ISSN: 0025-5742

THE MATHEMATICS STUDENT

Volume 95, Nos. 3-4, July - December (2026)
(Issued: November, 2026)

Editor-in-Chief
G. P. Youvaraj

EDITORS

Bruce C. Berndt	George E. Andrews	M. Ram Murty
B. Sury	Siddhartha Gadgil	Gadadhar Misra
Sukanta Pati	Kaushal Verma	Subhojoy Gupta
S. K. Tomar	Clare D′Cruz	L. Sunil Chandran
Aparna Dar	C. S. Aravinda	Atul Dixit
Indranil Biswas	Timothy Huber	T. S. S. R. K. Rao
Debjani Chakraborty	Safique Ahmad	T. Raja Sekhar

PUBLISHED BY
THE INDIAN MATHEMATICAL SOCIETY
Website: https://indianmathsoc.org

THE MATHEMATICS STUDENT

Edited by G. P. Youvaraj

In keeping with the current periodical policy, THE MATHEMATICS STUDENT seeks to publish material of interest not just to mathematicians with specialized interest but to the postgraduate students and teachers of mathematics in India and abroad. With this in view, it will ordinarily publish material of the following type:

1. research papers,
2. the texts (written in a way accessible to students) of the Presidential Addresses, the Plenary talks and the Award Lectures delivered at the Annual Conferences.
3. general survey articles, popular articles, expository papers and Book-Reviews.
4. problems and solutions of the problems,
5. new, clever proofs of theorems that graduate / undergraduate students might see in their course work, and
6. articles that arouse curiosity and interest for learning mathematics among readers and motivate them for doing mathematics.

Articles of the above type are invited for publication in THE MATHEMATICS STUDENT. Manuscripts intended for publication should be submitted online in the LaTeX and .pdf file including figures and tables to the Editor-in-Chief on the E-mail: msindianmathsociety@gmail.com.

Manuscripts (including bibliographies, tables, *etc.*) should be typed double spaced on A4 size paper with 1 inch (2.5 cm.) margins on all sides with font size 11 pt. in LaTeX. Sections should appear in the following order: Title Page, Abstract, Text, Notes and References. Comments or replies to previously published articles should also follow this format. In LaTeX the following preamble be used as is required by the Press:

\ documentclass[11 pt,a4paper,twoside,reqno]{amsart}
\ usepackage {amsfonts, amssymb, amscd, amsmath, enumerate, verbatim, calc}
\ renewcommand{\ baselinestretch}{1.2}
\ textwidth=12.5 cm
\ textheight=20 cm
\ topmargin=0.5 cm
\ oddsidemargin=1 cm
\ evensidemargin=1 cm
\ pagestyle{plain}

The details are available on Society's website: https://indianmathsoc.org

Authors of articles / research papers printed in the Mathematics Student as well as in the Journal shall be entitled to receive a *soft copy (PDF file)* of the paper published. There are no page charges for publication of articles in these journals.

All business correspondence should be addressed to Prof. M. M. Pawar, Treasurer, Indian Mathematical Society, Formerly at K.B.C. North Maharastra University, Jalgon, Maharashtra - 425001, India, on the E-mail: treasurerindianmathsociety@gmail.com.

In case of any query, one may contact the Editor through the e-mail.

Copyright of the published articles lies with the Indian Mathematical Society.

ISSN: 0025-5742

THE MATHEMATICS STUDENT

Volume 95, Nos. 3-4, July - December (2026)
(Issued: November, 2026)

Editor-in-Chief
G. P. Youvaraj

EDITORS

Bruce C. Berndt	**George E. Andrews**	**M. Ram Murty**
B. Sury	**Siddhartha Gadgil**	**Gadadhar Misra**
Sukanta Pati	**Kaushal Verma**	**Subhojoy Gupta**
S. K. Tomar	**Clare D′Cruz**	**L. Sunil Chandran**
Aparna Dar	**C. S. Aravinda**	**Atul Dixit**
Indranil Biswas	**Timothy Huber**	**T. S. S. R. K. Rao**
Debjani Chakraborty	**Safique Ahmad**	**T. Raja Sekhar**

PUBLISHED BY
THE INDIAN MATHEMATICAL SOCIETY
Website: https://indianmathsoc.org

ISSN: 0025-5742

© THE INDIAN MATHEMATICAL SOCIETY, 2026.

This volume or any part thereof should not be reproduced in any form without the written permission of the publisher.

This volume is not to be sold outside the Country to which it is consigned by the Indian Mathematical Society.

Member's copy is strictly for personal use. It is not intended for sale or circulation.

Published by Prof. M. M. Shikare for the Indian Mathematical Society, type set by Prof. G. P. Youvaraj, Former Director and Head, Ramanujan Institute for Advanced Study in Mathematics, Chepauk, Chennai - 600005, India and printed by Anant Barve at Parashuram Printers, A-17, Co-Operative Industrial Estate, Railway station MIDC, Chhatrapati Sambhajinagar-431 005, Maharashtra, India.

Printed in India.

The Mathematics Student ISSN: 0025-5742
Vol. 95, Nos. 3-4, July - December (2026)

CONTENTS

1.	Banerjee, D., Minamide, T. M., Tanigawa, Y.	Divisor Problem On Square-Free Integers	01–19
2.	Ghanwat, A., Pandit, S., Kute, S.	A Short Note On Open Books Of $\#_k S^1 \widetilde{\times} S^2$	20–27
3.	Vijayashankar, A. I., Vasuki, K. R.	Evaluation Of The Convolution Sums $W_8(n)$ and $W_{16}(n)$	28–37
4.	Kaur, S.	A Quadratic Residue In Some Finite Fields	38–41
5.	Thangavelu, G., Dimpi, K. M.	Hook Fusion Procedure For Direct Product Of Symmetric Groups	42–57
6.	Deep, A., Batra, R.	Some Remarks On The Metrizability Of Topological Group Valued Cone Metric Spaces	58–68
7.	Remil, M., Mahdi, F. Z., Yfrah, A., Daouadji, D. K.	Dynamics Of A Wave Equations With Nonlinear Damping Term And Critical Nonlinearity	69-108
8.	Saha, S., Pal, A. K., Roy, S.	Investigations On Weighted Value Sharing And Its Impact On Uniqueness Of Meromorphic Functions	109–125
9.	Dasireddy, N. S.	A Simple Proof Of An Identity Involving Skew-Harmonic Number And A Tail Of $\zeta(3)$	126–133
10.	Demiri, I., Rexhepi, S.	Comparison Of Some Series Containing Euler Number Via Apery's Constant	134–139
11.	Hariprasad, M.	A Note On Fiedler Vector Monotonicity Beyond R-matrices	140–149
12.	Bhandari, R., Bhandari, V.	Generating Infinite Series For A Unit Fraction	150–157

13.	Tran, Q. H.	Some Extensions Of A Problem Of Prasolov And Sharygin In Solid Geometry	158–169
14.	Brown, G. W., Brown, A. C.	Approximating The Sums Of Powers Of Reciprocals	170–176
15.	Patnaik, S., Sethi, R.	Block-Tridiagonal Matrix Form Of Bounded Operator, Compact Operator Commutator Spectrum Quasinilpotent Operator	177–194
16.	Baby, S. A.	Finsler Spaces With p-Power (α, β)-Metric	195–211
17.	Dhage, J. B., Dhage, B. C.	Local Convergence Analysis Of Initial Value Problems Of Nonlinear Riemann-Liouville Fractional Hybrid Differential Equations	212–241
18.	Rajagopal, S.	An Elementary Proof Of Question 359 Of Ramanujan	242–249
19.	Das, A., Das, J.	Modularity In Mathematics And Physics: From Modular Forms To Rational Conformal Field Theory	250–273
20.		Problem Section	274–284

****

The Mathematics Student ISSN: 0025-5742
Vol. 95, Nos. 3-4, July - December (2026), 01–19

DIVISOR PROBLEM ON SQUARE-FREE INTEGERS

DEBIKA BANERJEE, TOKUHON MAKOTO MINAMIDE
AND
YOSHIO TANIGAWA

(Received : 09 - 04 - 2023 ; Accepted : 11 - 09 - 2026)

Abstract. In this paper, we investigate the average of the number of divisors of square-free integers $n \leq x$, which was examined by Tenenbaum. First of all, we improve the error term appeared in the asymptotic formula using elementary methods. Moreover, we obtain a sharper estimate for the error term via a variant of the classical Perron formula which was proved by Liu and Ye.

1. Introduction

Let $\omega(n)$ denote the number of distinct prime divisors of a positive integer n, and $\zeta(s)$ the Riemann zeta function, that is, $\zeta(s) = \sum_{n=1}^{\infty} n^{-s}$ ($s = \sigma + it$, σ, $t \in \mathbb{R}$, $\sigma > 1$). In [17], Suryanarayana and Siva Rama Prasad investigated the error term $E(x)$ in the following asymptotic formula

$$\sum_{n \leq x} 2^{\omega(n)} = \frac{6}{\pi^2} x \log x + Cx + E(x), \tag{1.1}$$

as $x \to \infty$, where

$$C = \frac{6}{\pi^2} \left(2\gamma - 1 - 2\frac{\zeta'(2)}{\zeta(2)} \right), \tag{1.2}$$

and γ is the Euler constant. They obtained the following estimate

$$E(x) = O\left(x^{1/2} \exp\left(-c \frac{(\log x)^{3/5}}{(\log\log x)^{1/5}} \right) \right), \tag{1.3}$$

where $c > 0$ is a constant. The above result is an improvement of Suetuna's result [16]. Later the same result (1.3) was obtained by Ivić [9, p. 73] (see

2020 Mathematics Subject Classification: 11A25, 11N37, 11M06
Key words and phrases: divisor problem, square-free integers, convolution method, Perron's formula, zero-free region of the Riemann zeta function
This work is supported by JSPS KAKENHI Grant Numbers 19K03449, 18K03237.

© *Indian Mathematical Society, 2026.*

also [10, p. 394]). Since $2^{\omega(n)}$ stands for the number of square-free divisors of an integer n, we could consider their problem as another type of the Dirichlet divisor problem (see Titchmarsh [20], Ivić [10]). The first estimation of $E(x)$ in (1.1) was due to Mertens [13], who showed $E(x) = O\left(x^{1/2}\log x\right)$ (see Lemma 2.1, below). Later, Hölder [7] obtained

$$E(x) = O\left(x^{1/2}\right). \tag{1.4}$$

Gioia and Vaidya [6], and Saffari [15] rediscovered the same estimate in (1.4). In [2], [3], under the Riemann hypothesis, Baker showed that

$$E(x) = O\left(x^{4/11+\varepsilon}\right) \tag{1.5}$$

for any $\varepsilon > 0$.

In this paper we shall study $T(x)$ defined by

$$T(x) := \sum_{n\le x} |\mu(n)|2^{\omega(n)} = \sum_{\substack{n\le x \\ n:\ \text{square-free}}} 2^{\omega(n)},$$

where $\mu(\cdot)$ is the Möbius function. The function represents the average of the number of divisors of square-free integers $n \le x$. That is, in the sum $\sum_{n\le x} 2^{\omega(n)}$, we restrict the sum as n runs over square-free integers up to x. This was introduced by Tenenbaum in [18, p. 59]. It can be easily seen

$$T(x) = C_1 x\log x + O(x), \tag{1.6}$$

where $C_1 = \prod_p (1+2/p)(1-1/p)^2$ (p runs over all primes). However, the bound $O(x)$ can be further improved and is related to the study of two different kinds of divisor problem. First, we shall obtain a sharper estimate for the error term in (1.6) by an elementary method.

Theorem 1.1. *For any sufficiently large $x \ge 2$, we have*

$$T(x) = C_1 x\log x + C_2 x + O\left(x^{1/2}(\log x)^3\right),$$

where $C_2 = \frac{6}{\pi^2}\mathcal{H}'(1) + C\mathcal{H}(1)$ (C is defined by (1.2)) and

$$\mathcal{H}(s) = \prod_p \left(1+\frac{2}{p^s}\right)\frac{\left(1-\frac{1}{p^s}\right)}{\left(1+\frac{1}{p^s}\right)}.$$

We shall give a proof of this theorem in Section 2.

Next, using the theory of the Riemann zeta function $\zeta(s)$, especially its zero-free region, we shall improve Theorem 1.1 without using any hypothesis. In fact, we have

Theorem 1.2. *For any sufficiently large $x \geq e^e$, we have*

$$T(x) = C_1 x \log x + C_2^* x + O\left(x^{1/2} \exp\left(-\alpha_* \frac{(\log x)^{3/5}}{(\log\log x)^{1/5}}\right)\right),$$

where α_* is a positive constant, $C_2^* = \mathcal{A}'(1) + C_1(2\gamma - 1) = C_2$, and

$$\mathcal{A}(s) = \prod_p \left(1 - \frac{3}{p^{2s}} + \frac{2}{p^{3s}}\right).$$

We shall give a proof of this theorem in Section 3.

In the proofs of these theorems we make use of the asymptotic formulas of divisor functions, which we recall briefly here. Let $\tau(n) = \sum_{d_1 d_2 = n} 1$ (the number of pairs $\{d_1, d_2\}$ such that $d_1 d_2 = n$, where d_i are positive integers), and $\tau_3(n) = \sum_{d_1 d_2 d_3 = n} 1$ (the number of triples $\{d_1, d_2, d_3\}$ such that $d_1 d_2 d_3 = n$, where d_i are positive integers). These divisors functions are generated by the Riemann zeta function $\zeta(s)$ as

$$\zeta^2(s) = \sum_{n=1}^{\infty} \frac{\tau(n)}{n^s} \quad \textit{and} \quad \zeta^3(s) = \sum_{n=1}^{\infty} \frac{\tau_3(n)}{n^s} \quad \textit{for } \operatorname{Re} s > 1. \tag{1.7}$$

Dirichlet proved, by his celebrated hyperbola method that

$$\sum_{n \leq x} \tau(n) = x \log x + (2\gamma - 1)x + O\left(x^{1/2}\right). \tag{1.8}$$

The above bound $O\left(x^{1/2}\right)$ is improved as follows.

$$\sum_{n \leq x} \tau(n) = x \log x + (2\gamma - 1)x + O\left(x^{1/3}\right). \tag{1.9}$$

It is enough to apply the elementary result (1.8) for our proof of Theorem 1.1. However, to prove Theorem 1.2 we need (1.9) and

$$\sum_{n \leq x} \tau_3(n) = \frac{x}{2} \log^2 x + (3\gamma - 1)x \log x + (3\gamma_1 + 3\gamma^2 - 3\gamma + 1)x + O\left(x^{1/2}\right), \tag{1.10}$$

where γ and γ_1 are coefficients of the Laurent expansion of $\zeta(s)$:

$$\zeta(s) = \frac{1}{s-1} + \gamma + \gamma_1(s-1) + \cdots.$$

The result (1.9) is a consequence of van der Corput [5]. At present, the bound $O\left(x^{\frac{131}{416}+\varepsilon}\right)$ (any $\varepsilon > 0$) due to Huxley [8] is best. But for our purpose, (1.9) is enough. The result (1.10) is a consequence of Atkinson [1]. See, e.g., Titchmarsh [20, p. 317, p. 320, p. 325], Ivić [10, p. 353, p. 354, p. 381]. Moreover, it should be mentioned that we need a variant of Perron's formula [12] due to Liu and Ye. See Lemma 3.5 below.

2. Elementary method

In this section, we shall prove Theorem 1.1 by elementary methods. However, we use the absolute convergence of Dirichlet series and infinite products. In the last part of this section, from a view of complex analysis we will state a remark on the error term $O\left(x^{1/2}(\log x)^3\right)$ in Theorem 1.1.

Lemma 2.1 (Mertens [13], see [11], [4], [18, p. 59, Exercise 54], [14, p. 42, Exercise 2.1.1. 15]). *For $x \geq 1$ we have*

$$\sum_{n\leq x} 2^{\omega(n)} = \frac{6}{\pi^2} x \log x + Cx + O\left(x^{1/2}(1+\log x)\right), \tag{2.1}$$

where C is defined in (1.2).

Proof. From the identity ([20, p. 5, (1.2.8)])

$$\frac{\zeta^2(s)}{\zeta(2s)} = \sum_{n=1}^{\infty} \frac{2^{\omega(n)}}{n^s} \quad (\operatorname{Re} s > 1), \tag{2.2}$$

we have

$$2^{\omega(n)} = \sum_{d^2|n} \mu(d)\tau\left(\frac{n}{d^2}\right),$$

where $\tau(n)$ is the divisor function. Hence we have

$$\sum_{n\leq x} 2^{\omega(n)} = \sum_{d^2\delta\leq x} \mu(d)\tau(\delta) = \sum_{d\leq x^{1/2}} \mu(d) \sum_{\delta\leq x/d^2} \tau(\delta).$$

Using (1.8), $\sum_{n\leq x} \frac{\mu(n)}{n^2} = \frac{6}{\pi^2} + O\left(\frac{1}{x}\right)$, and

$$-2\sum_{n\leq x} \frac{\mu(n)\log n}{n^2} = 2\left(\frac{1}{\zeta(s)}\right)'\Big|_{s=2} + O\left(\frac{\log(1+x)}{x}\right)$$

we get (2.1). □

Lemma 2.2 ([18, p. 59]). *Define the function $\mathcal{Q}(s)$ by the Dirichlet series*

$$\mathcal{Q}(s) := \sum_{n=1}^{\infty} \frac{|\mu(n)|2^{\omega(n)}}{n^s} \quad \textit{for } \operatorname{Re} s > 1, \tag{2.3}$$

the function $\mathcal{H}(s)$ by the infinite product

$$\mathcal{H}(s) := \prod_{p} \left(1 - \frac{2}{p^{2s}} + \frac{2}{p^{3s}} - \cdots\right) \quad \textit{for } \operatorname{Re} s > \frac{1}{2}, \tag{2.4}$$

and the multiplicative function $h(n)$ by

$$h(n) := \begin{cases} 1 & (n = 1), \\ 0 & (n = p), \\ 2(-1)^{k-1} & (n = p^k, k \geq 2). \end{cases} \tag{2.5}$$

Then for $\operatorname{Re} s > 1/2$ we have

$$\mathcal{H}(s) = \sum_{n=1}^{\infty} \frac{h(n)}{n^s}. \tag{2.6}$$

Moreover we have for $\operatorname{Re} s > 1$

$$\mathcal{Q}(s) = \sum_{n=1}^{\infty} \frac{h(n)}{n^s} \cdot \sum_{n=1}^{\infty} \frac{2^{\omega(n)}}{n^s}, \tag{2.7}$$

from which we have the following expression

$$|\mu(n)|2^{\omega(n)} = \sum_{d|n} h(d) 2^{\omega(n/d)}. \tag{2.8}$$

We should note that the function $h(n)$ is introduced in Exercise 55 in Tenenbaum [18].

Proof. Note that the infinite product $\prod_p \left(1 + \frac{2}{p^{2\sigma}} + \frac{2}{p^{3\sigma}} + \cdots\right)$ is convergent for $\sigma > 1/2$. Hence the identity (2.6) follows immediately from the definition (2.5). The function $\mathcal{H}(s)$ defined in (2.4) coincides with $\mathcal{H}(s)$ introduced in Theorem 1.1.

Let $\operatorname{Re} s > 1$. Then we have

$$\begin{aligned} \mathcal{Q}(s) &= \sum_{n=1}^{\infty} \frac{|\mu(n)|2^{\omega(n)}}{n^s} = \prod_{p} \left(1 + \frac{2}{p^s}\right) \\ &= \prod_{p} \frac{\left(1 + \frac{2}{p^s}\right)\left(1 - \frac{1}{p^s}\right)}{\left(1 + \frac{1}{p^s}\right)} \frac{\zeta^2(s)}{\zeta(2s)} \end{aligned} \tag{2.9}$$

$$= \prod_p \left(1 - \frac{2}{p^{2s}} + \frac{2}{p^{3s}} - \cdots\right) \frac{\zeta^2(s)}{\zeta(2s)}.$$

Hence, by (2.6) and (2.2) we get (2.7) and reach the expression (2.8). □

Remark 2.3. If we put $s = 1/2$ in the definition (2.4) of $\mathcal{H}(s)$, the infinite product $\prod_p \left(1 - \frac{2}{p} + \frac{2}{p^{3/2}} - \frac{2}{p^2} + \cdots\right)$ is divergent to zero. It is because that $\sum_{p \le x} \frac{1}{p} \sim \log\log x$, $\sum_{p \le x} \frac{1}{p^{3/2}} \ll 1$, and $0 < \frac{1}{p} - \frac{1}{p^{3/2}} \frac{1}{1+\frac{1}{p^{1/2}}} < 1/2$ for any $p \ge 2$ (see [19, p. 15]).

We need to know some weighted averages of $|h(n)|$ and $h(n)$.

Lemma 2.4 ([10, p. 418])**.** *We define the function*

$$\mathcal{G}(s) := \prod_p \left(1 + \frac{2}{p^{3s}} - \frac{1}{p^{4s}} - \frac{2}{p^{5s}}\right) \quad \textit{for } \operatorname{Re} s > \frac{1}{3}, \tag{2.10}$$

and $g(n)$ by

$$\mathcal{G}(s) = \sum_{n=1}^{\infty} \frac{g(n)}{n^s} \quad \textit{for } \operatorname{Re} s > \frac{1}{3}. \tag{2.11}$$

Let $h(n)$ be the function defined by (2.5). For $\operatorname{Re} s > 1/2$ we have

$$\begin{aligned} \sum_{n=1}^{\infty} \frac{|h(n)|}{n^s} &= \prod_p \left(1 + \frac{2}{p^{2s}} \left(\frac{1}{1 - \frac{1}{p^s}}\right)\right) \\ &= \prod_p \left(\frac{1}{1 - \frac{1}{p^{2s}}}\right)^2 \left(1 + \frac{2}{p^{3s}} - \frac{1}{p^{4s}} - \frac{2}{p^{5s}}\right) = \zeta^2(2s)\mathcal{G}(s). \end{aligned} \tag{2.12}$$

Also, the Dirichlet series, the right hand side of (2.11) is convergent absolutely for $\operatorname{Re} s > 1/3$ and

$$|h(n)| = \sum_{d^2 | n} \tau(d) g\left(\frac{n}{d^2}\right). \tag{2.13}$$

Proof. In view of (2.5) and (2.6), we see that $\prod_p \left(1 + \frac{2}{p^{2s}} + \frac{2}{p^{3s}} + \cdots\right) = \sum_{n=1}^{\infty} \frac{|h(n)|}{n^s}$, hence we get the first equality in (2.12) for $\operatorname{Re} s > 1/2$, where the infinite product is convergent absolutely. The second equality in (2.12) follows from the following identity:

$$1 + \frac{2X^2}{1 - X} = \frac{\left(1 + 2X^3 - X^4 - 2X^5\right)}{\left(1 - X^2\right)^2}$$

with $X = 1/p^s$. Since the Euler product $\prod_p \left(1 + 2/p^{3\sigma} + 1/p^{4\sigma} + 2/p^{5\sigma}\right)$ is convergent for $\sigma > 1/3$, we see $\sum_{n=1}^{\infty} g(n)n^{-s}$ is convergent absolutely in the same region. From (2.12), (2.11) and (1.7), we get (2.13). □

Lemma 2.5 ([10, p. 418], [18, p. 59])**.** *As $x \to \infty$, we have*

(1) $\displaystyle\sum_{n\le x} |h(n)| = O\left(x^{1/2}\log x\right),$

(2) $\displaystyle\sum_{n\le x} \frac{|h(n)|}{n^{1/2}} = O\left(\log^2 x\right),$

(3) $\displaystyle\sum_{n>x} \frac{|h(n)|}{n} = O\left(\frac{\log x}{x^{1/2}}\right),$

(4) $\displaystyle\sum_{n>x} \frac{|h(n)|\log n}{n} = O\left(\frac{\log^2 x}{x^{1/2}}\right).$

Proof. The first estimate (1) is shown in [10, p. 418]. Using the first estimate and partial summation we have the others, easily. □

The next lemma is a consequence of (2.6) and Lemma 2.5.

Lemma 2.6. *As $x \to \infty$, we have*

(1) $\displaystyle\sum_{n\le x} \frac{h(n)}{n} = \mathcal{H}(1) + O\left(\frac{\log x}{x^{1/2}}\right),$

(2) $\displaystyle\sum_{n\le x} \frac{h(n)\log n}{n} = -\mathcal{H}'(1) + O\left(\frac{\log^2 x}{x^{1/2}}\right).$

Under these preparations we prove Theorem 1.1.

Proof of Theorem 1.1. By the expression (2.8) of $|\mu(n)|2^{\omega(n)}$ and the formula (2.1), we have

$$\begin{aligned}\sum_{n\le x} |\mu(n)|2^{\omega(n)} &= \sum_{d\delta\le x} h(d)2^{\omega(\delta)} = \sum_{d\le x} h(d) \sum_{\delta\le x/d} 2^{\omega(\delta)} \\ &= \frac{6}{\pi^2} x\log x \sum_{d\le x} \frac{h(d)}{d} - \frac{6}{\pi^2} x \sum_{d\le x} \frac{h(d)\log d}{d} \\ &\quad + Cx \sum_{d\le x} \frac{h(d)}{d} + \sum_{d\le x} h(d) E\left(\frac{x}{d}\right), \end{aligned} \tag{2.14}$$

where $E(x)$ is defined in (1.1), that is, $E(x) := \sum_{n\le x} 2^{\omega(n)} - \frac{6}{\pi^2} x\log x - Cx$ $(x \ge 1)$. By $E(x) = O\left(x^{1/2}\log x\right)$ and Lemma 2.5 (2) we observe that

$$\sum_{d\le x} h(d) E\left(\frac{x}{d}\right) = O\left(x^{1/2}\log^3 x\right). \tag{2.15}$$

Moreover using Lemma 2.6, we have

$$\sum_{n\le x}|\mu(n)|2^{\omega(n)} = \frac{6}{\pi^2}\mathcal{H}(1)x\log x + \left(\frac{6}{\pi^2}\mathcal{H}'(1) + C\mathcal{H}(1)\right)x + O\left(x^{1/2}\log^3 x\right). \tag{2.16}$$

Noting that $\frac{6}{\pi^2}\mathcal{H}(1) = \prod_p(1+2/p)(1-1/p)^2 = C_1$ by (2.9), we reach the assertion of Theorem 1.1. □

Remark 2.7. Instead of using the elementary bound $E(x) = O\left(x^{1/2}\log x\right)$ in (2.15), we can consider using the bound (1.4) ([7], [6], [15]). Then we get

$$\sum_{d\le x} h(d)E\left(\frac{x}{d}\right) = O\left(x^{1/2}\log^2 x\right).$$

Note that the bound (1.4) is a consequence of complex analysis, e.g., the zero-free region of $\zeta(s)$ or $\sum_{n\le x}\mu(n) = O\left(x/\log x\right)$. Furthermore even if we use the estimate (1.3) instead of (1.4), we might not be able to improve $O\left(x^{1/2}\log^2 x\right)$.

3. Complex analytic method

As we have seen in (2.14), if we attempt to improve $O\left(x^{1/2}\log^3 x\right)$ in Theorem 1.1 or $O\left(x^{1/2}\log^2 x\right)$ by the continuation of the previous method, we may have to rely on the Riemann Hypothesis or Baker's result (1.5). To get the unconditional sharper result, we take a different approach here.

We begin with introducing the following two functions.

Definition 3.1. For $\mathrm{Re}\, s > 1/2$ we define $\mathcal{A}(s)$ and $\mathcal{B}(s)$ as follows.

$$\mathcal{A}(s) := \prod_p\left(1 - \frac{3}{p^{2s}} + \frac{2}{p^{3s}}\right), \tag{3.1}$$

$$\mathcal{B}(s) := \prod_p\left(1 + \frac{3}{p^{2s}} + \frac{2}{p^{3s}}\right). \tag{3.2}$$

We note that at $s = 1/2$ the infinite product $\prod_p\left(1 - \left(\frac{3}{p} - \frac{2}{p^{3/2}}\right)\right)$ is divergent to zero, because $0 < \frac{3}{p} - \frac{2}{p^{3/2}} < 1$, $\sum_{p\le x}\frac{1}{p} \sim \log\log x$, and $\sum_{p\le x}\frac{1}{p^{3/2}} \ll 1$ (see [19, p. 15]).

Let $a(n)$ be the multiplicative function defined by

$$a(n) = \begin{cases} 1 & n = 1, \\ 0 & n = p, \\ -3 & n = p^2, \\ 2 & n = p^3, \\ 0 & n = p^k,\ (k \geq 4), \end{cases} \tag{3.3}$$

for any prime p. Then by (3.1), (3.2), and (3.3) we have

$$\mathcal{A}(s) = \sum_{n=1}^{\infty} \frac{a(n)}{n^s} \quad \textit{and} \quad \mathcal{B}(s) = \sum_{n=1}^{\infty} \frac{|a(n)|}{n^s} \quad \textit{for } \operatorname{Re} s > 1/2. \tag{3.4}$$

We have a relation between $\mathcal{A}(s)$ and $\mathcal{Q}(s)$ defined by (2.3).

Lemma 3.2. *For* $\operatorname{Re} s > 1$ *we have*

$$\mathcal{Q}(s) = \mathcal{A}(s)\zeta^2(s). \tag{3.5}$$

By this identity we have

$$|\mu(n)|2^{\omega(n)} = \sum_{d|n} a(d)\tau\left(\frac{n}{d}\right). \tag{3.6}$$

Proof. As for the right hand side of (3.1), since the polynomial $1-3X^2+2X^3$ has zeros at $X = -1/2$ and 1, we observe that $\mathcal{A}(s) = \prod_p (1 + 2/p^s)(1 - 1/p^s)^2$ ($\operatorname{Re} s > 1/2$). Hence, by (2.9), we have the assertion (3.5). □

Lemma 3.3. *We define the function* $\mathcal{C}(s)$ *by for* $\operatorname{Re} s > 1/3$

$$\mathcal{C}(s) := \prod_p \left(1 + \frac{2}{p^{3s}} - \frac{6}{p^{4s}} - \frac{6}{p^{5s}} + \frac{8}{p^{6s}} + \frac{6}{p^{7s}} - \frac{3}{p^{8s}} - \frac{2}{p^{9s}}\right)$$

and write $c(n)$ *by*

$$\mathcal{C}(s) = \sum_{n=1}^{\infty} \frac{c(n)}{n^s} \quad \textit{for } \operatorname{Re} s > \frac{1}{3}. \tag{3.7}$$

Let $a(n)$ *be an arithmetical function defined by (3.3). Then, we have*

$$\sum_{n \leq x} |a(n)| = B_1\sqrt{x}\log^2 x + B_2\sqrt{x}\log x + B_3\sqrt{x} + O\left(x^{1/3+\varepsilon}\right) \tag{3.8}$$

for any $0 < \varepsilon < 1/6$, *where* B_1, B_2, *and* B_3 *are some constants. In particular,* $B_1 = \mathcal{C}(1/2)/8 > 0$.

