2004), pp. 130–134, doi: 10.1080/0025570x.2004.11953238.

[11] E. Deutsch, S. Kitaev, J. Remmel: *Equidistribution of descents, adjacent pairs and place-value pairs on permutations*, J. Integer Seq. 12 (2009), Art. Num. 09.5.1.

[12] A. Goyt, D. Mathisen: *Permutation statistics and q-Fibonacci numbers*, Electron. J. Combin. 16 (2009), Art. Num. 101.

[13] R. Graham, D. Knuth, O. Patashnik: *Concrete Mathematics: A Foundation for Computer Science*, 2nd ed., Boston: Addison-Wesley, 1994.

[14] J. Hall, J. Remmel: *Counting descent pairs with prescribed tops and bottoms*, J. Combin. Theory Ser. A 115 (2008), pp. 693–725, doi: 10.1016/j.jcta.2007.09.001.

[15] S. Kitaev: *Patterns in Permutations and Words*, Monographs in Theoretical Computer Science - an EATCS series, Berlin: Springer-Verlag, 2011, doi: 10.1007/978-3-642-17333-2.

[16] T. Mansour, M. Shattuck: *Counting water cells in bargraphs of compositions and set partitions*, Appl. Anal. Discrete Math. 12 (2018), pp. 413–438, doi: 10.2298/aadm170428010m.

[17] A. Mendes, J. Remmel: *Permutations and words counted by consecutive patterns*, Adv. in Appl. Math. 37.4 (2006), pp. 443–480, doi: 10.1016/j.aam.2005.09.005.

[18] A. Robertson, D. Saracino, D. Zeilberger: *Refined restricted permutations*, Ann. Comb. 6 (2002), pp. 427–444, doi: 10.1007/s000260200015.

Annales Mathematicae et Informaticae
***50** (2019) pp. 57–70*
DOI: 10.33039/ami.2019.06.001
http://ami.uni-eszterhazy.hu

Fibonacci numbers which are products of two balancing numbers

Fatih Erduvan, Refik Keskin

Sakarya University, Mathematics Department, Sakarya, Turkey
erduvanmat@hotmail.com
rkeskin@sakarya.edu.tr

Submitted: September 11, 2018
Accepted: June 9, 2019
Published online: June 26, 2019

Abstract

The Fibonacci sequence (F_n) is defined by $F_0 = 0$, $F_1 = 1$ and $F_n = F_{n-1} + F_{n-2}$ for $n \geq 2$. The balancing number sequence (B_n) is defined by $B_0 = 0, B_1 = 1$ and $B_n = 6B_{n-1} - B_{n-2}$ for $n \geq 2$. In this paper, we find all Fibonacci numbers which are products of two balancing numbers. Also we found all balancing numbers which are products of two Fibonacci numbers. More generally, taking k, m, m as positive integers, it is proved that $F_k = B_m B_n$ implies that $(k, m, n) = (1, 1, 1), (2, 1, 1)$ and $B_k = F_m F_n$ implies that $(k, m, n) = (1, 1, 1), (1, 1, 2), (1, 2, 2), (2, 3, 4)$.

Keywords: Fibonacci number, balancing number, Diophantine equations, linear forms in logarithms.

MSC: 11B39, 11J86, 11D61

1. Introduction

The Fibonacci sequence (F_n) is defined as $F_0 = 0,\ F_1 = 1$ and $F_n = F_{n-1} + F_{n-2}$ for $n \geq 2$. F_n is called the n-th Fibonacci number. It well known that

$$F_n = \frac{\alpha^n - \beta^n}{\sqrt{5}}$$

for every $n \geq 0$, where $\alpha = \frac{1+\sqrt{5}}{2}$ and $\beta = \frac{1-\sqrt{5}}{2}$, which are the roots of the characteristic equations $x^2 - x - 1 = 0$. It is well known that

$$\alpha^{n-2} \leq F_n \leq \alpha^{n-1} \tag{1.1}$$

for all $n \geq 1$. The inequality (1.1) can be proved by induction. It can be seen that $1 < \alpha < 2$ and $-1 < \beta < 0$. For more information about the Fibonacci sequence and its applications, one can see [7]. A positive integer n is called a balancing number if the equation

$$1 + 2 + \cdots + (n-1) = (n+1) + \cdots + (n+r)$$

holds for some positive integer r. The sequence of balancing numbers (B_n) satisfys recurrence relation $B_n = 6B_{n-1} - B_{n-2}$ for $n \geq 2$ with initial conditions $B_0 = 0,\ B_1 = 1$. B_n is called the n-th balancing number. We have the Binet formula

$$B_n = \frac{\lambda^n - \delta^n}{4\sqrt{2}},$$

where $\lambda = 3 + 2\sqrt{2}$ and $\delta = 3 - 2\sqrt{2}$, which are the roots of the characteristic equations $x^2 - 6x + 1 = 0$. Therefore,

$$B_n < \frac{\lambda^n}{4\sqrt{2}}. \tag{1.2}$$

It can be seen that $5 < \lambda < 6$, $0 < \delta < 1$ and $\lambda\delta = 1$. Moreover, it holds that

$$\lambda^{n-1} \leq B_n < \lambda^n \tag{1.3}$$

for all $n \geq 1$. This inequality can be proved by noting the facts that $\lambda^n = \lambda B_n - B_{n-1}$ and $B_n - \lambda^{n-1} = B_n - (\lambda B_{n-1} - B_{n-2}) = 6B_{n-1} - B_{n-2} - (\lambda B_{n-1} - B_{n-2}) = (6-\lambda)B_{n-1} > 0$ for all $n \geq 2$. Clearly, the identity (1.3) holds for $n = 1$. For more information about the sequence of balancing numbers, see [6, 9, 10]. A different definition is given by Szakács [12]. A positive integer n is called a multiplying balancing number if the equation

$$1 \cdot 2 \cdots (n-1) = (n+1)(n+2) \cdots (n+r)$$

holds for some positive integer r. The number r is called the balancer corresponding to multiplying balancing number n. In [12], it is shown that the only multiplying balancing number is $n = 7$ with the balancer $r = 3$. For some other generalization of balancing numbers, the interested readers can consult [11] and the references there. In [3], the authors have found all Fibonacci numbers or Pell numbers which are products of two numbers from the other sequence. Taking k, m and n are positive integer, they showed that $F_k = P_m P_n$ implies that $k = 1, 2, 3, 5, 12$ and $P_k = F_m F_n$ implies that $k = 1, 2, 3, 7$, where (P_n) is the Pell sequence defined by $P_0 = 0, P_1 = 1$ and $P_n = 2P_{n-1} + P_{n-2}$ for $n \geq 2$. In this study, we determine all solutions of the equation

$$F_k = B_m B_n \tag{1.4}$$

and

$$B_k = F_m F_n \tag{1.5}$$

in positive integers k, n, m. More generally, taking k, m, m as positive integers, it is proved that $F_k = B_m B_n$ implies that $(k, m, n) = (1, 1, 1), (2, 1, 1)$ and $B_k = F_m F_n$ implies that $(k, m, n) = (1, 1, 1), (1, 1, 2), (1, 2, 2), (2, 3, 4)$.

Our study can be viewed as a continuation of the previous work on this subject. We follow the approach and the method presented in [3]. In Section 2, we introduce necessary lemmas and theorems. Then in Section 3, we prove our main theorem.

2. Auxiliary results

In [3], in order to solve Diophantine equations of the form (1.4) and (1.5), the authors have used Baker's theory of lower bounds for a nonzero linear form in logarithms of algebraic numbers. Since such bounds are of crucial importance in effectively solving of Diophantine equations of the similar form, we start with recalling some basic notions from algebraic number theory.

Let η be an algebraic number of degree d with minimal polynomial

$$a_0 x^d + a_1 x^{d-1} + \cdots + a_d = a_0 \prod_{i=1}^{d} \left(X - \eta^{(i)} \right) \in \mathbb{Z}[x],$$

where the a_i's are relatively prime integers with $a_0 > 0$ and $\eta^{(i)}$'s are conjugates of η. Then

$$h(\eta) = \frac{1}{d} \left(\log a_0 + \sum_{i=1}^{d} \log \left(\max \left\{ |\eta^{(i)}|, 1 \right\} \right) \right) \tag{2.1}$$

is called the logarithmic height of η. In particularly, if $\eta = a/b$ is a rational number with $\gcd(a, b) = 1$ and $b > 1$, then $h(\eta) = \log(\max\{|a|, b\})$.

The following properties of logarithmic height are found in many works stated in the references:

$$h(\eta \pm \gamma) \leq h(\eta) + h(\gamma) + \log 2, \tag{2.2}$$

$$h(\eta \gamma^{\pm 1}) \leq h(\eta) + h(\gamma), \tag{2.3}$$

$$h(\eta^s) = |s| h(\eta). \tag{2.4}$$

The following theorem is deduced from Corollary 2.3 of Matveev [8] and provides a large upper bound for the subscript n in the equations (1.4) and (1.5)(also see Theorem 9.4 in [2]).

Theorem 2.1. *Assume that $\gamma_1, \gamma_2, \ldots, \gamma_t$ are positive real algebraic numbers in a real algebraic number field $\mathbb{K}$ of degree D, $b_1, b_2, \ldots, b_t$ are rational integers, and*

$$\Lambda := \gamma_1^{b_1} \ldots \gamma_t^{b_t} - 1$$

is not zero. Then

$$|\Lambda| > \exp\left(-1.4 \cdot 30^{t+3} \cdot t^{4.5} \cdot D^2(1+\log D)(1+\log B)A_1A_2\dots A_t\right),$$

where

$$B \geq \max\{|b_1|,\dots,|b_t|\},$$

and $A_i \geq \max\{Dh(\gamma_i), |\log\gamma_i|, 0.16\}$ *for all* $i = 1,\dots,t$.

The following lemma was proved by Dujella and Pethő [5] and is a variation of a lemma of Baker and Davenport [1]. This lemma will be used to reduce the upper bound for the subscript n in the equations (1.4) and (1.5). In the following lemma, the function $||\cdot||$ denotes the distance from x to the nearest integer. That is, $||x|| = \min\{|x-n| : n \in \mathbb{Z}\}$ for any real number x.

Lemma 2.2. *Let* M *be a positive integer, let* p/q *be a convergent of the continued fraction of the irrational number* γ *such that* $q > 6M$, *and let* A, B, μ *be some real numbers with* $A > 0$ *and* $B > 1$. *Let* $\epsilon := ||\mu q|| - M||\gamma q||$. *If* $\epsilon > 0$, *then there exists no solution to the inequality*

$$0 < |u\gamma - v + \mu| < AB^{-w},$$

in positive integers u, v, *and* w *with*

$$u \leq M \quad \text{and} \quad w \geq \frac{\log(Aq/\epsilon)}{\log B}.$$

The following theorems are given in [2] and [4], respectively.

Theorem 2.3. *The only perfect powers in the Fibonacci sequence are* $F_0 = 0$, $F_1 = F_2 = 1$, $F_6 = 8$ *and* $F_{12} = 144$.

Theorem 2.4. *For any given positive integers* y *and* $l \geq 2$, *the equation* $B_m = y^l$ *has no solution for integers* $m \geq 2$.

3. Main theorems

Theorem 3.1. *The Diophantine equation* $F_k = B_mB_n$ *has only the solutions*

$$(k,m,n) = (1,1,1),(2,1,1)$$

in positive integers.

Proof. Assume that the equation $F_k = B_mB_n$ holds. If $m = n$, we have $F_k = B_n^2$, which is possible only for $k = 1, 2$, and $n = 1$ by Theorem 2.3. In this case, $(k,m,n) = (1,1,1),(2,1,1)$. Therefore, we assume that $1 \leq m < n$. Let $n \leq 30$. Then, by using the Mathematica program, we see that $k \leq 214$. In that case, with the help of Mathematica program, we obtain only the solutions $(k,m,n) =$

$(1,1,1),(2,1,1)$ in the range $1 \leq m < n \leq 30$. This takes a little time. From now on, assume that $n > 30$. Using the inequality (1.1) and (1.2) , we get the inequality

$$\alpha^{k-2} \leq F_k = B_m B_n < \lambda^{n+m}/32.$$

From this, it follows that

$$\alpha^k = \alpha^2 \alpha^{k-2} < 32\alpha^{k-2} < \lambda^{n+m} < (\alpha^4)^{n+m},$$

which yields to $k < 4(n+m) < 8n$. On the other hand, $\lambda^{m+n-2} \leq B_m B_n = F_k \leq \alpha^{k-1} < \lambda^{k-1}$ by (1.1) and (1.3). From this, we get $m+n-1 < k$, which implies that $k > n$.

Since

$$\frac{\alpha^k - \beta^k}{\sqrt{5}} = F_k = B_m B_n = \frac{\lambda^{n+m} + \delta^{n+m} - \lambda^n \delta^m - \lambda^m \delta^n}{32},$$

we get

$$\frac{\beta^k}{\sqrt{5}} - \frac{\lambda^n \delta^m + \lambda^m \delta^n - \delta^{n+m}}{32} = \frac{\alpha^k}{\sqrt{5}} - \frac{\lambda^{n+m}}{32}.$$

Taking absolute values, we obtain

$$\begin{aligned}
\left| \frac{\alpha^k}{\sqrt{5}} - \frac{\lambda^{n+m}}{32} \right| = \left| \frac{\beta^k}{\sqrt{5}} - \frac{\lambda^n \delta^m + \lambda^m \delta^n - \delta^{n+m}}{32} \right| &\leq \frac{|\beta|^k}{\sqrt{5}} + \frac{\lambda^n \delta^m + \lambda^m \delta^n + \delta^{n+m}}{32} \\
&= \frac{32\,|\beta|^k + \sqrt{5}\,(\lambda^{n-m} + \delta^{n-m} + \delta^{n+m})}{32\sqrt{5}} \\
&< \frac{\sqrt{5} + \sqrt{5}(\lambda^{n-m} + 2)}{32\sqrt{5}} < \frac{\sqrt{5} + 2\sqrt{5}\lambda^{n-m}}{32\sqrt{5}} \\
&< \frac{1 + 2\lambda^{n-m}}{32} < \frac{\lambda^{n-m+1}}{32},
\end{aligned}$$

where we have used the fact that $0 < \delta < 1$, $\lambda > 2$, $\lambda\delta = 1$, and $32\,|\beta|^k < \sqrt{5}$ for $k > n > 30$. If we divide both sides of the above inequality by $\frac{\lambda^{n+m}}{32}$, we get

$$\left| \frac{32}{\sqrt{5}} \alpha^k \lambda^{-(n+m)} - 1 \right| < \frac{1}{\lambda^{2m-1}}. \tag{3.1}$$

Now, let us apply Theorem 2.1 with $\gamma_1 := 32/\sqrt{5}$, $\gamma_2 := \alpha$, $\gamma_3 := \lambda$ and $b_1 := 1$, $b_2 := k$, $b_3 := -(n+m)$. Note that the numbers γ_1, γ_2, and γ_3 are positive real numbers and elements of the field $\mathbb{K} = \mathbb{Q}(\sqrt{2}, \sqrt{5})$. It is obvious that the degree of the field $\mathbb{K}$ is 4. So $D = 4$. Now, we show that $\Lambda_1 := \frac{32}{\sqrt{5}} \alpha^k \lambda^{-(n+m)} - 1$ is nonzero. For, if $\Lambda_1 = 0$, then we get

$$\alpha^k \lambda^{-(n+m)} = \alpha^k \delta^{n+m} = \sqrt{5}/32.$$

It is seen that $\sqrt{5}/32$ is not a algebraic integer although $\alpha^k\delta^{n+m}$ is an algebraic integer. This is a contradiction. Moreover, since

$$h(\gamma_1) = h(32/\sqrt{5}) = \frac{1}{2}(\log 5 + 2\log(32/\sqrt{5})) = 3.4657\ldots,$$
$$h(\gamma_2) = \frac{\log\alpha}{2} = \frac{0.4812\ldots}{2}$$

and

$$h(\gamma_3) = \frac{\log\lambda}{2} = \frac{1.76275\ldots}{2}$$

by (2.1), we can take $A_1 := 14,\ A_2 := 1$ and $A_3 = 3.6$. Also, since $k < 8n$, we can take $B := \max\{1, |k|, |-(n+m)|\} = 8n$. Thus, taking into account the inequality (3.1) and using Theorem 2.1, we obtain

$$\frac{1}{\lambda^{2m-1}} > |\Lambda_1| > \exp\left(-1.4\cdot 30^6\cdot 3^{4.5}\cdot 4^2(1+\log 4)(1+\log 8n)\,(14)\,(3.6)\right),$$

and so

$$(2m-1)\log\lambda < 1.4\cdot 30^6\cdot 3^{4.5}\cdot 4^2(1+\log 4)(1+\log 8n)\,(14)\,(3.6).$$

By a simple computation, it follows that

$$2m\log\lambda < 2.7554\cdot 10^{14}(1+\log 8n) + \log\lambda. \tag{3.2}$$

Now, we apply Theorem 2.1 a second time. Rearranging the equation $F_k = B_nB_m$ as

$$\frac{\beta^k}{\sqrt{5}B_m} - \frac{\delta^n}{4\sqrt{2}} = \frac{\alpha^k}{\sqrt{5}B_m} - \frac{\lambda^n}{4\sqrt{2}},$$

and taking absolute values, we obtain

$$\left|\frac{\alpha^k}{\sqrt{5}B_m} - \frac{\lambda^n}{4\sqrt{2}}\right| = \left|\frac{\beta^k}{\sqrt{5}B_m} - \frac{\delta^n}{4\sqrt{2}}\right| \le \frac{|\beta|^k}{\sqrt{5}B_m} + \frac{\delta^n}{4\sqrt{2}} < \frac{1}{\sqrt{5}B_m} + \frac{1}{4\sqrt{2}} < 1,$$

where we used the fact that $|\beta| < 1$ and $0 < \delta < 1$. Dividing both sides of the above inequality by $\lambda^n/4\sqrt{2}$, we get

$$\left|\frac{4\sqrt{2}\alpha^k\lambda^{-n}}{\sqrt{5}B_m} - 1\right| < \frac{4\sqrt{2}}{\lambda^n} < \frac{6}{\lambda^n}. \tag{3.3}$$

Taking $\gamma_1 := \alpha$, $\gamma_2 := \lambda$, $\gamma_3 := \sqrt{5}B_m/4\sqrt{2}$, and $b_1 := k$, $b_2 := -n$, $b_3 := -1$, we can apply Theorem 2.1. The numbers γ_1, γ_2, and γ_3 are positive real numbers and elements of the field $\mathbb{K} = \mathbb{Q}(\sqrt{2}, \sqrt{5})$ and so $D = 4$. In a similar manner, one can verify that $\Lambda_2 = 4\sqrt{2}\alpha^k\lambda^{-n}/B_m - 1 \neq 0$. Also, since $h(\gamma_1) = \frac{\log\alpha}{2} = \frac{0.4812\ldots}{2}$ and

$h(\gamma_2) = \frac{\log\lambda}{2} = \frac{1.76275\ldots}{2}$ by (2.1), we can take $A_1 := 1$ and $A_2 = 3.6$. The number $\sqrt{5}B_m/4\sqrt{2}$ is a root of the polynomial $32X^2 - 5B_m^2$. Thus, using the properties (2.2), (2.3) and (2.4), it is seen that

$$\begin{aligned} h(\gamma_3) &\le \frac{1}{2}\left(\log 32 + 2\log\left(\frac{\sqrt{5}B_m}{4\sqrt{2}}\right)\right) = \log(\sqrt{5}B_m) \le \log(\sqrt{5}\lambda^m/4\sqrt{2}) \\ &< m\log\lambda, \end{aligned}$$

by (1.2). So we can take $A_3 := 4m\log\lambda$. Since $k < 8n$, it follows that $B := 8n > \max\{|k|, |-n|, |-1|\}$. Thus, taking into account the inequality (3.3) and using Theorem 2.1, we obtain

$$\frac{6}{\lambda^n} > |\Lambda_2| > \exp\left((-C)(1+\log 4)(1+\log 8n)\,(3.6)\,4m\log\lambda\right),$$

or

$$n\log\lambda - \log 6 < C(1+\log 4)(1+\log 8n)\,(3.6)\,4m\log\lambda, \tag{3.4}$$

where $C = 1.4\cdot 30^6\cdot 3^{4.5}\cdot 4^2$. Inserting the inequality (3.2) into the last inequality, a computer search with Mathematica gives us that $n < 3.52\cdot 10^{31}$.

Now, let us try to reduce the upper bound on n by applying Lemma 2.2. Let

$$z_1 := k\log\alpha - (n+m)\log\lambda + \log(32/\sqrt{5}).$$

Then

$$|1 - e^{z_1}| < \frac{1}{\lambda^{2m-1}}$$

by (3.1). If $z_1 > 0$, then we have the inequality

$$|z_1| = z_1 < e^{z_1} - 1 = |1 - e^{z_1}| < \frac{1}{\lambda^{2m-1}}$$

since $x < e^x - 1$ for $x > 0$. If $z_1 < 0$, then

$$1 - e^{z_1} = |1 - e^{z_1}| < \frac{1}{\lambda^{2m-1}} < \frac{1}{2}.$$

From this, we get $e^{z_1} > \frac{1}{2}$ and therefore

$$e^{|z_1|} = e^{-z_1} < 2.$$

Consequently, we get

$$|z_1| < e^{|z_1|} - 1 = e^{|z_1|}\,|1 - e^{z_1}| < \frac{2}{\lambda^{2m-1}}.$$

In both cases, the inequality

$$|z_1| < \frac{2}{\lambda^{2m-1}}$$

holds. That is,

$$0 < \left| k \log \alpha - (n+m) \log \lambda + \log(32/\sqrt{5}) \right| < \frac{2}{\lambda^{2m-1}}.$$

Dividing this inequality by $\log \lambda$, we get

$$0 < \left| k \left(\frac{\log \alpha}{\log \lambda} \right) - (n+m) + \left(\frac{\log(32/\sqrt{5})}{\log \lambda} \right) \right| < 6.62 \cdot \lambda^{-2m}. \tag{3.5}$$

Take $\gamma := \frac{\log \alpha}{\log \lambda} \notin \mathbb{Q}$ and $M := 2.82 \cdot 10^{32}$. Then we found that q_{63}, the denominator of the 63^{th} convergent of γ exceeds $6M$. Moreover,

$$u := k < 8n < 8 \cdot 3.52 \cdot 10^{31} < M.$$

Now take

$$\mu := \frac{\log(32/\sqrt{5})}{\log \lambda}.$$

In this case, a quick computation with Mathematica gives us the inequality

$$0 < \epsilon = ||\mu q_{63}|| - M||\gamma q_{63}|| \leq 0.408068.$$

Let $A := 6.62$, $B := \lambda$ and $w := 2m$ in Lemma 2.2. Thus, with the help of Mathematica, we can say that the inequality (3.5) has no solution for

$$2m = w \geq \frac{\log(Aq_{63}/\epsilon)}{\log B} \geq 45.04933.$$

So

$$m \leq 22. \tag{3.6}$$

Substituting this upper bound for m into (3.4), we obtain $n < 7.255727 \cdot 10^{16}$.

Now, let

$$z_2 := k \log \alpha - n \log \lambda + \log \left(\frac{4\sqrt{2}}{\sqrt{5} B_m} \right).$$

In this case, taking into account that $n > 30$, it is seen that

$$|1 - e^{z_2}| < \frac{6}{\lambda^n} < \frac{1}{4} \tag{3.7}$$

by (3.3). If $z_2 > 0$, then

$$|z_2| = z_2 < e^{z_2} - 1 = |e^{z_2} - 1| < \frac{6}{\lambda^n}.$$

If $z_2 < 0$, then $1 - e^{z_2} = |1 - e^{z_2}| < \frac{1}{4}$. Therefore, we get $e^{z_2} > \frac{3}{4}$ and so $e^{|z_2|} = e^{-z_2} < \frac{4}{3}$. By using (3.7), we get

$$0 < |z_2| < e^{|z_2|} - 1 = e^{|z_2|} \, |1 - e^{z_2}| < \frac{4}{3} \cdot \frac{6}{\lambda^n} = \frac{8}{\lambda^n}.$$

Therefore, it holds that

$$|z_2| < \frac{8}{\lambda^n}.$$

That is,

$$0 < \left| k\log\alpha - n\log\lambda + \log\left(\frac{4\sqrt{2}}{\sqrt{5}B_m}\right)\right| < \frac{8}{\lambda^n}.$$

Dividing both sides of the above inequality by $\log\lambda$, we get

$$0 < \left| k\left(\frac{\log\alpha}{\log\lambda}\right) - n + \frac{\log\left(\frac{4\sqrt{2}}{\sqrt{5}B_m}\right)}{\log\lambda}\right| < 4.54\cdot\lambda^{-n}. \tag{3.8}$$

Putting $\gamma := \frac{\log\alpha}{\log\lambda}$ and taking $M := 5.81\cdot 10^{17}$, we found that q_{39}, the denominator of the 39^{th} convergent of γ exceeds $6M$. Note that $u := k < 8n < 8\cdot 7.25727\cdot 10^{16} < M$. Taking

$$\mu := \frac{\log\left(\frac{4\sqrt{2}}{\sqrt{5}B_m}\right)}{\log\lambda}$$

and considering the fact that $m \le 22$ by (3.6), a quick computation with Mathematica gives us the inequality

$$0 < \epsilon = ||\mu q_{39}|| - M||\gamma q_{39}|| \le 0.467267$$

for all $m \in [1, 22]$. Let $A := 4.54$, $B := \lambda$ and $w := n$ in Lemma 2.2. Thus, with the help of Mathematica, we can say that the inequality (3.8) has no solution for

$$n = w \ge \frac{\log(Aq_{39}/\epsilon)}{B} \ge \frac{\log(Aq_{39}/0.467267)}{B} \ge 25.6246.$$

Therefore $n \le 25$. This contradicts our assumption that $n > 30$. Thus, the proof is completed. □

Theorem 3.2. *The Diophantine equation $B_k = F_mF_n$ has only the solutions $(k, m, n) = (1, 1, 1), (1, 1, 2), (1, 2, 2), (2, 3, 4)$ in positive integers.*

Proof. Assume that $B_k = F_mF_n$ for some positive integers k, m, n. Let $n = m$. Then $B_k = F_m^2$. Therefore $k = 1$ by Theorem 2.4. So we get $(k, m, n) = (1, 1, 1), (1, 2, 2)$. Now assume that $1 \le m < n \le 107$. Then $k \le 58$ and we get the solutions $(k, m, n) = (1, 1, 1), (1, 1, 2), (1, 2, 2), (2, 3, 4)$ by using Mathematica. So assume that $n > 107$. Then $k \ge 59$. Since

$$(\alpha^3)^{k-1} < \lambda^{k-1} < B_k = F_mF_n \le \alpha^{n+m-2}$$

by (1.1) and (1.3), it follows that $3(k-1) < n + m - 2 < 2(n-1)$, which implies that $k < n$. In a similar manner, we see that $k > (m+n)/4 > 108/4 = 27$. Since $B_k = F_nF_m$, we get

$$\frac{\lambda^k}{4\sqrt{2}} - \frac{\alpha^{m+n}}{5} = \frac{\delta^k}{4\sqrt{2}} - \frac{\alpha^n\beta^m + \alpha^m\beta^n - \beta^{n+m}}{5}$$

$$= \frac{5\delta^k + 4\sqrt{2}(\alpha^n\beta^m + \alpha^m\beta^n - \beta^{n+m})}{20\sqrt{2}}.$$

Taking absolute values, it is seen that

$$\begin{aligned}\left|\frac{\lambda^k}{4\sqrt{2}} - \frac{\alpha^{m+n}}{5}\right| &\leq \frac{5\delta^k + 4\sqrt{2}(\alpha^n|\beta|^m + \alpha^m|\beta|^n + |\beta|^m)}{20\sqrt{2}} \\ &= \frac{4\sqrt{2}\alpha^{n-m} + 5\delta^k + 4\sqrt{2}(|\beta|^{n-m} + |\beta|^{n+m})}{20\sqrt{2}} \\ &\leq \frac{4\sqrt{2}\alpha^{n-m} + 4\sqrt{2}}{20\sqrt{2}} \\ &< \frac{4\sqrt{2}(\alpha^{n-m}+1)}{20\sqrt{2}} < \frac{\alpha^{n-m}+1}{5} < \frac{\alpha^{n-m+1}}{5},\end{aligned}$$

where we use the fact that $5\delta^k + 4\sqrt{2}(|\beta|^{n-m} + |\beta|^{n+m}) \leq 4\sqrt{2}$ for $k > 27$ and $n > 107$. Dividing both side of this inequality by $\alpha^{n+m}/5$, we get

$$\left|\frac{5\lambda^k\alpha^{-(n+m)}}{4\sqrt{2}} - 1\right| < \frac{1}{\alpha^{2m-1}}. \tag{3.9}$$

Now we apply Matheev's theorem. Let $\gamma_1 := \frac{5}{4\sqrt{2}}$, $\gamma_2 := \lambda$, $\gamma_3 := \alpha, b_1 := 1, b_2 := k, b_3 := -(n+m)$. The numbers $\gamma_1, \gamma_2, \gamma_3$ are real numbers and elements of the field $\mathbb{K} = \mathbb{Q}(\sqrt{2}, \sqrt{5})$. So $D = 4$. Now we show that $\Lambda_3 = (5\lambda^k\alpha^{-(n+m)})/4\sqrt{2} - 1$ is nonzero. For, if $\Lambda_3 = 0$, then $\lambda^k\alpha^{-(m+n)} = 4\sqrt{2}/5$. But this is impossible since $4\sqrt{2}/5$ is not an algebraic integer although $\lambda^k\alpha^{-(m+n)}$ is an algebraic integer. It can be seen that

$$h(\gamma_1) = h(5/4\sqrt{2}) = \frac{1}{2}(\log 32) = 1.7328\ldots,$$

$h(\gamma_2) = h(\lambda) = (1.76275)/2$ and $h(\gamma_3) = h(\alpha) = (0.4812)/2$. Therefore we can take $A_1 := 7, A_2 := 3.6, A_3 := 1$ and $B := 2n \geq \max\{1, |k|, |-(n+m)|\}$. Thus, taking into account the inequality (3.9) and using Theorem 2.1, we obtain

$$\frac{1}{\alpha^{2m-1}} > |\Lambda_3| > \exp\left((-1.4 \cdot 30^6 \cdot 3^{4.5} \cdot 4^2(1+\log 4)(1+\log 2n) \cdot 7 \cdot 3.6 \cdot 1\right),$$

and so

$$(2m-1)\log\alpha < (1.37767 \cdot 10^{14}) \cdot (1 + \log 2n).$$

Then it follows that

$$2m\log\alpha < (1.37767 \cdot 10^{14})(1+\log 2n) + \log\alpha. \tag{3.10}$$

Now, writing the equation $B_k = F_m F_n$ as

$$\frac{\lambda^k}{4\sqrt{2}F_m} - \frac{\alpha^n}{\sqrt{5}} = \frac{\delta^k}{4\sqrt{2}F_m} - \frac{\beta^n}{\sqrt{5}},$$

and taking absolute values, we get

$$\left|\frac{\lambda^k}{4\sqrt{2}F_m}-\frac{\alpha^n}{\sqrt{5}}\right|\leq\frac{\delta^k}{4\sqrt{2}F_m}+\frac{|\beta|^n}{\sqrt{5}}<1.$$

By dividing both side of this inequality by $\alpha^n/\sqrt{5}$, we obtain

$$\left|\frac{\lambda^k\sqrt{5}\alpha^{-n}}{4\sqrt{2}F_m}-1\right|<\frac{\sqrt{5}}{\alpha^n}<\frac{3}{\alpha^n}. \tag{3.11}$$

Take $\gamma_1:=\lambda$, $\gamma_2:=\alpha$, $\gamma_3:=(4\sqrt{2}F_m)/\sqrt{5}, b_1:=k, b_2:=-n, b_3:=-1$. Clearly, the numbers $\gamma_1,\gamma_2,\gamma_3$ are real numbers and elements of the field $\mathbb{K}=\mathbb{Q}(\sqrt{2},\sqrt{5})$ and so $D=4$. It can be seen that

$$\Lambda_4=\frac{\lambda^k\sqrt{5}\alpha^{-n}}{4\sqrt{2}F_m}-1$$

is nonzero. On the other hand, $h(\gamma_1)=h(\lambda)=(1.76275\dots)/2$ and $h(\gamma_2)=h(\alpha)=(0.4882\dots)/2$. Since $(4\sqrt{2}F_m)/\sqrt{5}$ is a root of the polynomial $5x^2-32F_m^2$, it follows that

$$\begin{aligned}h(\gamma_3)&\leq\frac{1}{2}\left(\log 5+2\log\left(4\sqrt{2}F_m/\sqrt{5}\right)\right)=\log(4\sqrt{2}F_m)=\log(4\sqrt{2})+\log F_m\\&<1.74+(m-1)\log\alpha<1.26+m\log\alpha,\end{aligned}$$

and so we can take $A_3:=4(1.26+m\log\alpha)$. Let $A_1:=3.6, A_2:=1$. Since $k<n$, we can take $B:=n=\max\{k,|-n|,|-1|\}$. Using the inequality (3.11) and Theorem 2.1, we get

$$\begin{aligned}\frac{3}{\alpha^n}&>|\Lambda_4|\\&>\exp\left(-1.4\cdot 30^6\cdot 3^{4.5}\cdot 4^2(1+\log 4)(1+\log n)\cdot 3.6\cdot 1\cdot 4(1.26+m\log\alpha)\right),\end{aligned}$$

or

$$n\log\alpha-\log 3<1.968\,1\times 10^{12}\cdot(1+\log n)\cdot(5.04+4m\log\alpha). \tag{3.12}$$

Inserting the inequality (3.10) into the last inequality, a computer search with Mathematica gives us that $n<6.26482\cdot 10^{31}$. Now we reduce this bound to a size that can be easily dealt. In order to do this, we use Lemma 2.2 again. Let $z_3=k\log\lambda-(n+m)\log\alpha+\log(5/4\sqrt{2})$. Then from the inequality (3.9), it follows that

$$|1-e^{z_3}|<\frac{1}{\alpha^{2m-1}}.$$

If $z_3>0$, then

$$|z_3|=z_3<e^{z_3}-1=|1-e^{z_3}|<\frac{1}{\alpha^{2m-1}}.$$

If $z_3<0$, then

$$1-e^{z_3}=|1-e^{z_3}|<\frac{1}{\alpha^{2m-1}}<\frac{2}{3}.$$

Thus $e^{-z_3} < 3$, which yields to

$$|z_3| < e^{|z_3|} - 1 = e^{|z_3|}|1 - e^{z_3}| < \frac{3}{\alpha^{2m-1}}.$$

Therefore, it holds that

$$|z_3| < \frac{3}{\alpha^{2m-1}}.$$

Then

$$\left|k \log \lambda - (n+m) \log \alpha + \log(5/4\sqrt{2})\right| < \frac{3}{\alpha^{2m-1}}.$$

Dividing both sides of this inequality by $\log \alpha$, we get

$$0 < \left|k \frac{\log \lambda}{\log \alpha} - (n+m) + \frac{\log(5/4\sqrt{2})}{\log \alpha}\right| < 10.08 \cdot \alpha^{-2m}. \tag{3.13}$$

Now, we apply Lemma 2.2. Take $\gamma =: \log \lambda / \log \alpha, \mu := \log(5/4\sqrt{2})/\log \alpha, A := 10.08, B := \alpha, w = 2m$ and $M = 6.26482 \cdot 10^{31}$. We see that q_{62}, the denominator of the 62^{th} convergent of γ exceeds $6M$. Note that $M = 6.26482 \cdot 10^{31} = n > k$. In this case, a quick computation with Mathematica gives us the inequality

$$0 < \epsilon = ||\mu q_{62}|| - M||\gamma q_{62}|| \leq 0.39276.$$

Thus, with the help of Mathematica, we can say that the inequality (3.13) has no solution for

$$2m = w \geq \frac{\log(Aq_{62}/\epsilon)}{\log B} \geq 163.277.$$

Therefore $m \leq 81$. Substituting this value of m into (3.12), we get $n < 2.70817 \cdot 10^{27}$. Now, let

$$z_4 := k \log \lambda - n \log \alpha + \log(\sqrt{5}/4\sqrt{2}F_m).$$

Then, from (3.11), we can write

$$|1 - e^{z_4}| < \frac{3}{\alpha^n} < \frac{1}{2}.$$

If $z_4 > 0$, then

$$|z_4| = z_4 < e^{z_4} - 1 = |1 - e^{z_4}| < \frac{3}{\alpha^n}.$$

If $z_4 < 0$, then $1 - e^{z_4} = |1 - e^{z_4}| < 1/2$ and we get $e^{|z_4|} < 2$. Thus,

$$|z_4| < e^{|z_4|} - 1 = e^{|z_4|}|1 - e^{z_4}| < \frac{6}{\alpha^n}.$$

In both cases, it holds that $|z_4| < 6/\alpha^n$. That is,

$$\left|k \log \lambda - n \log \alpha + \log(\sqrt{5}/4\sqrt{2}F_m)\right| < \frac{6}{\alpha^n}.$$

Dividing both sides of this inequality by $\log \alpha$, we get

$$0 < \left| k\frac{\log \lambda}{\log \alpha} - n + \frac{\log(\sqrt{5}/4\sqrt{2}F_m)}{\log \alpha} \right| < 12.46 \cdot \alpha^{-n}. \tag{3.14}$$

Now, we apply Lemma 2.2. Let $\gamma := \log \lambda / \log \alpha, \mu = \log(\sqrt{5}/4\sqrt{2}F_m)/\log \alpha, A := 12.46, B := k, w := n$ and $M := 2.70817 \cdot 10^{17}$. It is seen that q_{39}, the denominator of the 39^{th} convergent of γ exceeds $6M$. Moreover, $M = n > k$. In this case, a quick computation with Mathematica gives us the inequality

$$0 < \epsilon = ||\mu q_{39}|| - M||\gamma q_{39}|| \leq 0.493976$$

for all $m \in [1, 81]$. Thus, with the help of Mathematica, we can say that the inequality (3.14) has no solution for

$$n = w \geq \frac{\log(Aq_{39}/\epsilon)}{\log B} \geq \frac{\log(Aq_{39}/0.493976)}{\log B} \geq 105.224.$$

Therefore, $n \leq 105$. But this contradicts the assumption that $n > 107$. This completes the proof. □

References

[1] A. Baker, H. Davenport: *The equations* $3x^2 - 2 = y^2$ *and* $8x^2 - 7 = z^2$, Quart. J. Math. Oxford Ser. (2) 20.1 (1969), pp. 129–137, doi: 10.1093/qmath/20.1.129.

[2] Y. Bugeaud, M. Mignotte, S. Siksek: *Classical and modular approaches to exponential Diophantine equations I. Fibonacci and Lucas perfect powers*, Ann. of Math. 163.3 (2006), pp. 969–1018, doi: 10.4007/annals.2006.163.969.

[3] M. Ddamulira, F. Luca, M. Rakotomalala: *Fibonacci Numbers which are products of two Pell numbers*, The Fibonacci Quarterly 54.1 (2016), pp. 11–18.

[4] P. K. Dey, S. S. Rout: *Diophantine equations concerning balancing and Lucas balancing numbers*, Arch. Math. 108.1 (2017), pp. 29–43, doi: 10.1007/s00013-016-0994-z.

[5] A. Dujella, A. Pethő: *A generalization of a theorem of Baker and Davenport*, Quart. J. Math. Oxford Ser. (2) 49.3 (1998), pp. 291–306, doi: 10.1093/qmathj/49.3.291.

[6] R. Keskin, O. Karaatli: *Some New Properties of Balancing Numbers and SquareTriangular Numbers*, Journal of Integer Sequences (2012), Article 12.1.4, 13 pp.

[7] T. Koshy: *Fibonacci and Lucas Numbers With Applications*, New York: Wiley-Interscience Pub., 2001, doi: 10.1002/9781118033067.

[8] E. M. Matveev: *An Explicit lower bound for a homogeneous rational linear form in the logarithms of algebraic numbers II*, (Russian), Izv. Ross. Akad. Nauk Ser. Mat. 64.6 (2000), pp. 125–180, Translation in Izv. Math. 64.6 (2000) pp. 1217–1269, doi: 10.1070/im2000v064n06abeh000314.

[9] G. K. Panda: *Some fascinating properties of balancing numbers*, Proc. Eleventh Internat. Conference on Fibonacci Numbers and Their Applications, Cong. Numerantium 194 (2009), pp. 185–189.

[10] G. K. Panda, P. K. Ray: *Cobalancing numbers and cobalancers*, Int. J. Math. Math. Sci. 8 (2005), pp. 1189–1200, doi: 10.1155/ijmms.2005.1189.

[11] S. S. Rout: *Some Generalizations and Properties of Balancing Numbers*, Ph.D. Thesis, NIT Rourkela, 2015.

[12] T. Szakács: *Multiplying balancing numbers*, Acta Univ Sapientiae Mat. 3.1 (2011), pp. 90–96.

Annales Mathematicae et Informaticae
***50** (2019) pp. 71–87*
DOI: 10.33039/ami.2019.08.001
http://ami.uni-eszterhazy.hu

Topological loops with six-dimensional solvable multiplication groups having five-dimensional nilradical*

Ágota Figula, Kornélia Ficzere, Ameer Al-Abayechi

University of Debrecen, Institute of Mathematics, Hungary
figula@science.unideb.hu
ficzerelia@gmail.com
ameer@science.unideb.hu

Submitted: July 15, 2019
Accepted: August 4, 2019
Published online: August 14, 2019

Abstract

Using connected transversals we determine the six-dimensional indecomposable solvable Lie groups with five-dimensional nilradical and their subgroups which are the multiplication groups and the inner mapping groups of three-dimensional connected simply connected topological loops. Together with this result we obtain that every six-dimensional indecomposable solvable Lie group which is the multiplication group of a three-dimensional topological loop has one-dimensional centre and two- or three-dimensional commutator subgroup.

Keywords: multiplication group of a topological loop, connected transversals, linear representations of solvable Lie algebras

MSC: 22E25, 17B30, 20N05, 57S20, 53C30

1. Introduction

The multiplication group $Mult(L)$ and the inner mapping group $Inn(L)$ of a loop L are important tools for the investigations in loop theory since there are strong

*The paper was supported by the EFOP-3.6.1-16-2016-00022 project. This projects have been supported by the European Union, co-financed by the European Social Fund.

relations between the structure of the normal subloops of L and that of the normal subgroups of $Mult(L)$ (cf. [1, 2]). In [9] the authors have obtained necessary and sufficient conditions for a group G to be the multiplication group of L. These conditions say that one can use special transversals A and B with respect to a subgroup K of G. The subgroup K plays the role of the inner mapping group of L whereas the transversals A and B belong to the sets of left and right translations of L.

P. T. Nagy and K. Strambach in [8] investigate thoroughly topological and differentiable loops as continuous and differentiable sections in Lie groups. In this paper we follow their approach and study topological loops L of dimension 3 having a solvable Lie group as their multiplication group. Applying the criteria of [9] we obtained in [3] all solvable Lie groups of dimension ≤ 5 which are the multiplication group of a 3-dimensional connected simply connected topological proper loop. This classification has resulted only decomposable Lie groups as the group $Mult(L)$ of L. Hence we paid our attention to 6-dimensional solvable indecomposable Lie groups. If their Lie algebras have a 4-dimensional nilradical, then among the 40 isomorphism classes of Lie algebras there is only one class depending on a real parameter which consists of the Lie algebras of the group $Mult(L)$ of L (cf. [4]). This result has confirmed the observation that the condition for the multiplication group of a topological loop to be a (finite-dimensional) Lie group is strong. Since the 6-dimensional solvable indecomposable Lie algebras have 4 or 5-dimensional nilradical it remains to deal with the 99 classes of solvable Lie algebras having 5-dimensional nilradical (cf. [7, 10]). In [5] we proved that among them there are 20 classes of Lie algebras which satisfy the necessary conditions to be the Lie algebra of the group $Mult(L)$ of a 3-dimensional loop L. We determined there also the possible subalgebras of the corresponding inner mapping groups.

The purpose of this paper is to determine the indecomposable solvable Lie groups of dimension 6 which have 5-dimensional nilradical and which are the multiplication group of a 3-dimensional connected simply connected topological loop. To find a suitable linear representation of the simply connected Lie groups for the 20 classes of solvable Lie algebras given in [5] is the first step to achieve this classification (cf. Theorem 3.1). Applying the method of connected transversals we show that only those Lie groups G in Theorem 3.1 which have 2- or 3-dimensional commutator subgroup allow continuous left transversals A and B in the group G with respect to the subgroup K given in Theorem 3.1 such that A and B are K-connected and $A \cup B$ generates G (cf. Proposition 3.2 and Theorem 3.3). An arbitrary left transversal A to the 3-dimensional abelian subgroup K of G depends on three continuous real functions with three variables. The condition that the left transversals A and B are K-connected is formulated by functional equations. Summarizing the results of Theorem in [6], of Theorem 16 in [4] and of Theorem 3.3 we obtain that each 6-dimensional solvable indecomposable Lie group which is the multiplication group of a 3-dimensional topological loop has 1-dimensional centre and two- or three-dimensional commutator subgroup.

2. Preliminaries

A loop is a binary system $(L, \cdot)$ if there exists an element $e \in L$ such that $x = e \cdot x = x \cdot e$ holds for all $x \in L$ and the equations $x \cdot a = b$ and $a \cdot y = b$ have precisely one solution $x = b/a$ and $y = a\backslash b$. A loop is proper if it is not a group.

The left and right translations $\lambda_a = y \mapsto a \cdot y : L \to L$ and $\rho_a : y \mapsto y \cdot a : L \to L$, $a \in L$, are bijections of L. The permutation group $Mult(L) = \langle \lambda_a, \rho_a; a \in L \rangle$ is called the multiplication group of L. The stabilizer of the identity element $e \in L$ in $Mult(L)$ is called the inner mapping group $Inn(L)$ of L.

Let G be a group, let $K \leq G$, and let A and B be two left transversals to K in G. We say that A and B are K-connected if $a^{-1}b^{-1}ab \in K$ for every $a \in A$ and $b \in B$. The core $Co_G(K)$ of K in G is the largest normal subgroup of G contained in K. If L is a loop, then $\Lambda(L) = \{\lambda_a; a \in L\}$ and $R(L) = \{\rho_a; a \in L\}$ are $Inn(L)$-connected transversals in the group $Mult(L)$ and the core of $Inn(L)$ in $Mult(L)$ is trivial. In [9], Theorem 4.1, the following necessary and sufficient conditions are established for a group G to be the multiplication group of a loop L:

Proposition 2.1. *A group G is isomorphic to the multiplication group of a loop if and only if there exists a subgroup K with $Co_G(K) = 1$ and K-connected transversals A and B satisfying $G = \langle A, B \rangle$.*

A loop L is called topological if L is a topological space and the binary operations $(x, y) \mapsto x \cdot y$, $(x, y) \mapsto x\backslash y, (x, y) \mapsto y/x : L \times L \to L$ are continuous. In general the multiplication group of a topological loop L is a topological transformation group that does not have a natural (finite dimensional) differentiable structure. In this paper we deal with 3-dimensional connected simply connected topological loops L. We assume that the multiplication group of L is a 6-dimensional solvable indecomposable Lie group G such that its Lie algebra has 5-dimensional nilradical. Then L is homeomorphic to $\mathbb{R}^3$ (cf. [3, Lemma 5]). Since it has nilpotency class 2 (cf. [5, Theorem 3.1]) by Theorem 8 A in [2] the subgroup K in Proposition 2.1 is a 3-dimensional abelian Lie subgroup of G which does not contain any non-trivial normal subgroup of G, A and B are continuous K-connected left transversals to K in G such that $A \cup B$ generates G.

3. Six-dimensional solvable Lie multiplication groups with five-dimensional nilradical

Using necessary conditions we found in [5], Theorems 3.6, 3.7, those 6-dimensional solvable indecomposable Lie algebras with 5-dimensional nilradical which can occur as the Lie algebra $\mathbf{g}$ of the multiplication group of a 3-dimensional topological loop L. We obtained also the Lie subalgebras $\mathbf{k}$ of the inner mapping group of L. With the notation in [10] they are the following:

$\mathbf{g}_1 := \mathbf{g}_{6,14}^{a=b=0}$, $\mathbf{k}_{1,1} = \langle e_2, e_4 + e_1, e_5 \rangle$, $\mathbf{k}_{1,2} = \langle e_3, e_4 + e_1, e_5 \rangle$;

$\mathbf{g}_2 := \mathbf{g}_{6,22}^{a=0}$, $\mathbf{k}_2 = \langle e_3, e_4 + e_1, e_5 \rangle$,
$\mathbf{g}_3 := \mathbf{g}_{6,17}^{\delta=1,a=\varepsilon=0}$, $\mathbf{k}_{3,1} = \langle e_3, e_4, e_5 + e_1 \rangle$, $\mathbf{k}_{3,2} = \langle e_2, e_4, e_5 + e_1 \rangle$;
$\mathbf{g}_4 := \mathbf{g}_{6,51}^{\varepsilon=\pm1}$, $\mathbf{k}_4 = \langle e_1 + a_1 e_2, e_3 + e_2, e_4 \rangle$, $a_1 \in \mathbb{R}$;
$\mathbf{g}_5 := \mathbf{g}_{6,54}^{a=b=0}$, $\mathbf{k}_5 = \langle e_1 + e_2, e_3 + a_2 e_2, e_4 \rangle$, $a_2 \in \mathbb{R}$;
$\mathbf{g}_6 := \mathbf{g}_{6,63}^{a=0}$, $\mathbf{k}_6 = \langle e_1 + e_2, e_3 + a_2 e_2, e_4 \rangle$, $a_2 \in \mathbb{R}$;
$\mathbf{g}_7 := \mathbf{g}_{6,25}^{a=b=0}$, $\mathbf{k}_7 = \langle e_1 + e_5, e_2 + \varepsilon e_5, e_4 \rangle$, $\varepsilon = 0, 1$;
$\mathbf{g}_8 := \mathbf{g}_{6,15}^{a=0}$, $\mathbf{k}_8 = \langle e_1 + e_5, e_2 + a_2 e_5, e_4 + a_3 e_5 \rangle$, $a_3 \in \mathbb{R} \setminus \{0\}$, $a_2 \in \mathbb{R}$;
$\mathbf{g}_9 := \mathbf{g}_{6,21}^{a=0,0<|b|\leq1}$, $\mathbf{k}_9 = \langle e_3, e_4 + e_1, e_5 + e_1 \rangle$;
$\mathbf{g}_{10} := \mathbf{g}_{6,24}$, $\mathbf{k}_{10} = \langle e_3, e_4, e_5 + e_1 \rangle$;
$\mathbf{g}_{11} := \mathbf{g}_{6,30}$, $\mathbf{k}_{11} = \langle e_3, e_4 + a_2 e_1, e_5 + e_1 \rangle$, $a_2 \in \mathbb{R}$;
$\mathbf{g}_{12} := \mathbf{g}_{6,36}^{a=0,b\geq0}$, $\mathbf{k}_{12,1} = \langle e_3, e_4, e_5 + e_1 \rangle$, $\mathbf{k}_{12,2} = \langle e_3, e_4 + e_1, e_5 + a_3 e_1 \rangle$, $a_3 \in \mathbb{R}$;
$\mathbf{g}_{13} := \mathbf{g}_{6,16}$, $\mathbf{k}_{13} = \langle e_1 + e_5, e_2 + a_2 e_5, e_4 + a_3 e_5 \rangle$, $a_2, a_3 \in \mathbb{R}$;
$\mathbf{g}_{14} := \mathbf{g}_{6,27}^{a=1,b=\delta=0}$, $\mathbf{k}_{14} = \langle e_1 + e_5, e_2 + a_2 e_5, e_4 \rangle$, $a_2 \in \mathbb{R}$;
$\mathbf{g}_{15} := \mathbf{g}_{6,49}^{\varepsilon=0,\pm1}$, $\mathbf{k}_{15} = \langle e_1 + a_1 e_3, e_2 + e_3, e_4 + a_3 e_3 \rangle$, $a_1, a_3 \in \mathbb{R}$;
$\mathbf{g}_{16} := \mathbf{g}_{6,52}^{\varepsilon=0,\pm1}$, $\mathbf{k}_{16} = \langle e_1 + a_1 e_2, e_3 + e_2, e_4 \rangle$, $a_1 \in \mathbb{R}$;
$\mathbf{g}_{17} := \mathbf{g}_{6,57}^{a=0}$, $\mathbf{k}_{17} = \langle e_1 + e_2, e_3 + a_2 e_2, e_4 \rangle$, $a_2 \in \mathbb{R}$;
$\mathbf{g}_{18} := \mathbf{g}_{6,59}^{\delta=1}$, $\mathbf{k}_{18} = \langle e_1 + e_2, e_3 + a_2 e_2, e_4 \rangle$, $a_2 \in \mathbb{R}$;
$\mathbf{g}_{19} := \mathbf{g}_{6,17}^{\delta=\varepsilon=0,a\neq0}$, $\mathbf{k}_{19} = \langle e_1 + e_4, e_2 + a_2 e_4, e_5 + e_4 \rangle$, $a_2 \in \mathbb{R}$;
$\mathbf{g}_{20} := \mathbf{g}_{6,17}^{\delta=0,a=\varepsilon=1}$, $\mathbf{k}_{20} = \langle e_1 + e_4, e_2 + a_2 e_4, e_5 + a_3 e_4 \rangle$, $a_2, a_3 \in \mathbb{R}$.

