**HACETTEPE JOURNAL OF MATHEMATICS AND STATISTICS**

**Volume 50 Issue 1 (2021)**

E-ISSN: 2651-477X

# Contents

**Mathematics**

*Research Articles*

**Statistics**

*Research Articles*

Hacettepe Journal of
Mathematics & Statistics

Hacet. J. Math. Stat.
Volume 50 (1) (2021), 1 – 13
DOI : 10.15672/hujms.555416

Research Article

# A new class of generalized polynomials involving Laguerre and Euler polynomials

Nabiullah Khan[1], Talha Usman[2], Junesang Choi*[3]

[1] *Department of Applied Mathematics, Faculty of Engineering and Technology, Aligarh Muslim University, Aligarh 202002, India*
[2] *Department of Mathematics, School of Basic and Applied Sciences, Lingaya's Vidyapeeth, Faridabad 121002, Haryana, India*
[3] *Department of Mathematics, Dongguk University, Gyeongju 38066, Republic of Korea*

## Abstract

Motivated by their importance and potential for applications in a variety of research fields, recently, numerous polynomials and their extensions have been introduced and investigated. In this paper, we modify the known generating functions of polynomials, due to both Milne-Thomsons and Dere-Simsek, to introduce a new class of polynomials and present some involved properties. As obvious special cases of the newly introduced polynomials, we also introduce power sum-Laguerre-Hermite polynomials and generalized Laguerre and Euler polynomials and give certain involved identities and formulas. We point out that our main results, being very general, are specialised to yield a number of known and new identities involving relatively simple and familiar polynomials.

**Mathematics Subject Classification (2020).** 05A10, 05A15, 11B68

**Keywords.** Milne-Thomsons polynomials, Dere-Simsek polynomials, Laguerre polynomials, Hermite polynomials, Euler polynomials, generalized Laguerre-Euler polynomials, summation formulae, symmetric identities

## 1. Introduction and preliminaries

The two variable Laguerre polynomials $L_n(x,y)$ are generated by (see [8, 18])

$$\frac{1}{1-yt}\exp\left(\frac{-xt}{1-yt}\right)=\sum_{n=0}^{\infty}L_n(x,y)\,t^n\quad(|yt|<1). \tag{1.1}$$

Also, equivalently, the polynomials $L_n(x,y)$ are given by (see [9, 18])

$$\mathsf{e}^{yt}\,C_0(xt)=\sum_{n=0}^{\infty}L_n(x,y)\frac{t^n}{n!}, \tag{1.2}$$

*Corresponding Author.
Email addresses: nukhanmath@gmail.com (N.U. Khan), talhausman.maths@gmail.com (T. Usman), junesang@dongguk.ac.kr (J. Choi)
Received: 19.04.2019; Accepted: 05.04.2020

where $C_0(x)$ denotes the 0th order Tricomi function. The $n$th order Tricomi functions $C_n(x)$ are generated by

$$\exp\left(t-\frac{x}{t}\right)=\sum_{n=0}^{\infty} C_n(x)\, t^n \quad (t\in\mathbb{C}\setminus\{0\},\, x\in\mathbb{C}). \tag{1.3}$$

We have

$$C_n(x)=\sum_{r=0}^{\infty}\frac{(-1)^r x^r}{r!(n+r)!} \quad (n\in\mathbb{N}_0). \tag{1.4}$$

The Tricomi functions $C_n(x)$ are connected with the Bessel function of the first kind $J_n(x)$ (see [7]):

$$C_n(x)=x^{-\frac{n}{2}}J_n(2\sqrt{x}). \tag{1.5}$$

Here and throughout, we denote $\mathbb{C}$, $\mathbb{R}$, $\mathbb{R}^+$, $\mathbb{Z}$, and $\mathbb{N}$ by the sets of complex numbers, real numbers, positive real numbers, integers, and positive integers, respectively, and let $\mathbb{N}_0:=\mathbb{N}\cup\{0\}$.

From (1.2) and (1.4), we find

$$L_n(x,y)=n!\sum_{s=0}^{n}\frac{(-1)^s x^s y^{n-s}}{(s!)^2(n-s)!}=y^n L_n(x/y), \tag{1.6}$$

where $L_n(x)$ are the ordinary Laguerre polynomials (see, e.g., [1,26]). We thus have

$$L_n(x,0)=\frac{(-1)^n x^n}{n!},\quad L_n(0,y)=y^n,\quad L_n(x,1)=L_n(x). \tag{1.7}$$

Milne-Thomson [22] defined polynomials $\Phi_n^{(\alpha)}(x)$ of degree $n$ and order $\alpha$ by the following generating function

$$f(t,\alpha)\,\mathsf{e}^{xt+g(t)}=\sum_{n=0}^{\infty}\Phi_n^{(\alpha)}(x)\,\frac{t^n}{n!}, \tag{1.8}$$

where $f(t,\alpha)$ is a function of $t$ and $\alpha\in\mathbb{Z}$ and $g(t)$ is a function of $t$. Then, by choosing some explicit functions of $f(t,\alpha)$ and $g(t)$, Milne-Thomsons [22] presented several interesting properties for polynomials such as Bernoulli polynomials and Hermite polynomials.

Derre and Simsek [10] made a slight modification of the Milne-Thomson's polynomials $\Phi_n^{(\alpha)}(x)$ to give polynomials $\Phi_n^{(\alpha)}(x,\nu)$ of degree $n$ and order $\alpha$ by means of the following generating function

$$G(t,x;\alpha,\nu):=f(t,\alpha)\,\mathsf{e}^{xt+h(t,\nu)}=\sum_{n=0}^{\infty}\Phi_n^{(\alpha)}(x,\nu)\frac{t^n}{n!}, \tag{1.9}$$

where $f(t,\alpha)$ and $h(t,\nu)$ are functions of $t$ and $\alpha\in\mathbb{Z}$ and $t$ and $\nu\in\mathbb{N}_0$, respectively, which are analytic in a neighborhood of $t=0$. Observe that $\Phi_n^{(\alpha)}(x,0)=\Phi_n^{(\alpha)}(x)$ (see, for details, [22]).

By setting $f(t,\alpha)=\left(\frac{t}{\mathsf{e}^t-1}\right)^{\alpha}$ in (1.9), in [18], we introduced the polynomials $B_n^{(\alpha)}(x,\nu)$ defined by

$$\left(\frac{t}{\mathsf{e}^t-1}\right)^{\alpha}\mathsf{e}^{xt+h(t,\nu)}=\sum_{n=0}^{\infty}B_n^{(\alpha)}(x,\nu)\frac{t^n}{n!}. \tag{1.10}$$

Here, by choosing $f(t,\alpha)=\left(\frac{2}{\mathsf{e}^t+1}\right)^{\alpha}$ in (1.9), we introduce the following polynomials $E_n^{(\alpha)}(x,\nu)$ defined by

$$\left(\frac{2}{\mathsf{e}^t+1}\right)^{\alpha}\mathsf{e}^{xt+h(t,\nu)}:=\sum_{n=0}^{\infty}E_n^{(\alpha)}(x,\nu)\frac{t^n}{n!}. \tag{1.11}$$

We find that the polynomials $E_n^{(\alpha)}(x,\nu)$ are related to not only Euler polynomials but also the Hermite polynomials. For example, if $h(t,0)=0$ in (1.11), we have

$$E_n^{(\alpha)}(x,0)=E_n^{(\alpha)}(x)$$

where $E_n^{(\alpha)}(x)$ denote the Euler polynomials of higher order defined by means of the following generating function (see, e.g., [27, p. 88])

$$F_E(t,x;\alpha):=\left(\frac{2}{\mathrm{e}^t+1}\right)^{\alpha}\mathrm{e}^{xt}=\sum_{n=0}^{\infty}E_n^{(\alpha)}(x)\frac{t^n}{n!}. \tag{1.12}$$

We find

$$F_E(t,0;\alpha):=F_E(t;\alpha)=\left(\frac{2}{\mathrm{e}^t+1}\right)^{\alpha}=\sum_{n=0}^{\infty}E_n^{(\alpha)}\frac{t^n}{n!}, \tag{1.13}$$

where $E_n^{(\alpha)}$ are generalized Euler numbers. For more information about Euler numbers and Euler polynomials, we refer the reader, for example, to [3, 20, 21, 27].

Taking $h(t,y)=yt^2$ in (1.11), we get the generalized Hermite-Euler polynomials of two variables ${}_HE_n^{(\alpha)}(x,y)$ introduced by Pathan [23]:

$$\left(\frac{2}{\mathrm{e}^t+1}\right)^{\alpha}\mathrm{e}^{xt+yt^2}=\sum_{n=0}^{\infty}{}_HE_n^{(\alpha)}(x,y)\frac{t^n}{n!}. \tag{1.14}$$

Note that the polynomials ${}_HE_n^{(\alpha)}(x,y)$ generalize Euler numbers, Euler polynomials, Hermite polynomials, and Hermite-Euler polynomials ${}_HE_n(x,y)$ introduced by Dattoli et al. [6, p. 386, Eq. (1.6)]:

$$\frac{2}{\mathrm{e}^t+1}\,\mathrm{e}^{xt+yt^2}=\sum_{n=0}^{\infty}{}_HE_n(x,y)\frac{t^n}{n!}. \tag{1.15}$$

The sum of integer power (simply, power sum)

$$S_k(\mathtt{n}):=\sum_{j=0}^{\mathtt{n}}j^k \quad (k\in\mathbb{N}_0;\ \mathtt{n}\in\mathbb{N})$$

is generated by

$$\sum_{k=0}^{\infty}S_k(\mathtt{n})\frac{t^k}{k!}=1+\mathrm{e}^t+\mathrm{e}^{2t}+\cdots+\mathrm{e}^{\mathtt{n}t}=\frac{\mathrm{e}^{(\mathtt{n}+1)t}-1}{\mathrm{e}^t-1}. \tag{1.16}$$

Luo et al. [20, 21] introduced the generalized Euler numbers $E_n(a,b)$ generated by

$$\Phi(t;a,b)=\frac{2}{a^t+b^t}=\sum_{n=0}^{\infty}E_n(a,b)\frac{t^n}{n!} \tag{1.17}$$

$$\Big(|t|<2\pi;\ n\in\mathbb{N}_0;\ a,\,b\in\mathbb{R}^+\ \text{ with }\ a\neq b\Big).$$

Also, Luo et al. [20] introduced the generalized Euler polynomials $E_n(x;a,b,\mathrm{e})$ generated by

$$\Phi(x,t;a,b,\mathrm{e})=\frac{2\mathrm{e}^{xt}}{a^t+b^t}=\sum_{n=0}^{\infty}E_n(x;a,b,\mathrm{e})\frac{t^n}{n!} \tag{1.18}$$

$$\Big(|t|<2\pi;\ n\in\mathbb{N}_0;\ a,\,b\in\mathbb{R}^+\ \text{ with }\ a\neq b\Big).$$

The 2-variable Hermite-Kampé de Fériet polynomials $H_n(x,y)$ (see [2, 6]) are generated by

$$\mathrm{e}^{xt+yt^2}=\sum_{n=0}^{\infty}H_n(x,y)\frac{t^n}{n!}. \tag{1.19}$$

Note that

$$H_n(x,y) = n! \sum_{r=0}^{[\frac{n}{2}]} \frac{y^r x^{n-2r}}{r!(n-2r)!} \tag{1.20}$$

and $H_n(2x,-1) = H_n(x)$ are the ordinary Hermite polynomials (see, e.g., [2]; see also [26, Chapter 11]). Dere and Simsek [10] generalized the polynomials $H_n(x,y)$ in (1.19) to define two variable Hermite polynomials $H_n^{(\ell)}(x,y)$ by the following generating function

$$\mathsf{e}^{xt+yt^\ell} = \sum_{n=0}^{\infty} H_n^{(\ell)}(x,y)\, \frac{t^n}{n!} \quad (\ell \in \mathbb{N} \setminus \{1\}). \tag{1.21}$$

Very recently, Khan et al. [18, Eq. (20)] have introduced and investigated the following generalized Laguerre-Bernoulli polynomials

$$\left(\frac{t}{a^t - b^t}\right)^{\alpha} \mathsf{e}^{yt+zt^2}\, C_0(xt) = \sum_{n=0}^{\infty} {}_L B_n^{(\alpha)}(x,y,z;a,b,\mathsf{e}) \frac{t^n}{n!} \tag{1.22}$$

$$\left(\alpha,\, x,\, y,\, z \in \mathbb{C},\, a,\, b \in \mathbb{R}^+,\, a \neq b,\, |t| < \frac{2\pi}{|\ln a - \ln b|}\right).$$

Motivated by their importance and potential for applications in certain problems in number theory, combinatorics, classical and numerical analysis and other fields of applied mathematics, a number of certain numbers and polynomials, and their generalizations have recently been extensively investigated (see, e.g., [1–30]). Here, we also make a slight modification of Milne-Thomson polynomials $\Phi_n^{(\alpha)}(x)$ in (1.8) and Derre and Simsek polynomials $\Phi_n^{(\alpha)}(x,\nu)$ in (1.9) to define polynomials $\Phi_{n,\ell}^{(\alpha)}(x,y,\nu)$ by the following generating function

$$H(t,x,y;\alpha,\nu) := f(t,\alpha)\, \mathsf{e}^{xt+yt^\ell + h(t,\nu)} = \sum_{n=0}^{\infty} \Phi_n^{(\alpha,\ell)}\,(x,y,\nu)\, \frac{t^n}{n!} \tag{1.23}$$

$$(x,\, y \in \mathbb{C};\;\; \ell \in \mathbb{N} \setminus \{1\}),$$

where $f(t,\alpha)$ and $h(t,\nu)$ are functions of $t$ and $\alpha \in \mathbb{Z}$ and $t$ and $\nu \in \mathbb{N}_0$, respectively, which are analytic in a neighborhood of $t=0$. Obviously $\Phi_n^{(\alpha,\ell)}(x,0,\nu) = \Phi_n^{(\alpha)}(x,\nu)$. Then we establish various identities involving the polynomials $\Phi_n^{(\alpha,\ell)}(x,y,\nu)$. Also, as special cases of the generalized generating function in (1.23), we introduce two new polynomials: power sum-Laguerre-Hermite polynomials and generalized Laguerre-Euler polynomials and investigate some involved properties.

Some of the results presented here will include certain known identities and formulas involving relatively simple and familiar numbers and polynomials as particular cases, which are easy for the interested reader to check (see, e.g., [8, 12–17, 21, 23, 24, 29, 30]).

**Remark 1.1.** The substitution

$$f(t,\alpha) = \left(\frac{t}{a^t - b^t}\right)^{\alpha} C_0(xt), \quad h(t,\nu) = 0, \quad \text{and} \quad \ell = 2$$

in (1.23) yields (1.22). So it may imply that the polynomials in (1.23) are more general than those in (1.22). The process and methods used in this paper follow from those employed in such works as [5, 13, 15–17] including, in particular, the very recent work [18].

## 2. Some formulas involving the polynomials $\Phi_{n,\ell}^{(\alpha)}(x,y,\nu)$

Here, we present certain formulas associated with the polynomials $\Phi_{n,\ell}^{(\alpha)}(x,y,\nu)$. To do this, we recall some formal manipulations of double series in the following lemma (see, e.g., [4], [17], [26, pp. 56-57], and [28, p. 52]).

**Lemma 2.1.** *The following identities hold:*

$$\sum_{n=0}^{\infty}\sum_{k=0}^{\infty} A_{k,n} = \sum_{n=0}^{\infty}\sum_{k=0}^{[n/p]} A_{k,n-pk} \quad (p \in \mathbb{N}); \tag{2.1}$$

$$\sum_{n=0}^{\infty}\sum_{k=0}^{[n/p]} A_{k,n} = \sum_{n=0}^{\infty}\sum_{k=0}^{\infty} A_{k,n+pk} \quad (p \in \mathbb{N}); \tag{2.2}$$

$$\sum_{N=0}^{\infty} f(N)\frac{(x+y)^N}{N!} = \sum_{n,m=0}^{\infty} f(m+n)\frac{x^n}{n!}\frac{y^m}{m!}. \tag{2.3}$$

Here, the $A_{k,n}$ and $f(N)$ ($k$, $n$, $N \in \mathbb{N}_0$) are real or complex valued functions indexed by the $k$, $n$ and $N$, respectively, and $x$ and $y$ are real or complex numbers. Also, for possible rearrangements of the involved double series, all the associated series should be absolutely convergent.

**Theorem 2.2.** *Let $\alpha \in \mathbb{Z}$, $\nu \in \mathbb{N}_0$, and $\ell \in \mathbb{N} \setminus \{1\}$. Then*

$$\begin{aligned} \Phi_n^{(\alpha,\ell)}(x_1+x_2,y,\nu) &= \sum_{k=0}^{n}\binom{n}{k} x_1^k\, \Phi_{n-k}^{(\alpha,\ell)}(x_2,y,\nu) \\ &= \sum_{k=0}^{n}\binom{n}{k} x_2^k\, \Phi_{n-k}^{(\alpha,\ell)}(x_1,y,\nu) \quad (n \in \mathbb{N}_0,\ x_1,\ x_2,\ y \in \mathbb{C}); \end{aligned} \tag{2.4}$$

$$\begin{aligned} \Phi_n^{(\alpha,\ell)}(x,y_1+y_2,\nu) &= \sum_{k=0}^{\left[\frac{n}{\ell}\right]} \frac{n!\, y_1^k}{(n-\ell k)!\, k!}\, \Phi_{n-\ell k}^{(\alpha,\ell)}(x,y_2,\nu) \\ &= \sum_{k=0}^{\left[\frac{n}{\ell}\right]} \frac{n!\, y_2^k}{(n-\ell k)!\, k!}\, \Phi_{n-\ell k}^{(\alpha,\ell)}(x,y_1,\nu) \\ &(n \in \mathbb{N}_0,\ x,\ y_1,\ y_2 \in \mathbb{C}); \end{aligned} \tag{2.5}$$

$$\Phi_n^{(\alpha,\ell)}(x,y,\nu) = \sum_{k=0}^{n}\binom{n}{k} x^k\, \Phi_{n-k}^{(\alpha,\ell)}(0,y,\nu); \quad (n \in \mathbb{N}_0,\ x,\ y \in \mathbb{C}); \tag{2.6}$$

$$\Phi_n^{(\alpha,\ell)}(x,y,\nu) = \sum_{k=0}^{\left[\frac{n}{\ell}\right]} \frac{n!\, y^k}{(n-\ell k)!\, k!}\, \Phi_{n-\ell k}^{(\alpha,\ell)}(x,0,\nu) \quad (n \in \mathbb{N}_0,\ x,\ y \in \mathbb{C}); \tag{2.7}$$

$$\frac{\partial}{\partial x}\Phi_n^{(\alpha,\ell)}(x,y,\nu) = n\, \Phi_{n-1}^{(\alpha,\ell)}(x,y,\nu) \quad (n \in \mathbb{N},\ x,\ y \in \mathbb{C}); \tag{2.8}$$

$$\frac{\partial^r}{\partial x^r}\Phi_n^{(\alpha,\ell)}(x,y,\nu) = \frac{n!}{(n-r)!}\, \Phi_{n-r}^{(\alpha,\ell)}(x,y,\nu) \quad (n,\ r \in \mathbb{N} \text{ with } 1 \le r \le n;\ x,\ y \in \mathbb{C}); \tag{2.9}$$

$$\frac{\partial}{\partial y}\Phi_n^{(\alpha,\ell)}(x,y,\nu) = \frac{n!}{(n-\ell)!}\, \Phi_{n-\ell}^{(\alpha,\ell)}(x,y,\nu) \quad (n,\ \ell \in \mathbb{N} \text{ with } 1 \le \ell \le n;\ x,\ y \in \mathbb{C}); \tag{2.10}$$

$$\int_a^x \Phi_n^{(\alpha,\ell)}(u,y,\nu)\, du = \frac{\Phi_{n+1}^{(\alpha,\ell)}(x,y,\nu) - \Phi_{n+1}^{(\alpha,\ell)}(a,y,\nu)}{n+1} \quad (n \in \mathbb{N}_0,\ a,\ x \in \mathbb{R},\ y \in \mathbb{C}). \tag{2.11}$$

$$\int_a^y \Phi_n^{(\alpha,\ell)}(x,u,\nu)\,du = \frac{n!}{(n+\ell)!}\left\{\Phi_{n+\ell}^{(\alpha,\ell)}(x,y,\nu) - \Phi_{n+\ell}^{(\alpha,\ell)}(x,a,\nu)\right\} \tag{2.12}$$
$$(n \in \mathbb{N}_0,\, x \in \mathbb{C},\, a,\, y \in \mathbb{R}).$$

***Proof.*** From (1.23), we write

$$\sum_{n=0}^{\infty} \Phi_n^{(\alpha,\ell)}(x_1+x_2,y,\nu)\,\frac{t^n}{n!} = \mathsf{e}^{x_1 t}\cdot f(t,\alpha)\,\mathsf{e}^{x_2 t + y t^\ell + h(t,\nu)}.$$

Expanding $\mathsf{e}^{x_1 t}$ as the Maclaurin series and using (1.23) to expand the second factor, with the aid of (2.1) with $p = 1$, we find

$$\sum_{n=0}^{\infty} \Phi_n^{(\alpha,\ell)}(x_1+x_2,y,\nu)\,\frac{t^n}{n!} = \sum_{n=0}^{\infty}\sum_{k=0}^{n} \frac{x_1^k}{(n-k)!k!}\,\Phi_{n-k}^{(\alpha,\ell)}(x_2,y,\nu)\,t^n,$$

which, upon equating the coefficients of $t^n$, yields the first equality of (2.4). For the second equality of (2.4), we just change the role of $x_1$ and $x_2$ in the above proof.

Similarly as in the proof of (2.4), with the aid of (2.1) with $p = \ell$, we prove (2.5).

Setting $x_1 = x$ and $x_2 = 0$ in the first equality in (2.4), we obtain (2.6). Similarly, setting $y_1 = y$ and $y_2 = 0$ in the first equality in (2.5), we get (2.7).

Differentiating both sides of (2.6) with respect to the variable $x$, we have

$$\begin{aligned}\frac{\partial}{\partial x}\Phi_n^{(\alpha,\ell)}(x,y,\nu) &= \sum_{k=1}^{n} k\binom{n}{k} x^{k-1}\,\Phi_{n-k}^{(\alpha,\ell)}(0,y,\nu)\\ &= n\sum_{k=0}^{n-1}\binom{n-1}{k} x^k\,\Phi_{n-1-k}^{(\alpha,\ell)}(0,y,\nu)\\ &= n\,\Phi_{n-1}^{(\alpha,\ell)}(x,y,\nu),\end{aligned} \tag{2.13}$$

where the identity (2.6) is used for the last equality. This proves (2.8).

Then, differentiating both sides of (2.8) with respect to the variable $x$ by using the identity (2.8) on the right side of each resulting identity, consecutively, $r-1$ times, we obtain (2.9).

Differentiating both sides of (2.7) with respect to the variable $y$, we have

$$\frac{\partial}{\partial y}\Phi_n^{(\alpha,\ell)}(x,y,\nu) = \sum_{k=1}^{\left[\frac{n}{\ell}\right]} \frac{n!\,y^{k-1}}{(n-\ell k)!\,(k-1)!}\,\Phi_{n-\ell k}^{(\alpha,\ell)}(x,0,\nu). \tag{2.14}$$

Taking $k-1 = k'$ on the right side of (2.14) and considering

$$\left[\frac{n}{\ell}\right] - 1 = \left[\frac{n}{\ell} - 1\right] = \left[\frac{n-\ell}{\ell}\right],$$

we get

$$\frac{\partial}{\partial y}\Phi_n^{(\alpha,\ell)}(x,y,\nu) = \frac{n!}{(n-\ell)!}\sum_{k=0}^{\left[\frac{n-\ell}{\ell}\right]} \frac{(n-\ell)!\,y^k}{(n-\ell-\ell k)!\,k!}\,\Phi_{n-\ell-\ell k}^{(\alpha,\ell)}(x,0,\nu),$$

which, upon using (2.7), proves (2.10).

Replacing $x$ by $u$ in (2.8) and integrating both sides of the resulting identity with respect to the variable $u$ from $a$ to $x$ by using the fundamental theorem of calculus, and substituting $n+1$ for $n$ in the last resulting identity, we obtain (2.11).

Similarly as in getting (2.11), using (2.10), we get (2.12). □

## 3. Power sum-Laguerre-Hermite polynomials

Here, replacing $x$ by $y$ and $\nu$ by $z$ in (1.9) and setting $h(t,z) = z\,t^2$ and

$$f(x;t,\mathtt{n}) = \frac{\mathsf{e}^{(\mathtt{n}+1)t}-1}{\mathsf{e}^t-1}\,C_0(xt),$$

we introduce a new class of power sum-Laguerre-Hermite polynomials ${}_H^S L_n(x,y,z;\mathtt{n})$ by the following generating function:

$$\frac{\mathsf{e}^{(\mathtt{n}+1)t}-1}{\mathsf{e}^t-1}\,\mathsf{e}^{yt+z\,t^2}\,C_0(xt) = \sum_{n=0}^{\infty} {}_H^S L_n(x,y,z;\mathtt{n})\,\frac{t^n}{n!} \quad (|t|<2\pi). \tag{3.1}$$

Now, we present various implicit summation formulae for the power sum-Laguerre-Hermite polynomials.

**Theorem 3.1.** *The following implicit summation formulas for the power sum-Laguerre-Hermite polynomials hold.*

$${}_H^S L_n(x,y,0;\mathtt{n}) = \sum_{k=0}^{n} \binom{n}{k} L_{n-k}(x,y)\,S_k(\mathtt{n}) \quad (n\in\mathbb{N}_0;\ \mathtt{n}\in\mathbb{N})\,; \tag{3.2}$$

$${}_H^S L_n(x,y,z;\mathtt{n}) = n! \sum_{r=0}^{n}\sum_{k=0}^{n-r} \frac{(-1)^r\,x^r\,H_{n-k-r}(y,z)\,S_k(\mathtt{n})}{(r!)^2\,k!\,(n-k-r)!} \quad (n\in\mathbb{N}_0;\ \mathtt{n}\in\mathbb{N})\,; \tag{3.3}$$

$${}_H^S L_n(x,u+v,z;\mathtt{n}) = \sum_{k=0}^{n} \binom{n}{k} u^k\,{}_H^S L_{n-k}(x,v,z;\mathtt{n}) \quad (n\in\mathbb{N}_0;\ \mathtt{n}\in\mathbb{N})\,; \tag{3.4}$$

$${}_H^S L_n(x,y,a+b;\mathtt{n}) = \sum_{k=0}^{\left[\frac{n}{2}\right]} \frac{n!}{k!(n-2k)!}\,b^k\,{}_H^S L_{n-2k}(x,y,a;\mathtt{n}) \quad (n\in\mathbb{N}_0;\ \mathtt{n}\in\mathbb{N})\,. \tag{3.5}$$

***Proof.*** Setting $z=0$ in (3.1) and using (1.2) and (1.16) with the aid of (2.1) with $p=1$, we obtain

$$\sum_{n=0}^{\infty} {}_H^S L_n(x,y,z;\mathtt{n})\,\frac{t^n}{n!} = \sum_{n=0}^{\infty}\sum_{k=0}^{n} L_{n-k}(x,y)\,S_k(\mathtt{n})\,\frac{t^n}{(n-k)!k!},$$

which, upon equating the coefficients of $t^n$, yields the desired result (3.2).

The other identities can be proved as in the proof of (3.2). We omit the details. $\square$

## 4. Generalized Laguerre-Euler polynomials

Here, replacing $x$ by $y$ and $\nu$ by $z$ in (1.9) and $f(x;t,\alpha) = \left(\frac{2}{a^t+b^t}\right)^{\alpha} C_0(xt)$, we introduce a new class of the generalized Laguerre-Euler polynomials.

Let $\alpha \in \mathbb{R}$ or $\mathbb{C}$ be a parameter. Also, let $a,b\in\mathbb{R}^+$ with $a\neq b$. The generalized Euler polynomials $E_n^{(\alpha)}(x,y,z;a,b,\mathsf{e})$ are defined by the following generating function

$$\left(\frac{2}{a^t+b^t}\right)^{\alpha} \mathsf{e}^{yt+h(t,z)}C_0(xt) = \sum_{n=0}^{\infty} E_n^{(\alpha)}(x,y,z;a,b,\mathsf{e})\frac{t^n}{n!} \tag{4.1}$$

$$\left(x\in\mathbb{R};\ |t| < \frac{2\pi}{|\ln a - \ln b|}\right).$$

In particular, setting $h(t,z) = zt^2$ in (4.1), we get

Let $\alpha \in \mathbb{R}$ or $\mathbb{C}$ be a parameter. Also, let $a, b \in \mathbb{R}^+$ with $a \neq b$. The generalized Laguerre-Euler polynomials ${}_LE_n^{(\alpha)}(x, y, z; a, b, \mathsf{e})$ are defined by

$$\left(\frac{2}{a^t + b^t}\right)^\alpha \mathsf{e}^{yt+zt^2} C_0(xt) = \sum_{n=0}^{\infty} {}_LE_n^{(\alpha)}(x, y, z; a, b, \mathsf{e}) \frac{t^n}{n!} \tag{4.2}$$

$$\left(x \in \mathbb{R};\ |t| < \frac{2\pi}{|\ln a - \ln b|}\right).$$

We have

$${}_LE_n^{(\alpha)}(x, y, z; a, b, \mathsf{e}) = \sum_{m=0}^{n} \sum_{k=0}^{[\frac{m}{2}]} \frac{E_{n-m}^{(\alpha)}\ L_{m-2k}(x, y) z^k n!}{(m-2k)!k!(n-m)!}. \tag{4.3}$$

**Remark 4.1.** Consider some special cases of (4.2).

(i) The case $x = 0$ of (4.2) reduces to the known generalized Hermite-Bernoulli polynomials defined by (see [24])

$$\left(\frac{2}{a^t + b^t}\right)^\alpha \mathsf{e}^{yt+zt^2} = \sum_{n=0}^{\infty} {}_HE_n^{(\alpha)}(y, z; a, b, \mathsf{e}) \frac{t^n}{n!} \tag{4.4}$$

$$\left(|t| < \frac{2\pi}{|\ln a - \ln b|}\right).$$

(ii) The case $x = z = 0$ of (4.2) reduces to the known generalized Euler polynomials defined by (see [20])

$$\left(\frac{2}{a^t + b^t}\right)^\alpha \mathsf{e}^{yt} = \sum_{n=0}^{\infty} E_n^{(\alpha)}(y; a, b, \mathsf{e}) \frac{t^n}{n!} \tag{4.5}$$

$$\left(|t| < \frac{2\pi}{|\ln a - \ln b|}\right).$$

(iii) The case $x = y = z = 0$ of (4.2) reduces to the generalized Euler number $E_n^{(\alpha)}(a, b)$ defined by

$$\left(\frac{2}{a^t + b^t}\right)^\alpha = \sum_{n=0}^{\infty} E_n^{(\alpha)}(a, b) \frac{t^n}{n!} \tag{4.6}$$

$$\left(|t| < \frac{2\pi}{|\ln a - \ln b|}\right).$$

We find that $E_n^{(1)}(a, b) = E_n(a, b)$ in (1.17) and

$$E_n^{(\alpha+\beta)}(a, b) = \sum_{k=0}^{n} \binom{n}{k} E_k^{(\alpha)}(a, b)\, E_{n-k}^{(\beta)}(a, b) \quad (n \in \mathbb{N}_0). \tag{4.7}$$

Here, we present various implicit summation formulae for the generalized Laguerre-Euler polynomials.

**Theorem 4.2.** *Let $\alpha$, $\beta \in \mathbb{R}$ or $\mathbb{C}$ be parameters. Also, let $a, b \in \mathbb{R}^+$ with $a \neq b$. Further, let $u$, $v$, $w, x$, $y$, $z \in \mathbb{R}$, and $n \in \mathbb{N}_0$. Then the following implicit summation formulas for the generalized Laguerre-Euler polynomials in (4.2) hold:*

$$\begin{aligned} &{}_LE_{m+n}^{(\alpha)}(x, w, z; a, b, \mathsf{e}) \\ &= \sum_{s=0}^{m} \sum_{k=0}^{n} \binom{m}{s} \binom{n}{k} (w - y)^{s+k}\ {}_LE_{m+n-s-k}^{(\alpha)}(x, y, z; a, b, \mathsf{e}); \end{aligned} \tag{4.8}$$

$$ {}_LE_n^{(\alpha)}(x,y+\alpha,z;a,b,\mathsf{e}) = n! \sum_{j=0}^{[\frac{n}{2}]} \sum_{k=0}^{n-2j} \frac{(-1)^k x^k z^j E_{n-2j-k}^{(\alpha)}(y;\frac{a}{\mathsf{e}},\frac{b}{\mathsf{e}},\mathsf{e})}{(n-2j-k)!\,j!\,(k!)^2}; \tag{4.9} $$

$$ \begin{aligned} &{}_LE_n^{(\alpha+\beta)}(x,y+v,z;a,b,\mathsf{e}) \\ &\quad = \sum_{k=0}^{n} \binom{n}{k} {}_LE_{n-k}^{(\alpha)}(x,y,z;a,b,\mathsf{e}) E_k^{(\beta)}(v;a,b,\mathsf{e}); \end{aligned} \tag{4.10} $$

$$ \begin{aligned} &{}_LE_n^{(\alpha+\beta)}(x,y+z,v+u;a,b,\mathsf{e}) \\ &\quad = \sum_{k=0}^{n} \binom{n}{k} E_{n-k}^{(\alpha)}(x,z,v;a,b,\mathsf{e}) {}_HE_k^{(\beta)}(y,u;a,b,\mathsf{e}); \end{aligned} \tag{4.11} $$

$$ {}_LE_n^{(\alpha)}(x,y,z;a,b,\mathsf{e}) = n! \sum_{j=0}^{[\frac{n}{2}]} \sum_{k=0}^{n-2j} \frac{E_k^{(\alpha)}(a,b,\mathsf{e})\, L_{n-k-2j}(x,y)\, z^j}{k!\,j!\,(n-k-2j)!}. \tag{4.12} $$

***Proof.*** For (4.8), replacing $t$ by $t+u$ in (4.2) and using the binomial theorem, we have

$$ \begin{aligned} &\left(\frac{2}{a^{t+u}+b^{t+u}}\right)^{\alpha} \mathsf{e}^{y(t+u)+z(t+u)^2} C_0(x(t+u)) \\ &\quad = \sum_{n=0}^{\infty} {}_LE_n^{(\alpha)}(x,y,z;a,b,\mathsf{e}) \frac{(t+u)^n}{n!} \\ &\quad = \sum_{n=0}^{\infty} \sum_{m=0}^{n} {}_LE_n^{(\alpha)}(x,y,z;a,b,\mathsf{e}) \frac{t^{n-m}u^m}{(n-m)!\,m!}. \end{aligned} \tag{4.13} $$

Using (2.2) with $p=1$ in the last double summation in (4.13), we obtain

$$ \begin{aligned} &\left(\frac{2}{a^{t+u}+b^{t+u}}\right)^{\alpha} \mathsf{e}^{z(t+u)^2} C_0(x(t+u)) \\ &\quad = \mathsf{e}^{-y(t+u)} \sum_{n=0}^{\infty} \sum_{m=0}^{\infty} {}_LE_{n+m}^{(\alpha)}(x,y,z;a,b,\mathsf{e}) \frac{t^n u^m}{n!\,m!}. \end{aligned} \tag{4.14} $$

Since the left side of (4.14) is independent of the variable $y$, we introduce another variable $w$ for the variable $y$ in the right side of (4.14) and equate the two resulting identities to find

$$ \begin{aligned} &\sum_{n=0}^{\infty} \sum_{m=0}^{\infty} {}_LE_{n+m}^{(\alpha)}(x,w,z;a,b,\mathsf{e}) \frac{t^n u^m}{n!\,m!} \\ &\quad = \mathsf{e}^{(w-y)(t+u)} \sum_{n=0}^{\infty} \sum_{m=0}^{\infty} {}_LE_{n+m}^{(\alpha)}(x,y,z;a,b,\mathsf{e}) \frac{t^n u^m}{n!\,m!}. \end{aligned} \tag{4.15} $$

We use (2.3) to find

$$ \mathsf{e}^{(w-y)(t+u)} = \sum_{N=0}^{\infty} (w-y)^N \frac{(t+u)^N}{N!} = \sum_{k,s=0}^{\infty} (w-y)^{k+s} \frac{t^k\, u^s}{k!\,s!}. \tag{4.16} $$

Using (4.16) in the right side of (4.15) and applying (2.1) with $p=1$ in the resulting quadruple series, two times, we get

$$ \begin{aligned} &\sum_{n=0}^{\infty} \sum_{m=0}^{\infty} {}_LE_{n+m}^{(\alpha)}(x,w,z;a,b,\mathsf{e}) \frac{t^n u^m}{n!\,m!} \\ &= \sum_{n=0}^{\infty} \sum_{m=0}^{\infty} \sum_{k=0}^{n} \sum_{s=0}^{m} {}_LE_{n+m-s-k}^{(\alpha)}(x,y,z;a,b,\mathsf{e})\,(w-y)^{k+s} \frac{t^n\, u^m}{(n-k)!\,k!\,(m-s)!\,s!}. \end{aligned} \tag{4.17} $$

Finally, equating the coefficients of $t^n$ and $u^m$ in both sides of (4.17), consecutively, we obtain the identity (4.8).

For (4.9), we find from (4.2) that

$$\sum_{n=0}^{\infty} {}_L E_n^{(\alpha)}(x, y+\alpha, z; a, b, \mathsf{e}) \frac{t^n}{n!} = \left( \frac{2}{(\frac{a}{\mathsf{e}})^t + (\frac{b}{\mathsf{e}})^t} \right)^{\alpha} \mathsf{e}^{yt} \cdot \mathsf{e}^{zt^2} \cdot C_0(xt) \tag{4.18}$$

By using (4.5) and (2.1) with $p = 2$, we have

$$\begin{aligned}\left( \frac{2}{(\frac{a}{\mathsf{e}})^t + (\frac{b}{\mathsf{e}})^t} \right)^{\alpha} \mathsf{e}^{yt} \cdot \mathsf{e}^{zt^2} &= \sum_{n=0}^{\infty} E_n^{(\alpha)} \left( y; \frac{a}{\mathsf{e}}, \frac{b}{\mathsf{e}}, \mathsf{e} \right) \frac{t^n}{n!} \cdot \sum_{j=0}^{\infty} \frac{z^j\, t^{2j}}{j!} \\ &= \sum_{n=0}^{\infty} \sum_{j=0}^{[\frac{n}{2}]} E_{n-2j}^{(\alpha)} \left( y; \frac{a}{\mathsf{e}}, \frac{b}{\mathsf{e}}, \mathsf{e} \right) z^j \frac{t^n}{(n-2j)!\, j!}.\end{aligned} \tag{4.19}$$

Setting the result (4.19) in (4.18) and using (1.4) with $n = 0$, with the help of (2.1) with $p = 1$, we obtain

$$\begin{aligned}&\sum_{n=0}^{\infty} {}_L E_n^{(\alpha)}(x, y+\alpha, z; a, b, \mathsf{e}) \frac{t^n}{n!} \\ &= \sum_{n=0}^{\infty} \left\{ \sum_{j=0}^{[\frac{n}{2}]} \sum_{k=0}^{n-2j} E_{n-2j-k}^{(\alpha)} \left( y; \frac{a}{\mathsf{e}}, \frac{b}{\mathsf{e}}, \mathsf{e} \right) \frac{z^j x^k (-1)^k}{(n-2j-k)!\, j! (k!)^2} \right\} t^n.\end{aligned} \tag{4.20}$$

Finally, equating the coefficients of $t^n$ on both sides of (4.20), we get the identity (4.9).

Similarly as above, we can prove the other identities. We omit the details. □

## 5. Symmetry identities for the generalized Laguerre-Euler polynomials

A number of interesting symmetry identities for various polynomials have been presented (see, e.g., [12–18, 29, 30]). Here, we give symmetry identities for the generalized Laguerre-Euler polynomials ${}_L E_n^{(\alpha)}(x, y, z; a, b, \mathsf{e})$ in (4.2). To do this, we consider the following function:

$$\begin{aligned} g(t) := &\left\{ \frac{4}{(c^{at} + d^{at})(c^{bt} + d^{bt})} \right\}^{\alpha} \left\{ \frac{4}{(c^{at} + d^{at})(c^{bt} + d^{bt})} \right\}^{\beta} \\ &\times \mathsf{e}^{(a+b)(y_1+y_2)t + \left(a^2+b^2\right)(z_1+z_2)t^2} \\ &\times C_0(x_1 at)\, C_0(x_1 bt)\, C_0(x_2 at)\, C_0(x_2 bt). \end{aligned} \tag{5.1}$$

We see that the function $g(t)$ in (5.1) is symmetric with respect to $\alpha$ and $\beta$, $a$ and $b$, $c$ and $d$, $x_1$ and $x_2$, $y_1$ and $y_2$, $z_1$ and $z_2$, respectively. So, to make the generalized Laguerre-Euler polynomials in (4.2), we have 16 combinations. Then we will get 15 symmetry identities for the generalized Laguerre-Euler polynomials in (4.2), two of which will be asserted in the following theorem and the other 13 of which are left to the interested reader.

**Theorem 5.1.** *Let $\alpha$, $\beta \in \mathbb{R}$ or $\mathbb{C}$ be parameters. Also, let $c$, $d \in \mathbb{R}^+$ with $c \neq d$. Further, let $a$, $b$, $x_1$, $x_2$, $y_1$, $y_2$, $z_1$, $z_2 \in \mathbb{R}$ and $n \in \mathbb{N}_0$. Then*

$$\begin{aligned}&\sum_{r=0}^{n} \sum_{m=0}^{n-r} \sum_{s=0}^{r} {}_L E_{n-m-r}^{(\alpha)}(x_1, y_1, z_1; c, d, \mathsf{e})\, {}_L E_m^{(\alpha)}(x_1, y_1, z_1; c, d, \mathsf{e}) \\ &\times {}_L E_{r-s}^{(\beta)}(x_2, y_2, z_2; c, d, \mathsf{e})\, {}_L E_s^{(\beta)}(x_2, y_2, z_2; c, d, \mathsf{e}) \frac{a^{n-m-s}\, b^{m+s}}{(n-m-r)!\, m!\, (r-s)!\, s!}\end{aligned}$$

$$\begin{aligned}
&= \sum_{r=0}^{n} \sum_{m=0}^{n-r} \sum_{s=0}^{r} {}_LE^{(\alpha)}_{n-m-r}(x_2, y_2, z_2; c, d, \mathsf{e})\, {}_LE^{(\alpha)}_{m}(x_2, y_2, z_2; c, d, \mathsf{e}) \\
&\times {}_LE^{(\beta)}_{r-s}(x_1, y_1, z_1; c, d, \mathsf{e})\, {}_LE^{(\beta)}_{s}(x_1, y_1, z_1; c, d, \mathsf{e}) \, \frac{a^{n-m-s}\, b^{m+s}}{(n-m-r)!\, m!\, (r-s)!\, s!}
\end{aligned} \tag{5.2}$$

$$\begin{aligned}
&= \sum_{r=0}^{n} \sum_{m=0}^{n-r} \sum_{s=0}^{r} E^{(\beta)}_{n-m-r}(x_2, y_1, z_1; c, d, \mathsf{e})\, {}_LE^{(\beta)}_{m}(x_2, y_1, z_1; c, d, \mathsf{e}) \\
&\times {}_LE^{(\alpha)}_{r-s}(x_1, y_2, z_2; c, d, \mathsf{e})\, {}_LE^{(\alpha)}_{s}(x_1, y_2, z_2; c, d, \mathsf{e}) \, \frac{b^{n-m-s}\, a^{m+s}}{(n-m-r)!\, m!\, (r-s)!\, s!}.
\end{aligned} \tag{5.3}$$

***Proof.*** We try to combine $g(t)$ as follows:

$$\begin{aligned}
g(t) &= \left\{ \frac{2}{c^{at} + d^{at}} \right\}^{\alpha} \mathsf{e}^{a y_1 t + a^2 z_1 t}\; C_0(x_1 a t) \\
&\times \left\{ \frac{2}{c^{bt} + d^{bt}} \right\}^{\alpha} \mathsf{e}^{b y_1 t + b^2 z_1 t}\; C_0(x_1 b t) \\
&\times \left\{ \frac{2}{c^{at} + d^{at}} \right\}^{\beta} \mathsf{e}^{a y_2 t + a^2 z_2 t}\; C_0(x_2 a t) \\
&\times \left\{ \frac{2}{c^{bt} + d^{bt}} \right\}^{\beta} \mathsf{e}^{b y_2 t + b^2 z_2 t}\; C_0(x_2 b t),
\end{aligned} \tag{5.4}$$

which, upon using (4.2), gives

$$\begin{aligned}
g(t) &= \sum_{n=0}^{\infty} {}_LE^{(\alpha)}_{n}(x_1, y_1, z_1; c, d, \mathsf{e}) \frac{(at)^n}{n!} \\
&\times \sum_{m=0}^{\infty} {}_LE^{(\alpha)}_{m}(x_1, y_1, z_1; c, d, \mathsf{e}) \frac{(bt)^m}{m!} \\
&\times \sum_{r=0}^{\infty} {}_LE^{(\beta)}_{r}(x_2, y_2, z_2; c, d, \mathsf{e}) \frac{(at)^r}{r!} \\
&\times \sum_{s=0}^{\infty} {}_LE^{(\beta)}_{s}(x_2, y_2, z_2; c, d, \mathsf{e}) \frac{(bt)^s}{s!}
\end{aligned} \tag{5.5}$$

Now, we apply (2.1) with $p = 1$ to combine the first and second series into a single series and the third and fourth series into another single series. Then we use (2.1) with $p = 1$ to combine the two resulting single series into one series to find

$$\begin{aligned}
g(t) = \sum_{n=0}^{\infty} \Bigg\{ &\sum_{r=0}^{n} \sum_{m=0}^{n-r} \sum_{s=0}^{r} {}_LE^{(\alpha)}_{n-m-r}(x_1, y_1, z_1; c, d, \mathsf{e})\, {}_LE^{(\alpha)}_{m}(x_1, y_1, z_1; c, d, \mathsf{e}) \\
\times\, &{}_LE^{(\beta)}_{r-s}(x_2, y_2, z_2; c, d, \mathsf{e})\, {}_LE^{(\beta)}_{s}(x_2, y_2, z_2; c, d, \mathsf{e}) \, \frac{a^{n-m-s}\, b^{m+s}}{(n-m-r)!\, m!\, (r-s)!\, s!} \Bigg\} t^n.
\end{aligned} \tag{5.6}$$

Considering another combination of $g(t)$ as in (5.4), similarly as above, we can get another single series of $g(t)$ as in (5.6). Then, equating the coefficients of $t^n$ in both sides of the two single series, we can find 15 identities, two of which are recorded. $\square$

## 6. Concluding remarks

The results presented here, being very general, can be specialised to yield a number of known and new identities involving relatively simple and familiar polynomials. For example, setting $x = 0$ in (4.8), we have

$$\begin{aligned} &{}_H E^{(\alpha)}_{m+n}(w, z; a, b, \mathtt{e}) \\ &= \sum_{s=0}^{m} \sum_{k=0}^{n} \binom{m}{s} \binom{n}{k} (w - y)^{s+k} \, {}_H E^{(\alpha)}_{m+n-s-k}(y, z; a, b, \mathtt{e}). \end{aligned}$$

The power sum-Laguerre-Hermite polynomials ${}^{S}_{H}L_n(x, y, z; \mathtt{n})$ in (3) and the generalized Laguerre-Euler polynomials $E^{(\alpha)}_n(x, y, z; a, b, \mathtt{e})$ in (4.2) can be further extended and have their differential and integral formulas as in Theorem 2.2.

**Acknowledgment.** The authors would like to express their deep thanks for the reviewer whose useful comments improve this paper as it stands.

## References

[1] L.C. Andrews, *Special Functions for Engineer and Mathematician*, Macmillan Company, New York, 1985.
[2] E.T. Bell, *Exponential polynomials*, Ann. Math. **35** (2), 258–277, 1934.
[3] G. Betti and P.E. Ricci, *Multidimensional extensions of the Bernoulli and Appell polynomials*, Taiwanese J. Math. **8** (3), 415–428, 2004.
[4] J. Choi, *Notes on formal manipulations of double series*, Commun. Korean Math. Soc. **18** (4), 781–789, 2003.
[5] J. Choi, N.U. Khan and T. Usman, *A note on Legendre-based multi poly-Euler polynomials*, Bull. Iran. Math. Soc. **44**, 707–717, 2018.
[6] G. Dattoli, S. Lorenzutta and C. Cesarano, *Finite sums and generalized forms of Bernoulli polynomials*, Rend. Mat. **19**, 385–391, 1999.
[7] G. Dattoli and A. Torre, *Theory and Applications of Generalized Bessel Function*, Aracne, Rome, 1996.
[8] G. Dattoli and A. Torre, *Operational methods and two variable Laguerre polynomials*, Atti Acad. Torino **132**, 1–7, 1998.
[9] G. Dattoli, A. Torre and A.M. Mancho, *The generalized Laguerre polynomials, the associated Bessel functions and applications to propagation problems*, Radiat. Phys. Chem. **59**, 229–237, 2000.
[10] R. Dere and Y. Simsek, *Hermite base Bernoulli type polynomials on the umbral algebra*, Russian J. Math. Phys. **22** (1), 1–5, 2015.
[11] B.N. Guo and F. Qi, *Generalization of Bernoulli polynomials*, J. Math. Ed. Sci. Tech. **33** (3), 428–431, 2002.
[12] N.U. Khan and T. Usman, *A new class of Laguerre-based generalized Apostol polynomials*, Fasciculli. Math. **57**, 67–89, 2016.
[13] N.U. Khan and T. Usman, *A new class of Laguerre-based poly-Euler and multi poly-Euler polynomials*, J. Anal. Num. Theor. **4** (2), 113–120, 2016.
[14] N.U. Khan and T. Usman, *A new class of Laguerre poly-Bernoulli numbers and polynomials*, Adv. Stud. Contemporary Math. **27** (2), 229–241, 2017.
[15] N.U. Khan, T. Usman and A. Aman, *Generating functions for Legendre-Based poly-Bernoulli numbers and polynomials*, Honam Math. J. **39** (2), 217–231, 2017.
[16] N.U. Khan, T. Usman and J. Choi, *Certain generating function of Hermite-Bernoulli-Laguerre polynomials*, Far East J. Math. Sci. **101** (4), 893–908, 2017.

[17] N.U. Khan, T. Usman and J. Choi, *A new generalization of Apostol type Laguerre-Genocchi polynomials*, C. R. Acad. Sci. Paris, Ser. I, **355**, 607–617, 2017.

[18] N.U. Khan, T. Usman and J. Choi, *A new class of generalized polynomials associated with Laguerre and Bernoulli polynomials*, Turkish J. Math. **43**, 486–497, 2019.

[19] B. Kurt and Y. Simsek, *Notes on generalization of the Bernoulli type polynomials*, Appl. Math. Comput. **218**, 906–911, 2011.

[20] Q.-M. Luo, B.N. Guo, F. Qi and L. Debnath, *Generalization of Bernoulli numbers and polynomials*, Int. J. Math. Math. Sci. **59**, 3769–3776, 2003.

[21] Q.-M. Luo, F. Qi and L. Debnath, *Generalization of Euler numbers and polynomials*, Int. J. Math. Math. Sci. **61**, 3893–3901, 2003.

[22] L.M. Milne-Thomsons, *Two classes of generalized polynomials*, Proc. London Math. Soc. **35** (1), 514–522, 1933.

[23] M.A. Pathan, *A new class of generalized Hermite-Bernoulli polynomials*, Georgian Math. J. **19**, 559–573, 2012.

[24] M.A. Pathan and W.A. Khan, *A new class of generalized polynomials associated with Hermite and Euler polynomials*, Mediterr. J. Math. **13** (3), 913–928, 2016.

[25] F. Qi and B.N. Guo, *Generalization of Bernoulli polynomials*, RGMIA Res. Rep. Coll. **4** (4), Article 10, 691–695, 2001.

[26] E.D. Rainville, *Special Functions*, Macmillan Company, New York, 1960; Reprinted by Chelsea Publishing Company, Bronx, New York, 1971.

[27] H.M. Srivastava and J. Choi, *Zeta and q-Zeta Functions and Associated Series and Integrals*, Elsevier Science Publishers, Amsterdam, London and New York, 2012.

[28] H.M. Srivastava and H.L. Manocha, *A Treatise on Generating Functions*, Halsted Press (Ellis Horwood Limited, Chichester), John Wiley and Sons, New York, Chichester, Brisbane and Toronto, 1984.

[29] S. Yang, *An identity of symmetry for the Bernoulli polynomials*, Discrete Math. **308**, 550–554, 2008.

[30] Z. Zhang and H. Yang, *Several identities for the generalized Apostol Bernoulli polynomials*, Comput. Math. Appl. **56** (12), 2993–2999, 2008.

Hacettepe Journal of
Mathematics & Statistics

Hacet. J. Math. Stat.
Volume 50 (1) (2021), 14 – 23
DOI : 10.15672/hujms.558995

Research Article

# On the trace of powers of square matrices

Kailash M. Patil

*Department of Mathematics, Dharmsinh Desai University, Nadiad, Gujarat, INDIA 387001*

**Abstract**

Using Cayley-Hamilton equation for matrices, we obtain a simple formula for trace of powers of a square matrix. The formula becomes simpler in particular cases. As a consequence, we also demonstrate the formula for trace of negative powers of a matrix.

**Mathematics Subject Classification (2020).** 15A24, 15A45

**Keywords.** trace of a matrix, powers of a matrix, Cayley-Hamilton theorem, spectral radius

## 1. Introduction

With the advancement of highly complex computer network topologies and eternally growing number of nodes in the existing networks, certain applications require to find the number of cliques in the graph of a given network. Using the adjacency matrix $A$ of the graph, one clique of vertices $v_1, v_2, v_3$ contributes the 2 to each of the $a_{11}, a_{22}, a_{33}$. Thus the count of cliques will be $\frac{Tr(A^3)}{6}$ [2]. In [6], an identity involving the Eulerian congruence on trace of powers of integer matrices modulo $p^r$ is obtained, where $p$ is prime, and $r \in \mathbb{N}$. [4] makes a short survey of related results. For a square matrix $A = [a_{ij}]$, the *trace* of $A$ denoted by $Tr(A)$, is the sum of main diagonal entries of $A$, that is $Tr(A) = \sum_i a_{ii}$. [5] obtains the formula of computation of the eigenvalue with maximum modulus of a matrix using the trace of its higher powers. Our formula thus contributes to finding the spectral radius of a matrix. [1] also developes the similar formula for $n^{\text{th}}$ power of a $2 \times 2$ matrix. Our formula is a general one and does not require computation of entries of $n^{\text{th}}$ power.

The current paper is in the sequel of [3], wherein we have obtained the formula for the sum of the powers of matrices and its consequences. In Section 2, we set the required notations and recall the terminology. We also state the main result Theorem 2.1. The simplification of the long computations in the proofs are achieved by introducing the functions $l_m(n, k_0, k_1, \ldots, k_{m-2})$ used for finding trace of $n^{\text{th}}$ power of an $m \times m$ matrix $A$. The introduction of $l_m(\cdot)$ is motivated by the list of expression of $Tr(A^n)$ for a $3 \times 3$ matrix $A$ for first few powers of $A$. The jargon of notations, as one will be convinced, is used only for the proof to be simplified. However, the actual application of our formulae to real computation does not require much of knowledge except the definition of the Trace and a couple of related definitions. The proof of the main theorem is discussed in Section 3. In fact, a technical formula (3.1) for $l_m(\cdot)$ is obtained in a series of Lemmas using Mathematical Induction. Very important and useful particular cases are discussed

Email address: kailashpatil2111@gmail.com
Received: 29.04.2019; Accepted: 12.04.2020

in Section 4. Finally the formula for the trace of negative powers of nonsingular matrices is demonstrated in Section 5. To maintain the brevity, we restrict ourselves to $2\times 2$ matrices for negative powers. However, we should impress upon the reader that this restrictions can easily be done away with.

## 2. Main result

In what follows, $A = [a_{ij}]$ denotes an $m \times m$ matrix. For any integer $1 \le k \le m$ and the integers $1 \le i_1 \le i_2 \le i_3 \le \cdots \le i_k \le m$, the determinant of the $k \times k$ submatrix obtained by removing all rows except $i_1, i_2, i_3, \ldots, i_k$ rows and $i_1, i_2, i_3, \ldots, i_k$ columns is called a *principal minor* of $A$ of order $k$, thereby obtaining $\binom{m}{k}$ minors. We denote their sum as $S_k(A)$ or for $S_k$ for brevity whenever there is no confusion. Thus, $S_1$ will become the trace of the given matrix and $S_n$ will be the determinant of $A$.

The *characteristic equation* of $A$ is given by

$$\det(A - \lambda I) = 0,$$

where $I$ is $m \times m$ identity matrix. The roots of the characteristic equation are called the *characteristic roots* of $A$. We shall denote them by $\lambda_1, \lambda_2, \ldots, \lambda_m$.

The motivation for defining ingredients required for the formula of trace of powers of $A$ lies in the analysis of a $3 \times 3$ matrix, and hence, for time being, $A$ will denote a $3 \times 3$ matrix.

The characteristic equation of $A$ is

$$\lambda^3 - S_1\lambda^2 + S_2\lambda - S_3 = 0,$$

where $S_1 = Tr(A) = \lambda_1 + \lambda_2 + \lambda_3 = \sum_{i=1}^{3} a_{ii}$, $S_2 = \sum_{i\neq j} \lambda_i\lambda_j$ and $S_3 = \lambda_1\lambda_2\lambda_3 = det(A)$.

By the Cayley-Hamilton theorem, we have $A^3 - S_1A^2 + S_2A - S_3I = 0$. This, in turn, implies the following for $n \in \mathbb{N}$.

$$A^{n+3} - S_1A^{n+2} + S_2A^{n+1} - S_3A^n = 0. \tag{2.1}$$

Applying the trace, a linear operator, on (2.1) gives a recursive relation,

$$Tr(A^{n+3}) = S_1Tr(A^{n+2}) - S_2Tr(A^{n+1}) + S_3Tr(A^n), \tag{2.2}$$

which is central to this note. Observe that

$$\begin{aligned} Tr(A^2) &= \lambda_1^2 + \lambda_2^2 + \lambda_3^2 = (\lambda_1 + \lambda_2 + \lambda_3)^2 - 2\,(\lambda_1\lambda_2 + \lambda_2\lambda_3 + \lambda_1\lambda_3) \\ &= S_1^2 - 2S_2. \end{aligned}$$

Putting particular values of $n \in Z_+ \cup \{0\}$ in (2.2) and simplifying, we have the following.

$$Tr(A^3) = S_1^3 - 3S_1S_2 + 3S_3.$$

$$Tr(A^4) = S_1^4 - 4S_1^2S_2 + 2S_2^2 + 4S_1S_3.$$

$$Tr(A^5) = S_1^5 - 5S_1^3S_2 + 5S_1S_2^2 + (5S_1^2 - 5S_2)S_3.$$

$$Tr(A^6) = S_1^6 - 6S_1^4S_2 + 9S_1^2S_2^2 - 2S_2^3 + (6S_1^3 - 12S_1S_2)S_3 + 3(S_3)^2.$$

$$Tr(A^7) = S_1^7 - 7S_1^5S_2 + 14S_1^3S_2^2 - 7S_1S_2^3 + (7S_1^4 - 21S_1^2S_2 + 7S_2^2)S_3 + (7S_1)S_3^2.$$

It is quite apparent that the complexity of the formula increases as the power increases. Well within the ninth power, the formula really becomes highly involved.

$$\begin{aligned} Tr(A^9) = {} & S_1^9 - 9S_1^7S_2 + 27S_1^5S_2^2 - 30S_1^3S_2^3 + 9S_1S_2^4 + (9S_1^6 - 45S_1^4S_2 + 54S_1^2S_2^2 - 9S_2^3)S_3 \\ & + (18S_1^3 - 27S_1S_2)S_3^2 + 3S_3^3 \end{aligned}$$

$$= \sum_{k_1=0}^{\lfloor \frac{9}{3} \rfloor} \sum_{k_0=0}^{\lfloor \frac{9-3k_1}{2} \rfloor} \frac{(-1)^{k_0}}{k_0!k_1!} \begin{bmatrix} 9(9-k_0-2k_1-1)(9-k_0-2k_1-2)\cdots \\ (9-2k_0-3k_1+1) \end{bmatrix} \times \left[ S_1^{9-3k_1-2k_0} S_2^{k_0} S_3^{k_1} \right].$$

Before we conclude the general formula for $Tr(A^n)$, we define

$$l_3(n,k_0,k_1) = \begin{cases} \frac{1}{k_0!k_1!} n(n-k_0-2k_1-1)(n-k_0-2k_1-2) & \\ \times(n-k_0-2k_1-3)\cdots(n-2k_0-3k_1+1), & \text{if } k_0+k_1 \geq 2; \\ & \\ n, & \text{if } k_0+k_1 = 1; \\ 1, & \text{if } k_0+k_1 = 0. \end{cases}$$

The above definition is applied only when each $k_i \geq 0$. In the course of different order of matrices we get different $l_m(n, k_0, k_1, \cdots, k_{m-2})$. Throughout this note, we adopt the convention that if at least one $k_i < 0$, then we define $l_m(n, k_0, k_1, \cdots, k_{m-2}) = 0$. As a consequence, In general, for $m \times m$ matrix

$$l_m(n,k_0,k_1,\ldots,k_{m-2}) \quad = \quad \frac{n}{k_0!k_1!\cdots!k_{m-2}!} \begin{bmatrix} (n-k_0-2k_1-\cdots-(m-1)k_{m-2}-1) \\ \times(n-k_0-2k_1-\cdots-(m-1)k_{m-2}-2) \\ \times\cdots \\ \times(n-2k_0-3k_1-\cdots-mk_{m-2}+1) \end{bmatrix}.$$

To shorten the displayed identities, when $n, k_0, k_1, \ldots, k_{m-2}$ are already mentioned in the summation, we write $l_m$ for $l_m(n, k_0, k_1, \ldots, k_{m-2})$. Our main result in terms of a function $l_m$ is Theorem 2.1.

**Theorem 2.1.** *For a $m \times m$ matrix $A = [a_{ij}]$, we have*

$$Tr(A^n) = \sum_{k_j \geq 0} \sum_{k_0=0}^{\left\lfloor \frac{n-3k_1-4k_2-\cdots-mk_{m-2}}{2} \right\rfloor} (-1)^{k_0+k_2+k_4+\cdots k_{\lfloor \frac{m-2}{2} \rfloor}} l_m \times [S_1^{n-2k_0-3k_1-4k_2-\cdots-mk_{m-2}} S_2^{k_0} S_3^{k_1} S_4^{k_2} \cdots S_{m-1}^{k_{m-3}} S_m^{k_{m-2}}]. \tag{2.3}$$

For a nonsingular $m \times m$ matrix $A$, one observes that

$$S_1(A^{-1}) = Tr(A^{-1}) = \frac{1}{\lambda_1} + \frac{1}{\lambda_2} + \cdots + \frac{1}{\lambda_m} = \frac{S_{m-1}(A)}{S_m(A)}.$$

$$S_2(A^{-1}) = \sum_{\substack{i,j=1 \\ i<j}}^{m} \frac{1}{\lambda_i \lambda_j} = \frac{S_{m-2}(A)}{S_m(A)}.$$

$$\cdots = \cdots$$

$$S_{m-1}(A^{-1}) = \sum_{i=1}^{m} \frac{1}{\lambda_1\lambda_2\cdots\lambda_{i-1}\lambda_{i+1}\cdots\lambda_m} = \frac{S_1(A)}{S_m(A)}.$$

$$S_m(A^{-1}) = \frac{1}{\lambda_1\lambda_2\cdots\lambda_m} = \frac{1}{S_m(A)}.$$

Using all this, and replacing $A$ by $A^{-1}$ in the Theorem 2.1, the following is at once.

**Theorem 2.2.** *For a $m \times m$ nonsingular matrix $A = [a_{ij}]$, we have*

$$Tr(A^{-n}) = \frac{1}{[\det(A)]^n} \sum_{k_j \geq 0} \sum_{k_0=0}^{\left\lfloor \frac{n-3k_1-4k_2-\cdots-mk_{m-2}}{2} \right\rfloor} (-1)^{k_0+k_2+k_4+\cdots k_{\lfloor \frac{m-2}{2} \rfloor}} l_m \times [S_1^{k_{m-3}} S_2^{k_{m-4}} \cdots S_{m-1}^{n-2k_0-3k_1-\cdots-mk_{m-2}} S_m^{k_0+2k_1+3k_2+\cdots+(m-1)k_{m-2}}]. \tag{2.4}$$

## 3. Proof of the main theorem

In order to prove the main theorem, we first prove the following.

$$l_m(n,k_0,k_1,\ldots,k_{m-2}) = l_m(n-1,k_0,k_1,\ldots,k_{m-2}) + \sum_{i=2}^{m} l_m(n-i,k_0,k_1,\ldots,k_{i-2}-1,\ldots,k_{m-2}). \tag{3.1}$$

We establish (3.1) by applying mathematical induction on the order of the matrix $A=[a_{ij}]$. The proof is divided into a couple of Lemmas.

**Lemma 3.1.** $l_2(n,k_0) = l_2(n-1,k_0) + l_2(n-2,k_0-1)$.

***Proof.*** Since the cases $k_0 = 0$ and $k_0 = 1$ are trivial, we can assume that $k_0 \geq 2$. Now

$$\begin{aligned} l_2(n-1,k_0) + l_2(n-2,k_0-1) =& \frac{(n-1)(n-k_0-2)(n-k_0-3)\cdots(n-2k_0)}{k_0!} \\ &+ \frac{(n-2)(n-k_0-2)(n-k_0-3)\cdots(n-2k_0+1)}{(k_0-1)!} \\ =& \frac{(n-k_0-2)(n-k_0-3)\cdots(n-2k_0+1)}{(k_0-1)!} \\ &\times \left[\frac{(n-1)(n-2k_0)}{k_0} + n - 2\right] \\ =& \frac{(n-k_0-2)(n-k_0-3)\cdots(n-2k_0+1)}{(k_0-1)!} \\ &\times \left[\frac{n^2 - 2nk_0 - n + 2k_0 + nk_0 - 2k_0}{k_0}\right] \\ =& \frac{(n-k_0-2)(n-k_0-3)\cdots(n-2k_0+1)}{(k_0-1)!} \\ &\times \left[\frac{n(n-k_0-1)}{k_0}\right] \\ =& l_2(n,k_0). \end{aligned}$$

□

**Lemma 3.2.** $l_3(n,k_0,k_1) = l_3(n-1,k_0,k_1) + l_3(n-2,k_0-1,k_1) + l_3(n-3,k_0,k_1-1)$.

***Proof.*** If $k_1 = 0$, then $l_3(n,k_0,k_1) = l_2(n,k_0)$ and $l_3(n-3,k_0,k_1-1) = 0$. Consequently, our case reduces to the Lemma 3.1. For $k_0 = 0$ and $k_1 \geq 1$, we have,

$$\begin{aligned} R.H.S. =& l_3(n-1,0,k_1) + l_3(n-3,0,k_1-1) \\ =& \frac{(n-1)(n-2k_1-2)(n-2k_1-3)\cdots(n-3k_1)}{k_1!} \\ &+ \frac{(n-3)(n-2k_1-2)(n-2k_1-3)\cdots(n-3k_1+1)}{(k_1-1)!} \\ =& \frac{(n-2k_1-2)(n-2k_1-3)\cdots(n-3k_1+1)}{(k_1-1)!}\left[\frac{(n-1)(n-3k_1)}{k_1} + n - 3\right] \\ =& \frac{(n-2k_1-2)(n-2k_1-3)\cdots(n-3k_1+1)}{(k_1-1)}\left[\frac{n(n-2k_1-1)}{k_1}\right] \\ =& l_3(n,0,k_1) \\ =& L.H.S. \end{aligned}$$

Since the case $k_0 = 0 = k_1$ is trivial, we assume now $k_0, k \geq 1$.

$$\begin{aligned}
R.H.S. =& l_3(n-1,k_0,k_1) + l_3(n-2,k_0-1,k_1) + l_3(n-3,k_0,k_1-1)\\
=& \frac{(n-1)(n-k_0-2k_1-2)(n-k_0-2k_1-3)\cdots(n-2k_0-3k_1)}{k_0!k_1!}\\
&+ \frac{(n-2)(n-k_0-2k_1-2)(n-k_0-2k_1-3)\cdots(n-2k_0-3k_1+1)}{(k_0-1)!k_1!}\\
&+ \frac{(n-3)(n-k_0-2k_1-2)(n-k_0-2k_1-3)\cdots(n-2k_0-3k_1+1)}{k_0!(k_1-1)!}\\
=& \frac{(n-k_0-2k_1-2)(n-k_0-2k_1-3)\cdots(n-2k_0-3k_1+1)}{(k_0-1)!(k_1-1)!}\\
&\times \left[\frac{(n-1)(n-2k_0-3k_1)}{k_0k_1} + \frac{n-2}{k_1} + \frac{n-3}{k_0}\right]\\
=& \frac{(n-k_0-2k_1-2)(n-k_0-2k_1-3)\cdots(n-2k_0-3k_1+1)}{(k_0-1)!(k_1-1)!}\\
&\times \left[\frac{n(n-k_0-2k_1-1)}{k_0k_1}\right]\\
=& l_3(n,k_0,k_1)\\
=& L.H.S.
\end{aligned}$$

□

**Lemma 3.3.** *As an induction hypothesis, assume that*

$$\begin{aligned}
l_t(n,k_0,k_1,k_2,\ldots,k_{t-2}) =& l_{t-1}(n-1,k_0,k_1,k_2,\ldots,k_{t-2})\\
&+ \sum_{i=2}^{t} l_{t-1}(n-i,k_0,k_1,k_2,\ldots,k_{i-2}-1,\ldots,k_{t-2}) \qquad (3.2)
\end{aligned}$$

*for $t \leq m-1$. Then*

$$\begin{aligned}
l_m(n,k_0,\ldots,k_{m-2}) =& l_m(n-1,k_0,\ldots,k_{m-2})\\
&+ \sum_{i=2}^{m} l_m(n-i,k_0,k_1,\ldots,k_{i-2}-1,\ldots,k_{m-2}). \qquad (3.3)
\end{aligned}$$

***Proof.*** If $k_{m-2} = 0$, then $l_m(n,k_0,\ldots,k_{m-2}) = l_{m-1}(n,k_0,\ldots,k_{m-3})$ and $l_m(n,k_0,k_1,\ldots,k_{m-2}-1) = 0$. Therefore, (3.3) follows from the Induction Hypothesis (3.2). Let $k_j = 0$ for some $0 \leq j \leq m-1$. Then

$$\begin{aligned}
L.H.S. =& l_m(n-1,k_0,\ldots,k_{j-1},0,k_{j+1},\ldots,k_{m-2})\\
&+ \sum_{i=2, i\neq j+2}^{m} l_m(n-i,k_0,\ldots,k_{i-2}-1,\ldots,k_{m-2})\\
=& \frac{1}{k_0!\cdots k_{j-1}!k_{j+1}!\cdots k_{m-2}!}\\
&\times \begin{bmatrix} (n-1)(n-k_0-2k_1-\cdots-jk_{j-1}-(j+1)k_{j+1}-(m-1)k_{m-2}-2)\\ (n-k_0-2k_1-\cdots-jk_{j-1}-(j+2)k_{j+1}-\cdots-(m-1)k_{m-2}-3)\\ \cdots\\ (n-2k_0-3k_1-\cdots-(j+1)k_{j-1}-(j+3)k_{j+1}-\cdots-mk_{m-2}) \end{bmatrix}\\
&+ \sum_{i=2, i\neq j+2}^{m} \frac{1}{k_0!k_1!\cdots(k_{i-2}-1)!k_{i-1}!k_i!\cdots k_{m-2}!}
\end{aligned}$$

$$
\begin{aligned}
&\times\begin{bmatrix} (n-i)(n-k_0-2k_1-\cdots-(i-1)k_{i-2}-\cdots-(m-1)k_{m-2}-2) \\ (n-k_0-2k_1-\cdots-(i-1)k_{i-2}-\cdots-(m-1)k_{m-2}-3) \\ \cdots \\ (n-2k_0-3k_1-\cdots-(j+1)k_{j-1}-(j+3)k_{j+1}-\cdots-mk_{m-2}+1) \end{bmatrix} \\
=&\frac{1}{(k_0-1)!\cdots(k_{j-1}-1)!(k_{j+1}-1)!\cdots(k_{m-2}-1)!} \\
&\times\begin{bmatrix} (n-k_0-2k_1-\cdots jk_{j-1}-(j+2)k_{j+1}-\cdots-(m-1)k_{m-2}-2) \\ (n-k_0-2k_1-\cdots-jk_{j-1}-(j+2)k_{j+1}-\cdots-(m-1)k_{m-2}-3) \\ \cdots \\ (n-2k_0-3k_1-\cdots-(j+1)k_{j-1}-(j+3)k_{j+1}-\cdots-mk_{m-2}+1) \end{bmatrix} \\
&\times\begin{bmatrix} \dfrac{(n-1)(n-2k_0-3k_1-\cdots-(j+1)k_{j-1}-(j+3)k_{j+1}\cdots-mk_{m-2})}{k_0k_1\cdots k_{j-1}k_{j+1}\cdots k_{m-2}} \\ +\dfrac{n-2}{k_1k_2\cdots k_{j-1}k_{j+1}\cdots k_{m-2}}+\dfrac{n-3}{k_0k_2\cdots k_{j-1}k_{j+1}\cdots k_{m-2}}+\cdots \\ +\dfrac{n-m}{k_0k_1\cdots k_{j-1}k_{j+1}\cdots k_{m-3}} \end{bmatrix} \\
=&\frac{1}{k_0!k_1!\cdots k_{j-1}!k_{j+1}!\cdots k_{m-2}!} \\
&\times\begin{bmatrix} n(n-k_0-2k_1-\cdots-jk_{j-1}-(j+2)k_{j+1}-\cdots-(m-1)k_{m-2}-1) \\ (n-k_0-2k_1-\cdots-jk_{j-1}-(j+2)k_{j+1}-\cdots-(m-1)k_{m-2}-2) \\ \cdots \\ (n-2k_0-3k_1-\cdots-(j+1)k_{j-1}-(j+3)k_{j+1}-\cdots-mk_{m-2}+1) \end{bmatrix} \\
=&l_m(n,k_0,k_1,\cdots,k_{j-1},jk_{j+1},\cdots,k_{m-2}) \\
=&R.H.S.
\end{aligned}
$$

For other possibilities of more than one $k_i=0$, the proof is analogous to the previous case or follows from the induction hypothesis. The following takes care of the case when each $k_i \geq 1$:

$$
\begin{aligned}
&\frac{(n-1)(n-2k_0-3k_1-4k_2-\cdots-mk_{m-2})}{k_0k_1k_2\cdots k_{m-2}}+\frac{n-2}{k_1k_2\cdots k_{m-2}} \\
&+\frac{n-3}{k_0k_2k_3\cdots k_{m-2}}+\cdots+\frac{n-m}{k_0k_1k_2\cdots k_{m-3}} \\
&=\frac{1}{k_0k_1k_2\cdots k_{m-2}}\begin{bmatrix} (n^2-2nk_0-3nk_1-4nk_2-\cdots-mnk_{m-2}) \\ +(-n+2k_0+3k_1+4k_2+\cdots+mk_{m-2}) \\ nk_0-2k_0+nk_1-3k_1+\cdots+nk_{m-2}-mk_{m-2} \end{bmatrix} \\
&=\frac{n(n-k_0-2k_1-3k_2-\cdots-(m-1)k_{m-2}-1)}{k_0k_1k_2\cdots k_{m-2}}
\end{aligned}
$$

$\square$

***Proof of the Theorem 2.1.*** Let $\lambda_1,\lambda_2,\cdots,\lambda_m$ be the eigenvalues of $A$. We prove theorem by mathematical induction on the power of the matrix, that is, $n$. For $n=1$, it is trivial and for $n=2$,

$$
\begin{aligned}
Tr(A^2)&=\lambda_1^2+\lambda_2^2+\cdots+\lambda_m^2 \\
&=(\lambda_1+\lambda_2+\cdots+\lambda_m)^2-2\sum_{i\neq j}\lambda_i\lambda_j \\
&=S_1^2-2S_2.
\end{aligned}
$$

In the similar way, the direct computation using the manipulation of eigenvalues yields the proof of the identity (2.4) for $3\leq n\leq m-1$. Henceforth we assume that (2.4) holds for any positive integer less than $n$, where $n \geq m$. The characteristic equation of $A$ is

$$\lambda^m - S_1\lambda^{m-1} + S_2\lambda^{m-2} - S_3\lambda^{m-3} + \cdots + (-1)^m S_m = 0.$$

This, in turn, by the Cayley-Hamilton theorem implies the following:

$$A^m - S_1A^{m-1} + S_2A^{m-2} - S_3A^{m-3} + \cdots + (-1)^m S_m I = 0.$$

The trace being a linear operator, gives, a recursive relation,

$$\begin{aligned}
Tr(A^n) =& S_1Tr(A^{n-1}) - S_2Tr(A^{n-2}) + S_3Tr(A^{n-3}) - \cdots - (-1)^m S_m Tr(A^{n-m}) \\
=& \sum_{k_j\geq 0} \sum_{k_0=0}^{\left\lfloor \frac{n-1-3k_1-4k_2-\cdots-mk_{m-2}}{2}\right\rfloor} \begin{bmatrix} (-1)^{k_0+k_2+k_4+\cdots+k_{\lfloor\frac{m-2}{2}\rfloor}} \\ l_m(n-1, k_0, \cdots, k_{m-2}) \\ S_1^{n-2k_0-3k_1-\cdots-mk_{m-2}} S_2^{k_0}\cdots S_m^{k_{m-2}} \end{bmatrix} \\
&+ \sum_{k_j\geq 0} \sum_{k_0=1}^{\left\lfloor \frac{n-2-3k_1-4k_2-\cdots-mk_{m-2}}{2}+1\right\rfloor} \begin{bmatrix} (-1)^{k_0+k_2+k_4+\cdots+k_{\lfloor\frac{m-2}{2}\rfloor}} \\ l_m(n-2, k_0-1, k_1\cdots, k_{m-2}) \\ S_1^{n-2k_0-3k_1-\cdots-mk_{m-2}} S_2^{k_0}\cdots S_m^{k_{m-2}} \end{bmatrix} \\
&+ \sum_{\substack{k_j\geq 0\\ k_1\geq 1}} \sum_{k_0=0}^{\left\lfloor \frac{n-3k_1-4k_2-\cdots-mk_{m-2}}{2}\right\rfloor} \begin{bmatrix} (-1)^{k_0+k_2+k_4+\cdots+k_{\lfloor\frac{m-2}{2}\rfloor}} \\ l_m(n-3, k_0, k_1-1, k_2, \cdots, k_{m-2}) \\ S_1^{n-2k_0-3k_1-\cdots-mk_{m-2}} S_2^{k_0}\cdots S_m^{k_{m-2}} \end{bmatrix} \\
&+ \cdots \\
&+ \sum_{\substack{k_j\geq 0\\ k_{m-2}\geq 1}} \sum_{k_0=0}^{\left\lfloor \frac{n-3k_1-4k_2-\cdots-mk_{m-2}}{2}\right\rfloor} \begin{bmatrix} (-1)^{k_0+k_2+k_4+\cdots+k_{\lfloor\frac{m-2}{2}\rfloor}} \\ l_m(n-m, k_0, \cdots, k_{m-3}, k_{m-2}-1) \\ S_1^{n-2k_0-3k_1-\cdots-mk_{m-2}} S_2^{k_0}\cdots S_m^{k_{m-2}} \end{bmatrix}.
\end{aligned}$$

Taking certain terms out of the summations and using Lemma 3.3 the theorem follows. □

## 4. Particular cases

As the particular cases, we put on record some interesting observations in this section.

**Corollary 4.1.** *For a $2\times 2$ matrix $A = [a_{ij}]$,*

$$Tr(A^n) = \sum_{k_0=0}^{\lfloor\frac{n}{2}\rfloor} (-1)^{k_0} l_2(n, k_0)\, [Tr(A)]^{n-2k_0}\, [\det(A)]^{k_0}.$$

The following is an interesting fact stating that power and trace commute in case of a singular matrix.

**Corollary 4.2.** *If $A$ is a singular matrix, then $Tr(A^n) = [Tr(A)]^n$.*

**Corollary 4.3.** *If Tr(A)=0, then*

$$Tr(A^n) = \begin{cases} 2(-1)^{\frac{n}{2}}\,[\det(A)]^{\frac{n}{2}}, & \textit{if } n \textit{ is even;} \\ 0, & \textit{if } n \textit{ is odd.} \end{cases}$$

**Corollary 4.4.** *If $Tr(A) = 0 = \det(A)$, then $Tr(A^n) = 0$.*

Now, we apply our scheme of computation to a block matrix. It's noteworthy that in statistics block matrices play a crucial role.

**Corollary 4.5.** *For a block matrix $A$ of order $2m$ of the type*

$$A = \begin{bmatrix} A_1 & & \cdots & 0 \\ & A_2 & & \vdots \\ \vdots & & \ddots & \\ 0 & \cdots & & A_m \end{bmatrix}$$

$$Tr(A^n) = \sum_{r=1}^{m} \sum_{k_0=0}^{\lfloor \frac{n}{2} \rfloor} (-1)^{k_0} l_2(n, k_0) \left[Tr(A_r)\right]^{n-2k_0} \left[\det(A_r)\right]^{k_0}.$$

***Proof.*** Clearly $A^n = \begin{bmatrix} A_1^n & & \cdots & 0 \\ & A_2^n & & \vdots \\ \vdots & & \ddots & \\ 0 & \cdots & & A_m^n \end{bmatrix}$ for all $n \in \mathbb{N}$. Consequently,

$$\begin{aligned} Tr(A^n) &= \sum_{r=1}^{m} Tr(A_r^n) \\ &= \sum_{r=1}^{m} \sum_{k_0=0}^{\lfloor \frac{n}{2} \rfloor} (-1)^{k_0} l_2(n, k_0) \left[Tr(A_r)\right]^{n-2k_0} \left[\det(A_r)\right]^{k_0}. \end{aligned}$$

□

The following is an analogue of [3, Theorem 2.10].

**Proposition 4.6.** *If $A = \begin{bmatrix} a & b \\ b & c \end{bmatrix}$, with $a, b, c \geq 0$, then $2Tr(A^3) \geq Tr(A) \cdot Tr(A^2)$.*

***Proof.***

$$\begin{aligned} 2Tr(A^3) - Tr(A) \cdot Tr(A^2) &= 2\left[Tr(A)\right]^3 - 6Tr(A)\det(A) \\ &\quad - \left[Tr(A)\right]^3 + 2Tr(A)\det(A) \\ &= Tr(A)\left[\left[Tr(A)\right]^2 - 4\det(A)\right] \\ &= Tr(A)\left[(a+c)^2 - 4(ac - b^2)\right] \\ &= Tr(A)\left[(a-c)^2 + 4b^2\right] \geq 0. \end{aligned}$$

□

## 5. Trace of a negative power of $A$

The analogue of the formula (2.4) also holds for the trace of negative powers. We limit ourselves to the matrices of order $2 \times 2$, and hence, $A$ will denote a $2 \times 2$ matrices throughout the rest. The proof is on the same line following Lemma 3.3. The proofs are either direct evidence of the results in the previous sections or an obvious workout. From the characteristic equation and the linearity of the trace, we have

$$Tr(A^n) = \frac{1}{det(A)} \left[Tr(A)Tr(A^{n+1}) - Tr(A^{n+2})\right]. \tag{5.1}$$

For different values of $n$ in (5.1), we have the following

$$Tr(A^{-1}) = \frac{Tr(A)}{\det(A)} \tag{5.2}$$

$$Tr(A^{-2}) = \frac{1}{[\det(A)]^2}\left[[Tr(A)]^2 - 2\det(A)\right] \tag{5.3}$$

$$Tr(A^{-3}) = \frac{1}{[\det(A)]^3}\left[[Tr(A)]^3 - 3Tr(A)\det(A)\right] \tag{5.4}$$

$$Tr(A^{-4}) = \frac{1}{[\det(A)]^4}\left[[Tr(A)]^4 - 4\,[Tr(A)]^2\det(A) + 2\,[\det(A)]^2\right].$$

We conclude the following on the basis of the above observations.

**Theorem 5.1.** *If $A$ is nonsingular, then*

$$Tr(A^{-n}) = \frac{1}{[det(A)]^n}\sum_{k_0=0}^{\lfloor \frac{n}{2}\rfloor}(-1)^{k_0} l_2(n,k_0)\,[Tr(A)]^{n-2k_0}\,[\det(A)]^{k_0}.$$

***Proof.*** Follows from Lemma 3.3. □

**Corollary 5.2.** *If $A$ is nonsingular and $Tr(A) = 0$, then*

$$Tr(A^n) = \begin{cases} \frac{2(-1)^{\frac{n}{2}}}{[\det(A)]^{\frac{n}{2}}}, & \textit{if } n \textit{ is even;} \\ 0, & \textit{if } n \textit{ is odd.}\end{cases}$$

Now we obtain the inequality which is completely analogous to the Proposition 4.6.

**Proposition 5.3.** *For a nonsingular matrix $A = \begin{bmatrix} a & b \\ b & c\end{bmatrix}$ with $a, b, c \geq 0$,*

$$2Tr(A^{-3}) \geq Tr(A^{-1})\cdot Tr(A^{-2}); \quad if\det(A) > 0. \tag{5.5}$$

$$2Tr(A^{-3}) \leq Tr(A^{-1})\cdot Tr(A^{-2}); \quad if\det(A) < 0. \tag{5.6}$$

***Proof.*** From (5.2), (5.3) and (5.4),

$$\begin{aligned} 2Tr(A^{-3}) - Tr(A^{-1})\cdot Tr(A^{-2}) =& \frac{2}{[\det(A)]^3}\left[[Tr(A)]^3 - 3Tr(A)\det(A)\right] \\ &- \frac{Tr(A)}{[\det(A)]^3}\left[[Tr(A)]^2 - 2\det(A)\right] \\ =& \frac{Tr(A)}{[\det(A)]^3}\left[[Tr(A)]^2 - 4\det(A)\right] \\ =& \frac{Tr(A)}{[\det(A)]^3}\left[(a-c)^2 + 4b^2\right]. \end{aligned}$$

Hence, inequalities (5.5) and (5.6) follow. □

**Remark 5.4.** Similar observations could be made for $3\times 3$ and even higher order matrices. However, we have limited ourselves to just one order in this note.

**Acknowledgment.** The author is thankful to the referee for a careful reading and constructive suggestions making the paper readable to more people.

## References

[1] Z. Akyuz and S. Halici, *On some combinatorial identities involving the terms of generalized Fibonacci and Lucas sequences*, Hacet. J. Math. Stat. **42** (4), 431–435, 2013.

[2] H. Avron, *Counting triangles in large graphs using randomized matrix trace estimation*, Proceedings of Kdd-Ldmta'10, 2010.

[3] D.J. Karia, K.M. Patil and H.P. Singh, *On the sum of powers of square matrices*, Oper. Matrices **13** (1), 221–229, 2019.

[4] J.K. Merikoski, *On the trace and the sum of elements of a matrix*, Linear Algebra Appl. **60**, 177–185, 1984.

[5] V.P. Pugačev, *Application of the trace of a matrix to the calculation of its eigenvalues*, Ž. Vyčisl. Mat. i Mat. Fiz. **5**, 114–116, 1965.

[6] A.V. Zarelua, *On congruences for the traces of powers of some matrices*, Tr. Mat. Inst. Steklova, **263** (Geometriya, Topologiya i Matematicheskaya Fizika. I), 85–105, 2008.

Hacettepe Journal of
Mathematics & Statistics

Hacet. J. Math. Stat.
Volume 50 (1) (2021), 24 – 32
DOI : 10.15672/hujms.638900

Research Article

# The multiplicative norm convergence in normed Riesz algebras

Abdullah Aydın

*Department of Mathematics, Muş Alparslan University, Muş, Turkey*

## Abstract

A net $(x_\alpha)_{\alpha\in A}$ in an $f$-algebra $E$ is called multiplicative order convergent to $x \in E$ if $|x_\alpha - x| \cdot u \xrightarrow{o} 0$ for all $u \in E_+$. This convergence was introduced and studied on $f$-algebras with the order convergence. In this paper, we study a variation of this convergence for normed Riesz algebras with respect to the norm convergence. A net $(x_\alpha)_{\alpha\in A}$ in a normed Riesz algebra $E$ is said to be multiplicative norm convergent to $x \in E$ if $\||x_\alpha - x| \cdot u\| \to 0$ for each $u \in E_+$. We study this concept and investigate its relationship with the other convergences, and also we introduce the $mn$-topology on normed Riesz algebras.

**Mathematics Subject Classification (2020).** 46A40, 46E30

**Keywords.** $mn$-convergence, normed Riesz algebra, $mn$-topology, Riesz spaces, Riesz algebra, $mo$-convergence

## 1. Introduction and preliminaries

Let us recall some notations and terminologies used in this paper. An ordered vector space $E$ is said to be *vector lattice* (or, *Riesz space*) if, for each pair of vectors $x, y \in E$, the supremum $x \vee y = \sup\{x, y\}$ and the infimum $x \wedge y = \inf\{x, y\}$ both exist in E. For $x \in E$, $x^+ := x \vee 0$, $x^- := (-x) \vee 0$, and $|x| := x \vee (-x)$ are called the *positive* part, the *negative* part, and the *absolute value* of $x$, respectively. A vector lattice $E$ is called *order complete* if every nonempty bounded above subset has a supremum (or, equivalently, whenever every nonempty bounded below subset has an infimum). A vector lattice is order complete if and only if $0 \le x_\alpha \uparrow\le x$ implies the existence of the $\sup x_\alpha$. A partially ordered set $A$ is called *directed* if, for each $a_1, a_2 \in A$, there is another $a \in A$ such that $a \ge a_1$ and $a \ge a_2$ (or, equivalently, $a \le a_1$ and $a \le a_2$). A function from a directed set $A$ into a set $E$ is called a *net* in $E$. A net $(x_\alpha)_{\alpha\in A}$ in a vector lattice $E$ is *order convergent* (or *o-convergent*, for short) to $x \in E$, if there exists another net $(y_\beta)_{\beta\in B}$ satisfying $y_\beta \downarrow 0$, and for any $\beta \in B$ there exists $\alpha_\beta \in A$ such that $|x_\alpha - x| \le y_\beta$ for all $\alpha \ge \alpha_\beta$. In this case, we write $x_\alpha \xrightarrow{o} x$. An operator $T : E \to F$ between two vector lattices is called *order continuous* whenever $x_\alpha \xrightarrow{o} 0$ in $E$ implies $Tx_\alpha \xrightarrow{o} 0$ in $F$. A vector $e \ge 0$ in a vector lattice $E$ is said to be a *weak order unit* whenever the band generated by $e$ satisfies $B_e = E$, or equivalently, whenever for each $x \in E_+$ we have $x \wedge ne \uparrow x$; see much more information of vector lattices for example [1, 2, 16, 17]. Recall that a net $(x_\alpha)_{\alpha\in A}$ in a vector lattice $E$ is

Email address: aaydin.aabdullah@gmail.com
Received: 28.10.2019; Accepted: 17.04.2020

*unbounded order convergent* (or shortly, *uo-convergent*) to $x \in E$ if $|x_\alpha - x| \wedge u \xrightarrow{o} 0$ for every $u \in E_+$. In this case, we write $x_\alpha \xrightarrow{uo} x$, we refer the reader for an exposition on $uo$-convergence to [3, 5–11].

A vector lattice $E$ under an associative multiplication is said to be a *Riesz algebra* (or, shortly, $l$-algebra) whenever the multiplication makes $E$ an algebra (with the usual properties), and besides, it satisfies the following property: $x \cdot y \in E_+$ for every $x, y \in E_+$. A Riesz algebra $E$ is called *commutative* whenever $x \cdot y = y \cdot x$ for all $x, y \in E$. Also, a subset $A$ of an $l$-algebra $E$ is called *$l$-subalgebre* of $E$ whenever it is also an $l$-algebra under the multiplication operation in $E$

An $l$-algebra $X$ is called: *$d$-algebra* whenever $u \cdot (x \wedge y) = (u \cdot x) \wedge (u \cdot y)$ and $(x \wedge y) \cdot u = (x \cdot u) \wedge (y \cdot u)$ holds for all $u, x, y \in X_+$; *almost $f$-algebra* if $x \wedge y = 0$ implies $x \cdot y = 0$ for all $x, y \in X_+$; *$f$-algebra* if, for all $u, x, y \in X_+$, $x \wedge y = 0$ implies $(u \cdot x) \wedge y = (x \cdot u) \wedge y = 0$; *semiprime* whenever the only nilpotent element in $X$ is zero; *unital* if $X$ has a multiplicative unit. Moreover, any $f$-algebra is both $d$- and almost $f$-algebra (cf. [2, 12, 13, 17]). A vector lattice $E$ is called *Archimedean* whenever $\frac{1}{n}x \downarrow 0$ holds in $E$ for each $x \in E_+$. Every Archimedean $f$-algebra is commutative; see for example [13, p.7]. Assume $E$ is an Archimedean $f$-algebra with a multiplicative unit vector $e$. Then, by applying [17, Thm.142.1(v)], in view of $e = e \cdot e = e^2 \geq 0$, it can be seen that $e$ is a positive vector. On the other hand, since $e \wedge x = 0$ implies $x = x \wedge x = (x \cdot e) \wedge x = 0$, it follows that $e$ is a weak order unit (cf.[12, Cor.1.10]). In this article, unless otherwise, all vector lattices are assumed to be real and Archimedean and all $l$-algebras are assumed to be commutative.

A net $(x_\alpha)_{\alpha \in A}$ in an $f$-algebra $E$ is called *multiplicative order convergent* (or shortly, *mo-convergent*) to $x \in E$ whenever $|x_\alpha - x| \cdot u \xrightarrow{o} 0$ for all $u \in E_+$. Also, it is called *mo-Cauchy* if the net $(x_\alpha - x_{\alpha'})_{(\alpha, \alpha') \in A \times A}$ $mo$-converges to zero. $E$ is called *mo-complete* if every $mo$-Cauchy net in $E$ is $mo$-convergent, and it is also called *mo-continuous* if $x_\alpha \xrightarrow{o} 0$ implies $x_\alpha \xrightarrow{\text{mo}} 0$; see much more detail information [4]. Recall that a norm $\|\cdot\|$ on a vector lattice is said to be a *lattice norm* whenever $|x| \leq |y|$ implies $\|x\| \leq \|y\|$. A vector lattice equipped with a lattice norm is known as a *normed Riesz space* or *normed vector lattice*. Moreover, a normed complete vector lattice is called *Banach lattice*. A net $(x_\alpha)_{\alpha \in A}$ in a Banach lattice $E$ is *unbounded norm convergent* (or *un-convergent*) to $x \in E$ if $\||x_\alpha - x| \wedge u\| \to 0$ for all $u \in E_+$ (cf. [8–10, 15]). We routinely use the following fact: $y \leq x$ implies $u \cdot y \leq u \cdot x$ for all positive elements $u$ in $l$-algebras. So, we can give the following notion.

**Definition 1.1.** An $l$-algebra $E$ which is at the same time a normed Riesz space is called a *normed $l$-algebra* whenever $\|x \cdot y\| \leq \|x\|.\|y\|$ holds for all $x, y \in E$.

Motivated by the above definitions, we give the following notion.

**Definition 1.2.** A net $(x_\alpha)_{\alpha \in A}$ in a normed $l$-algebra $E$ is said to be *multiplicative norm convergent* (or shortly, *mn-convergent*) to $x \in E$ if $\||x_\alpha - x| \cdot u\| \to 0$ for all $u \in E_+$. Abbreviated as $x_\alpha \xrightarrow{\text{mn}} x$. If the condition holds only for sequences then it is called sequentially $mn$-convergence.

In this paper, we study only the $mn$- cases because the sequential cases are analogous in general.

**Remark 1.3.** (i) For a net $(x_\alpha)_{\alpha \in A}$ in a normed $l$-algebra $E$, $x_\alpha \xrightarrow{\text{mn}} x$ implies $x_\alpha \cdot y \xrightarrow{\text{mn}} x \cdot y$ for all $y \in E$ because of $\||x_\alpha \cdot y - x \cdot y| \cdot u\| \leq \||x_\alpha - x| \cdot |y| \cdot u\|$ for all $u \in E_+$; see for example [12, p.1]. The converse holds true in normed $l$-algebras with the multiplication unit. Indeed, assume $x_\alpha \cdot y \xrightarrow{\text{mn}} x \cdot y$ for each $y \in E$. Fix $u \in E_+$. So, $\||x_\alpha - x| \cdot u\| = \||x_\alpha \cdot e - x \cdot e| \cdot u\| \xrightarrow{\text{mn}} 0$.

(ii) In normed $l$-algebras, the norm convergence implies the $mn$-convergence. Indeed, by considering the inequality $\||x_\alpha - x| \cdot u\| \leq \|x_\alpha - x\|.\|u\|$ for any net $x_\alpha \xrightarrow{\mathrm{mn}} x$, we can get the desired result.

(iii) If a net $(x_\alpha)_{\alpha\in A}$ is order Cauchy and $x_\alpha \xrightarrow{\mathrm{mn}} x$ in a normed $l$-algebra then we have $x_\alpha \xrightarrow{\mathrm{mo}} x$. Indeed, since the order Cauchy norm convergent net is order convergent to its norm limit, we can get the desired result.

(iv) In order continuous normed $l$-algebras, it is clear that the $mo$-convergence implies the $mn$-convergence.

(v) In order continuous normed $l$-algebras, following from the inequality $\||x_\alpha - x|\cdot u\| \leq \|x_\alpha - x\|.\|u\|$, the order convergence implies the $mn$-convergence.

(vi) In atomic and order continuous Banach lattice $l$-algebras, an order bounded and $mn$-convergent to zero sequence is sequentially $mo$-convergent to zero; see [9, Lem.5.1.].

(vii) For an $mn$-convergent to zero sequence $(x_n)$ in a Banach lattice $l$-algebra, there is a subsequence $(x_{n_k})$ which sequentially $mo$-converges to zero; see [11, Lem.3.11.].

**Example 1.4.** Let $E$ be a Banach lattice. Fix an element $x \in E$. Then the principal ideal $I_x = \{y \in E : \exists\lambda > 0 \text{ with } |y| \leq \lambda x\}$, generated by $x$ in $E$ under the norm $\|\cdot\|_\infty$ which is defined by $\|y\|_\infty = \inf\{\lambda > 0 : |y| \leq \lambda x\}$, is an $AM$-space; see [2, Thm.4.21.].

Recall that a vector $e > 0$ is called order unit whenever for each $x$ there exists some $\lambda > 0$ with $|x| \leq \lambda e$ (cf. [1, p.20]). Thus, we have $(I_x, \|\cdot\|_\infty)$ is $AM$-space with the unit $|x|$. Since every $AM$-space with the unit, besides being a Banach lattice, has also an $l$-algebra structure (cf. [2, p.259]). So, we can say that $(I_x, \|\cdot\|_\infty)$ is a Banach lattice $l$-algebra. Therefore, for a net $(x_\alpha)_{\alpha\in A}$ in $I_x$ and $y \in I_x$, by applying [2, Cor.4.4.], we get $x_\alpha \xrightarrow{\mathrm{mn}} y$ in the original norm of $E$ on $I_x$ if and only if $x_\alpha \xrightarrow{\mathrm{mn}} y$ in the norm $\|\cdot\|_\infty$. In particular, take $x$ as the unit element $e$ of $E$. Then we have $E_e = E$. Thus, for a net $(x_\alpha)_{\alpha\in A}$ in $E$, we have $x_\alpha \xrightarrow{\mathrm{mn}} y$ in the $(E, \|\cdot\|_\infty)$ if and only if $x_\alpha \xrightarrow{\mathrm{mn}} y$ in the $(E, \|\cdot\|)$.

## 2. The $mn$-convergence on normed $l$-algebras

We begin the section with the next list of properties of $mn$-convergence which follows directly from the inequalities $|x - y| \leq |x - x_\alpha| + |x_\alpha - y|$ and $||x_\alpha| - |x|| \leq |x_\alpha - x|$ for arbitrary net in $(x_\alpha)_{\alpha\in A}$ in vector lattice.

**Lemma 2.1.** *Let $(x_\alpha)_{\alpha\in A}$ and $(y_\beta)_{\beta\in B}$ be two nets in a normed $l$-algebra $E$. Then the followings hold:*

(i) $x_\alpha \xrightarrow{\mathrm{mn}} x \Longleftrightarrow (x_\alpha - x) \xrightarrow{\mathrm{mn}} 0 \Longleftrightarrow |x_\alpha - x| \xrightarrow{\mathrm{mn}} 0$;

(ii) *if* $x_\alpha \xrightarrow{\mathrm{mn}} x$ *then* $y_\beta \xrightarrow{\mathrm{mn}} x$ *for each subnet* $(y_\beta)$ *of* $(x_\alpha)$;

(iii) *suppose* $x_\alpha \xrightarrow{\mathrm{mn}} x$ *and* $y_\beta \xrightarrow{\mathrm{mn}} y$, *then* $ax_\alpha + by_\beta \xrightarrow{\mathrm{mn}} ax + by$ *for any* $a, b \in \mathbb{R}$;

(iv) *if* $x_\alpha \xrightarrow{\mathrm{mn}} x$ *then* $|x_\alpha| \xrightarrow{\mathrm{mn}} |x|$.

The lattice operations in normed $l$-algebras are $mn$-continuous in the following sense.

**Proposition 2.2.** *Let $(x_\alpha)_{\alpha\in A}$ and $(y_\beta)_{\beta\in B}$ be two nets in a normed $l$-algebra $E$. If $x_\alpha \xrightarrow{\mathrm{mn}} x$ and $y_\beta \xrightarrow{\mathrm{mn}} y$ then $(x_\alpha \vee y_\beta)_{(\alpha,\beta)\in A\times B} \xrightarrow{\mathrm{mn}} x \vee y$.*

***Proof.*** Assume $x_\alpha \xrightarrow{\mathrm{mn}} x$ and $y_\beta \xrightarrow{\mathrm{mn}} y$. Then, for a given $\varepsilon > 0$, there exist indexes $\alpha_0 \in A$ and $\beta_0 \in B$ such that $\||x_\alpha - x| \cdot u\| \leq \frac{1}{2}\varepsilon$ and $\||y_\beta - y| \cdot u\| \leq \frac{1}{2}\varepsilon$ for every $u \in E_+$ and for all $\alpha \geq \alpha_0$ and $\beta \geq \beta_0$. It follows from the inequality $|a \vee b - a \vee c| \leq |b - c|$ in vector lattices (cf. [2, Thm.1.9(2)]) that

$$\begin{aligned}\||x_\alpha \vee y_\beta - x \vee y| \cdot u\| &\leq \||x_\alpha \vee y_\beta - x_\alpha \vee y| \cdot u + |x_\alpha \vee y - x \vee y| \cdot u\| \\ &\leq \||y_\beta - y| \cdot u\| + \||x_\alpha - x| \cdot u\| \leq \frac{1}{2}\varepsilon + \frac{1}{2}\varepsilon = \varepsilon\end{aligned}$$

for all $\alpha \geq \alpha_0$ and $\beta \geq \beta_0$ and for every $u \in E_+$. That is, $(x_\alpha \vee y_\beta)_{(\alpha,\beta)\in A\times B} \xrightarrow{\text{mn}} x \vee y$. $\square$

The following proposition is similar to [4, Prop.2.7.], and so we omit its proof.

**Proposition 2.3.** *Let $B$ be a projection band in a normed $l$-algebra $E$ and $P_B$ be the corresponding band projection. Then $x_\alpha \xrightarrow{\text{mn}} x$ in $E$ implies $P_B(x_\alpha) \xrightarrow{\text{mn}} P_B(x)$ in both $E$ and $B$.*

A positive vector $e$ in a normed vector lattice $E$ is called *quasi-interior point* if and only if $\|x - x \wedge ne\| \to 0$ for each $x \in E_+$. If $(x_\alpha)$ is a net in a vector lattice with a weak unit $e$ then $x_\alpha \xrightarrow{\text{uo}} 0$ if and only if $|x_\alpha| \wedge e \xrightarrow{\text{o}} 0$; see [10, Lem. 3.5]. Also, there exist some results for the quasi-interior point case in [9, Lem. 2.11] and for $p$-unit case in [5, Thm. 3.2]. We give an expansion to normed $l$-algebras with the $mn$-convergence for quasi-interior points in the next result.

**Proposition 2.4.** *Let $(x_\alpha)_{\alpha\in A}$ be a positive and decreasing net in a normed $l$-algebra $E$ with a quasi-interior point $e$. Then $x_\alpha \xrightarrow{\text{mn}} 0$ if and only if $(x_\alpha \cdot e)_{\alpha\in A}$ norm converges to zero.*

***Proof.*** The forward implication is immediate because of $e \in E_+$. For the converse implication, fix a positive vector $u \in E_+$ and $\varepsilon > 0$. Thus, for a fixed index $\alpha_1$, we have $x_\alpha \leq x_{\alpha_1}$ for all $\alpha \geq \alpha_0$ because of $(x_\alpha)_{\alpha\in A} \downarrow$. Then we have

$$x_\alpha \cdot u \leq x_\alpha \cdot (u - u \wedge ne) + x_\alpha \cdot (u \wedge ne) \leq x_{\alpha_1} \cdot (u - u \wedge ne) + n(x_\alpha \cdot e)$$

for all $\alpha \geq \alpha_1$ and each $n \in \mathbb{N}$. Hence, we get

$$\|x_\alpha \cdot u\| \leq \|x_{\alpha_1}\|.\|u - u \wedge ne\| + n\|x_\alpha \cdot e\|$$

for every $\alpha \geq \alpha_1$ and each $n \in \mathbb{N}$. So, we can find $n$ such that $\|u - u \wedge ne\| < \frac{\varepsilon}{2\|x_{\alpha_1}\|}$ because $e$ is a quasi-interior point. On the other hand, it follows from $x_\alpha \cdot e \xrightarrow{\|\cdot\|} 0$ that there exists an index $\alpha_2$ such that $\|x_\alpha \cdot e\| < \frac{\varepsilon}{2n}$ whenever $\alpha \geq \alpha_2$. Since index set $A$ is directed, there exists another index $\alpha_0 \in A$ such that $\alpha_0 \geq \alpha_1$ and $\alpha_0 \geq \alpha_2$. Therefore, we get

$$\|x_\alpha \cdot u\| < \|x_{\alpha_0}\| \frac{\varepsilon}{2\|x_{\alpha_0}\|} + n\frac{\varepsilon}{2n} = \varepsilon,$$

and so $\|x_\alpha \cdot u\| \to 0$. $\square$

**Remark 2.5.** A positive and decreasing net $(x_\alpha)_{\alpha\in A}$ in an order continuous Banach $l$-algebra $E$ with weak unit $e$ is $mn$-convergent to zero if and only if $x_\alpha \cdot e \xrightarrow{\|\cdot\|} 0$. Indeed, it is known that $e$ is a weak unit if and only if $e$ is a quasi-interior point in an order continuous Banach lattice; see for example [1, p.135]. Thus, following from Proposition 2.4, one can get the desired result.

The $mn$-convergence passes obviously to any normed $l$-subalgebra $Y$ of a normed $l$-algebra $E$, i.e., for any net $(y_\alpha)_{\alpha\in A}$ in $Y$ with $y_\alpha \xrightarrow{\text{mn}} 0$ in $E$ implies $y_\alpha \xrightarrow{\text{mn}} 0$ in $Y$. For the converse, we give the following theorem whose proof is similar to [4, Thm. 2.10], and so we omit it.

**Theorem 2.6.** *Let $Y$ be a normed $l$-subalgebra of a normed $l$-algebra $E$ and $(y_\alpha)_{\alpha\in A}$ be a net in $Y$. If $y_\alpha \xrightarrow{\text{mn}} 0$ in $Y$ then it $mn$-converges to zero in $E$ for both of the following cases hold;*

(i) *$Y$ is majorizing in $E$;*
(ii) *$Y$ is a projection band in $E$.*

It is known that every Archimedean vector lattice has a unique order completion; see [2, Thm. 2.24]. Moreover, Archimedean commutative $l$-algebra admits the unique extension multiplication to the order completion of it.

**Theorem 2.7.** *Let $E$ and $E^\delta$ be order continuous normed $l$-algebras with $E^\delta$ being order completion of $E$. Then, for a sequence $(x_n)$ in $E$, the followings hold true:*

(i) *If $x_n \xrightarrow{\mathrm{mn}} 0$ in $E$ then there is a subsequence $(x_{n_k})$ of $(x_n)$ such that $x_{n_k} \xrightarrow{\mathrm{mn}} 0$ in $E^\delta$;*

(ii) *If $x_n \xrightarrow{\mathrm{mn}} 0$ in $E^\delta$ then there is a subsequence $(x_{n_k})$ of $(x_n)$ such that $x_{n_k} \xrightarrow{\mathrm{mn}} 0$ in $E$.*

***Proof.*** Let $x_n \xrightarrow{\mathrm{mn}} 0$ in $E$, i.e., $|x_n| \cdot u \xrightarrow{\|\cdot\|} 0$ in $E$ for all $u \in E_+$. Now, let's fix $v \in E^\delta_+$. Then there exists $u_v \in E_+$ such that $v \le u_v$ because $E$ majorizes $E^\delta$. Since $|x_n| \cdot u_v \xrightarrow{\|\cdot\|} 0$, by the standard fact in [1, Exer.13., p.25], there exists a subsequence $(x_{n_k})$ of $(x_n)$ such that $(|x_{n_k}| \cdot u_v)$ order converges to zero in $E$. Thus, we get $|x_{n_k}| \cdot u_v \xrightarrow{o} 0$ in $E^\delta$; see [10, Cor.2.9.]. Then it follows from the inequality $|x_{n_k}| \cdot v \le |x_{n_k}| \cdot u_v$ that we have $|x_{n_k}| \cdot v \xrightarrow{o} 0$ in $E^\delta$. That is, $x_{n_k} \xrightarrow{\mathrm{mo}} 0$ in the order completion $E^\delta$ because $v \in E^\delta_+$ is arbitrary. It follows from the order continuous norm that $x_{n_k} \xrightarrow{\mathrm{mn}} 0$ in the order completion $E^\delta$.

For the converse, put $x_n \xrightarrow{\mathrm{mn}} 0$ in $E^\delta$. Then, for all $u \in E^\delta_+$, we have $|x_n| \cdot u \xrightarrow{\|\cdot\|} 0$ in $E^\delta$. In particular, for all $w \in E_+$, $\||x_n| \cdot w\| \to 0$ in $E^\delta$. Fix $w \in E_+$. Then, again by the standard fact in [1, Exer.13., p.25], we have a subsequence $(x_{n_k})$ of $(x_n)$ such that $(x_{n_k})$ is order convergent to zero in $E^\delta$. Thus, we get $|x_{n_k}| \cdot w \xrightarrow{o} 0$ in $E$. As a result, since $w$ is arbitrary, $x_{n_k} \xrightarrow{\mathrm{mo}} 0$ in $E$. Therefore, one can get the result by using order continuous norm. $\square$

Recall that a subset $A$ in a normed lattice $(E, |\cdot\|)$ is said to *almost order bounded* if, for any $\epsilon > 0$, there is $u_\epsilon \in E_+$ such that $\|(|x| - u_\epsilon)^+\| = \||x| - u_\epsilon \wedge |x|\| \le \epsilon$ for any $x \in A$. For a given normed $l$-algebra $E$, one can give the following definition: a subset $A$ of $E$ is called an *$l$-almost order bounded* if, for any $\epsilon > 0$, there is $u_\varepsilon \in E_+$ such that $\||x| - u_\epsilon \cdot |x|\| \le \epsilon$ for any $x \in A$. Similar to [11, Prop.3.7.], we give the following work.

**Proposition 2.8.** *Let $E$ be a normed $l$-algebra. If $(x_\alpha)_{\alpha\in A}$ is $l$-almost order bounded and $mn$-converges to $x$, then $(x_\alpha)_{\alpha\in A}$ converges to $x$ in norm.*

***Proof.*** Assume $(x_\alpha)_{\alpha\in A}$ is an $l$-almost order bounded net. Then the net $(|x_\alpha - x|)_{\alpha\in A}$ is also $l$-almost order bounded. For any fixed $\varepsilon > 0$, there exists $u_\varepsilon > 0$ such that

$$\||x_\alpha - x| - u_\epsilon \cdot |x_\alpha - x|\| \le \epsilon.$$

Since $x_\alpha \xrightarrow{\mathrm{mn}} x$, we have $\||x_\alpha - x| \cdot u_\varepsilon\| \to 0$. Therefore, following from Proposition 2.2, we get $\|x_\alpha - x\| \le \varepsilon$, i.e., $x_\alpha \to x$ in the norm. $\square$

**Proposition 2.9.** *In an order continuous Banach $l$-algebra, every $l$-almost order bounded mo-Cauchy net converges mn and in norm to the same limit.*

***Proof.*** Assume a net $(x_\alpha)_{\alpha\in A}$ is $l$-almost order bounded and *mo*-Cauchy in an order continuous Banach $l$-algebra $E$. Then the net $(x_\alpha - x_{\alpha'})_{(\alpha,\alpha')\in A\times A}$ is $l$-almost order bounded and is *mo*-convergent to zero. Thus, it *mn*-converges to zero by the order continuity of the norm. Hence, by applying Proposition 2.8, we get that the net $(x_\alpha - x_{\alpha'})_{(\alpha,\alpha')\in A\times A}$ converges to zero in the norm. It follows that the net $(x_\alpha)$ is norm Cauchy, and so it is norm convergent because $E$ is Banach lattice. As a result, we have that $(x_\alpha)$ *mn*-converges to its norm limit by Remark 1.3(*ii*). $\square$

The multiplication in normed $l$-algebra is *mn*-continuous in the following sense.

**Theorem 2.10.** *Let $E$ be a normed $l$-algebra, and $(x_\alpha)_{\alpha\in A}$ and $(y_\beta)_{\beta\in B}$ be two nets in $E$. If $x_\alpha \xrightarrow{\mathrm{mn}} x$ and $y_\beta \xrightarrow{\mathrm{mn}} y$ for some $x, y \in E$ and each positive element of $E$ can be written as a multiplication of two positive elements then we have $x_\alpha \cdot y_\beta \xrightarrow{\mathrm{mn}} x \cdot y$.*

***Proof.*** Assume $x_\alpha \xrightarrow{\mathrm{mn}} x$ and $y_\beta \xrightarrow{\mathrm{mn}} y$. Then $|x_\alpha - x| \cdot u \xrightarrow{\|\cdot\|} 0$ and $|y_\beta - y| \cdot u \xrightarrow{\|\cdot\|} 0$ for every $u \in E_+$. Let's fix $u \in E_+$ and $\varepsilon > 0$. So, there exist indexes $\alpha_0$ and $\beta_0$ such that $\||x_\alpha - x| \cdot u\| \le \varepsilon$ and $\||y_\beta - y| \cdot u\| \le \varepsilon$ for all $\alpha \ge \alpha_0$ and $\beta \ge \beta_0$.

Next, we show the $mn$-convergence of $(x_\alpha \cdot y_\beta)$ to $x \cdot y$. By considering the equality $|x \cdot y| \le |x| \cdot |y|$ (cf. [12, p.1]), we have

$$\begin{aligned}
\||x_\alpha \cdot y_\beta - x \cdot y|u\| &= \||x_\alpha \cdot y_\beta - x_\alpha \cdot y + x_\alpha \cdot y - x \cdot y| \cdot u\| \\
&\le \||x_\alpha| \cdot |y_\beta - y| \cdot u\| + \||x_\alpha - x| \cdot |y| \cdot u\| \\
&\le \||x_\alpha - x| \cdot |y_\beta - y| \cdot u\| + \||y_\beta - y| \cdot |x| \cdot u\| + \||x_\alpha - x| \cdot |y| \cdot u\|.
\end{aligned}$$

The second and the third terms in the last inequality both order converge to zero as $\beta \to \infty$ and $\alpha \to \infty$ respectively because of $|x| \cdot u, |y| \cdot u \in E_+$ and $x_\alpha \xrightarrow{\mathrm{mn}} x$ and $y_\beta \xrightarrow{\mathrm{mn}} y$. Now, let's show the $mn$-convergence of the first term of last inequality. For fixed $u$, we can find two positive elements $u_1, u_2 \in E_+$ such that $u = u_1 \cdot u_2$ because the positive element of $E$ can be written as a multiplication of two positive elements. So, we can get

$$\||x_\alpha - x| \cdot |y_\beta - y| \cdot u\| = \|(|x_\alpha - x| \cdot u_1) \cdot (|y_\beta - y| \cdot u_2)\| \le \||x_\alpha - x| \cdot u_1\|.\||y_\beta - y| \cdot u_2\|.$$

Therefore, we see $|x_\alpha - x| \cdot |y_\beta - y| \cdot u \xrightarrow{\|\cdot\|} 0$. Hence, we get $x_\alpha \cdot y_\beta \xrightarrow{\mathrm{mn}} x \cdot y$. □

In Theorem 2.10, the case of each positive element of $E$ can be written as a multiplication of two positive elements is called *the factorization property* for $f$-algebras in [13, Def.12.10]. But, instead of that property, we can give another easy condition in the following result.

**Corollary 2.11.** *Let $E$ be a normed $l$-algebra, and $(x_\alpha)_{\alpha\in A}$ and $(y_\beta)_{\beta\in B}$ be two nets in $E$. If $x_\alpha \xrightarrow{\mathrm{mn}} x$ and $y_\beta \xrightarrow{\mathrm{mn}} y$ for some $x, y \in E$ and at least one of two nets is eventually norm bounded then we have $x_\alpha \cdot y_\beta \xrightarrow{\mathrm{mn}} x \cdot y$.*

***Proof.*** Modify Theorem 2.10. □

We give some basic notions motivated by their analogies from vector lattice theory.

**Definition 2.12.** Let $(x_\alpha)_{\alpha\in A}$ be a net in a normed $l$-algebra $E$. Then

(1) $(x_\alpha)$ is said to be *mn-Cauchy* if the net $(x_\alpha - x_{\alpha'})_{(\alpha,\alpha')\in A\times A}$ $mn$-converges to 0,
(2) $E$ is called *mn-complete* if every $mn$-Cauchy net in $E$ is $mn$-convergent,
(3) $E$ is called *mn-continuous* if $x_\alpha \xrightarrow{o} 0$ implies that $x_\alpha \xrightarrow{\mathrm{mn}} 0$,

**Proposition 2.13.** *A normed $l$-algebra is mn-continuous if and only if $x_\alpha \downarrow 0$ implies $x_\alpha \xrightarrow{\mathrm{mn}} 0$.*

***Proof.*** Suppose any decreasing to zero net is $mn$-convergent to zero. We show $mn$-continuity. Let $(x_\alpha)_{\alpha\in A}$ be an order convergent to zero net in a normed $l$-algebra $E$. Then there exists another net $z_\beta \downarrow 0$ in $E$ such that, for any $\beta$ there exists $\alpha_\beta$ so that $|x_\alpha| \le z_\beta$, and so $\|x_\alpha\| \le \|z_\beta\|$ for all $\alpha \ge \alpha_\beta$. Since $z_\beta \downarrow 0$, by assumption, we have $z_\beta \xrightarrow{\mathrm{mn}} 0$, i.e., for fixed $\varepsilon > 0$ and $u \in E_+$, there is $\beta_0$ such that $\|z_\beta \cdot u\| < \varepsilon$ for all $\beta \ge \beta_0$. Thus, there exists an index $\alpha_{\beta_0}$ so that $\||x_\alpha| \cdot u\| \le \varepsilon$ for all $\alpha \ge \alpha_{\beta_0}$. Hence, $x_\alpha \xrightarrow{\mathrm{mn}} 0$. The other case is obvious. □

**Proposition 2.14.** *Let $E$ be an mn-continuous and mn-complete normed $l$-algebra. Then every $l$-almost order bounded and order Cauchy net is mn-convergent.*

***Proof.*** Let $(x_\alpha)_{\alpha\in A}$ be an $l$-almost order bounded order Cauchy net. Then the net $(x_\alpha - x_{\alpha'})_{(\alpha,\alpha')\in A\times A}$ is $l$-almost order bounded and is order convergent to zero. Since $E$ is $mn$-continuous, $x_\alpha - x_{\alpha'} \xrightarrow{\mathrm{mn}} 0$. By using Proposition 2.8, we have $x_\alpha - x_{\alpha'} \xrightarrow{\|\cdot\|} 0$. Hence, we get that $(x_\alpha)_{\alpha\in A}$ is $mn$-Cauchy, and so it is $mn$-convergent because of $mn$-completeness. $\square$

## 3. The $mn$-topology on normed $l$-algebra

In this section, we now turn our attention to topology on normed $l$-algebras. We show that the $mn$-convergence in a normed $l$-algebra is topological. While $mo$- and $uo$-convergence need not be given by a topology. But, it was observed in [9] that the $un$-convergence is topological. Motivated from that definition of the $mn$-convergence, we give the following construction of the $mn$-topology.

Let $\varepsilon > 0$ be given. For a non-zero positive vector $u \in E_+$, we put

$$V_{u,\varepsilon} = \{x \in E : \||x| \cdot u\| < \varepsilon\}.$$

Let $\mathcal{N}$ be the collection of all the sets of this form. We claim that $\mathcal{N}$ is a base of neighborhoods of zero for some Hausdorff linear topology. It is obvious that $x_\alpha \xrightarrow{\mathrm{mn}} 0$ if and only if every set of $\mathcal{N}$ contains a tail of this net, hence the $mn$-convergence is the convergence induced by the mentioned topology.

We have to show that $\mathcal{N}$ is a base of neighborhoods of zero. To show this we apply [14, Thm.3.1.10.]. First, note that every element in $\mathcal{N}$ contains zero. Now, we show that for every two elements of $\mathcal{N}$, their intersection is again in $\mathcal{N}$. Take any two set $V_{u_1,\varepsilon_1}$ and $V_{u_2,\varepsilon_2}$ in $\mathcal{N}$. Put $\varepsilon = \varepsilon_1 \wedge \varepsilon_2$ and $u = u_1 \vee u_2$. We show that $V_{u,\varepsilon} \subseteq V_{u_1,\varepsilon_1} \cap V_{u_2,\varepsilon_2}$. For any $x \in V_{u,\varepsilon}$, we have $\||x| \cdot u\| < \varepsilon$. Thus, it follows from $|x| \cdot u_1 \le |x| \cdot u$ that

$$\||x| \cdot u_1\| \le \||x| \cdot u\| < \varepsilon \le \varepsilon_1.$$

Thus, we get $x \in V_{u_1,\varepsilon_1}$. By a similar way, we also have $x \in V_{u_2,\varepsilon_2}$.

Next, it is not a hard job to see that $V_{u,\varepsilon} + V_{u,\varepsilon} \subseteq V_{u,2\varepsilon}$, so that for each $U \in \mathcal{N}$, there is another $V \in \mathcal{N}$ such that $V + V \subseteq U$. In addition, one can easily verify that, for every $U \in \mathcal{N}$ and every scalar $\lambda$ with $|\lambda| \le 1$, we have $\lambda U \subseteq U$.

Now, we show that, for each $U \in \mathcal{N}$ and each $y \in U$, there exists $V \in \mathcal{N}$ with $y + V \subseteq U$. Suppose $y \in V_{u,\varepsilon}$. We should find $\delta > 0$ and a non-zero $v \in E_+$ such that $y + V_{v,\delta} \subseteq V_{u,\varepsilon}$. Take $v := u$. Hence, since $y \in V_{u,\varepsilon}$, we have $\||y| \cdot u\| < \varepsilon$. Put $\delta := \varepsilon - \||y| \cdot u\|$. We claim that $y + V_{v,\delta} \subseteq V_{u,\varepsilon}$. Let's take $x \in V_{v,\delta}$. We show that $y + x \in V_{u,\varepsilon}$. Consider the inequality $|y + x| \cdot u \le |y| \cdot u + |x| \cdot u$. Then we have

$$\||y + x| \cdot u\| \le \||y| \cdot u\| + \||x| \cdot u\| < \||y| \cdot u\| + \delta = \varepsilon.$$

Finally, we show that this topology is Hausdorff. It is enough to show that $\bigcap \mathcal{N} = \{0\}$. Suppose that it is not hold true, i.e., assume that $0 \ne x \in V_{u,\varepsilon}$ for all non-zero $u \in E_+$ and for all $\varepsilon > 0$. In particular, take $x \in V_{|x|,\varepsilon}$. Thus, we have $\||x|^2\| < \varepsilon$. Since $\varepsilon$ is arbitrary, we get $|x|^2 = 0$, i.e., $x = 0$ by using [17, Thm.142.3.]; a contradiction.

Recall that the statement $V_{u,\varepsilon}$ is either contained in $[-u, u]$ or contains a non-trivial ideal holds true for the $un$-topology. However, it is not true for the $mn$-topology. To see this, we give the following counterexample.

**Example 3.1.** Consider the $l$-algebra $E = C[0, 1]$ with the sup-norm topology $\tau$. Take $a = \mathbb{1}$ and $A = B(0, 10)$. The set $U_{a,A} = \{x \in E : |x| \cdot a \in A\} = B(0, 10)$ is neither contained in $[-a, a] = [-\mathbb{1}, \mathbb{1}] = B(0, 1)$ nor contains a non-trivial ideal.

**Lemma 3.2.** *If $V_{u,\varepsilon}$ is contained in $[-u,u]$, then $u$ is a strong unit.*

***Proof.*** Take a positive element $x \in E_+$. Then we have a positive scalar $\lambda$ such that $(\lambda x)\cdot a \in A$. Thus we get $\lambda x \in U_{a,A}$ and so, $\lambda x \in [-a,a]$. Then one can see that $a$ is a strong unit. □

## 4. The $mn$-convergence on semiprime normed $f$-algebras

Recall that an element $x$ in an $f$-algebra $E$ is called *nilpotent* whenever $x^n = 0$ for some natural number $n \in \mathbb{N}$. The algebra $E$ is called *semiprime* if the only nilpotent element in $E$ is the null element ([17, p.670]). We begin the section with the next useful result.

**Proposition 4.1.** *Let $(x_\alpha)_{\alpha\in A}$ be a net in nilpotent elements of a normed $f$-algebra $E$. If $x_\alpha \xrightarrow{\text{mn}} x$ then $x$ is also a nilpotent element.*

***Proof.*** Take a fixed positive element $u \in E_+$. Then, by using [13, Prop.10.2(iii)] and [17, Thm.142.1(ii)], we get

$$\| |x_\alpha - x|\cdot u\| = \| |x_\alpha \cdot u - x\cdot u| \| = \|x_\alpha\cdot u - x\cdot u\| = \|x\cdot u\| \to 0.$$

Thus $\|x\cdot u\| = 0$ and hence $x\cdot u = 0$ for every $u \in X_+$. Then $y\cdot x = 0$ for all $y \in E$. It follows now from [12, p.157] that $x$ is nilpotent in $E$. □

**Remark 4.2.** By considering Proposition 4.1, it is easy to see that $mn$-convergence in normed $f$-algebra $E$ has an unique limit if and only if $E$ is semiprime normed $f$-algebra.

Unless stated otherwise, we will assume that $E$ is a semiprime normed $f$-algebra and all nets and vectors lie in $E$.

**Proposition 4.3.** *Let $(x_\alpha)_{\alpha\in A}$ be a net in $E$. Then we have that*

(i) $0 \le x_\alpha \xrightarrow{\text{mn}} x$ *implies* $x \in E_+$,

(ii) *if* $(x_\alpha)$ *is monotone and* $x_\alpha \xrightarrow{\text{mn}} x$ *then* $x_\alpha \xrightarrow{\text{o}} x$.

***Proof.*** (i) Assume $(x_\alpha)_{\alpha\in A}$ consists of non-zero elements and $mn$-converges to $x \in E$. Then, by using Proposition 2.2, we have $x_\alpha = x_\alpha^+ \xrightarrow{\text{mn}} x^+$. Also, following from Remark 4.2, we get $x^+ = x$. Therefore, we get $x \in E_+$.

(ii) For the order convergence of $(x_\alpha)_{\alpha\in A}$, it is enough to show that $x_\alpha \uparrow$ and $x_\alpha \xrightarrow{\text{mn}} x$ implies $x_\alpha \xrightarrow{\text{o}} x$. For a fixed index $\alpha$, we have $x_\beta - x_\alpha \in X_+$ for all $\beta \ge \alpha$. By applying (i), we can see $x_\beta - x_\alpha \xrightarrow{\text{mn}} x - x_\alpha \in X_+$ as $\beta \to \infty$. Therefore, $x \ge x_\alpha$ for the index $\alpha$. Since $\alpha$ is arbitrary, $x$ is an upper bound of $(x_\alpha)$. Assume $y$ is another upper bound of $(x_\alpha)$, i.e., $y \ge x_\alpha$ for all $\alpha$. So, $y - x_\alpha \xrightarrow{\text{mn}} y - x \in X_+$, or $y \ge x$, and so $x_\alpha \uparrow x$. □

**Theorem 4.4.** *The following statements are equivalent:*

(i) *$E$ is mn-continuous;*

(ii) *if $0 \le x_\alpha \uparrow \le x$ holds in $E$ then $(x_\alpha)$ is an mn-Cauchy net;*

(iii) *$x_\alpha \downarrow 0$ implies $x_\alpha \xrightarrow{\text{mn}} 0$ in $E$.*

***Proof.*** (i)⇒(ii) Take a net $0 \le x_\alpha \uparrow \le x$ in $E$. Then there exists another net $(y_\beta)$ in $E$ such that $(y_\beta - x_\alpha)_{\alpha,\beta} \downarrow 0$; see [2, Lem.4.8]. Thus, by applying Proposition 2.13, we have $(y_\beta - x_\alpha)_{\alpha,\beta} \xrightarrow{\text{mn}} 0$ because $E$ is $mn$-continuous. Therefore, the net $(x_\alpha)$ is $mn$-Cauchy because of $\|x_\alpha - x_{\alpha'}\|_{\alpha,\alpha'\in A} \le \|x_\alpha - y_\beta\| + \|y_\beta - x_{\alpha'}\|$.

(ii)⇒(iii) Put $x_\alpha \downarrow 0$ in $E$ and fix arbitrary $\alpha_0$. Thus, we have $x_\alpha \le x_{\alpha_0}$ for all $\alpha \ge \alpha_0$, and so we can get $0 \le (x_{\alpha_0} - x_\alpha)_{\alpha\ge\alpha_0} \uparrow \le x_{\alpha_0}$. Then it follows from $(ii)$ that the net $(x_{\alpha_0} - x_\alpha)_{\alpha\ge\alpha_0}$ is $mn$-Cauchy, i.e., $(x_{\alpha'} - x_\alpha) \xrightarrow{\text{mn}} 0$ as $\alpha_0 \le \alpha, \alpha' \to \infty$. Since $E$ is $mn$-complete, there exists an element $x \in E$ satisfying $x_\alpha \xrightarrow{\text{mo}} x$ as $\alpha_0 \le \alpha \to \infty$. It follows from Proposition 4.3 that $x_\alpha \downarrow 0$ because of $x_\alpha \downarrow$ and $x_\alpha \xrightarrow{\text{mn}} 0$, and so, following from Remark 4.2 tha we have $x = 0$. Therefore, we get $x_\alpha \xrightarrow{\text{mn}} 0$.

(iii)⇒(i) It is just the implication of Proposition 2.13. □

**Corollary 4.5.** *Every mn-continuous and mn-complete normed f-algebra E is order complete.*

***Proof.*** Suppose $E$ is $mn$-continuous and $mn$-complete. For $y \in E_+$, put a net $0 \leq x_\alpha \uparrow \leq y$ in $E$. By applying Theorem 4.4 $(ii)$, the net $(x_\alpha)$ is $mn$-Cauchy. Thus, there exists an element $x \in E$ such that $x_\alpha \xrightarrow{\text{mn}} x$ because of $mn$-completeness. Since $x_\alpha \uparrow$ and $x_\alpha \xrightarrow{\text{mo}} x$, it follows from Lemma 4.3 that $x_\alpha \uparrow x$. Therefore, $E$ is order complete. □

**Acknowledgment.** The author would like to thank Eduard Emelyanov and Mohamed Ali Toumi for improving the paper.

## References

[1] Y. Abramovich and C.D. Aliprantis, *An Invitation to Operator Theory*, American Mathematical Society, New York, 2003.
[2] C.D. Aliprantis and O. Burkinshaw, *Positive Operators*, Springer, Dordrecht, 2006.
[3] A. Aydın, *Unbounded $p_\tau$-convergence in vector lattice normed by locally solid vector lattices*, in: Academic Studies in Mathematics and Natural Sciences-2019/2, 118-134, IVPE, Cetinje-Montenegro, 2019.
[4] A. Aydın, *Multiplicative order convergence in f-algebras*, Hacet. J. Math. Stat. **49** (3), 998–1005, 2020.
[5] A. Aydın, E. Emel'yanov, N.E. Özcan, and M.A.A. Marabeh, *Compact-like operators in lattice-normed spaces*, Indag. Math. **2**, 633-656, 2018.
[6] A. Aydın, E. Emel'yanov, N.E. Özcan, and M.A.A. Marabeh, *Unbounded p-convergence in lattice-normed vector lattices*, Sib. Adv. Math. **29**, 153-181, 2019.
[7] A. Aydın, S.G. Gorokhova, and H. Gül, *Nonstandard hulls of lattice-normed ordered vector spaces*, Turkish J. Math. **42**, 155-163, 2018.
[8] Y.A. Dabboorasad, E.Y. Emelyanov, and M.A.A. Marabeh, *$u\tau$-Convergence in locally solid vector lattices*, Positivity **22**, 1065-1080, 2018.
[9] Y. Deng, M. O'Brien, and V.G. Troitsky, *Unbounded norm convergence in Banach lattices*, Positivity **21**, 963-974, 2017.
[10] N. Gao, V.G. Troitsky, and F. Xanthos, *Uo-convergence and its applications to Cesáro means in Banach lattices*, Israel J. Math. **220**, 649-689, 2017.
[11] N. Gao and F. Xanthos, *Unbounded order convergence and application to martingales without probability*, Math. Anal. Appl. **415**, 931-947, 2014.
[12] C.B. Huijsmans and B.D. Pagter, *Ideal theory in f-algebras*, Trans. Amer. Math. Soc. **269**, 225-245, 1982.
[13] B.D. Pagter, *f-Algebras and Orthomorphisms*, Ph.D. Dissertation, Leiden, 1981.
[14] V. Runde, *A Taste of Topology*, Springer, Berlin, 2005.
[15] V.G. Troitsky, *Measures of non-compactness of operators on Banach lattices*, Positivity **8**, 165-178, 2004.
[16] B.Z. Vulikh, *Introduction to the Theory of Partially Ordered Spaces*, Wolters-Noordhoff Scientific Publications, Groningen, 1967.
[17] A.C. Zaanen, *Riesz Spaces II*, The Netherlands: North-Holland Publishing Co., Amsterdam, 1983.

Hacettepe Journal of
Mathematics & Statistics

Hacet. J. Math. Stat.
Volume 50 (1) (2021), 33 – 40
DOI : 10.15672/hujms.519987

Research Article

# Sharp upper bounds of $A_\alpha$-spectral radius of cacti with given pendant vertices

Shaohui Wang[1], Chunxiang Wang*[2], Jia-Bao Liu[3]

[1] *Department of Mathematics, Louisiana College, Pineville, LA 71359, USA*
[2] *School of Mathematics and Statistics, Central China Normal University, Wuhan, 430079, P.R. China*
[3] *School of Mathematics and Physics, Anhui Jianzhu University, Hefei, 230601, P.R. China*

**Abstract**

For $\alpha \in [0,1]$, let $A_\alpha(G) = \alpha D(G) + (1-\alpha)A(G)$ be $A_\alpha$-matrix, where $A(G)$ is the adjacent matrix and $D(G)$ is the diagonal matrix of the degrees of a graph $G$. Clearly, $A_0(G)$ is the adjacent matrix and $2A_{\frac{1}{2}}$ is the signless Laplacian matrix. A connected graph is a cactus graph if any two cycles of $G$ have at most one common vertex. We first propose the result for subdivision graphs, and determine the cacti maximizing $A_\alpha$-spectral radius subject to fixed pendant vertices. In addition, the corresponding extremal graphs are provided. As consequences, we determine the graph with the $A_\alpha$-spectral radius among all the cacti with $n$ vertices; we also characterize the $n$-vertex cacti with a perfect matching having the largest $A_\alpha$-spectral radius.

**Mathematics Subject Classification (2020).** 05C50, 15A48

**Keywords.** adjacent matrix, trees, cacti, bounds

## 1. Introduction

Throughout this paper, we consider finite simple connected graph $G$ with vertex set $V(G)$ and edge set $E(G)$. The order of a graph is the number of vertices $|V(G)| = n$ and the size is the number of edges $|E(G)|$. Let $v \in V(G)$ be a vertex of $G$, $N(v) = N_G(v) = \{w \in V(G), vw \in E(G)\}$ be the neighborhood of $v$ , and $d_G(v)$ (or briefly $d_v$) be the degree of $v$ with $d_G(v) = |N(v)|$. If $e$ is an edge of $G$ and $G - e$ contains at least two components, then $e$ is a cut edge of $G$. If $P_k = v_1v_2 \cdots v_k$ is a subgraph of $G$ such that $v_1$ is a cut vertex of degree at least 3, $d(v_k) = 1$ and $d(v_i) = 2$ for $i \in [2, k-1]$, then $P_k$ is called a pendant path in $G$. For other undefined notations and terminologies, refer to [2].

It's known that $A(G)$ is the adjacency matrix and $D(G)$ is the diagonal matrix of the degrees of $G$. The signless Laplacian matrix of $G$ is $Q(G) = D(G) + A(G)$. For $\alpha \in [0,1]$, the $A_\alpha$-matrix

$$A_\alpha(G) = \alpha D(G) + (1-\alpha)A(G)$$

is given by Nikiforov [15]. Clearly, $A_0(G)$ is the adjacent matrix and $2A_{\frac{1}{2}}$ is the signless Laplacian matrix of $G$, respectively.

*Corresponding Author.
Email addresses: shaohuiwang@yahoo.com (S. Wang), wcxiang@mail.ccnu.edu.cn (C. Wang), liujiabaoad@163.com (J.-B. Liu)
Received: 31.01.2019; Accepted: 20.04.2020

The studies of the (adjacency, signless Laplacian) spectral radius are interesting and meaningful [7, 10–12, 19–23]. As examples, the spectral radius of trees are proposed by Lovász and J. Pelikán [14]. Feng et al.[10] studied the minimal Laplacian spectral radius of trees with given matching number. Chen [4] found the properties of spectra of graphs and their line graphs. Cvetković [8] explored the signless Laplacian spectra of graphs and a spectral theory in graphs. The bounds of signless Laplacian spectral radius and its hamiltonicity are studied by Zhou [24]. Lin and Zhou [13] obtained graphs with at most one signless Laplacian eigenvalue larger than three. In addition to the successful considerations of these spectral radius, $A_\alpha$-spectral radius is provided as a general version of adjacency and signless Laplacian radius, and this area would be challenging. For the $A_\alpha$-spectral radius, Nikiforov et al. [15, 16]introduced some properties of this spectral radius and provided the upper bounds on trees.

It is known that a tree is a noncyclic graph. If some vertices in a tree are replaced by cycles, then this graph has some cycles. The trees are extended as the definition that a cactus graph is a connected graph such that any two cycles have at most one common vertex. Denoted by $\mathcal{C}_n^k$ the set of all cacti with $n$ vertices and $k$ pendant vertices.

The cactus graphs have attracted many interests among the mathematical literature including algebra and graph theory. For instance, the properties of cacti with $n$ vertices [3] are explored by Borovićanin and Petrović. Chen and Zhou [5] investigated some upper bounds of the signless Laplacian spectral radius of cactus graphs. The signless Laplacian spectral radius of cacti with given matching number are obtained by Shen et al. [17]. Some results for spectral radius on cacti with $k$ pendant vertices are studied Wu et al. [18]. Ye et al. [22] gave the maximal adjacency or signless Laplacian spectral radius of graphs subject to fixed connectivity.

Motivated by the above results, in this paper, we generalize the results of $A_\alpha$-spectra from the trees to the cacti subject to fixed pendant vertices. For $\alpha \in [0,1]$, we first propose the result for subdivision graphs, and determine the cacti maximizing $A_\alpha$-spectral radius subject to fixed pendant vertices. In addition, the corresponding extremal graphs are determined. As consequences, we determine the graph with the $A_\alpha$-spectral radius among all the cacti with $n$ vertices; we also characterize the $n$-vertex cacti with a perfect matching having the largest $A_\alpha$-spectral radius.

## 2. Preliminary

In this section, we provide some important concepts and lemmas that will be used in the main proofs.

If $G$ is a graph with vertex set $V(G) = \{v_1, v_2, \cdots, v_n\}$ and edge set $E(G)$, then the $A_\alpha-$matrix $A_\alpha(G)$ of $G$ has the $(i,j)$-entry of $A_\alpha(G)$ is $1-\alpha$ if $v_iv_j \in E(G)$; $\alpha d(v_i)$ if $i = j$, and otherwise 0. For $\alpha \in [0,1]$, let $\lambda_1(A_\alpha(G)) \geq \lambda_2(A_\alpha(G)) \geq \cdots \geq \lambda_n(A_\alpha(G))$ be the eigenvalues of $A_\alpha(G)$. The $A_\alpha$-spectral radius of $G$ is considered as the maximal eigenvalue $\rho(G) := \lambda_1(A_\alpha(G))$. Let $X = (x_{v_1}, x_{v_2}, \cdots, x_{v_n})^T$ be a real vector of $\rho(G)$. By $A_\alpha(G) = \alpha D(G) + (1-\alpha)A(G)$, we have the quadratic formula of $X^TA_\alpha(G)X$ can be expressed that

$$X^TA_\alpha(G)X = \alpha \sum_{v_i \in V(G)} x_{v_i}^2 d_{v_i} + 2(1-\alpha) \sum_{v_iv_j \in E(G)} x_{v_i}x_{v_j}.$$

Because $A_\alpha(G)$ is a real symmetric matrix, and by Rayleigh principle, we have the formula $\rho(G) = max_{X \neq 0} \frac{X^TA_\alpha(G)X}{X^TX}$. Furthermore, if $X$ is a unit eigenvector of$A_\alpha(G)$ corresponding to $\rho(G)$, then we have the formula $\rho(G) = X^TA_\alpha(G)X$.

As we know that once $X$ is an eigenvector of $\rho(G)$ for a connected graph $G$, $X$ should be unique and positive. The corresponding eigenequations for $A_\alpha(G)$ is rewritten as

$$\rho(G)x_{v_i} = \alpha d_{v_i} x_{v_i} + (1-\alpha) \sum_{v_i v_j \in E(G)} x_{v_j}. \tag{2.1}$$

As $A_1(G) = D(G)$, we study the $A_\alpha-$matrix for $\alpha \in [0,1)$ below. Based on the definition of $A_\alpha$-spectral radius, we have

**Lemma 2.1** ([16, 21])**.** Denote by $A_\alpha(G)$ the $A_\alpha$-matrix of a connected graph $G$ with $\alpha \in [0,1)$, $v, w \in V(G)$, $u \in S \subset V(G)$ such that $S \subset N(v) \setminus (N(w) \cup \{w\})$. Let $H$ be a graph with vertex set $V(G)$ and edge set $E(G) \setminus \{uv, u \in S\} \cup \{uw, u \in S\}$, and $X$ a unit eigenvector to $\rho(A_\alpha(G))$. If $x_w \geq x_v$ and $|S| \neq 0$, then $\rho(H) \geq \rho(G)$.

**Lemma 2.2** ([22])**.** Let $A_\alpha(G)$ the $A_\alpha$-matrix of a connected graph $G$ with $\alpha \in [0,1)$, $s, t, u, v \in V(G)$, $st, uv \in E(G)$, $sv, tu \notin E(G)$. Let $H$ be a graph with vertex set $V(G)$ and edge set $E(G) \setminus \{uv, st\} \cup \{sv, ut\}$, and $X$ a unit eigenvector to $\rho(A_\alpha(G))$. If $(x_s - x_u)(x_v - x_t) \geq 0$, then $\rho(H) \geq \rho(G)$.

If $G$ is a connected graph, then $A_\alpha(G)$ is a nonnegative irreducible symmetric matrix. By the results of [1, 6, 15], if we add some edges to a connected graph, then $A_\alpha$-spectral radius will increase and the following lemma is straightforward.

**Lemma 2.3.** If $H$ is a proper subgraph of a connected graph $G$, and $\rho$ is the $A_\alpha$-spectral radius, then $\rho(H) < \rho(G)$.

Let $P_t = v_0 v_1 v_2 \cdots v_t$ be a subgraph of $G$. If $v_0$ is a cut vertex of degree at least 3, $d(v_t) = 1$ and $d(v_j) = 2$ with $j \in [1, t-1]$, then $P_t$ is called a pendant path in $G$. The following lemma is useful below.

**Lemma 2.4.** Let $G \in \mathcal{C}_n^k$. If $\rho(G)$ is maximal, then all pendant paths share a common vertex.

***Proof.*** Assume that $G$ is a cactus graph with $k$ pendant vertices and contains at least two pendant paths $P_t = v_0 v_1 \cdots v_t$ and $P_s = u_0 u_1 \cdots u_s$. Note that $d(u_0), d(v_0) \geq 3$. Without loss of generality, let $x_{v_0} \geq x_{u_0}$. Suppose that $u_0$ is a vertex in a cycle and this cycle contains at least one edge of the shortest path $P[u_0, v_0]$ between $u_0$ and $v_0$. Set $G_1$ to be a new graph with vertex set $V(G)$ and edge set $E(G) \setminus \{u_0 v, v \in N\} \cup \{v_0 v, v \in N\}$ with $N = N(u_0) \setminus \{w_1, w_2\}$, where $w_1$ is in $P[u_0, v_0]$, and $v_0, w_1, w_2$ are in the same cycle; if $u_0$ is not in any cycle, then let $G_2$ be a new graph with vertex set $V(G)$ and edge set $E(G) - \{u_0 v, v \in N\} \cup \{v_0 v, v \in N\}$ with $N = N(u_0) \setminus \{w_1, w_2\}$, where $w_1$ is in the shortest path between $v_0$ and $u_0$, and $w_2$ is another neighbor of $u_0$.

Note that both $G_1$ and $G_2$ are cacti with $k$ pendant vertices. By Lemma 2.1, we have $\rho(G_1) \geq \rho(G)$ and $\rho(G_2) \geq \rho(G)$. We can continue this process and move all pendant paths to a common vertex such that $\rho(G)$ is increasing. Then this lemma is proved. $\square$

**Lemma 2.5.** Let $G \in \mathcal{C}_n^k$. If $\rho(G)$ is maximal, then the length of any pendant path is at most 2, and there is at most one pendant path of the length 2.

***Proof.*** First we prove the length of any pendant path is at most 2. We prove it by a contradiction. Assume there are have a pendent path $P$, $P = v_0 v_1 \cdots v_m$, $m \geq 3$. Let $G_1$ be a new graph with vertex set $V(G)$ and $E(G) + v_1 v_{m-1}$, then $G_1$ is a cactus with $k$ pendent vertices and $\rho(G_1) > \rho(G)$ (by Lemma 2.3). Then there exists a contradicted graph. Thus, if $\rho(G)$ is maximal, then the length of any pendant path is at most 2. Next we prove there is at most one pendant path of length 2. Suppose there are $r$, $(r > 1)$ pendent path of the length 2. Without loss of generality $P_i = v_0 v_{i1} v_{i2}; (i = 1, 2, \cdots, r)$. Let $G_2$ be a new graph with vertex set $V(G)$ and $E(G) \cup \{v_{11} v_{21}, v_{31} v_{41}, \cdots, v_{(2\lfloor \frac{r}{2} \rfloor - 1)1} v_{(2\lfloor \frac{r}{2} \rfloor)1}\}$, then $G_2$ is a cactus with $k$ pendent vertices and $\rho(G_2) > \rho(G)$ (by Lemma 2.3). Then there exists a contradicted graph. Thus, if $\rho(G)$ is maximal, there is at most one pendant path of the length 2. This completes the proof. $\square$

**Lemma 2.6.** Let $G \in \mathcal{C}_n^k$. If $\rho(G)$ is maximal, then there does not exist an internal path such that it is built by cut edges.

***Proof.*** We prove it by a contradiction. Note that $d(v_0), d(v_t) \geq 3$. Let $P_t = v_0v_1 \cdots v_t$ be an internal path of $G$ such that every edge of $P_t$ is an cut edge. If $t \geq 2$, then let $G_1 = G + v_0v_t$. Then $G_1$ is a cactus with $k$ pendant vertices and $G$ is a proper subgraph of $G_1$. By Lemma 2.3, we have $\rho(G_1) > \rho(G)$, which is a contradiction. Next we consider $t = 1$. Without loss of generality, let $x_0 \geq x_1$ and $w \in N(v_1) \setminus \{v_0, v_1'\}$ such that $v_1'$ is a neighbor except for $v_0$. Denote a new graph $G_2$ with vertex set $V(G_2) = V(G)$ and edge set $E(G_2) = E(G) \setminus \{v_1w, w \in N(v_1) \setminus \{v_0, v_1'\}\} \cup \{v_0w, w \in N(v_1) \setminus \{v_0, v_1'\}\}$. Then $G_2$ is a cactus with $k$ pendant vertices and $\rho(G_2) \geq \rho(G)$ (by Lemma 2.1). These are contradictions and this lemma is proved. $\square$

**Lemma 2.7.** Let $G \in \mathcal{C}_n^k$. If $\rho(G)$ is maximal, then all cycles share a common vertex.

***Proof.*** Suppose that there are two cut vertices $v_0, v_1$ in $G$ such that not all cycles contain them. If there are only two cycles, then it is proved by Lemma 2.6: there does not exist an internal path such that it is built by cut edges. If there are more 3 cycles, then choose such $v_0$ and $v_1$ having the longest distance. Then $d(v_0), d(v_1) \geq 4$. Without loss of generality, let $x_{v_0} \geq x_{v_1}$ and $w \in N(v_1) \setminus \{v_0\}$. Denote a new graph $G_1$ with vertex set $V(G_1) = V(G)$ and edge set $E(G_1) = E(G) \setminus \{v_1w, w \in N(v_1) \setminus \{v_l, v_l'\}\} \cup \{v_0w, w \in N(v_1) \setminus \{v_l, v_l'\}\}$, where $v_l, v_l'$ are neighbors of $v_1$ and on a same cycle. Then $G_2$ is a cactus with $k$ pendant vertices and $\rho(G_1) \geq \rho(G)$ (by Lemma 2.1). We can continue this method to increase $\rho(G)$ until there exist a unique cut vertex sharing with all cycles. So, the result is proved. $\square$

**Lemma 2.8.** Let $G \in \mathcal{C}_n^k$. If $\rho(G)$ is maximal, then the length of any cycle is at most 4, and there is at most one cycle of length 4.

***Proof.*** Let $C_t = v_1v_2 \cdots v_tv_1$ be a cycle of length $t$ in $G$ and $v_1$ is a cut vertex. If $x_{v_1} \geq x_{v_3}$, we build a new graph $G_1$ such that $V(G_1) = V(G)$ and $E(G_1) = E(G) \setminus \{v_3v_4\} \cup \{v_1v_4\}$. Then $\rho(G) \leq \rho(G_1)$ (by Lemma 2.1). In addition, $G_1$ is a subgraph of $G_2 = G_1 \cup \{v_1v_3\}$, which yields that $\rho(G_1) < \rho(G_2)$ (by Lemma 2.3). If $x_{v_1} \leq x_{v_3}$, then we set up a graph $G_3$ such that $V(G_3) = V(G)$ and $E(G_3) = E(G) \setminus \{v_tv_1\} \cup \{v_tv_3\}$. We have $\rho(G) \leq \rho(G_3)$ (by Lemma 2.1). $G_4$ is a graph by connecting $v_1$ and $v_3$ from $G_3$. So, $G_3$ is a subgraph of $G_4$. By Lemma 2.3, we have $\rho(G_4) > \rho(G_3)$. Thus, if $G$ contains a cycle of length at least 5, then there exists a contradicted graph.

Next we show that there is at most one cycle of length 4. Suppose that there at at least two 4-cycles $C_1$ and $C_2$ in $G$. By Lemma 2.7, these two cycles share a common cut vertex. Let $C_1 = v_0v_1v_2v_3v_0$ and $C_2 = v_0u_1u_2u_3v_0$. If $x_{v_0} \geq min\{x_{v_1}, x_{v_3}\}$ and $x_{v_0} \geq min\{x_{u_1}, x_{u_3}\}$, say $x_{v_0} \geq x_{v_1}, x_{v_0} \geq x_{u_1}$, then we set a new graph $H_1$ such that $V(H_1) = V(G)$ and $E(H_1) = E(G) \setminus \{v_2v_1, u_2u_1\} \cup \{v_2v_0, u_2v_0\}$. By Lemma 2.1, we have $\rho(G) \leq \rho(H_1)$. Let $H_2$ be a graph from $H_1$ by connecting $u_1v_1$. Since $H_2$ is a proper subgraph of $H_1$, then $\rho(H_1) < \rho(H_2)$. This is a contradiction to the assumption that $\rho(G)$ is maximal.

If $x_{v_0} \leq min\{x_{v_1}, x_{v_3}\}$ and $x_{v_0} \leq min\{x_{u_1}, x_{u_3}\}$, say $x_{v_0} \leq x_{v_1}, x_{v_0} \leq x_{u_1}$, then we set new graphs $H_3$ with vertex set $V(H_3) = V(G)$ and $E(H_3) = E(G) \setminus \{v_3v_0, u_3u_0\} \cup \{v_3v_1, u_3u_1\}$, $H_4$ from $H_3$ by connecting $v_1u_1$. By Lemmas 2.1,2.3, we have $\rho(G) < \rho(H_3) < \rho(H_4)$. We can use Lemma 2.7 to find a graph in $\mathcal{C}_n^k$ with only one common vertex among cycles. This is a contradiction to the choice of $G$.

Lastly, without loss of generality, we consider the case of $max\{x_{u_1}, x_{u_3}\} \leq x_{v_0} \leq min\{x_{v_1}, x_{v_3}\}$, say $x_{u_1} \leq x_{v_0}$ and $x_{v_0} \leq x_{v_1}$. Let $H_5$ be a graph with $V(H_5) = V(G)$

and $E(H_5) = E(G) \setminus \{u_2u_1, v_3v_0\} \cup \{u_2v_0, v_3v_1\}$. By Lemma 2.1, $\rho(G) \leq \rho(H_5)$. We build a new graph $H_6$ by adding $v_1u_1$. Then $H_5$ is a proper subgraph of $H_6$ and $\rho(H_5) < \rho(H_6)$. We can use Lemma 2.7 to find a graph in $\mathcal{C}_n^k$ with only one common vertex among cycles. This is a contradiction to the choice of $G$. So, this lemma is true. □

## 3. Main results

In this section, we determine the cacti maximizing $A_\alpha$-spectral radius subject to fixed pendant vertices. In addition, we find the graph with the $A_\alpha$-spectral radius among all the cacti with $n$ vertices, and we also characterize the $n$-vertex cacti with a perfect matching having the largest $A_\alpha$-spectral radius.

Since $\mathcal{C}_n^k$ is the set of all cacti with $n > 0$ vertices and $k > 0$ pendant vertices, then let $C^e$ be a cactus graph in $\mathcal{C}_n^k$ such that $n-k-1$ is even and all cycles (if any) have length 3, that is, $C^e$ contains $\frac{n-k-1}{2}$ cycles $vv_1v_1'v, vv_2v_2'v, \cdots,$ $vv_{\frac{n-k-1}{2}}v'_{\frac{n-k-1}{2}}v$ and $k$ pendant edges (if any) $vu_1, vu_2, \cdots, vu_k$. Similarly, let $C^o$ be a cactus graph in $\mathcal{C}_n^k$ such that $n-k-1$ is odd and all cycles (if any) have length 3, that is $C^o$ contains $\frac{n-k-2}{2}$ cycles $vv_1v_1'v, vv_2v_2'v, \cdots, vv_{\frac{n-k-2}{2}}v'_{\frac{n-k-2}{2}}v$, $k-1$ pendant edges (if any) $vu_1, vu_2, \cdots, vu_{k-1}$ and 1 pendant path $vu_k'u_k$.

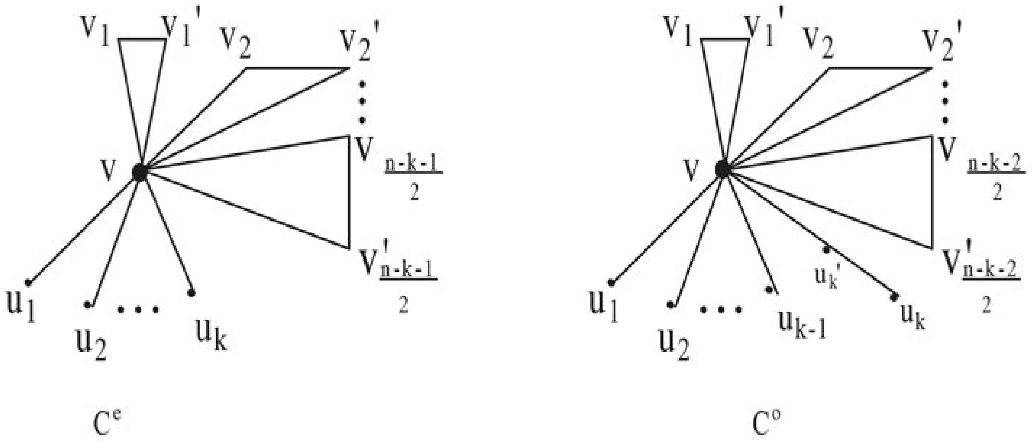



**Figure 1.** $C^e$: $n-k-1$ is even, contains $\frac{n-k-1}{2}$ cycles and $k$ pendant edges (if any); $C^o$: $n-k-1$ is odd, contains $\frac{n-k-2}{2}$ cycles,$k-1$ pendant edges (if any) and 1 pendant path.

**Theorem 3.1.** (i) *If $n-k$ is odd and $G$ is a graph with the maximum $A_\alpha-$spectral radius in $\mathcal{C}_n^k$, then $G \cong C^e$;*

(ii) *If $n-k$ is even and $G$ is a graph with the maximum $A_\alpha-$spectral radius in $\mathcal{C}_n^k$, then $G \cong C^o$.*

***Proof.*** Choose a cactus graph $G \in \mathcal{C}_n^k$ such that $\rho(G)$ is maximal. Assume $V(G) = \{v_0, v_2, \cdots, v_{n-1}\}$. By Lemma 2.4, we have all pendant paths share a common vertex. By Lemma 2.5 implies that the length of any pendant path is at most 2 and there is at most one pendant path of length 2. By Lemma 2.6 yields that there does not exist an internal path such that it is built by cut edges. By Lemma 2.8 all cycles share a common vertex. By Lemma 2.8 we have the length of any cycle is at most 4, and there is at most one cycle of length 4. In order to find the main results, we need the following two claims.

***Claim 1.*** The pendant paths and cycles share a common vertex.

*Proof.* Suppose that all cycles share a vertex $v$ and all pendant paths share a vertex $u$, $u, v \in \{v_0, v_1, \cdots, v_{n-1}\}$. Clearly, $u$ and $v$ is in a same cycle $C'$. Let $N'(u) = N(u) \setminus V(C')$

and $N'(v) = N(v) \setminus V(C')$. If $x_u \geq x_v$, then set a new graph $G_1$ with vertex set $V(G_1) = V(G) \setminus \{wv, \in w \in N'(v)\} \cup \{wu, \in w \in N'(v)\}$; Otherwise, if $x_u \leq x_v$, let a new graph $G_2$ with vertex set $V(G_2) = V(G) \setminus \{wu, \in w \in N'(u)\} \cup \{wv, \in w \in N'(u)\}$. By Lemma 2.1, we have $\rho(G) \leq \rho(G_1)$ or $\rho(G) \leq \rho(G_2)$. A contradiction yields this claim.