Proof. From the identity

$$\begin{aligned}&\left(1+3X^2+2X^3\right)\left(1-X^2\right)^3\\&=1+2X^3-6X^4-6X^5+8X^6+6X^7-3X^8-2X^9\end{aligned}$$

with $X=p^{-s}$ ($\operatorname{Re}s>1/2$), we have

$$\mathcal{B}(s)=\zeta^3(2s)\mathcal{C}(s). \tag{3.9}$$

Thus by (3.4), (3.9), (1.7), and (3.7) we have

$$|a(n)|=\sum_{d^2|n}\tau_3(d)c\left(\frac{n}{d^2}\right).$$

We shall apply the asymptotic formula (1.10) for $\sum_{n\le x}\tau_3(n)$ to estimate the sum $\sum_{n\le x}|a(n)|$. In fact we have

$$\begin{aligned}\sum_{n\le x}|a(n)|&=\sum_{n\le x}\sum_{\delta^2|n}\tau_3(\delta)c\left(\frac{n}{\delta^2}\right)=\sum_{d\le x}c(d)\sum_{\delta\le\sqrt{x/d}}\tau_3(\delta)\\&=\frac{1}{8}\sqrt{x}\log^2x\sum_{d\le x}\frac{c(d)}{d^{1/2}}\\&\quad-\frac{1}{4}\sqrt{x}\log x\sum_{d\le x}\frac{c(d)\log d}{d^{1/2}}+\frac{3\gamma-1}{2}\sqrt{x}\log x\sum_{d\le x}\frac{c(d)}{d^{1/2}}\\&\quad+\frac{1}{8}\sqrt{x}\sum_{d\le x}\frac{c(d)\log^2d}{d^{1/2}}-\frac{3\gamma-1}{2}\sqrt{x}\sum_{d\le x}\frac{c(d)\log d}{d^{1/2}}\\&\quad+\left(3\gamma_1+3\gamma^2-3\gamma+1\right)\sqrt{x}\sum_{d\le x}\frac{c(d)}{d^{1/2}}+O\left(x^{1/4}\sum_{d\le x}\frac{|c(d)|}{d^{1/4}}\right).\end{aligned}$$

Let ε be a real number such that $0<\varepsilon<1/6$. Since the Dirichlet series (3.7) is convergent absolutely for $\operatorname{Re}s>1/3$, we have

$$\sum_{d\le x}\frac{|c(d)|}{d^{1/4}}=\sum_{d\le x}\frac{|c(d)|}{d^{1/3+\varepsilon}d^{-1/12-\varepsilon}}\le x^{1/12+\varepsilon}\sum_{d=1}^{\infty}\frac{|c(d)|}{d^{1/3+\varepsilon}}\ll x^{1/12+\varepsilon},$$

$$\sum_{d\le x}\frac{c(d)}{d^{1/2}}=\mathcal{C}\left(\frac{1}{2}\right)+O\left(\sum_{d>x}\frac{|c(d)|}{d^{1/3+\varepsilon}d^{1/6-\varepsilon}}\right)=\mathcal{C}\left(\frac{1}{2}\right)+O\left(x^{-1/6+\varepsilon}\right),$$

and similarly $\sum_{d\le x}\frac{c(d)\log d}{d^{1/2}}=-\mathcal{C}'(1/2)+O\left(x^{-1/6+\varepsilon}\right)$, $\sum_{d\le x}\frac{c(d)\log^2d}{d^{1/2}}=\mathcal{C}''(1/2)+O\left(x^{-1/6+\varepsilon}\right)$. Therefore we obtain (3.8). Note that the implied constants depend on ε. □

Lemma 3.4. *For* $\operatorname{Re} s > 1/2$, *we define*

$$\mathcal{A}_1(s) := \prod_p \left(1 - \frac{3}{p^{2s}} + \frac{2}{p^{3s}}\right)\left(1 - \frac{1}{p^{2s}}\right)^{-3} = \mathcal{A}(s)\zeta^3(2s).$$

Then $\mathcal{A}_1(s)$ *is holomorphic for* $\operatorname{Re} s > 1/3$.

Proof. We observe the identity for $X^2 \neq 1$

$$\frac{1 - 3X^2 + 2X^3}{(1-X^2)^3} = 1 + 2X^3 + \frac{X^4(-3 + 6X + X^2 - 6X^3 + 2X^5)}{(1-X^2)^3}.$$

Setting $X = X_p = 1/p^s$ for each prime p (here, $\operatorname{Re} s > 1/2$), we have

$$\mathcal{A}(s)\zeta^3(2s) = \prod_p \left(1 + 2X_p^3 + \frac{X_p^4(-3 + 6X_p + X_p^2 - 6X_p^3 + 2X_p^5)}{(1-X_p^2)^3}\right)$$

for $\operatorname{Re} s > 1/2$. Since the infinite product is absolutely convergent for $\operatorname{Re} s > 1/3$, we see that $\mathcal{A}_1(s)$ is holomorphic in the region $\operatorname{Re} s > 1/3$. □

Here we shall recall a new type of Perron's formula due to [12].

Lemma 3.5 (Liu and Ye [12, p. 483, Theorem 2.1])**.** *Let* $f(s) = \sum_{n=1}^{\infty} \frac{a_n}{n^s}$ *be a Dirichlet series satisfying* $B(\sigma) := \sum_{n=1}^{\infty} \left|\frac{a_n}{n^s}\right| = \sum_{n=1}^{\infty} \frac{|a_n|}{n^\sigma}$ *is convergent for* $\sigma > \sigma_a$. *Choose* $b > \sigma_a$, $x \geq 2$, $T \geq 2$, $H \geq 2$. *Then we have*

$$\sum_{n \leq x} a_n = \frac{1}{2\pi i}\int_{b-iT}^{b+iT} f(s)\frac{x^s}{s}ds + O\left(\sum_{x - x/H < n \leq x + x/H} |a_n|\right) + O\left(\frac{x^b H B(b)}{T}\right).$$

Using Lemma 3.5 we shall prove the following estimate on the partial sum of $a(n)$ defined by (3.3).

Lemma 3.6. *For any sufficiently large* $x \geq e^e$, *there exists a positive constant* $\alpha_0 > 0$ *such that*

$$\sum_{n \leq x} a(n) = O\left(x^{1/2}\exp\left(-\alpha_0 \frac{(\log x)^{3/5}}{(\log\log x)^{1/5}}\right)\right). \tag{3.10}$$

Proof. In Lemma 3.5, we take $f(s) = \mathcal{A}(s)$, $B(\sigma) = \mathcal{B}(\sigma)$ of (3.4), $x \geq 4$, $T \geq 4$, $H \geq 2$, and $b = 1/2 + 1/\log x$. From (3.9) we have $\mathcal{B}(b) \ll \log^3 x$. Then we get

$$\sum_{n \leq x} a(n) = \frac{1}{2\pi i}\int_{b-iT/2}^{b+iT/2} \mathcal{A}(s)\frac{x^s}{s}ds$$

$$+O\left(\sum_{x-x/H<n\le x+x/H}|a(n)|\right)+O\left(\frac{x^{1/2}H\log^3 x}{T}\right). \quad (3.11)$$

Now we choose $H=\sqrt{T}$. Using (3.8) we find that

$$O\left(\sum_{x-x/H<n\le x+x/H}|a(n)|\right)=O\left(\frac{x^{1/2}\log^2 x}{\sqrt{T}}\right)+O\left(x^{1/3+\varepsilon}\right).$$

To evaluate the integral on the right hand side of (3.11), we shall apply the result of the zero-free region of $\zeta(s)$. According to Ivić [10, p. 310], we can find an absolute constant $A>0$ such that

$$\frac{1}{\zeta(2s)}\ll(\log T)^{2/3}(\log\log T)^{1/3} \quad (3.12)$$

in the region

$$\sigma>\frac{1}{2}-\frac{A}{(\log 2t)^{2/3}(\log\log 2t)^{1/3}},\quad t_0<t\le T/2. \quad (3.13)$$

Hereafter we write $b'=\frac{1}{2}-\frac{A}{2(\log T)^{2/3}(\log\log T)^{1/3}}$ for short. Take the rectangle with four vertices $b+i\frac{T}{2}$, $b'+i\frac{T}{2}$, $b'-i\frac{T}{2}$, and $b-i\frac{T}{2}$. Since by Lemma 3.4 $\mathcal{A}(s)=\zeta^{-3}(2s)\cdot\mathcal{A}_1(s)$ is holomorphic on this rectangle, we have by Cauchy's theorem

$$\begin{aligned}\frac{1}{2\pi i}\int_{b-iT/2}^{b+iT/2}\mathcal{A}(s)\frac{x^s}{s}ds=&-\frac{1}{2\pi i}\int_b^{b'}\mathcal{A}\left(\sigma+i\frac{T}{2}\right)\frac{x^{\sigma+iT/2}}{\sigma+i\frac{T}{2}}d\sigma\\&-\frac{1}{2\pi i}\int_{b'}^{b}\mathcal{A}\left(\sigma-i\frac{T}{2}\right)\frac{x^{\sigma-iT/2}}{\sigma-i\frac{T}{2}}d\sigma\\&-\frac{1}{2\pi i}\int_{T/2}^{-T/2}\mathcal{A}(b'+it)\frac{x^{b'+it}}{b'+it}idt. \quad (3.14)\end{aligned}$$

From the proof of Lemma 3.4 we observe that $|\mathcal{A}_1(s)|\le\exp\left(c\sum_p p^{-3\sigma}\right)$ for $\sigma>1/3$, here $c>0$ is a constant. Using this bound and (3.12), we have

$$\frac{1}{2\pi i}\int_b^{b'}\mathcal{A}\left(\sigma+i\frac{T}{2}\right)\frac{x^{\sigma+iT/2}}{\sigma+i\frac{T}{2}}d\sigma\ll\frac{x^b(\log T)^3}{T}.$$

Also we get the same estimate for the second integral. As for the third integral we see

$$\frac{1}{2\pi i}\int_{-T/2}^{T/2}\zeta^{-3}(2b'+i2t)\mathcal{A}_1(b'+it)\frac{x^{b'+it}}{b'+it}idt=O\left(x^{b'}\log^4 T\right).$$

Finally taking

$$T = \exp\left((\log x)^{3/5}(\log\log x)^{-1/5}\right), \tag{3.15}$$

we obtain the assertion (3.10). □

Remark 3.7. To prove Lemma 3.6 we might apply the usual Perron's formula e.g. [18, p. 220, Corollay 2.4]). However, in our case the abscissa of absolute convergence of $\sum_{n=1}^{\infty}\frac{a(n)}{n^s}$ is smaller than 1, and moreover we have an asymptotic formula of $\sum_{n\le x}|a(n)|$ (Lemma 3.3 (3.8)). In such a case it seems to be effective to use a new type Perron's formula of Liu and Ye (Lemma 3.5) for a proof of Lemma 3.6.

Using the bound (3.10) in Lemma 3.6 and noting $\mathcal{A}(1) = C_1$ (where C_1 is same in the formula (1.6)) and Lemma 3.3, we have the next lemma.

Lemma 3.8. *Keep α_0 as in* (3.10). *As $x\to\infty$, we have*

(1) $$\sum_{n\le x}\frac{a(n)}{n} = C_1 + O\left(x^{-1/2}\exp\left(-\alpha_0\frac{(\log x)^{3/5}}{(\log\log x)^{1/5}}\right)\right),$$

(2) $$\sum_{n\le x}\frac{a(n)\log n}{n} = -\mathcal{A}'(1) + O\left(x^{-1/2}\exp\left(-\frac{\alpha_0}{2}\frac{(\log x)^{3/5}}{(\log\log x)^{1/5}}\right)\right),$$

(3) $$\sum_{n\le x}\frac{|a(n)|}{n^{1/3}} = O\left(x^{1/6}\log^2 x\right).$$

Proof. Since $\sum_{n=1}^{\infty}\frac{a(n)}{n}$ is convergent in (3.4) we see that

$$\sum_{n\le x}\frac{a(n)}{n} = \mathcal{A}(1) - \lim_{X\to\infty}\sum_{x<n\le X}\frac{a(n)}{n}.$$

By partial summation and (3.10) in Lemma 3.6 we get the first assertion (1). Since $\sum_{n=1}^{\infty}\frac{a(n)\log n}{n} = -(\mathcal{A}(s))'|_{s=1}$ we have

$$\sum_{n\le x}\frac{a(n)\log n}{n} = -\mathcal{A}'(1) - \lim_{X\to\infty}\sum_{x<n\le X}\frac{a(n)\log n}{n}.$$

Similar to the above we get the second assertion (2). To show the third assertion (3) we use partial summation and Lemma 3.3. □

Applying the method of [17] (cf. [16]) and the above lemmas we shall improve Theorem 1.1.

Proof of Theorem 1.2. Let ρ be a function of x satisfying $0 < \rho = \rho(x) < 1$ and $\rho(x)\, x \to \infty$ as $x \to \infty$. By (3.6) in Lemma 3.2 we observe that

$$\begin{aligned}\sum_{n\le x}|\mu(n)|2^{\omega(n)} &= \left(\sum_{\substack{d\delta\le x\\ \delta\le\rho^{-1}}} + \sum_{\substack{d\le\rho x\\ \rho^{-1}<\delta\le x/d}}\right) a(d)\tau(\delta)\\ &=: T_1 + T_2 \quad \text{(say)}.\end{aligned}$$

We shall estimate these T_1 and T_2.

As for T_2, using the result (1.9) for $\sum_{n\le x}\tau(n)$ we have

$$\begin{aligned}T_2 = {}& x\log x \sum_{d\le\rho x}\frac{a(d)}{d} - x\sum_{d\le\rho x}\frac{a(d)\log d}{d} + (2\gamma-1)x\sum_{d\le\rho x}\frac{a(d)}{d}\\ &+ O\left(x^{1/3}\sum_{d\le\rho x}\frac{|a(d)|}{d^{1/3}}\right) - \rho^{-1}(\log\rho^{-1} + (2\gamma-1))\sum_{d\le\rho x}a(d)\\ &+ O\left(\rho^{-1/3}\left|\sum_{d\le\rho x}a(d)\right|\right).\end{aligned}$$

Then, by Lemmas 3.6 and 3.8 we have

$$\begin{aligned}T_2 = {}& C_1 x\log x + \big(C_1(2\gamma-1) + \mathcal{A}'(1)\big)\, x\\ &+ O\left(x^{1/2}\rho^{-1/2}(\log x)\exp\left(-\alpha_0\frac{(\log\rho x)^{3/5}}{(\log\log\rho x)^{1/5}}\right)\right)\\ &+ O\left(x^{1/2}\rho^{-1/2}\exp\left(-\frac{\alpha_0}{2}\frac{(\log\rho x)^{3/5}}{(\log\log\rho x)^{1/5}}\right)\right)\\ &+ O\left(x^{1/2}\rho^{1/6}(\log\rho x)^2\right)\\ &+ O\left(x^{1/2}\rho^{-1/2}(\log\rho^{-1}+1)\exp\left(-\alpha_0\frac{(\log\rho x)^{3/5}}{(\log\log\rho x)^{1/5}}\right)\right).\end{aligned}$$

We write $C_2^* = C_1(2\gamma-1) + \mathcal{A}'(1)$. Note that α_0 is same as in Lemma 3.6. Now, we choose $\rho = \exp\left(-\frac{\alpha_0}{4}\frac{(\log x)^{3/5}}{(\log\log x)^{1/5}}\right)$. Then we see that

$$\begin{cases}\log\rho^{-1} = \dfrac{\alpha_0}{4}\dfrac{(\log x)^{3/5}}{(\log\log x)^{1/5}} \ll \log x,\\ \rho^{-1/2} = \exp\left(\dfrac{\alpha_0}{8}\dfrac{(\log x)^{3/5}}{(\log\log x)^{1/5}}\right).\end{cases} \tag{3.16}$$

Therefore we obtain that

$$T_2 = C_1 x\log x + C_2^* x + O\left(x^{1/2}\rho^{1/6}(\log \rho x)^2\right) + O\left(x^{1/2}\exp\left(-\frac{\alpha_0}{2}\frac{(\log \rho x)^{3/5}}{(\log\log \rho x)^{1/5}} + \log\log x + \frac{\alpha_0}{8}\frac{(\log x)^{3/5}}{(\log\log x)^{1/5}}\right)\right).$$

To bound T_1 we note that the function $\frac{(\log x)^{3/5}}{(\log\log x)^{1/5}}$ $(x \geq e^e)$ is increasing, monotonically, and $\sum_{n\leq x}\frac{\tau(n)}{n^{1/2}} \ll x^{1/2}\log x$ which is deduced from partial summation and (1.8). Using these two facts, Lemma 3.6 and (3.16) we have

$$\begin{aligned} T_1 &= O\left(\sum_{\delta\leq\rho^{-1}}\tau(\delta)\cdot\left(\frac{x}{\delta}\right)^{1/2}\exp\left(-\alpha_0\frac{\left(\log\frac{x}{\delta}\right)^{3/5}}{\left(\log\log\frac{x}{\delta}\right)^{1/5}}\right)\right)\\ &= O\left(x^{1/2}\exp\left(-\alpha_0\frac{(\log \rho x)^{3/5}}{(\log\log \rho x)^{1/5}}\right)\sum_{\delta\leq\rho^{-1}}\frac{\tau(\delta)}{\delta^{1/2}}\right)\\ &= O\left(x^{1/2}\rho^{-1/2}\log\rho^{-1}\exp\left(-\alpha_0\frac{(\log \rho x)^{3/5}}{(\log\log \rho x)^{1/5}}\right)\right)\\ &= O\left(x^{1/2}\exp\left(-\frac{\alpha_0}{2}\frac{(\log \rho x)^{3/5}}{(\log\log \rho x)^{1/5}} + \log\log x + \frac{\alpha_0}{8}\frac{(\log x)^{3/5}}{(\log\log x)^{1/5}}\right)\right). \end{aligned}$$

That is, we obtain

$$\sum_{n\leq x}|\mu(n)|2^{\omega(n)} = C_1 x\log x + C_2^* x + O\left(x^{1/2}\rho^{1/6}(\log \rho x)^2\right) + O\left(x^{1/2}\exp\left(-\frac{\alpha_0}{2}\frac{(\log \rho x)^{3/5}}{(\log\log \rho x)^{1/5}} + \log\log x + \frac{\alpha_0}{8}\frac{(\log x)^{3/5}}{(\log\log x)^{1/5}}\right)\right). \tag{3.17}$$

For any sufficient large $x \geq e^e$, since $\rho x \geq x^{1/2}$ we observe that

(a) $\exp\left(-\frac{\alpha_0}{2}\frac{(\log \rho x)^{3/5}}{(\log\log \rho x)^{1/5}}\right) \leq \exp\left(-\frac{\alpha_0}{2\cdot 2^{3/5}}\frac{(\log x)^{3/5}}{(\log\log x)^{1/5}}\right)$,

(b) $\rho^{1/6}(\log \rho x)^2 \leq \exp\left(-\frac{\alpha_0}{48}\frac{(\log x)^{3/5}}{(\log\log x)^{1/5}}\right)$.

Now the assertion of Theorem 1.2 follows easily from the estimates (3.17), (a), and (b). Here note that $C_2^* = C_2$. Actually, since $\mathcal{H}(s) = \mathcal{A}(s)\zeta(2s)$,

we observe in (2.16) that

$$C_2 = \frac{6}{\pi^2}\mathcal{H}(1)\left(\frac{\mathcal{H}'(1)}{\mathcal{H}(1)} + 2\gamma - 1 - 2\frac{\zeta'(2)}{\zeta(2)}\right) = \frac{6}{\pi^2}\mathcal{H}(1)\left(\frac{\mathcal{A}'(1)}{\mathcal{A}(1)} + 2\gamma - 1\right)$$
$$= \frac{6}{\pi^2}\frac{\mathcal{H}(1)}{\mathcal{A}(1)}\mathcal{A}'(1) + C_1(2\gamma - 1) = C_2^*.$$

□

4. Revisiting the previous elementary method

It might be difficult to apply the method of proof of Theorem 1.2 to $T(x) = \sum_{d\delta \le x} h(d)2^{\omega(\delta)}$ in the proof of Theorem 1.1. In fact, it is a problem to improve (2.15) without using (1.3) and (1.5) (the latter assertion holds under the Riemann hypothesis). Applying the method in the proof of Theorem 1.2 to estimate $\sum_{n\le x} h(n)$, moreover, using the proof of Theorem 1.1 and the result of Theorem 1.2 we can improve (2.15) as follows.

Theorem 4.1. *Let $E(x)$ be the function defined by (1.1) and $h(n)$ the function defined by (2.5). For any sufficiently large $x \ge e^e$ there exists a positive constant $\beta_* > 0$ satisfying*

$$\sum_{n\le x} h(n)E\left(\frac{x}{n}\right) = O\left(x^{1/2}\exp\left(-\beta_*\frac{(\log x)^{3/5}}{(\log\log x)^{1/5}}\right)\right).$$

In order to prove this we shall prepare some lemmas.

Lemma 4.2. *Let $\varepsilon > 0$ be sufficiently small. We have*

$$\sum_{n\le x}|h(n)| = G_1\sqrt{x}\log x + G_2\sqrt{x} + O\left(x^{1/3+\varepsilon}\right), \tag{4.1}$$

where $G_1 = \mathcal{G}(1/2)/2$ and $G_2 = (2\gamma-1)\mathcal{G}(1/2)/2$ ($\mathcal{G}(s)$ is defined by (2.10)).

Proof. We go back to Lemma 2.4 and recall (2.13). By (1.8) we get

$$\sum_{n\le x}|h(n)| = \frac{1}{2}\sqrt{x}\log x\sum_{d\le x}\frac{g(d)}{d^{1/2}} - \frac{1}{2}\sqrt{x}\sum_{d\le x}\frac{g(d)\log d}{d^{1/2}}$$
$$+ (2\gamma-1)\sqrt{x}\sum_{d\le x}\frac{g(d)}{d^{1/2}} + O\left(x^{1/4}\sum_{d\le x}\frac{|g(d)|}{d^{1/4}}\right).$$

Since $\mathcal{G}(s) = \sum_{n=1}^{\infty} g(n)n^{-s}$ is convergent absolutely for $\operatorname{Re} s > 1/3$, we get the assertion (4.1) by a similar argument of the proof of Lemma 3.3. □

Lemma 4.3. *For* $\operatorname{Re} s > \frac{1}{2}$ *we define the function* $\mathcal{H}_1(s)$ *by*

$$\mathcal{H}_1(s) := \prod_p \left(1 - \frac{2}{p^{2s}} + \frac{2}{p^{3s}} - \cdots\right)\left(1 - \frac{1}{p^{2s}}\right)^{-2} = \mathcal{H}(s)\zeta^2(2s).$$

Then, $\mathcal{H}_1(s)$ *is holomorphic for* $\operatorname{Re} s > 1/3$.

Proof. Since we see that for $X^2 \neq 1$,

$$\left(1 - \frac{2X^2}{1+X}\right)\frac{1}{(1-X^2)^2} = 1 + \frac{2X^3 - X^4 - X^5}{(1+X)(1-X^2)^2}$$

we have that for $\operatorname{Re} s > 1/3$

$$\mathcal{H}_1(s) = \mathcal{H}(s)\zeta^2(2s) = \prod_p \left(1 + \frac{\frac{2}{p^{3s}} - \frac{1}{p^{4s}} - \frac{1}{p^{5s}}}{\left(1 + \frac{1}{p^s}\right)\left(1 - \frac{1}{p^{2s}}\right)^2}\right),$$

which implies $\mathcal{H}_1(s)$ is holomorphic for $\operatorname{Re} s > 1/3$. □

Lemma 4.4. *For any sufficiently large* $x \geq e^e$, *there exists a positive constant* $\beta_0 > 0$ *satisfying*

$$\sum_{n \leq x} h(n) = O\left(x^{1/2}\exp\left(-\beta_0 \frac{(\log x)^{3/5}}{(\log\log x)^{1/5}}\right)\right). \tag{4.2}$$

Proof. In Lemma 3.5, we choose $x \geq 4$, $H = \sqrt{T}$ $(T \geq 4)$, $f(s) = \mathcal{H}(s)$, $B(\sigma) = \sum_{n=1}^{\infty} \frac{|h(n)|}{n^\sigma}$, $b = 1/2 + 1/\log x$, then by Lemmas 2.4 and 4.2.

$$\sum_{n \leq x} h(n) = \frac{1}{2\pi i}\int_{b-iT/2}^{b+iT/2} \mathcal{H}(s)\frac{x^s}{s}ds + O\left(\frac{\sqrt{x}\log^2 x}{\sqrt{T}}\right) + O\left(x^{1/3+\varepsilon}\right),$$

where $\varepsilon > 0$ is sufficiently small. Since by Lemma 4.3 $\mathcal{H}(s) = \zeta^{-2}(2s)\cdot\mathcal{H}_1(s)$ is holomorphic in the region (3.13), we can rewrite the above integral as (3.14). Hence using (3.12), $\mathcal{H}_1(s) \ll \exp\left(c\sum_p \frac{1}{p^{3\sigma}}\right)$ $(\sigma > 1)$ and (3.15) we obtain the assertion (4.2). □

By Lemma 4.4 we have easily the following estimates.

Lemma 4.5. *Keep* β_0 *in (4.2). As* $x \to \infty$, *we have*

(1) $$\sum_{n \leq x} \frac{h(n)}{n} = \mathcal{H}(1) + O\left(x^{-1/2}\exp\left(-\beta_0 \frac{(\log x)^{3/5}}{(\log\log x)^{1/5}}\right)\right),$$

(2) $$\sum_{n \leq x} \frac{h(n)\log n}{n} = -\mathcal{H}'(1) + O\left(x^{-1/2}\exp\left(-\frac{\beta_0}{2} \frac{(\log x)^{3/5}}{(\log\log x)^{1/5}}\right)\right).$$

Finally, we shall show Theorem 4.1.

Proof of Theorem 4.1. In (2.14), we use Lemma 4.5, then we have

$$\sum_{d\le x} h(d)E\left(\frac{x}{d}\right) = T(x) - C_1 x\log x - C_2 x + O\left(x^{1/2}\exp\left(-\frac{\beta_0}{2}\frac{(\log x)^{3/5}}{(\log\log x)^{1/5}}\right)\right).$$

Moreover, we use Theorem 1.2 in the right hand side to obtain the assertion of Theorem 4.1. □

Acknowledgement: The second author would like to say thanks to Prof. Debdip Ganguly, IISER (India, Pune) and Dr. Tadaaki Igawa. This research was promoted in IISER when the second author was invited by Prof. Debdip Ganguly. Moreover Section 3 of this paper is a development from a discussion with Dr. Igawa for Master Thesis of Yamaguchi University. Finally, we say thanks to the referee of this paper for giving his or her careful and kind suggestion to us.

References

[1] Atkinson, F. V., A divisor problem, *Quart. J. Math.* **12** (1941), 193–200.
[2] Baker, R. C., The square-free divisor problem, *Quart. J. Math.* (2) **45** (1994), 269–277.
[3] Baker, R. C., The square-free divisor problem II, *Quart. J. Math.* (2) **47** (1996), 133–146.
[4] Cohen, E., The number of unitary divisors of an integer, *Amer. Math. Monthly* **67** (1960), 879–880.
[5] Corput, J. G. van der, Zum Teilerproblem, *Math. Ann.* **98** (1928), 697–716.
[6] Gioia, A. A. and Vaidya, A. M., The number of squarefree divisors of an integer, *Duke Math. J.* **33** (1966), 797–799.
[7] Hölder, O., Über ein asymptotische Ausdruck, *Acta Math.* **59** (1932), 89–97.
[8] Huxley, M. N., Exponential sums and lattice points III, *Proc. London Math. Soc.* **87** (2003), 591–609.
[9] Ivić, A., A convolution theorem with applications to some divisor problems, *Publ. Inst. Math. (Belgrade)* **24** (38) (1978), 67–78.
[10] Ivić, A., *The Riemann Zeta-Function*, Dover, 2003.
[11] Landau, E., *Handbuch der Leher von der Verteilung der Primzahlem*, Leipzig, 1909.
[12] Liu, J. and Ye, Y., Perron's formula and the prime number theorem for automorphic L-functions, *Pure Appl. Math. Quart.* **3** (2007), 481–497.
[13] Mertens, F., Über einige asymptotische Gesetze der Zahlentheorie, *J. Reine Angew. Math.* **77** (1874), 289–338.

[14] Montgomery, H. L. and Vaughan, R. C., *Multiplicative Number Theory I. Classical Theory*, Cambridge Univ. Press, 2007.
[15] Saffari, B., Sur le nombre de diviseurs "r-free" d'un entier. Cas $r = 2$ et $r = 3$, *C. R. Acad. Sci. Paris* **265** (1967), 705–707.
[16] Suetuna, Z., On the number of decompositions of an integer into two relatively prime factors, *Japanese J. Math.* **1** (1924), 5–7.
[17] Suryanarayana, D. and Siva Rama Prasad, V., The number of k-free divisors of an integer, *Acta Arith.* **17** (1970/1971), 345–354.
[18] Tenenbaum, G., *Introduction to Analytic and Probabilistic Number Theory* (third ed.), Graduate Studies in Mathematics **163** Amer. Math. Soc. 2015.
[19] Titchmarsh, E. C., *The Theory of Functions* (second edition), Oxford Univ. Press, 1932.
[20] Titchmarsh, E. C., *The Theory of the Riemann Zeta-function*, 2nd ed.rev. by Heath-Brown, D. R., Oxford Univ. Press, 1986.

Debika Banerjee
Department of Mathematics
Indraprastha Institute of Information Technology
Okhla Phase III, 110020, India
E-mail: devikabanerjee12@gmail.com

Tokuhon Makoto Minamide
Graduate School of Sciences and Technology for Innovation
Yamaguchi University
Yoshida 1677-1, Yamaguchi 753-8512, Japan
E-mail: minamide@yamaguchi-u.ac.jp

Yoshio Tanigawa
Nishizato 2-13-1, Meito, Nagoya 465-0084, Japan
E-mail: tanigawa@math.nagoya-u.ac.jp

The Mathematics Student ISSN: 0025-5742
Vol. 95, Nos. 3-4, July - December (2026), 20–27

A SHORT NOTE ON OPEN BOOKS OF $\#_K S^1\widetilde{\times}S^2$

ABHIJEET GHANWAT, SUHAS PANDIT, AND SANJAY KUTE
(Received : 04 - 10 - 2024 ; Accepted : 08 - 09 - 2026)

Abstract. In this short note, we give an application of a recent result of Miller and Naylor. We show the first Betti number of the page of an open book of $\#_k S^1\widetilde{\times}S^2$ agrees with k if and only if the monodromy of the open book is isotopic to the identity map.

1. Introduction

In 2023, Miller and Naylor [5], showed that for $g \geq k \geq 0$, any genus g Heegaard splitting of $M = \#_k S^1\widetilde{\times}S^2$ is equivalent to a Heegaard splitting of M obtained by stabilizing the standard genus k Heegaard splitting of M $g-k$ times. In particular, any two genus k Heegaard splittings of $\#_k S^1\widetilde{\times}S^2$ are equivalent. Using this, in Section 3, we prove that the first Betti number of the page of an open book of $\#_k S^1\widetilde{\times}S^2$ agrees with k if and only if the monodromy of the open book is isotopic to the identity map, see Theorem 3.3. This result is an analogue –in non-orientable setup– of a result of Ghiggini and Lisca [3].

2. Preliminaries

In this section, we recall the necessary definitions and results required for this article. In this article, all the diffeomorphisms and isotopies of a manifold with non-empty boundary are assumed to be the identity in a collar neighborhood of the boundary. Let us begin with the following definition.

Definition 2.1. An open book decomposition of a closed 3–manifold M is a pair (B, π), where

(1) B is a link in M with a trivial normal bundle,
(2) $\pi : M \setminus B \to S^1$ is a fiber bundle,

2010 Mathematics Subject Classification: 57R45, 57R15, 57R65
Key words and phrases: manifolds, open books, Heegaard decomposition
© *Indian Mathematical Society, 2026.*

(3) there is a trivialization of a tubular neighborhood $\mathcal{N}(B) = B \times D^2$ of B in M such that the map $\pi : B \times D^2 \setminus B \times \{0\} \to S^1$ sends (x, r, θ) to θ, where $x \in B$ and (r, θ) are the polar coordinates on D^2.
(4) For each $\theta \in S^1$, the closure of $\pi^{-1}(\theta)$ is a compact connected surface Σ with $\partial\Sigma = B$.

Here, B is called the binding, and Σ is called the page of the open book (B, π).

The page of an open book of an orientable (non-orientable) 3–manifold is orientable (non-orientable). Note that an open book (B, π) of M is completely determined up to diffeomorphism of M by the page Σ of the open book (B, π) and the monodromy ϕ of the fibration $\pi : M \setminus B \to S^1$ obtained from a flow on M which fixes the binding pointwise. More precisely, the monodromy ϕ is a diffeomorphism $\phi : \Sigma \to \Sigma$ that induces a diffeomorphism $F : M_{(\Sigma,\phi)} \to M$, where

$$M_{(\Sigma,\phi)} = \mathcal{M}T(\Sigma, \phi) \bigcup_{Id} \partial\Sigma \times D^2 \text{ with } \mathcal{M}T(\Sigma, \phi) = \frac{\Sigma \times [0, 2\pi]}{(x, 2\pi) \sim (\phi(x), 0)},$$

such that

(1) the map $\pi \circ F : \mathcal{M}T(\Sigma, \phi) \to S^1$ is given by $(\pi \circ F)(y, \theta) = e^{i\theta}$,
(2) the map $\pi \circ F : \partial\Sigma \times D^2 \setminus \partial\Sigma \times \{0\} \to S^1$ is given by $(\pi \circ F)(x, r, \theta) = \theta$, where x is a coordinate on $\partial\Sigma$ and (r, θ) are the polar coordinates on D^2.