In [11] a single matrix M is established depending on six variables such that the span of the matrices engenders the given Lie algebra in the list $\mathbf{g}_i$, $i = 1, \ldots, 20$. To obtain the matrix Lie group G_i of the Lie algebra $\mathbf{g}_i$ we exponentiate the space of matrices spanned by the matrix M. Simplifying the obtained exponential image we get a suitable simple form of a matrix Lie group such that by differentiating and evaluating at the identity its Lie algebra is isomorphic to the Lie algebra $\mathbf{g}_i$. In case of the Lie algebras $\mathbf{g}_j$, $j = 1, 2, 8, 9, 16$, we take in order the exponential image of the matrices:

$$M_1 = \begin{pmatrix} 0 & -s_3 & s_2 & 0 & -s_6 & 2s_1 \\ 0 & 0 & 0 & 0 & 0 & s_2 \\ 0 & 0 & 0 & 0 & 0 & s_3 \\ 0 & 0 & 0 & -s_6 & 0 & s_4 \\ 0 & 0 & 0 & 0 & 0 & 2s_5 \\ 0 & 0 & 0 & 0 & 0 & 0 \end{pmatrix}, \ s_i \in \mathbb{R}, i = 1, \ldots, 6,$$

$$M_2 = \begin{pmatrix} 0 & -s_3 & s_2 & 0 & -s_6 & 2s_1 \\ 0 & 0 & 0 & 0 & 0 & s_2 \\ 0 & -s_6 & 0 & 0 & 0 & s_3 \\ 0 & 0 & 0 & -s_6 & 0 & s_4 \\ 0 & 0 & 0 & 0 & 0 & 2s_5 \\ 0 & 0 & 0 & 0 & 0 & 0 \end{pmatrix}, \ s_i \in \mathbb{R}, i = 1, \dots, 6,$$

$$M_8 = \begin{pmatrix} -s_6 & -s_3 & -s_2 & 0 & 0 & 2s_1 \\ 0 & -s_6 & 0 & 0 & 0 & s_2 \\ 0 & 0 & 0 & 0 & 0 & -s_3 \\ 0 & -s_6 & 0 & -s_6 & 0 & s_4 \\ 0 & 0 & -s_6 & 0 & 0 & -s_5 \\ 0 & 0 & 0 & 0 & 0 & 0 \end{pmatrix}, \ s_i \in \mathbb{R}, i = 1, \dots, 6,$$

$$M_9 = \begin{pmatrix} 0 & -s_3 & s_2 & 0 & 0 & 2s_1 \\ 0 & 0 & 0 & 0 & 0 & s_2 \\ 0 & -s_6 & 0 & 0 & 0 & s_3 \\ 0 & 0 & 0 & -s_6 & 0 & s_4 \\ 0 & 0 & 0 & 0 & -bs_6 & s_5 \\ 0 & 0 & 0 & 0 & 0 & 0 \end{pmatrix}, \ s_i \in \mathbb{R}, i = 1, \dots, 6,$$

$$M_{16} = \begin{pmatrix} -s_6 & 0 & 0 & 0 & 0 & s_3 \\ 0 & 0 & 2s_5 & -\varepsilon s_6 & \varepsilon s_4 & 2s_2 \\ 0 & 0 & 0 & s_5 & 0 & -s_1 \\ 0 & 0 & 0 & 0 & s_5 & s_4 \\ 0 & 0 & 0 & 0 & 0 & s_6 \\ 0 & 0 & 0 & 0 & 0 & 0 \end{pmatrix}, \ s_i \in \mathbb{R}, \varepsilon = 0, \pm 1, i = 1, \dots, 6.$$

This procedure yields the following

Theorem 3.1. *The simply connected Lie group G_i and its subgroup K_i of the Lie algebra $\mathbf{g}_i$ and its subalgebra $\mathbf{k}_i$, $i = 1, \dots, 20$, is isomorphic to the linear group of matrices the multiplication of which is given by:*

For $i = 1$:

$$\begin{aligned} &g(x_1, x_2, x_3, x_4, x_5, x_6) g(y_1, y_2, y_3, y_4, y_5, y_6) \\ &= g(x_1 + y_1 + x_2 y_3 - x_3 y_2 - x_6 y_5, x_2 + y_2, x_3 + y_3, x_4 + y_4 e^{-x_6}, x_5 + y_5, x_6 + y_6), \\ &K_{1,1} = \{g(u_1, u_3, 0, u_1, u_2, 0); u_i \in \mathbb{R}, i = 1, 2, 3\}, \\ &K_{1,2} = \{g(u_1, 0, u_3, u_1, u_2, 0); u_i \in \mathbb{R}, i = 1, 2, 3\}, \end{aligned}$$

for $i = 2$:

$$\begin{aligned} &g(x_1, x_2, x_3, x_4, x_5, x_6) g(y_1, y_2, y_3, y_4, y_5, y_6) \\ &= g(x_1 + y_1 + x_2 y_3 - x_3 y_2 - x_6 (y_5 + x_2 y_2), \\ &\quad x_2 + y_2, x_3 + y_3 - x_6 y_2, x_4 + y_4 e^{-x_6}, x_5 + y_5, x_6 + y_6), \\ &K_2 = \{g(u_1, 0, u_3, u_1, u_2, 0); u_i \in \mathbb{R}, i = 1, 2, 3\}, \end{aligned}$$

for $i = 3$*:*

$$\begin{aligned}
&g(x_1,x_2,x_3,x_4,x_5,x_6)g(y_1,y_2,y_3,y_4,y_5,y_6)\\
&= g(x_1+y_1-x_6y_4+(\tfrac{1}{2}x_6^2+x_3)y_2,\\
&\qquad x_2+y_2, x_3+y_3, x_4+y_4-x_6y_2, x_5+y_5e^{-x_6}, x_6+y_6),\\
&\qquad K_{3,1} = \{g(u_2,u_3,0,u_1,u_2,0);\, u_i\in\mathbb{R}, i=1,2,3\},\\
&\qquad K_{3,2} = \{g(u_2,0,u_3,u_1,u_2,0);\, u_i\in\mathbb{R}, i=1,2,3\},
\end{aligned}$$

for $i = 4$*:*

$$\begin{aligned}
&g(x_1,x_2,x_3,x_4,x_5,x_6)g(y_1,y_2,y_3,y_4,y_5,y_6)\\
&= g(x_1+y_1+x_5y_4, x_2+y_2+x_5y_1+\varepsilon x_4y_6+\tfrac{1}{2}x_5^2y_4,\\
&\qquad x_3+y_3e^{-x_6}, x_4+y_4, x_5+y_5, x_6+y_6), \varepsilon=\pm1,\\
&K_4 = \{g(u_1, a_1u_1+u_2, u_2, u_3, 0, 0);\, u_i\in\mathbb{R}, i=1,2,3\}, a_1\in\mathbb{R},
\end{aligned}$$

for $i = 5$*:*

$$\begin{aligned}
&g(x_1,x_2,x_3,x_4,x_5,x_6)g(y_1,y_2,y_3,y_4,y_5,y_6)\\
&= g(x_1+(y_1+x_5y_3)e^{-x_6}, x_2+y_2+x_5y_4, x_3+y_3e^{-x_6}, x_4+y_4, x_5+y_5, x_6+y_6),\\
&K_5 = \{g(u_1, u_1+a_2u_2, u_2, u_3, 0, 0);\, u_i\in\mathbb{R}, i=1,2,3\}, a_2\in\mathbb{R},
\end{aligned}$$

for $i = 6$*:*

$$\begin{aligned}
&g(x_1,x_2,x_3,x_4,x_5,x_6)g(y_1,y_2,y_3,y_4,y_5,y_6)\\
&= g(x_1+(y_1+y_3x_5)e^{-x_6},\\
&\qquad x_2+y_2-(x_5+x_6)y_4, x_3+y_3e^{-x_6}, x_4+y_4, x_5+y_5, x_6+y_6),\\
&K_6 = \{g(u_1, u_1+a_2u_2, u_2, u_3, 0, 0);\, u_i\in\mathbb{R}, i=1,2,3\}, a_2\in\mathbb{R},
\end{aligned}$$

for $i = 7$*:*

$$\begin{aligned}
&g(x_1,x_2,x_3,x_4,x_5,x_6)g(y_1,y_2,y_3,y_4,y_5,y_6)\\
&= g(x_1+(y_1+y_2x_3)e^{-x_6}, x_2+y_2e^{-x_6}, x_3+y_3, x_4+y_4, x_5+y_5-x_4y_6, x_6+y_6),\\
&K_7 = \{g(u_1, u_2, 0, u_3, u_1+\varepsilon u_2, 0);\, u_i\in\mathbb{R}, i=1,2,3\},\ \varepsilon=0,1,
\end{aligned}$$

for $i = 8$*:*

$$\begin{aligned}
&g(x_1,x_2,x_3,x_4,x_5,x_6)g(y_1,y_2,y_3,y_4,y_5,y_6)\\
&= g(x_1+(y_1+y_2x_3)e^{-x_6}-y_3x_2,\\
&\qquad x_2+y_2e^{-x_6}, x_3+y_3, x_4+(y_4-y_2x_6)e^{-x_6}, x_5+y_5-x_6y_3, x_6+y_6),\\
&K_8 = \{g(u_1, u_2, 0, u_3, u_1+a_2u_2+a_3u_3, 0);\, u_i\in\mathbb{R}, i{=}1,2,3\}, a_3\in\mathbb{R}\setminus\{0\}, a_2\in\mathbb{R},
\end{aligned}$$

for $i = 9$*:*

$$\begin{aligned}
&g(x_1,x_2,x_3,x_4,x_5,x_6)g(y_1,y_2,y_3,y_4,y_5,y_6)\\
&= g(x_1+y_1+x_2y_3-(x_3+x_2x_6)y_2, x_2+y_2,\\
&\quad x_3+y_3-x_6y_2, x_4+y_4e^{-x_6}, x_5+y_5e^{-bx_6}, x_6+y_6),\ 0<|b|\le 1,\\
&\quad K_9=\{g(u_1+u_2,0,u_3,u_1,u_2,0); u_i\in\mathbb{R}, i=1,2,3\},
\end{aligned}$$

for $i = 10$*:*

$$\begin{aligned}
&g(x_1,x_2,x_3,x_4,x_5,x_6)g(y_1,y_2,y_3,y_4,y_5,y_6)\\
&= g(x_1+y_1-2x_6y_4+(x_6^2-x_2)y_3-(\tfrac{1}{3}x_6^3-x_2x_6-x_3)y_2, x_2+y_2,\\
&\quad x_3+y_3-x_6y_2, x_4+y_4-x_6y_3+\tfrac{1}{2}x_6^2y_2, x_5+y_5e^{-x_6}, x_6+y_6),\\
&\quad K_{10}=\{g(u_2,0,u_3,u_1,u_2,0); u_i\in\mathbb{R}, i=1,2,3\},
\end{aligned}$$

for $i = 11$*:*

$$\begin{aligned}
&g(x_1,x_2,x_3,x_4,x_5,x_6)g(y_1,y_2,y_3,y_4,y_5,y_6)\\
&= g(x_1+y_1+x_2y_3-\tfrac{1}{2}x_2^2y_6, x_2+y_2, x_3+y_3-x_2y_6,\\
&\quad x_4+y_4e^{-x_6}, x_5+y_5e^{-x_6}-x_4y_6, x_6+y_6),\\
&K_{11}=\{g(a_2u_1+u_2,0,u_3,u_1,u_2,0); u_i\in\mathbb{R}, i=1,2,3\}, a_2\in\mathbb{R},
\end{aligned}$$

for $i = 12$*:*

$$\begin{aligned}
&g(x_1,x_2,x_3,x_4,x_5,x_6)g(y_1,y_2,y_3,y_4,y_5,y_6)\\
&= g(x_1+y_1-x_2y_3+y_2(x_3+x_2x_6), x_2+y_2, x_3+y_3-x_6y_2,\\
&\quad x_4+y_4e^{-bx_6}\cos x_6+y_5e^{-bx_6}\sin x_6,\\
&\quad x_5-y_4e^{-bx_6}\sin x_6+y_5e^{-bx_6}\cos x_6, x_6+y_6), b\ge 0,\\
&K_{12,1}=\{g(u_2,0,u_3,u_1,u_2,0); u_i\in\mathbb{R}, i=1,2,3\},\\
&K_{12,2}=\{g(u_1+a_3u_2,0,u_3,u_1,u_2,0); u_i\in\mathbb{R}, i=1,2,3\}, a_3\in\mathbb{R},
\end{aligned}$$

for $i = 13$*:*

$$\begin{aligned}
&g(x_1,x_2,x_3,x_4,x_5,x_6)g(y_1,y_2,y_3,y_4,y_5,y_6)\\
&= g(x_1+[y_1-y_4x_6+y_2(\tfrac{1}{2}x_6^2+x_3)]e^{-x_6}-x_2y_3, x_2+y_2e^{-x_6},\\
&\quad x_3+y_3, x_4+(y_4-y_2x_6)e^{-x_6}, x_5+y_5-x_6y_3, x_6+y_6),\\
&K_{13}=\{g(u_1,u_2,0,u_3,u_1+a_2u_2+a_3u_3,0); u_i\in\mathbb{R}, i=1,2,3\}, a_2,a_3\in\mathbb{R},
\end{aligned}$$

for $i = 14$*:*

$$\begin{aligned}
&g(x_1,x_2,x_3,x_4,x_5,x_6)g(y_1,y_2,y_3,y_4,y_5,y_6)\\
&= g(x_1+y_1e^{-x_6}+x_2y_3, x_2+y_2e^{-x_6}, x_3+y_3,\\
&\quad x_4+y_4-x_6y_3, x_5+y_5-x_6y_4+\tfrac{1}{2}x_6^2y_3, x_6+y_6),
\end{aligned}$$

$$K_{14} = \{g(u_1, u_2, 0, u_3, u_1 + a_2u_2, 0); u_i \in \mathbb{R}, i = 1, 2, 3\}, a_2 \in \mathbb{R},$$

for $i = 15$*:*

$$\begin{aligned}
&g(x_1, x_2, x_3, x_4, x_5, x_6)g(y_1, y_2, y_3, y_4, y_5, y_6)\\
&= g(x_1 + y_1e^{-x_6} + x_4y_5, x_2 + (y_2 - 2\varepsilon y_4x_6 - y_1x_5)e^{-x_6} + (x_1 - x_4x_5)y_5,\\
&\quad x_3 + y_3 - x_6y_5, x_4 + y_4e^{-x_6}, x_5 + y_5, x_6 + y_6), \varepsilon = 0, \pm 1,\\
&K_{15} = \{g(u_1, u_2, a_1u_1 + u_2 + a_3u_3, u_3, 0, 0); u_i \in \mathbb{R}, i = 1, 2, 3\}, a_1, a_3 \in \mathbb{R},
\end{aligned}$$

for $i = 16$*:*

$$\begin{aligned}
&g(x_1, x_2, x_3, x_4, x_5, x_6)g(y_1, y_2, y_3, y_4, y_5, y_6)\\
&= g(x_1 + y_1 + x_5y_4 + \tfrac{1}{2}x_5^2y_6,\\
&\quad x_2 + y_2 + 2x_5y_1 + (x_5^2 - \varepsilon x_6)y_4 + (\tfrac{1}{3}x_5^3 + \varepsilon(x_4 - x_5x_6))y_6,\\
&\quad x_3 + y_3e^{-x_6}, x_4 + y_4 + x_5y_6, x_5 + y_5, x_6 + y_6), \varepsilon = 0, \pm 1,\\
&K_{16} = \{g(u_1, a_1u_1 + u_2, u_2, u_3, 0, 0); u_i \in \mathbb{R}, i = 1, 2, 3\}, a_1 \in \mathbb{R},
\end{aligned}$$

for $i = 17$*:*

$$\begin{aligned}
&g(x_1, x_2, x_3, x_4, x_5, x_6)g(y_1, y_2, y_3, y_4, y_5, y_6)\\
&= g(x_1 + (y_1 + x_5y_3)e^{-x_6}, x_2 + y_2 + x_5y_4 - \tfrac{1}{2}x_5^2y_6,\\
&\quad x_3 + y_3e^{-x_6}, x_4 + y_4 - x_5y_6, x_5 + y_5, x_6 + y_6),\\
&K_{17} = \{g(u_1, u_1 + a_2u_2, u_2, u_3, 0, 0); u_i \in \mathbb{R}, i = 1, 2, 3\}, a_2 \in \mathbb{R},
\end{aligned}$$

for $i = 18$*:*

$$\begin{aligned}
&g(x_1, x_2, x_3, x_4, x_5, x_6)g(y_1, y_2, y_3, y_4, y_5, y_6)\\
&= g(x_1 + (y_1 + y_3x_5)e^{-x_6}, x_2 + y_2 - (x_5 + x_6)y_4 - \tfrac{1}{2}(x_5 + x_6)^2y_5,\\
&\quad x_3 + y_3e^{-x_6}, x_4 + y_4 + (x_5 + x_6)y_5, x_5 + y_5, x_6 + y_6),\\
&K_{18} = \{g(u_1, u_1 + a_2u_2, u_2, u_3, 0, 0); u_i \in \mathbb{R}, i = 1, 2, 3\}, a_2 \in \mathbb{R},
\end{aligned}$$

for $i = 19$*:*

$$\begin{aligned}
&g(x_1, x_2, x_3, x_4, x_5, x_6)g(y_1, y_2, y_3, y_4, y_5, y_6)\\
&= g(x_1 + y_1e^{-ax_6} + x_3y_2, x_2 + y_2, x_3 + y_3e^{-ax_6},\\
&\quad x_4 + y_4 - x_6y_2, x_5 + y_5e^{-x_6}, x_6 + y_6), a \in \mathbb{R} \setminus \{0\},\\
&K_{19} = \{g(u_1, 0, u_2, u_1 + a_2u_2 + u_3, u_3, 0); u_i \in \mathbb{R}, i = 1, 2, 3\}, a_2 \in \mathbb{R},
\end{aligned}$$

for $i = 20$*:*

$$\begin{aligned}
&g(x_1, x_2, x_3, x_4, x_5, x_6)g(y_1, y_2, y_3, y_4, y_5, y_6)\\
&= g(x_1 + (y_1 - x_6y_5 + y_2x_3)e^{-x_6}, x_2 + y_2e^{-x_6},\\
&\quad x_3 + y_3, x_4 + y_4 - x_3y_6, x_5 + y_5e^{-x_6}, x_6 + y_6),\\
&K_{20} = \{g(u_1, u_2, 0, u_1 + a_2u_2 + a_3u_3, u_3, 0); u_i \in \mathbb{R}, i = 1, 2, 3\}, a_2, a_3 \in \mathbb{R}.
\end{aligned}$$

Among the Lie groups in Theorem 3.1 only the group G_1 has 2-dimensional commutator subgroup and the groups G_i, $i = 2, \dots, 7$, have 3-dimensional commutator subgroup. We show that among the 6-dimensional solvable indecomposable Lie groups with 5-dimensional nilradical precisely these Lie groups are the multiplication groups of three-dimensional connected simply connected topological loops.

Proposition 3.2. *There does not exist 3-dimensional connected topological proper loop L such that the Lie algebra $\mathbf{g}$ of the multiplication group of L is one of the Lie algebras $\mathbf{g}_i$, $i = 8, \dots, 20$.*

Proof. If L exists, then there exists its universal covering loop $\tilde{L}$ which is homeomorphic to $\mathbb{R}^3$. The pairs (G_i, K_i) in Theorem 3.1 can occur as the group $Mult(\tilde{L})$ and the subgroup $Inn(\tilde{L})$. We show that none of the groups G_i, $i = 8, \dots, 20$, satisfies the condition that there exist continuous left transversals A and B to K_i in G_i such that for all $a \in A$ and $b \in B$ one has $a^{-1}b^{-1}ab \in K_i$. By Proposition 2.1 the groups G_i, $i = 8, \dots, 20$, are not the multiplication group of a loop $\tilde{L}$. Hence no proper loop $\tilde{L}$ exists which yields that also no proper loop L exists. This proves the assertion.

Two arbitrary left transversals to the group K_i in G_i are:

For $i = 9, 10, 11, 12$,

$$A = \{g(u, v, h_1(u, v, w), h_2(u, v, w), h_3(u, v, w), w); u, v, w \in \mathbb{R}\},$$
$$B = \{g(k, l, f_1(k, l, m), f_2(k, l, m), f_3(k, l, m), m); k, l, m \in \mathbb{R}\},$$

for $i = 8, 13, 14, 15$,

$$A = \{g(h_1(u, v, w), h_2(u, v, w), u, h_3(u, v, w), v, w); u, v, w \in \mathbb{R}\},$$
$$B = \{g(f_1(k, l, m), f_2(k, l, m), k, f_3(k, l, m), l, m); k, l, m \in \mathbb{R}\},$$

for $i = 16, 17, 18$,

$$A = \{g(h_1(u, v, w), u, h_2(u, v, w), h_3(u, v, w), v, w); u, v, w \in \mathbb{R}\},$$
$$B = \{g(f_1(k, l, m), k, f_2(k, l, m), f_3(k, l, m), l, m); k, l, m \in \mathbb{R}\},$$

for $i = 19$

$$A = \{g(h_1(u, v, w), u, h_2(u, v, w), v, h_3(u, v, w), w); u, v, w \in \mathbb{R}\},$$
$$B = \{g(f_1(k, l, m), k, f_2(k, l, m), l, f_3(k, l, m), m); k, l, m \in \mathbb{R}\},$$

for $i = 20$

$$A = \{g(h_1(u, v, w), h_2(u, v, w), u, v, h_3(u, v, w), w); u, v, w \in \mathbb{R}\},$$
$$B = \{g(f_1(k, l, m), f_2(k, l, m), k, l, f_3(k, l, m), m); k, l, m \in \mathbb{R}\},$$

where $h_i(u, v, w) : \mathbb{R}^3 \to \mathbb{R}$ and $f_i(k, l, m) : \mathbb{R}^3 \to \mathbb{R}$, $i = 1, 2, 3$, are continuous functions with $f_i(0, 0, 0) = h_i(0, 0, 0) = 0$. Taking in G_i, $i = 9, 11, 12$, the elements

$$a = g(0, v, h_1(0, v, 0), h_2(0, v, 0), h_3(0, v, 0), 0) \in A,$$

$$b = g(0,0,f_1(0,0,m), f_2(0,0,m), f_3(0,0,m), m) \in B$$

and in G_{17} the elements

$$a = g(h_1(0,v,0),0,h_2(0,v,0),h_3(0,v,0),v,0) \in A,$$
$$b = g(f_1(0,0,m),0,f_2(0,0,m),f_3(0,0,m),0,m) \in B$$

one has $a^{-1}b^{-1}ab \in K_i$ if and only if
for $i = 9$

$$mv^2 - 2vf_1(0,0,m) = h_2(0,v,0)(1-e^m) + h_3(0,v,0)(1-e^{bm}), \tag{3.1}$$

for $i = 11$

$$\tfrac{1}{2}mv^2 + vf_1(0,0,m) = (e^m-1)(h_3(0,v,0) + a_2h_2(0,v,0)) - e^m m h_2(0,v,0), \tag{3.2}$$

for $i = 12$ and for $K_{12,1}$

$$2vf_1(0,0,m) - mv^2 = (1 - e^{bm}\cos m)h_3(0,v,0) - e^{bm}\sin m h_2(0,v,0), \tag{3.3}$$

for $i = 12$ and for $K_{12,2}$

$$\begin{aligned} 2vf_1(0,0,m) - mv^2 = {} & (1-e^{bm}\cos m)(h_2(0,v,0) + a_3h_3(0,v,0)) \\ & + e^{bm}\sin m(h_3(0,v,0) - a_3h_2(0,v,0)), \end{aligned} \tag{3.4}$$

for $i = 17$

$$\begin{aligned} -\tfrac{1}{2}mv^2 - vf_3(0,0,m) = {} & (1-e^m)[h_1(0,v,0) + (a_2 - v)h_2(0,v,0)] \\ & - e^m v f_2(0,0,m) \end{aligned} \tag{3.5}$$

is satisfied for all $m, v \in \mathbb{R}$. On the left hand side of equations (3.1), (3.2), (3.3), (3.4), (3.5) is the term mv^2 hence there does not exist any function $f_i(0,0,m)$ and $h_i(0,v,0)$, $i = 1,2,3$, satisfying these equations. Taking in G_{10} the elements

$$a = g(0,v,h_1(0,v,w),h_2(0,v,w),h_3(0,v,w),w) \in A$$
$$b = g(0,0,f_1(0,0,m),f_2(0,0,m),f_3(0,0,m),m) \in B,$$

respectively in G_{18} the elements

$$a = g(h_1(0,v,w),0,h_2(0,v,w),h_3(0,v,w),v,w) \in A,$$
$$b = g(f_1(0,0,m),0,f_2(0,0,m),f_3(0,0,m),0,m) \in B,$$

respectively in G_{16} the elements

$$a = g(h_1(0,v,0),0,h_2(0,v,0),h_3(0,v,0),v,0) \in A,$$
$$b = g(f_1(0,l,m),0,f_2(0,l,m),f_3(0,l,m),l,m) \in B$$

we obtain that $a^{-1}b^{-1}ab \in K_i$ if and only if in case $i = 10$ the equation

$$\begin{aligned} e^w(1-e^m)h_3(0,v,w) + e^m(e^w-1)f_3(0,0,m) \\ = (w^2+2v+2mw)f_1(0,0,m) + 2wf_2(0,0,m) \\ - (m^2+2wm)h_1(0,v,w) - 2mh_2(0,v,w) \\ - m^2wv - w^2mv - mv^2 - \tfrac{1}{3}vm^3, \end{aligned} \tag{3.6}$$

respectively in case $i = 18$ the equation

$$\begin{aligned} e^m(e^w-1)(f_1(0,0,m) + a_2f_2(0,0,m)) \\ + e^w(1-e^m)[h_1(0,v,w) + (a_2-v)h_2(0,v,w)] \\ = e^{m+w}vf_2(0,0,m) + (w+v)f_3(0,0,m) \\ - mh_3(0,v,w) + v^2m + \tfrac{1}{2}m^2v + wvm, \end{aligned} \tag{3.7}$$

respectively in case $i = 16$ the equation

$$\begin{aligned} -\tfrac{1}{3}v^3m - v^2lm - l^2vm - \tfrac{1}{2}a_1v^2m - \varepsilon m^2v - a_1vlm \\ = (1-e^m)h_2(0,v,0) - 2lh_1(0,v,0) + (l^2+2vl+a_1l+2\varepsilon m)h_3(0,v,0) \\ + 2vf_1(0,l,m) - (v^2+2vl+a_1v)f_3(0,l,m) \end{aligned} \tag{3.8}$$

holds for all $m, l, v, w \in \mathbb{R}$. Substituting into (3.6)

$$f_2(0,0,m) = f_2'(0,0,m) - mf_1(0,0,m), h_2(0,v,w) = h_2'(0,v,w) - wh_1(0,v,w),$$

respectively into (3.7)

$$f_1(0,0,m) = f_1'(0,0,m) - a_2f_2(0,0,m), h_1(0,v,w) = h_1'(0,v,w) + (v-a_2)h_2(0,v,w),$$

respectively into (3.8)

$$\begin{aligned} h_1(0,v,0) &= h_1'(0,v,0) + \left(v + \tfrac{1}{2}a_1\right)h_3(0,v,0), \\ f_1(0,l,m) &= f_1'(0,l,m) + \left(l + \tfrac{1}{2}a_1\right)f_3(0,l,m), \end{aligned}$$

we get in case $i = 10$

$$\begin{aligned} e^w(1-e^m)h_3(0,v,w) + e^m(e^w-1)f_3(0,0,m) \\ = (w^2+2v)f_1(0,0,m) - m^2h_1(0,v,w) + 2wf_2'(0,0,m) \\ - 2mh_2'(0,v,w) - m^2wv - w^2mv - mv^2 - \tfrac{1}{3}vm^3, \end{aligned} \tag{3.9}$$

respectively in case $i = 18$

$$\begin{aligned} e^m(e^w-1)f_1'(0,0,m) - e^{m+w}vf_2(0,0,m) + e^w(1-e^m)h_1'(0,v,w) \\ = (w+v)f_3(0,0,m) - mh_3(0,v,w) + v^2m + \tfrac{1}{2}m^2v + wvm, \end{aligned} \tag{3.10}$$

respectively in case $i = 16$

$$\begin{aligned}&(1-e^m)h_2(0,v,0)+(l^2+2\varepsilon m)h_3(0,v,0)\\&\quad-v^2f_3(0,l,m)-2lh_1'(0,v,0)+2vf_1'(0,l,m)\\&=-\tfrac{1}{3}v^3m-v^2lm-l^2vm-\tfrac{1}{2}a_1v^2m-\varepsilon m^2v-a_1vlm.\end{aligned}\tag{3.11}$$

Since on the right hand side of (3.9), respectively (3.10), respectively (3.11) there is the term $-\frac{1}{3}vm^3$, respectively $\frac{1}{2}m^2v$, respectively $-\frac{1}{3}v^3m$ there does not exist any function $f_i(0,0,m)$ and $h_i(0,v,w)$, $i=1,2,3$, respectively $f_i(0,l,m)$, $i=1,3$, and $h_j(0,v,0)$, $j=1,2,3$, satisfying equation (3.9), respectively (3.10), respectively (3.11).

Taking in G_i, $i = 8, 13, 14$, the elements

$$\begin{aligned}a&=g(h_1(0,0,w),h_2(0,0,w),0,h_3(0,0,w),0,w)\in A,\\b&=g(f_1(k,0,m),f_2(k,0,m),k,f_3(k,0,m),0,m)\in B,\end{aligned}$$

respectively in G_{19} the elements

$$\begin{aligned}a&=g(h_1(0,0,w),0,h_2(0,0,w),0,h_3(0,0,w),w)\in A,\\b&=g(f_1(k,0,m),k,f_2(k,0,m),0,f_3(k,0,m),m)\in B,\end{aligned}$$

respectively in G_{20} the elements

$$\begin{aligned}a&=g(h_1(0,0,w),h_2(0,0,w),0,0,h_3(0,0,w),w)\in A,\\b&=g(f_1(k,0,m),f_2(k,0,m),k,0,f_3(k,0,m),m)\in B\end{aligned}$$

we have $a^{-1}b^{-1}ab\in K_i$ precisely if for $i=8$ the equation

$$\begin{aligned}wk&=e^w(1-e^m)[(a_2+a_3w)h_2(0,0,w)+a_3h_3(0,0,w)+h_1(0,0,w)]\\&\quad+e^m(e^w-1)[(a_3m+a_2-k)f_2(k,0,m)+a_3f_3(k,0,m)+f_1(k,0,m)]\\&\quad+e^{m+w}[a_3wf_2(k,0,m)+(2k-a_3m)h_2(0,0,w)],\end{aligned}\tag{3.12}$$

for $i=13$ the equation

$$\begin{aligned}wk&=e^w(1-e^m)[(\tfrac{1}{2}w^2+a_2+a_3w)h_2(0,0,w)+(a_3+w)h_3(0,0,w)+h_1(0,0,w)]\\&+e^m(e^w-1)[(\tfrac{1}{2}m^2-k+a_3m+a_2)f_2(k,0,m)\\&\qquad+(m+a_3)f_3(k,0,m)+f_1(k,0,m)]\\&+e^{m+w}[((m+a_3)w+\tfrac{1}{2}w^2)f_2(k,0,m)+(2k-\tfrac{1}{2}m^2-(w+a_3)m)h_2(0,0,w)]\\&+e^{m+w}(wf_3(k,0,m)-mh_3(0,0,w)),\end{aligned}\tag{3.13}$$

for $i=14$ the equation

$$\begin{aligned}&\tfrac{1}{2}w^2k+mwk+wf_3(k,0,m)-mh_3(0,0,w)\\&=e^w(1-e^m)(h_1(0,0,w)+a_2h_2(0,0,w))\\&\quad+e^m(e^w-1)(f_1(k,0,m)+a_2f_2(k,0,m))-e^{m+w}kh_2(0,0,w),\end{aligned}\tag{3.14}$$

for $i = 19$ the equation

$$\begin{aligned} wk = {} & e^w(1-e^m)h_3(0,0,w) - e^m(1-e^w)f_3(k,0,m) - e^{a(m+w)}kh_2(0,0,w) \\ & + e^{aw}(1-e^{am})(h_1(0,0,w) + a_2h_2(0,0,w)) \\ & - e^{am}(1-e^{aw})(f_1(k,0,m) + a_2f_2(k,0,m)), \end{aligned} \tag{3.15}$$

for $i = 20$ the equation

$$\begin{aligned} -wk = {} & e^w(1-e^m)(h_1(0,0,w) + a_2h_2(0,0,w) + (w+a_3)h_3(0,0,w)) \\ & + e^m(1-e^w)((k-a_2)f_2(k,0,m) - f_1(k,0,m) - (m+a_3)f_3(k,0,m)) \\ & + e^{m+w}(kh_2(0,0,w) - mh_3(0,0,w) + wf_3(k,0,m)) \end{aligned} \tag{3.16}$$

is satisfied for all $k, m, w \in \mathbb{R}$, $a_2, a_3 \in \mathbb{R}$. Putting into (3.12)

$$\begin{aligned} h_1(0,0,w) &= h_1'(0,0,w) - (a_3w + a_2)h_2(0,0,w) - a_3h_3(0,0,w), \\ f_1(k,0,m) &= f_1'(k,0,m) + (k - a_3m - a_2)f_2(k,0,m) - a_3f_3(k,0,m), \end{aligned}$$

respectively into (3.13)

$$\begin{aligned} h_1(0,0,w) &= h_1'(0,0,w) - (\tfrac{1}{2}w^2 + a_3w + a_2)h_2(0,0,w) - (a_3+w)h_3(0,0,w), \\ f_1(k,0,m) &= f_1'(k,0,m) + (k - \tfrac{1}{2}m^2 - a_3m - a_2)f_2(k,0,m) - (m+a_3)f_3(k,0,m), \\ f_3(k,0,m) &= f_3'(k,0,m) - (m+a_3)f_2(k,0,m), \\ h_3(0,0,w) &= h_3'(0,0,w) - (w+a_3)h_2(0,0,w), \end{aligned}$$

respectively into (3.14)

$$\begin{aligned} h_1(0,0,w) &= h_1'(0,0,w) - a_2h_2(0,0,w), \\ f_3(k,0,m) &= f_3'(k,0,m) - mk, \\ f_1(k,0,m) &= f_1'(k,0,m) - a_2f_2(k,0,m), \end{aligned}$$

respectively into (3.15)

$$\begin{aligned} h_1(0,0,w) &= h_1'(0,0,w) - a_2h_2(0,0,w), \\ f_1(k,0,m) &= f_1'(k,0,m) - a_2f_2(k,0,m), \end{aligned}$$

respectively into (3.16)

$$\begin{aligned} h_1(0,0,w) &= h_1'(0,0,w) - a_2h_2(0,0,w) - (w+a_3)h_3(0,0,w), \\ f_1(k,0,m) &= f_1'(k,0,m) + (k-a_2)f_2(k,0,m) - (m+a_3)f_3(k,0,m) \end{aligned}$$

in order equations (3.12), (3.13), (3.14), (3.15), (3.16) reduce in case $i = 8$ to

$$\begin{aligned} wk = {} & e^w(1-e^m)h_1'(0,0,w) + e^m(e^w-1)f_1'(k,0,m) \\ & + e^{m+w}[a_3wf_2(k,0,m) + (2k - a_3m)h_2(0,0,w)], \end{aligned} \tag{3.17}$$

in case $i = 13$ to

$$\begin{aligned} wk = {} & e^w(1-e^m)h_1'(0,0,w) + e^m(e^w-1)f_1'(k,0,m) \\ & + e^{m+w}[\tfrac{1}{2}w^2 f_2(k,0,m) + (2k - \tfrac{1}{2}m^2)h_2(0,0,w) \\ & + wf_3'(k,0,m) - mh_3'(0,0,w)], \end{aligned} \tag{3.18}$$

in case $i = 14$ to

$$\begin{aligned} & \tfrac{1}{2}w^2k + wf_3'(k,0,m) - mh_3(0,0,w) \\ & = e^w(1-e^m)h_1'(0,0,w) + e^m(e^w-1)f_1'(k,0,m) - e^{m+w}kh_2(0,0,w), \end{aligned} \tag{3.19}$$

in case $i = 19$ to

$$\begin{aligned} wk = {} & e^w(1-e^m)h_3(0,0,w) - e^m(1-e^w)f_3(k,0,m) - e^{a(m+w)}kh_2(0,0,w) \\ & + e^{aw}(1-e^{am})h_1'(0,0,w) - e^{am}(1-e^{aw})f_1'(k,0,m), \end{aligned} \tag{3.20}$$

and in case $i = 20$ to

$$\begin{aligned} -wk = {} & e^w(1-e^m)h_1'(0,0,w) + e^m(e^w-1)f_1'(k,0,m) \\ & + e^{m+w}(kh_2(0,0,w) - mh_3(0,0,w) + wf_3(k,0,m)). \end{aligned} \tag{3.21}$$

Since on the left hand side of (3.17), (3.18), (3.20), (3.21), respectively of (3.19) is the term wk, respectively $\frac{1}{2}w^2k$ there does not exist any function $f_i(k,0,m)$, $h_i(0,0,w)$, $i = 1,2,3$, satisfying equation (3.17), (3.18), (3.20), (3.21), respectively (3.19).

Taking in G_{15} the elements

$$\begin{aligned} a &= g(h_1(0,0,w), h_2(0,0,w), 0, h_3(0,0,w), 0, w) \in A, \\ b &= g(f_1(0,l,m), f_2(0,l,m), 0, f_3(0,l,m), l, m) \in B \end{aligned}$$

the product $a^{-1}b^{-1}ab$ lies in K_{15} if and only if the equation

$$\begin{aligned} wl = {} & e^w(1-e^m)[h_2(0,0,w) + (a_3+2w\varepsilon)h_3(0,0,w) + a_1h_1(0,0,w)] \\ & + e^m(e^w-1)[f_2(0,l,m) + (l+a_1)f_1(0,l,m) + (a_3+2m\varepsilon)f_3(0,l,m)] \\ & + e^{m+w}[2w\varepsilon f_3(0,l,m) - 2lh_1(0,0,w) - (l^2+2m\varepsilon+a_1l)h_3(0,0,w)] \end{aligned} \tag{3.22}$$

is satisfied for all $m, l, w \in \mathbb{R}$. Substituting into (3.22)

$$\begin{aligned} h_1(0,0,w) &= h_1'(0,0,w) - \tfrac{1}{2}a_1h_3(0,0,w), \\ h_2(0,0,w) &= h_2'(0,0,w) - a_1h_1(0,0,w) - (a_3+2w\varepsilon)h_3(0,0,w), \\ f_2(0,l,m) &= f_2'(0,l,m) - (l+a_1)f_1(0,l,m) - (a_3+2m\varepsilon)f_3(0,l,m), \end{aligned}$$

we obtain

$$wl = e^w(1-e^m)h_2'(0,0,w) + e^m(e^w-1)f_2'(0,l,m)$$

$$+ e^{m+w}[2w\varepsilon f_3(0,l,m) - 2lh_1'(0,0,w) - (l^2+2m\varepsilon)h_3(0,0,w)]. \qquad (3.23)$$

On the left hand side of equation (3.23) is the term wl hence there does not exist any function $f_i(0,l,m)$, $i=2,3$, and $h_j(0,0,w)$, $j=1,2,3$ such that equation (3.23) holds. □

Theorem 3.3. *Let L be a connected simply connected topological proper loop of dimension 3 such that its multiplication group is a 6-dimensional solvable indecomposable Lie group having 5-dimensional nilradical. Then the pairs of Lie groups (G_i, K_i), $i=1,\dots,7$, are the multiplication groups $Mult(L)$ and the inner mapping groups $Inn(L)$ of L.*

Proof. The sets

$$A = \{g(k, 1-e^m, l, me^{-m}, 2l, m); k,l,m \in \mathbb{R}\},$$
$$B = \{g(u, w, v, 2ve^{-w}, 1-e^w, w); u,v,w \in \mathbb{R}\},$$

respectively

$$C = \{g(k, l, 1-e^m, me^{-m}, -2l, m); k,l,m \in \mathbb{R}\},$$
$$D = \{g(u, v, w, -2ve^{-w}, 1-e^w, w); u,v,w \in \mathbb{R}\}$$

are $K_{1,1}$-, respectively $K_{1,2}$-connected left transversals in G_1. The sets

$$A = \{g(k, l, l, me^{-m}, l^2-1+e^m, m); k,l,m \in \mathbb{R}\},$$
$$B = \{g(u, v, v, -we^{-w}, v^2+1-e^w, w); u,v,w \in \mathbb{R}\}$$

are K_2-connected left transversals in G_2. The sets

$$A = \{g(k, \tfrac{1}{2}m^2-l, l, e^m-1-m(\tfrac{1}{2}m^2-l), me^{-m}, m); k,l,m \in \mathbb{R}\},$$
$$B = \{g(u, \tfrac{1}{2}w^2-v, v, 1-e^w-w(\tfrac{1}{2}w^2-v), -we^{-w}, w); u,v,w \in \mathbb{R}\},$$

respectively

$$C = \{g(k, l, \tfrac{1}{2}m^2+e^m-1, -lm+m, le^{-m}, m); k,l,m \in \mathbb{R}\},$$
$$D = \{g(u, v, \tfrac{1}{2}w^2-e^w+1, -vw+w, -ve^{-w}, w); u,v,w \in \mathbb{R}\}$$

are $K_{3,1}$-, respectively $K_{3,2}$-connected left transversals in G_3. The sets

$$A = \{g((l+a_1)(1-e^m)+l, k, -e^{-m}(\tfrac{1}{2}l^2+\varepsilon m), 1-e^m, l, m); k,l,m \in \mathbb{R}\},$$
$$B = \{g((v+a_1)(e^w-1)+v, u, e^{-w}(\tfrac{1}{2}v^2+\varepsilon w), e^w-1, v, w); u,v,w \in \mathbb{R}\}$$

are K_4-connected left transversals in G_4. The sets

$$A = \{g(le^{-k}(a_2-l+1), m, -le^{-k}, 1-le^k-e^k, l, k); k,l,m \in \mathbb{R}\},$$
$$B = \{g(ve^{-u}(v-1-a_2), w, ve^{-u}, ve^u+e^u-1, v, u); u,v,w \in \mathbb{R}\}$$

are K_5-connected left transversals in G_5. The sets

$$A = \{g((l-a_2)l + (l+m)e^{-m}, k, l, e^m - 1, l, m); k, l, m \in \mathbb{R}\},$$
$$B = \{g((v-a_2)v - (v+w)e^{-w}, u, v, 1-e^w, v, w); u, v, w \in \mathbb{R}\}$$

are K_6-connected left transversals in G_6. The sets

$$A = \{g((\varepsilon - k)me^{-m}, -me^{-m}, k, -ke^m, l, m), k, l, m \in \mathbb{R}\},$$
$$B = \{g((u-\varepsilon)we^{-w}, we^{-w}, u, ue^w, v, w), u, v, w \in \mathbb{R}\}$$

are K_7-connected left transversals in G_7. For all $i = 1, \ldots, 7$, the sets A, B, respectively C, D generate the group G_i. According to Proposition 2.1 the pairs (G_i, K_i), $i = 1, \ldots, 7$, are multiplication groups and inner mapping groups of L which proves the assertion. □

Corollary 3.4. *Each* 3*-dimensional connected topological proper loop* L *having a solvable indecomposable Lie group of dimension* 6 *as the group* $Mult(L)$ *of* L *has* 1*-dimensional centre and* 2*- or* 3*-dimensional commutator subgroup.*

Proof. If L has a 6-dimensional indecomposable nilpotent Lie group as its multiplication group, then the assertion follows from case b) of Theorem in [6]. If it has a 6-dimensional indecomposable solvable Lie group with 4-dimensional nilradical, then the assertion is proved in Theorem 16 in [4]. If it has a 6-dimensional indecomposable solvable Lie group with 5-dimensional nilradical, then Theorems 3.6 and 3.7 in [5] and Theorem 3.3 give the assertion. □

References

[1] A. A. Albert: *Quasigroups I*, Trans. Amer. Math. Soc. 54 (1943), pp. 507–519.

[2] R. H. Bruck: *Contributions to the Theory of Loops*, Trans. Amer. Math. Soc. 60 (1946), pp. 245–354.

[3] Á. Figula: *Three-dimensional topological loops with solvable multiplication groups*, Comm. Algebra 42 (2014), pp. 444–468.

[4] Á. Figula, A. Al-Abayechi: *Topological loops having solvable indecomposable Lie groups as their multiplication groups*, submitted to Transform. Groups (2018).

[5] Á. Figula, A. Al-Abayechi: *Topological loops with solvable multiplication groups of dimension at most six are centrally nilpotent*, Int. J. Group Theory (2019), pp. 14, doi: 10.22108/ijgt.2019.114770.1522.

[6] Á. Figula, M. Lattuca: *Three-dimensional topological loops with nilpotent multiplication groups*, J. Lie Theory 25 (2015), pp. 787–805.

[7] G. M. Mubarakzyanov: *Classification of Solvable Lie Algebras in dimension six with one non-nilpotent basis element*, Izv. Vyssh. Uchebn. Zaved. Mat. 4 (1963), pp. 104–116.

[8] P. T. Nagy, K. Strambach: *Loops in Group Theory and Lie Theory (De Gruyter Expositions in Mathematics, 35)*, Berlin: Walter de Gruyter GmbH & Co. KG, 2002.

[9] M. Niemenmaa, T. Kepka: *On Multiplication Groups of Loops*, J. Algebra 135 (1990), pp. 112–122.

[10] A. Shabanskaya, G. Thompson: *Six-dimensional Lie algebras with a five-dimensional nil-radical*, J. Lie Theory 23 (2013), pp. 313–355.

[11] G. Thompson, C. Hettiarachchi, N. Jones, A. Shabanskaya: *Representations of Six-dimensional Mubarakazyanov Lie algebras*, J. Gen. Lie Theory Appl. 8.1 (2014), Art. ID 1000211, 10 pp.

Annales Mathematicae et Informaticae
***50** (2019) pp. 89–92*
DOI: 10.33039/ami.2019.03.002
http://ami.uni-eszterhazy.hu

Algorithm for the generation of complement-free sets*

Dániel Fülöp, Carolin Hannusch

Faculty of Informatics, University of Debrecen, Hungary
fulop.daniel9623@gmail.com
hannusch.carolin@inf.unideb.hu

Submitted: February 6, 2019
Accepted: March 29, 2019
Published online: April 6, 2019

Abstract

We introduce an algorithm for the generation of complement-free sets of binary m-tuples, where m is even. We also provide an implementation for this algorithm for $m = 12$. Such complement-free sets are needed for the generation of a new class of error-correcting codes, which were introduced by Hannusch and Lakatos. These codes build the fundamental improvement in the cryptographic system of Dömösi, Hannusch and Horváth. Therefore the generation of complement-free sets will be important for cryptographic applications. In the end of the paper we give some interesting facts about complement-free sets as combinatorial objects.

Keywords: algorithmic computation, discrete sets

MSC: 03D32, 97N70

1. Introduction and notation

Let m be an even number, thus $m = 2k$ for some $k \in \mathbb{N}$. Then let X be the set of all binary m-tuples with exactly k pieces of 1-s and k pieces of 0-s.

Definition 1.1. Let $x \in X$ be an arbitrary element. Further we denote the whole-1 tuple of length m by $\mathbf{1}$. Then we say that $\mathbf{1} - x$ is the *complement* of x.

*This work was supported by the construction EFOP-3.6.3-VEKOP-16-2017-00002. The project was supported by the European Union, co-financed by the European Social Fund.

Definition 1.2. Let $Y \subset X$, such that $y \in Y$ implies $\mathbf{1} - y \notin Y$. Then Y is called *complement-free subset* of X. If Y has order $\frac{1}{2}\binom{m}{k}$, then we say that Y is a *maximal complement-free subset*.

In this paper, we give an algorithm for generating a maximal complement-free set randomly. Such sets are used in [3] for the construction of self-dual error-correcting codes of length 2^m and with minimum distance 2^k. These codes are called HL-codes and they are used in the cryptographic system of Dömösi, Hannusch and Horváth in [1]. In order to develop an effective implementation of the DHH-cryptosystem [2], it is necessary to generate a complement-free set effectively.

The DHH-crypotosystem is using the HL-code for $m = 12$, therefore we provide an implementation of our algorithm for $m = 12$ in C++ under the following link:

`https://arato.inf.unideb.hu/hannusch.carolin/alg.cpp`

2. The algorithm

We fix $m = 2k$.

`Input:` number l with $0 \leq l \leq \frac{1}{2}\binom{m}{k} - 1$

`Output:` maximal complement-free set Y

Step 1:

- Let A be the list of all binary m-tuples with k pieces of 1-s, where the first coordinate is 1.
- Let B be the list of all binary m-tuples where $B[i] = \mathbf{1} - A[i]$.

Step 2: for i from 1 to $\frac{1}{2}\binom{m}{k} + l - 1 \mod \frac{1}{2}\binom{m}{k}$ do
$i := 0$ or 1 randomly; end for;

Step 3: if $i = 0$ then $Y[i] := A[i]$; else $Y[i] := B[i]$. end for;

Continue Step 2 until $order(Y) = \frac{1}{2}\binom{m}{k}$.

This algorithm provides one possibility to create a complement-free set. Further research step will be the use of this algorithm (esp. the implementation) in an implementation of the DHH-cryptosystem. A fast algorithm with low memory-need is a necessary part of a competetive DHH-cryptosystem. The provided algorithm generates 100 complement-free sets of order 462 in 2.7 seconds and 1000 complement-free sets of order 462 in 15.8 seconds on *Intel(R) Core(TM)2 Duo CPU* at 2.93 GHz.

3. Additional facts about complement-free sets

The ordering of the list A in Step 1 of the algorithm introduced in Section 2 should be kept secret. This will improve the security of the algorithm when it is used in Cryptography. For $m = 12$ the list A has 462 elements, which means there are 462! possible orders of the elements of A and since

$$462! > 10^{1032},$$

this cannot be brute-forced.

So, let us now assume that A is secret. For the random value of i in Step 2 of the algorithm we need a random generator with almost 50% possibility that if $i = 0$, then $i + 1 = 1$ and vice versa. Applying such a random generator we have a probability of $\left(\frac{1}{2}\right)^{462}$ that we generate the same complement-free set twice. A good random generator can be found e.g. in [4].

Some more interesting things can be investigated in relation to complement-free sets if we have a more detailed look at one set itself. Given a complement-free set Y, each element $y \in Y$ consists of m coordinates. We will count the 1-s in a fixed coordinate for all $y \in Y$. For example, let $Y = \{(1,1,0,0),(1,0,0,1),(0,1,0,1)\}$. Then we have two 1-s in each four positions. Thus we will say that Y is of type $(2,2,2,2)$ according to the following definition:

Definition 3.1. We say that the complement-free set Y is of type $\nu = (n_1, \ldots, n_m)$, if

$$n_i = \sum_{y \in Y} y_i,$$

i.e. n_i is the number of 1-s in the i-th coordinate of all binary strings in Y.

Remark 3.2. We have $\sum_{i=1}^{m} n_i = k \cdot \frac{1}{2}\binom{m}{k}$.

Let us denote $\sum_{i=1}^{m} n_i$ by N. Then it is clear, that if ν is the type of a complement-free set, then ν is also a partition of N. This statement is not true in the other way, since e.g. for $m = 6$ we have $N = 30$ and $(7,7,5,3,3,1)$ is a partition, but there is no complement-free set of such a type.

Proposition 3.3. *For fix $m = 2k$ there exist at least $\frac{1}{4}\binom{m}{k} + 1$ different types of complement-free sets.*

Proof. We may assume $n_1 \geq n_2 \geq \cdots \geq n_m$. Then there exists exactly one type with $n_1 = \frac{1}{2}\binom{m}{k}$ (namely the complement-free set consists of all elements of the list A in this case). Now imagine, that we change one element of the set from $A[i]$ to $B[i]$. Thus the new complement-free set has type $n_1 = \frac{1}{2}\binom{m}{k} - 1$. We continue this step until the descending order $n_1 \geq n_2 \geq \cdots \geq n_m$ can be fulfilled. Since $k \cdot \frac{1}{2}\binom{m}{k}$ is divisible by m there exists exactly one type with $n_1 = \frac{1}{4}\binom{m}{k}$. □

Computations of all types of complement-free sets for small values of m let us conjecture that the distribution of types with $\frac{1}{4}\binom{m}{k} \leq n_1 \leq \frac{1}{2}\binom{m}{k}$ is close to Gaussian distribution. Further, it turns out that computing all types of complement-free

sets for $m = 8$ needs a lot of computation and cannot be done fast. Thus we come to the following open problems.

Problem 3.4. *Determine all types of complement-free sets for fix m!*

Problem 3.5. *Show the distribution of complement-free sets with respect to the largest value in the type! (Is it Gaussian distribution?)*

References

[1] P. Dömösi, C. Hannusch, G. Horváth: *A cryptographic system based on a new class of binary error-correcting codes*, submitted.

[2] P. Dömösi, C. Hannusch, G. Horváth: *Public key cryptographic method and apparatus for data encryption and decryption based on error-correcting codes*, Budapest: Hungarian Intellectual Property Office, patent application, P1800038, 2018.

[3] C. Hannusch, P. Lakatos: *Construction of self-dual binary* $[2^{2k}, 2^{2k-1}, 2^k]$*-codes*, Algebra and Discrete Mathematics 21.1 (2016), pp. 59–68.

[4] T. Herendi: *Construction of uniformly distributed linear recurring sequences modulo powers of 2*, Uniform distribution theory 13.1 (2018), pp. 109–129, doi: 10.1515/udt-2018-0006.

Annales Mathematicae et Informaticae
***50** (2019) pp. 93–100*
DOI: 10.33039/ami.2019.11.001
http://ami.uni-eszterhazy.hu

On the caustics of Bézier curves

Imre Juhász

Department of Descriptive Geometry, University of Miskolc
imre.juhasz@uni-miskolc.hu

Submitted: June 6, 2019
Accepted: November 22, 2019
Published online: November 22, 2019

Abstract

We provide exact formulae for the rational Bézier representation of caustics of planar Bézier curves of degree greater than one.

Keywords: caustic, Bézier curve

MSC: 65D17, 68U07

1. Introduction

In optics a caustic is the envelope of light rays reflected or refracted by an object. We consider only that special case when the rays are parallel and are reflected by a planar curve.

Recently, caustics of control point based planar curves were studied in [6], however the special properties of basis functions in use were not exploited. In the present contribution we concentrate on the caustics of planar Bézier curves, the basis functions of which are the Bernstein polynomials.

2. Caustic curve

Without the loss of generality we can assume that the direction of the light rays is $\left[\begin{array}{cc} 1 & 0 \end{array}\right]^T$, since this only results in an isometric transformation of the curve. We consider the sufficiently smooth curve

$$\mathbf{r}(t) = \left[\begin{array}{c} r_x(t) \\ r_y(t) \end{array}\right], \; t \in [a,b].$$

The direction of the reflected ray at the point $\mathbf{r}(t)$ is

$$\mathbf{v}(t) = \begin{bmatrix} 1 - \frac{2\dot{r}_y^2(t)}{\dot{r}_x^2(t)+\dot{r}_y^2(t)} \\ \frac{2\dot{r}_x(t)\dot{r}_y(t)}{\dot{r}_x^2(t)+\dot{r}_y^2(t)} \end{bmatrix}, \ t \in [a,b]$$

and the caustic $\mathbf{c}$ of the curve $\mathbf{r}$ can be written in the form

$$c_x(t) = r_x(t) + \frac{\left(\dot{r}_x^2(t) - \dot{r}_y^2(t)\right)\dot{r}_y(t)}{2\left(\dot{r}_x(t)\,\ddot{r}_y(t) - \ddot{r}_x(t)\,\dot{r}_y(t)\right)},$$
$$c_y(t) = r_y(t) + \frac{\dot{r}_x(t)\,\dot{r}_y^2(t)}{\dot{r}_x(t)\,\ddot{r}_y(t) - \ddot{r}_x(t)\,\dot{r}_y(t)}.$$

An equivalent of the above formula for the caustic was also derived in [6].

The caustic may have point(s) at infinity, i.e., the curve can be composed of several branches. This happens where the denominator $\dot{r}_x(t)\,\ddot{r}_y(t) - \ddot{r}_x(t)\,\dot{r}_y(t)$ vanishes, i.e., where the curvature of $\mathbf{r}$ is zero. The asymptote at such a point is the reflected ray $\mathbf{r}(t) + \lambda\mathbf{v}(t)$, $\lambda \in \mathbb{R}$ itself. In Fig. 1 there is a quartic Bézier curve the caustic of which has two points at infinity.

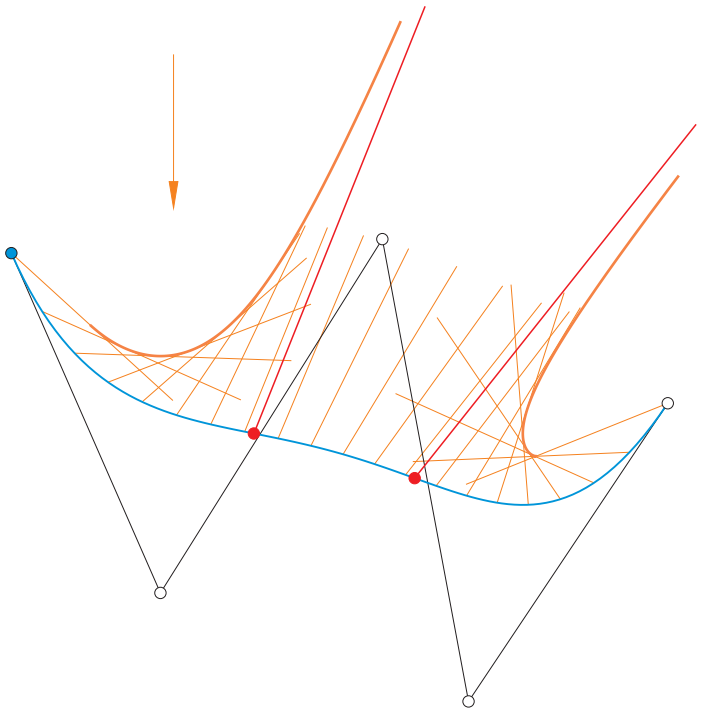


Figure 1: A quartic Bézier curve along with its caustic, which has two points at infinity. The arrow indicates the light direction.