***Claim 2.*** If there is a pendant path $P$ with the length at most 2, then there is no cycle of length 4.

*Proof.* Let $v_0v_1v_2v_3v_0$ be a cycle of length 4 and $P$ is a pendant path in $G$. By lemma 2.5 we know the length of $P$ is 1 or 2. Next we prove $x_{v_0} \geq max\{x_{v_1}, x_{v_2}, x_{v_3}\}$. Assume $x_{v_1} > x_{v_0}$. Let $S = N(v_0) \setminus \{v_1, v_3\}$, set a new graph $H$ with vertex set $V(G)$, $E(G) \setminus \{wv_0, w \in S\} \cup \{wv_1, w \in S\}$. Note that $H$ is a cactus graph with $k$ pendent vertices. By Lemma 2.1, we have $\rho(G) < \rho(H)$. It contradicts that $\rho(G)$ is maximal, thus, $x_{v_0} \geq x_{v_1}$. Similarity, we have $x_{v_0} \geq x_{v_2}$ and $x_{v_0} \geq x_{v_3}$. Thus, $x_{v_0} \geq max\{x_{v_1}, x_{v_2}, x_{v_3}\}$.

**Case 1.** $|P| = 2$. Assume $P = v_0v_4v_5$.

Let $H_1$ be a new graph with vertex set $V(G)$, $E(G)\setminus\{v_2v_3\} \cup \{v_0v_2\}$. Since $x_{v_0} \geq x_{v_3}$, then $\rho(G) \leq \rho(H_1)$ (by Lemma 2.1). Let $H_2$ be a new graph with vertex set $V(G)$, $E(H_1) + v_3v_4$. $H_1$ is proper subgraph of $H_2$ . By Lemma 2.3, we have $\rho(H_1) < \rho(H_2)$. Then, $\rho(G) < \rho(H_2)$. Note that $H_2$ is a cactus graph with $k$ pendent vertices.

**Case 2.** $|P| = 1$. Assume $P = v_0v_6$.

**Subcase 2.1.** $x_{v_2} \leq x_{v_6}$.

Let $H_3$ be a new graph with vertex set $V(G)$, $E(G)\setminus\{v_2v_3\} \cup \{v_3v_6\}$. Note that $H_3$ is a cactus graph with $k$ pendent vertices. By Lemma 2.1, we have $\rho(G) \leq \rho(H_3)$.

**Subcase 2.2.** $x_{v_2} > x_{v_6}$.

Let $H_4$ be a new graph with vertex set $V(G)$, $E(G)\setminus\{v_2v_3, v_0v_6\} \cup \{v_0v_2, v_3v_6\}$. Note that $H_4$ is a cactus graph with $k$ pendent vertices. Since $x_{v_0} \geq x_{v_3}$ and $x_{v_2} > x_{v_6}$, then $(x_{v_2} - x_{v_6})(x_{v_0} - x_{v_3}) \geq 0$. By Lemma 2.2, we have $\rho(G) \leq \rho(H_4)$. Note that $H_4$ is a cactus graph with $k$ pendent vertices. It is a contradiction and this claim is proved.

Therefore, if $n-k$ is odd, then $\rho(G) \leq \rho(C^e)$; if $n-k$ is even, then $\rho(G) \leq \rho(C^o)$. So, this theorem is proved. □

**Lemma 3.2** ([9]). Given a partition $\{1, 2, \cdots, n\} = \Delta_1 \cup \Delta_2 \cup \cdots \cup \Delta_m$ with $|\Delta_i| = n_i > 0$, $A$ be any matrix partitioned into blocks $A_{ij}$, where $A_{ij}$ is an $n_i \times n_j$ block. Suppose that the block $A_{ij}$ has constant row sums $b_{ij}$ , and let $B = (b_{ij})$. Then the spectrum of B is contained in the spectrum of A (taking into account the multiplicities of the eigenvalues).

Next we provide all eigenvalues of $C^e$ and $C^o$ in the proposition.

**Proposition 3.3.** *Let* $\alpha \in [0, 1)$. *The following statements hold.* $(i)$ *The maximum eigenvalues of* $A_\alpha(C^e)$ *satisfy the equation:* $f(\rho) = (\alpha-\rho)^3 + (n\alpha - 2\alpha + 1)(\alpha-\rho)^2 + [(1-n)\alpha^2 + (3n-4)\alpha + 1 - n](\alpha-\rho) - k(1-\alpha)^2 = 0$. $(ii)$ *The maximum eigenvalues of* $A_\alpha(C^o)$ *satisfy the equation:* $g(\rho) = (n\alpha - 2\alpha - \rho)(\alpha-\rho)(\alpha-\rho+1)(\rho^2 - 3\alpha\rho + \alpha^2 + 2\alpha - 1) - (k-1)(1-\alpha)^2(\alpha-\rho+1)(\rho^2 - 3\alpha\rho + \alpha^2 + 2\alpha - 1) - (n-k-2)(1-\alpha)^2(\alpha-\rho)(\rho^2 - 3\alpha\rho + \alpha^2 + 2\alpha - 1) - (1-\alpha)^2(\alpha-\rho)^2(\alpha-\rho+1) = 0$.

***Proof.*** Since the matrix $A_\alpha = \alpha D + (1-\alpha)A$, where $D$ has on the diagonal the vector $(n-1, 2, 1)$ and $A$ consists of the following three row-vectors, in the order: $(0, n-k-1, k)$; $(1, 1, 0)$; $(1, 0, 0)$. By Lemma 3.2, thus, the eigenvector $x$ of $\rho(A_\alpha(C^e))$ ( $C^e$, see Figure 1)is a constant value $\beta_2$ on the vertex set $\{v_1, v_1', v_2, v_2', \cdots, v_{\frac{n-k-1}{2}}, v'_{\frac{n-k-1}{2}}\}$, and constant value $\beta_3$ on the vertex set $\{u_1, u_2, \cdots, u_k\}$. Defining $x(v) =: \beta_1$, $\rho(C^e) =: \rho$, also by (1), we get $(\rho - (n-1)\alpha)\beta_1 = (1-\alpha)((n-k-1)\beta_2 + k\beta_3)$, $(\rho - 2\alpha)\beta_2 = (1-\alpha)(\beta_1 + \beta_2)$ and $(\rho - \alpha)\beta_3 = (1-\alpha)\beta_1)$.

Then we get:

$$f(\rho) = (\alpha-\rho)^3 + (n\alpha-2\alpha+1)(\alpha-\rho)^2 + [(1-n)\alpha^2 + (3n-4)\alpha + 1 - n](\alpha-\rho) - k(1-\alpha)^2 = 0.$$

Next we consider $A_\alpha(C^o)$ ( $C^o$, see Figure 1), since the matrix $A_\alpha = \alpha D + (1-\alpha)A$, where $D$ has on the diagonal the vector $(n-2,2,1,2,1)$ and $A$ consists of the following five row-vectors, in the order: $(0,n-k-2,k-1,1,0)$; $(1,1,0,0,0)$; $(1,0,0,0,0)$; $(1,0,0,0,1)$ $(0,0,0,1,0)$. By Lemma 3.2, thus, the eigenvector $x$ of $\rho(A_\alpha(C^o))$ is a constant value $\beta_2$ on the vertex set $\{v_1, v'_1, \cdots, v_{\frac{n-k-2}{2}}, v'_{\frac{n-k-2}{2}}\}$, and constant value $\beta_3$ on the vertex set $\{u_1, u_2, \cdots, u_{k-1}\}$. Defining $x(v) =: \beta_1$, and $x(u'_k) =: \beta_4$, and $x(u_k) =: \beta_5$. $\rho(C^e) =: \rho$, also by (1), similarly as above the computation of $A_\alpha(C^e)$, we obtain:

$g(\rho) = (n\alpha-2\alpha-\rho)(\alpha-\rho)(\alpha-\rho+1)(\rho^2-3\alpha\rho+\alpha^2+2\alpha-1)-(k-1)(1-\alpha)^2(\alpha-\rho+1)(\rho^2-3\alpha\rho+\alpha^2+2\alpha-1)-(n-k-2)(1-\alpha)^2(\alpha-\rho)(\rho^2-3\alpha\rho+\alpha^2+2\alpha-1)-(1-\alpha)^2(\alpha-\rho)^2(\alpha-\rho+1) = 0.$

Thus, our proof is finished. □

Denote by $\mathcal{C}_n^*$ be the set of all cacti with $n$ vertices. Let $C_n^{*1}$ be a cactus graph in $\mathcal{C}_n^*$ such that $n$ is odd and $C_n^{*1}$ contains $\frac{n-1}{2}$ cycles of length 3 (if any). Let $C_n^{*2}$ be a cactus graph in $\mathcal{C}_n^*$ such that $n$ is even and $C_n^{*2}$ contains $\frac{n-2}{2}$ cycles of length 3 (if any) and one pendant edge.

**Theorem 3.4.** (i) *If $n$ is odd and $G$ is a graph with the maximum $A_\alpha-$spectral radius in $\mathcal{C}_n^*$, then $G \cong C_n^{*1}$;*
(ii) *If $n$ is even and $G$ is a graph with the maximum $A_\alpha-$spectral radius in $\mathcal{C}_n^*$, then $G \cong C_n^{*2}$.*

***Proof.*** By the proof of Theorem 3.1, we have the sharp upper bounds of $A_\alpha-$spectral radius attain at $C^e$ and $C^o$. We can set up a new graph by connecting any two pendant vertices and the original graph is the proper subgraph of this new graph. By Lemma 2.2, we have $\rho(G)$ is increasing by this operation. Therefore, $\rho(G) \le \rho(C^{*1})$ if $n$ is odd, and $\rho(G) \le \rho(C^{*2})$ if $n$ is even. Since $C^{*1}$ is the cactus graph $C^e$ when $k=0$, and $C^{*2}$ is the cactus graph $C^o$ when $k=1$. Thus, this theorem is proved.

By Proposition 3.3, and letting $k = 0, 1$, we can also obtain their corresponding eigenvalues. □

Based on the above outcomes, we can determine the sharp upper bound for the $A_\alpha$-spectral radius of cacti with a perfect matching. Let $\mathcal{C}_{2k}^m$ be the set of all $2k$-vertex cacti with a perfect matching.

**Theorem 3.5.** *If $G$ is a graph with the maximum $A_\alpha-$spectral radius in $\mathcal{C}_{2k}^m$, then $G \cong C_{2k}^{*2}$.*

**Acknowledgment.** The work was partially supported by the National Natural Science Foundation of China under Grants 11771172 and 11571134. We would like to thank for Ting Huang's suggestions and improvements in our paper.

## References

[1] A. Berman and R.J. Plemmons, *Nonnegative Matrices in the Mathematical Sciences*, SIAM: Philadelphia, PA, USA, 1994.
[2] B. Bollobás, *Modern Graph Theory*, Springer: New York, NY, USA, 1998.
[3] B. Borovićanin and M. Petrović, *On the index of cactuses with n vertices*, Publ. Inst. Math **79** (93), 13-18, 2006.
[4] Y. Chen, *Properties of spectra of graphs and line graphs*, Appl. Math. J. Chinese Univ. Ser. B **17** (3), 371-376, 2002.
[5] M. Chen and B. Zhou, *On the Signless Laplacian Spectral Radius of Cacti*, Croat. Chem. Acta **89** (4), 493-498, 2016.
[6] L. Collatz and U. Sinogowitz, *Spektrcn endlicher Graten*, Abh. Math. Scm. Univ. Hamburg **21**, 63-77, 1957.

[7] L. Cui, Y.-Z. Fan, *The signless laplacian spectral radius of graphs with given number of cut vertices*, Discuss. Math. Graph Theory **30** (1), 85-93, 2010.

[8] D. Cvetković, P. Rowlinson and SK. Simić, *Signless Laplacians of finite graphs*, Linear Algebra Appl. **423** (3), 155-171, 2007.

[9] D. Cvetkovic, P. Rowlinson, S. Simic, *An Introduction to the Theory of Graph Spectra*, Cambridge University Press, 2009.

[10] L. Feng , Q. Li and X.-D. Zhang, *Minimizing the Laplacian spectral radius of trees with given matching number*, Linear Multilinear Algebra **55**, 199-207, 2007.

[11] J. Huang and S. Li, *On the Spectral Characterizations of Graphs*, Discuss. Math. Graph Theory **37**, 729-744, 2017.

[12] S. Li and M. Zhang, *On the signless Laplacian index of cacti with a given number of pendant vertices*, Linear Algebra Appl. **436**, 4400-4411, 2012.

[13] H. Lin and B. Zhou, *Graphs with at most one signless Laplacian eigenvalue exceeding three*, Linear Multilinear Algebra **63** (3), 377-383, 2015.

[14] L. Lovász and J. Pelikán, *On the eigenvalues of trees*, Period. Math. Hungar **3**, 175-182, 1973.

[15] V. Nikiforov, *Merging the A- and Q-spectral theories*, Appl. Anal. Discrete Math. **11**, 81-107, 2017.

[16] V. Nikiforov, G. Pastén, O. Rojo and R.L. Soto, *On the $A_\alpha$-spectra of trees*, Linear Algebra Appl. **520** (3), 286-305, 2017.

[17] Y. Shen, L. You, M. Zhang and S. Li, *On a conjecture for the signless Laplacian spectral radius of cacti with given matching number*, Linear Multilinear Algebra **65** (4), 457-474, 2017.

[18] J. Wu, H. Deng and Q. Jiang, *On the spectral radius of cacti with k-pendant vertices*, Linear Multilinear Algebra **58**, 391-398, 2010.

[19] T. Wu and H. Zhang, *Per-spectral characterizations of some bipartite graphs*, Discuss. Math. Graph Theory **37**, 935-951, 2017.

[20] R. Xing and B. Zhou, *On the least eigenvalue of cacti with pendant vertices*, Linear Algebra Appl. **438**, 2256-2273, 2013.

[21] J. Xue, H. Lin, S. Liu and J. Shu, *On the $A_\alpha$-spectral radius of a graph*, Linear Algebra Appl. **550**, 105-120, 2018.

[22] Y. Yan, C.Wang and S.Wang, *The $A_\alpha$-spectral radii of trees with specified maximum degree*, submitted.

[23] A. Yu, M. Lu and F. Tian, *On the spectral radius of graphs*, Linear Algebra Appl. **387**, 41-49, 2004.

[24] Bo. Zhou, *Signless Laplacian spectral radius and Hamiltonicity*, Linear Algebra Appl **423** (3), 566-570, 2010.

Hacettepe Journal of
Mathematics & Statistics

Hacet. J. Math. Stat.
Volume 50 (1) (2021), 41 – 45
DOI : 10.15672/hujms.622655

Research Article

# The nil-clean $2 \times 2$ integral units

Grigore Călugăreanu

*Babeş-Bolyai University, 1 Kogălniceanu street, Cluj-Napoca, Romania*

## Abstract

We prove that all trace 1, $2\times 2$ invertible matrices over $\mathbb{Z}$ are nil-clean and, up to similarity, that there are only two trace 1, $2 \times 2$ invertible matrices over $\mathbb{Z}$.

**Mathematics Subject Classification (2020).** 16U10, 16U60, 11E16

**Keywords.** nil-clean, clean, similarity, binary quadratic form, class number

## 1. Introduction

We first recall the following.

An element $a$ in a unital ring $R$ is *clean* (see [5]) if $a = e + u$ with an idempotent $e \in R$ and a unit $u \in R$, and, *nil-clean* (see [4]) if $a = e + t$ with an idempotent $e$ and a nilpotent $t$. It is *strongly* nil-clean if $et = te$. A nil-clean element is called *trivial* if $e \in \{0, 1\}$, the trivial idempotents. A unit $u$ is called *unipotent* if $u = 1 + t$, for some nilpotent $t$.

A ring is *clean* (or *nil-clean*) if so are all its elements. Via unipotent units, it is easy to see that nil-clean rings are clean.

Though all these notions are well-known for some time, very little is known about *which clean elements of a ring are nil-clean*. Actually, besides the unipotent units (indeed, *a unit is strongly nil-clean if and only if it is unipotent*), we do not know which units of a ring are nil-clean.

We can discard the *trivial nil-clean* elements. Indeed, if $e = 0$, then there is no unit which is nilpotent (unless $R = 0$), and if $e = 1$, $a = 1 + t$, are precisely the unipotent units. Over any *commutative domain*, such $2 \times 2$ matrices $M$, are easily characterized by $\det(M - I_2) = \mathrm{Tr}(M - I_2) = 0$.

In this note, using an adequate (but nontrivial) Number Theory machinery, we characterize the (nontrivial) nil-clean units in the matrix ring $\mathcal{M}_2(\mathbb{Z})$.

Notice that *non-trivial* nil-clean $2\times 2$ matrices over any commutative domain have trace 1.

As our main result, conversely, we show that *trace* $1$, $2 \times 2$ *units over* $\mathbb{Z}$ *are nil-clean*, that is, *a* $2 \times 2$ *unit over* $\mathbb{Z}$ *is non-trivial nil-clean if and only if it has trace 1.*

Up to similarity, we also prove that all trace 1, $2 \times 2$ units are similar to $\begin{bmatrix} 0 & 1 \\ -1 & 1 \end{bmatrix}$ or to $\begin{bmatrix} 2 & 1 \\ -1 & -1 \end{bmatrix}$.

Email address: calu@math.ubbcluj.ro
Received: 20.09.2019; Accepted: 22.04.2020

## 2. Binary quadratic forms preliminaries

The proof of our main result requires some preparation.

First consider a particular Diophantine equation, namely

$$(x+y)^2 + xy = m \qquad (*)$$

where $m$ is a positive integer.

**Lemma 2.1.** *For any divisor $m$ of a positive integer $A(A+1)-1$, $A>1$, the equation (*) is solvable.*

***Proof.*** From the general theory of quadratic binary forms, we know that the integer $m$ is represented by a binary quadratic form of discriminant $d$ only if the congruence $u^2 \equiv d(\bmod 4k)$ is solvable, where $k$ is the square-free part of $m$ (see [2], Theorem 7, p. 145). In our case, i.e. for the form $G(x,y) = (x+y)^2 + xy$, $d = 5$ and the class number of $\mathbb{Q}[\sqrt{5}]$ is 1, hence the above condition becomes necessary and sufficient. The solvability of the congruence $u^2 \equiv 5(\bmod 4k)$ is equivalent to the property that all prime factors of form $5s+2$ or $5s+3$ from the factorization of $m$ have even exponent.

Since we have to solve this equation for a divisor $m$ of $A(A+1)-1$, this reduces to show that if $m$ divides $A(A+1)-1$, then $m$ has this property. But this holds because if a prime $p$ divides $A(A+1)-1$, then it also divides $(2A+1)^2 - 5 = 4[A(A+1)-1]$, so 5 must be a quadratic residue modulo $p$.

On the other hand, denoting by $\left(\dfrac{a}{p}\right)$ the Legendre symbol, according to the Gauss reciprocity law (see [1], Theorem 9.1.3), $\left(\dfrac{5}{p}\right)\left(\dfrac{p}{5}\right) = (-1)^{\frac{p-1}{2}\cdot\frac{5-1}{2}} = 1$. Because $\left(\dfrac{5}{p}\right) = 1$, it follows $\left(\dfrac{p}{5}\right) = 1$ and so $p$ is a quadratic residue modulo 5, i.e., $p$ is congruent to $0, 1$ or 4 modulo 5, as desired. □

Next, we consider another particular Diophantine equation, namely

$$(x-y)^2 + xy = m \qquad (**)$$

where $m$ is a positive integer.

**Lemma 2.2.** *For any divisor $m$ of a positive integer $A(A+1)+1$, $A>1$, the equation (**) is solvable.*

***Proof.*** The proof is similar to the proof of the previous lemma. Just notice that now the discriminant is $-3$ and the corresponding class number is also 1. Moreover, if a prime $p$ divides $A(A+1)+1$, then it also divides $(2A+1)^2 + 3 = 4[A(A+1)+1]$, $-3$ must be a quadratic residue modulo $p$ and so on. □

Secondly, we need the following

**Proposition 2.3.** *Suppose $A(A+1)+BC = 1$ for integers $A, B, -C > 1$. We can always chose solutions $(b,d)$ and $(a,c)$ of the equation (*) with $m = B$ and $m = -C$, respectively, such that $ad - bc = 1$.*

***Proof.*** Again we use the theory of binary quadratic forms.

Consider the quadratic form $F(x,y) = Bx^2 + (2A+1)xy - Cy^2$.

Its discriminant is equal to $(2A+1)^2 + 4BC = 5$ (by our hypothesis). Using the reduction theory of quadratic forms, since the class number of $\mathbb{Q}[\sqrt{5}]$ is 1, it is well-known that (see [3]) all integer quadratic forms with discriminant 5 are $SL(2,\mathbb{Z})$-equivalent to

$G(x,y)=(x+y)^2+xy$, which has also discriminant 5. The equivalence means that there exist integers $a,b,c,d$ with $ad-bc=1$ such that $G(ax+by,cx+dy)=F(x,y)$.

If we set $x=1,y=0$ we get $G(a,c)=B$ and if we set $x=0,y=1$ we get $G(b,d)=-C$ and we are done. □

**Proposition 2.4.** *Suppose $A(A+1)+BC=-1$ for integers $A,B,-C>1$. We can always chose solutions $(b,d)$ and $(a,c)$ of the equation (**) with $m=B$ and $m=-C$, respectively, such that $ad-bc=1$.*

***Proof.*** We consider again the quadratic form $F(x,y)=Bx^2+(2A+1)xy-Cy^2$. Its discriminant is $(2A+1)^2+4BC=-3$ and so is the discriminant of $G(x,y)=(x-y)^2+xy$. Since the corresponding class number is 1, these are $SL(2,\mathbb{Z})$-equivalent, there exist integers $a,b,c,d$ with $ad-bc=1$ such that $G(ax+by,cx+dy)=F(x,y)$ and we complete the proof as for the previous proposition. □

## 3. The main result

By $E_{11}$ we denote the matrix with all entries zero, excepting the NW corner, which is 1. Recall that *over any principal ideal domain, every non-trivial $2\times 2$ idempotent matrix is similar to $E_{11}$*. The result holds also in a more general setting (see [6]), but this hypothesis suffices for our proof below.

We first give a characterization, up to similarity, of the non-trivial nil-clean units in $\mathcal{M}_2(\mathbb{Z})$.

**Proposition 3.1.** *An integral $2\times 2$ matrix $U$ is a non-trivial nil-clean unit iff it is similar to one of the following two matrices: $V_1=\begin{bmatrix}0&1\\-1&1\end{bmatrix}$, $V_{-1}=\begin{bmatrix}2&1\\-1&-1\end{bmatrix}$. More precisely, if $\det U=1$, it is similar to $V_1$ and if $\det U=-1$, it is similar to $V_{-1}$.*

***Proof.*** Since nil-clean and unit are invariant (properties) to conjugation, up to similarity, owing to the previous paragraph, we can suppose the idempotent in the nil-clean decomposition being $E_{11}$. Nilpotent matrices having zero trace and zero determinant, we deal with (nil-clean) matrices $M=\begin{bmatrix}a+1&b\\c&-a\end{bmatrix}$ such that $a^2+bc=0$. Since $\det M=-(a+1)a-bc=-a\in\{\pm 1\}$ we distinguish two cases.

**Case 1**. If $a=-1$ then $bc=-1$ which give two matrices: $V_1=E_{11}+\begin{bmatrix}-1&1\\-1&1\end{bmatrix}$ and transpose (which is similar to $V_1$: just conjugate by $\begin{bmatrix}1&0\\0&-1\end{bmatrix}$).

**Case 2**. If $a=1$ then $bc=-1$ which give two matrices: $V_{-1}=E_{11}+\begin{bmatrix}1&1\\-1&-1\end{bmatrix}$ and transpose (which is similar to $V_{-1}$: the same conjugation). □

**Example**. $A=\begin{bmatrix}8&5\\-11&-7\end{bmatrix}=\begin{bmatrix}9&6\\-12&-8\end{bmatrix}+\begin{bmatrix}-1&-1\\1&1\end{bmatrix}$. Here $U=\begin{bmatrix}3&2\\-4&-3\end{bmatrix}$ and $U^{-1}AU=\begin{bmatrix}2&1\\1&1\end{bmatrix}U=V_{-1}$, as stated.

Just taking the conjugates of these two matrices we can find the form of all the non-trivial nil-clean units in $\mathcal{M}_2(\mathbb{Z})$. This is

$$\begin{bmatrix}(a+c)(b+d)+ad & (b+d)^2+bd\\ -(a+c)^2-ac & -(a+c)(b+d)-bc\end{bmatrix}$$

for integers $a,b,c,d$ with $ad-bc=1$.

**Theorem 3.2.** *Trace 1, $2 \times 2$ units over $\mathbb{Z}$ are nil-clean.*

***Proof.*** In the sequel $M = \begin{bmatrix} A+1 & B \\ C & -A \end{bmatrix}$ denotes a trace 1, $2 \times 2$ integral matrix.

We first discuss the $\det M = -1$ case (i.e. $A(A+1) + BC = 1$) and (owing to the form of the non-trivial nil-clean units deduced above) prove that there are integers $a, b, c, d$ with $ad - bc = 1$ such that

$$M = \begin{bmatrix} (a+c)(b+d) + ad & (b+d)^2 + bd \\ -(a+c)^2 - ac & -(a+c)(b+d) - bc \end{bmatrix}.$$

Finding the integers $a, b, c, d$ amounts to solve the system

(i) $A = (a+c)(b+d) + bc$
(ii) $B = (b+d)^2 + bd$
(iii) $C = -(a+c)^2 - ac$
(iv) $1 = ad - bc$, with integer unknowns $a, b, c, d$.

First notice that $A(A+1) - 1 > 0$ with only two (integer) exceptions: $A = -1$ and $A = 0$. The case $A = 0$ reduces to $A = -1$, by conjugation with $\begin{bmatrix} 0 & 1 \\ 1 & 0 \end{bmatrix}$ and the case $A = -1$ was already settled as Case 1, Proposition 3.1.

Hence we can assume $BC < 0$ and even $B > 0$, $C < 0$ (otherwise we conjugate with $\begin{bmatrix} 1 & 0 \\ 0 & -1 \end{bmatrix}$), together with $A \geq 1$ (the case $A \leq -2$ also reduces to $A \geq 1$, by conjugation with $\begin{bmatrix} 0 & 1 \\ 1 & 0 \end{bmatrix}$).

Secondly observe that (ii) and (iii) are equations of type $(x+y)^2 + xy = m$, that is (*).

According to Proposition 2.3, the equations (ii), (iii) and (iv) have an integer solution.

Finally, we show that *any* solution of (ii), (iii) and (iv) (denoted again by $a, b, c, d$) also verifies (i) and we are done.

Indeed, $-BC = [(b+d)^2+bd][(a+c)^2+ac] = (b+d)^2(a+c)^2+ac(b+d)^2+bd(a+c)^2+abcd$ and so we have to check whether the degree 2 equation $A(A+1) = 1 + (b+d)^2(a+c)^2 + ac(b+d)^2 + bd(a+c)^2 + abcd$ has $A = (a+c)(b+d) + bc$ as one root, i.e.

$$(b+d)^2(a+c)^2+bc(bc+1)+(2bc+1)(a+c)(b+d) = 1+(b+d)^2(a+c)^2+ac(b+d)^2+bd(a+c)^2+abcd.$$

Equivalently $bc(bc+1-ad)+(2bc+1)(ab+ad+bc+cd) = 1+ab^2c+acd^2+a^2bd+bc^2d+4abcd$ or else $(bc+1-ad)(ab+cd+3bc-1) = 0$. This holds since $ad - bc = 1$.

Next, we settle the $\det M = 1$ case (i.e. $A(A+1) + BC = -1$) and prove that there are integers $a, b, c, d$ with $ad - bc = 1$ such that

$$M = \begin{bmatrix} (a-c)(b-d) + ad & (b-d)^2 + bd \\ -(a-c)^2 - ac & -(a-c)(b-d) - bc \end{bmatrix}.$$

Finding the integers $a, b, c, d$ amounts to solve the system

(i) $A = (a-c)(b-d) + bc$
(ii) $B = (b-d)^2 + bd$
(iii) $C = -(a-c)^2 - ac$
(iv) $1 = ad - bc$, with integer unknowns $a, b, c, d$.

Therefore now we deal with the equation (**). What remains for the proof is now deduced from Proposition 2.4 and a similar verification that any solution of (ii), (iii) and (iv) actually satisfies also (i). $\square$

In closing we mention that this result fails for higher dimensions of matrices. Here is a $3 \times 3$ example:

take $U = \begin{bmatrix} 1 & 0 & 0 \\ 0 & 1 & 2 \\ 0 & -1 & -1 \end{bmatrix}$ and $V = \begin{bmatrix} 1 & 0 & 1 \\ 1 & 0 & 0 \\ 0 & 1 & 0 \end{bmatrix}$, both with trace=determinant=1. Then $\mathrm{Tr}(U^2) = -1 \neq 1 = \mathrm{Tr}(V^2)$ and so the matrices $U$, $V$ have different characteristic polynomials. Consequently, $U$ and $V$ are not similar.

**Acknowledgment.** Thanks are due to D. Andrica for the proof of Lemma 2.1 and to F. Beukers for pointing out the use of the theory of binary quadratic forms in the proof of Proposition 2.3.

## References

[1] T. Andreescu and D. Andrica, *Number Theory. Structures, examples, and problems*, Birkhäuser Boston, Inc., Boston, MA, 2009.

[2] A.I. Borevich and I.R. Shafarevich, *Number Theory.* Pure and Applied Mathematics Vol. **20**, Academic Press, New York-London, 1966.

[3] J. Buchmann and U. Vollmer, *Binary Quadratic Forms.* Springer, Berlin 2007.

[4] A.J. Diesl, *Nil clean rings*, J. Algebra **383**, 197-211, 2013.

[5] W.K. Nicholson, *Lifting idempotents and exchange rings*, Trans. Amer. Math.Soc. **229**, 269-278, 1977.

[6] A. Steger, *Diagonalibility of Idempotent Matrices.* Pacific J. Math. **19** (3), 535-542, 1966.

Hacettepe Journal of Mathematics & Statistics

Hacet. J. Math. Stat.
Volume 50 (1) (2021), 46 – 62
DOI : 10.15672/hujms.550184

Research Article

# New Wilker-type and Huygens-type inequalities

Ling Zhu*[1], Branko Malešević[2]

[1] *Department of Mathematics, Zhejiang Gongshang University, Hangzhou City, Zhejiang Province, 310018, China*

[2] *School of Electrical Engineering, University of Belgrade, Bulevar kralja Aleksandra 73, 11000 Belgrade, Serbia*

## Abstract

In this paper, we first determine the relationships between the first Wilker's inequality, the second Wilker's inequality, the first Huygens inequality, and the second Huygens inequality for circular functions and for hyperbolic functions, respectively. Then, we establish new Wilker-type inequalities and Huygens-type inequalities for two function pairs, $x/\sin^{-1}x$ and $x/\tan^{-1}x$, $x/\sinh^{-1}x$ and $x/\tanh^{-1}x$. Finally, we obtain some more general conclusions than the first work of this paper, which reveal the absolute monotonicity of four functions involving the four inequalities mentioned above.

**Mathematics Subject Classification (2020).** 26D15, 42A10

**Keywords.** Wilker-type inequalities, Huygens-type inequalities, circular functions, hyperbolic functions, inverse circular functions, inverse hyperbolic functions

## 1. Introduction

Let $0 < x < \pi/2$. Then

$$\sin x < x < \tan x, \tag{1.1}$$

which can be rewrited as

$$\frac{\sin x}{x} < 1 < \frac{\tan x}{x}, \tag{1.2}$$

or

$$\frac{x}{\tan x} < 1 < \frac{x}{\sin x}. \tag{1.3}$$

When the functions involved in (1.2) are taken into account in two forms of size relations, two famous inequalities called the first Wilker's inequality (see [7, 21, 29, 30, 37, 39]), the first Huygens inequality (see [3–5, 8, 9, 11, 28, 32, 43]), it comes to the conclusions (1.4) and (1.5). The comparison of these two inequalities (see [6]) is shown as follows in (1.6).

$$\frac{1}{2}\left(\left(\frac{\sin x}{x}\right)^2 + \frac{\tan x}{x}\right) > 1, \tag{1.4}$$

$$\frac{1}{3}\left(\frac{2\sin x}{x} + \frac{\tan x}{x}\right) > 1, \tag{1.5}$$

*Corresponding Author.
Email addresses: zhuling0571@163.com (L. Zhu), branko.malesevic@etf.rs (B. Malešević)
Received: 06.04.2019; Accepted: 26.04.2020

$$\frac{1}{2}\left(\left(\frac{\sin x}{x}\right)^2+\frac{\tan x}{x}\right) > \frac{1}{3}\left(\frac{2\sin x}{x}+\frac{\tan x}{x}\right)>1. \tag{1.6}$$

Similar to (1.4) – (1.6), there are some conclusions (1.7) and (1.8) about the second Wilker's inequality (see [23, 24, 32, 43]), the second Huygens inequality (see [23, 24]), and the comparison of the two inequalities as follows.

$$\frac{1}{2}\left(\left(\frac{x}{\sin x}\right)^2+\frac{x}{\tan x}\right) > 1, \tag{1.7}$$

$$\frac{1}{3}\left(\frac{2x}{\sin x}+\frac{x}{\tan x}\right) > 1, \tag{1.8}$$

$$\frac{1}{2}\left(\left(\frac{x}{\sin x}\right)^2+\frac{x}{\tan x}\right) > \frac{1}{3}\left(\frac{2x}{\sin x}+\frac{x}{\tan x}\right)>1. \tag{1.9}$$

The last inequality chain is true due to

$$\begin{aligned}
&\frac{1}{2}\left(\left(\frac{x}{\sin x}\right)^2+\frac{x}{\tan x}\right)-\frac{1}{3}\left(\frac{2x}{\sin x}+\frac{x}{\tan x}\right)\\
=\ &\frac{1}{6}\left(2\left(\frac{x}{\sin x}\right)^2+\left(\frac{x}{\sin x}\right)^2+\frac{x}{\tan x}-4\frac{x}{\sin x}\right)\\
>\ &\frac{1}{6}\left(2\left(\frac{x}{\sin x}\right)^2+2-4\frac{x}{\sin x}\right)=\frac{1}{3}\left(1-\frac{x}{\sin x}\right)^2>0
\end{aligned}$$

and (1.8). At the same time, we find that the inequality (1.7) plays a key role in the above derivation. Furthermore, the relationships between the first and second Wilker's inequality (see [3, 42]), the first and second Huygens inequality (see [23, 24]) are given below.

$$\frac{1}{2}\left(\left(\frac{\sin x}{x}\right)^2+\frac{\tan x}{x}\right) > \frac{1}{2}\left(\left(\frac{x}{\sin x}\right)^2+\frac{x}{\tan x}\right)>1, \tag{1.10}$$

$$\frac{1}{3}\left(\frac{2\sin x}{x}+\frac{\tan x}{x}\right) > \frac{1}{3}\left(\frac{2x}{\sin x}+\frac{x}{\tan x}\right)>1. \tag{1.11}$$

The same case occurs in the hyperbolic functions (see [23, 24, 36, 39, 41–44]).

Now let's turn to the discussion of similar inequalities for inverse circular functions. Let $0<x<1$. Then

$$\tan^{-1}x<x<\sin^{-1}x, \tag{1.12}$$

which can be rewritten as

$$\frac{\tan^{-1}x}{x}<1<\frac{\sin^{-1}x}{x}, \tag{1.13}$$

or

$$\frac{x}{\sin^{-1}x}<1<\frac{x}{\tan^{-1}x}. \tag{1.14}$$

Chen and Cheung [6] obtained an important conclusion about the inverse circular functions as follows.

$$\left(\frac{x}{\sin^{-1}x}\right)^2+\frac{x}{\tan^{-1}x}<2,\ 0<x<1. \tag{1.15}$$

Then, they used the arithmetic–geometric–harmonic mean inequality to prove the following inequality chain for $x\in(0,1)$:

$$\frac{1}{2}\left(\left(\frac{\sin^{-1}x}{x}\right)^2+\frac{\tan^{-1}x}{x}\right)>\frac{1}{3}\left(\frac{2\sin^{-1}x}{x}+\frac{\tan^{-1}x}{x}\right) \tag{1.16}$$

$$> \left(\left(\frac{\sin^{-1} x}{x}\right)^2 \frac{\tan^{-1} x}{x}\right)^{1/3}$$

$$> \left(\frac{2}{1/\left(\left(\sin^{-1} x\right)/x\right)^2 + 1/\left(\left(\tan^{-1} x\right)/x\right)}\right)^{1/3} > 1.$$

They established the inverse hyperbolic version of above results for $x \in (0,1)$:

$$\left(\frac{x}{\sinh^{-1} x}\right)^2 + \frac{x}{\tanh^{-1} x} < 2, \tag{1.17}$$

and

$$\frac{1}{2}\left(\left(\frac{\sinh^{-1} x}{x}\right)^2 + \frac{\tanh^{-1} x}{x}\right) > \frac{1}{3}\left(\frac{2\sinh^{-1} x}{x} + \frac{\tanh^{-1} x}{x}\right) \tag{1.18}$$

$$> \left(\left(\frac{\sinh^{-1} x}{x}\right)^2 \frac{\tanh^{-1} x}{x}\right)^{1/3}$$

$$> \left(\frac{2}{1/\left(\left(\sinh^{-1} x\right)/x\right)^2 + 1/\left(\left(\tanh^{-1} x\right)/x\right)}\right)^{1/3} > 1.$$

The first task of this paper is to determine the relationship between the first Wilker's inequality, the second Wilker's inequality, the first Huygens inequality and the second Huygens inequality. The second one is to consider the results according to the form of the inequality (1.6) or (1.9) for two function pairs, $x/\sin^{-1} x$ and $x/\tan^{-1} x$, $x/\sinh^{-1} x$ and $x/\tanh^{-1} x$. Finally, we obtain some more general conclusions than the first work of this paper, which reveal the absolute monotonicity of four functions involving the above four inequalities.

## 2. Main results

This paper obtains the following main results.

**Theorem 2.1.** *Let $x \in (0, \pi/2)$. Then the inequality chain*

$$\begin{aligned} &\frac{1}{2}\left(\left(\frac{\sin x}{x}\right)^2 + \frac{\tan x}{x}\right) > \frac{1}{3}\left(\frac{2\sin x}{x} + \frac{\tan x}{x}\right) \\ &> \frac{1}{2}\left(\left(\frac{x}{\sin x}\right)^2 + \frac{x}{\tan x}\right) > \frac{1}{3}\left(\frac{2x}{\sin x} + \frac{x}{\tan x}\right) \\ &> 1. \end{aligned} \tag{2.1}$$

*holds.*

**Theorem 2.2.** *Let $x \in (0, \infty)$. Then the inequality chain*

$$\begin{aligned} &\frac{1}{2}\left(\left(\frac{\sinh x}{x}\right)^2 + \frac{\tanh x}{x}\right) > \frac{1}{3}\left(\frac{2\sinh x}{x} + \frac{\tanh x}{x}\right) \\ &> \frac{1}{2}\left(\left(\frac{x}{\sinh x}\right)^2 + \frac{x}{\tanh x}\right) > \frac{1}{3}\left(\frac{2x}{\sinh x} + \frac{x}{\tanh x}\right) \\ &> 1. \end{aligned} \tag{2.2}$$

*holds.*

**Theorem 2.3.** *Let $x \in (0,1)$. Then the inequality chain*

$$\frac{1}{2}\left(\left(\frac{x}{\sin^{-1} x}\right)^2 + \frac{x}{\tan^{-1} x}\right) < \frac{1}{3}\left(\frac{2x}{\sin^{-1} x} + \frac{x}{\tan^{-1} x}\right) < 1 \tag{2.3}$$

*holds.*

**Theorem 2.4.** *Let $x \in (0,1)$. Then the inequality chain*

$$\frac{1}{2}\left(\left(\frac{x}{\sinh^{-1} x}\right)^2 + \frac{x}{\tanh^{-1} x}\right) < \frac{1}{3}\left(\frac{2x}{\sinh^{-1} x} + \frac{x}{\tanh^{-1} x}\right) < 1 \tag{2.4}$$

*holds.*

Then we can obtain the following corollaries.

**Corollary 2.5.** *Let $x \in (0,1)$. Then*

$$\frac{1}{2}\left(\left(\frac{\sin^{-1} x}{x}\right)^2 + \frac{\tan^{-1} x}{x}\right) > \frac{1}{3}\left(\frac{2\sin^{-1} x}{x} + \frac{\tan^{-1} x}{x}\right) > 1 \tag{2.5}$$

$$> \frac{1}{3}\left(\frac{2x}{\sin^{-1} x} + \frac{x}{\tan^{-1} x}\right) > \frac{1}{2}\left(\left(\frac{x}{\sin^{-1} x}\right)^2 + \frac{x}{\tan^{-1} x}\right).$$

**Corollary 2.6.** *Let $x \in (0,1)$. Then*

$$\frac{1}{2}\left(\left(\frac{\sinh^{-1} x}{x}\right)^2 + \frac{\tanh^{-1} x}{x}\right) > \frac{1}{3}\left(\frac{2\sinh^{-1} x}{x} + \frac{\tanh^{-1} x}{x}\right) > 1 \tag{2.6}$$

$$> \frac{1}{3}\left(\frac{2x}{\sinh^{-1} x} + \frac{x}{\tanh^{-1} x}\right) > \frac{1}{2}\left(\left(\frac{x}{\sinh^{-1} x}\right)^2 + \frac{x}{\tanh^{-1} x}\right).$$

## 3. Proofs

### 3.1. Proof of Theorem 2.1

We shall complete the proof of Theorem 2.1 when proving second inequality of (2.1). Computing directly gives

$$\frac{1}{3}\left(\frac{2\sin x}{x} + \frac{\tan x}{x}\right) - \frac{1}{2}\left(\left(\frac{x}{\sin x}\right)^2 + \frac{x}{\tan x}\right) = \frac{\sin^2 x}{6x\cos x} F(x), \tag{3.1}$$

where

$$\begin{aligned} F(x) &= \frac{4\cos x \sin^3 x + 2\sin^3 x - 3x^3 \cos x - 3x^2 \cos^2 x \sin x}{\sin^4 x} \\ &= 4\cot x + 3x^2 \frac{1}{\sin x} - 3x^2 \frac{1}{\sin^3 x} + 2\frac{1}{\sin x} + x^3\left(-3\frac{\cos x}{\sin^4 x}\right). \end{aligned} \tag{3.2}$$

Since

$$\begin{aligned} \left(\frac{1}{\sin x}\right)' &= -\frac{\cos x}{\sin^2 x}, \\ \left(\frac{1}{\sin x}\right)'' &= \left(-\frac{\cos x}{\sin^2 x}\right)' = \frac{2}{\sin^3 x} - \frac{1}{\sin x}, \\ \left(\frac{1}{\sin^3 x}\right)' &= -3\frac{\cos x}{\sin^4 x}, \end{aligned}$$

from

$$\frac{1}{\sin x} = \frac{1}{x} + \sum_{n=1}^{\infty} \frac{2^{2n} - 2}{(2n)!}|B_{2n}|x^{2n-1}, \quad 0 < |x| < \pi, \quad \text{(see [12])} \tag{3.3}$$

we have

$$
\begin{aligned}
\frac{1}{\sin^3 x} &= \frac{1}{2}\left(\left(\frac{1}{\sin x}\right)'' + \frac{1}{\sin x}\right) \qquad (3.4)\\
&= \frac{1}{2}\left(\frac{2}{x^3} + \sum_{n=2}^{\infty} \frac{\left(2^{2n}-2\right)(2n-1)(2n-2)}{(2n)!}|B_{2n}|x^{2n-3}\right)\\
&\quad + \frac{1}{2}\left(\frac{1}{x} + \sum_{n=1}^{\infty} \frac{2^{2n}-2}{(2n)!}|B_{2n}|x^{2n-1}\right)\\
&= \frac{1}{x^3} + \sum_{n=2}^{\infty} \frac{\left(2^{2n}-2\right)(2n-1)(2n-2)}{2\,(2n)!}|B_{2n}|x^{2n-3}\\
&\quad + \frac{1}{2x} + \sum_{n=1}^{\infty} \frac{2^{2n}-2}{2\,(2n)!}|B_{2n}|x^{2n-1},
\end{aligned}
$$

and

$$
\begin{aligned}
-3\frac{\cos x}{\sin^4 x} &= \frac{1}{2}\left(-\frac{6}{x^4} + \sum_{n=2}^{\infty} \frac{\left(2^{2n}-2\right)(2n-1)(2n-2)(2n-3)}{(2n)!}|B_{2n}|x^{2n-4}\right)\\
&\quad + \frac{1}{2}\left(-\frac{1}{x^2} + \sum_{n=1}^{\infty} \frac{\left(2^{2n}-2\right)(2n-1)}{(2n)!}|B_{2n}|x^{2n-2}\right)\\
&= -\frac{3}{x^4} + \sum_{n=2}^{\infty} \frac{\left(2^{2n}-2\right)(2n-1)(2n-2)(2n-3)}{2\,(2n)!}|B_{2n}|x^{2n-4}\\
&\quad - \frac{1}{2x^2} + \sum_{n=1}^{\infty} \frac{\left(2^{2n}-2\right)(2n-1)}{2\,(2n)!}|B_{2n}|x^{2n-2}.
\end{aligned}
$$

We substitute the power series expansions of these functions into (3.2), and obtain

$$
\begin{aligned}
F(x) =& 4\cot x + 3x^2\frac{1}{\sin x} + 2\frac{1}{\sin x} - 3x^2\frac{1}{\sin^3 x} + x^3\left(-3\frac{\cos x}{\sin^4 x}\right)\\
=& 4\left(\frac{1}{x} - \sum_{n=1}^{\infty} \frac{2^{2n}}{(2n)!}|B_{2n}|x^{2n-1}\right) + 3x^2\left(\frac{1}{x} + \sum_{n=1}^{\infty} \frac{2^{2n}-2}{(2n)!}|B_{2n}|x^{2n-1}\right)\\
&+ 2\left(\frac{1}{x} + \sum_{n=1}^{\infty} \frac{2^{2n}-2}{(2n)!}|B_{2n}|x^{2n-1}\right)\\
&- 3x^2\left(\frac{1}{x^3} + \sum_{n=2}^{\infty} \frac{\left(2^{2n}-2\right)(2n-1)(2n-2)}{2\,(2n)!}|B_{2n}|x^{2n-3}\right)\\
&- 3x^2\left(\frac{1}{2x} + \sum_{n=1}^{\infty} \frac{2^{2n}-2}{2\,(2n)!}|B_{2n}|x^{2n-1}\right)\\
&+ x^3\left(-\frac{3}{x^4} + \sum_{n=2}^{\infty} \frac{\left(2^{2n}-2\right)(2n-1)(2n-2)(2n-3)}{2\,(2n)!}|B_{2n}|x^{2n-4}\right)\\
&+ x^3\left(-\frac{1}{2x^2} + \sum_{n=1}^{\infty} \frac{\left(2^{2n}-2\right)(2n-1)}{2\,(2n)!}|B_{2n}|x^{2n-2}\right)\\
=& \sum_{n=2}^{\infty} \frac{\left(2^{2n}-2\right)(2n-1)(2n-2)(2n-3)}{2\,(2n)!}|B_{2n}|x^{2n-1}\\
&- 3\sum_{n=2}^{\infty} \frac{\left(2^{2n}-2\right)(2n-1)(2n-2)}{2\,(2n)!}|B_{2n}|x^{2n-1}
\end{aligned}
$$

$$
\begin{aligned}
&+2\sum_{n=2}^{\infty}\frac{2^{2n}-2}{(2n)!}|B_{2n}|x^{2n-1}-4\sum_{n=2}^{\infty}\frac{2^{2n}}{(2n)!}|B_{2n}|x^{2n-1}\\
&+3\sum_{n=1}^{\infty}\frac{2^{2n}-2}{(2n)!}|B_{2n}|x^{2n+1}-3\sum_{n=1}^{\infty}\frac{2^{2n}-2}{2\,(2n)!}|B_{2n}|x^{2n+1}\\
&+\sum_{n=1}^{\infty}\frac{(2^{2n}-2)\,(2n-1)}{2\,(2n)!}|B_{2n}|x^{2n+1}\\
=&\sum_{n=2}^{\infty}\frac{(8n^3-36n^2+40n-16)\,2^{2n}-16n^3+72n^2+16-80n}{2\,(2n)!}|B_{2n}|x^{2n-1}\\
&+\sum_{n=1}^{\infty}\frac{2\,(n+1)\,(2^{2n}-2)}{2\,(2n)!}|B_{2n}|x^{2n+1}\\
=&\sum_{n=2}^{\infty}4\frac{2^{2n-1}\,(2n^3+10n-9n^2-4)-(2n^3+10n-9n^2-2)}{(2n)!}|B_{2n}|x^{2n-1}\\
&+\sum_{n=2}^{\infty}\frac{n\,(2^{2n-2}-2)}{(2n-2)!}|B_{2n-2}|x^{2n-1}\\
:=&\sum_{n=2}^{\infty}a_nx^{2n-1},
\end{aligned}
$$

where

$$
\begin{aligned}
a_n \;=\;& 4\frac{2^{2n-1}\,(2n^3-9n^2+10n-4)-(2n^3-9n^2+10n-2)}{(2n)!}|B_{2n}|\\
&+\frac{n\,(2^{2n-2}-2)}{(2n-2)!}|B_{2n-2}|
\end{aligned}
$$

for $n\geq 2$.

Since

$$|B_2|=\frac{1}{6},\ \ |B_4|=\frac{1}{30},\ \ |B_6|=\frac{1}{42},\ \ |B_8|=\frac{1}{30},$$

we first compute to obtain that

$$a_2=\frac{1}{6},\ a_3=\frac{17}{315},\ a_4=\frac{2509}{151\,200}.$$

Then using mathematical induction we can prove

$$2^{2n-1}\left(2n^3-9n^2+10n-4\right)-\left(2n^3-9n^2+10n-2\right)>0$$

or

$$2^{2n-1}>\frac{2n^3-9n^2+10n-2}{2n^3-9n^2+10n-4} \tag{3.5}$$

for $n\geq 4$. In fact, when $n=4$, the inequality (3.5) holds. Now, we assume that the (3.5) holds for $n=m$. Then, in order to complete the proof of (3.5) is also true for $n=m+1$ it suffices to show that

$$4\frac{2m^3-9m^2+10m-2}{2m^3-9m^2+10m-4}>\frac{2\,(m+1)^3-9\,(m+1)^2+10\,(m+1)-2}{2\,(m+1)^3-9\,(m+1)^2+10\,(m+1)-4},$$

which is true due to

$$
\begin{aligned}
&4\left(2m^3-9m^2+10m-2\right)\left(2\,(m+1)^3-9\,(m+1)^2+10\,(m+1)-4\right)\\
&-\left(2m^3-9m^2+10m-4\right)\left(2\,(m+1)^3-9\,(m+1)^2+10\,(m+1)-2\right)\\
=\;&12m^6-72m^5+129m^4-54m^3-3m^2-42m+12
\end{aligned}
$$

$$\begin{aligned} &= 12(m-4)^6 + 216(m-4)^5 + 1569(m-4)^4 + 5850(m-4)^3 \\ &\quad +11\,733(m-4)^2 + 11\,934(m-4) + 4788 \\ &> 0. \end{aligned}$$

So $a_n > 0$ for $n \geq 2$. This leads to $F(x) > 0$ for all $x \in (0, \pi/2)$. The proof of (2.1) is complete via (3.1).

## 3.2. Proof of Theorem 2.2

Similarly, if we can prove second inequality of (2.2), we then complete the proof of Theorem 2.2.

Computing gives

$$\frac{1}{3}\left(\frac{2\sinh x}{x} + \frac{\tanh x}{x}\right) - \frac{1}{2}\left(\left(\frac{x}{\sinh x}\right)^2 + \frac{x}{\tanh x}\right) := \frac{1}{24x\cosh x\sinh^3 x}G(x), \tag{3.6}$$

where

$$\begin{aligned} G(x) &= \cosh 4x - 3\cosh 3x - 4\cosh 2x + \cosh 5x + 2\cosh x \\ &\quad -\frac{3}{2}x^2\cosh 4x - 6x^3\sinh 2x + \frac{3}{2}x^2 + 3. \end{aligned} \tag{3.7}$$

Using the power series expansions of these hyperbolic functions, we have

$$\begin{aligned} G(x) &= \sum_{n=0}^{\infty}\frac{4^{2n}}{(2n)!}x^{2n} - 3\sum_{n=0}^{\infty}\frac{3^{2n}}{(2n)!}x^{2n} - 4\sum_{n=0}^{\infty}\frac{2^{2n}}{(2n)!}x^{2n} + \sum_{n=0}^{\infty}\frac{5^{2n}}{(2n)!}x^{2n} \\ &\quad + 2\sum_{n=0}^{\infty}\frac{1}{(2n)!}x^{2n} - \frac{3}{2}x^2\sum_{n=0}^{\infty}\frac{4^{2n}}{(2n)!}x^{2n} - 6x^3\sum_{n=0}^{\infty}\frac{2^{2n+1}}{(2n+1)!}x^{2n+1} \\ &\quad + \frac{3}{2}x^2 + 3 \\ &= \sum_{n=4}^{\infty}\frac{4^{2n} - 3\cdot 3^{2n} - 4\cdot 2^{2n} + 5^{2n} + 2}{(2n)!}x^{2n} \\ &\quad - \frac{3}{2}\sum_{n=3}^{\infty}\frac{4^{2n}}{(2n)!}x^{2n+2} - 6\sum_{n=2}^{\infty}\frac{2^{2n+1}}{(2n+1)!}x^{2n+4} \\ &= \sum_{n=4}^{\infty}\frac{4^{2n} - 3\cdot 3^{2n} - 4\cdot 2^{2n} + 5^{2n} + 2}{(2n)!}x^{2n} \\ &\quad - \sum_{n=4}^{\infty}\frac{3\cdot 4^{2n-2}}{2(2n-2)!}x^{2n} - \sum_{n=4}^{\infty}\frac{6\cdot 2^{2n-3}}{(2n-3)!}x^{2n} \\ &:= \sum_{n=4}^{\infty}\frac{1}{32(2n)!}b_n x^{2n}, \end{aligned}$$

where

$$\begin{aligned} b_n &= 32\cdot 5^{2n} - \left(6n^2 - 3n - 16\right)2^{4n+1} - 32\cdot 3^{2n+1} \\ &\quad - \left(6n^3 - 9n^2 + 3n + 4\right)2^{2n+5} + 64 \end{aligned}$$

for $n \geq 4$. We compute

$$\begin{aligned} c_n &:= b_{n+1} - 25b_n \\ &= 1536\cdot 3^{2n} + \left(108n^2 - 438n - 384\right)2^{4n} \\ &\quad + \left(126n^3 - 261n^2 + 63n + 84\right)2^{2n+5} - 1536 \end{aligned} \tag{3.8}$$

and obtain that

$$\begin{aligned} 108n^2-438n-384 &> 0, \\ \left(126n^3-261n^2+63n+84\right)2^{2n+5}-1536 &> 0 \end{aligned}$$

hold for all $n\geq 5$. So $c_n>0$ for $n\geq 5$. This together with $c_4=17\,940\,480>0$ gives that $c_n>0$ for $n\geq 4$. Then via (3.8) we have $b_{n+1}>25b_n$ holds for $n\geq 4$. This together with $b_4=860\,160>0$ gives that $b_n>0$ for $n\geq 4$. Then $G(x)>0$ for all $x\in(0,\pi/2)$. The proof of (2.2) is complete via (3.6).

### 3.3. Proof of Theorem 2.3

In order to prove Theorem 2.3 as simple as possible, we need a tool which offers a simple but efficient criterion to determine the sign of a kind of special power series, which we call as "*sign rule of a kind of special power series*".

**Lemma 3.1** ([34], [33]). *Let $\{a_k\}_{k=0}^{\infty}$ be a nonnegative real sequence with $a_m>0$ and $\sum_{k=m+1}^{\infty}a_k>0$ and let*

$$S(t)=-\sum_{k=0}^{m}a_kt^k+\sum_{k=m+1}^{\infty}a_kt^k$$

*be a convergent power series on the interval $(0,r)$ $(r>0)$. (i) If $S(r^-)\leq 0$ then $S(t)<0$ for all $t\in(0,r)$. (ii) If $S(r^-)>0$ then there is the unique $t_0\in(0,r)$ such that $S(t)<0$ for $t\in(0,t_0)$ and $S(t)>0$ for $t\in(t_0,r)$.*

(1) We first prove the left hand side of (2.3).

Let $\arcsin x=t$. Then the desired inequality is equivalent to

$$\frac{1}{2}\left(\frac{\sin t}{t}\right)^2-\frac{2}{3}\frac{\sin t}{t}+\frac{1}{6}\frac{\sin t}{\arctan(\sin t)}=\frac{\sin t}{6}\left(\frac{1}{\arctan(\sin t)}-\frac{4t-3\sin t}{t^2}\right)<0,$$

which is in turn equivalent to

$$H(t):=\frac{t^2}{4t-3\sin t}-\arctan(\sin t)<0$$

for $t\in(0,\pi/2)$. Differentiation yields

$$H'(t)=\frac{\sin^3 t}{(1+\sin^2 t)(4t-3\sin t)^2}h(t),$$

where

$$h(t)=4\frac{t^2}{\sin t}-6\frac{t}{\sin^2 t}-9\cot t-6t+4\frac{t^2}{\sin^3 t}+24t\frac{\cos t}{\sin^2 t}+3t^2\cot t-13t^2\frac{\cos t}{\sin^3 t}.$$

From

$$\cot x=\frac{1}{x}-\sum_{n=1}^{\infty}\frac{2^{2n}}{(2n)!}|B_{2n}|x^{2n-1},\quad 0<|x|<\pi,\quad \text{(see [10])} \tag{3.9}$$

and (3.3) we have

$$\begin{aligned} \frac{1}{\sin^2 t} &= -(\cot t)'=\frac{1}{t^2}+\sum_{n=1}^{\infty}\frac{(2n-1)\,2^{2n}}{(2n)!}|B_{2n}|t^{2n-2}, \qquad (3.10)\\ \frac{\cos t}{\sin^2 t} &= -\left(\frac{1}{\sin t}\right)'=\frac{1}{t^2}-\sum_{n=1}^{\infty}\frac{(2n-1)\left(2^{2n}-2\right)}{(2n)!}|B_{2n}|t^{2n-2}, \\ \frac{\cos t}{\sin^3 t} &= -\frac{1}{2}\left(\frac{1}{\sin^2 t}\right)'=\frac{1}{t^3}-\sum_{n=2}^{\infty}\frac{(2n-1)(n-1)\,2^{2n}}{(2n)!}|B_{2n}|t^{2n-3}. \end{aligned}$$

The above power series expansions and (3.4) give

$$\begin{aligned}
h\left(t\right) =&4t+4\sum_{n=1}^{\infty}\frac{2^{2n}-2}{\left(2n\right)!}|B_{2n}|t^{2n+1}-6t\left(\frac{1}{t^2}+\sum_{n=1}^{\infty}\frac{\left(2n-1\right)2^{2n}}{\left(2n\right)!}|B_{2n}|t^{2n-2}\right)\\
&-9\left(\frac{1}{t}-\sum_{n=1}^{\infty}\frac{2^{2n}}{\left(2n\right)!}|B_{2n}|t^{2n-1}\right)+4t^2\left(\frac{1}{2t}+\frac{1}{2}\sum_{n=1}^{\infty}\frac{2^{2n}-2}{\left(2n\right)!}|B_{2n}|t^{2n-1}\right)\\
&+4t^2\left(\frac{1}{t^3}+\frac{1}{2}\sum_{n=2}^{\infty}\frac{\left(2n-1\right)\left(2n-2\right)\left(2^{2n}-2\right)}{\left(2n\right)!}|B_{2n}|t^{2n-3}\right)\\
&+24t\left(\frac{1}{t^2}-\sum_{n=1}^{\infty}\frac{\left(2n-1\right)\left(2^{2n}-2\right)}{\left(2n\right)!}|B_{2n}|t^{2n-2}\right)\\
&+3t^2\left(\frac{1}{t}-\sum_{n=1}^{\infty}\frac{2^{2n}}{\left(2n\right)!}|B_{2n}|t^{2n-1}\right)\\
&-13t^2\left(\frac{1}{t^3}-\sum_{n=2}^{\infty}\frac{\left(2n-1\right)\left(n-1\right)2^{2n}}{\left(2n\right)!}|B_{2n}|t^{2n-3}\right)-6t\\
=&4t+4\sum_{n=1}^{\infty}\frac{2^{2n}-2}{\left(2n\right)!}|B_{2n}|t^{2n+1}-\frac{6}{t}-6\sum_{n=1}^{\infty}\frac{\left(2n-1\right)2^{2n}}{\left(2n\right)!}|B_{2n}|t^{2n-1}\\
&-\frac{9}{t}+9\sum_{n=1}^{\infty}\frac{2^{2n}}{\left(2n\right)!}|B_{2n}|t^{2n-1}+4t^2\left(\frac{1}{2t}+\frac{1}{t^3}\right)+2\sum_{n=1}^{\infty}\frac{2^{2n}-2}{\left(2n\right)!}|B_{2n}|t^{2n+1}\\
&+2\sum_{n=2}^{\infty}\frac{\left(2n-1\right)\left(2n-2\right)\left(2^{2n}-2\right)}{\left(2n\right)!}|B_{2n}|t^{2n-1}+\frac{24}{t}\\
&-24\sum_{n=1}^{\infty}\frac{\left(2n-1\right)\left(2^{2n}-2\right)}{\left(2n\right)!}|B_{2n}|t^{2n-1}+3t-3\sum_{n=1}^{\infty}\frac{2^{2n}}{\left(2n\right)!}|B_{2n}|t^{2n+1}\\
&-\frac{13}{t}+13\sum_{n=2}^{\infty}\frac{\left(2n-1\right)\left(n-1\right)2^{2n}}{\left(2n\right)!}|B_{2n}|t^{2n-1}-6t\\
=&\sum_{n=1}^{\infty}\frac{3\left(2^{2n}-4\right)}{\left(2n\right)!}\left|B_{2n}\right|t^{2n+1}\\
&+\sum_{n=2}^{\infty}\frac{\left(34n^2-111n+56\right)2^{2n}-8\left(2n-1\right)\left(n-7\right)}{\left(2n\right)!}\left|B_{2n}\right|t^{2n-1}\\
:=&\sum_{n=2}^{\infty}\frac{k_n\left|B_{2n-2}\right|+l_n\left|B_{2n}\right|}{\left(2n\right)!}t^{2n-1}:=\sum_{n=2}^{\infty}p_nt^{2n-1},
\end{aligned}$$

where

$$\begin{aligned}
k_n &= 24n\left(2n-1\right)\left(2^{2n-4}-1\right),\\
l_n &= \left(34n^2-111n+56\right)2^{2n}-8\left(2n-1\right)\left(n-7\right).
\end{aligned}$$

A simple computation shows that $p_2=-1/2$. We claim that $p_n>0$ for $n\geq 3$. In fact, $k_n>0$ for $n\geq 3$. Also, since $\left(34n^2-111n+56\right)>0$, so for $n\geq 3$,

$$l_n>\left(\left(34n^2-111n+56\right)8-8\left(2n-1\right)\left(n-7\right)\right)=8\left(32n\left(n-3\right)+49\right)>0.$$

These indicate that $p_n>0$ for $n\geq 3$.

On the other hand, we see that

$$h\left(\pi/2\right)=2\pi\left(\pi-3\right)>0.$$

By Lemma 3.1, there is a $t_0 \in (0, \pi/2)$ so that $h(t) < 0$ for $t \in (0, t_0)$ and $h(t) > 0$ for $t \in (t_0, \pi/2)$, which in turn implies that $H(t)$ is decreasing on $(0, t_0)$ and increasing on $(t_0, \pi/2)$. Consequently, we obtain

$$\begin{aligned} H(t) &< \lim_{t\to 0^+} H(t) = 0 \text{ for } t \in (0, t_0), \\ H(t) &< \lim_{t\to(\pi/2)^-} H(t) = -\frac{1}{4}\frac{\pi(\pi-3)}{2\pi-3} < 0 \text{ for } t \in (t_0, \pi/2), \end{aligned}$$

that is, $H(t) < 0$ for $t \in (0, \pi/2)$. This completes the proof of the left hand side of (2.3).

(2) We then prove the right hand side of (2.3).

The desired inequality is equivalent to

$$2\frac{x}{\sin^{-1} x} + \frac{x}{\tan^{-1} x} < 3.$$

Since

$$\begin{aligned} \frac{x}{\sin^{-1} x} &< \frac{2+\sqrt{1-x^2}}{3}, \quad \text{(see [15, 16, 21, 38])} \\ \frac{x}{\tan^{-1} x} &< 1 + \frac{1}{3}x^2, \quad \text{(see [6])} \end{aligned}$$

we have

$$2\frac{x}{\sin^{-1} x} + \frac{x}{\tan^{-1} x} < \frac{2\left(2+\sqrt{1-x^2}\right)}{3} + 1 + \frac{1}{3}x^2.$$

We can complete the proof of the right hand side of (2.3) as long as we can prove that

$$\frac{2\left(2+\sqrt{1-x^2}\right)}{3} + 1 + \frac{1}{3}x^2 < 3,$$

which is equivalent to $(1 - \sqrt{1-x^2})^2 > 0$.

### 3.4. Proof of Theorem 2.4

(1) We first prove the left hand side of (2.4).

Since

$$\begin{aligned} &\frac{1}{3}\left(\frac{2x}{\sinh^{-1} x} + \frac{x}{\tanh^{-1} x}\right) - \frac{1}{2}\left(\left(\frac{x}{\sinh^{-1} x}\right)^2 + \frac{x}{\tanh^{-1} x}\right) \\ = &\frac{x}{6}\left(\frac{4}{\sinh^{-1} x} - \frac{1}{\tanh^{-1} x} - 3\frac{x}{\left(\sinh^{-1} x\right)^2}\right), \end{aligned}$$

the desired inequality is equivalent to

$$\tanh^{-1} x > \frac{\left(\sinh^{-1} x\right)^2}{4\sinh^{-1} x - 3x}.$$

Let $\sinh^{-1} x = t$. Then $x = \sinh t$, the above inequality is equivalent to

$$\tanh^{-1}(\sinh t) > \frac{t^2}{4t - 3\sinh t}.$$

Let

$$Q(t) = \tanh^{-1}(\sinh t) - \frac{t^2}{4t - 3\sinh t}.$$

Then

$$Q'(t) = \frac{q(t)}{\left(1-\sinh^2 t\right)(4t - 3\sinh t)^2},$$

where

$$\begin{aligned} q(t) &= (\cosh t)(4t-3\sinh t)^2 - \left(1-\sinh^2 t\right)\left(3t^2\cosh t - 6t\sinh t + 4t^2\right) \\ &= \frac{9}{4}\cosh 3t - \frac{9}{4}\cosh t + 2t^2\cosh 2t + \frac{3}{4}t^2\cosh 3t + \frac{21}{2}t\sinh t \\ &\quad -12t\sinh 2t - \frac{3}{2}t\sinh 3t + \frac{49}{4}t^2\cosh t - 6t^2. \end{aligned}$$

Expanding in power series of the hyperbolic functions leads to

$$\begin{aligned} q(t) &= \frac{9}{4}\sum_{n=0}^{\infty}\frac{(3t)^{2n}}{(2n)!} - \frac{9}{4}\sum_{n=0}^{\infty}\frac{t^{2n}}{(2n)!} + 2t^2\sum_{n=0}^{\infty}\frac{(2t)^{2n}}{(2n)!} + \frac{3}{4}t^2\sum_{n=0}^{\infty}\frac{(3t)^{2n}}{(2n)!} \\ &\quad + \frac{21}{2}t\sum_{n=0}^{\infty}\frac{t^{2n+1}}{(2n+1)!} - 12t\sum_{n=0}^{\infty}\frac{(2t)^{2n+1}}{(2n+1)!} - \frac{3}{2}t\sum_{n=0}^{\infty}\frac{(3t)^{2n+1}}{(2n+1)!} \\ &\quad + \frac{49}{4}t^2\sum_{n=0}^{\infty}\frac{t^{2n}}{(2n)!} - 6t^2 \\ &= \sum_{n=2}^{\infty} r_n t^{2n+2}, \end{aligned}$$

where

$$r_n = \frac{4n^2-6n+17}{4(2n+2)!}3^{2n+1} + \frac{2n-11}{(2n+1)!}2^{2n+1} + \frac{196n^2+378n+173}{4(2n+2)!}.$$

We find that

$$r_2 = \frac{1}{2}, r_3 = \frac{11}{30}, r_4 = \frac{411}{5600}, r_5 = \frac{403}{50\,400},$$

and $r_n > 0$ for $n \geq 6$ due to $4n^2-6n+17>0$ and $2n-11>0$. So $r_n>0$ for $n\geq 2$. This leads to that $q(t)>0$. Then $Q'(t)>0$. So $Q(t)>Q(0^+)=0$, which completes the proof of the left hand side of (2.4).

(2) Then we prove the right hand side of (2.4).

The desired inequality is equivalent to

$$2\frac{x}{\sinh^{-1}x} + \frac{x}{\tanh^{-1}x} < 3.$$

Since

$$\begin{aligned} \frac{x}{\sinh^{-1}x} &< \frac{2+\sqrt{x^2+1}}{3}, \quad \text{(see [40])} \\ \frac{x}{\tanh^{-1}x} &< \frac{1+2\sqrt{1-x^2}}{3}, \quad \text{(see [6])} \end{aligned}$$

we have

$$2\frac{x}{\sinh^{-1}x} + \frac{x}{\tanh^{-1}x} < \frac{2\left(2+\sqrt{x^2+1}\right)}{3} + \frac{1+2\sqrt{1-x^2}}{3}.$$

In order to complete the proof of the right hand side of (2.4) it suffices to show

$$\frac{2\left(2+\sqrt{x^2+1}\right)}{3} + \frac{1+2\sqrt{1-x^2}}{3} < 3,$$

or

$$\begin{aligned} & 2\left(2+\sqrt{x^2+1}\right) + 1 + 2\sqrt{1-x^2} < 9 \\ \iff & \sqrt{x^2+1} < 2 - \sqrt{1-x^2} \\ \iff & x^2+1 < 4 - 4\sqrt{1-x^2} + 1 - x^2 \\ \iff & x^2 < 2 - 2\sqrt{1-x^2}. \end{aligned}$$

The last inequality is equivalent to $(1-\sqrt{1-x^2})^2 > 0$.

## 4. Further discussions

Let us consider a real function $f:(a,b) \longrightarrow \mathbb{R}$ in case when exist finite limits $f^{(k)}(a+) = \lim\limits_{x\to a+} f^{(k)}(x)$ (for $k = 0, 1, \ldots, n$ and $n \in \mathbb{N}_0$) and $f(b-) = \lim\limits_{x\to b-} f(x)$. We define

$$\begin{aligned} T_n^{f,\,a+}(x) &= \sum_{k=0}^{n} \frac{f^{(k)}(a+)}{k!}(x-a)^k, && (4.1)\\ R_n^{f,\,a+}(x) &= f(x) - T_n^{f,\,a+}(x), && (4.2) \end{aligned}$$

and

$$\mathbb{T}_n^{f;\,a+,\,b-}(x) = \begin{cases} T_{n-1}^{f,\,a+}(x) + \frac{1}{(b-a)^n} R_{n-1}^{f,\,a+}(b-)(x-a)^n &, \quad n \geq 1 \\ f(b-) &, \quad n = 0 \end{cases}. \tag{4.3}$$

Then the following statement is found to be true in [20, Theorem 3] and [18, Theorem 3].

**Theorem 4.1.** *Let $f : (a,b) \longrightarrow \mathbb{R}$ be real analytic function with the power series:*

$$f(x) = \sum_{k=0}^{\infty} c_k (x-a)^k, \tag{4.4}$$

*where $c_k \in \mathbb{R}$ and $c_k \geq 0$ for every $k \in \mathbb{N}_0$. Then,*

$$\begin{aligned} T_0^{f,\,a+}(x) &\leq \ldots \leq T_n^{f,\,a+}(x) \leq T_{n+1}^{f,\,a+}(x) \leq \ldots \\ \ldots &\leq f(x) \leq \ldots \\ \ldots &\leq \mathbb{T}_{n+1}^{f;\,a+,\,b-}(x) \leq \mathbb{T}_n^{f;\,a+,\,b-}(x) \leq \ldots \leq \mathbb{T}_0^{f;\,a+,\,b-}(x). \end{aligned} \tag{4.5}$$

*for every $x \in (a,b)$. If $c_k \in \mathbb{R}$ and $c_k \leq 0$ for every $k \in \mathbb{N}_0$, then the reversed inequality is true.*

Let us emphasize that previous theorem improves result of Theorem 2 from [31]. Inspired by [2, 13, 14, 17, 19, 22, 27], and [31], we obtain a conclusion more general than Theorem 2.1. The details are as follows.

**Theorem 4.2.** *Let us form the functions*

$$\begin{aligned} \varphi_1(x) &= \frac{1}{2}\left(\left(\frac{\sin x}{x}\right)^2 + \frac{\tan x}{x}\right) - \frac{1}{3}\left(\frac{2\sin x}{x} + \frac{\tan x}{x}\right) : \left(0, \frac{\pi}{2}\right) \longrightarrow R, \\ \varphi_2(x) &= \frac{1}{3}\left(\frac{2\sin x}{x} + \frac{\tan x}{x}\right) - \frac{1}{2}\left(\left(\frac{x}{\sin x}\right)^2 + \frac{x}{\tan x}\right) : \left(0, \frac{\pi}{2}\right) \longrightarrow R, \\ \varphi_3(x) &= \frac{1}{2}\left(\left(\frac{x}{\sin x}\right)^2 + \frac{x}{\tan x}\right) - \frac{1}{3}\left(\frac{2x}{\sin x} + \frac{x}{\tan x}\right) : \left(0, \frac{\pi}{2}\right) \longrightarrow R, \\ \varphi_4(x) &= \frac{1}{3}\left(\frac{2x}{\sin x} + \frac{x}{\tan x}\right) - 1 : \left(0, \frac{\pi}{2}\right) \longrightarrow R. \end{aligned}$$

*Then functions $\varphi_1(x), \varphi_2(x), \varphi_3(x), \varphi_4(x)$ are real analytic with power series*

$$\varphi_1(x) = \sum_{k=2}^{\infty} s_k^{(1)} x^{2k}, \ \varphi_2(x) = \sum_{k=2}^{\infty} s_k^{(2)} x^{2k}, \ \varphi_3(x) = \sum_{k=2}^{\infty} s_k^{(3)} x^{2k}, \ \varphi_3(x) = \sum_{k=2}^{\infty} s_k^{(3)} x^{2k}.$$

*with positive coefficients*

$$s_n^{(1)} = \frac{1}{2}\frac{(-1)^n 2^{2n+1}}{(2n+2)!} + \frac{1}{6}\frac{(2^{2n+2}-1)2^{2n+1}}{(2n+2)!}|B_{2n+2}| - \frac{2}{3}\frac{(-1)^n}{(2n+1)!} > 0,$$

$$
\begin{aligned}
s_n^{(2)} &= \frac{2\,(-1)^n}{3\,(2n+1)!} + \frac{(2^{2n+2}-1)\,2^{2n+2}}{3\,(2n+2)!}|B_{2n+2}| - \frac{(n-1)\,2^{2n}}{(2n)!}|B_{2n}| > 0,\\
s_n^{(3)} &= \frac{(3n-4)\,2^{2n}+4}{3\,(2n)!}|B_{2n}| > 0,\\
s_n^{(4)} &= \frac{2^{2n}-4}{3\,(2n)!}|B_{2n}| > 0
\end{aligned}
$$

*for $n = 2, 3, \ldots$. Let it be that $j \in \{1, 2, 3, 4\}$ and $c \in (0, \pi/2)$ fixed. Then the double inequality*

$$
\begin{aligned}
0 < T_2^{\varphi_j,0+}(x) &\le T_3^{\varphi_j,0+}(x) \ldots \le T_n^{\varphi_j,0+}(x) \le T_{n+1}^{\varphi_j,0+}(x) \le \ldots\\
\ldots &\le \varphi_j(x) \le \ldots \qquad (4.6)\\
\ldots &\le \mathbb{T}_{n+1}^{\varphi_j;0+,c-}(x) \le \mathbb{T}_n^{\varphi_j;0+,c-}(x) \le \ldots \mathbb{T}_3^{\varphi_j;0+,c-}(x) \le \mathbb{T}_2^{\varphi_j;0+,c-}(x)
\end{aligned}
$$

*holds for all $x \in (0, c)$.*

***Proof.*** For example, let us consider only case $j = 2$. Since

$$
\varphi(x) = \varphi_2(x) = \frac{1}{3}\left(\frac{2\sin x}{x} + \frac{\tan x}{x}\right) - \frac{1}{2}\left(\left(\frac{x}{\sin x}\right)^2 + \frac{x}{\tan x}\right) = \sum_{k=2}^{\infty} s_k x^{2k},
$$

where

$$
s_n = s_n^{(2)} = \frac{2\,(-1)^n}{3\,(2n+1)!} + \frac{(2^{2n+2}-1)\,2^{2n+2}}{3\,(2n+2)!}|B_{2n+2}| - \frac{(n-1)\,2^{2n}}{(2n)!}|B_{2n}|,\ n \ge 2.
$$

We can prove $s_n > 0$ holds for all $n \ge 2$. In [10, 1.3.1.4] or [46, 1.3.10], we can find the following power series expansion:

$$
\tan x = \sum_{n=1}^{\infty} \frac{2^{2n}-1}{(2n)!} 2^{2n} |B_{2n}| x^{2n-1}, \quad |x| < \frac{\pi}{2}. \qquad (4.7)
$$

Based on $(3.9)$, $(3.10)$, and $(4.7)$ follows

$$
\begin{aligned}
\varphi(x) =& \frac{2}{3}\frac{\sin x}{x} + \frac{1}{3}\frac{\tan x}{x} - \frac{1}{2}\left(\frac{x}{\sin x}\right)^2 - \frac{1}{2}\frac{x}{\tan x}\\
=& \frac{2}{3}\sum_{n=0}^{\infty}\frac{(-1)^n}{(2n+1)!}x^{2n} + \frac{1}{3}\sum_{n=1}^{\infty}\frac{2^{2n}-1}{(2n)!}2^{2n}|B_{2n}|x^{2n-2}\\
&- \frac{1}{2}\left[1 + \sum_{n=1}^{\infty}\frac{2^{2n}(2n-1)}{(2n)!}|B_{2n}|x^{2n}\right] - \frac{1}{2}\left[1 - \sum_{n=1}^{\infty}\frac{2^{2n}}{(2n)!}|B_{2n}|x^{2n}\right]\\
=& \frac{2}{3}\sum_{n=0}^{\infty}\frac{(-1)^n}{(2n+1)!}x^{2n} + \frac{1}{3}\sum_{n=0}^{\infty}\frac{2^{2n+2}-1}{(2n+2)!}2^{2n+2}|B_{2n+2}|x^{2n}\\
&- \frac{1}{2}\left[1 + \sum_{n=1}^{\infty}\frac{2^{2n}(2n-1)}{(2n)!}|B_{2n}|x^{2n}\right] - \frac{1}{2}\left[1 - \sum_{n=1}^{\infty}\frac{2^{2n}}{(2n)!}|B_{2n}|x^{2n}\right]\\
=& \sum_{n=2}^{\infty}\frac{2\,(-1)^n}{3\,(2n+1)!}x^{2n} + \sum_{n=2}^{\infty}\frac{(2^{2n+2}-1)\,2^{2n+2}}{3\,(2n+2)!}|B_{2n+2}|x^{2n}\\
&- \sum_{n=2}^{\infty}\frac{2^{2n-1}(2n-1)}{(2n)!}|B_{2n}|x^{2n} + \sum_{n=2}^{\infty}\frac{2^{2n-1}}{(2n)!}|B_{2n}|x^{2n}\\
=& \sum_{n=2}^{\infty}\frac{2\,(-1)^n}{3\,(2n+1)!}x^{2n} + \sum_{n=2}^{\infty}\frac{(2^{2n+2}-1)\,2^{2n+2}}{3\,(2n+2)!}|B_{2n+2}|x^{2n} - \sum_{n=2}^{\infty}\frac{(n-1)\,2^{2n}}{(2n)!}|B_{2n}|x^{2n}
\end{aligned}
$$

$$= \sum_{n=2}^{\infty} s_n x^{2n}.$$

Next, we shall prove that $s_n > 0$ for all $n \geq 2$.

(i) When $n$ is even,

$$s_n = \frac{2}{3\,(2n+1)!} + \frac{(2^{2n+2}-1)\,2^{2n+2}}{3\,(2n+2)!}|B_{2n+2}| - \frac{(n-1)\,2^{2n}}{(2n)!}|B_{2n}|,$$

we complete the proof of $s_n > 0$ as long as

$$\frac{(2^{2n+2}-1)\,2^{2n+2}}{3\,(2n+2)!}|B_{2n+2}| - \frac{(n-1)\,2^{2n}}{(2n)!}|B_{2n}| > 0,$$

or

$$\frac{|B_{2n+2}|}{|B_{2n}|} > \frac{\frac{(n-1)2^{2n}}{(2n)!}}{\frac{(2^{2n+2}-1)2^{2n+2}}{3(2n+2)!}} = \frac{(n-1)\,2^{2n}}{(2n)!}\frac{3\,(2n+2)!}{(2^{2n+2}-1)\,2^{2n+2}}.$$

Since

$$\frac{|B_{2n+2}|}{|B_{2n}|} > \frac{2^{2n-1}-1}{2^{2n+1}-1}\frac{(2n+2)(2n+1)}{\pi^2}, \text{ (see [1,25,26,35,45])}$$

we complete the proof when proving

$$\frac{2^{2n-1}-1}{2^{2n+1}-1}\frac{(2n+2)(2n+1)}{\pi^2} > \frac{(n-1)\,2^{2n}}{(2n)!}\frac{3\,(2n+2)!}{(2^{2n+2}-1)\,2^{2n+2}},$$

that is,

$$2^{2n} > \frac{6\left[\pi^2\,(n-1)+3\right]}{8} \text{ for } n \geq 2.$$

It is not difficult to prove the above formula by mathematical induction.

(ii) When $n$ is odd,

$$s_n = -\frac{2}{3\,(2n+1)!} + \frac{(2^{2n+2}-1)\,2^{2n+2}}{3\,(2n+2)!}|B_{2n+2}| - \frac{(n-1)\,2^{2n}}{(2n)!}|B_{2n}|.$$

By

$$\frac{2(2n)!}{(2\pi)^{2n}}\frac{1}{1-2^{-2n}} < |B_{2n}| < \frac{2(2n)!}{(2\pi)^{2n}}\frac{1}{1-2^{1-2n}}, n = 1, 2, \cdots, \text{ (see [1])}$$

we have

$$\begin{aligned}
s_n &> -\frac{2}{3\,(2n+1)!} + \frac{(2^{2n+2}-1)\,2^{2n+2}}{3\,(2n+2)!}\frac{2(2n+2)!}{(2\pi)^{2n+2}}\frac{1}{1-2^{-2n-2}} \\
&\quad -\frac{(n-1)\,2^{2n}}{(2n)!}\frac{2(2n)!}{(2\pi)^{2n}}\frac{1}{1-2^{1-2n}} \\
&= -\frac{2}{3\,(2n+1)!} + \frac{(2^{2n+2}-1)\,2^{2n+2}}{3\,(2n+2)!}\frac{2(2n+2)!}{(2\pi)^{2n+2}}\frac{2^{2n+2}}{2^{2n+2}-1} \\
&\quad -\frac{(n-1)\,2^{2n}}{(2n)!}\frac{2(2n)!}{(2\pi)^{2n}}\frac{2^{2n-1}}{2^{2n-1}-1} \\
&= \frac{2\cdot 2^{2n+2}}{3\pi^{2n+2}} - \frac{2^{2n}\,(n-1)}{(2^{2n-1}-1)\,\pi^{2n}} - \frac{2}{3\,(2n+1)!}.
\end{aligned}$$

Since

$$s_n > 0 \Longleftrightarrow \frac{2}{3}\frac{(4\cdot 2^{2n} - 3\pi^2 n + 3\pi^2 - 8)\,2^{2n}}{\pi^{2n}\pi^2\,(2^{2n}-2)} > \frac{2}{3\,(2n+1)!}$$

$$\iff \frac{(4\cdot 2^{2n}-3\pi^2 n+3\pi^2-8)\,2^{2n}}{\pi^{2n+2}\,(2^{2n}-2)} > \frac{1}{(2n+1)!}$$
$$\iff \left(4\cdot 2^{2n}-3\pi^2 n+3\pi^2-8\right)2^{2n}\,(2n+1)! > \pi^{2n+2}\left(2^{2n}-2\right),$$

and

$$n! > \left(\frac{n}{3}\right)^n,\ n\in\mathbb{N},$$

we have

$$(2n+1)! > \left(\frac{2n+1}{3}\right)^{2n+1} > 2^{2n+1},\ n\in\mathbb{N}_0,$$

and

$$\left(4\cdot 2^{2n}-3\pi^2 n+3\pi^2-8\right)2^{2n}\,(2n+1)! > \left(4\cdot 2^{2n}-3\pi^2 n+3\pi^2-8\right)2^{2n}2^{2n+1}.$$

Then we complete the proof when proving

$$\left(4\cdot 2^{2n}-3\pi^2 n+3\pi^2-8\right)2^{2n}2^{2n+1} > \pi^{2n+2}\left(2^{2n}-2\right),$$

that is,

$$\begin{aligned} t_n &= \left(4\cdot 2^{2n}-3\pi^2 n+3\pi^2-8\right)2^{2n}2^{2n+1}-\pi^{2n+2}\left(2^{2n}-2\right) \\ &= 8\cdot 8^{2n}-(2\pi)^{2n}\pi^2-2\cdot 4^{2n}\left[3\pi^2(n-1)+8\right]+2\pi^2\pi^{2n} \\ &> 0 \end{aligned}$$

for all $n\geq 2$. We find

$$t_2 = 28\,672-14\pi^6-1536\pi^2 = 52.839\ldots > 0,$$

and

$$\begin{aligned} t_{n+1}-64t_n &= \left[4\pi^2(4-\pi)(\pi+4)\,2^{2n}-\left(128\pi^2-2\pi^4\right)\right]\pi^{2n} \\ &\quad +96\cdot 4^{2n}\left(3\pi^2 n-4\pi^2+8\right) \\ &> 0. \end{aligned}$$

Then $t_n>0$ for all $n\geq 2$. □

**Remark 4.3.** Obviously, Theorem 2.1 is a simple corollary of Theorem 4.2.

**Acknowledgment.** The authors are grateful to anonymous referees for their careful corrections to and valuable comments on the original version of this paper.
The first author was supported by the National Natural Science Foundation of China (no. 61772025). The second author was supported in part by the Serbian Ministry of Education, Science and Technological Development, under projects ON 174032 and III 44006.

## References

[1] M. Abramowitz and I. A. Stegun (Eds), *Handbook of Mathematical Functions with Formulas, Graphs, and Mathematical Tables*, National Bureau of Standards, Applied Mathematics Series **55**, Dover Publications, 1972.

[2] B. Banjac, M. Makragić, and B. Malešević, *Some notes on a method for proving inequalities by computer*, Results Math. **69**, 161–176, 2016.

[3] A. Baricz and J. Sandor, *Extensions of the generalized Wilker inequality to Bessel functions*, J. Math. Inequal. **2** (3), 397–406, 2008.

[4] F. Cajori, *A History of Mathematics*, 2nd ed., New York, 1929.

[5] F.T. Campan, *The Story of Number* $\pi$, Ed. Albatros, Romania, 1977.

[6] C.-P. Chen and W.-S. Cheung, *Wilker- and Huygens-type inequalities and solution to Oppenheim's problem*, Int. Trans. Spec. Funct. **23** (5), 325–336, 2012.

[7] B.-N. Guo, B.-M. Qiao, F. Qi, and W. Li, *On new proofs of Wilker inequalities involving trigonometric functions*, Math. Inequal. Appl. **6** (1), 19–22, 2003.

[8] C. Huygens, *Oeuvres Completes*, Publiees Par la Societe Hollandaise des Science, Haga, 20 volumes, 1888–1940.

[9] A.P. Iuskevici, *History of Mathematics in 16th and 17th Centuries*, Moskva, 1961.

[10] A. Jeffrey, *Handbook of Mathematical Formulas and Integrals, 3rd ed.*, Elsevier Acad. Press, San Diego, CA, 2004.

[11] J.-C. Kuang, *Applied Inequalities, 3rd ed.* (in Chinese), Shangdong Science and Technology Press, Jinan City, Shangdong Province, China, 2004.

[12] J.-L. Li, *An identity related to Jordan's inequality*, Int. J. Math. Math. Sci. **6**, Article ID 76782, 2016.

[13] T. Lutovac, B. Malešević, and C. Mortici, *The natural algorithmic approach of mixed trigonometric-polynomial problems*, J. Inequal. Appl. **2017**, Article No: 116, 2017.

[14] T. Lutovac, B. Malešević, and M. Rašajski, *A new method for proving some inequalities related to several special functions*, Results Math. **73**, Article No: 100, 2018.

[15] B.J. Malešević, *Application of λ-method on Shafer–Fink's inequality*, Univ. Beograd. Publ. Elektrotehn. Fak. Ser. Mat. **8**, 103–105, 1997.

[16] B.J. Malešević, *An application of λ-method on inequalities of Shafer–Fink's type*, Math. Inequal. Appl. **10**, 529–534, 2007.

[17] B. Malešević, T. Lutovac, M. Rašajski, and C. Mortici, *Extensions of the natural approach to refinements and generalizations of some trigonometric inequalities*, Adv. Difference Equ. **2018**, Article No: 90, 2018.

[18] B. Malešević, M. Nenezić, L. Zhu, B. Banjac and M. Petrovic, *Some new estimates of precision of Cusa-Huygens and Huygens approximations*, accepted in Appl. Anal. Discrete Math., 2020.

[19] B. Malešević, M. Rašajski, and T. Lutovac, *Refinements and generalizations of some inequalities of Shafer-Fink's type for the inverse sine function*, J. Inequal. Appl. **2017**, Article No: 275, 2017.

[20] B. Malešević, M. Rašajski, and T. Lutovac, *Double-sided Taylor's approximations and their applications in Theory of analytic inequalities.* in: Differential and Integral Inequalities, Th.M. Rassias, D. Andrica (eds.), Optimization and Its Applications, vol. **151**, 569-582, 2019.

[21] D.S. Mitrinović, *Analytic Inequalities*, Springer-Verlag, New York, Berlin 1970.

[22] M. Nenezić and L. Zhu, *Some improvements of Jordan-Steckin and Becker-Stark inequalities*, Appl. Anal. Discrete Math. **12**, 244–256, 2018.

[23] E. Neuman, *Wilker and Huygens-type inequalities for Jacobian elliptic and theta functions*, Int. Trans. Spec. Funct. **25** (3), 240–248, 2014.

[24] E. Neuman and J. Sandor, *On some inequalities involving trigonometric and hyperbolic functions with emphasis on the Cusa-Huygens, Wilker, and Huygens inequalities*, Math. Inequal. Appl. **13** (4), 715–723, 2010.

[25] F. Qi, *Notes on a double inequality for ratios of any two neighbouring non-zero Bernoulli numbers*, Turkish J. Anal. Number Theory **6** (5), 129–131, 2018.

[26] F. Qi, *A double inequality for the ratio of two non-zero neighbouring Bernoulli numbers*, J. Comput. Appl. Math. **351**, 1–5, 2019.

[27] M. Rašajski, T. Lutovac, and B. Malešević, About some exponential inequalities related to the sinc function, J. Inequal. Appl. **2018**, Article No: 150, 2018.

[28] J. Sandor and M. Bencze, *On Huygens's trigonometric inequality*, RGMIA Research Report Collection **8** (3), Art. 14, 2005.

[29] J.S. Sumner, A.A. Jagers, M. Vowe, and J. Anglesio, *Inequalities involving trigonometric functions*, Amer. Math. Monthly **98**, 264–267, 1991.

[30] J.B. Wilker, *Problem E 3306*, Amer. Math. Monthly **96**, 55, 1989.

[31] S.-H. Wu, L. Debnath, *A generalization of L'Hôspital-type rules for monotonicity and its application*, Appl. Math. Lett. **22** (2), 284-290, 2009.

[32] S.-H. Wu, H.M. Srivastava, *A weighted and exponential generalization of Wilker's inequality and its applications*, Int. Trans. Spec. Funct. **18** (8), 529–535, 2008.

[33] Z.-H. Yang, W.-M. Qian, Y.-M. Chu and W. Zhang, *On approximating the arithmetic-geometric mean and complete elliptic integral of the first kind*, J. Math. Anal. Appl. **462**, 1714–1726, 2018.

[34] Z.-H. Yang and J.-F. Tian, *Convexity and monotonicity for the elliptic integrals of the first kind and applications*, Appl. Anal. Discrete Math. **13** (1), 240-260, 2019.

[35] Z.-H. Yang and J.-F. Tian, *Sharp bounds for the ratio of two zeta functions*, J. Comput. Appl. Math. **364**, 112359, 14 pages, 2020.

[36] L. Zhu, *On Shafer-Fink inequalities*, Math. Inequal. Appl. **8** (4), 571–574, 2005.

[37] L. Zhu, *A new simple proof of Wilker's inequality*, Math. Inequal. Appl. **8** (4), 749–750, 2005.

[38] L. Zhu, *A solution of a problem of Oppenheim*, Math. Inequal. Appl. **10**, 57–61, 2007.

[39] L. Zhu, *On Wilker-type inequalities*, Math. Inequal. Appl. **10** (4), 727–731, 2007.

[40] L. Zhu, *New inequalities of Shafer–Fink type for arc hyperbolic sine*, J. Inequal. Appl. **2008**, Article ID 368275, 2008.

[41] L. Zhu, *Some new inequalities of the Huygens type*, Comput. Math. Appl. **58**, 1180–1182, 2009.

[42] L. Zhu, *Some new Wilker-type inequalities for circular and hyperbolic functions*, Abstr. Appl. Anal. **2009**, Article ID 485842, 9 pages, 2009.

[43] L. Zhu, *A source of inequalities for circular functions*, Comput. Math. Appl. **58**, 1998–2004, 2009.

[44] L. Zhu, *Inequalities for hyperbolic functions and their applications*, *J. Inequal. Appl.* **2010**, Article ID 130821, 10 pages, 2010.

[45] L. Zhu, *New bounds for the ratio of two adjacent even-indexed Bernoulli numbers*, Rev. R. Acad. Cienc. Exactas Fís. Nat. Ser. A Mat. **114**, 2020.

[46] D. Zwillinger, *CRC Standard Mathematical Tables and Formulae*, CRC Press, 1996.

Hacettepe Journal of
Mathematics & Statistics

Hacet. J. Math. Stat.
Volume 50 (1) (2021), 63 – 78
DOI : 10.15672/hujms.612642

Research Article

# Some Laplace transforms and integral representations for parabolic cylinder functions and error functions

Dirk Veestraeten

*Amsterdam School of Economics, University of Amsterdam, Roetersstraat 11, 1018WB Amsterdam, The Netherlands*

**Abstract**

This paper uses the convolution theorem of the Laplace transform to derive new inverse Laplace transforms for the product of two parabolic cylinder functions in which the arguments may have opposite sign. These transforms are subsequently specialized for products of the error function and its complement thereby yielding new integral representations for products of the latter two functions. The transforms that are derived in this paper also allow to correct two inverse Laplace transforms that are widely reported in the literature and subsequently uses one of the corrected expressions to obtain two new definite integrals for the generalized hypergeometric function.

**Mathematics Subject Classification (2020).** 33B20, 33C05, 33C15, 33C20, 44A10, 44A35

**Keywords.** confluent hypergeometric function, convolution theorem, error function, Gaussian hypergeometric function, generalized hypergeometric function, Laplace transform, parabolic cylinder function

## 1. Introduction

The parabolic cylinder function is intensively used in various domains such as chemical physics [17], lattice field theory [8], astrophysics [30], finance [20], neurophysiology [5] and estimation theory [4]. Products of parabolic cylinder functions involving both positive and negative arguments arise in, for instance, problems of condensed matter physics [7,18] and the study of real zeros of parabolic cylinder functions [9–11]. The error function $\operatorname{erf}(x)$ and its complement $\operatorname{erfc}(x)$ emerge as special cases of the parabolic cylinder function and play a prominent role in, for instance, the conduction of heat [6], statistics and probability theory [15,23] and hydrology [2].

However, the extensive tables of inverse Laplace transforms [14,21,26] present relatively few expressions for products of parabolic cylinder functions especially when signs of the arguments differ. For example, [26] only specifies the following inverse Laplace transforms for such set–up

Email address: d.j.m.veestraeten@uva.nl
Received: 29.08.2019; Accepted: 28.04.2020

$$D_\nu\left(a\sqrt{p+\sqrt{p^2+b^2}}\right)\left\{D_\nu\left(-a\sqrt{\sqrt{p^2+b^2}-p}\right)\pm D_\nu\left(a\sqrt{\sqrt{p^2+b^2}-p}\right)\right\}$$
$$D_\nu\left(a\sqrt{p+\sqrt{p^2-b^2}}\right)\left\{D_\nu\left(-a\sqrt{p-\sqrt{p^2-b^2}}\right)\pm D_\nu\left(a\sqrt{p-\sqrt{p^2-b^2}}\right)\right\}$$

see Equations (3.11.4.9) and (3.11.4.10).

This paper uses the convolution theorem of the Laplace transform to derive inverse Laplace transforms for

$$p^i\exp\left(\tfrac{1}{2}p\,(y-x)\right)D_\mu\left(2^{1/2}y^{1/2}p^{1/2}\right)\left\{D_\nu\left(-2^{1/2}x^{1/2}p^{1/2}\right)\pm D_\nu\left(2^{1/2}x^{1/2}p^{1/2}\right)\right\}$$

with $i=0$ or $-\frac{1}{2}$, i.e. for expressions in which the arguments have opposite sign and differ, and where also the orders take on different values.

These results also offer inverse Laplace transforms for the product of (complementary) error functions as the parabolic cylinder function for order $-1$ specializes into the complementary error function. As a result, novel integral representations are obtained for products of the (complementary) error functions and, for instance, the integral representation for $1-\mathrm{erf}(a)^2$ in [19] can be generalized into $1-\mathrm{erf}(a)\mathrm{erf}(b)$.

The paper also corrects two inverse Laplace transforms that are reported in [14,21,26]. Combinations of one of the corrected results with the results derived in this paper are particularly interesting as they yield two definite integrals for the generalized hypergeometric function that are not reported in, for instance, the comprehensive overview in [16].

The remainder of this paper is organized as follows. Section 2 presents the relation between the parabolic cylinder function and the Kummer confluent hypergeometric function that is central to the subsequent derivations. Also, more detail is presented on the formulation of the convolution theorem for the Laplace transform given that the limits of integration in the integrals in the product differ. Section 3 presents the inverse Laplace transforms for products of the parabolic cylinder function and uses these results to obtain novel integral representations for products of (complementary) error functions. Section 4 corrects two widely-reported inverse Laplace transforms. Section 5 uses one of these corrected expressions together with the results of Section 3 to derive two novel definite integrals for the generalized hypergeometric function.

## 2. Notation and background

The parabolic cylinder function in the definition of Whittaker [29] is denoted by $D_\nu(z)$, where $\nu$ and $z$ represent the order and the argument, respectively. Equation (4) on p. 117 in [13] defines the parabolic cylinder function as follows

$$D_\nu(z)=2^{\nu/2}\exp\left(-\tfrac{1}{4}z^2\right)\left\{\frac{\Gamma[1/2]}{\Gamma[(1-\nu)/2]}\Phi\left(-\frac{\nu}{2};\frac{1}{2};\frac{1}{2}z^2\right)\right.$$
$$\left.+\frac{z}{2^{1/2}}\frac{\Gamma[-1/2]}{\Gamma[-\nu/2]}\Phi\left(\frac{1-\nu}{2};\frac{3}{2};\frac{1}{2}z^2\right)\right\}\qquad(2.1)$$

where $\Phi(a;b;z)$ is Kummer's confluent hypergeometric function

$$\Phi(a;b;z)=\sum_{n=0}^{\infty}\frac{(a)_n}{(b)_n}\frac{z^n}{n!},$$

$\Gamma[z]$ denotes the gamma function

$$\Gamma[z]=\int_0^\infty t^{z-1}\exp(-t)\,dt$$

and $(z)_n$ denotes the Pochhammer symbol

$$(z)_n = \frac{\Gamma[z+n]}{\Gamma[z]},$$

see Equation (1) on p. 434 in [24], Equations (6.1.1) and (6.1.22) in [1], respectively.

Note that the definition (2.1) holds for $z$ as well as $-z$ and adding the corresponding relation for $D_\nu(-z)$ to (2.1) then gives

$$D_\nu(-z) - D_\nu(z) = \frac{z 2^{(\nu+3)/2}\sqrt{\pi}}{\Gamma[-\nu/2]} \exp\left(-\tfrac{1}{4}z^2\right) \Phi\left(\frac{1-\nu}{2}; \frac{3}{2}; \frac{1}{2}z^2\right) \tag{2.2}$$

$$D_\nu(-z) + D_\nu(z) = \frac{2^{(\nu+2)/2}\sqrt{\pi}}{\Gamma[(1-\nu)/2]} \exp\left(-\tfrac{1}{4}z^2\right) \Phi\left(-\frac{\nu}{2}; \frac{1}{2}; \frac{1}{2}z^2\right) \tag{2.3}$$

see Equations (46:5:4) and (46:5:3) in [22].

The convolution theorem of the Laplace transform will be used to derive inverse Laplace transforms for products of two parabolic cylinder functions. The functions in the products are taken from inverse Laplace transforms for the parabolic cylinder function and the Kummer confluent hypergeometric function, respectively. The inverse Laplace transforms that will be used for $\Phi(a;b;z)$ and $D_\nu(z)$ are not both defined over the half–line $(0,\infty)$. As a result, the convolution theorem becomes somewhat more involved. The Laplace transforms of the original functions $f_1(t)$ and $f_2(t)$ are defined as

$$\overline{f}_1(p) = \int_{\alpha_1}^{\beta_1} \exp(-pt)\, f_1(t)\, dt \qquad \beta_1 > \alpha_1$$

$$\overline{f}_2(p) = \int_{\alpha_2}^{\beta_2} \exp(-pt)\, f_2(t)\, dt \qquad \beta_2 > \alpha_2$$

where $\operatorname{Re} p > 0$. The convolution theorem then can be specified, see [25], as

$$\overline{f}_1(p)\,\overline{f}_2(p) = \int_{\alpha_1+\alpha_2}^{\beta_1+\beta_2} \exp(-pt)\, f_1(t) * f_2(t)\, dt \tag{2.4}$$

where $f_1(t) * f_2(t)$ is the convolution of $f_1(t)$ and $f_2(t)$ that is to be obtained from

$$f_1(t) * f_2(t) = \int_{\max(\alpha_1;\, t-\beta_2)}^{\min(\beta_1;\, t-\alpha_2)} f_1(\tau)\, f_2(t-\tau)\, d\tau \tag{2.5}$$

## 3. Inverse Laplace transforms for products of parabolic cylinder functions

This section derives several inverse Laplace transforms for products of parabolic cylinder functions in which the sign of the arguments may differ and utilizes these results to obtain new integral representations for products of (complementary) error functions.

**Theorem 3.1.** *Let $\nu$ and $\mu$ be two complex numbers with $\operatorname{Re}\nu < 1$ and $\operatorname{Re}\mu < \min[1-\operatorname{Re}\nu, 2+\operatorname{Re}\nu]$. Then, the following inverse Laplace transform holds for $\operatorname{Re} p > 0$, $x > 0$, $|\arg y| < \pi$, $y > 0$*

$$\begin{aligned}
&p^{-1/2} \exp\left(\tfrac{1}{2}p(y-x)\right) D_\mu\left(2^{1/2}y^{1/2}p^{1/2}\right) \left\{D_\nu\left(-2^{1/2}x^{1/2}p^{1/2}\right) - D_\nu\left(2^{1/2}x^{1/2}p^{1/2}\right)\right\} \\
&= \frac{2^{(\mu-\nu)/2}\sqrt{\pi}}{\Gamma[1+(\nu-\mu)/2]\,\Gamma[-\nu]} \int_0^x \exp(-pt)\, t^{(\nu-\mu)/2} (x-t)^{-(1+\nu)/2} \\
&\quad \times (y+t)^{\mu/2}\, {}_2F_1\left(-\frac{\mu}{2}, \frac{1+\nu}{2}; 1+\frac{\nu-\mu}{2}; \frac{t(x-y-t)}{(x-t)(y+t)}\right) dt \\
&\quad + \frac{2^{2+(\mu+\nu)/2}\sqrt{\pi}\, y^{1/2} x^{1/2}}{\Gamma[-\mu/2]\,\Gamma[-\nu/2]} \int_x^\infty \exp(-pt)\, t^{(\nu-1)/2} (t-x)^{-(1+\mu+\nu)/2} \\
&\quad \times (y-x+t)^{(\mu-1)/2}\, {}_2F_1\left(\frac{1-\mu}{2}, \frac{1-\nu}{2}; \frac{3}{2}; \frac{xy}{t(y-x+t)}\right) dt
\end{aligned} \tag{3.1}$$

where ${}_2F_1(a,b;c;z)$ denotes the Gaussian hypergeometric function

$$ {}_2F_1(a,b;c;z) = \sum_{n=0}^{\infty} \frac{(a)_n (b)_n}{(c)_n} \frac{z^n}{n!} \qquad |z|<1, $$

see Equation (1) on p. 430 in [24].

***Proof.*** The inverse Laplace transform in Equation (5) on p. 290 in [14] is

$$ \Gamma[\nu] \exp\left(\tfrac{1}{2}ap\right) D_{-2\nu}\left(2^{1/2}a^{1/2}p^{1/2}\right) = \int_0^\infty \exp(-pt)\, 2^{-\nu} a^{1/2} t^{\nu-1} (t+a)^{-\nu-1/2}\, dt \tag{3.2} $$
$$ [\operatorname{Re} p > 0, \operatorname{Re}\nu > 0, |\arg a| < \pi] $$

and the inverse Laplace transform in Equation (3.33.2.2) in [26] is

$$ \exp(-xp)\, \Phi(a;b;xp) = \frac{x^{1-b}\Gamma[b]}{\Gamma[b-a]\,\Gamma[a]} \int_0^x \exp(-pt)\, t^{b-a-1} (x-t)^{a-1}\, dt \tag{3.3} $$
$$ [\operatorname{Re} p > 0, \operatorname{Re} b > \operatorname{Re} a > 0, x > 0] $$

These two inverse Laplace transforms, in the notation of Theorem 3.1, are rewritten as

$$ \Gamma[-\mu/2] \exp\left(\tfrac{1}{2}yp\right) D_{\mu}\left(2^{1/2}y^{1/2}p^{1/2}\right) = \int_0^\infty \exp(-pt)\, 2^{\mu/2} y^{1/2} t^{-\mu/2-1} (t+y)^{(\mu-1)/2}\, dt $$
$$ [\operatorname{Re} p > 0, \operatorname{Re}\mu < 0, |\arg y| < \pi] \tag{3.4} $$

and

$$ \begin{aligned} & x^{1/2}\frac{2}{\sqrt{\pi}}\Gamma[1+\nu/2]\,\Gamma[(1-\nu)/2] \exp(-xp)\, \Phi\left(\frac{1-\nu}{2};\frac{3}{2};px\right) \\ &\quad = \int_0^x \exp(-pt)\, t^{\nu/2} (x-t)^{-(1+\nu)/2}\, dt \\ &\qquad [\operatorname{Re} p > 0, -2 < \operatorname{Re}\nu < 1, x > 0] \end{aligned} \tag{3.5} $$

The original functions $f_1(t)$ and $f_2(t)$ are taken from the inverse Laplace transforms (3.4) and (3.5), respectively, with

$$ f_1(t) = 2^{\mu/2} y^{1/2} t^{-\mu/2-1} (t+y)^{(\mu-1)/2} \text{ and } f_2(t) = t^{\nu/2}(x-t)^{-(1+\nu)/2} $$

The integration limits in (2.4) and (2.5) are $\beta_1 = \infty, \beta_2 = x$ and $\alpha_1 = \alpha_2 = 0$ such that the convolution integral is given by

$$ \begin{aligned} f_1(t) * f_2(t) &= \int_0^t f_1(\tau)\, f_2(t-\tau)\, d\tau \qquad && t < x \\ &= \int_{t-x}^t f_1(\tau)\, f_2(t-\tau)\, d\tau && t > x \end{aligned} \tag{3.6} $$

First, the convolution integral for $t < x$ is

$$ f_1(t) * f_2(t) = \int_0^t 2^{\mu/2} y^{1/2} \tau^{-\mu/2-1} (\tau+y)^{(\mu-1)/2} (t-\tau)^{\nu/2} (x-(t-\tau))^{-(1+\nu)/2}\, d\tau $$

The substitution $\tau = tu$ allows to rewrite the integral as

$$ \begin{aligned} f_1(t) * f_2(t) &= 2^{\mu/2} t^{(\nu-\mu)/2} y^{\mu/2} (x-t)^{-(1+\nu)/2} \\ &\quad \times \int_0^1 u^{-\mu/2-1} \left(1+\frac{t}{y}u\right)^{(\mu-1)/2} (1-u)^{\nu/2} \left(1-\frac{t}{t-x}u\right)^{-(1+\nu)/2} du \end{aligned} $$

The integral in the latter equation will be expressed in terms of the Appell hypergeometric function $F_1(a, b_1, b_2; c; z_1, z_2)$, which is defined as

$$F_1(a, b_1, b_2; c; z_1, z_2) = \sum_{m=0}^{\infty} \sum_{n=0}^{\infty} \frac{(a)_{m+n} (b_1)_m (b_2)_n}{(c)_{m+n}} \frac{z_1^m z_2^n}{m!n!} \qquad \max\{|z_1|, |z_2|\} < 1,$$

see Equation (1) on p. 448 in [24]. In particular, the following integral representation of the Appell hypergeometric function $F_1(a, b_1, b_2; c; z_1, z_2)$ will be used

$$\frac{\Gamma[a]\Gamma[c-a]}{\Gamma[c]} F_1(a, b_1, b_2; c; z_1, z_2) = \int_0^1 u^{a-1} (1-u)^{c-a-1} (1-z_1 u)^{-b_1} (1-z_2 u)^{-b_2} \, du$$

for $\operatorname{Re} c > \operatorname{Re} a > 0$, see Equation (5) on p. 231 in [12]. This gives

$$f_1(t) * f_2(t) = 2^{\mu/2} t^{(\nu-\mu)/2} y^{\mu/2} (x-t)^{-(1+\nu)/2} \frac{\Gamma[-\mu/2]\Gamma[1+(\nu/2)]}{\Gamma[1+(\nu-\mu)/2]} \times F_1\left(-\frac{\mu}{2}, \frac{1+\nu}{2}, \frac{1-\mu}{2}; 1+\frac{\nu-\mu}{2}; \frac{t}{t-x}, -\frac{t}{y}\right)$$

The above Appell hypergeometric function can further be simplified into the Gaussian hypergeometric function given

$$F_1(a, b_1, b_2; b_1+b_2; z_1, z_2) = (1-z_2)^{-a} \, {}_2F_1\left(a, b_1; b_1+b_2; \frac{z_1-z_2}{1-z_2}\right)$$

see Equation (1) on p. 238 in [12]. The final expression for the convolution integral for $t < x$ then is

$$f_1(t) * f_2(t) = 2^{\mu/2} t^{(\nu-\mu)/2} (x-t)^{-(1+\nu)/2} (y+t)^{\mu/2} \frac{\Gamma[-\mu/2]\Gamma[1+(\nu/2)]}{\Gamma[1+(\nu-\mu)/2]} \times {}_2F_1\left(-\frac{\mu}{2}, \frac{1+\nu}{2}; 1+\frac{\nu-\mu}{2}; \frac{t(t+y-x)}{(t-x)(y+t)}\right) \qquad t < x \tag{3.7}$$

Second, the convolution integral for $t > x$ is given by

$$f_1(t) * f_2(t) = \int_{t-x}^{t} 2^{\mu/2} y^{1/2} \tau^{-\mu/2-1} (\tau+y)^{(\mu-1)/2} (t-\tau)^{\nu/2} (x-(t-\tau))^{-(1+\nu)/2} \, d\tau$$

The treatment of this convolution integral is similar to that of the integral for $t < x$ such that only the main steps are mentioned. The substitutions $\tau = s - x + t$ and $s = xu$ express the integral in terms of the Appell hypergeometric function $F_1(a, b_1, b_2; c; z_1, z_2)$ that again can be simplified into the Gaussian hypergeometric function. The convolution integral for $t > x$ then is given by

$$f_1(t) * f_2(t) = \frac{1}{\sqrt{\pi}} x^{1/2} y^{1/2} 2^{1+(\mu/2)} (t-x)^{-(1+\mu+\nu)/2} (y+t-x)^{(\mu-1)/2} t^{(\nu-1)/2} \times \Gamma[(1-\nu)/2]\Gamma[1+(\nu/2)] \, {}_2F_1\left(\frac{1-\mu}{2}, \frac{1-\nu}{2}; \frac{3}{2}; \frac{xy}{t(t+y-x)}\right) \qquad t > x \tag{3.8}$$

of which the derivation also used the following linear transformation formula

$${}_2F_1(a, b; c; z) = (1-z)^{-a} \, {}_2F_1\left(a, c-b; c; \frac{z}{z-1}\right)$$

see Equation (15.3.4) in [1].

Plugging (3.7) and (3.8) into the convolution integral (3.6) then gives

$$\exp\left(\tfrac{1}{2}py - px\right) D_\mu\left(2^{1/2} y^{1/2} p^{1/2}\right) \Phi\left(\frac{1-\nu}{2}; \frac{3}{2}; px\right) \tag{3.9}$$

$$
\begin{aligned}
&= \frac{2^{(\mu/2)-1}\sqrt{\pi}x^{-1/2}}{\Gamma\left[1+(\nu-\mu)/2\right]\Gamma\left[(1-\nu)/2\right]}\int_0^x \exp(-pt)\, t^{(\nu-\mu)/2}(x-t)^{-(1+\nu)/2} \\
&\quad \times (y+t)^{\mu/2}\,{}_2F_1\left(-\frac{\mu}{2},\frac{1+\nu}{2};1+\frac{\nu-\mu}{2};\frac{t(x-y-t)}{(x-t)(y+t)}\right)dt \\
&\quad + \frac{2^{\mu/2}y^{1/2}}{\Gamma\left[-\mu/2\right]}\int_x^\infty \exp(-pt)\, t^{(\nu-1)/2}(t-x)^{-(1+\mu+\nu)/2} \\
&\quad \times (y-x+t)^{(\mu-1)/2}\,{}_2F_1\left(\frac{1-\mu}{2},\frac{1-\nu}{2};\frac{3}{2};\frac{xy}{t(y-x+t)}\right)dt
\end{aligned}
$$

in which the recurrence and duplication formulas of the gamma function were employed to simplify expressions given that

$$
\Gamma\left[1+z\right] = z\Gamma\left[z\right], \quad \Gamma\left[2z\right] = \frac{1}{\sqrt{2\pi}}2^{2z-\frac{1}{2}}\Gamma\left[z\right]\Gamma\left[z+\tfrac{1}{2}\right],
$$

see Equations (6.1.15) and (6.1.18) in [1].

Finally, plugging the definition (2.2) into (3.9) and simplifying gives the inverse Laplace transform (3.1). □

The parabolic cylinder function specializes into the complementary error function when its order is at $-1$. The inverse Laplace transform (3.1) thus can be used to obtain an integral representation for the product of complementary error functions. However, this result will not be shown here as its integrand contains an inverse trigonometric function rather than the rational functions that are typical for existing integral representations, see for instance [16, 19]. Instead, the term $p^{-1/2}$ in inverse Laplace transforms such as (3.1) will be removed given that the resulting relations yield integrands in which such rational functions emerge. This will be illustrated in Theorem 3.2 and Corollary 3.3.

**Theorem 3.2.** *Let $\nu$ and $\mu$ be two complex numbers with $\operatorname{Re}\nu < 1$ and $\operatorname{Re}\mu < \min\left[1-\operatorname{Re}\nu, 2+\operatorname{Re}\nu\right]$. Then, the following inverse Laplace transform holds for $\operatorname{Re}p > 0$, $x > 0$, $|\arg y| < \pi$, $y > 0$*

$$
\begin{aligned}
&\exp\left(\tfrac{1}{2}p(y-x)\right)D_\mu\left(2^{1/2}y^{1/2}p^{1/2}\right)\left\{D_\nu\left(-2^{1/2}x^{1/2}p^{1/2}\right) - D_\nu\left(2^{1/2}x^{1/2}p^{1/2}\right)\right\} \qquad (3.10) \\
&= \frac{2^{(\mu-\nu)/2}\sqrt{\pi}y^{-1/2}}{\Gamma\left[(1-\mu+\nu)/2\right]\Gamma\left[-\nu\right]}\int_0^x \exp(-pt)\, t^{-(1+\mu-\nu)/2}(x-t)^{-(1+\nu)/2} \\
&\quad \times (y+t)^{(1+\mu)/2}\left\{{}_2F_1\left(-\frac{1+\mu}{2},\frac{1+\nu}{2};\frac{1-\mu+\nu}{2};\frac{t(x-y-t)}{(x-t)(y+t)}\right)\right. \\
&\quad \left. + \frac{\mu t}{(1-\mu+\nu)(y+t)}\,{}_2F_1\left(\frac{1-\mu}{2},\frac{1+\nu}{2};\frac{3-\mu+\nu}{2};\frac{t(x-y-t)}{(x-t)(y+t)}\right)\right\}dt \\
&\quad + \frac{2^{(4+\mu+\nu)/2}\sqrt{\pi}x^{1/2}}{\Gamma\left[-(1+\mu)/2\right]\Gamma\left[-\nu/2\right]}\int_x^\infty \exp(-pt)\, t^{(\nu-1)/2}(t-x)^{-(2+\mu+\nu)/2} \\
&\quad \times (y-x+t)^{\mu/2}\left\{{}_2F_1\left(-\frac{\mu}{2},\frac{1-\nu}{2};\frac{3}{2};\frac{xy}{t(y-x+t)}\right)\right. \\
&\quad \left. - \frac{\mu(t-x)}{(1+\mu)(y-x+t)}\,{}_2F_1\left(\frac{2-\mu}{2},\frac{1-\nu}{2};\frac{3}{2};\frac{xy}{t(y-x+t)}\right)\right\}dt
\end{aligned}
$$

***Proof.*** The recurrence relation of the parabolic cylinder function is given by

$$
zD_\mu(z) = D_{\mu+1}(z) + \mu D_{\mu-1}(z)
$$

see Equation (14) on p. 119 in [13]. Replacing $z$ by $2^{1/2}y^{1/2}p^{1/2}$ and multiplying by $p^{-1/2}\exp\left(\frac{1}{2}p(y-x)\right)\left\{D_\nu\left(-2^{1/2}x^{1/2}p^{1/2}\right) - D_\nu\left(2^{1/2}x^{1/2}p^{1/2}\right)\right\}$ gives

$$
2^{1/2}y^{1/2}\exp\left(\tfrac{1}{2}p(y-x)\right)D_\mu\left(2^{1/2}y^{1/2}p^{1/2}\right)\left\{D_\nu\left(-2^{1/2}x^{1/2}p^{1/2}\right) - D_\nu\left(2^{1/2}x^{1/2}p^{1/2}\right)\right\}
$$

$$= p^{-1/2}\exp\left(\tfrac{1}{2}p\left(y-x\right)\right)D_{\mu+1}\left(2^{1/2}y^{1/2}p^{1/2}\right)\left\{D_{\nu}\left(-2^{1/2}x^{1/2}p^{1/2}\right)\right. \tag{3.11}$$
$$\left.-D_{\nu}\left(2^{1/2}x^{1/2}p^{1/2}\right)\right\}+\mu p^{-1/2}\exp\left(\tfrac{1}{2}p\left(y-x\right)\right)D_{\mu-1}\left(2^{1/2}y^{1/2}p^{1/2}\right)$$
$$\times\left\{D_{\nu}\left(-2^{1/2}x^{1/2}p^{1/2}\right)-D_{\nu}\left(2^{1/2}x^{1/2}p^{1/2}\right)\right\}$$

Plugging the transform (3.1) into (3.11) and simplifying gives (3.10). □

**Corollary 3.3.** *The relation between the parabolic cylinder function and the complementary error function is given by*

$$D_{-1}\left(z\right)=\sqrt{\frac{\pi}{2}}\exp\left(\frac{z^2}{4}\right)\operatorname{erfc}\left(\frac{z}{\sqrt{2}}\right)$$

*see Equation (9.254.1) in* [16] *in which erfc($z$) denotes the complementary error function. Equations (E.3c) and (E.3d) in* [3] *specify the following relations between the error function and its complement*

$$\operatorname{erfc}\left(z\right)+\operatorname{erf}\left(z\right)=1$$
$$\operatorname{erfc}\left(-z\right)=1+\operatorname{erf}\left(z\right)$$

*and thus*

$$\operatorname{erfc}\left(-z\right)-\operatorname{erfc}\left(z\right)=2\operatorname{erf}\left(z\right) \tag{3.12}$$

*where erf($z$) denotes the error function. The below derivations also use the following properties of the Gaussian hypergeometric function*

$${}_2F_1\left(0,b;c;z\right)={}_2F_1\left(a,0;c;z\right)=1$$
$${}_2F_1\left(1,\frac{3}{2};\frac{3}{2};z\right)=\frac{1}{1-z}$$

*see Equations (15.1.1) and (15.1.8) in* [1]. *Plugging the transform (3.1) into (3.11), using $\mu=\nu=-1$ and (3.12) gives the following inverse Laplace transform for the product of two (complementary) error functions*

$$\exp\left(py\right)\ \operatorname{erfc}\left(y^{1/2}p^{1/2}\right)\ \operatorname{erf}\left(x^{1/2}p^{1/2}\right)= \tag{3.13}$$
$$\frac{1}{\pi}\int_0^x\exp\left(-pt\right)\frac{\sqrt{y}}{\sqrt{t}\left(y+t\right)}dt-\frac{1}{\pi}\int_x^\infty\exp\left(-pt\right)\frac{\sqrt{x}}{\sqrt{y-x+t}\left(y+t\right)}dt$$
$$\left[\operatorname{Re}p>0,\left|\arg y\right|<\pi,y>0,\left|\arg x\right|<\pi,x\geqslant 0\right]$$

*Using $p=1$ and setting $a$ and $b$ at $y^{1/2}$ and $x^{1/2}$, respectively, then gives the following integral representation*

$$\operatorname{erfc}\left(a\right)\ \operatorname{erf}\left(b\right)= \tag{3.14}$$
$$\frac{a\exp\left(-a^2\right)}{\pi}\int_0^{b^2}\frac{\exp\left(-t\right)}{\left(t+a^2\right)\sqrt{t}}dt-\frac{b\exp\left(-\left(a^2+b^2\right)\right)}{\pi}\int_0^\infty\frac{\exp\left(-t\right)}{\left(t+a^2+b^2\right)\sqrt{t+a^2}}dt$$
$$\left[\operatorname{Re}a>0,\operatorname{Re}b\geqslant 0\right]$$

*which is not present in, for instance, the extensive overview in* [19].

**Theorem 3.4.** *Let $\nu$ and $\mu$ be two complex numbers with $\operatorname{Re}\nu<1$ and $\operatorname{Re}\mu<\min\left[1-\operatorname{Re}\nu,\ 2+\operatorname{Re}\nu\right]$. Then, the following inverse Laplace transform holds for $\operatorname{Re}p>0$, $\left|\arg x\right|<\pi$, $x\geqslant 0$, $\left|\arg y\right|<\pi$, $y>0$*

$$p^{-1/2}\exp\left(\tfrac{1}{2}p\left(y-x\right)\right)D_{\mu}\left(2^{1/2}y^{1/2}p^{1/2}\right)\left\{D_{\nu}\left(-2^{1/2}x^{1/2}p^{1/2}\right)+D_{\nu}\left(2^{1/2}x^{1/2}p^{1/2}\right)\right\}$$
$$=\frac{2^{(\mu-\nu)/2}\sqrt{\pi}}{\Gamma\left[1+\left(\nu-\mu\right)/2\right]\Gamma\left[-\nu\right]}\int_0^x\exp\left(-pt\right)t^{(\nu-\mu)/2}\left(x-t\right)^{-(1+\nu)/2} \tag{3.15}$$

$$\times (y+t)^{\mu/2} {}_2F_1\left(-\frac{\mu}{2}, \frac{1+\nu}{2}; 1+\frac{\nu-\mu}{2}; \frac{t(x-y-t)}{(x-t)(y+t)}\right) dt$$
$$+ \frac{2^{1+(\mu+\nu)/2}\sqrt{\pi}}{\Gamma[(1-\nu)/2]\,\Gamma[(1-\mu)/2]} \int_x^\infty \exp(-pt)\, t^{\nu/2} (t-x)^{-(1+\mu+\nu)/2}$$
$$\times (y-x+t)^{\mu/2} {}_2F_1\left(-\frac{\mu}{2}, -\frac{\nu}{2}; \frac{1}{2}; \frac{xy}{t(y-x+t)}\right) dt$$

***Proof.*** The inverse Laplace transform in Equation (6) on p. 290 in [14] is

$$\Gamma[\nu]\, p^{-1/2} \exp\left(\tfrac{1}{2}ap\right) D_{1-2\nu}\left(2^{1/2}a^{1/2}p^{1/2}\right) = \int_0^\infty \exp(-pt)\, 2^{1/2-\nu} t^{\nu-1} (t+a)^{1/2-\nu}\, dt$$
$$[\operatorname{Re} p > 0, \operatorname{Re}\nu > 0, |\arg a| < \pi]$$

which in the notation of Theorem 3.3 gives

$$\Gamma[(1-\mu)/2]\, p^{-1/2} \exp\left(\tfrac{1}{2}yp\right) D_\mu\left(2^{1/2}y^{1/2}p^{1/2}\right)$$
$$= \int_0^\infty \exp(-pt)\, 2^{\mu/2} t^{-(\mu+1)/2} (t+y)^{\mu/2}\, dt$$
$$[\operatorname{Re} p > 0, \operatorname{Re}\mu < 1, |\arg y| < \pi] \tag{3.16}$$

The inverse Laplace transform (3.3) is specialized for $a = -\frac{\nu}{2}$ and $b = \frac{1}{2}$ and gives

$$\frac{x^{-1/2}}{\sqrt{\pi}} \Gamma[(1+\nu)/2]\, \Gamma[-\nu/2] \exp(-xp)\, \Phi\left(-\frac{\nu}{2}; \frac{1}{2}; xp\right)$$
$$= \int_0^x \exp(-pt)\, t^{(\nu-1)/2} (x-t)^{-(\nu/2)-1}\, dt$$
$$[\operatorname{Re} p > 0, -1 < \operatorname{Re}\nu < 0, x > 0] \tag{3.17}$$

The original functions $f_1(t)$ and $f_2(t)$ are taken from the inverse Laplace transforms (3.16) and (3.17), respectively

$$f_1(t) = 2^{\mu/2} t^{-(\mu+1)/2} (t+y)^{\mu/2} \quad \text{and} \quad f_2(t) = t^{(\nu-1)/2} (x-t)^{-(\nu/2)-1}$$

Using steps akin to those used in the proof of Theorem 3.1 then yields

$$p^{-1/2} \exp\left(\tfrac{1}{2}py - px\right) D_\mu\left(2^{1/2}y^{1/2}p^{1/2}\right) \Phi\left(-\frac{\nu}{2}; \frac{1}{2}; px\right) \tag{3.18}$$
$$= \frac{2^{\mu/2}\sqrt{\pi} x^{1/2} y^{1/2}}{\Gamma[1+(\nu-\mu)/2]\, \Gamma[-\nu/2]} \int_0^x \exp(-pt)\, t^{(\nu-\mu)/2} (x-t)^{-1-(\nu/2)}$$
$$\times (y+t)^{(\mu-1)/2} {}_2F_1\left(\frac{1-\mu}{2}, 1+\frac{\nu}{2}; 1+\frac{\nu-\mu}{2}; \frac{t(x-y-t)}{(x-t)(y+t)}\right) dt$$
$$+ \frac{2^{\mu/2}}{\Gamma[(1-\mu)/2]} \int_x^\infty \exp(-pt)\, t^{\nu/2} (t-x)^{-(1+\mu+\nu)/2}$$
$$\times (y-x+t)^{\mu/2} {}_2F_1\left(-\frac{\mu}{2}, -\frac{\nu}{2}; \frac{1}{2}; \frac{xy}{t(y-x+t)}\right) dt$$

The first integral in (3.18) can be rewritten via the following linear transformation formula for the Gaussian hypergeometric function

$${}_2F_1(a, b; c; z) = (1-z)^{c-a-b} {}_2F_1(c-a, c-b; c; z) \tag{3.19}$$

see Equation (15.3.3) in [1]. Combining the resulting expression for the transform (3.18) with the definition (2.3) then gives the inverse Laplace transform (3.15). □

Theorem 3.5 specifies the inverse Laplace transform for the product of two parabolic cylinder functions of which the arguments have opposite sign and Corollary 3.6 specializes this expression for a single parabolic cylinder function with negative sign in the argument.

**Theorem 3.5.** *Let $\nu$ and $\mu$ be two complex numbers with $\operatorname{Re}\nu < 1$ and $\operatorname{Re}\mu < \min[1-\operatorname{Re}\nu, 2+\operatorname{Re}\nu]$. Then, the following inverse Laplace transform holds for $\operatorname{Re}p > 0$, $x > 0$, $|\arg y| < \pi$, $y > 0$*

$$\begin{aligned}
&p^{-1/2}\exp\left(\tfrac{1}{2}p\,(y-x)\right)D_\mu\left(2^{1/2}y^{1/2}p^{1/2}\right)D_\nu\left(-2^{1/2}x^{1/2}p^{1/2}\right) \qquad (3.20)\\
&=\frac{2^{(\mu-\nu)/2}\sqrt{\pi}}{\Gamma\left[1+(\nu-\mu)/2\right]\Gamma\left[-\nu\right]}\int_0^x \exp(-pt)\,t^{(\nu-\mu)/2}(x-t)^{-(1+\nu)/2}\\
&\quad\times(y+t)^{\mu/2}\,{}_2F_1\left(-\frac{\mu}{2},\frac{1+\nu}{2};1+\frac{\nu-\mu}{2};\frac{t\,(x-y-t)}{(x-t)\,(y+t)}\right)dt\\
&\quad+\frac{2^{1+(\mu+\nu)/2}\sqrt{\pi}x^{1/2}y^{1/2}}{\Gamma\left[-\mu/2\right]\Gamma\left[-\nu/2\right]}\int_x^\infty \exp(-pt)\,t^{(\nu-1)/2}(t-x)^{-(1+\mu+\nu)/2}\\
&\quad\times(y-x+t)^{(\mu-1)/2}\left\{{}_2F_1\left(\frac{1-\mu}{2},\frac{1-\nu}{2};\frac{3}{2};\frac{xy}{t\,(y-x+t)}\right)\right.\\
&\quad\left.+\frac{\Gamma\left[-\mu/2\right]\Gamma\left[-\nu/2\right]}{\Gamma\left[(1-\mu)/2\right]\Gamma\left[(1-\nu)/2\right]}\left(\frac{t\,(y-x+t)}{4xy}\right)^{1/2}{}_2F_1\left(-\frac{\mu}{2},-\frac{\nu}{2};\frac{1}{2};\frac{xy}{t\,(y-x+t)}\right)\right\}dt
\end{aligned}$$

***Proof.*** The transform (3.20) is obtained by adding the inverse Laplace transforms (3.1) and (3.15) and simplifying the resulting expression. □

**Corollary 3.6.** *Using $y = 0$, the properties*

$$D_\mu(0) = \frac{2^{\mu/2}\sqrt{\pi}}{\Gamma\left[(1-\mu)/2\right]}$$

$${}_2F_1(a,b;c;1) = \frac{\Gamma[c]\,\Gamma[c-a-b]}{\Gamma[c-a]\,\Gamma[c-b]}$$

*see Equations (46:7:1) in [22] and (15.1.20) in [1], and $\mu = 0$ gives*

$$\begin{aligned}
&p^{-1/2}\exp\left(-\tfrac{1}{2}px\right)D_v\left(-2^{1/2}x^{1/2}p^{1/2}\right) = \qquad (3.21)\\
&\frac{2^{-\nu/2}\sqrt{\pi}}{\Gamma[-\nu]\,\Gamma[1+\nu/2]}\int_0^x \exp(-pt)\,t^{\nu/2}(x-t)^{-(1+\nu)/2}dt\\
&+\frac{2^{\nu/2}}{\Gamma\left[(1-\nu)/2\right]}\int_x^\infty \exp(-pt)\,t^{\nu/2}(t-x)^{-(1+\nu)/2}dt\\
&[\operatorname{Re}p > 0, \operatorname{Re}\nu < 1, x > 0]
\end{aligned}$$

**Theorem 3.7.** *Let $\nu$ and $\mu$ be two complex numbers with $\operatorname{Re}(\nu+\mu) < 1$. Then, the following inverse Laplace transform holds for $\operatorname{Re}p > 0$, $|\arg x| < \pi$, $x \geqslant 0$, $|\arg y| < \pi$, $y \geqslant 0$, $|\arg x + \arg y| < \pi$*

$$\begin{aligned}
&p^{-1/2}\exp\left(\tfrac{1}{2}p\,(y+x)\right)D_\mu\left(2^{1/2}y^{1/2}p^{1/2}\right)D_\nu\left(2^{1/2}x^{1/2}p^{1/2}\right) = \qquad (3.22)\\
&\frac{2^{(\mu+\nu)/2}}{\Gamma\left[(1-\mu-\nu)/2\right]}\int_0^\infty \exp(-pt)\,t^{-(1+\mu+\nu)/2}(y+t)^{\mu/2}(x+t)^{\nu/2}\\
&\times{}_2F_1\left(-\frac{\mu}{2},-\frac{\nu}{2};\frac{1-\mu-\nu}{2};\frac{t\,(x+y+t)}{(x+t)\,(y+t)}\right)dt
\end{aligned}$$

*which is identical to the transform in Equation (2.1) in [28].*

***Proof.*** Subtracting the inverse Laplace transform (3.1) from (3.10) gives

$$\begin{aligned}
&p^{-1/2}\exp\left(\tfrac{1}{2}p\,(y-x)\right)D_\mu\left(2^{1/2}y^{1/2}p^{1/2}\right)D_\nu\left(2^{1/2}x^{1/2}p^{1/2}\right) =\\
&+\frac{2^{(\mu+\nu)/2}}{\Gamma\left[(1-\mu-\nu)/2\right]}\int_x^\infty \exp(-pt)\,t^{\nu/2}(t-x)^{-(1+\mu+\nu)/2}
\end{aligned}$$

$$\times (y-x+t)^{\mu/2}\left\{\frac{\sqrt{\pi}\Gamma\left[(1-\mu-\nu)/2\right]}{\Gamma\left[(1-\mu)/2\right]\Gamma\left[(1-\nu)/2\right]}{}_2F_1\left(-\frac{\mu}{2},-\frac{\nu}{2};\frac{1}{2};\frac{xy}{t(y-x+t)}\right)\right.$$
$$\left.-\frac{\sqrt{\pi}\Gamma\left[(1-\mu-\nu)/2\right]}{\Gamma\left[-\mu/2\right]\Gamma\left[-\nu/2\right]}\left(\frac{4xy}{t(y-x+t)}\right)^{1/2}{}_2F_1\left(\frac{1-\mu}{2},\frac{1-\nu}{2};\frac{3}{2};\frac{xy}{t(y-x+t)}\right)\right\}dt$$

in which the linear transformation formula (3.19) was used. Subsequently, using the linear transformation formula

$$\begin{aligned}{}_2F_1(a,b;c;z) &= \frac{\Gamma[c]\Gamma[c-a-b]}{\Gamma[c-a]\Gamma[c-b]}{}_2F_1(a,b;a+b-c+1;1-z)\\ &+(1-z)^{c-a-b}\frac{\Gamma[c]\Gamma[a+b-c]}{\Gamma[a]\Gamma[b]}{}_2F_1(c-a,c-b;c-a-b+1;1-z)\end{aligned}$$

in Equation (15.3.6) in [1] gives

$$\begin{aligned}&p^{-1/2}\exp\left(\tfrac{1}{2}p(y-x)\right)D_\mu\left(2^{1/2}y^{1/2}p^{1/2}\right)D_\nu\left(2^{1/2}x^{1/2}p^{1/2}\right)=\\ &\frac{2^{(\nu+\mu)/2}}{\Gamma\left[(1-\mu-\nu]/2\right)}\int_x^\infty \exp(-pt)\,t^{\nu/2}(t-x)^{-(1+\mu+\nu)/2}(y-x+t)^{\mu/2}\\ &\quad\times{}_2F_1\left(-\frac{\mu}{2},-\frac{\nu}{2};\frac{1-\mu-\nu}{2};\frac{(t-x)(y+t)}{t(y-x+t)}\right)dt\end{aligned}$$

Multiplying both sides by $\exp(px)$, using the substitution $s=t-x$ and subsequently re-introducing $t$ then gives (3.22). □

As noted earlier, removing the term $p^{-1/2}$ from transforms such as (3.22) allows obtaining integral representations for (complementary) error functions in which the integrand contains rational functions. This is illustrated in Theorem 3.8 and Corollary 3.9 in which the integral representation for $1-\mathrm{erf}(a)^2$ in [19] is generalized into $1-\mathrm{erf}(a)\mathrm{erf}(b)$.

**Theorem 3.8.** *Let $\nu$ and $\mu$ be two complex numbers with $\mathrm{Re}(\nu+\mu)<1$. Then, the following inverse Laplace transform holds for $\mathrm{Re}\,p>0$, $|\arg x|<\pi$, $x>0$, $|\arg y|<\pi$, $y>0$, $|\arg x+\arg y|<\pi$*

$$\begin{aligned}&\exp\left(\tfrac{1}{2}p(y+x)\right)D_\mu\left(2^{1/2}y^{1/2}p^{1/2}\right)D_\nu\left(2^{1/2}x^{1/2}p^{1/2}\right)=\\ &\frac{2^{(\mu+\nu)/2}x^{-1/2}}{\Gamma\left[-(\mu+\nu)/2\right]}\int_0^\infty \exp(-pt)\,t^{-1-(\nu+\mu)/2}(y+t)^{\mu/2}\\ &\quad\times(x+t)^{(1+\nu)/2}\left\{{}_2F_1\left(-\frac{\mu}{2},-\frac{1+\nu}{2};-\frac{\mu+\nu}{2};\frac{t(x+y+t)}{(x+t)(y+t)}\right)\right.\\ &\quad\left.-\frac{\nu t}{(\mu+\nu)(x+t)}{}_2F_1\left(-\frac{\mu}{2},\frac{1-\nu}{2};1-\frac{\mu+\nu}{2};\frac{t(x+y+t)}{(x+t)(y+t)}\right)\right\}dt\end{aligned}\tag{3.23}$$

***Proof.*** The inverse Laplace transform (3.23) is obtained via the above recurrence relation of the parabolic cylinder function. Replacing $z$ by $2^{1/2}x^{1/2}p^{1/2}$ in the recurrence relation and multiplying by $p^{-1/2}\exp\left(\frac{1}{2}p(y+x)\right)D_\mu\left(2^{1/2}y^{1/2}p^{1/2}\right)$ gives

$$\begin{aligned}&\exp\left(\tfrac{1}{2}p(y+x)\right)D_\mu\left(2^{1/2}y^{1/2}p^{1/2}\right)D_\nu\left(2^{1/2}x^{1/2}p^{1/2}\right)=\\ &\quad 2^{-1/2}x^{-1/2}p^{-1/2}\exp\left(\tfrac{1}{2}p(y+x)\right)D_\mu\left(2^{1/2}y^{1/2}p^{1/2}\right)D_{\nu+1}\left(2^{1/2}x^{1/2}p^{1/2}\right)\\ &\quad+\nu 2^{-1/2}x^{-1/2}p^{-1/2}\exp\left(\tfrac{1}{2}p(y+x)\right)D_\mu\left(2^{1/2}y^{1/2}p^{1/2}\right)D_{\nu-1}\left(2^{1/2}x^{1/2}p^{1/2}\right)\end{aligned}$$

Plugging the transform (3.22) into the latter expression and simplifying the result via the linear transformation formula (3.19) gives (3.23). □

**Corollary 3.9.** *The below derivations employ the following property of the Gaussian hypergeometric function*

$$ {}_2F_1\left(1,\frac{1}{2};2;z\right) = {}_2F_1\left(\frac{1}{2},1;2;z\right) = \frac{2}{1+\sqrt{1-z}} $$

*see Equation (84) on p. 473 in* [24]*. Using $\mu = \nu = -1$ in (3.23) gives the following inverse Laplace transform for the product of two complementary error functions*

$$ \exp\left(p\left(x+y\right)\right)\ \operatorname{erfc}\left(y^{1/2}p^{1/2}\right)\ \operatorname{erfc}\left(x^{1/2}p^{1/2}\right) = \tag{3.24} $$
$$ \frac{1}{\pi}\int_0^\infty \exp\left(-pt\right)\frac{\sqrt{x}\sqrt{x+t}+\sqrt{y}\sqrt{y+t}}{\left(x+y+t\right)\sqrt{\left(x+t\right)\left(y+t\right)}}dt $$
$$ [\operatorname{Re} p > 0, |\arg y| < \pi, y \geqslant 0, |\arg x| < \pi, x \geqslant 0, |\arg x + \arg y| < \pi] $$

*Using $p = 1$, $y^{1/2} = a$ and $x^{1/2} = b$ then gives the following integral representation for the product of two complementary error functions*

$$ \operatorname{erfc}\left(a\right)\ \operatorname{erfc}\left(b\right) = \tag{3.25} $$
$$ \frac{1}{\pi}\exp\left(-\left(a^2+b^2\right)\right)\int_0^\infty \exp\left(-t\right)\frac{a\sqrt{t+a^2}+b\sqrt{t+b^2}}{\left(t+a^2+b^2\right)\sqrt{\left(t+a^2\right)\left(t+b^2\right)}}dt $$
$$ [\operatorname{Re} a > 0, \operatorname{Re} b > 0] $$

*which gives an alternative to the representation given on p. 70 in* [27]*. Using $a = 0$ and* $\operatorname{erfc}(0) = 1$*, see Equation (40:7) in* [22]*, gives*

$$ \operatorname{erfc}\left(b\right) = \frac{b}{\pi}\exp\left(-b^2\right)\int_0^\infty \frac{\exp\left(-t\right)}{\left(t+b^2\right)\sqrt{t}}dt $$
$$ [\operatorname{Re} b > 0] $$
$$ \operatorname{erf}\left(b\right) = 1 - \frac{b}{\pi}\exp\left(-b^2\right)\int_0^\infty \frac{\exp\left(-t\right)}{\left(t+b^2\right)\sqrt{t}}dt \tag{3.26} $$
$$ [\operatorname{Re} b > 0] $$

*The definition of the complementary error function gives* $\operatorname{erf}(a)\operatorname{erf}(b) = \operatorname{erf}(b) - \operatorname{erfc}(a)\operatorname{erf}(b)$ *such that plugging (3.26) and (3.14) into the latter relation gives*

$$ 1-\operatorname{erf}(a)\operatorname{erf}(b) = \tag{3.27} $$
$$ \frac{b}{\pi}\exp\left(-b^2\right)\int_0^\infty \exp\left(-t\right)\left\{\frac{1}{\left(t+b^2\right)\sqrt{t}} - \frac{\exp\left(-a^2\right)}{\left(t+a^2+b^2\right)\sqrt{t+a^2}}\right\}dt $$
$$ +\frac{a}{\pi}\exp\left(-a^2\right)\int_0^{b^2}\frac{\exp\left(-t\right)}{\left(t+a^2\right)\sqrt{t}}dt $$
$$ [\operatorname{Re} a > 0, \operatorname{Re} b > 0] $$

*which generalizes the expression for 1–erf$(a)^2$ in Equation (8) on p. 4 in* [19] *to differing arguments. Note that the representation in* [19] *can easily be obtained from (3.27) by using $a = b$ which gives*

$$ 1-\operatorname{erf}(a)^2 = \frac{2a}{\pi}\exp\left(-a^2\right)\int_0^{a^2}\frac{\exp\left(-t\right)}{\left(t+a^2\right)\sqrt{t}}dt $$

*The substitution $t = a^2s^2$ then gives*

$$ 1-\operatorname{erf}(a)^2 = \frac{4}{\pi}\exp\left(-a^2\right)\int_0^1\frac{\exp\left(-a^2s^2\right)}{\left(s^2+1\right)}ds $$

*which is the integral representation in* [19].

## 4. Correcting two inverse Laplace transforms

This Section utilizes the above results to correct two inverse Laplace transforms that are frequently found.

### 4.1. First correction

The following inverse Laplace transform is specified in Equation (3.11.4.3) in [26]

$$D_{\nu}\left(a\sqrt{p}\right)D_{-\nu-1}\left(a\sqrt{p}\right)=\int_{a}^{\infty}\exp\left(-pt\right)\frac{\left(t^{2}-a^{2}\right)^{-1/2}}{\sqrt{2t}}\cos\left[\left(\nu+\frac{1}{2}\right)\arccos\left[\frac{a^{2}}{2t}\right]\right]dt\quad **$$

where $**$ indicates that the expression is not correct. The corrected expression, however, can easily be obtained from the results in Section 3.

**Theorem 4.1.** *Let $\nu$ be a complex number. Then, the following inverse Laplace transform holds for* $\operatorname{Re}p>0$ *and* $\operatorname{Re}a>0$

$$D_{\nu}\left(a\sqrt{p}\right)D_{-\nu-1}\left(a\sqrt{p}\right)=\int_{\frac{1}{2}a^{2}}^{\infty}\exp\left(-pt\right)\frac{a\left(t^{2}-\frac{a^{4}}{4}\right)^{-1/2}}{\sqrt{2\pi t}}\cos\left[\left(2\nu+1\right)\arcsin\left[\sqrt{\frac{2t-a^{2}}{4t}}\right]\right]dt \tag{4.1}$$

***Proof.*** Using $a=2^{1/2}x^{1/2}=2^{1/2}y^{1/2}$ and $\mu=-\nu-1$ allows to rewrite (3.23) as follows

$$\exp\left(\tfrac{1}{2}a^{2}p\right)D_{\nu}\left(a\sqrt{p}\right)D_{-\nu-1}\left(a\sqrt{p}\right)=\frac{1}{a\sqrt{\pi}}\int_{0}^{\infty}\exp\left(-pt\right)t^{-1/2}\left\{{}_{2}F_{1}\left(-\frac{1+\nu}{2},\frac{1+\nu}{2};\frac{1}{2};\frac{4t\left(a^{2}+t\right)}{\left(a^{2}+2t\right)^{2}}\right)\right.$$
$$\left.+\frac{2\nu t}{a^{2}+2t}{}_{2}F_{1}\left(\frac{1-\nu}{2},\frac{1+\nu}{2};\frac{3}{2};\frac{4t\left(a^{2}+t\right)}{\left(a^{2}+2t\right)^{2}}\right)\right\}dt$$

Multiplying both sides by $\exp\left(-\frac{1}{2}a^{2}p\right)$, using the substitution $s=t+\frac{1}{2}a^{2}$ and subsequently re-introducing $t$ gives

$$D_{\nu}\left(a\sqrt{p}\right)D_{-\nu-1}\left(a\sqrt{p}\right)=\frac{2^{1/2}}{a\sqrt{\pi}}\int_{\frac{1}{2}a^{2}}^{\infty}\exp\left(-pt\right)\left(2t-a^{2}\right)^{-1/2}\left\{{}_{2}F_{1}\left(-\frac{1+\nu}{2},\frac{1+\nu}{2};\frac{1}{2};\frac{4t^{2}-a^{4}}{4t^{2}}\right)\right.$$
$$\left.+\frac{\nu\left(2t-a^{2}\right)}{2t}{}_{2}F_{1}\left(\frac{1-\nu}{2},\frac{1+\nu}{2};\frac{3}{2};\frac{4t^{2}-a^{4}}{4t^{2}}\right)\right\}dt$$

The quadratic transformation formula in Equation (15.3.22) in [1] states

$${}_{2}F_{1}\left(a,b;a+b+\frac{1}{2};z\right)={}_{2}F_{1}\left(2a,2b;a+b+\frac{1}{2};\frac{1}{2}-\frac{1}{2}\sqrt{1-z}\right)$$

Using the latter relation gives

$$D_{\nu}\left(a\sqrt{p}\right)D_{-\nu-1}\left(a\sqrt{p}\right)=\frac{2^{1/2}}{a\sqrt{\pi}}\int_{\frac{1}{2}a^{2}}^{\infty}\exp\left(-pt\right)\left(2t-a^{2}\right)^{-1/2}\left\{{}_{2}F_{1}\left(-1-\nu,1+\nu;\frac{1}{2};\frac{2t-a^{2}}{4t}\right)\right.$$
$$\left.+\frac{\nu\left(2t-a^{2}\right)}{2t}{}_{2}F_{1}\left(1-\nu,1+\nu;\frac{3}{2};\frac{2t-a^{2}}{4t}\right)\right\}dt$$

The latter result can be simplified on the basis of the relations (15.2.10) and (15.2.20) in [1], respectively

$$(c-a)\,{}_2F_1(a-1,b;c;z)+(2a-c-az+bz)\,{}_2F_1(a,b;c;z)$$
$$+a(z-1)\,{}_2F_1(a+1,b;c;z)=0$$
$$c(1-z)\,{}_2F_1(a,b;c;z)-c\,{}_2F_1(a-1,b;c;z)+(c-b)\,z\,{}_2F_1(a,b;c+1;z)=0$$

The latter two relations can be combined into

$$\left(ac-c^2\right){}_2F_1(a-1,b;c;z)+\left(c^2-ac+c(a-b)\,z\right){}_2F_1(a,b;c;z)$$
$$+a(b-c)\,z\,{}_2F_1(a+1,b;c+1;z)=0$$

which gives

$$\frac{a^2}{2t}\,{}_2F_1\left(1+\nu,-\nu;\frac{1}{2};\frac{2t-a^2}{4t}\right)={}_2F_1\left(-1-\nu,1+\nu;\frac{1}{2};\frac{2t-a^2}{4t}\right)$$
$$+\frac{\nu\left(2t-a^2\right)}{2t}\,{}_2F_1\left(1-\nu,1+\nu;\frac{3}{2};\frac{2t-a^2}{4t}\right)$$

This allows to rewrite the inverse Laplace transform as

$$D_\nu\left(a\sqrt{p}\right)D_{-\nu-1}\left(a\sqrt{p}\right)=$$
$$\frac{a}{\sqrt{2\pi}}\int_{\frac{1}{2}a^2}^{\infty}\exp(-pt)\,\frac{\left(2t-a^2\right)^{-1/2}}{t}\,{}_2F_1\left(1+\nu,-\nu;\frac{1}{2};\frac{2t-a^2}{4t}\right)dt$$

Equation (90) on p. 460 in [24] states

$${}_2F_1\left(a,1-a;\frac{1}{2};z\right)={}_2F_1\left(1-a,a;\frac{1}{2};z\right)=\frac{1}{\sqrt{1-z}}\cos\left[(2a-1)\arcsin\left[\sqrt{z}\right]\right]$$

Employing the latter property then gives (4.1). □

## 4.2. Second correction

The following inverse Laplace transform can be found in Equation (11) on p. 218 in [14], in Equation (16.7) on p. 379 in [21] and in Equation (3.11.5.1) in [26]

$$\exp\left(\tfrac{1}{4}a^2p^2\right)D_\mu(ap)\,D_\nu(ap)=$$
$$\frac{1}{\Gamma[-\mu-\nu]}\int_0^\infty\exp(-pt)\,a^{\mu+\nu}t^{-(1+\mu+\nu)}\exp\left(-\frac{t^2}{2a^2}\right)$$
$$\times\,{}_2F_2\left(-\mu,-\nu;-\frac{\mu+\nu}{2},\frac{1-\mu-\nu}{2};\frac{t^2}{4a^2}\right)dt\qquad **$$

**Theorem 4.2.** *Let $\nu$ and $\mu$ be two complex numbers with $\operatorname{Re}(\mu+\nu)<0$. Then, the following inverse Laplace transform holds for $\operatorname{Re}p>0$ and $\operatorname{Re}a>0$*

$$\exp\left(\tfrac{1}{2}a^2p^2\right)D_\mu(ap)\,D_\nu(ap)=\tag{4.2}$$
$$\frac{1}{\Gamma[-\mu-\nu]}\int_0^\infty\exp(-pt)\,a^{\mu+\nu}t^{-(1+\mu+\nu)}\exp\left(-\frac{t^2}{2a^2}\right)$$
$$\times\,{}_2F_2\left(-\mu,-\nu;-\frac{\mu+\nu}{2},\frac{1-\mu-\nu}{2};\frac{t^2}{4a^2}\right)dt$$

***Proof.*** From the specification of, for instance, the inverse Laplace transform (3.23), it is clear that the left-hand side of the expression in [14, 21, 26] contains a misprint as the exponential term should be $\exp\left(\frac{1}{2}a^2p^2\right)$ rather than $\exp\left(\frac{1}{4}a^2p^2\right)$. □

## 5. Two new definite integrals for the generalized hypergeometric function

The below definite integrals for the generalized hypergeometric function are derived from the inverse Laplace transform (4.2) in combination with two results from Section 3.

### 5.1. First integral

Using $a = 2^{1/2}x^{1/2}$ in (4.2) gives

$$\begin{aligned}\exp\left(p^2x\right) D_\mu\left(2^{1/2}x^{1/2}p\right) D_\nu\left(2^{1/2}x^{1/2}p\right) = \\ \frac{(2x)^{(\mu+\nu)/2}}{\Gamma[-\mu-\nu]}\int_0^\infty \exp(-pt)\, t^{-(1+\mu+\nu)}\exp\left(-\frac{t^2}{4x}\right) \\ \times {}_2F_2\left(-\mu,-\nu;-\frac{\mu+\nu}{2},\frac{1-\mu-\nu}{2};\frac{t^2}{8x}\right)dt\end{aligned} \tag{5.1}$$

and the inverse Laplace transform (3.23) for $y = x$ is

$$\begin{aligned}\exp(px) D_\mu\left(2^{1/2}x^{1/2}p^{1/2}\right) D_\nu\left(2^{1/2}x^{1/2}p^{1/2}\right) = \\ \frac{2^{(\mu+\nu)/2}x^{-1/2}}{\Gamma[-(\mu+\nu)/2]}\int_0^\infty \exp(-pt)\, t^{-1-(\nu+\mu)/2}(x+t)^{(1+\mu+\nu)/2} \\ \times\left\{{}_2F_1\left(-\frac{\mu}{2},-\frac{1+\nu}{2};-\frac{\mu+\nu}{2};\frac{t(2x+t)}{(x+t)^2}\right)\right. \\ \left.-\frac{\nu t}{(\mu+\nu)(x+t)}{}_2F_1\left(-\frac{\mu}{2},\frac{1-\nu}{2};1-\frac{\mu+\nu}{2};\frac{t(2x+t)}{(x+t)^2}\right)\right\}dt\end{aligned} \tag{5.2}$$

Let $f(t)$ be the original function in the Laplace transform (5.1) and $F(p)$ be the corresponding image function. Equation (26) on p. 4 of [26] states that the original function of the image function $F\left(p^{1/2}\right)$ then is related to $f(t)$ as follows

$$\frac{1}{2\sqrt{\pi t^3}}\int_0^\infty \tau\exp\left(-\frac{\tau^2}{4t}\right)f(\tau)\,d\tau \tag{5.3}$$

Hence, plugging the original function for the inverse Laplace transform (5.1) into the expression (5.3) gives the original function of expression (5.2). Straightforward simplifications and redefinitions of variables then give the following definite integral for the generalized hypergeometric function

$$\begin{aligned}\int_0^\infty t^{-(\mu+\nu)}\exp\left(-\frac{x+y}{4xy}t^2\right){}_2F_2\left(-\mu,-\nu;-\frac{\mu+\nu}{2},\frac{1-\mu-\nu}{2};\frac{t^2}{8x}\right)dt = \\ 2^{-(\mu+\nu)}\Gamma\left[\frac{1-\mu-\nu}{2}\right]y\left(\frac{x+y}{xy}\right)^{(1+\mu+\nu)/2}\left\{{}_2F_1\left(-\frac{\mu}{2},-\frac{1+\nu}{2};-\frac{\mu+\nu}{2};\frac{y(2x+y)}{(x+y)^2}\right)\right. \\ \left.-\frac{\nu y}{(\mu+\nu)(x+y)}{}_2F_1\left(-\frac{\mu}{2},\frac{1-\nu}{2};1-\frac{\mu+\nu}{2};\frac{y(2x+y)}{(x+y)^2}\right)\right\} \\ [\operatorname{Re}(\mu+\nu)<1,\operatorname{Re}x>0,\operatorname{Re}y>0]\end{aligned} \tag{5.4}$$

### 5.2. Second integral

Again, let $f(t)$ be the original function in the Laplace transform (5.1) and $F(p)$ be the corresponding image function. Equation (29) on p. 5 of [26] states that the original function of the image function $p^{-1/2}F\left(p^{1/2}\right)$ is given by

$$\frac{1}{\sqrt{\pi t}}\int_0^\infty \exp\left(-\frac{\tau^2}{4t}\right) f(\tau)\, d\tau \tag{5.5}$$

The property in (5.5) establishes a relation between the inverse Laplace transforms for $\exp(p^2x)\, D_\mu\left(2^{1/2}x^{1/2}p\right) D_\nu\left(2^{1/2}x^{1/2}p\right)$ and
$p^{-1/2}\exp(px)\, D_\mu\left(2^{1/2}x^{1/2}p^{1/2}\right) D_\nu\left(2^{1/2}x^{1/2}p^{1/2}\right)$. Equation (5.5) then allows us to obtain the following indefinite integral

$$\begin{aligned}&\int_0^\infty t^{-(1+\mu+\nu)}\exp\left(-\frac{x+y}{4xy}t^2\right) {}_2F_2\left(-\mu,-\nu;-\frac{\mu+\nu}{2},\frac{1-\mu-\nu}{2};\frac{t^2}{8x}\right)dt = \\ &2^{-(1+\mu+\nu)}\Gamma\left[-\frac{\mu+\nu}{2}\right]\left(\frac{x+y}{xy}\right)^{(\mu+\nu)/2} {}_2F_1\left(-\frac{\mu}{2},-\frac{\nu}{2};\frac{1-\mu-\nu}{2};\frac{y(2x+y)}{(x+y)^2}\right) \\ &\qquad [\operatorname{Re}(\mu+\nu)<0, \operatorname{Re}x\geqslant 0, \operatorname{Re}y\geqslant 0]\end{aligned} \tag{5.6}$$

**Acknowledgment.** I would like to thank the referee for the careful reading and very useful comments that greatly improved clarity and exposition of the main results.

## References


[1] M. Abramowitz and I.A. Stegun, *Handbook of Mathematical Functions with Formulas, Graphs, and Mathematical Tables*, Dover Publications, New York, 1972.

[2] A. Ali, T. Lebel and A. Mani, *Rainfall Estimation in the Sahel. Part I: Error Function*, J Appl. Meteorol. **44**, 1691–1706, 2005.

[3] J.V. Beck, K.D. Cole, A. Haji-Sheikh and B. Litkouhi, *Heat Conduction Using Green's Functions*, Hemisphere Publishing Corporation, London, 1992.

[4] J.-F. Bercher and C. Vignat, *On minimum Fisher information distributions with restricted support and fixed variance*, Inform. Sci. **179**, 3832–3842, 2009.

[5] R.M. Capocelli and L.M. Ricciardi, *Diffusion Approximation and First Passage Time Problem for a Model Neuron*, Kybernetik **8** (6), 214–223, 1971.

[6] H.S. Carslaw and J.C. Jaeger, *Conduction of Heat in Solids*, 2nd ed., Clarendon Press, Oxford, 1959.