Note that the monodromy ϕ of the open book is determined up to isotopy and conjugation by a self-diffeomorphism of Σ. We call the pair (Σ, ϕ) an *abstract open book* of M associated with the open book of (B, π). One can easily see that given an open book of a closed 3–manifold, we can associate an abstract open book and vice versa. Therefore, we will not distinguish between open books and abstract open books up to diffeomorphism. For more details on open books, see [1].

Given an open book (B, π) of a closed non-orientable 3–manifold M with the page Σ, we can associate a Heegaard splitting of M as follows: First, we recall the definition of a Heegarard splitting of closed non-orientable 3–manifold.

Definition 2.2. A Heegaard splitting of a closed non-orientable 3–manifold M is a decomposition of $M = H_g \cup H'_g$ into two genus g non-orinetable

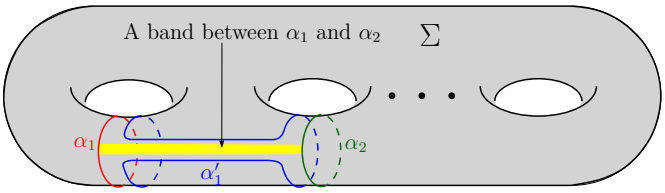

FIGURE 1. The curve α_1' is obtained by sliding α_1 over α_2 along the yellow band connecting α_1 and α_2.

handlebodies H_g and H_g', each diffeomorphic to $\natural_g S^1 \widetilde{\times} D^2$, such that $H_g \cap H_g' = \partial H_g = \partial H_g'$. The surface $S = \partial H_g$ is called the Heegaard surface of the Heegaard splitting.

Note that the Heegaard surface S is a non-orientable surface of genus $2g$, i.e., S is diffeomorphic to the connected sum of $2g$ copies of RP^2's, the projective planes.

Let θ_1 and θ_2 be two distinct points on S^1. Let C_1 and C_2 be two embedded arcs on S^1 with $C_1 \cap C_2 = \{\theta_1, \theta_2\}$. Note that $\mathcal{H}_g = \overline{\pi^{-1}(C_1)}$ and $\mathcal{H}_g' = \overline{\pi^{-1}(C_2)}$ are genus g handlebodies, where g is the first Betti number $b_1(\Sigma)$ of the page Σ. Now, it is easy to see that

$$M = \mathcal{H}_g \cup \mathcal{H}_g'$$

is a genus g Heegaard spltting of M associated with the open book (B, π). One can easily see that a Heegaard splitting $M = H_g \cup H_g'$ of a closed non-orientable 3–manifold M is determined by a triple (Σ, α, β), where

(1) Σ is the Heegaard surface of the Heegaard splitting,
(2) α and β are cut systems on Σ such that they bound properly embedded g disjoint discs in H_g and H_g', respectively.

Recall that by a *cut system* on a closed non-orientable surface Σ of genus $2g$ is a collection $\alpha = \{\alpha_1, \alpha_2, \ldots, \alpha_g\}$ of g disjoint non-separating two-sided simple closed curves on Σ such that α cuts Σ into a connected planar surface.

Definition 2.3. Let α_1 and α_2 be two disjoint two-sided simple closed curves in a surface Σ. We say that a simple closed curve α_1' on Σ is obtained by sliding α_1 over α_2 if α_1' is isotopic to a band connected sum of α_1 with α_2.

Definition 2.4. Let $\alpha = \{\alpha_1, \alpha_2, ..., \alpha_g\}$ and $\alpha' = \{\alpha'_1, \alpha'_2, ..., \alpha'_g\}$ be two cut systems on a closed non-orientable surface Σ of genus $2g$. We say that α and α' are handle slide equivalent if there exists a sequence of cut systems $\alpha^1, \alpha^2, \ldots, \alpha^n$ on Σ such that

(1) $\alpha^1 = \alpha$ and $\alpha^n = \alpha'$.
(2) α^{i+1} is obtained from α^i by sliding exactly one component of α^i over another component of α^i.

We call a cut system α on the boundary ∂H_g of handlebody H_g a *standard cut system* with respect to H_g, if the curves $\alpha_1, \alpha_2, \ldots, \alpha_g$ bounds properly embedded g disjoint discs in H_g. In Figure 2, the curves $\gamma_1, \gamma_2, \ldots, \gamma_k$ on the boundary of a handlebody H_k form a standard cut system on ∂H_k with respect to H_k. Now, we have the following easy proposition.

Proposition 2.5. *Let H_g be a non-orientable genus g handlebody. Let α and β be two cut systems on ∂H_g which are handle slide equivalent. If α is a standard cut system on ∂H_g with respect to H_g, then β is also a standard cut system on ∂H_g with respect to H_g.*

Theorem 2.6 (Johannson, [4])**.** *Let H_g be a non-orientable genus g handlebody. Let α and β be two standard cut systems of ∂H_g. Then α and β are handle slide equivalent.*

At the end of this section, we recall the equivalence of Heegaard splittings.

Definition 2.7. We say that two genus g Heegaard splittings $H_g \cup H'_g$ and $\mathcal{H}_g \cup \mathcal{H}'_g$ of a closed non-orientable 3–manifold M are equivalent if there exists a diffeomorphism $f : M \to M$ such that $f(H_g) = \mathcal{H}_g$ and $f(H'_g) = \mathcal{H}'_g$.

3. Main Theorem

In this section, we prove our main theorem of the article. Recall that the standard genus k Heegaard splitting of $\#_k S^1\widetilde{\times}S^2 = H_k \cup_{Id} H'_k$ is the double of a genus k handlebody $H_k = \natural_k S^1\widetilde{\times}D^2$, where H'_k is a copy of H_k and $Id : \partial H_k \to \partial H'_k$ is the identity map, see Figure 2. In order to prove the theorem, we require the following results:

Theorem 3.1 (Miller and Naylor, [5])**.** *For $g \geq k \geq 0$, any genus g Heegaard splitting of $M = \#_k S^1\widetilde{\times}S^2$ is equivalent to a Heegaard splitting of M*

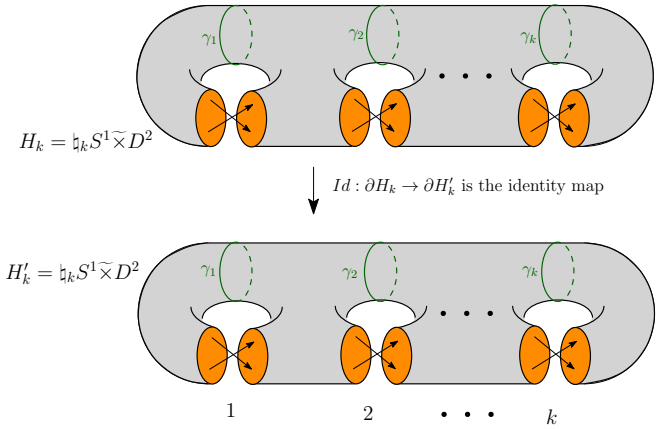

FIGURE 2. Standard genus k Heegaard splitting of $\natural_k S^1 \times S^2$.

obtained by stabilizing the standard genus k Heegaard splitting of M $g-k$ times. In particular, any two genus k Heegaard splittings of $\#_k S^1 \widetilde{\times} S^2$ are equivalent.

Theorem 3.2 (Epstein, [2])**.** *Let a and b be two properly embedded arcs in a compact surface Σ with the same end points such that they are homotopic relative to the boundary. Then they are ambient isotopic by an isotopy which is fixed on $\partial\Sigma$.*

Let (Σ, ϕ) be an abstract open book of a closed non-orientable 3–manifold $M = M_{(\Sigma,\phi)}$ with the first Betti number $b_1(\Sigma)$ of Σ equal to k. Let $\mathcal{H}_k = \Sigma \times [0,\pi], \mathcal{H}'_k = \Sigma \times [\pi, 2\pi]$,

$$\widetilde{\mathcal{H}_k} = \frac{\Sigma \times [0,\pi]}{(x,t) \sim (x,t') \text{ for } x \in \partial\Sigma, \text{ and } t,t' \in [0,\pi]} \text{ and}$$

$$\widetilde{\mathcal{H}'_k} = \frac{\Sigma \times [\pi,2\pi]}{(x,t) \sim (x,t') \text{ for } x \in \partial\Sigma, \text{ and } t,t' \in [\pi,2\pi]}$$

be genus k handlebodies with the natural quotient maps $q : \mathcal{H}_k \to \widetilde{\mathcal{H}_k}$ given by $q(x,t) = [(x,t)]$ and $q' : \mathcal{H}'_k \to \widetilde{\mathcal{H}'_k}$ given by $q'(x,t) = [(x,t)]$. Next, we observe that the Heegaard splitting of M associated with the abstract open book (Σ, ϕ) is:

$$M = \widetilde{\mathcal{H}'_k} \bigcup_f \widetilde{\mathcal{H}_k}, \text{ where}$$

$f : \partial\widetilde{\mathcal{H}'_k} = q'(\Sigma \times \{\pi, 2\pi\}) \to \partial\widetilde{\mathcal{H}_k} = q(\Sigma \times \{0,\pi\})$ such that $f : q'(\Sigma \times \{\pi\}) \to q(\Sigma \times \{\pi\})$ is the identity map and $f : q'(\Sigma \times \{2\pi\}) \to q(\Sigma \times \{0\})$

is the map ϕ. Note that the Heegaard surface of the Heegaard splitting $M = \widetilde{\mathcal{H}'_k} \bigcup_f \widetilde{\mathcal{H}_k}$ is $S = q(\Sigma \times \{0, \pi\}) = \partial\widetilde{\mathcal{H}_k}$.

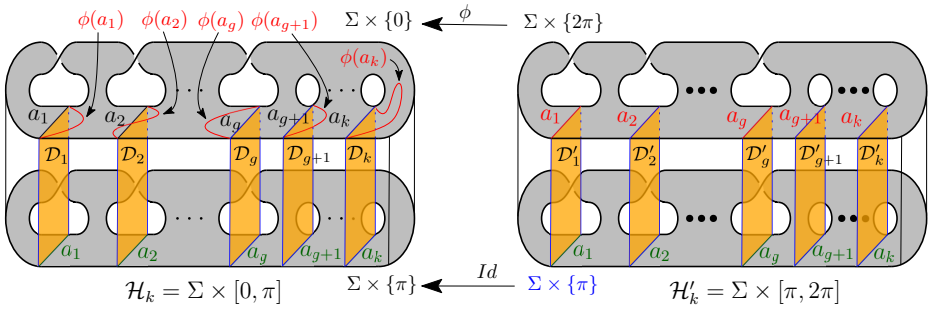

FIGURE 3. The figure depicts the two non-orientable genus k handlebodies $\mathcal{H}_k = \Sigma \times [0, \pi]$ and $\mathcal{H}'_k = \Sigma \times [\pi, 2\pi]$. The collection $\{a_1, a_2, \ldots, a_k\}$ depicts a basic arc system of Σ. The red arcs $\phi(a_i)$'s on $\Sigma \times \{0\} \subset \partial\mathcal{H}_k$ depicts the images of the arcs a_i's $\subset \Sigma \times \{2\pi\} \subset \partial\mathcal{H}'_k$ by the map ϕ. The discs $\mathcal{D}_i = a_i \times [0, \pi]$, $1 \leq i \leq k$ and $\mathcal{D}'_i = a_i \times [\pi, 2\pi]$, $1 \leq i \leq k$ is a collection of k disjoint discs in $\mathcal{H}_k$ and $\mathcal{H}'_k$, respectively.

Now, we see that, in particular, if the monodromy ϕ is isotopic to the identity map, the diffeomorphism $f : S' = \partial\widetilde{\mathcal{H}'_k} \to S = \partial\widetilde{\mathcal{H}_k}$ is isotopic to the identity map. Without loss of generality, we can assume that f is the identity map. Therefore, M is diffeomorphic to the double of the non-orientable handlebody $\mathcal{H}_k \cong \natural_k S^1\widetilde{\times}D^2$. Hence, M is diffeomorphic to $\#_k S^1\widetilde{\times}S^2$.

Now, we are ready to prove our main theorem.

Theorem 3.3. *Let (Σ, ϕ) be an abstract open book of $M = \#_k S^1\widetilde{\times}S^2$. Then, the monodromy ϕ is isotopic to the identity map of Σ if and only if $k = b_1(\Sigma)$.*

Proof. It is clear from the above discussion that if ϕ is isotopic to the identity map, then $k = b_1(\Sigma)$.

Now, let us prove the converse. Let $\mathcal{H}_k, \mathcal{H}'_k, \widetilde{\mathcal{H}_k}$ and $\widetilde{\mathcal{H}'_k}$ be the genus $k = b_1(\Sigma)$ handlebodies with the natural quotient maps $q : \mathcal{H}_k \to \widetilde{\mathcal{H}_k}$ and $q' : \mathcal{H}'_k \to \widetilde{\mathcal{H}'_g}$ as described above associated with the open book (Σ, ϕ). Therefore, the Heegaard splitting of $M = \#_k S^1\widetilde{\times}S^2$ associated with the abstract open book (Σ, ϕ) is:

$$M = \widetilde{\mathcal{H}'_k} \bigcup_f \widetilde{\mathcal{H}_k}, \text{ where}$$

$f : S' = \partial\widetilde{\mathcal{H}'_k} = q'(\Sigma \times \{\pi, 2\pi\}) \to S = \partial\widetilde{\mathcal{H}_k} = q(\Sigma \times \{0, \pi\})$ such that $f : q'(\Sigma \times \{\pi\}) \to q(\Sigma \times \{\pi\})$ is the identity map and $f : q'(\Sigma \times \{2\pi\}) \to q(\Sigma \times \{0\})$ is the map ϕ.

Recall that a *basic arc system* of a compact surface is a collection $\mathcal{A} = \{a_1, ..., a_k\}$ of disjoint properly embedded arcs in Σ such that $\mathcal{A}$ cuts Σ into a 2–disc. Let $\{a_1, a_2, \ldots, a_k\}$ be a basic arc system of Σ. We know that the diffeomorphism ϕ –up to isotopy– is determined by the basic arc system $\{\phi(a_1), \phi(a_2), \ldots, \phi(a_k)\}$ up to isotopy. Therefore, to achieve our goal, it is enough to show that $\phi(a_i)$ is isotopic to a_i for each i.

Note that the collections $\alpha = \{\alpha_1, \alpha_2, \ldots, \alpha_k\}$ and $\alpha' = \{\alpha'_1, \alpha'_2, \ldots, \alpha'_k\}$, where $\alpha_i = q(a_i \times \{0, \pi\})$ and $\alpha'_i = q(a_i \times \{\pi, 2\pi\})$, are cut systems of the Heegaard surfaces S and S', respectively, as $\{a_1, a_2, \ldots, a_k\}$ is a basic arc system of Σ. Since $f : S' \to S$ is diffeomorphism, the collection $\beta = \{\beta_1, \beta_2, \ldots, \beta_k\}$, where $\beta_i = f(\alpha'_i) = q(a_i \times \{0\} \cup \phi(a_i) \times \{\pi\})$, is a cut system of the Heegaard surface S. Consider $\mathcal{D}_i = a_i \times [0, \pi]$, $1 \leq i \leq k$ and $\mathcal{D}'_i = a_i \times [\pi, 2\pi]$, $1 \leq i \leq k$. Then $\{\mathcal{D}_i\}$ and $\{\mathcal{D}'_i\}$ are the collections of k disjoint discs in $\mathcal{H}_k$ and $\mathcal{H}'_k$, respectively. Observe that the curves α_i's bound disjoint discs $\widetilde{\mathcal{D}_i} = q(\mathcal{D}_i)$'s in $\widetilde{\mathcal{H}_k}$ and the curves β_i's bound disjoint discs $\widetilde{\mathcal{D}'_i} = q'(\mathcal{D}'_i)$'s in $\widetilde{\mathcal{H}'_k}$. For a schematic of this, refer Figure 3.

Let $M = \#_k S^1 \widetilde{\times} S^2 = H_k \cup_{Id} H'_k$ be the standard genus k Heegaard splitting of M, see Figure 2. Now, by Theorem 3.1, any two genera k Heegaard splittings of $\#_k S^1 \widetilde{\times} S^2$ are equivalent. Therefore, there is a diffeomorphism

$$h : \#_k S^1 \widetilde{\times} S^2 = \widetilde{\mathcal{H}'_k} \bigcup_{f} \widetilde{\mathcal{H}_k} \to \#_k S^1 \widetilde{\times} S^2 = H_k \bigcup_{Id} H'_k.$$

such that $h(\widetilde{\mathcal{H}_k}) = H_k$ and $h(\widetilde{\mathcal{H}'_k}) = H'_k$. Let $\{\eta_1, \eta_2, \ldots, \eta_k\}$ be a cut system of ∂H_k such that η_i's bound disjoint discs D_i's in H_k. Since $M = H_k \cup_{Id} H'_k$ is obtained by doubling H_k, the curves η_i's also bound disjoint discs D'_i's in H'_k. Therefore, the curves $\gamma_i = h^{-1}(\eta_i)$'s bound properly embedded disjoint discs $\bar{D}_i$'s in $\widetilde{\mathcal{H}_k}$ as well as bound disjoint discs $\bar{D}'_i$'s in $\widetilde{\mathcal{H}'_k}$. Thus, the cut systems α and $\gamma = \{\gamma_1, \gamma_2, \ldots, \gamma_k\}$ of $\partial\widetilde{\mathcal{H}_k}$ are standard with respect to $\widetilde{\mathcal{H}_k}$. Also, the cut system γ and β of $\partial\widetilde{\mathcal{H}'_k}$ are standard with respect to $\widetilde{\mathcal{H}'_k}$. Therefore, by Theorem 2.6, the cut systems α and γ are handle slide equivalent. Similarly, the cut systems γ and β are handle slide equivalent. Hence, the cut systems α and β are handle slide equivalent. Thus, by Proposition 2.5, as α_i's bound properly embedded disjoint discs $\mathcal{D}_i$'s in $\widetilde{\mathcal{H}_k}$,

the curves β_i's bound properly embedded disjoint discs $\widetilde{\mathcal{D}_i}$'s in $\widetilde{\mathcal{H}_k}$. Since $\beta_i = q(a_i \times \{0\} \cup \phi(a_i) \times \{\pi\})$, $\phi(a_i)$ is homotopic to a_i. Now it follows from Theorem 3.2 that $\phi(a_i)$ is isotopic to a_i for each i. Therefore, ϕ is isotopic to the identity map. This completes the proof. □

Acknowledgement: Abhijeet Ghanwat acknowledges United States-India Educational Foundation(USIEF), New Delhi and U.S. Department of State for the Fulbright Nehru Postdoctoral Research Fellowship, Award No. 3058 FNPDR /2024. Suhas Pandit is supported by MATRICS grant [MTR/ 2023/ 000438] of Science and Engineering Research Board Government of India. Sanjay Kute is supported by NBHM Ph.D Scholarship [Order Number: 0203/ 18/ 2022/ R&D-II/ 9396] India.

References

[1] Etnyre, J., Lectures on open book decompositions and contact structures, *Lecture notes for Clay Summer School at the Alfred Renyi Institute of Mathematics (Hungarian Academy of Sciences) Budapest, Hungary* (2004).

[2] Epstein, D., Curves on 2-manifolds and isotopies, *Acta Math.* **115**, (1966) 83-107.

[3] Ghiggini, P. and Lisca, P., Open book decompositions versus prime factorizations of closed, oriented 3–manifolds, *Geometry and Topology Monographs* **19**, (2015) 145–155.

[4] Johannson, K., Topology and combinatorics of 3-manifolds, *Lecture Notes in Mathematics 1599, Springer, Berlin* (1995)

[5] Miller, M. and Naylor, P., Trisections of non-orientable 4-manifolds, *Michigan Math. J.* **74** 403–447, (2023).

Abhijeet Ghanwat
Department of Mathematics, University of Georgia
Athens, GA 30602, United States
E-mail: ghanwata16@gmail.com

Suhas Pandit and Sanjay Kute
Indian Institute of Technology Madras,
IIT PO. Chennai- 600036, Tamil Nadu, India.
E-mail: suhas@iitm.ac.in, sanjaykutemjt@gmail.com.

The Mathematics Student ISSN: 0025-5742
Vol. 95, Nos. 3-4, July - December (2026), 28–37

EVALUATION OF THE CONVOLUTION SUMS $W_8(N)$ AND $W_{16}(N)$

VIJAYASHANKAR A. I. AND VASUKI K. R.

(Received : 02 - 02 - 2025 ; Accepted : 22 - 03 - 2026)

Abstract. In this paper, we evaluate the convolution sums $\sum_{l+8m=n} \sigma(l)\sigma(m)$ and $\sum_{l+16m=n} \sigma(l)\sigma(m)$ for all positive integers n, using Ramanujan's theta function identities and related Eisenstein series identities. Also we determine the partition functions for the terms in the convolution sums which are not expressible as divisor functions.

1. Introduction

For any positive integer k, the sum of divisors function is defined by

$$\sigma_k(n) = \begin{cases} \sum_{d|n} d^k, & \text{if } n \text{ is a positive integer} \\ 0, & \text{otherwise,} \end{cases}$$

where d runs through the positive divisors of n and set $\sigma_1(n) = \sigma(n)$. Let $i,\ j$ be any two positive integers with $i \leq j$. The convolution sum $W_{i,j}$ is defined by

$$W_{i,j}(n) := \sum_{il+jm=n} \sigma(l)\sigma(m)$$

and

$$W_{1,j}(n) := W_j(n).$$

K. S. Williams [29] evaluated the following convolution sum:

$$W_8(n) = \left(\frac{1}{24} - \frac{n}{32}\right)\sigma(n) + \left(\frac{1}{24} - \frac{n}{4}\right)\sigma\left(\frac{n}{8}\right) + \frac{1}{192}\sigma_3(n) + \frac{1}{64}\sigma_3\left(\frac{n}{2}\right) + \frac{1}{16}\sigma_3\left(\frac{n}{4}\right) + \frac{1}{3}\sigma_3\left(\frac{n}{8}\right) + \frac{3}{10}\sigma_3\left(\frac{n}{6}\right) + c_8(n), \quad (1.1)$$

2020 Mathematics Subject Classification: 11A25, 11E20, 11M36, 11F20
Key words and phrases: Sum of divisors function, convolution sums, Eisenstein series, theta functions

© *Indian Mathematical Society, 2026.*

K. S. Williams [29] evaluated the following convolution sum:

$$W_8(n) = \left(\frac{1}{24} - \frac{n}{32}\right)\sigma(n) + \left(\frac{1}{24} - \frac{n}{4}\right)\sigma\left(\frac{n}{8}\right) + \frac{1}{192}\sigma_3(n) + \frac{1}{64}\sigma_3\left(\frac{n}{2}\right) + \frac{1}{16}\sigma_3\left(\frac{n}{4}\right) + \frac{1}{3}\sigma_3\left(\frac{n}{8}\right) + \frac{3}{10}\sigma_3\left(\frac{n}{6}\right) + c_8(n), \quad (1.1)$$

where $\sum_{n=1}^{\infty} c_8(n)q^n = q\prod_{n=0}^{\infty}(1-q^{2n})^4(1-q^{4n})^4$ and A. Alaca, . Alaca and K. S. Williams [4] deduce that:

$$W_{16}(n) = \left(\frac{1}{24} - \frac{n}{64}\right)\sigma(n) + \left(\frac{1}{24} - \frac{n}{4}\right)\sigma\left(\frac{n}{16}\right) + \frac{1}{768}\sigma_3(n) + \frac{1}{256}\sigma_3\left(\frac{n}{2}\right) + \frac{1}{64}\sigma_3\left(\frac{n}{4}\right) + \frac{1}{16}\sigma_3\left(\frac{n}{8}\right) + \frac{1}{3}\sigma_3\left(\frac{n}{16}\right) + \frac{3}{10}\sigma_3\left(\frac{n}{6}\right) + c_{16}(n), \quad (1.2)$$

where

$$\begin{aligned}
&\sum_{n=1}^{\infty} c_{16}(n)q^n \\
&= \frac{1}{32}\prod_{n=1}^{\infty}(1+q^n)^{20}(1-q^n)^8(1-q^{4n-2})^{14} + \frac{3}{112}\prod_{n=1}^{\infty}(1+q^n)^{16}(1-q^n)^8(1-q^{4n-2})^{12} \\
&+ \frac{1}{224}\prod_{n=1}^{\infty}(1+q^n)^{12}(1-q^n)^8(1-q^{4n-2})^{10} - \frac{1}{32}\prod_{n=1}^{\infty}(1+q^n)^4(1-q^n)^8(1-q^{4n-2})^6 \\
&- \frac{3}{112}\prod_{n=1}^{\infty}(1-q^n)^8(1-q^{4n-2})^4 - \frac{1}{224}\prod_{n=1}^{\infty}(1+q^n)^{-4}(1-q^n)^8(1-q^{4n-2})^2.
\end{aligned}$$

They have used these convolutions sums to determine the number of representations of a positive integer n by the form

$$x_1^2 + x_2^2 + x_3^2 + x_4^2 + 2(x_5^2 + x_6^2 + x_7^2 + x_8^2)$$

and

$$x_1^2 + x_2^2 + x_3^2 + x_4^2 + 4(x_5^2 + x_6^2 + x_7^2 + x_8^2)$$

respectively, where $x_1,\ x_2,\ x_3,\ x_4,\ x_5,\ x_6,\ x_7$ and x_8 are positive integers.

In this paper, we give an alternative and simplified proof for the first convolution sum and derive a new formula for the latter convolution sum. These are evaluated by using Ramanujan's theta function identities and related Eisenstein series identities. Furthermore, we introduce two partition functions $p_8(n)$ and $p_{16}(n)$ for the convolution sums $W_8(n)$ and $W_{16}(n)$ respectively, in order to replace $c_8(n)$ and $c_{16}(n)$ and to write the formula

for $W_8(n)$ and $W_{16}(n)$ completely in terms of arithmetical functions. The following table gives the information regarding the convolution sums for some values of i and j:

Sl. No.	(i, j)	Authors
1.	(1,1).	M. Besge [9], J. W. L. Glaisher [14], S. Ramanujan [25].
2.	(1,2),(1,3),(1,4).	J. G. Huard et. al [16].
3.	(1,5),(1,7)	M. Lemire and K. S. Williams [17] and S. Cooper and P. C. Toh [13].
4.	(1,6),(2,3).	. Alaca and K. S. Williams [6].
5.	(1,8),(1,9).	K. S. Williams [29, 28].
6.	(1,10),(1,11),(1,13), (1,14).	E. Royer [27].
7.	(2,5),(4,5),(1,20).	S. Cooper and D. Ye [12].
8.	(1,12),(3,4),(1,16), (1,18),(2,9).	A. Alaca, . Alaca and K. S. Williams [1, 4, 2].
9.	(1,24),(3,8),(1,12), (3,4),(1,15),(3,5).	B. Ramakrishnan and B. Sahu[22, 21].
10.	(1,7),(2,7),(1,14), (3,7),(1,21),(4,7),(1,28).	B. Ramakrishnan, B. Sahu and A. K. Singh [23].
11.	(1,23)	H. H. Chan and S. Cooper [11].
12.	(1,25)	E. X. W. Xia, X. L. Tian and O. X. M. Yao [30].
13.	(1,27),(1,32)	. Alaca and Y. Kesicioğlu [5].
14.	(1,36),(4,9)	D. Ye [31].

In Section 2, we recall some preliminary definitions and results, and also construct some identities which are required to prove our main results and in Section 3, we prove these main results. Then we establish a relationship between some partition functions and the terms of the convolution sums which are not expressible in terms of divisor functions.

2. Preliminary results

Ramanujan's theta functions plays a key role in proving our results. This section consists of some preliminary definitions and identities which are used in the sequel. Let a, b be any two complex numbers with $|ab| < 1$.

The Ramanujan's general theta function is defined by

$$f(a,b) := \sum_{n=-\infty}^{\infty} a^{\frac{n(n+1)}{2}} b^{\frac{n(n-1)}{2}} = (-a;ab)_\infty (-b;ab)_\infty (ab;ab)_\infty,$$

where $(a;b)_\infty := \prod_{n=0}^{\infty}(1-ab^n)$. For $|q|<1$, Ramanujan [7, p. 36] considers following three special cases of $f(a,b)$:

$$\varphi(q) := f(q,q) = 1 + 2\sum_{n=1}^{\infty} q^{n^2} = \frac{(-q;q^2)_\infty (q^2;q^2)_\infty}{(q;q^2)_\infty (-q^2;q^2)_\infty},$$

$$\psi(q) := f(q,q^3) = \sum_{n=0}^{\infty} q^{\frac{n(n+1)}{2}} = \frac{(q^2;q^2)_\infty}{(q;q^2)_\infty},$$

and

$$f(-q) := f(-q,-q^2) = \sum_{n=0}^{\infty}(-1)^n q^{\frac{n(3n-1)}{2}} + \sum_{n=1}^{\infty}(-1)^n q^{\frac{n(3n+1)}{2}} = (q;q)_\infty.$$

If q is any complex number such that $|q|<1$, the Ramanujan Eisenstein series are defined by

$$P(q) := 1 - 24\sum_{n=1}^{\infty}\frac{nq^n}{1-q^n} \tag{2.1}$$

and

$$Q(q) := 1 + 240\sum_{n=1}^{\infty}\frac{n^3q^n}{1-q^n}. \tag{2.2}$$

For any positive integer k, set $P_k = P(q^k)$, $Q_k = Q(q^k)$ and $f_k = f(-q^k)$. The above Eisenstein series can be written in a form involving divisor functions as follows:

$$P_k = 1 - 24\sum_{n=1}^{\infty}\sigma\left(\frac{n}{k}\right)q^n \tag{2.3}$$

and

$$Q_k = 1 + 240\sum_{n=1}^{\infty}\sigma_3\left(\frac{n}{k}\right)q^n. \tag{2.4}$$

In his second notebook [24, p. 198], Ramanujan recorded the following theta function identities:

$$\varphi(q) + \varphi(-q) = 2\varphi(q^4), \tag{2.5}$$

$$\varphi(q)\varphi(-q) = \varphi(-q^2), \tag{2.6}$$

$$\varphi(q)\psi(q^2) = \psi^2(q), \tag{2.7}$$

$$\varphi^2(q) + \varphi^2(-q) = 2\varphi^2(q^2), \tag{2.8}$$

$$\varphi^2(q) - \varphi^2(-q) = 8q\psi(q^4) \tag{2.9}$$

and

$$\varphi^4(q) - \varphi^4(-q) = 16q\psi(q^2). \tag{2.10}$$

S. Cooper [10, p. 207, p. 486] deduced the following Eisenstein series identities that connects with Ramanujan's theta functions:

$$\varphi^8(-q) = \frac{1}{15}\left[-Q_1 + 16Q_2\right], \tag{2.11}$$

$$\psi^8(q) = \frac{1}{240}\left[Q_1 - Q_2\right], \tag{2.12}$$

$$-P_1 + P_2 - 2P_4 + 8P_8 = 6\varphi^2(q)\varphi^2(q^2) \tag{2.13}$$

and

$$-P_1 + 7P_2 - 14P_4 + 8P_8 = 24q\varphi^2(-q)\psi^2(q^4). \tag{2.14}$$

Now we deduce the following theta function identities by performing simple arithmetic operations on (2.5) to (2.10):

$$4\varphi^4(q)\varphi^4(q^2) = -\varphi^8(-q) - 3\varphi^8(-q^2) + 2\varphi^8(-q^4) + 288q^2\psi^8(q^2) - 512q^4\psi^8(q^4) + 16q(f_2f_4)^4, \tag{2.15}$$

$$16q^2\varphi^4(-q)\psi^4(q^4) = -4q\psi^8(q) + 48q^2\psi^8(q^2) + 256q^4\psi^8(q^4) + 4q(f_2f_4)^4, \tag{2.16}$$

$$\varphi^8(q) = \varphi^8(-q^2) + 16q\psi^8(q) \tag{2.17}$$

and

$$\psi^8(-q) = \psi^8(q) + 16q\psi^8(q^2). \tag{2.18}$$

Lemma 2.1. *If k is any positive integer, then*

$$P_k^2 = 1 + \sum_{n=1}^{\infty}\left(240\sigma_3\left(\frac{n}{k}\right) - \frac{288n}{k}\sigma\left(\frac{n}{k}\right)\right)q^n. \tag{2.19}$$

Proof. See proof [14, 15]. □

Lemma 2.2. *We have*

$$-P_1 + 8P_8 = 7\varphi^2(q)\varphi^2(q^2) - 4q\varphi^2(-q)\psi^2(q^4) \tag{2.20}$$

and

$$-P_1 + 16P_{16} = 3\varphi^4(q) + 12\varphi^4(q^4). \tag{2.21}$$

Proof. Proof of the first identity is easily obtained from (2.13) and (2.14). Using the identity $-P_1 + 4P_4 = 3\varphi^4(q)$, we deduce (2.21). □

3. Convolution sums

Let $p_{de}(n)$ denote the number of partitions of n into even number of parts such that the parts being even and distinct(including colored parts), with parts of the form $4k+2$ have four colors and parts of the form $4k$ have eight colors, and $p_{do}(n)$ denotes the number of partitions of n into odd number of parts such that the parts being even and distinct(including colored parts), with parts of the form $4k+2$ have four colors and parts of the form $4k$ have eight colors, where k is any positive integer. Then,

$$\sum_{n=1}^{\infty} (p_{de}(n) - p_{do}(n))\, q^n = (f_2 f_4)^4.$$

Let $p^*_{de}(n)$ be the number of partitions of n such that each part being even and distinct(including colored parts), parts of the form $8k \pm 2$ have four colors, parts of the form $8k+4$ have twelve colors, parts of the form $8k$ have eight colors and the number of parts of the form $8k$ and $8k+4$ being even. Similarly let $p^*_{do}(n)$ be the number of partitions of n such that each part being even and distinct(including colored parts), parts of the form $8k \pm 2$ have four colors, parts of the form $8k+4$ have twelve colors, parts of the form $8k$ are distinct and have eight colors and the number of parts of the form $8k$ and $8k+4$ being odd, where k is any positive integer. Then

$$\sum_{n=1}^{\infty} (p^*_{de}(n) - p^*_{do}(n))\, q^n = \frac{f_4^{16}}{f_2^4 f_8^4}.$$

Now define the partition functions $p_8(n)$ and $p_{16}(n)$ as follows:

$$p_8(n) = \frac{1}{64}\left(p_{de}(n-1) - p_{do}(n-1)\right) \tag{3.1}$$

and

$$p_{16}(n) = \frac{3}{128}\left(p_{de}(n-1) - p_{do}(n-1)\right) + \frac{3}{64}\left(p_{de}\left(\frac{n-2}{2}\right) - p_{do}\left(\frac{n-2}{2}\right)\right) + \frac{1}{32}\left(p^*_{de}(n-1) - p^*_{do}(n-1)\right), \tag{3.2}$$

where $p_{de}(n) = p_{do}(n) = p^*_{de}(n) = p^*_{do}(n) = 0$, if n is not a positive integer.