From here on, we will study the caustics of planar Bézier curves

$$\mathbf{r}(t) = \sum_{i=0}^{n} B_i^n(t)\,\mathbf{b}_i,\ t \in [0,1],\ n \geq 2,$$

where the sequence of points $\{\mathbf{b}_i\}_{i=0}^n$ are called control points and B_i^n denotes the ith Bernstein polynomial of degree n. (The case $n = 1$ is out of interest, since then the curve degenerates to a straight line segment and the reflected rays are parallel.)

3. Caustic of a Bézier curve

At first, we reformulate the caustic $\mathbf{c}$ to have a common denominator of the the coordinate functions, yielding

$$c_x(t) = \frac{2\left(\dot{r}_x(t)\,\ddot{r}_y(t) - \ddot{r}_x(t)\,\dot{r}_y(t)\right) r_x(t) + \dot{r}_x^2(t)\,\dot{r}_y(t) - \dot{r}_y^3(t)}{2\left(\dot{r}_x(t)\,\ddot{r}_y(t) - \ddot{r}_x(t)\,\dot{r}_y(t)\right)}, \tag{3.1}$$

$$c_y(t) = \frac{2\left(\dot{r}_x(t)\,\ddot{r}_y(t) - \ddot{r}_x(t)\,\dot{r}_y(t)\right) r_y(t) + 2\dot{r}_x(t)\,\dot{r}_y^2(t)}{2\left(\dot{r}_x(t)\,\ddot{r}_y(t) - \ddot{r}_x(t)\,\dot{r}_y(t)\right)}. \tag{3.2}$$

Obviously, numerators of the above expressions are polynomials of degree $3(n-1)$ and the common denominator is of degree $(2n-3)$, therefore these coordinate functions are rational functions of degree $3(n-1)$. In what follows we provide the rational Bézier representation of such caustics. We introduce notations

$$\dot{\mathbf{r}}(t) = \sum_{i=0}^{n-1} B_i^{n-1}(t)\,\mathbf{a}_i,\ t \in [0,1],\ \mathbf{a}_i = n(\mathbf{b}_{i+1} - \mathbf{b}_i),\ i = 0,1,\ldots,n-1,$$

$$\ddot{\mathbf{r}}(t) = \sum_{i=0}^{n-2} B_i^{n-2}(t)\,\mathbf{d}_i,\ t \in [0,1],\ \mathbf{d}_i = (n-1)(\mathbf{a}_{i+1} - \mathbf{a}_i),\ i = 0,1,\ldots,n-2.$$

Making use of the identity

$$B_i^n(t)\,B_j^m(t) = \frac{\binom{n}{i}\binom{m}{j}}{\binom{n+m}{i+j}} B_{i+j}^{n+m}(t)$$

of Bernstein polynomials (cf. [1]), we can derive an identity for the product of two linear combinations $\sum_{i=0}^n B_i^n(t)\,a_i$ and $\sum_{j=0}^m B_j^m(t)\,b_j$ can be written in the form

$$\sum_{i=0}^{n} B_i^n(t)\,a_i \sum_{j=0}^{m} B_j^m(t)\,b_j = \sum_{\ell=0}^{n+m} B_\ell^{n+m}(t)\,\frac{1}{\binom{n+m}{\ell}} \sum_{k=0}^{m} \binom{n}{\ell-k}\binom{m}{k} a_{\ell-k} b_k, \tag{3.3}$$

provided $n \geq m$.

Now, we study the common denominator. By means of identity 3.3, its first term is of the form

$$\begin{aligned}\dot{r}_x(t)\,\ddot{r}_y(t) &= \sum_{i=0}^{n-1} B_i^{n-1}(t)\,a_{x,i} \sum_{j=0}^{n-2} B_j^{n-2}(t)\,d_{y,j} \\ &= \sum_{\ell=0}^{2n-3} B_\ell^{2n-3}(t)\,\frac{1}{\binom{2n-3}{\ell}} \sum_{k=0}^{n-2} \binom{n-1}{\ell-k}\binom{n-2}{k} a_{x,\ell-k} d_{y,k}.\end{aligned}$$

The second term can analogously be expressed, yielding

$$\dot{r}_y(t)\,\ddot{r}_x(t) = \sum_{\ell=0}^{2n-3} B_\ell^{2n-3}(t)\,\frac{1}{\binom{2n-3}{\ell}} \sum_{k=0}^{n-2} \binom{n-1}{\ell-k}\binom{n-2}{k} a_{y,\ell-k} d_{x,k}$$

and the denominator has the form

$$2\left(\dot{r}_x(t)\,\ddot{r}_y(t) - \ddot{r}_x(t)\,\dot{r}_y(t)\right) = \sum_{\ell=0}^{2n-3} w_\ell B_\ell^{2n-3}(t), \tag{3.4}$$

where

$$w_\ell = \frac{2}{\binom{2n-3}{\ell}} \sum_{k=0}^{n-2} \binom{n-1}{\ell-k}\binom{n-2}{k} \left(a_{x,\ell-k} d_{y,k} - a_{y,\ell-k} d_{x,k}\right). \tag{3.5}$$

We elevate the degree of (3.4) by n, using the general degree elevation formula

$$\begin{aligned}\sum_{i=0}^{s} B_i^s(t)\,w_i &= \sum_{i=0}^{s+z} B_i^{s+z}(t)\,w_i^{[z]},\ z>0, \\ w_i^{[z]} &= w_i^{[z-1]} + \frac{i}{s+z}\left(w_{i-1}^{[z-1]} - w_i^{[z-1]}\right),\ i=0,1,\ldots,s+z \\ w_i^{[0]} &= w_i,\ i=0,1,\ldots,s.\end{aligned} \tag{3.6}$$

with substitutions $s = 3(n-1)$ and $z = n$.

The degree elevated denominator is

$$\sum_{\ell=0}^{3(n-1)} w_\ell^{[n]} B_\ell^{3(n-1)}(t),\ t \in [0,1].$$

The numerator of the x coordinate function is

$$2\left(\dot{r}_x(t)\,\ddot{r}_y(t) - \ddot{r}_x(t)\,\dot{r}_y(t)\right) r_x(t) + \dot{r}_x^2(t)\,\dot{r}_y(t) - \dot{r}_y^3(t).$$

Its first term can be expressed as

$$2\left(\dot{r}_x(t)\,\ddot{r}_y(t) - \ddot{r}_x(t)\,\dot{r}_y(t)\right) r_x(t)$$

$$= \sum_{\ell=0}^{3(n-1)} B_\ell^{3(n-1)}(t) \frac{1}{\binom{3(n-1)}{\ell}} \left(\sum_{k=0}^{n} \binom{2n-3}{\ell-k} \binom{n}{k} w_{\ell-k} b_{x,k} \right)$$

and the second one as

$$\dot{r}_x^2(t)\,\dot{r}_y(t) = \sum_{\ell=0}^{3(n-1)} B_\ell^{3(n-1)}(t) \frac{1}{\binom{3(n-1)}{\ell}} \times \sum_{k=0}^{n-1} \left(\binom{n-1}{k} a_{y,k} \sum_{z=0}^{n-1} \binom{n-1}{\ell-k-z} \binom{n-1}{z} a_{x,\ell-k-z} a_{x,z} \right)$$

while the third one as

$$\dot{r}_y^3(t) = \sum_{\ell=0}^{3(n-1)} B_\ell^{3(n-1)}(t) \frac{1}{\binom{3(n-1)}{\ell}} \times \sum_{k=0}^{n-1} \left(\binom{n-1}{k} a_{y,k} \sum_{z=0}^{n-1} \binom{n-1}{\ell-k-z} \binom{n-1}{z} a_{y,\ell-k-z} a_{y,z} \right).$$

Thus, the numerator of Eq. (3.1) can be written in the form

$$\sum_{\ell=0}^{3(n-1)} B_\ell^{3(n-1)}(t)\, q_{x,\ell},$$

where

$$q_{x,\ell} = \frac{1}{\binom{3(n-1)}{\ell}} \left(\sum_{k=0}^{n} \binom{2n-3}{\ell-k} \binom{n}{k} w_{\ell-k} b_{x,k} + \sum_{k=0}^{n-1} \binom{n-1}{k} a_{y,k} \times \sum_{z=0}^{n-1} \binom{n-1}{\ell-k-z} \binom{n-1}{z} \left(a_{x,\ell-k-z} a_{x,z} - a_{y,\ell-k-z} a_{y,z} \right) \right). \tag{3.7}$$

Analogously, we can obtain the numerator of (3.2) in the form

$$\sum_{\ell=0}^{3(n-1)} B_\ell^{3(n-1)}(t)\, q_{y,\ell},$$

where

$$q_{y,\ell} = \frac{1}{\binom{3(n-1)}{\ell}} \left(\sum_{k=0}^{n} \binom{2n-3}{\ell-k} \binom{n}{k} w_{\ell-k} b_{y,k} + 2 \sum_{k=0}^{n-1} \binom{n-1}{k} a_{x,k} \times \sum_{z=0}^{n-1} \binom{n-1}{\ell-k-z} \binom{n-1}{z} a_{y,\ell-k-z} a_{y,z} \right). \tag{3.8}$$

Finally, the rational Bézier representation of the caustic curve is

$$\mathbf{c}(t) = \sum_{\ell=0}^{3(n-1)} \frac{w_\ell^{[n]} B_\ell^{3(n-1)}(t)}{\sum_{k=0}^{3(n-1)} w_k^{[n]} B_k^{3(n-1)}(t)} \mathbf{q}_\ell, \quad t \in [0,1]$$

where weights $\left\{w_\ell^{[n]}\right\}_{\ell=0}^{3(n-1)}$ are the coefficients obtained by the degree elevation of the denominator, and control points are specified by

$$\mathbf{q}_\ell = \frac{1}{w_\ell^{[n]}} \begin{bmatrix} q_{x,\ell} \\ q_{y,\ell} \end{bmatrix}, \quad \ell = 0, 1, \ldots, 3(n-1). \tag{3.9}$$

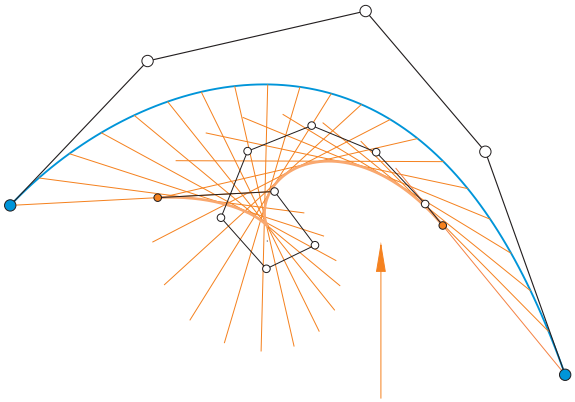


Figure 2: A quartic Bézier curve and its caustic along with the reflected rays. The control polygon of the rational Bézier representation (which is of degree 9) is also displayed. The arrow indicates the light direction.

Now, we summarize our results.

Proposition 3.1. *The caustic of a Bézier curve of degree n (if exists) is a rational Bézier curve of degree $3(n-1)$. Its weights and control points are specified by* (3.5), (3.6) *and* (3.7), (3.8), (3.9), *respectively.*

Remark 3.2. The caustic of a quadratic Bézier curve ($n = 2$) (if exists) is a cubic polynomial curve, since in this case the common denominator of (3.1) and (3.2) is the constant

$$\sum_{\ell=0}^{1} B_\ell^1(t) \left(2 \sum_{k=0}^{1} \frac{\binom{2-1}{\ell-k}\binom{0}{k}}{\binom{1}{\ell}} \left(a_{x,\ell-k} d_{y,k} - a_{y,\ell-k} d_{x,k}\right) \right)$$
$$= 2 \sum_{\ell=0}^{1} B_\ell^1(t) \left(a_{x,\ell} d_{y,0} - a_{y,\ell} d_{x,0}\right)$$

$$= 2\left(a_{x,0}a_{y,1} - a_{x,1}a_{y,0}\right)\left(B_0^1(t) + B_1^1(t)\right)$$
$$= 2\left(a_{x,0}a_{y,1} - a_{x,1}a_{y,0}\right),$$

therefore the caustic is a cubic polynomial curve. It is well-known that the caustic of a parabola is a Tschirnhausen cubic, if the rays are not parallel to the axis of the parabola, we have just obtained its Bézier representation.

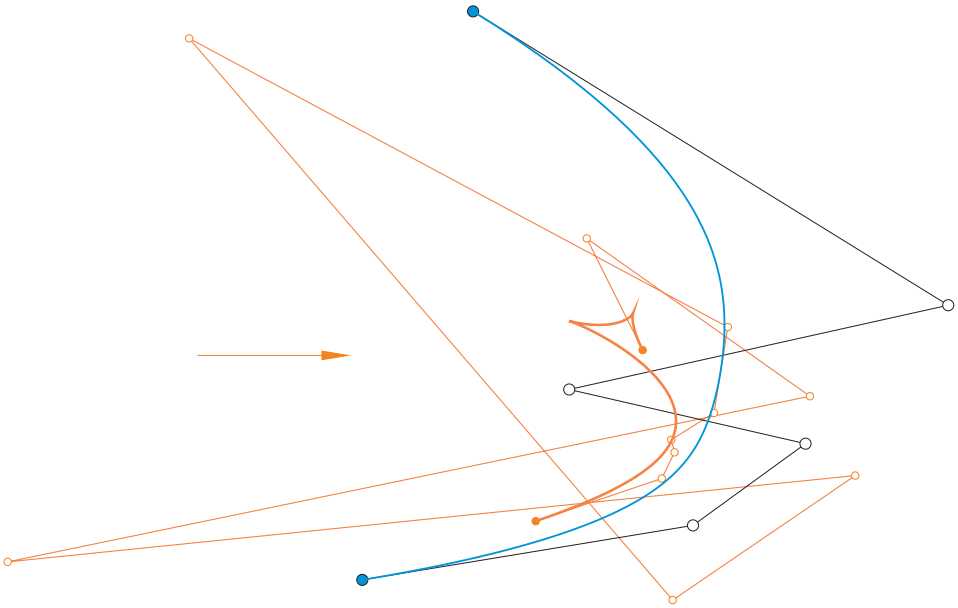


Figure 3: A quintic Bézier curve the caustic of which has two cusps. The control polygon of the rational Bézier representation of the caustic is also shown. The arrow indicates the light direction.

Remark 3.3. The cusp(s) of the caustic may be of interest. The caustic $\mathbf{c}$ has a cusp at $t_0 \in [0,1]$, if $\|\dot{\mathbf{c}}(t_0)\| = 0$, i.e., if the tangent vector vanishes, that we can find numerically. Actually, it is a root finding problem, which can be solved efficiently with high precision and stability, since the polynomials are specified in Bernstein basis (cf. [2, 5]).

In Fig. 2 there is a quartic Bézier curve and its caustic, along with the control polygon of the rational Bézier representation of the caustic. Fig. 3 shows such a quintic Bézier curve whose caustic has two cusps.

4. Conclusions

We have provided ready to implement exact formulae for the rational Bézier representation of caustics (if exist) of planar Bézier curves of degree greater than

one. This method can be extended to other control point based curves, i.e., curves described in the form

$$\mathbf{c}(t) = \sum_{i=0}^{n} F_i^n(t)\,\mathbf{d}_i,\ t \in [a,b].$$

We assume that function system $\mathcal{F} := \{F_i \mid F_i\colon [a,b] \to \mathbb{R}\}_{i=0}^n$ consists of sufficiently smooth non-negative functions, forming a partition of unity. Additional requirements are the existence of degree elevation and product formulae in the basis $\mathcal{F}$. These requirements are fulfilled by the B-basis of trigonometric (cf. [3]) and that of hyperbolic polynomials (cf. [4]), besides the Bernstein basis.

References

[1] G. Farin: *Curves and surfaces for CAGD: a practical guide*, 5th edition, Morgan Kaufmann, 2001, doi: 10.1016/B978-1-55860-737-8.X5000-5.

[2] R. T. Farouki, V. Rajan: *Algorithms for polynomials in Bernstein form*, Computer Aided Geometric Design 5.1 (1988), pp. 1–26, doi: 10.1016/0167-8396(88)90016-7.

[3] J. Sánchez-Reyes: *Harmonic rational Bézier curves, p-Bézier curves and trigonometric polynomials*, Computer Aided Geometric Design 15.9 (1998), pp. 909–923, doi: 10.1016/S0167-8396(98)00031-4.

[4] W.-Q. Shen, G.-Z. Wang: *A class of quasi Bézier curves based on hyperbolic polynomials*, Journal of Zhejiang University Science 6.1 (2005), pp. 116–123, doi: 10.1007/BF02887226.

[5] M. R. Spencer: *Polynomial real root finding in Bernstein form*, PhD thesis, Brigham Young University, 1994, url: https://scholarsarchive.byu.edu/etd/4246.

[6] E. Troll, M. Hoffmann: *Caustics of spline curves*, Annales Mathematicae et Informaticae 47 (2017), pp. 201–209, url: http://ami.ektf.hu/uploads/papers/finalpdf/AMI%5Ctextunderscore47%5Ctextunderscore%20from201to209.pdf.

Annales Mathematicae et Informaticae
***50** (2019) pp. 101–115*
DOI: 10.33039/ami.2019.09.001
http://ami.uni-eszterhazy.hu

Pillai's problem with the Fibonacci and Padovan sequences

Ana Cecilia García Lomelí[a*], Santos Hernández Hernández[a], Florian Luca[bcd†]

[a]Unidad Académica de Matemáticas
Universidad Autónoma de Zacatecas, Campus II
Zacatecas, Zac., México
aceciliagarcia.lomeli@gmail.com
shh@uaz.edu.mx

[b]School of Mathematics
University of the Witwatersrand
Johannesburg, South Africa

[c]Research Group in Algebraic Structures and Applications
King Abdulaziz University
Jeddah, Saudi Arabia

[d]Department of Mathematics
Faculty of Sciences, University of Ostrava
Ostrava, Czech Republic
Florian.Luca@wits.ac.za

Submitted: June 7, 2018
Accepted: September 6, 2019
Published online: September 23, 2019

Abstract

Let $(F_m)_{m\geqslant 0}$ and $(P_n)_{n\geqslant 0}$ be the Fibonacci and Padovan sequences given by the initial conditions $F_0 = 0, F_1 = 1, P_0 = 0, P_1 = P_2 = 1$ and the recurrence formulas $F_{m+2} = F_{m+1} + F_m, P_{n+3} = P_{n+1} + P_n$ for all $m, n \geqslant 0$, respectively. In this note we study and completely solve the Diophantine

*Supported by a CONACyT Doctoral Fellowship.
†Supported in part by grant CPRR160325161141 and an A-rated scientist award both from the NRF of South Africa and by grant no. 17-02804S of the Czech Granting Agency.

equation

$$P_n - F_m = P_{n_1} - F_{m_1}$$

in non-negative integers (n, m, n_1, m_1) with $(n, m) \neq (n_1, m_1)$.

Keywords: Fibonacci, Padovan sequences, Pillai's type problem, Linear form in logarithms.

MSC: 11B39, 11D45, 11D61, 11J86.

1. Introduction

Let a, b be fixed positive integers and consider the Diophatine equation

$$a^n - b^m = a^{n_1} - b^{m_1} \tag{1.1}$$

in positive integers n, m, n_1, m_1 with $(n, m) \neq (n_1, m_1)$. In particular, we look for the integers which can be written as a difference of a power of a and a power of b in at least two distinct ways. In [11], Herschfeld proved that in the case $(a, b) = (2, 3)$ equation (1.1) has only finitely many solutions. In [15], Pillai extended this result to the case a, $b \geqslant 2$ being coprime integers. Both results are ineffective. In [16], Pillai conjectured that in the case $(a, b) = (2, 3)$ the only solutions of equation (1.1) are $(3, 2, 1, 1)$, $(5, 3, 3, 1)$ and $(8, 5, 4, 1)$. This conjecture remained open for about 37 years and was confirmed in [20] by Stroeker and Tijdeman by using Baker's theory on linear forms in logarithms.

Recently, the above problem now known as the *Pillai problem*, was posed in the context of linear recurrence sequences. Namely, let $\mathbf{U} := (U_n)_{n \geqslant 0}$ and $\mathbf{V} := (V_m)_{m \geqslant 0}$ be two linearly recurrence sequences of integers and look at the diophantine equation

$$U_n - V_m = U_{n_1} - V_{n_1} \tag{1.2}$$

in positive integers n, m, n_1, m_1 with $(n, m) \neq (n_1, m_1)$. This reduces to determining the integers which can be written as a difference of an element of $\mathbf{U}$ and an element of $\mathbf{V}$ in at least two distinct ways. This version was started by Ddamulira, Luca and Rakotomalala in [8] where they considered $\mathbf{U}$ as being the Fibonacci sequence and $\mathbf{V}$ as being the sequence of powers of 2. Many other cases have been studied, see for example [3, 6, 7, 10, 12, 13]. In [5], there is a general result, namely that if $\mathbf{U}$ and $\mathbf{V}$ satisfy some natural conditions, then equation (1.2) has only finitely many solutions which furthermore are all effectively computable. We recall that the *Fibonacci sequence* $(F_m)_{m \geqslant 0}$ is given by $F_0 = 0$, $F_1 = 1$ and the recurrence formula

$$F_{m+2} = F_{m+1} + F_m \quad \text{for all} \quad m \geqslant 0.$$

Its first few terms are

$$0, 1, 1, 2, 3, 5, 8, 13, 21, 34, 55, 89, 144, 233, 377, 610, 987, 1597, \ldots$$

Now, let $(P_n)_{n\geqslant 0}$ be the *Padovan sequence*, named after the architect R. Padovan, given by $P_0 = 0$, $P_1 = P_2 = 1$ and the recurrence formula

$$P_{n+3} = P_{n+1} + P_n \quad \text{for all} \quad n \geqslant 0.$$

This is the sequence A000931 in [18]. Its first few terms are

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

In this note, we study another case of equation (1.2) namely with the Fibonacci and the Padovan sequences. More precisely, we solve the equation

$$P_n - F_m = P_{n_1} - F_{m_1} \tag{1.3}$$

in non-negative integers (n, m, n_1, m_1) with $(n, m) \neq (n_1, m_1)$. To avoid numerical repeated solutions we assume that $n \neq 1, 2, 4$ and $n_1 \neq 1, 2, 4$. That is whenever we think of 1 and 2 as members of the Padovan sequence que think of them as being P_3 and P_5, respectively. In the same way, $m \neq 1$ and $m_1 \neq 1$. With this conventions, our result is the following:

Theorem 1.1. *All non-negative integer solutions* (n, m, n_1, m_1) *of equation* (1.3) *belong to the set*

$$\left\{\begin{array}{ccccc}
(3,2,0,0), & (3,3,0,2), & (3,4,0,3), & (5,2,3,0), & (5,3,3,2),\\
(5,3,0,0), & (5,4,3,3), & (5,4,0,2), & (5,5,0,4), & (6,2,5,0),\\
(6,3,5,2), & (6,3,3,0), & (6,4,5,3), & (6,4,3,2), & (6,4,0,0),\\
(6,5,3,4), & (6,5,0,3), & (6,6,0,5), & (7,2,6,0), & (7,3,6,2),\\
(7,3,5,0), & (7,4,6,3), & (7,4,5,2), & (7,4,3,0), & (7,5,5,4),\\
(7,5,3,3), & (7,5,0,2), & (7,6,3,5), & (8,2,7,0), & (8,3,7,2),\\
(8,3,6,0), & (8,4,7,3), & (8,4,6,2), & (8,4,5,0), & (8,5,6,4),\\
(8,5,5,3), & (8,5,3,2), & (8,5,0,0), & (8,6,5,5), & (8,6,0,4),\\
(8,7,0,6), & (9,3,8,0), & (9,4,8,2), & (9,4.7,0), & (9,5,8,4),\\
(9,5,7,3), & (9,5,6,2), & (9,5,5,0), & (9,6,7,5), & (9,6,5,4),\\
(9,6,3,3), & (9,6,0,2), & (9,7,5,6), & (10,3,9,0), & (10,4,9,2),\\
(10,5,9,4), & (10,5,8,2), & (10,5,7,0), & (10,6,7,4), & (10,6,6,3),\\
(10,6,5,2), & (10,6,3,0), & (10,7,7,6), & (10,7,3,5), & (10,8,3,7),\\
(11,4,10,0), & (11,5,10,3), & (11,5,9,0), & (11,6,10,5), & (11,6,9,4),\\
(11,6,8,2), & (11,6,7,0), & (11,7,9,6), & (11,7,7,5), & (11,7,5,4),\\
(11,7,3,3), & (11,7,0,2), & (11,8,7,7), & (12,5,11,2), & (12,6,10,2),\\
(12,7,8,3), & (12,7,7,2), & (12,7,6,0), & (12,8,6,6), & (12,8,0,5),\\
(12,9,6,8), & (13,5,12,0), & (13,6,12,4), & (13,7,12,6), & (13,7,10,2),\\
(13,8,8,5), & (13,8,6,4), & (13,8,5,3), & (13,8,3,2), & (13,8,0,0),\\
(13,9,0,7), & (13,10,0,9), & (14,6,13,2), & (14,7,12,2), & (14,8,11,5),\\
(14,8,10,3), & (14,8,9,0), & (14,9,9,7), & (14,9,5,6), & (14,10,9,9),\\
(15,8,13,5), & (15,8,12,0), & (15,9,12,7), & (15,9,8,3), & (15,9,7,2),\\
(15,9,6,0), & (15,10,12,9), & (15,10,6,8), & (15,11,6,10), & (16,7,15,2),\\
(16,8,14,0), & (16,9,14,7), & (16,9,12,2), & (16,10,14,9), & (16,10,9,7),\\
(16,10,5,6), & (17,8,16,5), & (17,10,11,3), & (18,8,17,0), & (18,9,17,7),
\end{array}\right\}$$

$$\left\{\begin{array}{ccccc}(18,10,17,9), & (18,11,8,6), & (18,11,5,5), & (18,11,0,4), & (19,11,14,4),\\ (19,12,7,9), & (20,11,17,4), & (20,12,14,8), & (20,12,11,5), & (20,12,10,3),\\ (20,12,9,0), & (20,13,9,11), & (20,14,9,13), & (21,11,19,4), & (21,13,3,9),\\ (22,13,15,5), & (23,11,22,4), & (25,15,10,4), & (25,15,9,2) & \end{array}\right\}$$

The set of integers which can be written as the difference of a Padovan number and a Fibonacci number in at least two distinct ways is

$$\left\{\begin{array}{ccccccccc}-226, & -82, & -52, & -34, & -33, & -30, & -27, & -18, & -13,\\ -12, & -9, & -8, & -6, & -5, & -4, & -3, & -2, & -1,\\ 0, & 1, & 2, & 3, & 4, & 5, & 6, & 7, & 8,\\ 9, & 10, & 11, & 13, & 15, & 16, & 20, & 25, & 28,\\ 31, & 32, & 36, & 44, & 52, & 62, & 65, & 111, & 262.\end{array}\right\}.$$

All such representations of each of these numbers are

$$\begin{aligned}
-226 &= P_{20} - F_{14} = P_9 - F_{13};\\
-82 &= P_{20} - F_{13} = P_9 - F_{11};\\
-52 &= P_{15} - F_{11} = P_6 - F_{10};\\
-34 &= P_{13} - F_{10} = P_0 - F_9;\\
-33 &= P_{21} - F_{13} = P_3 - F_9;\\
-30 &= P_{19} - F_{12} = P_7 - F_9;\\
-27 &= P_{14} - F_{10} = P_9 - F_9;\\
-18 &= P_{12} - F_9 = P_6 - F_8 = P_{15} - F_{10};\\
-13 &= P_{13} - F_9 = P_0 - F_7;\\
-12 &= P_{10} - F_8 = P_3 - F_7;\\
-9 &= P_{11} - F_8 = P_7 - F_7;\\
-8 &= P_8 - F_7 = P_0 - F_6;\\
-6 &= P_{16} - F_{10} = P_{14} - F_9 = P_9 - F_7 = P_5 - F_6;\\
-5 &= P_{12} - F_8 = P_6 - F_6 = P_0 - F_5;\\
-4 &= P_{10} - F_7 = P_7 - F_6 = P_3 - F_5;\\
-3 &= P_{18} - F_{11} = P_8 - F_6 = P_5 - F_5 = P_0 - F_4;\\
-2 &= P_6 - F_5 = P_3 - F_4 = P_0 - F_3;\\
-1 &= P_{11} - F_7 = P_9 - F_6 = P_7 - F_5 = P_5 - F_4 = P_3 - F_3 = P_0 - F_2;\\
0 &= P_{13} - F_8 = P_8 - F_5 = P_6 - F_4 = P_5 - F_3 = P_3 - F_2 = P_0 - F_0;\\
1 &= P_{10} - F_6 = P_7 - F_4 = P_6 - F_3 = P_5 - F_2 = P_3 - F_0;\\
2 &= P_9 - F_5 = P_8 - F_4 = P_7 - F_3 = P_6 - F_2 = P_5 - F_0;\\
3 &= P_{15} - F_9 = P_{12} - F_7 = P_8 - F_3 = P_7 - F_2 = P_6 - F_0;\\
4 &= P_{11} - F_6 = P_{10} - F_5 = P_9 - F_4 = P_8 - F_2 = P_7 - F_0;\\
5 &= P_9 - F_3 = P_8 - F_0;\\
6 &= P_{25} - F_{15} = P_{10} - F_4 = P_9 - F_2;
\end{aligned}$$

$$
\begin{aligned}
7 &= P_{20} - F_{12} = P_{14} - F_8 = P_{11} - F_5 = P_{10} - F_3 = P_9 - F_0;\\
8 &= P_{13} - F_7 = P_{12} - F_6 = P_{10} - F_2;\\
9 &= P_{11} - F_4 = P_{10} - F_0;\\
10 &= P_{17} - F_{10} = P_{11} - F_3;\\
11 &= P_{12} - F_5 = P_{11} - F_2;\\
13 &= P_{13} - F_6 = P_{12} - F_4;\\
15 &= P_{16} - F_9 = P_{14} - F_7 = P_{12} - F_2;\\
16 &= P_{15} - F_8 = P_{13} - F_5 = P_{12} - F_0;\\
20 &= P_{14} - F_6 = P_{13} - F_2;\\
25 &= P_{19} - F_{11} = P_{14} - F_4;\\
28 &= P_{16} - F_8 = P_{14} - F_0;\\
31 &= P_{18} - F_{10} = P_{17} - F_9;\\
32 &= P_{22} - F_{13} = P_{15} - F_5;\\
36 &= P_{16} - F_7 = P_{15} - F_2;\\
44 &= P_{17} - F_8 = P_{16} - F_5;\\
52 &= P_{18} - F_9 = P_{17} - F_7;\\
62 &= P_{20} - F_{11} = P_{17} - F_4;\\
65 &= P_{18} - F_8 = P_{17} - F_0;\\
111 &= P_{21} - F_{11} = P_{19} - F_4;\\
262 &= P_{23} - F_{11} = P_{22} - F_4\,.
\end{aligned}
$$

In [19], Stewart notes that 3, 5 and 21 are both Fibonacci and Padovan numbers and asks whether there are any others. This problem was solved by De Weger in [21], where he proves that all integers which are both Fibonacci and Padovan numbers are 0, 1, 2, 3, 5, 21. Actually, he proves that the distance between Fibonacci and Padovan numbers growths exponentially. We remark that as a particular case of our result, we also have a solution of Stewart problem.

2. Tools

In this section, we gather the tools we need to prove Theorem 1.1. Let α be an algebraic number of degree d, let $a > 0$ be the leading coefficient of its minimal polynomial over $\mathbb{Z}$ and let $\alpha^{(1)}, \ldots, \alpha^{(d)}$ denote its conjugates. The *logarithmic height* of α is defined as

$$h(\alpha) = \frac{1}{d}\left(\log a + \sum_{i=1}^{d} \log \max\left\{|\alpha^{(i)}|, 1\right\}\right).$$

This height satisfies the following basic properties. For α, β algebraic numbers and $m \in \mathbb{Z}$ we have

- $h(\alpha+\beta) \leqslant h(\alpha)+h(\beta)+\log(2)$,
- $h(\alpha\beta) \leqslant h(\alpha)+h(\beta)$,
- $h(\alpha^m) = |m|h(\alpha)$.

Now, let $\mathbb{L}$ be a real number field of degree $d_{\mathbb{L}}$, $\alpha_1,\ldots,\alpha_\ell$ positive elements of $\mathbb{L}$ and $b_1,\ldots,b_\ell \in \mathbb{Z}\setminus\{0\}$. Let $B \geqslant \max\{|b_1|,\ldots,|b_\ell|\}$ and

$$\Lambda = \alpha_1^{b_1}\cdots\alpha_\ell^{b_\ell} - 1.$$

Let $A_1,\ldots,A_\ell$ be real numbers with

$$A_i \geqslant \max\{d_{\mathbb{L}}\, h(\alpha_i), |\log\alpha_i|, 0.16\}, \qquad i=1,2,\ldots,\ell.$$

The first tool we need is the following result due to Matveev in [14] (see also Theorem 9.4 in [4]).

Theorem 2.1. *Assume that $\Lambda \neq 0$. Then*

$$\log|\Lambda| > -1.4\cdot 30^{\ell+3}\cdot \ell^{4.5}\cdot d_{\mathbb{L}}^2\cdot(1+\log d_{\mathbb{L}})\cdot(1+\log B)A_1\cdots A_\ell.$$

In this note we always use $\ell = 3$. Further, $\mathbb{L} = \mathbb{Q}(\gamma,\alpha)$ has degree $d_{\mathbb{L}} = 6$, where γ and α are defined at the beginning of Section 3. Thus, once and for all we fix the constant

$$C := 1.43908\times 10^{13} \;>\; 1.4\cdot 30^{3+3}\cdot 3^{4.5}\cdot 6^2\cdot(1+\log 6)$$

The second one, is a version of the reduction method of Baker-Davenport based on Lemma in [1]. We shall use the one given by Bravo, Gómez and Luca in [2] (See also Dujella and Pethő [9]). For a real number x, we write $\|x\|$ for the distance from x to the nearest integer.

Lemma 2.2. *Let M be a positive integer. Let $\tau,\mu,A>0,B>1$ be given real numbers. Assume that p/q is a convergent of τ such that $q>6M$ and that $\varepsilon := \|q\,\mu\| - M\|q\,\tau\| > 0$. Then there is no solution to the inequality*

$$0 < |n\tau - m + \mu| < \frac{A}{B^w}$$

in positive integers n,m and w satisfying

$$n \leqslant M \quad \text{and} \quad w \geqslant \frac{\log(Aq/\varepsilon)}{\log B}.$$

Finally, the following result will be very useful. This is Lemma 7 in [17].

Lemma 2.3. *If $m \geqslant 1$, $T > (4m^2)^m$ and $T > x/(\log x)^m$. Then*

$$x < 2^m T(\log T)^m.$$

3. Proof of Theorem 1.1

We start with some basic properties of our sequences. For a complex number z we write $\overline{z}$ for its complex conjugate. Let $\omega \neq 1$ be a cubic root of 1. Put

$$\gamma := \sqrt[3]{\frac{9+\sqrt{69}}{18}} + \sqrt[3]{\frac{9-\sqrt{69}}{18}}, \quad \delta := \omega\sqrt[3]{\frac{9+\sqrt{69}}{18}} + \overline{\omega}\sqrt[3]{\frac{9-\sqrt{69}}{18}},$$

and

$$\alpha := \frac{1+\sqrt{5}}{2}, \quad \beta := \frac{1-\sqrt{5}}{2}.$$

It is clear that $\gamma, \delta, \overline{\delta}$ are the roots of the $\mathbb{Q}$-irreducible polynomial $X^3 - X - 1$. It can be proved, by induction for example, that the Binet formulas

$$F_n = \frac{\alpha^n - \beta^n}{\sqrt{5}} \quad \text{and} \quad P_n = c_1\gamma^n + c_2\delta^n + c_3\overline{\delta}^n \quad \text{hold for all} \quad n \geqslant 0, \tag{3.1}$$

where

$$c_1 = \frac{\gamma(\gamma+1)}{2\gamma+3}, \quad c_2 = \frac{\delta(\delta+1)}{2\delta+3}, \quad c_3 = \overline{c_2}.$$

The first formula in (3.1) is well known. The second one follows from the general theorem on linear recurrence sequences since the above polynomial is the characteristic polynomial of the Padovan sequence. Further, the inequalities

$$\alpha^{n-2} \leqslant F_n \leqslant \alpha^{n-1}, \quad \gamma^{n-3} \leqslant P_n \leqslant \gamma^{n-1} \tag{3.2}$$

also hold for all $n \geqslant 1$. These can be proved by induction. We note that

$$\gamma = 1.32471\ldots, \quad |\delta| = 0.86883\ldots, \quad c_1 = 0.54511\ldots, \quad |c_2| = 0.28241\ldots,$$

and

$$\alpha = 1.61803\ldots, \quad |\beta| = 0.61803\ldots$$

Now we start with the study of our equation (1.3) in non-negative integers (n, m, n_1, m_1) with $(n, m) \neq (n_1, m_1)$ where, as we have said, $n, n_1 \neq 1, 2, 4$, $m, m_1 \neq 1$. We note, if $m = m_1$ then $P_n = P_{n_1}$ which implies $n = n_1$, a contradiction. Thus, we assume that $m > m_1$. Rewriting equation (1.3) as

$$P_n - P_{n_1} = F_m - F_{m_1} \tag{3.3}$$

we observe the right-hand is positive. So, the left-hand side is also positive and therefore, $n > n_1$. Now, we compare both sides of (3.3) using (3.2). We have

$$\gamma^{n-8} \leqslant P_n - P_{n_1} = F_m - F_{m_1} \leqslant F_m \leqslant \alpha^{m-1}.$$

Indeed, the left-hand side inequality is clear if $n_1 = 0$. If $n_1 = 3, n \geqslant 5$. For $n = 5$ it is also clear and for $n \geqslant 6$ we have $P_n - P_{n_1} \geqslant P_n - P_{n-1} = P_{n-5} \geqslant \gamma^{n-8}$. Thus, $\gamma^{n-8} \leqslant \alpha^{m-1}$. In a similar way,

$$\gamma^{n-1} \geqslant P_n - P_{n_1} = F_m - F_{m_1} \geqslant \alpha^{m-4}.$$

where the inequality at the right-hand side is clear for both $m_1 = 0$ and $m_1 \neq 0$. Thus,

$$(n-8)\frac{\log\gamma}{\log\alpha} \leqslant m-1 \quad \text{and} \quad (n-1)\frac{\log\gamma}{\log\alpha} \geqslant m-4. \tag{3.4}$$

Since $\log\gamma/\log\alpha = 0.584357\ldots$ we have that if $n \leqslant 540$ then $m \leqslant 318$. A brute force search with *Mathematica* in the range $0 \leqslant n_1 < n \leqslant 540$, $0 \leqslant m_1 < m \leqslant 318$, with our conventions, we obtained all solutions listed in Theorem 1.1.

From now on, we assume that $n > 540$. Thus, from (3.4), we have that $m > 311$ and also that $n > m$. From Binet's formula (3.1), we rewrite our equation as

$$\left| c_1\gamma^n - \frac{\alpha^m}{\sqrt{5}} \right| \leqslant 2|c_2||\delta|^n + \frac{1}{\sqrt{5}} + \gamma^{n_1-1} + \alpha^{m_1-1} < \max\{\gamma^{n_1+6}, \alpha^{m_1+4}\}.$$

Dividing through by $\alpha^m/\sqrt{5}$ we get

$$\left| \sqrt{5}c_1\gamma^n\alpha^{-m} - 1 \right| < \max\{\gamma^{n_1-n+16}, \alpha^{m_1-m+6}\}, \tag{3.5}$$

where we have used $\gamma^{n-8} \leqslant \alpha^{m-1}$, $\sqrt{5} < \alpha\gamma^2$ and $\sqrt{5} < \alpha^2$. Let Λ be the expression inside the absolute value in the left-hand side of (3.5). Observe that $\Lambda \neq 0$. To see this, we consider the $\mathbb{Q}$-automorphism σ of the Galois extension $\mathbb{K} := \mathbb{Q}(\alpha, \gamma, \delta)$ over $\mathbb{Q}$ defined by $\sigma(\gamma) := \delta, \sigma(\delta) := \gamma$ and $\sigma(\alpha) := \alpha$. We note that $\sigma(\overline{\delta}) = \overline{\delta}$ and $\sigma(\beta) = \beta$. If $\Lambda = 0$ then $\sigma(\Lambda) = 0$ and we get

$$\frac{\alpha^m}{\sqrt{5}} = \sigma(c_1\gamma^n) = c_2\delta^n.$$

Thus,

$$\frac{\alpha^m}{\sqrt{5}} = |c_2||\delta|^n < 1,$$

which is absurd since $m > 311$. So, $\Lambda \neq 0$. We apply Matveev's inequality to Λ by taking

$$\alpha_1 = \sqrt{5}c_1,\ \alpha_2 = \gamma,\ \alpha_3 = \alpha, \quad b_1 = 1,\ b_2 = n,\ b_3 = -m.$$

Thus, $B = n$. Further, $h(\alpha_2) = \log\gamma/3$, $h(\alpha_3) = \log\alpha/2$. For α_1 we use the properties of the height to conclude

$$h(\alpha_1) \leqslant \log\gamma + 7\log 2.$$

So we take $A_1 = 30.8$, $A_2 = 0.57$, $A_3 = 1.45$. From Matveev's inequality we obtain

$$\log|\Lambda| > -C(1+\log n)\cdot 30.8\cdot 0.57\cdot 1.45 > -3.66336\times 10^{14}(1+\log n),$$

which, compared with (3.5) we obtain

$$\min\{(n-n_1)\log\gamma, (m-m_1)\log\alpha\} \leqslant 3.66337\times 10^{14}(1+\log n).$$

Now we study each one of these two possibilities.

Case 1. $\min\{(n-n_1)\log\gamma, (m-m_1)\log\alpha\} = (n-n_1)\log\gamma$.

In this case, using Binet's formulas (3.1), we rewrite our equation as

$$\left|c_1(\gamma^{n-n_1}-1)\gamma^{n_1} - \frac{\alpha^m}{\sqrt{5}}\right| \leqslant 4|c_2||\delta|^{n_1} + 1 + \alpha^{m_1-1} < 2\cdot\alpha^{m_1+2} \leqslant \alpha^{m_1+4}.$$

Thus,

$$\left|c_1\sqrt{5}(\gamma^{n-n_1}-1)\gamma^{n_1}\alpha^{-m} - 1\right| < \frac{1}{\alpha^{m-m_1-6}}. \tag{3.6}$$

Let Λ_1 be the expression inside the absolute value in the left-hand side of (3.6). We note that $\Lambda_1 \neq 0$. For if not, we apply the above σ to it and we have $\sigma(\Lambda_1) = 0$. Thus,

$$\frac{\alpha^m}{\sqrt{5}} = |\sigma(c_1)(\delta^n - \delta^{n_1})| \leqslant 2|c_2| < 1,$$

which is absurd since $m > 311$. We apply Matveev's inequality to Λ_1 and for this we take

$$\alpha_1 = \sqrt{5}c_1(\gamma^{n-n_1}-1),\ \alpha_2 = \gamma,\ \alpha_3 = \alpha, \quad b_1 = 1,\ b_2 = n_1,\ b_3 = -m.$$

We have $B = n$. The heights of α_2 and α_3 are already calculated. For α_1 we use the height properties and we get

$$h(\alpha_1) \leqslant \frac{3.66338\times 10^{14}(1+\log n)}{3}.$$

Thus, we can take $A_1 = 7.32676\times 10^{14}(1+\log n)$ and A_2, A_3 as above. From Matveev's inequality we obtain

$$\log|\Lambda_1| > -C(1+\log n)\cdot(7.32676\times 10^{14}(1+\log n))\cdot 0.57\cdot 1.45,$$

which compared with (3.6) gives

$$(m-m_1)\log\alpha < 8.71446\times 10^{27}(1+\log n)^2.$$

Case 2. $\min\{(n-n_1)\log\gamma, (m-m_1)\log\alpha\} = (m-m_1)\log\alpha$.

To this case, we rewrite our equation as

$$\left|c_1\gamma^n - \frac{(\alpha^{m-m_1}-1)\alpha^{m_1}}{\sqrt{5}}\right| < \gamma^{n_1-1} + 2|c_2| + 1 < \gamma^{n_1+4}.$$

Thus,

$$\left|1 - \left(\frac{\alpha^{m-m_1}-1}{\sqrt{5}c_1}\right)\gamma^{-n}\alpha^{m_1}\right| < \frac{1}{\gamma^{n-n_1-7}}, \tag{3.7}$$

where we have used $1 < c_1\gamma^3$. Let Λ_2 be the expression inside the absolute value in the left-hand side of (3.7). We note that $\Lambda_2 \neq 0$. Indeed, if it is not the case then by applying the above σ to it we obtain $\sigma(\Lambda_2) = 0$. Thus

$$1 < \frac{\alpha^{m-1}(\alpha-1)}{\sqrt{5}} \leqslant \frac{\alpha^m - \alpha^{m_1}}{\sqrt{5}} = \sqrt{5}|c_2||\delta|^n < \sqrt{5}|c_2| < 1,$$

where the left-hand side inequality holds since $m > 311$, which is absurd. So, $\Lambda_2 \neq 0$ and we apply Matveev's inequality to it. To do this, we take

$$\alpha_1 = \frac{\alpha^{m-m_1} - 1}{\sqrt{5}c_1},\ \alpha_2 = \gamma,\ \alpha_3 = \alpha, \quad b_1 = 1,\ b_2 = -n,\ b_3 = m_1.$$

Thus, $B = n$. The heights of α_2 and α_3 are already calculated. From the properties of the height for α_1 we obtain

$$h(\alpha_1) \leqslant \frac{3.66338 \times 10^{14}(1 + \log n)}{2}.$$

Thus, we can take $A_1 = 1.09901 \times 10^{15}(1 + \log n)$ and A_2, A_3 as above. Hence, from Matveev's inequality we obtain

$$\log|\Lambda_2| > -C(1 + \log n) \cdot \left(1.09901 \times 10^{15}(1 + \log n)\right) \cdot 0.57 \cdot 1.45,$$

which compared with (3.7) we get

$$(n - n_1)\log\gamma < 1.30717 \times 10^{28}(1 + \log n)^2.$$

So, from the conclusion of the two cases we have that

$$\max\{(n - n_1)\log\gamma, (m - m_1)\log 2\} < 1.30717 \times 10^{28}(1 + \log n)^2.$$

Now we get a bound on n. To do this we rewrite our equation as

$$\left| c_1(\gamma^{n-n_1} - 1)\gamma^{n_1} - \frac{(\alpha^{m-m_1} - 1)\alpha^{m_1}}{\sqrt{5}} \right| < 4|c_2| + 1 < 2.2.$$

Thus,

$$\left| \left(\sqrt{5}c_1 \frac{\gamma^{n-n_1} - 1}{\alpha^{m-m_1} - 1}\right)\gamma^{n_1}\alpha^{-m_1} - 1 \right| < \frac{2.2 \cdot \sqrt{5}}{\alpha^m - \alpha^{m_1}} \leqslant \frac{6.6 \cdot \sqrt{5}}{\alpha^m} < \frac{1}{\gamma^{n-16}}, \tag{3.8}$$

where we have used $\gamma^{n-8} < \alpha^{m-1}$ and $6.6 \cdot \sqrt{5} < \alpha\gamma^8$. Let Λ_3 be the expression inside the absolute value in the left-hand side of (3.8). As above, if $\Lambda_3 = 0$ we apply the above σ and we obtain $\sigma(\Lambda_3) = 0$. Then

$$1 < \frac{\alpha^{m-1}(\alpha - 1)}{\sqrt{5}} \leqslant \frac{\alpha^m - \alpha^{m_1}}{\sqrt{5}} = |c_2(\delta^n - \delta^{n_1})| \leqslant 2|c_2| < \frac{2}{3},$$

and as above, we get a contradiction. Thus, $\Lambda_3 \neq 0$ and we apply Matveev's inequality to it. To do this, we take

$$\alpha_1 = \sqrt{5}c_1 \frac{\gamma^{n-n_1} - 1}{\alpha^{m-m_1} - 1},\ \alpha_2 = \gamma,\ \alpha_3 = \alpha, \quad b_1 = 1,\ b_2 = n_1,\ b_3 = -m_1.$$

Hence, $B = n$. The height of α_2 and α_3 have already been calculated. For α_1 we use the properties of the height to conclude that

$$\begin{aligned} h(\alpha_1) &\leqslant \log\gamma + (n-n_1)\frac{\log\gamma}{3} + (m-m_1)\frac{\log\alpha}{2} + 9\log 2 \\ &< \frac{6.53586\times 10^{28}(1+\log n)^2}{6}. \end{aligned}$$

Thus, we can take $A_1 = 6.53586\times 10^{28}(1+\log n)^2$ and A_2, A_3 as above. From Matveev's inequality we get

$$\log|\Lambda_3| > -C\cdot\left((1+\log n)\cdot 6.53586\times 10^{28}(1+\log n)^2\right)\cdot 0.57\cdot 1.45,$$

which compared with (3.8) yields $n < 2.2116\times 10^{43}(\log n)^3$. Thus, from Lemma 2.3 we obtain

$$n < 1.75894\times 10^{50}. \tag{3.9}$$

Now we reduce this upper bound on n. To do this, let Γ be defined as

$$\Gamma = n\log\gamma - m\log\alpha + \log\left(\sqrt{5}\,c_1\right),$$

and we go to (3.5). Assume that $\min\{n-n_1, m-m_1\} \geqslant 20$. Observe that $e^{\Gamma}-1 = \Lambda \neq 0$. Therefore $\Gamma \neq 0$. If $\Gamma > 0$, then

$$0 < \Gamma < e^{\Gamma}-1 = |\Lambda| < \max\{\gamma^{n_1-n+16}, \alpha^{m_1-m+6}\}.$$

If $\Gamma < 0$, we then have $1-e^{\Gamma} = |e^{\Gamma}-1| = |\Lambda| < 1/2$. Thus, $e^{|\Gamma|} < 2$ and we get

$$0 < |\Gamma| < e^{|\Gamma|}-1 = e^{|\Gamma|}|\Lambda| < 2\max\{\gamma^{n_1-n+16}, \alpha^{m_1-m+6}\}.$$

So, in both cases we have

$$0 < |\Gamma| < 2\max\{\gamma^{n_1-n+16}, \alpha^{m_1-m+6}\}.$$

Dividing through $\log\alpha$ we get

$$0 < |n\tau - m + \mu| < \max\left\{\frac{374}{\gamma^{n-n_1}}, \frac{75}{\alpha^{m-m_1}}\right\},$$

where

$$\tau := \frac{\log\gamma}{\log\alpha}, \quad \mu := \frac{\log\left(\sqrt{5}\,c_1\right)}{\log\alpha}.$$

We apply Lemma 2.2. To do this we take $M := 1.75894\times 10^{50}$ which is the upper bound on n by (3.9). With the help of *Mathematica* we found that the convergent

$$\frac{p_{111}}{q_{111}} = \frac{1055018110290384419279582749015021525092270854503951799 7}{1805433708589770760526539129691547197889880925836949175 4}$$

of τ satisfies that $q_{111} > 6M$ and that $\varepsilon := \|q_{111}\mu\| - M\|q_{111}\tau\| = 0.450294 > 0$. Thus, by Lemma 2.2 with $A := 374$, $B := \gamma$ or $A := 75$, $B := \alpha$, we get that either

$$n - n_1 \leqslant 476 \quad \text{or} \quad m - m_1 \leqslant 275.$$

Now we study each one of these two cases. We first assume that $n - n_1 \leqslant 476$ and $m - m_1 \geqslant 20$. In this case, we consider

$$\Gamma_1 = n_1 \log\gamma - m\log\alpha + \log(\sqrt{5}c_1(\gamma^{n-n_1} - 1))$$

and we go to (3.6). We see that $e^{\Gamma_1} - 1 = \Lambda_1 \neq 0$. Thus, $\Gamma_1 \neq 0$ and, with a similar argument as the previous one we obtain

$$0 < |\Gamma_1| < \frac{2\alpha^6}{\alpha^{m-m_1}}.$$

Dividing through $\log\alpha$ we get

$$0 < |n_1\tau - m + \mu| < \frac{75}{\alpha^{m-m_1}},$$

where τ is the same one as above and

$$\mu := \frac{\log(\sqrt{5}c_1(\gamma^{n-n_1} - 1))}{\log\alpha}.$$

We note that $n_1 > 0$, since otherwise we would have $n \leqslant 476$ which contradicts $n > 540$. Thus, we can apply Lemma 2.2. Consider

$$\mu_k := \frac{\log(\sqrt{5}c_1(\gamma^k - 1))}{\log\alpha}, \quad k = 1, 2, \ldots, 476.$$

With the help of *Mathematica* we found that the denominator of the 111-th convergent above of τ is such that $q_{111} > 6M$ and $\varepsilon_k \geqslant 0.00129842 > 0$ for all $k = 1, 2, \ldots, 476$. Thus, by Lemma 2.2 with $A := 75$, $B := \alpha$ we obtain that the maximum value of $\log(q_{111} \cdot 75/\varepsilon_k)/\log\alpha$, $k = 1, 2, \ldots, 476$, is less than 287. Therefore $m - m_1 \leqslant 287$.

In a similar way we study the other case. Assume that $m - m_1 \leqslant 275$ and $n - n_1 \geqslant 20$. In this case we consider

$$\Gamma_2 = n\log\gamma - m_1\log\alpha + \log\left(\frac{\sqrt{5}c_1}{\alpha^{m-m_1} - 1}\right)$$

and we go to (3.7). Observe that $1 - e^{-\Gamma_2} = \Lambda_2 \neq 0$. Hence, $\Gamma_2 \neq 0$ and, with an argument as above we conclude that

$$0 < |\Gamma_2| < \frac{2\gamma^7}{\gamma^{n-n_1}},$$

Dividing through by $\log \alpha$ we get

$$0 < |n\tau - m_1 + \mu| < \frac{30}{\gamma^{n-n_1}}.$$

where τ is as above and

$$\mu := \frac{\log\left(\sqrt{5}c_1/(\alpha^{m-m_1} - 1)\right)}{\log \alpha}.$$

We note that $m_1 > 0$. Indeed, for if not, we get $m \leqslant 275$ which contradicts $m > 311$. Thus, we can apply Lemma 2.2 again. Consider

$$\mu_\ell := \frac{\log\left(\sqrt{5}c_1/(\alpha^\ell - 1)\right)}{\log \alpha}, \quad \ell = 1, \ldots, 275.$$

Again, with *Mathematica* we quickly found that the same 111-th convergent of τ satisfies $q_{111} > 6M$ and $\varepsilon_\ell > 0.000693865 > 0$ for all $\ell = 1, \ldots, 257$. Thus, from Lemma 2.2 with $A := 30$, $B := \gamma$ we obtain that the maximum value of $\log(q_{111} \cdot 30/\epsilon_\ell)/\log \gamma$, $\ell = 1, \ldots, 257$ is $\leqslant 490$. Hence, $n - n_1 \leqslant 490$.

Summarizing what we have done, we first got that either $n - n_1 \leqslant 476$ or $m - m_1 \leqslant 257$. Assuming the first one we obtained that $m - m_1 \leqslant 287$, and assuming the second one we obtained $n - n_1 \leqslant 490$. So, altogether we have that $n - n_1 \leqslant 490$, $m - m_1 \leqslant 287$. It remains to study this case.