[7] R. Combescot and T. Dombre, *Superfluid current in $^3$He-A at $T = 0$*, Phys. Rev. B **28** (9), 5140–5149, 1983.

[8] J.L. deLyra, S.K. Foong and T.E. Gallivan, *Finite lattice systems with true critical behavior*, Phys. Rev. D **46** (4), 1643–1657, 1992.

[9] L. Durand, *Nicholson-type integrals for products of Gegenbauer functions and related topics*, in: Theory and Applications of Special Functions, 353–374, Academic Press, New York, 1975.

[10] Á. Elbert and M.E. Muldoon, *Inequalities and monotonicity properties for zeros of Hermite functions*, Proc. Roy. Soc. Edinburgh Sect. A **129**, 57–75, 1999.

[11] Á Elbert and M.E. Muldoon, *Approximations for zeros of Hermite functions*, in: Special Functions and Orthogonal Polynomials, Contemporary Mathematics **471**, 117–126, American Mathematical Society, Providence, 2008.

[12] A. Erdélyi, W. Magnus, F. Oberhettinger and F.G. Tricomi, *Higher Transcendental Functions, Vol. 1*, McGraw-Hill, New York, 1953.

[13] A. Erdélyi, W. Magnus, F. Oberhettinger and F.G. Tricomi, *Higher Transcendental Functions, Vol. 2*, McGraw-Hill, New York, 1953.

[14] A. Erdélyi, W. Magnus, F. Oberhettinger and F.G. Tricomi, *Tables of Integral Transforms, Vol. 1*, McGraw-Hill, New York, 1954.

[15] W. Feller, *An Introduction to Probability Theory and Its Applications, Vol. 1*, 3rd ed., John Wiley & Sons, New York, 1968.

[16] I.S. Gradshteyn and I.M. Ryzhik, *Table of Integrals, Series, and Products*, 8th ed., edited by D. Zwillinger and V. Moll, Academic Press, New York, 2014.

[17] Yu.P. Kalmykov, W.T. Coffey and J.T. Waldron, *Exact analytic solution for the correlation time of a Brownian particle in a doublewell potential from the Langevin equation*, J. Chem. Phys. **105** (5), 2112–2118, 1996.

[18] C. Malyshev, *Higher corrections to the mass current in weakly inhomogeneous superfluid* $^3$*He-A*, Phys. Rev. B **59** (10), 7064–7075, 1999.

[19] E.W. Ng and M. Geller, *A Table of Integrals of the Error Functions*, J. Res. NBS **73B** (1), 1–20, 1969.

[20] Y. Nie and V. Linetsky, *Sticky reflecting Ornstein–Uhlenbeck diffusions and the Vasicek interest rate model with the sticky zero lower bound*, Stoch. Models, forthcoming.

[21] F. Oberhettinger and L. Badii, *Tables of Laplace Transforms*, Springer–Verlag, Berlin, 1973.

[22] K. Oldham, J. Myland and J. Spanier, *An Atlas of Functions*, 2nd ed., Springer–Verlag, Berlin, 2009.

[23] J.K. Patel and C.B. Read, *Handbook of the Normal Distribution*, Marcel Dekker, New York, 1982.

[24] A.P. Prudnikov, Yu.A. Brychkov and O.I. Marichev, *Integrals and Series. More Special Functions, Vol. 3*, Gordon and Breach, New York, 1990.

[25] A.P. Prudnikov, Yu.A. Brychkov and O.I. Marichev, *Integrals and Series. Direct Laplace Transforms, Vol. 4*, Gordon and Breach, New York, 1992.

[26] A.P. Prudnikov, Yu.A. Brychkov and O.I. Marichev, *Integrals and Series. Inverse Laplace Transforms, Vol. 5*, Gordon and Breach, New York, 1992.

[27] D. Veestraeten, *Some integral representations and limits for (products of) the parabolic cylinder function*, Integr. Transf. Spec. F. **27** (1), 64–77, 2016.

[28] D. Veestraeten, *An integral representation for the product of parabolic cylinder functions*, Integr. Transf. Spec. F. **28** (1), 15–21, 2017.

[29] E.T. Whittaker, *On the Functions Associated with the Parabolic Cylinder in Harmonic Analysis*, Proc. Lond. Math. Soc. **35**, 417–427, 1902.

[30] T.V. Zaqarashvili and K. Murawski, *Torsional oscillations of longitudinally inhomogeneous coronal loops*, Astron. Astrophys. **470**, 353–357, 2007.

Hacettepe Journal of
Mathematics & Statistics

Hacet. J. Math. Stat.
Volume 50 (1) (2021), 79 – 91
DOI : 10.15672/hujms.644665

Research Article

# Mean value theorem and semigroups of operators for interval-valued functions on time scales

Yonghong Shen

*School of Mathematics and Statistics, Tianshui Normal University, Tianshui 741001, P.R. China*


## Abstract

In this paper, a new version of mean value theorem for interval-valued functions on time scales is established. Meantime, some basic concepts and results associated with semigroups of operators for interval-valued functions on time scales are presented. As an application of semigroups of operators, under certain conditions, we consider the initial value problem for interval-valued differential equations on time scales. Finally, two issues worthy of further discussion are presented.

**Mathematics Subject Classification (2020).** 49J53, 54C60, 58C06

**Keywords.** interval-valued functions, mean value theorem, semigroup of operators, time scales

## 1. Introduction

In 1988, the notion of time scale was introduced by Hilger [6] to unify continuous and discrete analysis. There is no doubt that the time scale calculus provide a unified framework for the study of differential equations and difference equations. In practice, many problems involve various types of uncertainty. Usually, the knowledge about the parameters of a real world system is imprecise or uncertain because, generally, it is difficult to accurately observe or measure the true value of these parameters. In these cases, the value of a parameter cannot be characterized by an ordinary real number. Accordingly, interval numbers and fuzzy numbers are two important tools to deal with these problems. In fact, interval numbers can be regarded as a special case of fuzzy numbers. Taking into account the shortcoming of the difference of fuzzy numbers, it is necessary to carry out the study of interval analysis. More importantly, interval analysis can provide important methodologies and foundations for fuzzy analysis. In 1993, Markov [8] first studied the differentiability and integrability of interval-valued functions. Later, Stefanini and Bede [11] together with Chalco-Cano et al. [3] further extended the theory of calculus of interval-valued functions. In 2013, Lupulescu [7] introduced the differentiability and integrability for interval-valued functions on time scales by using the generalized Hukuhara differentiability.

The mean value theorem for real-valued functions has important and extensive application in the classical calculus. In [8], the mean value theorem for interval-valued functions was established. Afterwards, the work was extended to the interval-valued functions on time scales by Lupulescu [7]. One purpose of this paper is to give a new version of the

Email address: shenyonghong2008@hotmail.com
Received: 09.11.2019; Accepted: 02.05.2020

mean value theorem for interval-valued functions on time scales. In addition, semigroups of operators are very important in the study of differential equations. In 2005, the semigroups of operators on spaces of fuzzy-number-valued functions were proposed by Gal and Gal [4] and were applied to study fuzzy differential equations. Recently, Hamza and Oraby [5] developed the theory of semigroups of operators on time scales. Motivated by these works, the other purpose of the present paper is to present some basic concepts and results related to semigroups of operators for interval-valued functions on time scales.

## 2. Preliminaries

Let $\mathbb{Z}_0^+$, $\mathbb{R}_0^+$ and $\mathbb{R}$ denote the set of all nonnegative integers, nonnegative real numbers and real numbers, respectively. Denote by $\mathcal{K}$ the set of all nonempty compact convex subsets (i.e., bounded and closed intervals) of the real line $\mathbb{R}$. For $A = [a^-, a^+], B = [b^-, b^+] \in \mathcal{K}, \lambda \in \mathbb{R}$, the Minkowski addition $A+B$ and scalar multiplication $\lambda \cdot A$ (denoted by $\lambda A$) can be defined by

$$A + B = [a^-, a^+] + [b^-, b^+] = [a^- + b^-, a^+ + b^+]$$

and

$$\lambda \cdot A = \lambda A = \lambda[a^-, a^+] = [\min\{\lambda a^-, \lambda a^+\}, \max\{\lambda a^-, \lambda a^+\}].$$

It is well know that the addition is associative and commutative and with the neutral element $\{0\}$. Especially, if $\lambda = -1$, then the scalar multiplication gives the opposite $-A = (-1)A = [-a^+, -a^-]$. However, in general, $A + (-A) \neq \{0\}$. That is to say, the opposite of $A$ is not the inverse of $A$ with respect to the Minkowski addition, unless $A$ is a singleton.

Let $A, B \in \mathcal{K}$. If there exists $C \in \mathcal{K}$ such that $A = B + C$, then $C$ is called the Hukuhara difference (or H-difference) of $A$ and $B$, and it is denoted by $C := A \ominus B$. Note that the H-difference is unique, but it does not always exist for any two intervals. Given two intervals $A, B \in \mathcal{K}$, it is easy to know that the H-difference $A \ominus B$ exists if and only if $len(A) \geq len(B)$, where $len(\cdot)$ denotes the length of the interval, i.e., $len(A) = a^+ - a^-$. In order to overcome this shortcoming, the generalized difference is introduced as follows.

**Definition 2.1** (Markov [8], Stefanini [10])**.** Let $A, B \in \mathcal{K}$. The generalized Hukuhara difference (gH-difference for short) is defined as

$$A \ominus_g B = C \Leftrightarrow \begin{cases} (i)\ A = B + C \Leftrightarrow A \ominus B = C, \\ or\ (ii)\ B = A + (-C) \Leftrightarrow B \ominus A = -C. \end{cases}$$

According to Def. 2.1, if $A = [a^-, a^+], B = [b^-, b^+] \in \mathcal{K}$, then we have

$$\begin{aligned} A \ominus_g B &= [a^-, a^+] \ominus_g [b^-, b^+] \\ &= [\min\{a^- - b^-, a^+ - b^+\}, \max\{a^- - b^-, a^+ - b^+\}] \\ &= \begin{cases} [a^- - b^-, a^+ - b^+], & len(A) \geq len(B), \\ [a^+ - b^+, a^- - b^-], & len(A) < len(B). \end{cases} \end{aligned}$$

From [8, 10, 12], some basic properties of gH-difference can be summarized as follows.

(i) $A \ominus_g A = \{0\}$, $A \ominus_g \{0\} = A$, $\{0\} \ominus_g A = -A$;
(ii) $A \ominus_g B = (-B) \ominus_g (-A) = -(B \ominus_g A)$;
(iii) $A \ominus_g (-B) = B \ominus_g (-A)$, $(-A) \ominus_g B = (-B) \ominus_g A$;
(iv) $(A + B) \ominus_g B = A$, $A \ominus_g (A + B) = -B$;
(v) $(A \ominus_g B) + B = A$ if $len(A) \geq len(B)$, $A + (-1)(A \ominus_g B) = B$ if $len(A) < len(B)$;
(vi) $\lambda(A \ominus_g B) = \lambda A \ominus_g \lambda B$, $\lambda \in \mathbb{R}$;
(vii) $(\lambda + \mu)A = \lambda A + \mu A$ if $\lambda\mu \geq 0$, $(\lambda + \mu)A = \lambda A \ominus_g (-\mu A)$ if $\lambda\mu < 0$.

**Lemma 2.2.** *Let $A = [a^-, a^+], B = [b^-, b^+]$ and $C = [c^-, c^+]$ belong to $\mathcal{K}$. Then:*

(i) *If $len(A) \geq len(C)$, then $(A + B) \ominus_g C = (A \ominus_g C) + B$;*

(ii) *If $len(A) < len(C)$, then $(A + B) \ominus_g C = (A \ominus_g C) \ominus_g (-B)$.*

***Proof.*** For simplicity, we write $(A + B) \ominus_g C = D$, where $D = [d^-, d^+]$.
(i) If $len(A) \geq len(C)$, then $(A + B) \ominus_g C = (A + B) \ominus C$. Using the representation of endpoints, we have

$$\begin{aligned}(A + B) \ominus_g C &= (A + B) \ominus C \\ &= [a^- + b^- - c^-, a^+ + b^+ - c^+] \\ &= [a^- - c^-, a^+ - c^+] + [b^-, b^+] \\ &= (A \ominus C) + B \\ &= (A \ominus_g C) + B.\end{aligned}$$

(ii) If $len(A) < len(C)$, then $A \ominus_g C = -(C \ominus A)$. Therefore, we can infer from Definition 2.1 that

$$\begin{aligned}&(A \ominus_g C) \ominus_g (-B) \\ &= [a^+ - c^+, a^- - c^-] \ominus_g [-b^+, -b^-] \\ &= [\min\{a^+ - c^+ + b^+, a^- - c^- + b^-\}, \max\{a^+ - c^+ + b^+, a^- - c^- + b^-\}] \\ &= (A + B) \ominus_g C.\end{aligned}$$

$\square$

Now we define a functional $\|\cdot\| : \mathcal{K} \to [0, \infty)$ by $\|A\| = \max\{|a^-|, |a^+|\}$ for every $A = [a^-, a^+] \in \mathcal{K}$. It can easily be shown that $\|\cdot\|$ is a norm on $\mathcal{K}$, and thus the quadruple $(\mathcal{K}, +, \cdot, \|\cdot\|)$ is a normed quasilinear space [9].

Given two intervals $A = [a^-, a^+]$, $B = [b^-, b^+] \in \mathcal{K}$, the Hausdorff-Pompeiu metric between $A$ and $B$ is defined by $d_H(A, B) = \max\{|a^- - b^-|, |a^+ - b^+|\}$. It is well known that $(\mathcal{K}, d_H)$ is a complete and separable metric space. Furthermore, the following relationships exist between the Hausdorff-Pompeiu metric $d_H$ and the norm $\|\cdot\|$:

$$\|A\| = d_H(A, \{0\}), \qquad d_H(A, B) = \|A \ominus_g B\|.$$

In addition, for all $A, B, C, D \in \mathcal{K}$, the metric $d_H$ has the following properties:

(i) $d_H(A + B, A + C) = d_H(B, C)$,

(ii) $d_H(\lambda A, \lambda, B) = |\lambda| d_H(A, B)$, $\lambda \in \mathbb{R}$,

(iii) $d_H(A + C, B + D) \leq d_H(A, B) + d_H(C, D)$,

(iv) $d_H(A \ominus_g B, A \ominus_g C) \leq d_H(B, C)$.

Here, we briefly recall some basic notions related to the time scale. For more details, we recommend two excellent monographs [1,2] written by Bohner and Peterson. A time scale $\mathbb{T}$ is a nonempty closed subset of $\mathbb{R}$. For $t \in \mathbb{T}$, the forward jump operator $\sigma$ and the back jump operator $\rho$ are defined as $\sigma(t) := \inf\{s \in \mathbb{T} : s > t\}$ and $\rho(t) := \sup\{s \in \mathbb{T} : s < t\}$, respectively. Especially, $\inf \emptyset = \sup \mathbb{T}$, $\sup \emptyset = \inf \mathbb{T}$.

A point $t \in \mathbb{T}$ is said to be *right-scattered*, *right-dense*, *left-scattered* and *left-dense* if $\sigma(t) > t$, $\sigma(t) = t$, $\rho(t) < t$ and $\rho(t) = t$, respectively. Given a time scale $\mathbb{T}$, the *graininess function* $\mu : \mathbb{T} \to [0, \infty)$ is defined by $\mu(t) = \sigma(t) - t$. The set $\mathbb{T}^\kappa$ is derived from the time scale $\mathbb{T}$ as follows: If $\mathbb{T}$ has a left-scattered maximum $\gamma$, then $\mathbb{T}^\kappa = \mathbb{T} - \{\gamma\}$. Otherwise, $\mathbb{T}^\kappa = \mathbb{T}$. Especially, given a time scale interval $[a, b]_{\mathbb{T}} = \{t \in \mathbb{T} |\ a \leq t \leq b\}$, if $\rho(b) = b$, then $[a, b]^\kappa = [a, b]_{\mathbb{T}}$. Otherwise, $[a, b]^\kappa = [a, b)_{\mathbb{T}}$. In essence, $[a, b)_{\mathbb{T}} = [a, \rho(b)]_{\mathbb{T}}$.

Let $g : \mathbb{T} \to \mathbb{R}$ be a real-valued function and let $t \in \mathbb{T}^\kappa$. Given any $\varepsilon > 0$, if there exist a number $\alpha$ and a neighborhood $U$ of $t$ such that

$$|g(\sigma(t)) - g(s) - \alpha(\sigma(t) - s)| \leq \varepsilon |\sigma(t) - s|$$

for all $s \in U$, then we say that $g$ is delta differentiable (or in short: $\Delta$-differentiable) at $t$. Correspondingly, the number $\alpha$ is called the $\Delta$-derivative and it is denoted by $g^{\Delta}(t)$. More generally, the function $g$ is said to be delta differentiable ($\Delta$-differentiable) on $\mathbb{T}^{\kappa}$ provided the $\Delta$-derivative $g^{\Delta}(t)$ exists for all $t \in \mathbb{T}^{\kappa}$.

**Definition 2.3** (Lupulescu [7]). Let $F : \mathbb{T} \to \mathcal{K}$ be an interval-valued function. Then we say that $F$ is $l$-nondecreasing (or $l$-nonincreasing) on $\mathbb{T}$ if the real-valued function $t \to len(F(t))$ is nondecreasing (or nonincreasing) on $\mathbb{T}$. Generally, if $F$ is $l$-nondecreasing or $l$-nonincreasing on $\mathbb{T}$, then we say that $F$ is $l$-monotonic on $\mathbb{T}$.

**Definition 2.4** (Lupulescu [7]). Let $F : \mathbb{T} \to \mathcal{K}$ be an interval-valued function and let $A \in \mathcal{K}$. If for every $\varepsilon > 0$, there exists $\delta > 0$ such that $\|F(t) \ominus_g A\| = d_H(F(t), A) \leq \varepsilon$ for all $t \in U_{\mathbb{T}}(t_0, \delta)$ (i.e., $U_{\mathbb{T}}(t_0, \delta) = (t_0 - \delta, t_0 + \delta) \cap \mathbb{T}$), then we say that $A$ is the $\mathbb{T}$-limit of $F$ at $t_0 \in \mathbb{T}$. If $F$ has a $\mathbb{T}$-limit $A$ at $t_0$, then it is unique and is denoted by $A = \mathbb{T} - \lim\limits_{t \to t_0} F(t)$.

An interval-valued function $F : \mathbb{T} \to \mathcal{K}$ is called $rd$-continuous if it is continuous at all right-dense points in $\mathbb{T}$ and its left-sided $\mathbb{T}$-limits exist at all left-dense points in $\mathbb{T}$.

**Definition 2.5** (Lupulescu [7]). Let $F : \mathbb{T} \to \mathcal{K}$ be an interval-valued function and let $t \in \mathbb{T}^{\kappa}$. Then we define $F_{gH}^{\Delta}(t)$ to be the interval (provided it exists) with the property that for every $\varepsilon > 0$, there exists $\delta > 0$ such that

$$d_H(F(\sigma(t)) \ominus_g F(s), (\sigma(t) - s)F_{gH}^{\Delta}(t)) \leq \varepsilon|\sigma(t) - s|$$

for all $s \in U_{\mathbb{T}}(t, \delta)$. Here, $F_{gH}^{\Delta}(t)$ is called the delta generalized Hukuhara derivative ($\Delta_{gH}$-derivative for short) of $F$ at $t$. Meantime, if $F_{gH}^{\Delta}(t)$ exists for each $t \in \mathbb{T}^{\kappa}$, then we say that $F$ is delta generalized Hukuhara differentiable ($\Delta_{gH}$-differentiable for short) on $\mathbb{T}^{\kappa}$. In particular, the $\Delta_{gH}$-derivative $F_{gH}^{\Delta}$ degenerates into the $gH$-derivative $F_{gH}'$ if the time scale $\mathbb{T} = \mathbb{R}$.

**Theorem 2.6** (Lupulescu [7]). *Assume that $F : \mathbb{T} \to \mathcal{K}$ is an interval-valued function and let $t \in \mathbb{T}^{\kappa}$. Then, the following statements are true:*

(i) *If $F : \mathbb{T} \to \mathcal{K}$ is $\Delta_{gH}$-differentiable at $t \in \mathbb{T}^{\kappa}$, then it is continuous at $t$;*

(ii) *If $F$ is continuous at $t$ and $t$ is right-scattered, then $F$ is $\Delta_{gH}$-differentiable at $t$ with*

$$F_{gH}^{\Delta}(t) = \frac{F(\sigma(t)) \ominus_g F(t)}{\mu(t)};$$

(iii) *If $t$ is right-dense, then $F$ is $\Delta_{gH}$-differentiable at $t$ if and only if the $\mathbb{T}$-limit*

$$\mathbb{T} - \lim_{s \to t} \frac{F(t) \ominus_g F(s)}{t - s}$$

*exists as a closed interval. In this case*

$$F_{gH}^{\Delta}(t) = \mathbb{T} - \lim_{s \to t} \frac{F(t) \ominus_g F(s)}{t - s};$$

(iv) *If $F$ is $\Delta_{gH}$-differentiable at $t$, then*

$$F(\sigma(t)) \ominus_g F(t) = \mu(t) F_{gH}^{\Delta}(t).$$

Finally, the induction principle on time scales is provided, which is useful in the next section.

**Theorem 2.7** (Bohner and Peterson [7]). *Let $t_0 \in \mathbb{T}$ and let $\{S(t) : t \in [t_0, +\infty)\}$ be a family of statements satisfying:*

(I) *$S(t_0)$ is true;*

(II) *If $t \in [t_0, +\infty)$ is right-scattered and $S(t)$ is true, then $S(\sigma(t))$ is also true;*

(III) *If $t \in [t_0, +\infty)$ is right-dense and $S(t)$ is true, then there is a neighborhood $U$ of $t$ such that $S(s)$ is true for all $s \in U \cap (t, +\infty)$;*
(IV) *If $t \in (t_0, +\infty)$ is left-dense and $S(s)$ is true for all $s \in [t_0, t)$, then $S(t)$ is true.* *Then $S(t)$ is true for all $t \in [t_0, +\infty)$.*

## 3. Mean value theorem for interval-valued functions on time scales

Based on the works of Markov [8] and Lupulescu [7], in this section, we shall establish another version of the mean value theorem for interval-valued functions on time scales.

**Theorem 3.1** (Markov [8])**.** *Let $F$ be a continuous interval-valued function on $[a, b]$ and $gH$-differentiable in $(a, b)$. Then*

$$F(b) \ominus_g F(a) \subset (b-a)F'_{gH}([a, b]),$$

*where $F'_{gH}([a, b]) = \bigcup_{\xi \in [a,b]} F'_{gH}(\xi)$.*

**Remark 3.2.** In general, it is not true that there exists $\xi \in [a, b]$ such that $F(b) \ominus_g F(a) \subset (b-a)F'_{gH}(\xi)$.

**Theorem 3.3** (Lupulescu [7])**.** *Let $F$ be a continuous and $l$-monotonic interval-valued function on $[a, b]_{\mathbb{T}}$ and let $F$ be $\Delta_{gH}$-differentiable in $[a, b)_{\mathbb{T}}$. Then*

$$F(b) \ominus_g F(a) \subset (b-a)F^{\Delta}_{gH}([a, b)_{\mathbb{T}}),$$

*where $F^{\Delta}_{gH}([a, b)_{\mathbb{T}}) = \bigcup_{\xi \in [a,b)_{\mathbb{T}}} F^{\Delta}_{gH}(\xi)$.*

**Theorem 3.4.** *Let $F$ and $g$ be an interval-valued function and a real-valued function defined on $\mathbb{T}$, respectively. Assume that $F$ is $\Delta_{gH}$-differentiable and $g$ is $\Delta$-differentiable on $\mathbb{T}^\kappa$. If*

$$\|F^{\Delta}_{gH}(t)\| \leq g^{\Delta}(t)$$

*for all $t \in \mathbb{T}^\kappa$, then*

$$\|F(t) \ominus_g F(r)\| \leq g(t) - g(r)$$

*for all $t \in [r, s]_{\mathbb{T}}$ with $r, s \in \mathbb{T}$ and $r \leq s$.*

***Proof.*** Let $r, s \in \mathbb{T}$ with $r \leq s$. For any $\varepsilon > 0$, we can show by the induction principle as shown in Theorem 2.7 that

$$S(t) : \|F(t) \ominus_g F(r)\| \leq g(t) - g(r) + \varepsilon(t - r)$$

holds for all $t \in [r, s]_{\mathbb{T}}$. The proof is divided into four steps.

(I) If $t = r$, then the statement $S(r)$ is obviously true.
(II) Assume that $t$ is right-scattered and $S(t)$ is satisfied. According to Definition 2.1 and Theorem 2.6 (iv), we have the following two cases:
Case (a):

$$\begin{aligned}
\|F(\sigma(t)) \ominus_g F(r)\| &= d_H(F(\sigma(t)), F(r)) \\
&= d_H(F(t) + \mu(t)F^{\Delta}_{gH}(t), F(r)) \\
&\leq d_H(F(t), F(r)) + d_H(\mu(t)F^{\Delta}_{gH}(t), \{0\}) \\
&= d_H(F(t), F(r)) + \mu(t)d_H(F^{\Delta}_{gH}(t), \{0\}) \\
&= d_H(F(t), F(r)) + \mu(t)\|F^{\Delta}_{gH}(t)\| \\
&\leq d_H(F(t), F(r)) + \mu(t)g^{\Delta}(t) \\
&\leq g(t) - g(r) + \varepsilon(t - r) + g(\sigma(t)) - g(t) \\
&= g(\sigma(t)) - g(r) + \varepsilon(t - r) \\
&\leq g(\sigma(t)) - g(r) + \varepsilon(\sigma(t) - r).
\end{aligned}$$

Case (b):

$$\begin{aligned}
\|F(\sigma(t)) \ominus_g F(r)\| &= d_H(F(\sigma(t)), F(r)) \\
&= d_H(F(\sigma(t)) + (-1)\mu(t)F_{gH}^{\Delta}(t), F(r) + (-1)\mu(t)F_{gH}^{\Delta}(t)) \\
&= d_H(F(t), F(r) + (-1)\mu(t)F_{gH}^{\Delta}(t)) \\
&\leq d_H(F(t), F(r)) + d_H(\{0\}, (-1)\mu(t)F_{gH}^{\Delta}(t)) \\
&= d_H(F(t), F(r)) + \mu(t)d_H(\{0\}, F_{gH}^{\Delta}(t)) \\
&= d_H(F(t), F(r)) + \mu(t)\|F_{gH}^{\Delta}(t)\| \\
&\leq g(\sigma(t)) - g(r) + \varepsilon(\sigma(t) - r).
\end{aligned}$$

Thus, the statement $S(\sigma(t))$ is satisfied.

(III) Suppose that $S(t)$ holds and $t \neq s$ is right-dense. Clearly, $\sigma(t) = t$. Since $F$ is $\Delta_{gH}$-differentiable and $g$ is $\Delta$-differentiable at $t$, there exists a neighborhood $U_{\mathbb{T}}$ of $t$ such that

$$d_H(F(t) \ominus_g F(s), F_{gH}^{\Delta}(t)(t-s)) \leq \frac{\varepsilon}{2}|t-s|$$

for all $s \in U_{\mathbb{T}}$ and

$$|g(t) - g(s) - g^{\Delta}(t)(t-s)| \leq \frac{\varepsilon}{2}|t-s|$$

for all $s \in U_{\mathbb{T}}$. Therefore, we can obtain that

$$\begin{aligned}
d_H(F(t), F(s)) &= d_H(F(t) \ominus_g F(s), \{0\}) \\
&\leq d_H(F(t) \ominus_g F(s), F_{gH}^{\Delta}(t)(t-s)) + d_H(\{0\}, F_{gH}^{\Delta}(t)(t-s)) \\
&\leq \Big(\|F_{gH}^{\Delta}(t)\| + \frac{\varepsilon}{2}\Big)|t-s|
\end{aligned}$$

and

$$g(s) - g(t) - g^{\Delta}(t)(s-t) \geq -\frac{\varepsilon}{2}|t-s|$$

for all $s \in U_{\mathbb{T}}$. Hence, for all $s \in U_{\mathbb{T}} \cap (t, \infty)$, we have

$$\begin{aligned}
\|F(s) \ominus_g F(r)\| &= d_H(F(s), F(r)) \\
&\leq d_H(F(s), F(t)) + d_H(F(t), F(r)) \\
&\leq \Big(\|F_{gH}^{\Delta}(t)\| + \frac{\varepsilon}{2}\Big)|t-s| + d_H(F(t), F(r)) \\
&\leq \Big(g^{\Delta}(t) + \frac{\varepsilon}{2}\Big)|t-s| + d_H(F(t), F(r)) \\
&\leq \Big(g^{\Delta}(t) + \frac{\varepsilon}{2}\Big)|t-s| + g(t) - g(r) + \varepsilon(t-r) \\
&= g^{\Delta}(t)(s-t) + \frac{\varepsilon}{2}(s-t) + g(t) - g(r) + \varepsilon(t-r) \\
&\leq g(s) - g(t) + \frac{\varepsilon}{2}|t-s| + \frac{\varepsilon}{2}(s-t) + g(t) - g(r) + \varepsilon(t-r) \\
&= g(s) - g(r) + \varepsilon(s-r),
\end{aligned}$$

which implies that $S(s)$ holds for all $s \in U_{\mathbb{T}} \cap (t, \infty)$.

(IV) Let $t$ be left-dense and assume that $S(\tau)$ holds for all $\tau < t$. By the continuity of $F$ and $g$, we then obtain that

$$\begin{aligned}
\|F(t) \ominus_g F(r)\| &= \lim_{\tau \to t-} \|F(\tau) \ominus_g F(r)\| \\
&\leq \lim_{\tau \to t-} g(\tau) - g(r) + \varepsilon(\tau - r) \\
&= g(t) - g(r) + \varepsilon(t-r),
\end{aligned}$$

which means that the statement $S(t)$ is true.

Due to the arbitrariness of $\varepsilon$, we have obtained the desired result and completed the proof of this theorem. □

As an application of Theorem 3.4, we can obtain the following results.

**Corollary 3.5.** *Let $F, G : \mathbb{T} \to \mathcal{K}$ be two $\Delta_{gH}$-differentiable interval-valued functions on $\mathbb{T}^\kappa$. Then*

(i) *If $D$ is a compact interval with endpoints $r, s \in \mathbb{T}$, then*

$$\|F(s) \ominus_g F(r)\| \le \Big(\sup_{t\in D^\kappa} \|F^\Delta_{gH}(t)\|\Big)|s-r|.$$

(ii) *If $F^\Delta_{gH}(t) = \{0\}$ for all $t \in \mathbb{T}^\kappa$, then $F$ is a constant interval.*

(ii) *If both $F$ and $G$ are $l$-nondecreasing or $l$-nonincreasing, and $F^\Delta_{gH}(t) = G^\Delta_{gH}(t)$ for all $t \in \mathbb{T}^\kappa$, then*

$$F(t) \ominus_g G(t) = C$$

*for all $t \in \mathbb{T}$, where $C$ is a constant interval.*

(iv) *If $F$ and $G$ are such that one is $l$-nondecreasing and the other is $l$-nonincreasing, and $F^\Delta_{gH}(t) = -G^\Delta_{gH}(t)$ for all $t \in \mathbb{T}^\kappa$, then*

$$F(t) + G(t) = C$$

*for all $t \in \mathbb{T}$, where $C$ is a constant interval.*

***Proof.*** (i) Let $r, s \in \mathbb{T}$ with $r \le s$. Define

$$g(t) := \Big(\sup_{\tau\in[r,s]^\kappa} \|F^\Delta_{gH}(\tau)\|\Big)(t-r)$$

for $t \in \mathbb{T}$. Then, it is easy to know that

$$g^\Delta(t) = \sup_{\tau\in[r,s]^\kappa} \|F^\Delta_{gH}(\tau)\| \ge \|F^\Delta_{gH}(t)\|$$

for all $\tau \in [r, s]^\kappa$. By Theorem 3.4, the desired result can be obtained.
(ii) It is a direct consequence of part (i).
(iii) By Theorem 4 in [7], we have

$$(F(t) \ominus_g G(t))^\Delta_{gH} = F^\Delta_{gH}(t) \ominus_g G^\Delta_{gH}(t) = F^\Delta_{gH}(t) \ominus_g F^\Delta_{gH}(t) = \{0\}$$

for $t \in \mathbb{T}^\kappa$. The desired result follows immediately from (ii).
(iv) Similar to part (iii), since

$$(F(t) + G(t))^\Delta_{gH} = F^\Delta(t) \ominus_g (-G^\Delta_{gH}(t)) = F^\Delta_{gH}(t) \ominus_g F^\Delta_{gH}(t) = \{0\}$$

for $t \in \mathbb{T}^\kappa$. □

**Remark 3.6.** If $F$ and $G$ are differently $l$-monotonic in (iii) of Corollary 3.5, in general, there is no constant interval $C$ such that $F(t) \ominus_g G(t) = C$. Analogously, $F$ and $G$ are equally $l$-monotonic in (iv), then the result is not necessarily true.

**Remark 3.7.** The results (iii) and (iv) of Corollary 3.4 coincide with Corollary 2 in [7].

**Example 3.8.** (i) Let $\mathbb{T} = [0, 1]$ and let $F(t) = [t, 2t]$ and $G(t) = [2t - 1, t]$. Note that $len(F(t)) = t$ is nondecreasing on $\mathbb{T}$ and $len(G(t)) = 1-t$ is nonincreasing on $\mathbb{T}$. It is easy to check that $F(t)$ and $G(t)$ are $\Delta_{gH}$-differentiable on $\mathbb{T}^\kappa = [0, 1]$ and $F^\Delta_{gH}(t) = F'_{gH}(t) = [1, 2] = G'_{gH}(t) = G^\Delta_{gH}(t)$ for each $t \in [0, 1]$(Only consider the unilateral derivative at the endpoints 0 and 1). However, there is no constant interval $C$ such that $F(t) \ominus_g G(t) = C$.

(ii) Let $\mathbb{T} = [0, 1]$ and let $F(t) = [-t, 2t]$ and $G(t) = [t-1, 2(1-t)]$. Clearly, $len(F(t)) = 3t$ is nondecreasing and $len(G(t)) = 3(1-t)$ in nonincreasing on $\mathbb{T}$. It can easily be verified that $F(t)$ and $G(t)$ are $\Delta_{gH}$-differentiable on $\mathbb{T}^\kappa = [0, 1]$. Moreover, $F^\Delta_{gH}(t) = F'_{gH}(t) =$

$[-1,2]$, $G^{\Delta}_{gH}(t) = G'_{gH}(t) = [-2,1]$ for each $t \in [0,1]$. Then, we have $F^{\Delta}_{gH}(t) = -G^{\Delta}_{gH}(t)$ for each $t \in [0,1]$. By Corollary 3.5, there exists an interval $C = [-1,2]$ such that $F(t) + G(t) = [-1,2] = C$.

**Example 3.9.** Let $\mathbb{T} = h\mathbb{Z}_0^+ = \{hk : k \in \mathbb{Z}_0^+\}$, $h > 0$. Suppose $F(t) = [t, t^2]$ and $G(t) = [t+a, t^2+b]$, where $a$ and $b$ are two fixed constants with $a \le b$. Obviously, both $len(F(t)) = t(t-1)$ and $len(G(t)) = t(t-1)+b-a$ are $l$-nondecreasing on $\mathbb{T}$. By Theorem 2.6, we can obtain $F^{\Delta}_{gH}(t) = [1,2t] = G^{\Delta}_{gH}(t)$ for each $t \in \mathbb{T}$. Therefore, we can find an interval $C = [-b,-a]$ such that $F(t) \ominus_g G(t) = C$ on $\mathbb{T}$.

**Example 3.10.** Let $\mathbb{T} = \mathbb{R}$ and let $F(t) = [-2e^{-t}-1, e^{-t}+2]$ and $G(t) = [-2e^{-t}, e^{-t}+1]$. Obviously, $len(F(t)) = 3 + 3e^{-t}$ and $len(G(t)) = 1 + 3e^{-t}$ are nonincreasing on $\mathbb{R}$. It is easy to know that $F(t)$ and $G(t)$ are $\Delta_{gH}$-differentiable on $\mathbb{R}$ and $F^{\Delta}_{gH}(t) = F'_{gH}(t) = [-1,2]e^{-t} = G'_{gH}(t) = G^{\Delta}_{gH}(t)$ for each $t \in \mathbb{R}$. By Corollary 3.5, we can find an interval $C = [-1,1]$ such that $F(t) \ominus_g G(t) = C$.

**Example 3.11.** Let $\mathbb{T} = q^{\mathbb{Z}} = \{q^k | k \in \mathbb{Z}\}$, where $q > 1$. Assume $F(t) = [-t, 2t^2]$ and $G(t) = [-2t^2+1, t+2]$. According to Theorem 2.6, for each $t \in \mathbb{T}$, it follows that

$$\begin{aligned} F^{\Delta}_{gH}(t) &= \frac{F(\sigma(t)) \ominus_g F(t)}{\mu(t)} \\ &= \frac{F(qt) \ominus_g F(t)}{(q-1)t} \\ &= \frac{[-qt, 2q^2t^2] \ominus_g [-t, 2t^2]}{(q-1)t} \\ &= \frac{[-(q-1)t, 2(q^2-1)t^2]}{(q-1)t} \\ &= [-1, 2(q+1)t]. \end{aligned}$$

Using the similar method, we can obtain $G^{\Delta}_{gH}(t) = [-2(q+1)t, 1] = -F^{\Delta}_{gH}(t)$. However, $len(F(t)) = 2t^2 + t$ and $len(G(t)) = 2t^2 + t + 1$ are nondecreasing on $\mathbb{T}$. Therefore, the conditions of Corollary 3.5 are not satisfied. Indeed, there does not exist an interval $C$ such that $F(t) + G(t) = C$.

**Example 3.12.** Let $\mathbb{T} = \mathbb{N}_0^2 = \{n^2 | n \in \mathbb{N}_0\}$ and let $F(t) = [-\sqrt{t}, \sqrt{t}]$ and $G(t) = [\min\{1-\sqrt{t}, \sqrt{t}\}, \max\{1-\sqrt{t}, \sqrt{t}\}]$. For every $t \in \mathbb{T}$, it is easy to know that $t$ is right-scattered. By Theorem 2.6, we can obtain

$$\begin{aligned} F^{\Delta}_{gH}(t) &= \frac{F(\sigma(t)) \ominus_g F(t)}{\mu(t)} \\ &= \frac{F((\sqrt{t}+1)^2) \ominus_g F(t)}{2\sqrt{t}+1} \\ &= \frac{[-\sqrt{t}-1, \sqrt{t}+1] \ominus_g [-\sqrt{t}, \sqrt{t}]}{2\sqrt{t}+1} \\ &= \frac{1}{2\sqrt{t}+1}[-1,1]. \end{aligned}$$

Similarly, we can infer that $G^{\Delta}_{gH}(t) = \frac{1}{2\sqrt{t}+1}[-1,1] = F^{\Delta}_{gH}(t)$. Although $len(F(t)) = 2\sqrt{t}$ is nondecreasing on $\mathbb{T}$, $len(G(t)) = |2\sqrt{t} - 1|$ is not monotonic on $\mathbb{T}$. Therefore, the conditions of Corollary 3.5 are not satisfied. In fact, there does not exist an interval $C$ such that $F(t) \ominus_g G(t) = C$.

## 4. $C_0$-Semigroups for interval-valued functions on time scales

In this section, we shall introduce some basic notions and results associated with semigroups of operators for interval-valued functions on time scales.

**Definition 4.1.** Let $\widetilde{A} : \mathcal{K} \to \mathcal{K}$. $\widetilde{A}$ is said to be a linear operator on $\mathcal{K}$ if

$$\widetilde{A}(\alpha \cdot x + \beta \cdot y) = \alpha \cdot \widetilde{A}(x) + \beta \cdot \widetilde{A}(y)$$

for all $x, y \in \mathcal{K}$ and $\alpha, \beta \in \mathbb{R}$.

**Remark 4.2.** Unlike the property of linear operators on a linear space, it should be noticed that the continuity of a linear operator $\widetilde{A}$ at $\{0\} \in \mathcal{K}$ does not imply the continuity of $\widetilde{A}$ at each $x \in \mathcal{K}$, because $(\mathcal{K}, +, \cdot)$ is not a linear space, in general, the equality $x_0 = (x_0 \ominus_g x) + x$ does not hold, unless $len(x_0) \geq len(x)$.

**Lemma 4.3.** *Let $\widetilde{A}$ be a linear operator on $\mathcal{K}$. Then, for all $x, y \in \mathcal{K}$, we have*

$$\widetilde{A}(x \ominus_g y) = \widetilde{A}(x) \ominus_g \widetilde{A}(y).$$

***Proof.*** Let $z = x \ominus_g y$. Then, we get $x = y + z$ or $y = x + (-z)$. According to Definition 4.1, it follows that

$$\begin{cases} \widetilde{A}(x) = \widetilde{A}(y+z) = \widetilde{A}(y) + \widetilde{A}(z), \\ or\ \widetilde{A}(y) = \widetilde{A}(x + (-z)) = \widetilde{A}(x) + \widetilde{A}(-z), \end{cases}$$

which is equivalent to

$$\begin{cases} \widetilde{A}(x) = \widetilde{A}(y) + \widetilde{A}(z), \\ or\ \widetilde{A}(y) = \widetilde{A}(x) + (-1)\widetilde{A}(z). \end{cases}$$

Therefore, $\widetilde{A}(x \ominus_g y) = \widetilde{A}(z) = \widetilde{A}(x) \ominus_g \widetilde{A}(y)$. $\square$

$$L(\mathcal{K}) = \{\widetilde{A} : \mathcal{K} \to \mathcal{K} |\ \widetilde{A} \textit{ is linear and continuous at each } x \in \mathcal{K}\}.$$

Let us introduce the addition and scalar multiplication in $L(\mathcal{K})$ as follows

$$(\widetilde{A} + \widetilde{B})(x) = \widetilde{A}(x) + \widetilde{B}(x), \quad (\lambda \cdot \widetilde{A})(x) = \lambda \cdot \widetilde{A}(x),$$

for $\widetilde{A}, \widetilde{B} \in L(\mathcal{K})$ and $\lambda \in \mathbb{R}$. Consider the metric $D_H : L(\mathcal{K}) \times L(\mathcal{K}) \to [0, +\infty)$ defined by

$$D_H(\widetilde{A}, \widetilde{B}) = \sup\{d_H(\widetilde{A}(x), \widetilde{B}(x)) :\ \|x\| \leq 1\},$$

where $\|x\| = d_H(x, 0)$. From the properties of $d_H$, it can easily be verified that

(i) $D_H(\widetilde{A} + \widetilde{B}, \widetilde{C} + \widetilde{D}) \leq D_H(\widetilde{A}, \widetilde{C}) + D_H(\widetilde{B}, \widetilde{D})$;
(ii) $D_H(\lambda \cdot \widetilde{A}, \lambda \cdot \widetilde{B}) = |\lambda| D_H(\widetilde{A}, \widetilde{B})$;
(iii) $D_H(\widetilde{A}, \widetilde{B}) \leq D_H(\widetilde{A}, 0) + D_H(0, \widetilde{B}) = \|\widetilde{A}\| + \|\widetilde{B}\|$;
(iv) $D_H(\widetilde{A} + \widetilde{B}, \widetilde{C}) \leq D_H(\widetilde{A}, \widetilde{C}) + D_H(\widetilde{B}, \widetilde{C})$,

where $\widetilde{A}, \widetilde{B}, \widetilde{C}, \widetilde{D} \in L(\mathcal{K})$ and $\lambda \in \mathbb{R}$.

As a special case of Corollary 3.6 in [4], it is easy to know that $(L(\mathcal{K}), D_H)$ is a complete metric space.

**Definition 4.4.** Let $\mathbb{T} \subseteq \mathbb{R}_0^+$ be a semigroup time scale. A $C_0$-semigroup $T$ on $\mathcal{K}$ is a family of continuous linear operators $\{T(t) : t \in \mathbb{T}\} \subset L(\mathcal{K})$, which satisfies

(i) $T(0) = I$, $I$ is the identity operator on $\mathcal{K}$;
(ii) $T(t + s) = T(t)T(s)$ for every $t, s \in \mathbb{T}$;
(iii) $\lim_{t \to 0+} T(t)x = x$ for each $x \in \mathcal{K}$, i.e., $T(\cdot)x : \mathbb{T} \to \mathcal{K}$ is continuous at 0.

**Definition 4.5.** Let $T$ be a $C_0$-semigroup on $\mathcal{K}$. A linear operator $\widetilde{A}$ is called the generator of the $C_0$-semigroup $T$ if for all $x \in \mathcal{K}$, the limit

$$\lim_{s\to 0+} \frac{T(\mu(t))x \ominus_g T(s)x}{\mu(t)-s} = \widetilde{A}x$$

exists uniformly in $t$. Here the limit are considered in the metric $d_H$.

**Example 4.6.** Let $\mathbb{T} = h\mathbb{Z}_0^+ = \{hk : k \in \mathbb{Z}_0^+\}$, $h > 0$ and $\widetilde{A}$ be a continuous linear operator on $\mathcal{K}$. Then $\widetilde{A}$ is the generator of $T(t) = (I + t\widetilde{A})^{t/h}$ for $t \in h\mathbb{Z}_0^+$. In fact, for $x \in \mathcal{K}$, we have

$$\begin{aligned}
&\lim_{s\to 0+} \frac{T(\mu(t))x \ominus_g T(s)x}{\mu(t)-s}\\
&= \lim_{s\to 0+} \frac{T(h)x \ominus_g T(s)x}{h-s}\\
&= \frac{T(h)x \ominus_g Ix}{h}\\
&= \frac{(I+h\widetilde{A})x \ominus_g x}{h} = \widetilde{A}x.
\end{aligned}$$

**Lemma 4.7.** *Let $\widetilde{A} \in L(\mathcal{K})$ and $\widetilde{A}^0 = I$, $\widetilde{A}^{k+1} = \widetilde{A}^k\widetilde{A}$, $k = 0, 1, 2, \ldots$. Then the sequence of operators $\{S_n(t)\}$, $t \in \mathbb{R}_0^+$, is a Cauchy sequence in $L(\mathcal{K})$, where $S_n(t) = \sum_{k=0}^{n} \frac{t^k}{k!} \cdot \widetilde{A}^k$.*

***Proof.*** It is a direct consequence of Theorem 3.9 in [4]. □

In view of the completeness of $L(\mathcal{K})$ and Lemma 4.3, there exists $T(t) \in L(\mathcal{K})$ such that the sequence of operators $\{S_n(t)\}$ converges to $T(t)$ for each $t \in \mathbb{R}_0^+$. Formally, we denote $T(t)$ by

$$e^{t\cdot\widetilde{A}} \triangleq \sum_{k=0}^{\infty} \frac{t^k}{k!} \cdot \widetilde{A}^k.$$

**Lemma 4.8.** *Let $\mathbb{T} = \mathbb{R}_0^+$ and $\widetilde{A} \in L(\mathcal{K})$. Define $T(t) = e^{t\cdot\widetilde{A}}$, $t \in \mathbb{T}$, then*

(i) *$T(t+s) = T(t)T(s)$ for all $t, s \in \mathbb{T}$;*

(ii) *$\lim_{s\to 0+} \frac{T(s)x \ominus_g x}{s} = \widetilde{A}x$ for each $x \in \mathcal{K}$.*

***Proof.*** (i) By Theorem 3.9 (ii) in [4], it is obvious.
(ii) According to Proposition 5 in [7], this result can be proved in a similar way as in Theorem 3.9 in [4]. □

**Example 4.9.** Let $\mathbb{T} = \mathbb{R}_0^+$ and $\widetilde{A} \in L(\mathcal{K})$. Then $\widetilde{A}$ is the generator of $T(t) = e^{t\cdot\widetilde{A}}$ for $t \in \mathbb{R}_0^+$. In fact, by Lemma 4.8, for $x \in \mathcal{K}$, we have

$$\begin{aligned}
&\lim_{s\to 0+} \frac{T(\mu(t))x \ominus_g T(s)x}{\mu(t)-s}\\
&= \lim_{s\to 0+} \frac{T(s)x \ominus_g T(0)x}{s}\\
&= \widetilde{A}x.
\end{aligned}$$

**Lemma 4.10.** *Let $\mathbb{T} \subseteq \mathbb{R}_0^+$ be a semigroup time scale. Then for each $x \in \mathcal{K}$, the function $T(\cdot)x : t \mapsto T(t)x$ is continuous from $\mathbb{T}$ into $\mathcal{K}$.*

***Proof.*** Let $t \in \mathbb{T}$. For all $0 < s \in \mathbb{T}$, we get

$$\begin{aligned} d_H(T(t+s)x, T(t)x) &= d_H(T(t+s)x \ominus_g T(t)x, 0) \\ &= d_H(T(t)T(s)x \ominus_g T(t)x, 0) \\ &= \|T(t)(T(s)x \ominus_g x)\| \\ &\leq \|T(t)\|\|T(s)x \ominus_g x\|. \end{aligned}$$

Letting $s \to 0+$, $\|T(s)x \ominus_g x\| \to 0$, which implies the continuity of $T(t)x$ at $t \in \mathbb{T}$. $\square$

**Theorem 4.11.** *Let $\mathbb{T} \subseteq \mathbb{R}_0^+$ be a semigroup time scale with the constant graininess function $\mu(t) = h$. Suppose that $T$ is a $C_0$-semigroup on $\mathcal{K}$. Then $T(t)$ is $\Delta_{gH}$-differentiable in $t \in \mathbb{T}$, and*

$$T_{gH}^{\Delta}(t) = \widetilde{A}[T(t)].$$

***Proof.*** (i) If $\mu(t) = h > 0$, then $t$ is right-scattered. By Lemma 2.3 in [5], we know $T = h\mathbb{Z}_0^+$. Furthermore, according to Lemma 4.4, $T(t)$ is continuous at $t$, so $T(t)$ is $\Delta_{gH}$-differentiable. From Example 4.6, we can obtain

$$\begin{aligned} T_{gH}^{\Delta}(t) &= \frac{T(\sigma(t)) \ominus_g T(t)}{\mu(t)} \\ &= \frac{T(t+h) \ominus_g T(t)}{h} \\ &= \frac{T(h)T(t) \ominus_g T(t)}{h} \\ &= \widetilde{A}[T(t)]. \end{aligned}$$

(ii) If $\mu(t) = h = 0$, then $t$ is right-dense. In view of Lemma 2.3 in [5], $T = \mathbb{R}_0^+$. Based on Lemma 4.8, we can obtain the above result by using a similar argument as in Theorem 3.9 (iv) in [4]. $\square$

**Definition 4.12.** Let $\mathbb{T} \subseteq \mathbb{R}_0^+$ be a semigroup time scale and let $T$ be a $C_0$-semigroup on $\mathcal{K}$. We say that $T$ is a $l$-monotonic $C_0$-semigroup on $\mathcal{K}$ if the interval-valued function $T(\cdot)x : \mathbb{T} \to \mathcal{K}$ is $l$-monotonic for every $x \in \mathcal{K}$.

**Lemma 4.13.** *Let $\mathbb{T} \subseteq \mathbb{R}_0^+$ be a semigroup time scale and let $T$ be a $C_0$-semigroup on $\mathcal{K}$. Assume that $g : \mathbb{T} \to \mathcal{K}$ is rd-continuous on $\mathbb{T}$. Define $F(t) = \int_0^t T(t-s)g(s)\Delta s$. If $T$ is $l$-nondecreasing on $\mathcal{K}$, then $F$ is also $l$-nondecreasing on $\mathcal{K}$.*

***Proof.*** Let $t_1, t_2 \in \mathbb{T}$ with $t_1 < t_2$. Then, we have

$$\begin{aligned} T(t_2 - t_1)F(t_1) &= \int_0^{t_1} T(t_2 - s)g(s)\Delta s \\ &\subseteq \int_0^{t_1} T(t_2 - s)g(s)\Delta s + \int_{t_1}^{t_2} T(t_2 - s)g(s)\Delta s \\ &= \int_0^{t_2} T(t_2 - s)g(s)\Delta s = F(t_2). \end{aligned}$$

Since $T$ is $l$-nondecreasing on $\mathcal{K}$, it follows that

$$F(t_1) \subseteq T(t_2 - t_1)F(t_1) \subseteq F(t_2),$$

which implies $len(F(t_1)) \leq len(F(t_2))$. Namely, $F$ is $l$-nondecreasing on $\mathcal{K}$. $\square$

**Theorem 4.14.** *Let $\mathbb{T} \subseteq \mathbb{R}_0^+$ be a semigroup time scale with the constant graininess function $\mu(t) = h$. Assume that $x_0 \in \mathcal{K}$ and $g : \mathbb{T} \to \mathcal{K}$ is rd-continuous on $\mathbb{T}$. If $T$ is a $l$-nondecreasing $C_0$-semigroup on $\mathcal{K}$, then*

$$x(t) = T(t)(x_0) + \int_0^t T(t-s)g(s)\Delta s \tag{4.1}$$

*is $\Delta_{gH}$-differentiable on $\mathbb{T}^\kappa$. And then, $x(t)$ satisfies*

$$\begin{cases} x^{\Delta}_{gH}(t) = \widetilde{A}[x(t)] + T(\mu(t))(g(t)), \\ x(0) = x_0, \end{cases} \qquad t \in \mathbb{T}^\kappa, \tag{4.2}$$

*where the integral (including the integral in Lemma 4.13) for interval-valued functions defined on $[0,t)_{\mathbb{T}}$ is considered in the Riemann sense (the detailed definition can be seen in [7]).*

***Proof.*** For every $\mathbb{T}^\kappa$, we set

$$F(t) = \int_0^t T(t-s)g(s)\Delta s.$$

Since $T$ is a $l$-nondecreasing $C_0$-semigroup on $\mathcal{K}$, by Lemma 4.13, it is easy to know that $F(t)$ is $l$-nondecreasing on $\mathbb{T}^\kappa$. Now, we distinguish two cases.
(i) If $t \in \mathbb{T}^\kappa$ is right-scattered, then we get

$$\begin{aligned} F(\sigma(t)) = F(t+h) &= \int_0^{t+h} T(t-s+h)g(s)\Delta s \\ &= T(h)\Big(\int_0^{t+h} T(t-s)g(s)\Delta s\Big) \\ &= T(h)\Big(F(t) + \int_t^{t+h} T(t-s)g(s)\Delta s\Big) \\ &= T(h)(F(t)) + T(h)\Big(\int_t^{t+h} T(t-s)g(s)\Delta s\Big) \\ &= T(h)(F(t)) + T(h)(hT(0)(g(t))) \\ &= T(h)(F(t)) + hT(h)((g(t))) \end{aligned} \tag{4.3}$$

By Lemma 2.2, it follows from (4.3), Theorems 2.6 and 4.11 that

$$F^{\Delta}_{gH}(t) = \frac{F(t+h) \ominus_g F(t)}{h} = \widetilde{A}[F(t)] + T(h)g(t), \tag{4.4}$$

since $F$ is $l$-nondecreasing. By Theorem 4.11, we know that $x(t)$ is $\Delta_{gH}$-differentiable. Furthermore, we can infer from (4.1), (4.4) and Theorem 4 in [7] that

$$\begin{aligned} x^{\Delta}(t) &= \Big(T(t)(x_0) + \int_0^t T(t-s)g(s)\Delta s\Big)^{\Delta}_{gH} \\ &= \Big(T(t)(x_0)\Big)^{\Delta}_{gH} + \Big(\int_0^t T(t-s)g(s)\Delta s\Big)^{\Delta}_{gH} \\ &= \widetilde{A}[T(t)(x_0)] + \widetilde{A}[F(t)] + T(h)g(t) \\ &= \widetilde{A}[T(t)(x_0) + F(t)] + T(h)g(t) \\ &= \widetilde{A}[x(t)] + T(\mu(t))g(t), \end{aligned}$$

which means that $x(t)$ satisfies (4.2).
(ii) If $t \in \mathbb{T}^\kappa$ is right-dense, the proof is similar to Theorem 3.9 in [4] and so is omitted. □

**Remark 4.15.** From Lemma 4.13, we know that $F$ is $l$-nondecreasing on $\mathbb{T}$ if $T$ is $l$-nondecreasing $C_0$-semigroup on $\mathcal{K}$. Apparently, a question that deserves further consideration is whether $F$ is $l$-monotonic on $\mathbb{T}$ if $T$ is $l$-nonincreasing $C_0$-semigroup on $\mathcal{K}$. Furthermore, if $F$ is $l$-monotonic on $\mathbb{T}$, then we can consider another question from Theorem 4.14. In detail, what is the solution to the initial value problem (4.2) when $T$ is $l$-nonincreasing $C_0$-semigroup on $\mathcal{K}$?

**Acknowledgment.** This work was supported by the National Natural Science Foundation of China (No. 11701425).

## References


[1] M. Bohner and A. Peterson, *Dynamic Equations on Time Scales: An Introduction with Applications*, Birkhäuser Boston Inc., Boston, MA, 2001.

[2] M. Bohner and A. Peterson, *Advances in Dynamic Equations on Time Scales*, Birkhäuser, Boston, 2003.

[3] Y. Chalco-Cano, A. Rufián-Lizana, H. Román-Flores and M.D. Jiménez-Gamero, *Calculus for interval-valued functions using generalized Hukuhara derivative and applications*, Fuzzy Sets Syst. **219**, 49–67, 2013.

[4] C.G. Gal and S.G. Gal, *Semigroups of operators on spaces of fuzzy-number-valued functions with applications to fuzzy differential equations*, J. Fuzzy Math. **13**, 647–682, 2005.

[5] A.E. Hamza and K.M. Oraby, *Semigroups of operators and abstract dynamic equations on time scales*, Appl. Math. Comput. **270**, 334–348, 2013.

[6] S. Hilger, *Ein Makettenkalkül mit Anwendung auf Zentrumsmannigfaltigkeiten*, Ph.D. thesis, Universität Würzburg, 1988.

[7] V. Lupulescu, *Hukuhara differentiability of interval-valued functions and interval differential equations on time scales*, Inform. Sciences, **248**, 50–67, 2013.

[8] S. Markov, *Calculus for interval functions of a real variable*, Computing, **22**, 325–337, 1979.

[9] S. Markov, *On the algebraic properties of convex bodies and some applications*, J. Convex Anal. **7**, 129–166, 2000.

[10] L. Stefanini, *A generalization of Hukuhara difference and division for interval and fuzzy arithmetic*, Fuzzy Sets Syst. **161**, 1564–1584, 2010.

[11] L. Stefanini and B. Bede, *Generalized Hukuhara differentiability of interval-valued functions and interval differential equations*, Nonlinear Anal. **71**, 1311–1328, 2009.

[12] J. Tao and Z. Zhang, *Properties of interval-valued function space under the gH-difference and their application to semi-linear interval differential equations*, Adv. Differ. Equ. **2016**, Article number: 45, 2016.

Hacettepe Journal of Mathematics & Statistics

Hacet. J. Math. Stat.
Volume 50 (1) (2021), 92 – 109
DOI : 10.15672/hujms.638033

Research Article

# Depth and Stanley depth of the edge ideals of the strong product of some graphs

Zahid Iqbal[1]*, Muhammad Ishaq[1], Muhammad Ahsan Binyamin[2]

[1] School of Natural Sciences, National University of Sciences and Technology, Sector H-12, Islamabad 44000, Pakistan

[2] Department of Mathematics, Government College University Faisalabad, Pakistan

## Abstract

In this paper, we study depth and Stanley depth of the edge ideals and quotient rings of the edge ideals, associated with classes of graphs obtained by the strong product of two graphs. We consider the cases when either both graphs are arbitrary paths or one is an arbitrary path and the other is an arbitrary cycle. We give exact formula for values of depth and Stanley depth for some subclasses. We also give some sharp upper bounds for depth and Stanley depth in the general cases.

**Mathematics Subject Classification (2020).** Primary: 13C15, Secondary: 13F20, 05C38, 05E99

**Keywords.** depth, Stanley depth, Stanley decomposition, monomial ideal, edge ideal, strong product of graphs

## 1. Introduction

Let $S := K[x_1, \dots, x_n]$ be the polynomial ring over a field $K$. Let $M$ be a finitely generated $\mathbb{Z}^n$-graded $S$-module. A Stanley decomposition of $M$ is a presentation of $K$-vector space $M$ as a finite direct sum $\mathcal{D} : M = \bigoplus_{i=1}^{r} w_i K[A_i]$, where $w_i \in M$ is a homogeneous element in $M$, $A_i \subseteq \{x_1, \dots, x_n\}$ such that $w_i K[A_i]$ denote the $K$-subspace of $M$, which is generated by all elements $w_i u$, where $u$ is a monomial in $K[A_i]$. The $\mathbb{Z}^n$-graded $K$-subspace $w_i K[A_i] \subset M$ is called a Stanley space of dimension $|A_i|$, if $w_i K[A_i]$ is a free $K[A_i]$-module, where $|A_i|$ denotes the number of indeterminates of $A_i$. Define $\operatorname{sdepth}(\mathcal{D}) = \min\{|A_i| : i = 1, \dots, r\}$, and $\operatorname{sdepth}(M) = \max\{\operatorname{sdepth}(\mathcal{D}) : \mathcal{D}$ is a Stanley decomposition of $M\}$. The number $\operatorname{sdepth}(\mathcal{D})$ is called the Stanley depth of decomposition $\mathcal{D}$ and $\operatorname{sdepth}(M)$ is called the Stanley depth of $M$. For an introduction to Stanley depth, we refer the reader to [7, 10, 23]. Stanley conjectured in [26] that $\operatorname{sdepth}(M) \geq \operatorname{depth}(M)$ for any $\mathbb{Z}^n$-graded $S$-module $M$. This conjecture was disproved by Duval et al. [6]. However, there still looks to be a deep and interesting relationship between depth and Stanley depth, which is yet to be exactly understood. Also it is interesting to find new classes of modules which satisfy Stanley's inequality because in this case we have a lower bound for the Stanley depth.

*Corresponding Author.
Email addresses: 786zahidwarraich@gmail.com (Z. Iqbal), ishaq_maths@yahoo.com (M. Ishaq), ahsanbanyamin@gmail.com (M.A. Binyamin)
Received: 25.10.2019; Accepted: 02.05.2020

Let $I \subset J \subset S$ be monomial ideals, Herzog et al. [11] showed that the invariant Stanley depth of $J/I$ is combinatorial in nature. The strange thing about Stanley depth is that it shares some properties and bounds with homological invariant depth see ([11,15,22,24]). Until now mathematicians are not too much familiar with Stanley depth as it is hard to compute, for computation and some known results we refer the readers to ([1,12,16,17,19]). Let $P_n$ and $C_n$ represent path and cycle respectively on $n$ vertices and $\boxtimes$ represents the strong product of two graphs. The aim of this paper is to study depth and Stanley depth of the edge ideals and quotient ring of the edge ideals associated with classes of graphs $\mathcal{H} := \{P_n \boxtimes P_m : n, m \geq 1\}$ and $\mathcal{K} := \{C_n \boxtimes P_m : n \geq 3, m \geq 1\}$. In Section 3 we compute depth and Stanley depth of quotient ring of edge ideals associated with some subclasses of $\mathcal{H}$ and $\mathcal{K}$. For the monomial ideal $I \subset S$ it is clear that $\operatorname{depth}(I) = \operatorname{depth}(S/I)+1$, this means that once you know about $\operatorname{depth}(S/I)$ then you also know about $\operatorname{depth}(I)$ and vice versa, whereas for Stanley depth this is not the case. So far all examples show that $\operatorname{sdepth}(I) \geq \operatorname{sdepth}(S/I)$, as Herzog conjectured:

**Conjecture 1** ([10, Conjecture 64])**.** Let $I \subset S$ be a monomial ideal then $\operatorname{sdepth}(I) \geq \operatorname{sdepth}(S/I)$.

In Section 4 of this paper, we confirm the above conjecture for the edge ideals associated with some subclasses of $\mathcal{H}$ and $\mathcal{K}$. For recent works on the above conjecture, we refer the reader to [13, 14, 18]. In Section 5, we give sharp upper bounds for depth and Stanley depth of quotient ring of the edge ideals associated to $\mathcal{H}$ and $\mathcal{K}$. In the same section, we also propose some open questions. We gratefully acknowledge the use of the computer algebra system CoCoA ([5]) for our experiments.

## 2. Definitions and notations

In this section, we review some standard terminologies and notations from graph theory and algebra. For more details, one may consult [9, 28]. Let $G := (V(G), E(G))$ be a graph with vertex set $V(G) := \{x_1, x_2, \dots, x_n\}$ and edge set $E(G)$. The edge ideal $I(G)$ associated with $G$ is a squarefree monomial ideal of $S$, that is $I(G) = (x_i x_j : \{x_i, x_j\} \in E(G))$. A graph $G$ on $n \geq 2$ vertices is called a path on $n$ vertices if $E(G) = \{\{x_i, x_{i+1}\} : i = 1, 2 \dots, n-1\}$. We denote a path on $n$ vertices by $P_n$. A graph $G$ on $n \geq 3$ vertices is called a cycle if $E(G) = \{\{x_i, x_{i+1}\} : i = 1, 2, \dots, n-1\} \cup \{\{x_1, x_n\}\}$. A cycle on $n$ vertices is denoted by $C_n$. For vertices $x_i$ and $x_j$ of a graph $G$, the length of a shortest path from $x_i$ to $x_j$ is called the distance between $x_i$ and $x_j$ denoted by $\mathrm{d}_G(x_i, x_j)$. If no such path exists between $x_i$ and $x_j$, then $d_G(x_i, x_j) = \infty$. The diameter of a connected graph $G$ is $\operatorname{diam}(G) := \max\{\mathrm{d}_G(x_i, x_j) : x_i, x_j \in V(G)\}$. For a monomial $u$, $\operatorname{supp}(u) := \{x_i : x_i \mid u\}$.

**Definition 2.1** ([9])**.** The strong product $G_1 \boxtimes G_2$ of graphs $G_1$ and $G_2$ is a graph, with $V(G_1 \boxtimes G_2) = V(G_1) \times V(G_2)$ (the Cartesian product of sets), and for $(v_1, u_1), (v_2, u_2) \in V(G_1 \boxtimes G_2)$, $\{(v_1, u_1), (v_2, u_2)\} \in E(G_1 \boxtimes G_2)$, whenever

- $\{v_1, v_2\} \in E(G_1)$ and $u_1 = u_2$ or
- $v_1 = v_2$ and $\{u_1, u_2\} \in E(G_2)$ or
- $\{v_1, v_2\} \in E(G_1)$ and $\{u_1, u_2\} \in E(G_2)$.

Let $P_1$ denote the null graph on one vertex that is $V(P_1) := \{x_1\}$ and $E(P_1) := \emptyset$. Let $\mathcal{P}_{n,m} := P_n \boxtimes P_m \cong P_m \boxtimes P_n$, if $n = m = 1$, then $\mathcal{P}_{1,1} \cong P_1$, this trivial case is excluded. For $n \geq 3$ and $m \geq 1$, let $\mathcal{C}_{n,m} := C_n \boxtimes P_m \cong P_m \boxtimes C_n$.

**Remark 2.2.** $|V(\mathcal{P}_{n,m})| = nm$, $|E(\mathcal{P}_{n,m})| = 4(n-1)(m-1)+(n-1)+(m-1)$, $|V(\mathcal{C}_{n,m})| = nm$ and $|E(\mathcal{C}_{n,m})| = |E(\mathcal{P}_{n,m})| + 3(m-1) + 1$.

Since both graphs $\mathcal{P}_{n,m}$ and $\mathcal{C}_{n,m}$ are on $nm$ vertices, for the sake of convenience, we label the vertices of $\mathcal{P}_{n,m}$ and $\mathcal{C}_{n,m}$ by using $m$ sets of variables $\{x_{1j}, x_{2j}, \dots, x_{nj}\}$ where

$1 \leq j \leq m$. We set $S_{n,m} := K[\cup_{j=1}^{m}\{x_{1j}, x_{2j}, \ldots, x_{nj}\}]$. For examples of $\mathcal{P}_{n,m}$ and $\mathcal{C}_{n,m}$ see Fig 1.

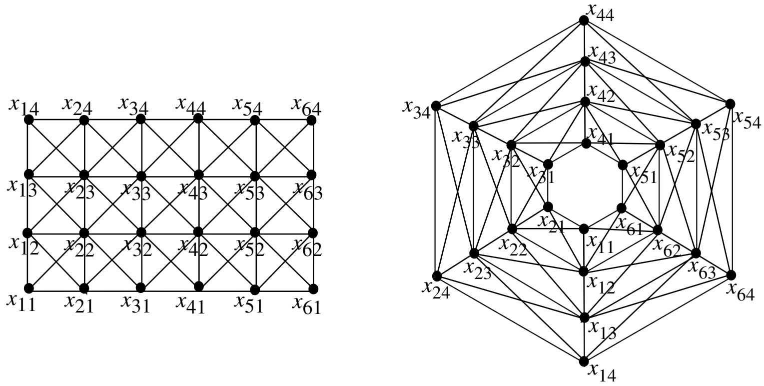



**Figure 1.** From left to right; $\mathcal{P}_{6,4}$ and $\mathcal{C}_{6,4}$.

**Remark 2.3.** Let $\mathcal{G}(I)$ denote the unique minimal set of monomial generators of the monomial ideal $I$.

(1) For positive integers $m, n$ such that $m$ and $n$ are not equal to 1 simultaneously, the minimal set of monomial generators of the edge ideal of $\mathcal{P}_{n,m}$ is given as:

$$\mathcal{G}(I(\mathcal{P}_{n,m})) = \cup_{i=1}^{n-1}\big\{ \cup_{j=1}^{m-1}\{x_{ij}x_{i(j+1)}, x_{ij}x_{(i+1)(j+1)}, x_{ij}x_{(i+1)j}, x_{(i+1)j}x_{i(j+1)}, x_{nj}x_{n(j+1)}\}, \\ x_{im}x_{(i+1)m}\big\}.$$

(2) For $n \geq 3$, $m \geq 1$, the minimal set of monomial generators for $I(\mathcal{C}_{n,m})$ is:

$$\mathcal{G}(I(\mathcal{C}_{n,m})) = \mathcal{G}(I(\mathcal{P}_{n,m})) \cup \big\{ \cup_{j=1}^{m-1}\{x_{1j}x_{n(j+1)}, x_{1j}x_{nj}, x_{1(j+1)}x_{nj}\}, x_{1m}x_{nm}\big\}.$$

(3) $\mathcal{P}_{n,1} \cong P_n$ and $\mathcal{C}_{n,1} \cong C_n$.

(4) For $n, m \geq 1$, $\mathcal{P}_{n,m} \cong \mathcal{P}_{m,n}$, so without loss of generality the strong product of two paths can be represented as $\mathcal{P}_{n,m}$ with $m \leq n$. Thus in some proofs by induction on $n$, whenever we are reduced to the case where we have $\mathcal{P}_{n',m}$ with $n' < m$, after a suitable relabeling of vertices we have $\mathcal{P}_{n',m} \cong \mathcal{P}_{m,n'}$. Therefore, we can simply replace $I(\mathcal{P}_{n',m})$ by $I(\mathcal{P}_{m,n'})$ and $S_{n',m}/I(\mathcal{P}_{n',m})$ by $S_{m,n'}/I(\mathcal{P}_{m,n'})$.

The method of Herzog et al. [11] for determining the Stanley depth of modules of the type $M = J/I$ (where $I \subset J \subset S$ are monomial ideals) using posets can be summarized in the following way. We define a natural partial order on $\mathbb{N}^n$ as follows: $a \leq b$ if and only if $a(l) \leq b(l)$ for $l = 1, \ldots, n$. Note that $x^a \mid x^b$ if and only if $a \leq b$. Here for $c \in \mathbb{N}^n$, $x^c$ denote the monomial $x_1^{c(1)}x_2^{c(2)}\cdots x_n^{c(n)}$. Let $J = (x^{a_1}, x^{a_2}, \ldots, x^{a_r})$ and $I = (x^{b_1}, x^{b_2}, \ldots, x^{b_t})$ where $a_i, b_j \in \mathbb{N}^n$. Let $h \in \mathbb{N}^n$ such that $h(l) = \max\{a_i(l), b_j(l)) : 1 \leq i \leq r, 1 \leq j \leq t\}$ (the component-wise maximum of the $a_i$ and $b_j$). Then the characteristic poset of $J/I$ with respect to $h$, denoted $P^h_{J/I}$, is the induced subposet of $\mathbb{N}^n$ with ground set

$$\{c \in \mathbb{N}^n | c \leq h, \text{ there is } i \text{ such that } c \geq a_i, \text{ and for all } j,\ c \ngeq b_j\}.$$

Let $x, y \in P^h_{J/I}$, $\alpha := [x, y] = \{z \in P^h_{J/I} : x \leq z \leq y\}$ be a subset of $P^h_{J/I}$ called interval and $\mathbf{P}$ be a partition of $P^h_{J/I}$ into intervals. Let $Z_\alpha := \{l : y(l) = h(l)\}$, define the Stanley depth of a partition $\mathbf{P}$ to be $\operatorname{sdepth}(\mathbf{P}) := \min_{\alpha\in\mathbf{P}} |Z_\alpha|$ and the Stanley depth of the poset $P^h_{J/I}$ to be $\operatorname{sdepth}(P^h_{J/I}) := \max_{\mathbf{P}} \operatorname{sdepth}(\mathbf{P})$, where the maximum is taken over all partitions $\mathbf{P}$ of $P^h_{J/I}$. Herzog et al. showed in [11] that $\operatorname{sdepth}(J/I) = \operatorname{sdepth}(P^h_{J/I})$. By considering all partitions of the characteristic poset, this correspondence provides an algorithm (albeit inefficient) to find the Stanley depth of $J/I$. Now we recall some known results that are heavily used in this paper.

**Lemma 2.4.** *(Depth Lemma) If $0 \to U \to M \to N \to 0$ is a short exact sequence of modules over a local ring $S$, or a Noetherian graded ring with local $S_0$, then*

(1) $\operatorname{depth}(M) \geq \min\{\operatorname{depth}(N), \operatorname{depth}(U)\}$.
(2) $\operatorname{depth}(U) \geq \min\{\operatorname{depth}(M), \operatorname{depth}(N) + 1\}$.
(3) $\operatorname{depth}(N) \geq \min\{\operatorname{depth}(U) - 1, \operatorname{depth}(M)\}$.

**Lemma 2.5** ([24, Lemma 2.2]). *Let $0 \to U \to V \to W \to 0$ be a short exact sequence of $\mathbb{Z}^n$-graded $S$-modules. Then* $\operatorname{sdepth}(V) \geq \min\{\operatorname{sdepth}(U), \operatorname{sdepth}(W)\}$.

**Remark 2.6.** Let $I \subset S$ be a monomial ideal. Then for $1 \leq i \leq n$ with $x_i \notin I$, the short exact sequence

$$0 \longrightarrow S/(I : x_i) \xrightarrow{\cdot x_i} S/I \longrightarrow S/(I, x_i) \longrightarrow 0,$$

implies that

$$\operatorname{depth}(S/I) \geq \min\{\operatorname{depth}(S/(I : x_i)), \operatorname{depth}(S/(I, x_i))\},$$
$$\operatorname{sdepth}(S/I) \geq \min\{\operatorname{sdepth}(S/(I : x_i)), \operatorname{sdepth}(S/(I, x_i))\}.$$

This will be used frequently throughout the paper.

**Lemma 2.7** ([11, Lemma 3.6]). *Let $I \subset J$ be monomial ideals of $S$ and $\bar{S} = S[x_{n+1}]$ be a polynomial ring in $n+1$ variables. Then*

$$\operatorname{depth}(J\bar{S}/I\bar{S}) = \operatorname{depth}(JS/IS) + 1 \quad \textit{and} \quad \operatorname{sdepth}(J\bar{S}/I\bar{S}) = \operatorname{sdepth}(JS/IS) + 1.$$

**Corollary 2.8** ([24, Corollary 1.3]). *Let $J \subset S$ be a monomial ideal. Then* $\operatorname{depth}(S/J) \leq \operatorname{depth}(S/(J : v))$ *for all monomials $v \notin J$.*

**Proposition 2.9** ([2, Proposition 2.7]). *Let $J \subset S$ be a monomial ideal. Then for all monomials $v \notin J$* $\operatorname{sdepth}(S/J) \leq \operatorname{sdepth}(S/(J : v))$.

Let $q \in \mathbb{Q}$, then $\lceil q \rceil$ denote the smallest integer greater than or equal to $q$, and $\lfloor q \rfloor$ denote the greatest integer less than or equal to $q$.

**Theorem 2.10** ([21, Theorem 2.3]). *Let $I \subset S$ be a monomial ideal of $S$ and $m$ be the number of minimal monomial generators of $I$, then* $\operatorname{sdepth}(I) \geq \max\{1, n - \lfloor \frac{m}{2} \rfloor\}$.

**Corollary 2.11** ([8, Corollary 3.2]). *Let $G$ be a connected graph of diameter $d \geq 1$ and let $I = I(G)$. Then* $\operatorname{depth}(S/I) \geq \lceil \frac{d+1}{3} \rceil$.

**Theorem 2.12** ([8, Theorem 4.18]). *Let $G$ be a graph with $p$ connected components, $I = I(G)$, and let $d = d(G)$ be the diameter of $G$. Then, for $1 \leq t \leq 3$ we have*

$$\operatorname{sdepth}(S/I^t) \geq \lceil \frac{d - 4t + 5}{3} \rceil + p - 1.$$

**Corollary 2.13.** *Let $G$ be a connected graph of diameter $d \geq 1$ and let $I = I(G)$. Then* $\operatorname{sdepth}(S/I) \geq \lceil \frac{d+1}{3} \rceil$.

## 3. Depth and Stanley depth of cyclic modules associated to $\mathcal{P}_{n,m}$ and $\mathcal{C}_{n,m}$ when $1 \leq m \leq 3$

Let $n \geq 2$ and $1 \leq i \leq n$, for convenience we take $x_i := x_{i1}$, $y_i := x_{i2}$ and $z_i := x_{i3}$, see Figures 2 and 3. We set $S_{n,1} := K[x_1, x_2, \dots, x_n]$, $S_{n,2} := K[x_1, x_2, \dots x_n, y_1, y_2, \dots, y_n]$ and $S_{n,3} := K[x_1, x_2, \dots x_n, y_1, y_2, \dots, y_n, z_1, z_2, \dots, z_n]$. Clearly $\mathcal{P}_{n,1} \cong P_n$ and $\mathcal{C}_{n,1} \cong C_n$, the minimal sets of monomial generators of the edge ideals of $\mathcal{P}_{n,2}$, $\mathcal{P}_{n,3}$, $\mathcal{C}_{n,2}$ and $\mathcal{C}_{n,3}$ are given as:

$$\mathcal{G}(I(\mathcal{P}_{n,2})) = \cup_{i=1}^{n-1}\{x_iy_i, x_iy_{i+1}, x_ix_{i+1}, x_{i+1}y_i, y_iy_{i+1}\} \cup \{x_ny_n\},$$

$$\mathcal{G}(I(\mathcal{P}_{n,3})) = \cup_{i=1}^{n-1}\{x_iy_i, x_iy_{i+1}, x_ix_{i+1}, x_{i+1}y_i, y_iy_{i+1}, y_iz_i, y_iz_{i+1}, y_{i+1}z_i, z_iz_{i+1}\} \cup \{x_ny_n, y_nz_n\},$$

$$\mathcal{G}(I(\mathcal{C}_{n,2})) = \mathcal{G}(I(\mathcal{P}_{n,2})) \cup \{x_1y_n, x_1x_n, y_1x_n, y_1y_n\} \text{ and}$$

$$\mathcal{G}(I(\mathcal{C}_{n,3})) = \mathcal{G}(I(\mathcal{P}_{n,3})) \cup \{x_1y_n, x_1x_n, y_1x_n, y_1y_n, y_1z_n, z_1y_n, z_1z_n\}.$$

In this section, we compute depth and Stanley depth of the cyclic modules $S_{n,m}/I(\mathcal{P}_{n,m})$ and $S_{n,m}/I(\mathcal{C}_{n,m})$, when $m = 1, 2, 3$.

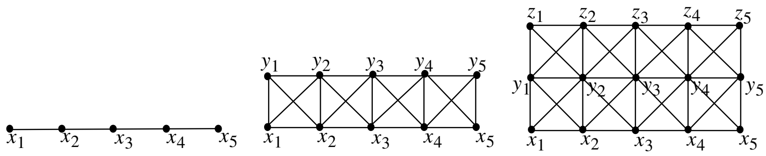



**Figure 2.** From left to right; $\mathcal{P}_{5,1}$, $\mathcal{P}_{5,2}$ and $\mathcal{P}_{5,3}$.

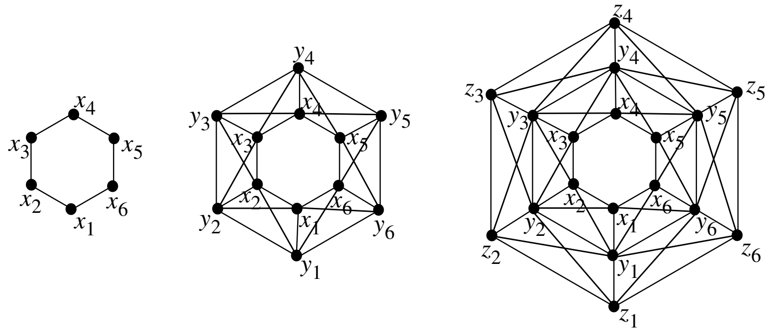



**Figure 3.** From left to right; $\mathcal{C}_{6,1}$, $\mathcal{C}_{6,2}$ and $\mathcal{C}_{6,3}$.

**Remark 3.1.** Note that for $n \geq 2$, $S_{n,1}/I(\mathcal{P}_{n,1}) \cong S/I(P_n)$, thus by [20, Lemma 2.8] and [27, Lemma 4] $\operatorname{depth}(S_{n,1}/I(\mathcal{P}_{n,1})) = \operatorname{sdepth}(S_{n,1}/I(\mathcal{P}_{n,1})) = \lceil \frac{n}{3} \rceil$. Let $n \geq 3$, then $S_{n,1}/I(\mathcal{C}_{n,1}) \cong S/I(C_n)$, and by [4, Propositions 1.3,1.8] $\operatorname{depth}(S_{n,1}/I(\mathcal{C}_{n,1})) = \lceil \frac{n-1}{3} \rceil \leq \operatorname{sdepth}(S_{n,1}/I(\mathcal{C}_{n,1})) \leq \lceil \frac{n}{3} \rceil$.

**Lemma 3.2.** *For $n \geq 1$ and $m = 2, 3$,* $\operatorname{depth}(S_{n,m}/I(\mathcal{P}_{n,m})) = \operatorname{sdepth}(S_{n,m}/I(\mathcal{P}_{n,m})) = \lceil \frac{n}{3} \rceil$.

***Proof.*** If $n = 1$, then proof follows from Remark 3.1. Let $n \geq 2$. First we prove the result for depth. If $(n, m) \in \{(2, 2), (3, 2), (3, 3)\}$ then the result is trivial. Let $n \geq 4$. Since $\operatorname{diam}(\mathcal{P}_{n,m}) = n - 1$, thus by Corollary 2.11 $\operatorname{depth}(S_{n,m}/I(\mathcal{P}_{n,m})) \geq \lceil \frac{n}{3} \rceil$. Now we prove that $\operatorname{depth}(S_{n,m}/I(\mathcal{P}_{n,m})) \leq \lceil \frac{n}{3} \rceil$, we prove this inequality by induction on $n$. Since $y_{n-1} \notin I(\mathcal{P}_{n,m})$, then by Corollary 2.8

$$\operatorname{depth}(S_{n,m}/I(\mathcal{P}_{n,m})) \leq \operatorname{depth}(S_{n,m}/(I(\mathcal{P}_{n,m}) : y_{n-1})).$$

As we can see that $S_{n,m}/(I(\mathcal{P}_{n,m}) : y_{n-1}) \cong S_{n-3,m}/I(\mathcal{P}_{n-3,m})[y_{n-1}]$, therefore by induction and Lemma 2.7 $\operatorname{depth}(S_{n,m}/(I(\mathcal{P}_{n,m}) : y_{n-1})) = \lceil \frac{n-3}{3} \rceil + 1 = \lceil \frac{n}{3} \rceil$. This completes the proof for depth.

Now we prove the result for Stanley depth. If $n = m = 2$, then $I(\mathcal{P}_{2,2})$ is a squarefree Veronese ideal of degree 2. Thus by [3, Theorem 1.1] we have $\operatorname{sdepth}(S_{n,2}/I(\mathcal{P}_{n,2})) = 1$, as required. If $n = 3$ and $m = 2$ or 3, then $\operatorname{diam}(\mathcal{P}_{3,m}) = 2$, thus by Corollary 2.13, we have $\operatorname{sdepth}(S_{3,m}/I(\mathcal{P}_{3,m})) \geq 1$. By Proposition 2.9 we have $\operatorname{sdepth}(S_{3,m}/I(\mathcal{P}_{3,m})) \leq \operatorname{sdepth}(S_{3,m}/(I(\mathcal{P}_{3,m}) : y_2))$ it is easy to see that $S_{3,m}/(I(\mathcal{P}_{3,m}) : y_2) \cong K[y_2]$, therefore $\operatorname{sdepth}(S_{3,m}/I(\mathcal{P}_{3,m})) \leq 1$, thus $\operatorname{sdepth}(S_{3,m}/I(\mathcal{P}_{3,m})) = 1$. Let $n \geq 4$, using Corollary 2.13 instead of Corollary 2.11 and Proposition 2.9 instead of Corollary 2.8, the proof for depth also works for Stanley depth. $\square$

**Theorem 3.3.** *For $n \geq 3$,* $\operatorname{sdepth}(S_{n,2}/I(\mathcal{C}_{n,2})) \geq \operatorname{depth}(S_{n,2}/I(\mathcal{C}_{n,2})) = \lceil \frac{n-1}{3} \rceil$.

***Proof.*** We first prove that $\operatorname{depth}(S_{n,2}/I(\mathcal{C}_{n,2})) = \lceil \frac{n-1}{3} \rceil$. For $n = 3, 4$ the result is trivial. For $n \geq 5$ using Remark 2.6 one has

$$\operatorname{depth}(S_{n,2}/I(\mathcal{C}_{n,2})) \geq \min\{\operatorname{depth}(S_{n,2}/(I(\mathcal{C}_{n,2}) : x_n)), \operatorname{depth}(S_{n,2}/(I(\mathcal{C}_{n,2}), x_n))\}.$$

$$(I(\mathcal{C}_{n,2}) : x_n) = \big(\cup_{i=2}^{n-3}\{x_iy_i, x_iy_{i+1}, x_ix_{i+1}, x_{i+1}y_i, y_iy_{i+1}\}, x_{n-2}y_{n-2}, x_1, y_1, x_{n-1}, y_{n-1}, y_n\big).$$

After renumbering the variables, we have $S_{n,2}/(I(\mathcal{C}_{n,2}) : x_n) \cong S_{n-3,2}/I(\mathcal{P}_{n-3,2})[x_n]$. Thus by Lemmas 3.2 and 2.7 $\operatorname{depth}(S_{n,2}/(I(\mathcal{C}_{n,2}) : x_n)) = \lceil \frac{n-3}{3} \rceil + 1 = \lceil \frac{n}{3} \rceil$. Let $J$ be a monomial ideal such that;

$$J = (I(\mathcal{C}_{n,2}), x_n) = \big(\cup_{i=1}^{n-2}\{x_iy_i, x_iy_{i+1}, x_ix_{i+1}, x_{i+1}y_i, y_iy_{i+1}\}, x_{n-1}y_{n-1}, x_n, x_{n-1}y_n,$$
$$y_{n-1}y_n, y_1y_n, x_1y_n\big) = (I(\mathcal{P}_{n-1,2}), x_n, x_{n-1}y_n, y_{n-1}y_n, y_1y_n, x_1y_n).$$

By Remark 2.6 we have $\operatorname{depth}(S_{n,2}/J) \geq \min\{\operatorname{depth}(S_{n,2}/(J : y_n)), \operatorname{depth}(S_{n,2}/(J, y_n))\}$. As $(J, y_n) = (I(\mathcal{P}_{n-1,2}), x_n, y_n)$ and $S_{n,2}/(J, y_n) \cong S_{n-1,2}/I(\mathcal{P}_{n-1,2})$. Therefore by Lemma 3.2 $\operatorname{depth}(S_{n,2}/(J, y_n)) = \lceil \frac{n-1}{3} \rceil$. Also

$$(J : y_n) = \big(\cup_{i=2}^{n-3}\{x_iy_i, x_iy_{i+1}, x_ix_{i+1}, x_{i+1}y_i, y_iy_{i+1}\}, x_{n-2}y_{n-2}, x_1, y_1, x_{n-1}, y_{n-1}, x_n\big).$$

After renumbering the variables, we get $S_{n,2}/(J : y_n) \cong S_{n-3,2}/I(\mathcal{P}_{n-3,2})[y_n]$. Therefore by Lemmas 3.2 and 2.7 $\operatorname{depth}(S_{n,2}/(J : y_n)) = \lceil \frac{n-3}{3} \rceil + 1 = \lceil \frac{n}{3} \rceil$. If $n \equiv 0 (\operatorname{mod} 3)$ or $n \equiv 2 (\operatorname{mod} 3)$, then $\operatorname{depth}(S_{n,2}/(I(\mathcal{C}_{n,2}) : x_n)) = \lceil \frac{n}{3} \rceil = \lceil \frac{n-1}{3} \rceil \leq \operatorname{depth}(S_{n,2}/(I(\mathcal{C}_{n,2}), x_n))$, thus Depth Lemma implies $\operatorname{depth}(S_{n,2}/I(\mathcal{C}_{n,2})) = \lceil \frac{n-1}{3} \rceil$, as required. Now for $n \equiv 1 (\operatorname{mod} 3)$, assume that $n \geq 7$, then we have the following $S_{n,2}$-module isomorphism:

$$(I(\mathcal{C}_{n,2}) : x_n)/I(\mathcal{C}_{n,2}) \cong x_1 \frac{K[x_3, \dots, x_{n-1}, y_3, \dots, y_{n-1}]}{\big(\bigcup_{i=3}^{n-2}\{x_iy_i, x_iy_{i+1}, x_ix_{i+1}, x_{i+1}y_i, y_iy_{i+1}\}, x_{n-1}y_{n-1}\big)}[x_1]$$
$$\oplus\, y_1 \frac{K[x_3, \dots, x_{n-1}, y_3, \dots, y_{n-1}]}{\big(\bigcup_{i=3}^{n-2}\{x_iy_i, x_iy_{i+1}, x_ix_{i+1}, x_{i+1}y_i, y_iy_{i+1}\}, x_{n-1}y_{n-1}\big)}[y_1]$$
$$\oplus\, y_n \frac{K[x_2, \dots, x_{n-2}, y_2, \dots, y_{n-2}]}{\big(\bigcup_{i=2}^{n-3}\{x_iy_i, x_iy_{i+1}, x_ix_{i+1}, x_{i+1}y_i, y_iy_{i+1}\}, x_{n-2}y_{n-2}\big)}[y_n]$$
$$\oplus\, x_{n-1} \frac{K[x_2, \dots, x_{n-3}, y_2, \dots, y_{n-3}]}{\big(\bigcup_{i=2}^{n-4}\{x_iy_i, x_iy_{i+1}, x_ix_{i+1}, x_{i+1}y_i, y_iy_{i+1}\}, x_{n-3}y_{n-3}\big)}[x_{n-1}]$$
$$\oplus\, y_{n-1} \frac{K[x_2, \dots, x_{n-3}, y_2, \dots, y_{n-3}]}{\big(\bigcup_{i=2}^{n-4}\{x_iy_i, x_iy_{i+1}, x_ix_{i+1}, x_{i+1}y_i, y_iy_{i+1}\}, x_{n-3}y_{n-3}\big)}[y_{n-1}].$$

Indeed, if $u \in (I(\mathcal{C}_{n,2}) : x_n)$ is a monomial such that $u \notin I(\mathcal{C}_{n,2})$. Then $u$ is divisible by at most one variable from the set $\{x_1, y_1, y_n, x_{n-1}, y_{n-1}\}$, if $u$ is divisible by two or more variables from $\{x_1, y_1, y_n, x_{n-1}, y_{n-1}\}$ then $u \in I(\mathcal{C}_{n,2})$, a contradiction. If $x_1 \mid u$ then $u = x_1^a w$ with $a \geq 1$, since $u \notin I(\mathcal{C}_{n,2})$ it follows that $w \in S' := K[x_3, \dots, x_{n-1}, y_3, \dots, y_{n-1}]$ and $w \notin J := \big(\bigcup_{i=3}^{n-2}\{x_iy_i, x_iy_{i+1}, x_ix_{i+1}, x_{i+1}y_i, y_iy_{i+1}\}, x_{n-1}y_{n-1}\big)$, thus $u \in x_1(S'/J)[x_1]$ which is the first summand in the direct sum. Let $S'' := S'[x_1]$ then $x_1(S'/J)[x_1] \cong x_1(S''/JS'')$, it is easy to see that $x_1$ is regular on $S''/JS''$, therefore we have the $S''$-module isomorphism $x_1(S''/JS'') = (S''/JS'')$. After a suitable renumbering of variables we have $(S''/JS'') \cong S_{n-3,2}/I(\mathcal{P}_{n-3,2})[x_n]$. If $y_1 \mid u$, then we get the second summand and if $y_n \mid u$ then we get the third summand. Proceeding in the same way one can easily show that these two summands are also isomorphic to $S_{n-3,2}/I(\mathcal{P}_{n-3,2})[x_n]$. If $x_{n-1} \mid u$ then we get the forth summand and if $y_{n-1} | u$ then we get the last summand. Similarly one can show that the last two summands are isomorphic to $S_{n-4,2}/I(\mathcal{P}_{n-4,2})[x_n]$. Thus by Lemmas 3.2 and 2.7, we have

$$\operatorname{depth}(I(\mathcal{C}_{n,2}) : x_n)/I(\mathcal{C}_{n,2})) = \min\{\lceil \frac{n-3}{3} \rceil + 1, \lceil \frac{n-4}{3} \rceil + 1\} = \lceil \frac{n-1}{3} \rceil.$$

Now by using Depth Lemma on the following short exact sequence we get the required result.

$$0 \longrightarrow (I(\mathcal{C}_{n,2}) : x_n)/I(\mathcal{C}_{n,2}) \xrightarrow{\cdot x_n} S_{n,2}/I(\mathcal{C}_{n,2}) \longrightarrow S_{n,2}/(I(\mathcal{C}_{n,2}) : x_n) \longrightarrow 0.$$

Now we prove the result for Stanley depth. If $n = 3$, then $I(\mathcal{C}_{3,2})$ is a squarefree Veronese ideal of degree 2. Thus by [3, Theorem 1.1] $\operatorname{sdepth}(S_{3,2}/I(\mathcal{C}_{3,2})) = 1$, as required. If $n = 4$, then by using [11] we have the following Stanley decomposition

$$S_{4,2}/I(\mathcal{C}_{4,2}) = K[x_1, x_3] \oplus y_1K[x_3, y_1] \oplus x_2K[x_2, x_4] \oplus y_2K[y_2, y_4] \oplus$$
$$y_3K[x_1, y_3] \oplus x_4K[x_4, y_2] \oplus y_4K[x_2, y_4] \oplus y_1y_3K[y_1, y_3].$$

Thus $\operatorname{sdepth}(S_{4,2}/I(\mathcal{C}_{4,2})) \geq 2$. For upper bound by Proposition 2.9 we have

$$\operatorname{sdepth}(S_{4,2}/I(\mathcal{C}_{4,2})) \leq \operatorname{sdepth}(S_{4,2}/(I(\mathcal{C}_{4,2}) : x_1x_3)),$$

since $S_{4,2}/(I(\mathcal{C}_{4,2}) : x_1x_3) \cong K[x_1, x_3]$, therefore $\operatorname{sdepth}(S_{4,2}/I(\mathcal{C}_{4,2})) \leq 2$, thus we get $\operatorname{sdepth}(S_{4,2}/I(\mathcal{C}_{4,2})) = 2$. Let $n \geq 5$, using Remark 2.6 we have

$$\operatorname{sdepth}(S_{n,2}/I(\mathcal{C}_{n,2})) \geq$$
$$\min\{\operatorname{sdepth}(S_{n,2}/(I(\mathcal{C}_{n,2}) : x_n)), \operatorname{sdepth}(S_{n,2}/(J : y_n)), \operatorname{sdepth}(S_{n,2}/(J, y_n))\} \geq \lceil \frac{n-1}{3} \rceil.$$

□

**Corollary 3.4.** *For $n \geq 3$, $\lceil \frac{n-1}{3} \rceil \leq \operatorname{sdepth}(S_{n,2}/I(\mathcal{C}_{n,2})) \leq \lceil \frac{n}{3} \rceil$.*

***Proof.*** Since $I(\mathcal{C}_{3,2})$ is a squarefree Veronese ideal, by using [3, Theorem 1.1], it follows that $\operatorname{sdepth}(S_{3,2}/I(\mathcal{C}_{3,2})) = 1$. For $n \geq 4$, by Proposition 2.9 $\operatorname{sdepth}(S_{n,2}/I(\mathcal{C}_{n,2})) \leq \operatorname{sdepth}(S_{n,2}/(I(\mathcal{C}_{n,2}) : x_n))$. Since $S_{n,2}/(I(\mathcal{C}_{n,2}) : x_n) \cong S_{n-3,2}/I(\mathcal{P}_{n-3,2})[x_n]$, using Lemmas 3.2 and 2.7, we have $\operatorname{sdepth}(S_{n,2}/(I(\mathcal{C}_{n,2}) : x_n)) = \lceil \frac{n-3}{3} \rceil + 1 = \lceil \frac{n}{3} \rceil$. □

For $n \geq 2$ we define a supergraph of $\mathcal{P}_{n,3}$ denoted by $\mathcal{P}^\star_{n,3}$ with the set of vertices $V(\mathcal{P}^\star_{n,3}) := V(\mathcal{P}_{n,3}) \cup \{z_{n+1}\}$ and edge set $E(\mathcal{P}^\star_{n,3}) := E(\mathcal{P}_{n,3}) \cup \{z_nz_{n+1}, y_nz_{n+1}\}$. Also we define a supergraph of $\mathcal{P}^\star_{n,3}$ denoted by $\mathcal{P}^{\star\star}_{n,3}$ with the set of vertices $V(\mathcal{P}^{\star\star}_{n,3}) := V(\mathcal{P}^\star_{n,3}) \cup \{z_{n+2}\}$ and edge set $E(\mathcal{P}^{\star\star}_{n,3}) := E(\mathcal{P}^\star_{n,3}) \cup \{z_1z_{n+2}, y_1z_{n+2}\}$. For examples of $\mathcal{P}^\star_{n,3}$ and $\mathcal{P}^{\star\star}_{n,3}$ see Fig. 4. Let $S^\star_{n,3} := S_{n,3}[z_{n+1}]$ and $S^{\star\star}_{n,3} := S_{n,3}[z_{n+1}, z_{n+2}]$ then we have the following lemma:

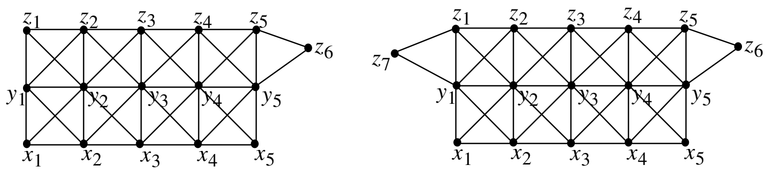



**Figure 4.** From left to right; $\mathcal{P}^\star_{5,3}$ and $\mathcal{P}^{\star\star}_{5,3}$.

**Lemma 3.5.** *For $n \geq 2$,*

(a) $\operatorname{depth}(S^\star_{n,3}/I(\mathcal{P}^\star_{n,3})) = \operatorname{sdepth}(S^\star_{n,3}/I(\mathcal{P}^\star_{n,3})) = \lceil \frac{n+1}{3} \rceil$.
(b) $\operatorname{depth}(S^{\star\star}_{n,3}/I(\mathcal{P}^{\star\star}_{n,3})) = \operatorname{sdepth}(S^{\star\star}_{n,3}/I(\mathcal{P}^{\star\star}_{n,3})) = \lceil \frac{n+2}{3} \rceil$.

***Proof.*** (a). First we prove the result for depth. Since $\operatorname{diam}(\mathcal{P}^\star_{n,3}) = n$, then by Corollary 2.11 we have $\operatorname{depth}(S^\star_{n,3}/I(\mathcal{P}^\star_{n,3})) \geq \lceil \frac{n+1}{3} \rceil$. Now we prove the reverse inequality, if $n = 2$ then the result is trivial. For $n \geq 3$, as $y_n \notin I(\mathcal{P}^\star_{n,3})$ so by Corollary 2.8 $\operatorname{depth}(S^\star_{n,3}/I(\mathcal{P}^\star_{n,3})) \leq \operatorname{depth}(S^\star_{n,3}/(I(\mathcal{P}^\star_{n,3}) : y_n))$. We have $S^\star_{n,3}/(I(\mathcal{P}^\star_{n,3}) : y_n) \cong (S_{n-2,3}/I(\mathcal{P}_{n-2,3}))[y_n]$. By Lemmas 3.2 and 2.7 $\operatorname{depth}(S^\star_{n,3}/(I(\mathcal{P}^\star_{n,3}) : y_n)) = \lceil \frac{n-2}{3} \rceil + 1 = \lceil \frac{n+1}{3} \rceil$. Thus $\operatorname{depth}(S^\star_{n,3}/I(\mathcal{P}^\star_{n,3})) \leq \lceil \frac{n+1}{3} \rceil$. Proof for Stanley depth is similar by using

Proposition 2.9 and Corollary 2.13.
(b). Clearly $\operatorname{diam}(\mathcal{P}^{\star\star}_{n,3}) = n+1$, by Corollary 2.11 we have $\operatorname{depth}(S^{\star\star}_{n,3}/I(\mathcal{P}^{\star\star}_{n,3})) \geq \lceil\frac{n+2}{3}\rceil$. Now we prove the reverse inequality, it is true when $n = 2, 3$. For $n \geq 4$, as $y_n \notin I(\mathcal{P}^{\star\star}_{n,3})$ so by Corollary 2.8 $\operatorname{depth}(S^{\star\star}_{n,3}/I(\mathcal{P}^{\star\star}_{n,3})) \leq \operatorname{depth}(S^{\star\star}_{n,3}/(I(\mathcal{P}^{\star\star}_{n,3}) : y_n))$. Since $S^{\star\star}_{n,3}/(I(\mathcal{P}^{\star\star}_{n,3}) : y_n) \cong (S^{\star}_{n-2,3}/I(\mathcal{P}^{\star}_{n-2,3}))[y_n]$. By (a) and Lemma 2.7 we obtain $\operatorname{depth}(S^{\star}_{n,3}/I(\mathcal{P}^{\star}_{n,3}) : y_n) = \lceil\frac{n-2+1}{3}\rceil + 1 = \lceil\frac{n+2}{3}\rceil$. Thus $\operatorname{depth}(S^{\star\star}_{n,3}/I(\mathcal{P}^{\star\star}_{n,3})) \leq \lceil\frac{n+2}{3}\rceil$. Similarly one can prove the result for Stanley depth by using Proposition 2.9 and Corollary 2.13. $\square$

**Theorem 3.6.** *For $n \geq 3$, and $n \equiv 0, 2 \,(\mathrm{mod}\, 3)$, $\operatorname{sdepth}(S_{n,3}/I(\mathcal{C}_{n,3})) = \lceil\frac{n-1}{3}\rceil = \operatorname{depth}(S_{n,3}/I(\mathcal{C}_{n,3}))$, and otherwise, $\lceil\frac{n-1}{3}\rceil \leq \operatorname{depth}(S_{n,3}/I(\mathcal{C}_{n,3})), \operatorname{sdepth}(S_{n,3}/I(\mathcal{C}_{n,3})) \leq \lceil\frac{n}{3}\rceil$.*

***Proof.*** We first prove the result for depth. For $n = 3, 4$ the result is clear. Let $n \geq 5$,

$$A := (I(\mathcal{C}_{n,3}) : x_n) = \big(\cup_{i=2}^{n-3}\{x_iy_i, x_iy_{i+1}, x_ix_{i+1}, x_{i+1}y_i, y_iy_{i+1}, y_iz_i, y_iz_{i+1}, y_{i+1}z_i, z_iz_{i+1}\},$$
$$x_{n-2}y_{n-2}, y_{n-2}z_{n-2}, x_1, y_1, x_{n-1}, y_{n-1}, y_n, z_nz_{n-1}, z_{n-1}z_{n-2}, y_{n-2}z_{n-1}, z_nz_1, z_1z_2, y_2z_1\big),$$

and

$$\overline{A} := (I(\mathcal{C}_{n,3}), x_n) = \big(\cup_{i=1}^{n-2}\{x_iy_i, x_iy_{i+1}, x_ix_{i+1}, x_{i+1}y_i, y_iy_{i+1}, y_iz_i, y_iz_{i+1}, y_{i+1}z_i, z_iz_{i+1}\}, x_n,$$
$$x_{n-1}y_{n-1}, y_{n-1}z_{n-1}, x_{n-1}y_n, y_{n-1}y_n, y_nz_{n-1}, y_{n-1}z_n, z_{n-1}z_n, y_nz_n, y_1y_n, x_1y_n, y_1z_n, y_nz_1, z_1z_n\big)$$
$$= (I(\mathcal{P}_{n-1,3}), x_n, x_{n-1}y_n, y_{n-1}y_n, y_nz_{n-1}, y_{n-1}z_n, z_{n-1}z_n, y_nz_n, y_1y_n, x_1y_n, y_1z_n, y_nz_1, z_1z_n),$$

then by Remark 2.6 we have

$$\operatorname{depth}(S_{n,3}/I(\mathcal{C}_{n,3})) \geq \min\{\operatorname{depth}(S_{n,3}/A), \operatorname{depth}(S_{n,3}/\overline{A})\}. \tag{3.1}$$

Since
$$(A, z_n) = \big(\cup_{i=2}^{n-3}\{x_iy_i, x_iy_{i+1}, x_ix_{i+1}, x_{i+1}y_i, y_iy_{i+1}, y_iz_i, y_iz_{i+1}, y_{i+1}z_i, z_iz_{i+1}\},$$
$$x_{n-2}y_{n-2}, y_{n-2}z_{n-2}, x_1, y_1, x_{n-1}, y_{n-1}, y_n, z_n, z_{n-1}z_{n-2}, y_{n-2}z_{n-1}, z_1z_2, y_2z_1\big),$$

after renumbering the variables we have $S_{n,3}/(A, z_n) \cong (S^{\star\star}_{n-3,3}/I(\mathcal{P}^{\star\star}_{n-3,3}))[x_n]$. Thus by Lemmas 3.5 and 2.7 $\operatorname{depth}(S_{n,3}/(A, z_n)) = \lceil\frac{n-3+2}{3}\rceil + 1 = \lceil\frac{n-1}{3}\rceil + 1$. Also

$$(A : z_n) = \big(\cup_{i=2}^{n-3}\{x_iy_i, x_iy_{i+1}, x_ix_{i+1}, x_{i+1}y_i, y_iy_{i+1}, y_iz_i, y_iz_{i+1}, y_{i+1}z_i, z_iz_{i+1}\}, x_{n-2}y_{n-2},$$
$$y_{n-2}z_{n-2}, x_1, y_1, x_{n-1}, y_{n-1}, y_n, z_{n-1}, z_1\big),$$

after renumbering the variables we get $S_{n,3}/(A : z_n) \cong (S_{n-3,3}/I(\mathcal{P}_{n-3,3}))[x_n, z_n]$. Thus by Lemmas 3.2 and 2.7 $\operatorname{depth}(S_{n,3}/(A : z_n)) = \lceil\frac{n-3}{3}\rceil + 2 = \lceil\frac{n}{3}\rceil + 1$. Using Remark 2.6

$$\operatorname{depth}(S_{n,3}/(A)) \geq$$
$$\min\{\operatorname{depth}(S_{n,3}/(A : z_n)), \operatorname{depth}(S_{n,3}/(A, z_n))\} = \min\{\lceil\frac{n}{3}\rceil + 1, \lceil\frac{n-1}{3}\rceil + 1\}. \tag{3.2}$$

As
$$(\overline{A} : y_n) = \big(\cup_{i=2}^{n-3}\{x_iy_i, x_iy_{i+1}, x_ix_{i+1}, x_{i+1}y_i, y_iy_{i+1}, y_iz_i, y_iz_{i+1}, y_{i+1}z_i, z_iz_{i+1}\},$$
$$x_{n-2}y_{n-2}, y_{n-2}z_{n-2}, x_n, x_1, y_1, z_1, x_{n-1}, y_{n-1}, z_{n-1}, z_n\big),$$

after renumbering the variables we get $S_{n,3}/(\overline{A} : y_n) \cong S_{n-3,3}/I(\mathcal{P}_{n-3,3})[y_n]$. Therefore by Lemmas 3.2 and 2.7 $\operatorname{depth}(S_{n,3}/(\overline{A} : y_n)) = \lceil\frac{n-3}{3}\rceil + 1 = \lceil\frac{n}{3}\rceil$. Now let

$$\widehat{A} := (\overline{A}, y_n) = (I(\mathcal{P}_{n-1,3}), x_n, y_n, y_{n-1}z_n, z_{n-1}z_n, y_1z_n, z_1z_n),$$

$$\operatorname{depth}(S_{n,3}/\overline{A}) \geq \min\{\operatorname{depth}(S_{n,3}/(\overline{A} : y_n)), \operatorname{depth}(S_{n,3}/\widehat{A})\}$$
$$= \min\{\lceil\frac{n}{3}\rceil, \operatorname{depth}(S_{n,3}/\widehat{A})\}. \tag{3.3}$$

Since $(\widehat{A}:z_n) = \big(\cup_{i=2}^{n-3}\{x_iy_i, x_iy_{i+1}, x_ix_{i+1}, x_{i+1}y_i, y_iy_{i+1}, y_iz_i, y_iz_{i+1}, y_{i+1}z_i, z_iz_{i+1}\},$
$x_{n-2}y_{n-2}, y_{n-2}z_{n-2}, z_1, y_1, z_{n-1}, y_{n-1}, y_n, x_n, x_{n-1}x_{n-2}, x_{n-1}y_{n-2}, x_1x_2, x_1y_2\big),$

after renumbering the variables, we have $S_{n,3}/(\widehat{A}:z_n) \cong (S^{\star\star}_{n-3,3}/I(\mathcal{P}^{\star\star}_{n-3,3}))[z_n]$. Thus by Lemmas 3.5 and 2.7 $\operatorname{depth}(S_{n,3}/(\widehat{A}:z_n)) = \lceil\frac{n-3+2}{3}\rceil + 1 = \lceil\frac{n-1}{3}\rceil + 1$. Also $S_{n,3}/(\widehat{A}, z_n) \cong S_{n-1,3}/I(\mathcal{P}_{n-1,3})$. Therefore by Lemma 3.2 $\operatorname{depth}(S_{n,3}/(\widehat{A}, z_n)) = \lceil\frac{n-1}{3}\rceil$. By Remark 2.6

$$\operatorname{depth}(S_{n,3}/\widehat{A}) \geq$$
$$\min\{\operatorname{depth}(S_{n,3}/(\widehat{A}:z_n))\operatorname{depth}(S_{n,3}/(\widehat{A}, z_n))\} = \min\{\lceil\frac{n-1}{3}\rceil + 1, \lceil\frac{n-1}{3}\rceil\} \quad (3.4)$$

Hence combining Eq. 3.1, Eq. 3.2, Eq. 3.3 and Eq. 3.4 we get $\operatorname{depth}(S_{n,3}/I(\mathcal{C}_{n,3})) \geq \lceil\frac{n-1}{3}\rceil$. By Corollary 2.8 we have $\operatorname{depth}(S_{n,3}/I(\mathcal{C}_{n,3})) \leq \operatorname{depth}(S_{n,3}/(I(\mathcal{C}_{n,3}):y_n))$. Since $(S_{n,3}/(I(\mathcal{C}_{n,3}):y_n)) \cong (S_{n-3,3}/(I(\mathcal{P}_{n-3,3}))[y_n]$, by Lemmas 3.2 and 2.7, we have $\operatorname{depth}(S_{n,3}/I(\mathcal{C}_{n,3})) \leq \lceil\frac{n}{3}\rceil$, if $n \equiv 0 (\operatorname{mod} 3)$ or $n \equiv 2 (\operatorname{mod} 3)$ then $\lceil\frac{n-1}{3}\rceil = \lceil\frac{n}{3}\rceil$. If $n \equiv 1 (\operatorname{mod} 3)$ then $\lceil\frac{n-1}{3}\rceil \leq \operatorname{depth}(S_{n,3}/I(\mathcal{C}_{n,3})) \leq \lceil\frac{n}{3}\rceil$.

Now we prove the result for Stanley depth. If $n = 3$, then by using [11] we have the following Stanley decomposition

$$S_{3,3}/I(\mathcal{C}_{3,3}) = K[x_1] \oplus y_1K[y_1] \oplus z_1K[z_1] \oplus x_2K[x_2] \oplus y_2K[y_2] \oplus z_2K[z_2] \oplus$$
$$\oplus x_3K[x_3] \oplus z_3K[z_3],$$

Thus $\operatorname{sdepth}(S_{3,3}/I(\mathcal{C}_{3,3})) \geq 1$. For upper bound by Proposition 2.9 we have

$$\operatorname{sdepth}(S_{3,3}/I(\mathcal{C}_{3,3})) \leq \operatorname{sdepth}(S_{3,3}/(I(\mathcal{C}_{3,3}):y_2)),$$

since $S_{3,3}/(I(\mathcal{C}_{3,3}):y_2) \cong K[y_2]$, therefore $\operatorname{sdepth}(S_{3,3}/I(\mathcal{C}_{3,3})) \leq 1$, as desired. For $n = 4$,

$$\text{let } T := K[x_1, z_1] \oplus y_1K[x_3, y_1] \oplus x_2K[x_2, z_1] \oplus y_2K[y_2, x_4] \oplus y_3K[x_1, y_3] \oplus x_4K[x_4, z_1]$$
$$\oplus y_4K[x_2, y_4] \oplus z_4K[x_1, z_4] \oplus z_2K[x_1, z_2] \oplus x_3K[x_1, x_3] \oplus z_3K[x_1, z_3],$$

if $u \in S_{4,3}/I(\mathcal{C}_{4,3})$ such that $u \notin T$, then $\deg(u_i) \geq 2$. It is easy to see that $S_{4,3}/I(\mathcal{C}_{4,3}) = T \oplus_u uK[\operatorname{supp}(u)]$, Thus $\operatorname{sdepth}(S_{4,3}/I(\mathcal{C}_{4,3})) \geq 2$. For upper bound by Proposition 2.9 we have $\operatorname{sdepth}(S_{4,3}/I(\mathcal{C}_{4,3})) \leq \operatorname{sdepth}(S_{4,3}/(I(\mathcal{C}_{4,3}):y_2y_4))$, since $S_{4,3}/(I(\mathcal{C}_{4,3}):y_2y_4) \cong K[y_2, y_4]$, therefore $\operatorname{sdepth}(S_{4,3}/I(\mathcal{C}_{4,3})) \leq 2$. Hence $\operatorname{sdepth}(S_{4,3}/I(\mathcal{C}_{4,3})) = 2$. Let $n \geq 5$, using Proposition 2.9 instead of Corollary 2.8 the proof for depth also works for Stanley depth. □

**Example 3.7.** *One can expect that* $\operatorname{depth}(S_{n,3}/I(\mathcal{C}_{n,3})) = \lceil\frac{n-1}{3}\rceil$ *as we have in [4, Proposition 1.3] and Theorem 3.3. But examples show that in the essential case when* $n \equiv 1 (\operatorname{mod} 3)$ *the upper bound in Theorem 3.6 is reached. For instance, when* $n = 4$, *then* $\operatorname{depth}(S_{4,3}/I(\mathcal{C}_{4,3})) = \operatorname{sdepth}(S_{4,3}/I(\mathcal{C}_{4,3})) = 2 = \lceil\frac{4}{3}\rceil$.

**Remark 3.8.** If $3 \leq n \leq 10$, then using SdepthLib:coc [25] we have $\operatorname{sdepth}(S_{n,3}/I(\mathcal{C}_{n,3})) = \lceil\frac{n}{3}\rceil$. Also for $3 \leq n \leq 6$, we have $\operatorname{depth}(S_{n,3}/I(\mathcal{C}_{n,3})) = \lceil\frac{n}{3}\rceil$ that is the upper bound in Theorem 3.6 is reached for both depth and Stanley depth in all known cases. In order to show that $\operatorname{sdepth}(S_{n,3}/I(\mathcal{C}_{n,3})) \geq \operatorname{depth}(S_{n,3}/I(\mathcal{C}_{n,3}))$ (Stanley's inequality) one needs to show that $\operatorname{sdepth}(S_{n,3}/I(\mathcal{C}_{n,3})) = \lceil\frac{n}{3}\rceil$, for all $n$. For this one needs to find a suitable Stanley decomposition which we don't know at the moment and could be hard to find.

## 4. Lower bounds for Stanley depth of $I(\mathcal{P}_{n,m})$ and $I(\mathcal{C}_{n,m})$ when $1 \leq m \leq 3$

In this section, we give some lower bounds for Stanley depth of $I(\mathcal{P}_{n,m})$ and $I(\mathcal{C}_{n,m})$, when $m \leq 3$. These bounds together with the results of the previous section allow us to give a positive answer to Conjecture 1 in some special cases. We begin this section with the following useful lemma:

**Lemma 4.1.** *Let $A$ and $B$ be two disjoint sets of variables, $I_1 \subset K[A]$ and $I_2 \subset K[B]$ be square free monomial ideals such that* $\operatorname{sdepth}_{K[A]}(I_1) > \operatorname{sdepth}(K[A]/I_1)$*. Then*

$$\operatorname{sdepth}_{K[A\cup B]}(I_1+I_2) \geq \operatorname{sdepth}(K[A]/I_1) + \operatorname{sdepth}_{K[B]}(I_2).$$

***Proof.*** By [2, Theorem 1.3(1)] we have

$$\operatorname{sdepth}_{K[A\cup B]}(I_1+I_2) \geq \min\{\operatorname{sdepth}_{K[A\cup B]}(I_1), \operatorname{sdepth}(K[A]/I_1) + \operatorname{sdepth}_{K[B]}(I_2)\}.$$

Now by Lemma 2.7 we have

$$\operatorname{sdepth}_{K[A\cup B]}(I_1+I_2) \geq \min\{\operatorname{sdepth}_{K[A]}(I_1) + |B|, \operatorname{sdepth}(K[A]/I_1) + \operatorname{sdepth}_{K[B]}(I_2)\}.$$

Since $|B| \geq \operatorname{sdepth}_{K[B]}(I_2)$, therefore

$$\operatorname{sdepth}_{K[A]}(I_1) + |B| > \operatorname{sdepth}(K[A]/I_1) + \operatorname{sdepth}_{K[B]}(I_2),$$

this proves the desired inequality. □

Now we introduce some notations for the case $m = 3$. For $3 \leq l \leq n-2$, let

$$J_l := (x_{n-l}, z_{n-l}, x_{n-l+1}, y_{n-l-1}, z_{n-l+1}, x_{n-l-1}, z_{n-l-1}),$$
$$I(P'_{l-1}) := (x_{n-l+2}x_{n-l+3}, \dots, x_{n-1}x_n),$$
$$I(P''_{l-1}) := (z_{n-l+2}z_{n-l+3}, \dots, z_{n-1}z_n),$$

be the monomial ideals of $S_{n,3}$. Consider the subsets of variables

$$D_l := \{x_{n-l+2}, x_{n-l+3}, \dots, x_{n-1}, x_n\},$$
$$D'_l := \{z_{n-l+2}, z_{n-l+3}, \dots, z_{n-1}, z_n\},$$
$$D''_l := \{x_{n-l}, z_{n-l}, x_{n-l+1}, y_{n-l-1}, z_{n-l+1}, x_{n-l-1}, z_{n-l-1}\}.$$

Let $L_l$ be a monomial ideal of $S_{n,3}$ such that $L_l = I(P'_{l-1}) + I(P''_{l-1}) + J_l$. With these notations we have the following lemma:

**Lemma 4.2.** *For $3 \leq l \leq n-2$,* $\operatorname{sdepth}_{K[D_l\cup D'_l\cup D''_l]}(L_l) \geq \lceil \frac{l+2}{3} \rceil + 1$.

***Proof.*** Since $L_l = I(P'_{l-1}) + I(P''_{l-1}) + J_l$, by [2, Theorem 1.3], we have

$$\begin{aligned}\operatorname{sdepth}_{K[D_l\cup D'_l\cup D''_l]}(L_l) \geq \min\big\{&\operatorname{sdepth}_{K[D_l\cup D'_l\cup D''_l]}(J_l), \min\{\operatorname{sdepth}_{K[D_l\cup D'_l]}(I(P'_{l-1})),\\ &\operatorname{sdepth}_{K[D_l]}(K[D_l]/I(P'_{l-1})) + \operatorname{sdepth}_{K[D'_l]}(I(P''_{l-1}))\}\big\}. \quad (4.1)\end{aligned}$$

By using [21, Theorem 2.3] and [22, Proposition 2.1], Eq. 4.1 implies that

$$\begin{aligned}\operatorname{sdepth}_{K[D_l\cup D'_l\cup D''_l]}(L_l) &\geq \min\{4+2(l-2), \min\{2l-2-\lfloor\frac{l-2}{2}\rfloor, \lceil\frac{l-1}{3}\rceil + l - 1 - \lfloor\frac{l-2}{2}\rfloor\}\}\\ &\geq \lceil\frac{l+2}{3}\rceil + 1.\end{aligned}$$

□

**Theorem 4.3.** *For $n \geq 1$ and $1 \leq m \leq 3$,*

$$\operatorname{sdepth}(I(\mathcal{P}_{n,m})) > \operatorname{sdepth}(S_{n,m}/I(\mathcal{P}_{n,m})) = \lceil\frac{n}{3}\rceil.$$

***Proof.*** By Lemma 3.2 and Remark 3.1 we have $\operatorname{sdepth}(S_{n,m}/I(\mathcal{P}_{n,m})) = \lceil \frac{n}{3} \rceil$, we use this fact frequently in the proof without referring it again and again.

(a) If $m = 1$, clearly $I(\mathcal{P}_{n,1}) \cong I(P_n)$, thus by [21, Theorem 2.3] and [22, Proposition 2.1] we have $\operatorname{sdepth}(I(\mathcal{P}_{n,1})) > \operatorname{sdepth}(S_{n,1}/I(\mathcal{P}_{n,1})) = \lceil \frac{n}{3} \rceil$.

(b) If $m = 2$, we prove the result by induction on $n$. If $n = 1$ then by (a) the required result follows. If $n = 2, 3$, then by [19, Lemma 2.1], $\operatorname{sdepth}(I(\mathcal{P}_{n,2})) > \lceil \frac{n}{3} \rceil$. Now assume that $n \geq 4$. Since $x_{n-1} \notin I(\mathcal{P}_{n,2})$, thus we have

$$I(\mathcal{P}_{n,2}) = I(\mathcal{P}_{n,2}) \cap S' \oplus x_{n-1}(I(\mathcal{P}_{n,2}) : x_{n-1})S_{n,2},$$

where $S' = K[x_1, x_2, \dots, x_{n-2}, x_n, y_1, y_2, \dots, y_n]$. Now

$$I(\mathcal{P}_{n,2}) \cap S' = (\mathcal{G}(I(\mathcal{P}_{n-2,2})), x_{n-2}y_{n-1}, y_{n-2}y_{n-1}, x_n y_n, y_{n-1}x_n, y_{n-1}y_n) \text{ and}$$

$$(I(\mathcal{P}_{n,2}) : x_{n-1})S_{n,2} = (\mathcal{G}(I(\mathcal{P}_{n-3,2})), x_{n-2}, y_{n-2}, y_{n-1}, x_n, y_n)S_{n,2}.$$

As $y_{n-1} \notin I(\mathcal{P}_{n,2}) \cap S'$, so we get

$$I(\mathcal{P}_{n,2}) \cap S' = (I(\mathcal{P}_{n,2}) \cap S') \cap S'' \oplus y_{n-1}(I(\mathcal{P}_{n,2}) \cap S' : y_{n-1})S',$$

where $S'' = K[x_1, \dots, x_{n-2}, x_n, y_1, \dots, y_{n-2}, y_n]$. Thus

$$I(\mathcal{P}_{n,2}) = (I(\mathcal{P}_{n,2}) \cap S') \cap S'' \oplus y_{n-1}(I(\mathcal{P}_{n,2}) \cap S' : y_{n-1})S' \oplus x_{n-1}(I(\mathcal{P}_{n,2}) : x_{n-1})S_{n,2},$$

where

$$(I(\mathcal{P}_{n,2}) \cap S') \cap S'' = (\mathcal{G}(I(\mathcal{P}_{n-2,2})), x_n y_n)S''$$

and

$$(I(\mathcal{P}_{n,2}) \cap S' : y_{n-1})S' = (\mathcal{G}(I(\mathcal{P}_{n-3,2})), x_{n-2}, y_{n-2}, x_n, y_n)S'.$$

By induction on $n$ and Lemma 4.1 we have

$$\operatorname{sdepth}((I(\mathcal{P}_{n,2}) \cap S') \cap S'') \geq \operatorname{sdepth}(S_{n-2,2}/I(\mathcal{P}_{n-2,2})) + \operatorname{sdepth}_{K[x_n,y_n]}(x_n y_n).$$

Again by induction on $n$, Lemma 4.1 and Lemma 2.7 we have

$$\operatorname{sdepth}((I(\mathcal{P}_{n,2}) \cap S' : y_{n-1})S') \geq \operatorname{sdepth}(S_{n-3,2}/I(\mathcal{P}_{n-3,2})) + \operatorname{sdepth}_T(x_{n-2}, y_{n-2}, x_n, y_n) + 1$$

and

$$\operatorname{sdepth}\left((I(\mathcal{P}_{n,2}) : x_{n-1})S_{n,2}\right) \geq \operatorname{sdepth}(S_{n-3,2}/I(\mathcal{P}_{n-3,2})) + \operatorname{sdepth}_R(x_{n-2}, y_{n-2}, y_{n-1}, x_n, y_n) + 1,$$

where $T = [x_{n-2}, y_{n-2}, x_n, y_n]$ and $R = K[x_{n-2}, y_{n-2}, y_{n-1}, x_n, y_n]$. Thus

$$\operatorname{sdepth}((I(\mathcal{P}_{n,2}) \cap S') \cap S'') > \lceil \frac{n}{3} \rceil$$

as $\operatorname{sdepth}_{K[x_n,y_n]}(x_n y_n) = 2$. By [1, Theorem 2.2] we have $\operatorname{sdepth}((I(\mathcal{P}_{n,2}) \cap S' : y_{n-1})S') > \lceil \frac{n}{3} \rceil$ and $\operatorname{sdepth}((I(\mathcal{P}_{n,2}) : x_{n-1})S_{n,2}) > \lceil \frac{n}{3} \rceil$. This completes the proof for $m = 2$.

(c) If $m = 3$, we proceed again by induction on $n$. If $n = 1$, then by (a) the required result follows. If $n = 2$, the result follows by (b). If $n = 3$ then by [19, Lemma 2.1] $\operatorname{sdepth}(I(\mathcal{P}_{3,3})) > \lceil \frac{3}{3} \rceil$. If $n \geq 4$, then we consider the following decomposition of $I(\mathcal{P}_{n,3})$ as a vector space:

$$I(\mathcal{P}_{n,3}) = I(\mathcal{P}_{n,3}) \cap R_1 \oplus y_n(I(\mathcal{P}_{n,3}) : y_n)S_{n,3}.$$

Similarly, we can decompose $I(\mathcal{P}_{n,3}) \cap R_1$ by the following:

$$I(\mathcal{P}_{n,3}) \cap R_1 = I(\mathcal{P}_{n,3}) \cap R_2 \oplus y_{n-1}(I(\mathcal{P}_{n,3}) \cap R_1 : y_{n-1})R_1.$$

Continuing in the same way for $1 \leq l \leq n-1$ we have

$$I(\mathcal{P}_{n,3}) \cap R_l = I(\mathcal{P}_{n,3}) \cap R_{l+1} \oplus y_{n-l}(I(\mathcal{P}_{n,3}) \cap R_l : y_{n-l})R_l,$$

where $R_l := K[x_1, x_2, \dots x_n, y_1, y_2, \dots, y_{n-l}, z_1, z_2, \dots, z_n]$. Finally, we get the following decomposition of $I(\mathcal{P}_{n,3})$:

$$I(\mathcal{P}_{n,3}) = I(\mathcal{P}_{n,3}) \cap R_n \oplus \oplus_{l=1}^{n-1} y_{n-l}(I(\mathcal{P}_{n,3}) \cap R_l : y_{n-l})R_l \oplus y_n(I(\mathcal{P}_{n,3}) : y_n)S_{n,3}.$$

Therefore

$$\operatorname{sdepth}(I(\mathcal{P}_{n,3})) \geq \min\Big\{ \operatorname{sdepth}(I(\mathcal{P}_{n,3}) \cap R_n), \operatorname{sdepth}((I(\mathcal{P}_{n,3}) : y_n)S_{n,3}), \min_{l=1}^{n-1}\{\operatorname{sdepth}((I(\mathcal{P}_{n,3}) \cap R_l : y_{n-l})R_l)\}\Big\}. \quad (4.2)$$

Since

$$I(\mathcal{P}_{n,3}) \cap R_n = ((x_1x_2, x_2x_3, \dots, x_{n-1}x_n)+(z_1z_2, z_2z_3, \dots, z_{n-1}z_n))K[x_1, \dots, x_n, z_1, \dots, z_n],$$

thus by [2, Theorem 1.3] and [22, Proposition 2.1] we have $\operatorname{sdepth}(I(\mathcal{P}_{n,3}) \cap R_n) > \lceil \frac{n}{3} \rceil$. As we can see that

$$(I(\mathcal{P}_{n,3}) : y_n)S_{n,3} = (\mathcal{G}(I(\mathcal{P}_{n-2,3})) + (x_n, z_n, x_{n-1}, z_{n-1}, y_{n-1}))[y_n].$$

Let $B := K[x_n, z_n, x_{n-1}, z_{n-1}, y_{n-1}]$ thus by induction on $n$, Lemmas 4.1 and 2.7

$$\operatorname{sdepth}((I(\mathcal{P}_{n,3}) : y_n)S_{n,3}) > \operatorname{sdepth}(S_{n-2,3}/I(\mathcal{P}_{n-2,3}))+\operatorname{sdepth}_B(x_n, z_n, x_{n-1}, z_{n-1}, y_{n-1})+1.$$

By [1, Theorem 2.2] we have $\operatorname{sdepth}((I(\mathcal{P}_{n,3}) : y_n)S_{n,3}) > \lceil \frac{n}{3} \rceil$.

**(1):** If $l = 1$, then $(I(\mathcal{P}_{n,3}) \cap R_1 : y_{n-1})R_1 = \Big(\mathcal{G}(I(\mathcal{P}_{n-3,3})) + J_1\Big)[y_{n-1}]$, where $J_1 := (x_{n-1}, z_{n-1}, x_n, y_{n-2}, z_n, x_{n-2}, z_{n-2})$, then by induction on $n$, Lemmas 4.1 and 2.7, we have

$$\operatorname{sdepth}((I(\mathcal{P}_{n,3}) \cap R_1 : y_{n-1})R_1) > \operatorname{sdepth}(S_{n-3,3}/I(\mathcal{P}_{n-3,3})) + \operatorname{sdepth}_{K[\operatorname{supp}(J_1)]}(J_1) + 1,$$

by [1, Theorem 2.2] we have $\operatorname{sdepth}((I(\mathcal{P}_{n,3}) \cap R_1 : y_{n-1})R_1) > \lceil \frac{n}{3} \rceil$.

**(2):** If $l = 2$ and $n \neq 4$, then

$$(I(\mathcal{P}_{n,3}) \cap R_2 : y_{n-2})R_2 = \Big(\mathcal{G}(I(\mathcal{P}_{n-4,3})) + J_2\Big)[y_{n-2}, x_n, z_n],$$

where $J_2 := (x_{n-2}, z_{n-2}, x_{n-1}, z_{n-1}, x_{n-3}, y_{n-3}, z_{n-3})$, using the same arguments as in case(1) we have $\operatorname{sdepth}((I(\mathcal{P}_{n,3}) \cap R_2 : y_{n-2})R_2) > \lceil \frac{n}{3} \rceil$.

**(3):** If $3 \leq l \leq n-3$, then $(I(\mathcal{P}_{n,3}) \cap R_l : y_{n-l})R_l = \Big(\mathcal{G}(I(\mathcal{P}_{n-(l+2),3})) + \mathcal{G}(L_l)\Big)[y_{n-l}]$, by induction on $n$, Lemmas 4.1 and 2.7, we have

$$\operatorname{sdepth}((I(\mathcal{P}_{n,3}) \cap R_l : y_{n-l})R_l) > \operatorname{sdepth}(S_{n-(l+2),3}/(I(\mathcal{P}_{n-(l+2),3}))) + \operatorname{sdepth}_{K[D_l \cup D'_l \cup D''_l]}(L_l) + 1, \quad (4.3)$$

By Eq. 4.3 and Lemma 4.2 we have

$$\operatorname{sdepth}((I(\mathcal{P}_{n,3}) \cap R_l : y_{n-l})R_l) > \lceil \frac{n-(l+2)}{3} \rceil + \lceil \frac{l+2}{3} \rceil + 1 + 1 > \lceil \frac{n}{3} \rceil.$$

**(4):** If $l = n-2$, then $(I(\mathcal{P}_{n,3}) \cap R_{n-2} : y_2)R_{n-2} = (\mathcal{G}(L_{n-2}))[y_2]$, by Lemmas 4.2 and 2.7 we have $\operatorname{sdepth}((I(\mathcal{P}_{n,3}) \cap R_{n-2} : y_2)R_{n-2}) > \lceil \frac{n}{3} \rceil$.

**(5):** If $l = n-1$, then

$$(I(\mathcal{P}_{n,3}) \cap R_{n-1} : y_1)R_{n-1} = (I(P'_{n-2}) + I(P''_{n-2}) + J_{n-1})K[D_{n-1} \cup D'_{n-1} \cup D''_{n-1} \cup \{y_1\}],$$

where $\mathcal{G}(J_{n-1}) = \{x_1, z_1, x_2, z_2\}$, $D_{n-1} = \{x_3, x_4, \dots, x_n\}$, $D'_{n-1} = \{z_3, z_4, \dots, z_n\}$ and $D''_{n-1} = \{x_1, z_1, x_2, z_2\}$. Using the proof of Lemma 4.2 and by Lemma 2.7

$$\operatorname{sdepth}_{K[D_{n-1} \cup D'_{n-1} \cup D''_{n-1} \cup \{y_1\}]}(I(P'_{n-2}) + I(P''_{n-2}) + J_{n-1}) > \lceil \frac{n}{3} \rceil,$$

that is $\operatorname{sdepth}((I(\mathcal{P}_{n,3}) \cap R_{n-1} : y_1)R_{n-1}) > \lceil \frac{n}{3} \rceil$.

Thus by Eq. 4.2 we get $\operatorname{sdepth}(I(\mathcal{P}_{n,3})) > \lceil \frac{n}{3} \rceil$.

□

**Proposition 4.4.** *For $n \geq 3$,* $\operatorname{sdepth}(I(\mathcal{C}_{n,2})/I(\mathcal{P}_{n,2})) \geq \lceil \frac{n+2}{3} \rceil$.

***Proof.*** For $3 \leq n \leq 5$, we use [11] to show that there exist Stanley decompositions of desired Stanley depth. When $n = 3$ or 4, then

$$I(\mathcal{C}_{n,2})/I(\mathcal{P}_{n,2}) = x_1x_nK[x_1, x_n] \oplus x_1y_nK[x_1, y_n] \oplus y_1x_nK[y_1, x_n] \oplus y_1y_nK[y_1, y_n].$$

If $n = 5$, then

$$\begin{aligned} I(\mathcal{C}_{5,2})/I(\mathcal{P}_{5,2}) = {} & x_1x_5K[x_1, x_3, x_5] \oplus x_1y_5K[x_1, x_3, y_5] \oplus y_1x_5K[y_1, x_3, x_5] \oplus y_1y_5K[y_1, x_3, y_5] \\ & \oplus x_1y_3x_5K[x_1, y_3, x_5] \oplus x_1y_3y_5K[x_1, y_3, y_5] \oplus y_1y_3y_5K[y_1, y_3, y_5] \oplus y_1y_3x_5K[y_1, y_3, x_5]. \end{aligned}$$

Let $n \geq 6$ and $T := \big(\bigcup_{i=3}^{n-3}\{x_iy_i, x_iy_{i+1}, x_ix_{i+1}, x_{i+1}y_i, y_iy_{i+1}\}, x_{n-2}y_{n-2}\big) \subset \tilde{S}$, where $\tilde{S} := K[x_3, x_4, \ldots, x_{n-2}, y_3, y_4 \ldots, y_{n-2}]$. Then we have the following $K$-vector space isomorphism:

$$I(\mathcal{C}_{n,2})/I(\mathcal{P}_{n,2}) \cong x_1x_n\frac{\tilde{S}}{T}[x_1, x_n] \oplus y_1y_n\frac{\tilde{S}}{T}[y_1, y_n] \oplus x_1y_n\frac{\tilde{S}}{T}[x_1, y_n] \oplus y_1x_n\frac{\tilde{S}}{T}[y_1, x_n].$$

Thus by Lemmas 3.2 and 2.7, we have $\operatorname{sdepth}(I(\mathcal{C}_{n,2})/I(\mathcal{P}_{n,2})) \geq \lceil \frac{n+2}{3} \rceil$. □

For $n \geq 6$, let $Q = \{x_1, y_1, x_2, y_2, x_n, y_n, x_{n-1}, y_{n-1}\}$. Consider a subgraph $\mathcal{C}^{\diamond}_{n,3}$ of $\mathcal{C}_{n,3}$ with vertex set $V(\mathcal{C}^{\diamond}_{n,3}) = V(\mathcal{C}_{n,3}) \setminus Q$ and edge set

$$E(\mathcal{C}^{\diamond}_{n,3}) = E(\mathcal{C}_{n,3}) \setminus \{e \in E(\mathcal{C}_{n,3}) : \text{ where } e \text{ has at least one end vertex in } Q\}.$$

For example of $\mathcal{C}^{\diamond}_{n,3}$ see Fig. 5.

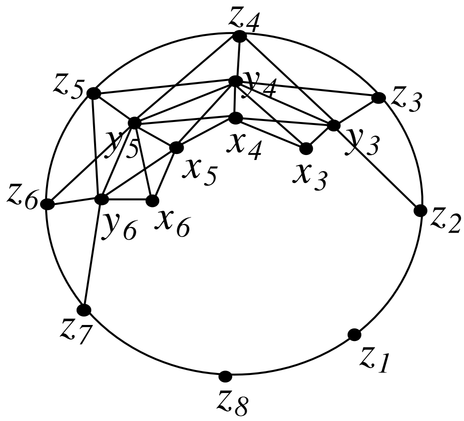



**Figure 5.** $\mathcal{C}^{\diamond}_{8,3}$.

**Lemma 4.5.** *Let $n \geq 6$, if $n \equiv 0 \,(\mathrm{mod}\, 3)$, then* $\operatorname{sdepth}(S^{\diamond}_{n,3}/I(C^{\diamond}_{n,3})) = \lceil \frac{n-2}{3} \rceil$. *Otherwise,* $\lceil \frac{n-2}{3} \rceil \leq \operatorname{sdepth}(S^{\diamond}_{n,3}/I(C^{\diamond}_{n,3})) \leq \lceil \frac{n}{3} \rceil$.

***Proof.*** By Remark 2.6

$$\operatorname{sdepth}(S^{\diamond}_{n,3}/I(C^{\diamond}_{n,3})) \geq \min\{\operatorname{sdepth}(S^{\diamond}_{n,3}/(I(C^{\diamond}_{n,3}) : z_1)), \operatorname{sdepth}(S^{\diamond}_{n,3}/(I(C^{\diamond}_{n,3}), z_1))\}. \tag{4.4}$$

$$\begin{aligned} \text{Since } (I(C^{\diamond}_{n,3}) : z_1) = \big(\big(\textstyle\bigcup_{i=3}^{n-3}\{ & x_iy_i, x_iy_{i+1}, x_ix_{i+1}, x_{i+1}y_i, y_iy_{i+1}, y_iz_i, y_iz_{i+1}, y_{i+1}z_i, z_iz_{i+1}\}, \\ & x_{n-2}y_{n-2}, y_{n-2}z_{n-2}\big), y_{n-2}z_{n-1}, z_{n-2}z_{n-1}, z_2, z_n\big), \end{aligned}$$

so after renumbering the variables we have $S^{\diamond}_{n,3}/(I(C^{\diamond}_{n,3}) : z_1) \cong S^{\star}_{n-4,3}/I(\mathcal{P}^{\star}_{n-4,3})[z_1]$. Therefore, by Lemmas 2.7 and 3.5,

$$\operatorname{sdepth}(S^{\diamond}_{n,3}/(I(C^{\diamond}_{n,3}) : z_1)) = \lceil \frac{n-4+1}{3} \rceil + 1 = \lceil \frac{n}{3} \rceil.$$

Now let

$$B := (I(C^{\diamond}_{n,3}), z_1) = ((\cup_{i=3}^{n-3}\{x_iy_i, x_iy_{i+1}, x_ix_{i+1}, x_{i+1}y_i, y_iy_{i+1}, y_iz_i, y_iz_{i+1}, y_{i+1}z_i, z_iz_{i+1}\},$$
$$x_{n-2}y_{n-2}, y_{n-2}z_{n-2}), y_{n-2}z_{n-1}, z_{n-2}z_{n-1}, z_{n-1}z_n, y_3z_2, z_2z_3, z_1),$$

so by Remark 2.6

$$\operatorname{sdepth}(S^{\diamond}_{n,3}/B) \geq \min\{\operatorname{sdepth}(S^{\diamond}_{n,3}/(B : z_n)), \operatorname{sdepth}(S^{\diamond}_{n,3}/(B, z_n))\}. \tag{4.5}$$

Since

$$(B : z_n) = ((\cup_{i=3}^{n-3}\{x_iy_i, x_iy_{i+1}, x_ix_{i+1}, x_{i+1}y_i, y_iy_{i+1}, y_iz_i, y_iz_{i+1}, y_{i+1}z_i, z_iz_{i+1}\},$$
$$x_{n-2}y_{n-2}, y_{n-2}z_{n-2}), y_3z_2, z_2z_3, z_1, z_{n-1}),$$

after renumbering the variables we have $S^{\diamond}_{n,3}/(B : z_n) \cong S^{\star}_{n-4,3}/I(\mathcal{P}^{\star}_{n-4,3})[z_n]$. Therefore by Lemmas 2.7 and 3.5, $\operatorname{sdepth}(S^{\diamond}_{n,3}/(B : z_n)) = \lceil\frac{n-4+1}{3}\rceil + 1 = \lceil\frac{n}{3}\rceil$. Now

$$(B, z_n) = ((\bigcup_{i=3}^{n-3}\{x_iy_i, x_iy_{i+1}, x_ix_{i+1}, x_{i+1}y_i, y_iy_{i+1}, y_iz_i, y_iz_{i+1}, y_{i+1}z_i, z_iz_{i+1}\},$$
$$x_{n-2}y_{n-2}, y_{n-2}z_{n-2}), y_{n-2}z_{n-1}, z_{n-2}z_{n-1}, y_3z_2, z_2z_3, z_1, z_n),$$

after renumbering the variables we have $S^{\diamond}_{n,3}/(B, z_n) \cong S^{\star\star}_{n-4,3}/I(\mathcal{P}^{\star\star}_{n-4,3})$. Therefore by Lemma 3.5, we have

$$\operatorname{sdepth}(S^{\diamond}_{n,3}/(B, z_n)) = \lceil\frac{n-4+2}{3}\rceil = \lceil\frac{n-2}{3}\rceil.$$

Combining Eq. 4.4 and Eq. 4.5 we get $\lceil\frac{n-2}{3}\rceil \leq \operatorname{sdepth}(S^{\diamond}_{n,3}/I(C^{\diamond}_{n,3}))$. For upper bound, as $z_1 \notin I(C^{\diamond}_{n,3})$ so by Proposition 2.9

$$\operatorname{sdepth}(S^{\diamond}_{n,3}/I(C^{\diamond}_{n,3})) \leq \operatorname{sdepth}(S^{\diamond}_{n,3}/(I(C^{\diamond}_{n,3}) : z_1)).$$

Since $(S^{\diamond}_{n,3}/(I(C^{\diamond}_{n,3}) : z_1)) \cong (S^{\star}_{n-4,3}/I(\mathcal{P}^{\star}_{n-4,3}))[z_1]$. Thus by Lemmas 2.7 and 3.5,

$$\operatorname{sdepth}(S^{\diamond}_{n,3}/I(C^{\diamond}_{n,3})) \leq \lceil\frac{n}{3}\rceil,$$

if $n \equiv 0(\text{mod}\, 3)$ then $\lceil\frac{n-2}{3}\rceil = \lceil\frac{n}{3}\rceil$. If $n \equiv 1(\text{mod}\, 3)$ or $n \equiv 2(\text{mod}\, 3)$ then

$$\lceil\frac{n-2}{3}\rceil \leq \operatorname{sdepth}(S^{\diamond}_{n,3}/I(C^{\diamond}_{n,3})) \leq \lceil\frac{n}{3}\rceil.$$

□

**Proposition 4.6.** *For $n \geq 3$,* $\operatorname{sdepth}(I(\mathcal{C}_{n,3})/I(\mathcal{P}_{n,3})) \geq \lceil\frac{n+2}{3}\rceil$.

***Proof.*** For $3 \leq n \leq 4$, as the minimal generators of $I(\mathcal{C}_{n,3})/I(\mathcal{P}_{n,3})$ have degree 2, so by [19, Lemma 2.1] $\operatorname{sdepth}(I(\mathcal{C}_{n,3})/I(\mathcal{P}_{n,3})) \geq 2 = \lceil\frac{n+2}{3}\rceil$. If $n = 5$ then we use [11] to show that there exist Stanley decompositions of desired Stanley depth. Let

$$H := x_1x_5K[x_1, x_3, x_5] \oplus x_1y_5K[x_1, x_3, y_5] \oplus y_1x_5K[x_3, x_5, y_1] \oplus y_1y_5K[x_3, y_1, y_5]$$
$$\oplus z_1y_5K[x_3, y_5, z_1] \oplus z_1z_5K[z_1, z_3, z_5] \oplus y_1z_5K[y_1, y_3, z_5]$$

Clearly, $H \subset I(\mathcal{C}_{5,3})/I(\mathcal{P}_{5,3})$. Let $v \in I(\mathcal{C}_{5,3})/I(\mathcal{P}_{5,3})$ be a sqaurefree monomial such that $v \notin H$ then $\deg(v) \geq 3$. Since

$$I(\mathcal{C}_{5,3})/I(\mathcal{P}_{5,3}) = H \oplus_v vK[\operatorname{supp}(v)],$$

thus we have $\operatorname{sdepth}(I(\mathcal{C}_{5,3})/I(\mathcal{P}_{5,3})) \geq 3 = \lceil\frac{5+2}{3}\rceil$. Now for $n \geq 6$, let

$$U := (\cup_{i=3}^{n-3}\{x_iy_i, x_iy_{i+1}, x_ix_{i+1}, x_{i+1}y_i, y_iy_{i+1}, y_iz_i, y_iz_{i+1}, y_{i+1}z_i, z_iz_{i+1}\}, x_{n-2}y_{n-2}, y_{n-2}z_{n-2})$$

be a squarefree monomial ideal of $R := K[x_3,\dots,x_{n-2},y_3,\dots,y_{n-2},z_3,\dots,z_{n-2}]$. Then we have the following $K$-vector space isomorphism:

$$\begin{aligned}
&I(\mathcal{C}_{n,3})/I(\mathcal{P}_{n,3}) \cong\\
&y_1y_n\frac{R}{U}[y_1,y_n]\oplus x_1y_n\frac{R[z_2]}{(\mathcal{G}(U),y_3z_2,z_2z_3)}[x_1,y_n]\oplus z_1y_n\frac{R[x_2]}{(\mathcal{G}(U),y_3x_2,x_2x_3)}[z_1,y_n]\\
&\oplus y_1x_n\frac{R[z_{n-1}]}{(\mathcal{G}(U),y_{n-2}z_{n-1},z_{n-2}z_{n-1})}[y_1,x_n]\oplus y_1z_n\frac{R[x_{n-1}]}{(\mathcal{G}(U),y_{n-2}x_{n-1},x_{n-2}x_{n-1})}[y_1,z_n]\\
&\oplus x_1x_n\frac{R[z_1,z_2,z_{n-1},z_n]}{(\mathcal{G}(U),y_{n-2}z_{n-1},z_{n-2}z_{n-1},z_{n-1}z_n,z_nz_1,z_1z_2,y_3z_2,z_2z_3)}[x_1,x_n]\\
&\oplus z_1z_n\frac{R[x_1,x_2,x_{n-1},x_n]}{(\mathcal{G}(U),y_{n-2}x_{n-1},x_{n-2}x_{n-1},x_{n-1}x_n,x_nx_1,x_1x_2,y_3x_2,x_2x_3)}[z_1,z_n].
\end{aligned}$$

Clearly we can see that $R/U \cong S_{n-4,3}/I(\mathcal{P}_{n-4,3})$,

$$\begin{aligned}
\frac{R[z_2]}{(\mathcal{G}(U),y_3z_2,z_2z_3)}\cong\frac{R[x_2]}{(\mathcal{G}(U),y_3x_2,x_2x_3)}\cong\frac{R[z_{n-1}]}{(\mathcal{G}(U),y_{n-2}z_{n-1},z_{n-2}z_{n-1})}\\
\cong\frac{R[x_{n-1}]}{(\mathcal{G}(U),y_{n-2}x_{n-1},x_{n-2}x_{n-1})}\cong S^{\star}_{n-4,3}/I(\mathcal{P}^{\star}_{n-4,3}),
\end{aligned}$$

and

$$\begin{aligned}
&\frac{R[z_1,z_2,z_{n-1},z_n]}{(\mathcal{G}(U),y_{n-2}z_{n-1},z_{n-2}z_{n-1},z_{n-1}z_n,z_nz_1,z_1z_2,y_3z_2,z_2z_3)}\\
&\quad\cong\frac{R[x_1,x_2,x_{n-1},x_n]}{(\mathcal{G}(U),y_{n-2}x_{n-1},x_{n-2}x_{n-1},x_{n-1}x_n,x_nx_1,x_1x_2,y_3x_2,x_2x_3)}\cong S^{\diamond}_{n,3}/I(\mathcal{C}^{\diamond}_{n,3}).
\end{aligned}$$

Thus by Lemmas 3.2, 3.5, 4.5 and 2.7 we have

$$\operatorname{sdepth}(I(\mathcal{C}_{n,3})/I(\mathcal{P}_{n,3}))\geq\min\Big\{\lceil\frac{n-4}{3}\rceil+2,\lceil\frac{n-4+1}{3}\rceil+2,\lceil\frac{n-2}{3}\rceil+2\Big\}=\lceil\frac{n+2}{3}\rceil.$$

□

**Theorem 4.7.** *For $1\leq m\leq 3$, $n\geq 3$, $\operatorname{sdepth}(I(\mathcal{C}_{n,m}))\geq\operatorname{sdepth}(S_{n,m}/I(\mathcal{C}_{n,m}))$.*

***Proof.*** For $m=1$, $I(\mathcal{C}_{n,1})=C_n$. Then the result follows by [4, Theorem 1.9] and [21, Theorem 2.3]. If $m=2$ or 3, consider the short exact sequence

$$0\longrightarrow I(\mathcal{P}_{n,m})\longrightarrow I(\mathcal{C}_{n,m})\longrightarrow I(\mathcal{C}_{n,m})/I(\mathcal{P}_{n,m})\longrightarrow 0,$$

then by Lemma 2.5, $\operatorname{sdepth}(I(\mathcal{C}_{n,m}))\geq\min\{\operatorname{sdepth}(I(\mathcal{P}_{n,m})),\operatorname{sdepth}(I(\mathcal{C}_{n,m})/I(\mathcal{P}_{n,m}))\}$. By Theorem 4.3 and we have $\operatorname{sdepth}(I(\mathcal{P}_{n,m}))\geq\lceil\frac{n}{3}\rceil+1$, and by Propositions 4.4 and 4.6, we have $\operatorname{sdepth}(I(\mathcal{C}_{n,m})/I(\mathcal{P}_{n,m}))\geq\lceil\frac{n+2}{3}\rceil=\lceil\frac{n-1}{3}\rceil+1$, this completes the proof. □

## 5. Upper bounds for depth and Stanley depth of cyclic modules associated to $\mathcal{P}_{n,m}$ and $\mathcal{C}_{n,m}$

Let $m\leq n$, in general, we don't know the values of depth and Stanley depth of $S_{n,m}/I(\mathcal{P}_{n,m})$. However, in the light of our observations, we propose the following question.

**Question 1.** Is $\operatorname{depth}(S_{n,m}/I(\mathcal{P}_{n,m}))=\operatorname{sdepth}(S_{n,m}/I(\mathcal{P}_{n,m}))=\lceil\frac{n}{3}\rceil\lceil\frac{m}{3}\rceil$?

Let $n\geq 2$, we have confirmed this question for the cases when $1\leq m\leq 3$ see Remark 3.1, and Lemma 3.2. If $m=4$, we make some calculations for depth and Stanley depth by using CoCoA, (for sdepth we use SdepthLib:coc [25]). Calculations give an affirmative answer to Question 1 in the case $(n,m)\in\{(4,4),(5,4),(6,4)\}$.

**Theorem 5.1.** *For $n \geq 2$,* $\operatorname{depth}(S_{n,m}/I(\mathcal{P}_{n,m})), \operatorname{sdepth}(S_{n,m}/I(\mathcal{P}_{n,m})) \leq \lceil \frac{n}{3} \rceil \lceil \frac{m}{3} \rceil$.

***Proof.*** Without loss of generality, we assume that $m \leq n$. We first prove the result for depth. When $m = 1$, then $I(\mathcal{P}_{n,1}) = I(P_n)$, we have the required result by Remark 3.1. For $m = 2, 3$ the result follows from Lemma 3.2. Let $m \geq 4$, we will prove this result by induction on $m$. Let $v$ be a monomial such that

$$v := \begin{cases} x_{2(m-1)}x_{5(m-1)} \cdots x_{(n-4)(m-1)}x_{(n-1)(m-1)}, & \text{if } n \equiv 0(\operatorname{mod} 3); \\ x_{1(m-1)}x_{4(m-1)} \cdots x_{(n-3)(m-1)}x_{n(m-1)}, & \text{if } n \equiv 1(\operatorname{mod} 3); \\ x_{2(m-1)}x_{5(m-1)} \cdots x_{(n-3)(m-1)}x_{n(m-1)}, & \text{if } n \equiv 2(\operatorname{mod} 3). \end{cases}$$

clearly $v \notin I(\mathcal{P}_{n,m})$ so by Corollary 2.8

$$\operatorname{depth}(S_{n,m}/I(\mathcal{P}_{n,m})) \leq \operatorname{depth}(S_{n,m}/(I(\mathcal{P}_{n,m}) : v)).$$

In all three cases $|\operatorname{supp}(v)| = \lceil \frac{n}{3} \rceil$ and $S_{n,m}/(I(\mathcal{P}_{n,m}) : v) \cong (S_{n,m-3}/I(\mathcal{P}_{n,m-3}))[\operatorname{supp}(v)]$, so by induction and Lemma 2.7

$$\operatorname{depth}(S_{n,m}/I(\mathcal{P}_{n,m})) \leq \operatorname{depth}(S_{n,m}/(I(\mathcal{P}_{n,m}) : v)) \leq \lceil \frac{n}{3} \rceil \lceil \frac{m-3}{3} \rceil + \lceil \frac{n}{3} \rceil = \lceil \frac{m}{3} \rceil \lceil \frac{n}{3} \rceil.$$

Similarly, we can prove the result for Stanley depth by using Proposition 2.9. □

**Remark 5.2.** For a positive answer to Question 1, one needs to prove that $\lceil \frac{n}{3} \rceil \lceil \frac{m}{3} \rceil$ is a lower bound for depth and Stanley depth of $S_{n,m}/I(\mathcal{P}_{n,m})$. The lower bound $\lceil \frac{\operatorname{diam}(P_{n,m})+1}{3} \rceil$ from Corollaries 2.11 and 2.13 which was helpful for the cases when $1 \leq m \leq 3$ is no more useful if $m \geq 4$. For instance, $\operatorname{depth}(S_{4,4}/I(\mathcal{P}_{4,4})) = \operatorname{sdepth}(S_{4,4}/I(\mathcal{P}_{4,4})) = 4$, but this lower bound shows that $\operatorname{depth}(S_{4,4}/I(\mathcal{P}_{4,4})) \geq 2 = \lceil \frac{\operatorname{diam}(P_{4,4})+1}{3} \rceil$ and $\operatorname{sdepth}(S_{4,4}/I(\mathcal{P}_{4,4})) \geq 2 = \lceil \frac{\operatorname{diam}(P_{4,4})+1}{3} \rceil$.

**Theorem 5.3.** *For $n \geq 3$ and $m \geq 1$,*

$$\operatorname{depth}(S_{n,m}/I(\mathcal{C}_{n,m})) \leq \begin{cases} \lceil \frac{n-1}{3} \rceil + (\lceil \frac{m}{3} \rceil - 1)\lceil \frac{n}{3} \rceil, & \textit{if } m \equiv 1, 2(\operatorname{mod} 3); \\ \lceil \frac{n}{3} \rceil \lceil \frac{m}{3} \rceil, & \textit{if } m \equiv 0(\operatorname{mod} 3). \end{cases}$$

***Proof.*** We prove this result by induction on $m$. If $m = 1$, then $I(\mathcal{C}_{n,1}) = I(C_n)$, by [4, Proposition 1.3], we have the required result. For $m = 2, 3$ the result follows by Theorems 3.3 and 3.6, respectively. Let $m \geq 4$,

$$u := \begin{cases} x_{3(m-1)}x_{6(m-1)} \cdots x_{(n-3)(m-1)}x_{n(m-1)}, & \text{if } n \equiv 0(\operatorname{mod} 3); \\ x_{1(m-1)}x_{4(m-1)} \cdots x_{(n-6)(m-1)}x_{(n-3)(m-1)}x_{(n-1)(m-1)}, & \text{if } n \equiv 1(\operatorname{mod} 3); \\ x_{2(m-1)}x_{5(m-1)} \cdots x_{(n-3)(m-1)}x_{n(m-1)}, & \text{if } n \equiv 2(\operatorname{mod} 3). \end{cases}$$

Clearly $u \notin I(\mathcal{C}_{n,m})$ and $S_{n,m}/(I(\mathcal{C}_{n,m}) : u) \cong (S_{n,m-3}/I(\mathcal{C}_{n,m-3}))[\operatorname{supp}(u)]$, since in all the cases $|\operatorname{supp}(u)| = \lceil \frac{n}{3} \rceil$, if $m \equiv 1, 2(\operatorname{mod} 3)$ so by induction and Lemma 2.7

$$\operatorname{depth}(S_{n,m}/(I(\mathcal{C}_{n,m}) : u)) \leq \lceil \frac{n-1}{3} \rceil + (\lceil \frac{m-3}{3} \rceil - 1)\lceil \frac{n}{3} \rceil + \lceil \frac{n}{3} \rceil = \lceil \frac{n-1}{3} \rceil + (\lceil \frac{m}{3} \rceil - 1)\lceil \frac{n}{3} \rceil.$$

Otherwise, by induction and Lemma 2.7 we have

$$\operatorname{depth}(S_{n,m}/(I(\mathcal{C}_{n,m}) : u)) \leq \lceil \frac{n}{3} \rceil \lceil \frac{m-3}{3} \rceil + \lceil \frac{n}{3} \rceil = \lceil \frac{n}{3} \rceil \lceil \frac{m}{3} \rceil.$$

□

**Theorem 5.4.** *For $n \geq 3$ and $m \geq 1$,* $\operatorname{sdepth}(S_{n,m}/I(\mathcal{C}_{n,m})) \leq \lceil \frac{n}{3} \rceil \lceil \frac{m}{3} \rceil$.