Theorem 3.1. *If n is any positive integer. Then*

$$W_8(n) = \left(\frac{1}{24} - \frac{n}{32}\right)\sigma(n) + \left(\frac{1}{24} - \frac{n}{4}\right)\sigma\left(\frac{n}{8}\right) + \frac{1}{192}\sigma_3(n) + \frac{1}{64}\sigma_3\left(\frac{n}{2}\right) + \frac{1}{16}\sigma_3\left(\frac{n}{4}\right) + \frac{1}{3}\sigma_3\left(\frac{n}{8}\right) + \frac{3}{10}\sigma_3\left(\frac{n}{6}\right) + p_8(n) \quad (3.3)$$

and

$$W_{16}(n) = \left(\frac{1}{24} - \frac{n}{64}\right)\sigma(n) + \left(\frac{1}{24} - \frac{n}{4}\right)\sigma\left(\frac{n}{16}\right) - \frac{1}{384}\sigma_3(n) + \frac{5}{128}\sigma_3\left(\frac{n}{2}\right) - \frac{1}{64}\sigma_3\left(\frac{n}{4}\right) + \frac{1}{16}\sigma_3\left(\frac{n}{8}\right) + \frac{1}{3}\sigma_3\left(\frac{n}{16}\right) + \frac{3}{10}\sigma_3\left(\frac{n}{6}\right) - p_{16}(n). \quad (3.4)$$

Proof of (3.3). From (2.20), we have

$$-P_1 + 8P_8 = 7\varphi^2(q)\varphi^2(q^2) - 4q\varphi^2(-q)\psi^2(q^4).$$

By squaring the above, we find that

$$P_1P_8 = \frac{1}{16}\left[P_1^2 + 64P_8^2 - 16q^2\varphi^4(-q)\psi^4(q^4) - 49\varphi^4(q)\varphi^4(q^2) + 56q(f_2f_4)^4\right], \quad (3.5)$$

where, we have used the fact that $\varphi(q) = \frac{f_2^5}{f_1f_4}$, $\varphi(-q) = \frac{f_1^2}{f_2}$ and $\psi(q) = \frac{f_2^2}{f_1}$. Substituting (2.3), (2.15) and (2.16) into the above , we obtain that

$$\sum_{n=1}^{\infty} W_8(n)q^n = -\frac{1}{576} + \frac{1}{24}\sum_{n=1}^{\infty}\frac{nq^n}{1-q^n} + \frac{1}{24}\sum_{n=1}^{\infty}\frac{nq^{8n}}{1-q^{8n}} + \frac{1}{9216}P_1^2 + \frac{1}{144}P_8^2 + \frac{49}{36864}\varphi^8(-q) - \frac{147}{36864}\varphi^8(-q^2) - \frac{49}{18432}\varphi^8(-q^4) + \frac{1}{2304}q\psi^8(q) - \frac{447}{1152}q^2\psi^8(q^2) + \frac{376}{576}q^4\psi^8(q^4) - \frac{9}{576}q(f_2f_4)^4. \quad (3.6)$$

Now using (2.19), (2.11), (2.12), (2.4) and (3.1), one can deduce the proof of (3.3). □

Proof of (3.4). Here also by squaring the equation (2.21), we find that

$$P_1P_{16} = \frac{1}{32}\left[P_1^2 + 256P_{16}^2 - 9q^2\varphi^8(q) - 144\varphi^8(q^4)\varphi^4(q^2) - 72\varphi^4(q)\varphi^4(q^4)\right]. \quad (3.7)$$

From (2.8), (2.9) and (2.5), one can deduce the following identities:

$$\varphi^4(q) = 2\varphi^4(q^2) - \varphi^4(-q^2) + 8q\psi^4(q^2); \quad (3.8)$$

$$16\varphi^4(q^4) = \varphi^4(q) + \varphi^4(-q) + 6\varphi^4(-q^2) + 8\varphi^4(-q^4). \quad (3.9)$$

By multiplying $4\varphi^4(q^4)$ on both sides of (3.8), and then by substituting (3.9) into the third summand and using (2.15) for the first and second summands by changing $q \to q^2$ and $q \to -q^2$ respectively:

$$4\varphi^4(q)\varphi^4(q^4) = -\varphi^8(-q^2)+3\varphi^8(-q^4)+2\varphi^8(-q^8)+8q\psi^8(q)-32q^2\psi^8(q^2) + 32q^4\psi^8(q^4) - 512q^8\psi^8(q^8) + 24q(f_2f_4)^4 + 48q^2(f_4f_8)^4 + 32q\frac{f_4^{16}}{f_2^4f_8^4}. \quad (3.10)$$

Now using (2.11), (2.12), (2.17), (2.19), (3.10), (3.7) and (3.2), we arrive at (3.4). □

Acknowledgement: The authors would like to thank anonymous referee for his/her valuable comments on this paper. The second author is supported by grant Ref. No.:191620127010/ (CSIR-UGC NET DEC.2019) by the funding agency University Grants Commission, Government of India under Joint CSIR-UGC JRF scheme. The author is grateful to funding agency.

References

[1] Alaca, A., Alaca, Ş and Williams, K. S., Evaluation of the Convolution sums $\sum_{l+12m=n}\sigma(l)\sigma(m)$ and $\sum_{3l+4m=n}\sigma(l)\sigma(m)$, *Adv. Theor. Appl. Math.,* **1**(1) (2006), 27-48.

[2] Alaca, A., Alaca, Ş and Williams, K. S., Evaluation of the Convolution sums $\sum_{l+18m=n}\sigma(l)\sigma(m)$ and $\sum_{2l+9m=n}\sigma(l)\sigma(m)$, *Int. Math. Forum,* **2** (2007), 45-68.

[3] Alaca, A., Alaca, Ş and Williams, K. S., Evaluation of the Convolution sums $\sum_{l+24m=n}\sigma(l)\sigma(m)$ and $\sum_{3l+8m=n}\sigma(l)\sigma(m)$, *Math. J. Okayama Univ.,* **49** (2007), 93-111.

[4] Alaca, A., Alaca, Ş and Williams, K. S., Evaluation of the Convolution sums $\sum_{m<n/16}\sigma(m)\sigma(n-16m)$, *Canad. Math. Bull.* **51**(1) (2008), 3-14.

[5] Alaca, Ş. and Kesicioğlu, Y., Evaluation of the Convolution sums $\sum_{l+27m=n}\sigma(l)\sigma(m)$ and $\sum_{l+32m=n}\sigma(l)\sigma(m)$, *Int. J. Number Theory,* **12**(1) (2016), 1-13.

[6] Alaca, Ş and Williams, K. S., Evaluation of the Convolution sums $\sum_{l+6m=n}\sigma(l)\sigma(m)$ and $\sum_{2l+3m=n}\sigma(l)\sigma(m)$, *J. Number Theory,* **124** (2007), 491-510.

[7] Berndt, B. C., Ramanujan's Notebooks: Part III, Springer New York, 1991.

[8] Berndt, B. C., Chan, H. H., Son, J. and Son, S. H., Eisenstein series in Ramanujan's lost notebook, *Trans. Amer. Math. Soc.,* **4**(1) (2000), 81-114.

[9] Besge, M., Extrait d'une lettre de M Besge á M Liouville., *J Math Pure Appl.,* **7** (1862), 256.

[10] Cooper, S., Ramanujan's theta functions, Springer, Cham, 2017.

[11] Chan H. H., and Cooper, S., Powers of theta functions, *Pacific J. Math.,* **235**(1) (2008), 1-14.

[12] Cooper, S. and Ye, D., Evaluation of the Convolution sums $\sum_{l+20m=n}\sigma(l)\sigma(m)$, $\sum_{4l+5m=n}\sigma(l)\sigma(m)$ and $\sum_{2l+5m=n}\sigma(l)\sigma(m)$, *J. Number Theory,* **124**(2007), 491-510.

[13] Cooper, S., and Toh, P. C., Quintic and septic Eisenstein series, *Ramanujan J.,* **19**(2) (2009), 163-181.

[14] Glaisher, J. W. L., On the square of the series in which the coefficients are the sum of the divisors of the exponents, *Mess. Math.,* **14** (1885), 156-163.

[15] Glaisher, J. W. L., Mathematical papers, 1883-1885, W. Metcalfe and son, Cambridge, 1885.

[16] Huard, J. G., Ou, Z. M., Spearman, B. K. and Williams, K. S., Elementary evaluation of certain convolution sums involving divisor functions, in *Number Theory for the Millennium, II(Urbana, IL, 2000)(A. K. Peters, Natick, MA, 2002),* 229-274.

[17] Lemire, M., and Williams, K. S., Evaluation of two convolution sums involving the sum of divisors function, *Bull. Austral. Math. Soc.,* **73**(1) (2006), 107-115.

[18] Lomadze, G. A., Representation of numbers by sums of the quadratic forms $x_1^2 + x_1x_2 + x_2^2$, *Acta Arith.,* **54**(1) (1989), 9-36.

[19] Ntienjem, E., Evaluation of convolution sums involving the sum of divisors function for levels 48 and 64, *Integers,* **17** (2017), 22.

[20] Ntienjem, E., Evaluation of convolution sums involving the sum of divisors function for 22, 44 and 52, *Open Math.,* **15**(1) (2017), 446-458.

[21] Ramakrishnan, B. and Sahu, B., Evaluation of the Convolution sums $\sum_{l+15m=n}\sigma(l)\sigma(m)$ and $\sum_{3l+5m=n}\sigma(l)\sigma(m)$, *Int. J. Number Theory,* **9**(3) (2013), 799-809.

[22] Ramakrishnan, B. and Sahu, B., Evaluation of convolution sums and some remarks on cusp forms of weight 4 and level 12, *Math. J. Okayama Univ.,* **59** (2017), 71-79.

[23] Ramakrishnan, B., Sahu, B., and Singh, A. K., On the number of representations of certain quadratic forms in 8 variables, in *Automorphic Forms and Related Topics, Samuele Anni et al. (eds.),* 215-224, *Contemp. Math.*, **732**, Amer. Math. Soc. Providence, RI, 2019.

[24] Ramanujan, S., Notebooks (volume 2), Tata Institute of Fundamental Research Bombay, 1957.

[25] Ramanujan, S., On certain arithmetical functions, *Trans. Cambridge Philos. Soc.*, **22** (1916), 159-184.

[26] Ramanujan, S.,The Lost Notebook and Other Unpublished Papers, Narosa, New Delhi, 1988.

[27] Royer, E., Evaluating convolution sums of the divisor function by quasimodular forms, *Int. J. Number Theory,* **3**(2) (2007), 231-261.

[28] Williams, K. S., Evaluation of the Convolution sums $\sum_{m<n/9}\sigma(m)\sigma(n-9m)$, *Int. J. Number Theory,* **1**(2) (2005), 193-205.

[29] Williams, K. S., Evaluation of the Convolution sums $\sum_{m<n/8}\sigma(m)\sigma(n-8m)$, *Pacific J. Math.,* **228**(2) (2006), 387-396.

[30] Xia, E. X. W., Tian, X. L. and Yao, O. X. M., Evaluation of the convolution sum $\sum_{i+25j=n}\sigma(i)\sigma(j)$, *Int. J. Number Theory,* **10**(6) (2014), 1421-1430.

[31] Ye, D., Evaluation of the convolution sum $\sum_{l+36m=n}\sigma(l)\sigma(m)$ and $\sum_{4l+9m=n}\sigma(l)\sigma(m)$, *Int. J. Number Theory,* **11**(1) (2015), 171-183.

Vijayashankar A. I.
Department of Studies in Mathematics
University of Mysore,
Mysuru, Karnataka, India.
E-mail: vijayshankarai3@gmail.com

Vasuki K. R.
Department of Studies in Mathematics
Senior Professor
University of Mysore,
Mysuru, Karnataka, India.
E-mail: vasuki_kr@hotmail.com

The Mathematics Student ISSN: 0025-5742
Vol. 95, Nos. 3-4, July - December (2026), 38–41

A QUADRATIC RESIDUE IN SOME FINITE FIELDS

SURINDER KAUR

(Received : 05 - 04 - 2025 ; Accepted: 18 - 06 - 2026)

Abstract. Let p be a prime with $3 \mid (p-1)$, and let F be a finite field of characteristic p. By analyzing the units of order 2 in the group ring of the cyclic group of order 3 over F, we prove that -3 is a quadratic residue in F.

1. Introduction

Let a be a non-zero element in a finite field F of characteristic p. If the equation $x^2 = a$ has a solution in F, then a is called a quadratic residue; otherwise, it is called a quadratic non-residue. It is well known that F has $\frac{|F|-1}{2}$ quadratic residues and $\frac{|F|-1}{2}$ quadratic non-residues. Determining quadratic residues in finite fields is a classical problem in number theory and finite field theory.

In this article, we show that -3 is a quadratic residue in any finite field of characteristic p, where $3 \mid p-1$. Although this fact can also be obtained from quadratic reciprocity, the aim here is to provide a proof using units of the group ring of the cyclic group of order 3 over F.

2. Preliminaries

Let G be a finite group. Then the group ring FG consists of all formal sums $\sum_{g\in G} \alpha_g g,\ \alpha_g \in F$. In particular, if $G = C_n$ is the cyclic group of order n and $(n,p) = 1$, then $FG \cong F[x]/\langle x^n - 1\rangle$, where the polynomial $(x^n - 1)$ is separable over F. Hence $(x^n - 1) = \prod_{i=1}^{r} f_i(x)$, where the polynomials $f_i(x)$ are pairwise co-prime irreducible factors of $(x^n - 1)$ in $F[x]$. Therefore, by

2020 Mathematics Subject Classification: 12E20, 16U60, 20C05
Key words and phrases: finite fields group rings unit group

© *Indian Mathematical Society, 2026.*

the Chinese Remainder Theorem,

$$FG \cong \bigoplus_{i=1}^{r} F[x]/\langle f_i(x)\rangle.$$

We recall that the map $\varepsilon : FG \to F$ defined by $\varepsilon\left(\sum_{g\in G} \alpha_g g\right) = \sum_{g\in G} \alpha_g$ is called the augmentation map. We denote by $\mathcal{U}(FG)$ the group of all units of the group ring FG. The subgroup

$$V(FG) = \{u \in \mathcal{U}(FG) \mid \varepsilon(u) = 1\}$$

consisting of all units of augmentation 1 is called the normalized unit group. Moreover, there is a canonical isomorphism

$$\mathcal{U}(FG) \cong F^\times \times V(FG),$$

where $F^\times = F \setminus \{0\}$ denotes the multiplicative group of the field F.

3. A quadratic residue in finite fields

Let F be a finite field of characteristic p such that $3 \mid (|F| - 1)$. Since $F^\times$ is cyclic, we write $F^\times = \langle\zeta\rangle$. Because $3 \mid (|F| - 1)$, the element

$$\eta = \zeta^{\frac{|F|-1}{3}}$$

is a primitive cube root of unity in F. Hence,

$$x^3 - 1 = (x - 1)(x - \eta)(x - \eta^2) \quad \text{in } F[x].$$

Therefore,

$$FC_3 \cong \frac{F[x]}{\langle x^3 - 1\rangle} \cong F \oplus F \oplus F.$$

Consequently,

$$V(FC_3) \cong F^\times \times F^\times.$$

Under this decomposition, an element of $V(FC_3)$ has order dividing 2 if and only if each coordinate has order dividing 2 in $F^\times$. Since $F^\times$ is cyclic, it contains a unique element of order 2, namely -1. Hence the elements of order 2 in $V(FC_3)$ correspond to

$$(1, -1, -1), \quad (1, -1, 1), \quad (1, 1, -1).$$

Thus, $V(FC_3)$ contains exactly three elements of order 2. Now we are ready to prove the main result:

Theorem 3.1. *Let F be a finite field of characteristic p, where $3 \mid p-1$. Then $-\frac{1}{3}$, and hence -3, is a quadratic residue in F.*

Proof. Let $C_3 = \langle b \rangle$. Assume that the element

$$v = 1 + \alpha(b-1) + \beta(b^2-1) \in V(FC_3),$$

where $\alpha, \beta \in F$, has order 2. Then $v^2 = 1$ gives

$$\alpha^2 + \beta^2 - 2\alpha - 2\beta + 4\alpha\beta = 0, \tag{3.1}$$

$$2\alpha^2 - \beta^2 - 2\alpha + 2\alpha\beta = 0. \tag{3.2}$$

Subtracting (3.2) from (3.1), we obtain

$$-\alpha^2 + 2\beta^2 - 2\beta + 2\alpha\beta = 0. \tag{3.3}$$

Now subtracting (3.3) from (3.2), one has

$$(\alpha - \beta)(3\alpha + 3\beta - 2) = 0.$$

Thus either

$$\alpha = \beta \qquad \text{or} \qquad \alpha + \beta = \frac{2}{3}.$$

Case 1: $\alpha = \beta$.

Substituting into (3.1), $6\alpha^2 - 4\alpha = 0$, which gives

$$\alpha = 0 \quad \text{or} \quad \alpha = \frac{2}{3}.$$

The solution $\alpha = \beta = 0$ corresponds to the trivial unit 1, while $\alpha = \beta = \frac{2}{3}$ gives the non-trivial unit

$$v = 1 + \frac{2}{3}(b + b^2 - 2).$$

Case 2: $\alpha \neq \beta$.

Then

$$\alpha + \beta = \frac{2}{3}. \tag{3.4}$$

Substituting this into (3.1),

$$\alpha^2 + \beta^2 + 4\alpha\beta - \frac{4}{3} = 0.$$

It implies that

$$(\alpha + \beta)^2 + 2\alpha\beta - \frac{4}{3} = \frac{4}{9} + 2\alpha\beta - \frac{4}{3} = 0,$$

and hence

$$\alpha\beta = \frac{4}{9}.$$

Using (3.1), we get

$$\alpha\Big(\frac{2}{3} - \alpha\Big) = \frac{4}{9}$$

and so

$$9\alpha^2 - 6\alpha - 4 = 0.$$

Solving, we get that $\alpha = \frac{1}{3} \pm \sqrt{-\frac{1}{3}}$ and since $\beta = \frac{2}{3} - \alpha$, the other two elements of order 2 are given by

$$1 + \Big(\frac{1}{3} + \sqrt{-\frac{1}{3}}\Big)(b-1) + \Big(\frac{1}{3} - \sqrt{-\frac{1}{3}}\Big)(b^2-1)$$

and

$$1 + \Big(\frac{1}{3} - \sqrt{-\frac{1}{3}}\Big)(b-1) + \Big(\frac{1}{3} + \sqrt{-\frac{1}{3}}\Big)(b^2-1).$$

It follows that $-\frac{1}{3}$ is a quadratic residue in F and since $-3 = \left(-\frac{1}{3}\right)^{-1}$, -3 is also a quadratic residue in F.

□

Acknowledgement: The author gratefully acknowledges the financial support of the National Board for Higher Mathematics (NBHM), Department of Atomic Energy, Government of India (Project grant Number: 02011/1/2025/ R & D II/4177).

References

[1] Milies, C.P., Sehgal, S.K., An Introduction to Group Rings, Kluwer Academic Publishers, Dordrecht, 2002.

[2] Wright, S., Quadratic Residues and Non-Residues, Lecture Notes in Mathematics 2171, Springer, Cham, 2016.

Surinder Kaur
Department of Mathematics,
SRM University-AP,
Mangalagiri Mandal, Guntur-522240,
Andhra Pradesh,
INDIA.
E-mail: skaur.math@gmail.com

The Mathematics Student ISSN: 0025-5742
Vol. 95, Nos. 3-4, July - December (2026), 42–57

HOOK FUSION PROCEDURE FOR DIRECT PRODUCT OF SYMMETRIC GROUPS

DIMPI K. M. AND GEETHA THANGAVELU

(Received : 02 - 04 - 2025 ; Accepted : 04 - 10 - 2026)

Abstract. In this work, we formulate a hook fusion procedure for the direct product $S_m \times S_n$ by applying Grime's construction to the constituent symmetric groups S_m and S_n. For each standard bi-tableau, we define a rational function with values in $\mathbb{C}[S_m \times S_n]$ whose restriction to the subspace determined by the principal hooks is regular at the simultaneous equal-parameter specialization. Its value yields the corresponding diagonal matrix element and, after normalization, a primitive idempotent. This provides a complete set of pairwise orthogonal primitive idempotents of $\mathbb{C}[S_m \times S_n]$. This procedure extends naturally to finite direct products of symmetric groups.

1. Introduction

The *Yang–Baxter equation* is a fundamental relation in mathematical physics and representation theory [13]. It appears in the study of quantum groups, integrable systems, and parameter-dependent representations of braid-type relations. Let V be a finite-dimensional complex vector space and let

$$R(u) \in \mathrm{End}(V \otimes V)$$

be a rational function of a complex parameter u. The operator $R(u)$ is said to satisfy the *Yang–Baxter equation* if

$$R_{12}(u)R_{13}(u+v)R_{23}(v) = R_{23}(v)R_{13}(u+v)R_{12}(u),$$

where $R_{ij}(u)$ acts as $R(u)$ on the i-th and j-th tensor components of $V^{\otimes 3}$ and as the identity on the remaining component. An important rational

2020 Mathematics Subject Classification: 16G10, Secondary 20C30, 05E10, 16D40
Key words and phrases: Yang–Baxter equation, Fusion procedure, Direct product of symmetric groups, Young tableau, Jucys–Murphy elements

© *Indian Mathematical Society, 2026.*

solution of the Yang–Baxter equation is the Yang R-matrix

$$R(u) = \mathrm{Id}_{V\otimes V} - \frac{P}{u},$$

where P is the permutation operator on $V \otimes V$, defined by

$$P(x \otimes y) = y \otimes x$$

for $x, y \in V$. For $1 \le i < j \le N$, let P_{ij} denote the operator that permutes the i-th and j-th tensor factors of $V^{\otimes N}$. Given complex parameters $u_1, \ldots, u_N$, one considers the ordered product

$$\mathcal{F}(u_1, \ldots, u_N) = \prod_{1\le i<j\le N}^{\longrightarrow} \left(\mathrm{Id}_{V^{\otimes N}} - \frac{P_{ij}}{u_i - u_j}\right),$$

where the pairs (i, j) are taken in lexicographic order. Products of this form are usually called *fusion functions*.

The relation between this fusion function and the representation theory of the symmetric group comes from the natural action of S_N on $V^{\otimes N}$. Under this action, a permutation acts by permuting the tensor factors, and the transposition $(i\,j)$ acts as the operator P_{ij}. This action constitutes one side of Schur–Weyl duality: the actions of S_N and the general linear group on $V^{\otimes N}$ commute. Consequently, the above operator-valued fusion function is the image of the following $\mathbb{C}[S_N]$-valued rational function:

$$\Phi(u_1, \ldots, u_N) = \prod_{1\le i<j\le N}^{\longrightarrow} \left(1 - \frac{(i\,j)}{u_i - u_j}\right) \in \mathbb{C}[S_N].$$

Thus, the Yang–Baxter equation gives rise to parameter-dependent elements of the symmetric group algebra through the permutation action on tensor space.

The fusion procedure for the symmetric group is concerned with suitable evaluations of this rational function. Let T be a *standard Young tableau* and let $c_i(T)$ denote the *content* of the node occupied by i. Direct substitution $u_i = c_i(T)$ into the rational function may lead to singular factors. The classical fusion procedure resolves these singularities by prescribing a suitable order of evaluation and produces the diagonal matrix elements associated with the Young basis. After normalization, these elements form a complete set of pairwise orthogonal primitive idempotents of $\mathbb{C}[S_N]$. This

construction originates in the work of Jucys [12] and was developed further by Cherednik [2] and Nazarov [16, 17, 18, 19]. Fusion procedures have subsequently been obtained for several other algebras, including Hecke algebras, Brauer algebras, and Birman–Murakami–Wenzl algebras [8, 9, 10]. Molev gave an alternative derivation of the fusion procedure for S_N from the Jucys–Murphy formulas for diagonal matrix units [14]; see also [1, 15, 20] for the Jucys–Murphy and Gelfand–Tsetlin approach to the representation theory of symmetric groups.

Grime introduced the hook fusion procedure for the symmetric group [5]. Instead of evaluating all auxiliary parameters independently, the parameters corresponding to entries lying in the same principal hook are identified. More precisely, one substitutes u_i by $z_i + c_i(T)$, restricts the resulting rational function to a subspace $\mathcal{H}_T$ determined by the principal hooks of T, and then specializes the remaining variables to a common value. Grime proved that the restricted function is regular at this specialization and that its value coincides with the diagonal matrix element corresponding to T. This procedure reduces the number of independent auxiliary parameters and highlights the role of the hook structure of the tableau. Grime later adapted the hook fusion construction to Hecke algebras [6].

There is also a representation-theoretic interpretation of the hook fusion procedure. Let $\nu_1, \ldots, \nu_d$ be the shapes of the principal hooks of $\lambda \vdash N$, and let $h_a = |\nu_a|$. Then

$$\mathrm{Ind}_{S_{h_1} \times \cdots \times S_{h_d}}^{S_N} (V_{\nu_1} \boxtimes \cdots \boxtimes V_{\nu_d}) \cong \bigoplus_{\mu \vdash N} V_\mu^{\oplus D_{\mu\lambda}},$$

where $D_{\mu\lambda}$ denotes the multiplicity of V_μ in the induced representation. In particular, $D_{\lambda\lambda} = 1$, so V_λ occurs with multiplicity one; see [5, Section 1]. For generic values of the principal-hook parameters, the left ideal generated by the specialized rational function affords the induced representation above, whereas the equal-parameter specialization yields the diagonal matrix element F_T, and hence $\mathbb{C}[S_N]F_T \cong V_\lambda$. Thus, this specialization singles out the distinguished irreducible component V_λ [5, Section 1]. This interpretation shows that hook fusion is not merely a regularization procedure, but rather connects Yang–Baxter-type rational functions, induction of representations, and the combinatorics of Young tableaux, and therefore suggests

analogous constructions for other groups and algebras admitting tableau-type bases.

In this paper, we apply the preceding ideas to the direct product $S_m \times S_n$. The natural isomorphism

$$\mathbb{C}[S_m \times S_n] \cong \mathbb{C}[S_m] \otimes \mathbb{C}[S_n]$$

shows that its representation theory is determined by those of S_m and S_n. At the level of tensor spaces, let V and W be finite-dimensional complex vector spaces. Then $S_m \times S_n$ acts naturally on $V^{\otimes m} \otimes W^{\otimes n}$, where S_m permutes the tensor factors in $V^{\otimes m}$ and S_n permutes the tensor factors in $W^{\otimes n}$. Consequently, the rational function associated with the direct product factors naturally into two independent rational functions, one belonging to $\mathbb{C}[S_m]$ and the other belonging to $\mathbb{C}[S_n]$. This factorization suggests that Grime's hook fusion procedure can be applied separately to the two components and subsequently combined.

The aim of this paper is to formulate this construction precisely and to show that it produces the diagonal matrix elements corresponding to the irreducible representations of $S_m \times S_n$. Although the tensor-product decomposition makes the factorization natural, the combined evaluation requires careful treatment of the indexing of the variables, the order of multiplication, and the hook restrictions associated with the two tableaux. We define a rational function with values in $\mathbb{C}[S_m \times S_n]$ together with a subspace determined by the principal hooks of both components. We then prove that the restriction of this function is regular at the simultaneous evaluation point and that its value coincides with the corresponding diagonal matrix element.

Thus, the paper provides a direct hook-fusion construction of the diagonal matrix elements, and hence of the corresponding primitive idempotents, for the direct product of two symmetric groups. The same argument extends naturally to arbitrary finite direct products of symmetric groups.

The paper is organized as follows. In Section 2, we recall the representation theory of $S_m \times S_n$, and the necessary facts concerning Young tableaux, Young orthogonal basis, and the Jucys–Murphy elements. In Section 3, we

define the rational function for the direct product and prove the main regularity and evaluation theorem. We also illustrate the procedure with an explicit example. Finally, in Section 4, we discuss the extension to arbitrary finite direct products of symmetric groups and present several directions for further study.

2. Representation theory of $S_m \times S_n$

We fix notation and recall the representation theory of the direct product $S_m \times S_n$.

Let $N = m + n$. We identify $S_m \times S_n$ with the subgroup of S_N that preserves the two subsets $\{1, \ldots, m\}$ and $\{m+1, \ldots, N\}$. Thus, the standard generators of $S_m \times S_n$ are $s_1, \ldots, s_{m-1}$ and $s_{m+1}, \ldots, s_{N-1}$, where $s_i = (i, i+1)$. Notice that s_m does not belong to $S_m \times S_n$, since it interchanges the two blocks. The group algebra of the direct product is naturally isomorphic to the tensor product of the two group algebras:

$$\mathbb{C}[S_m \times S_n] \cong \mathbb{C}[S_m] \otimes \mathbb{C}[S_n].$$

Under this isomorphism, an element $(g_1, g_2) \in S_m \times S_n$ corresponds to $g_1 \otimes g_2$.

We consider the following chain of subalgebras:

$$\begin{aligned} \mathbb{C} \cong \mathbb{C}[S_0 \times S_0] \subset \mathbb{C}[S_1 \times S_0] \subset \mathbb{C}[S_2 \times S_0] \subset \cdots \subset \mathbb{C}[S_m \times S_0] \\ \subset \mathbb{C}[S_m \times S_1] \subset \cdots \subset \mathbb{C}[S_m \times S_n], \end{aligned} \tag{2.1}$$

where S_0 denotes the trivial group. The branching rules associated with this chain are multiplicity-free and give rise to a Young basis indexed by suitable *standard bi-tableaux*; see [1, 20].

We recall the branching rules for $S_m \times S_n$, the Young orthogonal basis of irreducible representations and the action of the corresponding Jucys–Murphy elements. For general background on the representation theory of finite direct products and the realization of their irreducible representations as outer tensor products, we refer to [21, Chapter 3, Section 7] and [11, Chapter 19]. These constructions fix the basis with respect to which the *diagonal matrix elements* are defined. To formulate the branching rules and describe the indexing set of the Young orthogonal basis, we first introduce

some combinatorial notation concerning partitions, Young tableaux, and bi-tableaux.