Consider

$$\Gamma_3 = n_1 \log \gamma - m_1 \log \alpha + \log\left(\sqrt{5}c_1 \frac{\gamma^{n-n_1} - 1}{\alpha^{m-m_1} - 1}\right),$$

and we go to (3.8). Note that $e^{\Gamma_3} - 1 = \Lambda_3 \neq 0$. Thus, $\Gamma_3 \neq 0$ and since $n > 540$ with an argument as before we get

$$0 < |\Gamma_3| < \frac{2\gamma^{16}}{\gamma^n}.$$

Dividing through by $\log \alpha$ we obtain

$$o < |n_1\tau - m_1 - \mu| < \frac{374}{\gamma^n},$$

where τ is as above and

$$\mu := \frac{\log\left(\sqrt{5}c_1\left(\gamma^{n-n_1} - 1/\alpha^{m-m_1} - 1\right)\right)}{\log \alpha}.$$

As above we note that n_1 and m_1 are positives. We apply Lemma 2.2 again. Consider

$$\mu_{k,l} := \frac{\log\left(\sqrt{5}c_1\left(\gamma^k - 1/\alpha^\ell - 1\right)\right)}{\log \alpha}, \quad k = 1, \ldots, 490 \quad \ell = 1, \ldots, 287.$$

With *Mathematica* we find that the same 111-th convergent above of τ works again. That is, $q_{111} > 6M$ and $\varepsilon_{k,\ell} \geq 5.28933^{-8} > 0$ for all $k = 1, \ldots, 490$ and $\ell = 1, \ldots, 287$. Thus, by Lemma 2.2 with $A := 374$ and $B := \gamma$ we obtain that the maximum value of $\log(q_{111}374/\varepsilon_{k,\ell})/\log\gamma$, $k = 1, \ldots, 490$ and $\ell = 1, \ldots, 287$, is $\leqslant 533$. Thus, $n \leqslant 533$ which contradicts our assumption on n. This completes the proof of Theorem 1.1.

Acknowledgements. We thank the anonymous referee for valuable comments. The second author thanks Juan Manuel Pérez Díaz for helpful advice and kind support. He also thanks Lidia González García for valuable bibliography support. This paper started during a visit of the third author to the Universidad Autónoma de Zacatecas, in February 2018. He thanks this Institution for their hospitality.

References

[1] A. Baker, H. Davenport: *The equations* $3X^2 - 2 = Y^2$ *and* $8X^2 - 7 = Z^2$, Quart. J. Math. Oxford 20.2 (1969), pp. 129–137, doi: 10.1093/20.1.129.

[2] J. J. Bravo, C. A. Gómez, F. Luca: *Powers of two as sums of two k-Fibonacci numbers*, Miskolc Math. Notes 17.1 (2016), pp. 85–100, doi: 10.18514/MMN.2016.1505.

[3] J. J. Bravo, F. Luca, K. Yazán: *On Pillai's problem with Tribonacci numbers and Powers of 2*, Bull. Korean Math. Soc 54.3 (2017), pp. 1069–11080, doi: 10.4134/BKMS.b160486.

[4] Y. Bugeaud, M. Mignotte, S. Siksek: *Classical and modular approaches to exponential diophantine equations I: Fibonacci and Lucas perfect powers*, Ann. of Math. 163 (2006), pp. 269–1018, doi: 10.4007/annals.2006.163.969.

[5] K. C. Chim, I. Pink, V. Ziegler: *On a variant of Pillai's problem II*, J. Number Theory 183 (2018), pp. 269–290, doi: 10.1016/j.jnt.2017.07.016.

[6] K. Chim, I. Pink, V. Ziegler: *On a variant of Pillai's problem*, Int. J. Number Theory 7 (2017), pp. 1711–1727, doi: 10.1142/S1793042117500981.

[7] M. Ddamulira, C. A. Gómez, F. Luca: *On a problem of Pillai with k-generalized Fibonacci numbers and powers of 2*, Monatsh. Math. 187.4 (2018), pp. 635–664, doi: 10.1007/s00605-018-1155-1.

[8] M. Ddamulira, F. Luca, M. Rakotomalala: *On a problem of Pillai with Fibonacci and powers of 2*, Proc. Indian Acad. Sci. (Math. Sci.) 127.3 (2017), pp. 411–421, doi: 10.1007/s12044-017-0338-3.

[9] A. Dujella, A. Pethő: *A generalization of a theorem of Baker and Davenport*, Quart. J. Math. Oxford 49.3 (1998), pp. 291–306, doi: 10.1093/qmathj/49.3.291.

[10] S. H. Hernández, F. Luca, L. M. Rivera: *On Pillai's problem with the Fibonacci and Pell sequences*, Bol. Soc. Mat. Mex. (2018), doi: 10.1007/s40590-018-0223-9.

[11] A. Herschfeld: *The equation* $2^x - 3^y = d$, Bull. Amer. Math. Soc. 42 (1936), pp. 231–234, doi: 10.1090/S0002-9904-1936-06275-0.

[12] A. C. G. Lomelí, S. H. Hernández: *Pillai's problem with Padovan numbers and powers of two*, Revista Colombiana de Matemáticas 53.1 (2019), pp. 1–14, doi: 10.15446/recolma.v53n1.81034.

[13] A. C. G. Lomelí, S. H. Hernández, F. Luca: *Pillai's problem with the Padovan and tribonacci sequences*, Indian Journal of Mathematics 61.1 (2019), pp. 61–75.

[14] E. M. Matveev: *An explicit lower bound for a homogeneous rational linear form in the logarithms of algebraic numbers II*, Izv. Math. 64.6 (2000), pp. 1217–1269, doi: 10.1070/IM2000v064n06ABEH000314.

[15] S. S. Pillai: *On* $a^x - b^y = c$, J. Indian Math. Soc. 2 (1936), pp. 119–122.

[16] S. S. Pillai: *On the equation* $2^x - 3^y = 2^X + 3^Y$, Bull. Calcutta Math. Soc. 37 (1945), pp. 15–20.

[17] S. G. Sánchez, F. Luca: *Linear combinations of factorials and S-units in a binary recurrence sequence*, Ann. Math. Québec 38 (2014), pp. 169–188, doi: 10.1007/s40316-014-0025-z.

[18] N. J. A. Sloane: *The On-Line Encyclopedia of Integer Sequences*, https://oeis.org/, 1964.

[19] I. Stewart: *Mathematical Recreations: Tales of a neglected number*, Scientific American 274 (1996), pp. 92–93.

[20] R. J. Stroeker, R. Tijdeman: *Diophantine equations*, in: Computational methods in number theory, Part II, Math. Centrum, Amsterdam: Math. Centre Tracts, 1982, pp. 321–369.

[21] B. M. M. D. Weger: *Padua and Pisa are exponentially far apart*, Publ. Matemàtiques 41.2 (1997), pp. 631–651, doi: 10.5565/PUBLMAT_41297_23.

Annales Mathematicae et Informaticae
***50** (2019) pp. 117–130*
DOI: 10.33039/ami.2019.02.002
http://ami.uni-eszterhazy.hu

Volumetric flow rate reconstruction in great vessels*

Attila Lovas[a†], Róbert Nagy[b], Péter Sótonyi[c], Brigitta Szilágyi[d]

[a]Budapest University of Technology and Economics
Department of Analysis,
attila.lovas@gmail.com

[b]Budapest University of Technology and Economics
Department of Structural Mechanics
nagy.robert@epito.bme.hu

[c]Semmelweis University, Department of Vascular Surgery
sotonyi.peter@varosmajor.sote.hu

[d]Budapest University of Technology and Economics
Department of Geometry
szilagyibr@gmail.com

Submitted: November 12, 2018
Accepted: February 11, 2019
Published online: February 27, 2019

Abstract

We present a new algorithm to reconstruct the volumetric flux in the aorta. We study a simple 1D blood flow model without viscosity term and sophisticated material model. Using the continuity law, we could reduce the original inverse problem related to a system of PDEs to a parameter identification problem involving a Riccati-type ODE with periodic coefficients. We implemented a block-based optimization algorithm to recover the model

*This research was supported by the Higher Education Excellence Program of the Ministry of Human Capacities in the frame of Biotechnology research area of Budapest University of Technology and Economics (BME FIKP-BIO).

†The first author was also supperted by the Lendület grant LP 2015-6 of the Hungarian Academy of Sciences.

parameters. We tested our method on real data obtained using CG-gated CT angiography imaging of the aorta. Local flow rate was calculated in 10 cm long aorta segments which are located 1 cm below the heart. The reconstructed volumetric flux shows a realistic wave-like behavior, where reflections from arteria iliaca can also be observed. Our approach is suitable for estimating the main characteristics of pulsatile flow in the aorta and thereby contributing to a more accurate description of several cardiovascular lesions.

Keywords: Haemodynamics, pulse wave propagation, one-dimensional modeling, periodic Ricatti equation

MSC: 92C50, 92C10, 92C35, 76B99

1. Introduction

Pulsatile flow in blood vessels has been studied for more than 300 years. Euler initiated the theory of pressure wave propagation in the vascular system in 1775 [6]. The first modern mathematical model of pulsatile flow in blood vessels was developed by Korteweg and Lamb [10, 12]. It is widely accepted that, flow waveform carries valuable information about the physical properties of the circulatory system [1]. Moreover, it allows to calculate patient-specific estimates of haemodinamical quantities like blood pressure in the aorta that are difficult to measure non-invasively. Consequently, haemodynamical simulations have become increasingly popular in the last few decades. Blood flow modeling techniques can be divided into three main types: 0D or lumped parameter models, 1D and 3D models. Each of these has its own advantages and limitations. For example, 0D models are computationally inexpensive, but they are not suitable to study pulse wave propagation phenomena or complex flows [2]. Similarly, 3D are capable of representing complex velocity profiles. However, the main drawback of 3D simulations is their huge computational cost. The accuracy of these modeling frameworks have been evaluated against in vivo data [2]. It turned out that average relative errors are smaller than 7% between simulated and in vivo waveforms. The survey paper [1] gives a good overview with plenty information about numerical, theoretical and experimental efforts and recent developments made in this field.

The main objective of the present work is to demonstrate that the volumetric flux rate in the aorta can be reconstructed from the changes of the sectional area. Dumas demonstrated the possibility of the determination of volumetric flux by fitting a 1D model with results of 3D computations or with experimental values [4]. In this work we offer a new method independent from 3D simulations and thus keeping the computational cost down. State-of-the-art investigations of the disorders of human arterial sections apply 3D numerical fluid–structure interaction simulations involving the calculation of the blood flow field inside the lumen, the necessary boundary conditions of which are the time-dependent pressure profile at the outlet and volumetric flow rate at the inlet cross-section. The protocol to determine these functions non-invasively is the measurement of both on the arm followed by transforming them to the section under scrutiny by a 1D system

model of the circulatory system. Our method offers a different and easy to use procedure to formulate the inlet boundary condition without Doppler velocimetry, thus facilitating 3D simulations.

The paper is organized as follows. Section 2 deals with data acquisition technologies. Section 3 is divided into three main parts. In the first part, the 1D pulsatile blood flow model used in this study is introduced. The second part contains some analytical considerations about the existence and the stability of periodic solutions with prescribed bounds on the average volumetric flux. Furthermore, we show that the periodic solutions can be obtained by solving a first order Riccati type ODE. In the last part of this section, we present a new algorithm to solve the corresponding inverse problem. Section 4 is devoted to the presentation and discussions of numerical results.

2. Material and methods

In this Section, we present the techniques used in this study to acquire in vivo data. First, we give a brief overview of the ECG-gated computed tomography angiography. After this, we describe the details of the examinations and measurements. At least we specify the entire chain of data post-processing including the segmentation of 4D CT data and the strategies implemented to mitigate the impact of measurement errors.

2.1. Imaging of the aorta

The imaging of a pulsatile organ is a highly demanding application for any cross-sectional imaging modality. Computed tomography (CT) imaging of the heart became widely available with the introduction of multi-detector CT (MDCT) scanners with four-slice detector arrays and 500 ms minimum rotation time [3]. Still images of the moving heart was generated using retrospective ECG-gating: slow table motion during spiral scanning and simultaneous acquisition of the slices and the digital ECG trace provided oversampling of scan projections [3]. After the exposure, slices recorded in the same phase of the ECG trace are matched to generate a 3D dataset of the volume of interest, representing either systole or diastole. A drawback of this method is the higher radiation dose compared to the normal non-oversampled spiral acquisition [5]. An important advantage is the possibility to reconstruct multiphase datasets of the same volume, resulting in motion images of the same slices. This allows us the functional analysis of the moving organs such as the heart or the great vessels. Recent advances in CT technology (256–320 detector rows, 270 ms minimum rotation time) allow for rapid ECG-gated CTA of the whole aorta during a single breath-hold.

Imaging of the aorta was performed in 5 patients (5 men, mean age 68.2 ± 6.1 years, see Table 1) with a 256-slice MDCT (Philips Brilliance iCT, Koninklijke Philips N.V., Best, The Netherlands) using a retrospectively ECG-gated protocol tailored for the imaging of the aorta. Investigations were performed on images read-

ily available from patients with suspected aortic disease. Low-dose (tube voltage: 100 kV) native scan was followed by a retrospective ECG-gated CT angiography of the whole aorta (100 kV) with a reduced field of view to maximize spatial resolution. Nonionic contrast agent was injected into an antecubital vein at a flow rate of 4-5 ml/s using a power injector. Images were reconstructed using a sharp convolution kernel and iterative reconstruction algorithm (iDose4, Koninklijke Philips N.V., Best, The Netherlands) with a slice thickness of 1 mm and an increment of 1 mm. Multiphase images were reconstructed corresponding every 10% of the R-R cycle resulting in ten series of images for each patient. Patients gave written informed consent before the CT examination was performed. Experimental protocol and informed consent was approved by the Regional Ethical Committee of Semmelweis University (133/2011).

No.	Age	BMI	HR	DLP
1.	69	29.1	60	3504
2.	67	23.9	52	2690
3.	62	28.9	86	2756
4.	65	19.3	59	2161
5.	78	24.7	67	2300

Table 1: Patient data: Age (years), BMI (kg/m^2), HR (bpm), DLP (mGycm)

2.2. Data post-processing

In order to measure the section area at any longitudinal position and time, we have to locate the arterial lumen on a huge amount of bitmap images. This can be carried out by some sort of image segmentation algorithm. We wrote a program that implements a version of the Actice Contour Model (ACM) to perform this task. For more information about the Active Contours, we refer the reader to [9] and [14]. As the result of the segmentation process, we get the coordinates of the internal vessel wall for each slice perpendicular to z axis in the region of interest and for all phases in time.

To mitigate spatial measurement error, we fitted a bi-cubic smoothing spline to the point cloud resulting from the active contour method in each time step. This way the increase of temporal resolution also became possible exploiting the affine covariance of B–splines and ensuring the periodic movement of the control points by trigonometric approximation using the first 3 harmonics of the heart rate [13]. Generated meshes corresponding to the first phase are presented in Figure 1. The cross-sectional area is defined as the area of the section perpendicular to the center line of the fitted surface.

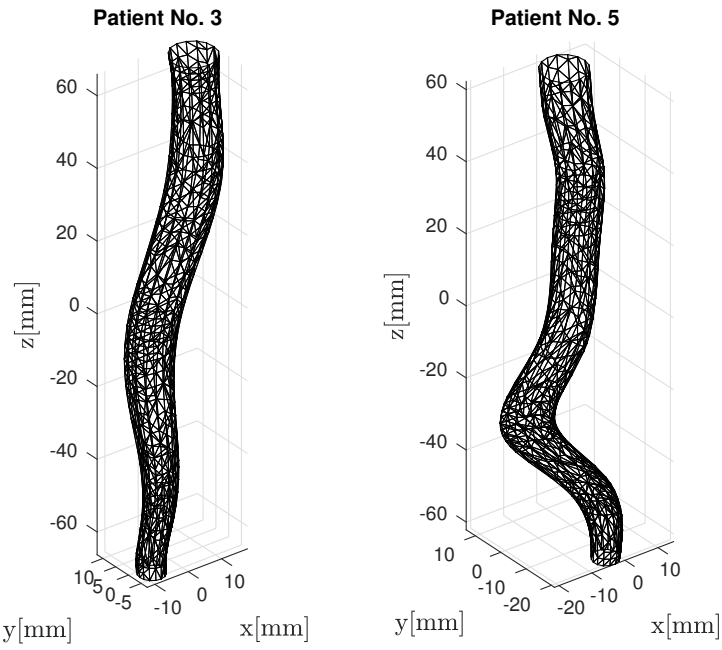


Figure 1: 3D mesh models of a non-branching vessel segments in the thoracic aorta

3. Presentation and resolution of the inverse problem

In this Section, we introduce the 1D blood flow model used in this study. We define the class of physically admissible solutions. We show that the volumetric flux can be calculated by solving a Riccati type ODE. We demonstrate that a physically admissible solution is not necessarily asymptotically stable. We will see that vessel wall motions do not provide enough information about the wall elasticity to recover the volumetric flux. We present a block-based optimization algorithm to resolve this problem.

3.1. Governing equations

To make our exposition self-contained and understandable for the largest possible audience, we present some laws of continuum mechanics: the conservation laws for mass and momentum. In addition, empirical constitutive laws are needed to relate certain unknown variables such as relations between stress and strain. Although there are one-dimensional models which also take into account fluid viscosity and wall viscoelasticity [15], we examine a simpler model using Euler's equation for the non-viscose case. However, our approach can be generalized to more sophisticated

models involving damping effects. We assume that the vessel wall is thin and elastic. Owing to the pressure gradient the artery wall deforms and the elastic restoring force of the wall makes it possible for waves to propagate and so it maintains a pulsating motion of the artery.

Now, we consider a non-branching cylindric vessel segment of length L. The section area $S(t,z)$ and averaged flow velocity $u(t,z)$ vary in time along the artery $z \in [0,L]$. We define the volumetric flux as $q = uS$. Assuming that blood is homogeneous and incompressible, we obtain from the law for conversation of mass that

$$\partial_t S + \partial_z q = 0.$$

Additionally, the law for conservation of momentum has the following form

$$\partial_t u + u\partial_z u = -\frac{\partial_z p}{\rho}, \tag{3.1}$$

where ρ denotes the blood density and $p(t,z)$ is the local blood pressure. We use the Hook's law to relate stress and strain rates. Let h be the vessel wall thickness, assumed to be much smaller than the vessel radius, and Young's modulus will be denoted by E. The change in tube radius must be caused by the blood pressure. The elastic strain due to the lengthening of the circumference is

$$\frac{r(t,z) - r_0(z)}{r_0(z)},$$

where $r_0(z)$ stands for the equilibrium radius. The change is elastic force must be balanced by the changing in pressure force $2r(t,z)p(t,z)$, hence the desired relation between pressure and radius has the following form.

$$p(t,z) = \left(\frac{1}{r_0(z)} - \frac{1}{r(t,z)}\right) hE$$

For the right-hand side in (3.1), we obtain

$$R_\alpha(t,z) \overset{\text{def.}}{=} -\frac{1}{\rho}\partial_z p = \alpha\left(\frac{S_0'(z)}{S_0(z)^{3/2}} - \frac{\partial_z S}{S(t,z)^{3/2}}\right)$$

where $\alpha = 0.5hE\pi^{1/2}\rho^{-1}$ is called wall compliance parameter and it is assumed to be constant during the cardiac cycle. By the conservation law for mass, the momentum transport equation can be expressed by means of the volumetric flux as follows:

$$\partial_t q + \partial_z\left(q^2/S\right) = S(t,z)R_\alpha(t,z). \tag{3.2}$$

3.2. Physically admissible solutions

Assuming that S, S_0 and α are known, we are looking for special solutions of (3.2). We will assume in the sequel that blood flows from $z = 0$ to $z = L$.

Definition 3.1. A solution q of (3.2) is said to be physiologically admissible if it is periodic with period T and

$$q_{\min} < \frac{1}{T}\int_0^T q(t,z)\,\mathrm{d}z < q_{\max}$$

holds for $z \in [0, L]$, where $q_{\min}$ and $q_{\max}$ are the minimal and maximal average flow rate which may occur under physiological conditions.

Using the continuity law again, we can express the volumetric flux as

$$q(t,z) = q(t,0) + \Phi(t,z),$$

where $Q(t) = q(t,0)$ and $\Phi(t,z) = -\int_0^z \partial_t S(t,y)\,\mathrm{d}y$. Note that the periodicity of S implies that q is physiologically admissible if and only if Q satisfies itself the conditions of admissibility.

After substituting the expression we have recently got for q into (3.2) and integrating both sides from 0 to L, we get a Riccati type ODE with periodic coefficients

$$\dot{Q} = AQ^2 + BQ + C, \tag{3.3}$$

where dot denotes the time derivative and for the coefficients we have

$$\begin{aligned}
A(t) &= -\frac{1}{L}\left[\frac{1}{S(t,L)} - \frac{1}{S(t,0)}\right] \\
B(t) &= -\frac{2}{L}\frac{\Phi(t,L)}{S(t,L)} \\
C(t) &= -\frac{1}{L}\left[\frac{\Phi^2(t,L)}{S(t,L)} + \int_0^L S(t,z) + R_\alpha(t,z)\,\mathrm{d}z\right]
\end{aligned}$$

It is a well known fact that the general solution of a scalar Riccati equation can be obtained by quadrature whenever at least one particular solution Q_0 is known. After substituting $Q = Q_0 - 1/W$ into the original equation, we get a linear ODE for W:

$$\dot{W} = -(2Q_0A + B)W + A$$

which general solution can be written as

$$W(t) = KW_1(t) + W_2(t),$$

where K is an arbitrary constant, $W_1(0) = 1$ and $W_2(0) = 0$. Periodicity of Q requires that $Q(0) = Q(T)$ which holds if and only if K solves the quadratic equation:

$$Q_0(T)K^2 + \left(Q_0(T)\frac{W_2(T)}{W_1(T)} - \frac{W_1(T)-1}{W_1(T)}\right)K - \frac{W_2(T)}{W_1(T)} = 0 \tag{3.4}$$

where we assumed that $Q_0(0) = 0$. We can conclude that the original Riccati equation has 0, 1 or 2 periodic solutions depending on the discriminant of equation (3.4). However, we need to know $W_1(T)$, $W_2(T)$ and $Q_0(T)$ in order to calculate the discriminant and that can be done just for a given case.

Leon Kotin demonstrated the existence and uniqueness of a positive and of a negative periodic solution of a periodic Riccati equation in which the coefficients satisfy certain general condition. Moreover, any solution which is everywhere continuous lies between these two solutions, and every solution is asymptotic to one of these as the independent variable increases or decreases [11]. More precisely, the following is true.

Theorem 3.2. *If coefficients in* (3.3) *are continuous everywhere, A is continuously differentiable and* $AC < 0$ *holds, then equation* (3.3) *has a unique positive solution* Q_+ *and a unique negative solution* Q_- *which are periodic with period* T *and any solution behaves asymptotically like* Q_+ *or* Q_- *as* $t \to \infty$.

It is clear that if conditions of Kotin's theorem are satisfied, then Q_+ can be the unique physically admissible solution. However, the average volumetric flux calculated from Q_+ may not fall into the acceptance interval. Moreover, nothing guarantees that Q_+ is asymptotically stable. For example, consider the case when $C < 0$ and Q is an arbitrary perturbation of Q_+. From the uniqueness of the positive solution, we can conclude that there exists $t_0 \in \mathbb{R}$ where Q vanishes and Q can have only one root, since at $Q = 0$ the right-hand side of (3.3) is equal to C which is negative. As a consequence, we get that the physically admissible solution is not necessarily asymptotically stable.

3.3. Calculation of model parameters

For any fixed α, we can calculate W_1, W_2 and Q_0 numerically thus for the periodic solutions of (3.3)

$$Q_{1,2}(t) = Q_0(t) - \frac{1}{K_{1,2}W_1(t) + W_2(t)}$$

yields, K_1 where and K_2 are solutions of (3.4). However, we do not have a priori information about the wall compliance parameter hence the problem is undetermined.

In order to reconstruct the volumetric flux from changes of the section area, we consider two adjacent vessel segment of length L that are narrow enough to neglect the longitudinal changes in the artery wall compliance parameter. Now, we follow the notations of Figure 2, where $z_C - z_B = z_B - z_A = L$ and changes in between and are negligible.

If α is fixed, then $q_{\alpha,1}$ denotes the volumetric flow rate calculated from changes of the section area between z_A and z_B while $q_{\alpha,2}$ stands for the volumetric flux obtained from the second vessel segment. For the true α, $q_{\alpha,1}(t, z_B)$ and $q_{\alpha,2}(t, z_B)$ should be equal to each other for $t \in [0, T]$. To measure the goodness of α, we

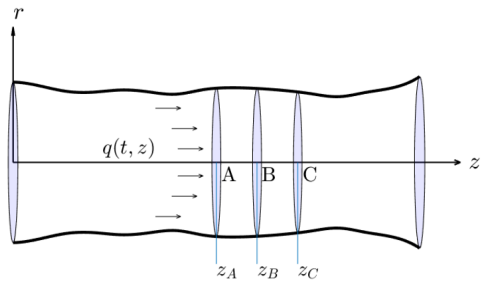


Figure 2: Blood flow in an arterial segment

introduce the so-called internal consistency functional

$$I_\alpha = \int_0^T \left(q_{\alpha,1}(t, z_B) - q_{\alpha,2}(t, z_B)\right)^2 \, \mathrm{d}t$$

that penalizes the difference between volumetric flow rates at z_B obtained for the first and second vessel segment. We also introduce the notation

$$V_\alpha = \frac{1}{2T} \int_0^T q_{\alpha,1}(t, z_B) + q_{\alpha,2}(t, z_B) \, \mathrm{d}t$$

for the average volumetric flux at z_B. Therefore, we can formulate the original problem as a constrained minimization of I_α.

$$\begin{aligned} \min_{\alpha} \quad & I_\alpha \\ \text{subject to} \quad & q_{\min} < V_\alpha < q_{\max}, \end{aligned}$$

where the prescribed constraint ensures the physical admissibility of the solution.

The domain of I_α consists of positive α values for which Riccati equations related to the first and second vessel segment admits periodic solutions i.e. the discriminant of the corresponding quadratic equation is positive. Obviously, V_α depends smoothly on α hence the feasibility set of the optimization problem is an open set in $(0, \infty)$ which is a collection of open intervals. At this point, we do not have any further information about the structure of this set. We just assume during the simulations that it is connected i.e. it is an open interval. We will see in the next section that this assumption is justified by simulation results. So, we first solve non-linear scalar equations

$$V_\alpha = q_{\min}, q_{\max}$$

for α and get $\alpha_{\min}$ and $\alpha_{\max}$. After this, we calculate $\alpha_{\text{opt}} = \underset{\alpha \in (\alpha_{\min}, \alpha_{\max})}{\arg\min} \; I_\alpha$. At least, the volumetric flux for the whole arterial segment can be obtained as

$$q(t, z) = \frac{1}{2} \left(q_{\alpha_{\text{opt},1}}(t, z) + q_{\alpha_{\text{opt},2}}(t, z) \right).$$

In our MATLAB implementation, we used the ode45 function to solve initial value problems, fzero function to solve non-linear equations for values corresponding to the minimal and maximal flow rate. At least, we applied the fminbnd function to find the global minimum of on. More information about these MATLAB solvers is available in the online MATLAB documentation: `https://www.mathworks.com/help/matlab/`.

4. Results

Numerical simulations were performed on a 10 cm long non-branching segment in the descending aorta, where the $z = 0$ level is located 1 cm below the heart. In order to minimize the disturbing effect of moving organs, we chose the last 2×1 cm region which means that we set $z_A = 8\,\text{cm}$, $z_B = 9\,\text{cm}$ and $z_C = 10\,\text{cm}$. According to the medical literature, cardiac output lies between $4\,\text{dm}^3/\text{min}$ and $6\,\text{dm}^3/\text{min}$ hence we set the maximal and minimal average flow rate to $66.7\,\text{cm}^3/\text{s}$ and $100\,\text{cm}^3/\text{s}$, respectively. We defined the equilibrium cross-sectional area as

$$S_0(z) = \min_{t \in [0,T]} S(t, z).$$

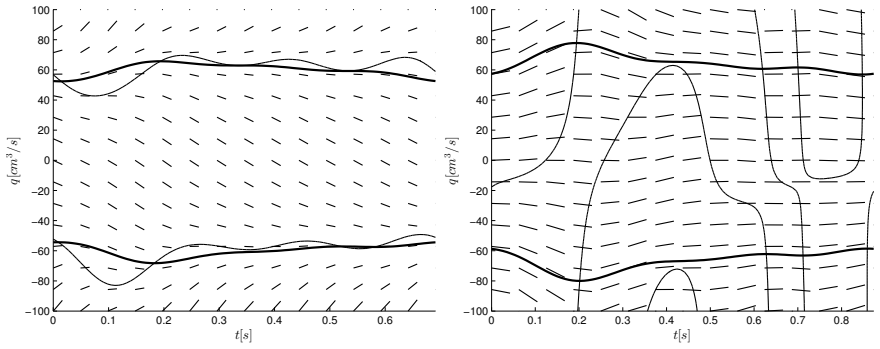
We summarize the simulation results in Table 2, where mean squared error is defined as $\text{MSE} = \left(I_{\alpha_{\text{opt}}}/T\right)^{1/2}$ and it characterizes the goodness of the optimum.

No.	$\alpha_{\min}$	$\alpha_{\max}$	α_{opt}	MSE	Kotin
1.	2.72	7.56	2.75	2.25	+
2.	2.47	6.87	2.47	3.42	+
3.	2.67	7.42	2.68	3.82	+
4.	2.81	7.99	2.88	3.06	+
5.	5.81	15.8	6.82	2.12	-

Table 2: Wall compliance parameter values ($10^3\text{cm}^3/\text{s}^2$) , mean squared error ($\text{cm}^3/\text{s}$) and conditions of Kotin's theorem – satisfied (+) or not (-)

We can see that conditions of Kotin's theorem are satisfied in all cases except for the oldest participant. We present phase portraits of the Riccati equation corresponding to the youngest and oldest participants (Figure 3A and 3B). Phase portrait presented in Figure 3a exhibits the typical behavior. We got similar results for patient No. 1, 2 and 4. In this case, we can realize that the only positive and physiologically admissible solution is unstable which is contrary to intuition. It is not clear at this point if exists any relationship between the qualitative behavior of solutions and arterial wall rigidity. Such connection would be helpful in order to gain information about the condition of the circulatory system and detect vascular diseases.

We calculated the volumetric flux at three different longitudinal position alongside the artery segment: $z_1 = 1\,\text{cm}$, $z_2 = 5\,\text{cm}$ and $z_3 = 9\,\text{cm}$. Simulation results



(a) Patient No. 3 (b) Patient No. 5

Figure 3: Phase portrait of the typical and non-typical behavior

for patient No. 3 are presented in Figure 4. As we are getting further from the beating heart and approaching the aortic bifurcation, initial peak in the volumetric flux slowly disappears and reflections from the arteria iliaca became even more dominant.

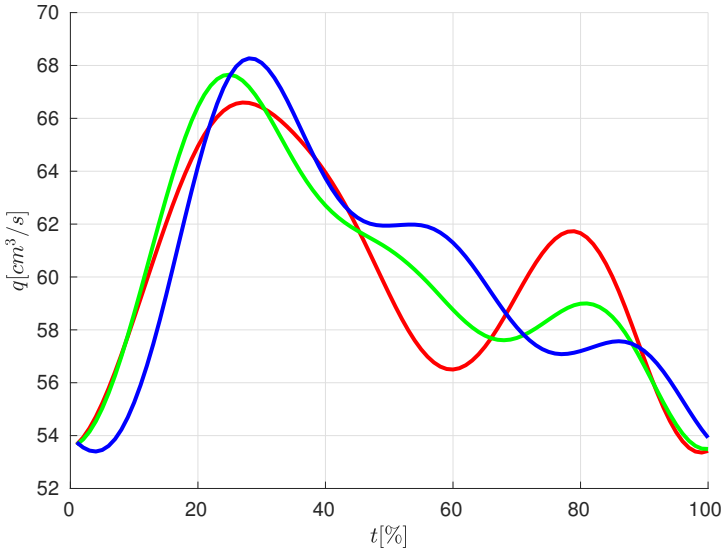


Figure 4: Volumetric flux $\left(\mathrm{cm}^3/\mathrm{s}\right)$ vs. time (% of the cardiac cycle): blue–$z_1 = 1$ cm, green–$z_2 = 5$ cm, red–$z_3 = 9$ cm

The remaining part of this section is devoted to the analysis of the velocity profile inside the aorta by means of Reynolds and Womersley numbers. For the sake of completeness, we give here the definitions of these dimensionless numbers. The Reynolds number is a dimensionless number in fluid mechanics that is defined

as the ratio of inertia forces to viscous forces, expressed in tubular flows as

$$\mathrm{Re} = \frac{2vr\rho}{\eta},$$

where v is the flow velocity, r is the vessel radius, ρ is the blood density and η is the dynamical viscosity of blood. The Womersley number in biofluid mechanics relates the transient inertial forces to viscous effects. It is defined by

$$\mathrm{Wh} = 2r\sqrt{\frac{\omega\rho}{\eta}},$$

where ω is the angular frequency of the oscillations. In accordance with the literature [7, 8], in our calculations blood density was set to $1.06\,\mathrm{g/cm^3}$ and the kinematic viscosity of blood to 3.5×10^{-3} Pas. Estimations for Reynolds and Womersley numbers along the artery segment are illustrated in common coordinate system in Figure 5.

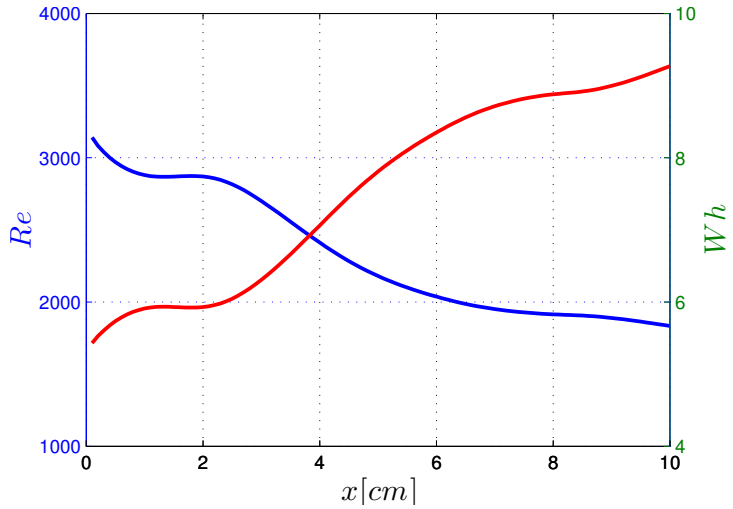


Figure 5: Time average of Reynolds and Womerseley numbers

The Womersley number is a dynamic similarity measure of oscillatory flows relating inertia and viscous forces. In rigid pipes for laminar incompressible flows small values (approx. $\mathrm{Wh} < 1$) allow the development of the parabolic velocity profile of the steady state solution and the flow is almost in phase with the pressure gradient, while large values (approx. $\mathrm{Wh} > 10$) indicate a flat velocity profile with a good approximation and the flow follows the pressure gradient by about 90 degrees in phase. In the presented examples – as seen in Figure 5 – the values lie in between, yielding a complex time-dependent velocity profile. The Reynolds number is another dynamic similarity measure relating inertia and viscous forces. In rigid pipes small values (approx. $\mathrm{Re} < 2100$) indicate laminar flow, while large

values (approx. Re $>$ 4000) correspond to turbulent flow. In the transition zone, where also our example is situated, the behavior strongly depends on the existing disturbances in the flow.

5. Conclusions

The inverse problem for volumetric flow rate reconstruction in large arteries has been successfully solved. We demonstrated that in the majority of cases periodic solutions are unstable even though the changes of the cross-sectional area is supposed to be periodic in time. Our approach makes possible to calculate the aortic flow on a routine ECG-gated CT angiography dataset. This is of huge clinical potential, as the knowledge of haemodynamic parameters could significantly improve the diagnostic performance of CT imaging in several cardiovascular pathologies, such as aortic coarctation or dissection. However, further verifications and comparative studies are needed to validate our method in a clinical cohort.

References

[1] J. Alastruey, K. H. Parker, S. J. Sherwin: *Arterial pulse wave haemodynamics*, in: 11th International Conference on Pressure Surges, Virtual PiE Led t/a BHR Group, 2012, pp. 401–443, isbn: 9781855981331.

[2] J. Alastruey, N. Xiao, H. Fok, T. Schaeffter, C. A. Figueroa: *On the impact of modelling assumptions in multi-scale, subject-specific models of aortic haemodynamics*, Journal of The Royal Society Interface 13.119 (2016), issn: 1742-5689, doi: 10.1098/rsif.2016.0073, eprint: http://rsif.royalsocietypublishing.org/content/13/119/20160073.full.pdf, url: http://rsif.royalsocietypublishing.org/content/13/119/20160073.

[3] C. R. Becker, B. M. Ohnesorge, U. J. Schoepf, M. F. Reiser: *Current development of cardiac imaging with multidetector-row CT*, European Journal of Radiology 36.2 (2000), pp. 97–103, doi: 10.1016/S0720-048X(00)00272-2.

[4] L. Dumas: *Inverse problems for blood flow simulation*, in: Proceedings of EngOpt2008 – International Conference on Engineering Optimization, 2008.

[5] J. P. Earls, E. L. Berman, B. A. Urban, et al.: *Prospectively Gated Transverse Coronary CT Angiography versus Retrospectively Gated Helical Technique: Improved Image Quality and Reduced Radiation Dose*, Radiology 246.3 (2008), pp. 742–753, doi: 10.1148/radiol.2463070989.

[6] L. Euler: *Principia pro motu sanguinis per arterias determinando*, Opera posthuma mathematica et physica anno 1844 detecta 2 (1775), pp. 814–823.

[7] H. Hinghofer-Szalkay, J. E. Greenleaf: *Continuous monitoring of blood volume changes in humans*, Journal of Applied Physiology 63.3 (1987), pp. 1003–1007, doi: 10.1152/jappl.1987.63.3.1003.

[8] J. M. Jung, D. H. Lee, K. T. Kim, et al.: *Reference intervals for whole blood viscosity using the analytical performance-evaluated scanning capillary tube viscometer*, Clinical Biochemistry 47.6 (2014), pp. 489–493, doi: 10.1016/j.clinbiochem.2014.01.021.

[9] M. Kass, A. P. Witkin, D. Terzopoulos: *Snakes: Active contour models*, International Journal of Computer Vision 1.4 (1988), pp. 321–331, doi: 10.1007/BF00133570.

[10] D. J. Korteweg: *Ueber die Fortpflanzungsgeschwindigkeit des Schalles in elastischen Röhren*, Annalen der Physik 241.12 (1878), pp. 525–542, doi: 10.1002/andp.18782411206.

[11] L. Kotin: *On Positive and Periodic Solutions of Riccati Equations*, SIAM Journal of Applied Mathematics 16.6 (1968), pp. 1227–1231, doi: 10.1137/0116103.

[12] H. Lamb: *On the velocity of sound in a tube, as affected by the elasticity of the walls*, Manchester Memoirs 42 (1898), pp. 1–16.

[13] R. Nagy, Cs. Csobay-Novák, A. Lovas, S. Péter, I. Bojtár: *Non-invasive in vivo time-dependent strain measurement method in human abdominal aortic aneurysms: Towards a novel approach to rupture risk estimation*, Journal of biomechanics 48.10 (2015), pp. 1876–1886, doi: 10.1016/j.jbiomech.2015.04.030.

[14] M. Sonka, V. Hlaváč, R. Boyle: *Image Processing, Analysis, and Machine Vision*, ISBN-10: 1133593607, Cengage Learning, 2008, doi: 10.1007/978-1-4899-3216-7.

[15] X. Wang, S. Nishi, M. Matsukawa, et al.: *Fluid friction and wall viscosity of the 1D blood flow model*, Journal of biomechanics 49 (4 2016), pp. 565–571, doi: 10.1016/j.jbiomech.2016.01.010.

Annales Mathematicae et Informaticae
***50** (2019) pp. 131–144*
DOI: 10.33039/ami.2018.12.002
http://ami.uni-eszterhazy.hu

On the X-coordinates of Pell equations which are rep-digits, II

Florian Luca[a], Sossa Victorin Togan[b], Alain Togbé[c]

[a]School of Mathematics, University of the Witwatersrand, South Africa and Department of Mathematics, Faculty of Sciences, University of Ostrava, Czech Republic
florian.luca@wits.ac.za

[b]Institut de Mathématiques et de Sciences Physiques, Porto-Novo, Bénin
tofils74@yahoo.fr

[c]Department of Mathematics, Statistics and Computer Science
Purdue University Northwest,Westville,USA
atogbe@pnw.edu

Submitted: June 27, 2018
Accepted: December 12, 2018
Published online: February 11, 2019

Abstract

For a positive integer d which is not a square, we show that there is at most one value of the positive integer X participating in the Pell equation $X^2 - dY^2 = \pm 4$ which is a rep-digit, that is all its base 10 digits are equal, except for $d = 2, 5, 13$.

Keywords: Pell equation, Rep-digit, Linear forms in complex logarithms.

MSC: 11A25 11B39, 11J86

1. Introduction

Let d be a positive integer which is not a perfect square. It is well-known that the Pell equation

$$X^2 - dY^2 = \pm 4 \tag{1.1}$$

has infinitely many positive integer solutions (X, Y). Furthermore, putting (X_1, Y_1) for the smallest such solution (solution with minimal value for X), all the positive

integer solutions are of the form (X_n, Y_n) for some positive integer n where

$$\frac{X_n + \sqrt{d}Y_n}{2} = \left(\frac{X_1 + \sqrt{d}Y_1}{2}\right)^n.$$

There are many papers in the literature which solve Diophantine equations involving members of the sequences $\{X_n\}_{n\geq 1}$ or $\{Y_n\}_{n\geq 1}$ being squares, or perfect powers of larger exponents of some other integers, etc. (see, for example, [4, 5]).

Let $g \geq 2$ be an integer. A natural number N is called a *base g rep-digit* if all of its base g-digits are equal; that is, if

$$N = a\left(\frac{g^m - 1}{g - 1}\right), \quad \text{for some} \quad m \geq 1 \text{ and } a \in \{1, 2, \ldots, g-1\}.$$

When $g = 10$, we omit the base and simply say that N is a rep-digit. Diophantine equations involving rep-digits were also considered in several papers which found all rep-digits which are perfect powers, or Fibonacci numbers, or generalized Fibonacci numbers, and so on (see [1–3, 7, 9, 11–15, 17] for a sample of such results). In this paper, we study when can X_n be a rep-digit. This reduces to the Diophantine equation

$$X_n = a\left(\frac{10^m - 1}{9}\right), \quad m \geq 1 \text{ and } a \in \{1, \ldots, 9\}. \tag{1.2}$$

Of course, for every positive integer X, there is a unique square-free integer $d \geq 2$ such that

$$X^2 - dY^2 = -4.$$

Namely d is the product of all prime factors of X^2+4 which appear at odd exponents in its factorization. In particular, taking $X = a(10^m - 1)/9$, we get that any rep-digit is the X-coordinate of the Pell equation (1.1) corresponding to some specific square-free integer d. If $X > 2$, we can instead look at $X^2 - 4$ and write it as dY^2 for some positive integers d and Y with d squarefree, and then

$$X^2 - dY^2 = 4.$$

In particular, we can take $X = a(10^m - 1)/9$ with $a \in \{1, \ldots, 9\}$ and $m \geq 1$, where we ask in addition that $a \geq 3$ when $m = 1$. Here, we study the square-free integers d such that the sequence $\{X_n\}_{n\geq 1}$ contains at least two rep-digits. Our result is the following.

Theorem 1.1. *Let $d \geq 2$ be square-free. The Diophantine equation*

$$X_n = a\left(\frac{10^m - 1}{9}\right), \quad m \geq 1 \text{ and } a \in \{1, \ldots, 9\} \tag{1.3}$$

has at most one positive integer solution n except when $d = 2, 5, 13$ for which we have

$$2^2 - 2 \cdot 2^2 = -4, \quad 6^2 - 2 \cdot 4^2 = 4,$$

$$1^2 - 5 \cdot 1^2 = -4, \quad 3^2 - 5 \cdot 1^2 = 4, \quad 4^2 - 5 \cdot 2^2 = -4 \quad 7^2 - 5 \cdot 3^2 = 4, \quad 11^2 - 5 \cdot 5^2 = -4,$$

and

$$3^2 - 13 \cdot 1^2 = -4, \quad 11^2 - 13 \cdot 3^2 = 4.$$

2. Linear forms in logarithms

We need some results from the theory of lower bounds for nonzero linear forms in logarithms of algebraic numbers. We start by recalling Theorem 9.4 of [4], which is a modified version of a result of Matveev [16]. Let $\mathbb{L}$ be an algebraic number field of degree $d_{\mathbb{L}}$. Let $\eta_1, \eta_2, \ldots, \eta_l \in \mathbb{L}$ not 0 or 1 and $d_1, \ldots, d_l$ be nonzero integers. We put

$$D = \max\{|d_1|, \ldots, |d_l|, 3\},$$

and

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

Let $A_1, \ldots, A_l$ be positive integers such that

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

where for an algebraic number η of minimal polynomial

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

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

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

The following consequence of Matveev's theorem is Theorem 9.4 in [4].

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

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

When $l = 2$ and $\eta_1,\ \eta_2$ are positive and multiplicatively independent, we can do better. Namely, let in this case $B_1,\ B_2$ be real numbers larger than 1 such that

$$\log B_i \geq \max\left\{h(\eta_i), \frac{|\log \eta_i|}{d_{\mathbb{L}}}, \frac{1}{d_{\mathbb{L}}}\right\} \quad i = 1, 2,$$

and put

$$b' := \frac{|d_1|}{d_{\mathbb{L}} \log B_2} + \frac{|d_2|}{d_{\mathbb{L}} \log B_1}.$$

Furthermore, let

$$\Lambda = d_1 \log \eta_1 + d_2 \log \eta_2.$$

Note that $\Lambda \neq 0$ when η_1 and η_2 are multiplicatively independent.

Theorem 2.2. *With the above notations, assuming that $\mathbb{L}$ is real, η_1, η_2 are positive and multiplicatively independent, then*

$$\log|\Lambda| > -24.34 d_{\mathbb{L}}^4 \left(\max\left\{\log b' + 0.14, \frac{21}{d_{\mathbb{L}}}, \frac{1}{2}\right\}\right)^2 \log B_1 \log B_2.$$

Note that $e^{\Lambda} - 1 = \Gamma$, so Γ is close to zero if and only if Λ is close to zero, which explains the relation between Theorems 2.1 and 2.2.

3. The Baker-Davenport lemma

Here, we recall the Baker-Davenport reduction method (see [8, Lemma 5a]), which turns out to be useful in order to reduce the bounds arising from applying Theorems 2.1 and 2.2.

Lemma 3.1. *Let $\kappa \neq 0$ and μ be real numbers. Assume that M is a positive integer. Let P/Q be the convergent of the continued fraction expansion of κ such that $Q > 6M$ and put*

$$\xi = \|\mu Q\| - M \cdot \|\kappa Q\|,$$

where $\|\cdot\|$ denotes the distance from the nearest integer. If $\xi > 0$, then there is no solution to the inequality

$$0 < |m\kappa - n + \mu| < AB^{-k}$$

in positive integers m, n and k with

$$\frac{\log(AQ/\xi)}{\log B} \leq k \quad \text{and} \quad m \leq M.$$

4. Bounding the variables

We assume that (X_1, Y_1) is the minimal solution of the Pell equation (1.1). Set

$$X_1^2 - dY_1^2 =: \pm 4$$

and

$$x_n = \frac{X_n}{2}, \quad y_n = \frac{Y_n}{2} \quad \text{for all} \quad n \geq 1.$$

We have

$$x_n^2 - dy_n^2 =: \varepsilon_n, \quad \varepsilon_n \in \{\pm 1\}.$$

Put

$$\delta := x_1 + \sqrt{x_1^2 - \varepsilon_1} = x_1 + \sqrt{d} y_1, \quad \eta := x_1 - \sqrt{d} y_1 = \varepsilon_1 \delta^{-1}, \quad \text{with} \quad \delta \geq (1+\sqrt{5})/2.$$

Then, we get

$$x_n = \frac{1}{2}(\delta^n + \eta^n),$$

or, equivalently,

$$X_n = \delta^n + \eta^n.$$

We start with some general considerations concerning equation (1.2). From equation (1.2), we have

$$X_n = a\left(\frac{10^m - 1}{9}\right) > a(1 + 10 + \cdots + 10^{m-1}) > 10^{m-1}.$$

We get

$$10^{m-1} \le X_n < 10^m. \tag{4.1}$$

Furthermore,

$$2\delta^n > \delta^n + \eta^n = X_n \ge \delta^n - \delta^{-n} \ge \frac{\delta^n}{2},$$

where the last inequality follows because $n \ge 1$ and $\delta \ge (1+\sqrt{5})/2 > \sqrt{2}$. So,

$$\frac{\delta^n}{2} \le X_n < 2\delta^n \quad \text{holds for all} \quad n \ge 1. \tag{4.2}$$

Using now the equations (4.1) and (4.2), we have

$$10^{m-1} \le X_n < 2\delta^n \quad \text{and} \quad \frac{\delta^n}{2} \le X_n \le 10^m.$$

Hence, we obtain

$$nc_1 \log \delta - c_2 \le m \le nc_1 \log \delta + c_2 + 1, \quad c_1 := 1/\log 10, \quad c_2 := c_1 \log 2. \tag{4.3}$$

From the left-hand side inequality of (4.3), we also deduce that

$$n \log \delta < m \log 10 + \log 2. \tag{4.4}$$

Since $\delta \ge (1+\sqrt{5})/2$, we get that

$$n \le m\frac{\log 10}{\log((1+\sqrt{5})/2)} + \frac{\log 2}{\log((1+\sqrt{5})/2)} < 4.8m + 2.$$

If $m \ge 2$, the last inequality above implies that $n < 6m$. If $m = 1$, then $X_n \le 9$, so $\delta^n \le 18$ by (4.2). Since $\delta \ge (1+\sqrt{5})/2$, we get that $n \le 6$, so the inequality $n \le 6m$ holds also when $m = 1$. We record this as

$$n \le 6m. \tag{4.5}$$

Next, using (1.3), we get

$$\delta^n + \eta^n = a\left(\frac{10^m - 1}{9}\right).$$

Put $b := a/9$. We have

$$\delta^n b^{-1} 10^{-m} - 1 = -b^{-1} 10^{-m} \eta^n - 10^{-m}.$$

Thus,

$$\begin{aligned} \left|\delta^n b^{-1} 10^{-m} - 1\right| &\leq \frac{1}{b 10^m \delta^n} + \frac{1}{10^m} = \frac{1}{10^m}\left(1 + \frac{9}{a\delta^n}\right) \\ &< \frac{6}{10^m}, \end{aligned}$$

using that $a \geq 1,\ n \geq 1$ and $\delta \geq (1+\sqrt{5})/2$. Thus,

$$\left|\delta^n b^{-1} 10^{-m} - 1\right| < \frac{6}{10^m}. \tag{4.6}$$

We now assume that $m \geq 2$ and search for an upper bound on it. Since $m \geq 2$, it follows that the right-hand side in (4.6) above is $< 1/2$. Put

$$\Lambda := n \log \delta - \log b - m \log 10.$$

Since $|e^{\Lambda} - 1| < 1/2$, it follows that

$$|\Lambda| < 2|e^{\Lambda} - 1| < \frac{12}{10^m}.$$

Let us return to (4.6) and put

$$\Gamma := e^{\Lambda} - 1 = \delta^n b^{-1} 10^{-m} - 1.$$

Note that Γ is nonzero. Indeed, if it were zero, then $\delta^n = b 10^m$. Hence, $\delta^n \in \mathbb{Q}$. Since δ is an algebraic integer and $n \geq 1$, it follows that $\delta^n \in \mathbb{Z}$. Since δ is a unit, we get that $\delta^n = 1$, so $n = 0$, which is a contradiction. Thus, $\Gamma \neq 0$. We apply Matveev's theorem. If $a \neq 9$ (so, $b \neq 1$), we then take

$$l = 3,\ \eta_1 = \delta,\ \eta_2 = b,\ \eta_3 = 10,\ d_1 = n,\ d_2 = -1,\ d_3 = -m,\ D = \max\{n, m\}.$$

Clearly, $\mathbb{L} = \mathbb{Q}[\sqrt{d}]$ contains all the numbers $\eta_1,\ \eta_2,\ \eta_3$ and has degree $d_{\mathbb{L}} = 2$. We have

$$h(\eta_1) = (1/2)\log \delta, \quad h(\eta_2) \leq \log 9 \quad \text{and} \quad h(\eta_3) = \log 10.$$

Thus, we can take

$$A_1 = \log \delta, \quad A_2 = 2 \log 9 \quad \text{and} \quad A_3 = 2 \log 10.$$

Now, Theorem 2.1 tells us that

$$\log |\Gamma| > -1.4 \times 30^6 \times 3^{4.5} \times 2^2 (1 + \log 2)(1 + \log D)(\log \delta)(2 \log 9)(2 \log 10).$$

Comparing the above inequality with (4.6), we get

$$m\log 10 - \log 6 < 1.4 \times 30^6 \times 3^{4.5} \times 2^4(1+\log 2)(1+\log D)(\log\delta)(\log 9)(\log 10).$$

Thus,

$$m < 1.4 \times 30^6 \times 3^{4.5} \times 2^4 \times (\log 9)(1+\log 2) \times (\log\delta)\cdot(1+\log D)$$

or

$$m < 8.6 \cdot 10^{12}(1+\log D)\log\delta.$$

Since $D \le 6m$ (see (4.5)), we get

$$m < 8.6 \cdot 10^{12}(1+\log(6m))\log\delta. \tag{4.7}$$

This was when $b \neq 1$. In case $b = 1$, we take $l = 2$ and apply the same inequality (except that now $\eta_2 := 1$ is no longer present) getting a better result. Finally, this was under the assumption that $m \ge 2$ but if $m = 1$ then inequality (4.7) also holds. Let us record what we have proved so far.

Lemma 4.1. *Denoting by* $\delta := x_1 + \sqrt{d}y_1$, *all positive integer solutions* (m,n) *of equation* (1.2) *satisfy*

$$m < 8.6 \cdot 10^{12}(1+\log(6m))\log\delta.$$

All this is for the equation $X_n = a(10^m - 1)/9$. Now we assume that

$$X_{n_1} = a_1\left(\frac{10^{m_1}-1}{9}\right) \quad \text{and} \quad X_{n_2} = a_2\left(\frac{10^{m_2}-1}{9}\right).$$

where $a_1,\ a_2 \in \{1,\dots,9\}$.

To fix ideas, we assume that $n_1 < n_2$, so $m_1 \le m_2$. We put as before $b_i := a_i/9$ for $i = 1, 2$. From the above analysis, assuming that $m_1 \ge 2$, we have that

$$|n_i\log\delta - \log b_i - m_i\log 10| < \frac{12}{10^{m_i}} \quad \text{holds for} \quad i \in \{1,2\}. \tag{4.8}$$

The argument proceeds in two steps according to whether $b_1b_2 < 1$ or $b_1b_2 = 1$.