***Proof.*** The proof is similar to the proof of Theorem 5.3 by using Corollary 3.4 instead of Theorems 3.3. □

**Remark 5.5.** The upper bounds for Stanley depth of $S_{n,m}/I(\mathcal{P}_{n,m})$ and $S_{n,m}/I(\mathcal{C}_{n,m})$ as proved in Theorems 5.1 and 5.4 are too sharp. On the bases of our observations, we formulate the following question. A positive answer to this question will prove Conjecture 1.

**Question 2.** Is $\operatorname{sdepth}(I(\mathcal{P}_{n,m})), \operatorname{sdepth}(I(\mathcal{C}_{n,m})) \geq \lceil \frac{n}{3} \rceil \lceil \frac{m}{3} \rceil$?

## References


[1] C. Bir, D.M. Howard, M.T. Keller, W.T. Trotter and S.J. Young, *Interval partitions and Stanley depth,* J. Combin. Theory Ser. A, **117**, 475–482, 2010.

[2] M. Cimpoeaş, *Several inequalities regarding Stanley depth*, Romanian Journal of Math. and Computer Science, **2**, 28–40, 2012.

[3] M. Cimpoeaş, *Stanley depth of squarefree Veronese ideals*, An. St. Univ. Ovidius Constanta, **21** (3), 67–71, 2013.

[4] M. Cimpoeaş, *On the Stanley depth of edge ideals of line and cyclic graphs*, Romanian Journal of Math. and Computer Science, **5** (1), 70–75, 2015.

[5] CoCoATeam, CoCoA: a system for doing Computations in Commutative Algebra, available at http://cocoa.dima.unige.it.

[6] A.M. Duval, B. Goeckner, C.J. Klivans and J.L. Martine, *A non-partitionable Cohen-Macaulay simplicial complex,* Adv. Math. **299**, 381–395, 2016.

[7] S.A.S. Fakhari, *On the Stanley Depth of Powers of Monomial Ideals,* Mathematics, **7**, 607, 2019.

[8] L. Fouli and S. Morey, *A lower bound for depths of powers of edge ideals*, J. Algebraic Combin. **42** (3), 829–848, 2015.

[9] R. Hammack, W. Imrich and S. Klavar, *Handbook of Product Graphs*, Second Edition, CRC Press, Boca Raton, FL, 2011.

[10] J. Herzog, *A survey on Stanley depth, In Monomial ideals, computations and applications*, Lecture Notes in Math. **2083**, Springer, Heidelberg, 3–45, 2013.

[11] J. Herzog, M. Vladoiu and X. Zheng, *How to compute the Stanley depth of a monomial ideal*, J. Algebra, **322** (9), 3151–3169, 2009.

[12] Z. Iqbal and M. Ishaq, *Depth and Stanley depth of edge ideals associated to some line graphs*, AIMS Mathematics, **4** (3), 686–698, 2019.

[13] Z. Iqbal and M. Ishaq, *Depth and Stanley depth of edge ideals of powers of paths and cycles*, An. Şt. Univ. Ovidius Constana, **27** (3), 113–135, 2019.

[14] Z. Iqbal, M. Ishaq and M. Aamir, *Depth and Stanley depth of edge ideals of square paths and square cycles*, Comm. Algebra, **46** (3), 1188–1198, 2018.

[15] M. Ishaq, *Upper bounds for the Stanley depth*, Comm. Algebra, **40** (1), 87–97, 2012.

[16] M. Ishaq, *Values and bounds for the Stanley depth*, Carpathian J. Math. **27** (2), 217–224, 2011.

[17] M. Ishaq and M.I. Qureshi, *Upper and lower bounds for the Stanley depth of certain classes of monomial ideals and their residue class rings*, Comm. Algebra, **41** (3), 1107–1116, 2013.

[18] M.T. Keller and S.J. Young, *Combinatorial reductions for the Stanley depth of $I$ and $S/I$*, Electron. J. Comb. **24** (3), #P3.48, 2017.

[19] M.T. Keller, Y. Shen, N. Streib and S.J. Young, *On the Stanley depth of squarefree veronese ideals*, J. Algebraic Combin. **33** (2), 313–324, 2011.

[20] S. Morey, *Depths of powers of the edge ideal of a tree*, Comm. Algebra, **38** (11), 4042–4055, 2010.

[21] R. Okazaki, *A lower bound of Stanley depth of monomial ideals*, J. Commut. Algebra, **3** (1), 83–88, 2011.

[22] M.R. Pournaki, S.A.S. Fakhari and S. Yassemi, *Stanley depth of powers of the edge ideals of a forest*, Proc. Amer. Math. Soc. **141** (10), 3327–3336, 2013.

[23] M.R. Pournaki, S.A.S. Fakhari, M. Tousi and S. Yassemi, *What is ... Stanley depth?* Not. Am. Math. Soc. **56**, 1106–1108, 2009.

[24] A. Rauf, *Depth and Stanley depth of multigraded modules*, Comm. Algebra, **38** (2), 773–784, 2010.

[25] G. Rinaldo, *An algorithm to compute the Stanley depth of monomial ideals*, Le Matematiche, LXIII(ii), 243–256, 2008.

[26] R.P. Stanley, *Linear Diophantine equations and local cohomology,* Invent. Math. **68** (2), 175–193, 1982.

[27] A. Stefan, *Stanley depth of powers of path ideal*, http://arxiv.org/pdf/1409.6072.pdf.

[28] R.H. Villarreal, *Monomial Algebras* in:Monographs and Textbooks in Pure and Applied Mathematics, Marcel Dekker, Inc., New York, **238**, 2011.

Hacettepe Journal of
Mathematics & Statistics

Hacet. J. Math. Stat.
Volume 50 (1) (2021), 110 – 119
DOI : 10.15672/hujms.589753

Research Article

# Rings whose total graphs have small vertex-arboricity and arboricity

Morteza Fatehi, Kazem Khashyarmanesh*, Abbas Mohammadian

*Department of Pure Mathematics, Ferdowsi University of Mashhad, P.O.Box 1159-91775, Mashhad, Iran*

## Abstract

Let $R$ be a commutative ring with non-zero identity, and $Z(R)$ be its set of all zero-divisors. The total graph of $R$, denoted by $T(\Gamma(R))$, is an undirected graph with all elements of $R$ as vertices, and two distinct vertices $x$ and $y$ are adjacent if and only if $x + y \in Z(R)$. In this article, we characterize, up to isomorphism, all of finite commutative rings whose total graphs have vertex-arboricity (arboricity) two or three. Also, we show that, for a positive integer $v$, the number of finite rings whose total graphs have vertex-arboricity (arboricity) $v$ is finite.

**Mathematics Subject Classification (2020).** Primary: 05C99, Secondary: 13A99

**Keywords.** total graph, arboricity, vertex-arboricity

## 1. Introduction

In [1], D.F. Anderson and A. Badawi introduced the total graph of ring $R$, denoted by $T(\Gamma(R))$, as the graph with all elements of $R$ as vertices, and for distinct $x, y \in R$, the vertices $x$ and $y$ are adjacent if and only if $x + y \in Z(R)$, where $Z(R)$ is the set of zero-divisors of $R$. They studied some graph theoretical parameters of $T(\Gamma(R))$ such as diameter and girth. In addition, they showed that the total graph of a commutative ring is connected if and only if $Z(R)$ is not an ideal of $R$. In [7], H.R. Maimani et al. gave the necessary and sufficient conditions for the total graphs of finite commutative rings to be planar or toroidal and in [5] T. Chelvam and T. Asir characterized all commutative rings such that their total graphs have genus two.

Suppose that $G$ is a graph, and let $V(G)$ and $E(G)$ be the vertex set and edge set of $G$, respectively. The *vertex-arboricity* of a graph $G$, denoted by $va(G)$, is the minimum positive integer $k$ such that $V(G)$ can be partitioned into $k$ sets $V_1, V_2 \ldots, V_k$ such that $G[V_i]$ is a forest for each $i \in \{1, 2, \ldots, k\}$, where $G[V_i]$ is the induced subgraph of $G$ whose vertex set is $V_i$ and its edge set consists of all of the edges in $E(G)$ that have both endpoints in $V_i$. This partition is called *acyclic partition*. The vertex-arboricity can be viewed as a vertex coloring $f$ with $k$ colors, where each color class $V_i$ induces a forest; namely, $G[f^{-1}(i)]$ is an acyclic graph for each $i \in \{1, 2, \ldots, k\}$. Vertex-arboricity, also known as point arboricity, was first introduced by G. Chartrand, H.V. Kronk, and C.E.

*Corresponding Author.
Email addresses: m.fatehi.h@gmail.com (M. Fatehi), khashyar@ipm.ir (K. Khashyarmanesh), abbasmohammadian1248@gmail.com (A. Mohammadian)
Received: 10.07.2019; Accepted: 03.05.2020

Wall [4] in 1968. Note that a graph with no cycles is a forest, and it has vertex-arboricity one.

Likewise, the arboricity of a graph $G$, denoted by $\nu(G)$, is the least number of line-disjoint spanning forests into which $G$ can be partitioned, that is, there is some collection of $\nu(G)$ subgraphs of $G$, where each subgraph is a forest and each edge in $G$ is in exactly one such subgraph. Arboricity of a graph was first introduced by C. St. J. A. Nash-Williams [4] in 1964.

The main purpose of this paper is to characterize all finite commutative rings whose total graph has vertex-arboricity (arboricity) two or three. In addition, we show that, for a positive integer $v$, there are only finitely many finite rings whose total graph has vertex-arboricity (arboricity) $v$.

Now, we recall some definitions of graph theory which are necessary in this article. Let $G = (V(G), E(G))$ be a graph with vertex set $V(G)$ and edge set $E(G)$. We use $n$ and $e$ to denote the number of vertices and the number of edges of $G$, respectively. A graph in which each pair of distinct vertices is joined by an edge is called a *complete graph*. We use $K_n$ to denote the complete graph with $n$ vertices. A *bipartite graph* $G$ is a graph whose vertex set $V(G)$ can be partitioned into two subsets $V_1$ and $V_2$ such that the edge set of such a graph consists of precisely those edges which join vertices in $V_1$ to vertices of $V_2$. In particular, if $E(G)$ consists of all possible such edges, then $G$ is called the *complete bipartite graph* and denoted by the symbol $K_{r,s}$, where $|V_1| = r$ and $|V_2| = s$. For a vertex $x \in V(G)$, $\deg(x)$ is the *degree of vertex* $x$, $\delta(G) = \min\{deg(x) : x \in V(G)\}$, $\Delta(G) = \max\{\deg(x) : x \in V(G)\}$. For a nonnegative integer $d$, a graph is called *$d$-regular* if every vertex has degree $d$. Let $S \subset V(G)$ be any subset of vertices of $G$. Then the *induced subgraph* $G[S]$ is the graph whose vertex set is $S$ and whose edge set consists of all of the edges in $E(G)$ that have both endpoints in $S$. A *spanning subgraph* for $G$ is a subgraph of $G$ which contains every vertex of $G$. A graph without any cycle is called *acyclic graph*. A *forest* is an acyclic graph. Let $G_1$ and $G_2$ be subgraphs of $G$, we say that $G_1$ and $G_2$ are *disjoint* if they have no vertex and no edge in common. The *union* of two disjoint graphs $G_1$ and $G_2$, which is denoted by $G_1 \cup G_2$ is a graph with $V(G_1 \cup G_2) = V(G_1) \cup V(G_2)$ and $E(G_1 \cup G_2) = E(G_1) \cup E(G_2)$. For any graph $G$, the disjoint union of $k$ copies of $G$ is denoted by $kG$. Graphs $G$ and $H$ are said to be *isomorphic* to one another, written $G \cong H$, if there exists a one-to-one correspondence $f : V(G) \to V(H)$ such that for each pair $x, y$ of vertices of $G$, $xy \in E(G)$ if and only if $f(x)f(y) \in E(H)$. Also, for a rational number $p$, $\lceil p \rceil$ is the first integer number greater than or equal to $p$, and $\lfloor p \rfloor$ is the first integer number less than or equal to $p$.

## 2. Basic properties

First of all, let us recall some of the basic facts about total graphs and vertex arboricity, which we shall use in the rest of the paper.

**Lemma 2.1** ([7, Lemma 1.1])**.** *Let $x$ be a vertex of $T(\Gamma(R))$. Then the following statements are true.*

(i) *If $2 \in Z(R)$, then $\deg(x) = |Z(R)| - 1$.*
(ii) *If $2 \notin Z(R)$, then $\deg(x) = |Z(R)| - 1$ for every $x \in Z(R)$ and $\deg(x) = |Z(R)|$ for every vertex $x \notin Z(R)$.*

**Remark 2.2.** It is clear that $va(G) = 1$ if and only if $G$ is acyclic. For a few classes of graphs, the vertex-arboricity is easily determined. For example, $va(C_n) = 2$, where $C_n$ is a cycle graph with $n$ vertices. If $n$ is even, $va(K_n) = \frac{n}{2}$; while if $n$ is odd, $va(K_n) = \frac{n+1}{2}$. So, in general, $va(K_n) = \lceil \frac{n}{2} \rceil$. Also, $va(K_{r,s}) = 1$ if $r = 1$ or $s = 1$, and $va(K_{r,s}) = 2$ otherwise.

**Lemma 2.3** ([3, Lemma 1])**.** *Let $G$ be the disjoint union of graphs $G_1, G_2, \ldots, G_k$. Then, for all $i$ with $1 \leq i \leq k$,*

$$va(G) = \max\ va(G_i).$$

Now, we are ready to show that for a positive integer $v$, there are only finitely many finite rings whose total graph has vertex-arboricity $v$.

**Theorem 2.4.** *For any positive integer $v$, the number of finite rings whose total graphs have vertex-arboricity $v$ is finite.*

***Proof.*** Let $R$ be a finite ring. We want to obtain a complete subgraph (with vertex set $T$) of $T(\Gamma(R))$. To achieve this, we consider the following two cases:

(a) $R$ is local. In this case $Z(R)$ is the maximal ideal of $R$ and $|R| \leq |Z(R)|^2$ [8]. In this situation, we put $T = Z(R)$.

(b) $R$ is not local. Then there is a natural number $n \geq 2$ and there are local rings $R_1, R_2, \ldots, R_n$ such that $R = R_1 \times R_2 \times \cdots \times R_n$. We may assume that $|R_1| \leq |R_2| \leq \cdots \leq |R_n|$. Now put $R_1^* = 0 \times R_2 \times \cdots \times R_n$. Since $|R| = |R_1||R_1^*|$, we have $|R| \leq |R_1^*|^2$. In this situation, we put $T = R_1^*$.

Now, it is easy to see that, for every elements $x$ and $y$ of $T$, $x$ is adjacent to $y$ in $T(\Gamma(R))$. Thus there is an induced subgraph $K_{|T|}$ in $T(\Gamma(R))$. Hence Remark 2.2 implies that $va(K_{|T|}) \leq v$, and so $\lceil \frac{|T|}{2} \rceil \leq v$. Thus $|R| \leq 4v^2$, and so the proof is complete. □

Let $Reg(\Gamma(R))$ be the induced subgraph of $T(\Gamma(R))$ with vertices $Reg(R) = R - Z(R)$, and $Z(\Gamma(R))$ be the induced subgraph of $T(\Gamma(R))$ with vertices $Z(R)$. Next, we record some facts concerning total graphs. If $Z(R)$ is an ideal of $R$, then $Z(\Gamma(R))$ is a complete subgraph of $T(\Gamma(R))$ and is disjoint from $Reg(\Gamma(R))$. Thus, the following theorem of D.F. Anderson and A. Badawi gives a complete description of $T(\Gamma(R))$.

**Theorem 2.5** ([1, Theorem 2.2])**.** *Let $R$ be a commutative ring such that $Z(R)$ is an ideal of $R$, and let $|Z(R)| = n$ and $|\frac{R}{Z(R)}| = m$. Then the following statements hold.*

(i) *If $2 \in Z(R)$, then $Reg(\Gamma(R))$ is the union of $m-1$ disjoint $K_n$'s.*
(ii) *If $2 \notin Z(R)$, then $Reg(\Gamma(R))$ is the union of $\frac{m-1}{2}$ disjoint $K_{n,n}$'s.*

**Theorem 2.6.** *Let $R$ be a finite commutative ring with identity and $I$ be a nontrivial ideal contained in $Z(R)$. Set $|I| = n$ and $|\frac{R}{I}| = m$. Then the following statements hold.*

(i) *If $2 \in I$, then $va(T(\Gamma(R))) \geq \lceil \frac{n}{2} \rceil$.*
(ii) *If $2 \notin I$, then $va(T(\Gamma(R))) \geq \max\{\lceil \frac{n}{2} \rceil, 2\}$.*

***Proof.*** Let $G$ be the spanning subgraph of $T(\Gamma(R))$ such that, for every two vertices $x, y \in R$, $x$ is adjacent to $y$ in $G$ if $x + y \in I$. Now, since $I$ is an ideal of $R$ contained in $Z(R)$, by making obvious modification to the proof of Theorem 2.5, one can show that

$$G = \begin{cases} mK_n & \text{if } 2 \in I \\ K_n \bigcup (\frac{m-1}{2})K_{n,n} & \text{if } 2 \notin I. \end{cases}$$

Now, by Remark 2.2 in conjunction with Lemma 2.3, we have the following equalities

$$va(G) = \begin{cases} \lceil \frac{n}{2} \rceil & \text{if } 2 \in I \\ \max\{\lceil \frac{n}{2} \rceil, 2\} & \text{if } 2 \notin I. \end{cases}$$

Now, since $G$ is a subgraph of $T(\Gamma(R))$, we have that $va(G) \leq va(T(\Gamma(R)))$, and so the proof is complete. □

The following corollary is immediate from Theorem 2.5.

**Corollary 2.7.** *Let $R$ be a finite commutative ring with identity, $Z(R)$ be nontrivial ideal of $R$ and set $|Z(R)| = n$ and $|\frac{R}{Z(R)}| = m$. Then the following statements hold.*

(i) *If $2 \in Z(R)$, then $va(T(\Gamma(R))) = \lceil \frac{n}{2} \rceil$.*
(ii) *If $2 \notin Z(R)$, then $va(T(\Gamma(R))) = \max\{\lceil \frac{n}{2} \rceil, 2\}$.*

## 3. The vertex-arboricity of the total graph

For any graph $G$, the girth of $G$, denoted by $gr(G)$, is the length of a shortest cycle in $G$ ($gr(G) = \infty$ if $G$ contains no cycles). The following Theorem of Anderson and Badawi implies that $T(\Gamma(R))$ has vertex-arboricity one if and only if either $R$ is an integral domain or $R$ is isomorphic to $\mathbb{Z}_4$ or $\frac{\mathbb{Z}_2[x]}{(x^2)}$.

**Theorem 3.1** ([2, Theorem 4.7]). *Let $R$ be a commutative ring. Then $gr(T(\Gamma(R))) \in \{3, 4, \infty\}$. Moreover,*

(i) *$gr(T(\Gamma(R))) = \infty$ if and only if either $R$ is an integral domain or $R$ is isomorphic to $\mathbb{Z}_4$ or $\frac{\mathbb{Z}_2[x]}{(x^2)}$,*

(ii) *$gr(T(\Gamma(R))) = 4$ if and only if $R$ is isomorphic to $\mathbb{Z}_2 \times \mathbb{Z}_2$, and*

(iii) *$gr(T(\Gamma(R))) = 3$ otherwise.*

Now, we will classify, up to isomorphism, all finite commutative rings whose total graphs have vertex-arboricity two or three. We begin with a following result which is essentially due to Raghavendran.

**Theorem 3.2** ([10, Theorem 2]). *Let $R$ be a finite commutative local ring with nonzero identity and $U(R)$ be the set of all unit elements of $R$. Then $|R| = p^{nr}$, $|Z(R)| = p^{(n-1)r}$ and $|U(R)| = p^{(n-1)r}(p^r - 1)$ for some prime $p$ and some positive integers $n$ and $r$.*

In sequel, we state two remarks which we will use throughout this paper.

**Remark 3.3.** Let $R_1$ and $R_2$ be two finite commutative rings with $|R_1| = m$, $|R_2| = n$ and $m \leq n$. It is easy to see that the subgraph of the total graph of $R_1 \times R_2$ induced by the set $\{0\} \times R_2$ is a copy of $K_n$.

**Remark 3.4.** Let $R_1, R_2, S_1$ and $S_2$ be finite commutative rings such that $T(\Gamma(R_1)) \cong T(\Gamma(R_2)$ and $T(\Gamma(S_1)) \cong T(\Gamma(S_2)$. Then $T(\Gamma(R_1 \times S_1)) \cong T(\Gamma(R_2 \times S_2)$. However, this property does not hold in general for other widely studied graphs associated to rings (for example, the zero-divisor graphs).

**Lemma 3.5.** $va(T(\Gamma(\mathbb{Z}_2 \times \mathbb{Z}_2 \times \mathbb{Z}_2))) = va(T(\Gamma(\mathbb{F}_4 \times \mathbb{F}_4))) = 3$.

***Proof.*** First of all, note that, in view of Remark 3.3, $va(T(\Gamma(\mathbb{Z}_2 \times \mathbb{Z}_2 \times \mathbb{Z}_2))) > 1$. Now, we show that $va(T(\Gamma(\mathbb{Z}_2 \times \mathbb{Z}_2 \times \mathbb{Z}_2))) > 2$. To this, we consider a set of vertices of the graph $T(\Gamma(\mathbb{Z}_2 \times \mathbb{Z}_2 \times \mathbb{Z}_2))$ of the form

$$A = \{(0,0,0), (1,0,0), (0,1,0), (0,0,1)\}.$$

Let the set $\{V_1, V_2\}$ be an acyclic partition of $V(T(\Gamma(\mathbb{Z}_2 \times \mathbb{Z}_2 \times \mathbb{Z}_2)))$. Since $G[A]$ is a complete graph isomorphic to $K_4$ and $G[V_i] (1 \leq i \leq 2)$ have no triangle, so $|A \cap V_1| = |A \cap V_2| = 2$. Without the loss of generality, we may assume that $(0,0,0), (1,0,0) \in V_1$ and $(0,1,0), (0,0,1) \in V_2$. Now, consider the vertex $(0,1,1)$ of $T(\Gamma(\mathbb{Z}_2 \times \mathbb{Z}_2 \times \mathbb{Z}_2))$. It is clear that $(0,1,1) \in V_1$. Therefore, each of the remaining vertex of the graph $T(\Gamma(\mathbb{Z}_2 \times \mathbb{Z}_2 \times \mathbb{Z}_2))$ forms a triangle with two vertices of $V_1$. Hence, all of these vertices must be in $V_2$, which is a contradiction.

Now, consider the partition of $V(T(\Gamma(\mathbb{Z}_2 \times \mathbb{Z}_2 \times \mathbb{Z}_2)))$ with sets $V_1 = \{(0,0,0), (0,1,0), (1,1,1)\}$, $V_2 = \{(1,0,0), (0,0,1), (0,1,1)\}$ and $V_3 = \{(1,0,1), (1,1,0)\}$. It is clear that the subgraphs of $T(\Gamma(\mathbb{Z}_2 \times \mathbb{Z}_2 \times \mathbb{Z}_2))$ induced by sets $V_1$, $V_2$ and $V_3$ are acyclic. Hence $va(T(\Gamma(\mathbb{Z}_2 \times \mathbb{Z}_2 \times \mathbb{Z}_2))) = 3$.

By Remark 3.3, we have $va(T(\Gamma(\mathbb{F}_4 \times \mathbb{F}_4))) > 1$. Assume that $B_y = \{(a, y) : a \in \mathbb{F}_4\}$ and $C_x = \{(x, b) : b \in \mathbb{F}_4\}$ for all $x, y \in \mathbb{F}_4$. Obviously, $\{B_y : y \in \mathbb{F}_4\}$ and $\{C_x : x \in \mathbb{F}_4\}$ both form partitions for $V(T(\Gamma(\mathbb{F}_4 \times \mathbb{F}_4)))$. Let $\{V_1, V_2\}$ be an acyclic partition of $V(T(\Gamma(\mathbb{F}_4 \times \mathbb{F}_4)))$. Since the subgraphs of $T(\Gamma(\mathbb{F}_4 \times \mathbb{F}_4))$ induced by sets $V_1$ and $V_2$ have no triangles, each of these sets has exactly two vertices of the sets $B_y$ and $C_x$ for all

$x, y \in \mathbb{F}_4$. Hence, each of the sets $V_1$ and $V_2$ has exactly two vertices such that their first components are the same and have exactly two vertices such that the second components are the same. So, each vertex in $V_1$ and $V_2$ has degree 2, which is a contradiction, since the subgraphs of $T(\Gamma(\mathbb{F}_4 \times \mathbb{F}_4))$ induced by the sets $V_1$ and $V_2$ are union of cycles. Thus we have $va(T(\Gamma(\mathbb{F}_4 \times \mathbb{F}_4))) > 2$.

Now, according to the Figure 1, we have $va(T(\Gamma(\mathbb{F}_4 \times \mathbb{F}_4))) = 3$.

□

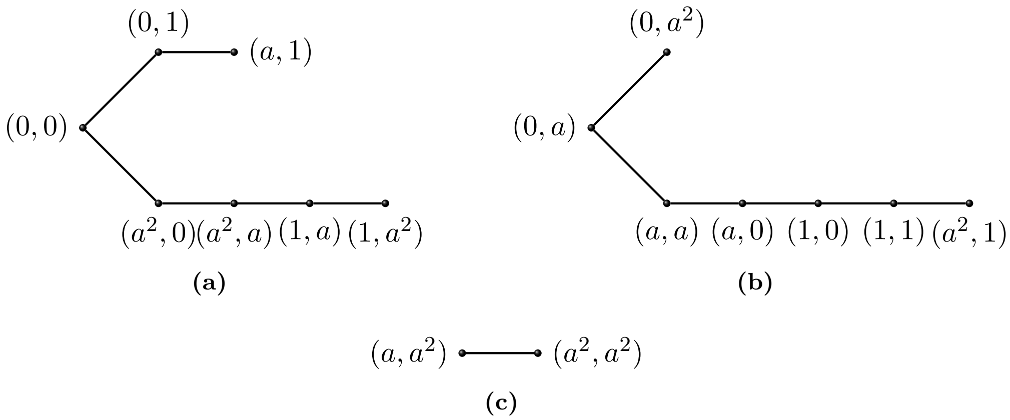



**Figure 1**

**Theorem 3.6.** *Let $R$ be a finite commutative ring such that $va(T(\Gamma(R))) = 2$. Then the following statements hold.*

(i) *If $R$ is local, then $R$ is isomorphic to one of the following rings:* $\mathbb{Z}_9$, $\frac{\mathbb{Z}_3[x]}{(x^2)}$, $\mathbb{Z}_8$, $\frac{\mathbb{Z}_2[x]}{(x^3)}$, $\frac{\mathbb{Z}_4[x]}{(2x,x^2-2)}$, $\frac{\mathbb{Z}_2[x,y]}{(x,y)^2}$, $\frac{\mathbb{Z}_4[x]}{(2,x)^2}$, $\frac{\mathbb{F}_4[x]}{(x^2)}$, $\frac{\mathbb{Z}_4[x]}{(x^2+x+1)}$.

(ii) *If $R$ is not local, then $R$ is isomorphic to one of the following rings:* $\mathbb{Z}_2 \times \mathbb{Z}_2$, $\mathbb{Z}_6$, $\mathbb{Z}_2 \times \mathbb{Z}_4$, $\mathbb{Z}_2 \times \frac{\mathbb{Z}_2[x]}{(x^2)}$, $\mathbb{Z}_2 \times \mathbb{F}_4$, $\mathbb{Z}_3 \times \mathbb{Z}_3$, $\mathbb{Z}_3 \times \mathbb{F}_4$.

***Proof.*** (i) Assume that $R$ is a local ring, and let $|Z(R)| = n$ and $|\frac{R}{Z(R)}| = m$. Then by Theorem 2.5, $T(\Gamma(R))$ has an induced subgraph isomorphic to $K_n$ and so by Remark 2.2, $|Z(R)| \leq 4$. Now, we consider the following two cases:

(a) If $2 \in Z(R)$, then by Theorem 3.2, $|R| = 2^k$ and $k \leq 4$. Since $va(T(\Gamma(R))) = 2$, Theorem 3.1 implies that $|R| = 16, 8$. According to Corbas and Williams [6] there are two non-isomorphic rings of order 16 with maximal ideals of order 4, namely $\frac{\mathbb{F}_4[x]}{(x^2)}$ and $\frac{\mathbb{Z}_4[x]}{(x^2+x+1)}$ (see also Redmond [11]), so for these rings have $T(\Gamma(R)) \cong 4K_4$. Therefore, by Remark 2.2, these rings have vertex-arboricity 2. In [6] it is also shown that there are 5 local rings of order 8 (except $\mathbb{F}_8$) as follows:

$$\mathbb{Z}_8, \frac{\mathbb{Z}_2[x]}{(x^3)}, \frac{\mathbb{Z}_4[x]}{(2x, x^2-2)}, \frac{\mathbb{Z}_2[x,y]}{(x,y)^2}, \frac{\mathbb{Z}_4[x]}{(2,x)^2}.$$

In all of these rings we have $|Z(R)| = 4$ and hence $T(\Gamma(R)) \cong 2K_4$. Then, by Remark 2.2, these rings have vertex-arboricity 2.

(b) If $2 \notin Z(R)$, then $|Z(R)| = 3$. According to [6], there are two rings of order 9 namely, $\mathbb{Z}_9$ and $\frac{\mathbb{Z}_3[x]}{(x^2)}$. For these rings, we have $T(\Gamma(R)) \cong K_3 \bigcup K_{3,3}$. Hence, by Corollary 2.7, these rings have vertex-arboricity 2.

(ii) Suppose that $R$ is not local. Since $R$ is finite, there are finite local rings $R_1, \ldots, R_t$ (with $t \geq 2$) such that $R = R_1 \times R_2 \times \cdots \times R_t$. Now, according to Remarks 2.2 and 3.3,

we have the following candidates:
$\mathbb{Z}_2 \times \mathbb{Z}_2$, $\mathbb{Z}_6$, $\mathbb{Z}_2 \times \mathbb{Z}_4$, $\mathbb{Z}_2 \times \frac{\mathbb{Z}_2[x]}{(x^2)}$, $\mathbb{Z}_2 \times \mathbb{F}_4$, $\mathbb{Z}_3 \times \mathbb{Z}_3$, $\mathbb{Z}_3 \times \mathbb{Z}_4$, $\mathbb{Z}_3 \times \frac{\mathbb{Z}_2[x]}{(x^2)}$, $\mathbb{Z}_3 \times \mathbb{F}_4$, $\mathbb{Z}_2 \times \mathbb{Z}_2 \times \mathbb{Z}_2$, $\mathbb{Z}_4 \times \mathbb{Z}_4$, $\mathbb{Z}_4 \times \frac{\mathbb{Z}_2[x]}{(x^2)}$, $\frac{\mathbb{Z}_2[x]}{(x^2)} \times \frac{\mathbb{Z}_2[x]}{(x^2)}$, $\mathbb{Z}_4 \times \mathbb{F}_4$, $\frac{\mathbb{Z}_2[x]}{(x^2)} \times \mathbb{F}_4$, $\mathbb{F}_4 \times \mathbb{F}_4$.
Now we examine each of the above rings.

The total graph of the ring $\mathbb{Z}_2 \times \mathbb{Z}_2$ is isomorphic to the cycle of size 4. We consider the acyclic partition $V_1 = \{(0,0),(1,0)\}$ and $V_2 = \{(0,1),(1,1)\}$ of $V(T(\Gamma(\mathbb{Z}_2 \times \mathbb{Z}_2)))$. Hence, the subgraphs of $T(\Gamma(\mathbb{Z}_2 \times \mathbb{Z}_2))$ induced by sets $V_1$ and $V_2$ are acyclic. Thus $va(T(\Gamma(\mathbb{Z}_2 \times \mathbb{Z}_2))) = 2$.

For $\mathbb{Z}_6$, by considering the acyclic partition $V_1 = \{0,1,3\}$ and $V_2 = \{2,4,6\}$ of $V(T(\Gamma(\mathbb{Z}_6)))$, we have $va(T(\Gamma(\mathbb{Z}_6))) = 2$.

For $\mathbb{Z}_2 \times \mathbb{Z}_4$, we put $V_1 = \{(0,0),(0,2),(1,1),(1,3)\}$ and $V_2 = \{(0,1),(0,3),(1,0),(1,2)\}$. Now, it is easy to see that $va(T(\Gamma(\mathbb{Z}_2 \times \mathbb{Z}_4))) = 2$. Since $T(\Gamma(\mathbb{Z}_4)) \cong T(\Gamma(\frac{\mathbb{Z}_2[x]}{(x^2)}))$, by Remark 3.4, we have $T(\Gamma(\mathbb{Z}_2 \times \mathbb{Z}_4)) \cong T(\Gamma(\mathbb{Z}_2 \times \frac{\mathbb{Z}_2[x]}{(x^2)}))$. Thus $va(T(\Gamma(\mathbb{Z}_2 \times \frac{\mathbb{Z}_2[x]}{(x^2)}))) = 2$.

For $\mathbb{Z}_2 \times \mathbb{F}_4$, by using the acyclic partition

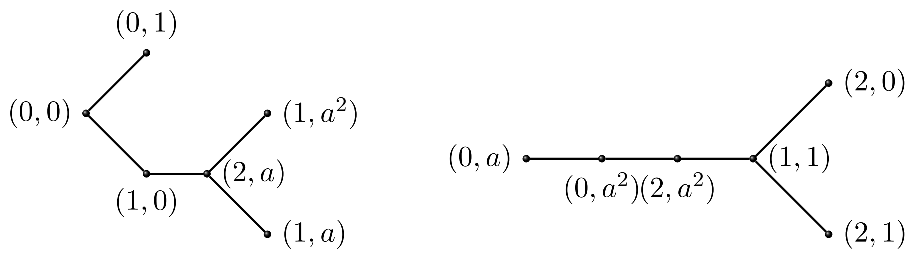
$$V_1 = \{(0,0),(0,1),(1,0),(1,a)\} \text{ and } V_2 = \{(0,a),(0,a^2),(1,1),(1,a^2)\}$$

of $V(T(\Gamma(\mathbb{Z}_2 \times \mathbb{F}_4)))$, we have $va(T(\Gamma(\mathbb{Z}_2 \times \mathbb{F}_4))) = 2$.

For $\mathbb{Z}_3 \times \mathbb{Z}_3$, we consider the acyclic partition $V_1 = \{(0,0),(0,1),(1,0),(1,1),(2,1)\}$ and $V_2 = \{(0,2),(2,0),(1,2),(2,2)\}$ of $V(T(\Gamma(\mathbb{Z}_3 \times \mathbb{Z}_3)))$. Hence $va(T(\Gamma(\mathbb{Z}_3 \times \mathbb{Z}_3))) = 2$.

For $\mathbb{Z}_3 \times \mathbb{Z}_4$, the graph $T(\Gamma(\mathbb{Z}_3 \times \mathbb{Z}_4))$ has a complete graph $K_6$ as a subgraph with vertex set $\{(0,0),(1,0),(2,0),(0,2),(1,2),(2,2)\}$, and so, by Remark 2.2, we have $va(T(\Gamma(\mathbb{Z}_3 \times \mathbb{Z}_4))) > 2$. Also by Remark 3.4, we have $T(\Gamma(\mathbb{Z}_3 \times \mathbb{Z}_4)) \cong T(\Gamma(\mathbb{Z}_3 \times \frac{\mathbb{Z}_2[x]}{(x^2)}))$. Thus $va(T(\Gamma(\mathbb{Z}_3 \times \frac{\mathbb{Z}_2[x]}{(x^2)}))) > 2$.

For $\mathbb{Z}_3 \times \mathbb{F}_4$, according to the Figure 2 we have $va(T(\Gamma(\mathbb{Z}_3 \times \mathbb{F}_4))) = 2$.




**(a)** **(b)**

**Figure 2**

For $\mathbb{Z}_2 \times \mathbb{Z}_2 \times \mathbb{Z}_2$, by Lemma 3.5, we have $va(T(\Gamma(\mathbb{Z}_2 \times \mathbb{Z}_2 \times \mathbb{Z}_2))) > 2$.

For $\mathbb{Z}_4 \times \mathbb{Z}_4$, the graph $T(\Gamma(\mathbb{Z}_4 \times \mathbb{Z}_4))$ has a $K_8$ as a subgraph with vertex set

$$\{(0,0),(1,0),(2,0),(3,0),(0,2),(1,2),(2,2),(3,2)\},$$

and so, by Remark 2.2, we have $va(T(\Gamma(\mathbb{Z}_4 \times \mathbb{Z}_4))) > 3$.

According to Remark 3.4, $T(\Gamma(\mathbb{Z}_4 \times \mathbb{Z}_4)) \cong T(\Gamma(\mathbb{Z}_4 \times \frac{\mathbb{Z}_2[x]}{(x^2)})) \cong T(\Gamma(\frac{\mathbb{Z}_2[x]}{(x^2)} \times \frac{\mathbb{Z}_2[x]}{(x^2)}))$. So the vertex-arboricity of graphs $T(\Gamma(\mathbb{Z}_4 \times \frac{\mathbb{Z}_2[x]}{(x^2)}))$ and $T(\Gamma(\frac{\mathbb{Z}_2[x]}{(x^2)} \times \frac{\mathbb{Z}_2[x]}{(x^2)}))$ is greater than three.

For $\mathbb{Z}_4 \times \mathbb{F}_4$, the graph $T(\Gamma(\mathbb{Z}_4 \times \mathbb{F}_4))$ has a $K_8$ as a subgraph with vertex set

$$\{(0,0),(0,1),(0,a),(0,a^2),(2,0),(2,1),(2,a),(2,a^2)\},$$

and so, by Remark 2.2, we have $va(T(\Gamma(\mathbb{Z}_4 \times \mathbb{F}_4))) > 3$. Also by Remark 3.4, $T(\Gamma(\mathbb{Z}_4 \times \mathbb{F}_4)) \cong T(\Gamma(\frac{\mathbb{Z}_2[x]}{(x^2)} \times \mathbb{F}_4))$. Therefore $va(T(\Gamma(\frac{\mathbb{Z}_2[x]}{(x^2)} \times \mathbb{F}_4))) > 3$.

For $\mathbb{F}_4 \times \mathbb{F}_4$, by Lemma 3.5, we have $va(T(\Gamma(\mathbb{F}_4 \times \mathbb{F}_4))) > 2$. □

**Lemma 3.7.** *For the ring $\mathbb{Z}_2 \times \mathbb{Z}_2 \times \mathbb{Z}_3$, $va(T(\Gamma(\mathbb{Z}_2 \times \mathbb{Z}_2 \times \mathbb{Z}_3))) = 4$.*

***Proof.*** First, by Remark 3.3, we have $va(T(\Gamma(\mathbb{Z}_2 \times \mathbb{Z}_2 \times \mathbb{Z}_3))) > 2$.

Now, let $T(\Gamma(\mathbb{Z}_2 \times \mathbb{Z}_2 \times \mathbb{Z}_3)) = G$ and $A = A_0 \cup A_1$, where $A_0 = \{(0,0,z) : z \in \mathbb{Z}_3\}$ and $A_1 = \{(0,1,z) : z \in \mathbb{Z}_3\}$. Also put $B = B_0 \cup B_1$, where $B_0 = \{(1,0,z) : z \in \mathbb{Z}_3\}$ and $B_1 = \{(1,1,z) : z \in \mathbb{Z}_3\}$. It is clear that the two sets $A$ and $B$ are partition for $V(G)$. Let $\{V_1, V_2, V_3\}$ be an acyclic partition for $V(G)$. If $|V_j| \geq 5$ for some $j \in \{1,2,3\}$, then $|A \cap V_j| \geq 3$ or $|B \cap V_j| \geq 3$, which is impossible, since $G[A]$ and $G[B]$ are complete graphs isomorphic to $K_6$ and $G[V_i]$ $(1 \leq i \leq 3)$ are acyclic induced subgraphs of $G$. Therefore $|V_i| = 4$ for some $i \in \{1,2,3\}$.

We know that every vertex of $G[A_0]$ $(G[A_1])$ are adjacent to every vertex of $G[B_0]$ $(G[B_1])$ and $G[V_i]$ $(1 \leq i \leq 3)$ are acyclic induced subgraphs of $G$. Hence without the loss of generality we can assume that $|A_0 \cap V_1| = |B_1 \cap V_1| = 2$ and $|A_1 \cap V_2| = |B_0 \cap V_2| = 2$. Then $V_3 = \{a_0, a_1, b_0, b_1 : a_s \in A_s, b_t \in B_t, 0 \leq s,t \leq 1\}$. It follows that $G[V_3]$ is a cycle of length 4, which is a contradiction and so $va(G) > 3$.

Now, by using the following partition of $V(G)$, we have that $va(G) = 4$.

$$V_1 = \{(0,0,0),(1,0,0),(1,1,2)\}, \qquad V_2 = \{(0,1,0),(1,1,1),(1,0,1)\},$$
$$V_3 = \{(0,1,2),(0,0,2),(1,0,2)\}, \qquad V_4 = \{(0,0,1),(0,1,1),(1,1,0)\}.$$

□

**Theorem 3.8.** *Let $R$ be a finite commutative ring such that $va(T(\Gamma(R))) = 3$. Then the following statements hold.*

(i) *If $R$ is local, then $R$ is isomorphic to $\mathbb{Z}_{25}$ or $\frac{\mathbb{Z}_5[x]}{(x^2)}$.*

(ii) *If $R$ is not local, then $R$ is isomorphic to one of the following rings:*

$\mathbb{Z}_3 \times \mathbb{Z}_4$, $\mathbb{Z}_3 \times \frac{\mathbb{Z}_2[x]}{(x^2)}$, $\mathbb{Z}_2 \times \mathbb{Z}_2 \times \mathbb{Z}_2$, $\mathbb{F}_4 \times \mathbb{F}_4$, $\mathbb{Z}_2 \times \mathbb{Z}_5$, $\mathbb{Z}_3 \times \mathbb{Z}_5$, $\mathbb{F}_4 \times \mathbb{Z}_5$, $\mathbb{Z}_5 \times \mathbb{Z}_5$.

***Proof.*** (i) Assume that $R$ is a local ring. We consider the following two cases:

(a) If $2 \in Z(R)$, then, by Theorem 2.5, we have $T(\Gamma(R)) \cong mK_n$. Hence, by Remark 2.2, $5 \leq |Z(R)| \leq 6$. But, in this situation $2 \in Z(R)$, and so, there are no such local rings.

(b) If $2 \notin Z(R)$, then, by Theorem 2.5, we have $T(\Gamma(R)) \cong K_n \bigcup (\frac{m-1}{2})K_{n,n}$. Hence, by Remark 2.2, $5 \leq |Z(R)| \leq 6$. Therefore $|Z(R)| = 5$ and so there exist two local rings, $\mathbb{Z}_{25}$ and $\frac{\mathbb{Z}_5[x]}{(x^2)}$ of order 25. For these rings we have $T(\Gamma(R)) \cong K_5 \bigcup 2K_{5,5}$. Hence, by Corollary 2.7, we have $va(T(\Gamma(R))) = 3$.

(ii) Suppose that $R$ is not a local ring. Arguments similar to those used in proof of Theorem 3.6 (ii), in conjunction with Remarks 2.2 and 3.3 show that we have the following candidates:

$\mathbb{Z}_2 \times \mathbb{Z}_2$, $\mathbb{Z}_6$, $\mathbb{Z}_2 \times \mathbb{Z}_4$, $\mathbb{Z}_2 \times \frac{\mathbb{Z}_2[x]}{(x^2)}$, $\mathbb{Z}_2 \times \mathbb{F}_4$, $\mathbb{Z}_3 \times \mathbb{Z}_3$, $\mathbb{Z}_3 \times \mathbb{Z}_4$, $\mathbb{Z}_3 \times \frac{\mathbb{Z}_2[x]}{(x^2)}$, $\mathbb{Z}_3 \times \mathbb{F}_4$, $\mathbb{Z}_2 \times \mathbb{Z}_2 \times \mathbb{Z}_2$, $\mathbb{Z}_4 \times \mathbb{Z}_4$, $\mathbb{Z}_4 \times \frac{\mathbb{Z}_2[x]}{(x^2)}$, $\frac{\mathbb{Z}_2[x]}{(x^2)} \times \frac{\mathbb{Z}_2[x]}{(x^2)}$, $\mathbb{Z}_4 \times \mathbb{F}_4$, $\frac{\mathbb{Z}_2[x]}{(x^2)} \times \mathbb{F}_4$, $\mathbb{F}_4 \times \mathbb{F}_4$, $\mathbb{Z}_2 \times \mathbb{Z}_5$, $\mathbb{Z}_2 \times \mathbb{Z}_2 \times \mathbb{Z}_3$, $\mathbb{Z}_3 \times \mathbb{Z}_5$, $\mathbb{Z}_4 \times \mathbb{Z}_5$, $\frac{\mathbb{Z}_2[x]}{(x^2)} \times \mathbb{Z}_5$, $\mathbb{F}_4 \times \mathbb{Z}_5$, $\mathbb{Z}_5 \times \mathbb{Z}_5$.

According to the proof of Theorem 3.6 (ii), we examine the following cases:

For $\mathbb{Z}_3 \times \mathbb{Z}_4$, we consider the partition

$$V_1 = \{(0,0),(1,1),(1,2),(1,3)\},$$
$$V_2 = \{(0,2),(2,0),(2,1),(2,3)\}$$

and

$$V_3 = \{(0,1),(0,3),(1,0),(2,2)\}$$

of $V(T(\Gamma(\mathbb{Z}_3 \times \mathbb{Z}_4)))$. The subgraphs of $T(\Gamma(\mathbb{Z}_3 \times \mathbb{Z}_4))$ induced by the sets $V_1$, $V_2$ and $V_3$ are acyclic graphs. Hence, we have $va(T(\Gamma(\mathbb{Z}_3 \times \mathbb{Z}_4))) = 3$. The Remark 3.4 implies that $T(\Gamma(\mathbb{Z}_3 \times \mathbb{Z}_4)) \cong T(\Gamma(\mathbb{Z}_3 \times \frac{\mathbb{Z}_2[x]}{(x^2)}))$ and so $va(T(\Gamma(\mathbb{Z}_3 \times \frac{\mathbb{Z}_2[x]}{(x^2)}))) = 3$.

For rings $\mathbb{Z}_2 \times \mathbb{Z}_2 \times \mathbb{Z}_2$ and $\mathbb{F}_4 \times \mathbb{F}_4$, by Lemma 3.5, we have $va(T(\Gamma(\mathbb{Z}_2 \times \mathbb{Z}_2 \times \mathbb{Z}_2))) = va(T(\Gamma(\mathbb{F}_4 \times \mathbb{F}_4))) = 3$.

For $\mathbb{Z}_2 \times \mathbb{Z}_5$, consider the acyclic partition $V_1 = \{(0,0),(0,1),(1,1),(1,2)\}$, $V_2 = \{(0,2),(0,3),(1,0),(1,4)\}$ and $V_3 = \{(0,4),(1,3)\}$ of $V(T(\Gamma(\mathbb{Z}_2 \times \mathbb{Z}_5)))$. Now, it is easy to see that $va(T(\Gamma(\mathbb{Z}_2 \times \mathbb{Z}_5))) = 3$.

For $\mathbb{Z}_2 \times \mathbb{Z}_2 \times \mathbb{Z}_3$, by Lemma 3.7, we have $va(T(\Gamma(\mathbb{Z}_2 \times \mathbb{Z}_2 \times \mathbb{Z}_3))) > 3$.

For $\mathbb{Z}_3 \times \mathbb{Z}_5$, by using the acyclic partition

$$V_1 = \{(0,4),(1,0),(1,3),(2,3)\},$$

$$V_2 = \{(0,0),(0,1),(1,2),(1,4),(2,1)\}$$

and

$$V_3 = \{(0,2),(0,3),(1,1),(2,0),(2,2),(2,4)\}$$

of $V(T(\Gamma(\mathbb{Z}_3 \times \mathbb{Z}_5)))$, we have $va(T(\Gamma(\mathbb{Z}_3 \times \mathbb{Z}_5))) = 3$.

For $\mathbb{Z}_4 \times \mathbb{Z}_5$, the graph $T(\Gamma(\mathbb{Z}_4 \times \mathbb{Z}_5))$ has a complete graph $K_{10}$ as a subgraph with vertex set $\{(0,0),(0,1),(0,2),(0,3),(0,4),(2,0),(2,1),(2,2),(2,3),(2,4)\}$, and so, we have $va(T(\Gamma(\mathbb{Z}_4 \times \mathbb{Z}_5))) \geq 5$. Also, Remark 3.4, $T(\Gamma(\mathbb{Z}_4 \times \mathbb{Z}_5)) \cong T(\Gamma(\frac{\mathbb{Z}_2[x]}{(x^2)} \times \mathbb{Z}_5))$ and so $va(T(\Gamma(\frac{\mathbb{Z}_2[x]}{(x^2)} \times \mathbb{Z}_5))) \geq 5$.

For $\mathbb{F}_4 \times \mathbb{Z}_5$, according to Figure 3, we have $va(T(\Gamma(\mathbb{F}_4 \times \mathbb{Z}_5))) = 3$.

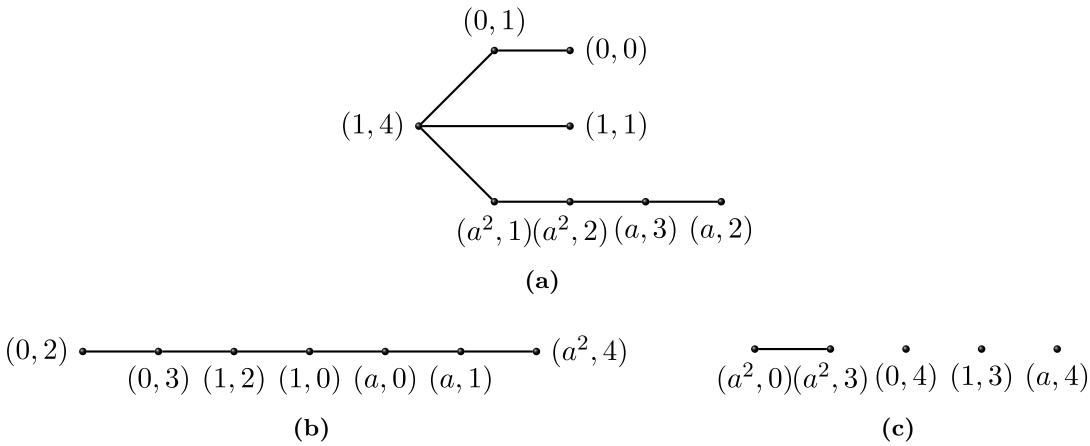



**Figure 3**

For $\mathbb{Z}_5 \times \mathbb{Z}_5$, by Figure 4, we conclude that $va(T(\Gamma(\mathbb{Z}_5 \times \mathbb{Z}_5))) = 3$.

Thus the proof is complete. $\square$

## 4. The arboricity of the total graph

In this section, we characterize all finite commutative rings whose total graph has arboricity two or three. In addition, we show that, for a positive integer $v$, there are only finitely many finite rings whose total graph has arboricity $v$. We begin the section with the following result of C. St. J. A. Nash-Williams.

**Theorem 4.1** ([9]). *For a graph $G$, $\nu(G) = \max\lceil \frac{e_H}{n_H - 1} \rceil$, where $n_H = |V(H)|$, $e_H = |E(H)|$ and $H$ ranges over all non-trivial induced subgraphs of $G$.*

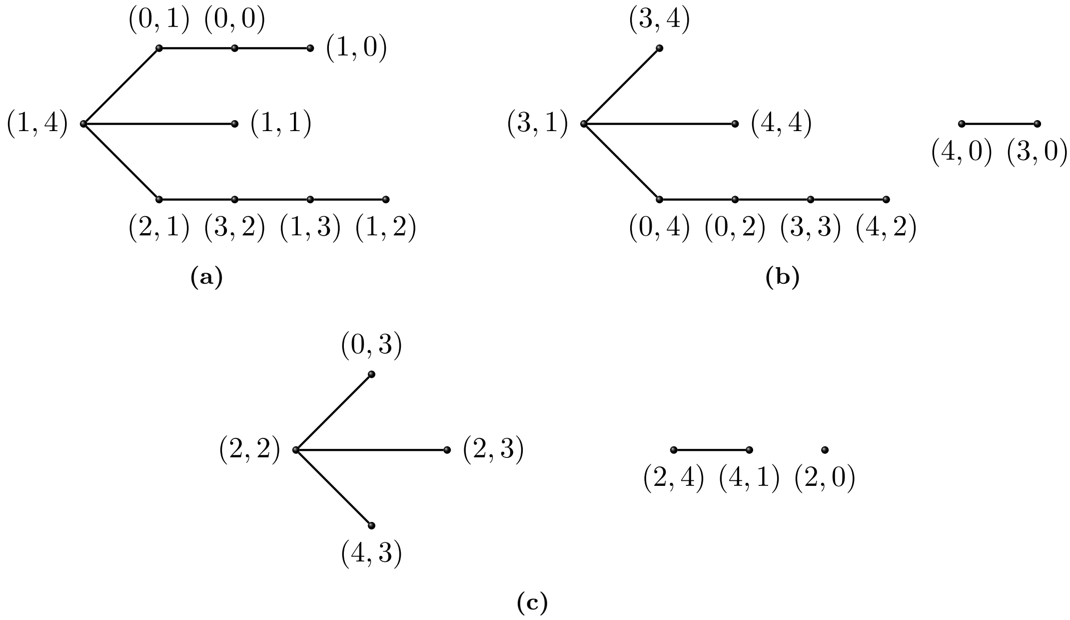



**Figure 4**

**Theorem 4.2.** *For a graph $G$, $\lceil \frac{\delta(G)+1}{2} \rceil \leq \nu(G) \leq \lceil \frac{\Delta(G)+1}{2} \rceil$. In particular, if $G$ is d-regular, then $\nu(G) = \lceil \frac{d+1}{2} \rceil = \lceil \frac{e}{n-1} \rceil$, where $n = |V(G)|$ and $e = |E(G)|$.*

***Proof.*** First, it is clear that, if $G$ has some isolated vertices, say $X = \{x_1, x_2, \ldots, x_k\}$, then $\nu(G) = \nu(G[V(G) \setminus X])$. So, we can assume that $G$ has no isolated vertices. Let $H$ be a subgraph of $G$ with $|V(H)| = n'$ and $|E(H)| = e'$. Then we have

$$\frac{e'}{n'-1} \leq \frac{\Delta(H)n'}{2(n'-1)} = \frac{1}{2}(\Delta(H) + \frac{\Delta(H)}{n'-1}).$$

Since $\Delta(H) \leq \min\{\Delta(G), n'-1\}$, we have $\frac{e'}{n'-1} \leq \frac{\Delta(G)+1}{2}$, and hence, by Theorem 4.1, $\nu(G) \leq \lceil \frac{\Delta(G)+1}{2} \rceil$. On the other hand $\frac{e}{n-1} \geq \frac{\delta(G)n}{2(n-1)} > \frac{\delta(G)}{2}$. Since $\nu(G)$ is an integer, $\nu(G) \geq \lceil \frac{\delta(G)+1}{2} \rceil$, as required. □

Clearly, in view of the above theorem, $\nu(K_n) = \lceil \frac{n}{2} \rceil$. So, by arguing as in the proof of Theorem 2.4, we have the following theorem.

**Theorem 4.3.** *For any positive integer $v$, the number of finite rings $R$ whose total graph has arboricity $v$ is finite.*

Theorem 3.1 implies that $T(\Gamma(R))$ has arboricity one if and only if either $R$ is an integral domain or $R$ is isomorphic to $\mathbb{Z}_4$ or $\frac{\mathbb{Z}_2[x]}{(x^2)}$. Now, we will classify, up to isomorphism, all the finite commutative rings whose total graph has arboricity two or three.

**Theorem 4.4.** *Let $R$ be a finite ring such that $\nu(T(\Gamma(R))) = 2$. Then the following statements hold.*

(i) *If $R$ is local, then $R$ is isomorphic to one of the following rings:* $\mathbb{Z}_9$, $\frac{\mathbb{Z}_3[x]}{(x^2)}$, $\mathbb{Z}_8$, $\frac{\mathbb{Z}_2[x]}{(x^3)}$, $\frac{\mathbb{Z}_4[x]}{(2x, x^2-2)}$, $\frac{\mathbb{Z}_2[x,y]}{(x,y)^2}$, $\frac{\mathbb{Z}_4[x]}{(2,x)^2}$, $\frac{\mathbb{F}_4[x]}{(x^2)}$, $\frac{\mathbb{Z}_4[x]}{(x^2+x+1)}$.

(ii) *If $R$ is not local, then $R$ is isomorphic to $\mathbb{Z}_2 \times \mathbb{Z}_2$ or $\mathbb{Z}_6$.*

***Proof.*** (i) Assume that $R$ is a local ring. If $2 \in Z(R)$, then, by Lemma 2.1 and Theorem 4.2, we have $|Z(R)| = 4$. Then by Theorem 3.2, $|R| = 16, 8$. Now, by same argument of

Theorem 3.6, $R$ is isomorphic to one of the following rings:

$$\mathbb{Z}_8, \frac{\mathbb{Z}_2[x]}{(x^3)}, \frac{\mathbb{Z}_4[x]}{(2x, x^2-2)}, \frac{\mathbb{Z}_2[x,y]}{(x,y)^2}, \frac{\mathbb{Z}_4[x]}{(2,x)^2}, \frac{\mathbb{F}_4[x]}{(x^2)}, \frac{\mathbb{Z}_4[x]}{(x^2+x+1)}.$$

If $2 \notin Z(R)$, then $|Z(R)| = 3$. So, $R$ is isomorphic to $\mathbb{Z}_9$ or $\frac{\mathbb{Z}_3[x]}{(x^2)}$.

(ii) If $R$ is not a local ring, then, by Theorem 4.2, we have $3 \leq |Z(R)| \leq 4$. When $|Z(R)| = 3$, it is clear that $R$ is isomorphic to $\mathbb{Z}_2 \times \mathbb{Z}_2$. Moreover, if $|Z(R)| = 4$, then $R$ is isomorphic to $\mathbb{Z}_6$, and so the proof is complete. □

By slight modifications in the proof of Theorem 4.4, one can prove the following theorem.

**Theorem 4.5.** *Let $R$ be a finite ring such that $\nu(T(\Gamma(R))) = 3$. Then the following statements hold.*

(i) *If $R$ is local, then $R$ is isomorphic to $\mathbb{Z}_{25}$ or $\frac{\mathbb{Z}_5[x]}{(x^2)}$.*

(ii) *If $R$ is not local, then $R$ is isomorphic to one of the following rings:*

$$\mathbb{Z}_2 \times \mathbb{F}_4, \mathbb{Z}_3 \times \mathbb{Z}_3, \mathbb{Z}_2 \times \mathbb{Z}_4, \mathbb{Z}_2 \times \frac{\mathbb{Z}_2[x]}{(x^2)}, \mathbb{Z}_2 \times \mathbb{Z}_5, \mathbb{Z}_3 \times \mathbb{F}_4.$$

In general, we can determine the arboricity of the total graph as in the following theorem.

**Theorem 4.6.** *Let $R$ be a finite ring.*

(i) *If $2 \in Z(R)$, then $\nu(T(\Gamma(R))) = \lceil \frac{|Z(R)|}{2} \rceil$.*

(ii) *If $2 \notin Z(R)$, then the following statements hold.*

  (1) *If $|Z(R)| = 2k+1$, then $\nu(T(\Gamma(R))) = k+1$.*

  (2) *If $|Z(R)| = 2k$, then $k \leq \nu(T(\Gamma(R))) \leq k+1$.*

***Proof.*** It follows from Lemma 2.1 and Theorem 4.2. □

## References


[1] D.F. Anderson and A. Badawi, *The total graph of a commutative ring*, J. Algebra, **320**, 2706–2719, 2008.

[2] D.F. Anderson and A. Badawi, *The total graph of a commutative ring without the zero element*, J. Algebra Appl. **11** 1–18 pages, 2012.

[3] G.J. Chang, C. Chen and Y. Chen, *Vertex and tree arboricities of graphs*, J. Comb. Optim. **8** 295–306, 2004.

[4] G. Chartrand, H.V. Kronk and C.E. Wall, *The point arboricity of a graph*, Israel J. Math. **6**, 169–175, 1968.

[5] T.T. Chelvam and T. Asir, *On the genus of the total graph of a commutative ring*, Comm. Algebra, **41**, 142–153, 2013.

[6] B. Corbas and G.D. Williams, *Ring of order p5. II. Local rings*, J. Algebra, **231** (2), 691–704, 2000.

[7] H.R. Maimani, C. Wickham and S. Yassemi, *Rings whose total graph have genus at most one*, Rocky Mountain J. Math. **42**, 1551–1560, 2012.

[8] B.R. McDonald, *Finite rings with identity* , Pure Appl. Math. **28**, Marcel Dekker, Inc., New York, 1974.

[9] C.St.J.A. Nash-Williams, *Decomposition of finite graphs into forests*, Journal London Math. Soc, **39**, 12, 1964.

[10] R. Raghavendran, iFinite associative rings, Compositio Math. **21**, 195–229, 1969.

[11] S.P. Redmond, *On zero-divisor graphs of small finite commutative rings*, Discrete Math. **307**, 1155–1166, 2007.

Hacettepe Journal of
Mathematics & Statistics

Hacet. J. Math. Stat.
Volume 50 (1) (2021), 120 – 134
DOI : 10.15672/hujms.613404

Research Article

# Quasi regular modules and trivial extension

Chillumuntala Jayaram[1], Ünsal Tekir*[2], Suat Koç[2]

[1] *The University of the West Indies, Department of CMP, P.O. Box 64, Bridgetown, Barbados*
[2] *Marmara University, Department of Mathematics, Ziverbey, Goztepe, 34722, Istanbul, Turkey*

## Abstract

Recall that a ring $R$ is said to be a quasi regular ring if its total quotient ring $q(R)$ is *von Neumann regular*. It is well known that a ring $R$ is quasi regular if and only if it is a reduced ring satisfying the property: for each $a \in R$, $ann_R(ann_R(a)) = ann_R(b)$ for some $b \in R$. Here, in this study, we extend the notion of quasi regular rings and rings which satisfy the aforementioned property to modules. We give many characterizations and properties of these two classes of modules. Moreover, we investigate the (weak) quasi regular property of trivial extension.

**Mathematics Subject Classification (2020).** 16E50, 13A15

**Keywords.** von Neumann regular ring, quasi regular ring, von Neumann regular module, quasi regular module, trivial extension

## 1. Introduction

In this paper, all rings are assumed to be commutative with $1 \neq 0$ and all modules are nonzero unital. Let $R$ always denote such a ring and $M$ always denote such an $R$-module. The concept of von Neumann regular rings has an important place in commutative algebra. There have been many generalizations and applications of von Neumann regular rings to other areas such as graph theory. See, for example, [2] and [10]. Previously, recall that a ring $R$ is said to be a *von Neumann regular* (for short, vn-regular) ring if for each $x \in R$, $x = x^2y$ for some $y \in R$ [14]. Note that a ring $R$ is vn-regular if and only if for each $x \in R$, $(x) = (e)$ for some idempotent element $e \in R$, where $(x)$ is the principal ideal generated by $x \in R$ if and only if it is a reduced and zero dimensional ring, i.e, every prime ideal is maximal if and only if the localization $R_P$ of $R$ at $P$ is a field for each prime ideal $P$ of $R$. Jayaram and Tekir extend the notion of vn-regular rings to modules in terms of $M$-regular elements [8]. Let $M$ be an $R$-module. Then $e \in R$ is said to be an *$M$-regular* (resp., a weak idempotent) element if $eM = e^2M$ (resp., $em = e^2m$ for each $m \in M$). Note that all idempotent elements are weak idempotent and these concepts are equal when $M$ is a faithful module. $M$ is called a *vn-regular $R$-module* if for each $m \in M$, there is an $e \in R$ such that $Rm = eM = e^2M$ [8]. It is well known that a finitely

*Corresponding Author.
Email addresses: jayaram.chillumu@cavehill.uwi.edu (C. Jayaram), utekir@marmara.edu.tr (Ü. Tekir), suat.koc@marmara.edu.tr (S. Koç)
Received: 30.08.2019; Accepted: 04.05.2020

generated $R$-module $M$ is a vn-regular module if and only if for each $m \in M$, there is a weak idempotent element $e \in R$ such that $Rm = eM$ [8, Lemma 5].

One of the generalization of vn-regular rings is quasi regular (sometimes called complemented) rings. A ring $R$ is called a *quasi regular ring* if its total quotient ring $q(R)$ is a vn-regular ring. In [4, Theorem 2.2], it was shown that a ring $R$ is a quasi regular ring if and only if $R$ is a reduced ring and satisfies the condition: for each $a \in R$, $ann_R(ann_R(a)) = ann_R(b)$ for some $b \in R$, where $ann_R(a) = \{x \in R : xa = 0\}$. Here, we call a ring $R$ *weak quasi regular* (for short, wq-regular) if for each $a \in R$, $ann_R(ann_R(a)) = ann_R(b)$ for some $b \in R$. Note that all quasi regular rings are wq-regular. But the converse is not true: just consider a non-reduced principal ideal ring. For instance, $\mathbb{Z}_4$ is a wq-regular ring, but is not a quasi regular ring.

Our aim in this article is to extend the notion of quasi regular rings and wq-regular rings to modules. For the sake of thoroughness we give some definitions which we will need throughout this study. For each submodules $N$ and $K$ of $M$, the residual of $N$ by $K$ is defined by $(N :_R K) = \{r \in R : rK \subseteq N\}$. In particular, if $N = 0$, we use $ann_R(K)$ to denote $(0 :_R K)$. Also for each cyclic submodule $Rm$, we use $ann_R(m)$ instead of $ann_R(Rm)$. Similarly, for each ideal $J$ of $R$ and each submodule $K$ of $N$, one can define residual of $N$ by $J$ as $(N :_M J) = \{m \in M : \ Jm \subseteq N\}$. In case $N = 0$, we will use $ann_M(J)$ instead of $(0 :_M J)$ and also for each $a \in R$, we denote $ann_M(Ra)$ by $ann_M(a)$.

Also the set $Z(M)$ of zero divisors on $M$ and the set $T(M)$ of all torsion elements of $M$ are defined as follows:

$$Z(M) = \{a \in R : ann_M(a) \neq 0\} \text{ and}$$
$$T(M) = \{m \in M : ann_R(m) \neq 0\}.$$

Note that $T(M)$ is not always a submodule of $M$ and similarly $Z(M)$ may not be an ideal of $R$. $M$ is called a torsion free module if $T(M) = 0$. Also if $T(M) = M$, then $M$ is called a torsion module. Otherwise, we call that $M$ is a non-torsion module. Assume that $S = R - Z(M)$. It is easily seen that $S$ is a multiplicatively closed subset (briefly m.c.s) of $R$. Also the localization $M_S$ is an $R_S$-module and it is called the total quotient module of $M$. We denote the total quotient module by $q(M)$. We call that $M$ is a *quasi regular $R$-module* if its total quotient module $q(M)$ is a vn-regular $R_S$ module, where $S = R - Z(M)$. Moreover, $M$ is said to be a *wq-regular module* if for each $m \in M$, there is an $a \in R$ such that

$$ann_M(ann_R(m)) = ann_M(a).$$

A submodule $N$ of $M$ is said to be a $*$-*submodule* if

$$N = O(S) = \{m \in M : sm = 0 \text{ for some } s \in S\}$$

for some m.c.s $S \subseteq R$. $N$ is said to be an $\alpha$-*submodule* if for each $m_1, m_2 \in N$ with $ann_R(m_1) \cap ann_R(m_2) = ann_R(m_3)$, we have $m_3 \in N$. Also $N$ is called an *annihilator submodule* if $ann_M(ann_R(N)) = N$. We study relations between these submodules and establish many characterizations of wq-regular modules in terms of $*$-submodules, $\alpha$-submodules and annihilator submodules (see Theorem 2.9-2.31). Also we prove that if $q(M)$ is a finitely generated multiplication module (not necessarily $M$ is) and $M$ is a non-torsion module, then $M$ is a quasi regular module if and only if $M$ is a reduced wq-regular module (compare the result [4, Theorem 2.2]). We also investigate whether the notion of wq-regular modules is invariant under homomorphism and direct products. In Section 3, we determine when the trivial extension $R \propto M$ (idealization) of $M$ is quasi regular and wq-regular, respectively (see Proposition 3.1 and Theoerem 3.4). In Section 4, we investigate the extension of wq-regular modules. In particular, we show that when polynomial modules and formal power series modules are wq-regular (see Theorem 4.6).

## 2. Characterizations of quasi regular modules

Throughout the section, we will examine $*$-submodules, $\alpha$-submodules, annihilator submodules and use them to characterize wq-regular modules.

**Definition 2.1.** Let $q(M)$ be the total quotient module of an $R$-module $M$. Then

(i) $M$ is called a *quasi regular module* if its total quotient module is vn-regular.

(ii) $M$ is called a *wq-regular module* if for each $m \in M$, there is an $a \in R$ such that $ann_M(ann_R(m)) = ann_M(a)$.

**Example 2.2.** (i) Every torsion free module is wq-regular. To see this, take a nonzero element $m \in M$. Then $ann_R(m) = 0$, and so $ann_M(ann_R(m)) = M = ann_M(0)$.

(ii) Every simple module is a wq-regular module. Assume $M$ is a simple $R$-module. Then $Rm = M$ or $Rm = 0$ for every $m \in M$. If $Rm = 0$, then $ann_M(ann_R(m)) = 0 = ann_M(1)$. Otherwise, we would have $ann_M(ann_R(m)) = M = ann_M(0)$.

(iii) Assume $R$ is a principal ideal ring. Then for any $m \in M$, $ann_R(m) = (a)$ for some $a \in R$. Then we can conclude that $ann_M(ann_R(m)) = ann_M(a)$. Hence every module over a principal ideal ring $R$ is wq-regular.

**Example 2.3.** (i) Every vn-regular module is a quasi-regular module. To see this, take a vn-regular $R$-module $M$. Let $\frac{m}{s} \in q(M)$ for some $m \in M, s \in S = R - Z(M)$. Then note that $R_S(\frac{m}{s}) = (Rm)_S$. Also we have $Rm = xM = x^2M$ for some $x \in R$ because $M$ is vn-regular. Then we can conclude that

$$\begin{aligned} R_S(\frac{m}{s}) &= (Rm)_S = (xM)_S = \frac{x}{1}q(M) \\ &= (x^2M)_S = (\frac{x}{1})^2 q(M). \end{aligned}$$

Hence, $M$ is quasi regular $R$-module.

(ii) Every simple module is vn-regular [8, Example 2], hence a quasi regular module by (i). In particular, the $\mathbb{Z}$-module $\mathbb{Z}_p$ is a quasi regular module for each prime number $p$.

(iii) Let $n > 1$ be a square free integer, i.e, $n = p_1p_2 \cdots p_r$, where $p_i$'s are distinct prime numbers. Consider the $\mathbb{Z}$-module $\mathbb{Z}_n$. Then by [8, Example 5], $\mathbb{Z}_n$ is vn-regular and thus a quasi regular module by (i).

(iv) Let $n > 1$ be a non-square free integer. We may assume that $n = p_1^{\alpha_1} p_2^{\alpha_2} \dots p_r^{\alpha_r}$ for some distinct prime numbers $p_1, p_2, \dots, p_r$, where $\alpha_1 \geq 2$ and $\alpha_2, \alpha_3, \dots, \alpha_r \geq 1$. Consider the $\mathbb{Z}$-module $\mathbb{Z}_n$. Then note that $Z(\mathbb{Z}_n) = p_1\mathbb{Z} \cup p_2\mathbb{Z} \cup \dots \cup p_r\mathbb{Z}$ is a union of prime ideals of $\mathbb{Z}$. Now, take $S = \mathbb{Z} - Z(\mathbb{Z}_n)$. Then it is clear that $q(\mathbb{Z}_n)$ is a finitely generated multiplication $\mathbb{Z}_S$-module. Since $\mathbb{Z}_n$ is not a reduced ring, by [4, Theorem 2.2] its total quotient ring is not vn-regular. Now, it can be easily verified that

$$\overline{S} = \pi(S) = \{a + n\mathbb{Z} : \gcd(a, p_i) = 1 \text{ for each } 1 \leq i \leq r\}$$

is the set of regular elements of $\mathbb{Z}/n\mathbb{Z}$, where $\pi : \mathbb{Z} \to \mathbb{Z}/n\mathbb{Z}$ is the canonical homomorphism. Furthermore,

$$ann_{\mathbb{Z}_S}(q(\mathbb{Z}_n)) = (ann_{\mathbb{Z}}(\mathbb{Z}_n))_S = (n\mathbb{Z})_S$$

and also $\mathbb{Z}_S/ann_{\mathbb{Z}_S}(q(\mathbb{Z}_n)) \cong (\mathbb{Z}/n\mathbb{Z})_{\overline{S}}$. Again by [4, Theorem 2.2], $\mathbb{Z}_S/ann_{\mathbb{Z}_S}(q(\mathbb{Z}_n))$ is not a vn-regular ring. Then by [8, Theorem 1], $q(\mathbb{Z}_n)$ is not a vn-regular $\mathbb{Z}_S$-module. Hence, $\mathbb{Z}_n$ is not a quasi regular $\mathbb{Z}$-module but wq-regular.

**Definition 2.4.** Let $N$ be a submodule of an $R$-module $M$. Then,

(i) $N$ is called a $*$*-submodule* if $N = O(S) = \{m \in M : sm = 0 \text{ for some } s \in S\}$, where $S \subseteq R$ is a m.c.s of $R$.

(ii) $(N :_R M)$ is a $*$*-ideal* if it is a $*$-submodule of the $R$-module $R$.

Let $N$ be a submodule of $M$. Then $N$ is called an *m-submodule* if $N = (N :_R M)M$. Note that an $R$-module $M$ is called a *multiplication module* if each submodule is an $m$-submodule [3].

**Lemma 2.5.** *(i) Let $M$ be a non-torsion module and $N$ a $*$-submodule of $M$. Then $(N : M)$ is a $*$-ideal of $R$.*

*(ii) Let $N$ be a prime $m$-submodule of $M$ in which $(N : M)$ is a $*$-ideal. Then $N$ is a $*$-submodule.*

***Proof.*** (i) Assume $N$ is a $*$-submodule of $M$. Then $N = O(S)$ for some m.c.s $S$ of $R$. As $M$ is non-torsion, we get $ann_R(m) = 0$ for some $m \in M$. Let $r \in (N :_R M)$. Then $rm \in N$, and so $s(rm) = 0$ for some $s \in S$. As $ann_R(m) = 0$, we have $sr = 0$. Now set $\overleftarrow{O(S)} = \{x \in R : sx = 0 \text{ for some } s \in S\}$. Note that $r \in \overleftarrow{O(S)}$, and so $(N :_R M) \subseteq \overleftarrow{O(S)}$. Let $x \in \overleftarrow{O(S)}$. Then $sx = 0$ for some $s \in S$. This implies that $s(xM) = 0$, and so $xM \subseteq O(S) = N$ and this yields $x \in (N :_R M)$. Accordingly, $(N :_R M) = \overleftarrow{O(S)}$ is a $*$-ideal of $R$.

(ii) Since $(N : M)$ is a $*$-ideal, $(N :_R M) = \overleftarrow{O(S)} = \{x \in R : sx = 0 \text{ for some } s \in S\}$, where $S$ is a m.c.s of $R$. Now, we will show that $N = O(S)$. Let $m \in N$. Since $N = (N :_R M)M$, we get $m = \sum_{i=1}^{n} a_i m_i$, $a_i \in (N :_R M)$ and $m_i \in M$. As $(N :_R M) = \overleftarrow{O(S)}$, there is $s_i \in S$ such that $s_i a_i = 0$ for each $i = 1, 2, ..., n$. Put $s = s_1 s_2 ... s_n$. Then note that $sm = \sum_{i=1}^{n} (sa_i)m_i = 0$, and so $m \in O(S)$. Then we conclude that $N \subseteq O(S)$. For the converse, take $m \in O(S)$. Then $sm = 0$ for some $s \in S$. It is clear that $S \cap (N :_R M) = \emptyset$ since $(N :_R M) = \overleftarrow{O(S)}$ and $0 \notin S$. This implies $s \notin (N :_R M)$, and so $m \in N$ as $N$ is a prime submodule. Accordingly, $N = O(S)$. $\square$

A submodule $N$ of an $R$-module $M$ is said to be a *Baer submodule* if for each $m \in N$, $ann_M(ann_R(m)) \subseteq N$.

**Definition 2.6.** A submodule $N$ of an $R$-module $M$ is said to be an *$\alpha$-submodule* if for each $m_1, m_2 \in N$ with $ann_R(m_1) \cap ann_R(m_2) = ann_R(m_3)$, we have $m_3 \in N$.

Baer ideals and $\alpha$-ideals are defined as Baer submodules and $\alpha$-submodules of the $R$-module $R$, respectively. In fact, $\alpha$-ideals are exactly strong Baer ideals of $R$ [7].

**Proposition 2.7.** (i) *Every $*$-submodule is a Baer submodule.*

(ii) *Assume $M$ is a module over a reduced ring $R$ satisfying the condition: for each $m \in M$, $ann_R(m) = ann_R(r)$ for some $r \in R$. Then every $\alpha$-submodule is a Baer submodule.*

(iii) *Every $*$-submodule is an $\alpha$-submodule.*

***Proof.*** (i) Assume that $N = O(S)$ for some m.c.s $S$ of $R$. Take $m \in N$. Then there is an $s \in S$ so that $sm = 0$. Let $m' \in ann_M(ann_R(m))$. Then we have $ann_R(m)m' = 0$, and so $sm' = 0$ since $s \in ann_R(m)$. This implies that $m' \in O(S) = N$. Thus $N$ is a Baer submodule.

(ii) Let $m' \in ann_M(ann_R(m))$ with $m \in N$. Then $ann_R(m)m' = 0$, and so $ann_R(m) \subseteq ann_R(m')$. By assumption, we have $ann_R(m) = ann_R(x)$ and $ann_R(m') = ann_R(y)$ for some $x, y \in R$. Then $ann_R(x) \subseteq ann_R(y)$. Since $R$ is a reduced ring, we have $ann_R(m') = ann_R(y) = ann_R(xy) = ann_R(ym)$. Since $N$ is an $\alpha$-submodule and $ym \in N$, we get $m' \in N$, and so $ann_M(ann_R(m)) \subseteq N$. Accordingly, $N$ is a Baer submodule.

(iii) Let $N$ be a $*$-submodule, i.e, $N = O(S)$ for some m.c.s $S$ of $R$. Assume that $ann_R(m) \cap ann_R(m') = ann_R(m'')$ with $m, m' \in N$ and $m'' \in M$. Then there are $s, s' \in S$ such that $sm = s'm' = 0$. Now put $t = ss'$. Then $t \in S$ and $t \in ann_R(m) \cap ann_R(m')$ and this yields that $tm'' = 0$. Thus we have $m'' \in O(S) = N$. Accordingly, $N$ is an $\alpha$-submodule of $M$. $\square$

Remember that $M$ is said to be a *reduced $R$-module* if for $r \in R, m \in M$ and $rm = 0$, we have $rM \cap Rm = 0$, or equivalently, $r^2 m = 0$ implies $rm = 0$ [9].

**Proposition 2.8.** (i) *Let $N$ be a prime $m$-submodule of a non-torsion module $M$. Then $N$ is a $*$-submodule if and only if $(N :_R M)$ is a $*$-ideal of $R$.*

(ii) *Let $M$ be a non-torsion reduced module over a quasi-regular ring $R$. Then any prime $m$-submodule $N$ of $M$ is a Baer submodule if and only if $(N :_R M)$ is a Baer ideal.*

***Proof.*** (i) It can be obtained from Lemma 2.5 (i) and (ii).

(ii) Assume $(N :_R M)$ is a Baer ideal and $N$ is a prime $m$-submodule of $M$. First note that $R$ is a reduced ring. By [7, Corollary 3], $(N :_R M)$ is a $*$-ideal of $R$. By Lemma 2.5 (ii), $N$ is a $*$-submodule. Then by Proposition 2.7 (i), $N$ is a Baer submodule of $M$. For the converse, assume $N$ is a Baer submodule. Let $r \in (N :_R M)$. As $M$ is non-torsion, we get $ann_R(m) = 0$ for some $m \in M$. Then note that $rm \in N$ and $ann_R(rm) = ann_R(r)$. As $N$ is a Baer submodule, we can conclude that $ann_M(ann_R(rm)) = ann_M(ann_R(r)) \subseteq N$. Now we will show that, for each ideal $I$ of $R$, $(ann_M(I) : M) = ann_R(I)$. The containment $ann_R(I) \subseteq (ann_M(I) : M)$ always holds. Let $x \in (ann_M(I) : M)$. Then $xM \subseteq ann_M(I)$, and so $I(xM) = 0$. This implies that $I(xm) = 0$, and so $Ix \subseteq ann_R(m) = 0$. Then we have $x \in ann_R(I)$, which yields $(ann_M(I) : M) = ann_R(I)$. Since $ann_M(ann_R(r)) \subseteq N$, we have $(ann_M(ann_R(r)) :_R M) = ann_R(ann_R(r)) \subseteq (N :_R M)$. Thus $(N :_R M)$ is a Baer ideal. □

We now characterize wq-regular modules in terms of $*$-submodules.

**Theorem 2.9.** *Let $M$ be a reduced faithful module. Then $M$ is a wq-regular module if and only if $ann_M(ann_R(m))$ is a $*$-submodule for each $m \in T(M)$.*

***Proof.*** Assume that $M$ is a wq-regular module. Take an element $m \in T(M)$. Then $ann_R(m) \neq 0$. As $M$ is a wq-regular module, $ann_M(ann_R(m)) = ann_M(r)$ for some $r \in R$. Since $M$ is faithful, $r \neq 0$. Otherwise, we would have $ann_R(m) = ann_R(M) = 0$, a contradiction. As $M$ is a reduced module, $R$ is a reduced ring, and so $S = \{r^n : n \in \mathbb{N}\}$ is an m.c.s of $R$. Also note that $ann_M(ann_R(m)) = ann_M(r) = O(S)$, and so $ann_M(ann_R(m))$ is a $*$-submodule. For the converse, assume $ann_M(ann_R(m))$ is a $*$-submodule for each $m \in T(M)$. Let $m \in M$. If $ann_R(m) = 0$, then $ann_M(ann_R(m)) = M = ann_M(0)$. Assume that $m \in T(M)$. By assumption, $ann_M(ann_R(m)) = O(S)$ for some m.c.s $S$ of $R$. This yields $rm = 0$ for some $r \in S$, which yields $ann_M(ann_R(m)) \subseteq ann_M(r)$. Let $m' \in ann_M(r)$. Then we have $rm' = 0$, and so $m' \in O(S) = ann_M(ann_R(m))$. Thus $ann_M(ann_R(m)) = ann_M(r)$. □

**Proposition 2.10.** *Let $M$ be a non-torsion wq-regular module. Then $R$ is a wq-regular ring and for each $m \in M$, there is an $r \in R$ such that $ann_R(m) = ann_R(r)$.*

***Proof.*** Let $r \in R$. Since $M \neq T(M)$, we get $ann_R(m) = 0$ for some $m \in M$ and also note that $ann_R(r) = ann_R(rm)$. As $M$ is wq-regular, there is an $s \in R$ such that $ann_M(s) = ann_M(ann_R(rm))$, and so $ann_M(s) = ann_M(ann_R(r))$. Then we conclude that

$$\begin{aligned} ann_R(ann_R(r)) &= (ann_M(ann_R(r)) :_R M) \\ &= (ann_M(s) :_R M) \\ &= ann_R(s). \end{aligned}$$

Therefore, $R$ is a wq-regular ring. Take an element $m^* \in M$. As $M$ is wq-regular, $ann_M(ann_R(m^*)) = ann_M(a)$ for some $a \in R$. This yields $ann_R(ann_R(m^*)) = ann_R(a)$, and so $ann_R(m^*) = ann_R(ann_R(a)) = ann_R(b)$ for some $b \in R$ because $R$ is a wq-regular ring. □

**Proposition 2.11.** *Assume $M$ is a non-torsion module and $ann_M(I)$ is an $m$-submodule of $M$ for each ideal $I$ of $R$. If $R$ is a wq-regular ring and for each $m \in M$, $ann_R(m) = ann_R(r)$ for some $r \in R$, then $M$ is a wq-regular module.*

***Proof.*** Assume $R$ is a wq-regular ring and for each $m \in M$, $ann_R(m) = ann_R(r)$ for some $r \in R$. Let $m \in M$. Then by assumption, $ann_R(m) = ann_R(r)$ for some $r \in R$. As $R$ is wq-regular, there is an $s \in R$ so that $ann_R(ann_R(r)) = ann_R(s)$, and so $(ann_M(ann_R(r)) :_R M) = ann_R(s)$. This yields that $(ann_M(ann_R(r)) :_R M) = (ann_M(s) :_R M)$. Since $ann_M(I)$ is an $m$-submodule for each ideal $I$ of $R$, we get

$$\begin{aligned} ann_M(ann_R(r)) &= (ann_M(ann_R(r)) :_R M)M \\ &= (ann_M(s) :_R M)M \\ &= ann_M(s). \end{aligned}$$

Accordingly, $M$ is a wq-regular module. □

The following example shows that an $R$-module satisfying all conditions in Proposition 2.11 may not be a multiplication module.

**Example 2.12.** Consider a torsion free module but not a multiplication module, e.g, a vector space $V$ over a field $F$ with $\dim_F(V) > 1$. Note that $V$ is a non-torsion module and for each $0 \neq m \in V$, $ann_F(m) = 0 = ann_F(1)$. Also it is easily seen that $ann_V(0) = V$ and $ann_V(F) = 0$ are m-submodules of $V$. But $V$ can not be a multiplication module.

The next Theorem 2.13 characterizes wq-regular modules in terms of wq-regular rings.

**Theorem 2.13.** *Let $M$ be a non-torsion module and $ann_M(I)$ is an $m$-submodule for each ideal $I$ of $R$. Then the followings are equivalent:*

(i) *$M$ is wq-regular module.*

(ii) *$R$ is wq-regular ring and for each $m \in M$, there is an $r \in R$ such that $ann_R(m) = ann_R(r)$.*

***Proof.*** It can be obtained from Proposition 2.10 and Proposition 2.11. □

**Definition 2.14.** Let $M$ be a finitely generated $R$-module. Then,

(i) $M$ is said to *satisfy the condition (#)* if $K$ is a minimal prime submodule, then $K = (K :_R M)M$.

(ii) $M$ is said to *satisfy the condition (P)* if $\bigcap(PM) = (\bigcap P)M$ for all prime ideals $P$ minimal over $ann_R(M)$.

(iii) $M$ is said to *satisfy the condition (##)* if it satisfies the condition (#) and (P).

Remark that a finitely generated multiplication module satisfies the conditions (#) and (##). But the converse is not true.

**Example 2.15.** Every finite dimensional vector space satisfies (#) and (##). In particular, consider the Euclidean Plane $\mathbb{R}$-module $\mathbb{R}^2$. Since 0 is a prime submodule, it is a minimal prime submodule. It is straightforward that the $\mathbb{R}$-module $\mathbb{R}^2$ satisfies (#) and (##). But it is not a multiplication module.

**Proposition 2.16.** *Let $M$ be a finitely generated module and $K$ be a submodule of $M$. Assume that $M$ satisfies the condition (#). Then*

(i) *If $P$ is a prime minimal over $ann_R(M)$, then $PM$ is a minimal prime submodule.*

(ii) *If $K$ is a minimal prime submodule, then $(K :_R M)$ is a prime ideal minimal over $ann_R(M)$.*

***Proof.*** (i) Assume $P$ is a prime ideal minimal over $ann_R(M)$. By [11, Proposition 8], $(PM :_R M) = P$. By [12, Theorem 3.3], $PM$ is contained in some prime submodule $N$ with $(N :_R M) = P$. Again by Zorn's Lemma, $PM$ is contained in $N_1$ where $N_1$ is a prime submodule minimal over $PM$ such that $(N_1 :_R M) = P$. The reader can easily verify that $N_1$ is a minimal prime submodule.

(ii) Assume $K$ is a minimal prime submodule. Thus $(K :_R M)$ is a prime ideal. Since $ann_R(M)$ is contained in $(K :_R M)$, there is a prime $P$ minimal over $ann_R(M)$ such that

$P$ is contained in $(K :_R M)$. So $PM$ is contained in $K$. By (i), $PM$ is a minimal prime submodule, thereby $PM = K$. Again $(K :_R M) = (PM :_R M) = P$ by [11, Proposition 8]. □

**Proposition 2.17.** *Let $M$ be a finitely generated module and $I$ be an ideal containing $ann_R(M)$. Assume that every prime submodule minimal over $IM$ is an $m$-submodule. Then*

(i) *If $P$ is minimal over $I$, then $PM$ is a prime minimal over $IM$.*
(ii) *If $K$ is minimal over $IM$, then $(K :_R M)$ is minimal over $I$.*

***Proof.*** The proof is similar to the proof of Proposition 2.16. □

We shall now prove several lemmas that we need.

**Lemma 2.18.** *Let $M$ be a non-torsion wq-regular module over a reduced ring $R$. Then $M$ satisfies annihilator condition, i.e, for any $m_1, m_2 \in M$, there is an $m_3 \in M$ such that*

$$ann_R(m_1) \cap ann_R(m_2) = ann_R(m_3).$$

***Proof.*** By Proposition 2.10, $ann_R(m_1) = ann_R(r_1)$ and $ann_R(m_2) = ann_R(r_2)$ for some $r_1, r_2 \in R$. Since $R$ is a reduced wq-regular ring, it is quasi regular, and so satisfies annihilator condition, i.e, $ann_R(r_1) \cap ann_R(r_2) = ann_R(r_3)$ for some $r_3 \in R$. Choose $m \in M - T(M)$. Then $ann_R(r_3) = ann_R(r_3 m)$. Put $m_3 = r_3 m$. So we have $ann_R(m_1) \cap ann_R(m_2) = ann_R(m_3)$. Thus $M$ satisfies annihilator condition. □

**Lemma 2.19.** *Let $N$ be a Baer submodule of an $R$-module $M$. If $ann_R(m) = ann_R(r)$ with $m \in N$, then $r \in (N :_R M)$.*

***Proof.*** Since $N$ is a Baer submodule, we have $ann_M(ann_R(m)) = ann_M(ann_R(r)) \subseteq N$, and so $(ann_M(ann_R(r)) :_R M) \subseteq (N :_R M)$. This yields $r \in (N :_R M)$. □

**Lemma 2.20.** *Assume that $M$ is a finitely generated module satisfying the condition (P) and $I$ is an ideal containing $ann_R(M)$. Assume that every prime submodule minimal over $IM$ is an $m$-submodule. Then $rad(IM) = rad(I)M$.*

***Proof.*** $rad(IM) = \bigcap_{N_\alpha \in Min(IM)} N_\alpha = [\bigcap (N_\alpha :_R M)M] = [\bigcap (N_\alpha :_R M)]M = \sqrt{I}M$. □

**Definition 2.21.** An $m$-submodule $N$ is said to be a *strong $m$-submodule* if all prime submodules minimal over $N$ are $m$-submodules.

Note that $M$ is a multiplication module if and only if every submodule is a strong $m$-submodule.

**Lemma 2.22.** *Assume that $M$ is a finitely generated reduced module and $N$ is a strong $m$-submodule which is also a Baer submodule. Then every prime submodule minimal over $N$ is a Baer submodule.*

***Proof.*** Let $N'$ be a minimal over $N$. Assume $ann_R(m) \subseteq ann_R(m')$ with $m \in N'$. By Proposition 2.17, $(N' :_R M)$ is a minimal over $(N :_R M)$. As $m \in N' = (N' :_R M)M$, $m = \sum_{i=1}^{n} a_i m_i$ for some $a_i \in (N' :_R M)$ and $m_i \in M$. Then there exist $b_i \notin (N' :_R M)$ and $n_i \in \mathbb{N}$ so that $a_i^{n_i} b_i \in (N :_R M)$. Since $N$ is a Baer submodule, $(N :_R M) = \sqrt{(N :_R M)}$, and so $a_i b_i \in (N :_R M)$. Put $b = b_1 b_2 .... b_n$. Then $b \notin (N' :_R M)$ and $a_i b \in (N :_R M)$, and so $a_i b m_i \in (N :_R M)M = N$, and we have $bm \in N$. Since $ann_R(bm) \subseteq ann_R(bm')$ and $N$ is a Baer submodule, $bm' \in N \subseteq N'$. As $b \notin (N' :_R M)$, we deduce $m' \in N'$, and so $N'$ is a Baer submodule. □

**Lemma 2.23.** *Let $M$ be a finitely generated reduced module satisfying the condition (P) and $N$ a strong $m$-submodule which is also a Baer submodule. Then $N$ is the intersection of prime Baer submodules.*

***Proof.*** It can be obtained from Lemma 2.20 and Lemma 2.22. □

**Lemma 2.24.** *Assume $M$ is a non-torsion reduced module and $N$ is a Baer submodule which is also a prime submodule. Then $(N :_R M)$ is a prime and Baer ideal.*

***Proof.*** We claim that $R$ is a reduced ring. Assume that $a^2 = 0$ for some $a \in R$. As $M$ is a non-torsion module, we have $ann_R(m) = 0$ for some $m \in M$. Then $a^2m = 0$ and thereby $am = 0$ since $M$ is reduced. This yields $a = 0$, and thus $R$ is a reduced ring. Let $ann_R(x) = ann_R(y)$ for some $x \in (N :_R M)$ and $y \in R$. Then $ann_R(xm) = ann_R(x) = ann_R(ym)$. Since $xm \in N$ and $N$ is a Baer submodule, we conclude that $ym \in N$. Also note that $m \notin N$. As $N$ is a prime submodule, $y \in (N :_R M)$. Then by [7, Lemma 1], $(N :_R M)$ is a Baer ideal. Since $N$ is a prime submodule, it follows that $(N :_R M)$ is a Baer and prime ideal. □

**Lemma 2.25.** *Let $M$ be a finitely generated reduced non-torsion wq-regular module. Further, assume $M$ satisfies the condition (P). Let $N$ be a strong m-submodule which is also a Baer submodule. Then $N$ is an $\alpha$-submodule.*

***Proof.*** Assume $ann_R(m_1) \cap ann_R(m_2) = ann_R(m_3)$ with $m_1, m_2 \in N$ but $m_3 \notin N$. By Lemma 2.23, there is a prime Baer submodule $N'$ with $m_3 \notin N'$. By Proposition 2.10, $ann_R(m_i) = ann_R(r_i)$ for some $r_i \in R$, $i = 1, 2, 3$. By Lemma 2.19, $r_1, r_2 \in (N' :_R M)$. Since $R$ is quasi-regular, there are $r'_1, r'_2 \in R$ so that $r_1r'_1 = 0 = r_2r'_2$ with $ann_R(r_1 + r'_1) = ann_R(r_2 + r'_2) = 0$. Since $r'_1r'_2m_3 = 0 \in N'$ and $m_3 \notin N'$, we have either $r'_1 \in (N' :_R M)$ or $r'_2 \in (N' :_R M)$. By Lemma 2.24, $(N' :_R M)$ is a Baer ideal and either $r_1 + r'_1 \in (N' :_R M)$ or $r_2 + r'_2 \in (N' :_R M)$, a contradiction. Thus $N$ is an $\alpha$-submodule. □

**Lemma 2.26.** *Let $M$ be a non-torsion wq-regular module over a reduced ring $R$. Then every $\alpha$-submodule is a $*$-submodule.*

***Proof.*** Let $N$ be an $\alpha$-submodule. Put $S = \{r \in R : ann_R(m) = ann_R(ann_R(r))$ for some $m \in N\}$. Note that by Proposition 2.10, for each $m \in M$, $ann_R(m) = ann_R(r)$ for some $r \in R$. Also by Proposition 2.7, $N$ is a Baer submodule. It can be easily seen that $S$ is a m.c.s. Let $m \in N$. Then $ann_R(m) = ann_R(a)$ for some $a \in R$. As $R$ is a wq-regular, $ann_R(a) = ann_R(ann_R(r))$ for some $r \in R$. So $ann_R(m) = ann_R(ann_R(r))$ and this implies that $rm = 0$ and $r \in S$. Then we have $m \in O(S)$, i.e, $N \subseteq O(S)$. Let $m' \in O(S)$. Then we have $r'm' = 0$ for some $r' \in S$. Also $ann_R(m) = ann_R(ann_R(r'))$ for some $m \in N$. As $R$ is wq-regular, $ann_R(m') = ann_R(a')$ for some $a' \in R$. Then $r' \in ann_R(a')$, and so $ann_R(ann_R(a')) = ann_R(ann_R(m')) \subseteq ann_R(r') = ann_R(ann_R(m))$. Since $m \in N$ and $N$ is a Baer submodule, we have $m' \in N$ and thus $N = O(S)$. Hence $N$ is a $*$-submodule. □

**Lemma 2.27.** *Let $M$ be an $R$-module. Assume that every $\alpha$-submodule is also a $*$-submodule. Then $M$ is a wq-regular.*

***Proof.*** First we prove that, $N = ann_M(ann_R(m))$ is an $\alpha$-submodule for each $m \in T(M)$. Let $ann_R(m') \cap ann_R(m'') = ann_R(m''')$ with $m', m'' \in N$. Then we have $ann_R(m) \subseteq ann_R(m')$ and $ann_R(m) \subseteq ann_R(m'')$. This implies that $ann_R(m) \subseteq ann_R(m') \cap ann_R(m'') = ann_R(m''')$ and this yields that $m''' \in ann_M(ann_R(m''')) \subseteq ann_M(ann_R(m)) = N$. Thus $N$ is an $\alpha$-submodule. The rest is similar to Theorem 2.9. □

The following Theorem 2.28 characterizes wq-regular modules in terms of $*$-submodules and $\alpha$-submodules.

**Theorem 2.28.** *Let $M$ be a non-torsion reduced module. Then $M$ is a wq-regular module if and only if every Baer submodule is a $*$-submodule if and only if every $\alpha$-submodule is a $*$-submodule.*

***Proof.*** It can be obtained from Lemma 2.26, Lemma 2.27, Proposition 2.7 and Theorem 2.9. □

**Definition 2.29.** Let $N$ be a submodule of $M$. Then $N$ is called an *annihilator submodule* if $ann_M(ann_R(N)) = N$. In particular, an annihilator ideal is an ideal $I$ of $R$ which is an annihilator submodule of the $R$-module $R$.

Note that a cyclic submodule $Rm$ is an annihilator submodule if and only if it is a Baer submodule.

**Lemma 2.30.** *Let $M$ be an $R$-module. Then,*

(i) *Every annihilator submodule is an $\alpha$-submodule.*

(ii) *Let $M$ be a non-torsion module and $N$ an annihilator submodule. Then $(N :_R M)$ is an annihilator ideal.*

***Proof.*** (i) Assume $N$ is an annihilator submodule, i.e, $N = ann_M(ann_R(N))$. Suppose $ann_R(m) \cap ann_R(m') = ann_R(m'')$ for some $m, m' \in N$. This yields $ann_R(N)m = 0 = ann_R(N)m'$, and so $ann_R(N) \subseteq ann_R(m) \cap ann_R(m')$. Then we can conclude that $ann_R(N) \subseteq ann_R(m'')$, and so $m'' \in ann_M(ann_R(m'')) \subseteq ann_M(ann_R(N)) = N$. So that $N$ is an $\alpha$-submodule.

(ii) Let $N = ann_M(ann_R(N))$. Since $M$ is non-torsion, $(N :_R M) = ann_R(ann_R(N))$. Let $r \in ann_R(N)$. Then $rN = 0$, and so $r(N :_R M)M = 0$. Choose $m \in M - T(M)$. This implies $r(N :_R M)m = 0$, and so $r(N :_R M) = 0$ and hence $r \in ann_R(N :_R M)$. Then we can conclude $ann_R(ann_R(N :_R M)) \subseteq ann_R(ann_R(N))$, and so $ann_R(ann_R(N :_R M)) \subseteq (N :_R M)$. This implies that $(N :_R M) = ann_R(ann_R(N :_R M))$. Consequently, $(N :_R M)$ is an annihilator ideal. □

The following Theorem 2.31 characterizes wq-regular modules in terms of annihilator submodules.

**Theorem 2.31.** *Let $M$ be a non-torsion module over a reduced ring $R$. Then $M$ is a wq-regular module if and only if every annihilator submodule is a $*$-submodule.*

***Proof.*** Assume $M$ is a wq-regular module. By Lemma 2.30, every annihilator submodule is an $\alpha$-submodule, and so by Lemma 2.26, every annihilator submodule is a $*$-submodule. For the converse, assume every annihilator submodule is a $*$-submodule. Let $m \in N$. Put $N = ann_M(ann_R(m))$. Then it is easily seen that $N$ is an annihilator submodule and thus a $*$-submodule. Then there is a m.c.s $S$ of $R$ so that $ann_M(ann_R(m)) = O(S)$. The rest is similar to Theorem 2.9. □

We now study quasi regular modules.

**Theorem 2.32.** *(i) Let $M$ be a non-torsion reduced wq-regular module. Assume that $q(M)$ is a multiplication module. Then $q(M)$ is a vn-regular module.*

*(ii) Assume that $q(M)$ is a finitely generated vn-regular module. Then $M$ is a reduced wq-regular module.*

***Proof.*** (i) Let $\frac{m}{t} \in q(M)$ and $S = R - Z(M)$. Put $N = R_S(\frac{m}{t})$. As $q(M)$ is a multiplication module and $N$ is a finitely generated submodule of $q(M)$, then $N = Jq(M)$ for some finitely generated ideal $J$ of $R_S$. Then there are $\frac{a_1}{s_1}, \ldots, \frac{a_n}{s_n} \in R_S$ such that $J = R_S(\frac{a_1}{s_1}) + \ldots + R_S(\frac{a_n}{s_n})$. Now, we will show that $R_S(\frac{a_1}{s_1}) = R_S(\frac{a_1}{s_1})^2$, and so $R_S(\frac{a_1}{s_1}) = R_S(\frac{e_1}{t_1})$ for some idempotent $\frac{e_1}{t_1} \in R_S$. As $M$ is non-torsion, we have $ann(m^*) = 0$ for some $m^* \in M$. Since $M$ is wq-regular, $ann_M(ann_R(a_1m^*)) = ann_M(b_1)$ and thereby $ann_R(ann_R(a_1m^*)) = ann_R(b_1)$. Note that $ann_R(a_1m^*) = ann_R(a_1)$, and so $ann_R(ann_R(a_1)) = ann_R(b_1)$. As $M$ is a reduced non-torsion module and $M$ is a wq-regular module, by Proposition 2.10 and [4, Theorem 2.1], $R$ is quasi-regular and thus $a_1 + b_1$ is a regular element and $a_1x = a_1^2$, where $x = a_1 + b_1$. Now we will show that $x \in S$. Let $m' \in M$ such that $xm' = 0$. Since

$M$ is wq-regular, $M$ satisfies the condition $ann_R(m') = ann_R(r)$ for some $r \in R$, and so $x \in ann_R(m') = ann_R(r)$ and this yields that $xr = 0$. Since $x$ is regular, $r = 0$, and so $ann_R(r) = R = ann_R(m')$ and thus $m' = 0$ and this yields $x \in S$. This implies that $R_S(\frac{a_1}{s_1})^2 = R_S(\frac{a_1^2}{s_1^2}) = R_S(\frac{a_1 x}{s_1^2}) = R_S(\frac{a_1}{s_1}\frac{x}{s_1}) = R_S(\frac{a_1}{s_1})$ since $\frac{x}{s_1}$ is a unit element of $R_S$. Thus we have $R_S(\frac{a_1}{s_1}) = R_S(\frac{e_1}{t_1})$ for some idempotent $\frac{e_1}{t_1} \in R_S$. Similarly, we get $R_S(\frac{a_i}{s_i}) = R_S(\frac{e_i}{t_i})$ for some idempotent $\frac{e_i}{t_i} \in R_S$, and so $J = R_S(\frac{e}{s})$ for some idempotent $\frac{e}{s} \in R_S$. Note that $\frac{e}{s}$ is weak idempotent $R_S$-module $q(M)$. Also $R_S(\frac{m}{t}) = Jq(M) = \frac{e}{s}q(M)$. Thus $q(M)$ is a vn-regular module.

(ii) By [8, Lemma 10], $q(M)$ is a reduced $R_S$-module, where $S = R - Z(M)$. Then it is easily seen that $M$ is reduced. Take an element $m \in M$. As $q(M)$ is a finitely generated vn-regular $R_S$-module, we deduce $R_S(\frac{m}{1}) = \frac{e}{s}q(M)$ for some weak idempotent $\frac{e}{s} \in R_S$. Note that $(1-\frac{e}{s})\frac{e}{s}q(M) = (1-\frac{e}{s})R_S(\frac{m}{1}) = 0$, and so $(1-\frac{e}{s})\frac{m}{1} = 0$ and this yields $(1-\frac{e}{s}) \in ann_{R_S}(\frac{m}{1})$ and thus we have $ann_{M_S}(ann_{R_S}(\frac{m}{1})) \subseteq ann_{M_S}(1-\frac{e}{s})$. Let $\frac{m^*}{s^*} \in ann_{M_S}(1-\frac{e}{s})$. Then we have $\frac{m^*}{s^*} = \frac{e}{s}\frac{m^*}{s^*}$. Take an element $\frac{r'}{s'} \in ann_{R_S}(\frac{m}{1})$. Then we conclude that $\frac{r'}{s'}\frac{e}{s}q(M) = 0$. Note that $\frac{m^*}{s^*} = \frac{e}{s}\frac{m^*}{s^*} \in \frac{e}{s}q(M)$, and so $\frac{r'}{s'}\frac{m^*}{s^*} = 0$ and this yields $\frac{m^*}{s^*} \in ann_{M_S}(ann_{R_S}(\frac{m}{1}))$. Then we conclude that

$$\begin{aligned} ann_{M_S}(ann_{R_S}(\frac{m}{1})) &= (ann_M(ann_R(m)))_S \\ &= ann_{M_S}(1-\frac{e}{s}) \\ &= (ann_M(s-e))_S. \end{aligned}$$

Then one can easily show that $ann_M(ann_R(m)) = ann_M(s-e)$. Accordingly, $M$ is a wq-regular module. □

Compare the following result with [4, Theorem 2.1 ].

**Corollary 2.33.** *Let $M$ be a non-torsion module in which $q(M)$ is a finitely generated multiplication module. The followings are equivalent:*

(i) *$M$ is a quasi regular module.*

(ii) *$M$ is a reduced wq-regular module.*

**Proposition 2.34.** *Assume $f : M \to M'$ is a monomorphism, where $M'$ is a wq-regular module. Then $M$ is wq-regular.*

***Proof.*** Take $m \in M$. As $M'$ is wq-regular, $ann_{M'}(ann_R(f(m)) = ann_{M'}(r)$ for some $r \in R$. Thus we have $rf(m) = f(rm) = 0$, and so $rm = 0$. This yields that $ann_M(ann_R(m)) \subseteq ann_M(r)$. Let $n \in ann_M(r)$. Then we have $rn = 0$, and so $rf(n) = f(rn) = 0$, i.e, $f(n) \in ann_{M'}(r) = ann_{M'}(ann_R(f(m))$. Thus we conclude that $ann_R(f(m))f(n) = 0$, and so $ann_R(m) \subseteq ann_R(n)$. This yields that $n \in ann_M(ann_R(n)) \subseteq ann_M(ann_R(m))$. Accordingly, we have $ann_M(ann_R(m)) = ann_M(r)$. □

**Corollary 2.35.** *Every submodule of a wq-regular module is wq-regular.*

**Proposition 2.36.** *Assume $M_i$ is an $R_i$-module for each $i \in \Delta$. Then $M = \prod_{i\in\Delta} M_i$ is a wq-regular $R = \prod_{i\in\Delta} R_i$-module if and only if $M_i$ is a wq-regular $R_i$-module for each $i \in \Delta$.*

***Proof.*** Assume that $M_i$ is a wq-regular $R_i$-module for each $i \in \Delta$. Let $(m_j)_{j\in\Delta} \in M$ and $(r_j)_{j\in\Delta} \in R$. For every $j \in \Delta$, $ann_{M_j}(ann_{R_j}(m_j)) = ann_{M_j}(r_j)$ for some $r_j \in R_j$. Also note that

$$ann_M(ann_R((m_j)_{j\in\Delta}) = \prod_{j\in\Delta} ann_{M_j}(ann_{R_j}(m_j)).$$

Thus we conclude that

$$ann_M(ann_R((m_j)_{j\in\Delta}) = \prod_{j\in\Delta} ann_{M_j}(r_j) = ann_M((r_j)_{j\in\Delta}).$$

Accordingly, $M$ is wq-regular. For the converse, assume $M$ is wq-regular. Let $m_i \in M_i$. Put the sequence

$$(n_j)_{j\in\Delta} = \begin{cases} m_i & ; \ j = i \\ 0 & ; \ j \neq i \end{cases}$$

Since $M$ is wq-regular, we have

$$\begin{aligned} ann_M(ann_R((n_j)_{j\in\Delta})) &= \prod_{j\in\Delta} ann_{M_j}(ann_{R_j}(n_j)) \\ &= ann_M((r_j)_{j\in\Delta}) \\ &= \prod_{j\in\Delta} ann_{M_j}(r_j) \end{aligned}$$

for some $(r_j)_{j\in\Delta} \in R$. This implies that $ann_{M_i}(ann_{R_i}(m_i)) = ann_{M_i}(r_i)$ for some $r_i \in R_i$ which shows that $M_i$ is a wq-regular $R_i$-module. $\square$

## 3. Trivial extension of weakly quasi regular modules

This section deals with trivial extension (idealization) of wq-regular modules. The trivial extension $R \propto M = R \oplus M$ of an $R$-module $M$ is a commutative ring with componentwise addition and multiplication $(a, m)(b, m') = (ab, am' + bm)$ for any $a, b \in R;\ m, m' \in M$ [13]. Also the nilradical of $R \propto M$ is characterized as

$$\sqrt{0_{R\propto M}} = \sqrt{0} \propto M$$

in [1] and [6]. So one can easily see that $R \propto M$ is reduced if and only if $R$ is reduced and $M = 0$ and hence $R \propto M \cong R$.

**Proposition 3.1.** *$R \propto M$ is a quasi regular ring if and only if $M = 0$ and $R$ is a quasi regular ring.*

***Proof.*** Follows from the fact that all quasi regular rings are reduced rings. $\square$

**Proposition 3.2.** (i) *Let $R \propto M$ be a wq-regular ring. Then $M$ is a wq-regular module.*

(ii) *Let $M$ be a non-torsion module in which $ann_M(I)$ is an m-submodule for all ideals $I$ of $R$. If $R \propto M$ is a wq-regular ring, then $R$ is a wq-regular ring.*

***Proof.*** (i) Take an element $m \in M$. Since $R \propto M$ is wq-regular, we can conclude $ann(ann(0, m)) = ann(r, m')$ for some $r \in R, m' \in M$. This yields $(0, m)(r, m') = (0, rm) = (0, 0)$, and so $r \in ann_R(m)$. This yields that $ann_M(ann_R(m)) \subseteq ann_M(r)$. Let $n \in ann_M(r)$. Then we have $rn = 0$ and thereby $(r, m')(0, n) = (0, 0)$, that is, $(0, n) \in ann(r, m') = ann(ann(0, m))$. Also note that $ann(0, m) = ann_R(m) \propto M$. Then we have $(0, n) \in ann(ann_R(m) \propto M)$, and so $ann_R(m)n = 0$. This gives $n \in ann_M(ann_R(m))$. Hence we have $ann_M(ann_R(m)) = ann_M(r)$, i.e, $M$ is a wq-regular module.

(ii) Let $a \in R$. Then $ann(a, 0) = \{(r, m') : (a, 0)(r, m') = (ar, am') = (0, 0)\} = ann_R(a) \propto ann_M(a)$. Then $(s, m') \in ann(ann(a, 0))$ if and only if $(s, m') \in ann(ann_R(a) \propto ann_M(a)))$ if and only if $sann_R(a) = 0$ and $sann_M(a) + ann_R(a)m' = 0$. As $M$ is non-torsion, we can conclude $(ann_M(a) : M) = ann_R(a)$, and so $sann_R(a) = 0$ implies that $s(ann_M(a) : M) = 0$. Thus by assumption, we also get $sann_M(a) = 0$. Then we get $ann_R(a)m' = 0$ and note that

$$ann(ann(a, 0)) = ann_R(ann_R(a)) \propto ann_M(ann_R(a)).$$

Since $R \propto M$ is wq-regular, we have $ann(ann(a, 0)) = ann(s, m)$ for some $s \in R, m \in M$. Thus we get $(a, 0)(s, m) = (sa, am) = (0, 0)$. This yields that $s \in ann(a)$, and so $ann_R(ann_R(a)) \subseteq ann_R(s)$. Now take $t \in ann_R(s)$. Then $st = 0$. Now choose $m^* \in M - T(M)$. Then note that $(s, m)(0, tm^*) = (0, 0)$, and so $(0, tm^*) \in ann(s, m)$. This yields that $tm^* \in ann_M(ann_R(a))$, and so $ann_R(a)tm^* = 0$. Therefore we conclude that $ann_R(a)t = 0$, and so $t \in ann_R(ann_R(a))$. Hence we get $ann_R(ann_R(a)) = ann_R(s)$, that is, $R$ is a wq-regular ring. $\square$

**Proposition 3.3.** *Let $R$ be a wq-regular ring and let $M$ be a non-torsion reduced module satisfying the condition $ann_R(m) = ann_R(r)$. Further assume that $ann_M(I)$ is an m-submodule of $M$ for each ideal $I$ of $R$. Then $R \propto M$ is a wq-regular ring.*

***Proof.*** Let $(r, m) \in R \propto M$. Then note that $(s, m') \in ann(r, m)$ implies that $sr = 0$ and $sm + rm' = 0$. So we conclude that $s(sm + rm') = s^2m = 0$. Since $M$ is reduced, we can conclude $sm = 0$, and hence $rm' = 0$. Thus we deduce

$$ann(r, m) = (ann_R(r) \cap ann_R(m)) \propto ann_M(r).$$

Since $R$ is quasi-regular, by assumption we have $ann_R(m) = ann_R(a)$ and so $ann_R(r) \cap ann_R(a) = ann_R(b)$ for some $b \in R$ by [5, Theorem 3.4]. So $ann(r, m) = ann_R(b) \propto ann_M(r)$. Then $(s, m') \in ann(ann(r, m))$ implies that $sann_R(b) = 0$ and $sann_M(r) + ann_R(b)m' = 0$. Thus we conclude that $s(sann_M(r)+ann_R(b)m') = 0$, and so $s^2ann_M(r) = 0$. Since $M$ is a reduced module, $sann_M(r) = 0$, and thus $ann_R(b)m' = 0$. So it follows that

$$ann(ann(r, m)) = (ann_R(ann_R(b)) \cap ann_R(ann_M(r))) \propto ann_M(ann_R(b)).$$

By assumption, $t \in ann_R(ann_M(r))$ if and only if $t(ann_M(r)) = t(ann_M(r) : M)M = t(ann_R(r))M = 0$ if and only if $t \in ann_R(ann_R(r))$. Since $R$ is quasi-regular, $ann_R(ann_R(b)) = ann_R(x)$ and also $ann_R(ann_R(r)) = ann_R(y)$ for some $x, y \in R$. Also note that $ann_M(ann_R(b)) = ann_M(x)$. Now choose $m^* \in M - T(M)$. Then we have $ann_R(y) = ann_R(ym^*)$, and so

$$\begin{aligned} ann(ann(r, m)) &= (ann_R(x) \cap ann_R(ym^*)) \propto ann_M(x) \\ &= ann(x, ym^*). \end{aligned}$$

Accordingly, $R \propto M$ is a wq-regular ring. □

**Theorem 3.4.** *Let $M$ be a non-torsion reduced module in which $ann_M(I)$ is an m-submodule of $M$ for all ideals $I$ of $R$. Then $R \propto M$ is a wq-regular ring if and only if $M$ is a wq-regular module.*

***Proof.*** It can be obtained from Proposition 3.3 and Proposition 3.2. □

## 4. Extension of weakly quasi regular modules

In this section, we study polynomial modules and power series modules. Let $M$ be an $R$-module and let $M[X]$ denote the set of all polynomials in indeterminate $X$ with coefficients in $R$. Then $M[X]$ becomes an $R[X]$-module. Note that if $M$ is a reduced module, then for any $m(X) = m_0 + m_1X + ... + m_nX^n \in M[X]$, where $m_i \in M$,

$$ann_{R[X]}(m(x)) = [\bigcap_{i=0}^{n} ann_R(m_i)][X].$$

**Proposition 4.1.** *Assume $M$ is a reduced non-torsion wq-regular module. Then $M[X]$ is a wq-regular $R[X]$ module.*

***Proof.*** Let $m(X) = m_0 + m_1X + ... + m_nX^n \in M[X]$. Since $M$ is reduced, we have $ann_{R[X]}(m(x)) = [\bigcap_{i=0}^{n} ann_R(m_i)][X]$. As $M$ is a non-torsion reduced module, $R$ is a reduced ring. To see this, take an element $a^2 = 0$. As $M$ is a non torsion module, there is an $m^* \in M$ with $ann_R(m^*) = 0$. Then note that $a^2m^* = 0$. As $M$ is a reduced module, we get $am^* = 0$, and thus $a = 0$. As $M$ is a non-torsion wq-regular module over a reduced ring $R$, by Lemma 2.18, $M$ satisfies annihilator condition, so that $\bigcap_{i=0}^{n} ann_R(m_i) = ann_R(m')$ for some $m' \in M$. Thus $ann_{R[X]}(m(X)) = (ann_R(m'))[X]$. Also it can be easily verified that $ann_{M[X]}(I[X]) = (ann_M(I))[X]$ for any ideal $I$ of $R$. Then we conclude that

$$\begin{aligned} ann_{M[X]}(ann_{R[X]}(m(X)) &= ann_{M[X]}((ann_R(m'))[X]) \\ &= [ann_M(ann_R(m'))][X]. \end{aligned}$$

Since $M$ is a quasi regular module, there is an $a \in R$ so that $ann_M(ann_R(m')) = ann_M(a)$, and so

$$ann_{M[X]}(ann_{R[X]}(m(X)) = (ann_M(a))[X].$$

Put $r(X) = a \in R[X]$. Then we have

$$ann_{M[X]}(ann_{R[X]}(m(X)) = ann_{M[X]}(r(X)).$$

Hence $M[X]$ is a wq-regular $R[X]$ module. □

**Proposition 4.2.** *Assume $M$ is a reduced non-torsion $R$-module in which $ann_M(I)$ is an m-submodule for each ideal $I$ of $R$. Further assume that $M$ satisfies annihilator condition and for each $m \in M$, $ann_R(m) = ann_R(r)$ for some $r \in R$. If $M[X]$ is a wq-regular $R[X]$ module, then $M$ is a wq-regular $R$-module.*

***Proof.*** Let $m \in M$. Put $m(X) = m \in M[X]$. As $M[X]$ is a wq-regular $R[X]$ module, we can conclude

$$ann_{M[X]}(ann_{R[X]}(m(X)) = [ann_M(ann_R(m))][X] = ann_{M[X]}(r(X)),$$

where $r(X) = r_0 + r_1X + ... + r_kX^k,\ r_i \in R$. Note that

$$ann_{M[X]}(r(X)) = [\bigcap_{i=0}^{k} ann_M(r_i)][X].$$

Now we will show that for any $a, b \in R$ there is $c \in R$ such that

$$ann_M(a) \cap ann_M(b) = ann_M(c).$$

Since $M$ is non-torsion, we have $ann(m^*) = 0$ for some $m^* \in M$, and so $ann_R(a) = ann_R(am^*), ann_R(b) = ann_R(bm^*)$. By annihilator condition, $ann_R(am^*) \cap ann_R(bm^*) = ann_R(m')$ for some $m' \in M$. By assumption, there is an $c \in R$ so that $ann_R(m') = ann_R(c)$. Since $M$ is non-torsion,

$$\begin{aligned} (ann_M(Ra + Rb) :_R M) &= ann_R(Ra + Rb) \\ &= ann_R(a) \cap ann_R(b) \\ &= ann_R(am^*) \cap ann_R(bm^*) \\ &= ann_R(c) = (ann_M(c) :_R M). \end{aligned}$$

This implies that

$$\begin{aligned} (ann_M(Ra + Rb) :_R M)M &= ann_M(Ra + Rb) \\ &= ann_M(a) \cap ann_M(b) \\ &= (ann_M(c) :_R M)M \\ &= ann_M(c). \end{aligned}$$

Then for $r_0, r_1, ..., r_k \in R$, $\bigcap_{i=0}^{k} ann_M(r_i) = ann_M(y)$ for some $y \in R$. This yields

$$\begin{aligned} ann_{M[X]}(ann_{R[X]}(m(X)) &= [ann_M(ann_R(m))][X] \\ &= (ann_M(y))[X]. \end{aligned}$$

Thus we have $ann_M(ann_R(m)) = ann_M(y)$. Accordingly, $M$ is a wq-regular $R$-module. □

Let $M$ be an $R$-module and let $M[[X]]$ denote the formal power series module over $R[[X]]$.

**Definition 4.3.** An $R$-module $M$ is said to *satisfy the countably annihilator condition* if for each family of $\{m_n\}_{n\in\mathbb{N}}$, then $\bigcap\limits_{i=1}^{\infty} ann_R(m_i) = ann_R(m)$ for some $m \in M$.

**Proposition 4.4.** *Assume $M$ is a reduced wq-regular module satisfying the countably annihilator condition. Then $M[[X]]$ is a wq-regular $R[[X]]$-module.*

***Proof.*** Let $f(X) = \sum\limits_{i=0}^{\infty} m_i X^i \in M[[X]]$. As $M$ is a reduced module, $ann_{R[[X]]}(f(X)) = (\bigcap\limits_{i=0}^{\infty} ann_R(m_i))[[X]]$. As $M$ satisfies the countably annihilator condition, $ann_{R[[X]]}(f(X)) = (ann_R(m))[[X]]$ for some $m \in M$. This yields

$$ann_{M[[X]]}(ann_{R[[X]]}(f(X))) = ann_{M[[X]]}((ann_R(m))[[X]]).$$

It is obvious that $ann_{M[[X]]}((ann_R(m))[[X]]) = (ann_M(ann_R(m))[[X]]$. As $M$ is wq-regular, $ann_M(ann_R(m)) = ann_M(a)$ for some $a \in R$. Thus

$$ann_{M[[X]]}(ann_{R[[X]]}(f(X))) = (ann_M(a))[[X]].$$

Now put $g(X) = a \in R[[X]]$ and note that $(ann_M(a))[[X]] = ann_{M[[X]]}(g(X))$. Accordingly, $M[[X]]$ is a wq-regular $R[[X]]$-module. □

**Proposition 4.5.** *Assume $M$ is a reduced non-torsion $R$-module in which $ann_M(I)$ is an $m$-submodule for each ideal $I$ of $R$. Further, suppose $M$ satisfies the countably annihilator condition and for each $m \in M$, $ann_R(m) = ann_R(r)$ for some $r \in R$. If $M[[X]]$ is a wq-regular $R[[X]]$-module, then $M$ is a wq-regular $R$-module.*

***Proof.*** Let $m \in M$. Put $f(X) = m \in M[[X]]$. Then $ann_{M[[X]]}(ann_{R[[X]]}(f(X))) = ann_{M[[X]]}(g(X))$ for some $g(X) = \sum\limits_{i=0}^{\infty} a_i X^i$, where $a_i \in R$. This implies that

$$(ann_M(ann_R(m)))[[X]] = (\bigcap_{i=0}^{\infty} ann_M(a_i))[[X]]$$

. As $M$ is non-torsion, we get $m^* \in M - T(M)$. Then $\bigcap\limits_{i=0}^{\infty} ann_R(a_i m^*) = ann_R(m')$ for some $m' \in M$ by the countably annihilator condition. By assumption, there is $b \in R$ so that $ann_R(m') = ann_R(b)$, and so

$$\begin{aligned}(ann_M(\sum_{i=0}^{\infty} Ra_i) :_R M) &= ann_R(\sum_{i=0}^{\infty} Ra_i)\\ &= ann_R(\sum_{i=0}^{\infty} Ra_i m^*) = ann_R(m')\\ &= ann_R(b) = (ann_M(b) :_R M).\end{aligned}$$

Then

$$\begin{aligned}(ann_M(\sum_{i=0}^{\infty} Ra_i) :_R M)M &= \bigcap_{i=0}^{\infty} ann_M(a_i)\\ &= (ann_M(b) :_R M)M = ann_M(b).\end{aligned}$$

This implies that

$$\begin{aligned}ann_{M[[X]]}(ann_{R[[X]]}(f(X))) &= (ann_M(ann_R(m)))[[X]]\\ &= (ann_M(b))[[X]],\end{aligned}$$

and so $ann_M(ann_R(m)) = ann_M(b)$. This gives that $M$ is a wq-regular $R$-module. □

**Theorem 4.6.** *Let $M$ be a reduced non-torsion module in which $ann_M(I)$ is an $m$-submodule for each ideal $I$ of $R$. Assume $M$ satisfies the countably annihilator condition and for each $m \in M$, $ann_R(m) = ann_R(r)$ for some $r \in R$. Then the following are equivalent:*

(i) *$M$ is a wq-regular $R$-module.*
(ii) *$M[X]$ is a wq-regular $R[X]$-module.*
(iii) *$M[[X]]$ is a wq-regular $R[[X]]$-module.*

***Proof.*** $(i) \Leftrightarrow (ii)$ It can be obtained from Proposition 4.1 and Proposition 4.2.

$(i) \Leftrightarrow (iii)$ It can be obtained from Proposition 4.4 and Proposition 4.5. □

**Acknowledgment.** We would like to thank the referee for his/her great effort in proofreading the manuscript.

## References

[1] D.D. Anderson and M. Winders, *Idealization of a module*, J. Commut. Algebra, **1** (1), 3–56, 2009.

[2] D.F. Anderson, R. Levy and J. Shapiro, *Zero-divisor graphs, von Neumann regular rings, and Boolean algebras*, J. Pure Appl. Algebra, **180** (3), 221–241, 2003.

[3] Z.A. El-Bast and P.F. Smith, *Multiplication modules*, Comm. Algebra, **16** (4), 755–779, 1988.

[4] M. Evans, *On commutative P.P. rings*, Pac. J. Math. **41** (3), 687–697, 1972

[5] M. Henriksen and M. Jerison, *The space of minimal prime ideals of a commutative ring*, Trans. Amer. Math. Soc. **115**, 110–130, 1965.

[6] J.A. Huckaba, *Commutative rings with zero divisors*, Marcel Dekker, New York, 1988.

[7] C. Jayaram, *Baer ideals in commutative semiprime rings*, Indian J. Pure Appl. Math. **15** (8) 855–864, 1984.

[8] C. Jayaram and Ü. Tekir, *von Neumann regular modules*, Comm. Algebra, **46** (5), 2205–2217, 2018.

[9] T.K. Lee and Y. Zhou, *Reduced modules, Rings, modules, algebras and abelian groups*, in:Lect. Notes Pure Appl. Math. New York, NY: Marcel Dekker, **236**, 365–377, 2004.

[10] R. Levy and J. Shapiro, *The zero-divisor graph of von Neumann regular rings*, Comm. Algebra, **30** (2), 745–750, 2002.

[11] C.P. Lu, *Prime submodules of modules*, Comment. Math. Univ. St. Pauli, **33** (1), 61–69, 1984.

[12] R.L. McCasland and M.E. Moore, *Prime submodules*, Comm. Algebra, **20** (6), 1803–1817, 1992.

[13] M. Nagata, *Local rings*, Interscience Publishers, New York, 1960.

[14] J. Von Neumann, *On regular rings*, Proc. Natl. Acad. Sci. **22** (12), 707–713, 1936.

Hacettepe Journal of Mathematics & Statistics

Hacet. J. Math. Stat.
Volume 50 (1) (2021), 135 – 143
DOI : 10.15672/hujms.659265

Research Article

# Connections on the rational Korselt set of $pq$

Nejib Ghanmi

*Preparatory Institute of Engineering Studies of Tunis, Tunis university, Tunisia*

## Abstract

For a positive integer $N$ and $\mathbb{A}$, a subset of $\mathbb{Q}$, let $\mathbb{A}$-$\mathcal{KS}(N)$ denote the set of $\alpha = \dfrac{\alpha_1}{\alpha_2} \in \mathbb{A}\backslash\{0, N\}$, where $\alpha_2 r - \alpha_1$ divides $\alpha_2 N - \alpha_1$ for every prime divisor $r$ of $N$. The set $\mathbb{A}$-$\mathcal{KS}(N)$ is called the set of $N$-Korselt bases in $\mathbb{A}$. Let $p, q$ be two distinct prime numbers. In this paper, we prove that each $pq$-Korselt base in $\mathbb{Z}\backslash\{q + p - 1\}$ generates at least one other in $\mathbb{Q}$-$\mathcal{KS}(pq)$. More precisely, we prove that if $(\mathbb{Q}\backslash\mathbb{Z})$-$\mathcal{KS}(pq) = \emptyset$, then $\mathbb{Z}$-$\mathcal{KS}(pq) = \{q + p - 1\}$.

**Mathematics Subject Classification (2020).** 11Y16, 11Y11, 11A51

**Keywords.** prime number, Carmichael number, squarefree composite number, Korselt base, Korselt number, Korselt set

## 1. Introduction

A Carmichael number [2] $N$ is a positive composite integer that satisfies $a^N \equiv 1 \pmod N$ for any $a$ with $\gcd(a, N) = 1$, it follows that a Carmichael number $N$ meets Korselt's criterion:

**Korselt's criterion 1.1** ([10]). A squarefree composite integer $N > 1$ is a Carmichael number if and only if $p - 1$ divides $N - 1$ for all prime factors $p$ of $N$.

In [1, 3], Bouallègue-Echi-Pinch introduced the notion of an $\alpha$-Korselt number, where $\alpha \in \mathbb{Z}\backslash\{0\}$, as a generalized Carmichael number when $\alpha = 1$ as follows:

**Definition 1.2.** An $\alpha$-Korselt number is a number $N$ such that $p - \alpha$ divides $N - \alpha$ for all prime divisors $p$ of $N$.

The $\alpha$-Korselt numbers for $\alpha \in \mathbb{Z}$ have been thoroughly investigated in recent years, especially in [1, 3, 4, 8, 9]. In [5], Ghanmi proposed another generalization for $\alpha = \dfrac{\alpha_1}{\alpha_2} \in \mathbb{Q}\backslash\{0\}$ by setting the following definitions:

**Definition 1.3.** Let $N \in \mathbb{N}\backslash\{0, 1\}$, $\alpha = \dfrac{\alpha_1}{\alpha_2} \in \mathbb{Q}\backslash\{0\}$ with $\gcd(\alpha_1, \alpha_2) = 1$ and $\mathbb{A}$ a subset of $\mathbb{Q}$. Then,

(1) $N$ is said to be an *$\alpha$-Korselt number* (*$K_\alpha$-number*) if $N \neq \alpha$ and $\alpha_2 p - \alpha_1$ divides $\alpha_2 N - \alpha_1$ for every prime divisor $p$ of $N$.

Email address: nejib.ghanmi@ipein.rnu.tn, neghanmi@yahoo.fr
Received: 13.12.2019; Accepted: 11.05.2020

(2) By the $\mathbb{A}$-*Korselt set* of a number $N$ (or the Korselt set of $N$ over $\mathbb{A}$), we mean the set $\mathbb{A}$-$\mathcal{KS}(N)$ of all $\beta \in \mathbb{A}\backslash\{0, N\}$ such that $N$ is a $K_\beta$-number.
(3) If $\mathbb{A}$-$\mathcal{KS}(N)$ has a finite number of elements, then its cardinality is the $\mathbb{A}$-*Korselt weight* of $N$. Otherwise, if the cardinality is infinite, we say that $N$ has an infinite weight over $\mathbb{A}$. The $\mathbb{A}$-*Korselt weight* of $N$ is simply denoted by $\mathbb{A}$-$\mathcal{KW}(N)$.

Carmichael numbers are exactly the 1-Korselt squarefree composite numbers. Furthermore, in [6, 7], Ghanmi defined the notion of Korselt bases as follows:

**Definition 1.4.** Let $N \in \mathbb{N}\backslash\{0, 1\}$, $\alpha \in \mathbb{Q}\backslash\{0\}$ and $\mathbb{B}$ be a subset of $\mathbb{N}$. Then,

(1) $\alpha$ is called an $N$-*Korselt base* ($K_N$-*base*) if $N$ is a $K_\alpha$-*number.*
(2) By the $\mathbb{B}$-*Korselt set* of base $\alpha$ (or the Korselt set of base $\alpha$ over $\mathbb{B}$), we mean the set $\mathbb{B}$-$\mathcal{KS}(B(\alpha))$ of all $M \in \mathbb{B}$ such that $\alpha$ is a $K_M$-base.
(3) If $\mathbb{B}$-$\mathcal{KS}(B(\alpha))$ has a finite number of elements, then its cardinality is called the $\mathbb{B}$-*Korselt weight* of base $\alpha$. Otherwise, if the cardinality is infinite, we say that $\alpha$ has an infinite weight over $\mathbb{B}$. The $\mathbb{B}$-*Korselt weight* of base $\alpha$ is denoted by $\mathbb{B}$-$\mathcal{KW}(B(\alpha))$.

The set $\mathbb{Q}$-$\mathcal{KS}(N)$ is simply called the rational Korselt set of $N$. In this paper, we are concerned only with a squarefree composite number $N$.

After extending the notion of a Korselt number to $\mathbb{Q}$, and in order to study the Korselt numbers and their Korselt sets over $\mathbb{Q}$, it is natural to ask about the existence of connections between the Korselt bases of a number $N$ over the sets $\mathbb{Z}$ and $\mathbb{Q}\backslash\mathbb{Z}$. The answer is affirmative for a squarefree composite number $N$ with two prime factors. Indeed, when we look deeply at a list of Korselt numbers and their Korselt sets (see Table 1 and Table 2), we note the absence of any squarefree composite number $N$ with two prime factors such that $\mathbb{Z}$-$\mathcal{KW}(N) \geq 2$ and $(\mathbb{Q}\backslash\mathbb{Z})$-$\mathcal{KS}(N) = \emptyset$. This finding inspired us to claim that such a relation between $\mathbb{Z}$-$\mathcal{KS}(N)$ and $(\mathbb{Q}\backslash\mathbb{Z})$-$\mathcal{KS}(N)$ exists. The case when $N$ is squarefree and has more than two prime factors remains untreated. To explain this (these) connection(s), we organize our work as follows. In Section 2, we give some numerical data showing connections between the Korselt bases of $N$ over $\mathbb{Z}$ and $(\mathbb{Q}\backslash\mathbb{Z})$. In Section 3, we prove that for each squarefree composite number $N$ with two prime factors, some $N$-Korselt bases in $\mathbb{Z}$ generate others in the same set $\mathbb{Z}$-$\mathcal{KS}(N)$. Finally, in Section 4, we show that for each squarefree composite number $N = pq$ with two prime factors, each $N$-Korselt base in $\mathbb{Z}\backslash\{q + p - 1\}$ generates a Korselt base in $\mathbb{Q}\backslash\mathbb{Z}$.

## 2. Preliminaries

The following data illustrate some cases of Korselt numbers and their Korselt sets. Table 1 provides all $N = pq$ and $\mathbb{Z}$-$\mathcal{KS}(N)$ with $p, q$ primes and $p < q \leq 53$ for which $(\mathbb{Q}\backslash\mathbb{Z})$-$\mathcal{KS}(N) = \emptyset$. Table 2 lists, for each integer $1 \leq i \leq 7$, the smallest squarefree composite number $N_i = pq$ with $p, q$ primes, $p < q < 10^3$ such that $\mathbb{Z}$-$\mathcal{KW}(N_i) = i$ and $(\mathbb{Q}\backslash\mathbb{Z})$-$\mathcal{KW}(N_i)$ is the smallest.

| $N$ | $\mathbb{Z}$-$\mathcal{KS}(N)$ | $N$ | $\mathbb{Z}$-$\mathcal{KS}(N)$ | $N$ | $\mathbb{Z}$-$\mathcal{KS}(N)$ |
|---|---|---|---|---|---|
| $2 \times 11$ | $\{12\}$ | $2 \times 31$ | $\{32\}$ | $5 \times 43$ | $\{47\}$ |
| $2 \times 13$ | $\{14\}$ | $3 \times 31$ | $\{33\}$ | $2 \times 47$ | $\{48\}$ |
| $2 \times 17$ | $\{18\}$ | $2 \times 37$ | $\{38\}$ | $3 \times 47$ | $\{49\}$ |
| $2 \times 19$ | $\{20\}$ | $3 \times 37$ | $\{39\}$ | $5 \times 47$ | $\{51\}$ |
| $3 \times 19$ | $\{21\}$ | $2 \times 41$ | $\{42\}$ | $13 \times 47$ | $\{59\}$ |
| $2 \times 23$ | $\{24\}$ | $3 \times 41$ | $\{43\}$ | $2 \times 53$ | $\{54\}$ |
| $3 \times 23$ | $\{25\}$ | $5 \times 41$ | $\{45\}$ | $3 \times 53$ | $\{55\}$ |
| $2 \times 29$ | $\{30\}$ | $2 \times 43$ | $\{44\}$ | $5 \times 53$ | $\{57\}$ |
| $3 \times 29$ | $\{31\}$ | $3 \times 43$ | $\{45\}$ | | |

**Table 2.** $\mathbb{Z}$-$\mathcal{KS}(N)$ where $N = pq; p, q$ primes , $p < q \leq 53$ and $(\mathbb{Q}\backslash\mathbb{Z})$-$\mathcal{KS}(N) = \emptyset$.

| $i$ | $N_i$ | $\mathbb{Z}$-$\mathcal{KS}(N_i)$ | $(\mathbb{Q}\backslash\mathbb{Z})$-$\mathcal{KW}(N_i)$ |
|---|---|---|---|
| 1 | $2\times 11$ | $\{12\}$ | 0 |
| 2 | $2\times 7$ | $\{6,8\}$ | 1 |
| 3 | $5\times 19$ | $\{15,20,23\}$ | 2 |
| 4 | $31\times 59$ | $\{29,60,62,89\}$ | 5 |
| 5 | $67\times 97$ | $\{64,75,91,99,163\}$ | 12 |
| 6 | $757\times 881$ | $\{755,773,797,845,867,1637\}$ | 17 |
| 7 | $37\times 61$ | $\{25,43,49,52,57,67,97\}$ | 22 |

**Table 2.** The smallest $N_i = pq$ with $p,q$ primes, $p<q<10^3$ such that $\mathbb{Z}$-$\mathcal{KW}(N_i)=i$ and $(\mathbb{Q}\backslash\mathbb{Z})$-$\mathcal{KW}(N_i)$ is the smallest.

Based on Table 1 and Table 2, we remark that there is no squarefree composite number $N$ with two prime factors such that $\mathbb{Z}$-$\mathcal{KW}(N)\geq 2$ and $(\mathbb{Q}\backslash\mathbb{Z})$-$\mathcal{KS}(N)=\emptyset$. This leads to the following result:

**Theorem 2.1 (Main Theorem).** *Let $N=pq$. If $(\mathbb{Q}\backslash\mathbb{Z})$-$\mathcal{KS}(N)=\emptyset$, then $\mathbb{Z}$-$\mathcal{KS}(N)=\{q+p-1\}$.*

Moreover, it appears that for numbers $N$ that satisfy Theorem 2.1, the sets $\mathbb{Z}$-$\mathcal{KS}(N)$ and $(\mathbb{Q}\backslash\mathbb{Z})$-$\mathcal{KS}(N)$ are somewhat related. To highlight this relation, we show that each $N$-Korselt base in $\mathbb{Z}\backslash\{p+q-1\}$ induces at least one other $N$-Korselt base in $(\mathbb{Q}\backslash\mathbb{Z})$-$\mathcal{KS}(N)$. Hence, the main theorem is deduced immediately.

For the rest of this paper, let $p<q$ be two primes and let $N=pq$ and $i,s$ be the integers given by the Euclidian division of $q$ by $p$: $q=ip+s$ with $s\in\{1,\ldots,p-1\}$.

Our work is based on the following result given by Echi-Ghanmi [4].

**Theorem 2.2.** [4, Theorem 14] *Let $N=pq$ such that $p<q$. Then, the following properties hold:*

(1) *If $q>2p^2$, then $\mathbb{Z}$-$\mathcal{KS}(N)=\{p+q-1\}$.*
(2) *If $p^2-p<q<2p^2$ and $p\geq 5$, then*
$$\mathbb{Z}\text{-}\mathcal{KS}(N)\subseteq\{ip,p+q-1\}.$$
(3) *If $4p<q<p^2-p$, then*
$$\mathbb{Z}\text{-}\mathcal{KS}(N)\subseteq\{ip,(i+1)p,p+q-1\}.$$
(4) *Suppose that $3p<q<4p$. Then, the following conditions are satisfied:*
 (*a*) *If $q=4p-3$, then the following properties hold:*
  (*i*) *If $p\equiv 1 \pmod 3$, then*
$$\mathbb{Z}\text{-}\mathcal{KS}(N)=\{4p,q-p+1,p+q-1\}.$$
  (*ii*) *If $p\not\equiv 1 \pmod 3$, then*
$$\mathbb{Z}\text{-}\mathcal{KS}(N)=\{q-p+1,p+q-1\}.$$
 (*b*) *If $q\neq 4p-3$, then*
$$\mathbb{Z}\text{-}\mathcal{KS}(N)\subseteq\{3p,4p,p+q-1\}.$$
(5) *If $2p<q<3p$, then*
$$\mathbb{Z}\text{-}\mathcal{KS}(N)\subseteq\{2p,3p,3q-5p+3,\frac{2p+q-1}{2},q-p+1,p+q-1\}.$$
(6) *If $p<q<2p$, then*
$$\mathbb{Z}\text{-}\mathcal{KS}(N)\subseteq\{q+p-1\}\cup[2,2p]\setminus\{p\}.$$

Next, we establish the following two results to serve us for the rest of the paper:

**Lemma 2.3.** *For each $N = pq$ with $p < q$ and both being prime, the set $\mathbb{Z}$-$\mathcal{KS}(N)$ is characterized by Theorem 2.2, except for $(p,q) \in \{(3,13),(3,17)\}$, where $\mathbb{Z}$-$\mathcal{KS}(3\times 13) = \{12,15\}$ and $\mathbb{Z}$-$\mathcal{KS}(3\times 17) = \{15,19\}$.*

***Proof.*** Let $N = pq$ with $p < q$ both being prime.

- If $p \geq 5$, then $\mathbb{Z}$-$\mathcal{KS}(N)$ is simply given by one of the six cases of Theorem 2.2.

- Suppose that $p = 2$. If $q < 8 = 4p$ (resp. $q > 8 = 2p^2$), then $\mathbb{Z}$-$\mathcal{KS}(N)$ is completely determined by one of states 4, 5, and 6 (resp. state 1) of Theorem 2.2.

- Similarly, for the case $p = 3$, if $q < 4p = 12$ (resp. $q > 2p^2 = 18$), then $\mathbb{Z}$-$\mathcal{KS}(N)$ is determined by one of cases 4, 5, and 6 (resp. case 1) of Theorem 2.2. Therefore, the remaining values for the prime number $q$ are 13 and 17, where $\mathbb{Z}$-$\mathcal{KS}(3\times 13) = \{12,15\}$ and $\mathbb{Z}$-$\mathcal{KS}(3\times 17) = \{15,19\}$ (see [4, Proposition 15]). $\square$

**Proposition 2.4.** [9, Corollary 3.6] *Let $p$ and $q$ be two prime numbers such that $p < q$ and $N = pq$. If $\alpha \in \mathbb{Z}$-$\mathcal{KS}(N)$, then the following statements hold:*

(1) $\gcd(\alpha, q) = 1$.
(2) $2 \leq q - p + 1 \leq \alpha \leq p + q - 1$.
(3) *If $p$ divides $\alpha$, then $\alpha \in \{ip, (i+1)p\}$.*

## 3. Connections in $\mathbb{Z}$-$\mathcal{KS}(N)$

In the following result, we prove that certain $N$-Korselt bases in $\mathbb{Z}$ induce others in the same set $\mathbb{Z}$-$\mathcal{KS}(N)$.

**Proposition 3.1.** *Suppose that $2p < q < 3p$. Then, the following statements hold:*

(1) $\dfrac{2p+q-1}{2} \in \mathbb{Z}$-$\mathcal{KS}(N)$ *if and only if* $q - p + 1 \in \mathbb{Z}$-$\mathcal{KS}(N)$.
(2) *If* $3q - 5p + 3 \in \mathbb{Z}$-$\mathcal{KS}(N)$, *then* $q - p + 1 \in \mathbb{Z}$-$\mathcal{KS}(N)$.

***Proof.*** First, since $q = 2p + s$, the integer $s$ must be odd, and therefore, $s < p - 1$.

(1) We have $\alpha = \dfrac{2p+q-1}{2} \in \mathbb{Z}$-$\mathcal{KS}(N)$ if and only if

$$\begin{cases} p - \alpha = \dfrac{-q+1}{2} & \mid \quad p(q-1) \\ q - \alpha = \dfrac{s+1}{2} & \mid \quad q(p-1) \end{cases}$$

which is equivalent to $s+1$ divides $2q(p-1)$.

However, we have $\gcd(q, s+1) = 1$ (as $s < p-1 < q-1$) and $2(p-1) = q - 1 - (s+1)$. Therefore, we conclude that

$$\frac{2p+q-1}{2} \in \mathbb{Z}\text{-}\mathcal{KS}(N) \quad \text{if and only if} \quad s+1 \mid q-1. \tag{3.1}$$

Similarly, $\beta = q - p + 1 \in \mathbb{Z}$-$\mathcal{KS}(N)$ is equivalent to

$$\begin{cases} p - \beta = -s - 1 & \mid \quad p(q-1) \\ q - \beta = p - 1 & \mid \quad q(p-1) \end{cases}$$

which is equivalent to $s+1$ divides $p(q-1)$.

However, we know that $\gcd(p, s+1) = 1$ since $s < p-1$, which shows that

$$q - p + 1 \in \mathbb{Z}\text{-}\mathcal{KS}(N) \quad \text{if and only if} \quad s+1 \mid q-1. \tag{3.2}$$

Therefore, by (3.1) and (3.2), we conclude that

$$\frac{2p+q-1}{2} \in \mathbb{Z}\text{-}\mathcal{KS}(N) \quad \text{if and only if} \quad q - p + 1 \in \mathbb{Z}\text{-}\mathcal{KS}(N).$$

(2) Suppose that $\gamma = 3q - 5p + 3 \in \mathbb{Z}\text{-}\mathcal{KS}(N)$. Then,

$$p - \gamma = 6p - 3q - 3 = -3(s+1) \mid p(q-1). \tag{3.3}$$

We consider two cases:

• If $p \neq 3$, then since $s < p - 1$, we have $\gcd(p, 3(s+1)) = 1$. Hence, by (3.3), $3(s+1)$ divides $q - 1$. Thus, by (3.2), $q - p + 1 \in \mathbb{Z}\text{-}\mathcal{KS}(N)$.

• Now, assume that $p = 3$. First, because $1 \leq s \leq p - 2 = 1$, we know that $s = 1$, $q = 2p + s = 7$ and $q - p + 1 = 5$. Therefore, we can easily check that $N = 3 \times 7 = 21$ is a 5-Korselt number. □

**Corollary 3.2.** *If $q > 2p$ and $q - p + 1 \notin \mathbb{Z}\text{-}\mathcal{KS}(N)$, then*

$$\mathbb{Z}\text{-}\mathcal{KS}(N) \subseteq \{ip, (i+1)p, p+q-1\}.$$

***Proof.*** By Theorem 2.2 and Lemma 2.3, the solution is straightforward when $q > 3p$.

Now, suppose that $2p < q < 3p$ (i.e., $i = 2$). Let $\beta \in \mathbb{Z}\text{-}\mathcal{KS}(N)$. Then, again by Theorem 2.2, we obtain

$$\beta \in \{2p, 3p, 3q - 5p + 3, \frac{2p+q-1}{2}, q - p + 1, p + q - 1\}.$$

However, since $q - p + 1 \notin \mathbb{Z}\text{-}\mathcal{KS}(N)$, using Proposition 3.1, we obtain $\beta \neq 3q - 5p + 3, \frac{2p+q-1}{2}$. Thus, $\beta \in \{2p, 3p, p+q-1\}$, as desired. □

## 4. Connections between $\mathbb{Z}\text{-}\mathcal{KS}(N)$ and $(\mathbb{Q}\backslash\mathbb{Z})\text{-}\mathcal{KS}(N)$

The following result concerns the case when $q < 2p$.

**Proposition 4.1.** *Suppose that $q < 2p$ and $\beta \in \mathbb{Z}\backslash\{0\}$ with $\beta \neq p+q-1$ and $\gcd(p,\beta) = \gcd(pq, p+q-\beta) = 1$. Then, $\beta \in \mathbb{Z}\text{-}\mathcal{KS}(N)$ if and only if $\frac{qp}{p+q-\beta} \in (\mathbb{Q}\backslash\mathbb{Z})\text{-}\mathcal{KS}(N)$.*

***Proof.*** Since $\gcd(p,\beta) = \gcd(pq, p+q-\beta) = 1$, we have

$$\begin{aligned}
\beta \in \mathbb{Z}\text{-}\mathcal{KS}(N) \quad &\Leftrightarrow \quad \begin{cases} p - \beta & \mid \quad q - 1 \\ q - \beta & \mid \quad p - 1 \end{cases} \\
&\Leftrightarrow \quad \begin{cases} (p+q-\beta)p - pq = (p-\beta)p & \mid \quad p(q-1) \\ (p+q-\beta)q - pq = (q-\beta)q & \mid \quad q(p-1) \end{cases} \\
&\Leftrightarrow \quad \frac{qp}{p+q-\beta} \in \mathbb{Q}\text{-}\mathcal{KS}(N).
\end{aligned}$$

Because $\beta \notin \{p, q\}$, $p+q-\beta \notin \{p, q\}$. Moreover, if $\beta \in \mathbb{Z}\text{-}\mathcal{KS}(N)$, then since $p < q < 2p$, we have $2 \leq \beta < 2p$ by Theorem 2.2; hence, $p+q-\beta \geq 2p-\beta+1 \geq 2$, that is, $p+q-\beta \neq 1$. Therefore, $\frac{qp}{p+q-\beta} \notin \mathbb{Z}$, and we conclude that $\frac{qp}{p+q-\beta} \in (\mathbb{Q}\backslash\mathbb{Z})\text{-}\mathcal{KS}(N)$. □

The next two results concern the case when $p$ divides $\beta$.

**Proposition 4.2.** *If $ip \in \mathbb{Z}\text{-}\mathcal{KS}(N)$, then there exists $k_1 \in \mathbb{N}\backslash\{0,1\}$ such that $\frac{(k_1+1)q}{ik_1+1} \in (\mathbb{Q}\backslash\mathbb{Z})\text{-}\mathcal{KS}(N)$ .*

***Proof.*** Let $ip \in \mathbb{Z}\text{-}\mathcal{KS}(N)$. Then,

$$\begin{cases} p - ip & \mid \quad pq - ip = p(q-1) + p - ip \\ q - ip & \mid \quad pq - ip = q(p-1) + q - ip. \end{cases}$$

As $\gcd(s,q)=1$, this is equivalent to

$$\begin{cases} i-1 & \mid & q-1 \\ s & \mid & p-1 \end{cases}$$

and hence, there exist $k_1$ and $k_2$ in $\mathbb{Z}$ such that

$$\begin{cases} q-1 & = & k_2(i-1) \\ p-1 & = & k_1 s. \end{cases}$$

As $q=ip+s$, $k_1q=ik_1p+k_1s=ik_1p+p-1$, and therefore,

$$(k_1+1)q-(ik_1+1)p=q-1. \tag{4.1}$$

Let $k=\gcd(k_1+1,ik_1+1)$, $\alpha_1^{'}=\dfrac{k_1+1}{k}$ and $\alpha_2=\dfrac{ik_1+1}{k}$. Therefore, using (4.1), we obtain

$$\alpha_1^{'}q-\alpha_2p=\frac{q-1}{k}. \tag{4.2}$$

Now, let us prove that $\alpha_2-\alpha_1^{'}$ divides $p-1$. First, note that

$$\alpha_2-\alpha_1^{'}=\frac{k_1}{k}(i-1). \tag{4.3}$$

Since $q-1=(i-1)p+(k_1+1)s$ and $i-1\mid q-1$, we deduce that $i-1\mid(k_1+1)s$. Furthermore, because $\gcd(k_1+1,i-1)=\gcd(k_1+1,ik_1+1)=k$, it follows that $m=\dfrac{i-1}{k}\mid\dfrac{k_1+1}{k}s$. However, $\gcd\left(\dfrac{k_1+1}{k},\dfrac{i-1}{k}\right)=1$; hence, $m\mid s$. Therefore, we conclude by (4.3) that

$$\alpha_2-\alpha_1^{'}=k_1m\mid k_1s=p-1. \tag{4.4}$$

Now, by (4.2) and (4.4), we obtain

$$\begin{cases} \alpha_2p-\alpha_1^{'}q & \mid & q-1 \\ \alpha_2-\alpha_1^{'} & \mid & p-1. \end{cases}$$

Thus,

$$\alpha=\frac{\alpha_1^{'}q}{\alpha_2}=\frac{(k_1+1)q}{ik_1+1}\in\mathbb{Q}\text{-}\mathcal{KS}(N).$$

As $\gcd(\alpha_1^{'},\alpha_2)=1$, $\gcd(q,\alpha_2)=1$ by (4.2) and $\alpha_2\neq1$, we conclude that $\dfrac{(k_1+1)q}{ik_1+1}\in(\mathbb{Q}\backslash\mathbb{Z})\text{-}\mathcal{KS}(N)$. □

In the following result, we need $(i+1)p\neq q+p-1$ (i.e., $s>1$) to show that $(i+1)p$ generates an element in $\mathbb{Q}\backslash\mathbb{Z})\text{-}\mathcal{KS}(N)$.

**Proposition 4.3.** *If $(i+1)p\in\mathbb{Z}\text{-}\mathcal{KS}(N)$ and $s>1$, then there exists $k_1\in\mathbb{N}\backslash\{0,1\}$ such that $\dfrac{(k_1-1)q}{(i+1)k_1-1}\in(\mathbb{Q}\backslash\mathbb{Z})\text{-}\mathcal{KS}(N)$.*

***Proof.*** If $(i+1)p\in\mathbb{Z}\text{-}\mathcal{KS}(N)$, then

$$\begin{cases} p-(i+1)p & \mid & pq-(i+1)p=p(q-1)+p-(i+1)p \\ q-(i+1)p & \mid & pq-(i+1)p=q(p-1)+q-(i+1)p. \end{cases}$$

This is equivalent to

$$\begin{cases} i & \mid & q-1 \\ p-s & \mid & p-1 \end{cases}$$

and hence, there exist $k_1$ and $k_2$ in $\mathbb{N}\backslash\{0\}$ such that

$$\begin{cases} q-1 & = & k_2i \\ p-1 & = & k_1(p-s). \end{cases}$$

First, as $s > 1$, it follows that $k_1 > 1$. Since $q = (i+1)p+s-p$, $k_1q = (i+1)k_1p-p+1$. Therefore, we can write

$$((i+1)k_1 - 1)p - (k_1 - 1)q = q - 1. \tag{4.5}$$

Let $k = \gcd(k_1 - 1, (i+1)k_1 - 1)$, $\alpha_1^{'} = \dfrac{k_1 - 1}{k}$ and $\alpha_2 = \dfrac{(i+1)k_1 - 1}{k}$.

Then, we use (4.5) to obtain

$$\alpha_2 p - \alpha_1^{'} q = \frac{q-1}{k}. \tag{4.6}$$

Next, let us prove that $\alpha_2 - \alpha_1^{'} \mid p - 1$. First, we have

$$\alpha_2 - \alpha_1^{'} = \frac{ik_1}{k}. \tag{4.7}$$

Since $i \mid q - 1 = ip + s - 1$, we know that $i \mid s - 1 = (k_1 - 1)(p - s)$. Moreover, as $\gcd(k_1-1, i) = \gcd(k_1-1, (i+1)k_1-1) = k$, it follows that $m = \dfrac{i}{k} \mid \dfrac{k_1 - 1}{k}(p-s)$. Hence, $m \mid p - s$ since $\gcd\left(\dfrac{k_1 - 1}{k}, \dfrac{i}{k}\right) = 1$. Therefore, we deduce by (4.7) that

$$\alpha_2 - \alpha_1^{'} = k_1 m \mid k_1(p - s) = p - 1. \tag{4.8}$$

Now, by (4.6) and (4.8), we obtain

$$\begin{cases} \alpha_2 p - \alpha_1^{'} q & \mid \quad q - 1 \\ \quad \alpha_2 - \alpha_1^{'} & \mid \quad p - 1. \end{cases}$$

Therefore,

$$\alpha = \frac{\alpha_1^{'} q}{\alpha_2} = \frac{(k_1 - 1)q}{(i+1)k_1 - 1} \in \mathbb{Q}\text{-}\mathcal{KS}(N).$$

As $\gcd(\alpha_1^{'}, \alpha_2) = 1$, $\gcd(q, \alpha_2) = 1$ by (4.6) and $\alpha_2 \neq 1$, we deduce that $\dfrac{(k_1 - 1)q}{(i+1)k_1 - 1} \in (\mathbb{Q}\backslash\mathbb{Z})\text{-}\mathcal{KS}(N)$. □

Now, it remains to prove that each $N$-Korselt base $\beta \in \mathbb{Z}$ generates an $N$-Korselt base in $(\mathbb{Q}\backslash\mathbb{Z})$, where $\gcd(\beta, p) = 1$, $2p < q < 4p$ and $\beta \neq q+p-1$. This is equivalent to discuss only the cases when $\beta \in \{3q - 5p + 3, \dfrac{2p + q - 1}{2}, q - p + 1\}$. It follows by Corollary 3.2 that we can restrain our work only for $\beta = q - p + 1$ with $\gcd(q + 1, p) = \gcd(\beta, p) = 1$.

**Proposition 4.4.** *Suppose that* $2p < q < 4p$ *with* $\gcd(q+1, p) = 1$. *If* $q-p+1 \in \mathbb{Z}\text{-}\mathcal{KS}(N)$, *then* $\dfrac{pq}{2p - 1} \in (\mathbb{Q}\backslash\mathbb{Z})\text{-}\mathcal{KS}(N)$.

***Proof.*** First, if $i = 3$, then by Theorem 2.2, we must have $q = 4p - 3$, and it is easy to verify that $\dfrac{pq}{2p-1}$ is an $N$-Korselt base. Furthermore, since $\gcd(pq, 2p - 1) = 1$ and $2p-1 \neq 1$, we know that $\dfrac{pq}{2p-1} \notin \mathbb{Z}$. Therefore, we conclude that $\dfrac{pq}{2p-1} \in (\mathbb{Q}\backslash\mathbb{Z})\text{-}\mathcal{KS}(N)$.

Next, assume that $q = 2p + s$. Then, $s$ is odd, so $s \neq p - 1$. If $q - p + 1 \in \mathbb{Z}\text{-}\mathcal{KS}(N)$, then $s + 1 \mid p(q - 1)$. However, we know that $\gcd(p, s + 1) = 1$ because $s < p - 1$, which implies that $s + 1 \mid q - 1$. Hence, by taking $\alpha_1^{''} = 1$ and $\alpha_2 = 2p - 1$, we show that $\alpha_2 p - \alpha_1^{''} pq = -p(s + 1) \mid p(q - 1)$. Thus, as $\alpha_2 q - \alpha_1^{''} pq = q(p - 1)$, we can write

$$\begin{cases} \alpha_2 p - \alpha_1^{''} pq & \mid \quad p(q - 1) \\ \alpha_2 q - \alpha_1^{''} pq & \mid \quad q(p - 1). \end{cases}$$

This implies that $\dfrac{pq}{2p-1}$ is an $N$-Korselt base.