A *partition* of a positive integer N, denoted by λ, is a sequence of positive integers $\lambda := (\lambda_1, \lambda_2, \ldots, \lambda_\ell)$ satisfying $\lambda_1 \geq \lambda_2 \geq \cdots \geq \lambda_\ell$ and $\lambda_1 + \lambda_2 + \cdots + \lambda_\ell = N$. We write $\lambda \vdash N$ to indicate that λ is a partition of N and set $|\lambda| = N$. The unique partition of 0 is the empty partition, denoted by $\varnothing$. This concept extends naturally to a *bi-partition* of N, which is an ordered pair $\boldsymbol{\lambda} = (\lambda^{(1)}, \lambda^{(2)})$ where $\lambda^{(1)}$ and $\lambda^{(2)}$ are partitions satisfying $|\lambda^{(1)}| + |\lambda^{(2)}| = N$. Let $\Lambda(N)$ denote the set of all bi-partitions of N. For the direct product $S_m \times S_n$, where $N = m + n$, we consider the subset

$$\Lambda_{m,n} = \{\boldsymbol{\lambda} = (\lambda^{(1)}, \lambda^{(2)}) \in \Lambda(N) \mid \lambda^{(1)} \vdash m, \lambda^{(2)} \vdash n\}.$$

Given a partition $\lambda = (\lambda_1, \ldots, \lambda_\ell) \vdash N$, the corresponding *Young diagram* consists of ℓ rows of boxes, where the i-th row contains λ_i left-justified boxes. A *Young tableau* of shape λ is obtained by filling the boxes of the Young diagram bijectively with the integers $1, \ldots, N$. A *standard tableau* is a Young tableau in which the entries increase from left to right along every row and from top to bottom along every column. Each box of the Young diagram is called a *node* and is identified by its coordinates $\alpha = (i, j)$, where i and j denote its row and column numbers, respectively. The *content* of the node α is defined by $c(\alpha) = j - i$. In the Young diagram of shape λ, the *hook* associated with a node $\alpha = (i, j)$ is the set $\{(i, j') \in \lambda \mid j' \geq j\} \cup \{(i', j) \in \lambda \mid i' \geq i\}$.

For a bi-partition $\boldsymbol{\lambda} = (\lambda^{(1)}, \lambda^{(2)}) \in \Lambda(N)$, the *Young diagram* of shape $\boldsymbol{\lambda}$ is the ordered pair of Young diagrams of shapes $\lambda^{(1)}$ and $\lambda^{(2)}$. A *bi-tableau* of shape $\boldsymbol{\lambda}$, which we also call a $\boldsymbol{\lambda}$-tableau, is an ordered pair $\boldsymbol{T} = (T_1, T_2)$ obtained by filling the boxes of the two Young diagrams bijectively with the integers $1, \ldots, N$. A node of the Young diagram of $\boldsymbol{\lambda}$ is denoted by $\boldsymbol{\alpha} = (\alpha, k)$, where $\alpha = (i, j)$ is a node of the Young diagram of $\lambda^{(k)}$ and $k \in \{1, 2\}$ indicates the component containing the node. The content of $\boldsymbol{\alpha}$ is defined by $c(\boldsymbol{\alpha}) := c(\alpha) = j - i$. For a bi-tableau $\boldsymbol{T}$ and $1 \leq r \leq N$, we denote by $c_r(\boldsymbol{T})$ the content of the node occupied by the entry r.

A $\boldsymbol{\lambda}$-tableau $\boldsymbol{T} = (T_1, T_2)$ is called *standard* if, in each component, the entries increase from left to right along every row and from top to bottom along every column. In general, the integers $1, \ldots, N$ may be distributed arbitrarily between the two components of a standard $\boldsymbol{\lambda}$-tableau. In the present paper, however, for $\boldsymbol{\lambda} = (\lambda^{(1)}, \lambda^{(2)}) \in \Lambda_{m,n}$, we consider only those standard $\boldsymbol{\lambda}$-tableaux for which T_1 contains precisely the integers $1, \ldots, m$ and T_2 contains precisely the integers $m+1, \ldots, N$. We denote the set of such tableaux by $\mathrm{Std}_{m,n}(\boldsymbol{\lambda})$. Throughout this paper, unless otherwise stated, the term standard $\boldsymbol{\lambda}$-tableau refers to an element of $\mathrm{Std}_{m,n}(\boldsymbol{\lambda})$.

Equivalently, a standard $\boldsymbol{\lambda}$-tableau may be identified with a sequence

$$\boldsymbol{\mu}_0 \longrightarrow \boldsymbol{\mu}_1 \longrightarrow \cdots \longrightarrow \boldsymbol{\mu}_N$$

of bi-partitions satisfying $\boldsymbol{\mu}_0 = (\varnothing, \varnothing)$ and $\boldsymbol{\mu}_N = \boldsymbol{\lambda}$. For each $t = 1, \ldots, N$, the bi-partition $\boldsymbol{\mu}_t$ is obtained from $\boldsymbol{\mu}_{t-1}$ by adding a single box to the first component when $1 \leq t \leq m$, and by adding a single box to the second component when $m+1 \leq t \leq N$. Thus, this sequence determines a path in the Bratteli diagram associated with the chain of group algebras (2.1). Conversely, every path from $(\varnothing, \varnothing)$ to $\boldsymbol{\lambda}$ satisfying these conditions determines a unique element of $\mathrm{Std}_{m,n}(\boldsymbol{\lambda})$.

Example 2.1. Consider the bi-partition $\boldsymbol{\lambda} = ((2,1),(2)) \in \Lambda_{3,2}$. The sequence

$$(\varnothing, \varnothing) \longrightarrow (\ydiagram{1}, \varnothing) \longrightarrow (\ydiagram{2}, \varnothing) \longrightarrow \left(\ydiagram{2,1}, \varnothing\right) \longrightarrow \left(\ydiagram{2,1}, \ydiagram{1}\right) \longrightarrow \left(\ydiagram{2,1}, \ydiagram{2}\right)$$

corresponds to the standard $\boldsymbol{\lambda}$-tableau

$$\boldsymbol{T} = \left(\begin{array}{|c|c|}\hline 1 & 2 \\ \hline 3 \\ \cline{1-1}\end{array}, \begin{array}{|c|c|}\hline 4 & 5 \\ \hline\end{array}\right)$$

The irreducible representations of $S_m \times S_n$ are indexed by the elements of $\Lambda_{m,n}$. For $\boldsymbol{\lambda} = (\lambda^{(1)}, \lambda^{(2)}) \in \Lambda_{m,n}$, we denote the corresponding irreducible representation by $V_{\boldsymbol{\lambda}} = V_{\lambda^{(1)}} \boxtimes V_{\lambda^{(2)}}$, where $V_{\lambda^{(1)}}$ and $V_{\lambda^{(2)}}$ are the irreducible representations of S_m and S_n, respectively, and $\boxtimes$ denotes the outer tensor product. As a vector space, $V_{\boldsymbol{\lambda}} = V_{\lambda^{(1)}} \otimes V_{\lambda^{(2)}}$. The branching rules for the chain (2.1) are multiplicity-free. To illustrate this, we consider the case of $\mathbb{C}[S_3 \times S_2]$; see Figure 1.

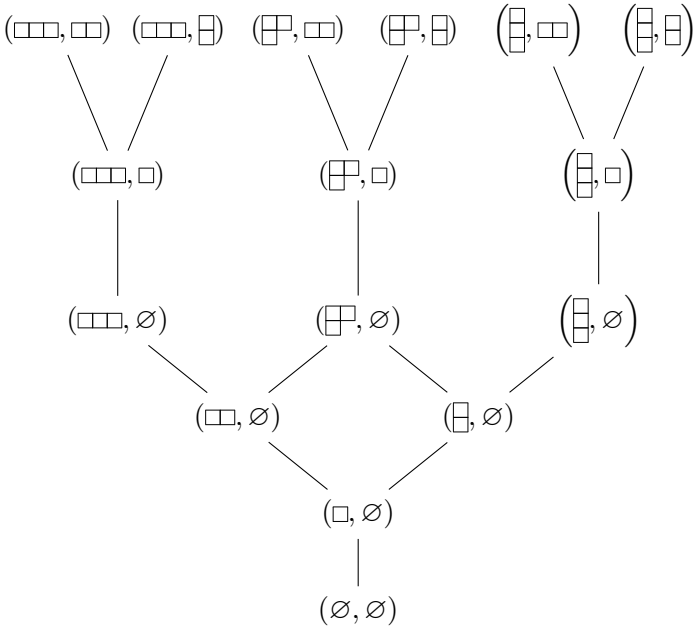

FIGURE 1. Bratteli diagram for $\mathbb{C}[S_3 \times S_2]$.

Since the branching rules associated with the chain (2.1) are multiplicity-free, every standard $\boldsymbol{\lambda}$-tableau $\boldsymbol{T} \in \mathrm{Std}_{m,n}(\boldsymbol{\lambda})$ determines a one-dimensional Gelfand–Tsetlin subspace $V_{\boldsymbol{T}} \subseteq V_{\boldsymbol{\lambda}}$. Choosing a unit vector $v_{\boldsymbol{T}}$ in each such subspace gives the Young orthogonal basis $\{v_{\boldsymbol{T}} \mid \boldsymbol{T} \in \mathrm{Std}_{m,n}(\boldsymbol{\lambda})\}$ of $V_{\boldsymbol{\lambda}}$. Let $\boldsymbol{T} = (T_1, T_2)$. Since $V_{\boldsymbol{\lambda}} = V_{\lambda^{(1)}} \otimes V_{\lambda^{(2)}}$, we may choose $v_{\boldsymbol{T}} = v_{T_1} \otimes v_{\widetilde{T}_2}$, where v_{T_1} is a vector in the Young orthogonal basis of $V_{\lambda^{(1)}}$, and $\widetilde{T}_2$ is the standard $\lambda^{(2)}$-tableau obtained from T_2 by replacing each entry $m+r$ by r, for $1 \leq r \leq n$. For simplicity, we write v_{T_2} in place of $v_{\widetilde{T}_2}$.

We next describe the action of the standard generators on the Young orthogonal basis. Let $\boldsymbol{T} \in \mathrm{Std}_{m,n}(\boldsymbol{\lambda})$, and let $i \in \{1, \ldots, m-1\} \cup \{m+1, \ldots, N-1\}$. Define $\boldsymbol{T}' = s_i\boldsymbol{T}$ to be the bi-tableau obtained from $\boldsymbol{T}$ by interchanging the entries i and $i+1$. Set $d_i(\boldsymbol{T}) = c_{i+1}(\boldsymbol{T}) - c_i(\boldsymbol{T})$. With respect to the Young orthogonal basis, the action of s_i is given by (2.2); see [1].

$$s_i v_{\boldsymbol{T}} = \begin{cases} \frac{1}{d_i(\boldsymbol{T})} v_{\boldsymbol{T}} + \sqrt{1 - \frac{1}{d_i(\boldsymbol{T})^2}}\, v_{\boldsymbol{T}'}, & \text{if } \boldsymbol{T}' \text{ is standard,} \\ \frac{1}{d_i(\boldsymbol{T})} v_{\boldsymbol{T}}, & \text{if } \boldsymbol{T}' \text{ is not standard.} \end{cases} \tag{2.2}$$

The group algebra $\mathbb{C}[S_m \times S_n]$ decomposes into a direct sum of matrix algebras, given by

$$\mathbb{C}[S_m \times S_n] \cong \bigoplus_{\boldsymbol{\lambda} \in \Lambda_{m,n}} \mathrm{End}_{\mathbb{C}}(V_{\boldsymbol{\lambda}}) \cong \bigoplus_{\boldsymbol{\lambda} \in \Lambda_{m,n}} \mathrm{Mat}_{d_{\boldsymbol{\lambda}}}(\mathbb{C}),$$

where $d_{\boldsymbol{\lambda}} = \dim(V_{\boldsymbol{\lambda}})$. The matrix units $E_{\boldsymbol{TT'}}$ in $\mathrm{Mat}_{d_{\boldsymbol{\lambda}}}(\mathbb{C})$ are indexed by pairs of standard $\boldsymbol{\lambda}$-tableaux $\boldsymbol{T}, \boldsymbol{T'} \in \mathrm{Std}_{m,n}(\boldsymbol{\lambda})$. An explicit isomorphism between these algebras is given by

$$E_{\boldsymbol{TT'}} = \frac{d_{\boldsymbol{\lambda}}}{|S_m \times S_n|} F_{\boldsymbol{TT'}},$$

where $F_{\boldsymbol{TT'}}$ represents the matrix coefficient associated with the vectors $v_{\boldsymbol{T}}$ and $v_{\boldsymbol{T'}}$ in the irreducible representation $V_{\boldsymbol{\lambda}}$, namely,

$$F_{\boldsymbol{TT'}} = \sum_{g \in S_m \times S_n} \langle v_{\boldsymbol{T}}, g v_{\boldsymbol{T'}} \rangle g \in \mathbb{C}[S_m \times S_n]. \tag{2.3}$$

We now restrict our attention to the diagonal matrix elements, and write $E_{\boldsymbol{T}} = E_{\boldsymbol{TT}}$ and $F_{\boldsymbol{T}} = F_{\boldsymbol{TT}}$.

For $g = (g_1, g_2) \in S_m \times S_n$, it can be observed that $\langle v_{\boldsymbol{T}}, g v_{\boldsymbol{T}} \rangle = \langle v_{T_1}, g_1 v_{T_1} \rangle \langle v_{T_2}, g_2 v_{T_2} \rangle$. It follows that

$$F_{\boldsymbol{T}} = \left(\sum_{g \in S_m} \langle v_{T_1}, g v_{T_1} \rangle \, g \right) \left(\sum_{g \in S_n} \langle v_{T_2}, g v_{T_2} \rangle \, g \right).$$

Thus, we have

$$F_{\boldsymbol{T}} = F_{T_1} F_{T_2}. \tag{2.4}$$

This factorization will be used in the proof of the hook fusion formula in the next section.

We now describe the Jucys–Murphy elements. In the group algebra $\mathbb{C}[S_m \times S_n]$, these elements are defined as follows:

$$X_1 = 0, \quad X_k = \sum_{i=1}^{k-1} (i\, k), \quad 2 \le k \le m,$$

$$X_{m+1} = 0, \quad X_l = \sum_{i=m+1}^{l-1} (i\, l), \quad m+2 \le l \le N.$$

The elements $X_1, \dots, X_N$ generate the maximal commutative subalgebra of $\mathbb{C}[S_m \times S_n]$ associated with the multiplicity-free chain (2.1), known as the *Gelfand–Tsetlin subalgebra*; see [1, 20]. It can be verified that for any $j \in \{1, \dots, N\}$, the element X_j commutes with all the elements of the subalgebra positioned at the $(j-1)$th level in the chain (2.1). This property plays a crucial role in understanding the action of the Jucys–Murphy elements on the Young orthogonal basis.

Proposition 2.2. *Let $\boldsymbol{\lambda} \in \Lambda_{m,n}$ and $\boldsymbol{T} \in \mathrm{Std}_{m,n}(\boldsymbol{\lambda})$. Then, for every $1 \leq j \leq N$, the Jucys–Murphy element X_j acts on the Young orthogonal basis vector $v_{\boldsymbol{T}}$ by*

$$X_j v_{\boldsymbol{T}} = c_j(\boldsymbol{T}) v_{\boldsymbol{T}}.$$

Proof. Write $\boldsymbol{T} = (T_1, T_2)$ and $v_{\boldsymbol{T}} = v_{T_1} \otimes v_{T_2}$. First suppose that $1 \leq j \leq m$. Then $X_j \in \mathbb{C}[S_m] \otimes 1$, and hence it acts only on the first tensor factor v_{T_1}. By the usual Jucys–Murphy eigenvalue formula for the symmetric group, X_j acts on v_{T_1} by the content of the node occupied by j; see, for example, [3, 15]; see also [1, 20]. Hence,

$$X_j v_{\boldsymbol{T}} = X_j(v_{T_1} \otimes v_{T_2}) = c_j(T_1) v_{T_1} \otimes v_{T_2} = c_j(\boldsymbol{T}) v_{\boldsymbol{T}}.$$

Now let $j = m+r$, where $1 \leq r \leq n$. Then $X_{m+r} \in 1 \otimes \mathbb{C}[S_n]$. Under the relabelling $m+q \mapsto q$ for $1 \leq q \leq r$, the element $X_{m+r} = \sum_{i=m+1}^{m+r-1} (i\ m+r)$ corresponds to the usual r-th Jucys–Murphy element $\sum_{q=1}^{r-1} (q\, r)$ of $\mathbb{C}[S_n]$. Therefore, X_{m+r} acts on v_{T_2} by the scalar equal to the content of the node occupied by $m+r$ in T_2. Thus,

$$X_{m+r} v_{\boldsymbol{T}} = c_{m+r}(\boldsymbol{T}) v_{\boldsymbol{T}}.$$

Therefore,

$$X_j v_{\boldsymbol{T}} = c_j(\boldsymbol{T}) v_{\boldsymbol{T}},$$

for $1 \leq j \leq N$. This completes the proof. □

3. Hook fusion procedure

Throughout this section, fix a bi-partition $\boldsymbol{\lambda} = (\lambda^{(1)}, \lambda^{(2)}) \in \Lambda_{m,n}$ and a standard $\boldsymbol{\lambda}$-tableau $\boldsymbol{T} = (T_1, T_2) \in \mathrm{Std}_{m,n}(\boldsymbol{\lambda})$. We set $c_i = c_i(\boldsymbol{T})$ for $1 \leq i \leq N$, where $N = m+n$. We define the subspace $\mathcal{H}_{\boldsymbol{T}}$ of $\mathbb{C}^N$ as

$$\mathcal{H}_{\boldsymbol{T}} = \left\{ (z_1, \dots, z_N) \in \mathbb{C}^N \mid z_i = z_j \text{ whenever } i \text{ and } j \text{ are in the same } \textit{principal hook} \text{ of } \boldsymbol{T} \right\}.$$

A *principal hook* is the hook associated with a diagonal node.

We now define a $\mathbb{C}[S_m \times S_n]$-valued rational function in the variables $z_1, z_2, \ldots, z_N$ by

$$\Phi_{\boldsymbol{T}}(z_1, \ldots, z_N) = \left(\overrightarrow{\prod_{1 \le i < j \le m}} \left(1 - \frac{(i\,j)}{z_i + c_i - (z_j + c_j)} \right) \right) \times \left(\overrightarrow{\prod_{m+1 \le i < j \le N}} \left(1 - \frac{(i\,j)}{z_i + c_i - (z_j + c_j)} \right) \right). \tag{3.1}$$

where $(i\,j)$ denotes the transposition interchanging i and j, and in both ordered products the factors are arranged according to the lexicographic order on the pairs (i, j).

Theorem 3.1. *When restricted to the subspace $\mathcal{H}_{\boldsymbol{T}}$, the rational function $\Phi_{\boldsymbol{T}}(z_1, \ldots, z_N)$ remains well-defined at $z_1 = \cdots = z_N$. Furthermore, the value of this restriction at $z_1 = \cdots = z_N$ coincides with the diagonal matrix element $F_{\boldsymbol{T}}$.*

Proof. Write

$$\Phi^1_{T_1}(z_1, \ldots, z_m) = \overrightarrow{\prod_{1 \le i < j \le m}} \left(1 - \frac{(i\,j)}{z_i + c_i - (z_j + c_j)} \right),$$

$$\Phi^2_{T_2}(z_{m+1}, \ldots, z_N) = \overrightarrow{\prod_{m+1 \le i < j \le N}} \left(1 - \frac{(i\,j)}{z_i + c_i - (z_j + c_j)} \right).$$

Then

$$\Phi_{\boldsymbol{T}}(z_1, \ldots, z_N) = \Phi^1_{T_1}(z_1, \ldots, z_m) \Phi^2_{T_2}(z_{m+1}, \ldots, z_N).$$

The first factor belongs to $\mathbb{C}[S_m] \otimes 1$, while the second belongs to $1 \otimes \mathbb{C}[S_n]$. Hence, the two factors commute. Let

$$\mathcal{H}_{T_1} = \left\{ (z_1, \ldots, z_m) \in \mathbb{C}^m \mid z_i = z_j \text{ whenever } i \text{ and } j \text{ lie in the same principal hook of } T_1 \right\}$$

and let

$$\mathcal{H}_{T_2} = \left\{ (z_{m+1}, \ldots, z_N) \in \mathbb{C}^n \mid z_i = z_j \text{ whenever } i \text{ and } j \text{ lie in the same principal hook of } T_2 \right\}.$$

Under the natural identification $\mathbb{C}^N \cong \mathbb{C}^m \times \mathbb{C}^n$, we have $\mathcal{H}_{\boldsymbol{T}} = \mathcal{H}_{T_1} \times \mathcal{H}_{T_2}$. By Grime's hook fusion procedure for the symmetric group [5], the

restriction of $\Phi^1_{T_1}(z_1, \ldots, z_m)$ to $\mathcal{H}_{T_1}$ is regular at $z_1 = \cdots = z_m$, and its value at this specialization is F_{T_1}. Let $\widetilde{T}_2$ be the standard $\lambda^{(2)}$-tableau obtained from T_2 by replacing each entry $m + r$ by r, for $1 \leq r \leq n$. Under the identification of S_n with the subgroup of S_N acting on the set $\{m+1, \ldots, N\}$, the function $\Phi^2_{T_2}(z_{m+1}, \ldots, z_N)$ is the shifted copy of the rational function associated with $\widetilde{T}_2$. Therefore, its restriction to $\mathcal{H}_{T_2}$ is regular at $z_{m+1} = \cdots = z_N$, and its value is F_{T_2}. It follows that the restriction of $\Phi_{\boldsymbol{T}}$ to $\mathcal{H}_{\boldsymbol{T}}$ is regular at $z_1 = \cdots = z_N$, and

$$\Phi_{\boldsymbol{T}}(z_1, \ldots, z_N)|_{\mathcal{H}_{\boldsymbol{T}},\, z_1 = \cdots = z_N} = F_{T_1} F_{T_2}.$$

By the factorization of diagonal matrix elements established in (2.4), we have $F_{\boldsymbol{T}} = F_{T_1} F_{T_2}$. Hence, $\Phi_{\boldsymbol{T}}(z_1, \ldots, z_N)|_{\mathcal{H}_{\boldsymbol{T}},\, z_1 = \cdots = z_N} = F_{\boldsymbol{T}}$. □

We now illustrate Theorem 3.1 using the standard bi-tableau introduced in Example 2.1.

Example 3.2. Let $m = 3$, $n = 2$, and consider

$$\boldsymbol{T} = \left(\begin{array}{|c|c|}\hline 1 & 2 \\ \hline 3 \\ \cline{1-1} \end{array}, \begin{array}{|c|c|}\hline 4 & 5 \\ \hline \end{array} \right)$$

of shape $\boldsymbol{\lambda} = ((2,1),(2))$. The contents of the entries are $c_1 = 0$, $c_2 = 1$, $c_3 = -1$, $c_4 = 0$, and $c_5 = 1$.

The entries $1, 2, 3$ lie in the same principal hook of T_1, while the entries $4, 5$ lie in the same principal hook of T_2. Therefore,

$$\mathcal{H}_{\boldsymbol{T}} = \left\{ (z_1, \ldots, z_5) \in \mathbb{C}^5 \;\middle|\; z_1 = z_2 = z_3, \quad z_4 = z_5 \right\}.$$

Using lexicographic order, the rational function associated with $\boldsymbol{T}$ is

$$\begin{aligned} \Phi_{\boldsymbol{T}}(z_1, \ldots, z_5) = & \left(1 - \frac{(1\,2)}{z_1 - z_2 - 1}\right) \left(1 - \frac{(1\,3)}{z_1 - z_3 + 1}\right) \\ & \times \left(1 - \frac{(2\,3)}{z_2 - z_3 + 2}\right) \left(1 - \frac{(4\,5)}{z_4 - z_5 - 1}\right). \end{aligned}$$

On $\mathcal{H}_{\boldsymbol{T}}$, write $z_1 = z_2 = z_3 = a$ and $z_4 = z_5 = b$. The restricted function is

$$\Phi_{\boldsymbol{T}}(z_1, \ldots, z_5)|_{\mathcal{H}_{\boldsymbol{T}}} = (1 + (1\,2))\,(1 - (1\,3)) \left(1 - \frac{1}{2}(2\,3)\right) (1 + (4\,5)).$$

This expression is independent of a and b. Hence, the restricted function is regular for all values of a and b, and in particular at $a = b$, which is precisely the specialization $z_1 = \cdots = z_5$. By Theorem 3.1, its value at this

specialization is the diagonal matrix element associated with $\boldsymbol{T}$. Thus,

$$F_{\boldsymbol{T}} = (1 + (1\,2))\,(1 - (1\,3))\left(1 - \frac{1}{2}(2\,3)\right)(1 + (4\,5))\,.$$

In particular, $F_{\boldsymbol{T}}$ is obtained directly by imposing the hook relations on the rational function, without computing the matrix coefficients $\langle v_{\boldsymbol{T}}, g v_{\boldsymbol{T}}\rangle$ for the elements $g \in S_3 \times S_2$.

Since $\dim V_{\boldsymbol{\lambda}} = \dim V_{(2,1)} \cdot \dim V_{(2)} = 2 \cdot 1 = 2$ and $|S_3 \times S_2| = 3!2! = 12$, the primitive idempotent associated with $\boldsymbol{T}$ is

$$E_{\boldsymbol{T}} = \frac{1}{6}\,(1 + (1\,2))\,(1 - (1\,3))\left(1 - \frac{1}{2}(2\,3)\right)(1 + (4\,5))\,.$$

Thus, the hook fusion procedure produces both the diagonal matrix element and the corresponding primitive idempotent by a direct evaluation of the single rational function associated with $\boldsymbol{T}$.

4. Conclusion and further directions

In this paper, we formulated a hook fusion procedure for the direct product $S_m \times S_n$. For each standard bi-tableau $\boldsymbol{T}$, we constructed a $\mathbb{C}[S_m \times S_n]$-rational function and proved that its restriction to the hook subspace $\mathcal{H}_{\boldsymbol{T}}$ is regular at the equal-parameter specialization. The value of this specialization is the diagonal matrix element $F_{\boldsymbol{T}}$, and hence, after normalization, multiplication by $\frac{d_{\boldsymbol{\lambda}}}{|S_m \times S_n|}$, it gives the primitive idempotent $E_{\boldsymbol{T}}$.

The construction extends naturally to an arbitrary finite direct product

$$S_{n_1} \times S_{n_2} \times \cdots \times S_{n_k}.$$

Indeed, if $N = n_1 + \cdots + n_k$, then the entries may be partitioned into k consecutive blocks of sizes $n_1, \cdots, n_k$, one block for each symmetric-group factor, and the corresponding rational function is obtained by the ordered product of the hook fusion functions associated with these blocks. The hook subspace is likewise the direct product of the hook subspaces associated with the component tableaux. Since the component rational functions lie in mutually commuting subalgebras and each component is regular under its corresponding hook specialization, their product is also regular. Its value is the product of the diagonal matrix elements associated with the component tableaux, which is precisely the diagonal matrix element for the

tensor-product representation. Thus, the same procedure yields the corresponding primitive idempotents for an arbitrary finite direct product of symmetric groups.

The hook fusion procedure has also been developed for the hyperoctahedral group in a recent preprint [4], and an analogous construction for the complex reflection groups $G(r, 1, N)$ has been obtained in unpublished work of the authors. Since these groups can be realized as the wreath products $C_2 \wr S_N$ and $C_r \wr S_N$, respectively, the available constructions suggest several natural questions for further study:

(1) Can the hook fusion procedures for the symmetric group, the hyperoctahedral group, and $G(r, 1, N)$ be placed in a uniform framework for the wreath product $G \wr S_N$, where G is an arbitrary finite group? Such a construction would require suitable analogues of the Jucys–Murphy elements, baxterized factors, and tableau parameters associated with the irreducible representations of G.
(2) Grime proposed a more general ribbon fusion framework for the symmetric group, in which the equalities among the auxiliary parameters are determined by suitable ribbon-type decompositions of a Young diagram [7, Chapter 5]. Can the corresponding regularity criterion be established, and can the resulting fusion procedure be extended componentwise to direct products and to the multipartition setting of $G(r, 1, N)$?
(3) Is there an analogue of the hook fusion procedure for diagram algebras, such as Brauer algebras, walled Brauer algebras, and Brauer algebras of type C? Since their natural bases are indexed by updown tableaux rather than ordinary Young tableaux, the principal-hook restriction does not carry over directly. The problem is therefore to determine an appropriate combinatorial constraint on the auxiliary parameters that simultaneously regularizes the fusion function and reduces the number of independent parameters.

Acknowledgement: The first author is supported by a PhD fellowship from the Indian Institute of Science Education and Research Thiruvananthapuram. The second author's research was partially supported by IISER Thiruvananthapuram and SERB-Power Grant SPG/2021/004200. This

work was also supported by an IGSTC WISER award granted to Geetha Thangavelu.

The authors are grateful to the editor and the referee for their careful reading of the manuscript and for their insightful comments and suggestions, which greatly improved the presentation of this work.

References

[1] Ceccherini-Silberstein, T., Scarabotti, F. and Tolli, F., Representation Theory of the Symmetric Groups: The Okounkov–Vershik Approach, Character Formulas, and Partition Algebras, Cambridge Studies in Advanced Mathematics, Cambridge University Press, 2010.

[2] Cherednik, I. V., Special bases of irreducible representations of a degenerate affine Hecke algebra, *Funktsional. Anal. i Prilozhen* **20** (1986), 87–88.

[3] Diaconis, P. and Greene, C., Applications of Murphy's elements, Stanford University Technical Report, No. 335, 1989.

[4] Dimpi, KM. and Thangavelu, G., Hook fusion procedure for hyper-octahedral groups, https://arxiv.org/pdf/2503.15981.pdf

[5] Grime, J., The hook fusion procedure, *Elect. J. Combin.* **12** (2005), R26.

[6] Grime, J., The hook fusion procedure for Hecke algebras, *J. Algebra* **309** (2007), 744–759.

[7] Grime, J., The Hook Fusion Procedure and Its Generalisations, https://arxiv.org/pdf/0708.1109.pdf

[8] Isaev, A. P., Molev, A. I. and Os'kin, A. F., On the idempotents of Hecke algebras, *Lett. Math. Phys.* **85** (2008), 79–90.

[9] Isaev, A. P. and Molev, A. I., Fusion procedure for the Brauer algebra, *Algebra i Analiz* **22** (2010), 142–154.

[10] Isaev, A. P., Molev, A. I. and Ogievetsky, O. V., Idempotents for Birman–Murakami–Wenzl algebras and reflection equation, *Adv. Theor. Math. Phys.* **18** (2014), 1–25.

[11] James, G. and Liebeck, M., Representations and Characters of Groups, 2nd ed., Cambridge University Press, 2001.

[12] Jucys, A. A., On the Young operators of the symmetric group, *Lietuvos Fizikos Rinkinys* **6** (1966), 163–180.

[13] Kulish, P. P., Reshetikhin, N. Yu. and Sklyanin, E. K., Yang–Baxter equation and representation theory: I, *Lett. Math. Phys.* **5** (1981), 393–403.

[14] Molev, A. I., On the fusion procedure for the symmetric group, *Rep. Math. Phys.* **61** (2008), 181–188.

[15] Murphy, G. E., A new construction of Young's seminormal representation of the symmetric groups, *J. Algebra* **69** (1981), 287–297.

[16] Nazarov, M., Young's symmetrizers for projective representations of the symmetric group, *Adv. Math.* **127** (1997), 190–257.

[17] Nazarov, M., Yangians and Capelli identities, Kirillov's Seminar on Representation Theory, *Amer. Math. Soc. Transl. Ser. 2* **181** (1998), 139–163.

[18] Nazarov, M., Representations of twisted Yangians associated with skew Young diagrams, *Selecta Math. (N.S.)* **10** (2004), 71–129.

[19] Nazarov, M. and Tarasov, V., On irreducibility of tensor products of Yangian modules associated with skew Young diagrams, *Duke Math. J.* **112** (2002), 343–378.

[20] Okounkov, A. and Vershik, A., A new approach to representation theory of symmetric groups, *Selecta Math. (N.S.)* **2** (1996), 581–605.

[21] Serre, J.-P., Linear Representations of Finite Groups, Springer-Verlag, 1977.