Suppose now that $b_1b_2 < 1$**.**

We multiply the equation (4.8) for $i = 1$ with n_2 and the one for $i = 2$ with n_1, subtract them and apply the absolute value inequality to get

$$\begin{aligned}
&|n_2\log b_1 - n_1\log b_2 + (n_2m_1 - n_1m_2)\log 10| \\
&= |n_1(n_2\log\delta - \log b_2 - m_2\log 10) - n_2(n_1\log\delta - \log b_1 - m_1\log 10)| \\
&\le n_1|n_2\log\delta - \log b_1 - m_2\log 10| + n_2|n_1\log\delta - \log b_1 - m_1\log 10| \\
&\le \frac{12n_1}{10^{m_2}} + \frac{12n_2}{10^{m_1}} \le \frac{24n_2}{10^{m_1}}.
\end{aligned} \tag{4.9}$$

If the right-hand side above is at least $1/2$, we then get

$$10^{m_1} \leq 48n_2 < 300m_2,$$

giving

$$m_1 < c_1 \log(300m_2). \tag{4.10}$$

Assume now that the right-hand side in (4.9) is smaller than $1/2$. Putting,

$$\Lambda_0 := n_2 \log b_1 - n_1 \log b_2 + (n_2 m_1 - n_1 m_2) \log 10,$$

we get $|\Lambda_0| < 1/2$. Putting

$$\Gamma_0 := b_1^{n_2} b_2^{-n_1} 10^{n_2 m_1 - n_1 m_2} - 1,$$

we get that

$$|\Gamma_0| = |e^{\Lambda_0} - 1| < 2|\Lambda_0| < \frac{48n_2}{10^{m_1}}, \tag{4.11}$$

where the middle inequality above follows from the fact that $|\Lambda_0| < 1/2$. We apply Matveev's theorem to estimate a lower bound on Γ_0. But first, let us see that it is nonzero. Assuming $\Gamma_0 = 0$, we get

$$b_1^{n_2} b_2^{-n_1} = 10^{n_2 m_1 - n_1 m_2}. \tag{4.12}$$

Assume first that $n_2 m_1 - n_1 m_2 = 0$. Then $b_1^{n_2} = b_2^{n_1}$. Thus, b_1 and b_2 are multiplicatively independent and they belong to the set

$$\left\{ \frac{1}{9}, \frac{2}{9}, \frac{1}{3}, \frac{4}{9}, \frac{5}{9}, \frac{2}{3}, \frac{7}{9}, \frac{8}{9}, 1 \right\}.$$

They are not both 1 and n_1 and n_2 are both positive. So, the only possibilities are that $b_1 = b_2$, or

$$\{b_1, b_2\} = \left\{ \frac{1}{9}, \frac{1}{3} \right\}, \left\{ \frac{2}{3}, \frac{4}{9} \right\}. \tag{4.13}$$

If $b_1 = b_2$, then $b_1^{n_1} = b_2^{n_2}$ implies $n_1 = n_2$, which together with $n_2 m_1 = n_1 m_2$ leads to $m_1 = m_2$. Thus, $(n_1, m_1) = (n_2, m_2)$ and $a_1 = a_2$ (because $b_1 = b_2$), and this is not convenient for us. If $\{b_1, b_2\}$ is one of the two sets from (4.13), then one of b_1, b_2 is the square of the other one. Thus, since $b_1^{n_1} = b_2^{n_2}$ and $n_2 > n_1$, we get $n_2 = 2n_1$. Since also $n_2 m_1 = n_1 m_2$, we have $m_2 = 2m_1$. Hence, also $b_2 = b_1^2$ and $b_1 \in \{1/3, 2/3\}$. So, we get the pair of equations

$$X_{n_1} = b_1 10^{m_1} - b_1 \quad \text{and} \quad X_{2n_1} = b_1^2 10^{2m_1} - b_1^2.$$

Since in fact

$$X_{2n} = \delta^{2n} + \eta^{2n} = (\delta^n + \eta^n)^2 - 2(\delta\eta)^n = X_n^2 \pm 2,$$

we get that

$$b_1^2 10^{2m_1} - b_1^2 = X_{2n_1} = X_{n_1}^2 \pm 2 = (b_1 10^{m_1} - b_1)^2 \pm 2 = b_1^2 10^{2m_1} - 2b_1^2 10^{m_1} + b_1^2 \pm 2,$$

which leads to

$$2b_1^2 10^{m_1} = 2b_1^2 \pm 2,$$

so

$$10^{m_1} = 1 \pm b_1^{-2}.$$

The last equation above is impossible for $m_1 \geq 2$. For $m_1 = 1$ we get $10 = 1 \pm b_1^{-2}$, which gives $b_1 = 1/3$. Hence,

$$X_{n_1} = \frac{10-1}{3} = 3, \quad \text{and} \quad X_{2n_1} = \frac{10^2-1}{9} = 11.$$

Since $X_{2n_1} = X_{n_1}^2 \pm 2$, it follows that the sign is $+$, so $X_{n_1}^2 - dY_{n_1}^2 = -4$, giving $dY_{n_1}^2 = 13$, so $d = 13$, $Y_1 = 1$, $n_1 = 1$. These solutions are among the ones mentioned in the statement of the main theorem.

This deals with the case when $n_2 m_1 - n_1 m_2 = 0$. Assume next that $n_2 m_1 - n_1 m_2 \neq 0$. Then in the right-hand side of (4.12), both primes 2 and 5 are involved at a nonzero exponent. Thus, they should be also involved with nonzero exponents in the left-hand side of (4.12). Thus, one of b_1, b_2 is $5/9$ and the other is in $\{2/9,\ 4/9,\ 2/3,\ 8/9\}$. A minute of reflection shows that in all cases the exponents of 2 and 5 in the left-hand side of (4.12) have opposite signs, whereas in the right they have the same sign, and this is impossible.

Thus, $\Gamma_0 \neq 0$. Hence, we are entitled to apply Matveev's theorem in order to find a lower bound on Γ_0. In case $b_1 \neq 1$ and $b_2 \neq 1$, we take

$$l = 3,\ \eta_1 = b_1,\ \eta_2 = b_2,\ \eta_3 = 10,\ d_1 = n_2,\ d_2 = -n_1,\ d_3 = n_2 m_1 - n_1 m_2.$$

Clearly, $\mathbb{L} = \mathbb{Q}$ contains all the numbers η_1, η_2, η_3 and has degree $d_{\mathbb{L}} = 1$. Further, $D = \max\{|d_1|, |d_2|, |d_3|\} \leq n_2 m_2 \leq 6m_2^2$. We have

$$h(\eta_1) \leq \log 9, \quad h(\eta_2) \leq \log 9 \quad \text{and} \quad h(\eta_3) = \log 10.$$

Thus, we can take

$$A_1 = \log 9, \quad A_2 = \log 9, \quad A_3 = \log 10.$$

Now, Theorem 2.1 tells us that

$$\log |\Gamma_0| > -1.4 \times 30^6 \times 3^{4.5}(1 + \log D)(\log 9)^2(\log 10).$$

Combining this with estimate (4.11) and using the fact that $48n_2 < 300m_2$ (see inequality (4.5)) we get

$$m_1 \log 10 \leq \log 300 + \log m_2 + 1.6 \times 10^{12}(1 + \log(6m_2^2)),$$

giving

$$m_1 < 7 \times 10^{11}(1 + \log(6m_2^2)). \tag{4.14}$$

The right-hand side of inequality (4.14) is larger than the right-hand side of inequality (4.10). So, regardless whether $24n_2/10^{m_1}$ is at least $1/2$ or smaller than $1/2$, estimate (4.14) holds. From equation (4.4), we get

$$\log \delta < (m_1 + 1) \log 10 < 1.7 \times 10^{12}(1 + \log(6m_2^2)),$$

which together with Lemma 4.1 gives

$$m_2 < \left(8.6 \times 10^{12}(1 + \log(6m_2))\right)\left(1.7 \times 10^{12}(1 + \log(6m_2^2))\right),$$

so

$$m_2 < 1.5 \times 10^{25}(1 + \log(6m_2))(1 + \log(6m_2^2)).$$

This gives $m_2 < 1.5 \times 10^{29}$. This was if both b_1 and b_2 are different than 1. If one of them is 1, we simply apply Matveev's theorem with $l = 2$ getting an even better bound for m_2.

Suppose now that $b_1 = b_2 = 1$.

We return to (4.11) getting that $8/9 \leq 24n_2/10^{m_1}$, which leads to (4.10), unless $n_1 m_2 = n_2 m_1$. In this last case, we get that $n_2/m_2 = n_1/m_1$. Thus, writing $n_1/m_1 = r/s$ in reduced terms, we get that $(n_1, m_1) = (\ell_1 r, \ell_1 s)$ and that $(n_2, m_2) = (\ell_2 r, \ell_2 s)$ for some positive integers $\ell_1 < \ell_2$. Hence, we have

$$X_{r\ell_1} = 10^{s\ell_1} - 1, \quad X_{r\ell_2} = 10^{s\ell_2} - 1.$$

The greatest common divisor of the right hand sides above is $10^s - 1 \geq 9$. The greatest common divisor of the left-hand sides above is X_r if $\ell_1\ell_2$ is odd and 1 or 2 otherwise. Thus, $\ell_1\ell_2$ must be odd and

$$X_r = 10^s - 1.$$

Consequently,

$$\delta^r - 10^s = -\eta^r - 1 \quad \text{and} \quad \delta^{\ell_2 r} - 10^{\ell_2 s} = -\eta^{\ell_2 r} - 1.$$

From the two equations above we get

$$\delta^{(\ell_2-1)r} + \delta^{(\ell_2-2)}10^s + \cdots + 10^{(\ell_2-1)s} = \frac{-\eta^{\ell_2 r} - 1}{-\eta^r - 1}.$$

The last relation above is impossible since its left-hand side is > 10 and its right hand side is

$$\leq \frac{2}{1 - \frac{2}{1+\sqrt{5}}} < 10,$$

a contradiction.

In conclusion, (4.10) holds, which is stronger than (4.14), and the above arguments imply that $m_2 < 1.5 \times 10^{29}$. Hence, we have the following result.

Lemma 4.2. *The inequality*

$$m_2 < 1.5 \times 10^{29}$$

holds.

Now one needs to apply LLL to the bound

$$|\Lambda_0| < \frac{24n_2}{10^{m_1}} < \frac{24 \times 6 \times 1.5 \times 10^{29}}{10^{m_1}} < \frac{1}{10^{m_1-32}}$$

to get a reasonably small bound on m_1.

- First, we will consider the case $b_1 = b_2 := b$; i.e., $a_1 = a_2 := a$ or

$$\{b_1, b_2\} \in \left\{\frac{1}{9}, \frac{1}{3}\right\}, \ \left\{\frac{2}{3}, \frac{4}{9}\right\}.$$

In

$$\Lambda_0 := n_2 \log b_1 - n_1 \log b_2 + (n_2 m_1 - n_1 m_2) \log 10, \tag{4.15}$$

we set $X := n_1 - n_2$ or $X := 2n_2 - n_1$, and $Y := n_2 m_1 - n_1 m_2$ and divide both sides by $Y \log b$ (with $b = b_1 = b_2 \in \{1/9, 2/9, 3/9, 4/9, 5/9, 6/9, 7/9, 8/9\}$) to get

$$\left| \frac{\log 10}{\log b} - \frac{X}{Y} \right| < \frac{1}{Y(\log(1/b))10^{m_1-32}}. \tag{4.16}$$

We assume that m_1 is so large that the right-hand side in (4.16) is smaller than $1/(2Y^2)$. This certainly holds if

$$10^{m_1-32} > 2(\log(1/b))^{-1} Y. \tag{4.17}$$

Since $|Y| < 1.5 \times 10^{59}$, it follows that the last inequality (4.17) holds provided that $m_1 \geq 92$ in all cases, which we now assume. In this case, X/Y is a convergent of the continued fraction of $\eta := \log 10 / \log b$ and $X < 1.5 \times 10^{59}$. Writing

$a = 1$, $\eta := [-2, 1, 19, 1, 5, 1, 6, 2, 5, 15, 3, \ldots, 7, 2, 121, 1, \ldots, 2, 569, 1, 2, 27, 7, \ldots]$

$a = 2$, $\eta := [-2, 2, 7, 1, 1, 2, 4, 2, 99, \ldots]1, 292, 1, 6, 1, 3, 3, 2, 2, 5, \ldots, 1, 1, 1, 42, \ldots]$

$a = 3$, $\eta := [-3, 1, 9, 2, 2, 1, 13, 1, 7, 18, \ldots, 2, 10, 3, 1, 1, 1, 1, 1, 6, \ldots, 1, 284, 2, \ldots]$

$a = 4$, $\eta := [-3, 6, 4, 2, 1, 1, 1, 1, 45, 89, 1, 6, 1, 9, 1, 2, 625, \ldots, 2, 2, 1, 1716, 1, 1, \ldots]$

$a = 5$, $\eta := [-4, 12, 9, 1, 1, 1, 1, 1, 2, 1, \ldots, 10, 1, 1, 12, 8860, 4, 13, 1, 1, 5, 3, 9, 1, \ldots]$

$a = 6$, $\eta := [-6, 3, 8, 1, 3, 3, 22, 1, 1, 44, \ldots, 1, 1, 38, 1, 5, 1, 857, 1, 3, 1, 3, 1, 2, 1, \ldots]$

$a = 7$, $\eta := [-10, 1, 5, 6, 118, 2, 8, 1, 2, 1, \ldots, 8, 23, 1, 30, 2, 2, 8, 1, 4, 2, 1, 1, 255, \ldots]$

$a = 8$, $\eta := [-20, 2, 4, 1, 1, 3, 2, 7, 1, 2, 1, 9, 2, 6, \ldots, 1, 2, 1332, 1, 12, 1, 5, 1, 1, 2, \ldots]$

for the continued fraction of η and p_k/q_k for the kth convergent, we get that $X/Y = p_j/q_j$ for some $j \leq 122$ in all cases. Furthermore, putting $M := \max\{a_j : 0 \leq j \leq 122\}$, we get $M = 8860$ (for $a = 5$). From the known properties of the continued fractions, we then get that

$$\frac{1}{8862Y^2} = \frac{1}{(M+2)Y^2} \leq \left|\eta - \frac{X}{Y}\right| < \frac{1}{Y(\log b)10^{m_1-32}},$$

giving

$$10^{m_1-32} < 8862(\log b)^{-1}Y < 8862(\log b)^{-1}(1.5 \times 10^{59}),$$

leading to $m_1 \leq 96$.

• We now consider the remaining cases. We transform the linear form (4.15) into one of the following forms:

$$\begin{aligned}\Lambda_1 = \quad & (m_1n_2 - m_2n_1 + \delta_1n_1 + \delta_2n_2)\log 2 + (\lambda_1n_1 + \lambda_2n_2)\log 3 \\ & + (m_1n_2 - m_2n_1 + \mu_1n_1 + \mu_2n_2)\log 5,\end{aligned}$$

$$\Lambda_2 := (\lambda_1n_1 + \lambda_2n_2)\log 3 + (\nu_1n_1 + \nu_2n_2)\log 7 + (m_1n_2 - m_2n_1)\log 10,$$

$$\begin{aligned}\Lambda_3 = \quad & (m_1n_2 - m_2n_1 + \delta_1n_1 + \delta_2n_2)\log 2 + (\lambda_1n_1 + \lambda_2n_2)\log 3 \\ & + (m_1n_2 - m_2n_1 + \mu_1n_1 + \mu_2n_2)\log 5 + (\nu_1n_1 + \nu_2n_2)\log 7,\end{aligned}$$

where $|\delta_i| \leq 3$, $|\lambda_i| \leq 2$, $|\mu_i| \leq 1$, $|\nu_i| \leq 1$, for $i = 1, 2$.

Now, we will estimate lower bounds for Λ_i, $i = 1, 2, 3$ via the LLL algorithm (see Proposition 2.3.20 in [6]). One knows that $\Lambda_i \neq 0$, $i = 1, 2, 3$ by what is done above. We set $X_1 = X_3 := 10^{60}$ as upper bounds for $|m_1n_2 - m_2n_1 + \delta_1n_1 + \delta_2n_2|$, $|m_1n_2 - m_2n_1 + \mu_1n_1 + \mu_2n_2|$ and $X_2 = X_4 := 10^{31}$ as upper bounds for $|\lambda_1n_1 + \lambda_2n_2|$, $|\nu_1n_1 + \nu_2n_2|$. We take $C := (3X_1)^3$ for Λ_1, Λ_2 and $C := (4X_1)^4$ for Λ_3. Moreover, we consider the lattice Ω spanned by

$$v_1 := (1, 0, \lfloor C\log 2\rfloor), \quad v_2 := (0, 1, \lfloor C\log 3\rfloor), \quad v_3 := (0, 0, \lfloor C\log 5\rfloor),$$

for Λ_1

$$v_1 := (1, 0, \lfloor C\log 3\rfloor), \quad v_2 := (0, 1, \lfloor C\log 7\rfloor), \quad v_3 := (0, 0, \lfloor C\log 10\rfloor),$$

for Λ_2

$$\begin{aligned}v_1 := (1, 0, 0, \lfloor C\log 2\rfloor), \quad v_2 := (0, 1, 0, \lfloor C\log 3\rfloor),\\ v_3 := (0, 0, 1, \lfloor C\log 5\rfloor), \quad v_4 := (0, 0, 0, \lfloor C\log 7\rfloor),\end{aligned}$$

for Λ_3. Then, we compute Q, T, c_1, m according to Proposition 2.3.20 in [6] and we obtain:

$$5.5 \cdot 10^{-122} < |\Lambda_1| < \frac{1}{10^{m_1-32}} \quad \Rightarrow \quad m_1 \leq 153;$$

$$3.2 \cdot 10^{-122} < |\Lambda_2| < \frac{1}{10^{m_1-32}} \quad \Rightarrow \quad m_1 \leq 153;$$

$$8.1 \cdot 10^{-183} < |\Lambda_3| < \frac{1}{10^{m_1-32}} \quad \Rightarrow \quad m_1 \le 214.$$

Hence, we have the following numerical result.

Lemma 4.3. *The estimate* $m_1 \le 214$ *holds.*

For $a_1 \in \{1, 2, \ldots, 9\}$, $1 \le n_1 \le 1284$, $1 \le m_1 \le 214$, we solve the equations

$$x_{n_1} = P_{n_1}(x_1) = a_1 \left(\frac{10^{m_1} - 1}{9} \right)$$

to see for which values of the triple (n_1, m_1) it has a solution $x_1 = X_1/2$ with positive integer X_1, where

$$x_n = P_n(X/2) = \left(\frac{X + \sqrt{X^2 \pm 4}}{2} \right)^n + \left(\frac{X - \sqrt{X^2 \pm 4}}{2} \right)^n.$$

We used a program written in Maple to see that $n_1 = 1$ in all cases. Here, $P_n(X)$ is one of the two polynomials giving x_n in terms of x_1 for the equation $x^2 - dy^2 = \pm 4$.

From equation (4.8), for $i = 2$ we get

$$\left| n_2 \frac{\log \delta}{\log 10} - \frac{\log b_2}{\log 10} - m_2 \right| < \frac{12}{(\log 10) 10^{m_2}}, \tag{4.18}$$

where $\delta = x_1 + y_1\sqrt{d} = x_1 + \sqrt{x_1^2 \pm 4}$, $x_1 = a_1(10^{m_1} - 1)/9$, and $b_2 = a_2/9$ with $a_1 \ne a_2$. To apply Lemma 3.1 to inequality (4.18), we put

$$\kappa = \frac{\log \delta}{\log 10}, \quad \mu = \frac{\log b_2}{\log 10}, \quad A = \frac{12}{\log 10}, \quad B = 10, \quad \text{and} \quad M = 1.5 \cdot 10^{29}.$$

The program was developed in PARI/GP running with 200 digits, for $1 \le m_1 \le 214$. For the computations, if the first convergent such that $q > 6M$ does not satisfy the condition $\eta > 0$, then we use the next convergent until we find the one that satisfies the conditions. In a few minutes, all the computations were done. In all cases, after the first run we obtained $m_2 \le 35$. We set $M = 35$ and the second run of the reduction method yields $m_2 \le 8$. In conclusion, we have

$$n_1 = 1, \quad 1 \le m_1 \le 8, \quad 1 \le m_2 \le 8, \quad 1 \le n_2 \le 48.$$

Now a verification by hand yields the final result.

Acknowledgements. F. L. was supported in part by grant CPRR160325161141 and an A-rated scientist award both from the NRF of South Africa and by grant no. 17-02804S of the Czech Granting Agency. This paper was finalized during a visit of A.T. at the School of Mathematics of Wits University in August 2017. This author thanks this institution for its hospitality and the CoEMaSS at Wits for support.

References

[1] S. D. Alvarado, F. Luca: *Fibonacci numbers which are sums of two repdigits*, in: Proceedings of the XIVth International Conference on Fibonacci numbers and their applications, vol. 20, Sociedad Matematica Mexicana, Aportaciones Matemáticas, Investigación, 2011, pp. 97–108.

[2] J. J. Bravo, F. Luca: *On a conjecture about repdigits in k-generalized Fibonacci sequences*, Publ. Math. Debrecen 82 (2013), pp. 623–639, doi: 10.5486/pmd.2013.5390.

[3] Y. Bugeaud, M. Mignotte: *On integers with identical digits*, Mathematika 46 (1999), pp. 411–417, doi: 10.1112/s0025579300007865.

[4] Y. Bugeaud, M. Mignotte, S. Siksek: *Classical and modular approaches to exponential Diophantine equations I. Fibonacci and Lucas perfect powers*, Annals of Mathematics 163 (2006), pp. 969–1018, doi: 10.4007/annals.2006.163.969.

[5] Y. Bugeaud, P. Mihăilescu: *On the Nagell–Ljunggren equation* $(x^n - 1)/(x - 1) = y^q$, Math. Scand. 101 (2007), pp. 177–183, doi: 10.7146/math.scand.a-15038.

[6] H. Cohen: *Number Theory, Vol. I: Tools and Diophantine Equations*, New York: Springer, 2007, doi: 10.5860/choice.45-2655.

[7] A. Dossavi-Yovo, F. Luca, A. Togbé: *On the x-coordinates of Pell equations which are rep-digits*, Publ. Math. Debrecen 88 (2016), pp. 381–391, doi: 10.5486/pmd.2016.7378.

[8] A. Dujella, A. Pethő: *A generalization of a theorem of Baker and Davenport*, Quart. J. Math. 49.195 (1998), pp. 291–306, doi: 10.1093/qjmath/49.195.291.

[9] B. Faye, F. Luca: *On x-coordinates of Pell equations which are repdigits*, Fibonacci Quart. 56 (2018), pp. 52–62.

[10] M. Laurent, M. Mignotte, Y. Nesterenko: *Formes linéaires en deux logarithmes et déterminants d'interpolation*, J. Number Theory 55 (1995), pp. 285–321, doi: 10.1006/jnth.1995.1141.

[11] F. Luca: *Fibonacci and Lucas numbers with only one distinct digit*, Port. Math. 57 (2000), pp. 243–254.

[12] F. Luca: *Repdigits which are sums of at most three Fibonacci number*, Math. Comm. 17 (2012), pp. 1–11.

[13] F. Luca, A. Togbé: *On the x-coordinates of Pell equations which are Fibonacci numbers*, Math. Scand. 122 (2018), pp. 18–30, doi: 10.7146/math.scand.a-97271.

[14] D. Marques, A. Togbé: *On repdigits as product of consecutive Fibonacci numbers*, Rend. Istit. Mat. Univ. Trieste 44 (2012), pp. 393–397.

[15] D. Marques, A. Togbé: *On terms of linear recurrence sequences with only one distinct block of digits*, Colloq. Math. 124 (2011), pp. 145–155, doi: 10.4064/cm124-2-1.

[16] E. M. Matveev: *An explicit lower bound for a homogeneous rational linear form in logarithms of algebraic numbers, II*, Izv. Math. 64 (2000), pp. 1217–1269, doi: 10.1070/im2000v064n06abeh000314.

[17] R. Obláth: *Une propriété des puissances parfaites*, Mathesis 65 (1956), pp. 356–364.

[18] W. R. Spickerman: *Binet's formula for the Tribonacci numbers*, Fibonacci Quart. 20 (1982), pp. 118–120.

[19] K. Yu: *p-adic logarithmic forms and group varieties II*, Acta Arith. 89 (1999), pp. 337–378, doi: 10.4064/aa-89-4-337-378.

Annales Mathematicae et Informaticae
***50** (2019) pp. 145–157*
DOI: 10.33039/ami.2019.03.003
http://ami.uni-eszterhazy.hu

Arithmetic subderivatives and Leibniz-additive functions

Jorma K. Merikoski[a], Pentti Haukkanen[a], Timo Tossavainen[b]

[a]Faculty of Information Technology and Communication Sciences,
FI-33014 Tampere University, Finland
jorma.merikoski@tuni.fi, pentti.haukkanen@tuni.fi

[b]Department of Arts, Communication and Education, Lulea University of Technology
SE-97187 Lulea, Sweden
timo.tossavainen@ltu.se

Submitted: June 15, 2018
Accepted: March 25, 2019
Published online: April 13, 2019

Abstract

We introduce the arithmetic subderivative of a positive integer with respect to a non-empty set of primes. This notion generalizes the concepts of the arithmetic derivative and arithmetic partial derivative. In order to generalize these notions a step further, we define that an arithmetic function f is Leibniz-additive if there is a nonzero-valued and completely multiplicative function h_f satisfying $f(mn) = f(m)h_f(n) + f(n)h_f(m)$ for all positive integers m and n. We study some basic properties of such functions. For example, we present conditions when an arithmetic function is Leibniz-additive and, generalizing the well-known bounds for the arithmetic derivative, we establish bounds for a Leibniz-additive function.

Keywords: arithmetic derivative, Leibniz rule, additivity, multiplicativity

MSC: 11A25, 11A05

1. Introduction

We let $\mathbb{P}$, $\mathbb{Z}_+$, $\mathbb{N}$, $\mathbb{Z}$, and $\mathbb{Q}$ stand for the set of primes, positive integers, nonnegative integers, integers, and rational numbers, respectively.

Let $n \in \mathbb{Z}_+$. There is a unique sequence $(\nu_p(n))_{p\in\mathbb{P}}$ of nonnegative integers (with only finitely many positive terms) such that

$$n = \prod_{p\in\mathbb{P}} p^{\nu_p(n)}. \tag{1.1}$$

We use this notation throughout.

Let $\emptyset \neq S \subseteq \mathbb{P}$. We define the *arithmetic subderivative* of n with respect to S as

$$D_S(n) = n'_S = n \sum_{p\in S} \frac{\nu_p(n)}{p}.$$

In particular, $n'_{\mathbb{P}}$ is the *arithmetic derivative* of n, defined by Barbeau [2] and studied further by Ufnarovski and Åhlander [10]. Another well-known special case is $n'_{\{p\}}$, the *arithmetic partial derivative* of n with respect to $p \in \mathbb{P}$, defined by Kovič [7] and studied further by the present authors and Mattila [4, 5].

We define the *arithmetic logarithmic subderivative* of n with respect to S as

$$\mathrm{ld}_S(n) = \frac{D_S(n)}{n} = \sum_{p\in S} \frac{\nu_p(n)}{p}.$$

In particular, $\mathrm{ld}_{\mathbb{P}}(n)$ is the *arithmetic logarithmic derivative* of n. This notion was originally introduced by Ufnarovski and Åhlander [10].

An arithmetic function g is *completely additive* (or *c-additive*, for short) if $g(mn) = g(m)+g(n)$ for all $m, n \in \mathbb{Z}_+$. It follows from the definition that $g(1) = 0$. An arithmetic function h is *completely multiplicative* (or *c-multiplicative*, for short) if $h(1) = 1$ and $h(mn) = h(m)h(n)$ for all $m, n \in \mathbb{Z}_+$. The following theorems recall that these functions are totally determined by their values at primes. The proofs are simple and omitted.

Theorem 1.1. *Let g be an arithmetic function, and let $(x_p)_{p\in\mathbb{P}}$ be a sequence of real numbers. The following conditions are equivalent:*

(a) *g is c-additive and $g(p) = x_p$ for all $p \in \mathbb{P}$;*

(b) *for all $n \in \mathbb{Z}_+$,*

$$g(n) = \sum_{p\in\mathbb{P}} \nu_p(n) x_p.$$

Theorem 1.2. *Let h be an arithmetic and nonzero-valued function, and let $(y_p)_{p\in\mathbb{P}}$ be a sequence of nonzero real numbers. The following conditions are equivalent:*

(a) *h is c-multiplicative and $h(p) = y_p$ for all $p \in \mathbb{P}$;*

(b) *for all $n \in \mathbb{Z}_+$,*

$$h(n) = \prod_{p\in\mathbb{P}} y_p^{\nu_p(n)}.$$

We say that an arithmetic function f is *Leibniz-additive* (or *L-additive*, for short) if there is a nonzero-valued and c-multiplicative function h_f such that

$$f(mn) = f(m)h_f(n) + f(n)h_f(m) \tag{1.2}$$

for all $m, n \in \mathbb{Z}_+$. Then $f(1) = 0$, since $h_f(1) = 1$. The property (1.2) may be considered a generalized Leibniz rule. Substituting $m = n = p \in \mathbb{P}$ and applying induction, we get

$$f(p^a) = af(p)h(p)^{a-1} \tag{1.3}$$

for all $p \in \mathbb{P}$, $a \in \mathbb{Z}_+$.

The arithmetic subderivative D_S is L-additive with $h_{D_S} = N$, where N is the identity function $N(n) = n$. A c-additive function g is L-additive with $h_g = E$, where $E(n) = 1$ for all $n \in \mathbb{Z}_+$. The arithmetic logarithmic subderivative ld_S is c-additive and hence L-additive.

This paper is a sequel to [6], where we defined L-additivity without requiring that h_f is nonzero-valued. We begin by showing how the values of an L-additive function f are determined in $\mathbb{Z}_+$ by the values of f and h_f at primes (Section 2) and then study under which conditions an arithmetic function f can be expressed as $f = gh$, where g is c-additive and h is nonzero-valued and c-multiplicative (Section 3). It turns out that the same conditions are necessary for L-additivity (Section 4). Finally, extending Barbeau's [2] and Westrick's [11] results, we present some lower and upper bounds for an L-additive function (Section 5). We complete our paper with some remarks (Section 6).

2. Constructing $f(n)$ and $h_f(n)$

An L-additive function f is not totally defined by its values at primes. Also, the values of h_f at primes must be known.

Theorem 2.1. *Let f be an arithmetic function, and let $(x_p)_{p\in\mathbb{P}}$ and $(y_p)_{p\in\mathbb{P}}$ be as in Theorems 1.1 and 1.2. The following conditions are equivalent:*

(a) *f is L-additive and $f(p) = x_p$, $h_f(p) = y_p$ for all $p \in \mathbb{P}$;*

(b) *for all $n \in \mathbb{Z}_+$,*

$$f(n) = \Big(\sum_{p\in\mathbb{P}} \nu_p(n)\frac{x_p}{y_p}\Big) \prod_{p\in\mathbb{P}} y_p^{\nu_p(n)}.$$

Proof. (a) $\Rightarrow$ (b). Since $f(1) = 0$, (b) holds for $n = 1$. So, let $n > 1$. Denoting

$$\{p_1, \dots, p_s\} = \{p \in \mathbb{P} \mid \nu_p(n) > 0\}$$

and

$$a_i = \nu_{p_i}(n), \quad i = 1, \dots, s,$$

we have

$$\begin{aligned}
f(n) &= \sum_{i=1}^{s} h_f(p_1)^{a_1} \cdots h_f(p_{i-1})^{a_{i-1}} f(p_i^{a_i}) h_f(p_{i+1})^{a_{i+1}} \cdots h_f(p_s)^{a_s} \\
&= \sum_{i=1}^{s} h_f(p_1)^{a_1} \cdots h_f(p_{i-1})^{a_{i-1}} a_i f(p_i) h_f(p_i)^{a_i-1} h_f(p_{i+1})^{a_{i+1}} \cdots h_f(p_s)^{a_s} \\
&= \sum_{p\in\mathbb{P}} \Big(\nu_p(n) f(p) h_f(p)^{\nu_p(n)-1} \prod_{\substack{q\in\mathbb{P} \\ q\neq p}} h_f(q)^{\nu_q(n)} \Big) \\
&= \sum_{p\in\mathbb{P}} \Big(\nu_p(n) \frac{f(p)}{h_f(p)} \prod_{q\in\mathbb{P}} h_f(q)^{\nu_q(n)} \Big) \\
&= \Big(\sum_{p\in\mathbb{P}} \nu_p(n) \frac{x_p}{y_p} \Big) \prod_{p\in\mathbb{P}} y_p^{\nu_p(n)}.
\end{aligned}$$

The first equation can be proved by induction on s, the second holds by (1.3), and the remaining equations are obvious.

(b) $\Rightarrow$ (a). We define now

$$h(n) = \prod_{p\in\mathbb{P}} y_p^{\nu_p(n)}.$$

Let $m, n \in \mathbb{Z}_+$. Then

$$\begin{aligned}
f(mn) &= \Big(\sum_{p\in\mathbb{P}} \nu_p(mn) \frac{x_p}{y_p} \Big) \prod_{p\in\mathbb{P}} y_p^{\nu_p(mn)} \\
&= \Big(\sum_{p\in\mathbb{P}} (\nu_p(m) + \nu_p(n)) \frac{x_p}{y_p} \Big) \prod_{p\in\mathbb{P}} y_p^{\nu_p(m)+\nu_p(n)} \\
&= \Big(\sum_{p\in\mathbb{P}} (\nu_p(m) + \nu_p(n)) \frac{x_p}{y_p} \Big) \Big(\prod_{p\in\mathbb{P}} y_p^{\nu_p(m)} \Big) \Big(\prod_{p\in\mathbb{P}} y_p^{\nu_p(n)} \Big) \\
&= \Big(\sum_{p\in\mathbb{P}} \nu_p(m) \frac{x_p}{y_p} \Big(\prod_{p\in\mathbb{P}} y_p^{\nu_p(m)} \Big) \Big) \Big(\prod_{p\in\mathbb{P}} y_p^{\nu_p(n)} \Big) \\
&\quad + \Big(\sum_{p\in\mathbb{P}} \nu_p(n) \frac{x_p}{y_p} \Big(\prod_{p\in\mathbb{P}} y_p^{\nu_p(n)} \Big) \Big) \Big(\prod_{p\in\mathbb{P}} y_p^{\nu_p(m)} \Big) \\
&= f(m)h(n) + f(n)h(m).
\end{aligned}$$

So, f is L-additive with $h_f = h$. It is clear that $f(p) = x_p$ and $h_f(p) = y_p$ for all $p \in \mathbb{P}$. □

Next, we construct h_f from f. Let us denote

$$U_f = \{p \in \mathbb{P} \mid f(p) \neq 0\}, \quad V_f = \{p \in \mathbb{P} \mid f(p) = 0\}.$$

If $f = \theta$, where $\theta(n) = 0$ for all $n \in \mathbb{Z}_+$, then any h_f applies. Hence, we now assume that $f \neq \theta$. Then $U_f \neq \emptyset$.

Since

$$f(p^2) = 2f(p)h_f(p)$$

by (1.3), we have

$$h_f(p) = \frac{f(p^2)}{2f(p)} \quad \text{for } p \in U_f.$$

The case $p \in V_f$ remains. Let $q \in \mathbb{P}$. Then (1.2) implies that

$$f(pq) = f(p)h_f(q) + f(q)h_f(p) = f(q)h_f(p).$$

Therefore,

$$h_f(p) = \frac{f(pq)}{f(q)} \quad \text{for } p \in V_f, \tag{2.1}$$

where $q \in U_f$ is arbitrary. Now, by Theorem 1.2,

$$h_f(n) = \Big(\prod_{p \in U_f} \Big(\frac{f(p^2)}{2f(p)} \Big)^{\nu_p(n)} \Big) \Big(\prod_{p \in V_f} \Big(\frac{f(pq)}{f(q)} \Big)^{\nu_p(n)} \Big), \tag{2.2}$$

where $q \in U_f$ is arbitrary. (If $V_f = \emptyset$, then the latter factor is the "empty product" one.) We have thus proved the following theorem.

Theorem 2.2. *If $f \neq \theta$ is L-additive, then h_f is unique and determined by* (2.2).

3. Decomposing $f = gh$

Let f be an arithmetic function and let h be a nonzero-valued and c-multiplicative function. By Theorem 2.1, f is L-additive with $h_f = h$ if and only if

$$f(n) = \Big(\sum_{p \in \mathbb{P}} \nu_p(n) \frac{f(p)}{h(p)} \Big) \prod_{p \in \mathbb{P}} h(p)^{\nu_p(n)} = \Big(\sum_{p \in \mathbb{P}} \nu_p(n) \frac{f(p)}{h(p)} \Big) h(n). \tag{3.1}$$

The function

$$g(n) = \sum_{p \in \mathbb{P}} \nu_p(n) \frac{f(p)}{h(p)}$$

is c-additive by Theorem 1.1.

We say that an arithmetic function f is *gh-decomposable* if it has a *gh decomposition*

$$f = gh,$$

where g is c-additive and h is nonzero-valued and c-multiplicative. We saw above that L-additivity implies gh-decomposability. Also, the converse holds.

Theorem 3.1. *Let f be an arithmetic function. The following conditions are equivalent:*

(a) *f is L-additive;*

(b) *f is gh-decomposable.*

Proof. (a) $\Rightarrow$ (b). We proved this above.
(b) $\Rightarrow$ (a). For all $m, n \in \mathbb{Z}_+$,

$$\begin{aligned} f(mn) &= g(mn)h(mn) = (g(m) + g(n))h(m)h(n) \\ &= g(m)h(m)h(n) + g(n)h(n)h(m) = f(m)h(n) + f(n)h(m). \end{aligned}$$

Consequently, f is L-additive with $h_f = h$. □

Corollary 3.2. *Let $f \neq \theta$ be an arithmetic function. The following conditions are equivalent:*

(a) *f is L-additive;*

(b) *f is uniquely gh-decomposable.*

Proof. In proving (a) $\Rightarrow$ (b), h_f is unique by Theorem 2.2. Since h_f is nonzero-valued, also $g = f/h_f$ is unique. □

For example, if $f = D_S$, then $g = \mathrm{ld}_S$ and $h = N$.

By Theorem 2.2, an L-additive function $f \neq \theta$ determines h_f uniquely. We consider next the converse problem: Given a nonzero-valued and c-multiplicative function h, find an L-additive function f such that $h_f = h$.

Theorem 3.3. *Let $(x_p)_{p\in\mathbb{P}}$ be a sequence of real numbers and let h be nonzero-valued and c-multiplicative. There is a unique L-additive function f with $h_f = h$ such that $f(p) = x_p$ for all $p \in \mathbb{P}$.*

Proof. If at least one $x_p \neq 0$, then apply Theorem 2.1 and Corollary 3.2. Otherwise, $f = \theta$. □

We can now characterize D_S and ld_S.

Corollary 3.4. *Let f be an arithmetic function and $\emptyset \neq S \subseteq \mathbb{P}$. The following conditions are equivalent:*

(a) *f is L-additive, $h_f = N$, $f(p) = 1$ for $p \in S$, and $f(p) = 0$ for $p \in \mathbb{P} \setminus S$;*

(b) *$f = D_S$.*

Corollary 3.5. *Let g be an arithmetic function and $\emptyset \neq S \subseteq \mathbb{P}$. The following conditions are equivalent:*

(a) *g is c-additive, $g(p) = 1/p$ for $p \in S$, and $g(p) = 0$ for $p \in \mathbb{P} \setminus S$;*

(b) *$g = \mathrm{ld}_S$.*

4. Conditions for L-additivity

Let $f \neq \theta$ be L-additive and $a, b \in \mathbb{N}$.

First, let $p \in \mathbb{P}$. By (1.3),

$$f(p^{a+1}) = (a+1)f(p)h_f(p)^a, \quad f(p^{b+1}) = (b+1)f(p)h_f(p)^b, \tag{4.1}$$

and, further,

$$f(p^{a+1})^b = (a+1)^b f(p)^b h_f(p)^{ab}, \ f(p^{b+1})^a = (b+1)^a f(p)^a h_f(p)^{ba}. \tag{4.2}$$

Assume now that $p \in U_f$. Then the right-hand sides of the equations in (4.1) are nonzero and $f(p^{a+1}), f(p^{b+1}) \neq 0$. Therefore, by (4.2),

$$\frac{f(p^{a+1})^b}{f(p^{b+1})^a} = \frac{(a+1)^b f(p)^b}{(b+1)^a f(p)^a}$$

or, equivalently,

$$\Big(\frac{f(p^{a+1})}{(a+1)f(p)}\Big)^b = \Big(\frac{f(p^{b+1})}{(b+1)f(p)}\Big)^a.$$

Second, assume that U_f has at least two elements. If $p, q \in U_f$, then (1.2) and (1.3) imply that

$$\begin{aligned} f(p^a q^b) &= f(p^a)h_f(q^b) + f(q^b)h_f(p^a) \\ &= f(p^a)h_f(q)^b + f(q^b)h_f(p)^a = \frac{f(p^a)f(q^{b+1})}{(b+1)f(q)} + \frac{f(q^b)f(p^{a+1})}{(a+1)f(p)}. \end{aligned}$$

Third, assume additionally that $V_f \neq \emptyset$. Let $p \in V_f$ and $q_1, q_2 \in U_f$. By (2.1) and the fact that h_f is nonzero-valued,

$$\frac{f(pq_1)}{f(q_1)} = \frac{f(pq_2)}{f(q_2)} \neq 0.$$

In other words, we can "cancel" p in

$$\frac{f(pq_1)}{f(pq_2)} = \frac{f(q_1)}{f(q_2)} \neq 0.$$

Fourth, both the nonzero-valuedness of h_f and (2.2) imply that

$$f(p^2) \neq 0 \quad \text{for all } p \in U_f.$$

We have thus found necessary conditions for L-additivity.

Theorem 4.1. *Let $f \neq \theta$ be L-additive and $a, b \in \mathbb{N}$.*

(i) *If $p \in U_f$, then*

$$\Big(\frac{f(p^{a+1})}{(a+1)f(p)}\Big)^b = \Big(\frac{f(p^{b+1})}{(b+1)f(p)}\Big)^a.$$

(ii) *If* $p, q \in U_f$, *then*

$$f(p^a q^b) = \frac{f(p^a)f(q^{b+1})}{(b+1)f(q)} + \frac{f(q^b)f(p^{a+1})}{(a+1)f(p)}.$$

(iii) *If* $p \in V_f$ *and* $q_1, q_2 \in U_f$, *then*

$$\frac{f(pq_1)}{f(pq_2)} = \frac{f(q_1)}{f(q_2)} \neq 0.$$

(iv) *If* $p \in U_f$, *then*

$$f(p^2) \neq 0.$$

The question about the sufficiency of these conditions remains open.

To find sufficient conditions for L-additivity, we study under which conditions we can apply the procedure described in the proof of Theorem 2.2 to a given arithmetic function $f \neq \theta$. The function h, defined as h_f in (2.2), must be (α) well-defined, (β) c-multiplicative, and (γ) nonzero-valued. Condition (α) follows from (iii), (β) is obvious, and (γ) follows from (iii) and (iv). If the function $g = f/h$ is also c-additive, then f is L-additive by Theorem 3.1. So, we have found sufficient conditions for L-additivity, and they are obviously also necessary.

Theorem 4.2. *An arithmetic function* $f \neq \theta$ *is L-additive if and only if* (iii) *and* (iv) *in Theorem* 4.1 *are satisfied and the function* f/h *is c-additive, where*

$$h(n) = \Big(\prod_{p \in U_f} \Big(\frac{f(p^2)}{2f(p)} \Big)^{\nu_p(n)} \Big) \Big(\prod_{p \in V_f} \Big(\frac{f(pq)}{f(q)} \Big)^{\nu_p(n)} \Big), \quad q \in U_f.$$

5. Bounds for an L-additive function

Let us express (1.1) as

$$n = q_1 \cdots q_r, \tag{5.1}$$

where $q_1, \ldots, q_r \in \mathbb{P}$, $q_1 \leq \cdots \leq q_r$. We first recall the well-known bounds for $D(n)$ using n and r only.

Theorem 5.1. *Let* n *be as in* (5.1). *Then*

$$r n^{\frac{r-1}{r}} \leq D(n) \leq \frac{rn}{2} \leq \frac{n \log_2 n}{2}. \tag{5.2}$$

Equality is attained in the upper bounds if and only if n *is a power of* 2, *and in the lower bound if and only if* n *is a prime or a power of* 2.

Proof. See [2, pp. 118–119], [10, Theorem 9]. □

The first upper bound can be improved using the same information. Westrick [11, Ineq. (6)] presented in her thesis the following bound without proof.

Theorem 5.2. *Let n be as in* (5.1). *Then*

$$D(n) \le \frac{r-1}{2}n + 2^{r-1}. \tag{5.3}$$

Equality is attained if and only if $n \in \mathbb{P}$ or $q_1 = \cdots = q_{r-1} = 2$.

Proof. If $r = 1$ (i.e., $n \in \mathbb{P}$), then (5.3) clearly holds with equality. So, assume that $r > 1$.

Case 1. $q_1 = \cdots = q_{r-1} = 2$. Then

$$D(n) = n\Big(\frac{r-1}{2} + \frac{1}{q_r}\Big) = \frac{r-1}{2}n + \frac{n}{n/2^{r-1}} = \text{rhs(5.3)},$$

where "rhs" is short for "the right-hand side".

Case 2. $q_1 = \cdots = q_{r-2} = 2$ (omit this if $r = 2$) and $q_{r-1} > 2$. Since

$$\frac{1}{q_{r-1}} + \frac{1}{q_r} = \frac{1}{2} + \frac{4 - (q_{r-1}-2)(q_r-2)}{2q_{r-1}q_r} < \frac{1}{2} + \frac{2}{q_{r-1}q_r},$$

we have

$$D(n) < n\Big(\frac{r-2}{2} + \frac{1}{2} + \frac{2}{q_{r-1}q_r}\Big) = \frac{r-1}{2}n + \frac{2n}{n/2^{r-2}} = \text{rhs(5.3)}.$$

Case 3. $q_{r-2} > 2$. Then $r \ge 3$ and

$$D(n) \le n\Big(\frac{r-3}{2} + \frac{1}{3} + \frac{1}{3} + \frac{1}{3}\Big) = \frac{r-1}{2}n < \text{rhs(5.3)}.$$

The claim with equality conditions is thus verified. Because

$$\frac{rn}{2} - \Big(\frac{r-1}{2}n + 2^{r-1}\Big) = \frac{n}{2} - 2^{r-1} \ge \frac{2^r}{2} - 2^{r-1} = 0,$$

the upper bound (5.3) indeed improves (5.2). □

We extend the upper bounds (5.2) and (5.3) under the assumption

$$h_f(p) \ge p \quad \text{for all } p \in U_f. \tag{5.4}$$

Let n in (5.1) have $q_{i_1}, \ldots, q_{i_s} \in U_f$. We denote

$$p_1 = q_{i_1}, \ \ldots, \ p_s = q_{i_s} \tag{5.5}$$

and

$$M = \max_{1 \le i \le r} f(q_i) = \max_{1 \le i \le s} f(p_i). \tag{5.6}$$

Theorem 5.3. *Let $f \ne \theta$ be nonnegative and L-additive satisfying* (5.4). *Then*

$$f(n) \le \frac{sM}{2} h_f(n) \le \frac{M \log_2 n}{2} h_f(n), \tag{5.7}$$

where s is as in (5.5) *and M is as in* (5.6). *Equality is attained if and only if n is a power of* 2.

Proof. By (3.1) and simple manipulation,

$$f(n) = h_f(n)\sum_{i=1}^{r}\frac{f(q_i)}{h_f(q_i)} = h_f(n)\sum_{i=1}^{s}\frac{f(p_i)}{h_f(p_i)} \le h_f(n)M\sum_{i=1}^{s}\frac{1}{p_i}$$
$$\le h_f(n)M\sum_{i=1}^{s}\frac{1}{2} = h_f(n)M\frac{s}{2} \le h_f(n)M\frac{r}{2} \le h_f(n)M\frac{\log_2 n}{2}.$$

The equality condition is obvious. □

Theorem 5.4. *Let $f \neq \theta$ be nonnegative and L-additive satisfying* (5.4). *Then*

$$f(n) \le \Big(\frac{s-1}{2}h_f(n) + h_f(2^{s-1})\Big)M, \tag{5.8}$$

where s is as in (5.5) *and M is as in* (5.6). *Equality is attained if and only if $n \in \mathbb{P}$ or $p_1 = \cdots = p_{s-1} = 2 = h_f(2)$.*

Proof. If $s = 1$ (i.e., $n \in \mathbb{P}$), then (5.8) clearly holds with equality. So, assume that $s > 1$.

Case 1. $p_1 = \cdots = p_{s-1} = 2$. Then

$$\begin{aligned} f(n) &= f(2^{s-1}p_s) = f(2^{s-1})h_f(p_s) + f(p_s)h_f(2^{s-1}) \\ &= (s-1)f(2)h_f(2^{s-2})h_f(p_s) + f(p_s)h_f(2^{s-1}) \\ &\le \big((s-1)(h_f(2^{s-2})h_f(p_s) + h_f(2^{s-1})\big)M \\ &\le \Big((s-1)h_f(2^{s-2})h_f(p_s)\frac{h_f(2)}{2} + h_f(2^{s-1})\Big)M \\ &= \Big(\frac{s-1}{2}h_f(n) + h_f(2^{s-1})\Big)M. \end{aligned}$$

Case 2. $p_1 = \cdots = p_{s-2} = 2$ (omit this if $s = 2$) and $p_{s-1} > 2$. If $s \ge 3$, then

$$\begin{aligned} f(n) &= f(2^{s-2}p_{s-1}p_s) = f(2^{s-2})h_f(p_{s-1}p_s) + f(p_{s-1}p_s)h_f(2^{s-2}) \\ &= (s-2)f(2)h_f(2^{s-3})h_f(p_{s-1}p_s) + f(p_{s-1}p_s)h_f(2^{s-2}) \\ &= \frac{s-2}{2}f(2)h_f(2^{s-2})h_f(p_{s-1}p_s) + \big(f(p_{s-1})h_f(p_s) + f(p_s)h_f(p_{s-1})\big)h_f(2^{s-2}) \\ &\le \Big(\frac{s-2}{2}h_f(2^{s-2})h_f(p_{s-1}p_s) + (h_f(p_{s-1}) + h_f(p_s))h_f(2^{s-2})\Big)M \\ &= \Big(\frac{s-2}{2}h_f(n) + (h_f(p_{s-1}) + h_f(p_s))h_f(2^{s-2})\Big)M \\ &= \Big(\frac{s-1}{2}h_f(n) + (h_f(p_{s-1}) + h_f(p_s))h_f(2^{s-2}) - \frac{1}{2}h_f(n)\Big)M. \end{aligned}$$

The last expression is obviously an upper bound for $f(n)$ also if $s = 2$. If

$$(h_f(p_{s-1}) + h_f(p_s))h_f(2^{s-2}) - \frac{1}{2}h_f(n) \le h_f(2^{s-1}),$$

i.e.,

$$2(h_f(p_{s-1}) + h_f(p_s)) - h_f(p_{s-1})h_f(p_s) \le 2h_f(2),$$

then (5.8) follows. Since

$$\begin{aligned} h_f(p_{s-1})h_f(p_s) - 2(h_f(p_{s-1}) + h_f(p_s)) + 4 &= (h_f(p_{s-1}) - 2)(h_f(p_s) - 2) \\ &\ge (p_{s-1} - 2)(p_s - 2) > 0, \end{aligned}$$

we actually have a stronger inequality

$$2(h_f(p_{s-1}) + h_f(p_s)) - h_f(p_{s-1})h_f(p_s) < 4.$$

Case 3. $p_{s-2} > 2$. Then $s \ge 3$ and

$$\begin{aligned} f(n) &= f(p_1)h_f(p_2 \cdots p_s) + f(p_2 \cdots p_s)h_f(p_1) \\ &= f(p_1)\frac{h_f(n)}{h_f(p_1)} + f(p_2 \cdots p_s)h_f(p_1) \\ &\le \frac{Mh_f(n)}{2} + f(p_2 \cdots p_s)h_f(p_1). \end{aligned}$$

Since

$$\begin{aligned} f(p_2 \cdots p_s)h_f(p_1) &= \big(f(p_2)h_f(p_3 \cdots p_s) + f(p_3 \cdots p_s)h_f(p_2)\big)h_f(p_1) \\ &= f(p_2)\frac{h_f(n)}{h_f(p_2)} + f(p_3 \cdots p_s)h_f(p_1p_2) \\ &\le \frac{Mh_f(n)}{2} + f(p_3 \cdots p_s)h_f(p_1p_2), \end{aligned}$$

we also have

$$f(n) \le 2\frac{Mh_f(n)}{2} + f(p_3 \cdots p_s)h_f(p_1p_2).$$

Similarly,

$$f(n) \le \frac{s-3}{2}Mh_f(n) + f(p_{s-2}p_{s-1}p_s)h_f(p_1 \cdots p_{s-3}). \tag{5.9}$$

Because

$$\begin{aligned} f(p_{s-2}p_{s-1}p_s) &= f(p_{s-2})h_f(p_{s-1}p_s) + f(p_{s-1})h_f(p_{s-2}p_s) + f(p_s)h_f(p_{s-2}p_{s-1}) \\ &\le Mh_f(p_{s-2}p_{s-1}p_s)\Big(\frac{1}{p_{s-2}} + \frac{1}{p_{s-1}} + \frac{1}{p_s}\Big) \\ &\le Mh_f(p_{s-2}p_{s-1}p_s)\Big(\frac{1}{3} + \frac{1}{3} + \frac{1}{3}\Big) = Mh_f(p_{s-2}p_{s-1}p_s), \end{aligned}$$

it follows from (5.9) that

$$f(n) \le \frac{s-3}{2}Mh_f(n) + Mh_f(n) = \frac{s-1}{2}Mh_f(n).$$

In other words, (5.8) holds strictly.

The proof is complete. It also includes the equality conditions. □

If we do not know s (but know r), we can substitute $s = r$ in (5.7) and (5.8). We complete this section by extending the lower bound (5.2).

Theorem 5.5. *Let f be nonnegative and L-additive, and let n be as in* (5.1) *with*

$$h_f(q_1), \dots, h_f(q_r) > 0.$$

Then

$$f(n) \geq r m h_f(n)^{\frac{r-1}{r}},$$

where

$$m = \min_{1 \leq i \leq r} f(q_i).$$

Equality is attained if and only if n is a prime or a power of 2.

Proof. By (3.1) and the arithmetic-geometric mean inequality,

$$\begin{aligned} f(n) &= h_f(n) \sum_{i=1}^{r} \frac{f(q_i)}{h_f(q_i)} \geq h_f(n) m \sum_{i=1}^{r} \frac{1}{h_f(q_i)} \geq h_f(n) m \frac{r}{(h_f(q_1) \cdots h_f(q_r))^{\frac{1}{r}}} \\ &= h_f(n) m \frac{r}{h_f(q_1 \cdots q_r)^{\frac{1}{r}}} = h_f(n) m \frac{r}{h_f(n)^{\frac{1}{r}}} = r h_f(n)^{1-\frac{1}{r}} m. \end{aligned}$$

The equality condition is obvious. □

6. Concluding remarks

According to the common custom, we credited in Section 1 the arithmetic derivative to Barbeau [2]. However, Mingot Shelly [8] considered it as early as in 1911. His paper has been overlooked for a long time and is found only recently [1, 9]. The only reference to it that we know from the past decades is in Dickson [3].

A nice introduction to the arithmetic derivative is Balzarotti and Lava [1] (written in Italian, but an English reader understands its formulas and mathematical terms). There is an extensive literature about this topic, but much work is still left to be done. For example, there is only a few results about "arithmetic integration" and, more generally, about "arithmetic differential equations".

For another example, let us define $D = D_{\mathbb{P}}$ as a function $\mathbb{Q} \to \mathbb{Q}$ by allowing $\nu_p(n) \in \mathbb{Z}$ in (1.1). What do we know about this function? Not much. We are currently investigating whether D (and, more generally, D_S) is discontinuous everywhere and, if so, how strongly.

The arithmetic partial derivative $D_p = D_{\{p\}}$ has received less attention than D and, according to our knowledge, the arithmetic subderivative D_S is a new concept. An overall question related to this notion is: Which properties of D and D_p can in some way be extended to D_S? Probably the cases of finite S and infinite S must then be studied separately.

As an extension of D_S, we defined the concept of an L-additive function f. For simplicity, we stated (contrary to [6]) that h_f must be nonzero-valued. If we allow

h_f to be zero, it turns out that we only meet extra work without gaining anything significant in results. Anyway, a very general question arises: Which properties of D_S can be extended to f? In Section 5, we found the generalizations of the classical upper and lower bounds of D. But what about other properties? This remains to be seen.

References

[1] G. Balzarotti, P. P. Lava: *La derivata aritmetica: Alla scoperta di un nuovo approccio alla teoria dei numeri*, Milan: Hoepli, 2013.

[2] E. J. Barbeau: *Remarks on an arithmetic derivative*, Canadian Mathematical Bulletin 4 (1961), pp. 117–122, doi: 10.4153/cmb-1961-013-0.

[3] L. E. Dickson: *History of the Theory of Numbers*, Washington: Carnegie Institution, 1919.

[4] P. Haukkanen, J. K. Merikoski, M. Mattila, T. Tossavainen: *The arithmetic Jacobian matrix and determinant*, Journal of Integer Sequences 20 (2017), Art. 17.9.2.