Now, as $\gcd(pq, 2p-1) = \gcd(q, q-1-s) = \gcd(q, s+1) = 1$ and $2p-1 \neq 1$, we deduce that $\dfrac{pq}{2p-1} \notin \mathbb{Z}$. Thus, $\dfrac{pq}{2p-1} \in (\mathbb{Q}\backslash\mathbb{Z})\text{-}\mathcal{KS}(N)$. $\square$

**Example 4.5.** Let $N = 2 \times 7$. Then, $\mathbb{Z}\text{-}\mathcal{KS}(N) = \{6, 8\}$ and $(\mathbb{Q}\backslash\mathbb{Z})\text{-}\mathcal{KS}(N) = \left\{\dfrac{7}{2}\right\}$ is exactly the set generated by $\mathbb{Z}\text{-}\mathcal{KS}(N)$. However, for $N = 3 \times 7$, we have $\mathbb{Z}\text{-}\mathcal{KS}(N) = \{5, 6, 9\}$ and $(\mathbb{Q}\backslash\mathbb{Z})\text{-}\mathcal{KS}(N) = \left\{\dfrac{7}{2}, \dfrac{7}{3}, \dfrac{21}{5}, \dfrac{21}{4}, \dfrac{15}{2}, \dfrac{33}{5}\right\}$, which is composed of more than the $N$-Korselt bases in $(\mathbb{Q}\backslash\mathbb{Z})$ generated by $\mathbb{Z}\text{-}\mathcal{KS}(N)$.

***Proof of the Main Theorem.*** Let $N = pq$, where $p < q$ are two prime numbers such that $(\mathbb{Q}\backslash\mathbb{Z})\text{-}\mathcal{KS}(N) = \emptyset$. Assume by contradiction that there exists $\beta \neq q + p - 1$ in $\mathbb{Z}\text{-}\mathcal{KS}(N)$. By Propositions 4.2 and 4.3, we know that $\beta \neq ip$ and $\beta \neq (i+1)p$, respectively. It follows that $\gcd(p, \beta) = \gcd(q, \beta) = 1$ by Proposition 2.4 and $q < 4p$ by Theorem 2.2.

Suppose that $q > 2p$. Then, by Corollary 3.2, we should have $\beta = q - p + 1$, and by Proposition 4.4, $\gcd(q+1, p) \neq 1$. However, since in our case, $2p < q = ip + s < 4p$ and $q$ is prime, this forces $q = 4p - 1$, and therefore, $\beta = q - p + 1 = 3p$, which contradicts $\gcd(p, \beta) = 1$.

Next, assume that $q < 2p$. Then, by Proposition 4.1, $\gcd(pq, p + q - \beta) \neq 1$; otherwise, $\beta$ generates an element in $(\mathbb{Q}\backslash\mathbb{Z})\text{-}\mathcal{KS}(N) = \emptyset$, which is impossible. This result implies that either $p$ or $q$ divides $p + q - \beta$, and one of the following holds:

- If $p$ divides $p + q - \beta$, then since $1 \leq p + q - \beta \leq 2p - 1$ by Proposition 2.4, we obtain $p = p + q - \beta$. Therefore, $\beta = q$, which is impossible.
- If $q$ divides $p + q - \beta$, then as $1 \leq p + q - \beta \leq 2p - 1 < 2q$ by Proposition 2.4, we obtain $q = p + q - \beta$. Hence, $\beta = p$, which is also impossible.

Thus, all cases lead to absurdity. Therefore, we conclude that $\beta = q + p - 1$ and $\mathbb{Z}\text{-}\mathcal{KS}(N) = \{q + p - 1\}$. $\square$

**Remark 4.6.** The converse of the main theorem is not true. For instance, if $N = 6 = 2\times 3$, then

$$\mathbb{Q}\text{-}\mathcal{KS}(N) = \left\{4, \frac{3}{2}, \frac{10}{3}, \frac{14}{5}, \frac{8}{3}, \frac{5}{2}, \frac{18}{7}, \frac{12}{5}, \frac{9}{4}\right\}.$$

This study motivates us to begin a deeper investigation of the rational Korselt set of a number $N$ with more than two prime factors. We believe that the study of a possible relation or relations between $(\mathbb{Q}\backslash\mathbb{Z})\text{-}\mathcal{KS}(N)$ and $\mathbb{Z}\text{-}\mathcal{KS}(N)$ can simplify this task, but not enough. The simple case when $N = pq$ is still full of unsolved problems. For instance, after examining the Korselt sets over $\mathbb{Q}$ of some values of $N = pq$, since $\mathbb{Q}\text{-}\mathcal{KW}(N)$ is finite (see [5, Theorem 2.3]), we state the following conjecture:

**Conjecture 4.7.** *For all $N = pq$, $\mathbb{Q}\text{-}\mathcal{KW}(N)$ is odd.*

**Acknowledgment.** I am grateful to the referee for his comments which have led to improvements in the paper.

## References

[1] K. Bouallegue, O. Echi and R. Pinch, *Korselt Numbers and Sets*, Int. J. Number Theory, **6**, 257–269, 2010.

[2] R.D. Carmichael, *On composite numbers P which satisfy the Fermat congruence* $a^{P-1} \equiv 1 \pmod{P}$, Amer. Math. Monthly, **19**, 22–27, 1912.

[3] O. Echi, *Williams Numbers*, C. R. Math. Acad. Sci. Soc. R. Can. **29**, 41–47, 2007.

[4] O. Echi and N. Ghanmi, *The Korselt Set of pq*, Int. J. Number Theory, **8** (2), 299–309, 2012.

[5] N. Ghanmi, $\mathbb{Q}$-*Korselt Numbers*, Turkish J. Math. **42**, 2752–2762, 2018.

[6] N. Ghanmi, *Rationel Korselt Bases of Prime Powers*, Stu. Sci. Math. Hungarica, 56 (4), 388-403 2019.
[7] N. Ghanmi, *The $\mathbb{Q}$-Korselt Set of pq*, Period. Math. Hungarica, **81**, 174–193, 2020.
[8] N.Ghanmi and I. Al-Rassasi, *On Williams Numbers With Three Prime Factors*, Missouri J. Math. Sci. **25** (2), 134–152, 2013.
[9] N. Ghanmi, O. Echi and I. Al-Rassasi, *The Korselt Set of a Squarefree Composite Number*, C. R. Math. Rep. Acad. Sci. Canada, **35** (1), 1–15, 2013.
[10] A. Korselt, *Problème chinois*, Interméd. Math. **6**, 142–143, 1899.

Hacettepe Journal of
Mathematics & Statistics

Hacet. J. Math. Stat.
Volume 50 (1) (2021), 144 – 145
DOI : 10.15672/hujms.674543

Research Article

# Addendum to "Ideal Rothberger spaces" [Hacet. J. Math. Stat. 47(1), 69-75, 2018]

Manoj Bhardwaj

*Department of Mathematics, University of Delhi, New Delhi-110007, India*

## Abstract

In this addendum we give an example to show that there is an error in Theorem 3.7 in "Ideal Rothberger spaces" [Hacet. J. Math. Stat. 47(1), 69-75, 2018]. We also prove the theorem with different hypothesis.

**Mathematics Subject Classification (2020).** 54D20, 54B20

**Keywords.** Rothberger modulo ideal spaces, perfect maps

We use notation and terminology from [2]. In [2], the author gave the following theorem for inverse invariant.

A function $f$ from a topological space $X$ to a space $Y$ is said to be perfect map [1] if

(1) $f$ is onto
(2) $f$ is continuous
(3) $f$ is closed map
(4) $f^{-1}(y)$ is compact in $X$ for each $y \in Y$.

**Theorem 1** ([2])**.** *Let $f : X \to Y$ be a perfect map and $\mathcal{I}$ be an ideal in $Y$. If $Y$ is Rothberger modulo $\mathcal{I}$, then $X$ is Rothberger modulo $f^{-1}(\mathcal{I})$.*

Here we give an example which contradicts the Theorem 1 given in [2].

**Example 2.** Let $\mathbb{R}$ be set of real numbers with usual topology and $\mathcal{I} = \{\phi\}$ be an ideal in $\{a\}$. Take a constant function $f$ from $[0,1]$ to one point Rothberger space or $\{a\}$, where $[0,1]$ is compact closed subspace of $\mathbb{R}$. Then $f$ is closed, open, onto and continuous map. Also $f^{-1}(a) = [0,1]$ is compact but $[0,1]$ is not Rothberger [3] since $\{a\}$ is Rothberger.

Now we give positive result regarding this and provide maps under which Rothberger modulo an ideal spaces are inverse invariant.

**Theorem 3.** *Let $f$ be an open bijective map from a space $X$ to $Y$ and $\mathcal{I}$ be an ideal in $Y$. If $Y$ is Rothberger modulo $\mathcal{I}$, then $X$ is Rothberger modulo $f^{-1}(\mathcal{I})$.*

***Proof.*** Let $< \mathcal{U}_n : n \in \omega >$ be a sequence of open covers of $X$. Then for each $n$,

$$\mathcal{V}_n = \{f(U) : U \in \mathcal{U}_n\}$$

Email address: manojmnj27@gmail.com
Received: 14.01.2020; Accepted: 20.05.2020

is an open cover of $Y$. Since $Y$ is Rothberger modulo $\mathcal{I}$, there are $J \in \mathcal{I}$ and a sequence $< \mathcal{W}_n : n \in \omega >$ such that for each $n$, $\mathcal{W}_n$ is a singleton subset of $\mathcal{U}_n$ and for each $y \in Y \setminus J$, belongs to $\bigcup \mathcal{W}_n$ for some $n$. Now assume that for each $n$,

$$\mathcal{W}_n = \{f(U_{n,1})\} \text{ and } \mathcal{G}_n = \{U_{n,1}\}.$$

Then $f^{-1}(J) \in f^{-1}(\mathcal{I})$ and sequence $< \mathcal{G}_n : n \in \omega >$ witnesses Rothberger modulo $f^{-1}(\mathcal{I})$ property of $X$ for the sequence $< \mathcal{U}_n : n \in \omega >$. Let $x \in X \setminus f^{-1}(J)$. Then

$$y = f(x) \in Y \setminus J \text{ and } y \in \bigcup \mathcal{W}_n \text{ for some } n.$$

This implies that $y \in f(U_{n,1})$. Since $f$ is one-to-one, $x \in U_{n,1}$. So $x \in \bigcup \mathcal{G}_n$ for some $n$. This completes the proof. □

## References


[1] R. Engelking, *General Topology, Revised and completed edition*, Heldermann Verlag Berlin, 1989.

[2] A. Güldürdek, *Ideal Rothberger spaces*, Hacet. J. Math. Stat. **47** (1), 69–75, 2018.

[3] M. Sakai and M. Scheepers, Combinatorics of open covers, in: K.P. Hart, J. van Mill, P. Simon (eds.), Recent Progress in General Topology III, pp. 751–799, Atlantis Press, Paris, 2014.

Hacettepe Journal of
Mathematics & Statistics

Hacet. J. Math. Stat.
Volume 50 (1) (2021), 146 – 158
DOI : 10.15672/hujms.597790

Research Article

# On new classes of chains of evolution algebras

Manuel Ladra*[1], Sherzod N. Murodov[2]

[1] Department of Mathematics & Institute of Mathematics, University of Santiago de Compostela, Santiago de Compostela, Spain

[2] Bukhara State Medical institute, Bukhara, Uzbekistan & Institute of Mathematics, University of Santiago de Compostela, Santiago de Compostela, Spain

**Abstract**

The paper is devoted to studying new classes of chains of evolution algebras and their time-depending dynamics and property transition.

**Mathematics Subject Classification (2020).** 17D92, 37C99, 15A24, 60J25

**Keywords.** evolution algebra, time, Chapman-Kolmogorov equation

## 1. Introduction

In the 1920s and 1930s, a new object, the general genetic algebra, was introduced into mathematics as a consequence of the synergy between Mendelian genetics and mathematics. Recognizing algebraic structures and properties in Mendelian genetics was one of the essential steps to start to study genetic algebras. Firstly, Mendel made use of some symbols [17], which expressed his genetic laws in an entirely algebraic manner. They were later named "Mendelian algebras" by several authors. Mendel's laws were mathematically formulated by Serebrowsky [25], who was the first to provide an algebraic interpretation of the sign "×", which suggested sexual reproduction. Later, Glivenkov [10] introduced the so-called Mendelian algebras. Independently, Kostitzin [15] also set forth a "symbolic multiplication" to express Mendel's laws. Etherington [6–8] made a systematic study of the algebras occurring in genetics and introduced the formal language of abstract algebra in the field of genetics. These algebras, in general, are non-associative.

The research on several classes of non-associative algebras (baric, evolution, Bernstein, train, stochastic, etc.) has rendered a notable enrichment to theoretical population genetics. Such classes have been defined at different times by various authors, and all algebras included in these classes are generally referred to as "genetic".

Essential contributions have also been made by Gonshor [11], Schafer [24], Holgate [13, 14], Heuch [12], Reiersöl [21], Abraham [1]. Until the 1980s, the most extensive reference in this area was Wörz-Busekros' book [28]. More recent results, such as evolution theory in genetic algebras, can be seen in Lyubich's book [16]. An excellent survey article is Reed's paper [20]. All algebras studied by these authors are generally called "genetic".

*Corresponding Author.
Email addresses: manuel.ladra@usc.es (M. Ladra), murodovs@yandex.ru (Sh. N. Murodov)
Received: 28.07.2019; Accepted: 20.05.2020

In the present days, non-Mendelian genetics has become an essential language for molecular geneticists. Some questions arise naturally in this context, such as what new subjects non-Mendelian genetics brings to mathematics, or what type of mathematics leads to a better understanding of non-Mendelian genetics. The systematic formulation of reproduction in non-Mendelian genetics as multiplication in algebras was introduced in [27], leading to the so-called "evolution algebras". These are algebras in which the multiplication tables are motivated by evolution laws of genetics.

Tian in [26] develops the framework of evolution algebra theory and applications in non-Mendelian genetics and Markov chains. The concept of evolution algebra is situated between algebras and dynamical systems. Evolution algebras associated with function spaces defined by graphs, state spaces, and Gibbs measures are studied in [23].

A notion of a chain of evolution algebras was introduced in [4], where the sequence of matrices of structural constants of the chain of evolution algebras satisfies an analogue of the Chapman-Kolmogorov equation. In [22], twenty-five distinct examples of chains of two-dimensional evolution algebras are constructed.

In this paper, we present examples of chains of two-dimensional evolution algebras other than those of [22], by studying the behavior of the baric property, of the set of absolute nilpotent elements and the time-depending dynamics of the set of idempotent elements.

The paper is organized as follows. In Section 2, we give the main concepts related to a chain of evolution algebras. In Section 3, we construct new chains of evolution algebras (CEAs) and study their time-depending dynamics. Finally, in Section 4, we analyze the property transitions of the new CEAs.

## 2. Chain of evolution algebras

Evolution algebras are defined as follows.

**Definition 2.1.** Let $(E,\cdot)$ be an algebra over a field $K$. If it admits a basis $\{e_1, e_2, \dots\}$, such that

$$e_i \cdot e_j = \begin{cases} 0, & \text{if } \ i \neq j; \\ \sum\limits_k a_{ik} e_k, & \text{if } \ i = j, \end{cases}$$

then this algebra is called an *evolution algebra*. The basis is called a natural basis.

The matrix $M = (a_{ij})$ is called the matrix of structural constants.

Evolution algebras have the following primary properties (see [26]). Evolution algebras are not associative, in general; they are commutative, flexible, but not power-associative, in general; direct sums of evolution algebras are also evolution algebras; Kronecker products of evolutions algebras are also evolution algebras.

Let $\{e_1, e_2\}$ be a basis of the two-dimensional evolution algebra $E$. It is evident that if $\dim E^2 = 0$, then $E$ is an abelian algebra, i.e. an algebra with all products equal to zero. The next theorem gives the classification of the real two-dimensional evolution algebras.

**Theorem 2.2** ([19])**.** *Any two-dimensional real evolution algebra $E$ is isomorphic to one of the following pairwise non-isomorphic algebras:*

(i) $\dim E^2 = 1$.

$E_1 : e_1e_1 = e_1, \quad e_2e_2 = 0, \quad$ *with matrix* $M_1 = \begin{pmatrix} 1 & 0 \\ 0 & 0 \end{pmatrix}$;

$E_2 : e_1e_1 = e_1, \quad e_2e_2 = e_1, \quad$ *with matrix* $M_2 = \begin{pmatrix} 1 & 1 \\ 0 & 0 \end{pmatrix}$;

$E_3 : e_1e_1 = e_1 + e_2, \quad e_2e_2 = -e_1 - e_2, \quad$ *with matrix* $M_3 = \begin{pmatrix} 1 & -1 \\ 1 & -1 \end{pmatrix}$;

$E_4 : e_1e_1 = e_2,\quad e_2e_2 = 0,\quad$ *with matrix* $M_4 = \begin{pmatrix} 0 & 0 \\ 1 & 0 \end{pmatrix}$;

$E_5 : e_1e_1 = e_2,\quad e_2e_2 = -e_2,\quad$ *with matrix* $M_5 = \begin{pmatrix} 0 & 0 \\ 1 & -1 \end{pmatrix}$;

(ii) $\dim E^2 = 2$.
$E_6(a_2, a_3) : e_1e_1 = e_1 + a_2e_2,\quad e_2e_2 = a_3e_1 + e_2,\ 1 - a_2a_3 \neq 0,\ a_2, a_3 \in \mathbb{R},$ *with matrix* $M_6 = \begin{pmatrix} 1 & a_3 \\ a_2 & 1 \end{pmatrix}$. *Moreover,* $E_6(a_2, a_3)$ *is isomorphic to* $E_6(a_3, a_2)$.

$E_7(a_4) : e_1e_1 = e_2,\quad e_2e_2 = e_1 + a_4e_2,$ *where* $a_4 \in \mathbb{R}$, *with matrix* $M_7 = \begin{pmatrix} 0 & 1 \\ 1 & a_4 \end{pmatrix}$.

Different authors performed the classification of two-dimensional evolution algebras over several fields. In [5] for the field of complex numbers, in [2] over a field that is closed under all square and cubic roots, and in [3, 9] without restrictions on the underlying field.

**Remark 2.3.** We notice that the classification of the two-dimensional real evolution algebras consists of an alternative of the complex case [5] or the case [3]. $E_5$ only appears in the real case. Observe that $E_5$ is isomorphic to the algebra with matrix $\begin{pmatrix} -1 & 1 \\ 0 & 0 \end{pmatrix}$. In the proof of [3, Theorem 3.3], case 1.2.2, the algebra $E_5$ does not appear since the author considers $c_1 \neq 0$, but if $c_1$ is negative there is no $\sqrt{c_1}$, and therefore there is one more case. Moreover, the cases (f), (g) and (h) of [3, Theorem 3.3] correspond to $E_6(0, a_3)$ with $a_3 \neq 0$, $E_6(0, 0)$, and $E_7(0)$, respectively.

Following [4] we consider a family $\left\{E^{[s,t]} :\ s, t \in \mathbb{R},\ 0 \leq s \leq t\right\}$ of $n$-dimensional evolution algebras over the field $\mathbb{R}$, with basis $e_1, \ldots, e_n$, and the multiplication table

$$e_ie_i = \sum_{j=1}^{n} a_{ij}^{[s,t]} e_j,\quad i = 1, \ldots, n; \qquad e_ie_j = 0,\quad i \neq j.$$

Here parameters $s, t$ are considered as time, and we define $\mathcal{T} = \{(s, t) : 0 \leq s \leq t,\ \text{where}\ s, t \in \mathbb{R}\}$.

Denote by $M^{[s,t]} = \left(a_{ij}^{[s,t]}\right)_{i,j=1,\ldots,n}$ the matrix of structural constants.

**Definition 2.4.** A family $\left\{E^{[s,t]} :\ s, t \in \mathbb{R},\ 0 \leq s \leq t\right\}$ of $n$-dimensional evolution algebras over the field $\mathbb{R}$ is called a *chain of evolution algebras* (CEA) if the matrix $M^{[s,t]}$ of structural constants satisfies the Chapman-Kolmogorov equation

$$M^{[s,t]} = M^{[s,\tau]} M^{[\tau,t]},\quad \text{for any}\quad s < \tau < t. \tag{2.1}$$

## 3. Construction of chains of evolution algebras

To construct a chain of two-dimensional evolution algebras, we need to solve equation (2.1) for the $2 \times 2$ matrix $\mathcal{M}^{[s,t]}$. This equation provides the following system of functional equations (with four unknown functions):

$$\begin{aligned}
a_{11}^{[s,t]} &= a_{11}^{[s,\tau]} a_{11}^{[\tau,t]} + a_{12}^{[s,\tau]} a_{21}^{[\tau,t]}, \\
a_{12}^{[s,t]} &= a_{11}^{[s,\tau]} a_{12}^{[\tau,t]} + a_{12}^{[s,\tau]} a_{22}^{[\tau,t]}, \\
a_{21}^{[s,t]} &= a_{21}^{[s,\tau]} a_{11}^{[\tau,t]} + a_{22}^{[s,\tau]} a_{21}^{[\tau,t]}, \\
a_{22}^{[s,t]} &= a_{21}^{[s,\tau]} a_{12}^{[\tau,t]} + a_{22}^{[s,\tau]} a_{22}^{[\tau,t]}.
\end{aligned} \tag{3.1}$$

But the general analysis of system (3.1) is complicated.

In [18] we studied the classification dynamics of known two-dimensional chains of evolution algebras constructed in [22] and showed that known chains of evolution algebras never contain an evolution algebra isomorphic to $E_4$ in any time $s,t$ (see Theorem 2.2). In this section, we will construct CEAs, including $E_4$ for some period of time.

To construct a CEA that will be isomorphic to $E_4$ at some time interval, we need the following theorem.

**Theorem 3.1** ([18]). *An evolution algebra $E_{\mathcal{M}}$ is isomorphic to $E_4$ if and only if $E_{\mathcal{M}}$ has the matrix of structural constants in the following form:*

$$\mathcal{M}_1 = \begin{pmatrix} 0 & \beta \\ 0 & 0 \end{pmatrix} \quad \text{or} \quad \mathcal{M}_2 = \begin{pmatrix} 0 & 0 \\ \gamma & 0 \end{pmatrix}, \quad \text{where} \quad \beta, \gamma \in \mathbb{R}. \tag{3.2}$$

Thus, we should construct CEAs with the matrix of structural constants that are listed in (3.2).

Consider (3.1) with $a_{11}^{[s,t]} = \alpha(s,t)$, $a_{12}^{[s,t]} = \beta(s,t)$, $a_{21}^{[s,t]} = \gamma(s,t)$, $a_{22}^{[s,t]} = \delta(s,t)$. Therefore, to find a CEA, we should solve the next equation:

$$\begin{pmatrix} \alpha(s,\tau) & \beta(s,\tau) \\ \gamma(s,\tau) & \delta(s,\tau) \end{pmatrix} \cdot \begin{pmatrix} \alpha(\tau,t) & \beta(\tau,t) \\ \gamma(\tau,t) & \delta(\tau,t) \end{pmatrix} = \begin{pmatrix} \alpha(s,t) & \beta(s,t) \\ \gamma(s,t) & \delta(s,t) \end{pmatrix}. \tag{3.3}$$

**Case 1.1.** If we consider in (3.3), $\alpha(s,t) = \gamma(s,t) \equiv 0, \beta(s,t) \neq 0, \delta(s,t) \neq 0$, then we have the following:

$$\begin{pmatrix} 0 & \beta(s,\tau) \\ 0 & \delta(s,\tau) \end{pmatrix} \cdot \begin{pmatrix} 0 & \beta(\tau,t) \\ 0 & \delta(\tau,t) \end{pmatrix} = \begin{pmatrix} 0 & \beta(s,t) \\ 0 & \delta(s,t) \end{pmatrix}. \tag{3.4}$$

From (3.4), we get the following system of functional equations:

$$\begin{cases} \beta(s,\tau)\delta(\tau,t) = \beta(s,t), \\ \delta(s,\tau)\delta(\tau,t) = \delta(s,t). \end{cases} \tag{3.5}$$

The second equation of system (3.5) is known as Cantor's second equation, which has the following solutions:

(1) $\delta(s,t) \equiv 0$;
(2) $\delta(s,t) = \frac{\phi(t)}{\phi(s)}$, where $\phi$ is an arbitrary function with $\phi(s) \neq 0$;
(3) $\delta(s,t) = \begin{cases} 1, & \text{if } 0 < s \leq t < a; \\ 0, & \text{if } t \geq a. \end{cases}$

Substituting these solutions into the first equation of (3.5), we find $\beta(s,t)$:

(1) $\beta(s,t) \equiv 0$;
(2) $\beta(s,t) = \rho(s)\phi(t)$, where $\rho$ is an arbitrary function;
(3) $\beta(s,t) = \begin{cases} \sigma(s), & \text{if } 0 < s \leq t < a; \\ 0, & \text{if } t \geq a, \end{cases}$

where $\sigma$ is an arbitrary function;

From these solutions, we have the following matrices of structural constants of CEAs:

$$\mathcal{M}_0^{[s,t]} = \begin{pmatrix} 0 & 0 \\ 0 & 0 \end{pmatrix},$$

$$\mathcal{M}_1^{[s,t]} = \begin{pmatrix} 0 & \rho(s)\phi(t) \\ 0 & \frac{\phi(t)}{\phi(s)} \end{pmatrix},$$

where $\rho, \phi$ are arbitrary functions, with $\phi(s) \neq 0$;

$$\mathcal{M}_2^{[s,t]} = \begin{cases} \begin{pmatrix} 0 & \sigma(s) \\ 0 & 1 \end{pmatrix}, & \text{if } 0 < s \leq t < a; \\ \begin{pmatrix} 0 & 0 \\ 0 & 0 \end{pmatrix}, & \text{if } t \geq a, \end{cases}$$

where $a > 0$ and $\sigma$ is an arbitrary function.

**Case 1.2.** Consider the case $\alpha(s,t) = \beta(s,t) \equiv 0, \gamma(s,t) \neq 0, \delta(s,t) \neq 0$. Then from (3.3), we have the following:

$$\begin{pmatrix} 0 & 0 \\ \gamma(s,\tau) & \delta(s,\tau) \end{pmatrix} \cdot \begin{pmatrix} 0 & 0 \\ \gamma(\tau,t) & \delta(\tau,t) \end{pmatrix} = \begin{pmatrix} 0 & 0 \\ \gamma(s,t) & \delta(s,t) \end{pmatrix}.$$

From the last equality, we have the following system of equations:

$$\begin{cases} \delta(s,\tau)\gamma(\tau,t) = \gamma(s,t), \\ \delta(s,\tau)\delta(\tau,t) = \delta(s,t). \end{cases} \tag{3.6}$$

The second equation (Cantor's second equation) of system (3.6) has the following solutions:

(1) $\delta(s,t) \equiv 0$;

(2) $\delta(s,t) = \frac{\varphi(t)}{\varphi(s)}$, where $\varphi$ is an arbitrary function with $\varphi(s) \neq 0$;

(3) $\delta(s,t) = \begin{cases} 1, & \text{if } 0 < s \leq t < a; \\ 0, & \text{if } t \geq a. \end{cases}$

Substituting these solutions into the first equation of (3.6), we find $b(s,t)$:

(1) $\gamma(s,t) \equiv 0$;

(2) $\gamma(s,t) = \frac{f(t)}{\varphi(s)}$, where $f$ is an arbitrary function;

(3) $\gamma(s,t) = \begin{cases} g(t), & \text{if } 0 < s \leq t < a; \\ 0, & \text{if } t \geq a. \end{cases}$ where $g$ is an arbitrary function.

From these solutions, we have the next matrices of structural constants of CEAs:

$$\mathcal{M}_0^{[s,t]} = \begin{pmatrix} 0 & 0 \\ 0 & 0 \end{pmatrix},$$

$$\mathcal{M}_3^{[s,t]} = \begin{pmatrix} 0 & 0 \\ \frac{f(t)}{\varphi(s)} & \frac{\varphi(t)}{\varphi(s)} \end{pmatrix},$$

where $f, \varphi$ are arbitrary functions, $\varphi(s) \neq 0$;

$$\mathcal{M}_4^{[s,t]} = \begin{cases} \begin{pmatrix} 0 & 0 \\ g(t) & 1 \end{pmatrix}, & \text{if } 0 < s \leq t < a; \\ \begin{pmatrix} 0 & 0 \\ 0 & 0 \end{pmatrix}, & \text{if } t \geq a, \end{cases}$$

where $a > 0$ and $g$ is an arbitrary function.

**Case 1.3.** Let us try to find the solution satisfying the following:

$$\begin{pmatrix} \alpha(s,\tau) & \beta(s,\tau) \\ \gamma(s,\tau) & \delta(s,\tau) \end{pmatrix} \cdot \begin{pmatrix} \alpha(\tau,t) & \beta(\tau,t) \\ \gamma(\tau,t) & \delta(\tau,t) \end{pmatrix} = \begin{pmatrix} 0 & \beta(s,t) \\ 0 & 0 \end{pmatrix}. \tag{3.7}$$

From (3.7) we have the next system of functional equations:

$$\begin{cases} \alpha(s,\tau)\alpha(\tau,t)+\beta(s,\tau)\gamma(\tau,t)=0, \\ \alpha(s,\tau)\beta(\tau,t)+\beta(s,\tau)\delta(\tau,t)=\beta(s,t), \\ \gamma(s,\tau)\alpha(\tau,t)+\delta(s,\tau)\gamma(\tau,t)=0, \\ \gamma(s,\tau)\beta(\tau,t)+\delta(s,\tau)\delta(\tau,t)=0. \end{cases} \tag{3.8}$$

Let $\alpha(s,t)=\gamma(s,t)=0$. Then we get:

$$\begin{cases} \beta(s,\tau)\delta(\tau,t)=\beta(s,t), \\ \delta(s,\tau)\delta(\tau,t)=0. \end{cases} \tag{3.9}$$

To find a non-zero solution of the system of equations (3.9), we should prove that the equation

$$\delta(s,\tau)\delta(\tau,t)=0, \quad \text{for all } \ s<\tau<t, \tag{3.10}$$

has a non-zero solution. Indeed, take $C>0$ and

$$\delta(s,t)=\begin{cases} 0, & \text{if } \ 0<C\leq s<t \ \text{ or } \ 0<s<t\leq C; \\ f(s,t), & \text{if } \ 0<s<C<t, \end{cases} \tag{3.11}$$

where $f(s,t)$ is an arbitrary non-zero function.

Now, we show that independently on $f(s,t)$ the function (3.11) satisfies (3.10): for a given $C>0$, we only have two possibilities by taking an arbitrary $\tau$ such that $s<\tau<t$:

**Case 1.3.1.** Let $\tau\leq C$. By the defined function (3.11), we have that $\delta(s,\tau)=0$ and for $\delta(\tau,t)$:

$$\delta(\tau,t)=\begin{cases} 0, & \text{if } \ t\leq C; \\ f(\tau,t), & \text{if } \ t>C, \end{cases} \tag{3.12}$$

where $f(\tau,t)$ is the function fixed in (3.11).

Therefore, $\delta(s,\tau)\delta(\tau,t)=0$.

**Case 1.3.2.** $\tau>C$. Also from (3.11), we have that $\delta(\tau,t)=0$ and for $\delta(s,\tau)$:

$$\delta(s,\tau)=\begin{cases} f(s,\tau), & \text{if } \ s<C; \\ 0, & \text{if } \ s\geq C, \end{cases}$$

where $f(s,\tau)$ is the function fixed in (3.11).

Therefore, $\delta(s,\tau)\delta(\tau,t)=0$.

Thus, we have proved that the function (3.11) satisfies equation (3.10).

Now we should find solutions to the first equation of system (3.9):

$$\beta(s,\tau)\delta(\tau,t)=\beta(s,t), \qquad s<\tau<t, \tag{3.13}$$

where $\delta(\tau,t)$ is given by (3.11).

To find a solution, we have the next possibilities:

**Case 1.3.3.** Let $\tau\leq C$. Then by the defined function (3.11) we have that $\delta(s,\tau)=0$ and from (3.12) in a period of time $t\leq C$, $\delta(\tau,t)=0$, and so from (3.13) we have $\beta(s,t)=0$. When $t>C$, $\delta(\tau,t)=f(\tau,t)$ and by (3.13) we have to solve the next equation:

$$\beta(s,\tau)f(\tau,t)=\beta(s,t), \qquad s<\tau<t. \tag{3.14}$$

We solve (3.14) for some particular cases:

**Case 1.3.3.1** Consider $\beta(s,t)=f(s,t)$. Then from (3.14), we have $f(s,\tau)f(\tau,t)=f(s,t)$, which is Cantor's second equation. As $f(s,t)$ is a non-zero function, then we have the next solution:

$$f(s,t) = \frac{\Phi(t)}{\Phi(s)},$$

where $\Phi$ is an arbitrary function, with $\Phi(s) \neq 0$.

Thus we have the next solution of system (3.8):

$$\begin{aligned}
&\alpha(s,t) \equiv 0,\\
&\beta(s,t) = \begin{cases} 0, & \text{if } \ s < t \leq C;\\ \frac{\Phi(t)}{\Phi(s)}, & \text{if } \ t > C,\end{cases}\\
&\gamma(s,t) \equiv 0,\\
&\delta(s,t) = \begin{cases} 0, & \text{if } \ 0 < C \leq s < t \ \text{ or } \ 0 < s < t \leq C;\\ \frac{\Phi(t)}{\Phi(s)}, & \text{if } \ s < C < t,\end{cases}
\end{aligned}$$

where $C > 0$ and $\Phi$ is an arbitrary function, with $\Phi(s) \neq 0$.

Then we have the next matrix of structural constants:

$$\mathcal{M}_5^{[s,t]} = \begin{cases} \begin{pmatrix} 0 & 0\\ 0 & 0\end{pmatrix}, & \text{if } \quad s < t \leq C;\\ \begin{pmatrix} 0 & \frac{\Phi(t)}{\Phi(s)}\\ 0 & 0\end{pmatrix}, & \text{if } \quad t > C,\end{cases}$$

where $C > 0$ and $\Phi$ is an arbitrary function, with $\Phi(t) \neq 0$.

**Case 1.3.3.2.** Let $\beta(s,t) \neq f(s,t)$. As $f(\tau,t)$ is an arbitrary non-zero function, consider $f(\tau,t) = \frac{\phi(\tau)}{\phi(t)}$, with $\phi(t) \neq 0$. Then from (3.14) we have the following:

$$\beta(s,\tau) \cdot \frac{\phi(\tau)}{\phi(t)} = \beta(s,t),$$
$$\beta(s,t)\phi(t) = \beta(s,\tau)\phi(\tau).$$

From the last equality, we can see $\beta(s,t)\phi(t)$ does not depend on $t$, i.e. there exists a function $\rho(s)$ such that $\beta(s,t)\phi(t) = \rho(s)$. Therefore, $\beta(s,t) = \frac{\rho(s)}{\phi(t)}$.

Then we get the next solution of system (3.8):

$$\begin{aligned}
&\alpha(s,t) \equiv 0,\\
&\beta(s,t) = \begin{cases} 0, & \text{if } \ s < t \leq C;\\ \frac{\rho(s)}{\phi(t)}, & \text{if } \ t > C,\end{cases}\\
&\gamma(s,t) \equiv 0,\\
&\delta(s,t) = \begin{cases} 0, & \text{if } \ 0 < C \leq s < t \ \text{ or } \ 0 < s < t \leq C;\\ \frac{\phi(s)}{\phi(t)}, & \text{if } \ s < C < t,\end{cases}
\end{aligned}$$

where $C > 0$ and $\phi, \rho$ are arbitrary functions with $\phi(t) \neq 0$.

Then we have, respectively, the next matrix of structural constants to the solution:

$$\mathcal{M}_6^{[s,t]} = \begin{cases} \begin{pmatrix} 0 & 0\\ 0 & 0\end{pmatrix}, & \text{if } \quad s < t \leq C;\\ \begin{pmatrix} 0 & \frac{\rho(s)}{\phi(t)}\\ 0 & 0\end{pmatrix}, & \text{if } \quad t > C,\end{cases}$$

where $C > 0$ and $\phi, \rho$ are arbitrary functions with $\phi(t) \neq 0$.

**Case 1.3.4.** When $\tau > C$, then by the defined function (3.11) we have that $\delta(\tau,t) = 0$. So from (3.13), we have $\beta(s,t) = 0$. Thus we get the trivial CEA.

**Case 1.4.** Let us try to find the solution satisfying:

$$\begin{pmatrix} \alpha(s,\tau) & \beta(s,\tau) \\ \gamma(s,\tau) & \delta(s,\tau) \end{pmatrix} \cdot \begin{pmatrix} \alpha(\tau,t) & \beta(\tau,t) \\ \gamma(\tau,t) & \delta(\tau,t) \end{pmatrix} = \begin{pmatrix} 0 & 0 \\ \gamma(s,t) & 0 \end{pmatrix}. \tag{3.15}$$

From equality (3.15) we have the next system of functional equations:

$$\begin{cases} \alpha(s,\tau)\alpha(\tau,t) + \beta(s,\tau)\gamma(\tau,t) = 0, \\ \alpha(s,\tau)\beta(\tau,t) + \beta(s,\tau)\delta(\tau,t) = 0, \\ \gamma(s,\tau)\alpha(\tau,t) + \delta(s,\tau)\gamma(\tau,t) = \gamma(s,t), \\ \gamma(s,\tau)\beta(\tau,t) + \delta(s,\tau)\delta(\tau,t) = 0. \end{cases}$$

Let $\alpha(s,t) = \beta(s,t) = 0$. Then we have the next system:

$$\begin{cases} \delta(s,\tau)\gamma(\tau,t) = \gamma(s,t), \\ \delta(s,\tau)\delta(\tau,t) = 0. \end{cases}$$

The analysis of this system is similar to (3.9), and we get the following CEAs:

$$\mathcal{M}_7^{[s,t]} = \begin{cases} \begin{pmatrix} 0 & 0 \\ \frac{\Psi(t)}{\Psi(s)} & 0 \end{pmatrix}, & \text{if} \quad s < C; \\ \begin{pmatrix} 0 & 0 \\ 0 & 0 \end{pmatrix}, & \text{if} \quad s \geq C, \end{cases}$$

where $C > 0$ and $\Psi$ is an arbitrary function, with $\Psi(t) \neq 0$;

$$\mathcal{M}_8^{[s,t]} = \begin{cases} \begin{pmatrix} 0 & 0 \\ \frac{\sigma(t)}{\varphi(s)} & 0 \end{pmatrix}, & \text{if} \quad s < C; \\ \begin{pmatrix} 0 & 0 \\ 0 & 0 \end{pmatrix}, & \text{if} \quad s \geq C, \end{cases}$$

where $C > 0$ and $\varphi, \sigma$ are arbitrary functions with $\varphi(s) \neq 0$.

Denote by $E_i^{[s,t]}$ the CEA with matrix $\mathcal{M}_i^{[s,t]}$.

**Remark 3.2.** We should note that from the CEAs $E_i^{[s,t]}, i = 1, \dots, 8$, only $E_3^{[s,t]}$ coincides with the CEA $E_{16}^{[s,t]}$ constructed in [22] and it has the same dynamic. All other CEAs are different from CEAs constructed in [22] and have different dynamics.

Now, we provide the time-depending dynamics of these CEAs:

**Theorem 3.3.** *For the next CEAs hold:*

$$E_1^{[s,t]} \simeq \begin{cases} E_1 & \textit{for all} \ \ (s,t) \in \{(s,t) : s < t, \quad \rho(s) = 0\}, \\ E_2 & \textit{for all} \ \ (s,t) \in \{(s,t) : s < t, \quad \rho(s) \neq 0\}; \end{cases}$$

$$E_2^{[s,t]} \simeq \begin{cases} E_1 & \textit{for all} \ \ (s,t) \in \{(s,t) : s < t < a, \quad \sigma(s) = 0\}, \\ E_2 & \textit{for all} \ \ (s,t) \in \{(s,t) : s < t < a, \quad \sigma(s) \neq 0\}, \\ E_0 & \textit{for all} \ \ (s,t) \in \{(s,t) : t \geq a\}; \end{cases}$$

$E_3^{[s,t]} \simeq E_1$ *for any* $(s,t) \in \mathfrak{T}$;

$$E_4^{[s,t]} \simeq \begin{cases} E_1 & \textit{for all} \ \ (s,t) \in \{(s,t) : s < t < a\}, \\ E_0 & \textit{for all} \ \ (s,t) \in \{(s,t) : t \geq a\}; \end{cases}$$

$$E_5^{[s,t]} \simeq \begin{cases} E_0 & \textit{for all} \ \ (s,t) \in \{(s,t) : s < t \leq C\}, \\ E_4 & \textit{for all} \ \ (s,t) \in \{(s,t) : t > C\}; \end{cases}$$

$$E_6^{[s,t]} \simeq \begin{cases} E_0 & \textit{for all} \ \ (s,t) \in \{(s,t) : s < t \leq C\}, \\ E_0 & \textit{for all} \ \ (s,t) \in \{(s,t) : t > C, \quad \rho(s) = 0\}, \\ E_4 & \textit{for all} \ \ (s,t) \in \{(s,t) : t > C, \quad \rho(s) \neq 0\}; \end{cases}$$

$$E_7^{[s,t]} \simeq \begin{cases} E_4 & \textit{for all} \ \ (s,t) \in \{(s,t) : s < C\}, \\ E_0 & \textit{for all} \ \ (s,t) \in \{(s,t) : s \geq C\}; \end{cases}$$

$$E_8^{[s,t]} \simeq \begin{cases} E_0 & \textit{for all} \ \ (s,t) \in \{(s,t) : s < C, \quad \sigma(t) = 0\}, \\ E_4 & \textit{for all} \ \ (s,t) \in \{(s,t) : s < C, \quad \sigma(t) \neq 0\}, \\ E_0 & \textit{for all} \ \ (s,t) \in \{(s,t) : s \geq C\}. \end{cases}$$

***Proof.*** When $\rho(s) = 0$, then $E_1^{[s,t]} \simeq E_1$, for all $s, t \in \mathfrak{T}$ by the change of basis $e_1' = e_1, \ e_2' = \frac{\phi(s)}{\phi(t)} e_2$, and when $\rho(s) \neq 0$, it is isomorphic to $E_2$, for all $s, t \in \mathfrak{T}$ by the change of basis $e_1' = \frac{1}{\rho(s)\phi(t)} e_1, \ e_2' = \frac{\phi(s)}{\phi(t)} e_2$.

When $\sigma(s) = 0$, then $E_2^{[s,t]} \simeq E_1$, for all $s, t \in \mathfrak{T}, \ s < t < a$, by the change of basis $e_1' = e_1, \ e_2' = e_2$, and when $\sigma(s) \neq 0$, it is isomorphic to $E_2$, for all $s, t \in \mathfrak{T}, \ s < t < a$, by the change of basis $e_1' = \frac{1}{\sigma(s)} e_1, \ e_2' = e_2$. In the period of time $t \geq a$, it will be isomorphic to the trivial evolution algebra $E_0$.

$E_3^{[s,t]} \simeq E_1$, for all $s, t \in \mathfrak{T}$ by the change of basis $e_2' = \frac{f(t)\varphi(s)}{\varphi^2(t)} e_1 + \frac{\varphi(s)}{\varphi(t)} e_2, \ e_2' = e_1$.

$E_4^{[s,t]} \simeq E_1$, for all $s, t \in \mathfrak{T}, \ s < t < a$, by the change of basis $e_1' = \sigma(t) e_1 + e_2, \ \ e_2' = e_1$, in the period of time $t \geq a$, it will be isomorphic to the trivial evolution algebra $E_0$.

$E_5^{[s,t]} \simeq E_4$, for all $s, t \in \mathfrak{T}, \ t > C$, by the change of basis $e_1' = \frac{\Phi(s)}{\Phi(t)} e_1, \ \ e_2' = e_2$, in the period of time $s < t \leq C$, it will be isomorphic to the trivial evolution algebra $E_0$.

When $\rho(s) \neq 0$, then $E_6^{[s,t]} \simeq E_4$, for all $s, t \in \mathfrak{T}, \ t > C$, by the change of basis $e_1' = \frac{\phi(t)}{\rho(s)} e_1, \ \ e_2' = e_2$, in the period of time $s < t \leq C$, and when $\rho(s) = 0$, then it will be isomorphic to the trivial evolution algebra $E_0$.

$E_7^{[s,t]} \simeq E_4$, for all $s, t \in \mathfrak{T}, \ s < C$, by the change of basis $e_1' = \frac{\Psi(s)}{\Psi(t)} e_1, \ \ e_2' = e_2$, in the period of time $s \geq C$, it will be isomorphic to the trivial evolution algebra $E_0$.

When $\sigma(t) \neq 0$, then $E_6^{[s,t]} \simeq E_4$, for all $s, t \in \mathfrak{T}, \ s < C$, by the change of basis $e_1' = \frac{\varphi(s)}{\sigma(t)} e_1, \ \ e_2' = e_2$, in the period of time $s \geq C$, and when $\sigma(t) = 0$, then it will be isomorphic to the trivial evolution algebra $E_0$. □

Thus, we proved that there exist CEAs that for some values of time will be isomorphic to $E_4$.

## 4. Property transition

In this section, we will study property transitions of the CEAs $E_i^{s,t}, i = 0. \dots, 8$.

In [4], we provided the ideas of property transition for CEAs. We recall these definitions.

**Definition 4.1.** Assume a CEA, $E^{[s,t]}$, has a property, say $P$, at pair of times $(s_0, t_0)$; one says that the CEA has $P$ *property transition* if there is a pair $(s,t) \neq (s_0, t_0)$ at which the CEA has no property $P$.

Denote

$$\begin{aligned} \mathcal{T} &= \{(s,t) : 0 \leq s \leq t\}; \\ \mathcal{T}_P &= \{(s,t) \in \mathcal{T} : E^{[s,t]} \text{ has property } P\}; \\ \mathcal{T}_P^0 &= \mathcal{T} \setminus \mathcal{T}_P = \{(s,t) \in \mathcal{T} : E^{[s,t]} \text{ has no property } P\}. \end{aligned}$$

The sets have the following meaning:

- $\mathcal{T}_P$-the duration of the property $P$;
- $\mathcal{T}_P^0$-the lost duration of the property $P$.

The partition $\{\mathcal{T}_P, \mathcal{T}_P^0\}$ of the set $\mathcal{T}$ is called the $P$ property diagram.

For example, if $P$ =commutativity, then we determine that any CEA has not commutativity property transition because any evolution algebra is commutative.

### 4.1. Baric property transition

A *character* for an algebra $A$ is a nonzero multiplicative linear form on $A$, i.e. a nonzero algebra homomorphism $\sigma \colon A \to \mathbb{R}$ (see [16]). Not every algebra carries a character. For example, an algebra with the zero multiplication has no character.

**Definition 4.2.** A pair $(A, \sigma)$ consisting of an algebra $A$ and a character $\sigma$ on $A$ is called a *baric algebra*. The homomorphism $\sigma$ is called the weight (or baric) function of $A$ and $\sigma(x)$ the weight (baric value) of $x$.

There is a character $\sigma(x) = \sum_i x_i$ for the evolution algebra of a free population (see [16]); therefore, that algebra is baric. But the evolution algebra $E$ introduced in [26] is not baric, in general. The following theorem provides a criterion for an evolution algebra $E$ to be baric.

**Theorem 4.3** ([4]). *An $n$-dimensional evolution algebra $E$, over the field $\mathbb{R}$, is baric if and only if there is a column $(a_{1i_0}, \dots, a_{ni_0})^T$ of its structural constants matrix $\mathcal{M} = (a_{ij})_{i,j=1,\dots,n}$, such that $a_{i_0i_0} \neq 0$ and $a_{ii_0} = 0$, for all $i \neq i_0$. Moreover, the corresponding weight function is $\sigma(x) = a_{i_0i_0}x_{i_0}$.*

Since an evolution algebra is not a baric algebra, in general, using Theorem 4.3, we can give the baric property diagram. Let us do this for the above-given chains $E_i^{[s,t]}$, $i = 0, \dots, 8$.

Denote by $\mathcal{T}_b^{(i)}$ the baric property duration of the CEA $E_i^{[s,t]}$, $i = 0, \dots, 8$.

**Theorem 4.4.**

(i) (There is no non-baric property transition) *The algebras $E_i^{[s,t]}$, $i = 0, 1, 2, 5, 6, 7, 8$, are not baric for any time $(s,t) \in \mathcal{T}$;*

(ii) (There is no baric property transition) *The algebra $E_3^{[s,t]}$ is baric for any time $(s,t) \in \mathcal{T}$;*

(iii) (There is baric property transition) *The CEA $E_4^{[s,t]}$ has baric property transition with baric property duration set as the following*

$$\mathcal{T}_b^{(4)} = \{(s,t) \in \mathcal{T} : s \leq t < a\}.$$

***Proof.*** By Theorem 4.3, a two-dimensional evolution algebra $E^{[s,t]}$ is baric if and only if $a_{11}^{[s,t]} \neq 0$, $a_{21}^{[s,t]} = 0$ or $a_{22}^{[s,t]} \neq 0$, $a_{12}^{[s,t]} = 0$. The assertions of the theorem are results of the meticulous checking of these conditions. □

## 4.2. Absolute nilpotent elements transition

Recall that the element $x$ of an algebra $A$ is called an *absolute nilpotent* if $x^2 = 0$.

Let $E = \mathbb{R}^n$ be an evolution algebra over the field $\mathbb{R}$ with structural constant coefficients matrix $\mathcal{M} = (a_{ij})$. Then for arbitrary $x = \sum_i x_i e_i$ and $y = \sum_i y_i e_i \in \mathbb{R}^n$, we have

$$xy = \sum_j \left( \sum_i a_{ij} x_i y_i \right) e_j, \qquad x^2 = \sum_j \left( \sum_i a_{ij} x_i^2 \right) e_j.$$

For an $n$-dimensional evolution algebra $\mathbb{R}^n$ consider the operator $V\colon \mathbb{R}^n \to \mathbb{R}^n$, $x \mapsto V(x) = x'$, defined as

$$x'_j = \sum_{i=1}^{n} a_{ij} x_i^2, \qquad j = 1, \dots, n. \tag{4.1}$$

This operator is called an *evolution operator* [16].

We have $V(x) = x^2$, hence the equation $V(x) = x^2 = 0$ is given by the following system

$$\sum_i a_{ij} x_i^2 = 0, \qquad j = 1, \dots, n. \tag{4.2}$$

In this section, we shall solve system (4.2) for $E_i^{[s,t]}$, $i = 0, \dots, 8$.

For a CEA $E_i^{[s,t]}$ with matrix $\mathcal{M}_i^{[s,t]}$ denote

$$\mathcal{T}_{nil}^{(i)} = \{(s,t) \in \mathcal{T} : E_i^{[s,t]} \text{ has a unique absolute nilpotent}\}, \quad \mathcal{T}_{nil}^{0} = \mathcal{T} \setminus \mathcal{T}_{nil}.$$

The following theorem answers the problem of the existence of "uniqueness of absolute nilpotent element" property transition.

**Theorem 4.5.**

(1) *There CEAs $E_i^{[s,t]}$, $i = 0, 3, 4, 5, 6, 7, 8$, have infinitely many of absolute nilpotent elements for any time $(s,t) \in \mathcal{T}$.*

(2) *The CEAs $E_i^{[s,t]}$, $i = 1, 2$, have "uniqueness of absolute nilpotent element" property transition with the property duration sets as the following*

$$\mathcal{T}_{nil}^{(1)} = \{(s,t) \in \mathcal{T} : \rho(s)\phi(s) > 0\},$$
$$\mathcal{T}_{nil}^{(2)} = \{(s,t) \in \mathcal{T} : s \le t < a,\ \sigma(s) > 0\}.$$

***Proof.*** The proof consists of the simple examination of the solutions of system (4.2) for each $E_i^{[s,t]}$, $i = 0, \dots, 8$. □

## 4.3. Idempotent elements transition

A element $x$ of an algebra $\mathcal{A}$ is called *idempotent* if $x^2 = x$. The idempotents of an evolution algebra are especially significant because they are the fixed points of the evolution operator $V$ (4.1), i.e. $V(x) = x$. We denote by $\mathcal{I}d(E)$ the set of idempotent elements of an algebra $E$. Using (4.1) the equation $x^2 = x$ can be written as

$$x_j = \sum_{i=1}^{n} a_{ij} x_i^2, \qquad j = 1, \dots, n. \tag{4.3}$$

The extensive analysis of the solutions of system (4.3) is very hard. We shall solve this problem for the CEAs $E_i^{[s,t]}$, $i = 0, \dots, 8$.

The following theorem provides the time-dynamics of the idempotent elements for the algebras $E_i^{[s,t]}$, $i = 0, \dots, 8$.

**Theorem 4.6.**

(1) *The algebras* $E_i^{[s,t]}$, $i = 0, 5, 6, 7, 8$, *have a unique idempotent* $(0, 0)$ *in any time* $(s, t) \in \mathfrak{T}$.

(2) *The algebra* $E_1^{[s,t]}$ *has two idempotents* $(0, 0)$, $(0, \frac{\phi(s)}{\phi(t)})$ *for all* $(s, t) \in \{(s, t) : s \leq t < a\}$.

(3) *The algebra* $E_2^{[s,t]}$ *has two idempotents* $(0, 0)$, $(0, 1)$ *in any time* $(s, t) \in \mathfrak{T}$.

(4) *The algebra* $E_3^{[s,t]}$ *has two idempotents* $(0, 0)$, $(\frac{f(t)\phi(s)}{\phi^2(t)}, \frac{\phi(s)}{\phi(t)})$ *in any time* $(s, t) \in \mathfrak{T}$.

(5) *The algebra* $E_4^{[s,t]}$ *has two idempotents* $(0, 0)$, $(g(t), 1)$ *for all* $(s, t) \in \{(s, t) : s \leq t < a\}$.

***Proof.*** The proof contains a precise analysis of the solutions of system (4.3) for each $E_i^{[s,t]}$, $i = 0, \ldots, 8$. □

**Acknowledgment.** We thank the referee for the helpful comments and suggestions that contributed to the improvement of this paper. We sincerely acknowledge Professor U.A. Rozikov for helpful discussions. This work was partially supported by Agencia Estatal de Investigación (Spain), grant MTM2016-79661-P and by Xunta de Galicia, grant ED431C 2019/10 (European FEDER support included, UE).

## References

[1] V.M. Abraham, *Linearizing quadratic transformations in genetic algebras*, Proc. London Math. Soc. (3), **40** (2), 346–363, 1980.

[2] Y. Cabrera Casado, *Evolution algebras*, Ph.D. thesis, Universidad de Málaga, 2016, http://hdl.handle.net/10630/14175.

[3] M.I. Cardoso Gonçalves, D. Gonçalves, D. Martín Barquero, C. Martín González and M. Siles Molina, *Squares and Associative Representations of two Dimensional Evolution Algebras*, J. Algebra Appl., 2020, doi: https://doi.org/10.1142/S0219498821500900.

[4] J.M. Casas, M. Ladra and U.A. Rozikov, *A chain of evolution algebras*, Linear Algebra Appl. **435** (4), 852–870, 2011.

[5] J.M. Casas, M. Ladra, B.A. Omirov and U.A. Rozikov, *On evolution algebras*, Algebra Colloq. **21** (2), 331–342, 2014.

[6] I.M.H. Etherington, *Genetic algebras*, Proc. Roy. Soc. Edinburgh, **59**, 242–258, 1939.

[7] I.M.H. Etherington, *Duplication of linear algebras*, Proc. Edinburgh Math. Soc. (2), **6**, 222–230, 1941.

[8] I.M.H. Etherington, *Non-associative algebra and the symbolism of genetics*, Proc. Roy. Soc. Edinburgh. Sect. B. **61**, 24–42, 1941.

[9] O.J. Falcón, R.M. Falcón and J. Núñez, *Classification of asexual diploid organisms by means of strongly isotopic evolution algebras defined over any field*, J. Algebra, **472**, 573–593, 2017.

[10] V. Glivenkov, *Algèbre Mendelienne*, C. R. (Doklady) Acad. Sci. URSS, **4**, 385–386, 1936.

[11] H. Gonshor, *Contributions to genetic algebras. II*, Proc. Edinburgh Math. Soc. (2), **18**, 273–279, 1973.

[12] I. Heuch, *Sequences in genetic algebras for overlapping generations*, Proc. Edinburgh Math. Soc. (2), **18**, 19–29, 1972.

[13] P. Holgate, *Sequences of powers in genetic algebras*, J. London Math. Soc. **42**, 489–496, 1967.

[14] P. Holgate, *Selfing in genetic algebras*, J. Math. Biology, **6**, 197–206, 1978.

[15] V.A. Kostitzin, *Sur les coefficients mendéliens d'hérédité*, C. R. Acad. Sci. Paris, **206**, 883–885, 1938.

[16] Y.I. Lyubich, *Mathematical structures in population genetics*, Springer-Verlag, Berlin, 1992.

[17] G. Mendel, *Experiments in plant-hybridization*, 1865. The Electronic Scholarly Publishing Project http://www.esp.org/foundations/genetics/classical/gm-65.pdf.

[18] Sh.N. Murodov, *Classification dynamics of two-dimensional chains of evolution algebras*, Internat. J. Math. **25** (2), 1450012, 23 pp., 2014.

[19] Sh.N. Murodov, *Classification of two-dimensional real evolution algebras and dynamics of some two-dimensional chains of evolution algebras*, Uzbek. Mat. Zh. **2014** (2), 102–111, 2014.

[20] M.L. Reed, *Algebraic structure of genetic inheritance*, Bull. Amer. Math. Soc. (N.S.), **34** (2), 107–130, 1997.

[21] O. Reiersöl, *Genetic algebras studied recursively and by means of differential operators*, Math. Scand. **10**, 25–44, 1962.

[22] U.A. Rozikov and Sh.N. Murodov, *Dynamics of two-dimensional evolution algebras*, Lobachevskii J. Math. **34** (4), 344–358, 2013.

[23] U.A. Rozikov and J.P. Tian, *Evolution algebras generated by Gibbs measures*, Lobachevskii J. Math. **32** (4), 270–277, 2011.

[24] R.D. Schafer, *An introduction to nonassociative algebras*, Academic Press, New York, 1966.

[25] A. Serebrowsky, *On the properties of the Mendelian equations*, C. R. (Doklady) Acad. Sci. URSS **2**, 33–39, 1934 (in Russian).

[26] J.P. Tian, *Evolution algebras and their applications*, Lecture Notes in Mathematics 1921, Springer-Verlag, Berlin, 2008.

[27] J.P. Tian and P. Vojtechovsky, *Mathematical concepts of evolution algebras in non-Mendelian genetics*, Quasigroups Related Systems **14**, 111–122, 2006.

[28] A. Wörz-Busekros, *Algebras in genetics*, Lecture Notes in Biomathematics 36, Springer-Verlag, Berlin-New York, 1980.

Hacettepe Journal of Mathematics & Statistics

Hacet. J. Math. Stat.
Volume 50 (1) (2021), 159 – 179
DOI : 10.15672/hujms.610834

Research Article

# Numerical investigation of dynamic Euler-Bernoulli equation via 3-Scale Haar wavelet collocation method

Ömer Oruç*[1], Alaattin Esen[2], Fatih Bulut[3]

[1] *Eğil Vocational and Technical Anatolian High School, Diyarbakır, Turkey*
[2] *İnönü University, Department of Mathematics, Malatya, Turkey*
[3] *İnönü University, Department of Physics, Malatya, Turkey*

## Abstract

In this study, we analyze the performance of a numerical scheme based on 3-scale Haar wavelets for dynamic Euler-Bernoulli equation, which is a fourth order time dependent partial differential equation. This type of equations governs the behaviour of a vibrating beam and have many applications in elasticity. For its solution, we first rewrite the fourth order time dependent partial differential equation as a system of partial differential equations by introducing a new variable, and then use finite difference approximations to discretize in time, as well as 3-scale Haar wavelets to discretize in space. By doing so, we obtain a system of algebraic equations whose solution gives wavelet coefficients for constructing the numerical solution of the partial differential equation. To test the accuracy and reliability of the numerical scheme based on 3-scale Haar wavelets, we apply it to five test problems including variable and constant coefficient, as well as homogeneous and non-homogeneous partial differential equations. The obtained results are compared wherever possible with those from previous studies. Numerical results are tabulated and depicted graphically. In the applications of the proposed method, we achieve high accuracy even with small number of collocation points.

**Mathematics Subject Classification (2020).** 65M70, 65T99

**Keywords.** 3-Scale Haar wavelets, vibrating beam, dynamic Euler-Bernoulli equation

## 1. Introduction

The fourth-order problem considered in this paper is

$$\mu(x)\frac{\partial^2 u}{\partial t^2} + \mathrm{EI}(x)\frac{\partial^4 u}{\partial x^4} = F(x,t), \quad a \le x \le b, \quad 0 \le t \le T, \tag{1.1}$$

*Corresponding Author.
Email addresses: omeroruc0@gmail.com (Ö Oruç), alaattin.esen@inonu.edu.tr (A. Esen), fatih.bulut@inonu.edu.tr (F. Bulut)
Received: 26.08.2019; Accepted: 26.05.2020

subject to the initial conditions

$$u(x,0) = \xi(x),$$
$$u_t(x,0) = \eta(x), \quad a \le x \le b,$$

and the boundary conditions of the form

$$u(a,t) = f_1(t),\ u(b,t) = f_2(t),$$
$$u_{xx}(a,t) = f_3(t),\ u_{xx}(b,t) = f_4(t), \quad 0 \le t \le T.$$

Such problems occur in the study of the transverse displacements of a flexible beam hinged at both ends. Here $u = u(x,t)$ is the transverse displacement of the beam, $t$ and $x$ are time and spatial variables, $\mu(x) > 0$ is the density of the beam, $\mathrm{EI}(x) > 0$ is the beam bending stiffness and $F(x,t)$ is dynamic driving force per unit mass. Such an equation is also called dynamic Euler-Bernoulli equation, and its solution is important in many applications such as control of large flexible space structures or the development of robotics designs [3, 28, 41, 50].

The analytic solutions of variable coefficient nonhomogeneous Euler–Bernoulli equation are obtained by Wazwaz [52] using the Adomian decomposition method. Some exact solutions of variable coefficient homogeneous and nonhomogeneous Euler–Bernoulli equation are obtained by Adomian method in [14]. Analytical solutions of partial differential equations are very useful. However, it is not always possible to obtain the analytical solutions or it is possible only for limited initial and boundary conditions. So it is crucial to develop efficient numerical methods. For obtaining numerical solutions of Eq. (1.1), finite difference methods are employed in [1, 7–13, 20, 25, 47, 51]. A fully Sinc-Galerkin method is used in [49] by Smith et al. for solving fourth-order partial differential equations. A three level scheme based on parametric quintic spline is proposed by Aziz et al. [2] for the solution of fourth-order parabolic partial differential equations with constant coefficients. Khan et al. used sextic splines for solving a fourth-order parabolic partial differential equation in [26].

Caglar and Caglar [4] have developed a fifth degree B-spline method to obtain the numerical solution of constant coefficient fourth-order parabolic partial differential equations. Free vibration of an Euler–Bernoulli beam is obtained by Liu and Gurram [32] using He's variational iteration method. For variable coefficient fourth order parabolic partial differential equations a new three level implicit method based on sextic spline is proposed by Rashidinia and Mohammadi [46]. Mittal and Jain [36] used cubic and quintic B-spline method with redefined basis functions for obtaining numerical solutions of fourth-order parabolic partial differential equations with constant coefficients. Recently, Mohammadi [41] developed a numerical method based on sextic B-splines to solve the fourth-order time dependent partial differential equations subjected to fixed and cantilever boundary conditions.

Due to attractiveness of Haar wavelets for their simplicity, accuracy, computational cost, and so on, in recent years they have got much attention in numerical solutions of differential equations. A brief review of the literature can be given as follows. Chen and Hsiao[5] used Haar wavelet method for solving lumped and distributed parameter systems. In [6], they also discussed an optimal control problem. Hsiao and Wang [16, 17] used Haar wavelets for solving singular bilinear and nonlinear systems and [18] investigated nonlinear stiff systems. Hsiao [15] showed that the Haar wavelet approach is also effective for solving variational problems. Lepik applied this method to some well known problems [29–31]. Zhi Shi et al. [48] applied Haar wavelets to solve 2D and 3D Poisson equations and biharmonic equations.

Jiwari [21] used a hybrid numerical scheme based on implicit Euler method, quasilinearization and uniform Haar wavelets for the numerical solutions of Burgers' equation. Kaur et al. [24] solved Lane-Emden equations arising in astrophysics with Haar

Wavelets. Pandit et al. [45] solved second-order hyperbolic telegraph type equations by Haar wavelets. Majak et al. [33–35] studied functionally graded material (FGM) beams by means of Haar wavelet discretization method and convergence of Haar wavelet method. An efficient numerical scheme based on uniform Haar wavelets and the quasilinearization process is proposed for the numerical simulation of time dependent nonlinear Burgers' equation by Jiwari [22].

Oruç et al. [42–44] solved modified Burgers' equation, coupled Schrödinger-KdV equations and regularized long wave equation with the help of a Haar wavelet based method. Vibration analysis of nanobeams is investigated by Haar wavelets in [27]. A new type of solutions was obtained for the MHD Falkner–Skan boundary layer flow problem using the Haar wavelet quasilinearization approach via Lie symmetric analysis by Jiwari et. al. [23]. Mittal and Pandit [38] used Haar wavelet operational matrix along with quasi-linearization to detect the spin flow of fractional Bloch equations. Mittal and Pandit [40] developed a novel algorithm based on Scale-3 Haar wavelets and quasilinearization for numerical solution of a dynamical system of ordinary differential equations. Recently, Scale-3 Haar wavelet-based algorithm has been extended to find numerical approximations of second order initial and boundary value problems by Mittal and Pandit [39]. Most of the papers mentioned above are based on classical Haar wavelets (2-scale Haar wavelets).

In this study our aim is to analyze the performance of the 3-scale Haar wavelet collocation method (HWCM), recently introduced by Mittal and Pandit in their paper [37], for fourth order partial differential equations with variable and constant coefficients. As far as we know, the 3-scale Haar wavelets have not been employed to solve high order partial differential equations such as Euler-Bernoulli problems, which motivates us for conducting this study. This paper is organized as follows. In Section 2, 3-scale Haar wavelets and their integrals are introduced. In Section 3, a method based on discretization of time and space variables is described. Numerical results and discussion are given in Section 4. Finally, we summarize our findings in Section 5.

## 2. 3-Scale Haar wavelets and their integrals

The 3-scale Haar wavelets are constructed from two wavelet functions, namely symmetric and antisymmetric wavelet functions. This is the main difference with the 2-scale Haar wavelets, which employ only one wavelet function. The 3-scale Haar wavelets have advantages over the 2-scale ones: they converge rapidly, they can be represented by sparse matrices, in numerical applications solutions can be found at any point in the range, and they can easily detect singularity and discontinuity [37].

Using the orthogonality properties of 3-scale Haar wavelets, one can express any square integrable function $f(x)$ on the interval $[0,1)$ as an infinite series in the following form [37,39]:

$$f(x) \approx c_1\phi_1(x) + \sum_{\text{even index}\, i,\, i\geq 2}^{\infty} c_i\psi_i^{(1)}(x) + \sum_{\text{odd index}\, i,\, i\geq 3}^{\infty} c_i\psi_i^{(2)}(x). \tag{2.1}$$

Herein, $\phi_1$, $\psi_i^{(1)}$and $\psi_i^{(2)}$ are given by

$$\phi_1(x) = \begin{cases} 1 & a \leq x \leq b, \\ 0 & \text{elsewhere}, \end{cases} \tag{2.2}$$

$$\psi_i^{(1)}(x) = \frac{1}{\sqrt{2}} \begin{cases} -1 & \alpha(i) \leq x < \beta(i), \\ 2 & \beta(i) \leq x < \gamma(i), \\ -1 & \gamma(i) \leq x < \delta(i), \end{cases} \tag{2.3}$$

$$\psi_i^{(2)}(x) = \sqrt{\frac{3}{2}} \begin{cases} 1 & \alpha(i) \leq x < \beta(i), \\ 0 & \beta(i) \leq x < \gamma(i), \\ -1 & \gamma(i) \leq x < \delta(i), \end{cases} \tag{2.4}$$

and

$$\alpha(i) = a + (b-a)\frac{k}{m},$$

$$\beta(i) = a + (b-a)\frac{k+1/3}{m},$$

$$\gamma(i) = a + (b-a)\frac{k+2/3}{m},$$

$$\delta(i) = a + (b-a)\frac{k+1}{m},$$

where $m$ is defined as $3^j$ $(j = 0, 1, ...)$, and integer $k = 0, 1, .., m-1$ is the translation parameter. The index $i$ in $\alpha(i)$, $\beta(i)$, $\gamma(i)$ and $\delta(i)$ shows the relation between wavelet level $m$ and translation parameter $k$. If $i = 1$, then we get scaling function $\phi_1(x)$ which is defined in (2.2) and shown in Fig. 1 for $[a, b] = [0, 1]$. In case of $i > 1$, the index $i$ is calculated according to formulae $i = m + 2k$ or $i = m + 2k + 1$. If $i$ is even then consider $\psi_i^{(1)}$, if $i$ is odd then consider $\psi_i^{(2)}$. In Figs. 2 and 3, first wavelets $\psi_i^{(1)}$ and $\psi_i^{(2)}$ are plotted for $[a, b] = [0, 1]$.

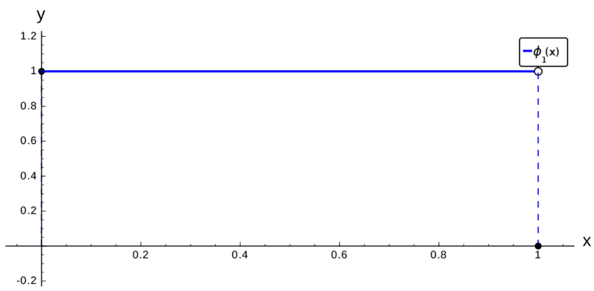



**Figure 1.** 3-scale Haar wavelet scaling function $\phi_1(x)$

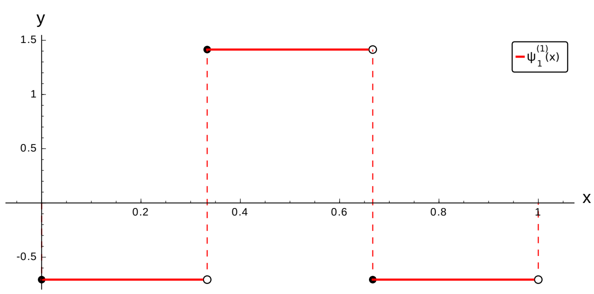



**Figure 2.** First symmetric wavelet $\psi_1^{(1)}(x)$

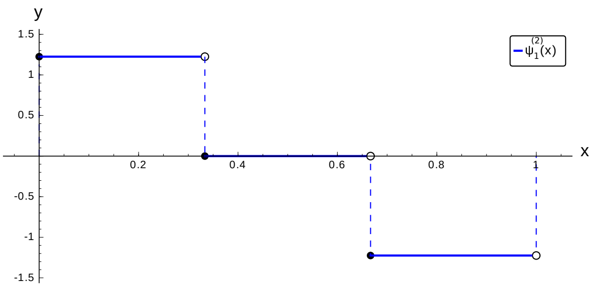



**Figure 3.** First anti-symmetric wavelet $\psi_1^{(2)}(x)$

Eq. (2.1) is an infinite series. We truncate this series to 3-scale Haar wavelets as [37]:

$$f(x) \approx c_1\phi_1(x) + \sum_{\text{even index } i,\, i\geq 2}^{3m} c_i\psi_i^{(1)}(x) + \sum_{\text{odd index } i,\, i\geq 3}^{3m} c_i\psi_i^{(2)}(x) = \boldsymbol{c}^T H_{3m}.$$

where $\boldsymbol{c}^T = [c_1, ..., c_{3m}]$ and $H_{3m} = [\phi_1(x), \psi_2^{(1)}(x), \psi_3^{(2)}(x), ..., \psi_{3m-1}^{(1)}(x), \psi_{3m}^{(2)}(x)]^T$ are in size of $1 \times 3m$.

In the solution process of a differential equation of any order, we need to integrate 3-scale Haar wavelets, that is we employ the integrals

$$\phi_{1,1}(x) = \int_0^x \phi_1(t)dt = \begin{cases} x & [a,b), \\ 0 & elsewhere, \end{cases}$$

$$\psi_{i,1}^{(1)}(x) = \int_0^x \psi_i^{(1)}(t)dt = \frac{1}{\sqrt{2}} \begin{cases} \alpha(i) - x & \alpha(i) \leq x < \beta(i), \\ 2x - 3\beta(i) + \alpha(i) & \beta(i) \leq x < \gamma(i), \\ \alpha(i) + 3\gamma(i) - 3\beta(i) - x & \gamma(i) \leq x < \delta(i), \end{cases}$$

$$\psi_{i,1}^{(2)}(x) = \int_0^x \psi_i^{(2)}(t)dt = \sqrt{\frac{3}{2}} \begin{cases} x - \alpha(i) & \alpha(i) \leq x < \beta(i), \\ \beta(i) - \alpha(i) & \beta(i) \leq x < \gamma(i), \\ \gamma(i) + \beta(i) - \alpha(i) - x & \gamma(i) \leq x < \delta(i). \end{cases}$$

Moreover, we introduce

$$\phi_{1,n+1}(x) = \int_0^x \phi_{1,n}(t)dt, \quad \psi_{1,n+1}^{(1)} = \int_0^x \psi_{1,n}^{(1)}(t)dt, \quad \psi_{1,n+1}^{(2)} = \int_0^x \psi_{1,n}^{(2)}(t)dt$$

which can explicitly be written as

$$\phi_{1,n+1}(x) = \begin{cases} \frac{x^{n+1}}{(n+1)!} & [a,b), \\ 0 & elsewhere, \end{cases}$$

$$\psi_{i,n+1}^{(1)}(x) = \frac{1}{\sqrt{2}} \begin{cases} \frac{-(x-\alpha(i))^{n+1}}{(n+1)!} & \alpha(i) \leq x < \beta(i), \\ \frac{3(x-\beta(i))^{n+1} - (x-\alpha(i))^{n+1}}{(n+1)!} & \beta(i) \leq x < \gamma(i), \\ \frac{3(x-\beta(i))^{n+1} - 3(x-\gamma(i))^{n+1} - (x-\alpha(i))^{n+1}}{(n+1)!} & \gamma(i) \leq x < \delta(i), \\ \frac{3(x-\beta(i))^{n+1} - 3(x-\gamma(i))^{n+1} - (x-\alpha(i))^{n+1} + (x-\delta(i))^{n+1}}{(n+1)!} & \delta(i) \leq x < 1, \end{cases}$$

$$\psi^{(2)}_{i,n+1}(x)=\sqrt{\frac{3}{2}}\begin{cases}\frac{(x-\alpha(i))^{n+1}}{(n+1)!} & \alpha(i)\le x<\beta(i),\\ \frac{(x-\alpha(i))^{n+1}-(x-\beta(i))^{n+1}}{(n+1)!} & \beta(i)\le x<\gamma(i),\\ \frac{(x-\alpha(i))^{n+1}-(x-\beta(i))^{n+1}-(x-\gamma(i))^{n+1}}{(n+1)!} & \gamma(i)\le x<\delta(i),\\ \frac{(x-\alpha(i))^{n+1}-(x-\beta(i))^{n+1}-(x-\gamma(i))^{n+1}+(x-\delta(i))^{n+1}}{(n+1)!} & \delta(i)\le x<1.\end{cases}$$

## 3. Discretization scheme for fourth order partial differential equations

To solve Eq. (1.1) we introduce a new variable, namely

$$v=\frac{\partial u}{\partial t}.$$

Now Eq. (1.1) can be rewritten as the system of partial differential equations that is first order in time given below.

$$\begin{aligned}u_t-v&=0,\\ \mu(x)v_t+\mathrm{EI}(x)u_{xxxx}&=F(x,t).\end{aligned}\tag{3.1}$$

We describe the discretization process of the equations above in the subsequent sections.

### 3.1. Time discretization

We use explicit finite difference schemes for time derivatives, as well as the time average for $v$ and $u_{xxxx}$ in Eq. (3.1). By doing so, we get

$$\begin{aligned}\frac{u^{j+1}-u^j}{\Delta t}-\frac{v^{j+1}+v^j}{2}&=0,\\ \mu(x)\frac{v^{j+1}-v^j}{\Delta t}+\mathrm{EI}(x)\frac{u^{j+1}_{xxxx}+u^j_{xxxx}}{2}&=F(x,t^{j+1}).\end{aligned}$$

The equations above can be rearranged as

$$\begin{aligned}u^{j+1}-\frac{\Delta t}{2}v^{j+1}&=u^j+\frac{\Delta t}{2}v^j,\\ \mu(x)v^{j+1}+\frac{\Delta t\mathrm{EI}(x)}{2}u^{j+1}_{xxxx}&=\mu(x)v^j-\frac{\Delta t\cdot\mathrm{EI}(x)}{2}u^j_{xxxx}+\Delta tF(x,t^{j+1}),\end{aligned}\tag{3.2}$$

with initial conditions

$$\begin{aligned}&u^0(x)=\xi(x),\\ &v^0(x)=\eta(x),\quad a\le x\le b\end{aligned}\tag{3.3}$$

and with the boundary conditions

$$\begin{aligned}&u^{j+1}(a)=f_1(t^{j+1}),\ u^{j+1}(b)=f_2(t^{j+1}),\\ &u^{j+1}_{xx}(a)=f_3(t^{j+1}),\ u^{j+1}_{xx}(b)=f_4(t^{j+1}),\end{aligned}\tag{3.4}$$

where $u^{j+1}$ and $v^{j+1}$ are the solutions of Eq. (3.2) at the $(j+1)$th time step and $t^{j+1}=\Delta t(j+1)$, $j=0,1,...,N-1$, $\Delta t\cdot N=T$.

### 3.2. Space discretization by Haar wavelets

Since Haar wavelets are generally defined for $[0,1]$. We have to transform the domain into unit interval. By introducing $y=(x-a)/L$ , $L=b-a$, the interval $a \le x \le b$ can be transformed into the unit interval $0 \le y \le 1$. Using this transformation, we can reduce a problem defined on $[a,b]$ to a problem defined on $[0,1]$. Hence, without loss of generality, the PDE we have at hand is defined over $[0,1]$ in space.

For the description of space discretization, we introduce notations

$$\sum_{i=1}^{3m} c_i h_i(x) := c_1\phi_1(x) + \sum_{\text{even index}\, i,\, i\ge 2}^{3m} c_i\psi_i^{(1)}(x) + \sum_{\text{odd index}\, i,\, i\ge 3}^{3m} c_i\psi_i^{(2)}(x)$$

$$\sum_{i=1}^{3m} c_i p_{i,j}(x) := c_1\phi_{1,j}(x) + \sum_{\text{even index}\, i,\, i\ge 2}^{3m} c_i\psi_{i,j}^{(1)}(x) + \sum_{\text{odd index}\, i,\, i\ge 3}^{3m} c_i\psi_{i,j}^{(2)}(x)$$

for $j=1,2,3,4$. Now we expand $u_{xxxx}^{j+1}(x)$ term in (3.2) into Haar wavelets, that is

$$u_{xxxx}^{j+1}(x) = \sum_{i=1}^{3m} c_i h_i(x). \tag{3.5}$$

By integrating the equation above from 0 to $x$, we get

$$u_{xxx}^{j+1}(x) = u_{xxx}^{j+1}(0) + \sum_{i=1}^{3m} c_i p_{i,1}(x) \tag{3.6}$$

We do not know the value of $u_{xxx}^{j+1}(0)$ term in Eq. (3.6), but we can calculate it by integrating Eq. (3.6) from 0 to 1 and using boundary conditions from Eq. (3.4) as follows:

$$u_{xxx}^{j+1}(0) = f_4(t^{j+1}) - f_3(t^{j+1}) - \sum_{i=1}^{3m} c_i p_{i,2}(1).$$

Now by integrating Eq. (3.6) from 0 to $x$ we obtain the second derivative $u_{xx}^{j+1}(x)$ as

$$u_{xx}^{j+1}(x) = \sum_{i=1}^{3m} c_i p_{i,2}(x) + f_3(t^{j+1}) + \left[f_4(t^{j+1}) - f_3(t^{j+1})\right] x - x\sum_{i=1}^{3m} c_i p_{i,2}(1). \tag{3.7}$$

By integrating Eq. (3.7) once again from 0 to $x$, we deduce

$$\begin{aligned} u_x^{j+1}(x) - u_x^{j+1}(0) = &\sum_{i=1}^{3m} c_i p_{i,3}(x) + x f_3(t^{j+1}) \\ &+ \left[f_4(t^{j+1}) - f_3(t^{j+1})\right]\frac{x^2}{2} - \frac{x^2}{2}\sum_{i=1}^{3m} c_i p_{i,2}(1), \end{aligned} \tag{3.8}$$

which we integrate again from 0 to 1 to obtain

$$\begin{aligned} u^{j+1}(1) - u^{j+1}(0) - u_x^{j+1}(0) = &\sum_{i=1}^{3m} c_i p_{i,4}(1) + \frac{1}{2} f_3(t^{j+1}) \\ &+ \left[f_4(t^{j+1}) - f_3(t^{j+1})\right]\frac{1}{6} - \frac{1}{6}\sum_{i=1}^{3m} c_i p_{i,2}(1). \end{aligned} \tag{3.9}$$

By exploiting the boundary conditions $u^{j+1}(1) = f_2(t^{j+1})$ and $u^{j+1}(0) = f_1(t^{j+1})$ in the equation above, we retrieve

$$\begin{aligned} u_x^{j+1}(0) =& f_2(t^{j+1}) - f_1(t^{j+1}) - \sum_{i=1}^{3m} c_i p_{i,4}(1) \\ & - \frac{1}{2} f_3(t^{j+1}) - \left[f_4(t^{j+1}) - f_3(t^{j+1})\right] \frac{1}{6} + \frac{1}{6} \sum_{i=1}^{3m} c_i p_{i,2}(1). \end{aligned}$$

Plugging the right-hand side of the equation above for $u_x^{j+1}(0)$ in Eq.(3.8), we have

$$\begin{aligned} u_x^{j+1}(x) =& \sum_{i=1}^{3m} c_i p_{i,3}(x) + f_2(t^{j+1}) - f_1(t^{j+1}) - \frac{1}{3} f_3(t^{j+1}) \\ & - \frac{1}{6} f_4(t^{j+1}) - \sum_{i=1}^{3m} c_i \left[p_{i,4}(1) - \frac{1}{6} p_{i,2}(1)\right] \qquad (3.10) \\ & + f_3(t^{j+1})x + \frac{x^2}{2} \left[f_4(t^{j+1}) - f_3(t^{j+1})\right] - \frac{x^2}{2} \sum_{i=1}^{3m} c_i p_{i,2}(1), \qquad (3.11) \end{aligned}$$

which in turn yields

$$\begin{aligned} u^{j+1}(x) =& \sum_{i=1}^{3m} c_i p_{i,4}(x) + f_1(t^{j+1}) + \left[f_2(t^{j+1}) - f_1(t^{j+1}) - \frac{1}{3} f_3(t^{j+1}) - \frac{1}{6} f_4(t^{j+1})\right] x \\ & - x \sum_{i=1}^{3m} c_i \left[p_{i,4}(1) - \frac{1}{6} p_{i,2}(1)\right] \\ & + f_3(t^{j+1}) \frac{x^2}{2} + \frac{x^3}{6} \left[f_4(t^{j+1}) - f_3(t^{j+1})\right] - \frac{x^3}{6} \sum_{i=1}^{3m} c_i p_{i,2}(1). \qquad (3.12) \end{aligned}$$

Additionally we express $v^{j+1}(x)$ in terms of Haar wavelets in the form

$$v^{j+1}(x) = \sum_{i=1}^{3m} d_i h_i(x). \qquad (3.13)$$

By plugging Eqs. (3.5), (3.12) and (3.13) into Eq. (3.2) and discretizing at collocation points $x_l = \frac{l-0.5}{3m}$, $l = 1, 2, ..., 3m$ yields a system of linear equations whose solution gives the wavelet coefficients $c_i$ and $d_i$. Then by plugging these wavelet coefficients into Eqs. (3.12) and (3.13) we can obtain the numerical solutions $u^{j+1}(x)$ and $v^{j+1}(x)$.

## 3.3. Convergence analysis of Haar wavelets

Let

$$u(x) = c_1 \phi_1(x) + \sum_{\text{even index}\, i,\, i \geq 2}^{\infty} c_i \psi_i^{(1)}(x) + \sum_{\text{odd index}\, i,\, i \geq 3}^{\infty} c_i \psi_i^{(2)}(x)$$

and

$$u_{3m}(x) = c_1 \phi_1(x) + \sum_{\text{even index}\, i,\, i \geq 2}^{3m} c_i \psi_i^{(1)}(x) + \sum_{\text{odd index}\, i,\, i \geq 3}^{3m} c_i \psi_i^{(2)}(x)$$

be exact and numerical solutions of Eq. (1.1) with $a = 0$ and $b = 1$. Furthermore, $E_J = u(x) - u_{3m}(x)$ with $J = 3m$ and $\left\|u(x)\right\| = \left(\int_0^1 |u(x)|^2 dx\right)^{1/2}$.

**Theorem 3.1.** [37] *Let the exact solution $u(x)$ be square integrable on $[0,1]$ with bounded derivatives on $(0,1)$. Then the error $E_J$ satisfies*

$$\|E_J\| \leq \frac{M}{\sqrt{24}}\frac{1}{3^J}$$

*for some constant $M$ independent of $J$.*

***Proof.*** See [37]. □

Theorem 3.1 implies that the error bound is inverse proportional to the level of resolution of scale-3 Haar wavelets. Therefore the error decreases as we increase $J$.

## 4. Numerical examples

Numerical computations have been done with python programming language and graphical outputs were generated by Matplotlib package [19].

In problem 1, we calculate the maximal absolute relative errors which are defined as follows:

$$E = \max_{i=1,\dots,3m}\left|\frac{u_i^{\text{exact}} - u_i^{\text{num}}}{u_i^{\text{exact}}}\right|.$$

In problems 2, 3, 4 and 5, for the sake of comparison with earlier studies, we calculate the absolute errors $|u(x) - u^{\text{num}}(x)|$ at the points $x = 0.1,\ 0.2,\ 0.3,\ 0.4,\ 0.5$, where $u(x)$ and $u^{\text{num}}(x)$ denote the exact and numerical solutions at $x$. Here we should note that, $u_i^{\text{exact}}$ and $u_i^{\text{num}}$ denote exact and numerical solutions at collocation points $x_i$ at a certain final time $t$. Since in the solution process we took the collocation points as $x_i = \frac{i-0.5}{3m}$, $i = 1, 2, \dots, 3m$, for calculating numerical results at the points $x = 0.1,\ 0.2,\ 0.3,\ 0.4,\ 0.5$ we have used interpolation techniques.

Also for every problem, at the bottom of the tables, we provide the error norm $L_\infty$ which is defined by

$$L_\infty(u,.) = \max_i \left|u_i^{\text{exact}} - u_i^{\text{num}}\right|,\ i = 1, 2, \dots, 3m.$$

Convergence rates are calculated according to the formula

$$\text{Rate} = \frac{\log\left(\frac{L_\infty(u,3\Delta x)}{L_\infty(u,\Delta x)}\right)}{\log\left(\frac{3\Delta x}{\Delta x}\right)} \tag{4.1}$$

where $\Delta x = \frac{1}{3m}$ is the step size of spatial variable $x$.

### 4.1. Problem 1

We consider

$$120x\frac{\partial^2 u}{\partial t^2} + \left(120 + x^5\right)\frac{\partial^4 u}{\partial x^4} = 0$$

subject to the initial conditions

$$u(x,0) = 0, \quad u_t(x,0) = 1 + \frac{x^5}{120}, \quad \frac{1}{2} \leq x \leq 1$$

and with the boundary conditions at $x = 1/2$ and $x = 1$ of the form

$$u(\frac{1}{2},t) = \frac{3841}{3840}\sin t, \quad u(1,t) = \frac{121}{120}\sin t,$$
$$u_{xx}(\frac{1}{2},t) = \frac{1}{48}\sin t, \quad u_{xx}(1,t) = \frac{1}{6}\sin t, \quad t \geq 0.$$

This equation is also studied by [46], [1] and [25]. The exact solution of this problem is

$$u(x,t) = \left(1 + \frac{x^5}{120}\right)\sin t.$$

In Table 1, to see convergence in time variable we set $3m = 27$ and compute the errors at $t = 0.01$ for decreasing values of $\Delta t$. From Table 1, it is obvious that as the values of $\Delta t$ are diminished, the error also decreases. Also to see convergence in space variable we fix $\Delta t = 0.00025$ and compute the errors at $t = 0.01$ for increasing values of collocation points in Table 2. It is clearly seen from Table 2 that the errors get smaller by increasing the number of collocation points. Using various values of $\Delta t$ and $t = 0.01$ we compared the maximum absolute relative errors of the present method with the results from existing methods in the literature in Table 3. We choose the number of collocation points as $3m = 9$ for the present method for comparison. Table 3 shows that the obtained results from the present method, are more accurate in comparison to the sextic spline method [46], A.D.I methods [1] and difference scheme method [25] for this problem. Numerical and exact solutions are plotted for $3m = 9$, $\Delta t = 0.0025$ at $t = 1$ in Fig. 4.

**Table 1.** Maximum absolute relative errors for different values of $\Delta t$ and $3m = 27$ at $t = 0.01$ for Problem 1

| | $\Delta t$ | $E$ |
|---|---|---|
| $3m = 27$ | 0.001 | 4.5780e-09 |
| | 0.0005 | 1.2246e-09 |
| | 0.00025 | 2.8163e-10 |
| | 0.000125 | 6.7723e-11 |
| | 6.25e-05 | 1.9763e-11 |
| | 3.125e-05 | 3.5833e-13 |

**Table 2.** Maximum absolute relative errors for different values of $3m$ and $\Delta t = 0.00025$ at $t = 0.01$ for Problem 1

| | $3m$ | $E$ |
|---|---|---|
| $\Delta t = 0.00025$ | 3 | 3.4277e-09 |
| | 9 | 8.8387e-10 |
| | 27 | 2.8162e-10 |
| | 81 | 9.4109e-11 |
| | 243 | 3.1088e-11 |
| | 729 | 1.0436e-11 |

**Table 3.** Maximum absolute relative errors at $t = 0.01$ in Problem 1

| | | Methods | | | |
|---|---|---|---|---|---|
| | | HWCM | Rashidinia and Mohammadi [46] | Andrade and Mckee [1] | Khaliq and Twizell [25] |
| Parameters | | $3m = 9$ | $h = 0.05$ | $h = 0.05$ | $h = 0.05$ |
| $E$ | $\Delta t = 0.000625$ | 5.8883e-009 | 3.51e-08 | 4.10e-07 | 3.30e-07 |
| | $\Delta t = 0.00025$ | 8.8387e-010 | 9.97e-08 | 7.20e-07 | 3.30e-07 |
| | $\Delta t = 0.000125$ | 2.2098e-010 | 5.33e-08 | 1.90e-06 | 3.30e-07 |

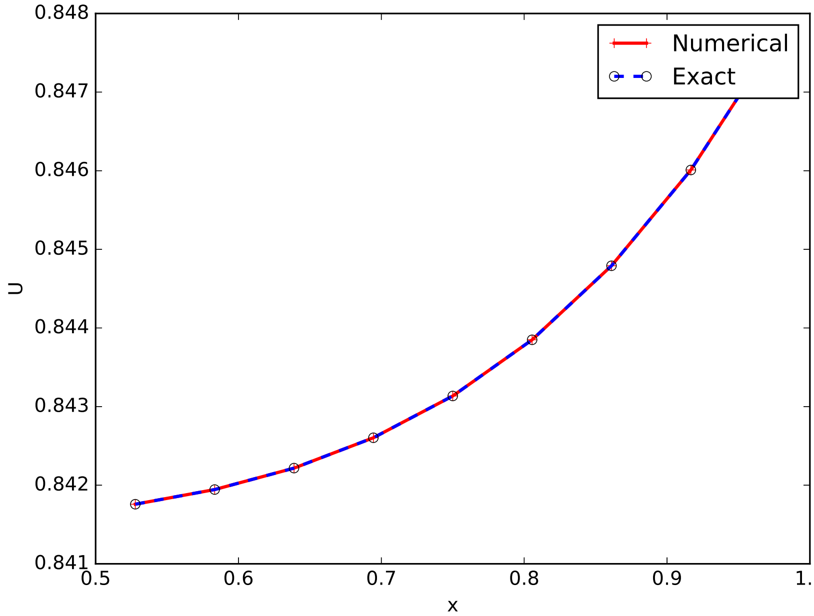



**Figure 4.** Exact solution versus numerical solution for $3m = 9$, $\Delta t = 0.0025$ at $t = 1$ in Problem 1

## 4.2. Problem 2

We consider

$$\sin x \frac{\partial^2 u}{\partial t^2} + (x - \sin x) \frac{\partial^4 u}{\partial x^4} = 0$$

subject to the initial conditions

$$u(x,0) = x - \sin x, \quad u_t(x,0) = -(x - \sin x), \quad 0 \leq x \leq 1$$

and with the boundary conditions

$$u(0,t) = 0, \quad u(1,t) = e^{-t}(1 - \sin 1),$$
$$u_{xx}(0,t) = 0, \quad u_{xx}(1,t) = e^{-t} \sin 1, \qquad t \geq 0.$$

This problem is also also studied in [46]. The exact solution for this problem is

$$u(x,t) = (x - \sin x) e^{-t}.$$

We solve the problem for $3m = 27$ and $\Delta t = 0.05$ with 10 and 16 time steps. We compared the approximate solutions obtained by the present method with exact solutions and tabulated the absolute errors for the present method and for the sextic spline method by Rashidinia and Mohammadi [46] at the points $x = 0.1,\ 0.2,\ 0.3,\ 0.4,\ 0.5$ and at times $t = 0.5$ and $t = 0.8$ in Table 4. It can be seen from the Table 4 that the present method gives more accurate results in comparison to [46] for all points. We plot the error with respect to $\Delta t$ in Fig. 5 for $3m = 27$ at $t = 1$. Also a plot of the error with respect to the number of collocation points is given in Fig. 6 for $\Delta t = 0.0025$ at $t = 1$. From Figs. 5-6 we can deduce that, for fixed $3m$, lowering the value of $\Delta t$ also reduces the error, and, for fixed $\Delta t$, increasing $3m$ decreases the error. Finally graphical representation of the exact solution and numerical solution are illustrated in Fig. 7 for $3m = 27$, $\Delta t = 0.005$ at $t = 0.08$. In Table 5 we tabulated the convergence rates in view of the errors calculated according to Eq. (4.1).

**Table 4.** $L_\infty$ and Absolute errors for Problem 2

| Methods | Time Steps | Parameters | $x = 0.1$ | $x = 0.2$ | $x = 0.3$ | $x = 0.4$ | $x = 0.5$ |
|---|---|---|---|---|---|---|---|
| HWCM | 10 | $3m = 27$ | 6.17e-11 | 3.55e-11 | 1.12e-09 | 8.03e-10 | 2.81e-10 |
| HWCM | 16 | $3m = 27$ | 5.77e-11 | 1.41e-10 | 1.31e-09 | 1.85e-09 | 6.58e-10 |
| [46] | 10 | $h = 0.05$ | 8.35e-08 | 4.51e-08 | 8.25e-08 | 2.33e-08 | 4.52e-08 |
| [46] | 16 | $h = 0.05$ | 8.42e-08 | 2.62e-08 | 5.32e-08 | 1.45e-08 | 2.89e-08 |
| HWCM | 10 | $3m = 27$ | $L_\infty = 3.0466e-09$ | | | | |
| HWCM | 16 | $3m = 27$ | $L_\infty = 4.2367e-09$ | | | | |

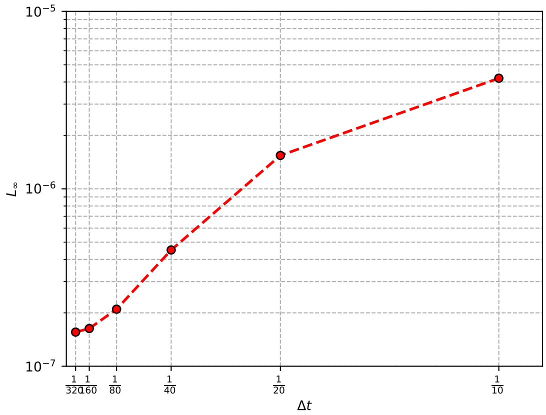



**Figure 5.** Error versus $\Delta t$ for $3m = 27$ at $t = 1$ in Problem 2

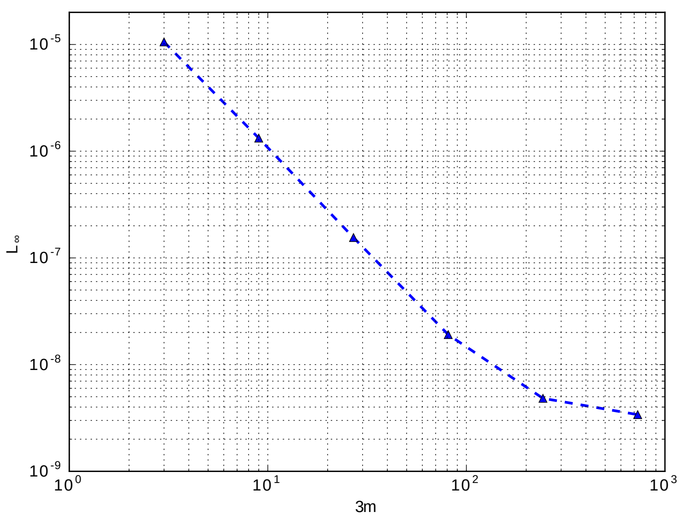



**Figure 6.** Error versus collocation points for $\Delta t = 0.0025$ at $t = 1$ in Problem 2

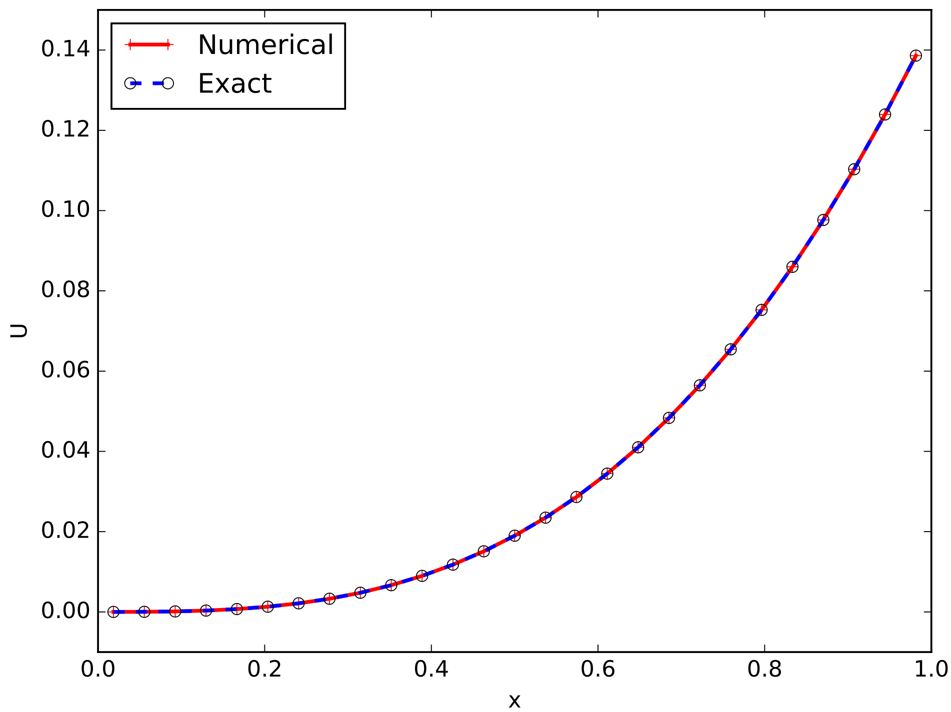



**Figure 7.** Exact solution versus numerical solution for $3m = 27$, $\Delta t = 0.005$ at $t = 0.08$ in Problem 2

**Table 5.** Convergence rates for $\Delta t = 0.005$ at time $t = 1$ in Problem 2

| | $L_\infty$ | Rate |
|---|---|---|
| $3m = 3$ | 1.0535e-05 | - |
| $3m = 9$ | 1.3257e-06 | 1.887 |
| $3m = 27$ | 1.5933e-07 | 1.928 |
| $3m = 81$ | 2.7831e-08 | 1.588 |

## 4.3. Problem 3

We consider a constant coefficient ($\mu(x) = \mathrm{EI}(x) = 1$) fourth order non-homogeneous parabolic partial differential equation given by

$$\frac{\partial^2 u}{\partial t^2} + \frac{\partial^4 u}{\partial x^4} = \left(\pi^4 - 1\right) \sin(\pi x) \cos t$$

subject to the initial conditions

$$u(x, 0) = \sin(\pi x), \quad u_t(x, 0) = 0, \quad 0 \leq x \leq 1$$

and with the boundary conditions

$$u(0, t) = u(1, t) = u_{xx}(0, t) = u_{xx}(1, t) = 0, \quad t \geq 0.$$

The exact solution for this problem is [12]

$$u(x, t) = \sin(\pi x) \cos t.$$

In Table 6, we give absolute errors at the points $x = 0.1, 0.2, 0.3, 0.4, 0.5$ using $3m = 27, 81$ and $\Delta t = 0.00125, 0.005$ at $t = 0.02, 0.05$. Also we give results from the previous studies for comparison. It can be seen from Table 6 that the present method gives more accurate results than AGE method [12], Fifth degree B-spline method [4], B-spline methods with redefined basis functions [36] and gives comparable results with other methods studied in [2, 26, 41, 46]. Note that $n$ stands for the number of collocation points in Table 6. Figure 8 shows the evolution of numerical solution in time during simulation for $3m = 81$ and $\Delta t = 0.05$.

**Table 6.** $L_\infty$ and Absolute errors for Problem 3

| Methods | Time | Parameters | $x = 0.1$ | $x = 0.2$ | $x = 0.3$ | $x = 0.4$ | $x = 0.5$ |
|---|---|---|---|---|---|---|---|
| HWCM | $t = 0.02$ | $3m = 81,\ \Delta t = 0.00125$ | 3.80e-07 | 7.22e-07 | 9.92e-07 | 1.16e-06 | 1.22e-06 |
| | $t = 0.05$ | $3m = 81,\ \Delta t = 0.005$ | 3.63e-06 | 6.91e-06 | 9.51e-06 | 1.12e-05 | 1.18e-05 |
| | $t = 0.02$ | $3m = 27,\ \Delta t = 0.00125$ | 3.23e-06 | 6.13e-05 | 8.75e-06 | 1.02e-05 | 1.04e-05 |
| | $t = 0.05$ | $3m = 27,\ \Delta t = 0.005$ | 2.04e-05 | 3.88e-05 | 5.37e-05 | 6.31e-05 | 6.60e-05 |
| Evans and Yousif [12] | $t = 0.02$ | $h = 0.05,\ \Delta t = 0.00125$ | 2.50e- 05 | 4.70e- 05 | 6.60e- 05 | 7.80e- 05 | 8.20e- 05 |
| | $t = 0.05$ | $h = 0.05,\ \Delta t = 0.005$ | 2.20e- 04 | 4.10e- 04 | 5.40e- 04 | 6.20e- 04 | 6.50e- 04 |
| Caglar and Caglar [4] | $t = 0.02$ | $n = 121,\ \Delta t = 0.005$ | 4.80e-06 | 9.70e-06 | 1.40e-05 | 1.90e-05 | 2.40e-05 |
| | $t = 0.02$ | $n = 191,\ \Delta t = 0.005$ | 5.20e-06 | 2.10e-06 | 3.10e-06 | 4.20e-06 | 5.20e-06 |
| Mittal and Jain [36] Method 1 | $t = 0.02$ | $n = 181,\ \Delta t = 0.005$ | 8.00e-06 | 1.52e-05 | 2.09e-05 | 2.46e-05 | 2.59e-05 |
| | $t = 0.05$ | $n = 181,\ \Delta t = 0.005$ | 8.97e-06 | 1.71e-05 | 2.35e-05 | 2.76e-05 | 2.90e-05 |
| Mittal and Jain [36] Method 2 | $t = 0.02$ | $n = 181,\ \Delta t = 0.005$ | 1.50e-07 | 2.90e-07 | 3.90e-07 | 4.60e-07 | 4.90e-07 |
| | $t = 0.05$ | $n = 181,\ \Delta t = 0.005$ | 1.10e-06 | 2.09e-06 | 2.88e-06 | 3.38e-06 | 3.56e-06 |
| Khan et al [26] | $t = 0.02$ | $h = 0.05,\ \Delta t = 0.00125$ | 9.07e-06 | 7.79e-06 | 2.75e-06 | 1.01e-06 | 2.59e-06 |
| | $t = 0.05$ | $h = 0.05,\ \Delta t = 0.005$ | 1.87e-06 | 2.13e-05 | 1.49e-05 | 8.60e-06 | 5.96e-06 |
| Rashidinia and Mohammadi [46] | $t = 0.02$ | $h = 0.05,\ \Delta t = 0.00125$ | 4.47e-07 | 2.66e-07 | 1.39e-07 | 1.55e-07 | 1.57e-07 |
| | $t = 0.05$ | $h = 0.05,\ \Delta t = 0.005$ | 2.91e-06 | 1.73e-06 | 1.60e-06 | 2.23e-06 | 2.60e-07 |
| Aziz et al. [2] | $t = 0.02$ | $h = 0.05,\ \Delta t = 0.00125$ | 9.20e-06 | 7.90e-06 | 2.80e-06 | 9.80e-07 | 2.50e-06 |
| | $t = 0.05$ | $h = 0.05,\ \Delta t = 0.005$ | 9.30e-06 | 8.00e-06 | 2.80e-06 | 1.00e-06 | 2.70e-06 |
| Mohammadi [41] | $t = 0.02$ | $h = 0.05,\ \Delta t = 0.00125$ | 4.29e-07 | 2.51e-07 | 1.24e-07 | 1.38e-07 | 1.40e-07 |
| | $t = 0.05$ | $h = 0.05,\ \Delta t = 0.005$ | 2.96e-06 | 1.77e-06 | 1.64e-06 | 2.28e-06 | 2.65e-07 |
| HWCM | $t = 0.02$ | $3m = 81,\ \Delta t = 0.00125$ | $L_\infty = 1.2239e-06$ | | | | |
| | $t = 0.05$ | $3m = 81,\ \Delta t = 0.005$ | $L_\infty = 1.1752e-05$ | | | | |

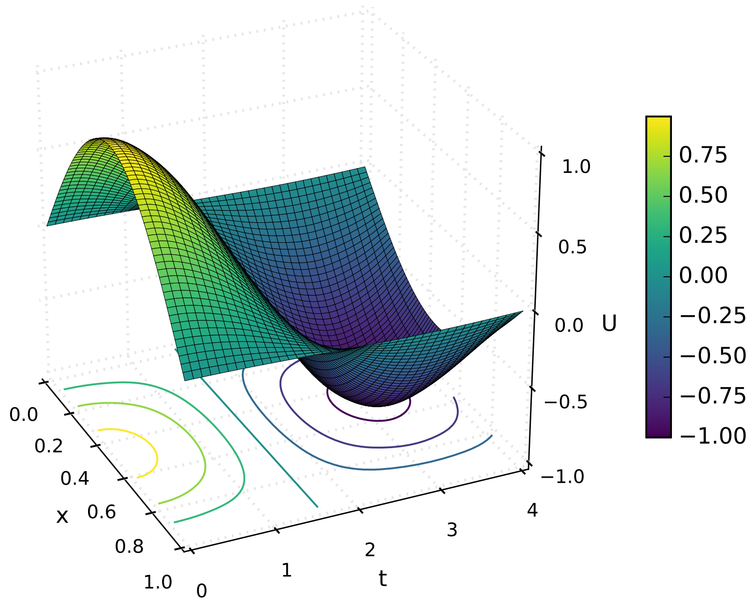



**Figure 8.** Evolution of numerical solution for $3m = 81$ and $\Delta t = 0.05$ from $t = 0$ to $t = 4$ in Problem 3

## 4.4. Problem 4

We consider a constant coefficient ($\mu(x) = \mathrm{EI}(x) = 1$) fourth order homogeneous parabolic partial differential equation given by

$$\frac{\partial^2 u}{\partial t^2} + \frac{\partial^4 u}{\partial x^4} = 0$$

subject to the initial conditions

$$u(x,0) = \frac{x}{12}\left(2x^2 - x^3 - 1\right), \quad u_t(x,0) = 0, \quad 0 \le x \le 1$$

and boundary conditions

$$u(0,t) = u(1,t) = u_{xx}(0,t) = u_{xx}(1,t) = 0, \quad t \ge 0.$$

The exact solution of this problem [11] is

$$u(x,t) = \sum_{s=1}^{\infty} a_s \sin(s\pi x)\cos(s^2\pi^2 t)$$

where

$$a_s = \frac{4}{s^5\pi^5}(\cos(s\pi) - 1).$$

For the sake of comparing our results with existing results, we choose the number of collocation points as $3m = 27$ and $3m = 81$. We observe from the Table 7 that for $3m = 27$ the present method gives more accurate results in comparison to existing methods except H.O.C.M. [13] at $t = 0.02$, and while at $t = 1$ the present method gives the best results among other methods. When we increase the number of collocation points to $3m = 81$, we see from the Table 7 that none of the existing methods can reach to the performance of the present method in terms of accuracy. In Fig. 9, evolution of numerical solution for $3m = 81$ and $\Delta t = 0.01$ from $t = 0$ to $t = 1$ is given. In Table 8 we tabulated the convergence rates in view of the errors calculated according to Eq. (4.1).

**Table 7.** $L_\infty$ and Absolute errors for Problem 4

| Methods | Time | Parameters | $x = 0.1$ | $x = 0.2$ | $x = 0.3$ | $x = 0.4$ | $x = 0.5$ |
|---|---|---|---|---|---|---|---|
| HWCM | $t = 0.02$ | $3m = 27,\ \Delta t = 0.00125$ | 3.33e-07 | 4.58e-07 | 1.45e-07 | 3.84e-07 | 1.97e-07 |
| HWCM | $t = 1$ | $3m = 27,\ \Delta t = 0.005$ | 2.04e-05 | 3.76e-05 | 2.16e-05 | 1.22e-05 | 2.45e-05 |
| HWCM | $t = 0.02$ | $3m = 81,\ \Delta t = 0.00125$ | 1.78e-07 | 1.35e-08 | 4.27e-07 | 4.07e-07 | 1.41e-07 |
| HWCM | $t = 1$ | $3m = 81,\ \Delta t = 0.005$ | 1.54e-05 | 1.06e-05 | 1.17e-05 | 3.13e-05 | 3.85e-05 |
| H.O.C.M. [13] | $t = 0.02$ | $h = 0.05,\ \Delta t = 0.00125$ | 1.40e-07 | 2.90e-07 | 5.60e-07 | 3.40e-07 | 1.70e-07 |
| H.O.C.M. [13] | $t = 1$ | $h = 0.05,\ \Delta t = 0.005$ | 2.59e-03 | 1.91e-03 | 7.17e-04 | 2.20e-03 | 6.65e-04 |
| Danea and Evans [10] | $t = 0.02$ | $h = 0.05,\ \Delta t = 0.00125$ | 2.50e-06 | 3.90e-06 | 1.37e-05 | 2.60e-06 | 9.80e-06 |
| Danea and Evans [10] | $t = 1$ | $h = 0.05,\ \Delta t = 0.005$ | 3.19e-03 | 2.73e-03 | 9.80e-03 | 1.25e-02 | 1.40e-02 |
| Evans [11] | $t = 0.02$ | $h = 0.05,\ \Delta t = 0.00125$ | 8.44e-06 | 1.42e-05 | 1.74e-05 | 1.40e-06 | 1.20e-05 |
| Evans [11] | $t = 1$ | $h = 0.05,\ \Delta t = 0.005$ | 3.20e-03 | 2.73e-03 | 9.80e-03 | 1.25e-02 | 1.40e-02 |
| Richtmyer [47] | $t = 0.02$ | $h = 0.05,\ \Delta t = 0.00125$ | 2.24e-04 | 3.67e-04 | 4.03e-04 | 3.64e-04 | 3.35e-04 |
| Richtmyer [47] | $t = 1$ | $h = 0.05,\ \Delta t = 0.005$ | 2.73e-03 | 9.48e-03 | 1.74e-02 | 2.30e-02 | 2.24e-02 |
| Semi-explicit [13] | $t = 0.02$ | $h = 0.05,\ \Delta t = 0.00125$ | 3.01e-05 | 6.19e-05 | 6.69e-05 | 5.10e-05 | 1.34e-05 |
| Semi-explicit [13] | $t = 1$ | $h = 0.05,\ \Delta t = 0.005$ | 2.74e-03 | 5.93e-03 | 4.48e-03 | 2.32e-03 | 6.51e-03 |
| Mittal and Jain[36] | $t = 0.02$ | $n = 181,\ \Delta t = 0.005$ | 1.14e-05 | 1.41e-05 | 9.70e-06 | 8.02e-06 | 1.92e-05 |
| Mittal and Jain[36] | $t = 1$ | $n = 181,\ \Delta t = 0.005$ | 7.33e-04 | 1.44e-03 | 2.04e-03 | 2.47e-03 | 2.63e-03 |
| HWCM | $t = 0.02$ | $3m = 81,\ \Delta t = 0.00125$ | $L_\infty$ = 4.4750e-07 | | | | |
| HWCM | $t = 1$ | $3m = 81,\ \Delta t = 0.005$ | $L_\infty$ = 3.8503e-05 | | | | |

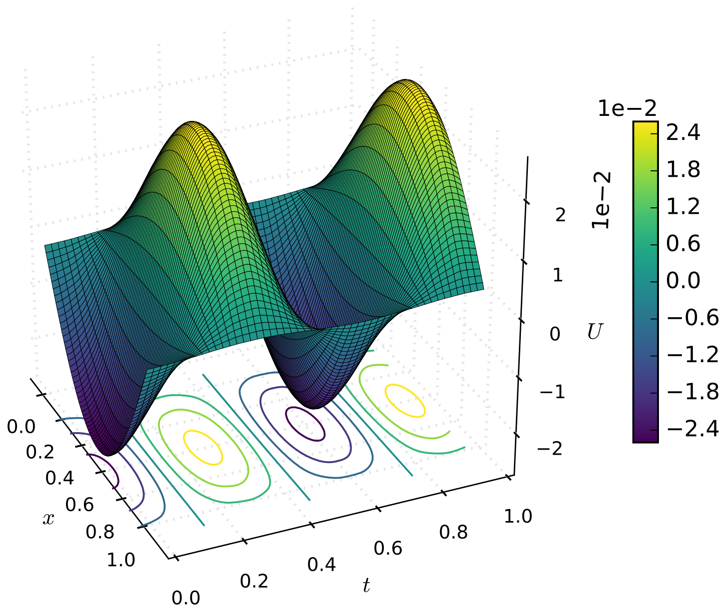



**Figure 9.** Evolution of numerical solution for $3m = 81$ and $\Delta t = 0.01$ from $t = 0$ to $t = 1$ in Problem 4

**Table 8.** Convergence rates for $\Delta t = 0.0001$ at final time $t = 1$ in Problem 4

| | $L_\infty$ | Rate |
|---|---|---|
| $3m = 9$ | 4.693704e-04 | - |
| $3m = 27$ | 4.495959e-05 | 2.135 |
| $3m = 81$ | 4.810131e-06 | 2.034 |
| $3m = 243$ | 6.716085e-07 | 1.792 |

## 4.5. Problem 5

We consider a constant coefficient ($\mu(x) = 1$, $\mathrm{EI}(x) = -1$) fourth order homogeneous parabolic partial differential equation which is also studied in [36]

$$\frac{\partial^2 u}{\partial t^2} = \frac{\partial^4 u}{\partial x^4}$$

subject to the initial conditions

$$u(x,0) = \sin(\pi x), \quad u_t(x,0) = -\pi^2 \sin(\pi x), \quad 0 \le x \le 1$$

and with boundary conditions

$$u(0,t) = u(1,t) = u_{xx}(0,t) = u_{xx}(1,t) = 0, \quad t \ge 0.$$

The exact solution of the problem is given by

$$u(x,t) = \sin(\pi x)e^{-\pi^2 t}.$$

In Table 9, we give computed results by the present method for $3m = 27$ and $\Delta t = 0.005$ at $t = 0.02,\ 0.05$. We also give the results of [36] for comparison. We observe in Table 9 that the present method gives more accurate results than B-spline methods with redefined basis functions [36].

**Table 9.** $L_\infty$ and Absolute errors for Problem 5

| Methods | Time | Parameters | $x = 0.1$ | $x = 0.2$ | $x = 0.3$ | $x = 0.4$ | $x = 0.5$ |
|---|---|---|---|---|---|---|---|
| HWCM | $t = 0.02$ | $3m = 27,\ \Delta t = 0.005$ | 7.74e-06 | 1.47e-05 | 2.05e-05 | 2.40e-05 | 2.50e-05 |
| HWCM | $t = 0.05$ | $3m = 27,\ \Delta t = 0.005$ | 5.99e-05 | 3.07e-05 | 2.89e-05 | 6.52e-05 | 8.15e-06 |
| Mittal and Jain[36] | $t = 0.02$ | $n = 31,\ \Delta t = 0.005$ | 2.80e-04 | 5.33e-04 | 7.33e-04 | 8.62e-04 | 9.06e-04 |
| Method 1 | $t = 0.05$ | $n = 31,\ \Delta t = 0.005$ | 2.62e-04 | 4.98e-04 | 6.86e-04 | 8.07e-04 | 8.48e-04 |
| Mittal and Jain [36] | $t = 0.02$ | $n = 31,\ \Delta t = 0.005$ | 1.08e-04 | 2.06e-04 | 2.83e-04 | 3.33e-04 | 3.50e-04 |
| Method 2 | $t = 0.05$ | $n = 31,\ \Delta t = 0.005$ | 6.13e-04 | 1.35e-03 | 1.95e-03 | 2.18e-03 | 2.20e-03 |
| HWCM | $t = 0.02$ | $3m = 27,\ \Delta t = 0.005$ | $L_\infty = 2.4987e-05$ | | | | |
| HWCM | $t = 0.05$ | $3m = 27,\ \Delta t = 0.005$ | $L_\infty = 6.7356e-05$ | | | | |

## 5. Conclusion

Our main goal in this study is to propose a new 3-scale Haar wavelet based method to high order partial differential equations and analyze the performance of the method. The comparisons of numerical solutions with exact solutions and the results from the previous studies that are based on numerical techniques such as finite differences, B-splines and high order spline methods indicate the power of the new 3-scale Haar wavelet based method in dealing with variable coefficient, constant coefficient, homogeneous and non-homogeneous partial differential equations. The implementation of the method is straight-forward and simpler than the existing methods. The advantages of the Haar wavelet based method can be listed as follows.

- High accuracy is attained even with small number of collocation points.
- Small computational costs are required, and the implementation of the method in computers is easy
- Coping with boundary conditions is very easy compared with other known methods.

We also note that the new 3-scale Haar wavelet based method introduced here with suitable modifications can be easily applied to similar problems.

## References

[1] C. Andrade and S. McKee, *High accuracy A.D.I. methods for fourth-order parabolic equations with variable coefficients,* J. Comput. Appl. Math. **3** (1), 11–14, 1977.

[2] T. Aziz, A. Khan and J. Rashidinia, *Spline methods for the solution of fourth-order parabolic partial differential equations*, Appl. Math. Comput. **167**, 153–166, 2005.

[3] H.T. Banks and K. Kunisch, *Estimation Techniques for Distributed Parameter Systems*, Birkhauser, Boston, 1989.

[4] H. Caglar and N. Caglar, *Fifth-degree B-spline solution for a fourth-order parabolic partial differential equations*, Appl. Math. Comput. **201**, 597–603, 2008.

[5] C. Chen and C.H. Hsiao, *Haar wavelet method for solving lumped and distributed parameter systems*, IEE Proc. Control Theory Appl. **144**, 87–94, 1997.

[6] C. Chen and C.H. Hsiao, *Wavelet approach to optimising dynamic systems*, IEE Proc. Control Theory Appl. **146**, 213–219, 1997.

[7] L. Collatz, *Hermitian methods for initial value problems in partial differential equations*, in: J.J.H. Miller (Ed.), Topics in Numerical Analysis, Academic Press, New York, 41–61, 1973.

[8] S.D. Conte, *A stable implicit finite difference approximation to a fourth order parabolic equation*, J. Assoc. Comput. Mech. **4**, 18–23, 1957.

[9] S.H. Crandall, *Numerical treatment of a fourth order partial differential equations*, J. Assoc. Comput. Mech. **1**, 111–118, 1954.

[10] A. Danaee, Arshad Khan, Islam Khan, Tariq Aziz and D.J. Evans, *Hopscotch procedure for a fourth-order parabolic partial differential equation*, Math. Comput. Simulat. **XXIV**, 326–329, 1982.

[11] D.J. Evans, *A stable explicit method for the finite difference solution of a fourth order parabolic partial differential equation*, Comput. J. **8**, 280–287, 1965.

[12] D.J. Evans and W.S.Yousif, *A note on solving the fourth order parabolic equation by the age method*, Int. J. Comput. Math. **40**, 93–97, 1991.

[13] G. Fairweather and A.R. Gourlay, *Some stable difference approximations to a fourth order parabolic partial differential equation*, Math. Comput. **21**, 1–11, 1967.

[14] H. Haddadpour, *An exact solution for variable coefficients fourth-order wave equation using the Adomian method*, Math. Comput. Model. **44**, 144–1152, 2006.

[15] C.H. Hsiao, *Haar wavelet direct method for solving variational problems*, Math. Comput. Simul. **64**, 569–585, 2004.

[16] C.H. Hsiao and W.J. Wang, *State analysis of time-varying singular nonlinear systems via Haar wavelets* Math. Comput. Simul. **51**, 91–100, 1999.

[17] C.H. Hsiao and W.J. Wang, *State analysis of time-varying singular bilinear systems via Haar wavelets*, Math. Comput. Simul. **52**, 11–20, 2000.

[18] C.H. Hsiao and W.J. Wang, *Haar wavelet approach to nonlinear stiff systems*, Math. Comput. Simul. **57**, 347–353, 2001.

[19] J.D. Hunter, *Matplotlib: A 2D graphics environment, Comput Sci Eng*, **9** (3), 90–95, 2007.

[20] M.K. Jain, S.R.K. Iyengar and A.G. Lone, *Higher order difference formulas for a fourth order parabolic partial differential equation*, Int. J. Numer. Methods Eng. **10**, 1357–1367, 1976.

[21] R. Jiwari, *Haar wavelet quasilinearization approach for numerical simulation of Burgers' equation*, Comput. Phys. Commun. **183**, 2413–2423, 2012.

[22] R. Jiwari, *A hybrid numerical scheme for the numerical solution of the Burgers' equation*, Comput. Phys. Commun. **188**, 59–67 2015.

[23] R. Jiwari, V. Kumar, R. Karan and A. S. Alshomrani, *Haar wavelet quasilinearization approach for MHD Falkner–Skan flow over permeable wall via Lie group method*, Int. J. Numer. Method H. **27** (6), 1332–1350, 2017.

[24] H. Kaur, R.C. Mittal and V. Mishra, *Haar wavelet approximate solutions for the generalized Lane–Emden equations arising in astrophysics*, Comput. Phys. Commun. **184**, 2169–2177, 2013.

[25] A.Q.M. Khaliq and E.H. Twizell, *A family of second order methods for variable coefficient fourth order parabolic partial differential equations*, Int. J. Comput. Math. **23**, 63–76, 1987.

[26] A. Khan, I. Khan, and T. Aziz, *Sextic spline solution for solving a fourth-order parabolic partial differential equation*, Int. J. Comput. Math. **82** (7), 871–879, 2005.

[27] M. Kirs, M. Mikola, A. Haavajõe, E. Õunapuu, B. Shvartsman and J. Majak, *Haar wavelet method for vibration analysis of nanobeams*, Waves Wavelets Fractals, **2** (1), 2016.

[28] K. Kunisch and E. Graif, *Parameter estimation for the Euler–Bernoulli beam*, Mat. Apficada Comput. **4**, 95–124, 1985.

[29] U. Lepik, *Numerical solution of differential equations using Haar wavelets*, Math. Comput. Simul. **68**, 127–143, 2005.

[30] U. Lepik, *Numerical solution of evolution equations by the Haar wavelet method*, Appl. Math. Comput. **185**, 695–704, 2007.

[31] U. Lepik, *Solving PDEs with the aid of two-dimensional Haar wavelets*, Comput. Math. with Appl. **61**, 1873–1879, 2011.

[32] Y. Liu and C.S. Gurram, *The use of He's variational iteration method for obtaining the free vibration of an Euler–Bernoulli beam*, Math. Comput. Model. **50**, 1545–1552, 2009.

[33] J. Majak, B. Shvartsman, M. Kirs, M. Pohlak and H. Herranen, *Convergence theorem for the Haar wavelet based discretization method*, Compos. Struct. **126**, 227–232, 2015.

[34] J. Majak, M. Pohlak, K. Karjust, M. Eerme, J. Kurnitski and B.S. Shvartsman, *New higher order Haar wavelet method: Application to FGM structures*, **201**, 72–78, 2018. https://doi.org/10.1016/j.compstruct.2018.06.013.

[35] J. Majak, B. Shvartsman, K. Karjust, M. Mikola, A. Haavajõe and M. Pohlak, *On the accuracy of the Haar wavelet discretization method*, Compos. B. Eng. **80**, 321–327, 2015.

[36] R.C. Mittal and R.K. Jain, *B-splines methods with redefined basis functions for solving fourth order parabolic partial differential equations*, Appl. Math. Comput. **217**, 9741–9755, 2011.

[37] R.C. Mittal and S. Pandit, *Sensitivity analysis of shock wave Burgers' equation via a novel algorithm based on scale-3 Haar wavelets*, Int. J. Comput. Math. **95** (3), 601–625, 2017.

[38] R.C. Mittal and S. Pandit, *A Numerical Algorithm to Capture Spin Patterns of Fractional Bloch NMR Flow Models*, J. Comput. Nonlinear Dynam. **14** (8), 2019.

[39] R.C. Mittal and S. Pandit, *New Scale-3 Haar Wavelets Algorithm for Numerical Simulation of Second Order Ordinary Differential Equations*, P. Natl. A. Sci. India A, **89**, 799–808, 2019.

[40] R.C. Mittal and S. Pandit, *Quasilinearized Scale-3 Haar Wavelets based Algorithm for Numerical Simulation of Fractional Dynamical System*, Eng. Computations. **35** (5), 1907–1931, 2018.

[41] R. Mohammadi, *Sextic B-spline collocation method for solving Euler–Bernoulli Beam Models*, Appl. Math. Comput. **241**, 151–166, 2014.

[42] Ö. Oruç, F. Bulut and A. Esen, *A Haar wavelet-finite difference hybrid method for the numerical solution of the modified Burgers equation*, J. Math. Chem. **53** (7), 1592–1607, 2015.

[43] Ö. Oruç, F. Bulut and A. Esen, *Numerical Solutions of Regularized Long Wave Equation By Haar Wavelet Method*, Mediterr. J. Math. **13** (5), 3235–3253, 2016.

[44] Ö. Oruç, A. Esen and F. Bulut, *A Haar wavelet collocation method for coupled nonlinear Schrödinger–KdV equations*, Int. J. Mod. Phys. C, **27** (9), 2016.

[45] S. Pandit, M. Kumar and S. Tiwari, *Numerical simulation of second-order hyperbolic telegraph type equations with variable coefficients*, Comput. Phys. Commun. **187**, 83–90, 2015.

[46] J. Rashidinia and R. Mohammadi, *Sextic spline solution of variable coefficient fourth-order parabolic equations*, Int. J. Comput. Math. **87** (15), 3443–3454, 2010.

[47] R.D. Richtmyer and K.W. Morton, *Difference methods for Initial value Problems*, second ed., John Wiley & Sons, 1967.

[48] Z. Shi, Y. Cao and Q.J. Chen, *Solving 2D and 3D Poisson equations and biharmonic equations by the Haar wavelet method*, Appl. Math. Model. **36**, 5143–5161, 2012.

[49] R.C. Smith, K.L. Bowers and J. Lund, *A fully Sinc–Galerkin method for Euler–Bernoulli Beam Models*, Numer. Methods Partial Diff. Equ. **8**, 171–202, 1992.

[50] S.P. Timoshenko and J.M. Gere, *Theory of Elastic Stability*, McGraw-Hill, New York, 1961.

[51] J. Todd, *A direct approach to the problem of stability in the numerical solution of partial differential equations*, Commun. Pure Appl. Math. **9**, 597–612, 1956.

[52] A.M. Wazwaz, *Analytic treatment for variable coefficient fourth-order parabolic partial differential equations*, Appl. Math. Comput. **123**, 219–227, 2001.

Hacettepe Journal of Mathematics & Statistics

Hacet. J. Math. Stat.
Volume 50 (1) (2021), 180 – 187
DOI : 10.15672/hujms.621072

Research Article

# On monotonic and logarithmic concavity properties of generalized $k$-Bessel function

İbrahim Aktaş

*Department of Mathematics, Kamil Özdağ Science Faculty, Karamanoğlu Mehmetbey University, Yunus Emre Campus, 70100, Karaman, Turkey*

**Abstract**

In this study, our main objective is to determine some monotonic and log-concavity properties of generalized $k$-Bessel function by using its Hadamard product representation and some earlier results on power series. In addition, by using the relationships between Bessel-type special functions and some basic functions, we present some specific examples related to the monotonic and log-concavity properties of some trigonometric and hyperbolic functions.

**Mathematics Subject Classification (2020).** 33E50

**Keywords.** $k$-Gamma functions, $k$-Bessel function, monotonicity, log-concavity

## 1. Introduction and preliminaries

In the recent years many geometric and monotonic properties of some special functions like Bessel, Struve, Lommel, Mittag-Leffler, Wright and their generalizations were investigated by many authors. Comprehensive information about these investigations can be found in [1–8, 10, 14] and references therein. Especially, some inequalities and monotonic properties of the above mentioned functions are usefull in engineering, physics, probability and statistics, and economics. It is known that log-concavity and log-convexity properties have a crucial role in economics. Comprehensive information about the log-concavity and the log-convexity properties can be found in [13] and its references. In this study, motivated by the some earlier results which are given in [14, 15], our main aim is to present some monotonic and log-concavity properties of generalized $k$-Bessel functions. Moreover, we give some specific examples regarding our obtained result by using the relationships between Bessel-type functions and elementary trigonometric and hyperbolic functions.

It is known that, most of special functions can be defined with the help of Euler's gamma function. Therefore, we would like to remind the definitions of gamma function and its $k$-generalization. The Euler's gamma function $\Gamma$ is defined by the following improper integral, for $x > 0$:

$$\Gamma(x) = \int_0^{\infty} t^{x-1} e^{-t} dt.$$

Email address: aktasibrahim38@gmail.com; ibrahimaktas@kmu.edu.tr
Received: 17.09.2019; Accepted: 28.05.2020

Also, the $k$-gamma function is defined by (see [12])

$$\Gamma_k(x) = \int_0^\infty t^{x-1} e^{\frac{-t^k}{k}} dt$$

for $k > 0$. We know that the $k$-gamma function $\Gamma_k$ reduces to the classical gamma function $\Gamma$ when $k \to 1$. In addition, Pochammer $k$-symbol is defined by

$$(\lambda)_{n,k} = \lambda(\lambda + k)(\lambda + 2k)\ldots((\lambda + (n-1)k))$$

for $\lambda \in \mathbb{C}, k \in \mathbb{R}$ and $n \in \mathbb{N}^+$. Other properties of Pochammer $k$-symbol and $k$-gamma function can be found in [12].

In this paper, we are considering the generalized $k$-Bessel function defined by the following series representation (see [14]):

$$W^k_{\nu,c}(x) = \sum_{n=0}^{\infty} \frac{(-c)^n}{n!\Gamma_k(nk + \nu + k)} \left(\frac{x}{2}\right)^{2n+\frac{\nu}{k}} \tag{1.1}$$

for $k > 0, \nu > -1$ and $c \in \mathbb{R}$. It is clear that the generalized $k$-Bessel function reduces to classical Bessel and modified Bessel functions for appropriate values of the parameters $k$ and $c$, respectively. More precisely, taking $k = c = 1$ and $k = -c = 1$ in (1.1), we have that

$$W^1_{\nu,1}(x) = \sum_{n=0}^{\infty} \frac{(-1)^n}{n!\Gamma(n + \nu + 1)} \left(\frac{x}{2}\right)^{2n+\nu} = J_\nu(x) \tag{1.2}$$

and

$$W^1_{\nu,-1}(x) = \sum_{n=0}^{\infty} \frac{1}{n!\Gamma(n + \nu + 1)} \left(\frac{x}{2}\right)^{2n+\nu} = I_\nu(x), \tag{1.3}$$

where $J_\nu(x)$ and $I_\nu(x)$ denote classical Bessel and modified Bessel functions of the first kind, respectively. In [15], the author studied some geomertric properties such as radii of starlikeness and convexity of generalized $k$-Bessel function. Also, the author gave an infinite product representation of generalized $k$-Bessel function by using Hadamard's theorem as follow (see [15, Lemma 1.1]):

$$W^k_{\nu,c}(x) = \frac{(\frac{x}{2})^{\frac{\nu}{k}}}{\Gamma_k(\nu + k)} \prod_{n\geq 1} \left(1 - \frac{x^2}{{}_kw^2_{\nu,c,n}}\right), \tag{1.4}$$

where ${}_kw_{\nu,c,n}$ denotes $n$th positive zero of generalized $k$-Bessel function $W^k_{\nu,c}(x)$.

Now, we would like to give the definition of logarithmic concavity of a function.

**Definition 1.1** ([13])**.** A function $f$ is said to be log-concave on interval $(a, b)$ if the function $\log f$ is a concave function on $(a, b)$.

To show log-concavity of a function $f$ on the interval $(a, b)$, it is sufficient to show one of the following two conditions:

**i.** $\frac{f'}{f}$ monotone decreasing on $(a, b)$.

**ii.** $\log f'' < 0$.

Also the following lemma due to Biernacki and Krzyż (see [11]) will be used in order to prove some monotonic properties of the mentioned functions.

**Lemma 1.2.** *Consider the power series* $f(x) = \sum_{n\geq 0} a_n x^n$ *and* $g(x) = \sum_{n\geq 0} b_n x^n$*, where* $a_n \in \mathbb{R}$ *and* $b_n > 0$ *for all* $n \in \{0, 1, \ldots\}$*, and suppose that both converge on* $(-r, r), r > 0$*. If the sequence* $\{\frac{a_n}{b_n}\}_{n\geq 0}$ *is increasing(decreasing), then the function* $x \mapsto \left(\frac{f(x)}{g(x)}\right)$ *is also increasing(decreasing) on* $(0, r)$*.*

It is important to note that the above result remains true for the even or odd functions.

The outcomes of our paper is as follow: In Section 2, we give our main results and their consequences, while the Section 3 is devoted for some applications of our main results.

## 2. Main results

In this section, we present our main results and their consequences.

**Theorem 2.1.** *Let $k > 0, k + \nu > 0, c \in \mathbb{R}$ and ${}_k w_{\nu,c,n}$ denote the nth positive zero of the generalized k-Bessel function $W^k_{\nu,c}(x)$. Further, consider the following sets:*

$$\delta_1 = \bigcup_{n\geq 1} \left({}_k w_{\nu,c,2n-1}, {}_k w_{\nu,c,2n}\right), \delta_2 = \bigcup_{n\geq 1} \left({}_k w_{\nu,c,2n}, {}_k w_{\nu,c,2n+1}\right) \text{ and } \delta_3 = [0, {}_k w_{\nu,c,1}) \cup \delta_2.$$

*The generalized k-Bessel function*

$$\Theta^k_{\nu,c}(x) = \Gamma_k(\nu + k) 2^{\frac{\nu}{k}} x^{-\frac{\nu}{k}} W^k_{\nu,c}(x) = \sum_{n=0}^{\infty} \frac{(-c)^n}{n!\,(\nu + k)_{n,k}} \left(\frac{x}{2}\right)^{2n} \tag{2.1}$$

*has the following properties:*

- **a.** *the function $x \mapsto \Theta^k_{\nu,c}(x)$ is negative on $\delta_1$ and it is positive on $\delta_3$,*
- **b.** *the function $x \mapsto \Theta^k_{\nu,c}(x)$ is a decreasing function on $[0, {}_k w_{\nu,c,1})$,*
- **c.** *the function $x \mapsto \Theta^k_{\nu,c}(x)$ is strictly log-concave on $\delta_3$.*

***Proof.*** **a.** If we consider the infinite product representation of generalized $k$-Bessel function $W^k_{\nu,c}(x)$ which is given by (1.4), then it can be easily seen that the function $\Theta^k_{\nu,c}(x)$ can be written by the following product representation:

$$\Theta^k_{\nu,c}(x) = \prod_{n\geq 1} \left(1 - \frac{x^2}{{}_k w^2_{\nu,c,n}}\right). \tag{2.2}$$

In order to investigate the sign of the function $x \mapsto \Theta^k_{\nu,c}(x)$ on the mentioned sets, we rewrite the function $x \mapsto \Theta^k_{\nu,c}(x)$ as

$$\Theta^k_{\nu,c}(x) = U_n V_n,$$

where

$$U_n = \prod_{n\geq 1} \frac{{}_k w_{\nu,c,n} + x}{{}_k w^2_{\nu,c,n}} \text{ and } V_n = \prod_{n\geq 1} \left({}_k w_{\nu,c,n} - x\right).$$

It is clear that $U_n > 0$ for all $x \in \mathbb{R}^+ \cup \{0\}$. On the other hand, since

$$0 < {}_k w_{\nu,c,1} < {}_k w_{\nu,c,2} < \cdots < {}_k w_{\nu,c,n} < \cdots,$$

we can say that, if $x \in ({}_k w_{\nu,c,2n-1}, {}_k w_{\nu,c,2n})$, then the first $(2n-1)$ terms of $V_n$ are strictly negative and remained terms are strictly positive. Also, if $x \in ({}_k w_{\nu,c,2n}, {}_k w_{\nu,c,2n+1})$, then the first $2n$ terms of $V_n$ are strictly negative and the rest is strictly positive. In addition, all the terms of $V_n$ are strictly positive for $x \in [0, {}_k w_{\nu,c,1})$. As a consequence, the function $x \mapsto \Theta^k_{\nu,c}(x)$ is negative on $\delta_1$ and it is positive on $\delta_3$.

**b.** We know from part ***a.*** that the function $x \mapsto \Theta^k_{\nu,c}(x)$ is positive on the interval $[0, {}_k w_{\nu,c,1})$. The logarithmic differentation of (2.2) implies that

$$\frac{\left(\Theta^k_{\nu,c}(x)\right)'}{\Theta^k_{\nu,c}(x)} = \sum_{n=1}^{\infty} \frac{2x}{x^2 - {}_k w^2_{\nu,c,n}}.$$

Thus, we get

$$\left(\Theta^k_{\nu,c}(x)\right)' = \Theta^k_{\nu,c}(x) \sum_{n=1}^{\infty} \frac{2x}{x^2 - {}_k w^2_{\nu,c,n}} < 0$$

for all $x \in [0, {}_k w_{\nu,c,1})$. As a result, the function $x \mapsto \Theta^k_{\nu,c}(x)$ is a decreasing function on $[0, {}_k w_{\nu,c,1})$.

**c.** In order to prove log-concavity of the function $x \mapsto \Theta^k_{\nu,c}(x)$, we need to show that

$$\frac{d^2}{dx^2}\left[\log \Theta^k_{\nu,c}(x)\right] < 0$$

for all $x \in \delta_3$. Now, by using the infinite product representation of the function $\Theta^k_{\nu,c}(x)$ which is given by (2.2) we infer that

$$\begin{aligned}
\frac{d^2}{dx^2}\left[\log \Theta^k_{\nu,c}(x)\right] &= \frac{d^2}{dx^2}\left[\log \prod_{n\geq 1}\left(1-\frac{x^2}{{}_k w^2_{\nu,c,n}}\right)\right] \\
&= \frac{d}{dx}\left[\frac{d}{dx}\sum_{n=1}^{\infty}\log\left(1-\frac{x^2}{{}_k w^2_{\nu,c,n}}\right)\right] \\
&= \frac{d}{dx}\sum_{n=1}^{\infty}\frac{-2x}{{}_k w^2_{\nu,c,n}-x^2} \\
&= -2\sum_{n=1}^{\infty}\frac{{}_k w^2_{\nu,c,n}+x^2}{\left({}_k w^2_{\nu,c,n}-x^2\right)^2} \\
&< 0
\end{aligned}$$

for $x \in \delta_3$. Thus, the proof is completed. $\square$

By setting $k = c = 1$ and $k = 1, c = -1$ in the Theorem 2.1 we have the following properties for the classical Bessel and modified Bessel functions, respectively.

**Corollary 2.2.** *Let $\nu > -1$ and $j_{\nu,n}$ denote the nth positive zero of the classical Bessel function $J_\nu(x)$. Further, consider the next sets:*

$$A_1 = \bigcup_{n\geq 1}\left(j_{\nu,2n-1}, j_{\nu,2n}\right), A_2 = \bigcup_{n\geq 1}\left(j_{\nu,2n}, j_{\nu,2n+1}\right) \text{ and } A_3 = [0, j_{\nu,1}) \cup A_2.$$

*The following assertions are true:*

- ***a.*** *the function $\Theta^1_{\nu,1}(x) = \Gamma(\nu+1)2^\nu x^{-\nu} J_\nu(x)$ is negative on $A_1$ and it is positive on $A_3$,*
- ***b.*** *the function $\Theta^1_{\nu,1}(x) = \Gamma(\nu+1)2^\nu x^{-\nu} J_\nu(x)$ is a decreasing function on $[0, j_{\nu,1})$,*
- ***c.*** *the function $\Theta^1_{\nu,1}(x) = \Gamma(\nu+1)2^\nu x^{-\nu} J_\nu(x)$ is strictly log-concave on $A_3$.*

**Corollary 2.3.** *Let $\nu > -1$ and $\epsilon_{\nu,n}$ denote the nth positive zero of the modified Bessel function $I_\nu(x)$. Further, consider the next sets:*

$$B_1 = \bigcup_{n\geq 1}\left(\epsilon_{\nu,2n-1}, \epsilon_{\nu,2n}\right), B_2 = \bigcup_{n\geq 1}\left(\epsilon_{\nu,2n}, \epsilon_{\nu,2n+1}\right) \text{ and } B_3 = [0, \epsilon_{\nu,1}) \cup B_2.$$

*The following assertions are true:*

- ***a.*** *the function $\Theta^1_{\nu,-1}(x) = \Gamma(\nu+1)2^\nu x^{-\nu} I_\nu(x)$ is negative on $B_1$ and it is positive on $B_3$,*
- ***b.*** *the function $\Theta^1_{\nu,-1}(x) = \Gamma(\nu+1)2^\nu x^{-\nu} I_\nu(x)$ is a decreasing function on $[0, \epsilon_{\nu,1})$,*
- ***c.*** *the function $\Theta^1_{\nu,-1}(x) = \Gamma(\nu+1)2^\nu x^{-\nu} I_\nu(x)$ is strictly log-concave on $B_3$.*

**Theorem 2.4.** *Let $k > 0, \nu > 0, c \in \mathbb{R}$ and ${}_k w_{\nu,c,n}$ denote the nth positive zero of the generalized k-Bessel function $W^k_{\nu,c}(x)$. Then, the function $x \mapsto W^k_{\nu,c}(x)$ is strictly log-concave on $(0, {}_k w_{\nu,c,1}) \cup \delta_2$.*

***Proof.*** It is known that the product of two strictly log-concave function is also strictly log-concave. By using this fact it is possible to prove the log-concavity of the generalized

$k$-Bessel function $W^k_{\nu,c}(x)$ on $\delta_3$. Hence, we rewrite the function $W^k_{\nu,c}(x)$ as follow:

$$W^k_{\nu,c}(x) = \frac{\left(\frac{x}{2}\right)^{\frac{\nu}{k}}}{\Gamma_k(\nu+k)}\Theta^k_{\nu,c}(x).$$

Since

$$\frac{d^2}{dx^2}\left[\log\left(\frac{x}{2}\right)^{\frac{\nu}{k}}\right] = -\frac{\nu}{kx^2} < 0$$

for $\nu > 0, k > 0$ and $x \in \mathbb{R}^+$, the function $x \mapsto \left(\frac{x}{2}\right)^{\frac{\nu}{k}}$ is strictly log-concave on $\mathbb{R}^+$. In addition, it is known from part ***c.*** of Theorem 2.1 that the function $\Theta^k_{\nu,c}(x)$ is strictly log-concave on $\delta_3$. As a result, the function $W^k_{\nu,c}(x)$ is strictly log-concave on $(0, {}_kw_{\nu,c,1}) \cup \delta_2$ as a product of two strictly log-concave functions. □

Now, by taking $k = c = 1$ and $k = 1, c = -1$ in Theorem 2.4, we deduce the following properties for the classical Bessel and modified Bessel functions, respectively.

**Corollary 2.5.** *The function $x \mapsto J_\nu(x)$ is strictly log-concave on $(0, j_{\nu,1}) \cup A_2$, while the function $x \mapsto I_\nu(x)$ is strictly log-concave on $(0, \epsilon_{\nu,1}) \cup B_2$.*

Our last main result is the following theorem.

**Theorem 2.6.** *The function $\Phi^k_{\nu,-1}(x) = \frac{x\left(\Theta^k_{\nu,-1}(x)\right)'}{\Theta^k_{\nu,-1}(x)}$ is increasing on $(0,\infty)$ for $v > -1$ and $\nu + k > 0$.*

***Proof.*** If we put $c = -1$ in definition of the function $\Theta^k_{\nu,c}(x)$, then we get the following infinite series representation for the function $\Theta^k_{\nu,-1}(x)$, that is,

$$\Theta^k_{\nu,-1}(x) = \sum_{n=0}^{\infty} \mathcal{P}_{n,\nu,k}x^{2n}, \tag{2.3}$$

where $\mathcal{P}_{n,\nu,k} = \frac{1}{n!4^n(\nu+k)_{n,k}}$. Differentiating both sides of the equality (2.3) and by multiplying by $x$ obtained equality, we get that

$$x\left(\Theta^k_{\nu,-1}(x)\right)' = \sum_{n=0}^{\infty} \mathcal{R}_{n,\nu,k}x^{2n},$$

where $\mathcal{R}_{n,\nu,k} = \frac{2n}{n!4^n(\nu+k)_{n,k}}$. According to Cauchy-Hadamard theorem for power series, it can be easily shown that both power series $\sum_{n=0}^{\infty} \mathcal{P}_{n,\nu,k}x^{2n}$ and $\sum_{n=0}^{\infty} \mathcal{R}_{n,\nu,k}x^{2n}$ are convergent on $(-\infty, \infty)$, since

$$\lim_{n\to\infty}\left|\frac{\mathcal{P}_{n,\nu,k}}{\mathcal{P}_{n+1,\nu,k}}\right| = \lim_{n\to\infty}\left|\frac{\mathcal{R}_{n,\nu,k}}{\mathcal{R}_{n+1,\nu,k}}\right| = \infty.$$

Here we used the equality $(\nu+k)_{n+1,k} = (\nu+k+nk)(\nu+k)_{n,k}$ for the Pochammer $k$-symbol. On the other hand, it can be easily seen that $\mathcal{R}_{n,\nu,k} \in \mathbb{R}$ and $\mathcal{P}_{n,\nu,k} > 0$ for all $n \in \{0,1,\dots\}, \nu > -1$ and $\nu + k > 0$. Now, if we consider the sequence

$$U_n = \frac{\mathcal{R}_{n,\nu,k}}{\mathcal{P}_{n,\nu,k}} = 2n,$$

then we have

$$\frac{U_{n+1}}{U_n} = \frac{n+1}{n} > 1.$$

So the sequence $\{U_n\}_{n\geq 0}$ is increasing. The proof is completed by applying Lemma 1.2. □

## 3. Applications

In this section, we want to give some applications of our main results. Therefore, we consider the relationships among of the functions $x \mapsto \Theta^k_{\nu,c}(x)$, $x \mapsto J_\nu(x)$ and $x \mapsto I_\nu(x)$. We know from (1.2) and (1.3) that, the following equalities

$$W^1_{\nu,1}(x) = J_\nu(x) \text{ and } W^1_{\nu,-1}(x) = I_\nu(x)$$

hold true for $k = c = 1$ and $k = 1, c = -1$, respectively. On the other hand, we know from [9] that some basic trigonometric and hyperbolic functions can be written in terms of Bessel and modified Bessel functions for some special values of $\nu$. Especially, for $\nu = -\frac{1}{2}, \nu = \frac{1}{2}$ and $\nu = \frac{3}{2}$ we have the following basic trigonometric and hyperbolic functions:

$$J_{-\frac{1}{2}}(x) = \sqrt{\frac{2}{\pi x}} \cos x, \quad J_{\frac{1}{2}}(x) = \sqrt{\frac{2}{\pi x}} \sin x, \quad J_{\frac{3}{2}}(x) = \sqrt{\frac{2}{\pi x}} \left( \frac{\sin x}{x} - \cos x \right)$$

and

$$I_{-\frac{1}{2}}(x) = \sqrt{\frac{2}{\pi x}} \cosh x, \quad I_{\frac{1}{2}}(x) = \sqrt{\frac{2}{\pi x}} \sinh x, \quad I_{\frac{3}{2}}(x) = -\sqrt{\frac{2}{\pi x}} \left( \frac{\sinh x}{x} - \cosh x \right).$$

By using above relationships, we have the followings:

$$\Theta^1_{-\frac{1}{2},1}(x) = \cos x, \quad \Theta^1_{\frac{1}{2},1}(x) = \frac{\sin x}{x}, \quad \Theta^1_{\frac{3}{2},1}(x) = 3\left( \frac{\sin x - x\cos x}{x^3} \right)$$

and

$$\Theta^1_{-\frac{1}{2},-1}(x) = \cosh x, \quad \Theta^1_{\frac{1}{2},-1}(x) = \frac{\sinh x}{x}, \quad \Theta^1_{\frac{3}{2},-1}(x) = 3\left( \frac{x\cosh x - \sinh x}{x^3} \right)$$

respectively.

Now, by using the above relationships in Corollary 2.2, Corollary 2.3, Corollary 2.5 and Theorem 2.6, respectively, we can give the following some interesting examples.

**Example 3.1.** The following assertions hold true.

- **i.** The function $x \mapsto \Theta^1_{-\frac{1}{2},1}(x) = \cos x$ is strictly log-concave on $\left[0, j_{-\frac{1}{2},1}\right) \cup T_1$, where $T_1 = \bigcup_{n\geq 1}\left(j_{-\frac{1}{2},2n}, j_{-\frac{1}{2},2n+1}\right)$ and $j_{-\frac{1}{2},n}$ denotes the $n$th positive zero of the equation $\cos x = 0$.
- **ii.** The function $x \mapsto \Theta^1_{\frac{1}{2},1}(x) = \frac{\sin x}{x}$ is strictly log-concave on $\left[0, j_{\frac{1}{2},1}\right) \cup T_2$, where $T_2 = \bigcup_{n\geq 1}\left(j_{\frac{1}{2},2n}, j_{\frac{1}{2},2n+1}\right)$ and $j_{\frac{1}{2},n}$ denotes the $n$th positive zero of the equation $\sin x = 0$.
- **iii.** The function $x \mapsto \Theta^1_{\frac{3}{2},1}(x) = 3\left(\frac{\sin x - x\cos x}{x^3}\right)$ is strictly log-concave on $\left[0, j_{\frac{3}{2},1}\right) \cup T_3$, where $T_3 = \bigcup_{n\geq 1}\left(j_{\frac{3}{2},2n}, j_{\frac{3}{2},2n+1}\right)$ and $j_{\frac{3}{2},n}$ denotes the $n$th positive zero of the equation $\tan x = x$.

**Example 3.2.** The following statements are valid.

- **i.** The function $x \mapsto \Theta^1_{-\frac{1}{2},-1}(x) = \cosh x$ is strictly log-concave on $\left[0, \epsilon_{-\frac{1}{2},1}\right) \cup S_1$, where $S_1 = \bigcup_{n\geq 1}\left(\epsilon_{-\frac{1}{2},2n}, \epsilon_{-\frac{1}{2},2n+1}\right)$ and $\epsilon_{-\frac{1}{2},n}$ denotes the $n$th positive zero of the equation $\cosh x = 0$.
- **ii.** The function $x \mapsto \Theta^1_{\frac{1}{2},-1}(x) = \frac{\sinh x}{x}$ is strictly log-concave on $\left[0, \epsilon_{\frac{1}{2},1}\right) \cup S_2$, where $S_2 = \bigcup_{n\geq 1}\left(\epsilon_{\frac{1}{2},2n}, \epsilon_{\frac{1}{2},2n+1}\right)$ and $\epsilon_{\frac{1}{2},n}$ denotes the $n$th positive zero of the equation $\sinh x = 0$.

**iii.** The function $x \mapsto \Theta^{1}_{\frac{3}{2},-1}(x) = 3\left(\frac{\sinh x - x\cosh x}{x^3}\right)$ is strictly log-concave on $\left[0, \epsilon_{\frac{3}{2},1}\right) \cup S_3$, where $S_3 = \bigcup_{n\geq 1}\left(\epsilon_{\frac{3}{2},2n}, \epsilon_{\frac{3}{2},2n+1}\right)$ and $\epsilon_{\frac{3}{2},n}$ denotes the $n$th positive zero of the equation $\tanh x = x$.

**Example 3.3.** The following assertions hold true.

**i.** The function $J_{-\frac{1}{2}}(x) = \sqrt{\frac{2}{\pi x}}\cos x$ is strictly log-concave on $\left[0, j_{-\frac{1}{2},1}\right) \cup T_1$.

**ii.** The function $J_{\frac{1}{2}}(x) = \sqrt{\frac{2}{\pi x}}\sin x$ is strictly log-concave on $\left[0, j_{\frac{1}{2},1}\right) \cup T_2$.

**iii.** The function $J_{\frac{3}{2}}(x) = \sqrt{\frac{2}{\pi x}}\left(\frac{\sin x}{x} - \cos x\right)$ is strictly log-concave on $\left[0, j_{\frac{3}{2},1}\right) \cup T_3$.

**iv.** The function $I_{-\frac{1}{2}}(x) = \sqrt{\frac{2}{\pi x}}\cosh x$ is strictly log-concave on $\left[0, \epsilon_{-\frac{1}{2},1}\right) \cup S_1$.

**v.** The function $I_{\frac{1}{2}}(x) = \sqrt{\frac{2}{\pi x}}\sinh x$ is strictly log-concave on $\left[0, \epsilon_{\frac{1}{2},1}\right) \cup S_2$.

**vi.** The function $I_{\frac{3}{2}}(x) = -\sqrt{\frac{2}{\pi x}}\left(\frac{\sinh x}{x} - \cosh x\right)$ is strictly log-concave on $\left[0, \epsilon_{\frac{3}{2},1}\right) \cup S_3$.

**Example 3.4.** The following functions

$$\Phi^{1}_{-\frac{1}{2},-1}(x) = x\tanh x, \qquad \Phi^{1}_{\frac{1}{2},-1}(x) = x\coth x - 1$$

and

$$\Phi^{1}_{\frac{3}{2},-1}(x) = \frac{(x^2+3)\sinh x - 3x\cosh x}{x\cosh x - \sinh x}$$

are increasing functions on $(0,\infty)$.

## References


[1] İ. Aktaş, *On some properties of hyper-Bessel and related functions*, TWMS J. App. and Eng. Math. **9** (1), 30–37, 2019.

[2] İ. Aktaş, *Partial sums of Hyper-Bessel function with applications*, Hacet. J. Math. Stat. **49** (1), 380–388, 2020.

[3] İ. Aktaş and Á. Baricz, *Bounds for the radii of starlikeness of some q-Bessel functions*, Results Math. **72** (1–2), 947–963, 2017.

[4] İ. Aktaş and H. Orhan, *Bounds for the radii of convexity of some q-Bessel functions*, Bull. Korean Math. Soc. **57** (2), 355–369, 2020.

[5] İ. Aktaş, Á. Baricz and H. Orhan, *Bounds for the radii of starlikeness and convexity of some special functions*, Turkish J. Math. **42** (1), 211–226, 2018.

[6] İ. Aktaş, Á. Baricz and S. Singh, *Geometric and monotonic properties of hyper-Bessel functions*, Ramanujan J. **51** (2), 275–295, 2020.

[7] İ. Aktaş, Á. Baricz and N. Yağmur, *Bounds for the radii of univalence of some special functions*, Math. Inequal. Appl. **20** (3), 825–843, 2017.

[8] Á. Baricz, *Geometric properties of generalized Bessel functions,* Publ. Math. Debrecen **73** (1–2), 155–178, 2008.

[9] Á. Baricz, *Generalized Bessel Functions of the First Kind*, Lecture Notes in Mathematics, Springer-Verlag, 2010.

[10] Á. Baricz and T.K. Pogány, *Functional inequalities of modified Struve functions*, Proc. Roy. Soc. Edinburgh Sect. A, **144** (5), 891–904, 2014.

[11] M. Biernacki and J. Krzyż, *On the monotonity of certain functionals in the theory of analytic functions*, Ann. Univ. Mariae Curie-Skłodowska Sect. A, **9**, 135–147, 1955.

[12] R. Díaz and E. Pariguan, *On hypergeometric functions and Pochhammer $k$-symbol*, Divulgaciones Mathemáticas, **15** (2), 179–192, 2007.

[13] G.R. Mohtasami Borzadaran and H.A. Mohtasami Borzadaran, *Log-concavity property for some well-known distributions*, Surv. Math. Appl. **6**, 203–219, 2011.

[14] S.R. Mondal and M.S. Akel, *Differential equation and inequalities of the generalized $k$-Bessel functions*, J. Inequal. Appl., 2018:175. doi: 10.1186/s13660-018-1772-1

[15] E. Toklu, *Radii of starlikeness and convexity of generalized $k$-Bessel functions*, arXiv:1902.09979, 2019.

Hacettepe Journal of
Mathematics & Statistics

Hacet. J. Math. Stat.
Volume 50 (1) (2021), 188 – 198
DOI : 10.15672/hujms.620711

Research Article

# A fixed point result for semigroups of monotone operators and a solution of discontinuous nonlinear functional-differential equations

Nabil Machrafi

*Mohammed V University in Rabat, Faculty of Sciences, Department of Mathematics, Team GrAAF, Laboratory LMSA, Center CeReMAR, B.P. 1014 RP, Rabat, Morocco*

**Abstract**

We improve some fixed point theorems by stating a fixed point result for semigroups of monotone operators in the setting of ordered Banach spaces with a normal cone. We illustrate the usefulness of our results by proving the existence and conditional unicity of a solution of an initial value problem for discontinuous nonlinear functional-differential equations under natural hypotheses involving the order structure of the underlying space.

**Mathematics Subject Classification (2020).** 47H10, 47H07, 34K05, 65L03

**Keywords.** fixed point, semitopological semigroup, ordered Banach space, normal cone, functional-differential equation, lower and upper solutions

## 1. Introduction

Since semigroups of self-mappings generalize powers of a self-mapping, it is natural to study their fixed points using the well-known technique of applying a contracting mapping principle to some power of that self-mapping. We will, in this paper, use the following generalized version of Banach contraction principle in the framework of partially ordered metric spaces; see also [13, Th. 2.1] for the first result given in this direction.