Dimpi, K. M., Geetha Thangavelu
Indian Institute of Science Education and Research
Thiruvananthapuram.
E-mail: dtyagi20@iisertvm.ac.in
E-mail: tgeetha@iisertvm.ac.in

The Mathematics Student ISSN: 0025-5742
Vol. 95, Nos. 3-4, July - December (2026), 58–68

SOME REMARKS ON THE METRIZABILITY OF TOPOLOGICAL GROUP VALUED CONE METRIC SPACES

AMAN DEEP AND RAKESH BATRA

(Received : 31 - 05 - 2025 ; Accepted : 22 - 06 - 2026)

Abstract. In this paper, we have given an example to show that topological group valued cone metric spaces need not be metrizable. We have also proven a sufficient condition for the metrizability of topological group valued cone metric spaces using a modification of classical arguments. Using our results, we have also proven that the hypothesis required to prove most of the fixed point theorems on recently introduced cone metric spaces over topological modules is also sufficient for the metrizability of these spaces.

1. Introduction

During the last decade, the study of cone metric spaces became very popular among fixed point theorists. Cone metric spaces are generalizations of metric spaces in which the metric d takes values in some partially ordered algebraic structure also having a compatible topology defined on it. Some of the most notable examples are cone metric spaces of Huang and Zhang [8], tvs-cone metric spaces of Beg *et al.* [2], complex valued metric spaces of Azam *et al.* [1], bicomplex valued metric spaces of Beg *et al.* [3] and cone metric spaces over Banach algebra of Liu and Xu [11]. Usually, the definition of such spaces and the definition of convergence on them follow a similar theme. All these spaces have been proven to be metrizable in the usual sense. For example, see [7], [9], [10].

Recently, Branga and Olaru [6] introduced the so-called cone metric spaces over topological modules. Branga also studied F-contraction mappings on such spaces and applied the results to integral equations (see [4]

2020 Mathematics Subject Classification: 54E35, 54D20, 54D10, 46A19
Keywords and phrases: metrizability, group valued cone metric space, cone metric spaces over topological modules

© *Indian Mathematical Society, 2026.*

and [5]). The main objective of this paper is to study the metrizability of cone metric spaces over topological modules and provide a sufficient condition for the same. In fact, we'll prove a metrization theorem in a much more general setting - when d takes values in a partially ordered topological group.

2. Preliminaries

Throughout the paper, $\mathbb{R}$, $\mathbb{R}_0^+$, $\mathbb{R}^+$ and $\mathbb{N}_+$ denote the sets of real numbers, nonnegative real numbers, positive real numbers and positive integers, respectively. Unless otherwise specified, the identity of a group G is denoted by 0. All the undefined terms related to topological spaces can be found in [12].

Definition 2.1. A group G is called a partially ordered group if there is a partial order $\leq$ defined on G such that for all $a_1, a_2, a_3, a_4 \in G$:

$$a_1 \leq a_2 \Rightarrow a_3 + a_1 + a_4 \leq a_3 + a_2 + a_4.$$

A ring R is called a partially ordered ring if there is a partial order $\preccurlyeq$ defined on R such that R as an additive group is a partially ordered group and for all $r_1, r_2 \in R$:

$$0 \preccurlyeq r_1, 0 \preccurlyeq r_2 \Rightarrow 0 \preccurlyeq r_1 r_2.$$

A left R-module G is called a partially ordered R-module if there are partial orders $\leq$ and $\preccurlyeq$ defined on G and R respectively, such that G becomes a partially ordered group, R becomes a partially ordered ring, and for all $r \in R, a \in G$:

$$0 \preccurlyeq r, \theta \leq a \Rightarrow \theta \leq ra$$

where θ is the identity of G and 0 is the identity of R.

If the identity of G is 0, then, we define, $G^+ = \{a \in G : 0 \leq a\}$, The set G^+ is called the cone of G. It is clear that $a_1 \leq a_2$ if and only if $a_2 - a_1 \in G^+$. The symbol $a_1 < a_2$ is used to denote that $a_2 - a_1 \in G^+ - \{0\}$. The symbol $a_1 \geq a_2$ is equivalent to $a_2 \leq a_1$. Now we recall the definitions of various topological algebraic structures.

Definition 2.2. A group G is called a topological group if it is nontrivial and there is a topology defined in G such that the mappings $G \times G \ni (a_1, a_2) :\to a_1 + a_2 \in G$ and $G \ni a :\to -a \in G$ are continuous; a ring R is called a topological ring if there is a topology defined on R such that its

additive group is a topological group and the mapping $R \times R \ni (r_1, r_2) :\to r_1 r_2 \in R$ is continuous; an R-module G is called a topological module if there are topologies defined on G are R such that G becomes a topological group, R becomes a topological ring and the mapping $R \times G \ni (r, a) :\to ra \in G$ is continuous, when all the concerned Cartesian products are equipped with the corresponding product topology.

Let G be a partially ordered topological group. We use the notation $a_1 \ll a_2$ if $a_2 - a_1 \in \text{Int}(G^+)$ and the notation $a_1 \not\ll a_2$ if $a_2 - a_1 \notin \text{Int}(G^+)$.

In [6] Branga and Olaru introduced the concept of group valued cone metric spaces over topological modules as follows:

Definition 2.3 ([6])**.** Let X be a nonempty set and G be a partially ordered topological left R-module such that G^+ has a nonempty interior and is closed. Suppose that a mapping $d : X \times X \to G$ satisfies:

(1) $d(x_1, x_2) \geq 0$ and $d(x_1, x_2) = \theta$ if and only if $x_1 = x_2$;
(2) $d(x_1, x_2) = d(x_2, x_1)$;
(3) $d(x_1, x_2) \leq d(x_1, x_3) + d(x_3, x_2)$.

Then, d is called a topological module valued cone metric on X and (X, G, d) is called a cone metric space over the topological left R-module G.

Branga and Olaru also introduced the notion of convergence of a sequence in a group valued cone metric space over a topological module.

Definition 2.4 ([6])**.** Let (X, G, d) be a cone metric space over a topological left R-module G. A sequence $\{x_i\}$ in X is said to converge to a point $x \in X$ if for all $\epsilon \gg 0$ there exists $i_0 \in \mathbb{N}_+$ such that $d(x_i, x) \ll \epsilon$ for all $i \geq i_0$.

We want to study the metrizability of such spaces in a slightly general setting. For convenience, we denote by $\mathcal{G}$ the class of all groups G such that

(1) G is Abelian;
(2) G is a partially ordered group with partial order $\leq$;
(3) G is a topological group;
(4) G^+ has a nonempty interior;
(5) G^+ is closed.

The following properties can be easily proved.

Proposition 2.5. *If $G \in \mathcal{G}$, then, for all $a_1, a_2, a_3, a_1', a_2' \in G$,*

(1) if $a_1 \leq a_2$ and $a_1' \ll a_2'$, then $a_1 + a_1' \ll a_2 + a_2'$;

(2) if $a_1 \leq a_2$ and $a_2 \ll a_3$, then $a_1 \ll a_3$;
(3) if $a_1 \ll a_2$ and $a_2 \leq a_3$, then $a_1 \ll a_3$;
(4) if $a_1 \ll a_2$ and $a_2 \ll a_3$, then $a_1 \ll a_3$;
(5) if $\{a_i\}$ is a sequence in G which converges to 0, then, for any $\epsilon \gg 0$, there exists $i_0 \in \mathbb{N}_+$ such that $a_i \ll \epsilon$ for all $i \geq i_0$;
(6) $0 \in Int(G^+)$ if and only if every subset of G is open;
(7) G is Hausdorff.

Definition 2.6. Let X be a nonempty set and $G \in \mathcal{G}$. A function $d : X \times X \to G$ is called a cone metric on X if for all $x_1, x_2, x_3 \in X$

(1) $d(x_1, x_2) \geq 0$ and $d(x_1, x_2) = 0$ if and only if $x_1 = x_2$;
(2) $d(x_1, x_2) = d(x_2, x_1)$;
(3) $d(x_1, x_2) \leq d(x_1, x_3) + d(x_3, x_2)$.

We say that (X, G, d) is a group valued cone metric space.

Although Definition 2.3 and Definition 2.6 are essentially the same (since the metric in Definition 2.3 does not use the ring R), it is still important to make a distinction between them. This is because we intend to not use the ring R in the proofs of our main results. The following is the analog of Definition 2.4 in the general setting.

Definition 2.7. Let (X, G, d) be a group valued cone metric space. A sequence $\{x_i\}$ in X is said to converge to a point $x \in X$ if for all $\epsilon \gg 0$ there exists $i_0 \in \mathbb{N}_+$ such that $d(x_i, x) \ll \epsilon$ for all $i \geq i_0$.

Keeping Definition 2.7 in mind, the notion of open balls usually has the following definition:

Definition 2.8. Let (X, G, d) be a group valued cone metric space, $x \in X$ and $\epsilon \gg 0$. The set $B(x; \epsilon) = \{y \in X : d(x, y) \ll \epsilon\}$ is called an open ball of radius ϵ centered at x.

When G is a topological vector space, one can show that the class of all open balls is basis for a topology on X [10]. However, when G has no additional algebraic structure on it, we must consent with making it a hypothesis. We denote by $\mathcal{X}_\tau$ the class of all group valued cone metric spaces (X, G, d) such that

(1) $G \in \mathcal{G}$;
(2) open balls form a basis for a topology on X.

For all group valued cone metric spaces (X, G, d) where $(X, G, d) \in \mathcal{X}_\tau$, the topology induced by open balls will be denoted by τ_d. A subset O of X is in τ_d if and only if for all $x \in O$ there exists $\epsilon \gg 0$ such that $B(x; \epsilon) \subset O$. The following proposition is obvious.

Proposition 2.9. *Let (X, G, d) be a group valued cone metric space such that $(X, G, d) \in \mathcal{X}_\tau$, then, a sequence $\{x_i\}$ in X converges to a point $x \in X$ with respect to the topology τ_d if and only if $\{x_i\}$ converges to x in the sense of Definition 2.7.*

3. Main Results

We start by giving a sufficient condition for the collection of open balls to be a basis for a topology.

Theorem 3.1. *Let (X, G, d) be a group valued cone metric space where $G \in \mathcal{G}$. Then, the class of all open balls is a basis for a topology on X if for all $\epsilon_1 \gg 0$ and $\epsilon_2 \gg 0$, there exists $\epsilon \gg 0$ such that $\epsilon \leq \epsilon_1$ and $\epsilon \leq \epsilon_2$.*

Proof. Denote the class of all balls by $\mathcal{B}$. Clearly, $x \in B(x; \epsilon)$ for all $x \in X$ and $\epsilon \in \text{Int}(G^+)$. Hence, $\mathcal{B}$ covers X. Let $x_1, x_2 \in X$ and $\epsilon_1, \epsilon_2 \in \text{Int}(G^+)$. Let $x \in B(x_1; \epsilon_1) \cap B(x_2; \epsilon_2)$. Hence, $\epsilon_1 - d(x, x_1) \gg 0$ and $\epsilon_2 - d(x, x_2) \gg 0$. Choose an $\epsilon \gg 0$ such that $\epsilon \leq \epsilon_1 - d(x, x_1)$ and $\epsilon \leq \epsilon_2 - d(x, x_2)$. We claim that $B(x; \epsilon) \subset B(x_1; \epsilon_1) \cap B(x_2; \epsilon_2)$. Let $y \in B(x; \epsilon)$, so that $d(y, x) \ll \epsilon$. By Proposition 2.5(3), $d(y, x) \ll \epsilon_1 - d(x, x_1)$ and $d(y, x) \ll \epsilon_2 - d(x, x_2)$, which, by Proposition 2.5(1), implies that $d(y, x) + d(x, x_1) \ll \epsilon_1$ and $d(y, x) + d(x, x_2) \ll \epsilon_2$. Hence, $d(y, x_1) \leq d(y, x) + d(x, x_1) \ll \epsilon_1$ and $d(y, x_1) \leq d(y, x) + d(x, x_2) \ll \epsilon_2$. So, by Proposition 2.5(2), $y \in B(x_1; \epsilon_1) \cap B(x_2; \epsilon_2)$. □

The following corollary of Theorem 3.1 is of utmost importance.

Corollary 3.2. *Let (X, d, G) be a group valued cone metric space. If there exists a sequence $\{\epsilon_i\}$ in Int(G^+) converging to 0, then, $(X, G, d) \in \mathcal{X}_\tau$.*

Proof. Choose $\epsilon' \gg 0$ and $\epsilon'' \gg 0$. Hence, $\epsilon', \epsilon'' \in \text{Int}(G^+)$. Hence, 0 is in interior point of $\epsilon' - G^+ \cap \epsilon'' - G^+$. Hence, there exists some $\epsilon_i \in \epsilon' - G^+ \cap \epsilon'' - G^+$. Let $\epsilon = \epsilon_i$. Clearly, $\epsilon \gg 0$ and $\epsilon \ll \epsilon'$ and $\epsilon \ll \epsilon''$. □

Now, we prove that the topology τ_d satisfies some desirable separation properties under an additional assumption that there exists a sequence $\{\epsilon_i\}$ in $\text{Int}(G^+)$ converging to 0.

Theorem 3.3. *Let (X, G, d) be a group valued cone metric space such that $G \in \mathcal{G}$. If there exists a sequence $\{\epsilon_i\}$ in Int(G^+) converging to 0, then, (X, τ_d) is a T_1 space.*

Proof. Let $x_1, x_2 \in X$ be such that $x_1 \neq x_2$. Let, if possible, $d(x_1, x_2) \ll \epsilon_i$ for all $i \in \mathbb{N}_+$. Hence, $\epsilon_i - d(x_1, x_2) \in \text{Int}(G^+) \subset G^+$ for all $i \in \mathbb{N}_+$. Since, the sequence $\{\epsilon_i - d(x_1, x_2)\}$ converges to $-d(x_1, x_2)$ and G^+ is closed, hence, $-d(x_1, x_2) \geq 0$. This gives, $d(x_1, x_2) = 0$, which is a contradiction. Thus, we can choose $i \in \mathbb{N}_+$ such that $d(x_1, x_2) \ll \epsilon_i$ does not hold. Hence, the open ball $B(x_1; \epsilon_i)$ is an open set containing x_1 but not containing x_2. □

Theorem 3.4. *Let (X, G, d) be a group valued cone metric space such that $G \in \mathcal{G}$. If there exists a sequence $\{\epsilon_i\}$ in Int(G^+) converging to 0, then, (X, τ_d) is a regular space.*

Proof. We have already proved that τ_d is T_1. Let $x \in X$ and C be a closed set in X not containing x. Hence, we can find an $\epsilon \gg 0$ such that $B(x; \epsilon) \cap C = \emptyset$. Since, the sequence $\{\epsilon_i + \epsilon_i\}$ converges to 0 and $\epsilon \gg 0$, hence, by Proposition 2.5(5), we can find $i \in \mathbb{N}_+$ such that $\epsilon_i + \epsilon_i \ll \epsilon$. By Proposition 2.5(2.5), $\epsilon_i + \epsilon_i \gg 0$ and clearly $B(x; \epsilon_i + \epsilon_i) \cap C = \emptyset$. Define

$$O = \bigcup_{y \in C} B(y; \epsilon_i).$$

Clearly, O is an open set containing C. We claim that $B(x; \epsilon_i) \cap O = \emptyset$. Suppose not. Hence, we can find a $y \in C$ and a $z \in B(y; \epsilon_i) \cap B(x; \epsilon_i)$. Hence, $d(x, y) \leq d(x, z) + d(z, y) \ll \epsilon_i + \epsilon_i$. Hence, $y \in B(x; \epsilon_i + \epsilon_i) \cap C$, which is a contradiction. □

Corollary 3.5. *Let (X, G, d) be a group valued cone metric space such that $G \in \mathcal{G}$. If there exists a sequence $\{\epsilon_i\}$ in Int(G^+) converging to 0, then, (X, τ_d) is a Hausdorff space.*

As we have proved that (X, τ_d) is a regular if we can find a sequence $\{\epsilon_i\}$ in Int(G^+) converging to 0; our next goal should be to show that every group valued cone metric space (X, G, d) with $(X, G, d) \in \mathcal{X}_\tau$ and satisfying this extra condition has a countably locally finite basis so that we may apply Nagata Simrnov metrization theorem.

Theorem 3.6. *Let (X, G, d) be a group valued cone metric space such that $G \in \mathcal{G}$. If there exists a sequence $\{\epsilon_i\}$ in Int(G^+) converging to 0, then,*

the family $\{B(x;\epsilon_i) : i \in \mathbb{N}_+\}$ is a local base at x. Hence, (X, τ_d) is a first countable space.

Proof. Let O be an open set in X and $x \in O$. Hence, we can find an $\epsilon \gg 0$ such that $B(x;\epsilon) \subset O$. By Proportion 2.5(5), we can find $i \in \mathbb{N}_+$ such that $\epsilon_i \ll \epsilon$. Hence, $B(x;\epsilon_i) \subset B(x;\epsilon) \subset O$. □

Note that if $G \in \mathcal{G}$ and there is a sequence in $\text{Int}(G^+)$ converging to 0, then, continuity of the function $(a_1, a_2, \ldots, a_j) :\to a_1 + a_2 + \ldots + a_j$ and Proposition 2.5(5) implies that for all $\epsilon \gg 0$ and $j \in \mathbb{N}_+$ there exists $\epsilon' \gg 0$ such that

$$\underbrace{\epsilon' + \epsilon' + \ldots + \epsilon'}_{j \text{ times}} \ll \epsilon.$$

Theorem 3.7. *Let, (X, G, d) be a group valued cone metric space such that $G \in \mathcal{G}$ and there exists a sequence $\{\epsilon_i\}$ in Int(G^+) converging to 0. Then, for every open covering $\mathcal{O}$ of X, there exists a countably locally finite open refinement of $\mathcal{O}$ covering X.*

Proof. Using well ordering theorem, we can assume that there is a well ordering $\prec$ defined on $\mathcal{O}$. Let $i \in \mathbb{N}_+$ be fixed. For each $O \in \mathcal{O}$, let us define

$$P_i(O) = \{x : B(x;\epsilon_i) \subset O\} - \bigcup_{U \prec O} U.$$

We claim that if $O_1, O_2 \in \mathcal{O}$ are distinct, then, for all $x_1 \in P_i(O_1)$ and $x_2 \in P_i(O_2)$, $d(x_1, x_2) \not\ll \epsilon_i$. To see this, assume that $O_1 \prec O_2$. Hence, $B(x_1;\epsilon_i) \subset O_1$ and $x_2 \notin O_1$. The claim now follows immediately.

Choose $\epsilon_i' \gg 0$ such that $\epsilon_i' + \epsilon_i' + \epsilon_i' \ll \epsilon_i$. For each $O \in \mathcal{O}$, let us define

$$Q_i(O) = \bigcup_{x \in P_i(O)} B(x;\epsilon').$$

We claim that if $O_1, O_2 \in \mathcal{O}$ are distinct, then, for all $x_1 \in Q_i(O_1)$ and $x_2 \in Q_i(O_2)$, $d(x_1, x_2) \not\ll \epsilon_i'$. If $x_1 \in Q_i(O_1)$ and $x_2 \in Q_i(O_2)$, then, we can find $y_1 \in P_i(O_1)$ and $y_2 \in P_i(O_2)$ such that $d(x_1, y_1) \ll \epsilon_i'$ and $d(x_2, y_2) \ll \epsilon_i'$. If $d(x_1, x_2) \ll \epsilon'$, then,

$$d(y_1, y_2) \leq d(y_1, x_1) + d(x_1, x_2) + d(y_2, x_2) \ll \epsilon_i' + \epsilon_i' + \epsilon_i' \ll \epsilon_i$$

which contradicts what we have already proved. For each $i \in \mathbb{N}_+$, let us define

$$\mathcal{Q}_i = \{Q_i(O) : O \in \mathcal{O}\}.$$

We claim that $\mathcal{Q}_i$ is a locally finite open refinement of $\mathcal{O}$. Clearly, each set in $\mathcal{Q}_i$ is open. We claim that $Q_i(O) \subset O$ for all $O \in \mathcal{O}$. This will show that $\mathcal{Q}_i$ is a refinement of $\mathcal{O}$. Let $x \in Q_i(O)$. Hence, there exists $y \in P_i(O)$ such that

$$x \in B(y; \epsilon_i') \subset B(y; \epsilon_i) \subset O.$$

Choose $\epsilon'' \gg 0$ such that $\epsilon'' + \epsilon'' \ll \epsilon'$. We claim that for all $x \in X$, the neighbourhood $B(x; \epsilon'')$ of x can intersect with at most two sets in $\mathcal{Q}_i$. Suppose not. Hence, we can find distinct $O_1, O_2 \in \mathcal{O}$ and

$$y_1 \in Q_i(O_1) \cap B(x; \epsilon''), y_2 \in Q_i(O_2) \cap B(x; \epsilon'').$$

Hence, we have

$$d(y_1, y_2) \leq d(y_1, x) + d(x, y_2) \ll \epsilon_i'' + \epsilon_i'' \ll \epsilon_i'$$

which is a contradiction. Hence, $\mathcal{Q}_i$ is locally finite. Consider the countably locally finite family

$$\mathcal{Q} = \bigcup_{i \in \mathbb{N}_+} \mathcal{Q}_i.$$

We claim that $\mathcal{Q}$ is a covering of X. Let $x \in X$. Let us define

$$O_x = \min\{O \in \mathcal{O} : x \in O\}.$$

Since O_x is an open set containing x, by Theorem 3.6, we can find $i \in \mathbb{N}_+$ such that $B(x; \epsilon_i) \subset O_x$. By the definition of $P_i(O_x)$ and the minimality of O_x, it follows that $x \in P_i(O) \subset Q_i(O)$. This completes the proof. □

Theorem 3.8. *Let, (X, G, d) be a group valued cone metric space such that $G \in \mathcal{G}$ and there exists a sequence $\{\epsilon_i\}$ in Int(G^+) converging to 0, then, (X, τ_d) is a metrizable space.*

Proof. We claim that X has a countably locally finite basis. For each $i \in \mathbb{N}_+$, consider the collection

$$\mathcal{O}_i = \{B(x; \epsilon_i) : x \in X\}.$$

By Theorem 3.7, we can find a countably locally finite open refinement $\mathcal{B}_i$ of $\mathcal{O}_i$ which covers X. Let,

$$\mathcal{B} = \bigcup_{i \in \mathbb{N}_+} \mathcal{B}_i.$$

Obviously, each $\mathcal{B}_i$ being countably locally finite implies that $\mathcal{B}$ is also countably locally finite. We claim that $\mathcal{B}$ is a basis for X. Let $x \in X$ and $\epsilon \gg 0$.

Choose, $i \in \mathbb{N}_+$ such that $\epsilon_i + \epsilon_i \ll \epsilon$. Since $\mathcal{B}_i$ is a covering of X, way can choose $B \in \mathcal{B}_i$ such that $x \in B$. We claim that $B \subset B(x;\epsilon)$. To see this, let $y \in B$. Since, $\mathcal{B}_i$ is a refinement of $\mathcal{O}_i$, hence, B is contained in an element of $\mathcal{O}_i$. Hence, $d(x,y) \ll \epsilon_i + \epsilon_i \ll \epsilon$.

By Theorem 3.4, and Nagata-Smirnov metrization theorem, it follows that (X, τ_d) is a metrizable topological space. □

Let I be an index set and for all $\alpha \in I$ let S_α be a set. Then we denote by $\prod_{\alpha\in I} S_\alpha$ the set of all functions $x : I \to \cup_{\alpha\in I} S_\alpha$ such that $x(\alpha) \in S_\alpha$. The value of x at α is denoted by x_α instead of $x(\alpha)$ and is called the αth component of x. The function x is denoted by $\{x_\alpha\}_{\alpha\in I}$. When $S_\alpha = S$ for all $\alpha \in I$, we denote the Cartesian product by S^I. The following example illustrates Theorem 3.8 and also shows that the additional hypothesis of Theorem 3.8 cannot be removed completely.

Example 3.9. Let I be an index set. Consider the set $G = \mathbb{R}^I$. The Cartesian product G is a partially ordered topological group under pointwise addition, box topology and pointwise partial order. It is easy to show that $G \in \mathcal{G}$ with

$$G^+ = \{x \in G : x_\alpha \in \mathbb{R}_0^+ \text{ for all } \alpha \in I\}$$

and

$$\text{Int}(G^+) = \{x \in G : x_\alpha \in \mathbb{R}^+ \text{ for all } \alpha \in I\}.$$

Let $|\cdot|$ denote the absolute value function of $\mathbb{R}$. Define, $d : G \times G \to G$ as

$$d(x,y) = \{|x_\alpha - y_\alpha|\}_{\alpha\in I}.$$

Easy arguments show that d is a cone metric on G. We claim that $(X, G, d) \in \mathcal{X}_\tau$. To avoid confusion, we'll denote the identity of G by θ and the zero of $\mathbb{R}$ by 0. Let $\epsilon' \gg \theta$ and $\epsilon'' \gg \theta$ be given. Let us define $\epsilon \in G$ whose α^{th} component is given by

$$\epsilon_\alpha = \min\{\epsilon'_\alpha, \epsilon''_\alpha\}.$$

Clearly, $\epsilon \gg \theta$ and $\epsilon \leq \epsilon'$ and $\epsilon \leq \epsilon''$. Hence, by Theorem 3.1, $(X, G, d) \in \mathcal{X}_\tau$. For all $x \in G$ and $\epsilon \gg \theta$, the equation

$$B(x;\epsilon) = \prod_{\alpha\in I} (x_\alpha - \epsilon_\alpha, x_\alpha + \epsilon_\alpha)$$

implies that the topology induced by d on G coincides with the box topology of G.

If the index set I is finite, then, G satisfies the additional hypothesis of Theorem 3.8, likewise the box topology of G is metrizable in this case. On the other hand, if the index set I is infinite, it is well known that the additional hypothesis of Theorem 3.8 is not satisfied and also the box topology on G is not metrizable.

If G is a module over a ring R, the identity of G will be denoted by θ and the additive identity of R will be denoted by 0. The partial order of R will be denoted by $\preccurlyeq$ and the partial order of G will be denoted by $\leq$. All rings will be assumed associative and with a multiplicative identity 1 such that $1 \succ 0$. We define,

$$U(R^+) = \{r \in R : r \succ 0 \text{ and } r \text{ is invertible}\}.$$

Theorem 3.10. *Let, (X, G, d) be a cone metric space over a topological left R-module G such that $G \in \mathcal{G}$. If there exists a sequence $\{r_i\}$ in $U(R^+)$ which converges to 0, then, (X, τ_d) is metrizable.*

Proof. Choose $\epsilon \gg \theta$. For each $i \in \mathbb{N}_+$, the mapping $a \to r_i a$ is a homeomorphism. Hence, $r_i\epsilon \in \text{Int}(G^+)$ for all $i \in \mathbb{N}_+$. Moreover, the sequence $\{r_i\epsilon\}$ converges to θ. Hence, by Theorem 3.8, the conclusion follows. □

The existence of a sequence $\{r_i\}$ in R converging to 0 is needed for the proofs of most of fixed point theorems in these spaces (e.g. see [4]). Theorem 3.10 shows that in all those cases, the topology induced is metrizable.

Acknowledgement: We are grateful to the referee for the comments which improved the quality of the paper.

References

[1] Azam, Akbar and Fisher, Brian and Khan, M, Common fixed point theorems in complex valued metric spaces, *Numerical Functional Analysis and Optimization* **32** (2011), 243–253.

[2] Beg, Ismat and Azam, Akbar and Arshad, Muhammad, Common fixed points for maps on topological vector space valued cone metric spaces, *International Journal of Mathematics and Mathematical Sciences* **2009** (2009), .

[3] Beg, Ismat and Datta, Sanjib Kumar and Pal, Dipankar, Fixed point in bicomplex valued metric spaces, *International Journal of Nonlinear Analysis and Applications* **12** (2021), 717–727.

[4] Branga, Adrian Nicolae, Fixed Point Results for F-Contractions in Cone Metric Spaces over Topological Modules and Applications to Integral Equations, *Fractal and Fractional* **6** (2021), 16.

[5] Branga, Adrian Nicolae, An application to a class of systems of nonlinear integral equations of F-contractions on cone metric spaces over topological modules, *Journal of Industrial and Management Optimization* **21** (2025), 2980–3000.

[6] Branga, Adrian Nicolae and Olaru, Ion Marian, Cone Metric Spaces over Topological Modules and Fixed Point Theorems for Lipschitz Mappings, *Mathematics* **8** (2020), 724.

[7] Du, Wei-Shih, A note on cone metric fixed point theory and its equivalence, *Nonlinear Analysis: Theory, Methods & Applications* **72** (2010), 2259–2261.

[8] Huang, Long-Guang and Zhang, Xian, Cone metric spaces and fixed point theorems of contractive mappings, *Journal of mathematical Analysis and Applications* **332** (2007), 1468–1476.

[9] Kadelburg, Zoran and Radenović, Stojan and Rakočević, Vladimir, A note on the equivalence of some metric and cone metric fixed point results, *Applied Mathematics Letters* **24** (2011), 370–374.

[10] Lin, Shou and Li, Kedian and Ge, Ying, On the metrizability of TVS-cone metric spaces, *Publications De L'institut Mathematique* **98** (2015), 271–279.

[11] Liu, Hao and Xu, Shaoyuan, Cone metric spaces with Banach algebras and fixed point theorems of generalized Lipschitz mappings, *Fixed point theory and applications* **2013** (2013), 1–10.

[12] Munkres, James Raymond, Topology, Pearson, 2018.

Aman Deep
Department of Mathematics
University of Delhi, Delhi, India.
E-mail: amansharmacpl001@gmail.com

Rakesh Batra
Department of Mathematics
Hansraj College, Delhi, India.
E-mail: rakeshbatra.30@gmail.com

The Mathematics Student ISSN: 0025-5742
Vol. 95, Nos. 3-4, July - December (2026), 69–108

DYNAMICS OF A WAVE EQUATIONS WITH NONLINEAR DAMPING TERM AND CRITICAL NONLINEARITY

MELOUKA REMIL, FATIMA ZOHRA MAHDI, ABED YFRAH, DJAMEL KOUIDER DAOUADJI

(Received : 10 - 06 - 2025 ; Accepted : 28 - 08 - 2026)

Abstract. This work investigates an evolution equation with critical nonlinearity, analyzing the existence and blow-up of solutions. The existence theory is developed both locally and globally, supported by appropriate energy functionals and a priori estimates. We also prove uniqueness under general assumptions. We study the generation of dynamical systems in appropriate phase spaces, analyze the asymptotic regularity of solutions, and establish the existence of global and exponential attractors. Our results contribute to the understanding of the global structure of the solution set and the mechanisms that govern dissipation and regularity in nonlinear PDEs.

1. Introduction

In many applications, the propagation of waves is influenced by nonlinear effects, which can significantly modify their behavior. Understanding these nonlinear dynamics is crucial for controlling wave propagation in fields such as nonlinear optics, materials science, and general relativity.

This work focuses on the study of the dynamics of a wave equation with critical nonlinearity, where the interaction between dispersion, damping, and the nonlinear term produces complex phenomena such as global existence, finite-time blow-up, and long-time asymptotic behavior. The

2020 Mathematics Subject Classification: 35L70, 35B41
Key words and phrases: wave equation, Nonlocal damping, Critical exponent, Asymptotic regularity, Uniform exponential attractor.

© *Indian Mathematical Society, 2026*.

equation under study is given by

$$u_{tt}(x,t) - \Delta u + |u_t(x,t)|^{q-1}u_t(x,t) + f(u(x,t)) = 0, \qquad \text{in } \Omega \times (0,+\infty), \tag{1.1}$$

with boundary and initial conditions

$$u|_{\partial\Omega} = 0, \qquad u(0) = u_0, \qquad u_t(0) = u_1. \tag{1.2}$$

Here, Ω is a bounded domain in $\mathbb{R}^n$, $n \geq 1$, with $|\Omega| < \infty$. The function $u(x,t)$ denotes the unknown solution, $|u_t|^{q-1}u_t$ is a nonlinear damping term, and $f(u)$ is the critical nonlinear term. The exponent $q > 1$ controls the damping effect.

The study of wave equations with critical nonlinearity presents a rich mathematical area with open problems related to global existence, compactness, and asymptotic behavior. By combining functional analysis and variational methods, this work aims to advance the understanding of nonlinear wave dynamics and contribute to both their theoretical and applied aspects.