[5] P. Haukkanen, J. K. Merikoski, T. Tossavainen: *On arithmetic partial differential equations*, Journal of Integer Sequences 19 (2016), Art. 16.8.6.

[6] P. Haukkanen, J. K. Merikoski, T. Tossavainen: *The arithmetic derivative and Leibniz-additive functions*, Notes on Number Theory and Discrete Mathematics 24.3 (2018), pp. 68–76, doi: 10.7546/nntdm.2018.24.3.68-76.

[7] J. Kovič: *The arithmetic derivative and antiderivative*, Journal of Integer Sequences 15 (2012), Art. 12.3.8.

[8] J. M. Shelly: *Una cuestión de la teoría de los números*, Asociación española, Granada (1911), pp. 1–12.

[9] N. J. A. Sloane: *The On-Line Encyclopedia of Integer Sequences, Seq. A003415.*

[10] V. Ufnarovski, B. Åhlander: *How to differentiate a number*, Journal of Integer Sequences 6 (2003), Art. 03.3.4.

[11] L. Westrick: *Investigations of the number derivative*, Student thesis, Massachusetts Institute of Technology, 2003.

Annales Mathematicae et Informaticae
***50** (2019) pp. 159–165*
DOI: 10.33039/ami.2019.11.002
http://ami.uni-eszterhazy.hu

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

Armand Noubissie[a], **Alain Togbé**[b]

[a]Institut de Mathématiques et de Sciences Physiques, Dangbo, Bénin
armand.noubissie@imsp-uac.org

[b]Department of Mathematics, Statistics and Computer Science
Purdue University Northwest, Westville, USA
atogbe@pnw.edu

Submitted: June 13, 2019
Accepted: November 19, 2019
Published online: December 1, 2019

Abstract

Let a and b be two distinct fixed positive integers such that $\min\{a, b\} > 1$. We show that the equation in the title with $b \equiv 3 \pmod{12}$ and a even has no solution in positive integers (n, x). This generalizes a result of Szalay [9]. Moreover, we show that this equation in the title with ($a \equiv 4 \pmod{10}$ and $b \equiv 0 \pmod{5}$) has no solution in positive integer (n, x). We give a necessary and sufficient condition for Diophantine equation $(a^n - 1)(b^n - 1) = x^2$ with ($a \equiv 4 \pmod{5}$ and $b \equiv 0 \pmod{5}$) or ($a \equiv 3 \pmod{4}$ and $b \equiv 0 \pmod{2}$) to have positive integer solutions. Finally, we prove that the equation with a even, $\vartheta_2(b-1) = 1$ and $5 \mid b$ has no solution in positive integer (n, x), where ϑ_2 is the 2-adic valuation.

Keywords: Pell equation, exponential Diophantine equation.

MSC: 11D41, 11D61

1. Introduction

Let $\mathbb{N}^+$ be the set of all positive integers. Let $a > 1$ and $b > 1$ be different fixed integers. The exponential Diophantine equation

$$(a^n - 1)(b^n - 1) = x^2, \ x, n \in \mathbb{N}^+ \tag{1.1}$$

has been studied by many authors in the literature since 2000. First, Szalay [9] studied equation (1.1) for $(a,b) = (2,3)$ and showed that this equation has no positive integer solutions. He also proved that equation (1.1) has only the positive integer solution $(n,x) = (1,2)$, for $(a,b) = (2,5)$ and there is no solution, for $(a,b) = (2,2^k)$ with $k \geq 2$ except when $n = 3$ and $k = 2$. Hajdu and Szalay [3] proved that equation (1.1) has no solution for $(a,b) = (2,6)$ and for $(a,b) = (a,a^k)$, there is no solution with $k \geq 2$ and $kn > 2$ except for the three cases $(a,n,k) = (2,3,2),(3,1,5),(7,1,4)$. So their result generalized Theorem 3 of [9]. This result was extended by Cohn [2] to the case $a^k = b^l$ (see RESULT 1). Cohn also proved that there are no solutions to (1.1) when $4 \mid n$, except for $\{a,b\} = \{13,239\}$ with $n = 4$. Walsh and Luca [7] proved equation (1.1) has finitely positive solutions for fixed (a,b) and showed that the equation has no solution with $n > 2$ for some pairs (a,b) in the range $1 < a < b \leq 100$. Theorem 1.1 completes this result ([7, Theorem 3.1]) for some special cases. Since then, many authors studied equation (1.1) by introducing some special contraints to a or b (see for examples [4, 6, 8, 10, 11]). Yuan and Zhang [11] showed that equation (1.1) has no solution with $n > 2$ if ($a \equiv 2 \pmod 3$ and $b \equiv 0 \pmod 3$) or ($a \equiv 4 \pmod 5$ and $b \equiv 0 \pmod 5$) or ($a \equiv 3 \pmod 4$ and $b \equiv 0 \pmod 2$). But this proof was not complete because Lemma 2 in their paper is not correct. The authors and Z. Zhang completed the proof of this theorem (see [8]). In 2013, Xiaoyan [10] showed that equation (1.1) has no solution with $n > 2$ and $2 \mid n$ when $\vartheta_2(a-1)$ and $\vartheta_2(b-1)$ have the opposite parity. In 2016, Ishii [4] gave a necessary and sufficient condition for equation (1.1) with the conditions ($a \equiv 5 \pmod 6$ and $b \equiv 0 \pmod 3$) to have positive integer solutions. Theorem 1.4 and Theorem 1.5 give a necessary sufficient condition for Diophantine equation (1.1) with ($a \equiv 4 \pmod 5$ and $b \equiv 0 \pmod 5$) or ($a \equiv 3 \pmod 4$ and $b \equiv 0 \pmod 2$) to have positive integer solutions with $n \geq 2$. In 2018, Keskin [5] showed that equation (1.1) has no solution in positive integer with $2 \mid n$ when a and b have the opposite parity. Recently, the authors of [8] showed that equation (1.1) has no solution in positive integer when a is even and $b \equiv 3 \pmod 8$ with b a prime number.

In this paper, we will show that equation (1.1) has no positive solution (n,x) under some contraints on a and b. Our main results are the following.

Theorem 1.1. *Let $a,\ b \in \mathbb{N}$ such that $a,\ b > 1$. Suppose that one of the following conditions is satisfied:*

- *$a \equiv 0 \pmod 2$ and $b \equiv 3 \pmod{12}$;*
- *a is even, $\vartheta_2(b-1) = 1$ and $5 \mid b$.*

Then, equation (1.1) has no solution in positive integers (n,x).

This result generalizes the main result of Szalay [9]. A consequence of the above theorem is the following result.

Corollary 1.2. *Let $b \in \{15,35,55,75,95,3,27,39,51,63,87,99\}$. Then the equation*

$$((2k)^n - 1)(b^n - 1) = x^2$$

has no solution in positive integers (k,n,x).

Theorem 1.3. *Let* $a,\ b\in\mathbb{N}$ *such that* $a,\ b>1$. *Suppose that* $a\equiv 4 \pmod{10}$ *and* $b\equiv 0 \pmod 5$. *Then, equation* (1.1) *has no solution in positive integers* (n,x).

Theorem 1.4. *Suppose that* $a\equiv 4 \pmod 5$ *and* $b\equiv 0 \pmod 5$. *Then equation* (1.1) *has a positive integer solution* (n,x) *if and only if* $(a,b)=(u_r,u_s)$ *with non-square* $d\equiv \pm 1 \pmod 5$ *satisfying* $u_1\equiv 0 \pmod 5, r\equiv 2 \pmod 4$ *and* s *is odd. In this case, the solution is* $(x,n)=(dv_rv_s,2)$.

Theorem 1.5. *Suppose that* $a\equiv 3 \pmod 4$ *and* $b\equiv 0 \pmod 2$. *Then equation* (1.1) *has positive integer solutions* (n,x) *if and only if* $(a,b)=(u_r,u_s)$ *with non-square* $d\equiv 3 \pmod 4$ *satisfying* $u_1\equiv 0 \pmod 2, r\equiv 2 \pmod 4$ *and* s *is odd. In this case, the solution is* $(x,n)=(dv_rv_s,2)$.

Remark 1.6. Using the Theorem 1.4 and the fact that there exist $d\equiv \pm 1 \pmod 5$ with $u_1\equiv 0 \pmod 5$. For example, $u_1=2543295$ for $d=94$. We deduce that

$$(a^2-1)(b^2-1)=x^2$$

has infinitely many solutions (a,b,x) with $a\equiv 4 \pmod 5$ and $b\equiv 0 \pmod 5$.

The proof of the first theorem using the method in [8]. We organize this paper as follows. To prove the above results, we need some results on divisibility properties of the solutions of Pell equations and some known results. See Section 2. The proof of Theorem 1.1 is done in Section 3. We prove Theorem 1.3 in Section 4 and the proof of Theorem 1.4 in Section 5. For similar reason, the proof of Theorem 1.5 is also left to the reader.

2. Preliminaries

In this section, we recall some results which will be very useful for the proofs.

Let d be a positive integer which is not a square. Then, by the theory of Pell equations, one knows that the equation

$$u^2-dv^2=1,\ u,v\in\mathbb{N}^+$$

has infinitely many solutions and all its positive solutions (u,v) are given by

$$u_n+v_n\sqrt{d}=(u_1+v_1\sqrt{d})^n,$$

for some positive integer n, where (u_1,v_1) is the smallest positive solution.

The following result is well-known. As a reference (see [6, Lemma 1].

Lemma 2.1. *Let* d *be a positive which is not square.*

1. *If* k *is even, then each prime factor* p *of* u_k *satisfies* $p\equiv \pm 1 \pmod 8$.

2. *If k is odd, then $u_1 \mid u_k$ and u_k/u_1 is odd.*

3. *If $q \in \{2, 3, 5\}$, then $q \mid u_k$ implies $q \mid u_1$.*

The following lemma can be deduced from [1, Proposition 1].

Lemma 2.2. *Let $p > 3$ be a prime. Then, the equation*

$$x^p = 2y^2 - 1, \quad x, y \in \mathbb{N}$$

has the only solution $(x, y) = (1, 1)$ in positive integers and the equation

$$x^3 = 2y^2 - 1, \quad x, y \in \mathbb{N}$$

has the only solutions $(x, y) = (1, 1), (23, 78)$ in positive integers.

The last result to recall is [10, Lemma 2.1].

Lemma 2.3. *For a fixed d, if $2 \mid u_r$ and $2 \nmid u_s$, then $2 \nmid r$ and $2 \mid s$.*

3. Proof of Theorem 1.1

We prove only the first part of the statement, the proofs of the other part is similar and left to the reader. Suppose that equation (1.1) has a solution in positive integer n, x with $a \equiv 0 \pmod 2$ and $b \equiv 3 \pmod{12}$. Then we have

$$a^n - 1 = Dy^2 \quad \textit{and} \quad b^n - 1 = Dz^2,$$

where $D = (a^n - 1, b^n - 1)$. D can be written as $D = dw^2$, with a square-free integer d. If $d = 1$, then n must be odd. Indeed, if n is even, then we obtain $(a^m)^2 - (yw)^2 = 1$ with $n = 2m$ and yw integers. This is impossible. So n is odd. As

$$b^n - 1 = (b - 1)(b^{n-1} + \cdots + b + 1) \tag{3.1}$$

and $2 \mid (b^n - 1)$, it follows that $2 \mid z^2$. This implies that $2 \mid z$. Hence, $4 \mid (b^n - 1)$, which is a contradiction to equation (3.1) (as $\vartheta_2(b-1) = 1$). So $d \geq 2$ and D is not square. Using the equation $a^n - 1 = Dy^2$ and the fact that a is even, we deduce that D is odd. Moreover, $2 \mid z^2$ by the equation $b^n - 1 = Dz^2$. This implies that $4 \mid (b^n - 1)$ and by equation (3.1) we conclude that n is even. Put now $n = 2m$, we obtain

$$(a^m)^2 - Dy^2 = 1 \quad \textit{and} \quad (b^m)^2 - Dz^2 = 1.$$

The pairs

$$\{(a^m, y), (b^m, z)\}$$

are two solutions of the corresponding Pell equation $u^2 - Dv^2 = 1$. So there exist distinct positive integers r and s such that

$$(a^m, y) = (u_r, v_r) \text{ and } (b^m, z) = (u_s, v_s),$$

where (u_1, v_1) is the fundamental solution of this Pell equation. Since $3 \mid b$ and 3 not congruent to ± 1 modulo 8. Lemma 2.1 tells us that s is odd. As $2 \mid a$ and $2 \nmid b$, it follows (by Lemma 2.3) that r is odd and s is even. This contradicts the fact that s is odd and thus completes the proof of Theorem 1.1.

4. Proof of Theorem 1.3

The aim of this section is to prove Theorem 1.3. Thus, let $a \equiv 4 \pmod{10}$ and $b \equiv 0 \pmod 5$. Suppose that (n, x) is a solution to equation (1.1). Put $D = (a^n - 1, b^n - 1)$. By this equation, we have

$$a^n - 1 = Dy^2,\ b^n - 1 = Dz^2,\ x = Dyz,\ D, y, z \in \mathbb{N}.$$

Since $5 \mid b$, by $b^n - 1 = Dz^2$, it follows that

$$D \equiv \pm 1 \pmod 5 \text{ and } 5 \nmid z.$$

Now, we consider two cases according to the fact that 5 divides y or not.

Case 1: Suppose that $5 \nmid y$. Then $y^2 \equiv \pm 1 \pmod 5$ and we get

$$a^n \equiv Dy^2 + 1 \equiv \pm D + 1 \equiv 0, 2 \pmod 5.$$

This contradicts the fact that $a \equiv 4 \pmod 5$.

Case 2: Assume now that $5 \mid y$. Since $a \equiv 4 \pmod 5$, by $a^n - 1 = Dy^2$, we obtain

$$4^n \equiv a^n \equiv Dy^2 + 1 \equiv 1 \pmod 5.$$

We deduce that n is even. Put $n = 2m$. Therefore, D cannot be a square and the pairs

$$\{(a^m, y), (b^m, z)\}$$

are two solutions of the corresponding Pell equation $u^2 - Dv^2 = 1$. Since $a \neq b$, there exist distinct positive integers r and s such that

$$(a^m, y) = (u_r, v_r) \text{ and } (b^m, z) = (u_s, v_s),$$

where (u_1, v_1) is the fundamental solution of this Pell equation. By Lemma 2.1 and as $5 \mid b$ and $b^m = u_s$, one can see that $2 \nmid s$ and $5 \mid u_1$. Therefore, $2 \mid a$ and $a^m = u_r$ implies that $2 \mid u_r$ and so r is odd. By Lemma 2.1 (2), it follows that $u_1 \mid u_r$. For above, we deduce that $5 \mid u_r$ and thus $5 \mid a$, which contradicts the fact that $a \equiv 4 \pmod 5$. This completes our proof.

5. Proof of Theorem 1.4

In this section, we will prove Theorem 1.4. Let $a \equiv 4 \pmod 5$ and $b \equiv 0 \pmod 5$ and suppose that (n, x) is a solution to equation (1.1). Put $D = (a^n - 1, b^n - 1)$. By this equation, we have

$$a^n - 1 = Dy^2,\ b^n - 1 = Dz^2,\ x = Dyz,\ D, y, z \in \mathbb{N}.$$

We similarly proceed as in the proof of Theorem 1.3 and obtain that n is even. Put $n = 2m$. Therefore, D cannot be a square and the corresponding Pell equation $u^2 - Dv^2 = 1$ has two solutions

$$(u, v) = (a^m, y), (b^m, z).$$

Since $a \neq b$, there exist distinct positive integers r and s such that

$$(a^m, y) = (u_r, v_r) \quad and \quad (b^m, z) = (u_s, v_s),$$

where (u_1, u_1) is the fundamental solution of this Pell equation. By Lemma 8 (1) and $5 \mid b$, we obtain that $2 \nmid s$ and $5 \mid u_1$. On the other hand, $a \equiv 4 \pmod 5$, which together with $5 \mid u_1$ and Lemma 8 (2), shows that $2 \mid r$. Put $r = 2t$, we get

$$a^m = u_{2t} = 2u_t^2 - 1.$$

Now we distinguish two cases. Firstly, if $2 \mid m$, then $4 \mid n$ and so RESULT 2 in [2] implies that $(a, b) = (13, 239)$, with contradicts $5 \mid b$. Now, we assume that $2 \nmid m$ and $m > 3$, Lemma 9 shows that we have a contradiction since $a > 1$. If $m = 3$, then we get $a^3 = 2u_t^2 - 1$ and by Lemma 9, we obtain $a = 23$ and $u_t = 78$, which contradicts the fact that $a \equiv 4 \pmod 5$. So $m = 1$, then $n = 2m = 2$. Now suppose that $r \equiv 0 \pmod 4$. Then t is even and hence u_t not congruent to 0 modulo 5 by Lemma 8 (1). Then $a = u_r = 2u_t^2 - 1 \equiv 1, 2 \pmod 5$, which contradicts that $a \equiv 4 \pmod 5$. Conversely, suppose that $(a, b) = (u_r, u_s)$ with $d \equiv \pm 1 \pmod 5, u_1 \equiv 0 \pmod 5, r \equiv 2 \pmod 4$ and s is odd. Therefore, equation (1.1) has the solution $(x, n) = (dv_r v_s, 2)$. Notice that $b \equiv u_t \equiv 0 \pmod 5$ by Lemma 8 (2) and hence $a = u_r = 2u_t^2 - 1 \equiv 4 \pmod 5$. This completes the proof of Theorem 1.4.

Acknowledgements. The authors are grateful to the anonymous referee's comments that lead to a more precise version of this paper.

References

[1] M. A. Bennett, C. M. Skinner: *Ternary Diophantine equation via Galois representations and modular forms*, Canad. J. Math 56 (2004), pp. 23–54, doi: 10.4153/cjm-2004-002-2.

[2] J. H. E. Cohn: *The Diophantine equation* $(a^n - 1)(b^n - 1) = x^2$, Period. Math. Hungar. 44.2 (2002), pp. 169–175, doi: 10.1023/a:1019688312555.

[3] L. Hajdu, L. Szalay: *On the Diophantine equation* $(2^n - 1)(6^n - 1) = x^2$ *and* $(a^n - 1)(a^{kn} - 1) = x^2$, Period. Math. Hungar. 40.2 (2000), pp. 141–145, doi: 10.1023/a:1010335509489.

[4] K. Ishii: *On the exponential Diophantine equation* $(a^n - 1)(b^n - 1) = x^2$, Publ. Math. Debrecen 89.1-2 (2016), pp. 253–256, doi: 10.5486/pmd.2016.7578.

[5] R. Keskin: *A Note On the Exponential Diophantine equation* $(a^n - 1)(b^n - 1) = x^2$, arXiv: 1801.04717v1.

[6] L. Lan, L. Szalay: *On the exponential Diophantine equation* $(a^n - 1)(b^n - 1) = x^2$, Publ. Math. Debrecen 77 (2010), pp. 1–6.

[7] F. Luca, P. G. Walsh: *The product of like-indexed terms in binary recurrences*, J. Number Theory 96.1 (2002), pp. 152–173, doi: 10.1016/s0022-314x(02)92794-0.

[8] A. Noubissie, A. Togbé, Z. Zhang: *On the Exponential Diophantine equation* $(a^n - 1)(b^n - 1) = x^2$, to appear in the Bulletin of the Belgian Mathematical Society – Simon Stevin.

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

[10] G. Xioyan: *A Note on the Diophantine equation* $(a^n - 1)(b^n - 1) = x^2$, Period. Math. Hungar. 66 (2013), pp. 87–93.

[11] P. Yuan, Z. Zhang: *On the Diophantine equation* $(a^n - 1)(b^n - 1) = x^2$, Publ. Math. Debrecen 80 (2012), pp. 327–331.

Annales Mathematicae et Informaticae
***50** (2019) pp. 167–177*
DOI: 10.33039/ami.2019.03.001
http://ami.uni-eszterhazy.hu

An exponential Diophantine equation related to the difference between powers of two consecutive Balancing numbers

Salah Eddine Rihane[a], Bernadette Faye[b], Florian Luca[c], Alain Togbé[d]

[a]Université des Sciences et de la Technologie Houari-Boumediène
Faculté de Mathématiques, Laboratoire d'Algèbre et Théorie des Nombres
Bab-Ezzouar Alger, Algérie
salahrihane@hotmail.fr

[b]Department of Mathematics, University Gaston Berger of Saint-Louis
Saint-Louis, Senegal
bernadette@aims-senegal.org

[c]School of Mathematics, University of the Witwatersrand, Johannesburg, South Africa
King Abdulaziz University, Jeddah, Saudi Arabia
Department of Mathematics, Faculty of Sciences
University of Ostrava, Ostrava, Czech Republic
Florian.Luca@wits.ac.za

[d]Department of Mathematics, Statistics, and Computer Science
Purdue University Northwest, Westville, USA
atogbe@pnw.edu

Submitted: November 6, 2018
Accepted: March 25, 2019
Published online: April 4, 2019

Abstract

In this paper, we find all solutions of the exponential Diophantine equation $B_{n+1}^x - B_n^x = B_m$ in positive integer variables (m, n, x), where B_k is the k-th term of the Balancing sequence.

Keywords: Balancing numbers, Linear form in logarithms, reduction method.

MSC: 11B39, 11J86

1. Introduction

The first definition of balancing numbers is essentially due to Finkelstein [3], although he called them numerical centers. A positive integer n is called a balancing number if

$$1 + 2 + \cdots + (n-1) = (n+1) + (n+2) + \cdots + (n+r)$$

holds for some positive integer r. Then r is called the *balancer* corresponding to the balancing number n. For example, 6 and 35 are balancing numbers with balancers 2 and 14, respectively. The n-th term of the sequence of balancing numbers is denoted by B_n. The balancing numbers satisfy the recurrence relation

$$B_n = 6B_{n-1} - B_{n-2}, \text{ for all } n \geq 2,$$

where the initial conditions are $B_0 = 0$ and $B_1 = 1$. Its first terms are

$$0, 1, 6, 35, 204, 1189, 6930, 40391, 235416, 1372105, \ldots$$

It is well-known that

$$B_{n+1}^2 - B_n^2 = B_{2n+2}, \text{ for any } n \geq 0.$$

In particular, this identity tells us that the difference between the square of two consecutive Balancing numbers is still a Balancing number. So, one can ask if this identity can be generalized?

Diophantine equations involving sum or difference of powers of two consecutive members of a given linear recurrent sequence $\{U_n\}_{n\geq 1}$ were also considered in several papers. For example, in [5], Marques and Togbé proved that if $s \geq 1$ an integer such that $F_m^s + F_{m+1}^s$ is a Fibonacci number for all sufficiently large m, then $s \in \{1, 2\}$. In [4], Luca and Oyono proved that there is no integer $s \geq 3$ such that the sum of sth powers of two consecutive Fibonacci numbers is a Fibonacci number. Later, their result has been extended in [8] to the generalized Fibonacci numbers and recently in [7] to the Pell sequence.

Here, we apply the same argument as in [4] to the Balancing sequence and prove the following:

Theorem 1.1. *The only nonnegative integer solutions* (m, n, x) *of the Diophantine equation*

$$B_{n+1}^x - B_n^x = B_m \tag{1.1}$$

are $(m, n, x) = (2n + 2, n, 2), (1, 0, x), (0, n, 0)$.

Our proof of Theorem 1.1 is mainly based on linear forms in logarithms of algebraic numbers and a reduction algorithm originally introduced by Baker and Davenport in [1]. Here, we will use a version due to Dujella and Pethő in [2, Lemma 5(a)].

2. Preliminary results

2.1. The Balancing sequences

Let $(\alpha, \beta) = (3 + 2\sqrt{2}, 3 - 2\sqrt{2})$ be the roots of the characteristic equation $x^2 - 6x + 1 = 0$ of the Balancing sequence $(B_n)_{n\geq 0}$. The Binet formula for B_n is

$$B_n = \frac{\alpha^n - \beta^n}{4\sqrt{2}}, \quad \text{for all } n \geq 0. \tag{2.1}$$

This implies that the inequality

$$\alpha^{n-2} \leq B_n \leq \alpha^{n-1} \tag{2.2}$$

holds for all positive integers n. It is easy to prove that

$$\frac{B_n}{B_{n+1}} \leq \frac{5}{29} \tag{2.3}$$

holds, for any $n \geq 2$.

2.2. Linear forms in logarithms

For any non-zero algebraic number γ of degree d over $\mathbb{Q}$, whose minimal polynomial over $\mathbb{Z}$ is $a\prod_{i=1}^{d}\left(X - \gamma^{(i)}\right)$, we denote by

$$h(\gamma) = \frac{1}{d}\left(\log|a| + \sum_{i=1}^{d}\log\max\left(1, \left|\gamma^{(i)}\right|\right)\right)$$

the usual absolute logarithmic height of γ.

With this notation, Matveev proved the following theorem (see [6]).

Theorem 2.1. *Let $\gamma_1, \ldots, \gamma_s$ be real algebraic numbers and let $b_1, \ldots, b_s$ be nonzero rational integer numbers. Let D be the degree of the number field $\mathbb{Q}(\gamma_1, \ldots, \gamma_s)$ over $\mathbb{Q}$ and let A_j be positive real numbers satisfying*

$$A_j = \max\{Dh(\gamma_j), |\log\gamma_j|, 0.16\}, \quad \textit{for} \quad j = 1, \ldots, s.$$

Assume that

$$B \geq \max\{|b_1|, \ldots, |b_s|\}.$$

If $\gamma_1^{b_1}\cdots\gamma_s^{b_s} - 1 \neq 0$, then

$$|\gamma_1^{b_1}\cdots\gamma_s^{b_s} - 1| \geq \exp(-1.4\cdot 30^{s+3}\cdot s^{4.5}\cdot D^2(1+\log D)(1+\log B)A_1\cdots A_s).$$

2.3. Reduction algorithm

Lemma 2.2. *Let M be a positive integer, let p/q be a convergent of the continued fraction expansion of the irrational γ such that $q > 6M$, and let A, B, μ be some real numbers with $A > 0$ and $B > 1$. Let*

$$\varepsilon = ||\mu q|| - M \cdot ||\gamma q||,$$

where $||\cdot||$ denotes the distance from the nearest integer. If $\varepsilon > 0$, then there is no solution of the inequality

$$0 < m\gamma - n + \mu < AB^{-k}$$

in positive integers m, n and k with

$$m \leq M \quad \text{and} \quad k \geq \frac{\log(Aq/\varepsilon)}{\log B}.$$

3. The proof of Theorem 1.1

3.1. An inequality for x versus m and n

The case $nx = 0$ is trivial so we assume that $n \geq 1$ and that $x \geq 1$. Observe that since $B_n < B_{n+1} - B_n < B_{n+1}$, the Diophantine equation (1.1) has no solution when $x = 1$.

When $n = 1$, we get $B_m = 6^x - 1$. In this case, we have that m is odd. Thus, using the Binet formula (2.1), we obtained the following factorization

$$6^x = B_m + 1 = B_m + B_1 = B_{(m+1)/2}C_{(m-1)/2},$$

where $\{C_m\}_{m\geq 1}$ is the Lucas Balancing sequence given by the recurrence $C_m = 6C_{m-1} - C_{m-2}$ with initial conditions $C_0 = 2$, $C_1 = 6$. The Binet formula of the Lucas Balancing sequence is given by $C_n = \alpha^n + \beta^n$. This shows that the largest prime factor of $B_{(m+1)/2}$ is 3 and by Carmichael's Primitive Divisor Theorem we conclude that $(m+1)/2 \leq 12$, so $m \leq 23$. Now, one checks all such m and gets no additional solution with $n = 1$.

So, we can assume that $n \geq 2$ and $x \geq 3$. Therefore, we have

$$B_m = B_{n+1}^x - B_n^x \geq B_3^3 - B_1^3 = 215,$$

which implies that $m > 4$. Here, we use the same argument from [4] to bound x in terms of m and n. Since most of the details are similar, we only sketch the argument.

Using inequality (2.2), we get

$$\alpha^{m-1} > B_m = B_{n+1}^x - B_n^x \geq B_n^x > \alpha^{(n-2)x}$$

and

$$\alpha^{m-2} < B_m = B_{n+1}^x - B_n^x < B_{n+1}^x < \alpha^{nx}.$$

Thus, we have

$$(n-2)x + 1 < m < nx + 2. \tag{3.1}$$

Estimate (3.1) is essential for our purpose.

Now, we rewrite equation (1.1) as

$$\frac{\alpha^m}{4\sqrt{2}} - B_{n+1}^x = -B_n^x + \frac{\beta^m}{4\sqrt{2}}. \tag{3.2}$$

Dividing both sides of equation (3.2) by B_{n+1}^x, taking absolute value and using the inequality (2.3), we obtain

$$\left|\alpha^m (4\sqrt{2})^{-1} B_{n+1}^{-x} - 1\right| < 2\left(\frac{B_n}{B_{n+1}}\right)^x < \frac{2}{5.8^x}. \tag{3.3}$$

Put

$$\Lambda_1 := \alpha^m (4\sqrt{2})^{-1} B_{n+1}^{-x} - 1. \tag{3.4}$$

If $\Lambda_1 = 0$, we get $\alpha^m = 4\sqrt{2}B_{n+1}^x$. Thus $\alpha^{2m} \in \mathbb{Z}$, which is false for all positive integers m, therefore $\Lambda_1 \neq 0$.

At this point, we will use Matveev's theorem to get a lower bound for Λ_1. We set $s := 3$ and we take

$$\gamma_1 := \alpha, \quad \gamma_2 := 4\sqrt{2}, \quad \gamma_3 := B_{n+1}, \quad b_1 := m, \quad b_2 := -1, \quad b_3 := -x.$$

Note that $\gamma_1, \gamma_2, \gamma_3 \in \mathbb{Q}(\sqrt{2})$, so we can take $D := 2$. Since $h(\gamma_1) = (\log \alpha)/2$, $h(\gamma_2) = (\log 32)/2$ and $h(\gamma_3) = \log B_{n+1} < n \log \alpha$, we can take $A_1 := \log \alpha$, $A_2 := \log 32$ and $A_3 := 2n \log \alpha$. Finally, inequality (3.1) implies that $m > (n-2)x \geq x$, thus we can take $B := m$. We also have $B := m \leq nx + 2 < (n+2)x$. Hence, Matveev's theorem implies that

$$\begin{aligned} \log|\Lambda_1| &\geq -1.4 \times 30^6 \times 3^{4.5} \times 2^2 \times (1 + \log 2)(\log \alpha)(\log 32)(2n \log \alpha)(1 + \log m) \\ &\geq -2.1 \times 10^{13} n(1 + \log m). \end{aligned} \tag{3.5}$$

The inequalities (3.3), (3.4) and (3.5) give that

$$x < 1.2 \times 10^{13} n(1 + \log m) < 2.1 \times 10^{13} n \log m,$$

where we used the fact that $1 + \log m < 1.7 \log m$, for all $m \geq 5$. Together with the fact that $m < (n+2)x$, we get that

$$x < 2.1 \times 10^{13} n \log((n+2)x).$$

3.2. Small values of n

Next, we treat the cases when $n \in [2, 37]$. In this case,

$$x < 2.1 \times 10^{13} n \log((n+2)x) < 7.8 \times 10^{14} \log(46x)$$

so $x < 4 \times 10^{16}$.

Now, we take another look at Λ_1 given by expression (3.4). Put

$$\Gamma_1 := m \log \alpha - \log(4\sqrt{2}) - x \log B_{n+1}.$$

Thus, $\Lambda_1 = e^{\Gamma_1} - 1$. One sees that the right-hand side of (3.2) is a number in the interval $[-B_n^x, -B_n^x + 1]$. In particular, Λ_1 is negative, which implies that Γ_1 is negative. Thus,

$$0 < -\Gamma_1 < \frac{2}{5.8^x},$$

so

$$0 < x \left(\frac{\log B_{n+1}}{\log \alpha} \right) - m + \left(\frac{\log(4\sqrt{2}}{\log \alpha} \right) < \frac{2}{5.8^x \log \alpha}. \tag{3.6}$$

For us, inequality (3.6) is

$$0 < x\gamma - m + \mu < AB^{-x},$$

where

$$\gamma := \frac{\log B_{n+1}}{\log \alpha}, \quad \mu = \frac{\log(4\sqrt{2})}{\log \alpha}, \quad A = \frac{2}{\log \alpha}, \quad B = 5.8.$$

We take $M := 4 \times 10^{16}$.

The program was developed in PARI/GP running with 200 digits. For the computations, if the first convergent such that $q > 6M$ does not satisfy the condition $\varepsilon > 0$, then we use the next convergent until we find the one that satisfies the condition. In one minute all the computations were done. In all cases, we obtained $x \leq 77$. A computer search with Maple revealed in less than one minute that there are no solutions to the equation (1.1) in the range $n \in [3, 37]$ and $x \in [3, 77]$.

3.3. An upper bound on x in terms of n

From now on, we assume that $n \geq 38$. Recall from the previous section that

$$x < 2.1 \times 10^{13} n \log((n+2)x). \tag{3.7}$$

Next, we give an upper bound on x depending only on n. If

$$x \leq n + 2, \tag{3.8}$$

then we are through. Otherwise, that is if $n + 2 < x$, we then have

$$x < 2.1 \times 10^{13} n \log x^2 = 4.2 \times 10^{13} n \log x,$$

which can be rewritten as

$$\frac{x}{\log x} < 4.2 \times 10^{13} n. \tag{3.9}$$

Using the fact that, for all $A \geq 3$

$$\frac{x}{\log x} < A \quad \text{yields} \quad x < 2A \log A,$$

and the fact that $\log(4.2 \times 10^{13} n) < 10 \log n$ holds for all $n \geq 38$, we get that

$$\begin{aligned} x &< 2(4.2 \times 10^{13} n) \log((4.2 \times 10^{13} n) \\ &< 8.4 \times 10^{13} n (10 \log n) \\ &< 8.4 \times 10^{14} n \log n. \end{aligned} \tag{3.10}$$

From (3.8) and (3.10), we conclude that the inequality

$$x < 8.4 \times 10^{14} n \log n \tag{3.11}$$

holds.

3.4. An absolute upper bound on x

Let us look at the element

$$y := \frac{x}{\alpha^{2n}}.$$

The above inequality (3.11) implies that

$$y < \frac{8.4 \times 10^{14} n \log n}{\alpha^{2n}} < \frac{1}{\alpha^n}, \tag{3.12}$$

where the last inequality holds for any $n \geq 23$. In particular, $y < \alpha^{-38} < 10^{-31}$. We now write

$$B_n^x = \frac{\alpha^{nx}}{32^{x/2}} \left(1 - \frac{1}{\alpha^{2n}}\right)^x$$

and

$$B_{n+1}^x = \frac{\alpha^{(n+1)x}}{32^{x/2}} \left(1 - \frac{1}{\alpha^{2(n+1)}}\right)^x.$$

We have

$$0 < \left(1 - \frac{1}{\alpha^{2n}}\right) < e^y < 1 + 2y,$$

because $y < 10^{-31}$ is very small. The same inequality holds if we replace n by $n+1$. Hence, we have that

$$\max\left\{\left|B_n^x - \frac{\alpha^{nx}}{32^{x/2}}\right|, \left|B_{n+1}^x - \frac{\alpha^{(n+1)x}}{32^{x/2}}\right|\right\} < \frac{2y\alpha^{(n+1)x}}{32^{x/2}}.$$

We now return to our equation (1.1) and rewrite it as

$$\begin{aligned}\frac{\alpha^m-\beta^m}{4\sqrt{2}} &= B_m = B^x_{n+1}-B^x_n \\ &= \frac{\alpha^{(n+1)x}}{32^{x/2}}-\frac{\alpha^{nx}}{32^{x/2}}+\left(B^x_{n+1}-\frac{\alpha^{(n+1)x}}{32^{x/2}}\right)-\left(B^x_n-\frac{\alpha^{nx}}{32^{x/2}}\right),\end{aligned}$$

or

$$\begin{aligned}\left|\frac{\alpha^m}{32^{1/2}}-\frac{\alpha^{nx}}{32^{x/2}}(\alpha^x-1)\right| &= \left|\frac{\beta^m}{32^{1/2}}+\left(B^x_{n+1}-\frac{\alpha^{(n+1)x}}{32^{x/2}}\right)-\left(B^x_n-\frac{\alpha^{nx}}{32^{x/2}}\right)\right| \\ &< \frac{1}{\alpha^m}+\left|B^x_{n+1}-\frac{\alpha^{(n+1)x}}{32^{x/2}}\right|+\left|B^x_n-\frac{\alpha^{nx}}{32^{x/2}}\right| \\ &< \frac{1}{\alpha^m}+2y\left(\frac{\alpha^{nx}(1+\alpha^x)}{32^{x/2}}\right).\end{aligned}$$

Thus, multiplying both sides by $\alpha^{-(n+1)x}32^{x/2}$, we obtain that

$$\begin{aligned}\left|\alpha^{m-(n+1)x}32^{(x-1)/2}-(1-\alpha^{-x})\right| &< \frac{32^{x/2}}{\alpha^{m+(n+1)x}}+2y(1+\alpha^{-x}) \\ &< \frac{1}{2\alpha^n}+\frac{396y}{197}<\frac{3}{\alpha^n},\end{aligned} \tag{3.13}$$

where we used the fact that $32^{x/2}/(\alpha^{(n+1)x}) \le (4\sqrt{2}/\alpha^{38})^x < 1/2$, $m \ge (n-2)x \ge n$ and $\alpha^x \ge \alpha^3 > 197$, as well as inequality (3.12). Hence, we conclude that

$$\left|\alpha^{m-(n+1)x}32^{(x-1)/2}-1\right| < \frac{1}{\alpha^x}+\frac{3}{\alpha^n} \le \frac{4}{\alpha^l}, \tag{3.14}$$

where $l := \min\{n, x\}$. We now set

$$\Lambda_2 := \alpha^{m-(n+1)x}32^{(x-1)/2}-1 \tag{3.15}$$

and observe that $\Lambda_2 \neq 0$. Indeed, for if $\Lambda_2 = 0$, then $\alpha^{2((n+1)x-m)} = 32^{x-1} \in \mathbb{Z}$ which is possible only when $(n+1)x = m$. But if this were so, then we would get $0 = \Lambda_2 = 32^{(x-1)/2}-1$, which leads to the conclusion that $x = 1$, which is not possible. Hence, $\Lambda_2 \neq 0$. Next, let us notice that since $x \ge 3$ and $m \ge 38$, we have that

$$|\Lambda_2| \le \frac{1}{\alpha^3}+\frac{1}{\alpha^{38}} < \frac{1}{2}, \tag{3.16}$$

so that $\alpha^{m-(n+1)x}32^{(x-1)/2} \in [1/2, 3/2]$. In particular,

$$(n+1)x-m < \frac{1}{\log\alpha}\left(\frac{(x-1)\log 32}{2}+\log 2\right) < x\left(\frac{\log 32}{2\log\alpha}\right) < x \tag{3.17}$$

and

$$(n+1)x-m > \frac{1}{\log\alpha}\left(\frac{(x-1)\log 32}{2}-\log 2\right) > 0.9x-1.4 > 0. \tag{3.18}$$

We lower bound the left-hand side of inequality (3.15) using again Matveev's theorem. We take

$$s := 2,\ \gamma_1 := \alpha,\ \gamma_2 := 4\sqrt{2},\ b_1 := m - (n+1)x,\ b_2 := x - 1,$$

$$D := 2,\ A_1 := \log\alpha,\ A_2 := \log 32,\ \text{ and } B := x.$$

We thus get that

$$\log|\Lambda_2| > -1.4 \times 30^5 \times 2^{4.5} \times 2^2(1+\log 2)(\log\alpha)(\log 32)(1+\log x). \tag{3.19}$$

The inequalities (3.14) and (3.19) give

$$l < 4 \times 10^{10} \log x.$$

Treating separately the case $l = x$ and the case $l = n$, following the argument in [4] we have that the upper bound

$$x < 7 \times 10^{28}$$

always holds.

3.5. Reducing the bound on x

Next, we take

$$\Gamma_2 := (x-1)\log(4\sqrt{2}) - ((n+1)x - m)\log\alpha.$$

Observe that $\Lambda_2 = e^{\Gamma_2} - 1$, where Λ_2 is given by (3.15). Since $|\Lambda_2| < \frac{1}{2}$, we have that $e^{|\Gamma_2|} < 2$. Hence,

$$|\Gamma_2| \le e^{|\Gamma_2|}\left|e^{\Gamma_2} - 1\right| < 2\,|\Lambda_2| < \frac{2}{\alpha^x} + \frac{6}{\alpha^n}.$$

This leads to

$$\left|\frac{\log(4\sqrt{2})}{\log\alpha} - \frac{(n+1)x - m}{x-1}\right| < \frac{1}{(x-1)\log\alpha}\left(\frac{2}{\alpha^x} + \frac{6}{\alpha^n}\right). \tag{3.20}$$

Assume next that $x > 100$. Then $\alpha^x > \alpha^{100} > 10^{33} > 10^4 x$. Hence, we get that

$$\frac{1}{(x-1)\log\alpha}\left(\frac{2}{\alpha^x} + \frac{6}{\alpha^n}\right) < \frac{8}{x(x-1)10^4\log\alpha} < \frac{1}{2200(x-1)^2}. \tag{3.21}$$

Estimates (3.20) and (3.21) lead to

$$\left|\frac{\log(4\sqrt{2})}{\log\alpha} - \frac{(n+1)x - m}{x-1}\right| < \frac{1}{2200(x-1)^2}. \tag{3.22}$$

By a criterion of Legendre, inequality (3.22) implies that the rational number $((n+1)x-m)/(x-1)$ is a convergent to $\gamma := \log(4\sqrt{2})/\log\alpha$. Let

$$[a_0, a_1, a_2, a_3, a_4, a_5, a_6, \ldots] = [0, 1, 57, 1, 234, 2, 1, \ldots]$$

be the continued fraction of γ, and let p_k/q_k be it's kth convergent. Assume that $((n+1)x-m)/(x-1) = p_k/q_k$ for some k. Then, $x-1 = dq_k$ for some positive integer d, which in fact is the greatest common divisor of $(n+1)x-m$ and $x-1$. We have the inequality

$$q_{54} > 7 \times 10^{28} > x-1.$$

Thus, $k \in \{0, \ldots, 53\}$. Furthermore, $a_k \leq 234$ for all $k = 0, 1, \ldots, 53$. From the known properties of the continued fraction, we have that

$$\left|\gamma - \frac{(n+1)x-m}{x-1}\right| = \left|\gamma - \frac{p_k}{q_k}\right| > \frac{1}{(a_k+2)q_k^2} \geq \frac{d^2}{236(x-1)^2} \geq \frac{1}{236(x-1)^2},$$

which contradicts inequality (3.22). Hence, $x \leq 100$.

3.6. The final step

To finish, we go back to inequality (3.13) and rewrite it as

$$\left|\alpha^{m-(n+1)x} 32^{(x-1)/2}(1-\alpha^{-x})^{-1} - 1\right| < \frac{3}{\alpha^n(1-\alpha^{-x})} < \frac{4}{\alpha^n}.$$

Recall that $x \in [3, 100]$ and from inequalities (3.17) and (3.18), we have that

$$0.9x - 1.4 < (n+1)x - m < x.$$

Put $t := (n+1)x-m$. We computed all the numbers $\left|\alpha^{-t} 32^{(x-1)/2}(1+\alpha^{-x})^{-1} - 1\right|$ for all $x \in [3, 100]$ and all $t \in [\lfloor 0.9x - 1.4\rfloor, \lfloor x \rfloor]$. None of them ended up being zero and the smallest of these numbers is $> 10^{-1}$. Thus, $1/10 < 3/\alpha^n$, or $\alpha^n < 30$, so $n \leq 3$ which is false.

Acknowledgements. We thank the referee for comments which improved the quality of this manuscript.

F. L. was supported in part by grant CPRR160325161141 and an A-rated scientist award both from the NRF of South Africa and by grant no. 17-02804S of the Czech Granting Agency.

B. F. worked on this paper during her visit to Purdue University Northwest, USA. She thanks the institution for the hospitality. She was also partially supported by a grant from the Simons Foundation.

A. T. was supported in part by Purdue University Northwest.

References

[1] A. Baker, H. Davenport: *The equations* $3x^2 - 2 = y^2$ *and* $8x^2 - 7 = z^2$, The Quarterly Journal of Mathematics 20.1 (1969), pp. 129–137, doi: 10.1093/qmath/20.1.129.

[2] A. Dujella, A. Pethő: *A generalization of a theorem of Baker and Davenport*, Quart. J. Math. Oxford Ser. (2) 49.195 (1998), pp. 291–306, doi: 10.1093/qmathj/49.3.291.

[3] R. P. Finkelstein: *The house problem*, American Math. Monthly 72 (1965), pp. 1082–1088, doi: 10.2307/2315953.

[4] F. Luca, R. Oyono: *An exponential Diophantine equation related to powers of two consecutive Fibonacci numbers*, Proc. Japan Acad. Ser. A 87 (2011), pp. 45–50, doi: 10.3792/pjaa.87.45.

[5] D. Marques, A. Togbé: *On the sum of powers of two consecutive Fibonacci numbers*, Proc. Japan Acad. Ser. A 86 (2010), pp. 174–176, doi: 10.3792/pjaa.86.174.

[6] E. M. Matveev: *An explicit lower bound for a homogeneous rational linear form in the logarithms of algebraic numbers, II*, Izv. Math. 64.6 (2000), pp. 1217–1269, doi: 10.1070/im2000v064n06abeh000314.

[7] S. E. Rihane, B. Faye, F. Luca, A. Togbé: *On the exponential Diophantine equation* $P_n^x + P_{n+1}^x = P_m$, To appear in Turkish Journal of Mathematics.

[8] C. A. G. Ruiz, F. Luca: *An exponential Diophantine equation related to the sum of powers of two consecutive k-generalized Fibonacci numbers*, Coll. Math. 137.2 (2014), pp. 171–188, doi: 10.4064/cm137-2-3.

Annales Mathematicae et Informaticae
***50** (2019) pp. 179–185*
DOI: 10.33039/ami.2019.12.001
http://ami.uni-eszterhazy.hu

A Marcinkiewicz–Zygmund type strong law of large numbers for non-negative random variables with multidimensional indices

Tibor Tómács*

Institute of Mathematics and Informatics
Eszterházy Károly University, Eger, Hungary
tomacs.tibor@uni-eszterhazy.hu

Submitted: September 2, 2019
Accepted: December 4, 2019
Published online: December 5, 2019

Abstract

In this paper a Marcinkiewicz–Zygmund type strong law of large numbers is proved for non-negative random variables with multidimensional indices, furthermore we give its an application for multi-index sequence of non-negative random variables with finite variances.

Keywords: Marcinkiewicz–Zygmund type strong law of large numbers, almost sure convergence, non-negative random variables, multidimensional indices

MSC: 60F15

1. Introduction

The Kolmogorov theorem and the Marcinkiewicz–Zygmund theorem are two famous theorems on the strong law of large numbers for X_n $(n \in \mathbb{N})$ sequence of independent identically distributed random variables (see e.g. Loève [8]). By Kolmogorov theorem, there exists a constant b such that $\lim_{n\to\infty} S_n/n = b$ almost surely if and only if $\mathrm{E}\,|X_1| < \infty$, where $S_n = \sum_{k=1}^{n} X_k$. If the latter condition is satisfied then $b = \mathrm{E}\,X_1$. By Marcinkiewicz–Zygmund theorem, if $0 < r < 2$ then

*The author's research was supported by the grant EFOP-3.6.1-16-2016-00001 ("Complex improvement of research capacities and services at Eszterhazy Karoly University").

$\lim_{n\to\infty}(S_n - bn)/n^{1/r} = 0$ almost surely if and only if $\mathrm{E}\,|X_1|^r < \infty$, where $b = 0$ if $0 < r < 1$, and $b = \mathrm{E}\,X_1$ if $1 \le r < 2$.

Etemadi [1] proved that the Kolmogorov theorem holds for identically distributed and pairwise independent random variables, furthermore Kruglov [7] extended the Marcinkiewicz–Zygmund theorem for pairwise independent case if $r < 1$.

Several papers are devoted to the study of the strong law of large numbers for multi-index sequence of random variables (see e.g. Gut [4], Klesov [5, 6], Fazekas [2], Fazekas, Tómács [3]). For example, Theorem 3.1 of Fazekas, Tómács [3] extends Theorem 2 of Kruglov [7] for multi-index case.

In this paper the main result is Theorem 3.1, which is a Marcinkiewicz–Zygmund type strong law of large numbers for non-negative random variables with multidimensional indices. It is a generalization of Theorem 3.1 of Fazekas, Tómács [3] in case $\mathbf{n} \to \infty$. Furthermore we give an application (see Theorem 4.1) for multi-index sequence of non-negative random variables with finite variances. A special case of this result gives Theorem of Petrov [9].

2. Notation

Let $\mathbb{N}^d$ be the positive integer d-dimensional lattice points, where d is a positive integer. For $\mathbf{n}, \mathbf{m} \in \mathbb{N}^d$, $\mathbf{n} \le \mathbf{m}$ is defined coordinate-wise, $(\mathbf{n}, \mathbf{m}] = (n_1, m_1] \times (n_2, m_2] \times \cdots \times (n_d, m_d]$ is a d-dimensional rectangle and $|\mathbf{n}| = n_1 n_2 \cdots n_d$, where $\mathbf{n} = (n_1, n_2, \ldots, n_d)$, $\mathbf{m} = (m_1, m_2, \ldots, m_d)$. $\sum_{\mathbf{n}}$ will denote the summation for all $\mathbf{n} \in \mathbb{N}^d$. We also use $\mathbf{1} = (1, 1, \ldots, 1) \in \mathbb{N}^d$ and $\mathbf{2} = (2, 2, \ldots, 2) \in \mathbb{N}^d$. Denote the integer part of x real number by $[x]$.

We shall say that $\lim_{\mathbf{n}\to\infty} a_{\mathbf{n}} = 0$, where $a_{\mathbf{n}}$ $(\mathbf{n} \in \mathbb{N}^d)$ are real numbers, if for all $\delta > 0$ there exists $\mathbf{N} \in \mathbb{N}^d$ such that $|a_{\mathbf{n}}| < \delta$ $\forall \mathbf{n} \ge \mathbf{N}$.

We shall assume that random variables $X_{\mathbf{n}}$ $(\mathbf{n} \in \mathbb{N}^d)$ are defined on the same probability space $(\Omega, \mathcal{F}, \mathrm{P})$. E and Var stand for the expectation and the variance.

Remark that a sum or a minimum over the empty set will be interpreted as zero (i.e. $\sum_{\mathbf{n}\in H} a_{\mathbf{n}} = \min_{\mathbf{n}\in H} a_{\mathbf{n}} = 0$ if $H = \emptyset$).

3. The result

The following result is a generalization of Theorem 3.1 of Fazekas, Tómács [3] in case $\mathbf{n} \to \infty$.

Theorem 3.1. *Let $X_{\mathbf{n}}$ $(\mathbf{n} \in \mathbb{N}^d)$ be a sequence of non-negative random variables, let $b_{\mathbf{n}}$ $(\mathbf{n} \in \mathbb{N}^d)$ be a sequence of non-negative numbers, $B_{\mathbf{n}} = \sum_{\mathbf{k}\le\mathbf{n}} b_{\mathbf{k}}$, $S_{\mathbf{n}} = \sum_{\mathbf{k}\le\mathbf{n}} X_{\mathbf{k}}$, $c > 0$, $K \in \mathbb{N}$ and $0 < r \le 1$. If*

$$B_{\mathbf{n}} - B_{\mathbf{m}} \le c(|\mathbf{n}| - |\mathbf{m}|) \quad \forall \mathbf{n}, \mathbf{m} \in \mathbb{N}^d, \mathbf{n} \ge \mathbf{m}, |\mathbf{n}| - |\mathbf{m}| \ge K \tag{3.1}$$

and

$$\sum_{\mathbf{n}} \frac{1}{|\mathbf{n}|} \mathrm{P}\left(|S_{\mathbf{n}} - B_{\mathbf{n}}| > \varepsilon |\mathbf{n}|^{1/r}\right) < \infty \quad \forall \varepsilon > 0, \tag{3.2}$$

then

$$\lim_{\mathbf{n}\to\infty} \frac{S_{\mathbf{n}} - B_{\mathbf{n}}}{|\mathbf{n}|^{1/r}} = 0 \quad \textit{almost surely.}$$

Proof. Let $\delta > 0$, $1 < \alpha < \left(\frac{\delta}{2c} + 1\right)^{1/3d}$ and $0 < \varepsilon < \frac{\delta}{2}\left(\frac{\delta}{2c} + 1\right)^{-1/r}$, which imply

$$\varepsilon \alpha^{3d/r} + c(\alpha^{3d} - 1) < \delta. \tag{3.3}$$

Let $k_n = [\alpha^n]$ $(n \in \mathbb{N})$ and $\mathbf{k_n} = (k_{n_1}, k_{n_2}, \ldots, k_{n_d})$, where $\mathbf{n} = (n_1, n_2, \ldots, n_d) \in \mathbb{N}^d$. It follows from the inequalities

$$\begin{aligned}
&\sum_{\mathbf{n}} \frac{1}{|\mathbf{n}|} \mathrm{P}\left(|S_{\mathbf{n}} - B_{\mathbf{n}}| > \varepsilon |\mathbf{n}|^{1/r}\right) \\
&\geq \sum_{\mathbf{n}} \sum_{\mathbf{h}\in(\mathbf{k_n},\mathbf{k_{n+1}}]} \frac{1}{|\mathbf{h}|} \mathrm{P}\left(|S_{\mathbf{h}} - B_{\mathbf{h}}| > \varepsilon |\mathbf{h}|^{1/r}\right) \\
&\geq \sum_{\mathbf{n}} \sum_{\mathbf{h}\in(\mathbf{k_n},\mathbf{k_{n+1}}]} \frac{1}{|\mathbf{k_{n+1}}|} \min_{\mathbf{k}\in(\mathbf{k_n},\mathbf{k_{n+1}}]} \mathrm{P}\left(|S_{\mathbf{k}} - B_{\mathbf{k}}| > \varepsilon |\mathbf{k}|^{1/r}\right) \\
&= \sum_{\mathbf{n}} \frac{|\mathbf{k_{n+1}} - \mathbf{k_n}|}{|\mathbf{k_{n+1}}|} \min_{\mathbf{k}\in(\mathbf{k_n},\mathbf{k_{n+1}}]} \mathrm{P}\left(|S_{\mathbf{k}} - B_{\mathbf{k}}| > \varepsilon |\mathbf{k}|^{1/r}\right)
\end{aligned}$$

and condition (3.2) that

$$\sum_{\mathbf{n}} \frac{|\mathbf{k_{n+1}} - \mathbf{k_n}|}{|\mathbf{k_{n+1}}|} \min_{\mathbf{k}\in(\mathbf{k_n},\mathbf{k_{n+1}}]} \mathrm{P}\left(|S_{\mathbf{k}} - B_{\mathbf{k}}| > \varepsilon |\mathbf{k}|^{1/r}\right) < \infty. \tag{3.4}$$

Since $\lim_{n\to\infty}\left(1 - \frac{1}{\alpha^{n+1}} - \frac{1}{\alpha}\right) = 1 - \frac{1}{\alpha} > 0$, so $\left(1 - \frac{1}{\alpha^{n+1}} - \frac{1}{\alpha}\right) > \frac{\alpha-1}{2\alpha}$ except for finitely many $n \in \mathbb{N}$. This implies that there exists $\mathbf{N}_0 \in \mathbb{N}^d$ such that

$$\begin{aligned}
0 < \left(\frac{\alpha-1}{2\alpha}\right)^d &< \prod_{i=1}^{d}\left(1 - \frac{1}{\alpha^{n_i+1}} - \frac{1}{\alpha}\right) = \prod_{i=1}^{d} \frac{\alpha^{n_i+1} - 1 - \alpha^{n_i}}{\alpha^{n_i+1}} \\
&\leq \prod_{i=1}^{d} \frac{[\alpha^{n_i+1}] - [\alpha^{n_i}]}{[\alpha^{n_i+1}]} = \frac{|\mathbf{k_{n+1}} - \mathbf{k_n}|}{|\mathbf{k_{n+1}}|} \quad \forall \mathbf{n} = (n_1, n_2, \ldots, n_d) \geq \mathbf{N}_0.
\end{aligned}$$

Hence

$$\begin{aligned}
&\left(\frac{\alpha-1}{2\alpha}\right)^d \sum_{\mathbf{n}\geq\mathbf{N}_0} \min_{\mathbf{k}\in(\mathbf{k_n},\mathbf{k_{n+1}}]} \mathrm{P}\left(|S_{\mathbf{k}} - B_{\mathbf{k}}| > \varepsilon |\mathbf{k}|^{1/r}\right) \\
&\leq \sum_{\mathbf{n}\geq\mathbf{N}_0} \frac{|\mathbf{k_{n+1}} - \mathbf{k_n}|}{|\mathbf{k_{n+1}}|} \min_{\mathbf{k}\in(\mathbf{k_n},\mathbf{k_{n+1}}]} \mathrm{P}\left(|S_{\mathbf{k}} - B_{\mathbf{k}}| > \varepsilon |\mathbf{k}|^{1/r}\right).
\end{aligned}$$

By this inequality and (3.4), it follows that

$$\sum_{\mathbf{n}\geq\mathbf{N}_0} \min_{\mathbf{k}\in(\mathbf{k_n},\mathbf{k_{n+1}}]} \mathrm{P}\left(|S_{\mathbf{k}} - B_{\mathbf{k}}| > \varepsilon|\mathbf{k}|^{1/r}\right) < \infty. \tag{3.5}$$

If $\mathbf{n} \geq \mathbf{N}_0$ then there exists $\mathbf{m_n} \in \mathbb{N}^d$ such that $\mathbf{m_n} \in (\mathbf{k_n}, \mathbf{k_{n+1}}]$ and

$$\mathrm{P}\left(|S_{\mathbf{m_n}} - B_{\mathbf{m_n}}| > \varepsilon|\mathbf{m_n}|^{1/r}\right) = \min_{\mathbf{k}\in(\mathbf{k_n},\mathbf{k_{n+1}}]} \mathrm{P}\left(|S_{\mathbf{k}} - B_{\mathbf{k}}| > \varepsilon|\mathbf{k}|^{1/r}\right).$$

Therefore, by (3.5) we have

$$\sum_{\mathbf{n}\geq\mathbf{N}_0} \mathrm{P}\left(|S_{\mathbf{m_n}} - B_{\mathbf{m_n}}| > \varepsilon|\mathbf{m_n}|^{1/r}\right) < \infty. \tag{3.6}$$

By the Borel–Cantelli lemma, (3.6) implies that there exist $\mathbf{N}_1 \in \mathbb{N}^d$ and $A \in \mathcal{F}$ such that $\mathbf{N}_1 \geq \mathbf{N}_0$, $\mathrm{P}(A) = 1$ and

$$\frac{|S_{\mathbf{m_n}}(\omega) - B_{\mathbf{m_n}}|}{|\mathbf{m_n}|^{1/r}} \leq \varepsilon \quad \forall \mathbf{n} \geq \mathbf{N}_1,\ \forall \omega \in A. \tag{3.7}$$

Henceforward let $\omega \in A$ be fixed.