**Theorem 1.1** ([12, Theorems 2.2–2.5]). *Let $(X,d)$ be a complete metric space endowed with a partial ordering $\leq$. Let $T : X \to X$ be a nondecreasing (order-preserving) mapping with the contraction condition*

$$\exists k \in (0,1) \quad \forall x,y \in X \quad (x \leq y \Rightarrow d(Tx,Ty) \leq kd(x,y)). \tag{1.1}$$

*Assume that $(X,d,\leq)$ is such that one of the the following conditions holds:*

$$\begin{gathered}\text{for any nondecreasing sequence } (x_n) \subset X, \text{ if } x_n \to x \text{ in } X, \text{ then } x_n \leq x \ \forall n \in \mathbb{N},\\ \text{and there exists } x_0 \in X \text{ with } x_0 \leq Tx_0;\end{gathered} \tag{1.2}$$

$$\begin{gathered}\text{for any nonincreasing sequence } (x_n) \subset X, \text{ if } x_n \to x \text{ in } X, \text{ then } x \leq x_n \ \forall n \in \mathbb{N},\\ \text{and there exists } x_0 \in X \text{ with } Tx_0 \leq x_0.\end{gathered} \tag{1.3}$$

Email address: n.machrafi@um5s.net.ma, nmachrafi@gmail.com
Received: 16.09.2019; Accepted: 29.05.2020

*Assume furthermore that every pair of elements of $X$ has a lower or an upper bound. Then, $T$ has a unique fixed point $x^*$ in $X$ and the iterative sequence $(T^n x)$ converges to $x^*$ for every $x \in X$.*

Conditions (1.2) and (1.3) hold in the setting of ordered Banach spaces $E$, in which we will improve the following two known fixed point theorems when we restrict our attention to monotone operators $T$ on a closed set $C \subset E$ (this is so common since we deal in this case with operators preserving the order structure) with a lower (resp. upper) fixed point, i.e., $x_0 \in C$ with $x_0 \leq Tx_0$ (resp. $Tx_0 \leq x_0$). Fixed point results for operators having lower or upper fixed points were considered in the literature to solve ordinary as well as functional-differential equations with lower or upper solutions; see for instance [6,8,10,12].

**Theorem 1.2** ([15, Theorem 1], [16, Theorem 1.2.12 ])**.** *Let $(E, \|.\|)$ be a (real) Banach space with a transitive binary relation $\prec$ and a mapping $m : E \to E$ satisfying the following conditions:*

(1) $\theta \prec m(x)\,,\ x \in E$ *and $\theta$ denotes the zero element in $E$.*
(2) $\|m(x)\| = \|x\|\,,\ x \in E.$

*Furthermore, assume that the norm on $E$ is monotone, that is*

$$\theta \prec x \prec y \Rightarrow \|x\| \leq \|y\|\,, \quad x, y \in E. \tag{1.4}$$

*Let the operator $T : E \to E$ be given with the following contraction condition:*

$$m(Tx - Ty) \prec Am(x - y)\,,\ x, y \in E \tag{1.5}$$

*for some bounded linear operator $A$ on $E$ with the following properties:*

(3) $\theta \prec x \prec y \Rightarrow Ax \prec Ay.$
(4) $r(A) < 1$, *where $r(A)$ stands for the spectral radius of $A$.*

*Then, $T$ has a unique fixed point $x^*$ in $E$ and the iterative sequence $(T^n x)$ converges to $x^*$ for every $x \in E$.*

**Theorem 1.3** ([8, Theorem 3.1.14])**.** *Let $E$ be an ordered Banach space with a normal generating cone $E^+$ and $T : E \to E$ be an operator. If there exists a positive linear bounded operator $A : E \to E$, $\|A\| < 1$ such that*

$$-A(x - y) \leq Tx - Ty \leq A(x - y)\,, \quad x, y \in E,\ y \leq x, \tag{1.6}$$

*then $T$ has a unique fixed point $x^*$ in $E$ and the iterative sequence $(T^n x)$ converges to $x^*$ for every $x \in E$.*

We will improve the above theorems through the followings:

- We will consider semigroups of operators instead of a single one. In this case, the notion of a lower (resp. upper) fixed point of an operator will be naturally extended to the existence of an element with a monotone orbit for that semigroup of operators.

- As a less restrictive contraction condition than (1.5) and (1.6), we will consider the following one:

$$-A(x - y) \leq Tx - Ty \leq A(x - y)\,, \quad x, y \in C,\ y \leq x, \tag{1.7}$$

where $A$ is some positive bounded linear operator on $E$ with $r(A) < 1$. While conditions of Theorem 1.2 and Theorem 1.3 (see for the latter theorem [8, p 118]) imply necessarily the uniform continuity of the operator $T$, such operator is not necessarily continuous under conditions of our main theorems (hence, our results are stated for discontinuous operators in general).

- Comparing (1.5) and (1.7), one observes that the structure of the underlying space is relaxed by avoiding the mapping $m$ on $E$. In this case, monotonicity of the norm of $E$, or its weak alternative, namely, the normality of the cone of $E$ will suffice to state our fixed point results. This fact is motivated by the following example from [2, Example 3].

Let us recall first that a cone $K$ of an ordered normed vector space $(E, \|.\|, \leq)$ is said to be normal, if there exists a constant $N > 0$ such that

$$\theta \leq x \leq y \Rightarrow \|x\| \leq N \|y\|, \quad x, y \in E,$$

equivalently, if $E$ admits an equivalent monotone norm, i.e., an equivalent norm satisfying condition (1.4) for the partial order relation of $E$; see [1, Theorem 2.38]. Moreover, $K$ is said to be generating if the vector subspace generated by $K$ coincides with $E$, i.e., $E = K - K$. Lattice cones of the classical function spaces that are Banach lattices are special examples of normal and generating cones. More details on cone theory can be found in [1,8].

**Example 1.4.** Let $l_2$ be equipped with its standard inner product norm $\|.\|$ and the ordering $\leq$ given by the closed positive cone,

$$K = \left\{(x_k)_{k=1}^{\infty} : x_{2k-1} \geq k x_{2k} \geq 0 \text{ for all } k\right\}.$$

It follows from [2, Example 3] that the ordered normed vector space $E = K - K$ is a vector lattice that admits no equivalent absolute norm $|||.|||$ (i.e. $|||\,|x|\,||| = |||x|||$, $x \in E$, where $|x| := x \vee -x$ the join of $\{x, -x\}$), and hence no equivalent norm satisfying condition (2) of Theorem 1.2, where $m : E \to E$ is given by $m(x) = |x|$ (which is the so common case in function spaces). However, since $K$ is a subset of the standard cone $l_2^+ \subset l_2$ with respect to which the norm $\|.\|$ is monotone, the latter is also monotone with respect to the cone $K$.

The last section of the paper is devoted to the application of our results in solving the order counterpart of the following initial value problem for nonlinear functional-differential equations:

$$\begin{cases} u'(t) = f\left(t, u\left(h_1(t)\right), ..., u\left(h_r(t)\right), u'(t)\right) \text{ for a.e. } t \in [0, R] \ \ (\text{resp. } \forall t \in [0, R]\,); \\ u(0) = 0, \end{cases} \tag{1.8}$$

where $R > 0$, the unknown $u$ belongs to $AC\,[0, R]$ (resp. $C_1\,[0, R]$) the space of real-valued absolutely continuous (resp. continuously differentiable) functions on $[0, R]$,

$$(t, x_1, ..., x_{r+1}) \to f(t, x_1, ..., x_{r+1})$$

is a given real-valued function defined on the set $[0, R] \times \mathbb{R}^{r+1}$ and Lebesgue measurable with respect to $t$ for all $(x_1, ..., x_{r+1}) \in \mathbb{R}^{r+1}$, and $h_i : [0, R] \to [0, R]$ are continuous functions. This means solving Problem (1.8) under suitable hypotheses involving the order structure of the underlying space, while the same problem has been studied in [15, p 183] under hypotheses that do not involve this structure; see also [16, p 49].

The essential order-type hypothesis here is the existence of a lower or an upper solution of Problem (1.8) that will generate its solution. This problem is said to have a lower solution if there exists $u_0 \in AC\,[0, R]$ (resp. $C_1\,[0, R]$) such that

$$\begin{cases} u_0'(t) \leq f\left(t, u_0\left(h_1(t)\right), ..., u_0\left(h_r(t)\right), u_0'(t)\right) \text{ for a.e. } t \in [0, R] \ \ (\text{resp. } \forall t \in [0, R]\,); \\ u_0(0) \leq 0. \end{cases}$$

An upper solution is defined similarly with the reversed inequalities. Assuming the existence of a lower (resp. upper) solution $u_0$ of Problem (1.8), we are able to localize its solution in the order interval of functions satisfying $u_0(t) \leq u(t)$, $t \in [0, R]$ (resp. $u(t) \leq u_0(t)$, $t \in [0, R]$). Solutions of nonlinear integro-differential equations having a lower or an upper solution have been studied in the literature in many works; see for instance [8,10,12].

Also, the assumption of continuity of the function $f$ in [15, Theorem 3] is replaced here with its increasing monotonicity with respect to $(x_1, ..., x_{r+1})$ on $\mathbb{R}^{r+1}$ (see Sec. 4). The lack of continuity in problems for nonlinear functional-differential equations may appear in many situations and motivations for this kind of problems which were developed in [3, Chap. 4].

As a consequence, we prove the existence of a positive solution of Problem (1.8) under some natural hypotheses. Positive solutions of nonlinear integro-differential equations have been, in their turn, studied intensively in the literature; see for instance [4,7,11,14].

## 2. Preliminaries

Throughout the paper, $C$ will denote a nonempty and closed subset of a (non-trivial) ordered Banach space $E$, i.e., a real Banach space $E$ with an ordering $\leq$ induced by a closed cone in $E$ that will be denoted by $E^+$. The norm of $E$ will be denoted by $\|.\|$. For $x \in E$, the intervals $[x)$, $(x]$ are the closed sets defined by $[x) = \{z \in E : x \leq z\}$, $(x] = \{z \in E : z \leq x\}$. For two vectors $x, y \in E$, if $x \leq y$ or $y \leq x$ then $x$ and $y$ are said to be comparable.

The term operator on $C$ will mean a self-mapping of $C$. An operator $T$ on $C$ is said to be monotone, if it is order-preserving, i.e., for every $x, y \in C$,

$$x \leq y \Rightarrow Tx \leq Ty.$$

Note that a linear operator $A$ on $E$ is monotone if and only if $A$ is a positive operator, i.e.,

$$\theta \leq x \Rightarrow \theta \leq Ax, \quad x \in E.$$

In the sequel, the Banach space of bounded linear operators on $E$ and the set of positive bounded linear operators on $E$ will be denoted by $B(E)$ and $B^+(E)$ respectively. The spectral radius of $A \in B(E)$ is defined by

$$r(A) = \max\{|\lambda| : \lambda \in \sigma(A)\}$$

where $\sigma(A) := \sigma(A_c)$ the spectrum of $A_c$ and $A_c \in B(E_c)$ is the complexification of $A$ defined on the complex Banach space $E_c$, the complexification of $E$, by

$$A_c(x + iy) = Ax + iAy, \quad x, y \in E.$$

The spectral radius of $A$ is given in terms of its norm via the following formula (well-known as Gelfand's formula):

$$r(A) = \lim_{n\to\infty} \|A^n\|^{\frac{1}{n}} = \inf_{n\in\mathbb{N}} \|A^n\|^{\frac{1}{n}}.$$

In the setting of ordered Banach spaces, it is more convenient to calculate the spectral radius of a positive operator $A \in B(E)$ through its local spectral radius $r(A, x)$ at some element $x \in E$. This is defined for an operator $A \in B(E)$ by

$$r(A, x) = \limsup_{n\to\infty} \|A^n x\|^{\frac{1}{n}}.$$

The details are in the following lemma which will be useful in proving some forthcoming results.

**Lemma 2.1** ([5, Proposition 5])**.** *Let the cone $E^+$ be normal and generating, $A \in B^+(E)$, and $x_0 \in E^+\backslash\{\theta\}$ such that $A$ is bounded from above by $x_0$, that is, for every $x \in E^+$ there is a positive number $n(x)$ with $Ax \leq n(x)x_0$. Then, $r(A) = r(A, x_0)$.*

Let us consider now a commutative semitopological semigroup $S$, i.e., a semigroup with a Hausdorff topology such that for each $s \in S$, the mapping $t \to st$ is continuous from $S$ into $S$. This includes particularly the discrete case $S = (\mathbb{N} \cup \{0\}, +)$. We will use the notation $s^n$ to mean the nth power of $s \in S$. Since $S$ is commutative, then $S$ will be directed by the binary relation $\preceq$ defined on $S$ by the following:

$$s \preceq t \quad \text{if} \quad \{s\} \cup \overline{sS} \supseteq \{t\} \cup \overline{tS}. \tag{2.1}$$

More on semitopological semigroups and their properties can be found in [9].

A family $\mathfrak{T} = \{T_i\}_{i\in S}$ of operators on $C$ is said to be a semigroup if it satisfies the following:

(1) $T_s T_t = T_{st}$ for all $s, t \in S$;
(2) the mapping $s \to T_s x$ is continuous from $S$ into $C$, for every $x \in C$.

For a family $\mathfrak{T} = \{T_i\}_{i\in S}$ of operators on a nonempty set $C$, an element $x \in C$ is said to be a fixed point of $\mathfrak{T}$ if it is a fixed point of $T_i$ for every $i \in S$, i.e., $T_i x = x$ for every $i \in S$.

## 3. Main results

We formulate the following lemma, generalizing the lemma in [15, p 179], that will be used in the proof of our main result. Its proof is simple and therefore omitted.

**Lemma 3.1.** *A sufficient condition for a commuting family $\mathfrak{T} = \{T_i\}_{i\in S}$ of operators on a nonempty set to have a unique fixed point $x^*$ is that $x^*$ is the unique fixed point of some operator from the family $T_{i_0}$, where $i_0 \in S$.*

**Theorem 3.2.** *Let the cone $E^+$ be normal, $S$ be a commutative semitopological semigroup and $\mathfrak{T} = \{T_s\}_{s\in S}$ be a semigroup of monotone operators on $C$. Assume that*

*(1) there exists $s_0 \in S$ such that $T_{s_0}$ satisfies the contraction condition (1.7) with respect to some operator $A \in B^+(E)$;*
*(2) there exists $x_0 \in C$ such that its orbit $\{T_s x_0\}_{s\in S}$ is an increasing (resp. decreasing) net.*

*Then $\mathfrak{T}$ has a unique fixed point $x^*$ in $C_0 = C \cap [y_0)$ (resp. $C_0 = C \cap (y_0]$), where $y_0 = T_{s_0} x_0$. Moreover, if $C$ is bounded, then $\lim_s \|T_s x - x^*\| = 0$ for every $x \in C$, $x$ and $x^*$ are comparable.*

***Proof.*** Assume that the net $\{T_s x_0\}_{s\in S}$ is increasing (the other case can be dealt in a similar way). Then for every $s \in S$, $T_s$ maps $C_0$ into itself. Indeed, since $s_0 \preceq s s_0$, $s \in S$ and $T_s$ is monotone, then

$$T_{s_0} x_0 \le T_{s s_0} x_0 \le T_s x,$$

so $T_s x \in C_0$ for every $x \in C_0$. Now, if $x, y \in C$ with $y \le x$, one has

$$\theta \le T_{s_0} x - T_{s_0} y \le A\,(x - y)\,. \tag{3.1}$$

Again, since $T_{s_0} y \le T_{s_0} x$, then

$$\theta \le T_{s_0}^2 x - T_{s_0}^2 y \le A\,(T_{s_0} x - T_{s_0} y)\,.$$

Applying the operator $A$ to the inequality (3.1), we get

$$\theta \le T_{s_0}^2 x - T_{s_0}^2 y \le A^2\,(x - y)\,.$$

Proceeding inductively, we have

$$\theta \le T_{s_0}^n x - T_{s_0}^n y \le A^n\,(x - y) \tag{3.2}$$

for each $n \in \mathbb{N}$. Since the cone $E^+$ is normal, we may assume that the norm $\|.\|$ is monotone. It follows that

$$\|T_{s_0}^n x - T_{s_0}^n y\| \le \|A^n\,(x - y)\| \le \|A^n\|\,\|x - y\|\,, \tag{3.3}$$

for each $n \in \mathbb{N}$ and for every $x, y \in C$ with $y \le x$. Since $r\,(A) < 1$, by Gelfand's formula there exists $n_0 \in \mathbb{N}$ such that $\|A^{n_0}\| < 1$. Assuming $0 < \|A^{n_0}\| < 1$, then we are in position to apply Theorem 1.1 for the mapping $T_{s_0}^{n_0}|_{C_0} : C_0 \to C_0$ to infer that $T_{s_0}^{n_0}$ has a unique fixed point $x^*$ in $C_0$, where $C_0$ is endowed with the metric induced by the norm of $E$ and $y_0 \le T_{s_0}^{n_0} y_0$ (as the net $\{T_s x_0\}_{s\in S}$ is increasing). Since $T_s$ maps $C_0$ into itself for every $s \in S$, then we infer from Lemma 3.1 that $x^*$ is the unique fixed point of $\mathfrak{T}$ in $C_0$. Now, if $A^{n_0} = 0$ then it follows from (3.2) that $T_{s_0}^{n_0}$ is the constant mapping on $C_0$ equal to $T_{s_0}^{n_0} y_0$.

Since $y_0 \leq T_{s_0}^{n_0} y_0$, then clearly $T_{s_0}^{n_0} y_0$ is the unique fixed point of $T_{s_0}^{n_0}$ in $C_0$. Therefore, by the same above argument $T_{s_0}^{n_0} y_0$ is the unique fixed point of $\mathcal{T}$ in $C_0$.

Assume now that $C$ is bounded with a diameter $M \geq 0$. Let $x \in C$, $x$ and $x^*$ be comparable, $t_0 = s_0^{n_0}$, and $k = \|A^{n_0}\|$. We will show that for every $\varepsilon > 0$ there exists $n \in \mathbb{N}$ such that

$$\left\| T_{t_0^n s} x - x^* \right\| < \varepsilon \text{ for every } s \in S. \tag{3.4}$$

Let $\varepsilon > 0$ and choose $n \in \mathbb{N}$ with $k^n M < \varepsilon$. Since the operators of $\mathcal{T}$ are monotone, for every $s \in S$, from (3.3) one has

$$\begin{aligned} \left\| T_{t_0^n s} x - x^* \right\| &= \left\| T_{t_0^n s} x - T_{t_0^n s} x^* \right\| \\ &\leq k^n \left\| T_s x - T_s x^* \right\| \\ &\leq k^n M < \varepsilon, \end{aligned}$$

as desired. Now, if $s \in S$ with $t_0^n \preceq s$, then $s \in \{t_0^n\} \cup \overline{t_0^n S}$. Therefore, it suffices to show the case $s \in \overline{t_0^n S}$. Let $(s_\alpha) \subset S$ be a net with $\lim_\alpha t_0^n s_\alpha = s$. It follows from (3.4) and the continuity of $s \to T_s x$ from $S$ into $C$ that $\|T_s x - x^*\| \leq \varepsilon$, that is $\lim_s \|T_s x - x^*\| = 0$. This ends the proof. □

**Remark 3.3.** (1) It is easy to see that in the particular case $S = (\mathbb{N} \cup \{0\}, +)$ and $T_n := T^n$, $T : C \to C$ is a monotone operator, condition (2) of the above theorem is equivalent to $x_0$ is a lower (resp. upper) fixed point of $T$, and hence it is a natural extension of the existence of a lower (resp. upper) fixed point of a single operator to the case of a semigroup of operators.

(2) The hypothesis of boundedness in the above theorem is realised if there exist two elements $x_0, z_0 \in C$, $x_0 \leq z_0$, such that the orbits $\{T_s x_0\}_{s \in S}$, $\{T_s z_0\}_{s \in S}$ are an increasing and a decreasing nets respectively. Indeed, by the arguments as shown before, for every $s \in S$, $T_s$ maps the (closed) order interval $[T_{s_0} x_0, T_{s_0} z_0] \cap C$ into itself, and in this case $\mathcal{T}$ has a unique fixed point $x^*$ in $C_0 = C \cap [T_{s_0} x_0, T_{s_0} z_0]$. Note that each order interval $[x, y]$ of $E$, $x \leq y$, is bounded since the cone $E^+$ is normal; see [1, Theorem 2.40].

As a consequence of our main theorem, taking the particular case $S = (\mathbb{N} \cup \{0\}, +)$ and $T_n := T^n$, $T : C \to C$, we get an improvement of Theorem 1.2 and Theorem 1.3 in case the operator $T$ is assumed to be monotone with a lower (resp. upper) fixed point.

**Corollary 3.4.** *Let the cone $E^+$ be normal, $T$ be a monotone operator on $C$ with a lower (resp. upper) fixed point $x_0 \in C$. Assume that there exists a positive integer $n_0$ such that the power $T^{n_0}$ satisfies the contraction condition* (1.7) *with respect to some operator $A \in B^+(E)$. Then, $T$ has a unique fixed point $x^*$ in $C_0 = C \cap [x_0)$ (resp. $C_0 = C \cap (x_0]$). Moreover, if $C$ is bounded, then the iterative sequence $(T^n x)$ converges to $x^*$ for every $x \in C$, $x$ and $x^*$ are comparable.*

## 4. An initial value problem for functional-differential equations

In this section, we illustrate the applicability of our results by using Corollary 3.4 to solve Problem (1.8) under some natural order-type hypotheses. So, we will assume that

$(H_1)$ Problem (1.8), $u \in AC[0, R]$ admits a lower solution $u_0$ with $u_0'(t) \geq a$ for almost all $t \in [0, R]$ and for some $a \in \mathbb{R}^+$, and the function

$$f(., u_0(h_1(.)) - u_0(0), ..., u_0(h_r(.)) - u_0(0), u_0'(.))$$

belongs to $L_1[0, R]$, the Lebesgue space of real-valued integrable functions on $[0, R]$.

Moreover, the function $f$ is assumed to be increasing with respect to $(x_1, ..., x_{r+1})$ on $\mathbb{R}^{r+1}$, that is

$(H_2)$ for all $(t, x_1, ..., x_{r+1}), (t, y_1, ..., y_{r+1}) \in [0, R] \times \mathbb{R}^{r+1}$ we have

$$x_1 \leq y_1,\ x_2 \leq y_2, ...,\ x_{r+1} \leq y_{r+1} \Rightarrow f(t, x_1, ..., x_{r+1}) \leq f(t, y_1, ..., y_{r+1}).$$

On the other hand, the hypothesis in [15, Theorem 3] consisting of the standard Lipschitz condition of $f$

$$\left|f\left(t,x_1,...,x_{r+1}\right)-f\left(t,y_1,...,y_{r+1}\right)\right|\leq\sum_{i=1}^{r+1}L_i\left(t\right)\left|x_i-y_i\right| \tag{4.1}$$

for all $\left(t,x_1,...,x_{r+1}\right),\left(t,y_1,...,y_{r+1}\right)\in[0,R]\times\mathbb{R}^{r+1}$, the $L_i$'s are continuous and positive functions on the interval $[0,R]$, will be weakened to the Lipschitz condition:

$(H_3)$ for all $\left(t,x_1,...,x_{r+1}\right),\left(t,y_1,...,y_{r+1}\right)\in[0,R]\times\mathbb{R}^{r+1}$ with $x_1\geq y_1\geq x_0$, $x_2\geq y_2\geq x_0,...,\ x_{r+1}\geq y_{r+1}\geq x_0$, we have

$$f\left(t,x_1,...,x_{r+1}\right)-f\left(t,y_1,...,y_{r+1}\right)\leq\sum_{i=1}^{r+1}L_i\left(t\right)\left(x_i-y_i\right) \tag{4.2}$$

where $x_0=\min\left(a,aH\right)$ and $H=\min_{i=1}^{r}\min_{t\in[0,R]}h_i\left(t\right).$

Finally, we make the estimate $h_i\left(t\right)\leq t,\ t\in[0,R]$ satisfying by the functions $h_i$ in [15, Theorem 3] less restrictive. This is

$(H_4)$ the functions $h_i$, $L_i$ satisfy the estimates
(a) $h\left(t\right):=\sup_{i=1}^{r}h_i\left(t\right)\leq ct^{\alpha},\ t\in[0,R]$, where $c>0,\ \alpha\in(0,1]$ are some constants;
(b) $L_r\left(1-\alpha\right)c^{\frac{1}{1-\alpha}}+L_{r+1}<1$ if $\alpha\neq1$ and $L_{r+1}<1$ if $\alpha=1$ and $c\leq1$, where $L_r:=\max_{i=1}^{r}\left(\max_{[0,R]}L_i\left(t\right)\right)r$ and $L_{r+1}:=\max_{[0,R]}L_{r+1}\left(t\right)$.

The following theorem provides a solution of Problem (1.8), $u\in AC\left[0,R\right]$ under the above-mentioned hypotheses.

**Theorem 4.1.** *Under the hypotheses $(H_1)-(H_4)$, Problem (1.8), $u\in AC\left[0,R\right]$ has a unique solution with $u'\left(t\right)\geq u_0'\left(t\right)$ for a.e. $t\in[0,R]$ (and hence $u\left(t\right)\geq u_0\left(t\right),\ t\in[0,R]$).*

In what follows, we let $E=L_1\left[0,R\right]$ be endowed with its standard norm and the ordering $\leq$ induced by the cone

$$E^+=\left\{u:u\left(t\right)\geq0 \text{ for a.e. } t\in[0,R]\right\}.$$

We will use the following lemma that provides an estimation of the spectral radius of a Voltera-type operator on $E$.

**Lemma 4.2.** *Let $A\in B\left(E\right)$ be the operator defined by*

$$Au\left(t\right)=L\int_0^{h(t)}u\left(s\right)ds,\quad t\in[0,R],$$

*where $L>0$ is some constant. Then, $r\left(A\right)\leq L\left(1-\alpha\right)c^{\frac{1}{1-\alpha}}$ if $\alpha\neq1$ and $r\left(A\right)=0$ if $\alpha=1$ and $c\leq1$.*

***Proof.*** Let $u_1\in E$ be the constant function equal to 1. Since the cone $E^+$ is normal and generating, $A\in B^+\left(E\right)$ and $Au\leq L\left\|u\right\|u_1$ for every $u\in E^+$, then by Lemma 2.1 $r\left(A\right)=r\left(A,u_1\right)$. Now, for $t\in[0,R]$ we see from $h\left(t\right)\leq ct^{\alpha}$ that

$$A\left(u_1\right)\left(t\right)=L\int_0^{h(t)}u_1\left(s\right)ds\leq L\int_0^{ct^{\alpha}}ds=Lct^{\alpha}.$$

Again, we have

$$A^2\left(u_1\right)\left(t\right)=L\int_0^{h(t)}Au_1\left(s\right)ds\leq L\int_0^{ct^{\alpha}}Lcs^{\alpha}ds=L^2\frac{c^{1+\alpha+1}}{\alpha+1}t^{\alpha(\alpha+1)},$$

and by induction, we have

$$A^n(u_1)(t) \le L^n \frac{c^{1+\alpha+1+\ldots+\alpha^{n-1}+\ldots+\alpha+1}}{(\alpha+1)(\alpha^2+\alpha+1)\ldots(\alpha^{n-1}+\ldots+\alpha+1)} t^{\alpha\left(\alpha^{n-1}+\ldots+\alpha+1\right)}$$

for every $n \ge 1$. Therefore, we have

$$\|A^n(u_1)\| \le L^n \frac{c^{1+\alpha+1+\ldots+\alpha^{n-1}+\ldots+\alpha+1}}{(\alpha+1)(\alpha^2+\alpha+1)\ldots(\alpha^{n}+\ldots+\alpha+1)} R^{\alpha^n+\ldots+\alpha+1}$$

for every $n \ge 1$. Let $a_n$ be the right hand side in the last inequality. If $\alpha \ne 1$, then

$$\frac{a_{n+1}}{a_n} = L \frac{c^{\alpha^n+\ldots+\alpha+1}}{\alpha^{n+1}+\ldots+\alpha+1} R^{\alpha^{n+1}} \to L(1-\alpha)c^{\frac{1}{1-\alpha}}$$

as $n \to \infty$, from which we get $a_n^{\frac{1}{n}} \to L(1-\alpha)c^{\frac{1}{1-\alpha}}$ as $n \to \infty$. Hence,

$$r(A,u_1) = \limsup_{n\to\infty} \|A^n(u_1)\|^{\frac{1}{n}} \le \lim_{n\to\infty} a_n^{\frac{1}{n}} = L(1-\alpha)c^{\frac{1}{1-\alpha}},$$

as desired. Similarly, we have $r(A) = 0$ if $\alpha = 1$ and $c \le 1$. □

**Remark 4.3.** The above lemma remains similarly true in the standard Banach lattice $E = C([0,R])$ of real-valued continuous functions on $[0,R]$, where the ordering of functions is the pointwise ordering.

***Proof of Theorem 4.1.*** It is easily shown that Problem (1.8), $u \in AC[0,R]$ and $u'(t) \ge u_0'(t)$ for a.e. $t \in [0,R]$ is equivalent to the following integral-functional equation:

$$\begin{cases} z(t) = f(t, \int_0^{h_1(t)} z(s)\,ds, \int_0^{h_2(t)} z(s)\,ds, \ldots, \int_0^{h_r(t)} z(s)\,ds, z(t)) \\ z(t) \ge z_0(t), \text{ for a.e. } t \in [0,R],\ z, z_0 \in E, \end{cases} \tag{4.3}$$

where $u(t) = \int_0^t z(s)\,ds$ and $u_0(t) = \int_0^t z_0(s)\,ds + u_0(0)$, $t \in [0,R]$. Define the operator $T$ on the interval $[z_0)$ of $E$ by

$$Tz(t) = f(t, \int_0^{h_1(t)} z(s)\,ds, \int_0^{h_2(t)} z(s)\,ds, \ldots, \int_0^{h_r(t)} z(s)\,ds, z(t)), \quad t \in [0,R]. \tag{4.4}$$

It follows easily from the hypotheses $(H_1)-(H_3)$ that $T$ is a monotone operator on $[z_0)$ with $z_0$ as a lower fixed point. Furthermore, for every $z, w \in [z_0)$ with $w \le z$, from $(H_3)$, one has

$$\begin{aligned} Tz(t) - Tw(t) &\le \sum_{i=1}^{r} L_i(t) \int_0^{h_i(t)} (z-w)(s)\,ds + L_{r+1}(z-w)(t) \\ &\le L_r \int_0^{h(t)} (z-w)(s)\,ds + L_{r+1}(z-w)(t) \\ &= (A + L_{r+1}I)(z-w)(t), \end{aligned}$$

for almost all $t \in [0,R]$, where $I$ is the identity operator of $E$ and $A \in B^+(E)$ is the operator of Lemma 4.2 with respect to the constant $L_r$. Since $\sigma(A + L_{r+1}I) = \sigma(A) + L_{r+1}$, it follows from Lemma 4.2 and the hypothesis $(H_4)$ that

$$r(A + L_{r+1}I) \le r(A) + L_{r+1} < 1.$$

Therefore, applying Corollary 3.4, we see that $T$ has a unique fixed point $z \in [z_0)$, that is $z$ is the unique solution of (4.3). This completes the proof. □

We get as a consequence a positive solution of Problem (1.8), $u \in AC\,[0,R]$ under natural hypotheses.

**Corollary 4.4.** *Assume that the hypotheses* $(H_2)$, $(H_4)$ *are satisfied, that the Lipschitz condition* (4.2) *is satisfied for all* $(t,x_1,...,x_{r+1})$, $(t,y_1,...,y_{r+1}) \in [0,R]\times\mathbb{R}_+^{r+1}$ *with* $x_1 \geq y_1$, $x_2 \geq y_2,...,$ $x_{r+1} \geq y_{r+1}$, *and that the function* $f\,(.,0,...,0)$ *belongs to* $(L_1[0,R])^+$. *Then, Problem* (1.8), $u \in AC\,[0,R]$ *has a unique solution with a positive derivative (and hence the solution* $u$ *is itself positive).*

***Proof.*** It follows from the hypotheses that Problem (1.8), $u \in AC\,[0,R]$ has the (everywhere) null function as a lower solution. The desired conclusion follows from Theorem 4.1. ☐

In case the function $f$ is assumed to be continuous on $[0,R]\times\mathbb{R}^{r+1}$, we get similar results for Problem (1.8), $u \in C_1\,[0,R]$. We omit the proofs since they follow by similar arguments applied in the setting of the standard Banach lattice $E = C\,[0,R]$.

**Theorem 4.5.** *Assume that* $f$ *is continuous on* $[0,R]\times\mathbb{R}^{r+1}$, *that Problem* (1.8), $u \in C_1\,[0,R]$ *has a lower solution* $u_0$ *with* $u_0'\,(t) \geq a$ *for every* $t \in [0,R]$ *and for some* $a \in \mathbb{R}^+$, *and that the hypotheses* $(H_2)-(H_4)$ *are satisfied. Then, Problem* (1.8), $u \in C_1\,[0,R]$ *has a unique solution with* $u'\,(t) \geq u_0'\,(t)$, $t \in [0,R]$ *(and hence* $u\,(t) \geq u_0\,(t)$, $t \in [0,R]$*).*

**Corollary 4.6.** *Assume that* $f$ *is continuous on* $[0,R]\times\mathbb{R}^{r+1}$, *that the hypotheses* $(H_2)$, $(H_4)$ *are satisfied, that the Lipschitz condition* (4.2) *is satisfied for all* $(t,x_1,...,x_{r+1})$, $(t,y_1,...,y_{r+1}) \in [0,R]\times\mathbb{R}_+^{r+1}$ *with* $x_1 \geq y_1$, $x_2 \geq y_2,...,$ $x_{r+1} \geq y_{r+1}$, *and that the function* $f\,(.,0,...,0)$ *belongs to* $(C[0,R])^+$. *Then, Problem* (1.8), $u \in C_1\,[0,R]$ *has a unique solution with a positive derivative (and hence the solution* $u$ *is itself positive).*

## 5. Concluding remarks

(1) The case $\alpha = 1$ and $c \leq 1$ in Theorem 4.5 is the order counterpart of [15, Theorem 3]. Moreover, since there are many functions $f$ which satisfy the Lipschitz condition (4.2) without the standard one (4.1), we see the need of Corollary 3.4 instead of Theorem 1.2 to get a fixed point of the operator defined by (4.4). Indeed, as a simple example, consider the discontinuous function $f : [0,R]\times\mathbb{R}^2 \to \mathbb{R}$ defined by

$$f(t,x,y) = \begin{cases} \frac{1}{2}x+1 & \text{if } x > -1, \\ 1-x^2 & \text{if } x \leq -1, \end{cases}$$

and let $h_1(t) = t$ for every $t \in [0,R]$. In this case, the null function on $[0,R]$ is a lower solution of Problem (1.8), $u \in AC\,[0,R]$, all the hypotheses $(H_1)-(H_4)$ are fulfilled, and Problem (1.8), $u \in AC\,[0,R]$ and $u'\,(t) \geq 0$ for a.e $t \in [0,R]$ reduces to the simple initial value problem

$$u'\,(t) = \frac{1}{2}u\,(t) + 1 \text{ for a.e. } t \in [0,R],\ u(0) = 0,$$

which has a unique solution $u \in AC\,[0,R]$ with a positive derivative.

(2) On the other hand, the following easy situation illustrates the need of Corollary 3.4 instead of Theorem 1.1 or Theorem 1.3. Let $\mathbb{R}^2$ be endowed with their Euclidean norm and coordinatewise ordering. Let $T : \mathbb{R}^2 \to \mathbb{R}^2$ be equal to $A = \begin{pmatrix} 0 & 1 \\ 0 & 0 \end{pmatrix}$. Clearly, all conditions of Corollary 3.4 are fulfilled and $(0,0)$ is the unique fixed point of $T$. In particular, the pair $T, A$ satisfies the contraction condition (1.7). However, it is easy to see that the contraction condition of Theorem 1.1 fails and that for any operator $B \in B^+\,(\mathbb{R}^2)$ with $\|B\| < 1$, the pair $T, B$ does not satisfy the contraction condition of Theorem 1.3.

(3) Theorems 4.1 and 4.5 can be stated under slight suitable modifications if we assume the existence of an upper solution instead of a lower solution of Problem (1.8).

(4) **The monotone iterative sequences of approximate solutions for Problem** (1.8)**.**

This is for the case when this problem admits simultaneously a lower and an upper solutions $u_0$ and $v_0$ with $a \leq u_0'(t) \leq v_0'(t) \leq b$ for almost all (resp. for all) $t \in [0, R]$ and for some $a, b \in \mathbb{R}^+$. If the other hypotheses of Theorem 4.1 (resp. 4.5) hold true for both the lower and the upper solutions $u_0$ and $v_0$ (with the suitable modifications for the upper solution $v_0$) and if we keep the notations of the proof of Theorem 4.1, then the operator $T$ is now defined on the order interval $[z_0, w_0]$ of $L_1[0, R]$ (resp. $C_1[0, R]$), where $w_0$ is generated similarly from the upper solution $v_0$, with $z_0$ and $w_0$ as a lower and an upper fixed points, respectively. In this case, the latter two theorems provide a unique solution $u$ of Problem (1.8) with $u_0'(t) \leq u'(t) \leq v_0'(t)$ for almost all (resp. for all) $t \in [0, R]$ (and hence $u_0(t) \leq u(t) \leq v_0(t)$, $t \in [0, R]$). Define the sequences of functions on $[0, R]$ $\left(f_{(n)}\right)$ and $\left(f^{(n)}\right)$ by $f_{(0)}(t) = z_0(t)$, $f^{(0)}(t) = w_0(t)$, and inductively by

$$\begin{aligned} f_{(n)}(t) &= f(t, \int_0^{h_1(t)} f_{(n-1)}(s)\, ds, ..., \int_0^{h_r(t)} f_{(n-1)}(s)\, ds, f_{(n-1)}(t)), \\ f^{(n)}(t) &= f(t, \int_0^{h_1(t)} f^{(n-1)}(s)\, ds, ..., \int_0^{h_r(t)} f^{(n-1)}(s)\, ds, f^{(n-1)}(t)). \end{aligned}$$

It follows easily from $f_{(n)} = T^n z_0$, $f^{(n)} = T^n w_0$, and Corollary 3.4 that the monotone sequences of functions $\left(\int_0^{\cdot} f_{(n)}(s)\, ds\right)$ and $\left(\int_0^{\cdot} f^{(n)}(s)\, ds\right)$ converge uniformly on $[0, R]$ to the solution of Problem $(1.8)$.

## References


[1] C.D. Aliprantis and R. Tourky, *Cones and Duality*, Graduate Studies in Mathematics, Vol. **84**, Amer. Math. Soc., Providence, Rhode Island, 2007.

[2] J.M. Borwein and D.T. Yost, *Absolute norms on vector lattices*, Proc. Edinb. Math. Soc. **27**, 215–222, 1984.

[3] S. Carl and S. Heikkilä, *Fixed Point Theory in Ordered Sets and Applications*, Springer, New York, 2011.

[4] L.H. Erbe, W. Krawcewicz, and D. Guo, *Positive solutions of two-point boudary value problems for nonlinear integro-differential equations in Banach spaces*, Differ. Equ. Dyn. Syst. **2**, 161–171, 1994.

[5] K.H. Förster and B. Nagy, *On the local spectral radius of a nonnegative element with respect to an irreducible operator*, Acta Sci. Math. (Szeged), **55**, 155–166, 1991.

[6] D. Guo and V. Lakshmikantham, *Nonlinear Problems in Abstract Cones*, Academic Press, Inc., Boston, 1988.

[7] D. Guo, *Multiple positive solutions of impulsive Fredholm integral equations and applications*, J. Math. Anal. Appl. **173**, 318–324, 1993.

[8] D. Guo, Y.J. Cho and J. Zhu, *Partial Ordering Methods in Nonlinear Problems*, Nova Science Publishers Inc., Hauppauge, 2004.

[9] R.D. Holmes and A.T. Lau, *Nonexpansive actions of topological semigroups and fixed points*, J. Lond. Math. Soc. (2), **5**, 330–336, 1972.

[10] G.S. Ladde, V. Lakshmikantham and A.S. Vatsala, *Monotone Iterative Techniques for Nonlinear Differential Equations*, Pitman, Boston, 1985.

[11] Z. Liang, *Some properties of nonlinear operators and positive solutions of a class of integral equations*, Acta Math. Sinica (Chin. Ser.), **40**, 345–350, 1997.

[12] J.J. Nieto and R. Rodriguez-Lopez, *Contractive mapping theorems in partially ordered sets and applications to ordinary differential equations*, Order, **22** 223–239, 2005.

[13] A.C.M. Ran and M.C.B. Reurings, *A fixed point theorem in partially ordered sets and some applications to matrix equations*, Proc. Amer. Math. Soc. **132**, 1435–1443, 2004.

[14] C.A. Stuart, *Positive solutions of a nonlinear integral equation*, Math. Ann. **192**, 119–124, 1971.

[15] M. Zima, *A certain fixed point theorem and its applications to integral-functional equations*, Bull. Aust. Math. Soc. **46**, 179–186, 1992.

[16] M. Zima, *Positive Operators in Banach Spaces and Their Applications*, Wydawnictwo Uniwersytetu Rzeszowskiego, Rzeszow, 2005.

Hacettepe Journal of Mathematics & Statistics

Hacet. J. Math. Stat.
Volume 50 (1) (2021), 199 – 215
DOI : 10.15672/hujms.683997

Research Article

# Weighted variable exponent grand Lebesgue spaces and inequalities of approximation

İsmail Aydın*[1], Ramazan Akgün[2]

[1] Sinop University, Faculty of Arts and Sciences, Department of Mathematics, Sinop, Turkey
[2] Balikesir University, Faculty of Arts and Sciences, Department of Mathematics, Balıkesir, Turkey

## Abstract

In this paper we discuss and investigate trigonometric approximation in weighted grand variable exponent Lebesgue spaces. We also prove the direct and inverse theorems in these spaces.

**Mathematics Subject Classification (2020).** 46E35, 43A15, 46E30

**Keywords.** weighted grand variable exponent Lebesgue, Sobolev and Lipschitz space, maximal operator, modulus of smoothness, best approximation, Jackson and inverse theorems, K-functional

## 1. Introduction

In 1992, T. Iwaniec and C. Sbordone [22] introduced the grand Lebesgue spaces $L^{p)}(\Omega)$, $1 < p < \infty$, on bounded sets $\Omega \subset \mathbb{R}^d$, with applications to differential equations. A generalized version $L^{p),\theta}(\Omega)$ appeared in L. Greco, T. Iwaniec and C. Sbordone [18]. During last years these spaces were intensively studied for various applications (see, e.g., [1, 16–18, 20, 22, 23]). The variable exponent Lebesgue spaces (or generalized Lebesgue spaces) $L^{p(.)}$ appeared in literature for the first time in 1931 with an article written by Orlicz [25]. Kováčik and Rákosník [24] introduced the variable exponent Lebesgue space $L^{p(.)}(\mathbb{R}^d)$ and Sobolev space $W^{k,p(.)}(\mathbb{R}^d)$ in higher dimensional Euclidean spaces. There are several applications of these spaces, such as, elastic mechanics, electrorheological fluids, image restoration and nonlinear degenerated partial differential equations (see [10,11,14]). The spaces $L^{p(.)}(\mathbb{R}^d)$ and $L^{p}(\mathbb{R}^d)$ have many common properties, such as Banach space, reflexivity, separability, uniform convexity, Hölder inequalities and embeddings. A crucial difference between $L^{p(.)}(\mathbb{R}^d)$ and $L^{p}(\mathbb{R}^d)$ is that the variable exponent Lebesgue space is not invariant under translation in general, see [13, Lemma 2.3] and [24, Example 2.9]. For more information see [10,14]. The grand variable exponent Lebesgue space $L^{p(.),\theta}(\Omega)$ was introduced and studied by Kokilasvili and Meski [23]. In their studies they established the boundedness of maximal and Calderon operators in these spaces. The space $L^{p(.),\theta}(\Omega)$ is not reflexive, separable, rearrangement invariant and translation invariant. There are several published papers about direct and inverse theorems of approximation theory in some function spaces weighted, variable or non-weighted, see, [2–8, 12, 19, 21].

*Corresponding Author.
Email addresses: iaydin@sinop.edu.tr (İ. Aydın), rakgun@balikesir.edu.tr (R. Akgün)
Received: 03.02.2020; Accepted: 29.05.2020

In this study we obtain some inequalities involving trigonometric polynomial approximation in a certain subspace of the weighted variable exponent grand Lebesgue space $L_w^{p(.),\theta}$. Also we give some basic properties of these spaces. Finally, we prove some direct and inverse theorems of approximation in $L_w^{p(.),\theta}$.

## 2. Notations and preliminaries

In this section, we give some essential definitions, theorems and remarks for weighted grand variable exponent Lebesgue spaces.

**Definition 2.1.** Let $\mathbb{T} := [0, 2\pi]$ and let $p(.) : \mathbb{T} \longrightarrow [1, \infty)$ be a measurable $2\pi$-periodic function such that

$$1 \leq p^- = \operatorname*{ess\,inf}_{x\in\mathbb{T}} p(x) \leq \operatorname*{ess\,sup}_{x\in\mathbb{T}} p(x) := p^+ < \infty.$$

Assume that $p(.)$ satisfies the local log-continuity condition, i.e., there exists a constant $C > 0$ such that the inequality

$$|p(x) - p(y)| \leq \frac{C}{-\log|x-y|}$$

holds for all $x, y \in \mathbb{T}$ with $|x - y| \leq \frac{1}{2}$ (briefly $p(.) \in P(\mathbb{T})$). We also define a subclass

$$P_0(\mathbb{T}) = \left\{p(.) \in P(\mathbb{T}) : 1 < p^-\right\}.$$

**Definition 2.2.** Let $p(.) \in P(\mathbb{T})$. Variable exponent Lebesgue space $L^{p(.)} := L^{p(.)}(\mathbb{T})$ is defined as the set of all measurable, $2\pi$-periodic functions $f$ on $\mathbb{T}$ such that $\varrho_{p(.)}(\lambda f) < \infty$ for some $\lambda > 0$, equipped with the Luxemburg norm

$$\|f\|_{p(.)} = \inf\left\{\lambda > 0 : \varrho_{p(.)}\left(\frac{f}{\lambda}\right) \leq 1\right\},$$

where $\varrho_{p(.)}(f) = \int\limits_{\mathbb{T}} |f(x)|^{p(x)}\,dx$. The space $L^{p(.)}$ is a Banach space with the norm $\|.\|_{p(.)}$. Moreover, the norm $\|.\|_{p(.)}$ coincides with the usual Lebesgue norm $\|.\|_p$ whenever $p(.) = p$ is a constant function. If $p^+ < \infty$, then $f \in L^{p(.)}$ if and only if $\varrho_{p(.)}(f) < \infty$.

**Definition 2.3.** A Lebesgue measurable and locally integrable function $w : \mathbb{T} \longrightarrow (0, \infty)$ is called a weight function. Suppose that $p(.) \in P(\mathbb{T})$. The weighted modular is defined by

$$\varrho_{p(.),w}(f) = \int\limits_{\mathbb{T}} |f(x)|^{p(x)}\,w(x)dx.$$

The weighted variable exponent Lebesgue space $L_w^{p(.)} := L_w^{p(.)}(\mathbb{T})$ consists of all measurable functions $f$ on $\mathbb{T}$ for which $\|f\|_{p(.),w} = \left\|fw^{\frac{1}{p(.)}}\right\|_{p(.)} < \infty$. Also, $L_w^{p(.)}$ is a uniformly convex Banach space, thus reflexive.

**Remark 2.4.** Let $w$ be a weight on $\mathbb{T}$ and $p(.) \in P(\mathbb{T})$.

(i) Relations between the modular $\varrho_{p(.),w}(.)$ and $\|.\|_{p(.),w}$ are as follows:

$$\min\left\{\varrho_{p(.),w}(f)^{\frac{1}{p^-}}, \varrho_{p(.),w}(f)^{\frac{1}{p^+}}\right\} \leq \|f\|_{p(.),w} \leq \max\left\{\varrho_{p(.),w}(f)^{\frac{1}{p^-}}, \varrho_{p(.),w}(f)^{\frac{1}{p^+}}\right\},$$

$$\min\left\{\|f\|_{p(.),w}^{p^+}, \|f\|_{p(.),w}^{p^-}\right\} \leq \varrho_{p(.),w}(f) \leq \max\left\{\|f\|_{p(.),w}^{p^+}, \|f\|_{p(.),w}^{p^-}\right\}.$$

(ii) If $0 < C \leq w$, then we have $L_w^{p(.)} \hookrightarrow L^{p(.)}$, since one gets easily that

$$C\int\limits_{\mathbb{T}} |f(x)|^{p(x)}\,dx \leq \int\limits_{\mathbb{T}} |f(x)|^{p(x)}\,w(x)dx$$

and $C\|f\|_{p(.)} \leq \|f\|_{p(.),w}$ (see [9]). Moreover, due to $|\mathbb{T}| < \infty$ and $1 \leq p(.)$ we have $L_w^{p(.)}(\mathbb{T}) \hookrightarrow L^{p(.)}(\mathbb{T}) \hookrightarrow L^1(\mathbb{T})$.

**Definition 2.5.** Let $\theta > 0$ and $p(.) \in P(\mathbb{T})$. The grand variable exponent Lebesgue space, $L^{p(.),\theta}$, is the class of all measurable functions $f$ for which

$$\|f\|_{p(.),\theta} := \sup_{0<\varepsilon<p^- -1} \varepsilon^{\frac{\theta}{p^- -\varepsilon}} \|f\|_{p(.)-\varepsilon} < \infty.$$

When $p(.) = p$ is a constant function, these spaces coincide with the grand Lebesgue spaces $L^{p),\theta}(\mathbb{T})$.

**Definition 2.6.** Let $w$ be a weight on $\mathbb{T}$ and $p(.) \in P(\mathbb{T})$. The weighted grand variable exponent Lebesgue spaces $L_w^{p(.),\theta} := L_w^{p(.),\theta}(\mathbb{T})$ is the class of all measurable functions $f$ for which

$$\|f\|_{p(.),w,\theta} := \sup_{0<\varepsilon<p^- -1} \varepsilon^{\frac{\theta}{p^- -\varepsilon}} \|f\|_{p(.)-\varepsilon,w} < \infty.$$

**Remark 2.7.** Let $w$ be a weight on $\mathbb{T}$ and $p(.) \in P(\mathbb{T})$.

(i) It is easy to see that the following continuous embeddings hold

$$L^{p(.)} \hookrightarrow L^{p(.),\theta} \hookrightarrow L^{p(.)-\varepsilon} \hookrightarrow L^1,\ 0 < \varepsilon < p^- - 1$$

due to $|\mathbb{T}| < \infty$ (see [12, 23]).

(ii) For $f \in L_w^{p(.),\theta}(\mathbb{T})$ the norm equality $\|f\|_{p(.),w,\theta} = \left\|fw^{\frac{1}{p(.)}}\right\|_{p(.),\theta}$ is not valid in $L_w^{p(.),\theta}(\mathbb{T})$ (see [17]).

**Example 2.8.** Let $\alpha > 0$, $\theta = 1$, $p(.) = p$ =constant and choose a weight $w(x) = x^\alpha$. If we take $f(x) = x^\beta$ for $\beta > -\alpha - 1$, then we have $f \in L_w^{p)}(0,1)$. But, $\left(fw^{\frac{1}{p}}\right)^{p-\varepsilon}$ is not integrable in $(0,1)$ for any $0 < \varepsilon < p - 1$ and so $fw^{\frac{1}{p}} \notin L^{p)}(0,1)$ (see [16]).

**Proposition 2.9** (Nesting Property). *If $0 < C \leq w$, $p(.) \in \mathrm{P}(\mathbb{T})$ and $\theta_1 < \theta_2$, then we have the following continuous embeddings*

$$L_w^{p(.)} \hookrightarrow L_w^{p(.),\theta_1} \hookrightarrow L_w^{p(.),\theta_2} \hookrightarrow L_w^{p(.)-\varepsilon} \hookrightarrow L^{p(.)-\varepsilon} \hookrightarrow L^1,\ 0 < \varepsilon < p^- - 1$$

*due to $|\mathbb{T}| < \infty$ (see [12, 23]).*

**Remark 2.10.** Let $w$ be a weight on $\mathbb{T}$ and $p(.) \in P(\mathbb{T})$. There are several differences between $L_w^{p(.)}$ and $L_w^{p(.),\theta}$. For instance, the set of the bounded functions is not dense in $L_w^{p(.),\theta}$, and the closure of $L^\infty(\mathbb{T})$ in the norm of $L_w^{p(.),\theta}$ can be characterized by the functions $f$ such that

$$\lim_{\varepsilon\to 0} \sup \varepsilon^{\frac{\theta}{p^- -\varepsilon}} \|f\|_{p(.)-\varepsilon,w} = 0$$

(see [1]). Moreover, the closure of simple functions is not dense in $L_w^{p(.),\theta}$. Also, the space $L_w^{p(.),\theta}$ is not reflexive, not separable and not rearrangement invariant. Since the closure of $L_w^{p(.)}$ in $L_w^{p(.),\theta}$ does not coincide with the latter space, that is, $L_w^{p(.)}$ is not dense in $L_w^{p(.),\theta}$, then we redefine this set in the following theorem as a subspace of $L_w^{p(.),\theta}$ (see [12, 23]).

**Theorem 2.11.** *Let $w$ be a weight on $\mathbb{T}$ and $p(.) \in \mathrm{P}(\mathbb{T})$. The following statements hold:*

(i) *The space $L_w^{p(.),\theta}$ is complete.*

(ii) *The closure of $L_w^{p(.)}$ in $L_w^{p(.),\theta}$ consists of functions $f$, which belong to $L_w^{p(.),\theta}$, for which $\lim_{\varepsilon\to 0} \varepsilon^{\frac{\theta}{p^- -\varepsilon}} \|f\|_{p(.)-\varepsilon,w} = 0$.*

***Proof.*** (i) Let $(f_n)_{n\in\mathbb{N}}$ be a Cauchy sequence in $L_w^{p(.),\theta}$. Then for all $\eta > 0$ there exists $N(\eta) > 0$ such that, whenever $n, m > N(\eta)$ we have

$$\varepsilon^{\frac{\theta}{p^- - \varepsilon}} \|f_n - f_m\|_{p(.)-\varepsilon,w} < \frac{\eta}{3} \tag{2.1}$$

for any $\varepsilon \in (0, p^- - 1)$. Therefore $(f_n)_{n\in\mathbb{N}}$ is a Cauchy sequence in $L_w^{p(.)-\varepsilon}$ for arbitrary $\varepsilon \in (0, p^- - 1)$. Then there is an $f$ in $L_w^{p(.)-\varepsilon}$ such that

$$\|f - f_n\|_{p(.)-\varepsilon,w} \to 0 \tag{2.2}$$

for every $\varepsilon \in (0, p^- - 1)$ (note that the function $f$ is unique for all $\varepsilon \in (0, p^- - 1)$, see [23]). For $n > N(\eta)$, there is an $\varepsilon_0(n) \in (0, p^- - 1)$ such that

$$\|f - f_n\|_{p(.),w,\theta} \leq \varepsilon_0(n)^{\frac{\theta}{p^- - \varepsilon}} \|f - f_n\|_{p(.)-\varepsilon_0(n),w} + \frac{\eta}{3} \tag{2.3}$$

by using the definition of the supremum. Moreover, there exists $N_1 \in \mathbb{N}$ such that for $m > N_1$ we have

$$\varepsilon^{\frac{\theta}{p^- - \varepsilon_0(n)}} \|f - f_m\|_{p(.)-\varepsilon_0(n),w} \leq \frac{\eta}{3} \tag{2.4}$$

due to (2.2). If we combine (2.3), (2.4) and (2.1), then we get

$$\begin{aligned}
\|f - f_n\|_{p(.),w,\theta} &\leq \varepsilon_0(n)^{\frac{\theta}{p^- - \varepsilon}} \|f - f_n\|_{p(.)-\varepsilon_0(n),w} + \frac{\eta}{3} \\
&\leq \varepsilon_0(n)^{\frac{\theta}{p^- - \varepsilon}} \|f_n - f_m\|_{p(.)-\varepsilon_0(n),w} + \varepsilon_0(n)^{\frac{\theta}{p^- - \varepsilon}} \|f - f_m\|_{p(.)-\varepsilon_0(n),w} + \frac{\eta}{3} \\
&\leq \frac{\eta}{3} + \frac{\eta}{3} + \frac{\eta}{3} = \eta
\end{aligned}$$

for $n > N(\eta)$ and $m > N_1$. This completes the proof of (i).

(ii) Denote by $\left[L_w^{p(.)}\right]_{p(.),w,\theta}$ the closure of $L_w^{p(.)}$ in $L_w^{p(.),\theta}$. For $f \in \left[L_w^{p(.)}\right]_{p(.),w,\theta}$ we can obtain that there is a sequence $(f_n)_{n\in\mathbb{N}}$ in $L_w^{p(.)}$ such that $\|f - f_n\|_{p(.),w,\theta} \to 0$ by the definition of the closure set. Then, for fixed $\delta > 0$, there exists $N = N(\delta) > 0$ such that, whenever $n > N(\delta)$ we obtain

$$\|f - f_n\|_{p(.),w,\theta} < \frac{\delta}{2}. \tag{2.5}$$

It is well-known that the continuous embedding $L_w^{q(.)}(\mathbb{T}) \hookrightarrow L_w^{p(.)}(\mathbb{T})$ holds if and only if $q(.) \geq p(.)$ because of $|\mathbb{T}| < \infty$ [24]. Hence we get

$$\varepsilon^{\frac{\theta}{p^- - \varepsilon}} \|f_n\|_{p(.)-\varepsilon,w} \leq (1 + |\mathbb{T}|)\varepsilon^{\frac{\theta}{p^- - \varepsilon}} \|f_n\|_{p(.),w} \to 0 \tag{2.6}$$

as $\varepsilon \to 0$. If we take $\varepsilon_0 > 0$ such that $0 < \varepsilon < \varepsilon_0$, then we can write

$$\varepsilon^{\frac{\theta}{p^- - \varepsilon}} \|f_n\|_{p(.)-\varepsilon,w} < \frac{\delta}{2}. \tag{2.7}$$

Finally, if we collect (2.5) and (2.7), then we have

$$\begin{aligned}
\varepsilon^{\frac{\theta}{p^- - \varepsilon}} \|f\|_{p(.)-\varepsilon,w} &\leq \varepsilon^{\frac{\theta}{p^- - \varepsilon}} \|f - f_n\|_{p(.)-\varepsilon,w} + \varepsilon^{\frac{\theta}{p^- - \varepsilon}} \|f_n\|_{p(.)-\varepsilon,w} \\
&\leq \|f - f_n\|_{p(.),w,\theta} + \frac{\delta}{2} \leq \delta
\end{aligned}$$

as $\varepsilon \to 0$. $\square$

**Definition 2.12.** We denote the closure of $L_w^{p(.)}$ by $L_{0,w}^{p(.),\theta}$. For $f \in L_{0,w}^{p(.),\theta}(\mathbb{T})$ we have

$$\lim_{\varepsilon \to 0} \varepsilon^{\frac{\theta}{p^- - \varepsilon}} \|f\|_{p(.)-\varepsilon,w} = 0$$

by the last theorem (see [12]).

**Proposition 2.13.** *Let $w$ be a weight on $\mathbb{T}$ and $p(.) \in \mathrm{P}(\mathbb{T})$. Then, $\left(L_w^{p(.),\theta}(\mathbb{T}), \|.\|_{p(.),w,\theta}\right)$ is a Banach function space (see [1]).*

We denote the Hardy-Littlewood maximal operator $Mf$ of $f$ by

$$Mf(x) = \sup_I \frac{1}{|I|} \int_I |f(t)|\, dt, \quad t \in \mathbb{T},$$

where the supremum is taken over all intervals $I$ whose length is less than $2\pi$.

The boundedness of the Hardy-Littlewood maximal operator $M$ on the space $L_W^{p(.),\theta}$, $\theta > 0$, $p(.) \in P_0(\mathbb{T})$, was proved in the following theorem for power weights of the form $W(x) = |x - x_0|^\gamma$, where $x_0 \in \mathbb{T}$, $-1 < \gamma < p(x_0) - 1$.

**Theorem 2.14.** ([17]) *Let $p(.) \in \mathrm{P}_0(\mathbb{T})$, $x_0 \in (-\pi, \pi)$, $\theta > 0$, and $-1 < \gamma < p(x_0) - 1$. Then the operator $M$ is bounded in $L_W^{p(.),\theta}$, i.e. for all $f \in L_W^{p(.),\theta}$ there exists a $C > 0$ such that the inequality*

$$\|Mf\|_{p(.),W,\theta} \le C \|f\|_{p(.),W,\theta}$$

*holds with $W(x) = |x - x_0|^\gamma$.*

In what follows, all weights $W$ considered will be power weight of the form $W(x) = |x - x_0|^\gamma$ satisfying the hypothesis of the last theorem.

Since $W(x) = |x - x_0|^\gamma$ satisfies the $A_{p(.)}$ condition of Muckenhoupt weights, then we have the continuous embedding $L_W^{p(.),\theta} \hookrightarrow L^1(\mathbb{T})$ [8]. This means that we can consider the corresponding Fourier series of $f \in L_W^{p(.),\theta}$ given by

$$f(x) \sim \frac{a_0(f)}{2} + \sum_{k=1}^{\infty} \left(a_k(f) \cos kx + b_k(f) \sin kx\right), \tag{2.8}$$

where $a_0(f) = \pi^{-1} \int_{\mathbb{T}} f(t)\, dt$ and

$$a_k(f) = \pi^{-1} \int_{\mathbb{T}} f(t) \cos kt dt, \quad b_k(f) = \pi^{-1} \int_{\mathbb{T}} f(t) \sin kt dt, \quad k = 1, 2, \dots .$$

The $n$-th partial sums of the series (2.8) is defined by

$$S_n(x, f) := \sum_{k=0}^{n} A_k(f)(x) = \frac{a_0(f)}{2} + \sum_{k=1}^{n} \left(a_k(f) \cos kx + b_k(f) \sin kx\right).$$

**Definition 2.15.** Let $W(x) = |x - x_0|^\gamma$, $\theta > 0$, $p(.) \in P_0(\mathbb{T})$, $r = 1, 2, ...$ and $f \in L_{0,W}^{p(.),\theta}$. Then the $r$-th modulus of smoothness $\Omega_r(f, .)_{p(.),W,\theta} : [0, \infty) \to [0, \infty)$ is defined as

$$\Omega_r(f, \delta)_{p(.),W,\theta} = \sup_{0<h\le\delta} \|\rho_h^r f\|_{p(.),W,\theta}, \; r \in \mathbb{N},$$

where

$$\rho_h^r f(x) := \frac{1}{h} \int_0^h \triangle_t^r f(x) dt,$$

$$\triangle_t^r f(x) := \sum_{s=0}^{r} (-1)^{r+s+1} b_{r,s} f(x + st), \quad t > 0,$$

and $b_{r,s}$ are binomial coefficients.

**Remark 2.16.** Using Theorem 2.14 we get

$$\sup_{0<h\leq\delta}\|\rho_h^r f\|_{p(.),W,\theta}\leq C\|f\|_{p(.),W,\theta}<\infty.$$

This shows that the function $\Omega_r(f,\delta)_{p(.),W,\theta}$ is well defined.

**Remark 2.17.** The modulus of smoothness $\Omega_r(f,\delta)_{p(.),W,\theta}$ has the following properties:

(i) $\Omega_r(f,\delta)_{p(.),W,\theta}$ is a non-negative, non-decreasing function of $\delta>0$.
(ii) $\Omega_r(f_1+f_2,.)_{p(.),W,\theta}\leq\Omega_r(f_1,.)_{p(.),W,\theta}+\Omega_r(f_2,.)_{p(.),W,\theta}$.
(iii) $\lim_{\delta\to 0}\Omega_r(f,\delta)_{p(.),W,\theta}=0$.

**Definition 2.18.** The best approximation error $E_n(f)_{p(.),W,\theta}$ of $f\in L_{0,W}^{p(.),\theta}$ is defined by

$$E_n(f)_{p(.),W,\theta}:=\inf\left\{\|f-T_n\|_{p(.),W,\theta}:T_n\in\Pi_n\right\}$$

where $\Pi_n$ is the set of trigonometric polynomials of degree at most $n$.

**Definition 2.19.** The Sobolev space $\mathcal{W}_{p(.),W,\theta}^r$ is the class of functions $f\in L_W^{p(.),\theta}$ such that $f^{(r)}\in L_W^{p(.),\theta}$ and

$$\|f\|_{p(.),W,\theta}^r=\|f\|_{p(.),W,\theta}+\left\|f^{(r)}\right\|_{p(.),W,\theta}<\infty,$$

for $r=1,2,\ldots$. Also the space $\mathcal{W}_{p(.),W,\theta}^r$ is a Banach space with respect to $\|.\|_{p(.),W,\theta}^r$. We define

$$\mathcal{W}_{0,p(.),W,\theta}^r=\left\{f:f\in L_{0,W}^{p(.),\theta}\cap\mathcal{W}_{p(.),W,\theta}^r\right\}.$$

## 3. Main results

The main results of this paper are the following theorems.

**Theorem 3.1.** *Let $W(x)=|x-x_0|^\gamma$, $\theta>0$, $p(.)\in\mathrm{P}_0(\mathbb{T})$ and $r,n\in\mathbb{N}$. If $f\in\mathcal{W}_{0,p(.),W,\theta}^r$, then*

$$E_n(f)_{p(.),W,\theta}\leq\frac{c}{n^r}E_n\left(f^{(r)}\right)_{p(.),W,\theta}$$

*with a constant $c>0$ independent of $n$.*

**Corollary 3.2.** *Under the conditions of Theorem 3.1,*

$$E_n(f)_{p(.),W,\theta}\leq\frac{c}{n^r}\left\|f^{(r)}\right\|_{p(.),W,\theta}$$

*with a constant $c>0$ independent of $n=0,1,2,3,\ldots$.*

**Theorem 3.3.** *Let $W(x)=|x-x_0|^\gamma$, $\theta>0$, $p(.)\in\mathrm{P}_0(\mathbb{T})$ and $r,n\in\mathbb{N}$. If $f\in L_{0,W}^{p(.),\theta}$, then*

$$E_n(f)_{p(.),W,\theta}\leq c\Omega_r\left(f,\frac{1}{n}\right)_{p(.),W,\theta}$$

*with a constant $c>0$ independent of $n$.*

**Theorem 3.4.** *Let $W(x)=|x-x_0|^\gamma$, $\theta>0$, $p(.)\in\mathrm{P}_0(\mathbb{T})$ and $r,n\in\mathbb{N}$. If $f\in L_{0,W}^{p(.),\theta}$, then*

$$\Omega_r\left(f,\frac{1}{n}\right)_{p(.),W,\theta}\leq\frac{c}{n^r}\sum_{k=0}^{n}(k+1)^{r-1}E_k(f)_{p(.),W,\theta}$$

*with a constant $c>0$ independent of $n$.*

To prove main results we need some lemmas and propositions given below.

**Lemma 3.5.** *Let $W(x)=|x-x_0|^{\gamma}$, $\theta>0$, $p(.)\in \mathrm{P}_0(\mathbb{T})$ and $r\in\mathbb{N}$. If $f\in \mathcal{W}^r_{0,p(.),W,\theta}$, then*

$$\Omega_r(f,\delta)_{p(.),W,\theta}\le c\delta^r\left\|f^{(r)}\right\|_{p(.),W,\theta}$$

*with a constant $c>0$ independent of $n$.*

***Proof.*** Since

$$\triangle_t^r f(.)=\int_0^t\int_0^t\dots\int_0^t f^{(r)}(.+t_1+\dots+t_r)\,dt_1\dots dt_r,$$

applying ($r$ times) the generalized Minkowski's inequality we get

$$\left\|\frac{1}{h}\int_0^h\triangle_t^r f\,dt\right\|_{p(.),W,\theta}\le\frac{c_1(p)}{h}\int_0^h\|\triangle_t^r f\|_{p(.),W,\theta}\,dt$$

$$\begin{aligned}
&\le h^r\frac{c_1(p)}{h^{r+1}}\int_0^h\left\|\int_0^t\dots\int_0^t f^{(r)}(.+t_1+\dots+t_r)\,dt_1\dots dt_r\right\|_{p(.),W,\theta}dt\\
&= h^r\frac{c_1(p)}{h}\int_0^h\left\|\frac{1}{h}\int_0^t\left|\frac{1}{h^{r-1}}\int_0^t\dots\int_0^t f^{(r)}(.+t_1+\dots+t_r)\,dt_1\dots dt_{r-1}\right|dt_r\right\|_{p(.),W,\theta}dt\\
&\le h^r\frac{c_2(p)}{h}\int_0^h\left\|\frac{1}{h^{r-1}}\int_0^t\dots\int_0^t f^{(r)}(.+t_1+\dots+t_{r-1})\,dt_1\dots dt_{r-1}\right\|_{p(.),W,\theta}dt\\
&\le \dots\le h^r\frac{c_3(p,r)}{h}\int_0^h\left\|\left\{\frac{1}{h}\int_0^h f^{(r)}(.+t_1)\,dt_1\right\}\right\|_{p(.),W,\theta}dt\\
&\le c_4(p,r)h^r\left\|f^{(r)}\right\|_{p(.),W,\theta}\frac{1}{h}\int_0^h dt=c_4(p,r)h^r\left\|f^{(r)}\right\|_{p(.),W,\theta},
\end{aligned}$$

and taking supremum on $0<h\le\delta$, we obtain the required inequality

$$\Omega_r(f,\delta)_{p(.),W,\theta}\le c\delta^r\left\|f^{(r)}\right\|_{p(.),W,\theta}.$$

□

**Definition 3.6.** Let $W(x)=|x-x_0|^{\gamma}$, $\theta>0$, $p(.)\in P_0(\mathbb{T})$, $r\in\mathbb{N}$ and $f\in L^{p(.),\theta}_{0,W}$. We define Peetre's $K$-functional as

$$K_r(f,\delta)_{p(.),W,\theta}:=\inf\left\{\|f-g\|_{p(.),W,\theta}+\delta^r\left\|g^{(r)}\right\|_{p(.),W,\theta}:g\in\mathcal{W}^r_{0,p(.),W,\theta},\ \delta>0\right\}.$$

**Theorem 3.7.** *Let $W(x)=|x-x_0|^{\gamma}$, $\theta>0$, $p(.)\in \mathrm{P}_0(\mathbb{T})$, $r\in\mathbb{N}$. If $f\in L^{p(.),\theta}_{0,W}$, then there are some constants $c_6$, $c_7>0$ independent of $\delta$ such that*

$$c_6\Omega_r(f,\delta)_{p(.),W,\theta}\le K_r(f,\delta)_{p(.),W,\theta}\le c_7\Omega_r(f,\delta)_{p(.),W,\theta}.$$

***Proof.*** Let $f\in L^{p(.),\theta}_{0,W}$ and $g\in\mathcal{W}^r_{0,p(.),W,\theta}$. By Lemma 3.5 and Remark 2.17,

$$\begin{aligned}
\Omega_r(f,\delta)_{p(.),W,\theta}&\le\Omega_r(f-g,\delta)_{p(.),W,\theta}+\Omega_r(g,\delta)_{p(.),W,\theta}\\
&\le c\left(\|f-g\|_{p(.),W,\theta}+\delta^r\left\|g^{(r)}\right\|_{p(.),W,\theta}\right),
\end{aligned}$$

and taking infimum with respect to $g \in \mathcal{W}^r_{0,p(.),W,\theta}$ in the last inequality we have

$$\Omega_r(f,\delta)_{p(.),W,\theta} \leq cK_r(f,\delta)_{p(.),W,\theta}.$$

In order to prove the reverse of the last inequality we define the function

$$f_{r,\delta}(x) = \frac{2}{\delta}\int_{\frac{\delta}{2}}^{\delta}\Big(\frac{1}{h^r}\sum_{s=0}^{r-1}(-1)^{r+s+1}\binom{r}{s}\int_0^h \dots \int_0^h f(x+\frac{r-s}{r}[t_1+\dots+t_r])dt_1\dots dt_r\Big)dh \quad (3.1)$$

for $\delta > 0$ and $r \geq 1$. Then, differentiating $r-1$ times and setting $t := \frac{r-s}{r}t_r$ we see that

$$\begin{aligned}
&\left\{\int_0^h \dots \int_0^h f\left(x+\frac{r-s}{r}[t_1+\dots+t_r]\right)dt_1\dots dt_r\right\}^{(r-1)} \\
=\ &\left\{\int_0^h \left(\frac{r}{r-s}\right)^{r-1}\sum_{m=0}^{r-1}\binom{r-1}{m}(-1)^{r+m} f(x+\frac{r-s}{r}t_r + m\frac{r-s}{r}h)dt_r\right\} \\
=\ &\int_0^h \left(\frac{r}{r-s}\right)^{r-1}\triangle^{r-1}_{\frac{r-s}{r}h} f(x+t)\,dt,
\end{aligned}$$

and then by (3.1)

$$f^{(r-1)}_{r,\delta}(x) := \frac{2}{\delta}\int_{\frac{\delta}{2}}^{\delta}\frac{1}{h^r}\left\{\sum_{s=0}^{r-1}\int_x^{x+\frac{r-s}{r}h}(-1)^{r+s+1}\binom{r}{s}\triangle^{r-1}_{\frac{r-s}{r}h}f(t)dt\right\}dh. \quad (3.2)$$

Now we prove $f^{(r)}_{r,\delta} \in L^{p(.),\theta}_{0,W}$. Differentiating the relation (3.2) we obtain

$$f^{(r)}_{r,\delta}(x) := \frac{2}{\delta}\int_{\frac{\delta}{2}}^{\delta}\frac{1}{h^r}\left\{\sum_{s=0}^{r-1}(-1)^{r+s+1}\binom{r}{s}\left(\frac{r}{r-s}\right)^r\triangle^r_{\frac{r-s}{r}h}f(x)\right\}dh$$

and denoting $t := \frac{r-s}{r}h$ we have

$$\begin{aligned}
\left|f^{(r)}_{r,\delta}(x)\right| &\leq \frac{2^{r+1}}{\delta^r}\sum_{s=0}^{r-1}\binom{r}{s}\left(\frac{r}{r-s}\right)^r\left|\frac{1}{\delta}\int_{\frac{\delta}{2}}^{\delta}\triangle^r_{\frac{r-s}{r}h}f(x)dh\right| \\
&= \frac{2^{r+1}}{\delta^r}\sum_{s=0}^{r-1}\binom{r}{s}\left(\frac{r}{r-s}\right)^r\left|\frac{1}{\frac{r-s}{r}\delta}\int_{\frac{r-s}{r}(\frac{\delta}{2})}^{\frac{r-s}{r}\delta}\triangle^r_t f(x)dt\right|
\end{aligned}$$

$$\leq \frac{2^{r+1}}{\delta^r}\sum_{s=0}^{r-1}\binom{r}{s}\left(\frac{r}{r-s}\right)^r\left\{\left|\frac{1}{\frac{r-s}{r}\delta}\int_0^{\frac{r-s}{r}\delta}\triangle^r_t f(x)dt\right| + \left|\frac{1}{\frac{r-s}{r}\delta}\int_0^{\frac{r-s}{r}(\frac{\delta}{2})}\triangle^r_t f(x)dt\right|\right\},$$

which implies the inequality

$$\left\|f^{(r)}_{r,\delta}\right\|_{p(.),W,\theta} \leq 2c(r)\delta^{-r}\Omega_r(f,\delta)_{p(.),W,\theta} \leq c_5(p,r)\|f\|_{p(.),W,\theta}. \quad (3.3)$$

Since $f \in L^{p(.),\theta}_{0,W}$, then $f^{(r)}_{r,\delta} \in L^{p(.),\theta}_{0,W}$.

Let $f \in L_{0,W}^{p(.),\theta}$. For $\delta > 0$ and $r = 1, 2, ...$, we have

$$|f_{r,\delta}(x) - f(x)| = \left| \frac{2}{\delta} \int\limits_{\frac{\delta}{2}}^{\delta} \left\{ \frac{1}{h^r} \int\limits_0^h \dots \int\limits_0^h \triangle^r_{\frac{t_1+\dots+t_r}{r}} f(x) dt_1 \dots dt_r \right\} dh \right|$$

and by the generalized Minkowski's inequality

$$\|f_{r,\delta} - f\|_{p(.),W,\theta} \le c_6(p,r) \frac{2}{\delta} \int\limits_{\frac{\delta}{2}}^{\delta} \left\{ \frac{1}{h^{r-1}} \int\limits_0^h \dots \int\limits_0^h \left\| \frac{1}{h} \int\limits_0^h \triangle^r_{\frac{t_1+\dots+t_r}{r}} f dt_1 \right\|_{p(.),W,\theta} dt_2 \dots dt_r \right\} dh$$

$$= c_6(p,r) \frac{2}{\delta} \int\limits_{\frac{\delta}{2}}^{\delta} \left\{ \frac{1}{h^{r-1}} \int\limits_0^h \dots \int\limits_0^h \left\| \frac{1}{h} \int\limits_{t_2+\dots+t_r}^{h+t_2+\dots+t_r} \triangle^r_{\frac{t}{r}} f dt \right\|_{p(.),W,\theta} dt_2 \dots dt_r \right\} dh. \tag{3.4}$$

Since

$$\begin{aligned} \left\| \frac{1}{h} \int\limits_{t_2+\dots+t_r}^{h+t_2+\dots+t_r} \triangle^r_{\frac{t}{r}} f dt \right\|_{p(.),W,\theta} &= \left\| \frac{1}{h} \left( \int\limits_0^{h+t_2+\dots+t_r} \triangle^r_{\frac{t}{r}} f dt - \int\limits_0^{t_2+\dots+t_r} \triangle^r_{\frac{t}{r}} f dt \right) \right\|_{p(.),W,\theta} \\ &\le \left\| \frac{1}{(h+t_2+\dots+t_r)/r} \int\limits_0^{(h+t_2+\dots+t_r)/r} \triangle^r_{\frac{t}{r}} f dt \right\|_{p(.),W,\theta} \\ &\quad + \left\| \frac{1}{(t_2+\dots+t_r)/r} \int\limits_0^{(t_2+\dots+t_r)/r} \triangle^r_{\frac{t}{r}} f dt \right\|_{p(.),W,\theta} \end{aligned}$$

$$\begin{aligned} &= \sup_{(h+t_2+\dots+t_r)/r \le \delta} \left\| \frac{1}{(h+t_2+\dots+t_r)/r} \int\limits_0^{(h+t_2+\dots+t_r)/r} \triangle^r_{\frac{t}{r}} f dt \right\|_{p(.),W,\theta} \\ &\quad + \sup_{(t_2+\dots+t_r)/r \le \delta} \left\| \frac{1}{(t_2+\dots+t_r)/r} \int\limits_0^{(t_2+\dots+t_r)/r} \triangle^r_{\frac{t}{r}} f dt \right\|_{p(.),W,\theta} \\ &= \Omega_r(f,\delta)_{p(.),W,\theta} + \Omega_r(f,\delta)_{p(.),W,\theta} = 2\Omega_r(f,\delta)_{p(.),W,\theta}, \end{aligned} \tag{3.5}$$

then combining (3.4) and (3.5) we have

$$\begin{aligned} \|f_{r,\delta} - f\|_{p(.),W,\theta} &\le c(p,r) \frac{2}{\delta} \int\limits_{\frac{\delta}{2}}^{\delta} \left\{ \frac{1}{h^{r-1}} \int\limits_0^h \dots \int\limits_0^h \Omega_r(f,\delta)_{p(.),W,\theta} dt_2 \dots dt_r \right\} dh \\ &\le c(p,r) \Omega_r(f,\delta)_{p(.),W,\theta} \frac{2}{\delta} \int\limits_{\frac{\delta}{2}}^{\delta} dh = c(p,r) \Omega_r(f,\delta)_{p(.),W,\theta} \end{aligned} \tag{3.6}$$

Finally, if we use (3.3) and (3.6), then we get

$$\begin{aligned} K_r(f,\delta)_{p(.),W,\theta} &\le \|f_{r,\delta} - f\|_{p(.),W,\theta} + \delta^r \left\| f^{(r)}_{r,\delta} \right\|_{p(.),W,\theta} \\ &\le c_7 \Omega_r(f,\delta)_{p(.),W,\theta}. \end{aligned}$$

This completes the proof. □

The following lemma is a Bernstein inequality for $L_W^{p(.),\theta}$.

**Lemma 3.8.** *Let $W(x)=|x-x_0|^{\gamma}$, $\theta>0$, $p(.)\in \mathrm{P}_0(\mathbb{T})$, $r\in\mathbb{N}$. If $T_n$ is a trigonometric polynomial of degree at most $n$, then*

$$\|T_n'\|_{p(.),W,\theta}\le cn\|T_n\|_{p(.),W,\theta}.$$

***Proof.*** It is well-known that

$$\sup_n|\sigma_n(x,f)|\le cMf(x)$$

with a constant $c>0$ independent of $f$ and $x\in\mathbb{T}$, where $\sigma_n(x,f)$ is the Cesàro means for a function $f\in L_W^{p(.),\theta}$ [27]. Using Theorem 2.14 we have

$$\left\|\sup_n|\sigma_n(.,f)|\right\|_{p(.),W,\theta}\le c\|f\|_{p(.),W,\theta}. \tag{3.7}$$

Since

$$T_n(x)=\frac{1}{\pi}\int_T T_n(t)D_n(t-x)dt,\text{ with }D_n(t)=\frac{1}{2}+\sum_{j=1}^{n}\cos jt,$$

it is well-known that

$$T_n'(x)=2n\sigma_{n-1}(x,T_n)$$

and, hence,

$$\|T_n'\|_{p(.),W,\theta}\le 2n\|\sigma_{n-1}(.,|T_n|)\|_{p(.),W,\theta}\le 2cn\|T_n\|_{p(.),W,\theta}.$$

This completes the proof. □

Lemma 3.8 can be generalized for $r$-th derivative of $T_n$. For this we need a Minkowski's inequality for integrals. The following results were proved, when $W\equiv1$, by Danelia and Kokilashvili [12, Proposition 2.4]. The same proof also suits our case below.

**Lemma 3.9.** *Let $W(x)=|x-x_0|^{\gamma}$, $\theta>0$, $p(.)\in\mathrm{P}_0(\mathbb{T})$, and $f\in L_{0,W}^{p(.),\theta}$. If $f(x,y)$ a measurable function on $\mathbb{T}\times\mathbb{T}$, then, the following integral inequality holds*

$$\left\|\int_{\mathbb{T}}f(.,y)dy\right\|_{p(.),W,\theta}\le C\int_{\mathbb{T}}\|f(.,y)\|_{p(.),W,\theta}dy.$$

As a corollary of the last two lemmas we get:

**Corollary 3.10.** *Let $W(x)=|x-x_0|^{\gamma}$, $\theta>0$, $p(.)\in\mathrm{P}_0(\mathbb{T})$ and $r\in\mathbb{N}$. If $T_n$ is a trigonometric polynomial of degree at most $n$, then*

$$\left\|T_n^{(r)}\right\|_{p(.),W,\theta}\le cn^r\|T_n\|_{p(.),W,\theta}.$$

## 4. Proof of main results

Let $n\in\mathbb{N}$ and

$$D_nf(x):=\frac{1}{\pi}\int_{\mathbb{T}}f(x-t)J_{2,\lfloor\frac{n}{2}\rfloor+1}(t)dt \tag{4.1}$$

be the Jackson operator (polynomial), where $\lfloor\frac{n}{2}\rfloor$ denotes the integer part of a real number $\frac{n}{2}$, and $J_{2,n}$ is the Jackson kernel

$$J_{2,n}(x):=\frac{1}{\varkappa_{2,n}}\left(\frac{\sin(nx/2)}{\sin(x/2)}\right)^4,\quad \varkappa_{2,n}:=\frac{1}{\pi}\int_{-\pi}^{\pi}\left(\frac{\sin(nt/2)}{\sin(t/2)}\right)^4dt.$$

It is known that ([15, p.147])

$$\frac{3}{2\sqrt{2}}n^3\le\varkappa_{2,n}\le\frac{5}{2\sqrt{2}}n^3.$$

Jackson kernel $J_{2,n}$ satisfies the relations

$$\left.\begin{array}{l}\frac{1}{\pi}\int_{\mathbb{T}}J_{2,n}(u)du=1,\\ \frac{1}{\pi}\int_0^{\pi}uJ_{2,n}(u)du\le\frac{1}{2n},\end{array}\right\} \tag{4.2}$$

**Lemma 4.1.** *Let $W(x) = |x - x_0|^\gamma$, $\theta > 0$, $p(.) \in \mathrm{P}_0(\mathbb{T})$, and $f \in L^{p(.),\theta}_{0,W}$. If $f \in \mathcal{W}^1_{0,p(.),W,\theta}$, then*

$$E_n(f)_{p(.),W,\theta} \leq \|f - D_n f\|_{p(.),W,\theta} \leq \frac{c}{n}\|f'\|_{p(.),W,\theta} \tag{4.3}$$

*holds for $n \in \mathbb{N}$.*

***Proof of Lemma 4.1.*** From (4.1), Theorem 2.14, and (4.2), we have

$$\begin{aligned}
\|f - D_n f\|_{p(.),W,\theta} &= \left\| \frac{1}{\pi}\int_{\mathbb{T}} (f(x) - f(x-t))(1/t) t J_{2,\lfloor \frac{n}{2}\rfloor+1}(t) dt \right\|_{p(.),W,\theta} \\
&= \left\| \frac{1}{\pi}\int_{\mathbb{T}} t J_{2,\lfloor \frac{n}{2}\rfloor+1}(t) \frac{1}{t}\int_{x-t}^{x} f'(\tau) d\tau dt \right\|_{p(.),W,\theta} \\
&\leq \frac{1}{\pi}\int_{\mathbb{T}} t J_{2,\lfloor \frac{n}{2}\rfloor+1}(t) \left\| \frac{1}{t}\int_{x-t}^{x} f'(\tau) d\tau \right\|_{p(.),W,\theta} dt \\
&\leq \|Mf'\|_{p(.),W,\theta} \frac{1}{\pi}\int_0^{\pi} t J_{2,\lfloor \frac{n}{2}\rfloor+1}(t) dt \\
&\leq \frac{C}{2\left(\lfloor \frac{n}{2}\rfloor + 1\right)} \|f'\|_{p(.),W,\theta} \leq \frac{c}{n}\|f'\|_{p(.),W,\theta}.
\end{aligned}$$

Hence (4.3) holds. □

***Proof of Theorem 3.1.*** Let $f \in \mathcal{W}^1_{0,p(.),W,\theta}$, $n \in \mathbb{N}$, $\Theta_n \in \mathcal{T}_n$, $E_n(f')_{p(.),W,\theta} = \|f' - \Theta_n\|_{p(.),W,\theta}$ and $\beta/2$ be the constant term of $\Theta_n$, namely,

$$\beta = \frac{1}{\pi}\int_{\mathbb{T}} \Theta_n(t)\, dt = \frac{1}{\pi}\int_{\mathbb{T}} \left(\Theta_n(t) - f'(t)\right) dt.$$

Then

$$\begin{aligned}
|\beta/2| &\leq \frac{1}{2\pi}\|f' - \Theta_n\|_{L_1} \\
&\leq \frac{c}{2\pi}\|f' - \Theta_n\|_{p(.),W,\theta} = \frac{c}{2\pi} E_n(f')_{p(.),W,\theta}.
\end{aligned}$$

On the other hand

$$\begin{aligned}
\|f' - (\Theta_n - \beta/2)\|_{p(.),W,\theta} &\leq E_n(f')_{p(.),W,\theta} + \|\beta/2\|_{p(.),W,\theta} \\
&\leq E_n(f')_{p(.),W,\theta} + \frac{c}{2\pi}\|W\|_{L_1} E_n(f')_{p(.),W,\theta} \\
&= \left(1 + \frac{c}{2\pi}\|W\|_{L_1}\right) E_n(f')_{p(.),W,\theta}.
\end{aligned}$$

Set $u_n \in \mathcal{T}_n$ so that $u_n' = \Theta_n - \beta/2$. Then

$$\begin{aligned}
E_n(f)_{p(.),W,\theta} &= E_n(f - u_n)_{p(.),W,\theta} \\
&\leq \frac{c}{n}\|f' - (\Theta_n - \beta/2)\|_{p(.),W,\theta} \\
&\leq \left(c + \frac{C}{2\pi}\|W\|_{L_1}\right)\frac{1}{n} E_n(f')_{p(.),W,\theta}
\end{aligned}$$

for all $f \in \mathcal{W}^1_{0,p(.),W,\theta}$. If $f \in \mathcal{W}^r_{0,p(.),W,\theta}$ for some $r$, the last inequality gives

$$\begin{aligned}
E_n(f)_{p(.),W,\theta} &\leq C\left(1 + \frac{c}{2\pi}\|W\|_{L_1}\right)^r \frac{1}{n^r} E_n\left(f^{(r)}\right)_{p(.),W,\theta} \\
&= \frac{c}{n^r} E_n\left(f^{(r)}\right)_{p(.),W,\theta}.
\end{aligned}$$

□

***Proof of Theorem 3.3.*** Let $f \in L_{0,W}^{p(.),\theta}$. Using Theorem 3.1 and Corollary 3.2 we have

$$\begin{aligned} E_n\left(f\right)_{p(.),W,\theta} &\leq E_n\left(f-g\right)_{p(.),W,\theta}+E_n\left(g\right)_{p(.),W,\theta} \\ &\leq c\left\{\left\|f-g\right\|_{p(.),W,\theta}+\delta^r\left\|g^{(r)}\right\|_{p(.),W,\theta}\right\}. \end{aligned}$$

for $g \in \mathcal{W}_{0,p(.),W,\theta}^r$ and $\delta = \frac{1}{n}$. Using Theorem 3.7 and taking infimum on $g \in \mathcal{W}_{0,p(.),W,\theta}^r$, we obtain

$$E_n\left(f\right)_{p(.),W,\theta} \leq c\Omega_r\left(f,\frac{1}{n}\right)_{p(.),W,\theta}, \; n \in \mathbb{N}.$$

□

***Proof of Theorem 3.4.*** Let $T_n$ be a best approximation trigonmetric polynomial for $f \in L_{0,W}^{p(.),\theta}$. For any $n \in \mathbb{N}$ we choose $n \in \mathbb{N}$ such that $2^m \leq n < 2^{m+1}$. If we use the subadditivity property of $\Omega_r\left(f,\delta\right)_{p(.),W,\theta}$, then we have

$$\Omega_r\left(f,\delta\right)_{p(.),W,\theta} \leq \Omega_r\left(f-T_{2^{m+1}},\delta\right)_{p(.),W,\theta}+\Omega_r\left(T_{2^{m+1}},\delta\right)_{p(.),W,\theta}. \tag{4.4}$$

On the other hand, it is well-known that

$$2^{(i+1)r}E_{2^i}(f)_{p(.),W,\theta} \leq 2^{2r}\sum_{j=2^{i-1}+1}^{2^i} j^{r-1}E_j(f)_{p(.),W,\theta} \tag{4.5}$$

by Theorem 3.1 in [26]. If we take $\delta = \frac{1}{n}$, then we get

$$\begin{aligned} \Omega_r\left(f-T_{2^{m+1}},\delta\right)_{p(.),W,\theta} &\leq c\left\|f-T_{2^{m+1}}\right\|_{p(.),W,\theta} \\ &= cE_{2^{m+1}}(f)_{p(.),W,\theta} \\ &\leq \frac{c}{n^r}2^{2(m+1)r}E_{2^m}(f)_{p(.),W,\theta} \\ &\leq c\delta^r 2^{2r}\sum_{k=2^{m-1}+1}^{2^m} k^{r-1}E_k(f)_{p(.),W,\theta}. \end{aligned} \tag{4.6}$$

Using Lemma 3.5, Lemma 3.8 and (4.5) one can find that

$$\begin{aligned} &\Omega_r\left(T_{2^{m+1}},\delta\right)_{p(.),W,\theta} \\ \leq\;& c\delta^r\left\|T_{2^{m+1}}^{(r)}\right\|_{p(.),W,\theta} \\ \leq\;& c\delta^r\left\{\left\|T_1^{(r)}+\sum_{i=0}^{m}\left(T_{2^{i+1}}^{(r)}-T_{2^i}^{(r)}\right)\right\|_{p(.),W,\theta}\right\} \\ \leq\;& c\delta^r\left\{\left\|T_1\right\|_{p(.),W,\theta}+\sum_{i=0}^{m}2^{(i+1)r}\left\|T_{2^{i+1}}^{(r)}-T_{2^i}^{(r)}\right\|_{p(.),W,\theta}\right\} \\ \leq\;& c\delta^r\left\{E_0(f)_{p(.),W,\theta}+\sum_{i=0}^{m}2^{(i+1)r}E_{2^i}(f)_{p(.),W,\theta}\right\} \\ =\;& c\delta^r\left\{E_0(f)_{p(.),W,\theta}+2^rE_1(f)_{p(.),W,\theta}+2^{2r}\sum_{i=1}^{m}\sum_{k=2^{i-1}+1}^{2^i}k^{r-1}E_k(f)_{p(.),W,\theta}\right\} \\ \leq\;& c\delta^r\left\{E_0(f)_{p(.),W,\theta}+\sum_{k=1}^{2^m}k^{r-1}E_k(f)_{p(.),W,\theta}\right\}. \end{aligned} \tag{4.7}$$

If we combine (4.4), (4.6) and (4.7), then we find

$$\Omega_r\left(f,\frac{1}{n}\right)_{p(.),W,\theta}\leq\frac{c}{n^r}\sum_{k=0}^{n}(k+1)^{r-1}E_k(f)_{p(.),W,\theta}\,,\ n\in\mathbb{N}.$$

□

The notation $\mathcal{O}$ indicates that $A=\mathcal{O}(B)$ if and only if there exists a positive constant $c$, independent of essential parameters, such that $A\leq cB$.

**Corollary 4.2.** *If $E_n(f)_{p(.),W,\theta}=\mathcal{O}(n^{-\alpha})$, $\alpha>0$, then under the conditions of Theorem 3.4 we have*

$$\Omega_r(f,\delta)_{p(.),W,\theta}=\begin{cases}\mathcal{O}(\delta^\alpha) & ,\ r>\alpha,\\ \mathcal{O}\left(\delta^\alpha\log\left(\frac{1}{\delta}\right)\right) & ,\ r=\alpha,\\ \mathcal{O}(\delta^r) & ,\ r<\alpha.\end{cases}$$

**Definition 4.3.** Let $W(x)=|x-x_0|^\gamma$, $\theta>0$, $p(.)\in P_0(\mathbb{T})$, $f\in L_{0,W}^{p(.),\theta}$, $\alpha>0$ and $r:=[\alpha]+1$ ( $[\alpha]$ is the integer part of $\alpha$ ). We define the generalized Lipschitz class as

$$Lip_{p(.),W,\theta}^{\alpha,r}=\left\{f\in L_W^{p(.),\theta}:\Omega_r(f,\delta)_{p(.),W,\theta}=\mathcal{O}(\delta^\alpha)\right\}.$$

**Corollary 4.4.** *Let $W(x)=|x-x_0|^\gamma$, $\theta>0$, $p(.)\in \mathrm{P}_0(\mathbb{T})$, $f\in L_{0,W}^{p(.),\theta}$ and $\alpha>0$. Then the following statements are equivalent:*

(i) $f\in Lip_{p(.),W,\theta}^{\alpha,r}$

(ii) $E_n(f)_{p(.),W,\theta}=O(n^{-\alpha})$, $n\in\mathbb{N}$.

**Theorem 4.5.** *Let $W(x)=|x-x_0|^\gamma$, $\theta>0$, $p(.)\in \mathrm{P}_0(\mathbb{T})$, $f\in L_{0,W}^{p(.),\theta}$ and $r\in\mathbb{N}$. If*

$$\sum_{k=1}^{\infty}k^{r-1}E_k(f)_{p(.),W,\theta}<\infty,$$

*then, $f\in\mathcal{W}_{p(.),0,W,\theta}^r$ and*

$$E_n\left(f^{(r)}\right)_{p(.),W,\theta}\leq c\left(n^rE_n(f)_{p(.),W,\theta}+\sum_{k=n+1}^{\infty}k^{r-1}E_k(f)_{p(.),W,\theta}\right)$$

*with a positive constant $c$ independent of $f$ and $n$.*

***Proof of Theorem 4.5.*** For the polynomial $T_n$ of the best approximation to $f$ we have by Lemma 3.8 that

$$\begin{aligned}\left\|T_{2^{i+1}}^{(r)}-T_{2^i}^{(r)}\right\|_{p(.),W,\theta} &\leq C(r)\,2^{(i+1)r}\left\|T_{2^{i+1}}-T_{2^i}\right\|_{p(.),W,\theta}\\ &\leq 2C(r)\,2^{(i+1)r}E_{2^i}(f)_{p(.),W,\theta}\,.\end{aligned}$$

Hence

$$\begin{aligned}\sum_{i=1}^{\infty}\left\|T_{2^{i+1}}-T_{2^i}\right\|_{p(.),W,\theta}^r &= \sum_{i=1}^{\infty}\left\|T_{2^{i+1}}^{(r)}-T_{2^i}^{(r)}\right\|_{p(.),W,\theta}+\sum_{i=1}^{\infty}\left\|T_{2^{i+1}}-T_{2^i}\right\|_{p(.),W,\theta}\\ &\leq c\sum_{m=2}^{\infty}m^{r-1}E_m(f)_{p(.),W,\theta}<\infty.\end{aligned}$$

Therefore

$$\left\|T_{2^{i+1}}-T_{2^i}\right\|_{p(.),W,\theta}^r\to 0 \text{ as } i\to\infty.$$

This means that $\{T_{2^i}\}$ is a Cauchy sequence in $L_W^{p(.),\theta}$. Since $T_{2^i}\to f$ in $L_W^{p(.),\theta}$ and $\mathcal{W}_{p(.),W,\theta}^r$ is a Banach space we obtain $f\in\mathcal{W}_{p(.),W,\theta}^r$.

On the other hand since

$$\left\|f^{(r)}-T_n^{(r)}\right\|_{p(.),W,\theta}\leq\left\|T_{2^{m+2}}^{(r)}-T_n^{(r)}\right\|_{p(.),W,\theta}+\sum_{k=m+2}^{\infty}\left\|T_{2^{k+1}}^{(r)}-T_{2^k}^{(r)}\right\|_{p(.),W,\theta}$$

for $2^m\leq n<2^{m+1}$, we have

$$\left\|T_{2^{m+2}}^{(r)}-T_n^{(r)}\right\|_{p(.),W,\theta}\leq c2^{(m+2)r}E_n\left(f\right)_{p(.),W,\theta}\leq c\left(n+1\right)^rE_n\left(f\right)_{p(.),W,\theta}.$$

Also we find

$$\begin{aligned}\sum_{k=m+2}^{\infty}\left\|T_{2^{k+1}}^{(r)}-T_{2^k}^{(r)}\right\|_{p(.),W,\theta} &\leq c\sum_{k=m+2}^{\infty}2^{(k+1)r}E_{2^k}\left(f\right)_{p(.),W,\theta}\\ &\leq c\sum_{k=m+2}^{\infty}\sum_{\mu=2^{k-1}+1}^{2^k}\mu^{r-1}E_\mu\left(f\right)_{p(.),W,\theta}\\ &= c\sum_{\nu=2^{m+1}+1}^{\infty}\nu^{r-1}E_\nu\left(f\right)_{p(.),W,\theta}\\ &\leq c\sum_{\nu=n+1}^{\infty}\nu^{r-1}E_\nu\left(f\right)_{p(.),W,\theta}.\end{aligned}$$

This completes the proof. □

A polynomial $T\in\Pi_n$ is said to be a *near best approximant* of $f\in L_{0,W}^{p(.),\theta}$ for $W(x)=|x-x_0|^\gamma$, $\theta>0$, $p(.)\in P_0\left(\mathbb{T}\right)$, if

$$\|f-T\|_{p(.),W,\theta}\leq cE_n\left(f\right)_{p(.),W,\theta},\quad n=1,2,\ldots.$$

**Theorem 4.6.** *Let $W(x)=|x-x_0|^\gamma$, $\theta>0$, $p(.)\in \mathrm{P}_0\left(\mathbb{T}\right)$, $r,n\in\mathbb{N}$. If $T_n\in\Pi_n$ is a near best approximant of $f\in\mathcal{W}_{p(.),W,\theta}^r$, then there exists a constant $c>0$ dependent only on $r,W$ and $p(.)$, such that*

$$\left\|f^{(r)}-T_n^{(r)}\right\|_{p(.),W,\theta}\leq cE_n\left(f^{(r)}\right)_{p(.),W,\theta}.$$

**Corollary 4.7.** *Suppose that $W(x)=|x-x_0|^\gamma$, $\theta>0$, $p(.)\in\mathrm{P}_0\left(\mathbb{T}\right)$, $r,n\in\mathbb{N}$, $f\in\mathcal{W}_{p(.),W,\theta}^{\alpha}$, and*

$$\sum_{\nu=1}^{\infty}\nu^{\alpha-1}E_\nu\left(f\right)_{p(.),W,\theta}<\infty$$

*for some $\alpha>0$. Hence there exists a constant $c>0$ dependent only on $\alpha$, $r$, $W$ and $p(.)$ such that*

$$\Omega_r(f^{(\alpha)},\frac{\pi}{n})_{p(.),W,\theta}\leq c\left\{\frac{1}{n^r}\sum_{\nu=0}^{n}(\nu+1)^{\alpha+r-1}E_\nu(f)_{p(.),W,\theta}+\sum_{\nu=n+1}^{\infty}\nu^{\alpha-1}E_\nu)f)_{p(.),W,\theta}\right\}.$$

***Proof of Theorem 4.6.*** We set $W_n(f):=W_n(x,f):=\frac{1}{n+1}\sum\limits_{\nu=n}^{2n}S_\nu(x,f),\quad n=0,1,2,\ldots.$ Since

$$W_n(.,f^{(\alpha)})=W_n^{(\alpha)}(.,f),$$

then we have

$$\begin{aligned}&\left\|f^{(\alpha)}(.)-T_n^{(\alpha)}(.,f)\right\|_{p(.),W,\theta}\leq\left\|f^{(\alpha)}(.)-W_n(.,f^{(\alpha)})\right\|_{p(.),W,\theta}\\ &+\left\|T_n^{(\alpha)}(.,W_n(f))-T_n^{(\alpha)}(.,f)\right\|_{p(.),W,\theta}+\left\|W_n^{(\alpha)}(.,f)-T_n^{(\alpha)}(.,W_n(f))\right\|_{p(.),W,\theta}\\ =\ &I_1+I_2+I_3.\end{aligned}$$

We denote by $T_n^*(x,f)$ the best approximating polynomial of degree at most $n$ to $f$ in $L_W^{p(.),\theta}$. In this case, from the boundedness of $W_n$ in $L_W^{p(.),\theta}$, we have

$$\begin{aligned} I_1 &\leq \left\|f^{(\alpha)}(.) - T_n^*(.,f^{(\alpha)})\right\|_{p(.),W,\theta} + \left\|T_n^*(.,f^{(\alpha)}) - W_n(.,f^{(\alpha)})\right\|_{p(.),W,\theta} \\ &\leq c\,(p,W,\theta)\,E_n\left(f^{(\alpha)}\right)_{p(.),W,\theta} + \left\|W_n(.,T_n^*(f^{(\alpha)}) - f^{(\alpha)})\right\|_{p(.),W,\theta} \\ &\leq c\,(p,W,\theta)\,E_n\left(f^{(\alpha)}\right)_{p(.),W,\theta}. \end{aligned}$$

From Lemma 3.8 we get

$$I_2 \leq c\,(p,W,\theta)\,n^{\alpha}\,\|T_n(.,W_n(f)) - T_n(.,f)\|_{p(.),W,\theta}$$

and

$$\begin{aligned} I_3 &\leq c\,(p,W,\theta)\,(2n)^{\alpha}\,\|W_n(.,f) - T_n(.,W_n(f))\|_{p(.),W,\theta} \\ &\leq c\,(p,W,\theta)\,(2n)^{\alpha}\,E_n\,(W_n(f))_{p(.),W,\theta}. \end{aligned}$$

Now we have

$$\|T_n(.,W_n(f)) - T_n(.,f)\|_{p(.),W,\theta} \leq \|T_n(.,W_n(f)) - W_n(.,f)\|_{p(.),W,\theta}$$

$$\begin{aligned} &+\|W_n(.,f) - f(.)\|_{p(.),W,\theta} + \|f(.) - T_n(.,f)\|_{p(.),W,\theta} \\ \leq\ & c\,(p,W,\theta)\,E_n\,(W_n(f))_{p(.),W,\theta} + c\,(p,W,\theta)\,E_n\,(f)_{p(.),W,\theta} \\ &+c\,(p,W,\theta)\,E_n\,(f)_{p(.),W,\theta}. \end{aligned}$$

Since

$$E_n\,(W_n(f))_{p(.),W,\theta} \leq c\,(p,W,\theta)\,E_n\,(f)_{p(.),W,\theta},$$

then we get

$$\begin{aligned} \left\|f^{(\alpha)}(.) - T_n^{(\alpha)}(.,f)\right\|_{p(.),W,\theta} &\leq c\,(p,W,\theta)\,E_n(f^{(\alpha)})_{p(.),W,\theta} \\ &\quad +c\,(p,W,\theta)\,n^{\alpha}E_n\,(W_n(f))_{p(.),W,\theta} \end{aligned}$$

$$\begin{aligned} &+c\,(p,W,\theta)\,n^{\alpha}E_n\,(f)_{p(.),W,\theta} + c\,(p,W,\theta)\,(2n)^{\alpha}\,E_n\,(W_n(f))_{p(.),W,\theta} \\ \leq\ & c\,(p,W,\theta)\,E_n\left(f^{(\alpha)}\right)_{p(.),W,\theta} + c\,(p,W,\theta)\,n^{\alpha}E_n\,(f)_{p(.),W,\theta}. \end{aligned}$$

Since, according to Theorem 3.1,

$$E_n\,(f)_{p(.),W,\theta} \leq \frac{c\,(p,W,\theta)}{(n+1)^{\alpha}}E_n\left(f^{(\alpha)}\right)_{p(.),W,\theta}, \tag{4.8}$$

we obtain

$$\left\|f^{(\alpha)}(.) - T_n^{(\alpha)}(.,f)\right\|_{p(.),W,\theta} \leq c\,(p,W,\theta)\,E_n\left(f^{(\alpha)}\right)_{p(.),W,\theta}$$

and the proof is completed. □

**Acknowledgment.** The authors would like to thank the referee for the helpful comments and suggestions.

## References

[1] G. Anatriello, *Iterated grand and small Lebesgue spaces*, Collect. Math. **65**, 273–284, 2014.

[2] R. Akgün, *Trigonometric approximation of functions in generalized Lebesgue spaces with variable exponent*, Ukrainian Math. J. **63** (1), 1–26, 2011.

[3] R. Akgün, *Approximating polynomials for functions of weighted Smirnov-Orlicz spaces*, J. Funct. Spaces Appl. **2012**, Art. ID 982360, 2012.

[4] R. Akgün and D.M. Israfilov, *Approximation and moduli of smoothness of fractional order in Smirnov-Orlicz spaces*, Glas. Mat. Ser. III, **42** (2), 121–136, 2008.

[5] R. Akgün and D.M. Israfilov, *Polynomial approximation in weighted Smirnov Orlicz space*, Proc. A. Razmadze Math. Inst. **139**, 89–92, 2005.

[6] R. Akgün and D.M. Israfilov, *Approximation by interpolating polynomials in Smirnov-Orlicz class*, J. Korean Math. Soc. **43** (2), 413–424, 2006.

[7] R. Akgün and D.M. Israfilov, *Simultaneous and converse approximation theorems in weighted Orlicz spaces*, Bull. Belg. Math. Soc. **17** (2), 13–28, 2010.

[8] R. Akgün and V. Kokilashvili, *On converse theorems of trigonometric approximation in weighted variable exponent Lebesgue spaces*, Banach J. Math. Anal. **5** (1), 70-82, 2011.

[9] I. Aydin, *Weighted variable Sobolev spaces and capacity*, J. Funct. Spaces Appl. **2012**, Art. ID 132690, 2012.

[10] D. Cruz-Uribe and A. Fiorenza, *Variable Lebesgue Spaces: Foundations and Harmonic Analysis (Applied and Numerical Harmonic Analysis)*, Birkhäuser/Springer, Heidelberg, 2013.

[11] D. Cruz-Uribe, L. Diening and P. Hästö, *The maximal operator on weighted variable Lebesgue spaces*, Fract. Calc. Appl. Anal. **14** (3), 361–374, 2011.

[12] N. Danelia and V. Kokilashvili, *Approximation by trigonometric polynomials in the framework of variable exponent Lebesgue spaces*, Georgian Math. J. **23** (1), 43–53, 2016.

[13] L. Diening, *Maximal function on generalized Lebesgue spaces $L^{p(\cdot)}$*, Math. Inequal. Appl. **7**, 245–253, 2004.

[14] L. Diening, P. Harjulehto, P. Hästö and M. Ruzicka, *Lebesgue and Sobolev Spaces with Variable Exponents*, Lecture Notes in Mathematics, **2017**, Springer, Heidelberg, 2011.

[15] V.K. Dzyadyk and I.A. Shevchuk, *Theory of Uniform Approximation of Functions by Polynomials*, Walter de Gruyter GmbH & Co. KG, 10785 Berlin, Germany, 2008.

[16] A. Fiorenza, B. Gupta and P. Jain, *The maximal theorem in weighted grand Lebesgue spaces*, Studia Math. **188** (2), 123–133, 2008.

[17] A. Fiorenza, V. Kokilashvili and A. Meskhi, *Hardy-Littlewood maximal operator in weighted grand variable exponent Lebesgue space*, Mediterr. J. Math. **14** (118), 2017.

[18] L. Greco, T. Iwaniec and C. Sbordone, *Inverting the p-harmonic operator*, Manuscripta Math. **92**, 249–258, 1997.

[19] D.M. Israfilov and R. Akgün, *Approximation in weighted Smirnov-Orlicz classes*, J. Math. Kyoto Univ. **46** (4), 755–770, 2006.

[20] D.M. Israfilov and A. Testici, *Approximation in weighted generalized grand Lebesgue spaces*, Colloq. Math. **143**, 113–126, 2016.

[21] D.M. Israfilov and A. Testici, *Approximation in weighted generalized grand Smirnov spaces*, Studia Sci. Math. Hungar. **54** (4), 471–488, 2017.

[22] T. Iwaniec and C. Sbordone, *On integrability of the Jacobien under minimal hypotheses*, Arch. Ration. Mech. Anal. **119**, 129–143, 1992.

[23] V. Kokilashvili and A. Meskhi, *Maximal and Calderon-Zygmund operators in grand variable exponent Lebesgue spaces*, Georgian Math. J. **21**, 447–461, 2014.

[24] O. Kováčik and J. Rákosník, *On spaces $L^{p(x)}$ and $W^{k,p(x)}$*, Czechoslovak Math. J. **41** (116), 592–618, 1991.
[25] W. Orlicz. *Über konjugierte exponentenfolgen*, Stud. Math. **3**, 200–211, 1931.
[26] R.A. De Vore and G.G. Lorentz, *Constructive Approximation*, Springer, 1993.
[27] A. Zygmund, *Trigonometric Series, Volume I-II*, Cambridge University Press, Cambridge, 1968.

Hacettepe Journal of Mathematics & Statistics

Hacet. J. Math. Stat.
Volume 50 (1) (2021), 216 – 223
DOI : 10.15672/hujms.685742

Research Article

# Rota-Baxter bialgebra structures arising from (co-)quasi-idempotent elements

Tianshui Ma*, Jie Li, Haiyan Yang

*School of Mathematics and Information Science, Henan Normal University, Xinxiang 453007, China*

## Abstract

In this note, we construct Rota-Baxter (coalgebras) bialgebras by (co-)quasi-idempotent elements and prove that every finite dimensional Hopf algebra admits nontrivial Rota-Baxter bialgebra structures and tridendriform bialgebra structures. We give all the forms of (co)-quasi-idempotent elements and related structures of tridendriform (co, bi)algebras and Rota-Baxter (co, bi)algebras on the well-known Sweedler's four-dimensional Hopf algebra.

**Mathematics Subject Classification (2020).** 16W99, 16T05

**Keywords.** Rota-Baxter bialgebras, (co-)quasi-idempotent element, tridendriform bialgebra

## 1. Introduction

Rota-Baxter algebras were introduced in [11] in the context of differential operators on commutative Banach algebras and since [1], intensively studied in probability and combinatorics, and more recently in mathematical physics, such as free Rota-Baxter algebras, Lie algebras, multiple zeta values, differential algebras and Connes-Kreimer renormalization theory in quantum field theory, see ([2–7], etc.). One can refer to the book [2] for the detailed theory of Rota-Baxter algebras.

In 2014, based on the dual method in the Hopf algebra theory, Jian and Zhang in [8] defined the notion of Rota-Baxter coalgebras and also provided various examples of the new object. Then Rota-Baxter bialgebras were presented in [9] whose examples can be constructed from the well-known Radford biproduct. In 2017, Jian construct quasi-idempotent Rota-Baxter operators by quasi-idempotent elements and show that every finite dimensional Hopf algebra admits nontrivial Rota-Baxter algebra structures and tridendriform algebra structures (see [7]).

So it is natural to consider if every finite dimensional Hopf algebra admits nontrivial Rota-Baxter bialgebra structure and tridendriform bialgebra structure. In this paper, we give a positive answer to this question. This is the motivation to write this paper.

This paper is organized as follows. In Section 2, we list some definitions that will be used later. In Section 3, we present the notions of tridendriform coalgebras, tridendriform

*Corresponding Author.
Email addresses: matianshui@yahoo.com (T. Ma), lijie_0224@163.com (J. Li), yhy3023551288@163.com (H. Yang)
Received: 06.02.2020; Accepted: 29.05.2020

bialgebras, and co-quasi-idempotent element in a coalgebra. We use (co-)quasi-idempotent element to construct Rota-Baxter coalgebras and bialgebras. And then we prove that every finite dimensional Hopf algebra admits nontrivial Rota-Baxter bialgebra structures and tridendriform bialgebra structures. All the forms of (co)-quasi-idempotent elements and related structures of tridendriform (co, bi)algebras and Rota-Baxter (co, bi)algebras on the well-known Sweedler's four-dimensional Hopf algebra are provided in Section 4.

## 2. Preliminaries

For simplicity, we fix our ground field to be the complex number field $\mathbb{C}$ throughout this paper. All the objects we discuss are defined over $\mathbb{C}$ unless otherwise specified. For an algebra $A$, we denote its multiplication $\mu_A$ (or simply $\mu$) by $\mu_A(a \otimes b) = ab$.

In what follows, we recall some useful definitions which will be used later (see [2,7,9]).

**Definition 2.1.** For $\lambda \in \mathbb{C}$, a **Rota-Baxter algebra of weight** $\lambda$ is an associative algebra $A$ together with a linear map $R : A \longrightarrow A$ such that

$$R(a)R(b) = R(aR(b)) + R(R(a)b) + \lambda R(ab) \tag{2. 1}$$

for all $a, b \in A$. Such a linear operator is called a **Rota-Baxter operator of weight** $\lambda$ on $A$.

**Remark 2.2.** If $R$ is a Rota-Baxter operator of weight 1, then $\lambda R$ is a Rota-Baxter operator of weight $\lambda$. Conversely, if $R$ is a Rota-Baxter operator of weight $\lambda$ and $\lambda$ is invertible, then $\lambda^{-1}R$ is a Rota-Baxter operator of weight 1.

**Definition 2.3.** Let $C$ be a vector space and $\Delta_C : C \longrightarrow C \otimes C$ (here we use Sweedler's notation and denote $\Delta_C(c)$ by $c_1 \otimes c_2$), $\varepsilon_C : C \longrightarrow \mathbb{C}$ two linear maps. Then $C$ is a coassociative coalgebra if

$$c_{11} \otimes c_{12} \otimes c_2 = c_1 \otimes c_{21} \otimes c_{22} \text{ and } \varepsilon_C(c_1)c_2 = c_1\varepsilon_C(c_2) = c$$

hold for all $c \in C$.

Let $\gamma$ be an element in $\mathbb{C}$. A pair $(C, Q)$ is called a **Rota-Baxter coalgebra of weight** $\gamma$ if $C$ is a coassociative coalgebra and $Q$ is a linear endomorphism of $C$ satisfying that for all $c \in C$,

$$Q(c_1) \otimes Q(c_2) = Q(c)_1 \otimes Q(Q(c)_2) + Q(Q(c)_1) \otimes Q(c)_2 + \gamma Q(c)_1 \otimes Q(c)_2. \tag{2. 2}$$

The map $Q$ is called a **Rota-Baxter operator weight** $\gamma$ on $C$.

**Remark 2.4.** If $Q$ is a Rota-Baxter operator of weight 1, then $\gamma Q$ is a Rota-Baxter operator of weight $\gamma$. Conversely, if $Q$ is a Rota-Baxter operator of weight $\gamma$ and $\gamma$ is invertible, then $\gamma^{-1}Q$ is a Rota-Baxter operator of weight 1.

**Definition 2.5.** Let $H$ be a vector space. $H$ is a bialgebra if $(H, \mu_H)$ is an associative algebra and $(H, \Delta_H)$ is a coassociative coalgebra such that $\Delta_H$ and $\varepsilon_H$ are algebra maps.

Let $\lambda$, $\gamma$ be elements in $\mathbb{C}$ and $H$ a bialgebra (maybe without unit and counit). A triple $(H, R, Q)$ is called a **Rota-Baxter bialgebra of weight** $(\lambda, \gamma)$ if $(H, R)$ is a Rota-Baxter algebra of weight $\lambda$ and $(H, Q)$ is a Rota-Baxter coalgebra of weight $\gamma$.

**Remark 2.6.** If $(H, R, Q)$ is a Rota-Baxter bialgebra of weight $(1, 1)$, then $(H, \lambda R, \gamma Q)$ is a Rota-Baxter bialgebra of weight $(\lambda, \gamma)$. Conversely, if $(H, R, Q)$ is a Rota-Baxter bialgebra of weight $(\lambda, \gamma)$ and $\lambda, \gamma$ are invertible, then $(H, \lambda^{-1}R, \gamma^{-1}Q)$ is a Rota-Baxter bialgebra of weight $(1, 1)$.

**Definition 2.7.** Let $A$ be an associative algebra and $\lambda \in \mathbb{C}$. A linear endomorphism $\phi$ of $A$ is called a **quasi-idempotent operator of weight** $\lambda$ **on** $A$ if $\phi^2 = -\lambda\phi$. A nonzero element $\xi \in A$ is called a **quasi-idempotent element of weight** $\lambda$ if $\xi^2 = -\lambda\xi$.

**Definition 2.8.** Let $V$ be a vector space, and $\prec, \succ, \cdot : V \otimes V \longrightarrow V$ be three linear maps. The quadruple $(V, \prec, \succ, \cdot)$ is called a **tridendriform algebra** if the following conditions are satisfied: for all $x, y, z \in V$,

$$\begin{aligned}
&(x \prec y) \prec z = x \prec (y * z), \quad (x \succ y) \prec z = x \succ (y \prec z),\\
&(x * y) \succ z = x \succ (y \succ z), \quad (x \succ y) \cdot z = x \succ (y \cdot z),\\
&(x \prec y) \cdot z = x \cdot (y \succ z), \quad (x \cdot y) \prec z = x \cdot (y \prec z), \quad (x \cdot y) \cdot z = x \cdot (y \cdot z),
\end{aligned}$$

where $x * y = x \prec y + x \succ y + x \cdot y$.

**Remark 2.9.** Given a Rota-Baxter algebra $(A, R)$ of weight 1, we define

$$a \prec b = a \cdot R(b), \qquad a \succ b = R(a) \cdot b,$$

for all $a, b \in A$. Then $(V, \prec, \succ, \mu_A)$ is a tridendriform algebra.

## 3. Construction of tridendriform co(bi)algebra and Rota-Baxter bialgebras

In this section, based on the dual method in Hopf algebra theory, we define tridendriform co(bi)algebras, co-quasi-idempotent elements, then construct tridendriform co(bi)algebras and Rota-Baxter co(bi)algebras through (co-)quasi-idempotent elements.

**Definition 3.1.** Let $V$ be a vector space, and $\Delta_\prec, \Delta_\succ, \Delta_\cdot : V \longrightarrow V \otimes V$ be three linear maps (write $\Delta_\prec(x) = x^1 \otimes x^2, \Delta_\succ(x) = x^{(1)} \otimes x^{(2)}, \Delta_\cdot(x) = x^{[1]} \otimes x^{[2]}$). The quadruple $(V, \Delta_\prec, \Delta_\succ, \Delta_\cdot)$ is called a **tridendriform coalgebra** if the following conditions are satisfied: for all $x \in V$,

$$\begin{gathered}
x^{11} \otimes x^{12} \otimes x^2 = x^1 \otimes (x^{21} \otimes x^{22} + x^{2(1)} \otimes x^{2(2)} + x^{2[1]} \otimes x^{2[2]}),\\
x^{1(1)} \otimes x^{1(2)} \otimes x^2 = x^{(1)} \otimes x^{(2)1} \otimes x^{(2)2},\\
(x^{(1)1} \otimes x^{(1)2} + x^{(1)(1)} \otimes x^{(1)(2)} + x^{(1)[1]} \otimes x^{(1)[2]}) \otimes x^{(2)} = x^{(1)} \otimes x^{(2)(1)} \otimes x^{(2)(2)},\\
x^{[1](1)} \otimes x^{[1](2)} \otimes x^{[2]} = x^{(1)} \otimes x^{(2)[1]} \otimes x^{(2)[2]},\\
x^{[1]1} \otimes x^{[1]2} \otimes x^{[2]} = x^{[1]} \otimes x^{[2](1)} \otimes x^{[2](2)},\\
x^{1[1]} \otimes x^{1[2]} \otimes x^2 = x^{[1]} \otimes x^{[2]1} \otimes x^{[2]2},\\
x^{[1][1]} \otimes x^{[1][2]} \otimes x^{[2]} = x^{[1]} \otimes x^{[2][1]} \otimes x^{[2][2]}.
\end{gathered}$$

Rota-Baxter coalgebras are closely related to tridendriform coalgebras.

**Lemma 3.2.** Given a Rota-Baxter coalgebra $(C, Q)$ of weight 1, we define

$$\Delta_\prec(c) = c_1 \otimes Q(c_2), \qquad \Delta_\succ(c) = Q(c_1) \otimes c_2.$$

Then $(C, \Delta_\prec, \Delta_\succ, \Delta_C)$ is a tridendriform coalgebra.

***Proof.*** It can be proved by direct computation. □

**Definition 3.3.** Let $V$ be a vector space. A seven-tuple $(V, \prec, \succ, \cdot, \Delta_\prec, \Delta_\succ, \Delta_\cdot)$ is called a **tridendriform bialgebra** if $(V, \prec, \succ, \cdot)$ is a tridendriform algebra and at the same time $(V, \Delta_\prec, \Delta_\succ, \Delta_\cdot)$ is a tridendriform coalgebra.

**Proposition 3.4.** Let $H$ be a bialgebra and $(H, R, Q)$ a Rota-Baxter bialgebra of weight $(1, 1)$. Define

$$\begin{gathered}
x \prec y = xR(y), \qquad x \succ y = R(x)y,\\
\Delta_\prec(x) = x_1 \otimes Q(x_2), \qquad \Delta_\succ(x) = Q(x_1) \otimes x_2,
\end{gathered}$$

for all $x, y \in H$. Then $(V, \prec, \succ, \mu_H, \Delta_\prec, \Delta_\succ, \Delta_H)$ is a tridendriform bialgebra.

***Proof.*** It is a consequence of Lemma 3.2 and the Remark 2.9. □

**Definition 3.5.** Let $C$ be a coassociative coalgebra and $\gamma \in \mathbb{C}$. A linear endomorphism $\vartheta$ of $C$ is called a **quasi-idempotent operator of weight** $\gamma$ **on** $C$ if $\vartheta^2 = -\gamma\vartheta$. A nonzero element $\tau \in C^*$ is called a *co-quasi-idempotent element of weight* $\gamma$ if $\tau(c_1)\tau(c_2) = -\gamma\tau(c)$ for all $c \in C$.

**Proposition 3.6.** Let $C$ be a coalgebra. Given a co-quasi-idempotent element $\tau \in C^*$ of weight $\gamma \neq 0$. Three linear maps $\Delta_\prec, \Delta_\succ, \Delta_\cdot : C \longrightarrow C \otimes C$ defined below endow a tridendriform coalgebra structure on $C$: for all $c \in C$,

$$\Delta_\prec(c) = \gamma^{-1} c_1 \otimes \tau(c_2)c_3,\ \Delta_\succ(c) = \gamma^{-1}\tau(c_1)c_2 \otimes c_3,\ \Delta_\cdot(c) = c_1 \otimes c_2.$$

***Proof.*** We only check the first equality in the definition of tridendreform coalgebra as follows. For all $c \in C$, we can get

$$\begin{aligned}
&c^1 \otimes (c^{21} \otimes c^{22} + c^{2(1)} \otimes c^{2(2)} + c^{2[1]} \otimes c^{2[2]})\\
&\quad = \gamma^{-2} c_1\tau(c_2)\tau(c_{32}) \otimes c_{31} \otimes c_{33} + \gamma^{-2} c_1\tau(c_2)\tau(c_{31}) \otimes c_{32} \otimes c_{33}\\
&\qquad + \gamma^{-1} c_1\tau(c_2) \otimes c_{31} \otimes c_{32}\\
&\quad = \gamma^{-2} c_1\tau(c_2)\tau(c_{32}) \otimes c_{31} \otimes c_{33} - \gamma^{-1} c_1\tau(c_2) \otimes c_{31} \otimes c_{32}\\
&\qquad + \gamma^{-1} c_1\tau(c_2) \otimes c_{31} \otimes c_{32}\\
&\quad = \gamma^{-2} c_1\tau(c_2)\tau(c_{32}) \otimes c_{31} \otimes c_{33}\\
&\quad = c^{11} \otimes c^{12} \otimes c^2,
\end{aligned}$$

finishing the proof. □

**Theorem 3.7.** Let $H$ be a bialgebra. Given a quasi-idempotent element $\xi \in H$ of weight $\lambda \neq 0$ and a co-quasi-idempotent element $\tau \in H^*$ of weight $\gamma \neq 0$. Six linear maps $\prec, \succ, \cdot : H \otimes H \longrightarrow H$ and $\Delta_\prec, \Delta_\succ, \Delta_\cdot : H \longrightarrow H \otimes H$ defined below endow a tridendriform bialgebra structure on $H$: for all $x, y \in H$,

$$x \prec y = \lambda^{-1} x\xi y,\ x \succ y = \lambda^{-1}\xi xy,\ x \cdot y = xy,$$

and

$$\Delta_\prec(x) = \gamma^{-1} x_1 \otimes \tau(x_2)x_3,\ \Delta_\succ(x) = \gamma^{-1}\tau(x_1)x_2 \otimes x_3,\ \Delta_\cdot(x) = x_1 \otimes x_2.$$

***Proof.*** We can finish the proof by [7, Corollary 2.4] and Proposition 3.6. □

Now we use co-quasi-idempotent elements to construct quasi-idempotent Rota-Baxter operators.

**Proposition 3.8.** For a fixed co-quasi-idempotent element $\tau \in C^*$ of weight $\gamma$, we define linear map $Q_\tau : C \longrightarrow C$ by $Q_\tau(c) = \tau(c_1)c_2$ for any $c \in C$. Then $Q_\tau$ is a quasi-idempotent Rota-Baxter operator of weight $\gamma$ on $C$.

***Proof.*** It is direct to prove that $Q_\tau^2 = -\gamma Q_\tau$ by the definition of co-quasi-idempotent element. Next for any $c \in C$, we have

$$\begin{aligned}
&Q_\tau(c)_1 \otimes Q_\tau(Q_\tau(c)_2) + Q_\tau(Q_\tau(c)_1) \otimes Q_\tau(c)_2 + \gamma Q_\tau(c)_1 \otimes Q_\tau(c)_2\\
&\quad = \tau(c_1)c_{21} \otimes \tau(c_{221})c_{222} + \tau(c_1)\tau(c_{211})c_{212} \otimes c_{22} + \gamma\tau(c_1)c_{21} \otimes c_{22}\\
&\quad = \tau(c_1)c_{21} \otimes \tau(c_{221})c_{222} - \gamma\tau(c_1)c_{21} \otimes c_{22} + \gamma\tau(c_1)c_{21} \otimes c_{22}\\
&\quad = \tau(c_{11})c_{12} \otimes \tau(c_{21})c_{22}\\
&\quad = Q_\tau(c_1) \otimes Q_\tau(c_2),
\end{aligned}$$

finishing the proof. □

**Theorem 3.9.** Let $H$ be a bialgebra. Suppose that $\xi \in H$ is a quasi-idempotent of weight of $\lambda$ and $\tau \in H^*$ is a co-quasi-idempotent element of weight $\gamma$, then $(H, R_\xi, Q_\tau)$ is a Rota-Baxter bialgebra of weight $(\lambda, \gamma)$, where

$$R_\xi(x) = \xi x, \qquad Q_\tau(x) = \tau(x_1)x_2,$$

for all $x \in H$.

***Proof.*** By [7, Prosition 2.2] and Proposition 3.8, we can finish the proof. $\square$

Let recall the following result from [10] on finite dimensional Hopf algebra. As we know, a Hopf algebra $H$ is a bialgebra $H$ with an antipode $S$, where the linear map $S : H \longrightarrow H$ is the convolution inverse of identity map $\mathrm{id}_H$ in convolution algebra $Hom(H, H)$.

Let $H$ be a finite dimensional Hopf algebra. Then there is a unique element $x_H$ such that

$$\langle a^*, x_H\rangle = Tr(l_{a^*}),\ \forall\ a^* \in H^*.$$

Furthermore, the element $x_H$ has the following properties.

$$\varepsilon(x_H) = \dim(H), \qquad x_H^2 = \varepsilon(x_H)x_H.$$

that is to say, $x_H \in H$ is a quasi-idempotent element of weight $-\dim(H)$ on $H$.

When $H$ is finite dimensional, $H^*$ is also a finite dimensional Hopf algebra and $\dim(H^*)$=$\dim(H)$. So using the above result to finite dimensional Hopf algebra $H^*$, we can get: there is a unique element $\chi_H \in H^*$ such that

$$\langle \chi_H, a\rangle = Tr(l_a),\ \forall\ a \in H.$$

Furthermore, the element $\chi_H$ has the following properties.

$$\varepsilon_{H^*}(\chi_H) = \langle \chi_H, 1_H\rangle = \dim(H), \qquad \chi_H^2 = \varepsilon_{H^*}(\chi_H)\chi_H$$

$$i.e., \chi_H(a_1)\chi_H(a_2) = \langle \chi_H, 1_H\rangle\chi_H(a) = \dim(H)\chi_H(a),$$

that is to say, $\chi \in H^*$ is a co-quasi-idempotent element of weight $-\dim(H)$ on $H$.

Also we know the integral $\Lambda$ and cointegral $\bigwedge$ (i.e. integral of $H^*$) for finite dimensional Hopf algebra $H$ must exist, and $\Lambda$ is a quasi-idempotent element and $\bigwedge$ is a co-quasi-idempotent element.

By combining the discussions above, we see that $R_{x_H}$, $R_\Lambda$ and $Q_\chi$, $Q_{\bigwedge}$ are Rota-Baxter operators on $H$. As a consequence, we have

**Theorem 3.10.** Every finite dimensional Hopf algebra admits nontrivial Rota-Baxter coalgebra and bialgebra structures and tridendriform coalgebra and bialgebra structures.

## 4. An example

The well-known Sweedler's four-dimensional Hopf algebra $H_4$ is a very popular example in the theory of Hopf algebras, and many researchers pay their attention to it because there are many nice properties on it. In this section, we will apply the above results in Section 3 to $H_4$, and give all the forms of (co)-quasi-idempotent elements and related structures of tridendriform (co, bi)algebras and Rota-Baxter (co, bi)algebras.

Let $H_4$ be the algebra generated by two elements $x$ and $y$ subject to

$$x^2 = 1, \quad y^2 = 0, \quad yx = -xy.$$

Then $H_4$ is a four-dimensional algebra with a linear basis $\{1, x, y, xy\}$ (see [10, 12]), explicitly, its multiplication is

| $\mu_{H_4}$ | $1$ | $x$ | $y$ | $xy$ |
|---|---|---|---|---|
| $1$ | $1$ | $x$ | $y$ | $xy$ |
| $x$ | $x$ | $1$ | $xy$ | $y$ |
| $y$ | $y$ | $-xy$ | $0$ | $0$ |
| $xy$ | $xy$ | $-y$ | $0$ | $0$ |

.

Moreover it is a Hopf algebra equipped with the following operations:

$$\Delta(x) = x \otimes x, \quad \Delta(y) = 1 \otimes y + y \otimes x,$$

$$\varepsilon(x) = 1, \quad \varepsilon(y) = 0,$$

$$S(x) = x, \quad S(y) = xy.$$

Denote by $\{f_1, f_2, f_3, f_4\}$ the dual basis of $\{1, x, y, xy\}$, i.e.,

| | $1$ | $x$ | $y$ | $xy$ |
|---|---|---|---|---|
| $f_1$ | 1 | 0 | 0 | 0 |
| $f_2$ | 0 | 1 | 0 | 0 |
| $f_3$ | 0 | 0 | 1 | 0 |
| $f_4$ | 0 | 0 | 0 | 1 |

Then the multiplication of ${H_4}^*$ is

| $\mu_{{H_4}^*}$ | $f_1$ | $f_2$ | $f_3$ | $f_4$ |
|---|---|---|---|---|
| $f_1$ | $f_1$ | 0 | $f_3$ | 0 |
| $f_2$ | 0 | $f_2$ | 0 | $f_4$ |
| $f_3$ | 0 | $f_3$ | 0 | 0 |
| $f_4$ | $f_4$ | 0 | 0 | 0 |

Thus by the definitions of (co-)quasi-idempotent element, we have

| | quasi-idempotent element $\xi$ | weight $\lambda$ |
|---|---|---|
| $\xi_1$ | $l_1(1+x) + l_2 y + l_3 xy$ | $-2l_1$ |
| $\xi_2$ | $l_1(1-x) + l_2 y + l_3 xy$ | $-2l_1$ |
| $\xi_3$ | $l_1 1$ | $-l_1$ |

| | co-quasi-idempotent element $\tau$ | weight $\gamma$ |
|---|---|---|
| $\tau_1$ | $k_1 f_2 + k_2 f_3 + k_3 f_4$ | $-k_1$ |
| $\tau_2$ | $k_1 f_1 + k_2 f_3 + k_3 f_4$ | $-k_1$ |
| $\tau_3$ | $k_1 f_1 + k_1 f_2$ | $-k_1$ |
| $\tau_4$ | $k_1 f_3 + k_2 f_4$ | 0 |

where $k_i, l_j \in \mathbb{C}, i, j = 1, 2, 3$.

Next we assume that $k_1 \neq 0$ and $l_1 \neq 0$.

By [7, Corollary 2.4], if we set $l = (-2l_1)^{-1}$, then the tridendriform algebra structures on $H_4$ are given by $(H_4, \prec_i, \succ_i, \mu_{H_4}), i = 1, 2, 3$, where

| $\prec_1$ | $1$ | $x$ | $y$ | $xy$ |
|---|---|---|---|---|
| $1$ | $l\xi_1$ | $l(l_1(1+x) - l_3 y - l_2 xy)$ | $-\frac{1}{2}(y+xy)$ | $-\frac{1}{2}(y+xy)$ |
| $x$ | $l(l_1(1+x) + l_3 y + l_2 xy)$ | $l(l_1(1+x) - l_2 y - l_3 xy)$ | $-\frac{1}{2}(y+xy)$ | $-\frac{1}{2}(y+xy)$ |
| $y$ | $-\frac{1}{2}(y-xy)$ | $-\frac{1}{2}(y-xy)$ | 0 | 0 |
| $xy$ | $\frac{1}{2}(y+xy)$ | $\frac{1}{2}(y+xy)$ | 0 | 0 |

| $\succ_1$ | $1$ | $x$ | $y$ | $xy$ |
|---|---|---|---|---|
| $1$ | $l\xi_1$ | $l(l_1(1+x) - l_3 y - l_2 xy)$ | $-\frac{1}{2}(y+xy)$ | $-\frac{1}{2}(y+xy)$ |
| $x$ | $l(l_1(1+x) - l_3 y - l_2 xy)$ | $l\xi_1$ | $-\frac{1}{2}(y+xy)$ | $-\frac{1}{2}(y+xy)$ |
| $y$ | $-\frac{1}{2}(y+xy)$ | $\frac{1}{2}(y+xy)$ | 0 | 0 |
| $xy$ | $-\frac{1}{2}(y+xy)$ | $\frac{1}{2}(y+xy)$ | 0 | 0 |

| $\prec_2$ | $1$ | $x$ | $y$ | $xy$ |
|---|---|---|---|---|
| $1$ | $l\xi_2$ | $l(l_1(-1+x)-l_3y-l_2xy)$ | $-\frac{1}{2}(y-xy)$ | $-\frac{1}{2}(-y+xy)$ |
| $x$ | $l(l_1(-1+x)+l_3y+l_2xy)$ | $l(l_1(1-x)-l_2y-l_3xy)$ | $\frac{1}{2}(y-xy)$ | $-\frac{1}{2}(y-xy)$ |
| $y$ | $-\frac{1}{2}(y+xy)$ | $\frac{1}{2}(y+xy)$ | $0$ | $0$ |
| $xy$ | $-\frac{1}{2}(y+xy)$ | $\frac{1}{2}(y+xy)$ | $0$ | $0$ |

| $\succ_2$ | $1$ | $x$ | $y$ | $xy$ |
|---|---|---|---|---|
| $1$ | $l\xi_2$ | $l(l_1(-1+x)-l_3y-l_2xy)$ | $-\frac{1}{2}(y-xy)$ | $\frac{1}{2}(y-xy)$ |
| $x$ | $l(l_1(-1+x)-l_3y-l_2xy)$ | $l\xi_2$ | $\frac{1}{2}(y-xy)$ | $-\frac{1}{2}(y-xy)$ |
| $y$ | $-\frac{1}{2}(y-xy)$ | $-\frac{1}{2}(y-xy)$ | $0$ | $0$ |
| $xy$ | $\frac{1}{2}(y-xy)$ | $\frac{1}{2}(y-xy)$ | $0$ | $0$ |

and $\prec_3=\succ_3=\mu_{H_4}$.

By Proposition 3.6, if we set $k=(-k_1)^{-1}$, then the tridendriform coalgebra structures on $H_4$ are given by $(H_4,\Delta_{\prec_j},\Delta_{\succ_j},\Delta_{H_4}), j=1,2,3$, where

$$\left|\begin{array}{c} \Delta_{\prec_1}(1)=0\\ \Delta_{\prec_1}(x)=-x\otimes x\\ \Delta_{\prec_1}(y)=lk_2 1\otimes x-y\otimes x\\ \Delta_{\prec_1}(xy)=-x\otimes xy+lk_3x\otimes 1\end{array}\right|\begin{array}{c} \Delta_{\succ_1}(1)=0\\ \Delta_{\succ_1}(x)=-x\otimes x\\ \Delta_{\succ_1}(y)=lk_2x\otimes x\\ \Delta_{\succ_1}(xy)=-x\otimes xy-xy\otimes 1+lk_3 1\otimes 1\end{array}\right|$$

$$\left|\begin{array}{c} \Delta_{\prec_2}(1)=-1\otimes 1\\ \Delta_{\prec_2}(x)=0\\ \Delta_{\prec_2}(y)=-1\otimes y+lk_2 1\otimes x\\ \Delta_{\prec_2}(xy)=lk_3x\otimes 1-xy\otimes 1\end{array}\right|\begin{array}{c} \Delta_{\succ_2}(1)=-1\otimes 1\\ \Delta_{\succ_2}(x)=0\\ \Delta_{\succ_2}(y)=-1\otimes y-y\otimes x+lk_2x\otimes x\\ \Delta_{\succ_2}(xy)=lk_3 1\otimes 1\end{array}\right|$$

and

$$\begin{aligned} \Delta_{\prec_3}(1)&=\Delta_{\succ_3}(1)=-1\otimes 1,\\ \Delta_{\prec_3}(x)&=\Delta_{\succ_3}(x)=-x\otimes x,\\ \Delta_{\prec_3}(y)&=\Delta_{\succ_3}(y)=-1\otimes y-y\otimes x,\\ \Delta_{\prec_3}(xy)&=\Delta_{\succ_3}(xy)=-x\otimes xy-xy\otimes 1.\end{aligned}$$

With notations above, then by Theorem 3.7, the tridendriform bialgebra structures on $H_4$ are given by $(H_4,\prec_i,\succ_i,\mu_{H_4},\Delta_{\prec_j},\Delta_{\succ_j},\Delta_{H_4}), i,j=1,2,3$.

By [7, Prosition 2.2], $(H,R_{\xi_i}), i=1,2,3$ are Rota-Baxter algebras of weight $\lambda_i, i=1,2,3$, where $\lambda_1=\lambda_2=-2l_1$, $\lambda_3=-l_1$ and

| | $R_{\xi_1}$ | $R_{\xi_2}$ | $R_{\xi_3}$ |
|---|---|---|---|
| $1$ | $\xi_1$ | $\xi_2$ | $\xi_3$ |
| $x$ | $l_1(1+x)-l_3y-l_2xy$ | $l_1(-1+x)-l_3y-l_2xy$ | $l_1x$ |
| $y$ | $l_1(y+xy)$ | $l_1(y-xy)$ | $l_1y$ |
| $xy$ | $l_1(y+xy)$ | $l_1(-y+xy)$ | $l_1xy$ |

.

By Proposition 3.8, $(H,Q_{\tau_j}), j=1,2,3,4$ are Rota-Baxter coalgebras of weight $\gamma_j, j=1,2,3,4$, where $\gamma_1=\gamma_2=\gamma_3=-k_1$, $\gamma_4=0$ and

| | $Q_{\tau_1}$ | $Q_{\tau_2}$ | $Q_{\tau_3}$ | $Q_{\tau_4}$ |
|---|---|---|---|---|
| $1$ | $0$ | $k_1 1$ | $k_1 1$ | $0$ |
| $x$ | $k_1x$ | $0$ | $k_1x$ | $0$ |
| $y$ | $k_2x$ | $k_1y$ | $k_1y$ | $0$ |
| $xy$ | $k_1xy+k_3 1$ | $k_3 1$ | $k_1xy$ | $k_2 1$ |

.

With notations above, then by Theorem 3.9, $(H, R_{\xi_i}, Q_{\tau_j}), i = 1,2,3, j = 1,2,3,4$ are Rota-Baxter bialgebras of weight $(\lambda_i, \gamma_j), i = 1,2,3, j = 1,2,3,4$.

**Acknowledgment.** The authors are deeply indebted to the referees for their very useful suggestions and some improvements to the original manuscript. This work was partially supported by 2020 Research and innovation funding project for Postgraduates of Henan Normal University, Natural Science Foundation of Henan Province (No. 20A110019) and National Natural Science Foundation of China (No. 11801150). T. Ma is grateful to the Erasmus Mundus project FUSION for supporting the postdoctoral fellowship visiting to Mälardalen University, Västeras, Sweden and to the Division of Applied Mathematics at the School of Education, Culture and Communication for cordial hospitality.

## References

[1] G. Baxter, *An analytic problem whose solution follows from a simple algebraic identity*, Pacific J. Math. **10**, 731–742, 1960.

[2] L. Guo, *An Introduction to Rota-Baxter Algebra*, Surveys of Modern Mathematics, 4. International Press, Somerville, MA; Higher Education Press, Beijing, 2012.

[3] L. Guo, *Properties of free Baxter algebras*, Adv. Math. **151**, 346–374, 2000.

[4] L. Guo and W. Keigher, *Baxter algebras and shuffle products*, Adv. Math. **150**, 117–149, 2000.

[5] L. Guo and B. Zhang, *Polylogarithms and multiple zeta values from free Rota-Baxter algebras*, Sci. China Math. **53** (9), 2239–2258, 2010.

[6] L. Guo, J.-Y. Thibon and H. Yu, *Weak composition quasi-symmetric functions, Rota-Baxter algebras and Hopf algebras*, Adv. Math. **344**, 1–34, 2019.

[7] R.Q. Jian, *Quasi-idempotent Rota-Baxter operators arising from quasi-idempotent elements*, Lett. Math. Phys. **107**, 367–374, 2017.

[8] R.Q. Jian and J. Zhang, *Rota-Baxter coalgebras*, arXiv:1409.3052.

[9] T.S. Ma and L.L. Liu, *Rota-Baxter coalgebras and Rota-Baxter bialgebras*, Linear Multilinear Algebra, **64** (5), 968–979, 2016.

[10] D.E. Radford, *Hopf Algebras, KE Series on Knots and Everything*, World Scientific, Vol. **49**, New Jersey, 2012.

[11] G.C. Rota, *Baxter algebras and combinatorial identities I, II*, Bull. Amer. Math. Soc. **75** (2), 325–329, 330–334, 1969.

[12] E.J. Taft, *The order of the antipode of finite dimensional Hopf algebra*, Proc. Nat. Acad. Sci. USA. **68**, 2631–2633, 1971.

Hacettepe Journal of
Mathematics & Statistics

Hacet. J. Math. Stat.
Volume 50 (1) (2021), 224 – 234
DOI : 10.15672/hujms.703437

Research Article

# A higher version of Zappa products for monoids

Ahmet Sinan Cevik*[1,3], Suha Ahmad Wazzan[1], Firat Ates[2]

[1] *Department of Mathematics, KAU King Abdulaziz University, Science Faculty, 21589, Jeddah-Saudi Arabia*
[2] *Department of Mathematics, Science and Art Faculty, Balikesir University, Campus, 10100, Balikesir, Turkey*
[3] *Department of Mathematics, Faculty of Science, Selcuk University, Campus, 42075, Konya, Turkey*

**Abstract**

For arbitrary monoids $A$ and $B$, a presentation for the restricted wreath product of $A$ by $B$ that is known as the semi-direct product of $A^{\oplus B}$ by $B$ has been widely studied. After that a presentation for the Zappa product of $A$ by $B$ was defined which can be thought as the mutual semidirect product of given these two monoids under a homomorphism $\psi : A \to \mathcal{T}(B)$ and an anti-homomorphism $\delta : B \to \mathcal{T}(A)$ into the full transformation monoid on $B$, respectively on $A$. As a next step of these above results, by considering the monoids $A^{\oplus B}$ and $B^{\oplus A}$, we first introduce an extended version (generalization) of the Zappa product and then we prove the existence of an implicit presentation for this new product. Furthermore we present some other outcomes of the main theories in terms of finite and infinite cases, and also in terms of groups. At the final part of this paper we point out some possible future problems related to this subject.

**Mathematics Subject Classification (2020).** 20E22, 20F05, 20L05, 20M05

**Keywords.** Knit products, Wreath products, Zappa products, presentations

## 1. Introduction

Study on the product of groups have received much attention in the literature. During these studies, people investigated this group product which is constructed by subgroups either in terms of permutability (cf. [6, 9, 17]) or in terms of an extension (cf. [5, 24]). Nevertheless, direct, semidirect and (standard) wreath products are the most famous structures among these extension constructions (see, for instance, [10, 14, 18, 20, 25]). As a next step of these products, some other people also studied *Zappa* (or Zappa-Szép) *products* ([13, 16, 27, 28]) which is also referred as bilateral semidirect products ([22]), general products ([23]) or knit products ([1, 26]). Unlikely semi-direct products, none of the factor is normal in the Zappa product of any two groups. In other words, for a group $G$ with subgroups $A$ and $B$ that satisfy $A \cap B = \{1_G\}$ and $G = AB$, we know that each element $g \in G$ is expressible (uniquely) as $g = ab$ with $a \in A$ and $b \in B$. Now to reserve

*Corresponding Author.
Email addresses: ahmetsinancevik@gmail.com (A.S. Cevik), swazzan@kau.edu.sa (S.A. Wazzan), firat@balikesir.edu.tr (F. Ates)
Received: 13.03.2020; Accepted: 31.05.2020

certain products, let us consider an element $ba \in G$. In fact there must be unique elements $b' \in B$ and $a' \in A$ such that $ba = a'b'$. This actually implies two functions

$$(b, a) \longmapsto b^a \in B, \qquad (b, a) \longmapsto b.a = {}^b a \in A \tag{1.1}$$

which are unique and so satisfy

$$ba = (b.a)(b^a) = {}^b a b^a, \tag{1.2}$$

for all $b \in B$ and $a \in A$.

According to the references [13, 22, 23, 25], by considering the action given (1.1), the monoid version of the Zappa product of any two monoids can be defined as follows.

For any two monoids $A$ and $B$, let us consider a homomorphism $\psi : A \to \mathcal{T}(B)$ and an anti-homomorphism $\delta : B \to \mathcal{T}(A)$ such that $\mathcal{T}(.)$ denotes the full transformation monoid. For $a \in A,\ b \in B$, denote the operation of $(a)\psi$ on $B$ by $b \longmapsto (a)\psi = b^a$ and the operation of $(b)\delta$ on $A$ by $a \longmapsto (a)\delta_b = {}^b a$. For every elements $a, a_1, a_2 \in A,\ b, b_1, b_2 \in B$, suppose that the conditions

$$\begin{aligned} b^{1_A} = b, \quad 1_B^a = 1_B, \qquad & (1_A)\delta_b = 1_A, \quad (a)\delta_{1_B} = a, \\ b^{(a_1 a_2)} = (b^{a_1})^{a_2}, \qquad & (a)\delta_{b_1 b_2} = ((a)\delta_{b_2})\delta_{b_1}, \\ (b_1 b_2)^a = b_1^{(a)\delta_{b_2}} b_2^a \quad \text{and} \quad & (a_1 a_2)\delta_b = (a_1)\delta_b (a_2)\delta_{b^{a_1}} \end{aligned}$$

are all true. Then the set $A \times B$ defines the Zappa product $A_\delta \times_\psi B$ (cf. [13, 22]) of $A$ and $B$ which is of course a monoid with respect to the multiplication,

$$(a_1, b_1)(a_2, b_2) = (a_1 (a_2)\delta_{b_1}, b_1^{a_2} b_2)\,. \tag{1.3}$$

Assume that $A$ has a monoid presentation $\mathcal{P}_A = [X; R]$ while $B$ has $\mathcal{P}_B = [Y; S]$. Then, by [23, Theorem 2], a presentation for $A_\delta \times_\psi B$ with the structure defined by (1.3) on the set $A \times B$ is given as $\mathcal{P} = [X,\, Y\,;\, R,\, S,\, T]$ in which the relator $T$ consists of all ordered elements $(ba, {}^b a b^a)$, as given in (1.2), for $(b, a) \in B \times A$.

Since there are some difficulties in the meaning of embedding for the factors in the product unless they are not taken as identities, throughout in this paper we will not attempt to study the cases of Zappa products for semigroups.

To give another preliminary material for the next section, let us recall the fundamentals of standard wreath products of any two monoids $A$ and $B$. First let us consider the monoid $A^{\oplus B}$ which is the direct product of the number of $B$ copies of $A$. In fact $A^{\oplus B}$ can be thought as the set of all functions $f$ having finite support. Suppose that $\psi : A^{\oplus B} \to \mathcal{T}(B)$ is a homomorphism and $\delta : B \to \mathcal{T}(A^{\oplus B})$ is an anti-homomorphism where $\mathcal{T}(.)$ is the full transformation monoid on $B$ and $A^{\oplus B}$, respectively, as previously. For $g \in A^{\oplus B}$ and $b \in B$, let us denote the operation of $(g)\psi$ on $B$ by $b \longmapsto b$ and operation of $(b)\delta$ on $A^{\oplus B}$ by $g \longmapsto (g)\delta_b = {}^b g$. Then the set $A^{\oplus B} \times B$ defines a monoid $A \wr B$ (namely the (standard) wreath product of $A$ by $B$) with the operation $(f, b_1)(g, b_2) = (f\ {}^{b_1}g, b_1 b_2)$, and the identity is $(I, 1_B)$, where $(x)I = 1_A$ (cf. [14, 18, 20, 22]). It is clear that $A \wr B$ is actually the semidirect product of $A^{\oplus B}$ by $B$ and notated by $A^{\oplus B} \times_\delta B$. Now, by taking into account the same presentations $\mathcal{P}_A$ and $\mathcal{P}_B$ for the monoids $A$ and $B$ as in above, for each $b \in B$, let us assume the set $X_b = \{x_b \ :\ x \in X\}$ is a copy of $X$ and the set $R_b$ is the corresponding copy of $R$. So, for $x, x' \in X$, $y \in Y$, $b, e \in B$, $b \neq e$, the monoid $A \wr B$ has a presentation

$$\left[ X_b, Y\,;\, R_b,\, S,\, x_b x'_e = x'_e x_b,\, y x_b = (\prod_{m \in by^{-1}} x_m) y \right] \tag{1.4}$$

(cf. [2, 14, 18, 25]).

## 2. A higher version of the Zappa product

By combining the definitions of Zappa and (standard) wreath products, the main purposes of this section are to define and study a generalized version of the Zappa product of $A^{\oplus B}$ by $B^{\oplus A}$, namely restricted generalized Zappa product $A^{\oplus B}{}_{\delta}\times_{\psi}B^{\oplus A}$ with an operation adapted from (1.3). Additionally, by considering the presentation in (1.4), we will prove the existence of an implicit presentation for this product (see Theorem 2.2 below). Moreover, by taking into account a special case $A^{\oplus B}{}_{\delta}\times_{\psi}B$ of this new product, we will state and prove some consequences of this theorem.

Let $A$ and $B$ be monoids, and let the set $A^{\times B}$ denotes the Cartesian product of the number of $B$ copies of the monoid $A$ while the set $A^{\oplus B}$ denotes the corresponding direct product as in wreath products. Recall that $A^{\oplus B}$ can be thought as the set of whole functions $f$ with finite support (in other words, functions with the property $(x)f = 1_A$ for all but finitely many $x$ in $B$). Hence a generalization of restricted and unrestricted Zappa products of the monoid $A^{\oplus B}$ by the monoid $B^{\oplus A}$ are defined on $A^{\times B} \times B^{\times A}$ and $A^{\oplus B} \times B^{\oplus A}$, respectively, with the multiplication

$$(f,h)(g,k) = (f\ (g)\delta_h, (h)\psi_g k) = (f\ {}^{h}g, h^{g}k), \tag{2.1}$$

where $\delta : B^{\oplus A} \to \mathfrak{T}(A^{\oplus B})$, $(g)\delta_h = {}^{h}g$ and $\psi : A^{\oplus B} \to \mathfrak{T}(B^{\oplus A})$, $(h)\psi_g = h^g$ are defined by, for $a \in A$ and $b \in B$,

$${}^{h}g = {}^{(h^a)}g \quad \text{ and } \quad h^g = h^{({}^{b}g)}.$$

Also, for $x \in A$ and $y \in B$, we define

$$(x)h^a = (ax)h \quad \text{ and } \quad (y)\ {}^{b}g = (yb)g \tag{2.2}$$

such that, for all $d \in B$, $c \in A$,

$$(d)\ {}^{(h^a)}g = (dh^a)g \quad \text{ and } \quad (c)h^{({}^{b}g)} = ({}^{b}gc)h.$$

Both these restricted and unrestricted generalized Zappa products are monoids under the multiplication defined in (2.1) with the identity $(\overline{1}, \tilde{1})$, where $\overline{1} : B \to A$, $(b)\overline{1} = 1_A$ and $\tilde{1} : A \to B$, $(a)\tilde{1} = 1_B$, for all $a \in A$ and $b \in B$.

Throughout this paper all generalized Zappa products will be assumed to be restricted and so we will use the notation $A^{\oplus B}{}_{\delta}\times_{\psi}B^{\oplus A}$ for it. It is clear that the sets $\{(f,\tilde{1}) : f \in A^{\oplus B}\}$ and $\{(\overline{1},k) : k \in B^{\oplus A}\}$ are the submonoids of $A^{\oplus B}{}_{\delta}\times_{\psi}B^{\oplus A}$ which are isomorphic to $A^{\oplus B}$ and $B^{\oplus A}$, respectively. Moreover, for $f \in A^{\oplus B}$ and $k \in B^{\oplus A}$, we definitely have $(f,\tilde{1})(\overline{1},k) = (f,k)$.

For $a \in A$ and $b \in B$, we now define $\overline{a_b} : B \to A$ and $\widetilde{b_a} : A \to B$ as

$$(m)\overline{a_b} = \begin{cases} a, & b = m \\ 1_A, & \text{otherwise} \end{cases} \quad \text{ and } \quad (n)\widetilde{b_a} = \begin{cases} b, & a = n \\ 1_B, & \text{otherwise} \end{cases}.$$

Notice that if $f : B \to A$ and $k : A \to B$ have finite supports, then

$$f = \prod_{b \in B} \overline{((b)f)_b} \quad \text{ and } \quad k = \prod_{a \in A} \widetilde{((a)k)_a}.$$

Also notice that if the monoid $A$ is generated by a set $X$ (so that every $a$ in $A$ is expressible as a finite product $x_1x_2\cdots x_n$ of elements of $X$) and if the monoid $B$ is generated by $Y$ (so every $b$ in $B$ is expressible as a finite product $y_1y_2\cdots y_m$), then

$$\overline{a_b} = \overline{x_{1_b}}\,\overline{x_{2_b}}\cdots\overline{x_{n_b}} \quad \text{ and } \quad \widetilde{b_a} = \widetilde{y_{1_a}}\widetilde{y_{2_a}}\cdots\widetilde{y_{m_a}}.$$

After all, we have the following lemma which is actually a generalization of [18, Lemma 2.1].

**Lemma 2.1.** *Assume that the sets $X$ and $Y$ generate the monoids $A$ and $B$, respectively. Further, let $\overline{X_b} = \{(\overline{x_b}, \widetilde{1}) : b \in B, x \in X\}$ and $\widetilde{Y_a} = \{(\overline{1}, \widetilde{y_a}) : a \in A, y \in Y\}$. Then the product $A^{\oplus B}{}_{\delta}\times_{\psi} B^{\oplus A}$ is generated by the set $(\bigcup_{b\in B} \overline{X_b}) \cup (\bigcup_{a\in A} \widetilde{Y_a})$.*

In general, the generating set given in Lemma 2.1 is the best possible for the monoids $A$ and $B$. If $B$ has an indecomposable identity (in other words, for all $b, c \in B$, $bc = 1_B \Rightarrow b = c = 1_B$), then any generating set of $A^{\oplus B}{}_{\delta}\times_{\psi} B^{\oplus A}$ must contain elements from the generating set of the submonoid $A^{\oplus B} \cong \{(f, \widetilde{1}) : f \in A^{\oplus B}\}$ and, in fact, $\bigcup_{b\in B} \overline{X_b}$ is the smallest such a set. One may discuss same arguments for $\bigcup_{a\in A} \widetilde{Y_a}$ as well.

For simplicity, we will denote the set $\{m \in B : b = my\}$ with only $by^{-1}$ (where $b, y \in B$) and will denote the set $\{n \in A : a = xn\}$ with only $x^{-1}a$ (where $a, x \in A$).

The following theorem generalizes the result presented in [13].

**Theorem 2.2.** *Suppose that the monoids $A$ and $B$ are presented by $[X; R]$ and $[Y; S]$, respectively. For each $b \in B$, let $X_b = \{x_b : x \in X\}$ denote a copy of $X$, and let $R_b$ denote the corresponding copy of $R$. Similarly, for each $a \in A$, let $Y_a = \{y_a : y \in Y\}$ be a copy of $Y$, and let $S_a$ be the corresponding copy of $S$. Then the (restricted) generalized Zappa product $A^{\oplus B}{}_{\delta}\times_{\psi} B^{\oplus A}$ is defined by the generators $(\bigcup_{b\in B} X_b) \cup (\bigcup_{a\in A} Y_a)$ and relations*

$$R_b, S_a, \quad (a \in A,\ b \in B)\,; \tag{2.3}$$

$$x_b x'_e = x'_e x_b, \quad (x, x' \in X,\ b, e \in B,\ b \neq e)\,; \tag{2.4}$$

$$y_a y'_s = y'_s y_a, \quad (y, y' \in Y,\ a, s \in A,\ a \neq s)\,; \tag{2.5}$$

$$y_a x_b = \left(\prod_{m\in by'^{-1}} x_m\right)\left(\prod_{n\in x'^{-1}a} y_n\right) \tag{2.6}$$

*such that the elements $x'$ and $y'$ in Eq. (2.6) are defined as*

$$x' = \prod_{m\in by^{-1}} x_m \quad \text{and} \quad y' = \prod_{n\in x^{-1}a} y_n\,,$$

*respectively.*

***Proof.*** We first recall that, for a set of alphabet $\mathfrak{M}$, the monoid of all words in $\mathfrak{M}$ is notated by $\mathfrak{M}^*$.