The study of this wave equation presents several mathematical difficulties. One of the main challenges is to determine the conditions for global existence and finite-time blow-up. Some initial data configurations may lead to singularities, requiring accurate energy estimates to understand the behavior of the solution. In addition, critical nonlinearities cause the failure of standard Sobolev embeddings, resulting in a loss of compactness. To overcome these difficulties, advanced techniques are needed. Another key issue is to understand the long-term dynamics of solutions, namely whether they stabilize, disperse, or oscillate. Energy methods play a crucial role in addressing these concerns by establishing inequalities that control the behavior of the solution.

The aim of this work is to study wave equations with critical nonlinearity from several perspectives. First, we establish the existence and uniqueness of solutions by constructing appropriate functional spaces and variational frameworks. Next, we investigate the mechanisms leading to blow-up by identifying critical energy thresholds. Finally, we analyze the asymptotic behavior of solutions using suitable dynamical systems techniques.

2. Existence and Uniqueness of Solutions

The goal is to show that there exists a unique solution (u, v) in a suitable functional space. For the equation (1.1) the associated energy functional is

$$E(t) = \frac{1}{2}\int_\Omega \left(u_t^2 + |\nabla u|^2\right)\, dx + \int_\Omega F(u)\, dx.$$

Differentiating $E(t)$ with respect to time yields

$$\frac{dE}{dt} = \int_\Omega \left(u_t u_{tt} + \nabla u \cdot \nabla u_t\right)\, dx + \int_\Omega f(u) u_t\, dx.$$

$$u_{tt} = \Delta u - |u_t|^{q-1}u_t - f(u),$$

we obtain

$$\frac{dE}{dt} = \int_\Omega \left(u_t \Delta u + \nabla u \cdot \nabla u_t\right)\, dx - \int_\Omega |u_t|^{q+1}\, dx.$$

Using Green's formula and the boundary condition $u = 0$ on $\partial\Omega$, it follows that:

$$\frac{dE}{dt} = -\int_\Omega |u_t|^{q+1}\, dx \le 0.$$

Therefore, $E(t) \le E(0), \qquad t \ge 0.$

Theorem 2.1. *Let $\Omega \subset \mathbb{R}^n$ be a bounded domain with smooth boundary $\partial\Omega$. Assume that the initial data satisfy $u_0 \in H^2(\Omega) \cap H_0^1(\Omega), u_1 \in L^2(\Omega)$, with $q > 1$, and the nonlinear term $f(u)$ is locally Lipschitz continuous. Then there exists a unique solution $u(t)$ to problem (1.1) such that*

$$u \in C([0,T]; H^2(\Omega) \cap H_0^1(\Omega)), \qquad u_t \in C([0,T]; L^2(\Omega)).$$

Proof. Let $\{w_j\}_{j=1}^\infty$ be an orthonormal basis of $H^2(\Omega) \cap H_0^1(\Omega)$ consisting of eigenfunctions of the Laplacian $-\Delta$ with Dirichlet boundary conditions. Define the finite-dimensional subspace $V_k = \operatorname{span}\{w_1, w_2, \ldots, w_k\}$. Construct the approximate solution $u_k(t)$ in the form

$$u_k(t) = \sum_{j=1}^k g_{jk}(t) w_j,$$

where $g_{jk}(t)$ are time-dependent coefficients. The initial conditions are

$$u_k(0) = \sum_{j=1}^k \langle u_0, w_j\rangle w_j, \qquad \partial_t u_k(0) = \sum_{j=1}^k \langle u_1, w_j\rangle w_j.$$

Multiplying equation (1.1) by w_j and integrating over Ω, the coefficients $g_{jk}(t)$, $j = 1, \dots, k$, satisfy the following system of ordinary differential equations:

$$\int_\Omega \left(u_{k,tt} w_j - \Delta u_k w_j + |u_{k,t}|^{q-1} u_{k,t} w_j + f(u_k) w_j\right) \, dx = 0.$$

Define the energy functional

$$E_k(t) = \frac{1}{2}\|u_{k,t}\|_{L^2}^2 + \frac{1}{2}\|\nabla u_k\|_{L^2}^2 + \int_\Omega F(u_k)\, dx,$$

where

$$F(u) = \int_0^u f(s)\, ds.$$

Differentiating $E_k(t)$ with respect to t, we obtain

$$\frac{d}{dt}E_k(t) = \int_\Omega \left(u_{k,t} u_{k,tt} + \nabla u_k \cdot \nabla u_{k,t} + f(u_k) u_{k,t}\right) \, dx.$$

Using the same procedure as in the proof of the energy decay, we obtain

$$\frac{d}{dt}E_k(t) \le -C\|u_{k,t}\|_{L^{q+1}}^{q+1}.$$

This implies that $E_k(t)$ is uniformly bounded with respect to k, and

$$\|u_k\|_{L^\infty(0,T;H^1(\Omega))} \le C, \qquad \|u_{k,t}\|_{L^\infty(0,T;L^2(\Omega))} \le C.$$

By the Banach–Alaoglu theorem, there exists a subsequence $\{u_k\}$ such that

$$u_k \rightharpoonup u \quad \text{weakly-* in } L^\infty(0,T;H^1(\Omega)),$$

$$u_{k,t} \rightharpoonup u_t \quad \text{weakly-* in } L^\infty(0,T;L^2(\Omega)).$$

Using the Aubin–Lions lemma, we obtain the strong convergence

$$u_k \to u \quad \text{strongly in } L^2(0,T;L^2(\Omega)).$$

Passing to the limit in the weak formulation shows that u satisfies

$$u_{tt} - \Delta u + |u_t|^{q-1} u_t + f(u) = 0. \qquad \square$$

Theorem 2.2. *Under the assumptions of Theorem 2.1, assume additionally that the nonlinear term $f(u)$ satisfies*

$$|f(u)| \le C(1 + |u|^p), \qquad \text{for some } p \ge 1,$$

and that the potential energy $F(u)$ satisfies $F(u) \ge -C_1|u|^2 - C_2$, for some constants $C_1, C_2 \ge 0$. Then the solution $u(t)$ exists globally in time, that is, for all $t \ge 0$, and satisfies $u \in L^\infty(0,\infty; H^2(\Omega) \cap H_0^1(\Omega))$, and $u_t \in L^\infty(0,\infty; L^2(\Omega)) \cap L^{q+1}(\Omega \times (0,\infty))$.

Proof. The proof relies on uniform a priori estimates and a continuation argument. Recall the energy functional defined for the Galerkin approximate solution: $E_k(t) = \frac{1}{2}\|u_{k,t}(t)\|^2_{L^2(\Omega)} + \frac{1}{2}\|\nabla u_k(t)\|^2_{L^2(\Omega)} + \int_\Omega F(u_k(t))\,dx$, where $F(u) = \int_0^u f(s)\,ds$. We have that $E_k(t)$ is non-increasing in time and satisfies

$$E_k(t) \le E_k(0), \qquad \text{for all } t \in [0,T].$$

By the growth assumption on F, we obtain

$$\int_\Omega F(u_k(t))\,dx \ge -C_1\|u_k(t)\|^2_{L^2(\Omega)} - C_2|\Omega|.$$

Combining this estimate with the energy inequality yields

$$\frac{1}{2}\|u_{k,t}(t)\|^2_{L^2} + \frac{1}{2}\|\nabla u_k(t)\|^2_{L^2} \le E_k(0) + C_1\|u_k(t)\|^2_{L^2} + C_2|\Omega|.$$

Using Poincaré's inequality, we obtain

$$\|\nabla u_k(t)\|^2_{L^2} \le C\big(E_k(0)+1\big), \quad \|u_{k,t}(t)\|^2_{L^2} \le C\big(E_k(0)+1\big).$$

Thus, the solution and its time derivative are uniformly bounded with respect to k and $t \in [0,T]$. Moreover, the dissipation term satisfies

$$\int_0^T \|u_{k,t}(s)\|^{q+1}_{L^{q+1}(\Omega)}\,ds \le E_k(0).$$

As in the proof of local existence, the uniform estimates allow us to extract weakly-* convergent subsequences:

$$\begin{aligned} u_k &\rightharpoonup u && \text{weakly-* in } L^\infty(0,T;H^1_0(\Omega)),\\ \partial_t u_k &\rightharpoonup \partial_t u && \text{weakly-* in } L^\infty(0,T;L^2(\Omega)). \end{aligned}$$

Furthermore,

$$u_{k,t} \rightharpoonup u_t \quad \text{weakly in } L^{q+1}(0,T;L^{q+1}(\Omega)).$$

Moreover, by the Aubin–Lions compactness lemma, we obtain the strong convergence

$$u_k \to u \quad \text{strongly in } L^2(0,T;L^2(\Omega)).$$

This allows us to pass to the limit in the nonlinear terms

$$f(u_k) \to f(u), \qquad |u_{k,t}|^{q-1}u_{k,t} \rightharpoonup |u_t|^{q-1}u_t,$$

using weak–strong convergence arguments. The local solution $u(t)$ exists up to some maximal time $T^* > 0$. If $T^* < \infty$, then the uniform estimates established above imply that

$$\sup_{t\in[0,T^*)} \left(\|u(t)\|^2_{H^1(\Omega)} + \|u_t(t)\|^2_{L^2(\Omega)} \right) < \infty.$$

Therefore, no finite-time blow-up occurs, and the solution can be continued beyond T^*, contradicting the maximality of T^*. Hence, $T^* = \infty$, and the solution exists globally in time. Passing to the limit in the Galerkin approximation and using the uniform estimates, we obtain

$$u \in L^\infty(0,\infty; H_0^1(\Omega)), \qquad u_t \in L^\infty(0,\infty; L^2(\Omega)) \cap L^{q+1}(\Omega \times (0,\infty)).$$

Finally, by elliptic regularity, since $-\Delta u = u_{tt} + |u_t|^{q-1}u_t + f(u) \in L^2(\Omega)$, we conclude that

$$u \in L^\infty(0,\infty; H^2(\Omega) \cap H_0^1(\Omega)).$$

Therefore, the solution exists for all $t \geq 0$ and satisfies the desired regularity properties. □

Theorem 2.3. *Let $\Omega \subset \mathbb{R}^n$, $n \geq 1$, be a bounded domain with smooth boundary, and let $q > 1$. Assume that f is Lipschitz continuous. Then problem (1.1) admits at most one weak solution.*

Proof. Assume that u_1 and u_2 are two weak solutions of problem (1.1) with the initial data $u_1(0) = u_2(0) = u_0, \qquad \partial_t u_1(0) = \partial_t u_2(0) = u_1$. Let $w = u_1 - u_2$, then w satisfies

$$w_{tt} - \Delta w + \left(|u_{1,t}|^{q-1}u_{1,t} - |u_{2,t}|^{q-1}u_{2,t} \right) + (f(u_1) - f(u_2)) = 0,$$

with the initial conditions

$$w(0) = 0, \qquad w_t(0) = 0.$$

Multiplying the above equation by w_t and integrating over Ω, we obtain

$$\int_\Omega w_t w_{tt}\, dx - \int_\Omega w_t \Delta w\, dx + \int_\Omega w_t \left(|u_{1,t}|^{q-1}u_{1,t} - |u_{2,t}|^{q-1}u_{2,t} \right) dx + \int_\Omega w_t \big(f(u_1) - f(u_2) \big)\, dx = 0.$$

Using integration by parts together with the boundary condition $w = 0$ on $\partial\Omega$, we obtain

$$\frac{1}{2}\frac{d}{dt}\left(\|w_t\|^2_{L^2(\Omega)} + \|\nabla w\|^2_{L^2(\Omega)}\right) + \int_\Omega w_t\left(|u_{1,t}|^{q-1}u_{1,t} - |u_{2,t}|^{q-1}u_{2,t}\right)\,dx + \int_\Omega w_t\big(f(u_1) - f(u_2)\big)\,dx = 0.$$

Since f is Lipschitz continuous, we have

$$\left|\int_\Omega w_t\left(f(u_1) - f(u_2)\right)\,dx\right| \leq L\int_\Omega |w_t||w|\,dx \leq \frac{L}{2}\left(\|w_t\|^2_{L^2} + \|w\|^2_{L^2}\right).$$

Moreover, the damping term satisfies

$$\left(|u_{1,t}|^{q-1}u_{1,t} - |u_{2,t}|^{q-1}u_{2,t}\right)(u_{1,t} - u_{2,t}) \geq 0,$$

by the monotonicity of the mapping $s \longmapsto |s|^{q-1}s$, then

$$\int_\Omega w_t\left(|u_{1,t}|^{q-1}u_{1,t} - |u_{2,t}|^{q-1}u_{2,t}\right)\,dx \geq 0.$$

Combining the previous estimates, we obtain

$$\frac{d}{dt}\left(\|w_t\|^2_{L^2} + \|\nabla w\|^2_{L^2}\right) \leq L\left(\|w_t\|^2_{L^2} + \|w\|^2_{L^2}\right).$$

Using Poincaré's inequality, we obtain

$$\frac{d}{dt}\left(\|w_t\|^2_{L^2} + \|\nabla w\|^2_{L^2}\right) \leq L(1+C)\left(\|w_t\|^2_{L^2} + \|\nabla w\|^2_{L^2}\right).$$

Define $E_w(t) = \|w_t\|^2_{L^2} + \|\nabla w\|^2_{L^2}$, then

$$\frac{dE_w(t)}{dt} \leq L(1+C)E_w(t).$$

Applying Gronwall's inequality and using the fact that $E_w(0) = 0$, we obtain

$$E_w(t) = 0, \qquad \forall\, t \geq 0.$$

Therefore, $w = 0$, which implies that $u_1 = u_2$. □

3. Long-time Dynamics

To generate a dynamical system associated with problem (1.1)–(1.2), we first introduce a suitable functional. We consider the energy space

$$\mathcal{H} = H^1_0(\Omega) \cap H^2(\Omega) \times L^2(\Omega).$$

Let $A = -\Delta$, with domain $D(A) = H^2(\Omega) \cap H_0^1(\Omega)$. For the family of Hilbert spaces $D(A^{s/2})$, where $s \in \mathbb{R}$, the corresponding inner products and norms are defined by

$$\langle \cdot, \cdot \rangle_{D(A^{s/2})} = \langle A^{s/2} \cdot, A^{s/2} \cdot \rangle, \qquad \| \cdot \|_{D(A^{s/2})} = \| A^{s/2} \cdot \|.$$

Thus,

$$D(A^{s/2}) \subset D(A^{r/2}), \qquad \text{for all } s > r.$$

Moreover, the following continuous embedding holds:

$$D(A^{s/2}) \hookrightarrow L^{\frac{6}{3-2s}}(\Omega), \qquad \text{for all } s \in \left[0, \frac{3}{2}\right).$$

The following interpolation inequality also holds. Given $s > r > q'$, for every $\varepsilon > 0$, there exists a positive constant $C_\varepsilon = C_\varepsilon(s, r, q')$ such that

$$|A^{r/2} u| \leq \varepsilon |A^{s/2} u| + C_\varepsilon |A^{q'/2} u|, \qquad \forall\, u \in D(A^{s/2}).$$

Define the operator $\mathcal{A} : D(\mathcal{A}) \subset \mathcal{H} \longrightarrow \mathcal{H}$ by

$$\mathcal{A}(u, v) = \left(v, \Delta u - |v|^{q-1} v - f(u)\right).$$

The domain of $\mathcal{A}$ is defined by

$$D(\mathcal{A}) = \left\{(u, v) \in \mathcal{H} \,:\, |v|^{q-1} v \in L^2(\Omega),\ f(u) \in L^2(\Omega)\right\}.$$

It can be shown that $\mathcal{A}$ generates a continuous semigroup on $\mathcal{H}$. More precisely, for any initial datum $U_0 \in \mathcal{H}$, there exists a unique weak solution $U(t) = \big(u(t), u_t(t)\big)$ of problem (1.1) satisfying:

$$U \in C([0, T]; \mathcal{H}), \quad u \in L^\infty(0, T; V), \qquad u_t \in L^\infty(0, T; H) \cap L^q(0, T; L^q(\Omega)).$$

Therefore, the mapping $S(t) : \mathcal{H} \longrightarrow \mathcal{H}, \qquad S(t)(u_0, u_1) = (u(t), u_t(t))$, defines a nonlinear dynamical system, or more precisely, a strongly continuous semigroup $\{S(t)\}_{t \geq 0}$ on the energy space $\mathcal{H}$. This semigroup describes the time evolution of the solution and provides the framework for studying qualitative properties such as global attractors, asymptotic compactness, and long-time behavior. We require the following assumptions.

(1) Assume that $f \in C^1(\mathbb{R}), \qquad f(0) = 0$, and

$$|f'(r)| \leq C_1(1 + |r|^q), \tag{3.1}$$

together with

$$\liminf_{|r|\to\infty} \frac{f(r)}{r} > -\lambda_1. \tag{3.2}$$

(2) There exist positive constants $c_1, c_2 > 0$ and $p > 1$ such that

$$|f(s)| \le c_1(1 + |s|^p),$$

and

$$F(s) \ge -c_2, \qquad \forall\, s \in \mathbb{R}. \tag{3.3}$$

(3) Assume that $g \in C(\mathbb{R})$.

(4) There exist positive constants $c_1, c_2 > 0$ such that

$$c_1|s|^{p+1} \le g(s) \le c_2|s|^{p+1}, \qquad \forall\, s \in \mathbb{R}. \tag{3.4}$$

The natural norm on $\mathcal{H}$ is defined by $\|(u,v)\|^2_{\mathcal{H}} = \|u\|^2_V + \|v\|^2_{L^2}$. We define the state variable $U(t) = \big(u(t), u_t(t)\big) \in \mathcal{H}$. Then, system (1.1)–(1.2) can be reformulated as the following first-order evolution equation in $\mathcal{H}$

$$\frac{d}{dt}U(t) = \mathcal{A}(U(t)), \qquad U(0) = (u_0, u_1). \tag{3.5}$$

In this subsection, we prove the existence of a bounded uniformly absorbing set for the dynamical system $\{U(t)\}_{t\ge 0}$.

Theorem 3.1. *Suppose that the nonlinear damping term $|u_t|^{q-1}u_t$ and the nonlinear function f satisfy assumptions (3.1)–(3.4), and assume that $u_1 \neq 0$. Then the semigroup generated by problem (3.5) possesses a bounded uniformly absorbing set $\widetilde{\mathcal{B}}$ in $\mathcal{H}$. More precisely, there exists a positive constant N such that, for every bounded subset $\mathcal{B} \subset \mathcal{H}$, there exists a time $T_{\mathcal{B}} > 0$ satisfying: $\|U(t,0)\mathcal{B}\|_{\mathcal{H}} \le N, \qquad \forall\, t \ge T_{\mathcal{B}}$.*

Proof. We introduce the modified velocity field

$$\widetilde{u} = u_t + \varepsilon u, \qquad 0 < \varepsilon \ll 1.$$

We define the modified energy functional by

$$\Phi(t) = \|\widetilde{u}(t)\|^2 + \|\nabla u(t)\|^2 + 2\int_\Omega F(u)\,dx,$$

and assume that it satisfies the energy identity

$$\frac{d}{dt}\Phi(t) + 2\Psi(t) = 0. \tag{3.6}$$

Our goal is to compute explicitly the dissipation functional $\Psi(t)$. Since $\widetilde{u} = u_t + \varepsilon u$, we have

$$\frac{d}{dt}\|\widetilde{u}\|^2 = 2\int_\Omega \left(u_t u_{tt} + \varepsilon u u_{tt} + \varepsilon u_t^2 + \varepsilon^2 u u_t\right) dx,$$
$$\frac{d}{dt}\|\nabla u\|^2 = 2\int_\Omega \nabla u \cdot \nabla u_t \, dx = -2\int_\Omega \Delta u \, u_t \, dx.$$

Moreover,

$$\frac{d}{dt}\left(2\int_\Omega F(u)\, dx\right) = 2\int_\Omega f(u) u_t \, dx.$$

Combining the above identities yields

$$\frac{d}{dt}\Phi(t) = 2\Bigg[\int_\Omega u_t u_{tt}\, dx + \varepsilon\int_\Omega u u_{tt}\, dx + \varepsilon\int_\Omega u_t^2\, dx + \varepsilon^2\int_\Omega u u_t\, dx - \int_\Omega \Delta u\, u_t\, dx + \int_\Omega f(u) u_t\, dx\Bigg]. \tag{3.7}$$

Using equation (1.1), we obtain

$$\frac{d}{dt}\Phi(t) = 2\Bigg[- \|\nabla u_t\|^2_{L^2(\Omega)} - \|u_t\|^{q+1}_{L^{q+1}(\Omega)} - \int_\Omega f(u) u_t\, dx + \varepsilon\Big(- \|\nabla u\|^2_{L^2(\Omega)} - \int_\Omega u|u_t|^{q-1}u_t\, dx - \int_\Omega u f(u)\, dx\Big) + \varepsilon\|u_t\|^2_{L^2(\Omega)} + \varepsilon^2\int_\Omega u u_t\, dx - \int_\Omega \Delta u\, u_t\, dx + \int_\Omega f(u) u_t\, dx\Bigg].$$

Substituting the previous identity into equation (3.7), we obtain

$$\frac{d}{dt}\Phi(t) = 2\Bigg[- \|\nabla u_t\|^2_{L^2(\Omega)} - \|u_t\|^{q+1}_{L^{q+1}(\Omega)} + \varepsilon^2\int_\Omega u u_t\, dx + \varepsilon\Big(- \|\nabla u\|^2_{L^2(\Omega)} - \int_\Omega u|u_t|^{q-1}u_t\, dx - \int_\Omega u f(u)\, dx + \|u_t\|^2_{L^2(\Omega)}\Big)\Bigg].$$

Therefore, we define

$$\Psi(t) = \|\nabla u_t\|^2_{L^2(\Omega)} + \|u_t\|^{q+1}_{L^{q+1}(\Omega)} - \varepsilon^2\int_\Omega u u_t\, dx + \varepsilon\Big(\|\nabla u\|^2_{L^2(\Omega)} + \int_\Omega u|u_t|^{q-1}u_t\, dx + \int_\Omega u f(u)\, dx - \|u_t\|^2_{L^2(\Omega)}\Big). \tag{3.9}$$

Indeed, assumption (3.2) implies that, for every $\delta > 0$, there exists a constant $C_\delta > 0$ such that:

$$f(u)u \geq -(\lambda_1 - \delta)|u|^2 - C_\delta. \tag{3.10}$$

Moreover,

$$F(u) \geq -\frac{1}{2}(\lambda_1 - \delta)|u|^2 - C_\delta. \tag{3.11}$$

From the estimate on $F(u)$ and Poincaré's inequality, we obtain

$$2\int_\Omega F(u)\,dx \geq -\frac{\lambda_1 - \delta}{\lambda_1}\|\nabla u\|^2 - 2C_\delta|\Omega|.$$

Substituting this estimate into $\Phi(t)$, we obtain

$$\Phi(t) \geq \|\widetilde{u}(t)\|^2 + \frac{\delta}{\lambda_1}\|\nabla u(t)\|^2 - 2C_\delta|\Omega|. \tag{3.12}$$

To estimate $\displaystyle\int_\Omega F(u)\,dx$, we first note that

$$|f(u)| = \left|\int_0^u f'(s)\,ds\right| \leq \int_0^{|u|} |f'(s)|\,ds.$$

Using assumption (3.1), we obtain

$$|f(u)| \leq C_1\int_0^{|u|}(1+s^q)\,ds = C_1\left(|u| + \frac{|u|^{q+1}}{q+1}\right).$$

Integrating this estimate yields

$$\int_\Omega |F(u)|\,dx \leq C\int_\Omega \left(|u|^2 + |u|^{q+2}\right)\,dx.$$

By the Sobolev embedding theorem together with Poincaré's inequality, we obtain

$$\int_\Omega |F(u)|\,dx \leq C\left(\|u\|_{L^2}^2 + \|u\|_{L^{q+2}}^{q+2}\right) \leq C\left(\|\nabla u\|^2 + \|\nabla u\|^{q+2}\right).$$

Combining the above estimates, we obtain

$$\Phi(t) \leq C_f\left(\|\widetilde{u}(t)\|^2 + \|\nabla u(t)\|^2 + \|\nabla u(t)\|^{q+2}\right).$$

In particular, when $q = 2$, we obtain

$$\Phi(t) \leq C_f\left(\|\widetilde{u}(t)\|^2 + \|\nabla u(t)\|^2 + \|\nabla u(t)\|^4\right). \tag{3.13}$$

Using inequality (3.10), we obtain

$$\langle f(u), u\rangle \geq \frac{\lambda_1 - \delta}{\lambda_1}\|\nabla u(t)\|^2 - C_\delta|\Omega|.$$

Next, define

$$h(t) = \epsilon\delta\|u(t)\|^2 - 2\epsilon\left(\epsilon + |u_t|^{q-1}\right)\|u(t)\|\|\widetilde{u}(t)\| + \left(|u_t|^{q-1} - 2\epsilon\right)\|\widetilde{u}(t)\|^2.$$

Using the energy identity (3.6), we obtain

$$\Psi(t) = \left(|u_t|^{q-1} - 2\epsilon\right) \|\widetilde{u}(t)\|^2 + \epsilon\langle f(u), u\rangle + h(t) - \left(|u_t|^{q-1} - 2\epsilon\right) \|\widetilde{u}(t)\|^2.$$

Using the estimate

$$\langle f(u), u\rangle \geq \frac{\lambda_1 - \delta}{\lambda_1}\|\nabla u(t)\|^2 - C_\delta|\Omega|,$$

and assuming that $\delta \leq \frac{\lambda_1}{2}$, we obtain $\epsilon\langle f(u), u\rangle \geq \frac{\epsilon\delta}{2\lambda_1}\|\nabla u(t)\|^2 - \epsilon C_\delta|\Omega|$. Therefore,

$$\Psi(t) \geq \left(|u_t|^{q-1} - 2\epsilon\right) \|\widetilde{u}(t)\|^2 + \frac{\epsilon\delta}{2\lambda_1}\|\nabla u(t)\|^2 - \epsilon C_\delta|\Omega| + h(t).$$

Assuming that $|u_t|^{q-1} \geq 4\epsilon$, it follows that:

$$\Psi(t) \geq \frac{\epsilon\delta}{2\lambda_1}\|\nabla u(t)\|^2 + \frac{1}{2}|u_t|^{q-1}\|\widetilde{u}(t)\|^2 + \frac{1}{2}h(t) - C_\delta|\Omega|.$$

Choose $\epsilon_1 = \min\left\{\frac{1}{2}, \frac{(\alpha-2)\delta}{(\beta+1)^2}\right\}$, sufficiently small so that $h(t) \geq 0$. Then,

$$\Psi(t) \geq \frac{\epsilon\delta}{2\lambda_1}\|\nabla u(t)\|^2 + \frac{1}{2}\alpha\|\widetilde{u}(t)\|^2 - \epsilon C_\delta|\Omega|. \tag{3.14}$$

Substituting inequality (3.14) into the energy identity (3.6), and setting $\mu = 2\epsilon C_\delta|\Omega|$, we obtain

$$\frac{d}{dt}\Phi(t) + \frac{\epsilon\delta}{\lambda_1}\|\nabla u(t)\|^2 + \frac{1}{2}\alpha\|\widetilde{u}(t)\|^2 \leq \mu.$$

Now let $\epsilon = \min\left\{\epsilon_1, \sqrt{\frac{\lambda_1}{2}}\right\}$, $\vartheta = \min\left\{\frac{\epsilon\delta}{2\lambda_1}, \frac{\alpha}{4}\right\}$. Here, $\epsilon_1 = \min\left\{\frac{1}{2}, \frac{(\alpha-2)\delta}{(\beta+1)^2}\right\}$ is chosen sufficiently small so that $h(t) \geq 0$, where

- $\alpha = \inf |u_t(x,t)|^{q-1}, \qquad \alpha > 2,$
- $\beta = \sup |u_t(x,t)|^{q-1},$
- δ is a geometric parameter (typically, $\delta = 1$),
- the damping coefficient satisfies $0 < \alpha \leq |u_t|^{q-1} \leq \beta$.

Then we have

$$\frac{d}{dt}\Phi(t) + \vartheta(\|\nabla u(t)\|^2 + \|u_t(t)\|^2) \leq \mu. \tag{3.15}$$

Applying Gronwall inequality, and combining with (3.12), (3.13) and (3.15), we get

$$\Phi(U(t)) \leq \sup_{U=(u,u_t)\in\mathcal{H}} \{\Phi(U) : \vartheta\|U\|_{\mathcal{H}}^2 \leq 2\mu\}, \quad \forall U(t) \in \mathcal{H},$$

provided that

$$t \geq T_B = \frac{2C_\delta|\Omega| + 2C_f(\|B\|_{\mathcal{H}}^2 + \|B\|_{\mathcal{H}}^4)}{\mu}.$$

Then B be a bounded subset in Banach space E.
Using (3.12) and $\epsilon \leq \sqrt{\frac{\lambda_1}{2}}$, we obtain that the set

$$\tilde{\mathcal{B}} = \left\{U \in \mathcal{H} : \|U\|_{\mathcal{H}}^2 \leq 2M + 4C_\delta|\Omega|\right\}, \tag{3.16}$$

is a bounded uniformly absorbing set for the process $\{U(t,0)\}$ corresponding to (3.5). □

Remark 3.2.

(1) In order to ensure that the bounded uniformly absorbing set is semi-invariant, our candidate for the absorbing set is

$$\hat{\mathcal{B}} = \overline{\bigcup_{t \geq T_{\tilde{\mathcal{B}}}} U(t,0)\tilde{\mathcal{B}}} \subset \tilde{\mathcal{B}}, \tag{3.17}$$

where $\tilde{\mathcal{B}}$ *is from* (3.16).

(2) Integrating inequality (3.15) *over* $[0,t]$, $t \leq T(\mathcal{B})$, *we have*

$$\Phi(t) \leq \Phi(0) + \mu(t).$$

Recalling (3.5), (3.12) *and* (3.13), *we get*

$$\|U(t,0)\mathcal{B}\|_{\mathcal{H}} \leq \mathcal{Q}_0(\|\mathcal{B}\|_{\mathcal{H}}), \quad \forall t \geq 0, \tag{3.18}$$

where the monotonic function $\mathcal{Q}_0$ *is independent of* t.

Lemma 3.3. *For all* $t \geq 0$, *the skew product flow defined by*

$$S(t)(u,0) = (U(t,0)u, T(t)), \quad (u,0) \in \mathbb{E} = \mathcal{H} \times \mathbb{R}^n,$$

is Lipschitz continuous on $\mathcal{B}_S = \hat{B} \times \mathbb{R}^n$, *so there exists a constant* $\alpha_0 > 0$ *such that* $\|S(t)x - S(t)y\|_{\mathbb{E}} \leq e^{\alpha_0 t}\|x - y\|_{\mathbb{E}}$, $\forall x, y \in \mathcal{B}_S$, *where* α_0 *depends on* $\mathcal{B}_S$.

Proof. Let $x = (u_1, z_1)$, $y = (u_2, z_2) \in \mathcal{B}_S = \hat{B} \times \mathbb{R}^n$. We want to estimate

$$\|S(t)x - S(t)y\|_{\mathbb{E}} = \|(U(t,0)u_1, T(t)z_1) - (U(t,0)u_2, T(t)z_2)\|_{\mathcal{H}\times\mathbb{R}^n}.$$

Using the product norm on $\mathbb{E}$

$$\|S(t)x - S(t)y\|_{\mathbb{E}}^2 = \|U(t,0)u_1 - U(t,0)u_2\|_{\mathcal{H}}^2 + \|T(t)z_1 - T(t)z_2\|_{\mathbb{R}^n}^2.$$

Since $\hat{B} \subset \mathcal{H}$ is bounded and the solution operator $U(t,0)$ is assumed to be differentiable or Lipschitz on $\hat{B}$, there exists $\alpha_1 > 0$ such that

$$\|U(t,0)u_1 - U(t,0)u_2\|_{\mathcal{H}} \leq e^{\alpha_1 t}\|u_1 - u_2\|_{\mathcal{H}}.$$

Similarly, if $T(t)$ is Lipschitz continuous, then there exists $\alpha_2 \geq 0$ such that

$$\|T(t)z_1 - T(t)z_2\|_{\mathbb{R}^n} \leq e^{\alpha_2 t}\|z_1 - z_2\|_{\mathbb{R}^n}.$$

Putting both estimates together

$$\|S(t)x - S(t)y\|_{\mathbb{E}}^2 \leq e^{2\alpha_1 t}\|u_1 - u_2\|_{\mathcal{H}}^2 + e^{2\alpha_2 t}\|z_1 - z_2\|_{\mathbb{R}^n}^2.$$

Let $\alpha_0 = \max(\alpha_1, \alpha_2)$. Then we obtain

$$\|S(t)x - S(t)y\|_{\mathbb{E}} \leq e^{\alpha_0 t}\|x - y\|_{\mathbb{E}}. \qquad \square$$

Theorem 3.4. *Let f and $|u_t|^{q-1}$ satisfy the Assumption (3.1)-(3.4), and assume further $\alpha > 0$. Then the semigroup $\{S(t)\}_{t\geq 0}$ induced by solutions of (1.1)-(1.2) on $\mathbb{E}$ has a bounded uniform exponentially attracting set. More precisely, there exist a bounded set $\mathcal{M} \times \mathbb{R}^n \subset \mathbb{E}^1$, a constant $\alpha > 0$ and a monotonically increasing function $\mathcal{Q}(\cdot)$, such that for any bounded set $\mathcal{B} \subset \mathbb{E}$, it holds*

$$dist_{\mathbb{E}}\left(S(t)\mathcal{B}, \mathcal{M} \times \mathbb{R}^n\right) \leq \mathcal{Q}(\|\mathcal{B}\|_{\mathbb{E}})e^{-\alpha t}, \quad \text{for all } t \geq 0, \tag{3.19}$$

where $\mathbb{E} = \mathcal{H} \times \mathbb{R}^n$, $\mathbb{E}^s = \mathcal{H}^s \times \mathbb{R}^n$, $s \in (0,1]$, and α is independent of $\mathcal{B}$.