If $\mathbf{n} \geq \mathbf{N}_1$ and $\mathbf{t} \in (\mathbf{k_{n+1}}, \mathbf{k_{n+2}}]$, then by $\mathbf{t} \in (\mathbf{m_n}, \mathbf{m_{n+2}}]$, (3.7) and

$$|\mathbf{m_{n+2}}|^{1/r} \geq |\mathbf{m_n}|^{1/r} \geq |\mathbf{m_n}|$$

we have

$$\begin{aligned}
\frac{S_{\mathbf{t}}(\omega) - B_{\mathbf{t}}}{|\mathbf{t}|^{1/r}} &\geq \frac{S_{\mathbf{m_n}}(\omega) - B_{\mathbf{m_{n+2}}}}{|\mathbf{m_{n+2}}|^{1/r}} \\
&= \frac{S_{\mathbf{m_n}}(\omega) - B_{\mathbf{m_n}}}{|\mathbf{m_n}|^{1/r}} \frac{|\mathbf{m_n}|^{1/r}}{|\mathbf{m_{n+2}}|^{1/r}} - \frac{B_{\mathbf{m_{n+2}}} - B_{\mathbf{m_n}}}{|\mathbf{m_{n+2}}|^{1/r}} \\
&\geq -\varepsilon - \frac{B_{\mathbf{m_{n+2}}} - B_{\mathbf{m_n}}}{|\mathbf{m_n}|}.
\end{aligned} \tag{3.8}$$

If $\mathbf{n} = (n_1, n_2, \ldots, n_d) \geq \mathbf{N}_0$ and $\mathbf{m_n} = (\mathbf{m_n^{(1)}}, \mathbf{m_n^{(2)}}, \ldots, \mathbf{m_n^{(d)}})$ then

$$[\alpha^{n_i}] < \mathbf{m_n^{(i)}} \leq [\alpha^{n_i+1}].$$

On the other hand $\mathbf{m_n^{(i)}} \in \mathbb{N}$, hence we get

$$\alpha^{n_i} < \mathbf{m_n^{(i)}} \leq \alpha^{n_i+1}. \tag{3.9}$$

This inequality implies

$$|\mathbf{m_{n+2}}| - |\mathbf{m_n}| > \prod_{i=1}^{d} \alpha^{n_i+2} - \prod_{i=1}^{d} \alpha^{n_i+1}$$

$$= (\alpha^d - 1) \prod_{i=1}^{d} \alpha^{n_i+1}$$
$$> (\alpha^d - 1)\alpha^{n_1} \quad \forall \mathbf{n} = (n_1, n_2, \ldots, n_d) \geq \mathbf{N}_0.$$

Since $\lim_{n\to\infty} \alpha^n = \infty$, therefore $\alpha^n \geq K(\alpha^d - 1)^{-1}$ except for finitely many values of $n \in \mathbb{N}$. Hence there exists $\mathbf{N}_2 \in \mathbb{N}^d$ such that $\mathbf{N}_2 \geq \mathbf{N}_1$ and

$$|\mathbf{m_{n+2}}| - |\mathbf{m_n}| > (\alpha^d - 1)\frac{K}{\alpha^d - 1} = K \quad \forall \mathbf{n} \geq \mathbf{N}_2.$$

This inequality implies by (3.1), that

$$B_{\mathbf{m_{n+2}}} - B_{\mathbf{m_n}} \leq c(|\mathbf{m_{n+2}}| - |\mathbf{m_n}|) \quad \forall \mathbf{n} \geq \mathbf{N}_2. \tag{3.10}$$

Using (3.9) we have

$$\frac{|\mathbf{m_{n+2}}|}{|\mathbf{m_n}|} \leq \prod_{i=1}^{d} \frac{\alpha^{n_i+3}}{\alpha^{n_i}} = \alpha^{3d} \quad \forall \mathbf{n} = (n_1, n_2, \ldots, n_d) \geq \mathbf{N}_2. \tag{3.11}$$

Hence (3.8), (3.10), (3.11) and (3.3) imply, that if $\mathbf{n} \geq \mathbf{N}_2$ and $\mathbf{t} \in (\mathbf{k_{n+1}}, \mathbf{k_{n+2}}]$, then

$$\frac{S_{\mathbf{t}}(\omega) - B_{\mathbf{t}}}{|\mathbf{t}|^{1/r}} \geq -\varepsilon - \frac{B_{\mathbf{m_{n+2}}} - B_{\mathbf{m_n}}}{|\mathbf{m_n}|} \geq -\varepsilon - c\left(\frac{|\mathbf{m_{n+2}}|}{|\mathbf{m_n}|} - 1\right)$$
$$\geq -\varepsilon - c(\alpha^{3d} - 1) \geq -\varepsilon\alpha^{3d/r} - c(\alpha^{3d} - 1) > -\delta. \tag{3.12}$$

If $\mathbf{n} \geq \mathbf{N}_2$ and $\mathbf{t} \in (\mathbf{k_{n+1}}, \mathbf{k_{n+2}}]$, then by $\mathbf{t} \in (\mathbf{m_n}, \mathbf{m_{n+2}}]$, $|\mathbf{m_n}|^{1/r} \geq |\mathbf{m_n}|$, (3.7), (3.11), (3.10) and (3.3), we have

$$\frac{S_{\mathbf{t}}(\omega) - B_{\mathbf{t}}}{|\mathbf{t}|^{1/r}} \leq \frac{S_{\mathbf{m_{n+2}}}(\omega) - B_{\mathbf{m_n}}}{|\mathbf{m_n}|^{1/r}}$$
$$= \frac{S_{\mathbf{m_{n+2}}}(\omega) - B_{\mathbf{m_{n+2}}}}{|\mathbf{m_{n+2}}|^{1/r}} \frac{|\mathbf{m_{n+2}}|^{1/r}}{|\mathbf{m_n}|^{1/r}} + \frac{B_{\mathbf{m_{n+2}}} - B_{\mathbf{m_n}}}{|\mathbf{m_n}|^{1/r}}$$
$$\leq \frac{S_{\mathbf{m_{n+2}}}(\omega) - B_{\mathbf{m_{n+2}}}}{|\mathbf{m_{n+2}}|^{1/r}} \frac{|\mathbf{m_{n+2}}|^{1/r}}{|\mathbf{m_n}|^{1/r}} + \frac{B_{\mathbf{m_{n+2}}} - B_{\mathbf{m_n}}}{|\mathbf{m_n}|}$$
$$\leq \varepsilon\alpha^{3d/r} + c\left(\frac{|\mathbf{m_{n+2}}|}{|\mathbf{m_n}|} - 1\right) \leq \varepsilon\alpha^{3d/r} + c(\alpha^{3d} - 1) < \delta.$$

This inequality and (3.12) imply

$$\frac{|S_{\mathbf{t}}(\omega) - B_{\mathbf{t}}|}{|\mathbf{t}|^{1/r}} < \delta \quad \forall \mathbf{n} \geq \mathbf{N}_2, \mathbf{t} \in (\mathbf{k_{n+1}}, \mathbf{k_{n+2}}]. \tag{3.13}$$

If $\mathbf{t} \geq \mathbf{k_{N_2+1}} + \mathbf{1}$, then there exists $\mathbf{n} \geq \mathbf{N}_2$ such that $\mathbf{t} \in (\mathbf{k_{n+1}}, \mathbf{k_{n+2}}]$. Hence (3.13) implies

$$\frac{|S_{\mathbf{t}}(\omega) - B_{\mathbf{t}}|}{|\mathbf{t}|^{1/r}} < \delta \quad \forall \mathbf{t} \geq \mathbf{k_{N_2+1}} + \mathbf{1}.$$

Therefore the statement is proved. □

4. An application for multi-index sequence of non-negative random variables with finite variances

In this section we give an application of Theorem 3.1. In case $d = r = 1$, this result gives Theorem of PETROV [9].

Theorem 4.1. *Let $X_{\mathbf{n}}$ $(\mathbf{n} \in \mathbb{N}^d)$ be a sequence of non-negative random variables with finite variances, $S_{\mathbf{n}} = \sum_{\mathbf{k} \le \mathbf{n}} X_{\mathbf{k}}$, $c > 0$, $K \in \mathbb{N}$ and $0 < r \le 1$. If*

$$\mathrm{E}\, S_{\mathbf{n}} - \mathrm{E}\, S_{\mathbf{m}} \le c(|\mathbf{n}| - |\mathbf{m}|) \quad \forall \mathbf{n}, \mathbf{m} \in \mathbb{N}^d, \mathbf{n} \ge \mathbf{m}, |\mathbf{n}| - |\mathbf{m}| \ge K \tag{4.1}$$

and

$$\sum_{\mathbf{n}} \frac{\operatorname{Var} S_{\mathbf{n}}}{|\mathbf{n}|^{1+2/r}} < \infty, \tag{4.2}$$

then

$$\lim_{\mathbf{n} \to \infty} \frac{S_{\mathbf{n}} - \mathrm{E}\, S_{\mathbf{n}}}{|\mathbf{n}|^{1/r}} = 0 \quad \textit{almost surely.}$$

Proof. With notation $b_{\mathbf{k}} = \mathrm{E}\, X_{\mathbf{k}}$ and $B_{\mathbf{n}} = \sum_{\mathbf{k} \le \mathbf{n}} b_{\mathbf{k}} = \mathrm{E}\, S_{\mathbf{n}}$, (4.1) implies (3.1). On the other hand, if $\varepsilon > 0$, then the Chebyshev inequality and (4.2) imply

$$\sum_{\mathbf{n}} \frac{1}{|\mathbf{n}|} \mathrm{P}\left(|S_{\mathbf{n}} - B_{\mathbf{n}}| > \varepsilon |\mathbf{n}|^{1/r}\right) \le \sum_{\mathbf{n}} \frac{1}{|\mathbf{n}|} \frac{\operatorname{Var} \frac{S_{\mathbf{n}}}{|\mathbf{n}|^{1/r}}}{\varepsilon^2} = \varepsilon^{-2} \sum_{\mathbf{n}} \frac{\operatorname{Var} S_{\mathbf{n}}}{|\mathbf{n}|^{1+2/r}} < \infty.$$

Therefore (3.2) holds. Hence, using Theorem 3.1, we have that the statement is true. □

References

[1] N. ETEMADI: *An elementary proof of the strong law of large numbers*, Z. Wahrscheinlichkeitstheorie Verw. Gebiete 55.1 (1981), pp. 119–122, DOI: 10.1007/bf01013465.

[2] I. FAZEKAS: *Convergence rates in the Marcinkiewicz strong law of large numbers for Banach space valued random variables with multidimensional indices*, Publicationes Mathematicae, Debrecen 32 (1985), pp. 203–209.

[3] I. FAZEKAS, T. TÓMÁCS: *Strong laws of large numbers for pairwise independent random variables with multidimensional indices*, Publicationes Mathematicae, Debrecen 53.1-2 (1998), pp. 149–161.

[4] A. GUT: *Marcinkiewicz laws and convergence rates in the law of large numbers for random variables with multidimensional indices*, The Annals of Probability 6.3 (1978), pp. 469–482, DOI: 10.1214/aop/1176995531.

[5] O. I. KLESOV: *Strong law of large numbers for random fields with orthogonal values*, Dokl. Akad. Nauk. Ukr. SSR Ser. A 7 (1982), pp. 9–12.

[6] O. I. KLESOV: *The law of large numbers for multiple sums of independent identically distributed random variables*, Theor. Probab. Math. Statist. 50 (1995), pp. 77–87.

[7] V. M. Kruglov: *Strong law of large numbers*, in: Stability Problems for Stochastic Models: Proceedings of the Fifteenth Perm Seminar, Perm, Russia, June 2-6, 1992, Moscow, Utrecht, Tokyo: TVP/VSP, 1994, pp. 139–150, isbn: 90-6764-159-6.

[8] M. Loève: *Probability Theory I.* New York: Springer-Verlag, 1977.

[9] V. V. Petrov: *On the strong law of large numbers for a sequence of nonnegative random variables*, Zapiski Nauchnnykh Seminarov POMI 384 (2010), pp. 182–184, doi: 10.1007/s10958-011-0411-x.

Annales Mathematicae et Informaticae
***50** (2019) pp. 187–202*
DOI: 10.33039/ami.2019.01.002
http://ami.uni-eszterhazy.hu

Optimization of hadoop cluster for analyzing large-scale sequence data in bioinformatics

Ádám Tóth, Ramin Karimi

Faculty of Informatics
University of Debrecen
adamtoth102@gmail.com
raminkm2000@yahoo.ca

Submitted: June 13, 2017
Accepted: January 23, 2019
Published online: February 27, 2019

Abstract

Unexpected growth of high-throughput sequencing platforms in recent years impacted virtually all areas of modern biology. However, the ability to produce data continues to outpace the ability to analyze them. Therefore, continuous efforts are also needed to improve bioinformatics applications for a better use of these research opportunities. Due to the complexity and diversity of metagenomics data, it has been a major challenging field of bioinformatics. Sequence-based identification methods such as using DNA signature (unique k-mer) are the most recent popular methods of real-time analysis of raw sequencing data. DNA signature discovery is compute-intensive and time-consuming.

Hadoop, the application of parallel and distributed computing is one of the popular applications for the analysis of large scale data in bioinformatics. Optimization of the time-consumption and computational resource usages such as CPU consumption and memory usage are the main goals of this paper, along with the management of the Hadoop cluster nodes.

Keywords: hadoop, optimization, next-Generation Sequencing, DNA signature, resource management

1. Introduction

Since the announcement of the human genome project completion in 2003 [8], Next-Generation Sequencing (NGS) technologies have revolutionized exploration of the secrets in the life science. Due to extraordinary progress in this field, massively parallel sequencing of the microbial genomes in the complex communities has led the advent of metagenomics techniques.

Metagenomics, a high-throughput culture-independent technique has provided the ability to investigate the entire community of microorganisms of an environment with analyzing their genetic content obtained directly from their natural residence [9, 15].

Along with technical advances of sequencing, mining the enormous and ever-growing amount of data generated by sequencing technologies is now one of the fastest growing fields of big data science, but there are still a lot of difficulties and challenges ahead. Real-time identification of microorganisms from raw read sequencing data is one of the key problems of current metagenomics and next-generation sequencing analysis. It has a pivotal role for pathogenic diagnostics assays to consider an early treatment.

Sequence-based identification of the species can be classified into two groups: Assembly and alignment-based approaches and alignment-free approaches [7].

Due to the difficulties, technical challenges, and computational complexity of alignment and assembly-based approaches, they are not applicable for complex sequencing data such as metagenomics data. Moreover, they are expensive and time-consuming.

Reads generated by high-throughput sequencing technology are short in length and large in volume, very noisy and partial, with too many missing parts [9, 19]. They contain sequencing errors caused by the sequencer machines. Another challenge is repetitive elements in the DNA sequence of species. As an example, about half of the human genome is covered by repeats [16]. These challenges cause incapability and unreliability of the results in alignment-based identification. Thus, it is necessary to develop efficient alignment-free methods for phylogenetic analysis and rapid identification of species in Metagenomics and clinical diagnostics assays, based on the sequence reads. Several alignment-free methods have been proposed in the literature to address this problem. One of the latest methods is using DNA signature that plays the role of fingerprints for microbial species.

DNA signature is a unique short fragment (k-mer) of DNA, which is specific for a species that is selected from a target genome database. DNA signature can be obtained for every individual species in the genome databases by screening and counting the frequency of k-mers through the entire database. The term k-mer refers to the existence of all the possible substrings of length k in a genome sequence. Each k-mer that appears once in a genome database is a unique DNA signature for the related sequence containing that k-mer. Since comparing k-mers frequencies are computationally easier than sequence alignment, the method can also be used as a first stage analysis before an alignment [10].

Considering the large size of genome databases, searching the unique DNA signatures (k-mers) needs powerful computational resources and it is still time-consuming. This problem can be solved by incorporating parallel and distributed computing.

Several tools and algorithms of k-mers frequency counting and DNA signature discovery have been proposed in the literature. Some of them use the applications of parallel and distributed computing. Hadoop and MapReduce are among these applications. The Apache Hadoop software [18], is a platform for parallel and distributed computing of large data set and MapReduce [3] is a programming model for parallel and distributed data processing.

In this paper, we have proposed optimization techniques to reduce time-consuming and to use less computational resources. Accordingly, we designed the nodes and managed the performance of the nodes in the Hadoop cluster. Managing the CPU consumption and memory usage according to the size of data and the number of maps, monitoring and comparing the running time of the maps were other issues that we have considered in this study.

The aim of this research is to enhance the possibility of using ordinary computers as a distributed system, in order to allow the process of searching DNA signatures and k-mers frequency to be applicable for the entire research community.

2. Background

This section contains a brief overview of the basic concepts that are used in this paper.

2.1. Next-Generation Sequencing (NGS)

Next-generation sequencing (NGS) refers to methods that have emerged in the last decade. NGS technologies allow simultaneous determination of nucleotide sequences of a variety of different DNA strands. It provides reading of billions of nucleotides per day. NGS is also known as massive parallel sequencing. The NGS has dramatically improved in recent years, making the number of bases that can be sequenced per unit price has grown exponentially. Therefore the new platforms are distinguished by their ability to sequence millions of DNA fragments parallel to a much cheaper price per base. In this technology, experimental advances in chemistry, engineering, molecular biology and nanotechnology are integrated with high-performance computing to increase the speed at which data is obtained. Its potential has allowed the development of new applications and biological diagnostic assays that will revolutionize, in the near future, the diagnosis of genetic and pathogenic diseases [1, 12–14, 17, 20].

2.2. Metagenomics

For the first time, the word "metagenomics" appeared in 1998 in the article written by Jo Handelsman [6]. Metagenomics is the study of the entire genetic material of microorganisms obtained directly from the environment. The main goals of metagenomics are to determine the taxonomic (phylogenetic) and functional composition of microbial communities and understanding that how they interact with each other in the term of metabolism. Historically, the bacterial composition was determined by culturing bacterial cells, but the majority of bacteria are simply not cultivated. You can isolate a separate species and study its genome, but this is long, since there are many species together as a diverse community [5, 19].

2.3. Alignment-based analysis

One of the most effective and convenient methods for classification and identification of sequence similarity is an alignment method. Alignment of the new sequences with already well-studied makes it possible to quantify the level of similarity of these sequences, as well as to indicate the most likely regions of similarity structures. The most commonly used programs for the comparison of sequences are BLAST, FASTA and the Smith-Waterman (SW) is the most sensitive and popular algorithm is used. The applications of sequence alignment are limited to apply for closely related sequences, but when the sequences are divergent, the results cannot be reliable. Alignment-based approaches are computationally complex, expensive, and time-consuming and therefore aligning large-scale sequence data is another limitation [11]. Inability of this method becomes more visible when facing massive sequencing reads in metagenomics.

2.4. Alignment-free analysis

Alignment-free methods are an alternative to overcome various difficulties of traditional sequence alignment approaches, they are increasingly used in NGS sequence analysis, such as searching sequence similarity, clustering, classification of sequences, and more recently in phylogeny reconstruction and taxonomic assignments [2, 4]. They are much faster than alignment-based methods. The recent most common alignment-free methods are based on k-mer/word frequency. DNA signature is a unique short fragment (k-mer) of DNA, which is specific for a species that is selected from a target genome database. DNA signature can be obtained for every individual species in the genome databases by screening and counting the frequency of k-mers through the entire database [10].

3. System model

In our investigation we create a cluster which consists of three computers (N1, N2, N3). The next table (see Table 1) shows the main components of them.

Node	Processor	Memory	Hard disk
N1	Intel Core i3-2120 (3.3 GHz)	4 GB	ST1500DL003-9VT16L (1.5 TB)
N2	Intel Core i7-3770K (3.5 GHz)	16 GB	WDC WD20EZRX-00DC0B0 (2 TB)
N3	Intel Core i7-4771 (3.5 GHz)	16 GB	TOSHIBA DT01ACA200 (2 TB)

Table 1: Main components of the nodes

To run Hadoop, Java is required to be installed and we use version of 1.8.0-111. Throughout the investigation on each computer runs Ubuntu 14.04.5 Server (64-bit) and to collect information of utilization of cpu, memory, I/O of the nodes we choose collectl tool which is suitable for benchmarking, monitoring a system's general heath, providing lightweight collection of device performance information. In this paper the input data is part of the genome database which contains k-mers. In our case k is 18, so that each file contains lines of 18 character lengths. We configure block size as the input for running the maps to be exactly 1GB so after loading the data into HDFS it will be divided into 1GB parts. We considered 1GB of RAM to each map process. We downloaded the bacterial genome database in FASTA format from the National Center for Biotechnology Information (NCBI) database. The size of this database is 9.7 GB after decompression. In order to generate k-mers from the FASTA files, we used GkmerG software to generate all the possibilities of 18-mers from individual bacterial genomes of the whole database with a total size of 177.35 GB. The result is a file containing a single column of k-mers with length 18. The length of k-mers can differ according to the needs. Since, the aim of this paper is optimizing the process of searching the frequency of k-mers, the large file must be split in the same size (1GB) for each map to compare the running time of the maps, therefore we assume that the large file is pre-sorted and the same k-mers is not located in different files after splitting. For the real data processing we have to split the large file in an intelligence manner.

During our investigation we use one of the example program on the input data called grep that extracts matching strings from text files and counts how many time they occurred.

Hadoop is designed to scale up from single servers to thousands of machines, each proposing local computation and storage, giving the opportunity to the system to be highly-available. Usually in practice the chosen machine, which is the master node, does not store any data so it does not do the job of datanode. In our case there is only one "real" datanode (N3) which is opposite of Hadoop's principles. However, in our scenario in that way we can focus on the efficiency of the utilization of resources during running Hadoop and also with that configuration we decrease network traffic as low as possible. Because YARN only supports CPU and memory reservation with our setting we can discover the crucial point of the system. The nodes of the cluster are configured in the following way: N1 is the master node (Resourcemanager runs on it), on every occasion only the applicationmaster (AM) runs on N2 because 15 GB RAM is given to AM and on N3 the default setting remains (1536MB) and Hadoop chooses N2 to start running AM. 14GB is given to yarn on N3 meaning that maximum 14 maps can run simultaneously. Every block

is located physically on N3 and every container (so every map process) runs on N3 except the AM.

The following table (see Table 2) includes the main features of *Cases*.

Case	Number of map	Size of the block	Size of dataset	Given memory to N3
a	1	1 GB	1 GB	14 GB
b	2	1 GB	2 GB	14 GB
c	3	1 GB	3 GB	14 GB
d	4	1 GB	4 GB	14 GB
e	5	1 GB	5 GB	14 GB
f	6	1 GB	6 GB	14 GB
g	7	1 GB	7 GB	14 GB
h	8	1 GB	8 GB	14 GB
i	9	1 GB	9 GB	14 GB
j	10	1 GB	10 GB	14 GB
k	11	1 GB	11 GB	14 GB
l	12	1 GB	12 GB	14 GB
m	13	1 GB	13 GB	14 GB
n	14	1 GB	14 GB	14 GB

Table 2: Scenario A

4. Numerical results

4.1. Scenario A

As it is indicated earlier we use collectl to get information about the utilization of resources of the nodes. It works under linux operating system and basically reads data from /proc and writes its results into a file or on the terminal. It is capable of monitoring any of a broad set of subsystems which currently include buddyinfo, cpu, disk, inodes, infiniband, lustre, memory, network, nfs, processes, quadrics, slabs, sockets and tcp. Collectl output can also be saved in a rolling set of logs for later playback or displayed interactively in a variety of formats. The command can be run with lots of arguments and it can be freely customized in which mode collectl runs or how its output is saved. Below (see Table 3) we can see some results about running times. The first column indicates the whole running time of the application, the second one shows the mean running time of the map jobs, the third one the initialization period which means after hadoop starts some time is needed before launching map jobs (e.g. deciding which node will be the master node). The last column represents the whole running time divided by the number of map jobs.

It can be seen that as more map processes are running in parallel the average processing time of 1 GB data starts to decrease then it remains around a constant value.

	Running time	Average running time of the maps	Initialization period	Average processing time of 1 GB data
a	33	22	11	33 sec
b	39	26	13	19.5 sec
c	61	49	12	20.33 sec
d	78	65	13	19.5 sec
e	92	80	12	18.4 sec
f	109	96	13	18.17 sec
g	129	117	12	18.43 sec
h	150	137	13	18.75 sec
i	161	149	12	17.89 sec
j	187	174	13	18.7 sec
k	204	191	13	18.54 sec
l	226	213,4	12	18.83 sec
m	242	230	12	18.61 sec
n	267	255	12	19.07 sec

Table 3: Results in connection with running times

4.1.1. Results in connection with the node where AM is located

As mentioned earlier the AM runs on this node on every occasion. Because of the rounding and the way of configuring collectl figures might not demonstrate the exact beginning of the running process but the deviation is very little. Because of the lots of data the graphs would be unclear so the achieved results are divided into two groups according to the number of processed blocks:

- Number of processed blocks are odd (1,3,5,7,9,11,13)
- Number of processed blocks are even (2,4,6,8,10,12,14)

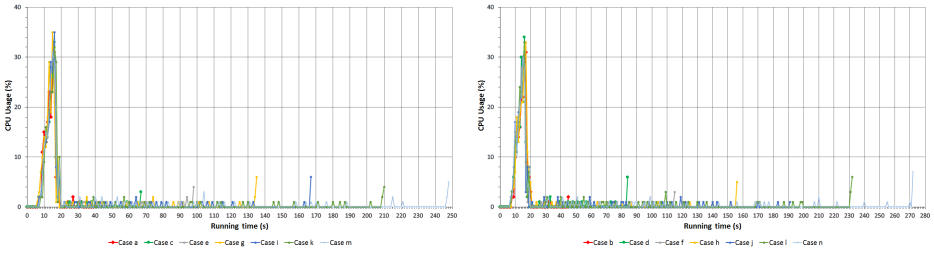


Figure 1: CPU usage of N2

On Figure 1 the data of cpu usage of N2 node is shown. As on N2 just the AM runs it can be seen after the jobs are initiated CPU usage increases to about 35%, it lasts for a while then it remains almost 0% till the application runs. Some jumps can be observable at the end of the *Cases* which are caused by the fact that map jobs come to an end.

Figure 2 shows the utilization of disk capacity on N2 node. Similarly to cpu usage when the jobs are initiated usage of disk rises then it drops independently of the number of map processes and also some jumps occur at the end of *Cases* when map jobs are finished.

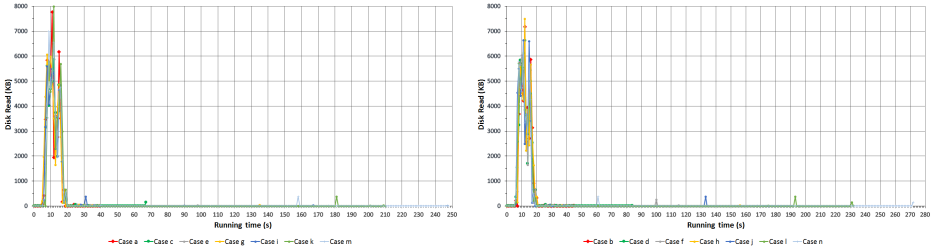


Figure 2: Speed of disk reading of N2

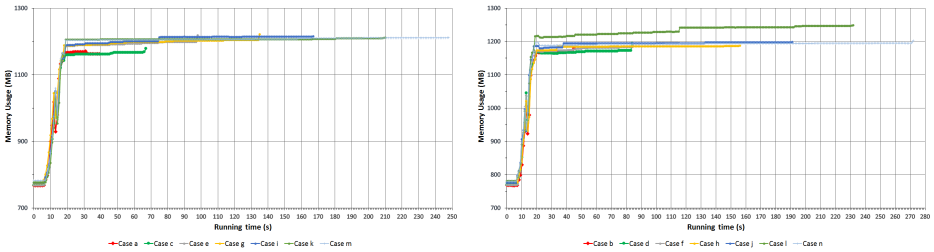


Figure 3: Size of reserved memory of N2

Figure 3 displays the memory usage of N2 node. Size of the reserved memory is independent of the number of initiated map processes.

4.1.2. Results in connection with the node where the "real" datanode is located

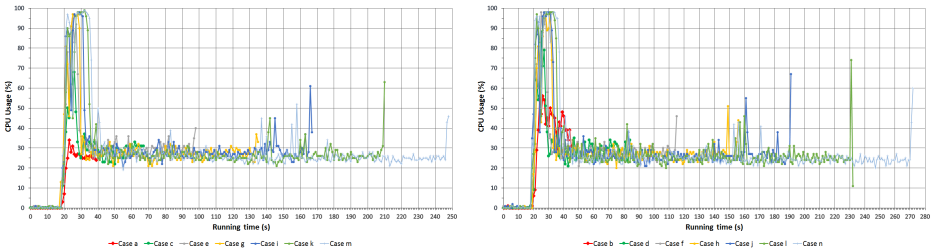


Figure 4: Cpu usage of N3

Figure 4 represents the cpu usage of N3 node. When one map is running (*Case a*) it reserves one of the four cores so cpu usage barely passes 25%. When two maps are running (*Case b*) it reserves two of the four cores so the maximum cpu usage can not step over 50% but it is around 40%. Whenever three or more maps are running simultaneously apart from the initial jump cpu usage stabilizes around

30%. The reason for this is the limit of disk reading capability as Figure 6 will prove that statement.

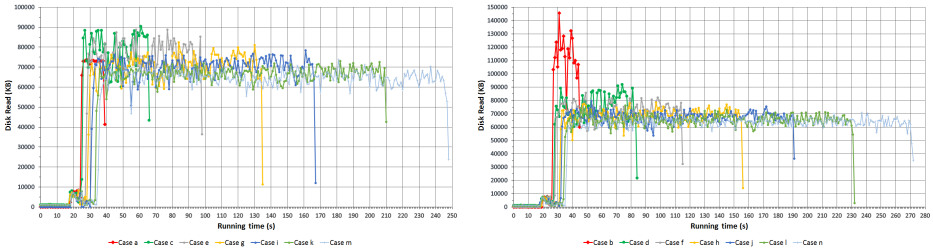


Figure 5: Speed of disk reading of N3

Figure 5 represents the utilization of capability of disk reading of N3 node.

In case of 1 map to read 1 GB it uses approximately the half of disk capability because only 1 core is reserved which is fully loaded.

When two maps are running two cpu cores are reserved and around 3/4 of disk capability is used. To read 2 GB into the memory lasts almost the same as in the first case but the speed of disk reading is almost twice as much as in *Case a*. Furthermore whenever three or more maps are running the limit of I/O arises because of the emerging congestion in the system. That is why cpu usage does not increase after *Case b*.

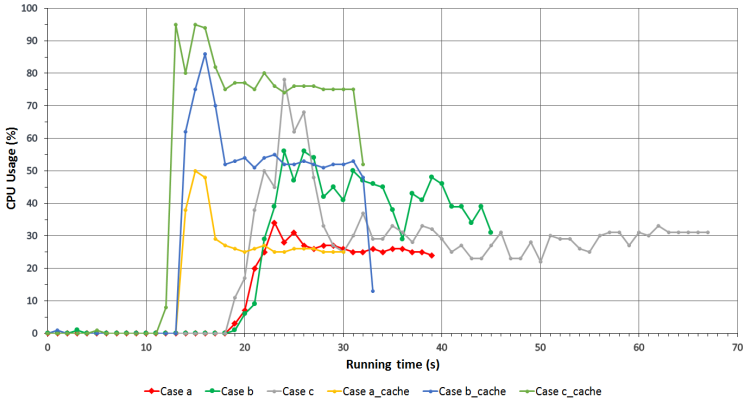


Figure 6: Comparison of cpu usage

We investigate the scenario when we preload the necessary dataset into the memory so there is no I/O procedure during the running of the map tasks and the program reaches the data directly from the memory. From Figure 6 it can be observed when the data is reachable from the memory CPU usage reach the theoretical maximum in all *Cases* (*Case_a_cache, Case_b_cache and Case_c_cache*)

so it is clear that the cross section point is the I/O capability (reading).

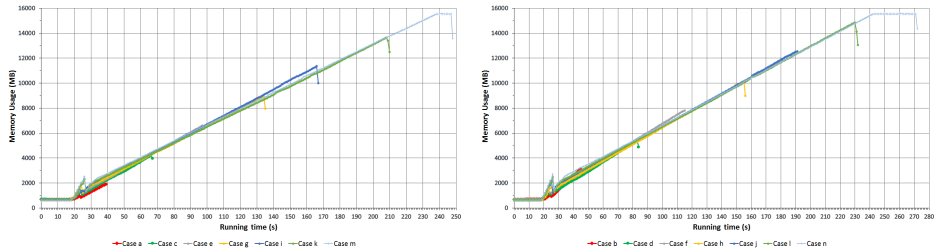


Figure 7: Size of reserved memory of N3

Figure 7 shows the size of reserved memory of N3. As time goes by the amount of memory usage increases.

4.2. Scenario B

From the obtained results it appears that in *Case a* limit of cpu usage arises while in the other *Cases* limit of I/O capability restricts the performance. So in the next scenario the system is changed a little bit. Almost everything is the same as in Scenario A except that instead of 14 GB 2 GB RAM is given to yarn on N3 (see Table 4). This little modification has a remarkable effect on the operation of Hadoop, in Scenario B 2 maps can run in parallel at the same time altogether.

Case	Number of parallel maps	Number of total maps	Size of dataset	Given memory to N3
a	1	1	1 GB	2 GB
b	2	2	2 GB	2 GB
c	2	3	3 GB	2 GB
d	2	4	4 GB	2 GB
e	2	5	5 GB	2 GB
f	2	6	6 GB	2 GB
g	2	7	7 GB	2 GB
h	2	8	8 GB	2 GB
i	2	9	9 GB	2 GB
j	2	10	10 GB	2 GB
k	2	11	11 GB	2 GB
l	2	12	12 GB	2 GB
m	2	13	13 GB	2 GB
n	2	14	14 GB	2 GB

Table 4: Scenario B

Below (see Table 5) some results about running times can be noticeable, this table is the same as Table 3 with the results of Scenario B:

The same tendency takes place here as in Scenario B namely as more map processes are running in parallel the average processing time of 1 GB data starts to decrease then it remains around a constant value.

	Running time	Average running time of a map	Initialization period	Average processing time of 1 GB data
a	34	22	12	34 sec
b	48	36	12	24 sec
c	58	25	12	19.33 sec
d	68	28	12	17 sec
e	85	26,4	11	17 sec
f	92	26,1666667	12	15.33 sec
g	115	27	11	16.43 sec
h	126	28,5	11	15.75 sec
i	133	24,333	13	14.77 sec
j	142	25,5	11	14.2 sec
k	157	23,818181	13	14.27 sec
l	173	26,08333	13	14.42 sec
m	190	25,384615	12	14.62 sec
n	191	24,47143	13	13.64 sec

Table 5: Results in connection with running times

4.2.1. Results in connection with the node where AM is located

As mentioned earlier the AM runs on this node on every occasion.

We use the same style as previously so the achieved results are divided into two groups according to the number of processed blocks:

- Number of processed blocks are odd (1,3,5,7,9,11,13)
- Number of processed blocks are even (2,4,6,8,10,12,14)

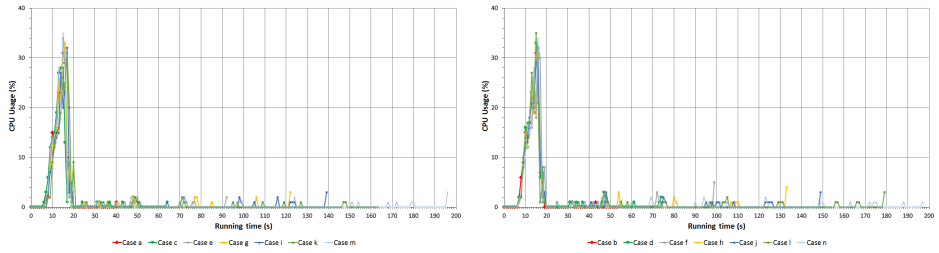


Figure 8: CPU usage of N2

On Figure 8 we can see the data of CPU usage of N2. We get back almost the same result as in Scenario A even the values are practically identical.

We can see the speed of disk reading of N2 on Figure 9. The situation is the same as in case of Scenario A. These figures also imply the fact that the little modification does not change the utilization of the resources of N2.

Figure 10 present the memory usage of N2. It is evident that the size of reserved memory depends a little on the number of launched map processes.

4.2.2. Results in connection with the node where the "real" datanode is located

Figure 11 demonstrates the CPU usage of N3. It is noticeable that in particular intervals the CPU usage boosts. These phenomena can by explained by the fact

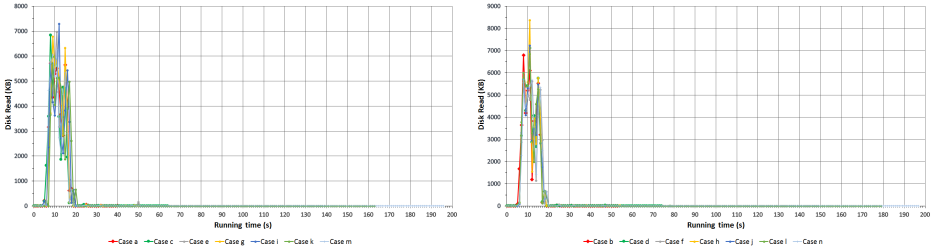


Figure 9: Speed of disk reading of N2

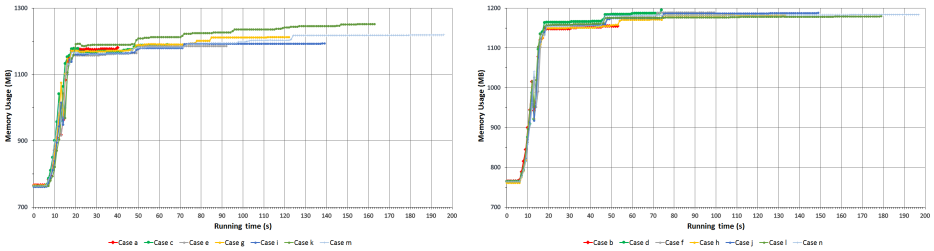


Figure 10: Size of reserved memory of N2

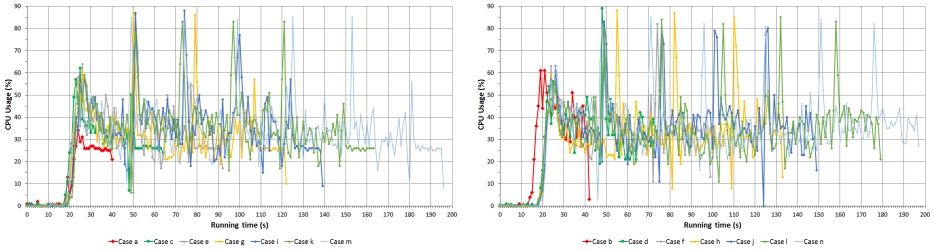


Figure 11: CPU usage of N3

that 2 map processes can run in parallel at the same time so when 3 GB or more data are processed launching a new map process requires some time. In these intervals the CPU usage is greater. Another interesting situation is observable in cases of odd numbered processed blocks because at the end of the running of the last map process CPU usage decreases to about 25% as only one map is running at that time.

Figure 12 shows the speed of disk reading of N3. The same tendency can be observed as in CPU usage. Here we reserve less resources compare to the other scenario. Just 2 maps can run at the same time in parallel so the factor of congestion is smaller but we do lose some time whenever a map finishes/starts. Despite that fact it still results greater speed of disk reading in overall compared to Scenario A.

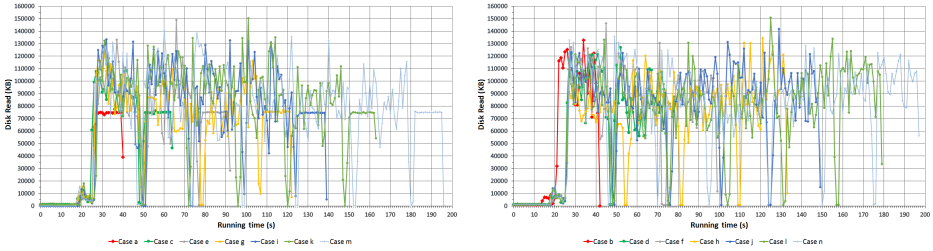


Figure 12: Speed of disk reading of N3

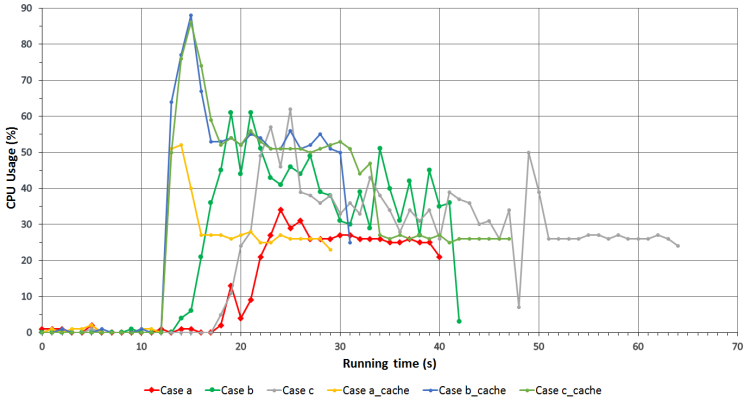


Figure 13: Comparison of cpu usage

Figure 13 presents the situation when the necessary dataset is available from memory. The difference is still there among the appropriate *Cases* (like between *Case b* and *Case_ b_ cache* or *Case c* and *Case_ c_ cache*) but it is smaller than in Scenario A.

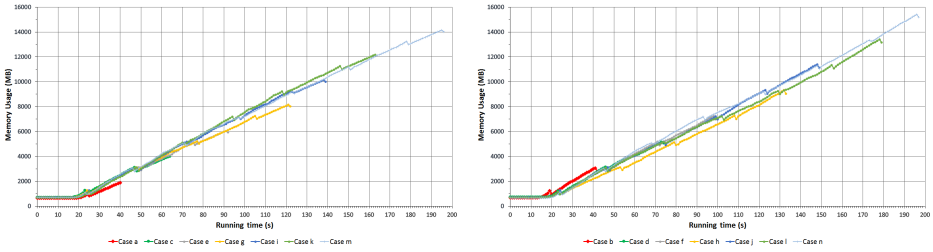


Figure 14: Size of reserved memory of N3

Figure 14 show the size of reserved memory of N3. As time goes by the amount

of memory usage increases.

Let introduce the following notations:

- n(t): at time t the number of parallelly running maps
- T: entire running time of the map in seconds
- g: the cost of required resource of one map
- K: the cost of the entire running time

Then

$$\sum_{t=0}^{T}(g * n(t)).$$

In the Table 6 we used the following formula: $K = T * n * g$, where n is the maximum number of executable maps.

The next table compares the investigated scenarios where $g = 0.9$:

- We gave 14 GB memory to YARN so the whole dataset can be executable in parallel. The number of parallelly executable maps are 14 (MAP14).
- We gave 2 GB memory to YARN. The number of parallelly executable maps are 2 (MAP2).

Number of blocks	Running time (Scenario A)	Cost (Scenario A)	Running time (Scenario B)	Cost (Scenario B)
1	33	29,7	34	30,6
2	39	70,2	48	86,4
3	61	164,7	58	104,4
4	78	280,8	68	122,4
5	92	414	85	153
6	109	588,6	92	165,6
7	129	812,7	115	207
8	150	1080	126	226,8
9	161	1304,1	133	239,4
10	187	1683	142	255,4
11	204	2019,6	157	282,6
12	226	2440,8	173	311,4
13	242	2831,4	190	342
14	267	3364,2	191	343,8

Table 6: Comparison of the scenarios

5. Conclusion

This paper addressed optimization techniques for reducing the time-consumption and computational resource usage in a Hadoop cluster. Running time, CPU usage, memory usage, and Speed of disk reading of the nodes are the subjects that have been screened in the Hadoop cluster to examine the proposed optimization techniques for searching the frequency of DNA signatures (k-mers) in the genomic

data. The obtained results show that using all the resources (CPU, memory) is not always the best solution and our scenarios is a prime example of it. Managing the maps according to the size of data and memory is critical. Our results also indicate that the speed of I/O greatly affects the effectiveness of performance. To get a better and faster operation, optimizing the configurations and parameters of Hadoop is also required in order to reduce the data transfer and communication between nodes of the cluster. Comparing two Scenarios show another remarkable result; running the maps in parallel causes shorter processing time. The aim of this research is to enhance the possibility of using ordinary computers as a distributed computing system for the entire research community to analyze large-scale dataset.

References

[1] L. Barzon, E. Lavezzo, V. Militello, S. Toppo, G. Palù: *Applications of Next-Generation Sequencing Technologies to Diagnostic Virology*, International Journal of Molecular Sciences 12.11 (2011), pp. 7861–7884, doi: 10.3390/ijms12117861.

[2] C. X. Chan, M. A. Ragan: *Next-generation phylogenomics*, Biology Direct 8.3 (2013), doi: 10.1186/1745-6150-8-3.

[3] J. Dean, S. Ghemawat: *MapReduce: Simplified Data Processing on Large Clusters*, Communications of the ACM 51.1 (2008), pp. 107–113, doi: 10.1145/1327452.1327492.

[4] M. Domazet-Lošo, B. Haubold: *Alignment-free detection of local similarity among viral and bacterial genomes.* Bioinformatics. 27.11 (2011), pp. 1466–1472, doi: 10.1093/bioinformatics/btr176.

[5] J. A. Gilbert, C. L. Dupont: *Microbial metagenomics: beyond the genome.* Annual review of marine science. 3 (2011), pp. 347–371, doi: 10.1146/annurev-marine-120709-142811.

[6] J. Handelsmanl, M. R. Rondon, S. F. Brady, J. Clardy, R. M. Goodman: *Molecular biological access to the chemistry of unknown soil microbes: a new frontier for natural products.* Chemistry & biology 5.10 (1998), R245–R249.

[7] B. Haubold, F. Reed, P. Pfaffelhuber: *Alignment-free estimation of nucleotide diversity*, PLOS Computational Biology 27.4 (2011), pp. 449–455, doi: 10.1093/bioinformatics/btq689.

[8] *Human Genome Project*, https://web.ornl.gov/sci/techresources/Human_Genome/index.shtml.

[9] R. Karimi, L. Bellatreche, P. Girard, A. Boukorca, A. Hajdu: *BINOS4DNA: Bitmap Indexes and NoSQL for Identifying Species with DNA Signatures through Metagenomics Samples*, in: Information Technology in Bio- and Medical Informatics, Switzerland: Springer, Cham, 2014, pp. 1–14, doi: 10.1007/978-3-319-10265-8_1.

[10] R. Karimi, A. Hajdu: *HTSFinder: Powerful Pipeline of DNA Signature Discovery by Parallel and Distributed Computing*, Evolutionary bioinformatics online 12 (2016), pp. 73–85, doi: 10.4137/EBO.S35545.

[11] C. Kemena, C. Notredame: *Upcoming challenges for multiple sequence alignment methods in the high-throughput era.* Bioinformatics. 25.19 (2009), pp. 2455–2465, doi: 10.1093/bioinformatics/btp452.

[12] S. Moorthie, C. J. Mattocks, C. F. Wright: *Review of massively parallel DNA sequencing technologies*, The HUGO journal 5.1-4 (2011), pp. 1–12, doi: 10.1007/s11568-011-9156-3.

[13] C. S. Pareek, R. Smoczynski, A. Tretyn: *Sequencing technologies and genome sequencing.* Journal of applied genetics. 52.4 (2011), pp. 413–435, doi: 10.1007/s13353-011-0057-x.

[14] E. Pettersson, J. Lundeberg, A. Ahmadian: *Generations of sequencing technologies.* Genomics. 94.2 (2009), pp. 105–111, doi: 10.1016/j.ygeno.2008.10.003.

[15] T. Thomas, J. Gilbert, F. Meyer: *Metagenomics - a guide from sampling to data analysis*, Microbial Informatics and Experimentation 2.3 (2012), doi: 10.1186/2042-5783-2-3.

[16] T. J. Treangen, S. L. Salzberg: *Repetitive DNA and next-generation sequencing: computational challenges and solutions.* Nature reviews. Genetics. 13.1 (2011), pp. 36–46, doi: 10.1038/nrg3117.

[17] K. V. Voelkerding, S. A. Dames, J. D. Durtschi: *Next-generation sequencing: from basic research to diagnostics.* Clinical chemistry. 55.4 (2009), pp. 641–658, doi: 10.1373/clinchem.2008.112789.

[18] T. White: *Hadoop: The Definitive Guide*, Sebastopol, California, USA: O'Reilly Media, 2015.

[19] J. C. Wooley, A. Godzik, I. Friedberg: *A Primer on Metagenomics*, PLOS Computational Biology 6.2 (2010), doi: 10.1371/journal.pcbi.1000667.

[20] J. Zhang, R. Chiodini, A. Badr, G. Zhang: *The impact of next-generation sequencing on genomics*, J Genet Genomics. 38.3 (2011), pp. 95–109, doi: 10.1016/j.jgg.2011.02.003.

Methodological papers

Annales Mathematicae et Informaticae
***50** (2019) pp. 205–216*
DOI: 10.33039/ami.2019.01.003
http://ami.uni-eszterhazy.hu

Situation games to ease transition between abstract and real life mathematics for primary school student teachers*

Erika Gyöngyösi-Wiersum[a], Zita Makó Czapné[b], Gregory Makrides[c]

[a]Sárospataki Comenius Campus, Eszterházy Károly University, Sárospatak, Hungary
wiersumne.erika@uni-eszterhazy.hu

[b]Eszterházy Károly University, Eger, Hungary
mako.zita@uni-eszterhazy.hu

[c]Predisent of Cyprus Mathematical Society
Mathematical Society of South-Eastern Europe
THALES Foundation, Member of Edu Committee
European Mathematical Society
Nicosia, Cyprus
makrides.g@eaecnet.com

Submitted: May 19, 2018
Accepted: January 15, 2019
Published online: February 27, 2019

Abstract

With the accelerated development of science and innovation, as well as the invasion of digital systems there is a growing need for science teachers who can provide short, precise and clear explanations on scientific issues. In addition, it is essential for teachers to know how to use new systems, information technology and how to help their students in evaluating and sharing information responsibly. They need to become active data explorers who can plan for, acquire, manage, analyse, and infer from data. The goal is to use data to describe the world and answer puzzling questions while playing roles in different situations so students can playfully prepare for today's data-driven

*This research was supported by the Ministry of Human Resources that awarded the first author of the paper with a research scholarship within national excellence programmes.

society. On the other hand the time of students and teachers is precious; hence, one of the teacher's crucial tasks is to find methods and techniques in order to motivate students to learn and to make the learning as effective as possible. The freedom in teaching enables teachers to develop an innovative learning environment and effective teaching techniques for students to work well together and be successful at learning. Our approach is to explore new forms of teaching and learning to allow students to think critically without relying on their teacher's answers. In this research, using methods that also improve communication skills in the form of situation games with the help of drama pedagogy and observing what makes the method more effective can help in productive innovation.

Keywords: mathematics education, innovative learning environment, primary teacher training, didactical methodology, situation games, drama pedagogy.

1. Introduction

The problem of low mathematical skills of students is caused by many reasons such as unsuitable teaching and learning environment, few teaching methods, negative attitude of pupils and parents towards mathematics, shortage of good quality teaching and learning materials, negative interaction between teachers and pupils to mention few (Michael, 2013).

All learners are different. However, most educational materials are the same for all. In terms of teaching methods, a teacher has plenty of great possibilities to use, learning environment arrangement, visual aids, etc. This freedom in teaching provides a good chance to use an enormous number of ideas in the classrooms (Boumová, 2008).

According to a report by Open University in 2017, ten innovations are already in currency but have not yet had a profound influence on education. A short list of these new pedagogies is: spaced learning, learners making science, open textbooks, people need to be able to evaluate and share information responsibly, intergroup empathy, immersive learning, student-led analytics, big-data inquiry, learning with internal values, humanistic knowledge-building communities (Ferguson, 2017).

2. Theoretical background

Since the time of Socrates, philosophers have questioned the purpose of education and they have proposed four answers. Education is intended to train people for employment, to develop good citizens, to socialise people within a community, and to develop happy, rounded individuals.

As Kenneth T. Henson claims, some teaching purposes such as understanding, knowledge transfer usually determines the methods to use; however, there never exists the best method for everything (Henson, 1980).

Mathematics can provide the necessary knowledge and skills to empower a person to process a mass of information every day. Students are required to learn a

considerable amount of complex and diverse mathematical knowledge accumulated during thousands of years. However, instead of expanding the curriculum another dimension such as the didactical point of view is to be considered and integrated into it. Students need to be engaged in activities encouraging learning and investigation. Carefully designed teaching methodology and entirely new organisations of tasks provide opportunities for students to develop their epistemic value (also Artigue, 2010, p. 467) and to take part in problem solving activities while learning how to apply their knowledge to real-life situations. In the present project, this means that students' work with the designed situations should be related to, and support, their work with theory and so facilitate the transition between abstract and real life mathematics.