For $x \in X, b \in B, y \in Y, a \in A$, the mapping $\rho$ from the monoid $\left((\bigcup_{b\in B} X_b) \cup (\bigcup_{a\in A} Y_a)\right)^*$, say $M$, to the product $A^{\oplus B}{}_{\delta}\times_{\psi} B^{\oplus A}$ defined by $(x_b)\rho = (\overline{x_b}, \widetilde{1})$ and $(y_a)\rho = (\overline{1}, \widetilde{y_a})$ is surjective as a result of Lemma 2.1. Furthermore, relations in (2.3), (2.4) and (2.5) are all held in $A^{\oplus B}{}_{\delta}\times_{\psi} B^{\oplus A}$ by the equalities and explanations presented just before Lemma 2.1.

Now the next step is to obtain relation (2.6). We easily deduce from (2.1) that

$$(\overline{1}, \widetilde{y_a})(\overline{x_b}, \widetilde{1}) = ({}^{\widetilde{y_a}}\overline{x_b}, \widetilde{y_a}^{\overline{x_b}}).$$

Now by considering (2.2), for each $x \in X$, we can write

$${}^{\widetilde{y_a}}\overline{x_b} = {}^{(\widetilde{y_a}^{x})}\overline{x_b}\,,$$

where for $d \in A$,

$$\begin{aligned} (d)\widetilde{y_a}^{x} = (xd)\widetilde{y_a} \quad &= \begin{cases} y, & a = xd \\ 1_B, & \text{otherwise} \end{cases} = \begin{cases} y, & d \in x^{-1}a \\ 1_B, & \text{otherwise} \end{cases} \\ &= \prod_{n\in x^{-1}a} (d)\widetilde{y_n} = (d) \prod_{n\in x^{-1}a} \widetilde{y_n}. \end{aligned}$$

So we have $\widetilde{y_a}^x = \prod_{n\in x^{-1}a} \widetilde{y_n}$. For simplicity, let us denote $\prod_{n\in x^{-1}a} \widetilde{y_n}$ by only $y'$. As a result, we obtain

$$^{(\widetilde{y_a}^x)}\overline{x_b} = {}^{y'}\overline{x_b}.$$

Moreover, for $e \in B$,

$$\begin{aligned}(e)^{y'}\overline{x_b} = (ey')\overline{x_b} &= \begin{cases} x, & b = ey' \\ 1_A, & \text{otherwise} \end{cases} = \begin{cases} x, & e \in by'^{-1} \\ 1_A, & \text{otherwise} \end{cases} \\ &= \prod_{m\in by'^{-1}} (e)\overline{x_m} = (e) \prod_{m\in by'^{-1}} \overline{x_m}.\end{aligned}$$

Therefore ${}^{y'}\overline{x_b} = \prod_{m\in by'^{-1}} \overline{x_m}$ and finally we have

$$^{\widetilde{y_a}}\overline{x_b} = {}^{(\widetilde{y_a}^x)}\overline{x_b} = {}^{y'}\overline{x_b} = \prod_{m\in by'^{-1}} \overline{x_m}\,.$$

Additionally, for each $y \in Y$, by taking into account the second part of (2.2) and its attachments, since

$$\widetilde{y_a}^{\overline{x_b}} = \widetilde{y_a}^{(^y\overline{x_b})}\,,$$

we clearly obtain

$$\widetilde{y_a}^{\overline{x_b}} = \widetilde{y_a}^{(^y\overline{x_b})} = \widetilde{y_a}^{x'} = \prod_{n\in x'^{-1}a} \widetilde{y_n}\,,$$

where $x' = \prod_{m\in by^{-1}} \overline{x_m}$.

Therefore, if we write all above results together, then we get

$$(\overline{1}, \widetilde{y_a})(\overline{x_b}, \tilde{1}) = \left(\prod_{m\in by'^{-1}} x_m\right)\left(\prod_{n\in x'^{-1}a} y_n\right)\,,$$

as required. As a result of all these above findings, we deduce that $\rho$ defines actually an epimorphism $\overline{\rho}$ from the monoid $M$ obtained by relations (2.3), (2.4), (2.5) and (2.6) onto the monoid $A^{\oplus B}{}_\delta\times_\psi B^{\oplus A}$.

Now we need to prove that $\rho$ is a monomorphism. Let $w$ be a word representing an element of $M$. By using relations (2.4), (2.5) and (2.6), it is easy to show that there exist words $(b)w$ in $X^*$ $(b \in B)$ and $(a)w$ in $Y^*$ $(a \in A)$ such that

$$w = (\prod_{b\in B}((b)w)_b)(\prod_{a\in A}((a)w)_a)$$

in $M$. (We note that if $z \in X^*$, $t \in Y^*$ then $z_b$ and $t_a$ are the corresponding words in $X_b^*$ and $Y_a^*$, respectively). Now, for each $w \in X^* \cup Y^*$, $c \in B$ and $d \in A$, we have

$$(c)\overline{w_b} = \begin{cases} w, & b = c \\ 1, & \text{otherwise} \end{cases} \quad \text{and} \quad (d)\widetilde{w_a} = \begin{cases} w, & a = d \\ 1, & \text{otherwise} \end{cases}.$$

Hence we get

$$(c)(\prod_{b\in B}\overline{((b)w)_b}) = \prod_{b\in B}(c)\overline{((b)w)_b} = (c)w\,, \tag{2.7}$$

$$(d)(\prod_{a\in A}\widetilde{((a)w)_a}) = \prod_{a\in A}(d)\widetilde{((a)w)_a} = (d)w\,, \tag{2.8}$$

for all $c \in B$ and $d \in A$.

For any two words $u$, $v$ in $((\bigcup_{b\in B} X_b)\cup(\bigcup_{a\in A} Y_a))^*$, we have

$$\begin{aligned}
(u)\rho=(v)\rho \;&\Rightarrow\; ((\prod_{b\in B}((b)u)_b)(\prod_{a\in A}((a)u)_a))\rho=((\prod_{b\in B}((b)v)_b)(\prod_{a\in A}((a)v)_a))\rho\\
&\Rightarrow\; ((\prod_{b\in B}((b)u)_b)\rho(\prod_{a\in A}((a)u)_a))\rho=((\prod_{b\in B}((b)v)_b)\rho(\prod_{a\in A}((a)v)_a))\rho\\
&\Rightarrow\; (\prod_{b\in B}(\overline{((b)u)_b},\tilde{1}))(\prod_{a\in A}(\overline{1},\widetilde{((a)u)_a}))=(\prod_{b\in B}(\overline{((b)v)_b},\tilde{1}))(\prod_{a\in A}(\overline{1},\widetilde{((a)v)_a}))\\
&\Rightarrow\; (\prod_{b\in B}\overline{((b)u)_b},\prod_{a\in A}\widetilde{((a)u)_a})=(\prod_{b\in B}\overline{((b)v)_b},\prod_{a\in A}\widetilde{((a)v)_a})\,.
\end{aligned}$$

Now from the equality of the first and second components and using equalities (2.7)-(2.8), we deduce that $(c)u=(c)v$ in $A$ (for all $c\in B$) and $(d)u=(d)v$ in $B$ (for all $d\in A$). Also, relations given in (2.3) imply $u=v$ in the monoid $M$. Therefore $\overline{\rho}$ is injective.

These complete the proof. □

**Remark 2.3.** For $d\in x^{-1}a$ and $e\in by^{-1}$, since $(d)\widetilde{y_a}^{x}=y$ and $(e)^{y}\overline{x_b}=x$, we have seen in the above proof there exist equalities

$$^{(\widetilde{y_a}^{x})}\overline{x_b}={}^{y}\overline{x_b}={}^{y'}\overline{x_b}\quad\text{ and }\quad\widetilde{y_a}^{({}^{y}\overline{x_b})}=\widetilde{y_a}^{x}=\widetilde{y_a}^{x'}\,.$$

Therefore, by omitting the bar and tilde signs, another version of the relation given in (2.6) can be stated as

$$y_ax_b={}^{\left(\prod\limits_{n\in x^{-1}a}y_n\right)}x_b\,y_a^{\left(\prod\limits_{m\in by^{-1}}x_m\right)}\,. \tag{2.9}$$

We have the following consequence of Theorem 2.2.

**Corollary 2.4.** *Let $A$ and $B$ be monoids with the conditions given in Theorem* 2.2 *hold. Then the standard presentation for $A^{\oplus B}{}_{\delta}\!\times_{\psi}B^{\oplus A}$ is given by*

$$\begin{aligned}
[\,X_b,Y_a\;\;;\;\;&R_b,S_a\quad(a\in A,\,b\in B)\,,\\
&x_bx'_e=x'_ex_b\quad(x,x'\in X,\;b,e\in B,\;b\neq e)\,,\\
&y_ay'_s=y'_sy_a\quad(y,y'\in Y,\;a,s\in A,\;a\neq s)\,,\\
&y_ax_b={}^{\left(\prod\limits_{n\in x^{-1}a}y_n\right)}x_b\,y_a^{\left(\prod\limits_{m\in by^{-1}}x_m\right)}\;]\,.
\end{aligned}$$

At the rest of this section, as a special case of Theorem 2.2 (and also Corollary 2.4), we will only consider the generalized Zappa product $A^{\oplus B}\;{}_{\delta}\times_{\psi}B$ for defining a presentation on it.

For an arbitrary monoid $A$ with a presentation $[X;R]$ and an arbitrary monoid $B$ with a presentation $[Y;S]$, let us consider

$$\begin{aligned}
\delta\;:\;\; &B\to\mathfrak{T}(A^{\oplus B})\quad\text{ and }\quad\psi\;:\;\; A^{\oplus B}\to\mathfrak{T}(B)\\
&b\mapsto(g)\delta_b={}^{b}g\qquad\qquad\qquad\quad g\mapsto(b)\psi_g=b^{g}
\end{aligned}$$

such that $(x)\,{}^{b}g=(xb)g$ for $x\in B$ and $b^{g}=b^{({}^{b'}g)}$ for $b'\in B$. Then the generalized Zappa product $A^{\oplus B}\;{}_{\delta}\times_{\psi}B$ is defined on the set $A^{\oplus B}\times B$ with a multiplication $(f,b)(g,b')=(f\,{}^{b}g,b^{g}b')$.

**Theorem 2.5.** *A presentation for $A^{\oplus B}\ {}_{\delta}\times_{\psi} B$ is defined by*

$$[\, X_b, Y\ ;\ R_b,\ S,\ x_b x'_e = x'_e x_b,\ y x_b = (\prod_{m\in by^{-1}} x_m) y^{(\prod_{m\in by^{-1}} x_m)}\ ], \tag{2.10}$$

*where $x, x' \in X,\ y \in Y,\ b, e \in B,\ b \neq e$.*

***Proof.*** Let us consider the presentation given in Corollary 2.4. Since we have just one copy of $B$ in the product $A^{\oplus B}\ {}_{\delta}\times_{\psi} B$, we must have $Y$ instead of $Y_a$ in the generating set and also $S$ instead of $S_a$ in the relators set of the requiring presentation. Moreover, by the same reason, the relator $y_a y'_s = y'_s y_a$ $(y, y' \in Y,\ a, s \in A,\ a \neq s)$ will be disappeared.

For the last relator, again let us consider the multiplication $(\overline{1}, y)(\overline{x_b}, 1_B) = ({}^{y}\overline{x_b}, y^{\overline{x_b}})$, where $x \in X$, $y \in Y$ and $b \in B$. Recall that, in the proof of Theorem 2.2, we obtained the equation

$${}^{y}\overline{x_b} = \prod_{m\in by^{-1}} \overline{x_m}\,.$$

Hence, by considering both (2.6) and (2.9) with the fact that there exists a single $B$ in the product $A^{\oplus B}\ {}_{\delta}\times_{\psi} B$, we obtain

$$(\overline{1}, y)(\overline{x_b}, 1_B) = (\prod_{m\in by^{-1}} x_m) y^{(\prod_{m\in by^{-1}} x_m)},$$

as required.

Notice that presentation in (2.10) is a generalization of the presentation given in (1.4) since it presents a product having mutual actions. □

As a consequence of Theorem 2.5, we can get a much nicer presentation in the case of $B$ is a group which is actually a generalization of the presentation defined in [18, Corollary 2.3].

**Corollary 2.6.** *Assume that $A$ is a monoid but $B$ is a group. Now consider their monoid presentations $[X; R]$ and $[Y; S]$, respectively. Thus $A^{\oplus B}\ {}_{\delta}\times_{\psi} B$ has a presentation*

$$\left[X, Y\ ;\ R, S, x(b^{-1}x'b^{x''}) = (b^{-1}x'b^{x''})x\right],$$

*where $x, x', x'' \in X,\ b \in B$.*

***Proof.*** Recall from (1.2), for any $a \in A$ and $b \in B$, the action satisfies $ba = {}^{b}ab^{a}$. So, for $x_b \in A^{\oplus B}$ and $b \in B$, we get

$$b x_b = {}^{b}x_b b^{x_b}. \tag{2.11}$$

Now, by replacing $b$ instead of $y$ in equations ${}^{y}\overline{x_b} = \prod_{m\in by^{-1}} \overline{x_m}$ and $y^{\overline{x_b}} = y^{({}^{y}\overline{x_b})}$, where $m \in B$, which are obtained in Theorems 2.2 and 2.5 and also by writing those new equations in (2.11), we obtain the relation

$$b x_b = x_{1_B} b^{\prod_{m\in by^{-1}} x_m}$$

in $A^{\oplus B}\ {}_{\delta}\times_{\psi} B$. For just simplicity, if we write $x'$ instead of $x_{1_B}$ and $x''$ instead of $\prod_{m\in by^{-1}} x_m$, then this above last relation becomes

$$x_b = b^{-1}x'b^{x''}. \tag{2.12}$$

Further, by using (2.12), if we eliminate the element $x_b$ (where $x \in X, b \in B - \{1_B\}$) from the relations in presentation (2.10), the last relator of this presentation becomes trivial while the relations $R_b$ and $x_b x'_e = x'_e x_b$ are actually consequences of the relations $R$ and

$x(b^{-1}x'b^{x''}) = (b^{-1}x'b^{x''})x$, respectively, in the meaning of Tietze transformations, where $x, x', x'' \in X$, $b \in B$.

Hence this completes the proof. $\square$

By taking into account both $A$ and $B$ as any groups, Corollary 2.6 can be expressed as in the following.

**Corollary 2.7.** *Assume that both $A$ and $B$ are groups with their monoid presentations $[X; R]$ and $[Y; S]$, respectively. Hence the presentation*

$$[\, X, Y \ ; \ R,\ S,\ a(b^{-1}a'b^{a''}) = (b^{-1}a'b^{a''})a \quad (b \in B,\ a, a', a'' \in A)\,]$$

*defines $A^{\oplus B}\ {}_{\delta}\times_{\psi} B$.*

***Proof.*** As in the proof of Corollary 2.6, for $a \in A$ and $b \in B$, we can easily see that

$$a_b = b^{-1}a_{1_B}b^{\prod\limits_{m \in by^{-1}} a_m}$$

holds in $A^{\oplus B}\ {}_{\delta}\times_{\psi} B$. For simplicity, let us replace $a_{1_B}$ by $a'$ and $\prod\limits_{m \in by^{-1}} a_m$ by $a''$. Then the above equality becomes $a_b = b^{-1}a'b^{a''}$. Therefore, by replacing $a_b$ in presentation (2.10), we obtain the required presentation given in the statement of corollary. $\square$

## 3. Some applications

By considering the presentation defined in Theorem 2.5 for $A^{\oplus B}{}_{\delta}\times_{\psi} B$, we will give some examples while $A$ and $B$ are taken as some special monoids.

### 3.1. Finite case

In this section we will study on finite cyclic monoids (cf. [19]). In fact some examples and applications over other extensions for these monoids have been investigated, for instance, in [3, 4, 15].

Suppose that $A = [x \ ; \ x^k = x^l \ (k > l)]$ and $B = [y \ ; \ y^s = y^t \ (s > t)]$ are finite cyclic monoids, and consider $\delta$ and $\psi$ as given in Theorem 2.5. We then have the following result.

**Corollary 3.1.** *Let $A$ and $B$ be finite cyclic monoids as in above. Then*

$$\begin{aligned}
[\, x^{(0)}, x^{(1)}, \cdots, x^{(s-1)}, y \quad ; \quad & y^s = y^t,\ x^{(i)}x^{(j)} = x^{(j)}x^{(i)} \quad (0 \le i < j \le s-1),\\
& {x^{(i)}}^k = {x^{(i)}}^l \quad (0 \le i \le s-1),\\
& yx^{(i)} = x^{(i-1)}y^{x^{(i-1)}} \quad (1 \le i \le s-1),\\
& yx^{(t)} = x^{(s-1)}y^{x^{(s-1)}}\,]
\end{aligned}$$

*is a presentation for the product $A^{\oplus B}\ {}_{\delta}\times_{\psi} B$.*

***Proof.*** By considering $A$ and $B$ are finite cyclic monoids, we just need to convert presentation (2.10) in Theorem 2.5. For all $y^i \in B$, let us label each $x_{y^i}$ by $x^{(i)}$, where $0 \le i \le s-1$, for simplicity. Therefore the set of the generators for the monoid $A^{\oplus B}\ {}_{\delta}\times_{\psi} B$ is $\{x^{(i)}, y\}$. Further, since $A^{\oplus B}$ is a direct product, we must have $x^{(i)}x^{(j)} = x^{(j)}x^{(i)}$ $(0 \le i < j \le s-1)$ and ${x^{(i)}}^k = {x^{(i)}}^l$ as relations in our presentation.

Now let us consider the relator

$$yx_b = (\prod_{m \in by^{-1}} x_m)y^{(\prod\limits_{m \in by^{-1}} x_m)}$$

in presentation (2.10). In this relator, by taking $1, y, y^2, \cdots, y^{s-1}$ instead of each $b \in B$ and replacing each $x_b$ by related $x^{(i)}$ where $0 < i \leq s-1$, we obtain the relator $yx^{(i)} = x^{(i-1)}y^{x^{(i-1)}}$. Moreover, for the monoid $B$, since we have $y^s = y^t$ as a relator, we can write this relator as $y^t = y^{s-1}y$ which implies that, for $b = y^t$ and $m = y^{s-1}$, $yx^{(t)} = x^{(s-1)}y^{x^{(s-1)}}$ by keeping same idea as in the previous sentence.

Hence this completes the proof. □

We can also give the following application which is a consequence of Corollary 2.6.

**Corollary 3.2.** *Let $A$ be a finite monoid (not necessarily cyclic) and let $B$ be a cyclic group of order $s$. If $\mathcal{P}_A = [X; R]$ and $\mathcal{P}_B = [y\,;\, y^s = y^t\ (s > t)]$ are their monoid presentations, respectively, then the presentation*

$$[X, y\,;\, R,\ y^s = y^t,\ x(y^{-i}x'(y^i)^{x'}) = (y^{-i}x'(y^i)^{x'})x \quad (x, x' \in X,\ 0 < i \leq (s-t)-1)]$$

*defines the product $A^{\oplus B}\,{}_{\delta} \times_{\psi} B$.*

***Proof.*** From Corollary 2.6, we have the relations $bx_b = x_{1_B}b^{x_{1_B}}$, for $b \in B$, $x \in X$. If we take $1, y, y^2, \cdots, y^{(s-t)-1}$ instead of for each $b$, we obtain $x^{(i)} = y^{-i}x^{(0)}(y^i)^{x^{(0)}}$ where $0 < i \leq (s-t)-1$. Also let us replace $x'$ by $x^{(0)}$. Thus we have $x^{(i)} = y^{-i}x'(y^i)^{x'}$. Hence this completes the proof. □

### 3.2. Infinite case

In this subcase, let $A$ be the free Abelian monoid rank 2 and let $B$ be the finite cyclic monoid. As a consequence of Theorem 2.5, we have the following result which can be proved quite similarly as in Corollary 3.1.

**Corollary 3.3.** *Let $\mathcal{P}_A = [x_1, x_2\,;\, x_1x_2 = x_2x_1]$ and $\mathcal{P}_B = [y\,;\, y^s = y^t\,(s > t)]$ be monoid presentations for the above monoids $A$ and $B$. Therefore, the monoid $A^{\oplus B}{}_{\delta} \times_{\psi} B$ has a presentation with generators*

$$x_1^{(0)}, x_1^{(1)}, \cdots, x_1^{(s-1)}, x_2^{(0)}, x_2^{(1)}, \cdots, x_2^{(s-1)}, y$$

*and relators*

$$\begin{aligned} y^s = y^t,\ x_i^{(m)}x_j^{(n)} &= x_j^{(n)}x_i^{(m)} && (i, j \in \{1, 2\}, \quad 0 \leq m, n \leq s-1), \\ yx_1^{(m)} &= x_1^{(m-1)}y^{x_1^{(m-1)}} && (0 < m \leq s-1), \\ yx_2^{(n)} &= x_2^{(n-1)}y^{x_2^{(n-1)}} && (0 < n \leq s-1), \\ yx_1^{(t)} &= x_1^{(s-1)}y^{x_1^{(s-1)}}, && yx_2^{(t)} = x_2^{(s-1)}y^{x_2^{(s-1)}}. \end{aligned}$$

We note that Corollary 3.3 can be easily generalized for an arbitrary free abelian monoid $A$ with rank greater than 2.

On the other hand another infinite case application of Theorem 2.5 is the following:

Let $A$ be the free monoid with a presentation $\mathcal{P}_A = [x\ ;\ ]$ and let $B$ be the monoid $\mathbb{Z}_s \times \mathbb{Z}_m$ with a presentation

$$\mathcal{P}_B = [y_1,\, y_2\,;\, y_1^s = y_1^t,\ y_2^m = y_2^n\ (s > t,\ m > n),\ y_1y_2 = y_2y_1].$$

For a representive element $y_1^iy_2^j$ in the monoid $B$, let us label $x_{y_1^iy_2^j}$ by $x^{(i,j)}$ where $0 \leq i \leq s-1$, $0 \leq j \leq m-1$. Then, for each element in $B$, we have a generating set $\{x^{(i,j)}, y_1, y_2\}$ for the monoid $A^{\oplus B}{}_{\delta} \times_{\psi} B$. Therefore, by suitable changes in presentation (2.10), we obtain the following result.

**Corollary 3.4.** *Let $A$ and $B$ be as above. Then*

$$\begin{aligned}[\, x^{(i,j)}, y_1, y_2 \quad ; \quad & y_1^s = y_1^t,\ y_2^m = y_2^n\ (s > t,\ m > n),\ y_1y_2 = y_2y_1, \\ & x^{(i,j)}x^{(l,k)} = x^{(l,k)}x^{(i,j)} \quad (0 \le i \le s-1,\ 0 \le j \le m-1,\ (i,j) < (l,k)), \\ & y_1x^{(i,j)} = x^{(i-1,j)}y_1^{x^{(i-1,j)}} \quad (1 \le i \le s-1,\ 0 \le j \le m-1), \\ & y_2x^{(i,j)} = x^{(i,j-1)}y_2^{x^{(i,j-1)}} \quad (0 \le i \le s-1,\ 1 \le j \le m-1), \\ & y_1x^{(t,j)} = x^{(s-1,j)}y_1^{x^{(s-1,j)}} \quad (0 \le j \le m-1), \\ & y_2x^{(i,n)} = x^{(i,m-1)}y_2^{x^{(i,m-1)}} \quad (0 \le i \le s-1)\end{aligned}$$

*is a presentation for $A^{\oplus B}{}_\delta \times_\psi B$.*

## 4. Conclusions and future problems

In this paper, we first introduced a new monoid $A^{\oplus B}{}_\delta{\times}_\psi B^{\oplus A}$ under the name of *a higher version of Zappa products* or *generalized Zappa products* of the monoid $A^{\oplus B}$ by the monoid $B^{\oplus A}$ which is obtained by a combination of Zappa and wreath products. Then we defined a presentation on this new Theorem 2.2. After that, by taking $A$ and $B$ as finite (or infinite) monoid examples and also taking them as groups with their monoid presentations, we presented some consequences of Theorem 2.2.

It is clear that to define a presentation on an algebraic structure is an important tool in geometric group theory since this implies new studying areas over this structure. So, by considering the presentation defined in Theorem 2.2 or the presentations defined in corollaries of Theorem 2.2, one may study Gröbner-Shirshov bases (see, for instance, [12, 21]) over these presentations since the normal forms obtained by Gröbner-Shirshov bases implies the solvability of word problems ([11]). Furthermore the existence of other decision problems, specially the isomorphism problem, over the monoid $A^{\oplus B}{}_\delta{\times}_\psi B^{\oplus A}$ can be studied for a future project. Additionally, with the help of Theorem 2.2, the subjects Green‘s relations, periodicity and local finiteness may also be studied on $A^{\oplus B}{}_\delta{\times}_\psi B^{\oplus A}$.

Another future research on $A^{\oplus B}{}_\delta{\times}_\psi B^{\oplus A}$ would be the adaptation of the results presented in [7] and [8], that is, to investigate whether there exists a bijective correspondence between formations of the monoid $A^{\oplus B}{}_\delta{\times}_\psi B^{\oplus A}$ with formations of languages.

**Acknowledgment.** The authors would like to thank referees for their valuable suggestions and comments. This work was funded by the Deanship of Scientific Research (DSR), King Abdulaziz University, Jeddah, under grant No (G. 1711-130-1440). The authors, therefore, acknowledge with thanks DSR technical and financial support.

## References

[1] F. Ates and A.S. Cevik, *Knit products of finite cyclic groups and their applications*, Rend. Sem. Mat. Univ. Padova, **121**, 1–12, 2009.

[2] H. Ayik, C.M. Campbell, J.J. O'Connor and N. Ruskuc, *On the efficiency of wreath products of groups*, Groups-Korea 98, in: Proceedings of the International Conference held at Pusan National University, Pusan, Korea, August 10-16, 1998, Walter de Gruyter, 39–51, 2000.

[3] H. Ayik, C.M. Campbell, J.J. O'Connor and N. Ruskuc, *Minimal presentations and efficiency of semigroups*, Semigroup Forum, **60**, 231–242, 2000.

[4] H. Ayik, F. Kuyucu and B. Vatansever, *On semigroup presentations and efficiency*, Semigroup Forum, **65**, 329–335, 2002.

[5] A. Ballester-Bolinches, E. Cosme-Llopez and R. Esteban-Romero, *Group extensions and graphs*, Expo. Math. **34** (3), 327–334, 2016.

[6] A. Ballester-Bolinches, R. Esteban-Romero and M. Asaad, *Products of Finite Groups*, de Gruyter Exp. Math. **53**, Walter de Gruyter, 2010.

[7] A. Ballester-Bolinches, J.E. Pin and X. Soler-Escriva, *Formations of finite monoids and formal languages: Eilenberg's variety theorem revisited*, Forum Math. **26** (6), 1737–1761, 2014.

[8] A. Ballester-Bolinches, E. Cosme-Llopez, R. Esteban-Romero and J.J.M.M. Rutten, *Formations of monoids, congruences, and formal languages*, Sci. Ann. Comput. Sci. **25** (2), 171–209, 2015.

[9] A. Ballester-Bolinches, L.M. Ezquerro, A.A. Heliel and M.M. Al-Shomrani, *Some results on products of finite groups*, Bull. Malays. Math. Sci. Soc. **40** (3), 1341–1351, 2017.

[10] G. Baumslag, *Wreath products and finitely presented groups*, Math. Z. **75**, 22–28, 1961.

[11] L.A. Bokut, *Unsolvability of the word problem, and subalgebras of finitely presented Lie algebras*, Izv. Akad. Nauk. SSSR Ser. Math. **36**, 1173–1219, 1972.

[12] L.A. Bokut, Y. Chen and X. Zhao, *Gröbner-Shirshov bases for free inverse semigroups*, Internat. J. Algebra Comput. **19** (2), 129–143, 2009.

[13] M.G. Brin, *On the Zappa-Szép product*, Comm. Algebra **33**, 393–424, 2005.

[14] A.S. Cevik, *The efficiency of standard wreath product*, Proc. Edinburgh Math. Soc. **43** (2), 415–423, 2000.

[15] A.S. Cevik, *Minimal but inefficient presentations of the semi-direct product of some monoids*, Semigroup Forum, **66** (1), 1–17, 2003.

[16] N.D. Gilbert and S. Wazzan, *Zappa-Szép products of bands and groups*, Semigroup Forum, **77**, 438–455, 2008.

[17] A.A. Heliel, A. Ballester-Bolinches, R. Esteban-Romero and M.O. Almestady, $\mathfrak{Z}$-*permutable subgroups of finite groups*, Monat. Math. **179** (4), 523–534, 2016.

[18] J.M. Howie and N. Ruskuc, *Constructions and presentations for monoids*, Comm. Algebra, **22** (15), 6209–6224, 1994.

[19] J.M. Howie, *Fundamentals of Semigroup Theory*, London Math. Soc. Monographs, Oxford University Press, 1995.

[20] D.L. Johnson, *Presentation of Groups*, London Math. Soc. Lecture Note Series **15**, Cambridge University Press, 1990.

[21] C. Kocapinar, E.G. Karpuz, F. Ates and A.S. Cevik, *Gröbner-Shirshov bases of the generalized Bruck-Reilly* $*$*-extension*, Algebra Colloq. *19*, 813–820, 2012.

[22] M. Kunze, *Zappa products*, Acta Math. Hung. **41**, 225–239, 1983.

[23] T.G. Lavers, *Presentations of general products of monoids*, J. Algebra **204**, 733–741, 1998.

[24] S. MacLane, *Homology*, Classics in Mathematics, Springer Verlag, 1975.

[25] J.D.P. Meldrum, *Wreath Products of Groups and Semigroups*, Monographs and Surveys in Pure and Applied Mathematics (Book 74), Chapman and Hall/CRC; First Edition, 1995.

[26] P.W. Michor, *Knit products of graded Lie algebras and groups*, Rend. Circ. Mat. Palermo (2), **22**, 171–175, 1989.

[27] J. Szép, *On the structure of groups which can be represented as the product of two subgroups*, Acta Sci. Math. Szeged, **12**, 57–61, 1950.

[28] G. Zappa, *Sulla costruzione dei gruppi prodotto di due sottogruppi permutabili tra loro*, in: Atti Secondo Congresso Un. Ital., Bologna 1940. Edizioni Rome: Cremonense, 119–125, 1942.

Hacettepe Journal of Mathematics & Statistics

Hacet. J. Math. Stat.
Volume 50 (1) (2021), 235 – 242
DOI : 10.15672/hujms.554500

Research Article

# Some characterizations of rectifying curves in the 3-dimensional hyperbolic space $\mathbb{H}^3(-r)$

Buddhadev Pal, Akhilesh Yadav*

*Department of Mathematics, Institute of Science, Banaras Hindu University, Varanasi-221005, India*

## Abstract

In this paper, we study the geometry of rectifying curves in the 3-dimensional hyperbolic space $\mathbb{H}^3(-r)$. Further we obtain the distance function in terms of arc length when the rectifying curve lying in the upper half plane. Then we find the distance function and also give the general equations of the curvature and torsion of rectifying general helices in $\mathbb{H}^3(-r)$.

**Mathematics Subject Classification (2020).** 53C50, 53C40, 53A04, 53A05

**Keywords.** rectifying curve, general helix, geodesic, hyperbolic space $\mathbb{H}^3(-r)$

## 1. Introduction

In [4], B.Y. Chen gave the idea that the ratio of torsion and curvature of a regular curve is a linear function of arc length $s$, i.e., $(\tau/\kappa)(s) = c_1 s + c_2$ for some constants $c_1$ and $c_2$. If $c_1 = 0$, we obtain generalized helices; otherwise, we obtain rectifying curves. A space curve whose position vector always lies in its rectifying plane is called rectifying curve. So, a curve $\gamma$ is said to be rectifying curve if there exist a point $r$ in $\mathbb{R}^3$ such that $\gamma(s) - r = C_1 B(s) + C_2 T(s)$, where $C_1, C_2$ are some function of arc length $s$. Now the Frenet frame: $T = \gamma', N, B = T \times N$ of a unit speed curve $\gamma$ in $\mathbb{R}^3$ satisfies the Serret-Frenet equations:

$$\begin{pmatrix} T' \\ N' \\ B' \end{pmatrix} = \begin{pmatrix} 0 & \kappa & 0 \\ -\kappa & 0 & \tau \\ 0 & -\tau & 0 \end{pmatrix} \begin{pmatrix} T \\ N \\ B \end{pmatrix},$$

where the function $\kappa(s) > 0$ and $\tau(s)$ are called the curvature and the torsion of the curve and the above matrix is skew-symmetric. Therefore at each point of the curve we always get three planes namely: {T,N}-osculating plane, {N,B}-normal plane, {B,T}-rectifying plane and the equations of the corresponding planes are $(R-r).B = 0, (R-r).T = 0, (R-r).N = 0$, where $R$- position vector of any point on the respective plane, $r$-position vector of a specified point of the given curve. To know more about the characterization of rectifying curve we refer the reader to see [1, 2, 6]. In [7], P. Lucas and J.A.O. Yagues, studied rectifying curves in the three-dimensional hyperbolic space, and obtain some results of characterization and classification for such kind of curves.

*Corresponding Author.
Email addresses: pal.buddha@gmail.com (B. Pal), akhilesh_mathau@rediffmail.com (A. Yadav)
Received: 16.04.2019; Accepted: 01.06.2020

In [5], S. Izumiya and N. Takeuchi introduced the notion of slant helix, if the principle normal lines of $\gamma$ makes a constant angle with a fixed direction, also they found a necessary and sufficient condition for a curve $\gamma$ with $\kappa(s) > 0$ to be a slant helix is that function $\sigma = \frac{\kappa^2}{(\kappa^2+\tau^2)^{3/2}}(\frac{\tau}{\kappa})'$ be constant. Further in [8], P. Lucas and J.A.O Yagues studied slant helices in the three dimensional sphere. Also in [3], M. Barros gave the definition of Lancret curve (general helix), the principle normal lines are perpendicular to a fixed direction. Thus a general helix is the special case of a slant helix. It is clear that if $\sigma \equiv 0$ then $\gamma$ is a general helix. Also M. Barros gave a theorem that, a curve $\gamma$ in $\mathbb{H}^3$ is a slant helix if and only if either $\gamma$ is a curve in some unit hyperbolic plane $\mathbb{H}^2 \subset \mathbb{H}^3$ with $\tau \equiv 0$ or $\gamma$ is a helix in $\mathbb{H}^3$.

Thus motivated sufficiently we study general helices in the 3-dimensional hyperbolic space $\mathbb{H}^3(-r)$ and obtain several results corresponding to the rectifying general helix and characterization of rectifying curve in $\mathbb{H}^3(-r)$. Our work is organized as follows: using the Gauss formula and the definition of rectifying curve in $\mathbb{H}^3(-r)$, we find expressions of ${T^0}'_\gamma$, ${N^0}'_\gamma$, ${B^0}'_\gamma$, ${T^0}'_{\phi_s}.{T^0}'_{\bar{\gamma}}$, ${N^0}'_{\phi_s}.{N^0}'_{\bar{\gamma}}$, ${B^0}'_{\phi_s}.{B^0}'_{\bar{\gamma}}$ etc. Here we take dot product because it gives the geometrical interpretation of curve. Further we obtain the distance function in $\mathbb{H}^3(-r)$ in terms of $\lambda$ and $\mu$, which satisfy some differential equation. We also find distance function in terms of arc length when the rectifying curve lying in the upper half plane. Next we find some characterizations of rectifying curve in $\mathbb{H}^3(-r)$. Finally we give the general equations of the curvature and torsion of a rectifying general helix.

## 2. Preliminaries

Let $\mathbb{H}^3(p,-r) = \{x = (x_1,x_2,x_3,x_4) \in \mathbb{R}^4_1 | < x-p, x-p >= -r^2, x_1 > 0\} \subset \mathbb{R}^4_1$ be the hyperbolic space with centered at $p \in \mathbb{R}^4_1$ and radius $r > 0$, where $\mathbb{R}^4_1$ is the four dimensional Lorentzian manifold with flat metric $g = -dx_1^2 + dx_2^2 + dx_3^2 + dx_4^2$. Also we denote $\mathbb{H}^3(0,-r) \equiv \mathbb{H}^3(-r) = \{x \in \mathbb{R}^4_1 | -x_1^2 + x_2^2 + x_3^2 + x_4^2 = -r^2, x_1 > 0\} \subset \mathbb{R}^4_1$ and $\mathbb{H}^3(0,-1) \equiv \mathbb{H}^3$.

We know that if $\bar{\nabla}$ and $\nabla^\circ$ denote the Levi-Civita connections on $\mathbb{H}^3(-r)$ and $\mathbb{R}^4_1$ respectively then they are related by the Gauss formula, $\nabla^\circ_X Y = \bar{\nabla}_X Y + \frac{1}{r^2} < X, Y > \phi$, where $\phi : \mathbb{H}^3(-r) \to \mathbb{R}^4_1$ denotes the position vector and $X$, $Y$ are vector fields tangent to $\mathbb{H}^3(-r)$. Let us consider a unit speed curve $\gamma : I \subset \mathbb{R} \to \mathbb{H}^3(-r)$ and assume that $\gamma$ is not a geodesic curve then we always get $\nabla^\circ_{T_\gamma} T_\gamma = \kappa_\gamma N_\gamma + \frac{1}{r^2}\gamma$, $\nabla^\circ_{T_\gamma} N_\gamma = -\kappa_\gamma T_\gamma + \tau_\gamma B_\gamma$, $\nabla^\circ_{T_\gamma} B_\gamma = -\tau_\gamma N_\gamma$, where two functions $\kappa_\gamma > 0$ and $\tau_\gamma$ are curvature and torsion of the curve $\gamma$. It is also well-known that the *principle normal geodesic* in $\mathbb{H}^3(-r)$ starting at $\gamma(s)$ of the curve $\gamma$ can be defined as the geodesic curve parameterized by $\phi_s(t) = \exp_{\gamma(s)}(tN_\gamma(s)) = \cosh(\frac{t}{r})\gamma(s) + r\sinh(\frac{t}{r})N_\gamma(s), t \in \mathbb{R}$.

In [7], authors gave two equivalent definitions of rectifying curve in the three dimensional hyperbolic space.

**Definition 2.1.** A unit speed curve $\gamma = \gamma(s)(s \in I)$ in $\mathbb{H}^3(-r)$, with $\kappa_\gamma > 0$, is said to be rectifying curve if there exists a point $p \in \mathbb{H}^3(-r)$ such that $p$ is not belongs to $Im(\gamma) \equiv \gamma(I)$ and the geodesics connecting $p$ with $\gamma(s)$ are orthogonal to the principle normal geodesics at $\gamma(s)$, for all $s$.

**Definition 2.2.** The geodesics connecting $p$ with $\gamma(s)$ are tangent to the rectifying plane of $\gamma$ i.e., the planes generated by $\{T_\gamma(s), B_\gamma(s)\}$.

Also in [7], two characterization theorems for rectifying curves are given.

**Theorem 2.3.** *Let $\gamma = \gamma(s)(s \in I)$ be a unit speed curve in $\mathbb{H}^3(-r)$. Then, $\gamma$ is a rectifying curve if and only if the ratio of torsion and curvature of the curve is given by $\frac{\tau_\gamma}{\kappa_\gamma}(s) = c_1 \sinh(\frac{s+s_0}{r}) + c_2 \cosh(\frac{s+s_0}{r})$, for some constants $c_1$, $c_2$ and $s_0$, with $1 - {c_1}^2 + {c_2}^2 <$ 0.*

**Theorem 2.4.** *Let $p \in H^3(-r)$ and consider a unit speed curve $V(t)$ in $S^2(1) \subset T_pH^3(-r)$. Then, for any nonzero function $\rho(t)$, the curvature $\kappa_\gamma$ and the speed $v$ of the curve $\gamma(t) = \exp_p(\rho(t)V(t))$, and the geodesic curvature $\kappa_V$ of $V$ satisfy the inequality $\kappa_V^2 \leq \dfrac{v^4\kappa_\gamma^2}{r^2\sinh^2(\rho/r)}$, with the equality sign holding identically if and only if $\gamma$ is a rectifying curve.*

## 3. Main results

**Theorem 3.1.** *Let $\gamma : I \subset \mathbb{R} \to \mathbb{H}^3(-r)$ be a unit speed rectifying curve in $\mathbb{H}^3(-r)$. If $\{T_\gamma, N_\gamma, B_\gamma\}$ is the Frenet frame along $\gamma$ and $\bar{\nabla}$ and $\nabla^\circ$ denote the Levi-Civita connections on $\mathbb{H}^3(-r)$ and $\mathbb{R}^4_1$ respectively then by using the Gauss formula the Frenet equations of $\gamma$ can be written as follows:*

$$T^{\circ'}{}_\gamma = \kappa_\gamma N_\gamma + 1/r^2\gamma, N^{\circ'}{}_\gamma = -\kappa_\gamma T_\gamma + \kappa_\gamma\psi B_\gamma, B^{\circ'}{}_\gamma = -\kappa_\gamma\psi N_\gamma,$$

*where $\kappa_\gamma, \tau_\gamma$ denote the curvature and torsion of $\gamma$, which satisfy any of the following conditions:*

(1) $T^{\circ'}{}_{\phi_s}.T^{\circ'}{}_{\bar{\gamma}} = \frac{1}{r^2}(\kappa_{\phi_s}N_{\phi_s}.\bar{\gamma} + \kappa_{\bar{\gamma}}\phi_s.N_{\bar{\gamma}} + \frac{1}{r^2}\phi_s.\bar{\gamma})$,
$N^{\circ'}{}_{\phi_s}.N^{\circ'}{}_{\bar{\gamma}} = \lambda_1\tau_{\phi_s}\tau_{\bar{\gamma}}$,
$B^{\circ'}{}_{\phi_s}.B^{\circ'}{}_{\bar{\gamma}} = 0.$

(2) $T^{\circ'}{}_{\phi_s}.T^{\circ'}{}_{\bar{\gamma}} = \lambda_4\kappa_{\phi_s}\kappa_{\bar{\gamma}} + \frac{1}{r^2}(\lambda_4\kappa_{\phi_s}\bar{\gamma} + \phi_s\kappa_{\bar{\gamma}}).N_{\bar{\gamma}} + \frac{1}{r^4}\phi_s.\bar{\gamma}$, $N^{\circ'}{}_{\phi_s}.N^{\circ'}{}_{\bar{\gamma}} = -\lambda_2\tau_{\phi_s}\kappa_{\bar{\gamma}} - \lambda_3\kappa_{\phi_s}\tau_{\bar{\gamma}}$, $B^{\circ'}{}_{\phi_s}.B^{\circ'}{}_{\bar{\gamma}} = -\lambda_4\tau_{\phi_s}\tau_{\bar{\gamma}}$.

(3) $T^{\circ'}{}_{\phi_s}.T^{\circ'}{}_{\bar{\gamma}} = \frac{1}{r^2}(\kappa_{\phi_s}N_{\phi_s}.\bar{\gamma} + \kappa_{\bar{\gamma}}\phi_s.N_{\bar{\gamma}} + \frac{1}{r^2}\phi_s.\bar{\gamma})$, $N^{\circ'}{}_{\phi_s}.N^{\circ'}{}_{\bar{\gamma}} = -d_1\tau_{\phi_s}\kappa_{\bar{\gamma}}$,
$B^{\circ'}{}_{\phi_s(t)}.B^{\circ'}{}_{\bar{\gamma}} = 0.$

(4) $T^{\circ'}{}_{\phi_s}.T^{\circ'}{}_{\bar{\gamma}} = \frac{1}{r^2}(\kappa_{\phi_s}N_{\phi_s}.\bar{\gamma} + \kappa_{\bar{\gamma}}\phi_s.N_{\bar{\gamma}} + \frac{1}{r^2}\phi_s.\bar{\gamma})$, $N^{\circ'}{}_{\phi_s}.N^{\circ'}{}_{\bar{\gamma}} = -d_2\kappa_{\phi_s}\tau_{\bar{\gamma}}$,
$B^{\circ'}{}_{\phi_s(t)}.B^{\circ'}{}_{\bar{\gamma}} = 0,$

*where $\lambda_1, \lambda_2, \lambda_3, \lambda_4, d_1, d_2 \in \mathbb{R}$.*

***Proof.*** Let $\gamma : I \subset \mathbb{R} \to \mathbb{H}^3(-r)$ be a unit speed rectifying curve in $\mathbb{H}^3(-r)$. If $\{T_\gamma, N_\gamma, B_\gamma\}$ be the Frenet frame along $\gamma$ and $\bar{\nabla}$ and $\nabla^\circ$ denote the Levi-Civita connections on $\mathbb{H}^3(-r)$ and $\mathbb{R}^4_1$ respectively then the Frenet equations of $\gamma$ are

$$\bar{\nabla}_{T_\gamma}T_\gamma = \kappa_\gamma N_\gamma, \bar{\nabla}_{T_\gamma}N_\gamma = -\kappa_\gamma T_\gamma + \tau_\gamma B_\gamma, \bar{\nabla}_{T_\gamma}B_\gamma = -\tau_\gamma N_\gamma, \tag{3.1}$$

where functions $\kappa_\gamma > 0$ and $\tau_\gamma$ are curvature and torsion of the curve $\gamma$. After using the Gauss formula in (3.1), we get

$$\nabla^\circ_{T_\gamma}T_\gamma = \kappa_\gamma N_\gamma + \frac{1}{r^2}\gamma, \nabla^\circ_{T_\gamma}N_\gamma = -\kappa_\gamma T_\gamma + \tau_\gamma B_\gamma, \nabla^\circ_{T_\gamma}B_\gamma = -\tau_\gamma N_\gamma. \tag{3.2}$$

Then from ([7], Theorem 3.), using the relation of $\tau_\gamma$ and $\kappa_\gamma$ for rectifying curve we obtain,

$$\nabla^\circ{}_{T_\gamma}T_\gamma = \kappa_\gamma N_\gamma + \frac{1}{r^2}\gamma, \nabla^\circ{}_{T_\gamma}N_\gamma = -\kappa_\gamma T_\gamma + \kappa_\gamma\psi B_\gamma, \nabla^\circ{}_{T_\gamma}B_\gamma = -\kappa_\gamma\psi N_\gamma, \tag{3.3}$$

where $\psi(s) = c_1f(s) + c_2g(s)$. Now, we write the equation (3.3) in the following notation

$$T^{\circ'}{}_\gamma = \kappa_\gamma N_\gamma + \frac{1}{r^2}\gamma, N^{\circ'}{}_\gamma = -\kappa_\gamma T_\gamma + \kappa_\gamma\psi B_\gamma, B^{\circ'}{}_\gamma = -\kappa_\gamma\psi N_\gamma. \tag{3.4}$$

Now, using Definition 2.1, let $\phi_s(t)$ be geodesics connecting $p$ with $\gamma(s)$ are orthogonal to the principle normal geodesics $\bar{\gamma}$ at $\gamma(s)$, for all $s$. Then we get,

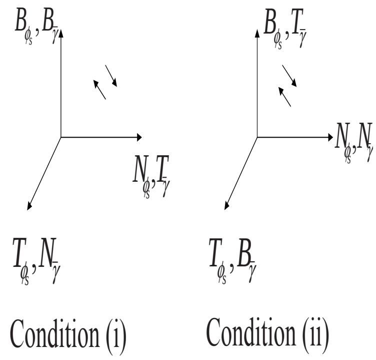
$$\begin{aligned} T^{\circ'}{}_{\phi_s(t)} &= \kappa_{\phi_s(t)}N_{\phi_s(t)} + \frac{1}{r^2}\phi_s(t), \\ N^{\circ'}{}_{\phi_s(t)} &= -\kappa_{\phi_s(t)}T_{\phi_s(t)} + \tau_{\phi_s(t)}B_{\phi_s(t)}, \\ B^{\circ'}{}_{\phi_s(t)} &= -\tau_{\phi_s(t)}N_{\phi_s(t)}, \end{aligned} \tag{3.5}$$

and

$$
\begin{aligned}
T^{\circ'}{}_{\bar{\gamma}} &= \kappa_{\bar{\gamma}} N_{\bar{\gamma}} + \frac{1}{r^2}\bar{\gamma}, \\
N^{\circ'}{}_{\bar{\gamma}} &= -\kappa_{\bar{\gamma}} T_{\bar{\gamma}} + \tau_{\bar{\gamma}} B_{\bar{\gamma}}, \\
B^{\circ'}{}_{\bar{\gamma}} &= -\tau_{\bar{\gamma}} N_{\bar{\gamma}}.
\end{aligned}
\tag{3.6}
$$

Now for the case of rectifying curve, $\phi_s(t)$ and $\bar{\gamma}(s)$ are orthogonal at $\gamma(s)$ for all $s$ i.e., $T_{\phi_s(t)}.T_{\bar{\gamma}} = 0$ and we get two cases corresponding to the Frenet frame of the curves $\phi_s$ and $\bar{\gamma}$.

**Case 1.**




Then using Condition (i) in the equations (3.5) and (3.6), we get

$$
T^{\circ'}{}_{\phi_s}.T^{\circ'}{}_{\bar{\gamma}} = \frac{1}{r^2}(\kappa_{\phi_s} N_{\phi_s}.\bar{\gamma} + \kappa_{\bar{\gamma}} \phi_s.N_{\bar{\gamma}} + \frac{1}{r^2}\phi_s.\bar{\gamma}),
$$

$$
N^{\circ'}{}_{\phi_s}.N^{\circ'}{}_{\bar{\gamma}} = \lambda_1 \tau_{\phi_s} \tau_{\bar{\gamma}} B_{\bar{\gamma}}.B_{\bar{\gamma}} = \lambda_1 \tau_{\phi_s} \tau_{\bar{\gamma}}, B^{\circ'}{}_{\phi_s(t)}.B^{\circ'}{}_{\bar{\gamma}} = 0,
$$

where $B_{\phi_s} = \lambda_1 B_{\bar{\gamma}}$. By using Condition (ii) in the equations (3.5) and (3.6), we obtain

$$
\begin{aligned}
T^{\circ'}{}_{\phi_s}.T^{\circ'}{}_{\bar{\gamma}} &= \kappa_{\phi_s}\kappa_{\bar{\gamma}} N_{\phi_s}.N_{\bar{\gamma}} + \frac{1}{r^2}(\kappa_{\phi_s} N_{\phi_s}.\bar{\gamma} + \kappa_{\bar{\gamma}}\phi_s.N_{\bar{\gamma}} + \frac{1}{r^2}\phi_s.\bar{\gamma}) \\
&= \lambda_4 \kappa_{\phi_s}\kappa_{\bar{\gamma}} + \frac{1}{r^2}(\lambda_4 \kappa_{\phi_s}\bar{\gamma} + \kappa_{\bar{\gamma}}\phi_s).N_{\bar{\gamma}} + \frac{1}{r^4}\phi_s.\bar{\gamma},
\end{aligned}
$$

$$
N^{\circ'}{}_{\phi_s}.N^{\circ'}{}_{\bar{\gamma}} = -\lambda_2 \tau_{\phi_s}\kappa_{\bar{\gamma}} T_{\bar{\gamma}}.T_{\bar{\gamma}} - \lambda_3 \kappa_{\phi_s}\tau_{\bar{\gamma}} B_{\bar{\gamma}}.B_{\bar{\gamma}} = -\lambda_2 \tau_{\phi_s}\kappa_{\bar{\gamma}} - \lambda_3 \kappa_{\phi_s}\tau_{\bar{\gamma}},
$$

$$
B^{\circ'}{}_{\phi_s}.B^{\circ'}{}_{\bar{\gamma}} = \lambda_4 \tau_{\phi_s}\tau_{\bar{\gamma}},
$$

where $B_{\phi_s} = \lambda_2 T_{\bar{\gamma}}$, $T_{\phi_s} = \lambda_3 B_{\bar{\gamma}}$ and $N_{\phi_s} = \lambda_4 N_{\bar{\gamma}}$. Now we know that $T_\gamma$ can be written as $T_\gamma = c_1 T_{\bar{\gamma}} + c_2 N_{\bar{\gamma}} + c_3 B_{\bar{\gamma}}$, and $T_\gamma = c'_1 T_{\phi_s} + c'_2 N_{\phi_s} + c'_3 B_{\phi_s}$. Also we know that $T_\gamma.T_\gamma = 1$, therefore after using Condition (ii), we get

$$
\begin{gathered}
c_1 c'_3 T_{\bar{\gamma}}.B_{\phi_s} + c_2 c'_2 N_{\bar{\gamma}}.N_{\phi_s} + c_3 c'_1 B_{\bar{\gamma}}.T_{\phi_s} = 1, \\
\Rightarrow c_1 c'_3 \lambda_2 + c_2 c'_2 \lambda_4 + c_3 c'_1 \lambda_3 = 1. \\
\Rightarrow c_1 c'_3 \lambda_2 + c_3 c'_1 \lambda_3 = 1 - c_2 c'_2 d_3,
\end{gathered}
$$

where we consider $\lambda_4 = d_3 \in \mathbb{R}$. Thus we get

$$
c\lambda_2 + d\lambda_3 = n, \tag{3.7}
$$

where $c = c_1 c'_3, d = c_3 c'_1, n = 1 - c_2 c'_2 d_3$.

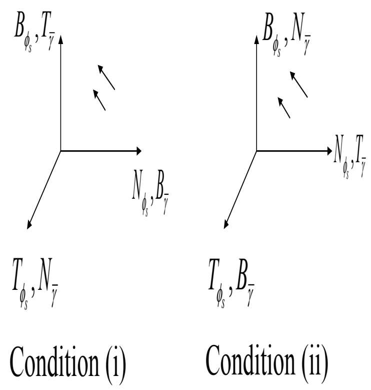
On the other hand we can write $N_{\bar{\gamma}} = b_1 T_{\phi_s} + b_2 N_{\phi_s} + b_3 B_{\phi_s}$ and $N_{\phi_s} = b_1^{'} T_{\bar{\gamma}} + b_2^{'} N_{\bar{\gamma}} + b_3^{'} B_{\bar{\gamma}}$. Now, taking the dot product of $N_{\bar{\gamma}}$ and $N_{\phi_s}$, and then using Condition (ii), we get $b_3 b_1^{'} \lambda_2 + b_1 b_3^{'} \lambda_3 = (1 - b_2 b_2^{'}) d_3 = m$, which implies

$$a\lambda_2 + b\lambda_3 = m, \tag{3.8}$$

where $a = b_3 b_1^{'}$, $b = b_1 b_3^{'}$, $m = (1 - b_2 b_2^{'}) d_3$ and $c_1, c_2, c_3, c_1^{'}, c_2^{'}, c_3^{'}, b_1, b_2, b_3, b_1^{'}, b_2^{'}, b_3^{'}$, $a, b, c, d, m, n, \lambda_1, \lambda_2, \lambda_3, \lambda_4 \in \mathbb{R}$.

On solving the equations (3.7) and (3.8), we get $\lambda_2 = \frac{dm-bn}{ad-cb}, \lambda_3 = \frac{cm-an}{cb-ad}$. Similarly, using Condition (i), $\lambda_1$ can also be calculated.

**Case 2.**




Then using Condition (i) in the equations (3.5) and (3.6), we get

$$T^{\circ'}{}_{\phi_s} . T^{\circ'}{}_{\bar{\gamma}} = \frac{1}{r^2}(\kappa_{\phi_s} N_{\phi_s} . \bar{\gamma} + \kappa_{\bar{\gamma}} \phi_s . N_{\bar{\gamma}} + \frac{1}{r^2} \phi_s . \bar{\gamma}),$$

$$N^{\circ'}{}_{\phi_s} . N^{\circ'}{}_{\bar{\gamma}} = -\tau_{\phi_s} \kappa_{\bar{\gamma}} T_{\bar{\gamma}} . B_{\phi_s} = -d_1 \tau_{\phi_s} \kappa_{\bar{\gamma}}, B^{\circ'}{}_{\phi_s(t)} . B^{\circ'}{}_{\bar{\gamma}} = 0,$$

where $B_{\phi_s} = d_1 T_{\bar{\gamma}}$. By using Condition (ii) in the equations (3.5) and (3.6), we get

$$T^{\circ'}{}_{\phi_s} . T^{\circ'}{}_{\bar{\gamma}} = \frac{1}{r^2}(\kappa_{\phi_s} N_{\phi_s} . \bar{\gamma} + \kappa_{\bar{\gamma}} \phi_s . N_{\bar{\gamma}} + \frac{1}{r^2} \phi_s . \bar{\gamma}),$$

$$N^{\circ'}{}_{\phi_s} . N^{\circ'}{}_{\bar{\gamma}} = -\kappa_{\phi_s} \tau_{\bar{\gamma}} T_{\phi_s} . B_{\bar{\gamma}} = -d_2 \kappa_{\phi_s} \tau_{\bar{\gamma}}, B^{\circ'}{}_{\phi_s(t)} . B^{\circ'}{}_{\bar{\gamma}} = 0,$$

where $T_{\phi_s} = d_2 B_{\bar{\gamma}}$. Then from above procedure we can find the values of $d_1, d_2 \in \mathbb{R}$. Thus, we obtain the required results. ☐

**Theorem 3.2.** *Let $\gamma = \gamma(s)$ be a unit speed rectifying curve in $\mathbb{H}^3(-r)$. Then the distance function $\rho = \|\gamma\|$ satisfies $\rho^2 = -\lambda^2 + \mu^2$, where $\lambda$ and $\mu$ satisfy the equation $(1-\lambda^{'})aT_{\bar{\gamma}} - (b - b\lambda^{'} + \mu^{'})B_\gamma + \frac{\lambda\gamma}{r^2} = \lambda T_\gamma^{\circ'} + \mu B_\gamma^{\circ'}$ and $a, b \in \mathbb{R}$.*

***Proof.*** Let $\gamma = \gamma(s)$ be a unit speed rectifying curve in $\mathbb{H}^3(-r)$. Then position vector $\gamma$ of a curve satisfies the equation

$$\gamma(s) = \lambda(s) T_\gamma(s) + \mu(s) B_\gamma(s), \tag{3.9}$$

where $\lambda(s)$ and $\mu(s)$ are differential functions. Now, differentiating the equation (3.9) with respect to $s$ and using Frenet equations, we get $T_\gamma(s) = \lambda^{'}(s) T_\gamma(s) + \lambda(s)(T_\gamma^{\circ'} - \frac{1}{r^2}\gamma) + \mu^{'}(s) B_\gamma(s) + \mu B_\gamma^{\circ'}$, which implies

$$(1 - \lambda^{'}) T_\gamma - \mu^{'} B_\gamma - \lambda T_\gamma^{\circ'} - \mu B_\gamma^{\circ'} + \frac{\lambda\gamma}{r^2} = 0. \tag{3.10}$$

Then using Definition 2.1 of rectifying curve in $\mathbb{H}^3(-r)$, $T_\gamma$ can be written in the form, $T_\gamma = aT_{\bar{\gamma}} - bB_\gamma$, where $\bar{\gamma}$ is the geodesics connecting $p$ with $\gamma(s)$ are tangent to the rectifying plane of $\gamma$ i.e., the planes generated by $\{T_\gamma(s), B_\gamma(s)\}$. Therefore the equation (3.10) can be rewritten as

$$(1-\lambda^{'})aT_{\bar{\gamma}} - (b - b\lambda^{'} + \mu^{'})B_\gamma + \frac{\lambda\gamma}{r^2} = \lambda T_\gamma^{\circ'} + \mu B_\gamma^{\circ'}. \tag{3.11}$$

Also from the equation (3.9), it is clear that the distance function $\rho^2 = \|\gamma\|^2 = |g(\gamma,\gamma)| = -\lambda^2 + \mu^2$, where $\lambda$ and $\mu$ satisfy the equation (3.11). Thus the proof is completed. □

**Theorem 3.3.** *Let $\gamma = \gamma(s)$ be a unit speed rectifying curve in $\mathbb{H}^3(-r)$, lies in the upper half plane $U^2$. Then the distance function $\rho = \|\gamma\|$ satisfies $\rho^2 = |as^2 + bs + c|$ or $\rho^2 = 1 + f^2(s)$, where $f(s) = c_1 \sinh(\frac{s+s_0}{r}) + c_2 \cosh(\frac{s+s_0}{r})$ and $a, b, c \in \mathbb{R}$.*

***Proof.*** Let $\gamma = \gamma(s)$ be a unit speed rectifying curve in $\mathbb{H}^3(-r)$. Now, we know that

$$\gamma(s) = \lambda(s)T_\gamma(s) + \mu(s)B_\gamma(s), \tag{3.12}$$

where $\lambda(s)$ and $\mu(s)$ are differentiable functions.

Now we know that $T_\gamma(s)$ and $B_\gamma(s)$ are generating a plane, let it be a subset of upper half plane. Therefore $\gamma(s) = (\lambda(s), \mu(s))$ be a curve in $U^2$. Then after differentiating the equation (3.12) and using Frenet formulas for $\gamma$, we obtain $(1-\lambda^{'})T_\gamma + (\mu\tau_\gamma - \lambda\kappa_\gamma)N_\gamma - \mu^{'}(s)B_\gamma = 0$, which implies

$$\lambda^{'} = 1, \mu^{'} = 0, \mu\tau_\gamma - \lambda\kappa_\gamma = 0. \tag{3.13}$$

Therefore $\lambda(s) = s + d_1$, $\mu(s) = d_2$, $\mu(s)\tau_\gamma(s) = \lambda(s)\kappa_\gamma(s)$. Thus the distance function $\rho^2 = |g(\gamma,\gamma)| = |\frac{\lambda^2+\mu^2}{\mu^2}| = |\frac{(s+d_1)^2+d_2^2}{d_2^2}| = |as^2 + bs + c|$, where $a = \frac{1}{d_2^2}, b = \frac{2d_1}{d_2^2}, c = \frac{d_1^2+d_2^2}{d_2^2}, d_1, d_2 \in \mathbb{R}$. Also from the equation (3.13), we get $\frac{\lambda(s)}{\mu(s)} = \frac{\tau_\gamma}{\kappa_\gamma}$. Now we know that $\frac{\tau_\gamma}{\kappa_\gamma} = c_1 \sinh(\frac{s+s_0}{r}) + c_2 \cosh(\frac{s+s_0}{r}) = f(s)$, from [7]. Hence $\frac{\lambda}{\mu} = f$. Therefore the distance function, $\rho^2 = |g(\gamma,\gamma)| = |\frac{\lambda^2+\mu^2}{\mu^2}| = |1 + f^2|$. Thus, $\rho^2 = 1 + f^2(s)$. This proves the theorem. □

**Note.** Now, we know that $\gamma(s) = \lambda(s)T_\gamma(s) + \mu(s)B_\gamma(s)$, where $\lambda(s)$ and $\mu(s)$ are differential functions.

(i) Therefore, $g(\gamma, T_\gamma) = \lambda(s) = s + d_1$. This is the tangential component of $\gamma(s)$.

(ii) The normal component of $\gamma(s) = \mu(s)B_\gamma(s)$. Therefore, $\|\gamma^N\| = d_2 \neq 0$ i.e.,the normal component component of $\gamma(s)$ has a constant length.

(iii) The binormal component of $\gamma(s)$, $g(\gamma(s), B_\gamma(s)) = \mu(s) = d_2$, is constant.

**Theorem 3.4.** *Let $\psi(t)$ be a unit speed curve in $\mathbb{R}^4_1$ and $\gamma$ be a rectifying curve in $\mathbb{H}^3(-r)$ with upper half plane as rectifying plane then it has up to a parametrization given by $\gamma(t) = \psi(t)\phi(t)$, or $\gamma(t) = \psi(t)h(t)$.*

***Proof.*** Now from Theorem 3.3, we know that $\rho^2 = as^2 + bs + c$ or $\rho^2 = 1 + f^2(s)$. Let $\rho^2 = |\frac{(s+d_1)^2+d_2^2}{d_2^2}|$, we apply a translation to $s$, such that $\rho^2 = as^2 + 1$. Now we define a curve $\psi(t)$ in $\mathbb{R}^4_1$ by $\psi(s) = \frac{\gamma(s)}{\rho(s)}$, $\Rightarrow \gamma(s) = \psi(s)\sqrt{as^2+1}$. Then differentiating with respect to $s$, we get

$$T_{\gamma(s)} = \psi(s)\frac{as}{\sqrt{as^2+1}} + \psi^{'}(s)\sqrt{as^2+1}. \tag{3.14}$$

Since, $g(\psi,\psi) = 1$, it follows that $g(\psi,\psi^{'}) = 0$. Therefore from the equation (3.14), we obtain $1 = g(T_\gamma, T_\gamma) = g(\psi^{'},\psi^{'})(as^2+1) + \frac{a^2s^2}{as^2+1}$, which implies

$$g(\psi^{'},\psi^{'}) = \frac{as^2(1-a)+1}{(as^2+1)^2}. \tag{3.15}$$

Thus, $\|\psi^{'}(s)\| = \frac{\sqrt{as^2(1-a)+1}}{as^2+1}$. Let $t = \int_0^s \|\psi^{'}(u)\|du = \int_0^s \frac{\sqrt{as^2(1-a)+1}}{as^2+1}du = \varphi(s)$. Therefore $t = \varphi(s)$ or $s = \varphi^{-1}(t)$. Put this into $\gamma(s) = \psi(s)\sqrt{as^2+1}$, we get $\gamma(t) = \psi(t)\eta(\varphi^{-1}(t)) = \psi(t)\phi(t)$, where $\eta(s) = \sqrt{as^2+1}, \phi = \eta \circ \varphi^{-1}$. Hence $\gamma(t) = \psi(t)\phi(t)$. Similarly if we take $\rho^2 = 1+f^2(s)$ then up to parametrization for $\gamma$ is in the form $\psi(t)h(t)$, which completes the proof. $\square$

**Theorem 3.5.** *Let $\gamma = \gamma(s)$ be a unit speed rectifying curve in $\mathbb{H}^3(-r)$. Then $T_{\bar{\gamma}}$ can be written in the form, $T_{\bar{\gamma}} = \alpha(s)N_\gamma + \beta(s)B\gamma$, where $\alpha(s) = \frac{\lambda\kappa_\gamma - \mu\tau_\gamma}{a-a\lambda}$, $\beta(s) = \frac{b-b\lambda+\mu^{'}}{a-a\lambda}$ and $a, b \in \mathbb{R}$.*

***Proof.*** Let us consider $\gamma = \gamma(s)$ be a unit speed rectifying curve in $\mathbb{H}^3(-r)$. Then position vector $\gamma$ of a curve satisfies the equation,

$$\gamma(s) = \lambda(s)T_\gamma(s) + \mu(s)B_\gamma(s), \tag{3.16}$$

where $\lambda(s)$ and $\mu(s)$ are differentiable functions. On differentiating the equation (3.16), we obtain $T_\gamma = \lambda^{'}T_\gamma + \mu^{'}(s)B_\gamma + \lambda\kappa_\gamma N_\gamma - \mu\tau_\gamma N_\gamma$, which implies

$$\Rightarrow (1-\lambda^{'})T_\gamma + (\mu\tau_\gamma - \lambda\kappa_\gamma)N_\gamma - \mu^{'}(s)B_\gamma = 0. \tag{3.17}$$

Since $\gamma = \gamma(s)$ is a unit speed rectifying curve in $\mathbb{H}^3(-r)$ therefore $T_\gamma = aT_{\bar{\gamma}} - bB_\gamma$, where $a, b \in \mathbb{R}$. Thus from the equation (3.17), we get $(a - a\lambda)T_{\bar{\gamma}} + (\mu\tau_\gamma - \lambda\kappa_\gamma)N_\gamma - (b - b\lambda + \mu^{'})B_\gamma = 0$, which gives

$$T_{\bar{\gamma}} = \alpha(s)N_\gamma + \beta(s)B_\gamma, \tag{3.18}$$

where $\alpha(s) = \frac{\lambda\kappa_\gamma - \mu\tau_\gamma}{a-a\lambda}$ and $\beta(s) = \frac{b-b\lambda+\mu^{'}}{a-a\lambda}$, $a, b \in \mathbb{R}$. This completes the proof. $\square$

**Theorem 3.6.** *Let $\gamma = \gamma(s)$ be a unit speed curve in $\mathbb{H}^3(-r)$. Then $\gamma$ is a rectifying general helix if and only if the torsion and curvature of the curve are given by*

*(i)*$\tau_\gamma^2(s) = \sinh^2(\frac{\rho}{r})\cosh^2(\frac{s+s_0}{r})[A\tanh^2(\frac{s+s_0}{r}) + C\tanh(\frac{s+s_0}{r}) + B]$,
*where* $A = \frac{c_1^2\kappa_V^2 r^2}{v^4}, B = \frac{c_2^2\kappa_V^2 r^2}{v^4}, C = \frac{2c_1c_2\kappa_V^2 r^2}{v^4}$,

*(ii)* $\kappa_\gamma^2(s) = \sinh^2(\frac{\rho}{r})$, *if* $A = c_1^2, B = c_2^2, C = 2c_1c_2$.

***Proof.*** By using Theorem 2.3 and Theorem 2.4, we obtain

$$\tau_\gamma^2(s) = \frac{\kappa_V^2 r^2\sinh^2(\rho/r)}{v^4}(c_1\sinh(\frac{s+s_0}{r}) + c_2\cosh(\frac{s+s_0}{r}))^2,$$

which implies

$$\tau_\gamma^2(s) = A\sinh^2(\rho/r)\sinh^2(\frac{s+s_0}{r}) + C\sinh^2(\rho/r)\sinh(\frac{s+s_0}{r})\cosh(\frac{s+s_0}{r})$$
$$+B\sinh^2(\rho/r)\cosh^2(\frac{s+s_0}{r}),$$

where $A = \frac{c_1^2\kappa_V^2 r^2}{v^4}, B = \frac{c_2^2\kappa_V^2 r^2}{v^4}, C = \frac{2c_1c_2\kappa_V^2 r^2}{v^4}$. Thus

$$\tau_\gamma^2(s) = \sinh^2(\rho/r)\cosh^2(\frac{s+s_0}{r})[A\frac{\sinh^2(\frac{s+s_0}{r})}{\cosh^2(\frac{s+s_0}{r})} + C\frac{\sinh(\frac{s+s_0}{r})\cosh(\frac{s+s_0}{r})}{\cosh^2(\frac{s+s_0}{r})} + B],$$

$$\Rightarrow \tau_\gamma^2(s) = \sinh^2(\rho/r)\cosh^2(\frac{s+s_0}{r})[A\tanh^2(\frac{s+s_0}{r}) + C\tanh(\frac{s+s_0}{r}) + B].$$

Also, again by using Theorem 2.3 and Theorem 2.4, we obtain

$$\kappa_\gamma^2(s) = \frac{\tau_\gamma^2}{(c_1\sinh(\frac{s+s_0}{r}) + c_2\cosh(\frac{s+s_0}{r}))^2},$$

$$\Rightarrow \kappa_\gamma^2(s) = \frac{\sinh^2(\rho/r)\cosh^2(\frac{s+s_0}{r})[A\tanh^2(\frac{s+s_0}{r}) + C\tanh(\frac{s+s_0}{r}) + B]}{\cosh^2(\frac{s+s_0}{r})[c_1^2\tanh^2(\frac{s+s_0}{r}) + 2c_1c_2\tanh(\frac{s+s_0}{r}) + c_2^2]}.$$

Thus $\kappa_\gamma^2(s) = \sinh^2(\rho/r)$ if $A = c_1^2$, $B = c_2^2$ and $C = 2c_1c_2$, which concludes the theorem. □

**Corollary 3.7.** *The geodesic curvature $\kappa_V$ of rectifying general helix in $H^3(-r)$ is given by $\kappa_V = \frac{v^2}{r}$, where $v$ is the speed of rectifying general helix.*

***Proof.*** The proof is obtained from Theorem 3.6. □

**Theorem 3.8.** *A curve $\gamma(s) = \exp(\rho(s)V(s))$ in $H^3(-r)$ is a rectifying general helix with geodesic curvature $\kappa_V(t) = c(cos^2(t+t_0) - a^2)^{-3/2}$ and torsion $\tau(s) = d_1\sinh((s+s_0)/r) + d_2\cosh((s+s_0)/r)$ then its curvature $\kappa_\gamma$ is of the form $\kappa_\gamma = \frac{d_1}{c_1}$ if and only if*

$$\begin{vmatrix} c_1 & c_2 \\ d_1 & d_2 \end{vmatrix} = 0.$$

***Proof.*** By using Corollary 9 of [7], we obtain

$$\kappa_\gamma = \frac{d_1\sinh((s+s_0)/r) + d_2\cosh((s+s_0)/r)}{c_1\sinh(\frac{s+s_0}{r}) + c_2\cosh(\frac{s+s_0}{r})},$$

$$\Rightarrow \kappa_\gamma = \frac{d_1(\tanh(s+s_0)/r) + A)}{c_1(\tanh(\frac{s+s_0}{r}) + B)},$$

where $A = \dfrac{d_2}{d_1}$ and $B = \dfrac{c_2}{c_1}$.

Thus $\kappa_\gamma = \dfrac{d_1}{c_1}$ if and only if $A = B$ i.e.

$$\begin{vmatrix} c_1 & c_2 \\ d_1 & d_2 \end{vmatrix} = 0.$$

□

**Acknowledgment.** We thank to the referees for their valuable suggestions to improve the paper.

## References

[1] P. Alegre, K. Arslan, A. Carriazo, C. Murathan and G. Ozturk, *Some special types of developable ruled surface*, Hacet. J. Math. Stat. **39** (3), 319–325, 2010.

[2] B. Altunkaya and L. Kula, *On spacelike rectifying slant helices in Minkowski 3-space*, Turkish J. Math. **42**, 1098–1110, 2018.

[3] M. Barros, *General helices and a theorem of Lancret*, Proc. Amer. Math. Soc. **125** (5), 1503–1509, 1997.

[4] B.Y. Chen, *When does the position vector of a space curve always lie in its rectifying plane?*, Amer. Math. Monthly **110**, 147–152, 2003.

[5] S. Izumiya and N. Takeuchi, *New special curves and developable surfaces*, Turkish J. Math. **28** (2), 153–163, 2004.

[6] K. Ilarslan, E. Nesovic and M.P. Torgasev, *Some characterization of rectifying curves in the Minkowski 3-space*, Novi Sad J. Math. **33** (2), 23–32, 2003.

[7] P. Lucas and J.A.O. Yagues, *Rectifying curves in the three dimensional hyperbolic space*, Mediterr. J. Math. **13**, 2199–2214, 2016.

[8] P. Lucas and J.A.O. Yagues, *Slant helices in the three dimensional sphere*, J. Korean Math. Soc. **54** (4), 1331–1343, 2017.

Hacettepe Journal of Mathematics & Statistics

Hacet. J. Math. Stat.
Volume 50 (1) (2021), 243 – 254
DOI : 10.15672/hujms.605105

Research Article

# Mappings between the lattices of saturated submodules with respect to a prime ideal

Morteza Noferesti, Hosein Fazaeli Moghimi*, Mohammad Hossein Hosseini

*Department of Mathematics, University of Birjand, P.O.Box 97175-615, Birjand, Iran*

## Abstract

Let $\mathfrak{S}_p({}_RM)$ be the lattice of all saturated submodules of an $R$-module $M$ with respect to a prime ideal $p$ of a commutative ring $R$. We examine the properties of the mappings $\eta : \mathfrak{S}_p({}_RR) \to \mathfrak{S}_p({}_RM)$ defined by $\eta(I) = S_p(IM)$ and $\theta : \mathfrak{S}_p({}_RM) \to \mathfrak{S}_p({}_RR)$ defined by $\theta(N) = (N : M)$, in particular considering when these mappings are lattice homomorphisms. It is proved that if $M$ is a semisimple module or a projective module, then $\eta$ is a lattice homomorphism. Also, if $M$ is a faithful multiplication $R$-module, then $\eta$ is a lattice epimorphism. In particular, if $M$ is a finitely generated faithful multiplication $R$-module, then $\eta$ is a lattice isomorphism and its inverse is $\theta$. It is shown that if $M$ is a distributive module over a semisimple ring $R$, then the lattice $\mathfrak{S}_p({}_RM)$ forms a Boolean algebra and $\eta$ is a Boolean algebra homomorphism.

**Mathematics Subject Classification (2020).** 13C13, 06B99, 06E99, 13C99

**Keywords.** saturated submodules with respect to a prime ideal, $\eta$-modules, $\theta$-modules, $\mathfrak{S}$-distributive modules, semisimple rings

## 1. Introduction

We assume throughout this paper that all rings are commutative with nonzero identity and all modules are unitary. Let $R$ be a ring and $M$ be an $R$-module. For any submodule $N$ of $M$, we denote the annihilator of the $R$-module $M/N$ by $(N : M)$, i.e., $(N : M) = \{r \in R \mid rM \subseteq N\}$.

It is well-known that the collection of all submodules of $M$ forms a lattice with respect to the operations $\vee$ and $\wedge$ defined by

$$L \vee N = L + N \text{ and } L \wedge N = L \cap N.$$

Note that this lattice, denoted $\mathcal{L}({}_RM)$, is bounded with the least element (0) and greatest element $M$. Recently, P.F. Smith has studied several mappings between $\mathcal{L}({}_RR)$ and $\mathcal{L}({}_RM)$ [22–24]. For instance, in [22], he examined conditions under which the mappings $\lambda : \mathcal{L}({}_RR) \to \mathcal{L}({}_RM)$ defined by $\lambda(I) = IM$ and $\mu : \mathcal{L}({}_RM) \to \mathcal{L}({}_RR)$ defined by $\mu(N) = (N : M)$ are injective, surjective or lattice homomorphisms. An $R$-module $M$ is called a $\lambda$-*module* (respectively $\mu$-*module*), if $\lambda$ (respectively $\mu$) is a lattice homomorphism.

*Corresponding Author.
Email addresses: morteza_noferesti@birjand.ac.ir (M. Noferesti), hfazaeli@birjand.ac.ir (H.F. Moghimi), mhhosseini@birjand.ac.ir (M.H. Hosseini)
Received: 12.08.2019; Accepted: 03.06.2020

The study of the mappings $\lambda$ and $\mu$ continued in [23], considering when these mappings are complete lattice homomorphisms.

A proper submodule $P$ of $M$ is called a *prime submodule* if for $r \in R$ and $x \in M$, $rx \in P$ implies that $r \in (P : M)$ or $x \in P$ (see, for example, [2,6,18,19]). For a proper submodule $N$ of an $R$-module $M$, the intersection of all prime submodules of $M$ containing $N$ is called the *radical* of $N$ and denoted by $\operatorname{rad} N$; if there are no such prime submodules, $\operatorname{rad} N$ is $M$ (see, for example, [11,14,17]). A submodule $N$ of $M$ is called a *radical submodule* if $\operatorname{rad} N = N$. The collection of all radical submodules of $M$ which is denoted by $\mathcal{R}(_RM)$ forms a lattice with respect to the following operations:

$$L \vee N = \operatorname{rad}(L + N) \text{ and } L \wedge N = L \cap N.$$

Note that $\mathcal{R}(_RM)$ is a bounded lattice with the least element $\operatorname{rad}(0)$ and the greatest element $M$.

In [20], H.F. Moghimi and J.B. Harehdashti have studied the properties of the mappings $\rho : \mathcal{R}(_RR) \to \mathcal{R}(_RM)$ defined by $\rho(I) = \operatorname{rad}(IM)$ and $\sigma : \mathcal{L}(_RR) \to \mathcal{L}(_RM)$ defined by $\sigma(N) = (N : M)$, in particular considering when these mappings are lattice monomorphisms or epimorphisms. Later in [9], they investigated conditions under which these mappings are complete homomorphisms. Note that $\rho$ is always a lattice homomorphism, but not necessarily a complete lattice homomorphism. An $R$-module $M$ is called a *$\sigma$-module* if $\sigma$ is a lattice homomorphism.

Let $M$ be an $R$-module. For a prime ideal $p$ of $R$ and a submodule $N$ of $M$, the set $S_p(N) = \{m \in M \mid cm \in N \text{ for some } c \in R \setminus p\}$ is called the *saturation* of $N$ with respect to $p$. It is clear that $N \subseteq S_p(N)$. It is said that $N$ is *saturated* with respect to $p$, if $N = S_p(N)$. It is easily seen that $S_p(N)$ is a saturated submodule of $M$ (see [15,16], for more details about saturation of submodules). The collection of all saturated submodules of an $R$-module $M$ with respect to a fixed prime ideal $p$ of $R$ is a lattice with the following operations:

$$L \vee N = S_p(L + N) \quad \text{and} \quad L \wedge N = L \cap N.$$

We shall denote this lattice by $\mathfrak{S}_p(_RM)$, or by $\mathfrak{S}_p(M)$ if there is no ambiguity about $R$. Note that $\mathfrak{S}_p(M)$ is bounded, with the least element $S_p(0)$ and the greatest element $M$.

Let $R$ be a ring, $p$ a fixed prime ideal of $R$ and $M$ an $R$-module. Now consider the mappings $\eta : \mathfrak{S}_p(R) \to \mathfrak{S}_p(M)$ defined by

$$\eta(I) = S_p(IM),$$

for every saturated ideal $I$ of $R$, and $\theta : \mathfrak{S}_p(M) \to \mathfrak{S}_p(R)$ defined by

$$\theta(N) = (N : M),$$

for every saturated submodule $N$ of $M$. It will be convenient for us to call the module $M$ an $\eta$-module (resp. a $\theta$-module) in case the above mapping $\eta$ (resp. $\theta$) is a lattice homomorphism.

In this paper, we investigate conditions under which $\eta$ and $\theta$ are lattice homomorphisms, in particular considering when $\eta$ and $\theta$ are Boolean algebra homomorphisms. It is shown that modules over Prüfer domains (Corollary 2.4), projective modules (Corollary 2.6) and semisimple $R$-modules (Corollary 2.7) are three classes of $\eta$-modules. It is proved that if $M$ is a faithful multiplication $R$-module, then $\eta$ is a lattice epimorphism, and in particular $\mathfrak{S}_p(M)$ is isomorphic to a quotient of $\mathfrak{S}_p(R)$ (Theorem 2.8) for all prime ideals $p$ of $R$. It is shown that a finitely generated module $M$ is a $\theta$-module if and only if it is a multiplication module (Corollary 2.11). In particular, every cyclic $R$-module is a $\theta$-module (Corollary 2.10). Moreover, if $M$ is a finitely generated faithful multiplication $R$-module then $\eta$ and $\theta$ are lattice isomorphisms (Corollary 2.17).

An $R$-module $M$ is called *distributive* if $\mathcal{L}(_RM)$ is a distributive lattice (see, for example,

[8]). A ring $R$ is called *arithmetical* if it is a distributive $R$-module. We say that an $R$-module $M$ is $\mathfrak{S}$-*distributive* with respect to a prime ideal $p$ of $R$ if $\mathfrak{S}_p(M)$ is a distributive lattice. It is proved that an $R$-module $M$ is distributive if and only if it is $\mathfrak{S}$-distributive with respect to any prime ideal of $R$ (Corollary 3.4). In particular, every multiplication module over an arithmetical ring $R$ is $\mathfrak{S}$-distributive with respect to any prime ideal of $R$ (Corollary 3.5). It is shown that if $M$ is a distributive module over a semisimple ring $R$, then $\mathfrak{S}_p(M)$ forms a Boolean algebra (Theorem 3.7) and $\eta$ is a Boolean algebra homomorphism (Theorem 3.13). In particular, if $M$ is a multiplication module over a semisimple ring $R$, then $\eta$ is a Boolean algebra epimorphism (Corollary 3.14).

## 2. $\eta$-modules and $\theta$-modules

We start with a lemma which collects some facts about saturation of submodules.

**Lemma 2.1.** *Let $R$ be a ring, $p$ a prime ideal of $R$ and $M$ an $R$-module. Then*

(1) $S_p(L \cap N) = S_p(L) \cap S_p(N)$ *for all submodules $L$ and $N$ of $M$;*

(2) $S_p(S_p(IM) + S_p(JM)) = S_p(S_p(I+J)M) = S_p(IM+JM)$ *for all ideals $I$ and $J$ of $R$.*

***Proof.*** (1) Clear.
(2) Since $IM \subseteq (I+J)M \subseteq S_p(I+J)M$, we conclude that $S_p(IM) \subseteq S_p(S_p(I+J)M)$. Similarly, $S_p(JM) \subseteq S_p(S_p(I+J)M)$. Therefore, we have $S_p(IM)+S_p(JM) \subseteq S_p(S_p(I+J)M)$. Hence we have $S_p(S_p(IM)+S_p(JM)) \subseteq S_p(S_p(I+J)M)$. Now, let $x \in S_p(S_p(I+J)M)$. Then there exists $c \in R \setminus p$ such that $cx \in S_p(I+J)M$. Therefore $cx = \sum_{i=1}^k r_i x_i$ for some $r_i \in S_p(I+J)$ and $x_i \in M$ $(1 \le i \le k)$. Thus there are $c_i \in R \setminus p$ $(1 \le i \le k)$ such that $c_i r_i \in I+J$, and so $c_1 \dots c_k cx \in (I+J)M$. It follows that $x \in S_p((I+J)M)$. Hence we have $S_p(S_p(I+J)M) \subseteq S_p(IM+JM)$. It is also clear that $S_p(IM+JM) \subseteq S_p(S_p(IM)+S_p(JM))$. □

**Theorem 2.2.** *Let $R$ be a ring, $p$ a prime ideal of $R$ and $M$ an $R$-module. Then the following statements are equivalent:*

(1) *$M$ is an $\eta$-module over $R$;*

(2) $S_p((I \cap J)M) = S_p(IM) \cap S_p(JM)$ *for all ideals $I$ and $J$ of $R$;*

(3) $(I_p \cap J_p)M_p = I_pM_p \cap J_pM_p$ *for all ideals $I$ and $J$ of $R$;*

(4) *$M_p$ is a $\lambda$-module over $R_p$.*

***Proof.*** (1) ⇒ (2) By definition.
(2) ⇒ (1) Let $I, J \in \mathfrak{S}_p(R)$. By the assumption, $\eta(I \wedge J) = \eta(I) \wedge \eta(J)$.
By using Lemma 2.1, we have

$$\begin{aligned}\eta(I \vee J) &= S_p((I \vee J)M) = S_p(S_p(I+J)M)\\ &= S_p(S_p(IM)+S_p(JM))\\ &= S_p(IM) \vee S_p(JM)\\ &= \eta(I) \vee \eta(J).\end{aligned}$$

(2) ⇒ (3) Let $z \in I_pM_p \cap J_pM_p$. Then $z = \sum_{i=1}^k a_i x_i/s_i = \sum_{i=1}^k b_i y_i/t_i$ for some $a_i \in I$, $b_i \in J$, $x_i, y_i \in M$, $s_i, t_i \in R \setminus p$. Hence we have $s_1 \dots s_k t_1 \dots t_k z \in IM \cap JM$ which follows that $z \in S_p(IM) \cap S_p(JM)$. Therefore by (2), $z \in S_p((I \cap J)M)$. Thus $cz \in (I \cap J)M$ for some $c \in R \setminus p$, and so $z \in (I_p \cap J_p)M_p$ as desired. The reverse inclusion is clear.
(3) ⇒ (2) Let $x \in S_p(IM) \cap S_p(JM)$. Then $cx \in IM$ and $dx \in JM$ for some $c, d \in R \setminus p$. Therefore $cx = \sum_{i=1}^k c_i x_i$ and $dx = \sum_{j=1}^k d_j x'_j$ for some $c_i \in I$, $d_j \in J$ and $x_i, x'_j \in M$ $(1 \le i, j \le k)$. Thus $c_1 dx = \sum_{j=1}^k c_1 d_j x'_j$ and hence $c_1 dx \in (I \cap J)M$ such that $c_1 d \in R \setminus p$. Thus $x \in S_p((I \cap J)M)$. The reverse inclusion is clear.
(3) ⇔ (4) Follows from [22, Lemma 2.1 (ii)]. □

Let $R$ be a domain with the field of fractions $K$. A non-zero ideal $I$ of $R$ is called *invertible* provided $I^{-1}I = R$ where $I^{-1} = \{k \in K : kI \subseteq R\}$. A domain $R$ is called *Prüfer* if every non-zero finitely generated ideal of $R$ is invertible (see, for more details, [13]).

**Corollary 2.3.** *Let $R$ be a domain, $p$ a prime ideal of $R$ and $M$ an $R$-module. Then the following statements are equivalent:*

(1) *$R_p$ is Prüfer;*
(2) *Every $R_p$-module is a $\lambda$-module;*
(3) *Every $R$-module is an $\eta$-module.*

***Proof.*** (1) $\Leftrightarrow$ (2) By [22, Theorem 2.3].
(2) $\Leftrightarrow$ (3) By Theorem 2.2. □

**Corollary 2.4.** *Let $R$ be any Prüfer domain. Then every $R$-module is an $\eta$-module.*

***Proof.*** Let $R$ be a Prüfer domain and $p$ be a prime ideal of $R$. Then by [13, Theorem 6.6], $R_p$ is a valuation ring. Thus by [22, Proposition 2.4], every $R_p$-module is a $\lambda$-module and hence by Corollary 2.3, every $R$-module is an $\eta$-module. □

**Theorem 2.5.** *Let $R$ be any ring. Then*

(1) *Every direct summand of an $\eta$-module is an $\eta$-module.*
(2) *Every direct sum of $\lambda$-modules is an $\eta$-module.*

***Proof.*** (1) Let $K$ be a direct summand of an $\eta$-module $M$. Let $I$ and $J$ be any ideals of $R$ and $p$ be a prime ideal of $R$. Then by Lemma 2.1 (1) and Theorem 2.2, we have

$$\begin{aligned} S_p(IK) \cap S_p(JK) &= S_p(K \cap IM) \cap S_p(K \cap JM) \\ &= S_p(K) \cap S_p(IM) \cap S_p(JM) \\ &= S_p(K) \cap S_p((I \cap J)M) \\ &= S_p(K \cap (I \cap J)M) \\ &= S_p((I \cap J)K). \end{aligned}$$

Thus by Theorem 2.2, $K$ is an $\eta$-module.
(2) Let $M_i$ ($i \in \mathfrak{I}$) be any collection of $\lambda$-modules and let $M = \oplus_{i\in\mathfrak{I}} M_i$. Given any ideals $I$ and $J$ of $R$, by [22, Lemma 2.1], we have

$$\begin{aligned} S_p(IM) \cap S_p(JM) &= S_p(\oplus_{i\in\mathfrak{I}} IM_i) \cap S_p(\oplus_{i\in\mathfrak{I}} JM_i) \\ &= S_p(\oplus_{i\in\mathfrak{I}} IM_i \cap \oplus_{i\in\mathfrak{I}} JM_i) \\ &= S_p(\oplus_{i\in\mathfrak{I}} (IM_i \cap JM_i)) \\ &= S_p(\oplus_{i\in\mathfrak{I}} (I \cap J)M_i) \\ &= S_p((I \cap J)M). \end{aligned}$$

Thus by Theorem 2.2, $M$ is an $\eta$-module. □

**Corollary 2.6.** *For any ring $R$, every projective $R$-module is an $\eta$-module.*

***Proof.*** By [22, Lemma 2.1], every ring $R$ is a $\lambda$-module. Thus by [10, Theorem IV.2.1] and Theorem 2.5(2), every free $R$-module is an $\eta$-module, and therefore by [10, Theorem IV.3.4] and Theorem 2.5(1), every projective $R$-module is an $\eta$-module. □

**Corollary 2.7.** *For any ring $R$, every semisimple $R$-module is an $\eta$-module.*

***Proof.*** Clearly every simple module is a $\lambda$-module. Since any semisimple module is a direct sum of a family of simple submodules, the result follows from Theorem 2.5(2). □

An $R$-module $M$ is called a *multiplication* module if the mapping $\lambda$ is surjective, i.e., for each submodule $N$ of $M$ there exist an ideal $I$ of $R$ such that $N = IM$. In this case, we can take $I = (N : M)$ (see, for example, [4, 7]).

**Theorem 2.8.** *Let $M$ be a faithful multiplication $R$-module. Then $\eta$ is a lattice epimorphism.*
*In particular, $\mathfrak{S}_p(M)$ is isomorphic to a quotient of $\mathfrak{S}_p(R)$ for all prime ideals $p$ of $R$.*

***Proof.*** Since $M$ is a faithful multiplication $R$-module, $M$ is a $\lambda$-module by [22, Theorem 2.12]. Thus by [22, Lemma 2.1], $(I \cap J)M = IM \cap JM$ for all ideals $I$ and $J$ of $R$. It follows that, by Lemma 2.1 (1),

$$S_p((I \cap J)M) = S_p(IM \cap JM) = S_p(IM) \cap S_p(JM)$$

for all ideals $I$ and $J$ and prime ideals $p$ of $R$. Hence by Theorem 2.2, $\eta$ is a lattice homomorphism. Now, let $p$ be a prime ideal of $R$ and $N \in \mathfrak{S}_p(M)$. Since $M$ is a multiplication module, we have

$$\eta((N : M)) = S_p((N : M)M) = S_p(N) = N$$

and therefore $\eta$ is an epimorphism. Now, we define the relation $\sim$ on $\mathfrak{S}_p(R)$ by

$$I \sim J \Leftrightarrow S_p(IM) = S_p(JM).$$

It is evident that $\sim$ is an equivalence relation on $\mathfrak{S}_p(R)$. We show that $\sim$ is a congruence relation. Assume that $I_1 \sim J_1$ and $I_2 \sim J_2$. Thus we have $S_p(I_1M) = S_p(J_1M)$ and $S_p(I_2M) = S_p(J_2M)$. Since $M$ is a faithful multiplication module,

$$\begin{aligned} S_p((I_1 \cap J_1)M) &= S_p(I_1M) \cap S_p(J_1M) \\ &= S_p(I_2M) \cap S_p(J_2M) \\ &= S_p((I_2 \cap J_2)M), \end{aligned}$$

and therefore $I_1 \wedge J_1 \sim I_2 \wedge J_2$. Also, by Lemma 2.1 (2),

$$\begin{aligned} S_p(S_p(I_1 + J_1)M) &= S_p(S_p(I_1M) + S_p(J_1M)) \\ &= S_p(S_p(I_2M) + S_p(J_2M)) \\ &= S_p(S_p(I_2 + J_2)M) \end{aligned}$$

which follows that $I_1 \vee J_1 \sim I_2 \vee J_2$. Thus $\mathfrak{S}_p(R)/\sim$, the set of equivalence classes with respect to $\sim$, is a lattice with the following operations:

$$I/\sim \ \tilde{\vee}\ J/\sim = I \vee J/\sim \ \text{ and } \ I/\sim \ \tilde{\wedge}\ J/\sim = I \wedge J/\sim.$$

Now, the mapping $\bar{\eta} : \mathfrak{S}_p(R)/\sim \to \mathfrak{S}_p(M)$ given by $\bar{\eta}(I/\sim) = \eta(I) = S_p(IM)$ is a lattice isomorphism. □

Recall that $\theta : \mathfrak{S}_p(M) \to \mathfrak{S}_p(R)$ defined by $\theta(N) = (N : M)$ is the restriction of the mapping $\mu : \mathcal{L}({}_RM) \to \mathcal{L}({}_RR)$ to $\mathfrak{S}_p(M)$ given in [22]. Thus every $\mu$-module is a $\theta$-module.

**Theorem 2.9.** *Let $R$ be a ring and $M$ an $R$-module. Consider the following statements:*

(1) *$M$ is a $\theta$-module over $R$;*
(2) *$(L + N : M) = (L : M) + (N : M)$ for all saturated submodules $L$ and $N$ of $M$;*
(3) *$(L_p + N_p : M_p) = (L_p : M_p) + (N_p : M_p)$ for all submodules $L$ and $N$ of $M$ and for all prime ideals $p$ of $R$;*
(4) *$(L + N : M) = (L : M) + (N : M)$ for all submodules $L$ and $N$ of $M$;*
(5) *$M$ is a $\mu$-module over $R$.*

*Then (1) $\Leftrightarrow$ (2) and (4) $\Leftrightarrow$ (5).*
*In particular, if $M$ is a finitely generated $R$-module, then all of the above statements are equivalent.*

***Proof.*** (1) $\Leftrightarrow$ (2) Follows from definition.
(4) $\Leftrightarrow$ (5) Follows from [22, Lemma 3.1].
(4) $\Rightarrow$ (2) Clear.
(2) $\Rightarrow$ (3) Suppose that $M$ is finitely generated. Then $M = Rm_1 + \ldots + Rm_k$ for some $m_i \in M$ $(1 \le i \le k)$. Let $L$ and $N$ be two submodules of $M$. First we show that $(S_p(L)+S_p(N) : M)_p = ((L+N)_p : M_p)$ for all prime ideals $p$ of $R$. Let $p$ be a prime ideal of $R$ and assume that $r/1 \in (S_p(L) + S_p(N) : M)_p$. It follows that $rM \subseteq S_p(L) + S_p(N)$. Thus $rm_i = x_i + y_i$ for some $x_i \in S_p(L)$, $y_i \in S_p(N)$ $(1 \le i \le k)$. Therefore $c_i x_i \in L$ and $d_i y_i \in N$ for some $c_i, d_i \in R \setminus p$ $(1 \le i \le k)$. Now, since $c_1 \ldots c_k d_1 \ldots d_k rM \subseteq L + N$, we have $r/1 \in ((L + N)_p : M_p)$, as requested. Hence, by using [15, Theorem 2.1], we have

$$\begin{aligned}(L_p : M_p) + (N_p : M_p) &= (S_p(L) : M)_p + (S_p(N) : M)_p \\ &= ((S_p(L) : M) + (S_p(N) : M))_p \\ &= (S_p(L) + S_p(N) : M)_p \\ &= ((L + N)_p : M_p) \\ &= (L_p + N_p : M_p).\end{aligned}$$

(3) $\Rightarrow$ (4) Follows from [3, Proposition 3.8 and Corollaries 3.4 and 3.15].
(4) $\Rightarrow$ (3) Follows from [3, Corollary 3.4 and Corollary 3.15]. □

**Corollary 2.10.** *For any ring $R$, every cyclic $R$-module is a $\theta$-module.*

***Proof.*** Follows from [22, Corollary 3.7] and Theorem 2.9. □

**Corollary 2.11.** *Let $M$ be a finitely generated $R$-module. Then the following statements are equivalent:*

(1) *$M$ is a $\theta$-module over $R$;*
(2) *$M_p$ is a $\theta$-module over $R_p$ for every prime ideal $p$ of $R$;*
(3) *$M_m$ is a $\theta$-module over $R_m$ for every maximal ideal $m$ of $R$;*
(4) *$M$ is a $\mu$-module over $R$;*
(5) *$M$ is a $\sigma$-module over $R$;*
(6) *$M$ is a multiplication module over $R$.*

***Proof.*** (1) $\Leftrightarrow$ (4) By Theorem 2.9.
(4) $\Leftrightarrow$ (5) $\Leftrightarrow$ (6) By [20, Theorem 2.11 and Theorem 2.19].
(6) $\Leftrightarrow$ (2) $\Leftrightarrow$ (3) By [4, Lemma 2 (ii)], [20, Theorem 2.11] and Theorem 2.9. □

**Corollary 2.12.** *Let $R$ be a ring. If $M$ is a finitely generated $\theta$-module over $R$ and $((0) : M) = Re$ for some idempotent $e$ of $R$, then $M$ is an $\eta$-module over $R$. In particular, every finitely generated faithful $\theta$-module is an $\eta$-module.*

***Proof.*** By Corollary 2.11 $M$ is a multiplication $R$-module, and then by [21, Theorem 11] $M$ is a projective $R$-module. Thus by Corollary 2.6, $M$ is an $\eta$-module over $R$. □

Now, we investigate conditions under which $\eta$ and $\theta$ are injective or surjective.

**Theorem 2.13.** *Let $\eta$ and $\theta$ be as before. Then*

(1) *$\eta\theta\eta = \eta$;*
(2) *$\theta\eta\theta = \theta$.*

***Proof.*** (1) Let $p$ be a prime ideal of $R$ and $I \in \mathfrak{S}_p(R)$. Since $\eta\theta\eta(I) = S_p((S_p(IM) : M)M)$, we must show that $S_p((S_p(IM) : M)M) = S_p(IM)$. First note that, since $I \subseteq (S_p(IM) : M)$, we have $IM \subseteq (S_p(IM) : M)M$ and thus $S_p(IM) \subseteq S_p((S_p(IM) : M)M)$. The reverse inclusion follows from

$$S_p((S_p(IM) : M)M) \subseteq S_p(S_p(IM)) = S_p(IM).$$

(2) Let $p$ be a prime ideal of $R$ and $N \in \mathfrak{S}_p(M)$. Now, since $\theta\eta\theta(N) = (S_p((N:M)M) : M)$, we must show that $(S_p((N:M)M) : M) = (N:M)$. Since $(N:M)M \subseteq S_p((N:M)M)$, we have $(N:M) \subseteq (S_p((N:M)M) : M)$. The reverse inclusion follows from

$$(S_p((N:M)M) : M) \subseteq (S_p(N) : M) = (N:M).$$

$\square$

**Corollary 2.14.** *Let $\eta$ and $\theta$ be as before, and $p$ be a prime ideal of $R$. Then the following statements are equivalent:*

(1) $\eta : \mathfrak{S}_p(R) \to \mathfrak{S}_p(M)$ *is a surjection;*
(2) $\eta\theta = 1$*;*
(3) $S_p((N:M)M) = N$ *for all* $N \in \mathfrak{S}_p(M)$*;*
(4) $\theta : \mathfrak{S}_p(M) \to \mathfrak{S}_p(R)$ *is an injection.*

***Proof.*** (1) $\Rightarrow$ (2) and (4) $\Rightarrow$ (2) follows from Theorem 2.13.
(2) $\Leftrightarrow$ (3), (2) $\Rightarrow$ (1) and (2) $\Rightarrow$ (4) are clear. $\square$

**Corollary 2.15.** *Let $\eta$ and $\theta$ be as before, and $p$ be a prime ideal of $R$. Then the following statements are equivalent:*

(1) $\eta : \mathfrak{S}_p(R) \to \mathfrak{S}_p(M)$ *is an injection;*
(2) $\theta\eta = 1$*;*
(3) $(S_p(IM) : M) = I$ *for all* $I \in \mathfrak{S}_p(R)$*;*
(4) $\theta : \mathfrak{S}_p(M) \to \mathfrak{S}_p(R)$ *is a surjection.*

***Proof.*** (1) $\Rightarrow$ (2) and (4) $\Rightarrow$ (2) follows from Theorem 2.13.
(2) $\Leftrightarrow$ (3), (2) $\Rightarrow$ (1) and (2) $\Rightarrow$ (4) are clear. $\square$

**Corollary 2.16.** *Let $\eta$ and $\theta$ be as before. Then $\eta$ is a bijection if and only if $\theta$ is a bijection. In this case $\eta$ and $\theta$ are inverse of each other.*

***Proof.*** By Corollaries 2.14 and 2.15. $\square$

**Corollary 2.17.** *Let $R$ be a ring and $M$ be a finitely generated faithful multiplication $R$-module. Then the mappings $\eta$ and $\theta$ are lattice isomorphisms. In particular, $\eta$ and $\theta$ are inverse of each other, and therefore $\mathfrak{S}_p(R)$ and $\mathfrak{S}_p(M)$ are isomorphic lattices for all prime ideals $p$ of $R$.*

***Proof.*** Since $M$ is a faithful multiplication $R$-module, $\eta$ is an epimorphism by Theorem 2.8, and hence $\theta$ is a monomorphism by Corollary 2.14 and [22, Theorem 3.8]. On the other hand, by [15, Proposition 3.2], we have

$$(S_p(IM) : M) = S_p(IM : M) = S_p(I) = I,$$

for all prime ideals $p$ of $R$ and $I \in \mathfrak{S}_p(R)$. Hence, by Corollary 2.15, $\eta$ is an injection and $\theta$ is a surjection. Hence $\eta$ is an isomorphism and its inverse is $\theta$. $\square$

## 3. $\mathfrak{S}_p(M)$ as a Boolean algebra

We start this section by recalling the following basic definition.

**Definition 3.1.** Let $R$ be a ring and $p$ be a prime ideal of $R$. An $R$-module $M$ is called a $\mathfrak{S}$-*distributive* module with respect to $p$, if $\mathfrak{S}_p(M)$ is a distributive lattice.

First note the following simple fact.

**Lemma 3.2.** *Let $R$ be a ring, $p$ a prime ideal of $R$ and $M$ be an $R$-module. Then the following statements are equivalent:*

(1) *$M$ is $\mathfrak{S}$-distributive with respect to $p$;*
(2) $K \cap S_p(L+N) = S_p((K \cap L) + (K \cap N))$ *for all* $K, L, N \in \mathfrak{S}_p(M)$*;*

(3) *$S_p(K+(L\cap N)) = S_p(K+L)\cap S_p(K+N)$ for all $K, L, N \in \mathfrak{S}_p(M)$.*

***Proof.*** By [5, Theorem I.3.2]. □

The following example shows that a ring $R$ may be $\mathfrak{S}$-distributive with respect to a prime ideal and not with respect to another one.

**Example 3.3.** Let $R = K[X, Y]$ be the ring of polynomials with independent indeterminates $X$ and $Y$ over a field $K$. It is evident that $R$ is $\mathfrak{S}$-distributive with respect to (0), since $\mathfrak{S}_{(0)}(R) = \{(0), R\}$. However, $R$ is not $\mathfrak{S}$-distributive with respect to $m = RX+RY$. Let $p_1 = RX$, $p_2 = RY$, $p_3 = R(X+Y)$ . Since $p_1$, $p_2$ and $p_3$ are prime ideals of $R$, these ideals are saturated with respect to $m$ and hence $p_3\cap p_1$ and $p_3\cap p_2$ are saturated with respect to $m$ by Lemma 2.1 (1). Now, since $p_3\cap(p_1+p_2)\nsubseteq(p_3\cap p_1)+(p_3\cap p_2)$, $R$ is not $\mathfrak{S}$-distributive with respect to $m$ by Lemma 3.2.

It is remarked that some classes of $R$-modules are characterized by using the localization with respect to all prime ideal of $R$ (see for example [1]). In the next result, it is seen that the class of distributive modules has this property.

**Corollary 3.4.** *Let $R$ be a ring and $M$ be an $R$-module. Then the following conditions are equivalent:*

(1) *$M$ is a distributive $R$-module;*
(2) *$M$ is $\mathfrak{S}$-distributive with respect to any prime ideal $p$ of $R$;*
(3) *$M_p$ is a distributive $R_p$-module for all prime ideals $p$ of $R$.*

***Proof.*** (1) ⇒ (2) Let $p$ be a prime ideal of $R$ and $K, L, N \in \mathfrak{S}_p(M)$. By Lemma 2.1 (1) and the assumption, we have

$$S_p(K+L)\cap S_p(K+N) = S_p((K+L)\cap(K+N)) = S_p(K+(L\cap N)).$$

Thus, the result follows from Lemma 3.2 (3).
(2) ⇒ (3) Let $p$ be a prime ideal of $R$ and $K$, $L$ and $N$ be submodules of $M$. It suffices to show that $(K_p+L_p)\cap(K_p+N_p)\subseteq(K_p+(L_p\cap N_p))$ or equivalently, by [3, Corollary 3.4], $((K+L)\cap(K+N))_p\subseteq(K+(L\cap N))_p$. For this, let $x/s\in((K+L)\cap(K+N))_p$. Thus there are elements $k_1, k_2\in K$, $l\in L$, $n\in N$ and $s_1, s_2\in R\setminus p$ such that $x/s = (k_1+l)/s_1 = (k_2+n)/s_2$. It follows that $uss_1s_2x = (k_1+l) = (k_2+n)$ for some $u\in R\setminus p$ so that $x\in S_p(K+L)\cap S_p(K+N)$. Hence by (2), $x\in S_p(K+(L\cap N))$. Therefore $cx\in K+(L\cap N)$ for some $c\in R\setminus p$ which implies that $x/s = cx/cs\in(K+(L\cap N))_p$, as required.
(3) ⇒ (1) Follows from [3, Corollary 3.4 and Proposition 3.8].

□

**Corollary 3.5.** *Let $R$ be an arithmetical ring, and $M$ be a multiplication $R$-module. Then $M$ is a $\mathfrak{S}$-distributive $R$-module with respect to any prime ideal of $R$.*

***Proof.*** By [8, Proposition 1.2] and Corollary 3.4. □

Our next example shows that $M$ being a multiplication module is needed in Corollary 3.5.

**Example 3.6.** Let $K$ be a field and $V = K\oplus K$ be the usual two-dimensional vector space over $K$. It is easy to see that every subspace of $V$ is saturated with respect to (0). Now if $W_1 = K(1,0)$, $W_2 = K(0,1)$ and $W_3 = K(1,1)$. Then $W_3\cap(W_1+W_2) = W_3$ while $(W_3\cap W_1)+(W_3\cap W_2) = K(0,0)$. Thus $V$ is not $\mathfrak{S}$-distributive

We recall that a distributive lattice $(L, \vee, \wedge)$ is a Boolean algebra if there is a unary operation $'$ on $L$ and two constants 0 and 1 such that $x\wedge x' = 0$ and $x\vee x' = 1$.

Let $M$ be a semisimple $R$-module and $N$ a submodule of $M$. Then, by definition, there is a submodule $L$ of $M$ such that $M = N\oplus L$. We define the unary operation $'$ on $\mathfrak{S}_p(M)$ by $N' = S_p(L)$.

**Theorem 3.7.** *Let $R$ be a semisimple ring, $p$ a prime ideal of $R$ and $M$ a distributive $R$-module. Then the lattice $\mathfrak{S}_p(M)$ is a Boolean algebra with the unary operation $'$ defined above, $\mathbf{0} = S_p(0)$ and $\mathbf{1} = M$.*

***Proof.*** By Corollary 3.4, $M$ is a $\mathfrak{S}$-distributive $R$-module. By using Lemma 2.1 (1),

$$N \wedge N' = N \cap N' = S_p(N) \cap S_p(L) = S_p(N \cap L) = S_p(0) = \mathbf{0}.$$

Moreover, $M = N + L \subseteq S_p(N) + S_p(L) \subseteq S_p(S_p(N) + S_p(L))$, which implies

$$N \vee N' = S_p(N + N') = S_p(S_p(N) + S_p(L)) = M.$$

Hence $\mathfrak{S}_p(M)$ is a Boolean algebra. □

From now on, $\mathfrak{S}_p(M)$ is assumed to be a Boolean algebra with the above assumptions.

**Corollary 3.8.** *For any semisimple ring $R$, $\mathfrak{S}_p(R)$ is a Boolean algebra with respect to any prime ideal $p$ of $R$.*

***Proof.*** Let $R$ be a semisimple ring and $p$ a prime ideal of $R$. By [12, Exercise 1.2.5] $R$ is an arithmetical ring. Thus by Theorem 3.7, $\mathfrak{S}_p(R)$ is a Boolean algebra. □

**Corollary 3.9.** *Let $R$ be a semisimple ring and $M$ be a distributive $R$-module. Then $\mathfrak{S}_p(M)$ is a Boolean ring with the following operations:*

$$L + N = S_p(L \cap S_p(\tilde{N}) + S_p(\tilde{L}) \cap N) \text{ and } L \cdot N = L \cap N,$$

*where $M = L \oplus \tilde{L} = N \oplus \tilde{N}$.*

***Proof.*** Follows from Theorem 3.7 and [5, Theorem IV.2.3]. □

**Corollary 3.10.** *Let $R$ be a semisimple ring, $p$ a prime ideal of $R$ and $M$ a multiplication $R$-module. Then $M$ is cyclic and the lattice $\mathfrak{S}_p(M)$ is a Boolean algebra.*

***Proof.*** Since $R$ is a semisimple ring, by [12, Corollary 2.6], $R$ is an Artinian ring. Hence $M$ is cyclic by [7, Corollary 2.9]. Also, by [12, Exercise 1.2.5], $R$ is an arithmetical ring. Thus by [8, Proposition 1.2], $M$ is a distributive $R$-module. Hence by Theorem 3.7, $\mathfrak{S}_p(M)$ is a Boolean algebra with respect to any prime ideal $p$ of $R$. □

**Theorem 3.11.** *Let $R$ be a ring, $p$ a prime ideal of $R$, $M$ an $R$-module and $N$ a submodule of $M$. Then the followings hold:*

(1) *For any submodule $L$ containing $N$, $S_p(L/N) = S_p(L)/N$. In particular, the assignment $L \mapsto L/N$ is a one to one corresponding between the set $\{L \mid L \in \mathfrak{S}_p(M), L \supseteq N\}$ and $\mathfrak{S}_p(M/N)$;*
(2) *If $M$ is a $\mathfrak{S}$-distributive lattice over $R$ with respect to $p$, then $M/N$ is $\mathfrak{S}$-distributive over $R$ with respect to $p$;*
(3) *If $R$ is a semisimple ring and $M$ a distributive $R$-module, then $\mathfrak{S}_p(M/N)$ is a Boolean algebra.*

***Proof.*** (1) Clear.
(2) Let $\mathfrak{S}_p(M)$ be a distributive lattice with the operations $\vee$ and $\wedge$ and $\mathfrak{S}_p(M/N)$ be a lattice with the operations $\tilde{\vee}$ and $\tilde{\wedge}$. It is seen that $\tilde{\vee}$ and $\tilde{\wedge}$ are expressed by $\vee$ and $\wedge$ respectively as follows:

$$\begin{aligned} L/N \;\tilde{\vee}\; K/N &= S_p(L/N + K/N) \\ &= S_p((L+K)/N) \\ &= S_p(L+K)/N \\ &= (L \vee K)/N, \end{aligned}$$

and

$$L/N \;\tilde{\wedge}\; K/N = L/N \cap K/N = (L \cap K)/N = (L \wedge K)/N.$$

By these statements, the distributivity of $\mathfrak{S}_p(M/N)$ follows immediately from the distributivity of $\mathfrak{S}_p(M)$.
(3) Follows from Theorem 3.7 and (2). $\square$

**Theorem 3.12.** *Let $R$ be a ring, $T$ a multiplicatively closed subset of $R$, $M$ an $R$-module and $N$ a submodule of $M$. Then the followings hold:*

1. *$S_{T^{-1}p}(T^{-1}N) = T^{-1}(S_p(N))$ for all prime ideals $p$ disjoint from $T$. In particular, $N \in \mathfrak{S}_p(M)$ if and only if $T^{-1}N \in \mathfrak{S}_{T^{-1}p}(T^{-1}M)$ for all prime ideals $p$ disjoint from $T$;*
2. *If $M$ is a $\mathfrak{S}$-distributive lattice over $R$ with respect to a prime ideal $p$ of $R$ such that $p \cap T = \emptyset$, then $T^{-1}M$ is $\mathfrak{S}$-distributive over $T^{-1}R$ with respect to $T^{-1}p$;*
3. *If $R$ is a semisimple ring, $p$ a prime ideal of $R$ with $p \cap T = \emptyset$ and $M$ a distributive $R$-module, then $\mathfrak{S}_{T^{-1}p}(T^{-1}M)$ is a Boolean algebra.*

***Proof.*** (1) Clear.
(2) Let $p$ be a prime ideal of $R$ such that $p \cap T = \emptyset$. Let $\mathfrak{S}_p(M)$ be a distributive lattice with the operations $\vee$ and $\wedge$ and $\mathfrak{S}_{T^{-1}p}(T^{-1}M)$ be a lattice with the operations $\tilde{\vee}$ and $\tilde{\wedge}$ . It is seen that $\tilde{\vee}$ and $\tilde{\wedge}$ are expressed by $\vee$ and $\wedge$ respectively as follows:

$$\begin{aligned} T^{-1}L \;\tilde{\vee}\; T^{-1}N &= S_{T^{-1}p}(T^{-1}L + T^{-1}N) \\ &= S_{T^{-1}p}(T^{-1}(L+N)) \\ &= T^{-1}(S_p(L+N)) \\ &= T^{-1}(L \vee N), \end{aligned}$$

and

$$\begin{aligned} T^{-1}L \;\tilde{\wedge}\; T^{-1}N &= T^{-1}L \cap T^{-1}N \\ &= T^{-1}(L \cap N) \\ &= T^{-1}(L \wedge N). \end{aligned}$$

By these statements, the distributivity of $\mathfrak{S}_{T^{-1}p}(T^{-1}M)$ follows immediately from the distributivity of $\mathfrak{S}_p(M)$.
(3) Since $R$ is a semisimple ring, then so is $T^{-1}R$. Thus the result follows from Theorem 3.7 and (2). $\square$

Let $A$ and $B$ be Boolean algebras. A function $f : A \to B$ is called a *Boolean algebra homomorphism*, if $f$ is a lattice homomorphism, $f(\mathbf{0}) = \mathbf{0}$, $f(\mathbf{1}) = \mathbf{1}$ and $f(a') = f(a)'$ for all $a \in A$. It is easily proved that a lattice homomorphism $f$ preserves $\mathbf{0}$ and $\mathbf{1}$ if and only if it preserves $'$. Thus, in order to show that a function $f$ between two Boolean algebras is a Boolean algebra homomorphism, it suffices to check that $f$ preserves lattice operations $\vee$ and $\wedge$ and constants $\mathbf{0}, \mathbf{1}$.

**Theorem 3.13.** *Let $R$ be a semisimple ring, $p$ a prime ideal of $R$ and $M$ a distributive $R$-module. Then $\eta : \mathfrak{S}_p(R) \to \mathfrak{S}_p(M)$ is a Boolean algebra homomorphism.*

***Proof.*** First note that $\mathfrak{S}_p(M)$ and $\mathfrak{S}_p(R)$ are Boolean algebras, by Theorem 3.7 and Corollary 3.8 respectively. By Corollary 2.7, $\eta$ is a lattice homomorphism. Also,

$$\eta(\mathbf{0}) = \eta(S_p(0)) = S_p(S_p(0)M) = S_p(0) = \mathbf{0},$$

and

$$\eta(\mathbf{1}) = \eta(R) = S_p(RM) = S_p(M) = M = \mathbf{1}.$$

Hence, as noted above, $\eta$ is a Boolean algebra homomorphism. $\square$

**Corollary 3.14.** *Let $R$ be a semisimple ring, $p$ a prime ideal of $R$ and $M$ a multiplication $R$-module. Then $\eta : \mathfrak{S}_p(R) \to \mathfrak{S}_p(M)$ is a Boolean algebra epimorphism.*

***Proof.*** By Corollaries 3.8 and 3.10, $\mathfrak{S}_p(R)$ and $\mathfrak{S}_p(M)$ are Boolean algebras respectively. Also, by the proof of Corollary 3.10, $M$ is distributive. Thus by Theorem 3.13, $\eta$ is a Boolean algebra homomorphism. Moreover, if $N \in \mathfrak{S}_p(M)$, then $(N : M) \in \mathfrak{S}_p(R)$ and

$$\eta(N : M) = S_p((N : M)M) = S_p(N) = N.$$

Thus, $\eta$ is an epimorphism. □

Finally, we remark that if $M$ is a faithful multiplication module over a semisimple ring $R$, then since $M$ is cyclic by Corollary 3.10, we conclude that $M$ is isomorphic to $R$. So it clearly follows that $\eta$ and $\theta$ are Boolean algebra isomorphisms.

**Acknowledgment.** The authors would like to thank the referee for his/her helpful comments.

## References

[1] M. Alkan and Y. Tiras, *On invertible and dense submodules*, Comm. Algebra, **32** (10), 3911–3919, 2004.
[2] M. Alkan and Y. Tiras, On prime submodules, Rocky Mountain J. Math. **37** (3), 709–722, 2007.
[3] M.F. Atiyah and I.G. Macdonald, *Introduction to Commutative Algebra*, Addison-Wesley, London, 1969.
[4] A. Barnard, *Multiplication modules*, J. Algebra, **71** (1), 174–178, 1981.
[5] S. Burris and H.P. Sankappanavar, *A Course in Universal Algebra*, Springer-Verlag, New York, 1981.
[6] J. Dauns, *Prime submodules*, J. Reine Angew. Math. **298**, 156–181, 1978.
[7] Z.A. El-Bast and P.F. Smith, *Multiplication modules*, Comm. Algebra, **16** (4), 755–799, 1988.
[8] V. Erdogdu, *Multiplication modules which are distributive*, J. Pure Appl. Algebra, **54**, 209–213, 1988.
[9] J.B. Harehdashti and H.F. Moghimi, *Complete homomorphisms between the lattices of radical submodules*, Math. Rep. **20(70)** (2), 187–200, 2018.
[10] T.W. Hungerford, *Algebra*, Springer-Verlag, New York, 1974.
[11] J. Jenkins and P.F. Smith, *On the prime radical of a module over a commutative ring*, Comm. Algebra, **20** (12), 3593–3602, 1992.
[12] T.Y. Lam, *A First Course in Noncommutative Rings*, Springer-Verlag, New York, 1991.
[13] M.D. Larsen and P.J. McCarthy, *Multiplicative Theory of Ideals*, Academic Press, New York, 1971.
[14] C.P. Lu, *M-radical of submodules in modules.* Math. Japonica, **34** (2), 211–219, 1989.
[15] C.P. Lu, *Saturations of submodules*, Comm. Algebra, **31** (6), 2655–2673, 2003.
[16] C.P. Lu, *A module whose prime spectrum has the surjective natural map*, Houston J. Math. **33** (1), 125–143, 2007.
[17] R.L. McCasland and M.E. Moore, *On radicals of submodules*, Comm. Algebra, **19** (5), 1327–1341, 1991.
[18] R.L. McCasland and M.E. Moore, *Prime submodules*, Comm. Algebra, **20** (6), 1803–1817, 1992.
[19] R.L. McCasland, M.E. Moore and P.F. Smith, *On the spectrum of a Module over a commutative ring*, Comm. Algebra, **25** (1), 79–103, 1997.
[20] H.F. Moghimi and J.B. Harehdashti, *Mappings between lattices of radical submodules*, Int. Electron. J. Algebra, **19**, 35–48, 2016.
[21] P.F. Smith, *Some remarks on multiplication modules*, Arch. Math. **50**, 223–235, 1988.

[22] P.F. Smith, *Mappings between module lattices*, Int. Electron. J. Algebra, **15**, 173–195, 2014.
[23] P.F. Smith, *Complete homomorphisms between module lattices*, Int. Electron. J. Algebra, **16**, 16–31, 2014.
[24] P.F. Smith, *Anti-homomorphisms between module lattices*, J. Commut. Algebra, **7**, 567–591, 2015.

Hacettepe Journal of Mathematics & Statistics

Hacet. J. Math. Stat.
Volume 50 (1) (2021), 255 – 274
DOI : 10.15672/hujms.671806

Research Article

# Modeling under or over-dispersed binomial count data by using extended Altham distribution families

Senay Asma

*McMaster University, 1280 Main St W, Hamilton, Ontario, Canada*

## Abstract

While aiming particularly at handling under-dispersion, we explore a type of models constructed conservatively using the minimum information of first two moments for the fitting of binomial count data, which could have under, equal or over-dispersion. The extended Altham distribution (EAD) families were presented in this study. The extended Altham families are very close to the binomial distribution under equal dispersion setting, implying that they are alternative models of the binomial distribution. The feature that extended Altham families can reach the full range of dispersion outperforms some commonly used models such as extended beta-binomial and quasi-binomial which have restricted ranges of dispersion. Moreover, the extended Altham family can have double peaks at two boundaries, indicating they are feasible for fitting the double tail inflation phenomenon. This study illustrated the modeling using extended Altham families for both under-dispersed and over-dispersed binomial data resulted from disease cases within the same family.

**Mathematics Subject Classification (2020).** 62H10

**Keywords.** Binomial count data, Kullback-Leibler, exponential family, binomial distribution, dispersion index

## 1. Introduction

Binomial count data, a type of count data with bounded supports, arise from many disciplines such as toxicological study, medical research, ecology, agriculture, logistics management, linguistics, electronic engineering, political science, and so on. This type of data are often associated with an important quantity called proportion which is the study purpose. For the binomial count data, the most commonly used model is binomial distribution. The binomial random variable (rv) is the sum of independent and identically distributed (iid) Bernoulli rv's which have a fixed success probability for value 1. This success probability is the interested population proportion.

However, the above binomial setting is too ideal and simple. In reality, there could exist more complicated situations. For example, the success probability may be a rv instead of a fixed constant, or the Bernoulli rv's may positively or negative correlated (corresponding to attraction or repulsion). The data could even result from an aggregation of subsets

Email address: asmas@mcmaster.ca
Received: 07.07.2020; Accepted: 17.10.2020

with varying upper bounds. Thus, observations could appear to be over-dispersion or under-dispersion relative to the binomial distribution.

Handling over-dispersion has received a great deal of attention and is quite mature. The common way is to us binomial mixture. Allowing varying success probabilities in the binomial distribution can yield a mixture with over-dispersion relative to the binomial distribution. A widely used model is the beta-binomial which is the binomial mixture of the beta distribution, i.e., the success probability follows a beta distribution. Refer to Wilcox [18] for a review of beta-binomial and its extensions.

However, in reality, the under-dispersion can occur, especially in the repulsion situation. Bailey [4] reported the repulsion examples of function word counts, which are under-dispersed relative to binomial due to the nature that a function word can not follow itself in general. Assuming negative correlation among Bernoulli rv's, the sum of them will result in a distribution of under-dispersion relative to the binomial. See Theorem 7.1 in Joe [12]. Viveros-Aguilera, Balasubramanian and Balakrishnan [17] constructed a concrete example using the homogeneous Markov chain for binary response. In addition, quasi-binomial and its variations prescribe non-homogeneous dependence mechanisms for successive trials by Chakraborty and Das [5]. We show another possibility leading to under-dispersion in Section 2, which is a mixture of varying upper bounds of supports.

Prentice [15] extended the beta-binomial to allow limited under-dispersion. Consul [8] proposed the quasi-binomial (type I) using an urn model, in which the success probability of the $i$-th trial has an additional part proportional to $i$ ($i > 1$). This additional part in the success probability can be negative or positive, resulting in under-dispersion or over-dispersion, but both under- and over-dispersions are bounded. Some extensions of quasi-binomial can be found in Mishra, Tiwary and Singh [14], Dobson, Carreras and Newman [9], Chakraborty and Das [5] and some advanced studies in Altham [2,3]. Other models using particular mechanisms like Bailey [4] were practised in the literature too. Although there are many attempts to handle the under-dispersion case, none of them becomes a mature tool for a general case.

The descriptive statistics are not always as easy as might be expected, particularly when data exhibit skewness and/or outliers. A relevant example is given by Chatfield [6] which involves the number of issues of a particular monthly magazine read by 20 people in a year. In this example, the data has bimodal U-shape which is even more difficult to summarize than a skewed distribution. Therefore, the sample mean and standard deviation are potentially very misleading. The proportion of regular readers is a useful statistic, but it may be sensible to describe the data in words rather than with summary statistics.

Since binomial count data can arise from complex situations, none of existing models provides a unified way to handle them. Thus, there is a need to develop a unified model capable of handling various dispersion situations. To this end, we construct models with specified mean and variance using the entropy method. The resulted two-parameter models can reach the full range of dispersion, providing a unified way for modelling binomial count data with different dispersion case.Also, numerical comparison shows that the proposed models are quite close to the binomial distribution in the equal-dispersion setting. Hence, they are alternative to the binomial model in the equal-dispersion setting.

In summary, the proposed two-parameter models have the ability to better fit various binomial count data in a unified way. Based on our proposed models, we have found that the Altham distribution [1] is a special case by reparametrization. Thus, this finding uncovers the feature of full dispersion of the Altham distribution. To credit Altham, the models we proposed are named as the extended Altham distribution families (EAD).

The remainder of this paper is organized as follows. We define new exponential families in Section 2, with the computational algorithm for the probability mass function (pmf).

MLEs are derived in Section 3. We conduct simulation study and illustrate data examples in Section 4. A brief discussion is given in Section 5.

## 2. Model construction

In this section, we shall present the construction of extended Altham distribution families by using Kullback-Leibler (KL) divergence measure. KL is non-symmetric measure defined by

$$KL(p_i||q_i) = \sum_{i=0}^{M} p_i \log(\frac{p_i}{q_i}) \tag{2.1}$$

and it gives the distance between two probability distributions, P and Q, where Q is given distribution and P is unknown probability distribution. The distribution Q is known as a priori distribution.

For example, if a priori distribution Q is considered as a discrete uniform distribution assigns equal probability $1/(M+1)$ to every point in the support, then the closest distribution in sense of KL measure will be the distribution that has the maximum uncertainty in the support, leading to the maximum entropy. For a discrete distribution, denote the probability mass function (pmf) as $\Pr[X = i] = p_i$, $(i = 0, 1, \ldots, M)$, the mean as $\mu$ and variance as $\sigma^2$. Given the information of a priori distribution $Q$, mean and variance, KL optimization defines the distribution which obtain the probability distribution which satisfy minimum KL distance. Encouraging probability assignment in the support as even as possible, thus, taking advantage of given information in a minimum and conservative sense. That is

$$\min \left\{ \sum_{i=0}^{M} p_i \log(\frac{p_i}{q_i}) \right\}, \tag{2.2}$$

subject to three constrains

$$\sum_{i=0}^{M} p_i = 1, \qquad \sum_{i=0}^{M} i p_i = \mu, \qquad \sum_{i=0}^{M} i^2 p_i = \sigma^2 + \mu^2. \tag{2.3}$$

There is no explicit form of pmf in terms of parameters $\mu$ and $\sigma^2$, however, there is an explicit form in terms of Lagrangian multipliers $\beta$'s:

$$p_i = q_i C(\beta_1, \beta_2) e^{i\beta_1 + i^2\beta_2}, i = 0, 1, \ldots, M, \tag{2.4}$$

where $C(\beta_1, \beta_2)$ is the normalizing constant.

Note that if Q is considered as a binomial distribution, the pmf will be Altham distribution [1]. Thus, we call this family as extended Altham distribution family. In the following, we give a formal definition.

**Definition 2.1. (extended Altham distribution family):** A rv $X$ is said to be from the extended Altham distribution family, denoted as extended Altham$(M, h, \beta_1, \beta_2)$ where $-\infty < \beta_1, \beta_2 < \infty$, if its probability mass function (pmf) is of form:

$$p_i \propto h_i \exp(\beta_1 i + \beta_2 i^2), \qquad i = 0, 1, \ldots, M, \tag{2.5}$$

where $h_i$ is an arbitrary function with positive values and $\beta_1$ and $\beta_2$ are real parameters and satisfy

$$\sum_{i=0}^{M} \Pr[X = i] = \sum_{i=0}^{M} h_i C(\beta_1, \beta_2) \exp(i\beta_1 + i^2\beta_2) = 1, \tag{2.6}$$

$$E[X] = \sum_{i=0}^{M} i \Pr[X = i] = \sum_{i=0}^{M} ih_i C(\beta_1, \beta_2) \exp(i\beta_1 + i^2\beta_2) = \mu, \tag{2.7}$$

$$E[X^2] = \sum_{i=0}^{M} i^2 \Pr[X = i] = \sum_{i=0}^{M} i^2 h_i C(\beta_1, \beta_2) \exp(i\beta_1 + i^2\beta_2) = \sigma^2 + \mu^2, \tag{2.8}$$

where $C(\beta_1, \beta_2)$ is the normalizing constant.

$\beta_1$ and $\beta_2$ seem to govern the increasing or decreasing speed of pmf, but no direct connection with the mean and variance. The parametrization in terms of $\mu$ and $\sigma^2$ has clear explanation, however, no analytical pmf available. But this can be compensated by numerical solution.

Since constrain (2.7) implies

$$C^{-1}(\beta_1, \beta_2) = \sum_{i=0}^{M} h_i e^{i\beta_1 + i^2\beta_2} \tag{2.9}$$

hence, there are only two independent parameters: $\beta_1$ and $\beta_2$. For any discrete distribution on the support $\{0, 1, \ldots, M\}$, since

$$\mu = \sum_{i=1}^{M} ip_i = E[1 \times X] \le E[X^2] \le E[M \times X] = M \sum_{i=1}^{M} ip_i = M\mu, \tag{2.10}$$

the natural ranges of $\mu$ and $\sigma^2$ are

$$0 \le \mu \le M, \qquad \max(0, \mu - \mu^2) \le \sigma^2 = E[X^2] - \mu^2 \le M\mu - \mu^2. \tag{2.11}$$

There is no restriction for parameters $\mu$ and $\sigma^2$, thus, these two parameters can vary in their full ranges shown in (2.14). However, the ranges of $\beta_1$ and $\beta_2$ can not be determined in explicit forms.

When $M = 1$, the rv $X$ degenerates to the Bernoulli case, and only one parameter is needed. Thus, we exclude this extreme case for the upper bound of the support, and only consider $M \ge 2$.

When $M = i$, the pmf can be expressed in terms of $h_i$ and $\beta = (\beta_1, \beta_2)$:

$$p_i = log(h_i) + \beta[(i + 1)^\alpha - i^\alpha], \tag{2.12}$$

where $\alpha > 0$ and $h_i$ are arbitrary positive valued function.

The extended Altham distribution has only two independent parameters: $\beta_1$ and $\beta_2$.

For any discrete distribution on the support $\{0, 1, \ldots, M\}$, since

$$\mu = \sum_{i=1}^{M} ip_i = E[1 \times X] \le E[X^2] \le E[M \times X] = M \sum_{i=1}^{M} ip_i = M\mu, \tag{2.13}$$

the natural ranges of $\mu$ and $\sigma^2$ are

$$0 \le \mu \le M, \qquad \max(0, \mu - \mu^2) \le \sigma^2 = E[X^2] - \mu^2 \le M\mu - \mu^2. \tag{2.14}$$

There is no restriction for parameters $\mu$ and $\sigma^2$, thus, these two parameters can vary in their full ranges shown in (2.14). However, the ranges of $\beta_1$ and $\beta_2$ can not be determined in explicit forms.

The binomial distribution is usually referred as the equally-dispersed distribution. Assume $Y \sim binomial(M, p)$ which has pmf

$$p_i = \binom{M}{i} p^i (1 - p)^{M-i}, \qquad 0 \le p \le 1, \quad i = 0, 1, \ldots, M. \tag{2.15}$$

Then $E[Y] = Mp$ and $Var[Y] = Mp(1-p)$. The ratio of variance to mean is $\frac{Var[Y]}{E[Y]} = 1-p = 1-\frac{E[Y]}{M}$. A discrete distribution on the same support is said to be under-dispersed or over-dispersed if its ratio is smaller or bigger than that of the binomial distribution of the same mean. That is, the comparison is regarded to the binomial distribution of the same mean.

For convenience, we define the dispersion index for discrete distribution on the support $\{0, 1, \ldots, M\}$ as follows

$$D = \frac{Var[Y]}{E[Y]\,(1 - E[Y]/M)}. \tag{2.16}$$

Then, a discrete distribution on the support $\{0, 1, \ldots, M\}$ is said to be under-dispersed, equally-dispersed or over-dispersed if its dispersion index defined in (2.16) is smaller than, equal to or bigger than 1 respectively. Obviously, the binomial distribution is equally-dispersed. However, other distributions can be equally-dispersed too.

According to (2.14), the full range of dispersion is

$$\max\left(0, \frac{1-\mu}{1-\mu/M}\right) = \frac{\max(0, \mu(1-\mu))}{\mu(1-\mu/M)} \le D \le \frac{M\mu - \mu^2}{\mu(1-\mu/M)} = M. \tag{2.17}$$

Note that the lower bound is $\frac{1-\mu}{1-\mu/M} > 0$ when $0 \le \mu < 1$, and 0 otherwise. When $M$ is large, the interval $(0, 1)$ for under-dispersion is very narrow comparing with the interval $(1, M)$ for over-dispersion, one might uses $\log(D)$ as the dispersion index. But to keep consistent with the convention, we use (2.16).

The over-dispersion is usually explained by a mixture of binomial, say the beta-binomial. We have found that under-dispersion could be caused by a mixture too, but of varying upper bounds of supports. Here we illustrate using a simple example of two-component binomial mixture.

Let $X_1 \sim \text{binomial}(M_1, p_1)$ and $X_2 \sim \text{binomial}(M_2, p_2)$, where $M_1 < M_2$. Assume $E[X_1] = E[X_2] = \mu < M_1$. Denote $I \sim \text{Bernoulli}(p)$, and define $Y$ conditional on $I$ as follows

$$[Y|I=1] \;\sim\; \text{binomial}(M_1, p_1), \qquad [Y|I=0] \;\sim\; \text{binomial}(M_2, p_2). \tag{2.18}$$

Note that the support of $Y$ is $\{0, 1, \ldots, M_2\}$. Then

$$E[Y] = E\{E[Y|I]\} = pE[X_1] + (1-p)E[X_2] = \mu, \tag{2.19}$$

$$\begin{aligned} Var[Y] &= E[(Y-\mu)^2] = E\{E[(Y-\mu)^2|I]\} \\ &= pVar[X_1] + (1-p)Var[X_2] \\ &= p\mu(1-\mu/M_1) + (1-p)\mu(1-\mu/M_1) \\ &= \mu\{1 - [p\mu/M_1 + (1-p)\mu/M_2]\} \\ &< \mu(1-\mu/M_2). \end{aligned} \tag{2.20}$$

Thus

$$D = \frac{\mu\{1 - [p\mu/M_1 + (1-p)\mu/M_2]\}}{\mu(1-\mu/M_2)} < \frac{\mu(1-\mu/M_2)}{\mu(1-\mu/M_2)} = 1, \tag{2.21}$$

implying that $Y$ is under-dispersed.

In order to illustrate the extended Altham distribution family, we considered the following models with different $h_i$ functions:

$$\text{Model 1.} \quad h_i = 1; \quad \textit{flat (Discrete Uniform)} \tag{2.22}$$

$$\text{Model 2.} \quad h_i = log(M - X + 1) + 1; \quad \textit{decreasing} \tag{2.23}$$

$$\text{Model 3.} \quad h_i = \frac{M!}{(M - X)!(X)!}; \quad \textit{convex (Weighted Binomial)} \tag{2.24}$$

$$\text{Model 4.} \quad h_i = \frac{(M - X)!(X)!}{M!}; \quad \textit{concave} \tag{2.25}$$

$$\text{Model 5.} \quad h_i = X + 1; \quad \textit{increasing} \tag{2.26}$$

$$\text{Model 6.} \quad h_i = \frac{1}{(X + 1)}; \quad \textit{decreasing} \tag{2.27}$$

$$\text{Model 7.} \quad h_i = M - X + 1; \quad \textit{decreasing} \tag{2.28}$$

$$\text{Model 8.} \quad h_i = \frac{1}{(M - X + 1)}; \quad \textit{increasing} \tag{2.29}$$

$$\text{Model 9.} \quad h_i = X(M - X) + 1; \quad \textit{convex} \tag{2.30}$$

$$\text{Model 10.} \quad h_i = \frac{1}{X(M - X) + 1}; \quad \textit{concave} \tag{2.31}$$

$$\text{Model 11.} \quad h_i = log(X + 1) + 1; \quad \textit{increasing} \tag{2.32}$$

$$\text{Model 12.} \quad h_i = \frac{1}{log(X + 1) + 1}; \quad \textit{decreasing.} \tag{2.33}$$

The dispersion index for extended Altham$(\mu, \sigma^2)$ is $D = \frac{\sigma^2}{\mu(1-\mu/M)}$, which can reach the full range of dispersion because of no restriction on parameters $\mu$ and $\sigma^2$. Since $\sigma^2$ is independent of $\mu$, $D$ could be smaller than, equal to or bigger than 1. Therefore, the extended Altham family covers all dispersion situations. The extended Altham distribution family given by 2.5 includes Binomial distribution when the function $h_i = 1$,

Weighted Binomial distributions Zelterman [19] when the function $h_i$ is the binomial coefficient and so Altham distribution [1] because it is known to be an example of a weighted binomial model.

For comparison purposes, we need reparametrization so that we can fix $(\mu, \sigma^2)$. Figure 1 and 2 displays the pmf profiles of the extended Altham distributions with $h_i$ functions given by 2.22-2.33, mean $\mu = 5$ and various dispersions using the developed numerical algorithm. Comparing with Binomial distribution (red line), the under-dispersed extended Altham distributions (green lines) seem to have larger probability masses around the mean, while the over-dispersed extended Altham distributions (blue lines) attempt to have more masses at two boundaries. When the dispersion large enough, the pmf shows U-shape, like that of the beta-binomial distributions.

Since the extended Altham distribution can have equal dispersion, it is natural to compare it with the binomial distributions under the same means.

Figure 3 and 4 demonstrate some of them on the support $\{0, 1, \ldots, 40\}$. We see that both pmf's are very close when the mean is not close to the two boundaries. When the mean close to two boundaries, there are slight differences among two distributions, and the extended Altham distribution assigns more masses at 0 or $M$. For many values of $M$, we check the maximum absolute difference of pmf of two distributions under the same mean, and find that this maximum is no more than 3% when the mean close to boundaries, and becomes smaller when the mean close to the center of the support. The larger the $M$, the smaller the maximum of probability difference. From the viewpoint of distribution theory, this suggests that the binomial distribution can be approximated by the extended Altham distribution. On the other hand, for the distribution constructed using the minimum information of mean and equal-dispersion, the binomial distribution is

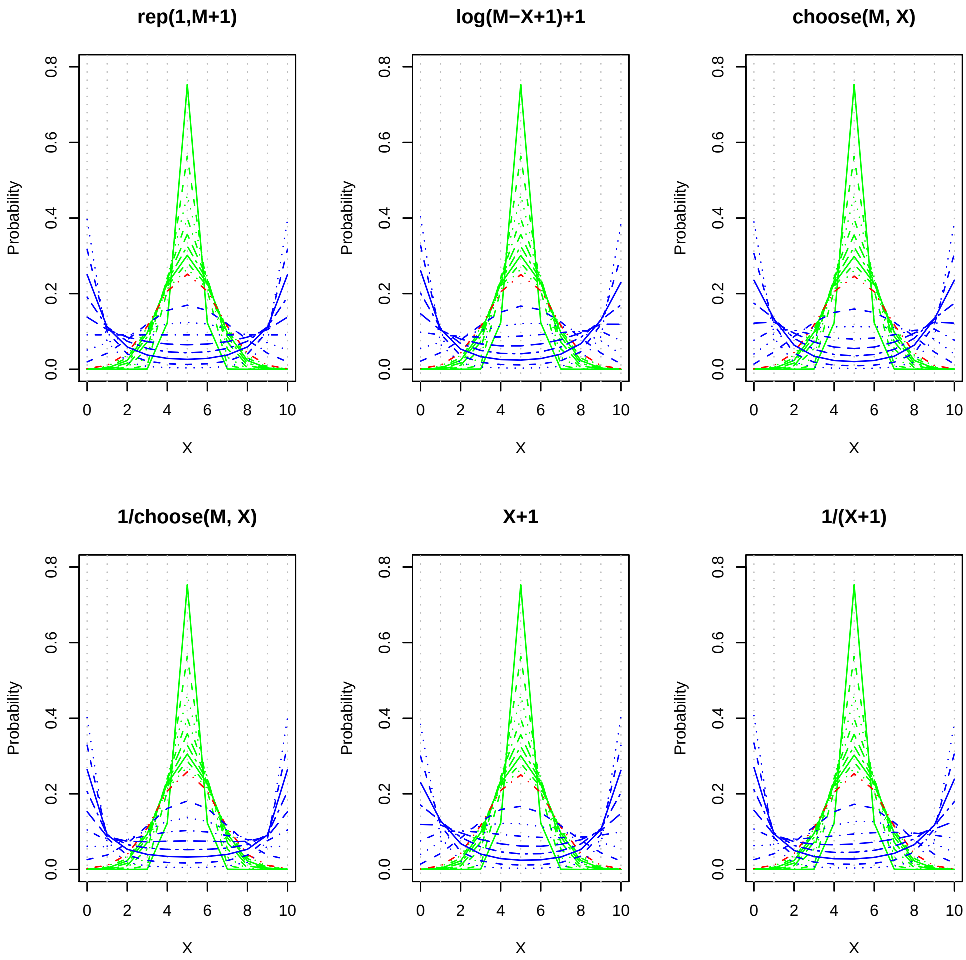



**Figure 1.** Probability profiles of the extended Altham distributions of mean $\mu = 5$ and various dispersions regarding to $h_i$ given by 2.22 and 2.27 The red line indicates the equal dispersion. The blue lines correspond to over dispersions of $2, 3, \ldots, 9$, while the green lines shows under dispersions of $0.1, 0.2, \ldots, 0.9$. The most centered extended Altham distributions with the largest mass at 5 has dispersion 0.1, and the most spread extended Altham distributions with largest masses at two boundaries has dispersion 9.

very close to it. Thus, from the aspect of modelling, such a fact implies that the extended Altham distribution could be an alternative of the binomial distribution if the mean is not extremely small or large.

Note that the extended beta-binomial and quasi-binomial can handle both under-dispersion and over-dispersion too. The beta-binomial distribution is constructed using mixture. Assume the success probability in binomial distribution $p \sim \text{beta}(a, b)$ $(a > 0, b > 0)$, the pmf of beta-binomial$(M, a, b)$ is

$$p_i = \binom{M}{i} \frac{B(a+i, b+M-i)}{B(a,b)}, \qquad i = 0, 1, \ldots, M, \tag{2.34}$$

where $B(x, y)$ is the complete beta function. See Hasemann and Kupper [11].

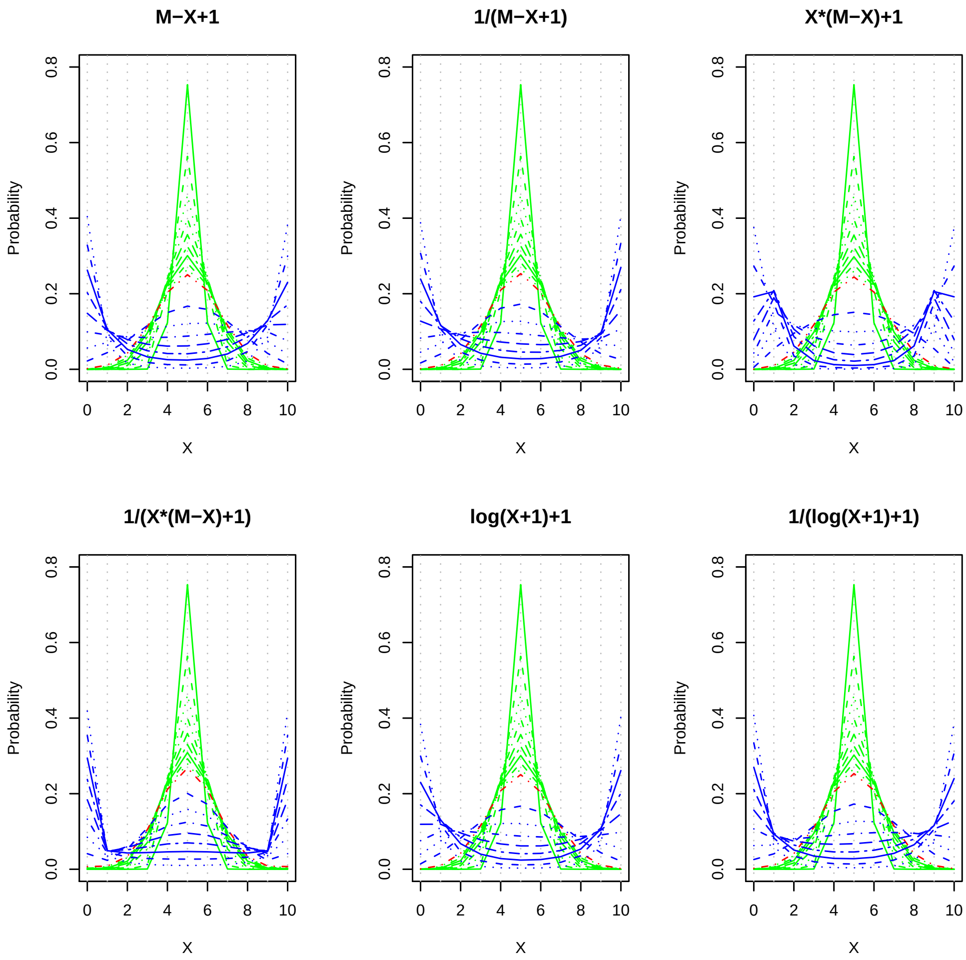



**Figure 2.** Probability profiles of the extended Altham distributions of mean $\mu = 5$ and various dispersions regarding to $h_i$ given by 2.28 and 2.33. The red line indicates the equal dispersion. The blue lines correspond to over dispersions of $2, 3, \ldots, 9$, while the green lines shows under dispersions of $0.1, 0.2, \ldots, 0.9$. The most centered extended Altham distributions with the largest mass at 5 has dispersion 0.1, and the most spread extended Altham distributions with largest masses at two boundaries has dispersion 9.

The mean and variance are

$$E[X] = \frac{Ma}{a+b}, \qquad Var[X] = \frac{Mab(a+b+M)}{(a+b)^2(a+b+1)}, \tag{2.35}$$

and it is over-dispersed. Prentice [15] extended the beta-binomial, denoted as EBB$(M; p, \delta)$, using the following reparametrized pmf form

$$p_i = \binom{M}{i} \prod_{j=0}^{i-1}(p+\gamma j) \prod_{j=0}^{M-i-1}(1-p+\gamma j) \Big/ \prod_{j=0}^{M-1}(1+\gamma j), \qquad i = 0, 1, \ldots, M, \tag{2.36}$$

where $0 \le p \le 1$, $\gamma = \frac{\delta}{1-\delta}$ and

$$\delta = \gamma(1+\gamma)^{-1} \ge \max\left(\frac{-p}{M-p-1}, \; \frac{-q}{M-q-1}\right), \quad q = 1-p. \tag{2.37}$$

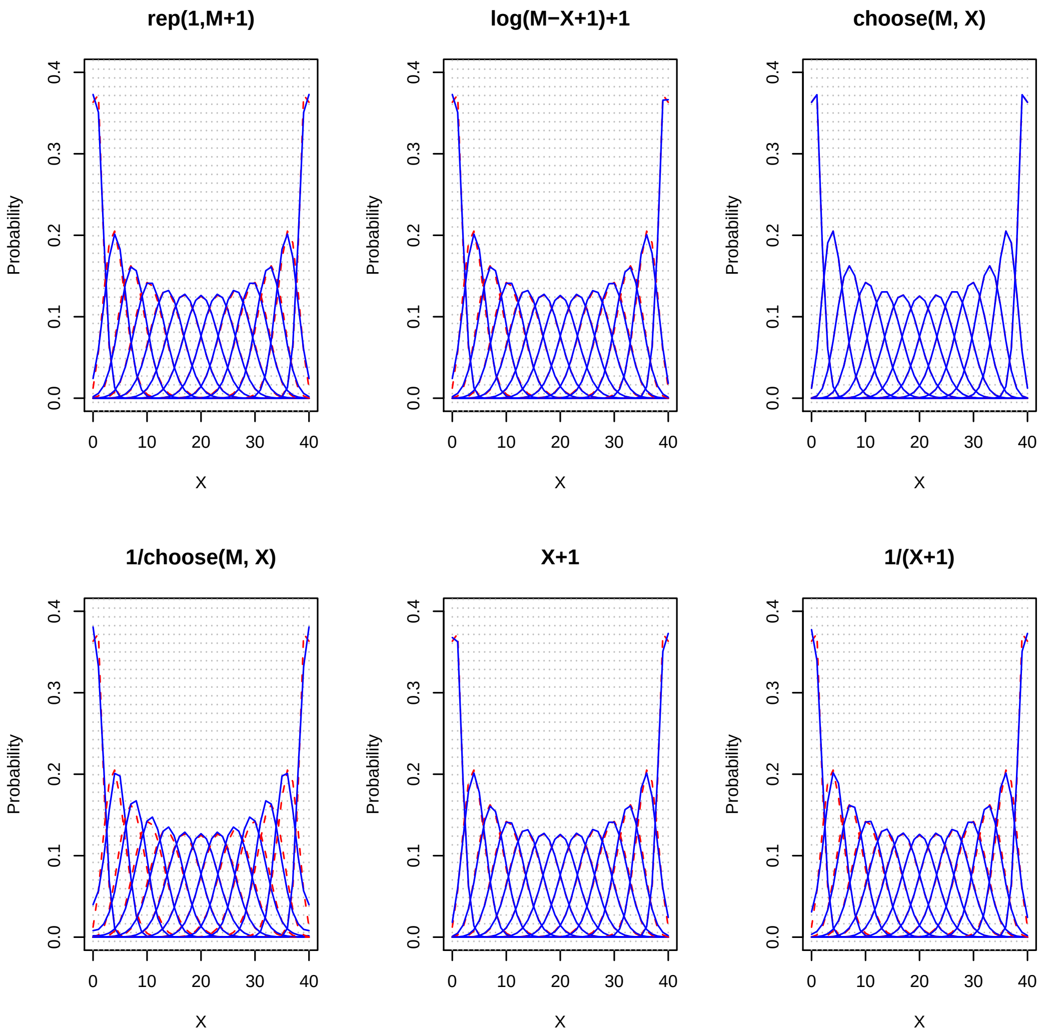



**Figure 3.** Comparison of probability profiles between extended Altham distributions with $h_i$ given by 2.22 - 2.27 and binomial distributions under the same means. The blue lines indicate the extended Altham distributions, while red lines correspond to binomial distributions. Any close pair of the extended Altham and binomial distributions has the same mean.

The mean and variance are

$$E[X] = Mp, \qquad Var[X] = Mp(1-p)[1 + (M-1)\delta]. \tag{2.38}$$

The extended beta-binomial allows under-dispersion, but bounded when $\delta$ reaches it lower bound. For example, if $M = 10$ and $p = 0.5$, then the lower bound of $\delta$ is $-1/17$, and the lower bound of dispersion is approximately $D = 0.4706$.

Consul [8] proposed the quasi-binomial distribution, later termed as type I QBD$(M; p, \phi)$, with pmf

$$p_i = \binom{M}{i} p(p + i\phi)^{i-1}(1 - p - i\phi)^{M-i}, \qquad i = 0, 1, \ldots, M, \tag{2.39}$$

where $0 \le p \le 1$ and $-p/M < \phi < (1-p)/M$.

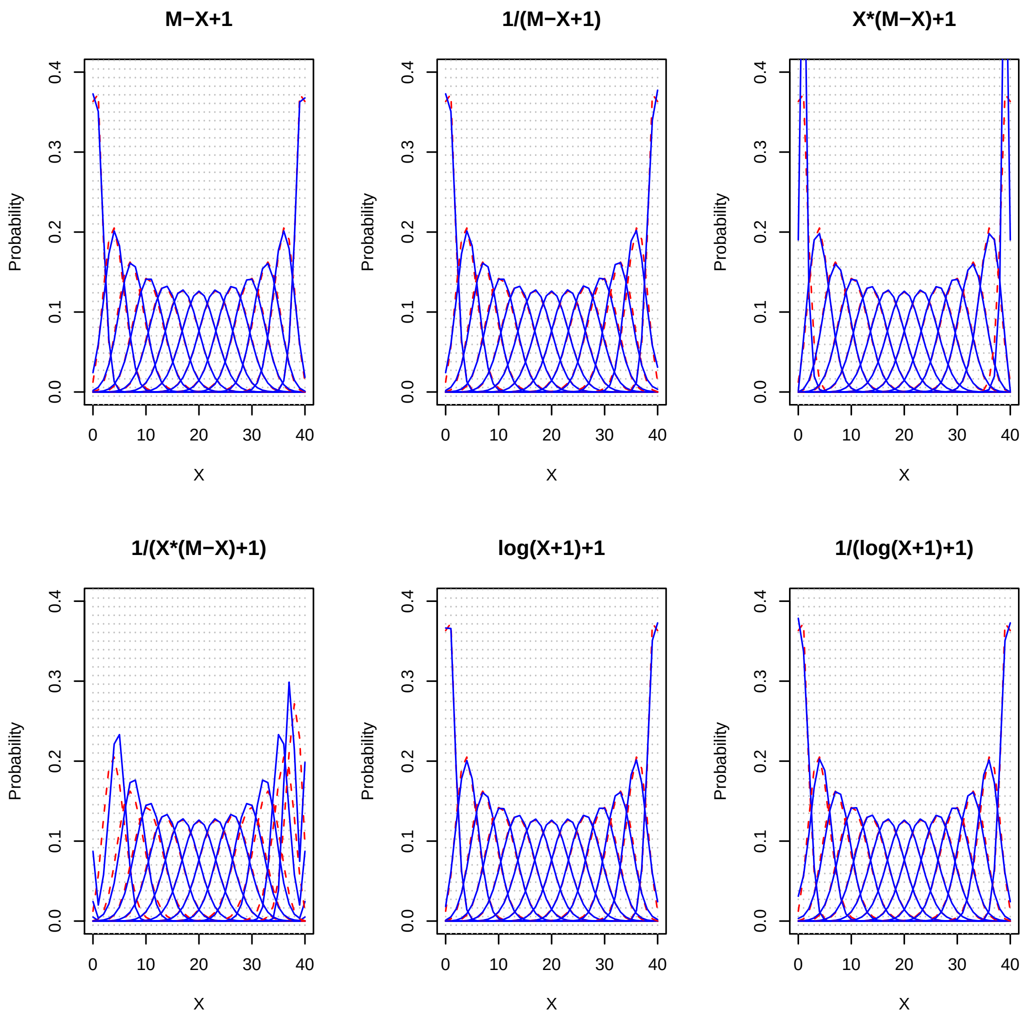



**Figure 4.** Comparison of probability profiles between extended Altham distributions with $h_i$ given by 2.28 - 2.33 and binomial distributions under the same means. The blue lines indicate the extended Altham distributions, while red lines correspond to binomial distributions. Any close pair of the extended Altham and binomial distributions has the same mean.

As pointed by Mishra, Tiwary and Singh [14], the most unfortunate result of this distribution (and other types QBD) is that the moments are series which are not possible to be summed. When $\phi \neq 0$, the probability of success in the $i$-th trial becomes $p + i\phi$. Positive or negative $\phi$ indicates attraction or repulsion of a trial to previous trials. This quasi-binomial distribution has lower bound for the under-dispersion and upper bound for the over-dispersion when $\phi$ reaches its lower and upper bounds respectively. For example, let $M = 10$ and $p = 0.5$. The lower and upper bounds of $\phi$ will be $-0.05$ and 0.05 respectively, and the lower and upper bounds of dispersion $D$ will be approximately 0.4518 and 3.1847 respectively.

The range of dispersion for both extended beta-binomial and quasi-binomial distributions can be numerically displayed. However, both can not cover the full range of dispersion like the extended Altham. Since the extended beta-binomial distribution can be reparametrized in terms of mean and variance analytically, we make numerical comparison of pmf under the same mean and dispersion between this distribution and the extended

Altham distribution, and find that they are different, matching the fact that they are constructed from different angles.

## 3. Comparison and statistical inference

The pmf (2.5) is explicit in $(\beta_1, \beta_2)$ and is implicit in $(\mu, \sigma^2)$. So, for MLE, we can solve it either by parametrization $(\beta_1, \beta_2)$ or $(\mu, \sigma^2)$. Since extended Altham distribution is a member of general exponential family, the MLEs for $(\beta_1, \beta_2)$ can be obtained by using the form given by

$$p(x|\theta) = h(x)c(\theta)e^{\sum_{i=1}^{k} w_i(\theta)t_i(x)}. \tag{3.1}$$

Then, the Log-likelihood function is,

$$L(\theta) = \sum_{j=1}^{N} log[h(x_j)c(\theta)e^{\sum_{i=1}^{k} w_i(\theta)t_i(x_j)}] \tag{3.2}$$

and the corresponding derivative is

$$\frac{\partial L(\theta)}{\partial \theta} = N\frac{c'(\theta)}{c(\theta)} + \sum_{j}^{N}\sum_{i=1}^{k} w_i(\theta)t_i(x_j). \tag{3.3}$$

Since $p(x|\theta)$ is a probability distribution, we can write

$$\int p(x|\theta) = \int h(x)c(\theta)e^{\sum_{i=1}^{k} w_i(\theta)t_i(x)}dx = 1 \tag{3.4}$$

and we can get

$$c(\theta) = \frac{1}{\int h(x)e^{\sum_{i=1}^{k} w_i(\theta)t_i(x)}dx} \tag{3.5}$$

$$c'(\theta) = -c(\theta)E[\sum_{i}^{k} \frac{\partial w_i(\theta)}{\partial \theta}t_i(x)] \tag{3.6}$$

If $c'(\theta)$ is replaced in the derivative of the log-likelihood function,

$$-NE[\sum_{i=1}^{k} \frac{\partial w_i(\theta)}{\partial \theta}t_i(x)] + \sum_{j}^{N}\sum_{i=1}^{k} w_i(\theta)t_i(x_j) = 0. \tag{3.7}$$

Finally, maximum likelihood estimator of extended Altham distribution family is found as

$$E[\sum_{i=1}^{k} \frac{\partial w_i(\theta)}{\partial \theta}t_i(x)] = \frac{\sum_{j}^{N}\sum_{i=1}^{k} w_i(\theta)t_i(x_j)}{N}, \tag{3.8}$$

which means the MLE of extended Altham distribution family coincide the moment estimator.