One of the most important principles is gaining experience based on specific activity, using tools, and inserting games and playful activities in classroom lessons. The emphasis should be on understanding, on the process and on creating efficient learners rather than on the product (Carr, 2011).

The Hungarian endeavours characterised by names such as Zoltán Dienes and Tamás Varga were part of the worldwide education reform, but in many respects, they diverged from the dominant foreign trends. "According to Tamás Varga, young people are able to learn new topics if it is done playfully. Teaching tools were recommended for primary school teachers, for example, how to improve space-vision with the building game Babylon, or the Dienes-set for teaching number systems" (Reményi, 2007).

According to Jim Scrivener, the teacher's main role is to "help learning to happen," which means "involving" students in what is going on "by enabling them to work at their own speed, by not giving long explanations, by encouraging them to participate, talk, interact, do things, etc." (Scrivener, 2005).

Another vital aspect is context and purpose. This is supported by the opinion expressed by Jill and Charles Hadfield who claim, that activities which mirror real life situations and which have a goal, for example finding a rule, are "more interesting and motivating for the learners (Hadfield, 2003).

Significant changes are needed in the pedagogical support of the university curriculum, filling it with teaching methods providing the training of future specialists with the required comprehensive result. Modern education should focus on students' independent activities, the organisation of self-learning environments and experimental and practical training that encourage students' interest in the profession, promote the efficient acquisition of training materials, form patterns of conduct, provide high motivation, strengthen knowledge, team spirit and freedom of expression, and most importantly, contribute to the complex competence of future specialists (Nadezhda, 2014).

3. Framework of the research

It has long been known that we learn facts better in a series of short sections of the education material with gaps between them, rather than in a long teaching session

such as a lecture. Situation games can provide gaps necessary for successful spaced learning.

Learners need the skills and knowledge to solve problems, evaluate evidence and make sense of data encountered in a complex and constantly changing world. A strong understanding of Science, Technology, Engineering, Arts and Mathematics (STEAM) topics can develop these skills. These changes can be achieved through participation and contribution to science activities in different situations that are personally relevant, help critical thinking and reflection. In situation games, learners can experience how science is made and can enhance their content knowledge.

In situation games, students from different backgrounds interact with each other. This means that skills such as communication, teamwork and empathy are important. Drama pedagogy provides the theoretical framework of situations used for educational purposes. Activities designed to promote intergroup empathy can provide effective responses and help to reduce tensions.

Learning based on experience in learning situations and exploration can be intensified through immersion. It can enable people to experience a situation as if they were there, applying their knowledge to solve a problem or practice a skill. The learning comes from integrating vision, sound and movement. Immersion requires learners to act out scenarios or take part in investigations, pretending to be actors to stimulate reality.

Learning should be rooted in students' own needs and interests and shaped by their internal values. However, students need to learn a set of external values from the national curriculum. We have made efforts to design and develop situations that can meet this challenge. The main approach is to offer students a choice of what role they can play and how they learn. At the same time, it equips them with means to develop appropriate skills and way of thinking in order to support their learning.

Another goal of situation games is to help students become open to experience, creative and self-directed. This is a person centred approach. The curriculum contains collective knowledge of a community. This is an idea-centred approach. We focus on combining the two approaches. Research shows that students who can find a balance between the two approaches develop their knowledge in integrated and transformative ways.

Through a discussion of the results of the present action research, we can share some interesting first results with practicing primary school student teachers.

The new direction in curriculum development is to link methodology of teaching mathematics to subjects taught in primary schools. In this way, students can gain practical knowledge in their future teaching job and they see what mathematics topics are necessary to teach and how to teach those in primary schools.

The first author teaches mostly mathematics and methodology in a teacher training college (now part of a teacher training university) since 2004 in a small town Sárospatak, in Hungary. In 2017 she became responsible for the practical training of teaching students in the second and third year. Therefore the first author carried out research work with students teaching mathematics at II. Rákóczi

Ferenc elementary training school collecting potentials and disadvantages of situation games in teaching. Data were collected by student questionnaires.

The second author teaches mathematics combined with methodology at Eszterházy Károly University in Eger. In this experiment she taught Functions, elements of analysis for students in the second year. In the previous terms students did not use situation games during their mathematics lessons. However, they used them when they were learning Functions, elements of analysis. We compare students' average results at the end of the first three terms with those at the end of the last term.

The third author is the inventor of the new communication methods closely related to situation games, such as the MATHeatre method and the MATHFactor method and recently the SCIENCEtheatre and the SCIENCEFactor (Makrides, 2017). He is also the founder of the new THALES programme for developing analytical skills in pupils of ages 8-15 through a short programme that involved word problems relating to real life situation, mathematics communication and memory development actions. The later includes also the new THALES^{TM} testing for competence (C test) and for mathematical ability (M test).

All three of us trust in applying innovative teaching methods building on pupils' and university students' activities. Love of mathematics and interest motivate learning more than any other factor. It is important to differentiate in the teaching process, to take into account differences of individuals, to let make mistakes without punishment, to play games at home and in the lessons for pedagogical purposes.

4. Findings and interpretations

Our aim is to educate university students to become proficient learners and later on teachers. These skills include understanding the nature of knowledge, assessing the validity of claims, and forming sound arguments. They include the development of reliable processes and strategies for making sense of the world – such as the scientific method. They include the ability to empathise with others and to judge the merit of different perspectives and narratives. Recent research in neuroscience has uncovered the detail of how we produce long-term memories. A study of spaced learning shows a significant increase in learning compared to a typical lesson. (Ferguson, 2017). This has led us to design a similar teaching method of spaced repetition that occurs in the following order for university students: (1) the lesson begins with a revision of 5 minutes (2) the teacher gives information for 20 minutes; (3) students take a break of 10 minutes to participate in a connected practical activity such as playing situation games, modelling; (4) students are asked to recall key information for 10 minutes with the help of situations where they applied their new knowledge. The first two authors applied situation games during their lessons for university students and built on students' active participation in problem solving activities. Students acted a certain problem in a situation and then solved it by creating a mathematical model. The principles of teaching management include the methodological diversity, the encouragement of group and individual work instead

of the previously dominating frontal form of work.

Some examples of problems solved with the help of situation games are shown below. We used problems of elementary level to show university students how to make learning mathematics more experimental, effective and to raise their pupils' motivation level. We tried to find less abstract examples being useful for future elementary school teachers to apply in their teaching practice. The first problem concerns sets and set operations. Function is involved in the second and geometric sequences in the third.

Example 1

For the situation game we choose five students. Four will be shopkeepers and one of them will be the customer. There are four shops a bookshop, a music shop, a shoe shop and a bicycle shop. The customer wants to buy something from each of these shops. The customer lives in a small village and needs to travel to a neighbouring town to buy these things in those shops. The customer knows when they are open and he/she needs to find when all four shops are open at the same time and to find out how much time he/she has to get to each of these shops. The shopkeepers will tell the customer the opening hours of the four shops and what they sell and then they play their roles:

1. 8.00–14.30
2. 9.00–15.30
3. 9.30–16.30
4. 8.00–12.00 and 13.00–17.30

This task found in a textbook has been transformed into a situation game.

Example 2

Two students want to go on a treasure hunt tour. They got a small map with the ratio of the zoom and they need to find out the distances in reality. They calculate that if they leave at 9 o'clock in the morning they need to walk 3 hours in a forest to reach a view point 10 km from the starting point. They have a rest for an hour and have lunch. After lunch they walk 18 km further for another 4 hours, and they get to a tree where they need to dig to find the treasure. The other students are drawing the graph of the trip as a function of time.

Example 3

The following problems can be solved in groups. Some students are chosen to be bank managers giving offers to a student who has money and wants to find the most advantageous investment offer. Geometric sequences can be practiced with these real world problems. If we want to challenge students we can even ask them

to find the current best investment. They can use the internet during problem solving. In this way everybody who has some capital is interested to find the best offer to gain as much money as possible.

1. Which investment is the most advantageous, A, B or C if we want to put a given amount of money to gain interest in a bank,

   A: at 4% interest, compounded annually at the end of 3 years;

   B: at 12% simple interest at the end of a year;

   C: at 6% interest, compounded annually at the end of 2 years?

   Solution:

   In the case A you need to multiply the given amount of money by 1.04^3 which is approximately 1.125.
   In the case B: the same amount is multiplied by 1.12.
   In the case C: you need to multiply the same amount of money by 1.06^2 which is approximately 1.124.
   Therefore the most advantageous investment is in the case A.

2. We have 1,000,000 Ft and two investment opportunities to invest our capital for five years:

   A: at 10% interest compounded annually;

   B: at a simple interest gaining 120,000 Ft annually.

   Is the investment B more advantageous?

   Solution:

   In the case A $1{,}000{,}000 \cdot 1.1^5 = 1{,}610{,}510$ Ft is the returned amount.
   In the case B $1{,}000{,}000 + 120{,}000 \cdot 5 = 1{,}600{,}000$ Ft. Hence the statement is false, B is not more advantageous than A.

3. Which investment is more advantageous A or B if we place a given amount of money in a bank.

   A: at 4% interest compounded annually in the first 3 years then the grown amount is put at 6% interest compounded annually at the end of the next 3 years;

   B: at 5% interest compounded annually in 6 years;

   C: at 6% interest compounded annually in the first 3 years then the grown amount is put at 4% interest compounded annually in the next 3 years?

   Solution: B

Figure 1 presents mathematics result of the same group of students in the last 2 years. The second author was their teacher in the last term. 17 students took part in the research and they learnt Thinking methods (Sets, logic and combinatory), Algebra and number theory, Geometry, Functions, elements of analysis.

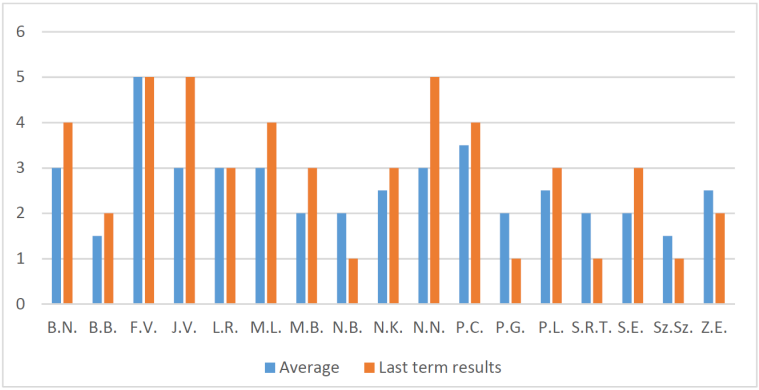


Figure 1: Results of students in Mathematics

Figure 2 shows that 59% of students improved their results at mathematics. Hence, most students progressed during the teaching experiment. Seeing these results we wanted to find out why the results of 29% of students got worse. In order to answer this question we asked students to fill in a questionnaire.

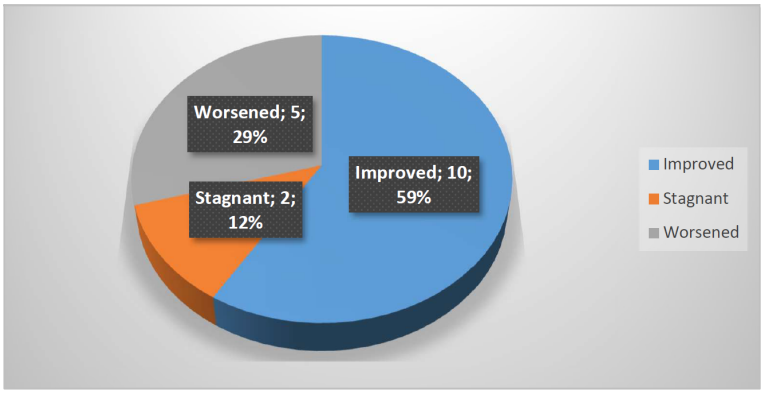


Figure 2: Number of students

5. Student questionnaire

It is not surprising that students with an overall low performance will not be considerably improved. Our aim was not to make easy tasks but to include tasks

in the form of situation games. The easier alternative, structured interviews based on selected topics, had to be given up because there were not a sufficient number of volunteers (exam period followed the term). However, interesting evidence is found in student questionnaires, done in writing during the exam period and with student replies being anonymous. Three questions concerned the lessons where situation games were used as a teaching method during the term. In the first, students were asked if they improved their results in the last term. Then in the second and third questions, students were asked to indicate reasons for the changes in their results and to comment how situation games contributed to their progress. Students could give open field responses. As many as 17 students have responded; they cannot be considered – representative but they represent strong opinion on the matter. Here are some examples of the students' responses (translated from Hungarian):

"My results did not improve. Situation games make lessons fun and interactive. I intend to use this method in the future."

"My results improved as I am more motivated to get a scholarship. I can use situation games in my future teaching practice."

"My results improved as I attended lessons more often. Situation games made mathematics more understandable. As the group is small we had more time to talk about what we did not understand."

"My results improved. Situation games involved us more into the topic. Mathematics makes much more sense, it we are playing an everyday situation."

"My results did not improve, however, I liked situation games. I find them useful and funny. The curriculum was processed differently and not with the usual boring methods."

"My results improved. We could learn Mathematics while playing and having a good time and everything was more understandable."

"My results did not get better, however, I found situation games very useful. Not only the numeracy skills, but other important competences like problem solving in real life situations can be improved."

The first author collected benefits and disadvantages of the method situation games from students in their third year of training performing teaching in a practicing primary school. In the first term the method of situation games was not focused on, however, it was shown to students via videos on the internet. In the second term students were asked to write lesson plans containing at least one mathematics problem to solve with situation games. Most students used this method during their teaching practice at the elementary practicing school during this research. There were two questions on the questionnaires as follows.

1. What are the benefits of the method situation games?

2. What are the disadvantages of it?

Here are some examples of the comments of 15 students (translated from Hungarian):

Benefits	**Disadvantages**
Situation games make lessons more enjoyable for pupils. This method makes mathematics more practical, less abstract. Situation games help the learner to understand and memorise concepts. Pupils can have practical life education. Thus, more competence areas can be developed.	It is time consuming and requires more preparation, a deeper reflection on the part of the teacher.
It is fun and easy to build team spirit. It helps to overcome shyness. It helps to learn how to share tasks and to prepare for real life. It teaches behavioural rules and classmates get to know each other better.	Good planning is needed.
It improves logical thinking and problem solving strategies. It helps socialisation. With experience-based situation games it is easier to gain knowledge.	Shy pupils find it more difficult to play rules and show themselves.
Attention-grabbing, interesting, and fun for the students. Furthermore, many interesting things can be built in the lesson (fairy tale stories, animals, etc.)	it requires lots of extra work, imagination and time. I do not see a lot of disadvantages in it.
Students love to play. They are actively learn and not just passively listen to the teacher. The teacher can differentiate while giving problems or choosing rules for students to play. A game can be designed for several problems or for different topics in the curriculum. A concept can be easily recalled with the game.	The success of situation games cannot be predicted for certain (students can argue about the rules or they do not understand the problem).
Students who never liked mathematics can develop a positive attitude toward mathematics.	I do not think that situation games would be a disadvantage, perhaps certain situation games require more space than we have in the classroom.
With life-like situations it is an extremely effective element of the lesson.	

Table 1

6. Concluding remarks

This paper contributes to our understanding of how difficult it is to master mathematics concepts, problem solving strategies and methodology for students to become successful primary school teachers. Reasons for this difficulty are cited in the first two sections.

In today's education, the knowledge-centred approach is still dominant, often lacking a system approach, real-life applications. In different situations, students interact with a real or simulated world to support their learning process. Their mind and body work together so that physical and mental activities reinforce learning. In a classroom or lecture theatre, the context enables students to learn from experience. By interpreting new information in the context of where and when it occurs and relating it to what we already learnt, students come to understand its relevance and meaning.

Situation games can be played formally or informally. An effective method is for a teacher to propose a question in the classroom, then for learners to explore that question at home or on a field trip collecting information, then share their findings back in the class to produce individual or group answers. Learning in informal settings can link educational content with everyday life.

Students can advance their understanding of science and mathematics by arguing in ways similar to professional scientists and mathematicians in different situations with educational purposes. Argumentation helps students attend to contrasting ideas, which can deepen their learning.

Incidental learning is unplanned or unintentional learning. It may occur while carrying out an activity that is seemingly unrelated to what is learned. Situation games provide many opportunities for incidental learning.

Another advantage of teaching in the form of situation games is to provide time for teachers to observe their students. Eye tracking and facial recognition help teachers in analysing how students learn. One thing is certain students are hungry for wisdom but they need to see how the educational content is benefit for them. If students are lacking the content they are about to learn how it would hurt their future success? The better a student is at Mathematics the better they'll be able to solve problems of everyday life or to invest their capital, create innovative businesses and reach their aims.

During role-playing games students can collect data about players' actions and strategies in order to present new challenges. This idea of applying knowledge in a simulated learning environment is now one of the 10 innovative learning strategies for modern pedagogy.

Results of our research show that mathematical knowledge of 59% of students has improved and even those students whose results got worse or remained stagnant find the innovative method 'situation games' very useful. We offered modern teaching methods developing students' critical thinking, problem solving and decision-making skills. However, repetition and memorisation of information to educate students cannot be avoided in today's education. As with most things, it

is all about balance. We need to understand when different methods work best and when it is right to try new and innovative approaches. The needs and work of students have to be studied more intensively than we were able to do it in this study.

References

[1] *10 Learning Strategies For Modern Pedagogy by TeachThought Staff*, 2019, URL: https://www.teachthought.com/the-future-of-learning/10-innovative-learning-strategies-for-modern-pedagogy/.

[2] M. Artigue: *The future of teaching and learning with digital technologies*, in: Mathematics education and technology-rethinking the terrain, New York, NY, USA: Springer, 2010, pp. 463–475.

[3] J. Carr: *Maths in Primary School*, 2011, URL: https://www.into.ie/ROI/Publications/MathsPrimarySchool.pdf.

[4] R. Ferguson, S. Barzilai, D. Ben-Zvi, et al.: *Innovating Pedagogy 2017*, 2018, URL: https://iet.open.ac.uk/file/innovating-pedagogy-2017.pdf.

[5] J. Gordon Győri, M. Halmos, K. Munkácsy, J. Pálfalvi: *A matematikatanítás mestersége - Mestertanárok a matematikatanításról/ The Art of Mathematics Teaching - Master Teachers in Mathematics Education*, Budapest: Gondolat Kiadó, 2007.

[6] J. Hadfield, C. Hadfield: *Simple Speaking Activities*, Oxford, United Kingdom: Oxford University Press, 2000.

[7] K. T. Henson: *Teaching Methods: History and Status*, Theory into Practice 1.2 (1980), pp. 2–5, DOI: 10.1080/00405848009542864.

[8] G. Makrides, A. Demetriou: *THALES^{TM} C & M – New programme for the development of competences and skills in mathematics: experimental results*, in: Proceedings of the 34th Pan-Hellenic Conference in Mathematics, New York, NY, USA, 2017.

[9] T. Michael: *Teaching Methods for Pupils with Low Mathematical Skills in Primary Schools*, Master's Thesis, Faculty of Educational Sciences, University of Oslo, 2013, URL: https://www.duo.uio.no/bitstream/handle/10852/36635/THESISxTULIA.pdf?sequence=1&isAllowed=y.

[10] J. Scrivener: *Learning Teaching*, Oxford: Macmillan, 2005.

[11] N. O. Yakovleva, E. V. Yakovlev: *Interactive teaching methods in contemporary higher education*, Pacific Science Review 16.2 (2014), pp. 75–80, DOI: 10.1016/j.pscr.2014.08.016.

Annales Mathematicae et Informaticae
***50** (2019) pp. 217–236*
DOI: 10.33039/ami.2019.11.003
http://ami.uni-eszterhazy.hu

Proving skills in geometry of secondary grammar school leavers specialized in mathematics

Ákos Győry[a], **Eszter Kónya**[b]

[a]Földes Ferenc Secondary School
gyoryakos@gmail.com

[b]University of Debrecen
eszter.konya@science.unideb.hu

Submitted: September 4, 2019
Accepted: November 19, 2019
Published online: December 17, 2019

Abstract

We examined the evolution of the van Hiele level of some study groups specialized in mathematics from 2015 to 2018, then selected two of these groups and measured the students' proof skills by Zalman Usiskin's proof test. We examined whether students were able to read from the text of the statement the given fact and the fact to be proved, whether they were able to draw a figure and, using the labels, whether they were able to perform a simple proof requiring 2-3 steps.

Keywords: van Hiele levels, reasoning and proving, specialized mathematics education, given fact, fact to be proved, role of figures.

MSC: D74, E54, G44.

1. Introduction

The elementary geometry is one of the most appropriate areas of mathematics for developing students' proving abilities, because it is complex and expressive. Constructing a reasoning chain consisting of 1-2 steps does not require hard abstraction; that is why this area can be studied and developed from grades 6-7.

Teaching geometry curriculum in secondary school includes many challenges for teachers and students alike, as solving geometric problems often goes beyond recalling well-practiced patterns. The first author's own experience shows that the most difficult part of geometry education is the development of students' reasoning and proving skills. He currently works as a secondary school mathematics teacher in Hungary and teaches students specialized in mathematics. The number of their mathematics lessons are more than average, and in addition to the normal requirements, they also acquire special topics. These students, who are particularly interested in mathematics, have to take an entrance exam in this form of education. Although they are talented in mathematics, they still have difficulty in solving problems that require proof. Their educational program will be discussed later.

Since 2005, there has been a two-tier graduation system in Hungary. This means that students can take higher or standard level school leaving examination. One of the main goals of the specialized mathematics education is to provide students with the adequate knowledge to pass an advanced school leaving examination. In this examination there have often appeared such tasks that require construction of a short, simple proof. Instructions of graduation and the framework of mathematics curriculum require from students to be able to produce an exact logical chain by means of their thoughts and aquired knowledge to solve some simple problems and to conceive and write the solution in a clear form. Therefore, it is of high priority to examine and analyze the difficulties that students face in solving tasks which aquire proof in order to be able to integrate experiences into the teaching process.

The present study is a part of a longitudinal research. In a previous study we followed up the evolution of van Hiele level of students specialized in mathematics from 2015 to 2018 (Győry & Kónya, 2018). Two of the study groups were selected for deeper examination, in which the mean of the students' van Hiele level reached 4, that is, according to the theory they were able to implement a few-step proof. Now we examine what proving ability they actually have, how they can formulate their thoughts in writing, and what typical mistakes they make. This article is about the obtained results which we will use in the future to accomplish a developmental teaching experiment.

2. Theoretical background

"The teaching of mathematical proof appears to be a failure in almost all countries, no matter how this teaching is organized... in USA mathematical proofs are taught only to students who take the geometry course..." (Balacheff, 2017, p. 1) Because of the importance of the topic, a number of studies has been made on the examination of students' reasoning-proving skills (Stylianides, 2008; G. Hanna & M. de Villiers, 2008; Balacheff, 2017, D. S. Hong & K. M. Choi, 2018). Perhaps the main difficulty is that most students consider the proofs construction only a necessary task required by the teacher. It is a long teaching process until the construction of proof becomes an intrinsic need of students, rather than just meeting

the theacher's expectation.

In our opinion one of the biggest mistakes a mathematics teacher can make is to teach concepts and statements without motivating which establishes them. Accordingly students need to be led to see definitions and statements not only as things prescribed by mathematicians, but to try to look for the root causes of them. To achieve this, the teacher has to educate the students on their own independent and critical thinking, which is esential for them to create a mathematical proof successfully. In connection with this, Lakatos wrote quite sharply in the '70s: "It has not yet been sufficiently realised that present mathematical and scientific education is a hotbed of authoritarianism and is the worst enemy of independent and critical thought." (Lakatos, 1976, p. 152)

Teaching of proofs is also a hard task because it requires, on the one hand, students to have a thorough knowledge of previous learning, additionally to be able to make a corollary based on some facts and to operate with abstract concepts. According to Ambrus (2004), the teaching of proof has three phases: (1) assumption of statements; (2) finding and executing an idea of proof based on previous proofs' strategies and methods; (3) description of the proof.

In Hungarian mathematics education the last two phases are the most emphasized, but in the special mathematics educational form the first phase is also expected. In this paper we deal with the last two phases in more detail. In elementary school, mathematical statements are often considered by students as absolute truth, which later makes it difficult to understand the need for mathematical proof. Only a minority of the students are then concerned with the question "Why?". It aggrevates the difficulty of the mathematics teacher that he/she needs use statements without proof in the teaching process, as students' understanding of mathematical concepts and their thinking are often not ready for the proper execution of a proof. The use of these so-called "school axioms" is essential for the proper development of thinking. (Szendrei, 2005) It is an additional problem for students to read clearly from the text of the statement the conditions and the fact to be proved. This is especially problematic when the statement is not given in the form "If ... then ...". (Ambrus, 2004)

Geometry tasks often include figures, but in most cases students have to draw their own figures. In the latter case, the student must correctly represent the concepts used in the task. This can be a pitfall in solving the problem, because a poorly drawn figure can steer the process of thinking in the wrong direction. A frequent mistake is that the student is able to spell out the definition correctly, but is unable to apply it in the solution. (Fishbein, 2012) If a figure is attached to the task, it serves to understand the task in some cases, that is to say it is an integral part of the task, but there are also cases where it serves merely to illustrate the task by reason of better understanding. However, even in the latter case, it is possible that the figure may lead to an unestablished or inaccurate conclusion for students. (Dvora, Dreyfus, 2004) We did not intend to investigate these aspects in the current research, but we will briefly touch on it in relation to one of the problems that suggests a false conclusion.

Balacheff distinguishes three categories of proof that are built on each other (Balacheff, 2017). These are the following:

- explanation: an individual intends to establish for somebody else the validity of a statement;
- proof: an explanation which is accepted by a community at a given time;
- mathematical proof: a proof which is accepted by mathematicians.

In secondary school, generally accepted proof is the second type described by Balacheff.

Stylianides categorized reasoning and proving in a so-called analytic framework in which he describes two types of proofs (Stylianides, 2008):

- generic example: it is a proof that uses a particular case seen as representative of the general case;
- demonstration: it is a proof that does not rely on the "representativeness" of a particular case.

In secondary school we use both methods of proving.

We relied on the van Hiele theory to examine proof skills. According to van Hiele theory, students' geometric thinking is evolved through sequential and hierarchical levels. Five different levels are distinguished and somebody can only reach the level n if he or she has achieved the criteria of all the levels below level n. In many papers, van Hiele levels are scaled from 0 to 4, but we will scale from 1 to 5. Achieving level 4 is a requirement for proving in secondary school. This is the level of formal deduction. At this level students understand the meaning of deduction. (Usiskin, 1982) According to the theory, at this level students are able to formulate causality, construct simpler proofs, and realise the need of proofs. They are aware of the concept of generalization, and they know and use different methods of proof (constructive, contrapositive, induction etc.). They are able to distinguish between necessary and sufficient condition. They are not yet able to provide a full proof and tend to treat statements requiring proof as fundamental truths. Based on our previous research (Győry & Kónya, 2018) the mean of van Hiele level of the examined students was around 4, so grounded in the theory and Usiskin's results we assumed that they were able to perform a 2-3 step proof.

3. Research question

In our research, we were curious about the reasoning ability of secondary school students talented in mathematics, how they express themselves in writing, and what typical mistakes they make.

4. Research methodology

4.1. The circumstances of the survey

The students took part in special mathematics program. This is an unusual form of education progresses in ten secondary schools in Hungary at the present time. Following the entrance examination, students study mathematics in small groups (12-20 students per group). A further feature of this program is that students study mathematics on average 6-8 hours per week, usually with two subject teachers per study group. One of the teachers teaches geometry and the other teaches algebra. One of the main aims of this form of education is to teach solving mathematical problems on their own. The students could be said to be motivated, talented, and many of them extremely outstanding. Students in the specialized mathematics program, on the one hand, acquire deeper mathematics skills and, on the other hand, learn certain topics faster than students in normal mathematics training. This form of education thus contains elements of the so-called acceleration and enrichment programs designed specifically for talented students. (Poli, 2018) We mention as an important difference that the Hungarian education system does not provide an opportunity for a specially talented student to learn subject of higher grade.

In a previous study we kept track of the Hiele levels of several study groups from 2015 to 2018. Two of the study groups were selected for deeper analysis. One group took part in a four-grade (Grade 9-12) and the other a six-grade (Grade 7-12) system of education. Hereinafter we will denote the four-class group with N, while the six-class group will be denoted by H. The number of examined students was 27, 14 of them from the group N and 13 from the group H. Initially, the members of the groups were more, but by reason of the longitudinal monitoring we only considered the performance of the students who wrote each test. The van Hiele levels of the two groups were already measured in grade 9, in 2015. The results achieved by the students were averaged. The mean for the group N was 3.80 and the mean for the group H was 4.25. (We scaled the van Hiele levels from 1 to 5.)

The result obtained shows that the students are roughly at level 4 of the hierarchical theory, which is the level of formal deduction. We repeatedly measured the van Hiele level of the same students over the years, and each time we got a mean of 3.5-4 (Győry & Kónya, 2018). Since the mean of van Hiele levels of the two groups was quite similar, the results of the measurements will not be devided into groups, but will be aggregated.

To assess proof skills we took as a basis the proof test can be found in the paper van Hiele Levels and Achievement in Secondary School Geometry by Zalman Usiskin (1982). The proof test from this article was conducted in April 2018, when both groups were in 12th grade in some weeks before the final examination.

4.2. The proof test

Usiskin and his colleagues, who dealt extensively with the van Hiele theory, were curious about how this theory can describe and predict the geometric achievement of secondary school students, including proving, and at what level students are able to describe their proof. They found that at level 4 students are able to independently create simpler proofs, whereas in case of lower van Hiele levels they are not. That is, the van Hiele level is a good predictor. In our work, we do not aim to discover the relationship between the 4th van Hiele level and proof skills. Based on the results of Usiskin, we assumed that our students would be able to perform some simple proofs consisting of one or two steps at this level. After selecting one of the 3 proof tests in Usiskin's paper (Usiskin, 1982, pp. 173–177.), we examined systematic mistakes, proof ideas, how to draw figures and the written communication.

The test consists of 6 exercises, and requires the following prior knowledges from geometry curriculum in Hungary up to grade 10.

- Knowledge of angle pairs.
- Basic properties of triangles. Basic cases of congruence and similarity of triangles.
- Knowledge and use of Pythagorean theorem.
- Knowledge of properties of parallelograms.

It should be noted that the first task of the original test asked students to make solution in the two-column style prevalently used in the US. As this method is not well known in Hungary and is completely unknown to students, we did not ask for a two-column description of the proof in the Hungarian translation in contrast with the original version.[1]

The writing conditions were very similar to those of the Usiskin. We differred only in one case from them: our working time was 45 minutes in contratry to the 35 minutes. We did this because, on the one hand, we did not give help steps in our first task, and on the other hand, we feared that students might run out of time, which would have significantly affected the in-depth analysis of the tasks.

5. Discussion

We will discuss in detail three of the six tasks: Task 2, Task 4 and Task 6. Why did we just choose these tasks?

1. The second of the six tasks is the only one that does not have to be proved, and no figure is attached to it. Furthermore, the formulation of this statement differs from all other tasks' one. The condition and the corollary are not given

[1]The English version of the test we have written can be found in the Appendix.

in two separate sentences, but in a single "If ... then ..." type sentence. In case of all other tasks have been attached figures and the statement must be proved.

2. Task 4 is the only one where the attached figure is only for illustrative purposes and does not play a role in understanding and may even be omitted.

3. The structure of tasks 1, 3, 5 and 6 is completely the same. In these statements, the condition and the corollary are included in separate sentences, and there is a figure in all of them that is necessary to understand the text of the statement. We chose Task 6 of these because the figure may even suggest false information.

Task 2

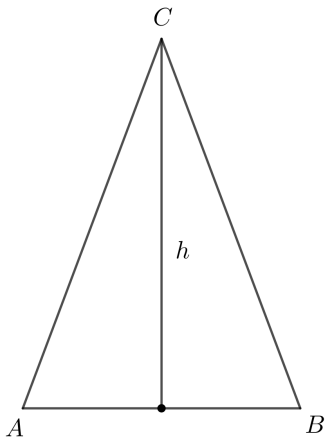
Statement: If an altitude is drawn to the base of an isosceles triangle, then it bisects the vertex angle.

a. Draw a figure and label it.

b. Write, in terms of your figure, what is given and what is to be proved in this statement.

The solution

a.



b. Given: $CA = CB$; $CT \perp AB$.
To prove: $ACT\sphericalangle = BCT\sphericalangle$.

What were we able to examine by means of this task?

(1) Is the student able to draw a figure for the task?

(2) Is the student able to read from the text what is given and what is to be proved, i.e. what fact do we infer from which fact (height $\Longrightarrow$ bisector)?

(3) Is the student able to use symbols properly?

We can see that Task 2 was given in the form "If ... then ...", so the condition and the corollary were detached well. This made the students' job a little easier.

Results

About a quarter of the students could only answer correctly (Table 1).

	Good solution	Wrong solution
Number of students	7	20

Table 1: Results (Task 2)

Let's look at the statistics for the parts of the task one after the other.

(1) The students did not have a problem with drawing the figures and introducing the labels.

(2) Reading the given fact (condition) from the text (Table 2).

	Good solution	Wrong solution
Number of students	10	17

Table 2: What is given? (Task 2)

Only about one-third of the students were able to solve this part. We found a typical mistake: more than two-thirds of the students who gave the wrong solution could not separate the condition, the corollary and the statement itself, and confused them (Table 3).

	To be written what to be proved	To be written the whole statement	Other
Number of students	6	6	5

Table 3: Wrong sulitions (Task 2)

(3) Reading the fact must be proved from the text (Table 4).

	Good solution	Wrong solution
Number of students	18	9

Table 4: What to prove? (Task 2)

Already two thirds of the students have succeeded in this section.

As a typical mistake we could note that most students confuse the concept of the given fact, the fact to be proved and the statement, as to be shown in Ádám's solution below.

Given: "If the triangle ABC is isosceles, then m_c bisects the side "c" (splits it into 2 pieces of "x" parts); The legs are of equal length."[2]

To be proved: "The above assumption is true only if the angle at C is bisects by altitude (2 pieces of α angle are created)."

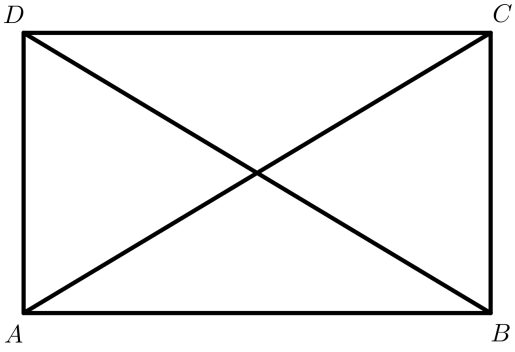
Conclusion

1. There is a need to clarify the distinction between the condition and what is to be proved from each other and from the statement itself.
2. Emphasis should be placed on transforming a statement into an "If ... then ..." sentence throughout the teaching process.

Task 4

We know about the following figure: $ABCD$ is a rectangle.

Prove that the diagonals are congruent.



Use the labels to describe what is given.

Use the labels to describe what is to be proved.

[2]In Hungary m_c denotes the altitude to the side c.

A possible solution:

Given: $AB = CD$ and $DAB\sphericalangle = 90°$.

To be proved: $AC = BD$.

Proof:
The DAB and CBA triangles are congruent because two pairs of sides of these triangles and the included angles are equal ($AB = BA$, $AD = BC$, $DAB\sphericalangle = CBA\sphericalangle$).

So the third sides of the two triangles are equal in length: $DB = AC$, which we just wanted to prove.

What were we able to examine by means of this task?

1. Is the student able to read from the text of the statement what is given and what is to be proved, i.e. what fact do we infer from which fact and
2. is he/she able to describe these with his/her own notions?
3. Is the student able to create a simple reasoning?

Results

1. Reading the given fact from the text (Table 5).

	Good solution	Incompleted solution
Number of students	17	10

Table 5: What is given? (Task 4)

Students who did not solve this part flawlessly, without exception, forgot about the angles (that is, each angle of the quadrilateral ABCD is a right angle), but they demanded the equality of the opposite sides. Probably this little mistake was made due to the figure attached to the problem, because of it they considered the equality of angles as obvious fact.

It is important to note that this task is very different in its formulation from Task 2, as in this case the condition and the corollary are given in separate sentences. Based on our findings, it can be claimed that if the task is set in this style, the students are able to read the condition from the text of the task.

2. Reading the fact must be proved from the text. (Table 6)

	Good solution	Wrong solution
Number of students	24	3

Table 6: What to prove? (Task 4)

Overall, this part did well for the students probably due to the style of the formulation.

3. Execution of proof (Table 7).

	Good solution with Pythagorean theorem	Good solution with elementary geometry way	Nearly good solution	Wrong solution
Number of students	17	4	2	4

Table 7: The way of proving (Task 4)

We can say that the students passed this subtask successfully, because 21 out of 27 students reasoned correctly (Table 7). It is interesting to note that only 4 of the 21 students chose purely geometric proofs, while the other 17 reasoned with Pythagorean theorem, suggesting an algebraic approach. We conceive it is due to the fact that students associate the Pythagoras theorem with a right triangle immediately because of its central role in secondary school calculations. Kitti's solution demonstrates this way of proving:

"$BD^2 = BC^2 + DC^2, AC^2 = AB^2 + BC^2$.
Since $DC = AB \Longrightarrow AC^2 = DC^2 + BC^2 \Longrightarrow BD^2 = AC^2$.
Since BD and AC are also positive, hence $BD = AC$."

We considered two solutions as "nearly good", because the students stopped, though there was only one step left to complete the proof.

Of the 3 students who were not able to read the corollary from the statement (Table 6), 2 students were not able to successfully complete the proof either. The additional student proved also well, but he described his proof to the part to be proved. So he understood what the task demanded from him as proof, but he was probably careless or the formulation was perhaps unusual to him. Of the 4 students , who could not prove the statement (Table 7), 2 students were able to separate the fact to be proved from the statement well. We have thus got that it is essential for the correct implementation of the proof that the individual can sharply separate the part to be proved from the statement.

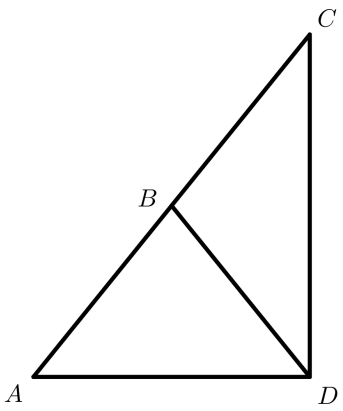
Of the students who were unable to prove the statement (4 in total, Table 7), two students abandoned the solution after one step, one student initially started from the fact to be proved, and one student tried to write trigonometric relations, no avail.

Conclusion

- If we give a statement for the student in "separate" style, that is, the given fact and the fact to be proven are expressed in separate sentences, they will be able to read out the condition and the corollary from the statement. This means that the formulation "If ... then ..." is also worth splitting at first into two sentences.
- The students use his labels well in the formulation of the condition and the corollary.
- A simple reasoning of no more than 2-3 steps was not a problem for the students. This ascertaining is consistent with the result of Usiskin, who obtained similar results for students at van Hiele level 4.

Task 6

We know about the following figure: B is the midpoint of the AC section. $AB = BD$. Prove that the angle CDA is a right angle.



A possible solution:

According to the conditions: $AB = BD = BC$.

This means that the triangles ABD and DCB are isosceles.

Since in an isosceles triangle the angles opposite the legs are equal, accordingly $BAD\sphericalangle = BDA\sphericalangle$ and $BDC\sphericalangle = BCD\sphericalangle$.

From this we can conclude that

$$CDA\sphericalangle = BDA\sphericalangle + BDC\sphericalangle = \frac{2 \cdot BDA\sphericalangle + 2 \cdot BDC\sphericalangle}{2} =$$
$$= \frac{BAD\sphericalangle + BDA\sphericalangle + BDC\sphericalangle + BCD\sphericalangle}{2} = \frac{CAD\sphericalangle + ADC\sphericalangle + ACD\sphericalangle}{2} =$$

$$= \frac{180^\circ}{2} = 90^\circ.$$

The structure of the text is similar to that of Task 4, that is, the given fact and the fact to be proved are clearly separated in two sentences.

What were we able to examine by means of this task?

- Is the student able to interpret the figure?
- Is the student able to construct a proof which consists more than 3 simple steps?
- How do the students reason? What do they refer to?
- What level of detail do the students reason? How do they use the language of mathematics?

Results

How did they manage to create the proof? (Table 8)

	Good solution	Wrong solution
Number of students	13	14

Table 8: Creating the proof (Task 6)

Overall, about half of the students succeeded in making the proof. In order to draw conclusions, we analyzed the good and the wrong solutions.

Good solutions (13 students)

The proof was correctly described by 10 students. It is important to note that none of these students referred to the fact that "there are equal angles opposite congruent sides". This is the point, where the problem already mentioned in the introduction arises, which is one of the dilemmas of mathematics education in secondary school. Namely, how to distinguish between statements requiring proof and so-called basic truths to be treated as fact.

One of the 13 students described the proof in a fairly short way: he added to the given figure a semicircle with centre B and with radius BC (which lies on the points A, D, and C due to the condition), and then wrote that "the angle CDA is right angle by reason of Thales' theorem."

There were 2 students who worked on the figure. In such cases the train of thought of a proof is difficult to follow, not to mention that in the case of a false reasoning it is difficult to determine where a failure was commetted.

Figure 1 shows István's solution, where the chain of reasoning is relatively easy to follow.

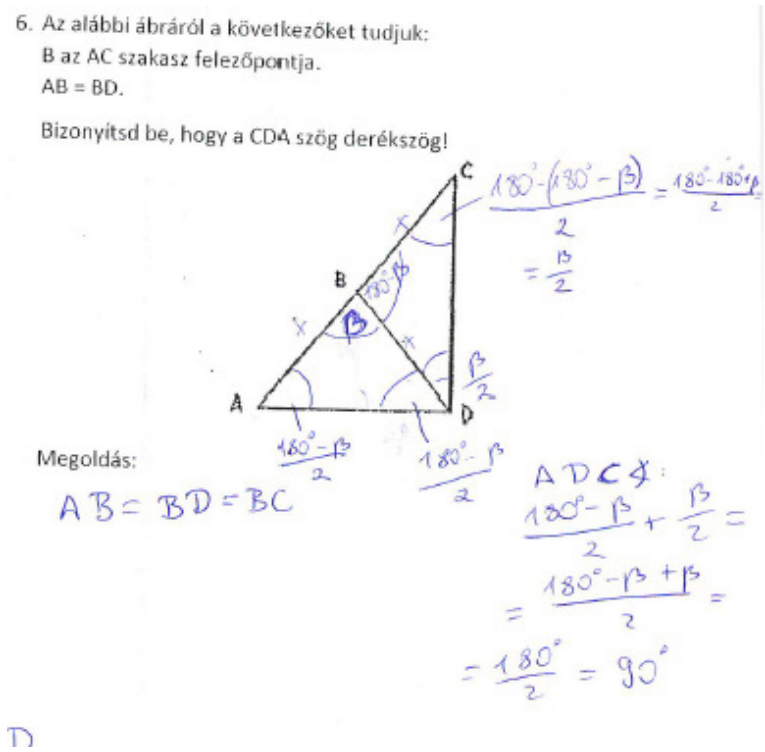


Figure 1: Solution on the figure

The student's steps could have been as follows (repeatedly using the sum of the interior angles of a triangle is 180°):

1. He described the given fact: $AB = BD = BC$, then denoted the length of these segments in the figure by x.
2. He denoted the measure of the angle ABD by β.
3. Since the triangle ADB is isosceles, its angles at A and D are equal: $\frac{180^\circ-\beta}{2}$.
4. Because the angles at B are adjacent angles, hence $DBC\sphericalangle = 180^\circ - \beta$.
5. The triangle DCB is isosceles, so its angles at D and C are equal: $\frac{180^\circ-(180^\circ-\beta)}{2} = \frac{\beta}{2}$.
6. Since he had already determined the two angles at D, he wrote that $ADC\sphericalangle = \frac{180^\circ-\beta}{2} + \frac{\beta}{2} = \frac{180^\circ-\beta+\beta}{2} = \frac{180^\circ}{2} = 90^\circ$.

Although he did not write it, he proved the claim.

Wrong solutions (14 students)

Three of the 14 students did not begin the solution, left the sheet blank. There was one student who began the solution in a good way, but after a certain point, he stopped reasoning. In the case of two students, the train of thought could not be followed and the reasoning described was chaotic. However, for the remaining eight students, we observed a typical, repeating mistake. These students drew a false inference grounded in the figure, namely equality of certain angles ($BAD\sphericalangle = BDA\sphericalangle = BDC\sphericalangle = BCD\sphericalangle$). This is illustrated by Éva's solution in Figure 2.

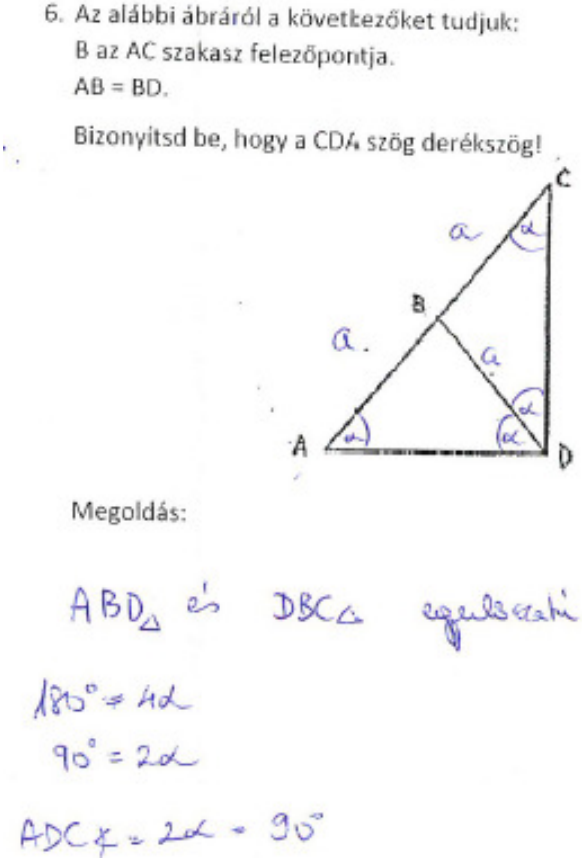


Figure 2: A typical wrong solution[3]

[3] ABD and DBC triangles are isosceles.

Conclusions

- During the teaching process, the teacher must make sure that the students do not just write the proof in the figure, but use the appropriate notations to describe their reasoning in adequate detail.
- It is necessary to clarify with the students, what facts require reasoning and what facts do not in a proof. Anyway this is a very problematic element of the teaching process.
- It should be emphasized that information suggested by the figure should not be accepted without doubt. This is well illustrated by the typical mistake of solutions of Task 6 (Figure 2).

6. Summary

The research described in this article is a part of a comprehensive series of experiments. As a first phase, we examined the evolution of the Hiele level of several study groups specialized in mathematics from 2015 to 2018. Two of the study groups were selected and we measured the students' proof skills, in 2018, by Usiskin's proof test.

During teaching proofs, strong emphasis must be placed on the ability of students to pick out the given fact and the fact to be proven from the statement. To do this, it seems to be to follow the sequence below.

(1) The teacher should give the statement to the students in such a form that the given fact and the fact to be proven are in two separate sentences. For example: "Let's consider a rhombus. Prove that its diagonals are perpendicular to each other."

(2) The teacher should give the statements in the "If... then..." structure to the students and should clarify where the given fact and where the fact to be proved appear. It is worthwhile to reword the statements, which are given in such a way like we saw in (1), with the students themselves. "If a quadrilateral is a rhombus, then its diagonals are perpendicular to each other."

Of course, statements made with the help of quantifiers and other means also occur in lessons. ("Every rhombus' diagonals are perpendicular to each other." "The diagonals of a rhombus are perpendicular to each other." Etc.). The study of these cases will be a part of a later study.

Students in specialized mathematics education, reaching van Hiele level 4, are able to complete a simple proof, which requires at most 2 or 3 steps, but many of them have problems with multi-step proofs, so these need much more practice.

The teacher should pay special attention to the figures, too, in teaching of proofs, as well. It should be made clear to the students that they ought to make a clear figure with appropriate labels. The teacher should emphasize that students ought not to accept the facts suggested by the figure unless they have actually

reasoned the validity of these facts. The important thing is that the teacher should make sure that the students do not write the proof only in the figure, but that they formulate their thoughts based on the labels in the figure and write them down using mathematical language.

One of the most difficult parts of teaching of proofs is that the teacher clarifies with students what fact needs to be reasoned and what fact does not in secondary school. This is, of course, a long process, which demands a huge problem-solving routine that can only be achieved through lots of practice.

Based on the results, we also designed a series of developmental experiments in a study group, specialized in mathematics, at grade 9 (this group takes part in 4-form education, ie. this is its first school year in the secondary school).

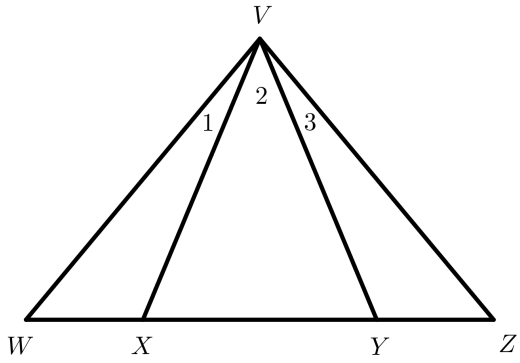
7. Appendix

Tasks of our version of the proof test.

1. From the figure below we know the following:

   The angle at W is equal to the angle at Z and the length of the segment WX is equal to the length of the segment YZ.

   Prove that the angle denoted by 1 is equal to the angle denoted by 3!



   Solution:

2. *Statement*: If an altitude is drawn to the base of an isosceles triangle, then it bisects the vertex angle.

   a. Draw a figure and label it.

   b. Write, in terms of your figure, what is given and what is to be proved in this statement.

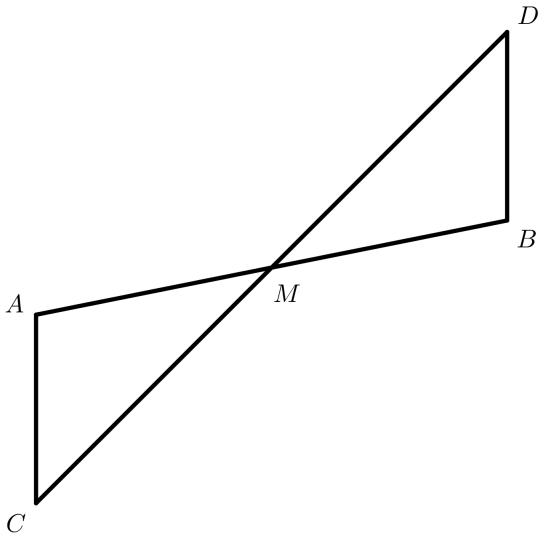
   Figure:

   Given:

To prove:

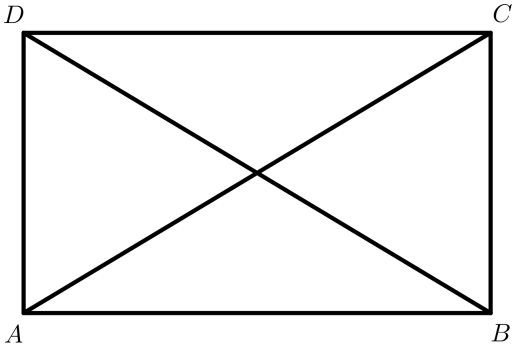
Do not prove the statement.

3. From the figure below we know the following:
   M is the midpoint of the segment AB.
   M is the midpoint of the segment CD.
   Prove that the triangles ACM and BDM are similar!



Solution:

4. We know about the following figure: $ABCD$ is a rectangle
   Prove that the diagonals are congruent.



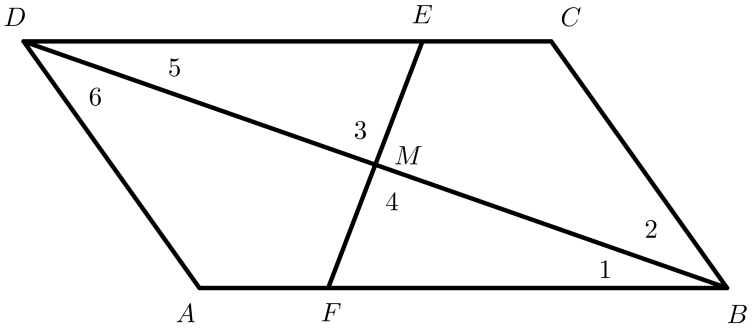
Solution:

Use the labels to describe what is given.

Use the labels to describe what is to be proved.

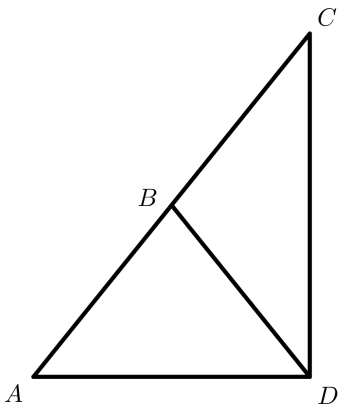
Proof:

5. From the following figure we know: $AB = DC, AD = BC$. M is the midpoint of the segment DB. M lies on the segment EF. Prove that $FM = ME$!



Solution:

6. We know about the following figure: B is the midpoint of the AC section. $AB = BD$. Prove that the angle CDA is a right angle.



Solution:

References

[1] A. Ambrus: *Introduction to Didactics of Mathematics (In Hungarian: Bevezetés a matematikadidaktikába)*, Budapest: ELTE Eötvös Kiadó, 1995.

[2] N. Balacheff: *A Study of Students' Proving Processes At The Junior High School Level*, 66th NCTM Annual Meeting, 2017.

[3] T. Dvora, T. Dreyfus: *Unjustified Assumptions Based On Diagrams In Geometry*, Proceedings of the 28th Conference of the International, Group for the Psychology of Mathematics Education 2 (2004), pp. 311–318.

[4] E. Fischbein: *The Theory of Figural Concepts*, Educational Studies in Mathematics 24.2 (1993), pp. 139–162.

[5] E. Győry Á. Herendiné Kónya: *Development of high school students' geometric thinking with particular emphasis on mathematically talented students*, Teaching Mathematics and Computer Science 16.1 (2018), pp. 93–110.

[6] G. Hanna, M. de Villiers: *ICMI Study 19: Proof and Proving In Mathematics Education*, ZDM MAthematics Education 40.2 (2008), pp. 329–336.

[7] E. Herendiné Kónya: *The characterictics of the geometric thinking of teacher trainees (In Hungarian: A tanítójelöltek geometriai gondolkodásának jellegzetességei)*, Iskolakultúra 12 (2003), pp. 51–61.

[8] D. S. Hong, K. M. Choi: *Reasoning and Proving Opportunities in Textbooks: A Comparative Analysis Reasoning and proving opportunities in textbooks: A comparative analysis*, International Journal of Research in Education and Science (IJRES) 4.1 (2018), pp. 82–97.

[9] T. Kettler, M. Curliss: *Mathematical Acceleration In A Mixed-Ability Classroom*, Gifted Child Today 26.1 (Winter 2003).

[10] I. Lakatos: *Proofs and refutations. The logic of mathematical discovery*, Cambridge: Cambridge University Press, 1976.

[11] M. D. Poli: *Successful Programs and Strategies for Secondary Students Who Are Gifted and in Mathematics Classes: A Qualitative Study*, Doctoral dissertation, Indiana University of Pennsylvania, USA, 2018.

[12] G. J. Stylianides: *An Analytic Framework of Reasonin-and-Proving*, For the Learning of Mathematics, FLM Publishing Association 28.1 (2008), pp. 9–16.

[13] L. Surányi: *"I have experienced few such a motivating atmosphere" The 50th anniversary of special mathematics education (In Hungarian: "Kevés ilyen inspiráló légkört tapasztaltam" Ötvenéves a speciális matematika tagozat.)* Természet Világa 143.6 (2012).

[14] Z. Usiskin: *Van Hiele Levels and Achievement in Secondary School Geometry*, University of Chicago (1982).