On the other hand, we need the reparametrization of extended Altham distribution with respect to $\mu$ and $\sigma^2$ in order to be able to make appropriate comparision. First, we derive the MLE of parameter vector by employing the maximum likelihood method. $\beta = (\beta_1, \beta_2)^T$, and its asymptotic normality. Then we obtain the MLE of $\boldsymbol{\theta} = (\mu, \sigma^2)^T$ and its asymptotic normality according to (2.7) and (2.8). Note that the normalizing constant is the function of $\beta_1$ and $\beta_2$. We establish the following key results for MLEs and their asymptotic covariance matrix. Denote the moment $m_j = E[X^j]$ for $j = 1, 2, 3, 4$.

**Lemma 3.1.**

$$\frac{\partial C(\beta_1,\beta_2)}{\partial \beta_1} = -m_1, \qquad \frac{\partial C(\beta_1,\beta_2)}{\partial \beta_2} = -m_2,$$

$$\frac{\partial^2 C(\beta_1,\beta_2)}{\partial \beta_1^2} = m_2 - m_1^2, \quad \frac{\partial^2 C(\beta_1,\beta_2)}{\partial \beta_2^2} = m_4 - m_2^2, \quad \frac{\partial^2 C(\beta_1,\beta_2)}{\partial \beta_1 \partial \beta_2} = m_3 - m_2 m_1.$$

***Proof.*** Taking the first and second order partial derivatives with respect to $\beta_1$ and $\beta_2$ respectively for both sides of $C(\beta_1, \beta_2)$, and then simplifying the equations will yield the results. For instance,

$$e^{C(\beta_1,\beta_2)} \times \frac{\partial C(\beta_1,\beta_2)}{\partial \beta_1} = \frac{\partial}{\partial \beta_1}\left(\sum_{i=0}^{M} h_i e^{i\beta_1+i^2\beta_2}\right) = -\sum_{i=0}^{M} i h_i e^{i\beta_1+i^2\beta_2}, \tag{3.9}$$

$$e^{C(\beta_1,\beta_2)} \times \frac{\partial C(\beta_1,\beta_2)}{\partial \beta_2} = \frac{\partial}{\partial \beta_2}\left(\sum_{i=0}^{M} h_i e^{i\beta_1+i^2\beta_2}\right) = -\sum_{i=0}^{M} i^2 h_i e^{i\beta_1+i^2\beta_2}, \tag{3.10}$$

$$e^{C(\beta_1,\beta_2)} \times \frac{\partial^2 C(\beta_1,\beta_2)}{\partial \beta_1 \partial \beta_2} + e^{C(\beta_1,\beta_2)} \times \frac{\partial C(\beta_1,\beta_2)}{\partial \beta_2} \times \frac{\partial C(\beta_1,\beta_2)}{\partial \beta_1} \tag{3.11}$$

$$= \frac{\partial}{\partial \beta_2}\left(-\sum_{i=0}^{M} i h_i e^{i\beta_1+i^2\beta_2}\right) = \sum_{i=0}^{M} i^3 h_i e^{i\beta_1+i^2\beta_2}, \tag{3.12}$$

thus

$$\frac{\partial C(\beta_1,\beta_2)}{\partial \beta_1} = -\sum_{i=0}^{M} i h_i e^{i\beta_1+i^2\beta_2} e^{-C(\beta_1,\beta_2)} = -E[X] = -m_1, \tag{3.13}$$

$$\frac{\partial^2 C(\beta_1,\beta_2)}{\partial \beta_1 \partial \beta_2} = \sum_{i=0}^{M} i^3 h_i e^{i\beta_1+i^2\beta_2} e^{C(\beta_1,\beta_2)} - \frac{\partial C(\beta_1,\beta_2)}{\partial \beta_2} \times \frac{\partial C(\beta_1,\beta_2)}{\partial \beta_1} \tag{3.14}$$

$$= E[X^3] - E[X^2]E[X] = m_3 - m_2 m_1. \tag{3.15}$$

□

Suppose the observations are $x_1, x_2, \ldots, x_n$. The log-likelihood is

$$\log L(\beta \mid x_1,\ldots,x_n) = \sum_{k=1}^{n} \log(\Pr[X_k = x_k]) \tag{3.16}$$

$$= -nC(\beta_1,\beta_2) - \beta_1 \sum_{k=1}^{n} x_k - \beta_2 \sum_{k=1}^{n} x_k^2.$$

The score functions are

$$\frac{\partial \log L}{\partial \beta_1} = -n\frac{\partial C(\beta_1,\beta_2)}{\partial \beta_1} - \sum_{k=1}^{n} x_k, \qquad \frac{\partial \log L}{\partial \beta_2} = -n\frac{\partial C(\beta_1,\beta_2)}{\partial \beta_2} - \sum_{k=1}^{n} x_k^2, \tag{3.17}$$

leading to estimating equations

$$\sum_{i=0}^{M} i h_i e^{i\beta_1+i^2\beta_2} \Big/ \sum_{i=0}^{M} h_i e^{i\beta_1+i^2\beta_2} = \frac{1}{n}\sum_{k=1}^{n} x_k = \bar{X}, \tag{3.18}$$

$$\sum_{i=0}^{M} i^2 h_i e^{i\beta_1+i^2\beta_2} \Big/ \sum_{i=0}^{M} h_i e^{i\beta_1+i^2\beta_2} = \frac{1}{n}\sum_{k=1}^{n} x_k^2. \tag{3.19}$$

Applying the quasi-Newton method used before, we can obtain the MLE $\widehat{\beta}$ numerically. Under regularity conditions, for $\beta$ in the interior of the parameter space, the asymptotic normality holds as follows:

$$\sqrt{n}\left(\widehat{\boldsymbol{\beta}} - \boldsymbol{\beta}\right) \to N\left(\mathbf{0}, \boldsymbol{\Sigma}^{-1}\right), \qquad \text{as } n \to \infty, \tag{3.20}$$

where the Hessian matrix is

$$\boldsymbol{\Sigma} = \begin{pmatrix} -E\left[\frac{\partial^2 \log L}{\partial \beta_1^2}\right] & -E\left[\frac{\partial^2 \log L}{\partial \beta_1 \partial \beta_2}\right] \\ -E\left[\frac{\partial^2 \log L}{\partial \beta_1 \partial \beta_2}\right] & -E\left[\frac{\partial^2 \log L}{\partial \beta_2^2}\right] \end{pmatrix} = n \begin{pmatrix} m_2 - m_1^2 & m_3 - m_2 m_1 \\ m_3 - m_2 m_1 & m_4 - m_2^2 \end{pmatrix}. \tag{3.21}$$

Although $\widehat{\boldsymbol{\beta}}$ does not have an explicit form, the MLE of $\boldsymbol{\theta} = (\mu, \sigma^2)^T$ has an explicit form. From score functions (3.17), we also obtain estimating equations for $\mu$ and $\sigma^2$:

$$\mu = \bar{X}, \qquad \sigma^2 + \mu^2 = \frac{1}{n}\sum_{k=1}^{n} x_k^2, \tag{3.22}$$

leading to the MLEs

$$\hat{\mu} = \bar{X}, \qquad \hat{\sigma}^2 = \frac{1}{n}\sum_{k=1}^{n} x_k^2 - \bar{X}^2 = \frac{1}{n}\sum_{k=1}^{n}(x_k - \bar{X})^2. \tag{3.23}$$

Constrains (2.7) and (2.8) imply that $\mu$ and $\sigma^2$ are functions of $\beta_1$ and $\beta_2$ respectively. Denote

$$\mathbf{A} = \begin{pmatrix} \frac{\partial\mu}{\partial\beta_1} & \frac{\partial\mu}{\partial\beta_2} \\ \frac{\partial\sigma^2}{\partial\beta_1} & \frac{\partial\sigma^2}{\partial\beta_2} \end{pmatrix}, \tag{3.24}$$

where

$$\begin{aligned}
\frac{\partial\mu}{\partial\beta_1} &= \frac{\partial}{\partial\beta_1}\left(\sum_{i=0}^{M} i h_i e^{C(\beta_1,\beta_2)+i\beta_1+i^2\beta_2}\right) = -\frac{\partial C(\beta_1,\beta_2)}{\partial\beta_1} \times E[X] - E[X^2] = m_1^2 - m_2, \\
\frac{\partial\mu}{\partial\beta_2} &= \frac{\partial}{\partial\beta_2}\left(\sum_{i=0}^{M} i h_i e^{C(\beta_1,\beta_2)+i\beta_1+i^2\beta_2}\right) = -\frac{\partial log(C(\beta_1,\beta_2))}{\partial\beta_2} \times E[X] - E[X^3] \\
&= m_1 m_2 - m_3,
\end{aligned}$$

and

$$\begin{aligned}
\frac{\partial\sigma^2}{\partial\beta_1} &= \frac{\partial}{\partial\beta_1}\left(\sum_{i=0}^{M} i^2 h_i e^{C(\beta_1,\beta_2)+i\beta_1+i^2\beta_2}\right) - 2\mu\frac{\partial\mu}{\partial\beta_1} \\
&= -\frac{\partial C(\beta_1,\beta_2)}{\partial\beta_1} \times E[X^2] - E[X^3] - 2m_1(m_1^2 - m_2) = 3m_1m_2 - 2m_1^3 - m_3, \\
\frac{\partial\sigma^2}{\partial\beta_2} &= \frac{\partial}{\partial\beta_2}\left(\sum_{i=0}^{M} i^2 h_i e^{C(\beta_1,\beta_2)+i\beta_1+i^2\beta_2}\right) - 2\mu\frac{\partial\mu}{\partial\beta_2} \\
&= -\frac{\partial C(\beta_1,\beta_2)}{\partial\beta_2} \times E[X^2] - E[X^4] - 2m_1(m_1^2 - m_2) \\
&= m_2^2 - m_4 - 2m_1^2 m_2 + 2m_1 m_3.
\end{aligned}$$

Then,

$$\sqrt{n}\left(\widehat{\boldsymbol{\theta}} - \boldsymbol{\theta}\right) \;\to\; N\left(\mathbf{0}, \mathbf{A}\boldsymbol{\Sigma}^{-1}\mathbf{A}^{\mathbf{T}}\right), \qquad \text{as } n \to \infty. \tag{3.25}$$

Matrix $\mathbf{A}$ and $\boldsymbol{\Sigma}$ can be estimated by replacing $m_j$'s as their estimates $\hat{m}_j$'s. Standard errors of $\hat{\mu}$ and $\hat{\sigma}^2$ can be obtained as the square root of diagonal elements of the estimated covariance matrix. There are two approaches to estimate $m_j$:

(1) using the sample only, $\widehat{m}_j = \frac{1}{n}\sum_{k=1}^{n} x_k^j$, or
(2) using the MLEs $\widehat{\beta}$, $\widehat{m}_j = \sum_{i=0}^{M} i^j p_i(\widehat{\beta})$.

The former has large variation when the sample size is not large. Thus, for small sample size, the latter is recommended.

The closed form MLEs of parameters $\mu$ and $\sigma^2$ simplifies the model fitting using the extended Altham distribution.

Under the extended Altham model, the MLE of dispersion index $D$ is $\hat{D} = \frac{\hat{\sigma}^2}{\hat{\mu}(1-\hat{\mu}/M)}$. Denote $\mathbf{B} = \left(\frac{\partial D}{\partial \mu}, \frac{\partial D}{\partial \sigma^2}\right)$ ,where

$$\begin{aligned}
\frac{\partial D}{\partial \mu} &= \frac{\partial}{\partial \mu}\left[\frac{\sigma^2}{M^2}\left(\frac{1}{M-\mu}+\frac{1}{\mu}\right)\right] = \frac{\sigma^2}{M^2}\left(\frac{1}{(M-\mu)^2}-\frac{1}{\mu^2}\right) \\
&= \frac{\sigma^2(2\mu-M)}{M\mu^2(M-\mu)^2} = \frac{(m_2-m_1^2)(2m_1-M)}{Mm_1^2(M-m_1)^2}, \\
\frac{\partial D}{\partial \sigma^2} &= \frac{1}{\mu(1-\mu/M)} = \frac{1}{m_1(1-m_1/M)}.
\end{aligned}$$

Then,

$$\sqrt{n}\left(\hat{D}-D\right) \rightarrow N\left(0, \mathbf{BA\Sigma}^{-1}\mathbf{A}^T\mathbf{B}^T\right), \qquad \text{as } n \to \infty. \tag{3.26}$$

Let $s_D^2$ be the estimate of asymptotic variance $\mathbf{BA\Sigma}^{-1}\mathbf{A}^T\mathbf{B}^T$. The standard error of $\hat{D}$ is $s_D$, and an asymptotic CI of significant level $\alpha$ for dispersion index $D$ is $\hat{D} \pm z_{\alpha/2}s_D$, where $z_{\alpha/2}$ is the $100(1-\alpha/2)\%$ quantile of the standard normal distribution.
Let $P$ denote the extended Altham family (2.5),

$$P = \left\{f_\theta(x) \propto h(x)e^{\beta_1 x+\beta_2 x^2} | \theta = (\beta_1, \beta_2) : -\infty < \beta_1, \beta_2 < \infty, h = 1, 2, \ldots, 12\right\} \tag{3.27}$$

where $X = 0, 1, 2, \ldots, M$. Assume $f_\theta = f_{\tilde{\theta}}$, then the expression

$$log(\frac{h(x)}{\tilde{h}(x)}) + (\beta_1 - \tilde{\beta}_1)x + (\beta_2 - \tilde{\beta}_2)x^2 = 0 \tag{3.28}$$

is satisfied for all $x$ only when all its coefficients are equal to zero, which is only possible when $h = \tilde{h}$, $\beta_1 = \tilde{\beta}_1$ and $\beta_2 = \tilde{\beta}_2$. Hence, we conclude that the extended Altham family is identifiable iff $log(\frac{h(x)}{\tilde{h}(x)}) \neq \beta_1 x + \beta_2 x^2, \beta_1, \beta_2 \neq 0$.

## 4. Simulation study and examination of existing examples

In the literature, some scholars tried different models. Bailey [4] proposed a particular probabilistic model based on the Markov property to study the author's writing style by investigation of occurrences of function word in 5-word and 10-word samples. Two data sets from Macaulay's 'Essay on Milton' [13] and from Chesterton's essay 'About the workers' [7] respectively were fitted. Chakraborty and Das [5] fitted QBD I and QBD II models for four data sets from other authors, these examples were actually truncated count data, not from true binomial experiments. The observed and expected frequencies, as well as the values of goodness-of-fit of fitted models were reported in both papers, thus, we can compare the fitting of the extended Altham models with theirs using the the quantity of the goodness-of-fit under the same data grouping schemes. Dispersion investigation shows that all examples are under-dispersed in Bailey [4], and over-dispersed in Chakraborty and Das [5]. The comparison results are reported in Table 1 and Table 2. Table 1 gives the fitting comparison of extended Altham models with the model proposed by Bailey [4] for 5-word and 10-word samples of function word occurrence from two authors (Macaulay's work, Chesterton's work*). Data sets (see Appendix Table A1) and original fittings are referred to Bailey [4].

**Table 1.** Fitting comparison of extended Altham models with the model proposed by Bailey [4] for 5-word and 10-word samples

| Model | 5-word $(0.61, 0.35, 0.66)$ | 10-word $(1.05, 0.64, 0.68)$ | 5-word* $(0.61, 0.35, 0.66)$ | 10-word* $(1.05, 0.64, 0.68)$ |
|---|---|---|---|---|
| Bailey's model | 8.16 | 6.38 | 2.76 | 4.93 |
| 1 | 0.0819 | 0.4887 | 0.3432 | 2.0268 |
| 2 | 0.0809 | 0.4869 | 0.3404 | 2.0263 |
| 3 | 0.1027 | 1.0299 | 0.3356 | **1.6936** |
| 4 | 0.0659 | **0.1549** | 0.3440 | 2.4571 |
| 5 | 0.0974 | 0.7956 | 0.3367 | 1.8095 |
| 6 | 0.0695 | 0.2599 | 0.3431 | 2.2801 |
| 7 | 0.0804 | 0.4847 | 0.3405 | 2.0307 |
| 8 | 0.0843 | 0.4925 | 0.3396 | 2.0232 |
| 9 | 0.2073 | 6.4423 | **0.3199** | 1.7605 |
| 10 | **0.0324** | 0.8428 | 0.3536 | 5.4180 |
| 11 | 0.1029 | 0.9111 | 0.3355 | 1.7519 |
| 12 | 0.0657 | 0.2019 | 0.3441 | 2.3670 |

**Table 2.** Fitting comparison of extended Altham models and the fitted QBD I and QBD II models by Chakraborty and Das [5] for four data sets

| Model | Example 1 $(0.41, 0.51, 1.39)$ | Example 2 $(0.68, 0.81, 1.37)$ | Example 3 $(2.50, 3.37, 2.70)$ | Example 4 $(0.92, 0.93, 1.23)$ |
|---|---|---|---|---|
| QBD I | 0.075 | 3.608 | 0.457 | 0.941 |
| QBD II | **0.067** | 3.618 | 0.324 | 0.944 |
| 1 | 0.8834 | 4.0709 | 0.3488 | 2.1207 |
| 2 | 0.4713 | 4.3125 | 0.4443 | 2.4330 |
| 3 | 1.7661 | 2.5235 | 0.5243 | **0.7481** |
| 4 | 0.3429 | 6.0471 | 0.2100 | 4.5916 |
| 5 | 2.1230 | 2.8038 | 0.4936 | 0.9950 |
| 6 | 0.1998 | 5.6710 | 0.3157 | 4.1748 |
| 7 | 0.4870 | 4.3679 | 0.4160 | 2.5023 |
| 8 | 1.4354 | 3.8007 | 0.3871 | 1.8306 |
| 9 | 6.9637 | **0.5060** | 1.3459 | 3.6989 |
| 10 | 0.9193 | 15.1880 | **0.0016** | 20.2011 |
| 11 | 2.6102 | 2.4775 | 0.5451 | 0.7993 |
| 12 | 0.0988 | 6.2147 | 0.3025 | 4.9262 |

Table 2 gives the fitting comparison of extended Altham models and the Chakraborty and Das [5] fitted QBD I and QBD II models for four data sets (see Appendix Tables A2-A5). Data sets and original fittings are referred to Chakraborty and Das [5]. The $\chi^2$-values of goodness-of-fit are obtained under the same data grouping schemes. ($\bar{x}$, $s^2$, $\hat{D}$) are given for each example, where $\bar{x}$ is sample mean, $s^2$ is sample variance and $\hat{D}$ is sample dispersion index. In all examples in Bailey [4], the extended Altham models fits better than the model proposed by Bailey. Refering to samples from Macaulay's work, for the 5-word and the 10-word samples we get the appropriate extended Altham models (2.31) with $\chi^2 = 0.0324$ and (2.25) with $\chi^2 = 0.1549$, respectively.

Regarding to Chesterton's work, we found that the appropriate models for the 5-word* and for the 10*-word samples are extended Altham models (2.30) with $\chi^2 = 0.3199$ and (2.24) with $\chi^2 = 0.16936$, respectively.

In fact, most of the extended Altham models beat the Bailey's model. Moreover, the $\chi^2$ testing at significant level 10% will accept the extended Altham model, but reject the Bailey's model. This might indicate that the original setting of probabilistic mechanism needs further adjustment or refinement.

In Example 2, 3 and 4 in Chakraborty and Das [5], the extended Altham model is better than QBD I and QBD II, while in first example the QBD I and QBD II are better than the extended Altham model. However, the results of acceptance or rejection from the $\chi^2$ test at significant level 10% for all three models are the same. The above examination shows that the extended Altham model can be a safe tool in explorative analysis without special preference in model specification, and also can be an alternative model if other favoured models do not fit data well.

Now we apply the proposed extended Altham model to over-dispersed binomial data resulted from a survey of deaths of children in northest Brazil and the counts the frequencies of 430 childhood deaths in 2946 families of sizes up to eight children. Maternity histories were collected on women aged 15 to 44 over a 3-month period in 1986. The original data was published by Sastry [16] and later it was used for demonstration of different weighted binomial models by Zelterman [19]. We get the sample data regarding to families that has more than three siblings (see Appendix Table A6). From this point of view, the results of extended Altham modelling are given in Table 3.

**Table 3.** Fitting extended Altham models for the childhood death in Brazilian family data

| Model | Number of siblings (n) | | | | |
|---|---|---|---|---|---|
| | 4 | 5 | 6 | 7 | 8 |
| | $\bar{x} = 0.49$ $s^2 = 0.52$ $\hat{D} = 1.13$ | $\bar{x} = 0.99$ $s^2 = 1.16$ $\hat{D} = 1.34$ | $\bar{x} = 1.34$ $s^2 = 1.78$ $\hat{D} = 1.59$ | $\bar{x} = 1.80$ $s^2 = 1.48$ $\hat{D} = 1.06$ | $\bar{x} = 2.33$ $s^2 = 1.72$ $\hat{D} = 1.04$ |
| 1 | 0.3147 | 1.7954 | 4.6315 | 1.4215 | 0.9545 |
| 2 | 0.3120 | 1.7737 | 4.6929 | 1.4044 | 0.9534 |
| 3 | 0.5310 | 2.2675 | **3.1773** | 2.0239 | **0.9450** |
| 4 | 0.2573 | 2.0807 | 7.2788 | 1.0064 | 1.0830 |
| 5 | 0.4297 | 2.0184 | 3.6284 | 1.7884 | 0.9473 |
| 6 | **0.2631** | 1.8905 | 6.2074 | 1.1332 | 1.0152 |
| 7 | 0.3100 | **1.7623** | 4.7541 | 1.3936 | 0.9555 |
| 8 | 0.3195 | 1.8329 | 4.5469 | 1.4509 | 0.9561 |
| 9 | 3.4693 | 9.4540 | 3.8744 | 5.2578 | 1.7549 |
| 10 | 1.1462 | 7.7364 | 25.3762 | **0.6005** | 1.8558 |
| 11 | 0.4698 | 2.0462 | 3.4754 | 1.8721 | 0.9594 |
| 12 | 0.2636 | 2.0075 | 6.5180 | 1.0801 | 1.0180 |

According to Table 3, it is obvious that we have huge improvement over the previously examined models. Moreover, extended Altham model has the advantage of having only two parameters.

The last example that we consider is the data that was collected on the sex of the first four children carried out at the A Maxwell Evans Clinic by Elwood and Coldman [10] on 1022 newly diagnosed women with primary breast cancer who had four or fewer children and for whom the sex of each child was known. The data shows mean ages at diagnosis by number and sex of children. Elwood and Coldman [10] made the analysis in order to observe a possible relationship between the age at diagnosis in women with breast cancer and the sex of their offspring.

**Table 4.** Fitting extended Altham model for diagnosis of breast cancer by number and sex of children

| Model | Number of siblings (n) | |
|---|---|---|
| | 3 | 4 |
| | $\bar{x} = 1.51$<br>$s^2 = 0.77$<br>$\hat{D} = 0.82$ | $\bar{x} = 1.93$<br>$s^2 = 1.15$<br>$\hat{D} = 1.16$ |
| 1 | 3.2216 | 4.4021 |
| 2 | **2.4989** | **3.0160** |
| 3 | 3.4059 | 4.1389 |
| 4 | 3.2205 | 4.7523 |
| 5 | 4.0819 | 5.5094 |
| 6 | 2.6270 | 3.4797 |
| 7 | 2.6185 | 3.2717 |
| 8 | 3.9468 | 5.7224 |
| 9 | 5.2482 | 3.8781 |
| 10 | 4.4652 | 6.4805 |
| 11 | 4.3434 | 5.7529 |
| 12 | 2.5426 | 3.3352 |

Actually, they didn't mention any models for their data. Since their data includes under-dispersed and over-dispersed cases in the same experiment, we decided to use their data (see Appendix Table A7). The number of siblings bigger than two is considered. The summary results of fitting extended Altham model is given in Table 4. In Table 4, we can see that the distribution of the number of diagnosis of breast cancer in the family that has 3 children is under-dispersed ($\hat{D} = 0.82$) and the similar distribution for the family that has 4 children is over-dispersed ($\hat{D} = 1.16$). And extended Altham model (2.23) is best fit for the both cases.

## 5. Discussion

The extended Altham distribution family is constructed by Kullback-Leibler divergence measure. It turns out to be a particular type of extended Altham distribution, with simple form of pmf from the parametrization of Lagrangian multipliers, which may rendered it to be overlooked previously. Since the construction is very conservative, it is relatively safer than the binomial as well other models developed based on particular probabilistic mechanisms.

The capability to reach the full range of dispersion makes the extended Altham a flexible model for binomial data of various dispersion situations. Thus, it can serve as an explorative model first to avoid wrong specification (say using the binomial model). Because of the conservative feature of the extended Altham, its fitting can be refined or improved by a better model like QBD or EBB, based on revealed dispersion information.

The closed form MLEs simplify the fitting for data, thus, facilitating the application for general end-users, although the calculation of pmf requires the numerical algorithm. The development of a regression framework is in progress.

**Acknowledgment.** The author thanks Dr. Rong Zhu, Dr.Renjun Ma and Dr. Abdel H. El-Shaarawi for their valuable comments and suggestions regarding to this work which led to a great improvement of the results and the presentation of the article.

## References

[1] P.M.E. Altham, *Two Generalizations of the binomial distribution*, Appl. Statist. **27** (2), 162-167, 1978.

[2] P.M.E. Altham and R.K.S. Hankin, *Multivariate generalizations of the multiplicative binomial distribution: Introducing the MM package*, J. Stat. Softw **46** (12), 1-23, 2012.

[3] P.M.E. Altham and J.K. Lindsey, *Analysis of the human sex ratio by using overdispersion models*, Appl. Statist. **47** (1), 149-157, 1998.

[4] B.J.R. Bailey, *A model for function word counts*, Appl. Statist. **39** (1), 107-114, 1990.

[5] S. Chakraborty and K.K. Das, *On some properties of a class of weighted quasi-binomial distributions*, J. Statist. Plann. Inference **136** (1), 159-182, 2006.

[6] C. Chatfield, *Problem Solving: A Statistician's Guide*, Chapman & Hall, London, 1988.

[7] G.K. Chesterton, *Selected Essays of G. K. Chesterton*, London: Methuen, 1949.

[8] P.C. Consul, *A simple urn model dependent upon predetermined strategy*, Sankhya A Series B **36** (4), 391-399, 1974.

[9] I. Dobson, B.A. Carreras and D.E. Newman, *A probabilistic loading-dependent model of cascading failure and possible implications for blackouts*, Proceedings of the 36th Annual Hawaii International Conference on System Sciences (HICSS'03), 2003.

[10] M. Elwood and A. Coldman, *Age of mothers with breast cancer and sex of their children*, Br Med J **282** (6265), 734, 1981.

[11] J.K. Haseman and L.L. Kupper, *Analysis of dichotomous response data from certain toxicological experiments*, Biometrics **35** (1), 281-293, 1979.

[12] H. Joe, *Multivariate Models and Dependence Concepts*, Chapman & Hall, London, 1997.

[13] L. Macaulay, *Literary Essays Contributed to the Edinburgh Review*, London: Oxford University Press, 1923.

[14] A. Mishra, D. Tiwary and S.K. Singh, *A class of quasi-binomial distributions*, Sankhya A Series B **54** (1), 67-76, 1992.

[15] R.L. Prentice, *Binary regression using an extended beta-binomial distribution, with discussion of correlation induced by covariate measurement errors*, J. Amer. Statist. Assoc. **81** (394), 321-327, 1986.

[16] N. Sastry, *A nested frailty model for survival data, with an application to the study of child survival in northest Brazil*, J. Amer. Statist. Assoc. **92** (438), 426-434, 1997.

[17] R. Viveros-Aguilera, K. Balasubramanian and N. Balakrishnan, *Binomial and negative binomial analogues under correlated Bernoulli trials*, Amer. Statist. **48** (3), 243-247, 1994.

[18] R.R. Wilcox, *A review of the beta-binomial model and its extensions*, J. Educ. Stat. **6** (1), 3-32, 1981.

[19] D. Zelterman, *Discrete Distributions: Applications in the Health Sciences*, John Wiley & Sons, West Sussex, 2004.

## Appendix

**Table A1.** Underdispersed word counts [4]

| Occurences | 0 | 1 | 2 | 3 |
|---|---|---|---|---|
| 5-word | 45 | 49 | 6 | 0 |
| 10-word | 27 | 44 | 26 | 3 |

**Table A2.** Observed and expected frequencies of European Corn borer in 1296 Corn plants [5]

| No. of borers per plant | 0 | 1 | 2 | 3 | $\geq 4$ |
|---|---|---|---|---|---|
| Observed no. of plants | 907 | 275 | 88 | 23 | 3 |

**Table A3.** Distribution of yeast cells per square in a haemacytometer [5]

| No. of cells per square | 0 | 1 | 2 | 3 | 4 | 5 |
|---|---|---|---|---|---|---|
| Observed no. of squares | 213 | 128 | 37 | 18 | 3 | 1 |

**Table A4.** Distribution of number of seeds by time of day [5]

| Time | 0 | 1 | 2 | 3 | 4 | 5 |
|---|---|---|---|---|---|---|
| Observed no. seeds | 7 | 4 | 5 | 5 | 4 | 7 |

**Table A5.** Distribution of number of hits per square [5]

| No. of hits | 0 | 1 | 2 | 3 | 4 | 5 |
|---|---|---|---|---|---|---|
| No. of 1/4 km squares | 229 | 211 | 93 | 35 | 7 | 1 |

**Table A6.** The frequency of childhood deaths in Brazilian families [19]

| Number of siblings $n$ | Number of families $f_n$ | Number of deaths $m_n$ | Number of affected siblings i | | | | | | | |
|---|---|---|---|---|---|---|---|---|---|---|
| | | | 0 | 1 | 2 | 3 | 4 | 5 | 6 | 7+ |
| 1 | 267 | 12 | 255 | 12 | | | | | | |
| 2 | 285 | 48 | 239 | 44 | 2 | | | | | |
| 3 | 202 | 80 | 143 | 41 | 15 | 3 | | | | |
| 4 | 110 | 54 | 69 | 30 | 9 | 2 | 0 | | | |
| 5 | 104 | 103 | 43 | 34 | 15 | 9 | 3 | 0 | | |
| 6 | 50 | 67 | 15 | 18 | 8 | 5 | 3 | 0 | 1 | |
| 7 | 21 | 38 | 4 | 4 | 7 | 4 | 2 | 0 | 0 | 0 |
| 8 | 12 | 28 | 1 | 2 | 4 | 3 | 1 | 1 | 0 | 0 |
| Totals | 2946 | 430 | | | | | | | | |

**Table A7.** Diagnosis of breast cancer by number and sex of children [10]

| No of Children | 0 | 1 | | 2 | | | 3 | | | | 4 | | | | |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| No of boys | 0 | 0 | 1 | 0 | 1 | 2 | 0 | 1 | 2 | 3 | 0 | 1 | 2 | 3 | 4 |
| No of patients | 284 | 93 | 71 | 65 | 134 | 83 | 26 | 71 | 75 | 26 | 11 | 21 | 30 | 28 | 4 |

Hacettepe Journal of
Mathematics & Statistics

Hacet. J. Math. Stat.
Volume 50 (1) (2021), 275 – 288
DOI : 10.15672/hujms.758925

Research Article

# Comparative analysis between FAR and ARL based control charts with runs rules

Rashid Mehmood[1], Muhammad Hisyam Lee*[1], Iftikhar Ali[2], Muhammad Riaz[3]

[1] Department of Mathematical Sciences, Universiti Teknologi Malaysia, Malaysia
[2] Department of Mathematics, University of Hafr Al Batin, Saudi Arabia
[3] Department of Mathematics and Statistics, King Fahad University of Petroleum and Minerals, Saudi Arabia

## Abstract

In this study, we have conducted comparative analysis between false alarm rate (FAR) and average run length (ARL) based control charts with runs rules. In this regard, we have considered various univariate and multivariate control charts which include mean, standard deviation, variance, Hotelling, and generalized variance. For evaluation purpose, we have used actual false alarm rate, power, in-control actual average run length, and out-of-control average run length as performance indicators. Furthermore, the performance indicators are calculated through Monte Carlo simulation procedures. Results revealed that performance order of runs rules with FAR based control charts are persistent whereas, performance order of runs rules with ARL based control charts are dependent on the circumstances, that is, sample size, size of shift, type of control chart, and side of control limit (upper-sided and lower-sided). Besides, we have provided a real life example using the data on electrical resistance of insulation. In this approach, we have determined that behavior of FAR and ARL based control charts using the real data is recorded similar to the behavior using the statistical performance indicators.

**Mathematics Subject Classification (2020).** 62N05, 62F10, 62F12, 62F15, 62F25, 62F40

**Keywords.** Average run length, control chart, false alarm rate, performance indicators, power, probability of single point, runs rules

## 1. Introduction

The theory of control charts was first proposed by Walter A. Shewhart in 1931 [17] for the detection of assignable causes of variations in a parameter (location and dispersion) of a process characteristics. The assignable causes of variations are unnaturally appeared in an ongoing process, and they are usually occurred due to improper adjustment of controller, operators error, and low quality of batch material. A control chart based on the concept of Shewhart [17] is often known as Shewhart-type control chart. The Shewhart-type control chart based on classical runs rule (any single point out-of-control) is generally considered

*Corresponding Author.
Email addresses: mr.rashid_mehmood@yahoo.com (R. Mehmood), mhl@utm.my (M.H. Lee), iftikharali4u@gmail.com (I. Ali), riaz76qau@yahoo.com (M. Riaz)
Received: 27.06.2020; Accepted: 25.10.2020

less efficient for detection of small variations in a parameter [13]. However, to increase the ability of Shewhart control charts towards detection of small variations, Western [21] recommended sensitizing rules or runs rules (also known as decision rules). With passage of time, various authors introduced new forms of sensitizing rules as well as explored their behavior in forms of actual in-control average run length (abbr. as AIARL and denoted as $ARL_{act}$) and out-of-control average run length (abbr. as OARL and denoted as $ARL_1$) such as [3–6, 10, 15, 18–20]. The AIARL is an actual value of the average number of sample points that stayed in-control before declaring a process out-of-control on the basis of decision points when in-fact process is in-control. Furthermore, OARL is the average number of sample points that stayed in-control before declaring a process out-of-control on the basis of decision points when actually process is out-of-control.

Champ and Woodall [3] investigated the AIARL as well as OARL of different sensitizing rules. In addition, they used Markov Chain approach as computational technique. Their results showed that although simultaneously implementation of sensitizing rules enhanced the detection ability of Shewhart type control chart but at the same time generated another issue. The issue stated as AIARL deviated from intended level, that is, substantially degraded. To overcome the issue of sensitizing or runs rules, many authors recommended to incorporate the correct value of in-control probability of single point (abbr. as IPSP and denoted as $p_0$) into the design structure of Shewhart type control chart [4, 5, 8, 10, 15, 22]. The IPSP is defined as the probability of an out-of-control signal when in-fact a process is in-control. Furthermore, IPSP is generally computed through involving an appropriate method by taking into account an independent choice of runs rules and prefix value of FAR (denoted as $\alpha$) or in-control ARL (denoted as $ARL_0$). The prefix value of $\alpha$ can be defined as the prefix value of probability of decision points for a given choice of runs rule when in-fact a process is in-control. On the other hand, $ARL_0$ is the prefix value of the average number of sample points that should be stayed in-control before declaring a process out-of-control on the basis of decision points when in-fact process is in-control.

The appropriate method for computing the IPSP is considered important in designing of Shewhart-type control charts. For instance, Klein [5] computed IPSP based on Markov chain approach for designing and evaluating the mean ($\bar{X}$) control chart. Khoo [4] established graphical plots based on Markov chain approach to obtain the IPSP of existing and proposed runs rules. In addition, he applied the probabilities of single point in the construction of $\bar{X}$ control chart. Shepherd et al. [16] computed the IPSP based on Markov chain approach for designing and evaluation of attribute control chart under runs rules. In continuation, Riaz et al. [15] utilized the proposed equation for designing the FAR based upper-sided mean (symbolized as $\bar{X}_U$), variance ($S_U^2$), standard deviation ($S_U$) and range ($R_U$) control charts. In addition, they showed that proposed equations play its role to maintain the AFAR of FAR based $\bar{X}_U$, $S_U^2$, $S_U$ and $R_U$ control charts under runs rules at $\alpha$. The applications of polynomial equation by [15] can be seen in various studies such as [9, 11, 12, 22]. In this particular research direction, Mehmood et al. [8] offered new polynomial equation alternative to the study by [15] for increasing the detection ability of two sided Shewhart-type control chart under runs rules.

The aforementioned literature review is representing the FAR and ARL based control charts. A control chart depends on the $\alpha$ is termed as FAR based control chart such as [15, 22]. Likewise, ARL based control chart depends on the $ARL_0$ such as [4, 5]. It is valuable to mention that numbers of studies have been seen on the topic of FAR and ARL based control charts separately. In this research direction, it is very rare to find study on the comparative analysis between FAR and ARL based control charts. This has taken as the motivation of current study.

This study aims to conduct comparative analysis between FAR and ARL based control charts with runs rules. To achieve the goal, we will construct design structures of upper-sided and lower-sided univariate and multivariate control charts with runs rules.

The upper-sided and lower-sided univariate control charts include mean ($\bar{X}_U$ and $\bar{X}_L$), variance ($S_U^2$ and $S_L^2$), and standard deviation ($S_U$ and $S_L$). Furthermore, upper-sided and lower-sided multivariate control charts contain generalized variance ($|S|_U$, and $|S|_L$) and Hotelling's ($T_U^2$). Besides, we will evaluate the performance of FAR and ARL based control charts by considering the AFAR, power (denoted as $P_1$), AIARL, and OARL as performance measures. The $P_1$ is defined as the probability of the decision points for a given choice of runs rule that are declared out-of-control when in-fact the process is out-of-control. In addition, for computation of the performance measures, we will illustrate and also employ the Monte Carlo simulation procedures without loss of generality. Furthermore, we will conduct comparative analysis on the behavior of FAR and ARL based control charts under classical and additional runs rules. All of the prescribed methods for comparative analysis cover the statistical aspects of current study. To highlight the practical significance of the study, a real life example will be presented using the data on electrical resistance of insulation.

Rest of the article is organized as follows: In Section 2, we will construct different design structures of FAR and ARL based control charts with classical and additional runs rules. In Section 3, we will discuss Monte Carlo simulation procedure for computing different performance measures of each control chart under consideration, and also conduct comparative analysis. In Section 4, we will give a real life example using the data on electrical resistance of insulation to compare the behavior of FAR and ARL based control charts with runs rules. Lastly, we will summarize and conclude the whole study in Section 5.

## 2. Design structures of FAR and ARL based Shewhart-type control charts under runs rules

In this section, we construct FAR and ARL based design structures of the Shewhart-type control charts under runs rules. Now assume that a process characteristic $X$ follows a normal distribution and characteristics ($Y_1$, $Y_2$) follow bivariate normal distribution.

### 2.1. $\bar{X}_U$ control chart

Let $\bar{X}_j$, $j = 1,2,3,...$ denote the $j$th plotting statistic of sample of size $n$. Thus, a process is said be out-of-control if $k/k$ or $k/k + r$ consecutive statistic $\bar{X}_j$ falling above the control limit $U_{\bar{X}}$. The $\bar{X}_j$ and $U_{\bar{X}}$ are formulated as follows:

$$\bar{X}_j = \frac{1}{n}\sum_{i=1}^{n} X_{ij}, \quad U_{\bar{X}} = \mu_0 + Z_{(1-p_0)}\frac{\sigma_0}{\sqrt{n}},$$

where $\mu_0$ and $\sigma_0$ are known in-control mean and standard deviation of $X$, $Z_{(1-p_0)}$ is $(1-p_0)$th percentile of standard normal distribution [13]. Furthermore, choice of $p_0$ depends on the prefix value of $k/k$ or $k/k+r$ runs rules and $\alpha$ or $ARL_0$. The correct value of $p_0$ is desired to sustain the $\alpha_{act}$ or $ARL_{act}$ of a control chart at $\alpha$ or $ARL_0$, respectively. To compute the required $p_0$ value, one of the best solutions provided by [15] in the form of a polynomial equation for handling the FAR based control charts. Riaz et al. [15] introduced exact polynomial equation for computing the required $p_0$ value as per the given choice of $k/k$ or $k/k+r$ and $\alpha$. Thus, polynomial equation for computing the $p_0$ as per the given choice of $k/k$ and $\alpha$ or $ARL_0$ are given as:

$$\begin{cases} p_0 = \sqrt[k]{\alpha}, & \text{if } \alpha \text{ is given,} \\ ARL_0(1-p_0)p_0^k + p_0^k - 1 = 0, & \text{if } ARL_0 \text{ is given.} \end{cases} \tag{2.1}$$

To cover the case of $k$ out of $k+r$ (denoted as $k/k+r$, $r \geq 1$) runs rules, expressions to obtain $p_0$ for the given value of $\alpha$ or $ARL_0$ are as follows:

$$\begin{cases} \alpha = \binom{k+r}{k} p_0^k (1-p_0)^r, & \text{if } \alpha \text{ is given,} \\ p_0 = R(k|k+r, ARL_0), & \text{if } ARL_0 \text{ is given,} \end{cases} \tag{2.2}$$

where $R(k|k+r, ARL_0)$ denote a constant, lies between zero and one, and it depends on the given value of $k/k+r$ and $ARL_0$. Besides, a control chart dependent on $\alpha$ is termed as FAR based control chart. Similarly, a control chart contingent on $ARL_0$ is called ARL based control chart. The theoretical justification of Eqs.(2.1)–(2.2) when $\alpha$ given can be seen in [8]. In addition, theoretical illustration of Eq.(2.1) when $ARL_0$ given is as follows: The probability distribution (also called run length distribution) of $k/k$ consecutive statistics breached the control limit is generalized geometric distribution of order $k$ with parameter $p_0$ [2]. As our interest is to find out correct value of $p_0$ so that $ARL_{act}$ of a Shewhart-type control remains equal to $ARL_0$. Therefore, we equate the mean of generalized geometric distribution of order $k$ with parameter $p_0$ to $ARL_0$. Note that value of $p_0$ in Eq. (2.2) when $ARL_0$ given is hard to obtain by analytical approach. However, one may calculate using a computational technique (e.g. Monte Carlo simulation) with a condition that $ARL_{act}$ remains equal to $ARL_0$.

## 2.2. $\bar{X}_L$ control chart

Let $\bar{X}_j$, $j = 1,2,3,...$ denote the $j$th plotting statistic of sample of size $n$. Thus, a process is said be out-of-control if $k/k$ or $k/k+r$ consecutive $\bar{X}_j$ falling below the $L_{\bar{X}}$. The $\bar{X}_j$ and $L_{\bar{X}}$ are formulated as follows:

$$\bar{X}_j = \frac{1}{n}\sum_{i=1}^{n} X_{ij}, \quad L_{\bar{X}} = \mu_0 + Z_{p_0}\frac{\sigma_0}{\sqrt{n}},$$

where $Z_{p_0}$ is $p_0$th percentiles of standard normal distribution [13]. Rest of the discussion remained similar to Section 2.1.

## 2.3. $S_U^2$ and $S_L^2$ control charts

Let $S_j^2$, $j = 1,2,3,...$ denote the $j$th plotting statistic of sample of size $n$. Thus, a process is said be out-of-control if $k/k$ or $k/k+r$ consecutive $S_j^2$ crossed the control limit ($U_{S^2}$ for $S_U^2$ or $L_{S^2}$ for $S_L^2$ control chart). The $S_j^2$, $U_{S^2}$, and $L_{S^2}$ are formulated as follows:

$$S_j^2 = \frac{1}{n-1}\sum_{i=1}^{n}(X_{ij} - \bar{X}_j), \quad U_{S^2} = \frac{w_U \sigma_0^2}{n-1}, \quad L_{S^2} = \frac{w_L \sigma_0^2}{n-1},$$

where $w_U$ and $w_L$ are $(1-p_0)$th and $p_0$th percentiles of chi-squared distribution with $n-1$ degree of freedom, and $\sigma_0^2$ is known in-control variance of $X$.

## 2.4. $S_U$ and $S_L$ control charts

Let $S_j$, $j = 1,2,3,...$ denote the $j$th plotting statistic of sample of size $n$. Thus, a process is said be out-of-control if $k/k$ or $k/k+r$ consecutive $S_j$ falls outside the control limit ($U_S$ for $S_U$ or $L_S$ for $S_L$ control chart). The $S_j$, $U_S$, and $L_S$ are formulated as follows:

$$S_j = \sqrt{\frac{\sum_{i=1}^{n}(X_{ij} - \bar{X}_j)}{n-1}}, \quad U_S = \frac{m_U \sigma_0}{\sqrt{n-1}}, \quad L_S = \frac{m_L \sigma_0}{\sqrt{n-1}},$$

where $m_U$ and $m_L$ are $(1-p_0)$th and $p_0$th percentiles of chi distribution with $n-1$ degree of freedom, and $\sigma_0$ is known in-control standard deviation of $X$.

### 2.5. Bivariate $T_U^2$ control chart

Let $T_j^2$, $j = 1,2,3,...$ denote the $j$th plotting statistic of sample of size $n$. Thus, a process is said be out-of-control if $k/k$ or $k/k+r$ consecutive $T_j^2$ lies beyond the $U_{T^2}$. The $T_j^2$ and $U_{T^2}$ are formulated as follows:

$$T_j^2 = n(M_j - \underline{\mu_0})^t \Sigma_0^{-1}(M_j - \underline{\mu_0}), \quad U_{T^2} = t_U^2,$$

where $M_j = (\bar{Y}_{1j}, \bar{Y}_{2j})^t$ is the $j$th sample mean vector, $\underline{\mu_0} = (\mu_{10}, \mu_{20})^t$ is known in-control mean vector of $Y_1$ and $Y_2$, $\Sigma_0$ is variance-covariance matrix of $M_j$, and $t_U^2$ is $(1-p_0)$th percentile of chi-squared distribution with two degree of freedom.

### 2.6. Bivariate $|S|_U$ and $|S|_L$ control charts

Let $|S|_j$, $j = 1,2,3,...$ denote the $j$th plotting statistic of sample of size $n$. Thus, a process is said be out-of-control if $k/k$ or $k/k+r$ consecutive $|S|_j$ falls outside the control limit ($U_{|S|}$ for $|S|_U$ or $L_{|S|}$ for $|S|_L$ control chart). The $S_j$, $U_S$, and $L_S$ are formulated as follows:

$$|S|_j = S_{1j}^2 S_{2j}^2 - S_{12j}^2, \quad U_{|S|} = \frac{|\Sigma_0|\ b_U^2}{4(n-1)^2}, \quad L_{|S|} = \frac{|\Sigma_0|\ b_L^2}{4(n-1)^2},$$

where $S_{1j}^2$ and $S_{2j}^2$ are $j$th sample variance of size $n$, $S_{12j}^2$ is sample covariance between process characteristics ($Y_1$ and $Y_2$), $b_U$ and $b_L$ are $(1-p_0)$th and $p_0$th percentiles of chi-squared distribution with $2n-4$ degree of freedom, and $|\Sigma_0|$ is the determinants of $\Sigma_0$.

## 3. Computation of performance measures and comparative analysis

In this section we are intended to provide Monte Carlo simulation procedure [7, 15] for computing the performance measures of upper-sided and lower-sided control charts under runs rules (see Sec. 2), and also conduct comparative analysis. The performance measures are $\alpha_{act}$, $P_1$, $ARL_{act}$, and $ARL_1$, and their further details are given in Sec. 1. A control chart for different choices of runs rules is said to be best if $\alpha_{act}$ or $ARL_{act}$ is equal to $\alpha$ or $ARL_o$, respectively. Likewise, a control chart under different choices of runs rules can be announced best for a certain choice of runs rule if it attains minimum $ARL_1$ or maximum $P_1$ given that the control chart has same $ARL_0$ or $\alpha$ respectively.

### 3.1. $\bar{X}_U$ and $\bar{X}_L$ control charts

To compute the $P_1$ of $\bar{X}_U$ control chart, generate $10^5$ random samples of size $n$ from normal distribution with out-of-control mean $\mu^* = \mu_0 + \delta_1\sigma_0$ (where $\delta_1 \geq 0$ represents amount of upward shift) and in-control standard deviation $\sigma_0$ followed by calculating the plotting statistics ($\bar{X}_j$) and comparing them with $U_{\bar{X}}$ to count the number of statistics falling above the $U_{\bar{X}}$. Finally, proportion of plotting statistics falling above the $U_{\bar{X}}$ is reported as $P_1$. Similarly, one may proceed for $\bar{X}_L$ control chart by considering $L_{\bar{X}}$ with $\mu^* = \mu_0 + \delta_2\sigma_0$ (where $\delta_2 \leq 0$ represents amount of downward shift). Furthermore, for computing the $ARL_1$, generate a random sample of size $n$ from normal distribution followed by calculating the statistics to compare with the $U_{\bar{X}}$ or $L_{\bar{X}}$ for deciding either process is in-control or out-of-control. Afterwards, repeat the prescribed procedure until the process is declared out-of-control and then record the sample number (run length). Likewise, repeat the aforementioned procedure $10^5$ times to attain the vector of run length. Ultimately, average of the vector of run length is required $ARL_1$. Note that $\alpha_{act}$ and $ARL_{act}$ is the special case of $P_1$ and $ARL_1$, respectively when $\delta_1 = \delta_2 = 0$. Based on the aforesaid procedures, we have attained $\alpha_{act}$, $ARL_{act}$, $P_1$ and $ARL_1$ of $\bar{X}_U$ and $\bar{X}_L$control charts for some selective choices of $\delta_1$, $\delta_2$, $n$, $\alpha = 0.0027$, $ARL_0 = 370$, $k/k$ and $k/k+r$ (see Tables 1–3). Thus, the results are discussed as follows:

- The $\alpha_{act}$ and $ARL_0$ of mean control charts ($\bar{X}_U$ and $\bar{X}_L$) are obtained equal to $\alpha$ and $ARL_0$ (i.e. $\alpha_{act} = \alpha = 0.0027$ and $ARL_{act} = ARL_0 = 370$) for classical and additional runs rules (see Table 1). This means that Eqs.(2.1)–(2.2) plays its role for resolving the issue of Shewhart-type control charts under runs rules. The details about the issue of Shewhart-type control charts are given in Sec. 1.
- Behavior of FAR based mean control charts with runs rules are sustained in terms of $P_1$ (see Tables 2–3). Similarly, we have observed for the case of ARL based mean control charts in terms of $ARL_1$. These outcomes can be interpreted as detection ability of $\bar{X}_U$ control chart is similar to the $\bar{X}_L$ control chart when in-control process mean is shifted to new level with same magnitude of distance.
- The detection ability of FAR based mean control charts are observed uniformly higher for all choices of shifts ($\delta_1 > 0$ and $\delta_2 < 0$) in terms of $P_1$ when additional runs rules are employed as compared to 1/1 runs rule (see Tables 2–3). In continuation, detection ability of ARL based mean control charts are found higher for only small-to-moderate shifts (e.g. $0 < \delta_1 < 1$) in terms of $ARL_1$ when additional runs rules are implemented relative to classical runs rule. This implies that ARL based mean control charts are efficient towards detection of small-to-moderate shifts when additional runs rules are considered, and also efficient for large shifts when classical runs rule is incorporated.
- There are relationships between detection ability and choices of $k/k$, $k/k+r$, $n$, $\delta_1$ and $\delta_2$ (see Tables 2–3). For instance, detection ability of FAR based mean control charts uniformly increase as value of $k/k$ increases. This remains valid for all choices of $n$, $\delta_1$ and $\delta_2$. Also, detection ability of ARL based mean control charts increase as value of $k/k$ increases.
- Among variant choices of runs rules, the 3/4 with mean control charts is proved efficient towards detection of small-to-moderate shifts relative to the other choices. Also, based on the detection ability in terms of $ARL_1$ and $P_1$, performance order of runs rules with mean control charts is 3/4, 3/3, 2/4, 2/2, 2/3, and 1/1.

**Table 1.** $\alpha_{act}$ and $ARL_{act}$ at $\alpha = 0.0027$, $ARL_0 = 370$, $\delta_1 = 0$, $\delta_2 = 0$, $\delta_3 = 1$, $\delta_4 = 1$, $d^* = 1$, $d = 0$, $k/k$ and $k/k+r$

| | 1/1 | | 2/2 | | 3/3 | | 2/3 | | 2/4 | | 3/4 | |
|---|---|---|---|---|---|---|---|---|---|---|---|---|
| | $\alpha_{act}$ | $ARL_{act}$ | $\alpha_{act}$ | $ARL_{act}$ | $\alpha_{act}$ | $ARL_{act}$ | $\alpha_{act}$ | $ARL_{act}$ | $\alpha_{act}$ | $ARL_{act}$ | $\alpha_{act}$ | $ARL_{act}$ |
| $\bar{X}_U$ | 0.0027 | 370.37 | 0.0027 | 370.17 | 0.0027 | 370.14 | 0.0027 | 370.18 | 0.0027 | 370.10 | 0.0027 | 370.15 |
| $\bar{X}_L$ | 0.0027 | 370.37 | 0.0027 | 370.17 | 0.0027 | 370.13 | 0.0027 | 370.43 | 0.0027 | 370.53 | 0.0027 | 370.10 |
| $S_U^2$ | 0.0027 | 370.17 | 0.0027 | 370.60 | 0.0027 | 370.40 | 0.0027 | 370.63 | 0.0027 | 370.13 | 0.0027 | 370.72 |
| $S_L^2$ | 0.0027 | 370.17 | 0.0027 | 370.17 | 0.0027 | 370.14 | 0.0027 | 370.18 | 0.0027 | 370.10 | 0.0027 | 370.43 |
| $S_U$ | 0.0027 | 370.23 | 0.0027 | 370.31 | 0.0027 | 370.13 | 0.0027 | 370.43 | 0.0027 | 370.53 | 0.0027 | 371.20 |
| $S_L$ | 0.0027 | 370.21 | 0.0027 | 370.28 | 0.0027 | 370.40 | 0.0027 | 370.63 | 0.0027 | 370.13 | 369.71 | 372.42 |
| $\lvert S\rvert_U$ | 0.0027 | 370.25 | 0.0027 | 370.17 | 0.0027 | 370.14 | 0.0027 | 370.18 | 0.0027 | 370.10 | 0.0027 | 371.31 |
| $\lvert S\rvert_L$ | 0.0027 | 370.37 | 0.0027 | 370.17 | 0.0027 | 370.13 | 0.0027 | 370.43 | 0.0027 | 370.41 | 0.0027 | 372.31 |
| $T_U^2$ | 0.0027 | 370.37 | 0.0027 | 370.11 | 0.0027 | 370.40 | 0.0027 | 370.63 | 0.0027 | 370.20 | 0.0028 | 371.25 |

### 3.2. $S_U^2$, $S_L^2$, $S_U$, $S_L$, $|S|_L$ and $|S|_U$ control charts

The mechanism for computing $P_1$ and $ARL_1$ of $S_L^2$ and $S_U^2$ control charts is similar to $\bar{X}_L$ and $\bar{X}_U$ control charts except in-control mean is stable $\mu_0$, whereas in-control variance $\sigma_0^2$ is out-of-control, that is, $\sigma_1^2 = (\delta_3\sigma_0)^2$ and $\sigma_1^2 = (\delta_4\sigma_0)^2$, where $\delta_3 \geq 1$ and $\delta_4 \leq 1$ are upward and downward shift. Likewise, for $S_U$ and $S_L$ control charts, assume that the in-control mean is stable, whereas standard deviation is out-of-control $\sigma_1 = \delta_3\sigma_0$ and $\sigma_1 = \delta_4\sigma_0$. Besides, procedures for computing the power and out-of-control average run length of $|S|_L$ and $|S|_U$ control charts is to assume the $\underline{\mu_0}$ is stable, whereas $\Sigma_0$ is out-of-control

**Table 2.** $P_1$ and $ARL_1$ of $\bar{X}_U$ control chart at $n = 5$, $\alpha = 0.0027$, $ARL_0 = 370$, $k/k$, $k/k + r$ and $\delta_1$

| | $\bar{X}_U$ | | | | | | | | | | | |
|---|---|---|---|---|---|---|---|---|---|---|---|---|
| | 1/1 | | 2/2 | | 3/3 | | 2/3 | | 2/4 | | 3/4 | |
| $\delta_1$ | $P_1$ | $ARL_1$ | $P_1$ | $ARL_1$ | $P_1$ | $ARL_1$ | $P_1$ | $ARL_1$ | $P_1$ | $ARL_1$ | $P_1$ | $ARL_1$ |
| 0 | 0.0027 | 370.37 | 0.0027 | 370.76 | 0.0027 | 370.03 | 0.0026 | 372.02 | 0.0028 | 371.70 | 0.0027 | 369.47 |
| 0.05 | 0.0038 | 263.16 | 0.0042 | 240.47 | 0.0045 | 228.97 | 0.0043 | 243.52 | 0.0047 | 233.21 | 0.0048 | 225.25 |
| 0.1 | 0.0053 | 188.68 | 0.0065 | 159.79 | 0.0074 | 146.52 | 0.0069 | 155.64 | 0.0077 | 151.97 | 0.0082 | 136.01 |
| 0.15 | 0.0072 | 138.89 | 0.0097 | 108.75 | 0.0117 | 96.92 | 0.0107 | 106.06 | 0.0122 | 102.20 | 0.0136 | 89.33 |
| 0.2 | 0.0098 | 102.04 | 0.0142 | 75.73 | 0.018 | 66.16 | 0.0164 | 71.05 | 0.019 | 69.54 | 0.0219 | 59.05 |
| 0.25 | 0.0131 | 76.34 | 0.0204 | 53.9 | 0.027 | 46.6 | 0.0244 | 50.62 | 0.0288 | 48.99 | 0.034 | 41.87 |
| 0.3 | 0.0174 | 57.47 | 0.0288 | 39.23 | 0.0392 | 33.84 | 0.0355 | 36.90 | 0.0426 | 34.92 | 0.0513 | 29.75 |
| 0.35 | 0.0228 | 43.86 | 0.0398 | 29.19 | 0.0556 | 25.28 | 0.0505 | 27.64 | 0.0615 | 26.00 | 0.0747 | 21.88 |
| 0.4 | 0.0295 | 33.9 | 0.0539 | 22.19 | 0.0767 | 19.41 | 0.0702 | 20.55 | 0.0863 | 19.74 | 0.1056 | 17.12 |
| 1 | 0.2925 | 3.42 | 0.5316 | 3.22 | 0.6708 | 3.88 | 0.7029 | 3.11 | 0.8065 | 3.14 | 0.8406 | 3.611 |

**Table 3.** $P_1$ and $ARL_1$ of $\bar{X}_L$ control chart at $n = 5$, $\alpha = 0.0027$, $ARL_0 = 370$, $k/k$, $k/k + r$ and $\delta_2$

| | $\bar{X}_L$ | | | | | | | | | | | |
|---|---|---|---|---|---|---|---|---|---|---|---|---|
| | 1/1 | | 2/2 | | 3/3 | | 2/3 | | 2/4 | | 3/4 | |
| $\delta_2$ | $P_1$ | $ARL_1$ | $P_1$ | $ARL_1$ | $P_1$ | $ARL_1$ | $P_1$ | $ARL_1$ | $P_1$ | $ARL_1$ | $P_1$ | $ARL_1$ |
| 0 | 0.0027 | 370.37 | 0.0027 | 370.76 | 0.0027 | 370.03 | 0.0026 | 370.15 | 0.0028 | 369.27 | 0.0027 | 367.20 |
| -0.05 | 0.0038 | 263.16 | 0.0042 | 240.47 | 0.0045 | 228.97 | 0.0043 | 239.73 | 0.0047 | 232.81 | 0.0048 | 219.60 |
| -0.1 | 0.0053 | 188.68 | 0.0065 | 159.79 | 0.0074 | 146.52 | 0.0069 | 158.33 | 0.0077 | 148.88 | 0.0082 | 138.42 |
| -0.15 | 0.0072 | 138.89 | 0.0097 | 108.75 | 0.0117 | 96.92 | 0.0107 | 106.96 | 0.0122 | 101.41 | 0.0136 | 89.84 |
| -0.2 | 0.0098 | 102.04 | 0.0142 | 75.73 | 0.018 | 66.16 | 0.0164 | 71.94 | 0.019 | 68.35 | 0.0219 | 59.87 |
| -0.25 | 0.0131 | 76.34 | 0.0204 | 53.9 | 0.027 | 46.6 | 0.0244 | 51.56 | 0.0288 | 48.12 | 0.034 | 42.35 |
| -0.3 | 0.0174 | 57.47 | 0.0288 | 39.23 | 0.0392 | 33.84 | 0.0355 | 37.01 | 0.0426 | 34.98 | 0.0513 | 30.04 |
| -0.35 | 0.0228 | 43.86 | 0.0398 | 29.19 | 0.0556 | 25.28 | 0.0505 | 27.25 | 0.0615 | 26.00 | 0.0747 | 22.00 |
| -0.4 | 0.0295 | 33.9 | 0.0539 | 22.19 | 0.0767 | 19.41 | 0.0702 | 20.38 | 0.0863 | 20.11 | 0.1056 | 16.98 |
| -1 | 0.2925 | 3.42 | 0.5316 | 3.22 | 0.6708 | 3.88 | 0.7029 | 3.1292 | 0.8065 | 3.12 | 0.8406 | 3.60 |

(say $\Sigma_1$), that is,

$$\underline{\mu_0} = \begin{bmatrix} \mu_{10} \\ \mu_{20} \end{bmatrix} \quad \text{and} \quad \Sigma_1 = \begin{bmatrix} \delta_5^2\sigma_{10}^2 & \delta_5\delta_6\rho\sigma_{10}\sigma_{20} \\ \delta_5\delta_6\rho\sigma_{10}\sigma_{20} & \delta_6^2\sigma_{20}^2 \end{bmatrix},$$

where $\delta_5^2 \geq 1$ and $\delta_6^2 \geq 1$ are amount of shifts in the in-control variances ($\sigma_{10}^2$ and $\sigma_{20}^2$), $\rho$ is the amount of correlation between $Y_1$ and $Y_2$. After that, generate random sample from bivariate normal distribution with $\underline{\mu_0}$ and $\Sigma_1$ followed by calculating the $|S|_U$ and comparing with the control limit ($U_{|S|}$ or $L_{|S|}$) to decide whether the process is in-control or out-of-control. Rest of the steps for computing $P_1$ and $ARL_1$ of $|S|_L$ and $|S|_U$ control charts are identical to $\bar{X}_L$ and $\bar{X}_U$ control charts. It is worthy to mention that detection ability of $|S|_L$ and $|S|_U$ control charts are dependent on the product of shifts $d^{*2} = \delta_5^2\delta_6^2$ and $n$ in respective of the choice of other quantities such as $\underline{\mu_0}$, and $\Sigma_1$. This property is termed as invariance property. Therefore, one may consider the product value of shift instead of assuming each shift separately. For comparative purpose, we have obtained $\alpha_{act}$, $ARL_{act}$, $P_1$ and $ARL_1$ of $S_U^2$, $S_L^2$, $S_U$, $S_L$, $|S|_L$ and $|S|_U$ control charts at $\alpha = 0.0027$, $ARL_0 = 370$, various choices of $k/k$, $k/k + r$, $\delta_3$, $\delta_4$ and $d^*$ (see Tables 4–9). Note that $\alpha_{act}$ and $ARL_{act}$ is the special case of $P_1$ and $ARL_1$, respectively when $\delta_3 = \delta_4 = d^* = 1$. Now discussions on the behavior of $S_U^2$, $S_L^2$, $S_U$, $S_L$, $|S|_L$ and $|S|_U$ control charts are given in the following points:

- The $\alpha_{act}$ and $ARL_1$ of $S_U^2$, $S_L^2$, $S_U$, $S_L$, $|S|_L$ and $|S|_U$ control charts are obtained equal to prefix values of $\alpha$ and $ARL_0$ (i.e. $\alpha_{act} = \alpha = 0.0027$ and $ARL_{act} = ARL_0 = 370$) for classical and additional runs rules (see Table 1).

- The detection ability of FAR based $S_U^2$ and $S_U$ control charts uniformly increases for small $n$ (e.g. $n < 5$) in terms of $P_1$ as value of $k/k$ increases. In comparison, detection ability of ARL based $S_U^2$ and $S_U$ control charts decreases for small $n$ in terms of $ARL_1$ as value of $k/k$ increases. This may illustrate as the $k/k$ runs rules are useful for FAR based $S_U^2$ and $S_U$ control charts at any choice of $n$ relative to 1/1 runs rule, whereas $k/k$ runs rules are not beneficial for ARL based $S_U^2$ and $S_U$ control charts when $n$ is small. However, for $n \geq 5$, detection ability of ARL based $S_U^2$ and $S_U$ control charts with 2/2 and 3/3 runs rules are seen higher at wide range of shifts relative to 1/1 runs rule (see Tables 4 & 6). Between runs rules, 2/2 results in higher detection ability of ARL based $S_U^2$ and $S_U$ control charts as compared to 3/3.
- The diagnosing ability of FAR based $|S|_U$ control chart uniformly increases for small $n$ (e.g. $n < 5$) in terms of $P_1$ as value of $k/k$ increases. In contrast, detection ability of ARL based $|S|_U$ control chart reduces for small $n$ (e.g. $n < 5$) in terms of $ARL_1$ as $k/k$ increases. This may illustrate as $k/k$ runs rules are useful for FAR based $|S|_U$ control chart relative to 1/1 runs rule at any choice of $n$, whereas $k/k$ runs rules are not useful for ARL based $|S|_U$ control chart when $n$ is small. However, for $n \geq 5$, detection ability of ARL based $|S|_U$ control chart under $k/k$ runs rules is seen higher than 1/1 runs rule (see Table 8) at various choices of shifts ($1 < d^* < 1.50$). Among $k/k$ runs rules, 3/3 results in highest detection ability of ARL based $|S|_U$ control chart for $1 < d^* \leq 1.20$ relative to 2/2. Similarly, 2/2 results into highest detection ability of ARL based $|S|_U$ control chart for $1.20 < d^* \leq 2.5$ relative to 3/3.
- The detection ability of FAR based $S_L^2$, $S_L$, and $|S|_L$ control charts are uniformly higher when additional runs rules are applied relative to classical runs rule (see Tables 5,7 & 9). Similarly, detection ability of ARL based $S_L^2$, $S_L$, and $|S|_L$ control charts are observed maximum for small-to-moderate shifts when additional runs rules are employed.
- The $n$, $\delta_3$, $\delta_4$ and $d^*$ have an effect on the detection ability of $S_U^2$, $S_L^2$, $S_U$, $S_L$, $|S|_L$ and $|S|_U$ control charts. In simple words, detection ability of FAR and ARL based control charts increases in terms of $P_1$ and $ARL_1$ as size of $n$, $\delta_3$, $\delta_4$ and/or $d^*$ increases (see Tables 4–9).
- At several choices of small-to-moderate shifts ($\delta_3$, $\delta_4$ and $d^*$), either 2/4 or 3/4 runs rule is proved efficient with dispersion control charts relative to $k/k$ runs rules in general. In terms of $ARL_1$ and $P_1$, performance order of various runs rules with dispersion control charts is as follows: 2/4, 3/4, 2/3, 2/2, 3/3, 1/1 when $S_U^2$ and $S_U$; 3/4 3/3, 2/2, 2/3 or 2/4, 1/1 when $S_L^2$ and $S_L$. Also, for $|S|_U$ and $|S|_L$ control charts, pattern of various runs rules are almost similar to $S_U^2$ and $S_L^2$ control charts.

**Table 4.** $P_1$ and $ARL_1$ of $S_U^2$ control chart at $n = 5$, $\alpha = 0.0027$, $ARL_0 = 370$, $k/k$, $k/k + r$ and $\delta_3$

| | $S_U^2$ | | | | | | | | | | | |
|---|---|---|---|---|---|---|---|---|---|---|---|---|
| | 1/1 | | 2/2 | | 3/3 | | 2/3 | | 2/4 | | 3/4 | |
| $\delta_3$ | $P_1$ | $ARL_1$ | $P_1$ | $ARL_1$ | $P_1$ | $ARL_1$ | $P_1$ | $ARL_1$ | $P_1$ | $ARL_1$ | $P_1$ | $ARL_1$ |
| 1 | 0.0027 | 370.37 | 0.0027 | 370.76 | 0.0027 | 370.02 | 0.0026 | 370.30 | 0.0028 | 369.33 | 0.0027 | 368.29 |
| 1.10 | 0.0096 | 104.13 | 0.0104 | 101.86 | 0.0109 | 103.72 | 0.0125 | 96.65 | 0.0149 | 88.38 | 0.0137 | 92.38 |
| 1.21 | 0.0255 | 39.261 | 0.029 | 39.24 | 0.0313 | 41.48 | 0.0398 | 34.36 | 0.0509 | 32.29 | 0.0452 | 34.14 |
| 1.32 | 0.0542 | 18.44 | 0.0632 | 19.36 | 0.0692 | 21.39 | 0.0946 | 17.09 | 0.1256 | 15.56 | 0.1086 | 17.26 |
| 1.44 | 0.0977 | 10.23 | 0.115 | 11.44 | 0.1263 | 13.21 | 0.1802 | 10.00 | 0.2417 | 9.42 | 0.2056 | 10.54 |
| 1.56 | 0.1552 | 6.44 | 0.1824 | 7.71 | 0.1998 | 9.28 | 0.2901 | 6.77 | 0.3853 | 6.50 | 0.3265 | 7.50 |
| 1.69 | 0.2235 | 4.47 | 0.2609 | 5.72 | 0.2841 | 7.14 | 0.4116 | 5.08 | 0.5335 | 4.91 | 0.4555 | 5.80 |
| 1.82 | 0.2985 | 3.35 | 0.3446 | 4.56 | 0.3724 | 5.87 | 0.5313 | 4.12 | 0.6665 | 3.988 | 0.5777 | 4.89 |
| 1.96 | 0.3757 | 2.66 | 0.4284 | 3.83 | 0.459 | 5.06 | 0.6391 | 3.48 | 0.7734 | 3.44 | 0.6835 | 4.31 |
| 4 | 0.9074 | 1.10 | 0.9302 | 2.11 | 0.9401 | 3.12 | 0.9941 | 2.08 | 0.9995 | 2.08 | 0.9958 | 3.06 |

**Table 5.** $P_1$ and $ARL_1$ of $S_L^2$ control chart at $n = 5$, $\alpha = 0.0027$, $ARL_0 = 370$, $k/k$ and $\delta_4$

| | $S_L^2$ | | | | | | | | | | | |
|---|---|---|---|---|---|---|---|---|---|---|---|---|
| | 1/1 | | 2/2 | | 3/3 | | 2/3 | | 2/4 | | 3/4 | |
| $\delta_4$ | $P_1$ | $ARL_1$ | $P_1$ | $ARL_1$ | $P_1$ | $ARL_1$ | $P_1$ | $ARL_1$ | $P_1$ | $ARL_1$ | $P_1$ | $ARL_1$ |
| 0.04 | 1 | 7 | 1 | 1 | 1 | 3 | 1 | 2 | 1 | 2 | 1 | 3 |
| 0.36 | 0.1158 | 8.63 | 0.5922 | 2.94 | 0.8862 | 3.21 | 0.66 | 3.11 | 0.7173 | 3.25 | 0.9442 | 3.21 |
| 0.42 | 0.0676 | 14.79 | 0.3642 | 4.31 | 0.6797 | 3.80 | 0.41 | 4.57 | 0.4567 | 4.80 | 0.7674 | 3.73 |
| 0.49 | 0.0401 | 24.96 | 0.1995 | 7.06 | 0.4285 | 5.17 | 0.2238 | 7.43 | 0.2494 | 7.79 | 0.4996 | 5.08 |
| 0.56 | 0.0242 | 41.27 | 0.1015 | 12.57 | 0.2269 | 8.17 | 0.1115 | 13.04 | 0.1239 | 13.36 | 0.2646 | 7.98 |
| 0.64 | 0.015 | 66.75 | 0.0496 | 23.74 | 0.1053 | 14.75 | 0.0531 | 24.55 | 0.0586 | 24.92 | 0.1202 | 14.64 |
| 0.72 | 0.0095 | 105.62 | 0.0238 | 46.47 | 0.0447 | 29.78 | 0.0249 | 47.61 | 0.0272 | 46.70 | 0.0494 | 29.39 |
| 0.81 | 0.0061 | 163.65 | 0.0114 | 92.65 | 0.0179 | 65.52 | 0.0116 | 93.81 | 0.0126 | 93.50 | 0.0192 | 64.05 |
| 0.90 | 0.004 | 248.51 | 0.0055 | 185.72 | 0.007 | 152.93 | 0.0055 | 186.30 | 0.0059 | 187.79 | 0.0072 | 152.44 |
| 1 | 0.0027 | 370.37 | 0.0027 | 370.76 | 0.0027 | 370.02 | 0.0026 | 377.00 | 0.0028 | 368.91 | 0.0027 | 371.57 |

**Table 6.** $P_1$ and $ARL_1$ of $S_U$ control chart at $n = 5$, $\alpha = 0.0027$, $ARL_0 = 370$, $k/k$, $k/k + r$ and $\delta_3$

| | $S_U$ | | | | | | | | | | | |
|---|---|---|---|---|---|---|---|---|---|---|---|---|
| | 1/1 | | 2/2 | | 3/3 | | 2/3 | | 2/4 | | 3/4 | |
| $\delta_3$ | $p_1$ | $ARL_1$ | $p_1$ | $ARL_1$ | $p_1$ | $ARL_1$ | $p_1$ | $ARL_1$ | $p_1$ | $ARL_1$ | $p_1$ | $ARL_1$ |
| 1 | 0.0027 | 370.37 | 0.0027 | 370.76 | 0.0027 | 370.03 | 0.0026 | 373.71 | 0.0028 | 368.81 | 0.0027 | 368.77 |
| 1.05 | 0.0053 | 189.71 | 0.0055 | 186.51 | 0.0057 | 187.42 | 0.006 | 178.51 | 0.0068 | 171.72 | 0.0064 | 174.27 |
| 1.1 | 0.0094 | 106.93 | 0.0101 | 104.6 | 0.0106 | 106.43 | 0.0121 | 97.71 | 0.0144 | 90.02 | 0.0132 | 93.83 |
| 1.15 | 0.0153 | 65.22 | 0.017 | 64.13 | 0.0182 | 66.31 | 0.0219 | 58.28 | 0.027 | 54.83 | 0.0245 | 55.96 |
| 1.2 | 0.0235 | 42.49 | 0.0267 | 42.32 | 0.0288 | 44.57 | 0.0363 | 37.09 | 0.0462 | 34.82 | 0.0412 | 36.70 |
| 1.25 | 0.0342 | 29.25 | 0.0393 | 29.68 | 0.0428 | 31.87 | 0.056 | 25.72 | 0.0729 | 24.12 | 0.064 | 26.12 |
| 1.3 | 0.0474 | 21.09 | 0.055 | 21.89 | 0.0602 | 23.98 | 0.0814 | 19.12 | 0.1075 | 17.64 | 0.0933 | 19.00 |
| 1.35 | 0.0632 | 15.82 | 0.0739 | 16.85 | 0.0811 | 18.81 | 0.1123 | 14.67 | 0.1498 | 13.64 | 0.1288 | 15.26 |
| 1.4 | 0.0815 | 12.27 | 0.0956 | 13.43 | 0.1051 | 15.28 | 0.1482 | 11.62 | 0.1987 | 10.93 | 0.1696 | 12.32 |
| 2 | 0.3976 | 2.52 | 0.4515 | 3.68 | 0.4828 | 4.89 | 0.6667 | 3.36 | 0.7986 | 3.31 | 0.7099 | 4.17 |

**Table 7.** $P_1$ and $ARL_1$ of $S_L$ control chart at $n = 5$, $\alpha = 0.0027$, $ARL_0 = 370$, $k/k$, $k/k + r$ and $\delta_4$

| | $S_L$ | | | | | | | | | | | |
|---|---|---|---|---|---|---|---|---|---|---|---|---|
| | 1/1 | | 2/2 | | 3/3 | | 2/3 | | 2/4 | | 3/4 | |
| $\delta_4$ | $P_1$ | $ARL_1$ | $P_1$ | $ARL_1$ | $P_1$ | $ARL_1$ | $P_1$ | $ARL_1$ | $P_1$ | $ARL_1$ | $P_1$ | $ARL_1$ |
| 0.2 | 0.5616 | 1.78 | 0.9977 | 2 | 1 | 3 | 0.9997 | 2.00 | 0.9999 | 2.016 | 1 | 3.00 |
| 0.6 | 0.0191 | 52.43 | 0.0716 | 17.09 | 0.1577 | 10.76 | 0.0777 | 17.48 | 0.086 | 17.67 | 0.1824 | 10.75 |
| 0.65 | 0.0141 | 70.76 | 0.0453 | 25.78 | 0.0951 | 16.01 | 0.0483 | 27.41 | 0.0533 | 26.43 | 0.1081 | 15.43 |
| 0.7 | 0.0107 | 93.65 | 0.029 | 38.76 | 0.0566 | 24.46 | 0.0305 | 39.11 | 0.0334 | 40.06 | 0.0631 | 24.44 |
| 0.75 | 0.0082 | 121.81 | 0.0188 | 57.9 | 0.0335 | 38.05 | 0.0195 | 59.30 | 0.0212 | 59.07 | 0.0366 | 36.74 |
| 0.8 | 0.0064 | 156.02 | 0.0124 | 85.76 | 0.0199 | 59.81 | 0.0127 | 87.37 | 0.0137 | 86.16 | 0.0213 | 58.85 |
| 0.85 | 0.0051 | 197.07 | 0.0083 | 125.74 | 0.0119 | 94.52 | 0.0084 | 127.79 | 0.009 | 127.99 | 0.0125 | 93.10 |
| 0.9 | 0.0041 | 245.86 | 0.0056 | 182.36 | 0.0072 | 149.48 | 0.0056 | 185.15 | 0.006 | 184.65 | 0.0074 | 148.47 |
| 0.95 | 0.0033 | 303.3 | 0.0039 | 261.5 | 0.0044 | 235.78 | 0.0038 | 264.10 | 0.0041 | 264.93 | 0.0044 | 237.43 |
| 1 | 0.0027 | 370.37 | 0.0027 | 370.76 | 0.0027 | 370.03 | 0.0026 | 373.48 | 0.0028 | 372.11 | 0.0027 | 367.01 |

**Table 8.** $P_1$ and $ARL_1$ of $|S|_U$ control chart at $n = 5$, $\alpha = 0.0027$, $ARL_0 = 370$, $k/k$, $k/k+r$ and $d^*$

| | $\|S\|_U$ | | | | | | | | | | | |
|---|---|---|---|---|---|---|---|---|---|---|---|---|
| | 1/1 | | 2/2 | | 3/3 | | 2/3 | | 2/4 | | 3/4 | |
| $d^*$ | $P_1$ | $ARL_1$ | $P_1$ | $ARL_1$ | $P_1$ | $ARL_1$ | $P_1$ | $ARL_1$ | $P_1$ | $ARL_1$ | $P_1$ | $ARL_1$ |
| 1 | 0.0027 | 370.37 | 0.0027 | 370.76 | 0.0027 | 370.03 | 0.0026 | 370 | 0.0028 | 370 | 0.0027 | 370 |
| 1.05 | 0.0040 | 250.79 | 0.0042 | 244.81 | 0.0043 | 243.16 | 0.0043 | 238.69 | 0.0048 | 239.53 | 0.0046 | 234.54 |
| 1.1 | 0.0057 | 176.56 | 0.0061 | 169.22 | 0.0064 | 167.88 | 0.0067 | 161.21 | 0.0076 | 155.46 | 0.0073 | 156.97 |
| 1.15 | 0.0078 | 128.55 | 0.0086 | 121.66 | 0.0092 | 120.91 | 0.01 | 116.82 | 0.0116 | 107.89 | 0.0111 | 109.15 |
| 1.2 | 0.0104 | 96.37 | 0.0118 | 90.51 | 0.0128 | 90.29 | 0.0142 | 84.36 | 0.017 | 79.37 | 0.0161 | 79.05 |
| 1.25 | 0.0135 | 74.130 | 0.0157 | 69.36 | 0.0172 | 69.58 | 0.0196 | 62.81 | 0.0238 | 58.89 | 0.0225 | 60.14 |
| 1.3 | 0.0171 | 58.31 | 0.0203 | 54.55 | 0.0225 | 55.1 | 0.0262 | 49.84 | 0.0324 | 46.04 | 0.0305 | 47.18 |
| 1.35 | 0.0214 | 46.8 | 0.0256 | 43.88 | 0.0287 | 44.67 | 0.0341 | 38.97 | 0.0429 | 37.42 | 0.0401 | 37.07 |
| 1.4 | 0.0262 | 38.23 | 0.0318 | 36.02 | 0.0358 | 36.98 | 0.0434 | 32.20 | 0.0552 | 30.25 | 0.0514 | 30.76 |
| 2 | 0.1234 | 8.11 | 0.1572 | 8.74 | 0.1805 | 10.01 | 0.2447 | 7.67 | 0.3238 | 7.40 | 0.2889 | 8.23 |

**Table 9.** $P_1$ and $ARL_1$ of $|S|_L$ control chart at $n = 5$, $\alpha = 0.0027$, $ARL_0 = 370$, $k/k$, $k/k+r$ and $d^*$

| | $\|S\|_L$ | | | | | | | | | | | |
|---|---|---|---|---|---|---|---|---|---|---|---|---|
| | 1/1 | | 2/2 | | 3/3 | | 2/3 | | 2/4 | | 3/4 | |
| $d^*$ | $P_1$ | $ARL_1$ | $P_1$ | $ARL_1$ | $P_1$ | $ARL_1$ | $P_1$ | $ARL_1$ | $P_1$ | $ARL_1$ | $P_1$ | $ARL_1$ |
| 0.2 | 0.1553 | 6.44 | 0.6141 | 2.87 | 0.8693 | 3.26 | 0.7119 | 2.93 | 0.7767 | 3.07 | 0.9426 | 3.22 |
| 0.6 | 0.0109 | 91.36 | 0.0265 | 42.13 | 0.0471 | 28.6 | 0.0287 | 43.38 | 0.0319 | 41.87 | 0.0542 | 27.44 |
| 0.65 | 0.0088 | 113.24 | 0.019 | 57.38 | 0.0318 | 39.98 | 0.0202 | 57.79 | 0.0224 | 57.09 | 0.0358 | 38.11 |
| 0.7 | 0.0072 | 138.38 | 0.0139 | 77.26 | 0.0216 | 55.79 | 0.0145 | 79.31 | 0.016 | 77.62 | 0.0239 | 53.49 |
| 0.75 | 0.006 | 167 | 0.0102 | 102.89 | 0.0149 | 77.58 | 0.0106 | 104.58 | 0.0116 | 102.30 | 0.0161 | 76.19 |
| 0.8 | 0.005 | 199.34 | 0.0077 | 135.57 | 0.0104 | 107.36 | 0.0078 | 139.23 | 0.0085 | 132.88 | 0.011 | 104.68 |
| 0.85 | 0.0042 | 235.62 | 0.0058 | 176.81 | 0.0073 | 147.72 | 0.0059 | 174.74 | 0.0063 | 172.15 | 0.0076 | 146.35 |
| 0.9 | 0.0036 | 276.07 | 0.0045 | 228.36 | 0.0052 | 201.98 | 0.0044 | 232.46 | 0.0048 | 228.26 | 0.0053 | 194.40 |
| 0.95 | 0.0031 | 320.91 | 0.0035 | 292.25 | 0.0037 | 274.33 | 0.0034 | 294.21 | 0.0037 | 286.50 | 0.0038 | 278.27 |
| 1 | 0.0027 | 370.37 | 0.0027 | 370.76 | 0.0027 | 370.03 | 0.0026 | 374.14 | 0.0028 | 369.12 | 0.0027 | 370.25 |

## 3.3. $T_U^2$ control chart

The procedure for computing the $P_1$ and $ARL_1$ of $T_U^2$ control chart is similar to $|S|_U$ and $|S|_U$ control charts except difference is at least one elements of $\underline{\mu_0}$ is shifted (say $\underline{\mu_1}$), whereas $\Sigma_0$ is stable, that is,

$$\underline{\mu_1} = \begin{bmatrix} \delta_7 \\ \delta_8 \end{bmatrix}, \ \delta_7 = \delta_8, \text{ and } \Sigma_0 = \begin{bmatrix} \sigma_{10}^2 & \rho\sigma_{10}\sigma_{20} \\ \rho\sigma_{10}\sigma_{20} & \sigma_{20}^2 \end{bmatrix}$$

where $\delta_7 \in \Re$ and $\delta_8 \in \Re$ represent amount of shift in the in-control process means, thats is, $\mu_{10}$ and $\mu_{20}$ respectively. After that, generate random sample from bivariate normal distribution with $\underline{\mu_0}$ and $\Sigma_1$ followed by calculating the $T_j^2$ and comparing with $U_{T^2}$ to decide whether the process is in-control or out-of-control. Rest of the steps for computing $P_1$ and $ARL_1$ of $T_U^2$ control chart are similar to $\bar{X}_U$ control chart. It is valuable to mention that detection ability of $T_U^2$ control chart is dependent on the Mahalanobis distance $d$, that is,

$$d = \sqrt{(\underline{\mu_1} - \underline{\mu_0})^t \Sigma_0^{-1} (\underline{\mu_0} - \underline{\mu_1})},$$

and $n$ in respective of the choice of other quantities ($\underline{\mu_1}$, and $\Sigma_0$). This property is termed as directional invariance [14]. Therefore, we have considered shift in form of $d$ as can be seen in many existing studies such as Mehmood et al. [7] and Pignatillo and Runger [14]. Also, for $\alpha = 0.0027$, $ARL_0 = 370$, and some choices of $n$, $d$, $k/k$ and $k/k+r$, we have provided $\alpha_{act}$, $ARL_{act}$, $P_1$ and $ARL_1$ in Table 10. Similarly one may proceed for other choices of $\alpha$, $ARL_0$, $n$, $d$, $k/k$ and $k/k+r$. Furthermore, results are described in following points:

- The $\alpha_{act}$ and $ARL_1$ of $T_U^2$ control chart are determined equal to $\alpha$ and $ARL_0$ (i.e. $\alpha_{act} = \alpha = 0.0027$ and $ARL_{act} = ARL_0 = 370$), respectively for classical and additional runs rules (see Table 1).
- The detection ability of FAR based $T_U^2$ control chart is uniformly outstanding at various choices of $d$ when additional runs rules are plugged relative to classical runs rule in general. In comparison, ARL based $T_U^2$ control chart is noted superior for small-to-moderate $d$ when additional runs rules are integrated relative to classical rule (see Table 10).
- The $n$ and $d$ are associated with the detection ability of $T_U^2$ control chart. It is summarized as detection ability of $T_U^2$ control chart in terms of $P_1$ and $ARL_1$ increases as size of $n$ and/or $d$ increases (see Table 10).
- The 2/4 runs rule is performed superb with $T_U^2$ control chart for detection of small-to-moderate $d$ relative to the other runs rules schemes. Also, performance order of various runs rules is as follows: 2/4 is ranked at 1st position followed by 3/4 at 2nd, 2/3 at 3rd, 3/3 at 4th, 2/2 at 5th, and 1/1 at last.

**Table 10.** $P_1$ and $ARL_1$ of $T_U^2$ control chart at $n = 5$, $\alpha = 0.0027$, $ARL_0 = 370$, $k/k$, $k/k+r$ and $d$.

| | $T_U^2$ | | | | | | | | | | | |
|---|---|---|---|---|---|---|---|---|---|---|---|---|
| | 1/1 | | 2/2 | | 3/3 | | 2/3 | | 2/4 | | 3/4 | |
| $d$ | $P_1$ | $ARL_1$ | $P_1$ | $ARL_1$ | $P_1$ | $ARL_1$ | $P_1$ | $ARL_1$ | $P_1$ | $ARL_1$ | $P_1$ | $ARL_1$ |
| 0 | 0.0027 | 370.37 | 0.0027 | 370.76 | 0.0027 | 370.03 | 0.0026 | 374.76 | 0.0028 | 367.76 | 0.0027 | 367.97 |
| 0.4 | 0.0094 | 106.89 | 0.011 | 97.04 | 0.0119 | 96.42 | 0.0173 | 72.709 | 0.0208 | 67.74 | 0.0198 | 69.14 |
| 0.5 | 0.0149 | 67.11 | 0.019 | 57.89 | 0.0215 | 57.15 | 0.0332 | 41.84 | 0.0411 | 37.78 | 0.0403 | 38.57 |
| 0.6 | 0.0236 | 42.34 | 0.0327 | 35.04 | 0.0387 | 34.54 | 0.0614 | 24.60 | 0.0778 | 22.67 | 0.0781 | 22.74 |
| 0.8 | 0.056 | 17.85 | 0.0884 | 14.35 | 0.1113 | 14.52 | 0.1756 | 9.74 | 0.2265 | 9.26 | 0.2324 | 9.31 |
| 0.9 | 0.0829 | 12.06 | 0.1362 | 9.86 | 0.1739 | 10.25 | 0.2682 | 6.85 | 0.3434 | 6.72 | 0.352 | 6.76 |
| 1 | 0.1191 | 8.4 | 0.1999 | 7.12 | 0.2558 | 7.67 | 0.3817 | 5.9 | 0.4802 | 4.10 | 0.4893 | 5.56 |
| 1.1 | 0.1657 | 6.03 | 0.2792 | 5.4 | 0.3542 | 6.05 | 0.5076 | 4.99 | 0.6214 | 3.15 | 0.628 | 4.41 |
| 1.2 | 0.2233 | 4.48 | 0.3714 | 4.29 | 0.4627 | 5.01 | 0.6335 | 3.78 | 0.7494 | 2.1 | 0.7515 | 3.72 |
| 1.4 | 0.3681 | 2.72 | 0.5728 | 3.05 | 0.6767 | 3.86 | 0.8383 | 2.8 | 0.9206 | 2 | 0.9161 | 2.15 |

## 4. Real life example

In this section, we conduct a comparative analysis between FAR and ARL based control charts with runs rules by using the practical data sets. The purpose of comparative analysis with aid of practical data sets is to know whether the behavior of FAR and ARL based control charts remains similar as described in Section 3 using the statistical performance indicators. To achieve the purpose, we consider a data set from Alwan [1] which refers back to [17] containing the data on 204 consecutive measurement on the electrical resistance of insulation in megohms. The data set is normally distributed with mean=4498.076 and the standard deviation=328. Afterwards, we have developed a code in R language to implement the FAR and ARL based $\bar{X}_U$ control charts for $k/k = 1/1, 2/2, 3/3, \alpha = 0.0027$, and $ARL_0 = 370$ (see Figures 1–2).

The FAR based $\bar{X}_U$ control chart shows 3, 5 and 6 out-of-control signals for the 1/1, 2/2 and 3/3 runs rules, respectively (see Figure 1). It is worthy to mention that numbers of out-of-control signals given by ARL based $\bar{X}_U$ control chart with varying choices of runs rules are equal to the case of FAR based $\bar{X}_U$ control chart (see Figure 2). This indicates that behavior of FAR and ARL based $\bar{X}_U$ control charts are identical. Also, additional runs rules contributes towards detection of small and moderate variations. This comparative discussion is in accordance with the statistical results provided in Section 3.1. On the similar lines, one may attempt for the other choices of control charts.

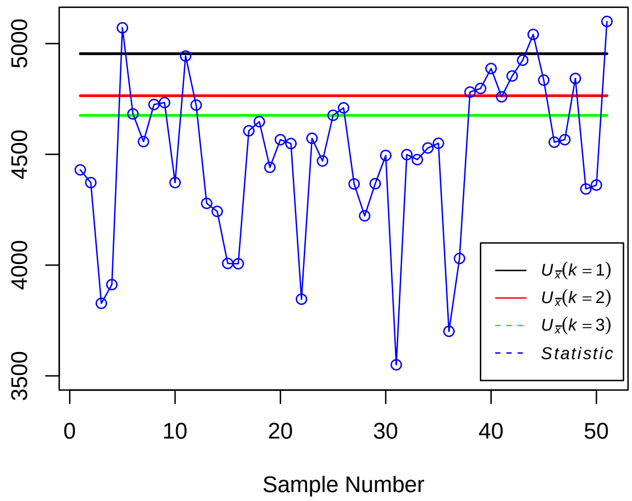



**Figure 1.** FAR based $\bar{X}_U$ control chart for varying choices of runs rules ($k = 1, 2, 3$) and $\alpha = 0.0027$

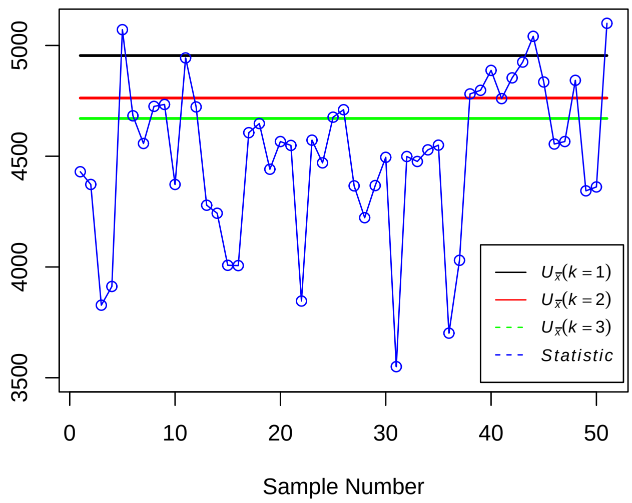



**Figure 2.** ARL based $\bar{X}_U$ control chart for varying choices of runs rules ($k = 1, 2, 3$) and $ARL_0 = 370$

## 5. Summary, conclusions and future recommendations

In this article, we have described comparative behavior of false alarm rate (FAR) and average run length (ARL) based control charts with runs rules. In the list of univariate and multivariate control charts, we have included upper-sided and lower-sided mean, variance, standard deviation, generalized variance, and Hotelling's. For comparative analysis and discussions, we have included actual false alarm rate, power, in-control actual average run length, and out-of-control average run length as performance measures. Furthermore, performance measures are computed by using Monte Carlo simulation procedures as computation methodology. Besides, diverse results are presented by taking into account numbers of factors. The detection ability of FAR based lower-sided and upper-sided control charts are remained uniformly higher when additional runs rule are incorporated relative to classical runs rule. Also, detection ability of ARL based lower-sided control charts are recorded outstanding for small-to-moderate shifts when additional runs rules are employed relative to classical runs rules. In brief, performance order of decision rules with FAR based lower-sided and upper-sided control charts are persistent, whereas performance order of decision rules with ARL based control charts are dependent on the circumstances, that is, sample size, size of shift, class of control chart (location and dispersion), and side of control limit (upper-sided and lower-sided). Lastly, we have provided a real life example using the data on electrical resistance of insulation. In the real life example, we have recorded that behavior of FAR and ARL based control charts using the real data sets are similar to the behavior using the statistical measures.

The scope of current study covers the processes in which characteristics follows normal distribution and parameters are known. It is often that process distribution is non-normal or unknown, and parameters are unknown. Therefore, it would be excellent to conduct an efficient study in future for non-normal distribution and parameters are unknown. Likewise, one may contribute a study by involving robust techniques (e.g. robust estimators and non-parametric) with control charts. An interesting study can be added on the topic of comparative analysis between cumulative sum (CUSUM) and exponentially weighted moving average (EWMA) control charts with runs rules.

**Acknowledgment.** This study was partially funded by the Ministry of Higher Education, Malaysia, through the UTM Encouragement Research Grant (grant number 19J38).

## References

[1] L.C. Alwan, *Statistical Process Analysis*, McGraw-Hill International Editions, Singapore, 2000.

[2] S. Chakraborti and S. Eryilmaz, *A nonparametric Shewhart-type signed-rank control chart based on runs*, Comm. Statist. Simulation Comput. **36** (2), 335–356, 2007.

[3] C.W. Champ and W.H. Woodall, *Exact results for Shewhart control charts with supplementary runs rules*, Technometrics **29** (4), 393–399, 1987.

[4] M.B. Khoo, *Design of runs rules schemes*, Qual. Eng. **16** (1), 27–43, 2003.

[5] M. Klein, *Two alternatives to the shewhart x control chart*, J. Qual. Technol. **32** (4), 427–431, 2000.

[6] J.C. Malela-Majika, S.C. Shongwe and P. Castagliola, *One-sided precedence monitoring schemes for unknown shift sizes using generalized 2-of-(h+1) and w-of-w improved runs-rules*, Comm. Statist. Theory Methods, 1–35, 2020.

[7] R. Mehmood, M.H. Lee, M. Riaz, B. Zaman and I. Ali, *Hotelling $T^2$ control chart based on bivariate ranked set schemes*, Comm. Statist. Simulation Comput., 1–28, 2019.

[8] R. Mehmood, M.S. Qazi and M. Riaz, *On the performance of $\bar{X}$ control chart for known and unknown parameters supplemented with runs rules under different probability distributions*, J. Stat. Comput. Simul. **88** (4), 675–711, 2018.

[9] R. Mehmood, M. Riaz and R.J.M.M. Does, *Control charts for location based on different sampling schemes*, J. Appl. Stat. **40** (3), 483–494, 2013.

[10] R. Mehmood, M. Riaz and R.J.M.M. Does, *Efficient power computation for r out of m runs rules schemes*, Comput. Statist. **28** (2), 667–681, 2013.

[11] R. Mehmood, M. Riaz and R.J.M.M. Does, *Quality quandaries: on the application of different ranked set sampling schemes*, Qual. Eng. **26** (3), 370–378, 2014.

[12] R. Mehmood, M. Riaz, T. Mahmood, S.A. Abbasi and N. Abbas, *On the extended use of auxiliary information under skewness correction for process monitoring*, Trans. Inst. Meas. Control. **39** (6), 883–897, 2017.

[13] D.C. Montgomery, *Introduction to Statistical Quality Control*, John Wiley Sons, New York, 2009.

[14] Jr, J.J. Pignatiello and G.C. Runger, *Comparisons of multivariate cusum charts*, J. Qual. Technol. **22** (3), 173–186, 1990.

[15] M. Riaz, R. Mehmood and R.J.M.M. Does, *On the performance of different control charting rules*, Qual. Reliab. Eng. **27** (8), 1059–1067, 2011.

[16] D.K. Shepherd, S.E. Rigdon and C.W. Champ, *Using runs rules to monitor an attribute chart for a markov process*, QTQM **9** (4), 383–406, 2012.

[17] W.A. Shewhart, *Economic Control of Quality of Manufactured Product*, ASQ Quality Press, 1931.

[18] S.C. Shongwe, *On the design of nonparametric runs-rules schemes using the markov chain approach*, Qual. Reliab. Eng. **36** (5), 1604–1621, 2020.

[19] S.C. Shongwe, J.C. Malela-Majika and T. Molahloe, *One-sided runs rules schemes to monitor autocorrelated time series data using a first-order autoregressive model with skip sampling strategies*, Qual. Reliab. Eng. **35** (6), 1973–1997, 2019.

[20] S. Shongwe, J.C. Malela-Majika and E. Rapoo, *One-sided and two-sided w-of-w runs-rules schemes: An overall performance perspective and the unified run-length derivations*, J. Probab. Stat., 1–20, 2019.

[21] E.C. Western, *Statistical Quality Control Handbook*, Western Electric Company, Indianapolis, 1956.

[22] B. Zaman, M. Riaz and S.A. Abbasi, *On the efficiency of runs rules schemes for process monitoring*, Qual. Reliab. Eng. **32** (2), 663–671, 2016.

Hacettepe Journal of Mathematics & Statistics

Hacet. J. Math. Stat.
Volume 50 (1) (2021), 289 – 303
DOI : 10.15672/hujms.734212

Research Article

# Robust regression estimation and variable selection when cellwise and casewise outliers are present

Onur Toka*[1], Meral Çetin[1], Olcay Arslan[2]

[1] *Hacettepe University, Faculty of Science, Department of Statistics, Ankara, Turkey*
[2] *Ankara University, Faculty of Science, Department of Statistics, Ankara, Turkey*

**Abstract**

Two main issues regarding a regression analysis are estimation and variable selection in presence of outliers. Popular robust regression estimation methods are combined with variable selection methods to simultaneously achieve robust estimation and variable selection. However, recent works showed that the robust estimation methods used in those estimation and variable selection procedures are only resistant to the casewise (rowwise) outliers in the data. Therefore, since these robust variable selection methods may not be able to cope with cellwise outliers in the data, some extra care should be taken when cellwise outliers are present along with the casewise outliers. In this study, we proposed a robust estimation and variable selection method to deal with both cellwise and casewise outliers in the data. The proposed method has three steps. In the first step, cellwise outliers were identified, deleted and marked with NA sign in each explanatory variable. In the second step, the cells with NA signs were imputed using a robust imputation method. In the last step, robust regression estimation methods were combined with the variable selection method LASSO (Least Angle Solution and Selection Operator) to estimate the regression parameters and to select remarkable explanatory variables. The simulation results and real data example revealed that the proposed estimation and variable selection procedure perform well in the presence of cellwise and casewise outliers.

**Mathematics Subject Classification (2020).** 62F35 , 62F07, 62J07

**Keywords.** Robust variable selection, outliers, cellwise outlier, LASSO

## 1. Introduction

One of the challenging problems in a regression analysis is to obtain estimators for the regression parameters that are robust against outliers in data sets. Until recently, outliers are defined as the observations that are not follow the model of the majority of the data. In a regression analysis, there are two types of outliers. One type is the outliers that may occur in the response variable and the other type of outliers occur in exploratory variables, which are usually called leverage points. Compared to the outliers in response variable,

*Corresponding Author.
Email addresses: onur.toka@hacettepe.edu.tr (O. Toka), meral@hacettepe.edu.tr (M. Çetin), oarslan@ankara.edu.tr (O. Arslan)
Received: 08.05.2020; Accepted: 23.11.2020

outliers in exploratory variables have a much greater influence on classical estimation procedures. If $X_{n\times p}$ is the data matrix formed by using the observations on the explanatory variables (rows as cases and columns as variables) the outliers in explanatory variables are used to be considered as the entire cases that correspond to the entire rows of $X_{n\times p}$. These outliers are called as casewise or rowwise outliers. Most of the robust regression methods, which are proposed against Huber-Tukey contaminated model, proceed by downweighting the entire rows that are considered as outliers (in response and/or casewise). Note that, in practice, the Huber-Tukey contaminated model corresponds to the casewise outliers [2]. However, in recent years, it has been realized that the observations considered as casewise outliers may not be completely contaminated. These observations may only have few contaminated cells and the rest of the cells may contain important information. These type of outliers are called as cellwise outliers [20]. That is, the cellwise outlier is a cell-deviated observation, so only outlier in one observation and one variable at the same time. The cellwise outliers may be the result of an independent contaminated model (ICM) [2]. In the presence of cellwise outliers, using ordinary robust regression estimation methods (for example using high breakdown point regression estimation methods) may be caused some loss of information since those methods try to downweight the entire row without considering non-contaminated cells in the outlying observations. Therefore, in recent papers new robust regression estimation methods have been proposed to take some extra care if cellwise and casewise outliers are present [1,6,17]. Debruyne *et al.* [7] argued that these outliers identification tools can be a thrilling topics. In order to compare outlier detection methods in the presence of cellwise and casewise outliers, Unwin [25] plotted the O3 graph, new visualization technique which is coded in a new R package called "cellWise" [19].

Another challenging problem in a regression analysis is to select a group of remarkable explanatory variables. To this extend, many variable selection methods have been proposed [11,24,31]. However, the popular ones are the methods that combine estimation and selection procedures together. These combined methods are also very effective for the high dimensional data sets. In particular, these methods are used for the regression problems involving data sets that have number of dimensions greater than the number of observes. The LASSO proposed by [24] is the first method in this direction. After the definition of LASSO, many other methods such as SCAD and bridge have been proposed to carry on simultaneous estimation and variable selection in a regression problem. Since LASSO and the other variable selection methods are based on the classical methods the researchers have been developed robust versions of these methods by using robust regression methods instead of the classical ones [3,4,8,15,28]. Since, the popular robust methods are designed to deal with the casewise outliers the combined robust estimation and variable selection methods, such as robust LASSO and robust SCAD, can only deal with the casewise outliers. However, recent works [1,9] show that the popular robust estimation methods may not be very successful when cellwise outliers are present. Especially, if we have high dimensional data and if the number of observations is rather small relative to the dimension of the data downweighting entire rows as casewise outliers may cause loss of information. Instead of doing so, monitoring those outliers and taking care only the outlying cells may reduce loss of information and improve estimation procedure.

Therefore, in recent papers, researchers have started concerning cellwise outliers and have proposed robust methods to deal with the cellwise outliers along with the casewise outliers. Some of these works are as follows. Raymaekers and Rousseeuw [18] proposed new identification technique which is based on LASSO regression with a stepwise application of constructed cutoff values for cellwise outliers. Leung *et al.* [12] proposed robust regression estimation methods under cellwise and casewise outliers contamination. However, there are few proposals for the robust estimation and variable selection in the presence of cellwise and casewise outliers [14]. In this paper, we will consider the robust estimation and the variable selection in linear regression models when cellwise and casewise outliers are present in the data. Our proposal will have three steps. In the first step, we will try to identify the cellwise outliers in each explanatory variable. This will be done by independently monitoring each explanatory variable using outlier detection methods. After identifying cellwise outliers in each explanatory variable these outliers will be removed from the data and those cells will be marked by NA sign as it is done in [1, 13]. Then, in the second step, these cells will be regarded as missing observations and will be imputed by using the robust imputation method proposed by [5]. These two steps will make our explanatory data matrix as cellwise outliers free, but we may still have casewise outliers in the data. Finally, in the third step, we will combine robust regression estimation methods with LASSO, the variable selection method, to estimate the regression parameters and to select the remarkable explanatory variables without suffering from the casewise outliers. Our simulation results and real data example showed that the proposed estimation and selection method work well when casewise and cellwise outliers are possible in the data sets.

The rest of the paper is organized as follows. In Section 2 we will provide the details of the proposed method. In Section 3 the simulation and the real data examples will be given. The paper will be finalized with a conclusion section.

## 2. Three step robust regression estimation and variable selection in the presence of cellwise outliers

Consider the linear regression model

$$y_i = \alpha + \mathbf{x}_i^T \beta + \varepsilon_i, \quad i = 1, 2, 3, ..., n \tag{2.1}$$

where $y_i \in R$ is the response variable; $\mathbf{x}_i = (x_{i1}, x_{i2}, ..., x_{ip})^T$ is the $p-$dimensional vector of the explanatory variables; $\beta = (\beta_1, \beta_2, ..., \beta_p)^T$ is the vector of regression parameters in $R^p$; and $\varepsilon_i$'s are the iid random errors with zero mean, $\sigma^2$ variance and the distribution function $F$. Note that, distribution function $F$ is symmetric distributions. Without loss of generality, we assume that $\alpha = 0$ and consider the model

$$y_i = \mathbf{x}_i^T \beta + \varepsilon_i, \quad i = 1, 2, 3, ..., n \tag{2.2}$$

The regression equation given in Equation (2.2) can also be written in matrix notation as

$$Y = X\beta + \varepsilon \tag{2.3}$$

where $X_{n\times p}$ is the design matrix, $Y$ is the response vector, and $\varepsilon$ is the vector of $\varepsilon_i$. Throughout this study, $\beta_0 = (\beta_{01}, \beta_{02}, ..., \beta_{0p})^T$ denotes the true parameter vector and $\Omega \subset R^p$ will denote the parameter space.

In this paper, our main aim is to estimate the regression parameters and select the important regressors under cellwise and casewise contaminations. As we have already mentioned, the casewise outliers can be identified using robust methods [16, 21] and are easily dealt with using robust variable selection methods if the variable selection is a concern. All of these can be done using combined robust estimation and variable selection methods. However, extra care should be taken to detect the cellwise outliers since they are not identified by examining the whole data matrix $X$. Each explanatory variable, that is; each column of $X$ should be monitored to detect the cellwise outliers. Thus, before preforming estimation and variable selection each variable should be scanned in terms of cellwise outliers. As it is proposed by [1] and [13] after detecting the cellwise outlier, those cells should be imputed using robust imputation methods. Then, robust methods related to the problem of interests can be used to handle the casewise outliers. In the following subsections, starting from the identification of the cellwise outliers, we will describe the three steps of the proposed robust estimation and variable selection method when cellwise and casewise outliers are present.

### 2.1. Identifying cellwise outlier

Cellwise outlier (introduced in [2]) is not a big problem when the proportion of outliers compared to the sample size is not high. However, Alqallaf *et al.* [2] observed that even if there is a very small percent of outliers in every variables, but if the dimension of the data is large, popular robust estimators with high breakdown point will easily reach their possible breakdown point. In recent years, researchers have become aware of cellwise outliers and they have proposed several methods to deal with this problem. Most of the proposed methods first identifies the cellwise outliers and regard them as missing observations by changing them with NA sign [1, 9, 13]. That is, the outlier problem is transferred to a missing data problem. In order to obtain cellwise outliers, there is a new methodology which combines LASSO regression with a stepwise application of constructed cutoff values [18]. In this paper, following the same strategy, we will try to identify the cellwise outliers by using the outlier detection method described in [20]. First, we have to obtain robust estimates for the location and scale of each column. In this paper, we will use the sample median for location and MAD for scale. These estimates will be used as initial robust estimates to obtained the one-step M estimates for location and scale computed as

$$\begin{aligned} \hat{\mu}_M &= \frac{\sum w_i x_i}{\sum w_i} \\ \hat{\sigma}_M^2 &= \frac{1}{n}\sum w_i \left(x_i - \hat{\mu}_M\right)^2 \end{aligned} \tag{2.4}$$

where weights $w\left(t\right) = \frac{\rho'(t)}{t}$ are computed using Tukey biweight $\rho$ function. Note that, $w_i$ is weights for $i_{th}$ observation and $W$ is a diagonal weight matrix. After robust location and scale estimates are computed, each column will be standardized using these robust estimates. Let $z_i$ denote these standardized columns. Then, the observations $x_i$ will be considered as outliers if

$$|z_i| \geq \sqrt{\chi^2_{1,q}} \tag{2.5}$$

where $q$ is $q - th$ quantile of the chi-squared distribution. After screening all the columns and identifying all the cellwise outliers those cells will be replaced by NA signs, and hence the cellwise outlier problem will be transfered into the missing observation problem. This will be the first step of our proposed robust variable selection method. In next subsection we will describe the robust imputation algorithm to impute the observations that are flagged as NA.

### 2.2. Bypassing cellwise outlier: Robust imputation

After identifying cellwise outliers and replace them with NA, we have created a missing value problem. Thus, these missing values have to be imputed using some imputation methods. There are several procedures to deal with missing observations in the data. These procedures are classified according to the missingness patterns in the data. Cellwise outliers are considered as randomly occurred outliers. Therefore, deleting the cellwise outliers in the data causes the missingness case called as missing completely at random (MCAR). This type of missing data can be easily imputed using mean or median imputation method. In this paper we will use the robust imputation (ROBimpute) method proposed by [5]. Actually, the robust imputation method is a robust alternative to the sequential imputation (SEQimpute) method proposed by [26] and it can be summarized as follows. Let $X_c$ be the completely observed part and $X_m$ be the missing part of our explanatory data matrix $X$ which contains missing observations. $x^*$ be a row in $X_m$ defined as $x^* = \left[(x_m^*)^T (x_o^*)^T\right]^T$, where $x_m^*$ and $x_o^*$ are the missing and observed part of that row, respectively. As described in [26], let the matrix $C$ defined as in Equation (2.6) be the inverse of the covariance matrix of $X_c$ and let $X^*$ be $\left[X_c^T, x^*\right]^T$. Further, let $\bar{x}_c$ be the rowwise sample mean of the complete data. Now minimizing the equation given in Equation (2.7), which can be also written as in Equation (2.8), will be an estimate for $x^*$. After finding $x_m^*$ in $X^*$, it will be used instead of $x^*$ in $X^*$ to form new completed data. Then we have to take care the next missing observations. This procedure should be continued after all the missingness are imputed. The detailed information about SEQimpute can be found in [26].

$$C = \begin{bmatrix} C_{m,m} & C_{m,o} \\ C_{m,o}^T & C_{m,m} \end{bmatrix} \tag{2.6}$$

$$D\left(x^*\right) = \left(x^* - \bar{x}_c\right)^T \left(cov\left(X_c\right)\right)^{-1} \left(x^* - \bar{x}_c\right) \tag{2.7}$$

$$x_m^* = \left(\bar{x}_c\right)_m - \left(C_{m,m}\right)^{-1} C_{m,o}\left(x_o^* - \left(\bar{x}_c\right)_o\right) \tag{2.8}$$

However, since this SEQimpute algorithm is based on sample mean and sample covariance, it is not robust against the outliers in the whole dataset. Therefore, even a single outlier can badly ruin the algorithm and the imputed value for the missing observations will be far from the expected value. For this reason, robust alternative to the SEQimpute has been proposed in [5]. They use robust covariance estimator and the robust location estimator instead of sample mean and the sample covariance matrix. In particular, they use minimum covariance determinant (MCD) estimator as the covariance estimator and the sample median for the mean estimator. The rest of the imputation will be same as in the classical one described above. This imputation is called ROBimpute and the detail of the algorithm is found in [5]. In this paper we will use the ROBimpute to impute the missing cells that are created deleting the cellwise outliers.

### 2.3. Variable selection with robust LASSO

In this section, we will describe the third step of our proposal. Namely, we will explore the variable selection for the regression model using refined data. Variable selection methods are one of the most important part of modeling aspect. In particular, in regression methods, we are interested in the most important variables and the subsets of full model. Robust variable selection, such as LASSO, is the robust versions of the classical ones in the presence of outliers. In this paper, we used LASSO to carry on our variable selection. LASSO is a well known method which minimizes OLS loss function $\left(\boldsymbol{y} - X\boldsymbol{\beta}\right)^T\left(\boldsymbol{y} - X\boldsymbol{\beta}\right)$ under the restriction $\sum_{j=1}^{p} |\boldsymbol{\beta_j}| \leq t$. Hence, this minimization problem with respect to $\beta$ can be carried on using lagrange multiplier method. That is we have to minimize the following objective function,

$$Q_N = \left(\boldsymbol{y} - X\boldsymbol{\beta}\right)^T\left(\boldsymbol{y} - X\boldsymbol{\beta}\right) + \lambda \sum\nolimits_{j=1}^{p} |\boldsymbol{\beta_j}| \tag{2.9}$$

where $\lambda$ is regularization parameter.

Using LASSO, parameter estimation and variable selection can be simultaneously obtained. Since the classical LASSO is based on OLS criterion, the resulting estimators will be sensitive to the outliers, the robust version of LASSO have been proposed in literature [3, 27]. In robust versions, OLS loss functions have been replaced with robust version of loss functions such as Huber or Tukey $\rho$ functions.

Several algorithms have been proposed to obtain LASSO estimators. One of these algorithms to solve the robust LASSO problem is proposed by [28] and it is called using semi-smooth Newton coordinate descent (SNCD) algorithm. In this paper, we will use this algorithm to obtain robust LASSO estimates when we have outlier in y direction or we have heavy-tailed error distribution. The algorithm is provided in the same paper and it is available as R packages named "hqreg".

By using robust LASSO, we will get estimators that are resistant to the outliers in y direction. However, if we have casewise outliers in x direction, the robust LASSO obtained using Huber or Tukey $\rho$ functions will be badly affected from the casewise outliers in x direction. Therefore, we have to modified the robust LASSO method to deal with the casewise outliers.

Concerning the casewise outliers, we will use the MM regression estimation method proposed by [29]. The MM estimation method will be used as follows. We will first obtain the MM estimators for the regression parameters. Then, using these MM estimators, we will compute the weights $w_i$ for $i = 1, 2, ..., n$ for each observations using the weight function obtained from the Tukey $\rho$ function (see e.g. [16], page 30). Then, we will form $W = diag(w_1, ..., w_n)$ matrix and transform our $X$ and $Y$ using $W$ matrix as $X^* = W^{1/2}X$ and $Y^* = W^{1/2}Y$, respectively. Now we can apply classical LASSO to transform data to do variable selection.

Finally, these three steps can be combined to obtain robust parameter estimation and variable selection in the presence of cellwise and casewise outlier. The following algorithm will be used to carry on all of these procedures. In our simulation and real-data example, this algorithm will be implemented to demonstrate the performance of the proposed method. If it is followed from the algorithm, it will see that robust methods with robust imputation are preferred when there are both cellwise and casewise outliers. If there are only casewise outliers robust LASSO methods are preferred. If there are no outliers in dataset, classical LASSO method is preferred.

**Algorithm 1:** Variable Selection in the presence of cellwise and casewise outliers

**Starting of Algorithm.**
**Data** Obtain data (Generating data in simulation or use data from real world example)
**If** you suspect any Cellwise outliers, then Run
**STEP 1: Identification of Cellwise Outliers**
    **Loop 1.** $i = 1, 2, \cdots, p$ (For each regressors)
    Identify cellwise outliers using the procedure described in Section 2.1 and change them with NA
**End Loop 1**
**STEP 2: Robust Imputation of NA**
    **Loop 2.** $m = 1, 2, \cdots, M$ (For each NA)
    Impute the NA's by using robust imputation methods described in Section 2.2
**End Loop 2**
    **ElseIf** Any Casewise Outlier
**STEP 3: Robust Estimation and Variable Selection**
    Apply Robust LASSO described in Section 2.3
    **ElseIf** No Outlier
    **Apply LASSO**
**End If**
**End of Algorithm.**

## 3. Numerical studies

In the application part, we considered simulation study in R to compare the performance of variable selection methods in the presence of cellwise outliers. We considered the regression model given in Section 2. The explanatory variables were independently generated from the normal distribution $N(m, 1)$ with $m$ coming from discrete uniform distribution randomly between zero and five. In the simulation study, the dimension of the parameter vector was taken as 7,15 and 30 and the sample sizes were taken as 50, 100 and 250. For the regression model, we took the regression parameters as $[1,0,1,0,1,1,0]'$ for dimension 7. For the dimensions 15, we formed $\beta$ as follows: first five entries were taken as one and the others are zero. Similarly, for the dimension 30, the first 10 entries of $\beta$ were one and the rest of the entries are as zero. In the regression model, we used three different error distributions. We first took the standard normal distribution ($N(0,1)$) to explore the case without outliers in y-direction. The other two error distributions were $0.9N(0,1) + 0.1N(3,1)$ and $t_3$. With these distributions we guaranteed the outliers in y-direction. For the outliers in x direction, we generated randomly observations from $N(50,1)$ and combined these observations with the major part of the data.

In this simulation study, cellwise outliers were generated as follows. We first generated explanatory variables and form our $X$ matrix. Using missingmat() function in ForIMP R package (see [10]), we created missing observations which were completely at random and replaced the missing observations with NA signs. Now, we would apply three different imputation procedures to the $X$ matrix. First, we used ROBimpute method to robustly impute this missing observations. Second, we used SEQimpute method to impute the missing observation in classical way (for the functions for imputation given in [5, 26] are used). Finally, to have data with cellwise outliers, we imputed the NAs with the values calculated by $\max(x_i)+2\sigma_{x_i}$. In this ad-hoc method, we easily obtained cellwise outliers in simulated data. To sum up, we had three different $X$ matrices. One had cellwise outliers, the other ones had missingness wich were imputed by robust and the classical imputation methods. The proportion of cellwise outliers were 1%, 5% and 10%. Note that, when cellwise outliers were constructed, the proportion was calculated using $n \times p$, not just $n$. After we designed our data, we applied three different combination of LASSO methods using the glmnet [22] and hqregraw functions [28] in R. Note that, for the casewise outlier in x-direction we used glmnet function for the modified dataset described in previous section.

In the simulation results the methods were compared in three different ways. We randomly divided data in two subsections. We used one part for estimation and variable selection (training ; 80% of dataset) and the other part is testing (20 % of dataset). After we did estimation and variable selection, we counted the number of true zero- beta selection and we also calculated proportion of true model selection. Then, using the testing part of data, we computed the prediction error $\frac{1}{T}\sum_{i=1}^{T}\sum_{j=1}^{n}(y_j - \hat{y}_j)^2/n$ where $n$ is the number of observation and $T$ is the number of iteration in testing data. We also provided some boxplot illustrations for estimated betas.

The simulation results were summarized in Tables 1-5. Tables 1-3 contained prediction errors. In Table 1, we displayed the results for the case normally distributed errors with cellwise and casewise outliers for the sample size $n = 50$. If we only had cellwise outlier, we observed the smallest prediction error for the case robust imputed data using classical LASSO (ROB-LASSO) and sequentially imputed data using classical LASSO (SEQ-LASSO). Therefore, we could say that robust imputation gave a better estimation for cellwise outliers. We also observed that when the number of cellwise outliers increased, the prediction errors for LASSO and robust imputed LASSO also increased. Overall, ROB-LASSO and SEQ-LASSO had superiority over the other methods for this case. When casewise outliers were introduced to the data, we observed that robust imputed robust LASSO (ROB-RLASSO) seems better performance for most of the cases compared to the other methods.

In Table 2, we gave the simulation results for the contaminated error distribution and we observed similar behavior for ROB-RLASSO. That is, the results for the ROB-RLASSO was superior to the other methods. In Table 3, simulation results for $t_3$ distributed error case were summarized. Concerning this case, without casewise outlier ROB-RLASSO gave smaller prediction errors for almost all the cases. However, when the casewise outliers were introduced in the data, the performance of the ROB-RLASSO was getting worse compare to the robust LASSO (RLASSO).

**Table 1.** Prediction error for $n = 50$ and $\varepsilon \sim N(0, 1)$

| pr-casew | p | pr-cellw | LASSO | RLASSO | ROB-LASSO | ROB-RLASSO | SEQ-LASSO | SEQ-RLASSO |
|---|---|---|---|---|---|---|---|---|
| 0 | 7 | 0.01 | 5.902 | 5.154 | 1.793 | 1.898 | 1.799 | 1.900 |
| 0 | 7 | 0.05 | 18.289 | 19.410 | 2.160 | 2.273 | 2.153 | 2.249 |
| 0 | 7 | 0.10 | 21.716 | 25.832 | 2.427 | 2.521 | 2.409 | 2.519 |
| 0 | 15 | 0.01 | 3.808 | 3.345 | 1.013 | 1.085 | 1.017 | 1.079 |
| 0 | 15 | 0.05 | 11.166 | 10.145 | 1.438 | 1.510 | 1.449 | 1.491 |
| 0 | 15 | 0.10 | 12.549 | 12.068 | 5.937 | 6.017 | 3.951 | 3.712 |
| 0 | 30 | 0.01 | 4.146 | 3.808 | 1.002 | 1.087 | 1.004 | 1.098 |
| 0 | 30 | 0.05 | 14.497 | 11.570 | 1.062 | 1.152 | 1.046 | 1.162 |
| 0 | 30 | 0.10 | 14.333 | 11.610 | 1.289 | 1.434 | 1.311 | 1.440 |
| 0.05 | 7 | 0.01 | 813.074 | 541.661 | 2.987 | 3.059 | 2.717 | 2.817 |
| 0.05 | 7 | 0.05 | 5346.473 | 3577.792 | 8.556 | 8.174 | 8.979 | 9.233 |
| 0.05 | 7 | 0.10 | 3616.270 | 9062.238 | 27.115 | 17.156 | 14.483 | 15.828 |
| 0.05 | 15 | 0.01 | 469.709 | 356.820 | 2.054 | 3.333 | 2.275 | 6.391 |
| 0.05 | 15 | 0.05 | 4307.260 | 2233.865 | 17.964 | 23.400 | 6.736 | 11.948 |
| 0.05 | 15 | 0.10 | 4108.178 | 4861.238 | 218.290 | 217.778 | 431.289 | 297.443 |
| 0.05 | 30 | 0.01 | 493.541 | 982.032 | 2.069 | 9.900 | 1.271 | 10.643 |
| 0.05 | 30 | 0.05 | 5886.633 | 4772.21 | 7.783 | 19.457 | 6.780 | 23.559 |
| 0.05 | 30 | 0.10 | 10434.33 | 8941.705 | 98.562 | 123.690 | 128.721 | 221.843 |

p: Number of parameters; pr-cellw: Cellwise outlier proportion; pr-casew: x direction outlier proportion; LASSO: Classical LASSO; RLASSO: Robust LASSO; ROB-LASSO: Robust imputed LASSO; ROB-RLASSO: Robust imputed Robust LASSO; SEQ-LASSO:Sequential imputed LASSO; SEQ-RLASSO: Sequential imputed Robust LASSO.

**Table 2.** MSE of beta for $n = 50$ and $\varepsilon \sim N(0, 1) + N(3, 1)$

| pr-casew | p | pr-cellw | LASSO | RLASSO | ROB-LASSO | ROB-RLASSO | SEQ-LASSO | SEQ-RLASSO |
|---|---|---|---|---|---|---|---|---|
| 0 | 7 | 0.01 | 7.022 | 5.963 | 3.050 | 3.115 | 3.072 | 3.113 |
| 0 | 7 | 0.05 | 20.945 | 20.988 | 3.579 | 3.586 | 3.582 | 3.595 |
| 0 | 7 | 0.10 | 25.313 | 30.345 | 4.063 | 4.052 | 4.056 | 4.059 |
| 0 | 15 | 0.01 | 4.397 | 4.244 | 1.716 | 1.766 | 1.700 | 1.777 |
| 0 | 15 | 0.05 | 12.502 | 10.808 | 2.088 | 2.139 | 2.073 | 2.105 |
| 0 | 15 | 0.10 | 12.339 | 11.738 | 7.121 | 6.931 | 4.446 | 4.483 |
| 0 | 30 | 0.01 | 4.733 | 4.435 | 1.517 | 1.596 | 1.517 | 1.597 |
| 0 | 30 | 0.05 | 15.298 | 12.179 | 1.632 | 1.671 | 1.706 | 1.666 |
| 0 | 30 | 0.10 | 14.507 | 12.227 | 1.850 | 1.972 | 1.876 | 1.956 |
| 0.05 | 7 | 0.01 | 38.653 | 661.211 | 3.318 | 3.841 | 3.513 | 5.331 |
| 0.05 | 7 | 0.05 | 6562.070 | 3857.103 | 3.816 | 5.648 | 3.672 | 5.011 |
| 0.05 | 7 | 0.10 | 3933.421 | 7910.298 | 77.336 | 37.420 | 18.466 | 28.275 |
| 0.05 | 15 | 0.01 | 450.921 | 367.050 | 1.660 | 2.110 | 1.662 | 3.253 |
| 0.05 | 15 | 0.05 | 3807.158 | 2180.618 | 23.511 | 23.329 | 3.181 | 8.353 |
| 0.05 | 15 | 0.10 | 3984.041 | 5059.752 | 260.543 | 257.044 | 83.259 | 384.752 |
| 0.05 | 30 | 0.01 | 606.122 | 841.965 | 9.369 | 41.843 | 11.931 | 64.211 |
| 0.05 | 30 | 0.05 | 7971.026 | 4716.980 | 32.025 | 50.330 | 46.235 | 103.099 |
| 0.05 | 30 | 0.10 | 9336.502 | 8642.231 | 87.116 | 141.860 | 141.060 | 235.443 |

p: Number of parameters; pr-cellw: Cellwise outlier proportion; pr-casew: x direction outlier proportion; LASSO: Classical LASSO; RLASSO: Robust LASSO; ROB-LASSO: Robust imputed LASSO; ROB-RLASSO: Robust imputed Robust LASSO; SEQ-LASSO:Sequential imputed LASSO; SEQ-RLASSO: Sequential imputed Robust LASSO.

**Table 3.** MSE of beta for $n = 50$ and $\varepsilon \sim t_3$

| pr-casew | p | pr-cellw | LASSO | RLASSO | ROB-LASSO | ROB-RLASSO | SEQ-LASSO | SEQ-RLASSO |
|---|---|---|---|---|---|---|---|---|
| 0 | 7 | 0.01 | 9.058 | 8.095 | 4.960 | 4.736 | 4.954 | 4.746 |
| 0 | 7 | 0.05 | 19.349 | 19.948 | 4.663 | 4.515 | 4.667 | 4.494 |
| 0 | 7 | 0.10 | 24.775 | 30.729 | 5.549 | 5.301 | 5.537 | 5.279 |
| 0 | 15 | 0.01 | 5.228 | 4.868 | 2.843 | 2.701 | 2.818 | 2.709 |
| 0 | 15 | 0.05 | 11.125 | 9.156 | 2.599 | 2.512 | 2.589 | 2.542 |
| 0 | 15 | 0.10 | 13.388 | 12.450 | 7.489 | 7.695 | 5.626 | 5.877 |
| 0 | 30 | 0.01 | 5.056 | 4.747 | 2.061 | 1.999 | 2.064 | 1.987 |
| 0 | 30 | 0.05 | 18.032 | 12.510 | 2.175 | 2.086 | 2.152 | 2.099 |
| 0 | 30 | 0.10 | 15.820 | 12.237 | 2.181 | 2.206 | 2.247 | 2.203 |
| 0.05 | 7 | 0.01 | 898.064 | 523.740 | 6.539 | 7.146 | 6.532 | 8.383 |
| 0.05 | 7 | 0.05 | 7288.761 | 4026.641 | 7.124 | 8.047 | 6.625 | 7.848 |
| 0.05 | 7 | 0.10 | 4171.419 | 8775.199 | 18.291 | 29.342 | 10.470 | 17.674 |
| 0.05 | 15 | 0.01 | 428.620 | 342.041 | 4.010 | 4.376 | 3.979 | 4.375 |
| 0.05 | 15 | 0.05 | 4510.567 | 2382.344 | 16.507 | 18.087 | 6.570 | 10.137 |
| 0.05 | 15 | 0.10 | 3913.375 | 5047.254 | 254.423 | 267.726 | 116.708 | 145.420 |
| 0.05 | 30 | 0.01 | 634.455 | 866.216 | 2.777 | 32.245 | 2.731 | 40.746 |
| 0.05 | 30 | 0.05 | 6331.625 | 4951.489 | 39.471 | 105.542 | 47.720 | 114.237 |
| 0.05 | 30 | 0.10 | 8828.778 | 8993.255 | 97.325 | 179.917 | 136.942 | 251.600 |

p: Number of parameters; pr-cellw: Cellwise outlier proportion; pr-casew: x direction outlier proportion; LASSO: Classical LASSO; RLASSO: Robust LASSO; ROB-LASSO: Robust imputed LASSO; ROB-RLASSO: Robust imputed Robust LASSO; SEQ-LASSO:Sequential imputed LASSO; SEQ-RLASSO: Sequential imputed Robust LASSO.

**Table 4.** Percents of true model selection - I

| | cellwise pr | True Choice Pr. | LASSO | RLASSO | ROB-LASSO | ROB-RLASSO | SEQ-LASSO | SEQ-RLASSO |
|---|---|---|---|---|---|---|---|---|
| $\varepsilon \sim N(0,1)$ | 0.01 | $\beta_2 = 0$ | 57.2 | 70 | 69.6 | 79.6 | 63.2 | 65.2 |
| | | $\beta_4 = 0$ | 53.6 | 70 | 70 | 80.8 | 54.8 | 54.8 |
| | | $\beta_7 = 0$ | 58.4 | 69.6 | 70.8 | 84 | 60.4 | 61.2 |
| | | True Model | 16.4 | 36.4 | 38.4 | 51.2 | 24.8 | 27.2 |
| $\varepsilon \sim N(0,1)$ | 0.05 | $\beta_2 = 0$ | 29.6 | 66.4 | 63.6 | 82.4 | 56.8 | 57.6 |
| | | $\beta_4 = 0$ | 31.6 | 62.8 | 65.2 | 80.8 | 55.6 | 54.8 |
| | | $\beta_7 = 0$ | 30 | 61.6 | 62.4 | 74.4 | 53.2 | 52.4 |
| | | True Model | 2.4 | 31.2 | 31.2 | 32.4 | 20.4 | 19.6 |
| $\varepsilon \sim N(0,1)$ $+N(3,1)$ | 0.01 | $\beta_2 = 0$ | 51.6 | 65.6 | 65.2 | 81.6 | 58.8 | 60.4 |
| | | $\beta_4 = 0$ | 50.4 | 62.4 | 63.2 | 79.2 | 56.4 | 56.8 |
| | | $\beta_7 = 0$ | 50.8 | 62 | 61.6 | 76.8 | 53.2 | 56 |
| | | True Model | 13.2 | 26.8 | 27.6 | 49.2 | 20.8 | 22 |
| $\varepsilon \sim N(0,1)$ $+N(3,1)$ | 0.05 | $\beta_2 = 0$ | 30.8 | 64 | 65.6 | 74.4 | 56.8 | 56.8 |
| | | $\beta_4 = 0$ | 30 | 65.2 | 63.2 | 78.8 | 53.2 | 53.6 |
| | | $\beta_7 = 0$ | 28.4 | 62 | 63.6 | 76.8 | 52.4 | 51.2 |
| | | True Model | 1.6 | 31.2 | 32 | 27.2 | 18.4 | 17.6 |
| $\varepsilon \sim t_3$ | 0.01 | $\beta_2 = 0$ | 56.8 | 62 | 64 | 77.6 | 60 | 62 |
| | | $\beta_4 = 0$ | 57.6 | 57.6 | 62 | 76 | 59.2 | 58.4 |
| | | $\beta_7 = 0$ | 50.8 | 59.2 | 60.8 | 74.8 | 53.2 | 52 |
| | | True Model | 18 | 26.4 | 28 | 46.4 | 22.8 | 24 |
| $\varepsilon \sim t_3$ | 0.05 | $\beta_2 = 0$ | 30.8 | 58 | 57.6 | 70 | 55.6 | 54.4 |
| | | $\beta_4 = 0$ | 31.2 | 60.8 | 63.2 | 77.2 | 57.2 | 58 |
| | | $\beta_7 = 0$ | 30.8 | 61.6 | 62.8 | 77.2 | 53.2 | 54.8 |
| | | True Model | 2 | 24.8 | 27.2 | 26.8 | 19.6 | 21.6 |
| $\varepsilon \sim N(0,1)$ $+5\%$ *casewise* | 0.01 | $\beta_2 = 0$ | 65.2 | 82.8 | 82.4 | 90.0 | 80.0 | 81.2 |
| | | $\beta_4 = 0$ | 62.4 | 81.6 | 82.0 | 90.4 | 77.6 | 75.6 |
| | | $\beta_7 = 0$ | 66.4 | 83.2 | 84.0 | 88.4 | 83.2 | 82.0 |
| | | True Model | 2.7 | 21.4 | 21.8 | 28.0 | 20.6 | 20.5 |
| $\varepsilon \sim N(0,1)$ + 5% casewise | 0.05 | $\beta_2 = 0$ | 87.2 | 79.6 | 80 | 98.8 | 79.6 | 79.6 |
| | | $\beta_4 = 0$ | 86.8 | 77.6 | 78.8 | 96.4 | 79.6 | 80.8 |
| | | $\beta_7 = 0$ | 89.2 | 77.6 | 78.8 | 96.8 | 74.8 | 77.6 |
| | | True Model | 0 | 48 | 48 | 71.6 | 46.8 | 50.8 |

LASSO: Classical LASSO; RLASSO: Robust LASSO; ROB-LASSO: Robust imputed LASSO; ROB-RLASSO: Robust imputed Robust LASSO; SEQ-LASSO:Sequential imputed LASSO; SEQ-RLASSO: Sequential imputed Robust LASSO.

In Tables 4-5, we displayed the correctly selected number of zero betas and the correctly selected true models for $p = 7$ and $n = 50$ (Table 4) and $n = 250$ (Table 5). We observed that robust imputed robust LASSO performed the best correctly choosing zero betas and the correctly choosing true model. Robust LASSO seemed the second best among the others for identifying zero betas and the correct model. We observed that the other methods were broke-down for correctly choosing zero betas and correct model in the presence of cellwise and casewise outliers.

**Table 5.** Percents of true model selection - II

| | | True Choice Pr. | LASSO | RLASSO | ROB-LASSO | ROB-RLASSO | SEQ-LASSO | SEQ-RLASSO |
|---|---|---|---|---|---|---|---|---|
| $\varepsilon \sim N(0,1)$ | 0.01 | $\beta_2 = 0$ | 26 | 93.2 | 92.8 | 96.4 | 91.6 | 92 |
| | | $\beta_4 = 0$ | 32 | 90.8 | 90.8 | 95.6 | 90 | 89.6 |
| | | $\beta_7 = 0$ | 26.4 | 92 | 93.2 | 97.2 | 89.2 | 90 |
| | | True Model | 4.8 | 78 | 79.2 | 89.2 | 73.2 | 74.4 |
| $\varepsilon \sim N(0,1)$ | 0.05 | $\beta_2 = 0$ | 2.4 | 87.6 | 88 | 90.8 | 86 | 87.2 |
| | | $\beta_4 = 0$ | 7.2 | 88.8 | 89.2 | 92.8 | 88.8 | 89.2 |
| | | $\beta_7 = 0$ | 4.4 | 90.4 | 90 | 92.8 | 90 | 89.6 |
| | | True Model | 0 | 71.6 | 71.6 | 77.2 | 68.4 | 69.2 |
| $\varepsilon \sim N(0,1)$ $+N(3,1)$ | 0.01 | $\beta_2 = 0$ | 26.8 | 82 | 81.6 | 92.4 | 82 | 82.4 |
| | | $\beta_4 = 0$ | 28.4 | 79.2 | 76 | 89.6 | 83.2 | 82.8 |
| | | $\beta_7 = 0$ | 26.4 | 80.8 | 78.8 | 92 | 84.8 | 84.8 |
| | | True Model | 2 | 55.2 | 52.4 | 76 | 58.8 | 58 |
| $\varepsilon \sim N(0,1)$ $+N(3,1)$ | 0.05 | $\beta_2 = 0$ | 5.6 | 79.2 | 78.4 | 84.4 | 85.6 | 86.4 |
| | | $\beta_4 = 0$ | 8 | 74.8 | 76 | 80.8 | 77.2 | 77.2 |
| | | $\beta_7 = 0$ | 5.2 | 80.8 | 80 | 85.6 | 84.4 | 84.4 |
| | | True Model | 0 | 50 | 48.4 | 58.4 | 55.6 | 56.4 |
| $\varepsilon \sim t_3$ | 0.01 | $\beta_2 = 0$ | 30.8 | 81.2 | 81.6 | 93.2 | 87.2 | 87.6 |
| | | $\beta_4 = 0$ | 32.8 | 79.6 | 79.2 | 89.6 | 84.8 | 84.4 |
| | | $\beta_7 = 0$ | 29.6 | 79.2 | 79.6 | 89.2 | 85.6 | 86.4 |
| | | True Model | 3.2 | 54.8 | 56.4 | 76 | 63.2 | 64 |
| $\varepsilon \sim t_3$ | 0.05 | $\beta_2 = 0$ | 6 | 74.8 | 72.4 | 79.6 | 86 | 84.8 |
| | | $\beta_4 = 0$ | 4.4 | 78.4 | 76.4 | 82.8 | 84 | 84.4 |
| | | $\beta_7 = 0$ | 6.4 | 74.8 | 74.4 | 80.8 | 83.2 | 83.2 |
| | | True Model | 0 | 48.4 | 45.2 | 53.2 | 59.6 | 59.2 |
| $\varepsilon \sim N(0,1)$ $+5\%$ *casewise* | 0.01 | $\beta_2 = 0$ | 64 | 91.6 | 91.6 | 97.6 | 91.6 | 91.6 |
| | | $\beta_4 = 0$ | 65.2 | 93.6 | 93.6 | 98.8 | 92.8 | 92.8 |
| | | $\beta_7 = 0$ | 66.4 | 95.6 | 95.6 | 98.4 | 92.8 | 92.8 |
| | | True Model | 0.8 | 81.2 | 81.2 | 94.8 | 78.0 | 78.0 |
| $\varepsilon \sim N(0,1)$ $+5\%$ *casewise* | 0.05 | $\beta_2 = 0$ | 1.6 | 94.8 | 94.8 | 100.0 | 93.6 | 92.4 |
| | | $\beta_4 = 0$ | 2.0 | 94.0 | 94.0 | 100.0 | 94.0 | 93.6 |
| | | $\beta_7 = 0$ | 2.8 | 93.2 | 93.2 | 100.0 | 91.2 | 91.6 |
| | | True Model | 0.0 | 82.8 | 82.8 | 98.8 | 79.2 | 78.8 |

LASSO: Classical LASSO; RLASSO: Robust LASSO; ROB-LASSO: Robust imputed LASSO; ROB-RLASSO: Robust imputed Robust LASSO; SEQ-LASSO:Sequential imputed LASSO; SEQ-RLASSO: Sequential imputed Robust LASSO.

Concerning the results given in Table 5, we observed exactly the similar performance of the methods. Robust imputed Robust LASSO had the excellent behavior for correctly choosing zero betas and for identifying the correct models. Comparing to the results given in Table 4, we noticed that the performances were getting better. For example, when the sample size was small for normally distributed error with 5% cellwise and 5% casewise outliers (see the 8th case in Table 4 and Table 5), the ratio choosing the corrected model is 71.6% . However, that ratio was 98.8% in Table 5. Therefore, increasing sample size affected for choosing correct model and correct zero betas.

Further to illustrate performance of the methods for higher dimensional cases, we gave boxplots of the some of the estimated zero betas (Mainly, we took last three zeros for simplicity). These boxplots were given in Figures 1-3. In these figures, dimension of the regression parameter is 15. We considered different outliers configurations in these figures. In Figures 1 and 2, heavy-tailed error distribution with cellwise outliers. On the other hand, in Figure 3, we had cellwise outlier and casewise outlier with normally distributed errors. We observed that robust imputed robust LASSO superior to the other methods in terms of correctly choosing zero betas almost all the cases. Compare to the others, variability seemed smaller.

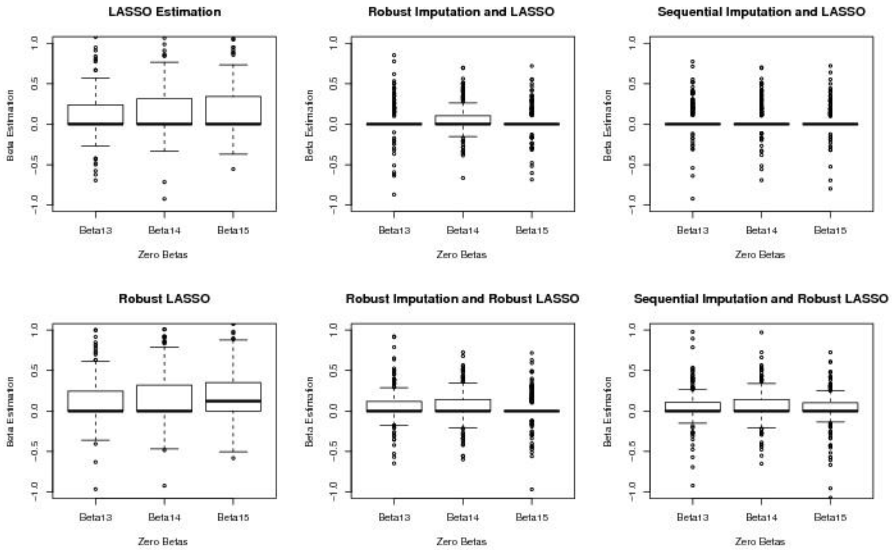



**Figure 1.** Results for $p = 15, n = 50, cellwise - pr = 0.05$, and $\varepsilon \sim N(0, 1) + N(3, 1)$

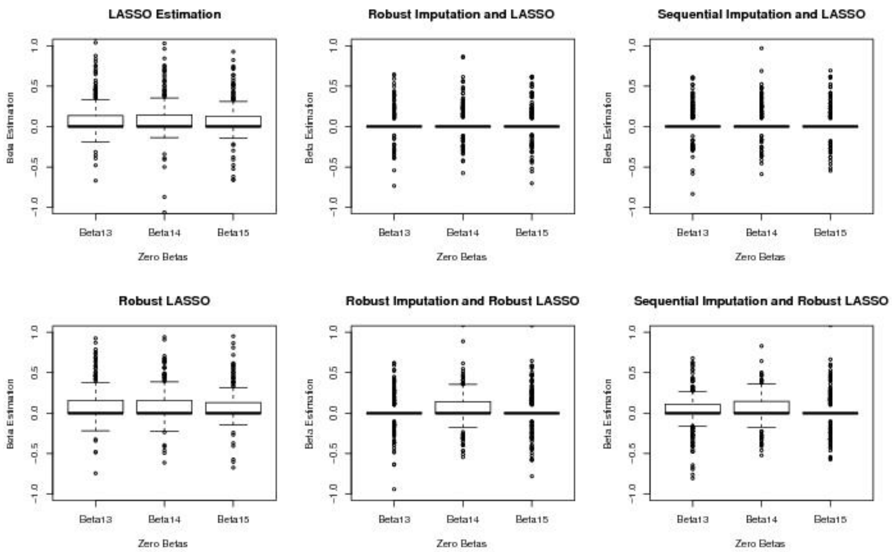



**Figure 2.** Results for $p = 15, n = 50, cellwise - pr = 0.01$ and $\varepsilon \sim t_3$

## 4. Real data example

To compare the methods in real data example, the most known model selection data, the prostate cancer data in [23] was examined. There are 97 observations collected from men who were about to receive a radical prostatectomy. The response variable was log(prostate specific antigen) (lpsa). The explanatory variables were log (cancer volume) $x_1$ : lcavol), log(prostate weight) ($x_2$ : lweight), age($x_3$ ), log(benign prostatic hyperplasia amount) ($x_4$ : lbph), seminal vesicle invasion ($x_5$: svi), log(capsular penetration) ($x_6$: lcp), Gleason

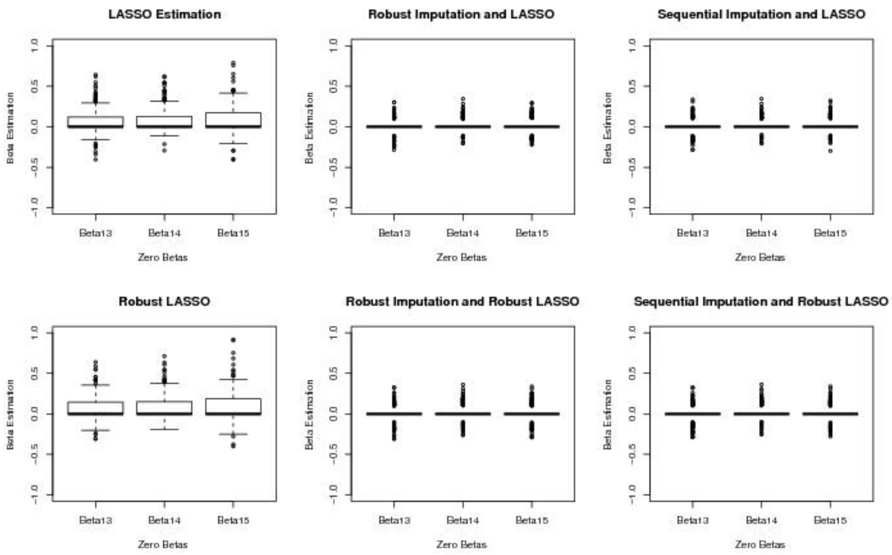



**Figure 3.** Results for $p = 15, n = 100, cellwise - pr = 0.05$, $casewise - pr = 0.05$ and $\varepsilon \sim N(0, 1)$

score ($x_7$: gleason) and percentage Gleason scores 4 or 5 ($x_8$ : pgg45). In literature, this dataset has been extensively used to access the performance of the model selection methods [24, 30]. In those papers, the variables $x_1, x_4, x_5$ were found the most important variables. In the applications of [24, 30], explanatory variable $x_3$ was also found significant. In our paper, we compared the methods in terms of correctly selected non-significant betas (zero betas) and true model selection. We also checked the prediction errors for testing dataset which was randomly chosen 20% of the real dataset in each iteration. The results were given in Table 6 and Figure 4. All of these results confirmed that robust imputed robust LASSO was the best according to the criteria we were using. We also noticed that sequential imputed Robust LASSO had the similar behavior to the robust imputed robust LASSO.

## 5. Conclusion

After introducing cellwise outlier or independent contamination model, some problems occured in estimation even robust ones. Especially in high dimension, breakdown points of estimation will be exceeded even though there is very small proportion cellwise outliers. In this paper, we considered cellwise and the casewise outlier problem in a regression analysis when parameter estimation and variable selection is a concern. We used robust imputation method to deal with the cellwise outlier and we combined the robust regression estimation method with LASSO to deal with the variable selection in the presence of cellwise and casewise outliers. We did this procedure in three steps. In the first step, we had identified the cellwise outliers and in the second step, we had dealt with the cellwise outliers and use robust imputation to get rid of the cellwise outliers. Finally, in the last step, we combined robust estimation with LASSO to dealt with casewise outliers if they are in present. We provided an extensive simulation study to illustrate the performance of proposed method and observed that the proposed method has comparable results among the methods that have similar proposal. We had also explored the real data example using prostate cancer data which have been extensively used in literature to show the performance of the model selection methods. The result of the real data example have also confirm the simulation results in terms of the proposed method.

**Table 6.** Real data examples: prostate cancer data results

| MSE of Beta for Prostate Cancer Data | | | | | | |
|---|---|---|---|---|---|---|
| pr-cellw | LASSO | RLASSO | ROB-LASSO | ROB-RLASSO | SEQ-LASSO | SEQ-RLASSO |
| 0.01 | 4.477 | 5.199 | 1.308 | 1.200 | 1.303 | 1.199 |
| 0.05 | 14.845 | 14.047 | 1.272 | 1.192 | 1.272 | 1.195 |
| 0.10 | 13.520 | 12.898 | 0.970 | 0.906 | 0.952 | 0.896 |
| Zero Beta Selection for Prostate Cancer Data | | | | | | |
| pr-cellw | LASSO | RLASSO | ROB-LASSO | ROB-RLASSO | SEQ-LASSO | SEQ-RLASSO |
| | 100.00 | 100.00 | 100.00 | 100.00 | 100.00 | 100.00 |
| | 33.60 | 87.60 | 85.60 | 90.00 | 38.80 | 38.00 |
| 0.01 | 63.60 | 94.40 | 93.20 | 99.60 | 80.40 | 81.20 |
| | 98.00 | 96.80 | 97.60 | 100.00 | 100.00 | 100.00 |
| | 100.00 | 100.00 | 100.00 | 100.00 | 100.00 | 100.00 |
| | 100.00 | 100.00 | 100.00 | 100.00 | 100.00 | 100.00 |
| | 66.80 | 95.20 | 95.20 | 100.00 | 86.00 | 85.60 |
| 0.05 | 98.40 | 96.00 | 95.60 | 100.00 | 76.80 | 78.80 |
| | 73.60 | 90.80 | 90.80 | 100.00 | 96.00 | 95.60 |
| | 100.00 | 100.00 | 100.00 | 100.00 | 100.00 | 100.00 |
| | 100.00 | 100.00 | 100.00 | 100.00 | 100.00 | 100.00 |
| | 84.40 | 72.00 | 70.80 | 87.20 | 49.20 | 55.60 |
| 0.10 | 99.60 | 36.00 | 21.60 | 99.20 | 26.00 | 19.60 |
| | 5.60 | 11.60 | 4.40 | 98.00 | 31.20 | 17.60 |
| | 100.00 | 100.00 | 100.00 | 100.00 | 100.00 | 100.00 |
| True Model Selection for Prostate Cancer Data | | | | | | |
| pr-cellw | LASSO | RLASSO | ROB-LASSO | ROB-RLASSO | SEQ-LASSO | SEQ-RLASSO |
| 0.01 | 8.80 | 81.20 | 79.20 | 88.80 | 31.20 | 31.20 |
| 0.05 | 0.80 | 88.40 | 88.00 | 98.40 | 65.20 | 66.4 |
| 0.10 | 0.00 | 6.40 | 1.60 | 78.40 | 6.40 | 5.60 |

LASSO: Classical LASSO; RLASSO: Robust LASSO; ROB-LASSO: Robust imputed LASSO; ROB-RLASSO: Robust imputed Robust LASSO; SEQ-LASSO:Sequential imputed LASSO; SEQ-RLASSO: Sequential imputed Robust LASSO.

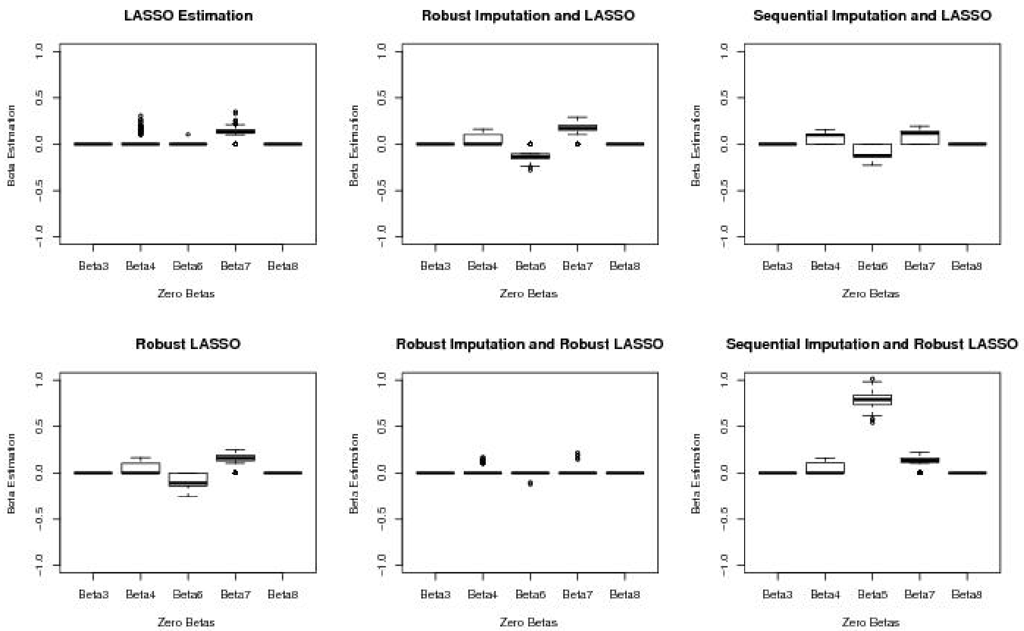



**Figure 4.** Results for prostate cancer data

## References


[1] C. Agostinelli, A. Leung, V.J. Yohai and R.H. Zamar, *Robust estimation of multivariate location and scatter in the presence of cellwise and casewise contamination*, Test, **24** (3), 441-461, 2015.

[2] F. Alqallaf, S. Van Aelst, V.J. Yohai and R.H. Zamar, *Propagation of Outliers in Multivariate Data*, Ann. Statist. **37** (1), 311-331, 2009.

[3] O. Arslan, *Weighted LAD-LASSO method for robust parameter estimation and variable selection in regression*, Comput. Statist. Data Anal. **56** (6), 1952-1965, 2012.

[4] O. Arslan, *Penalized MM regression estimation with L $\gamma$ penalty: a robust version of bridge regression*, Statistics **50** (6), 1236-1260, 2016.

[5] K.V. Branden and S. Verboven, *Robust data imputation*, Comput. Biol. Chem. **33** (1), 7-13, 2009.

[6] M. Danilov, *Robust estimation of multivariate scatter in non-affine equivariant scenarios*, University of British Columbia, 2010.

[7] M. Debruyne, S. Höppner, S. Serneels and T. Verdonck, *Outlyingness: Which variables contribute most?*, Stat. Comput. **29** (4), 707-723, 2019.

[8] J. Fan, Y. Fan and E. Barut, *Adaptive robust variable selection*, Ann. Statist. **42** (1), 324-351, 2014.

[9] A. Farcomeni, *Snipping for robust k-means clustering under component-wise contamination*, Stat. Comput. **24** (6), 907-919, 2014.

[10] P.A. Ferrari, P. Annoni, A. Barbiero and G. Manzi, *An imputation method for categorical variables with application to nonlinear principal component analysis*, Comput. Statist. Data Anal. **55** (7), 2410-2420, 2011.

[11] A.E. Hoerl and R.W. Kennard, *Ridge regression Biased estimation for nonorthogonal problems*, Technometrics **12** (1), 55-67, 1970.

[12] A. Leung, H. Zhang and R. Zamar, *Robust regression estimation and inference in the presence of cellwise and casewise contamination*, Comput. Statist. Data Anal. **99**, 1-11, 2016.

[13] A. Leung, V. Yohai and R. Zamar, *Multivariate location and scatter matrix estimation under cellwise and casewise contamination*, Comput. Statist. Data Anal. **111**, 59-76, 2017.

[14] J. Machkour, B. Alt, M. Muma and A.M. Zoubir, *The outlier-corrected-data-adaptive Lasso: A new robust estimator for the independent contamination model*, 25th European Signal Processing Conference (EUSIPCO), IEEE, 1649-1653, 2017.

[15] R.A. Maronna, *Robust ridge regression for high-dimensional data*, Technometrics **53** (1), 44-53, 2011.

[16] R.A. Maronna, R.D. Martin, V.J. Yohai and S.B. Matías, *Robust statistics: theory and methods (with R)*, John Wiley & Sons, 2019.

[17] V. Ollerer, A. Andreas and C. Croux, *The shooting S-estimator for robust regression*, Comput. Statist. **31** (3), 829-844, 2016.

[18] J. Raymaekers and P.J. Rousseeuw, *Flagging and handling cellwise outliers by robust estimation of a covariance matrix*, arXiv preprint arXiv:1912.12446, 2019.

[19] J. Raymaekers, P.J. Rousseeuw, W. Van den Bossche and M. Hubert, *cellWise: Analyzing Data with Cellwise Outliers*, CRAN, R package version: 2.0.9, 2019.

[20] P.J. Rousseeuw and W. Van den Bossche, *Detecting deviating data cells*, Technometrics **60** (2), 135-145, 2018.

[21] P.J. Rousseeuw and A. M. Leroy, *Robust regression and outlier detection*, John Wiley & Sons, 2005.

[22] N. Simon, J. Friedman, T. Hastie and R. Tibshirani, *Regularization paths for Coxs proportional hazards model via coordinate descent*, J. Stat. Softw. **39** (5), 1-13, 2011.

[23] T.A. Stamey, J.N. Kabalin, J.E. McNeal, I. Johnstone, M. Iain, F. Freiha, E.A. Redwine and N. Yang, *Prostate specific antigen in the diagnosis and treatment of adenocarcinoma of the prostate. II. Radical prostatectomy treated patients*, J. Urol. **141** (5), 1076-1083, 1989.

[24] R. Tibshirani, *Regression shrinkage and selection via the lasso*, J. R. Stat. Soc. Ser. B. Stat. Methodol. **58** (1), 267-288, 1996.

[25] A. Unwin, *Multivariate outliers and the O3 Plot*, J. Comput. Graph. Statist. **28** (3), 635-643, 2019.

[26] S. Verboven, K.V. Branden and P. Goos, *Sequential imputation for missing values*, Comput. Biol. Chem. **33** (5-6), 320-327, 2007.

[27] H. Xu, C. Caramanis and S. Mannor, *Robust regression and LASSO*, Adv Neural Inf Process Syst, 1801-1808, 2009.

[28] C. Yi and J. Huang, *Semismooth newton coordinate descent algorithm for elastic-net penalized huber loss regression and quantile regression*, J. Comput. Graph. Statist. **26** (3), 547-557, 2017.

[29] J.V. Yohai, *High breakdown-point and high efficiency robust estimates for regression*, Ann. Statist. **15** (2), 642-656, 1987.

[30] L. Zeng and J. Xie, *Regularization and variable selection for data with interdependent structures*, 2008.

[31] H. Zou and T. Hastie, *Regularization and variable selection via the elastic net*, J. R. Stat. Soc. Ser. B. Stat. Methodol. **67** (2), 301-320, 2005.