Decomposition of the equation. For the nonlinear function f satisfying Assumption $(3.1) - (3.4)$, we get that f allows the following decomposition $f = f_0 + f_1$, where $f_0, f_1 \in C^1(\mathbb{R})$ and satisfy

$$|f_0(u)| \leq C|u|^q, \quad \text{for all } u \in \mathbb{R}, \tag{3.20}$$

$$f_0(u)u \geq 0, \quad \text{for all } u \in \mathbb{R}, \tag{3.21}$$

$$|f_1'(u)| \leq C, \quad \text{for all } u \in \mathbb{R}, \tag{3.22}$$

$$\liminf_{|u|\to\infty} \frac{f_1(u)}{u} > -\lambda_1.$$

In order to obtain the regularity estimates later, we decompose the solution $u(t)$ of equation (1.1)-(1.2) with the initial value $(u(0), u_t(0))$ into the sum

$$U(t,0)(u(0), u_t(0)) = Z(t,0)(u(0), u_t(0)) + Y(t,0)(u(0), u_t(0)), \tag{3.23}$$

for all $t \geq 0$ and all $(u(0), u_t(0)) \in \mathcal{H}$, where $Z(t,0)(u(0), u_t(0)) = (z(t), z_t(t))$ and $Y(t,0)(u(0), u_t(0)) = (y(t), y_t(t))$ are the solution of the following equations:

$$\begin{cases} z_{tt} - \Delta z + |u_t|^{q-1} z_t + f_0(z) = 0, \quad x \in \Omega, \\ z|_{\partial\Omega} = 0, \\ z(x,0) = u(0), \quad z_t(x,0) = u_t(0), \end{cases} \tag{3.24}$$

and

$$\begin{cases} y_{tt} - \Delta y + |u_t|^{q-1} y_t + f(u) - f_0(z) = 0, \quad x \in \Omega, \\ y|_{\partial\Omega} = 0, \\ y(x,0) = 0, \quad y_t(x,0) = 0. \end{cases} \tag{3.25}$$

The first a priori estimate. In this subsection, we establish some a priori estimates for the solutions of equations (3.24) and (3.25).

Lemma 3.5. *There exists a positive constant M such that for any $U(t) = (u(t), u_t(t)) \in \mathcal{H}$,*

$$\|(z(t), z_t(t))\|_{\mathcal{H}}^2 + \|(y(t), y_t(t))\|_{\mathcal{H}}^2 \le M \quad \forall t \ge 0. \tag{3.26}$$

Proof. The function $z(t)$ satisfies (3.24), we define the energy functional

$$E_Z(t) = \frac{1}{2}\|z_t(t)\|_{L^2}^2 + \frac{1}{2}\|\nabla z(t)\|_{L^2}^2 + \int_\Omega F_0(z(t))\, dx,$$

where $F_0(s) = \displaystyle\int_0^s f_0(\sigma)\, d\sigma$.

Multiplying the equation (3.27) by z_t and integrating over Ω, we get

$$\frac{d}{dt}E_Z(t) + \int_\Omega |u_t|^{q-1} z_t^2\, dx = 0.$$

Therefore,

$$E_Z(t) + \int_0^t \int_\Omega |u_t|^{q-1} z_t^2\, dx\, ds = E_Z(0).$$

Under the assumptions $f_0(z)z \ge 0$ and $|f_0(z)| \le C|z|^q$, we have $F_0(z) \ge 0$, so

$$E_Z(t) \le E_Z(0) \le C\left(\|u(0)\|_{H_0^1}^2 + \|u_t(0)\|_{L^2}^2\right).$$

Hence,

$$\|(z(t), z_t(t))\|_{\mathcal{H}}^2 \le C, \quad \forall t \ge 0. \tag{3.28}$$

The function $y(t)$ satisfies (3.24), we define the energy functional

$$E_Y(t) = \frac{1}{2}\|y_t(t)\|_{L^2}^2 + \frac{1}{2}\|\nabla y(t)\|_{L^2}^2.$$

Multiplying the equation (3.29) by y_t and integrating over Ω, we obtain

$$\frac{d}{dt}E_Y(t) + \int_\Omega |u_t|^{q-1} y_t^2\, dx = -\int_\Omega (f(u) - f_0(z)) y_t\, dx.$$

Using Young's inequality

$$\left|\int_\Omega (f(u)-f_0(z))y_t\,dx\right| \le \frac{1}{2}\int_\Omega |u_t|^{q-1}y_t^2\,dx + C\int_\Omega \frac{|f(u)-f_0(z)|^2}{|u_t|^{q-1}}\,dx.$$

By assumptions$|f(u)| \le C(1+|u|^p), \quad |f_0(z)| \le C|z|^q$, we get

$$|f(u)-f_0(z)|^2 \le C(1+|u|^{2p}+|z|^{2q}).$$

Since $u = z + y$, and z is bounded , while $u(t)$ lies in the absorbing set $\tilde{\mathcal{B}} \subset \mathcal{H}$, the integral is bounded. Therefore

$$\frac{d}{dt}E_Y(t) \le C \quad \Rightarrow \quad E_Y(t) \le Ct.$$

Because the initial energy is zero and $f(u)-f_0(z) \in L^2(0,T;L^2(\Omega))$, Gronwall's inequality gives

$$\|(y(t),y_t(t))\|_{\mathcal{H}}^2 \le C, \quad \forall t \ge 0. \tag{3.30}$$

From (3.28) and (3.30) we find

$$\|(z(t),z_t(t))\|_{\mathcal{H}}^2 + \|(y(t),y_t(t))\|_{\mathcal{H}}^2 \le M, \quad \forall t \ge 0,$$

for some constant $M > 0$ depending on the initial data . □

Lemma 3.6. *Under the following assumptions:*

(1) The system generates a solution semigroup $\{Z(t,0)\}_{t\ge 0}$ on the energy space $\mathcal{H}$.

(2) The initial data $(u(0),u_t(0)) \in \mathcal{H}$.

(3) There exists a damping mechanism satisfying

$$\int_0^t \|u_t(r)\|^2 dr \ge \gamma\|(u(0),u_t(0))\|_{\mathcal{H}}^2,$$

for some $\gamma > 0$ and all $t \ge t_0 > 0$.

Then the energy decays exponentially:

$$\|Z(t,0)(u(0),u_t(0))\|_{\mathcal{H}}^2 \le \mathcal{Q}_1(\|(u(0),u_t(0))\|_{\mathcal{H}})e^{-k_0 t} \quad \forall t \ge 0, \tag{3.31}$$

where $k_0 > 0$ is a constant and $\mathcal{Q}_1(\cdot)$ is monotone increasing function.

Proof. Consider the following Lyapunov function

$$\mathcal{E}(z,z_t) = \frac{1}{2}\|\nabla z\|^2 + 2\theta\langle z,z_t\rangle + \frac{1}{2}\|z_t\|^2 + \int_\Omega F_0(z)dx,$$

where

$$\theta = \min\left\{\frac{\sqrt{\lambda_1}}{4}, \frac{\alpha}{2}, \frac{\lambda_1\alpha}{2(\beta^2+\lambda_1)}\right\}, \tag{3.32}$$

$$F_0(u) = \int_0^u f_0(r)dr.$$

From (3.32), and we apply the Cauchy-Schwarz inequality followed by Young's inequality: $2\theta(z, z_t) \leq 2\theta\|z\|\|z_t\| \leq \varepsilon\|z\|^2 + \dfrac{\theta^2}{\varepsilon}\|z_t\|^2$.
Assuming and Poincaré-type inequality holds, we get

$$2\theta(z, z_t) \leq \varepsilon C\|\nabla z\|^2 + \frac{\theta^2}{\varepsilon}\|z_t\|^2.$$

Choosing ε appropriately, using the above, we estimate

$$\frac{1}{2}\|\nabla z\|^2 + 2\theta(z, z_t) + \frac{1}{2}\|z_t\|^2 \leq \frac{3}{4}(\|\nabla z\|^2 + \|z_t\|^2). \tag{3.33}$$

Using Young's inequality again, we get

$$\frac{1}{2}\|\nabla z\|^2 + 2\theta(z, z_t) + \frac{1}{2}\|z_t\|^2 \geq \frac{1}{4}(\|\nabla z\|^2 + \|z_t\|^2). \tag{3.34}$$

From (3.33) and (3.34) we find

$$\frac{1}{4}(\|\nabla z\|^2 + \|z_t\|^2) \leq \frac{1}{2}\|\nabla z\|^2 + 2\theta\langle z, z_t\rangle + \frac{1}{2}\|z_t\|^2 \leq \frac{3}{4}(\|\nabla z\|^2 + \|z_t\|^2).$$

If $(z(t), z_t(t))$ is a solution of (3.24), using (3.21) and (3.32), we have Differentiating $\mathcal{E}(z(t), z_t(t))$ with respect to time, we obtain

$$\frac{d}{dt}\mathcal{E}(z(t), z_t(t)) = \langle\nabla z, \nabla z_t\rangle + \langle z_t, z_{tt}\rangle + 2\theta\langle z_t, z_t\rangle + 2\theta\langle z, z_{tt}\rangle + \int_\Omega f_0(z)z_t\,dx.$$

Using the equation $z_{tt} = \Delta z - |u_t|^{q-1}z_t - f_0(z)$, we get

$$\frac{d}{dt}\mathcal{E}(z(t), z_t(t)) = -|u_t|^{q-1}\|z_t\|^2 + 2\theta\|z_t\|^2 - 2\theta\int_\Omega |u_t|^{q-1}zz_t\,dx$$
$$- 2\theta\int_\Omega f_0(z)z\,dx.$$

Using Cauchy-Schwarz Young's, and Poincaré inequalities, and assuming $|u_t|^{q-1} \geq \alpha$, we obtain

$$\frac{d}{dt}\mathcal{E}(z(t), z_t(t)) + \theta\|\nabla z(t)\|^2 + \left(\frac{1}{2}|u_0|^{q-1} - \theta\right)\|z_t(t)\|^2 \leq 0.$$

Using (3.20) and (3.26), we can conclude that there exists a $k_0 > 0$ small enough such that

$$\frac{d}{dt}\mathcal{E}(z(t), z_t(t)) + k_0\mathcal{E}(z(t), z_t(t)) \leq 0.$$

Applying Gronwall's inequality and using (3.35), we obtain

$$\|Z(t,0)(u(0), u_t(0))\|^2_{\mathcal{H}} \leq \mathcal{Q}_1(\|(u(0), u_t(0))\|_{\mathcal{H}})e^{-k_0 t} \quad \forall t \geq 0. \tag{3.1}$$

□

The next estimate is about higher regularity of the solution of (3.25).

Lemma 3.7. *There exists $k_1 > 0$ such that for any $t \geq 0$,*

$$\|Y(t,0)B_0\|^2_{\mathcal{H}^{\frac{1}{4}}} \leq \mathcal{Q}_2(\|B_0\|_{\mathcal{H}})e^{k_1(t)},$$

where $\mathcal{Q}_2(\cdot)$ is a monotone increasing function.

Proof. Take the inner product of (3.25) with $A^{\frac{1}{4}}y_t$, we deduce

$$\frac{1}{2}\frac{d}{dt}\Big(\|A^{\frac{1}{8}}y(t)\|^2 + \|A^{\frac{5}{8}}y(t)\|^2 + |u(t)|^{p-1}\|A^{\frac{1}{8}}y(t)\|^2\Big) + \Big\langle f(u) - f_0(z), A^{\frac{1}{4}}y(t)\Big\rangle = 0. \tag{3.36}$$

With the nonlinear term

$$\langle f(u) - f_0(z), A^{\frac{1}{4}}y(t)\rangle = \langle f_1(z), A^{\frac{1}{4}}y(t)\rangle + \langle f(u) - f(z), A^{\frac{1}{4}}y(t)\rangle.$$

Note that

$$\langle f_1(z), A^{\frac{1}{4}}y(t)\rangle = \frac{d}{dt}\langle f_1(z), A^{\frac{1}{4}}y(t)\rangle - \langle f_1'(z), A^{\frac{1}{4}}y_t(t)\rangle.$$

By (3.14), and Hölder's inequality we obtain

$$\begin{aligned}\Big|\Big\langle f_1'(z)z_t(t), A^{\frac{1}{4}}y(t)\Big\rangle\Big| &\leq C\int_\Omega |z_t(t)|\,|A^{\frac{1}{4}}y(t)|\,dx \\ &\leq C\|z_t(t)\| \cdot \|A^{\frac{1}{4}}y(t)\|,\end{aligned} \tag{3.37}$$

and

$$\Big|\langle f_1(z), A^{\frac{1}{4}}y(t)\rangle\Big| \leq C\int_\Omega |z(t)||A^{\frac{1}{4}}y(t)|dx \leq C\|z(t)\| \cdot \|A^{\frac{1}{4}}y(t)\|. \tag{3.38}$$

Secondly,

$$\begin{aligned}\langle f(u)-f(z),A^{\frac{1}{4}}y(t)\rangle &= \frac{d}{dt}\langle f(u)-f(z),A^{\frac{1}{4}}y(t)\rangle\\ &\quad-\left\langle f'(u)u_t(t)-f'(z)z_t(t),A^{\frac{1}{4}}y(t)\right\rangle\\ &\quad-\left\langle f(u)-f(z),A^{\frac{1}{4}}y_t(t)\right\rangle. \end{aligned}\tag{3.39}$$

And then recall (3.1) and Hölder's inequality to find

$$\begin{aligned}\left|\langle f'(u)z_t(t)-f'(z)z_t(t),A^{\frac{1}{4}}y(t)\rangle\right| &\le C\int_\Omega |z_t(t)||y(t)|(1+|u(t)|\\ &\quad+|z(t)|)A^{\frac{1}{4}}y(t)dx \le C\|A^{\frac{5}{8}}y(t)\|^2.\end{aligned}$$

Similarly, we get

$$\left|\langle f'(u)y_t(t),A^{\frac{1}{4}}y(t)\rangle\right| \le C\left(\|A^{\frac{5}{8}}y(t)\|^2+\|A^{\frac{1}{8}}y(t)\|^2\right),\tag{3.40}$$

and

$$\begin{aligned}\left|\langle f(u)-f(z),A^{\frac{1}{4}}y(t)\rangle\right| &\le C\|y(t)\|_{\mathcal{H}^1}\cdot\Big(1+\|u(t)\|^2_{\mathcal{H}^1}\\ &\quad+\|z(t)\|^2_{\mathcal{H}^1}\Big)\cdot\|A^{\frac{1}{4}}y(t)\|.\end{aligned}\tag{3.41}$$

Finally, combining (3.36)–(3.37) and (3.38)–(3.39), we obtain

$$\begin{aligned}&\frac{d}{dt}\left(\|A^{\frac{5}{8}}y(t)\|^2+\|A^{\frac{1}{8}}y(t)\|^2+\mathcal{F}(t)\right)\\ &\quad\le C\left(\|A^{\frac{5}{8}}y(t)\|^2+\|A^{\frac{1}{8}}y(t)\|^2\right),\end{aligned}\tag{3.42}$$

where $\mathcal{F}(t)=2\langle f(u)-f(z),A^{\frac{1}{4}}y(t)\rangle+2\langle f_1(z),A^{\frac{1}{4}}y(t)\rangle$. Applying Young's inequality in (3.38) and (3.41), we get

$$|\mathcal{F}(t)|\le\varepsilon\|A^{\frac{5}{8}}y(t)\|^2+C_{\varepsilon M},\tag{3.43}$$

where $\varepsilon>0$ is sufficiently small. Combining now (3.42) and (3.43), we conclude

$$\begin{aligned}\frac{d}{dt}\left(\|A^{\frac{5}{8}}y(t)\|^2+\|A^{\frac{1}{8}}y(t)\|^2+\mathcal{F}(t)\right) &\le k_1\left(\|A^{\frac{5}{8}}y(t)\|^2+\|A^{\frac{1}{8}}y(t)\|^2+\mathcal{F}(t)\right)\\ &\quad+C_{\varepsilon M},\end{aligned}$$

where k_1 depend on $\|z_t(t)\|$, $\|z(t)\|_{\mathcal{H}^1}$ and $\|u(t)\|_{\mathcal{H}^1}$.
Applying Gronwall's inequality and recalling

$$\|U(t,0)\mathcal{B}\|_{\mathcal{H}}\le\mathcal{Q}_0(\|\mathcal{B}\|_{\mathcal{H}}),\quad\forall t\ge0,$$

using (3.31) and (3.43), we have

$$\|A^{\frac{5}{8}}y(t)\|^2+\|A^{\frac{1}{8}}y_t(t)\|^2\le\mathcal{Q}_2(\|(y(0),y_t(0))\|_{\mathcal{H}})e^{-\gamma t}+C_{\varepsilon M}.$$

Finally, since $\|Y(t,0)B_0\|^2_{\mathcal{H}^{\frac{1}{4}}} \sim \|A^{\frac{1}{8}}y_t\|^2 + \|A^{\frac{5}{8}}y\|^2$, we conclude that

$$\|Y(t,0)B_0\|^2_{\mathcal{H}^{\frac{1}{4}}} \leq \mathcal{Q}_2(\|B_0\|_{\mathcal{H}})e^{k_1 t}. \qquad \square$$

Similarly to [2], [10], based on Lemma 3.5 and Lemma 3.6, we can now decompose the solution $u(t)$ of $(1.1)-(1.2)$ as follows:

Lemma 3.8. *Let $(u(t), u_t(t))$ be the solution of $(1.1)-(1.2)$ corresponding to the initial data $(u(0), u_t(0)) \in B_0$. Then, for any $\eta > 0$ there exist positive constants K_η and C_η such that $u(t) = z_1(t) + y_1(t)$ for all $t \geq 0$, where $z_1(t)$ and $y_1(t)$ satisfy the following estimates:*

$$\|y_1(t)\|^2_{\mathcal{H}^{\frac{5}{4}}} \leq K_\eta \quad \text{for all } t \geq 0,$$

$$\int_s^t \|z_1(r)\|^2_{\mathcal{H}^1} dr \leq \eta(t-s) + C_\eta \quad \text{for all } t \geq s \geq 0. \tag{3.44}$$

Proof. Let $\chi_R : \mathbb{R} \to [0,1]$ be a smooth function such that

$$\chi_R(s) = \begin{cases} 1, & |s| \leq R, \\ 0, & |s| \geq 2R, \end{cases} \quad \text{and} \quad \|\chi_R'\|_\infty \leq C.$$

Define $y_1(t)$ as the solution to the regularized problem

$$\begin{cases} y_{1,tt} - \Delta y_1 + |y_{1,t}|^{q-1}y_{1,t} + \chi_R(u)f(u) = 0, \\ y_1|_{\partial\Omega} = 0, \\ y_1(0) = u_0, \quad y_{1,t}(0) = u_1. \end{cases}$$

Then $z_1(t) = u(t) - y_1(t)$ satisfies

$$\begin{cases} z_{1,tt} - \Delta z_1 + \left(|u_t|^{q-1}u_t - |y_{1,t}|^{q-1}y_{1,t}\right) + (f(u) - \chi_R(u)f(u)) = 0, \\ z_1|_{\partial\Omega} = 0, \\ z_1(0) = 0, \quad z_{1,t}(0) = 0. \end{cases} \tag{3.45}$$

Let us analyze the properties of the source term:

- The function $\chi_R(u)$ is smooth, bounded between 0 and 1, and has compact support in the sense that $\chi_R(u(x)) = 0$ whenever $|u(x)| \geq 2R$, and $\chi_R(u(x)) = 1$ when $|u(x)| \leq R$.
- Since $f(u)$ is assumed to have at most polynomial growth, say $|f(u)| \leq C(1 + |u|^p)$, and since $\chi_R(u)$ vanishes when $|u| \geq 2R$,

we have

$$|\chi_R(u(x))f(u(x))| \leq C_R \quad \text{for all } x \in \Omega,$$

for some constant C_R depending on R. Therefore, $\chi_R(u)f(u) \in L^\infty(\Omega)$ uniformly in time.

Now observe that the equation for y_1 is a semilinear wave equation with a globally bounded source term:

$$y_{1,tt} - \Delta y_1 + |y_{1,t}|^{q-1}y_{1,t} = -\chi_R(u)f(u) \in L^\infty(\Omega).$$

From classical regularity theory for damped wave equations (see, [47,48]), we have the following result:

$$\|y_1(t)\|^2_{\mathcal{H}^s} = \|y_1(t)\|^2_{H^s} + \|y_{1,t}(t)\|^2_{H^{s-1}} \leq K_\eta, \quad \forall t \geq 0,$$

for all $s < 2$. Hence, choosing $s = \frac{5}{4} \in (1, 2)$, we deduce

$$\|y_1(t)\|^2_{\mathcal{H}^{5/4}} \leq K_\eta, \quad \forall t \geq 0.$$

Define the energy of z_1 by $E_{z_1}(t) = \dfrac{1}{2}\left(\|z_{1,t}(t)\|^2 + \|\nabla z_1(t)\|^2\right)$.
Then using (3.45) we have

$$\begin{aligned}\frac{d}{dt}E_{z_1}(t) = &-\int_\Omega \left(|u_t|^{q-1}u_t - |y_{1,t}|^{q-1}y_{1,t}\right) z_{1,t}\,dx \\ &-\int_\Omega \left(f(u) - \chi_R(u)f(u)\right) z_{1,t}\,dx.\end{aligned}$$

By the monotonicity of the function $\phi(s) = |s|^{q-1}s$, we have

$$\int_\Omega \left(|u_t|^{q-1}u_t - |y_{1,t}|^{q-1}y_{1,t}\right) z_{1,t}\,dx \geq c\|z_{1,t}\|^{q+1}_{L^{q+1}}.$$

Since $f(u) - \chi_R(u)f(u)$ is supported where $|u| \geq R$. If $f(u)$ has polynomial growth, then

$$\left|\int_\Omega \left(f(u) - \chi_R(u)f(u)\right) z_{1,t}\,dx\right| \leq \varepsilon\|z_{1,t}\|^2 + \frac{C}{R^\theta}.$$

Using Gronwall-type estimate and Combining the previous steps, we get

$$\frac{d}{dt}E_{z_1}(t) + c\|z_{1,t}\|^{q+1}_{L^{q+1}} \leq \varepsilon\|z_{1,t}\|^2 + \frac{C}{R^\theta}.$$

Integrating over $[s,t]$, and using Poincaré and Young inequalities, we conclude

$$\int_s^t \|z_1(r)\|^2_{\mathcal{H}^1}\, dr \leq \eta(t-s) + C_\eta,$$

by choosing R sufficiently large depending on η. □

Now we begin to establish the asymptotic regularity of solutions.

Lemma 3.9. *For all bounded set $B \subset \mathcal{H}$, there exists a positive constant $N_{\|B\|_{\mathcal{H}}}$ which depends only on the $\mathcal{H}$-bounds of B, such that:*

$$\|Y(t,0)u_0\|^2_{\mathcal{H}^{\frac{1}{4}}} \leq N_{\|B\|_{\mathcal{H}}} \quad \textit{for all } t \geq 0 \textit{ and } u_0 = (u(0), u_t(0)) \in B.$$

Proof. We define $\tilde{y}(t) := y_t(t) + \varepsilon y(t)$. and the modified energy functional by $\mathcal{E}_y(t) = \|\tilde{y}(t)\|^2_{\mathcal{H}^{1/4}} + \|y(t)\|^2_{\mathcal{H}^{5/4}}$, where $\mathcal{H}^s = D(A^{s/2})$.
We compute

$$\frac{d}{dt}\mathcal{E}_y(t) = 2\left\langle A^{1/4}\tilde{y}, A^{1/4}\tilde{y}_t\right\rangle + 2\left\langle A^{5/4}y, y_t\right\rangle.$$

Then, we get

$$\frac{d}{dt}\mathcal{E}_y(t) = 2\left\langle A^{1/4}\tilde{y}, y_{tt}\right\rangle + 2\varepsilon\left\langle A^{1/4}\tilde{y}, y_t\right\rangle + 2\left\langle A^{5/4}y, y_t\right\rangle.$$

From the equation (3.29), we have

$$\begin{aligned}\frac{d}{dt}\mathcal{E}_y(t) = &-2\langle A^{1/4}\tilde{y}, Ay\rangle - 2\langle A^{1/4}\tilde{y}, |u_t|^{q-1}y_t\rangle + 2\langle A^{1/4}\tilde{y}, f_0(z) - f(u)\rangle \\ &+ 2\varepsilon\langle A^{1/4}\tilde{y}, y_t\rangle + 2\langle A^{5/4}y, y_t\rangle.\end{aligned}$$

Using the identity $A^{1/4}\tilde{y} = A^{1/4}y_t + \varepsilon A^{1/4}y$, then we obtain the identity

$$\begin{aligned}\frac{d}{dt}\mathcal{E}_y(t) + \varepsilon\|y(t)\|^2_{\mathcal{H}^{5/4}} + \left(|u_t|^{q-1} - \varepsilon\right)\|\tilde{y}(t)\|^2_{\mathcal{H}^{1/4}} \\ + h_y(t) + 2\left\langle f(u) - f_0(z), A^{1/4}\tilde{y}(t)\right\rangle = 0.\end{aligned}$$

We can choose $\mathcal{E} = \min\left\{\frac{\lambda_1\alpha}{1+\beta}, \frac{\alpha}{2}, \frac{1}{2}\right\}$, we find

$$\frac{d}{dt}\mathcal{E}_y(t) + \epsilon y_y(t) + 2\langle f(u) - f_0(z), A^{\frac{1}{4}}\tilde{y}(t)\rangle \leq \frac{\epsilon}{4}\|\tilde{y}(t)\|^2_{\mathcal{H}^{\frac{1}{4}}}. \tag{3.46}$$

Indeed, we have

$$\begin{aligned}\langle f(u) - f_0(z), A^{\frac{1}{4}}\tilde{y}(t)\rangle = &\frac{d}{dt}\langle f(u) - f(z), A^{\frac{1}{4}}y(t)\rangle + \\ &\epsilon\langle f(u) - f(z), A^{\frac{1}{4}}y(t)\rangle + \frac{d}{dt}\langle f_1(z), A^{\frac{1}{4}}y(t)\rangle + \\ &\epsilon\langle f_1(z), A^{\frac{1}{4}}y(t)\rangle - \mathcal{I}_1 - \mathcal{I}_2 - \mathcal{I}_3,\end{aligned} \tag{3.47}$$

$$\mathcal{I}_1 = \langle f'(u)z_t(t) - f'(z)z_t(t), A^{\frac{1}{4}}y(t)\rangle,$$
$$\mathcal{I}_2 = \langle f'(u)y_t(t), A^{\frac{1}{4}}y(t)\rangle,$$
$$\mathcal{I}_3 = \langle f_1(z)z_t(t), A^{\frac{1}{4}}y(t)\rangle,$$

where we will estimate the integrals $\mathcal{I}_1$, $\mathcal{I}_2$ and $\mathcal{I}_3$. Due to assumption (3.1)-(3.4) and Lemma 3.5, we get

$$\begin{aligned}|\mathcal{I}_1| &= \left|\langle f'(u)z_t(t) - f'(z)z_t(t), A^{\frac{1}{4}}y(t)\rangle\right| \\ &\leq C\int_\Omega |z_t(t)|y(t)|(1+|u(t)|+|z(t)|)A^{\frac{1}{4}}y(t)|dx \\ &\leq C\int_\Omega |z_t(t)|y(t)|(1+|z_1(t)|+|y_1(t)|+|z(t)|)A^{\frac{1}{4}}y(t)|dx. \qquad (3.48)\end{aligned}$$

Now, using Hölder's and Sobolev-type inequality, we have

$$\int_\Omega |z_t(t)||y(t)||A^{\frac{1}{4}}y(t)|dx \leq C_{\|z_t(t)\|,\|y(t)\|_{\mathcal{H}^1}}\|y(t)\|_{\mathcal{H}^{\frac{5}{4}}},$$
$$\int_\Omega |z_t(t)||y(t)||z_1(t)||A^{\frac{1}{4}}y(t)|dx \leq C_{\|z_t(t)\|}\|\|z_1(t)\|_{\mathcal{H}^1}\|y(t)\|^2_{\mathcal{H}^{\frac{5}{4}}},$$
$$\int_\Omega |z_t(t)||y(t)||y_1(t)||A^{\frac{1}{4}}y(t)|dx \leq C_{\|z_t(t)\|,\|y(t)\|_{\mathcal{H}^1}}K_\eta\|y(t)\|_{\mathcal{H}^{\frac{5}{4}}},$$
$$\int_\Omega |z_t(t)||y(t)||z(t)||A^{\frac{1}{4}}y(t)|dx \leq C_M\|z_t(t)\|_{\mathcal{H}^1}\|y(t)\|^2_{\mathcal{H}^{\frac{5}{4}}}.$$

Thanks to Lemma 3.2.3, we can take T_1 large enough so that

$$\|z(t)\|_{\mathcal{H}^1} \leq \frac{\mathcal{E}}{C} \quad \text{for all } t \geq T_1. \qquad (3.49)$$

Combining (3.48)-(3.49), using Young's inequality, we find

$$|\mathcal{I}_1| \leq \frac{\epsilon}{24}\|y(t)\|^2_{\mathcal{H}^{\frac{5}{4}}} + C_{\|z_t(t)\|,\epsilon}\|z_1(t)\|^2_{\mathcal{H}^1}\cdot\|y(t)\|^2_{\mathcal{H}^{\frac{5}{4}}} + C_{\epsilon K_{\eta,M}}. \qquad (3.50)$$

Similarly, we have

$$|\mathcal{I}_2| = \left|\langle f'(u)y_t(t), A^{\frac{1}{4}}y(t)\rangle\right| \leq \int_\Omega |y_t(t)|(1+|y_1(t)|^2+|z_1(t)|^2)|A^{\frac{1}{4}}y(t)|dx, \qquad (3.51)$$

and

$$\int_\Omega |y_t(t)||A^{\frac{1}{4}}y(t)|dx \leq C_{\|y_t(t)\|}\|y(t)\|_{\mathcal{H}^{\frac{5}{4}}},$$

$$\begin{aligned}\int_\Omega |y_t(t)||y_1(t)|^2|A^{\frac{1}{4}}y(t)|dx &\leq C_{\|y_t(t)\|,K_\eta}\|y(t)\|_{\mathcal{H}^{\frac{5}{4}}},\\ \int_\Omega |y_t(t)||z_1(t)|^2|A^{\frac{5}{4}}y(t)|dx &\leq \int_\Omega (|\tilde{y}(t)|+\epsilon|y(t)|)|z_1(t)|^2|A^{\frac{1}{4}}y(t)|dx\\ &\leq C_{\|z_1(t)\|_{\mathcal{H}^1},\|y(t)\|_{\mathcal{H}^1}}\|y(t)\|_{\mathcal{H}^{\frac{5}{4}}} + \frac{\epsilon}{24}\|\tilde{y}(t)\|^2_{\mathcal{H}^{\frac{1}{4}}}\\ &\quad + C_\epsilon\|z_1(t)\|^2_{\mathcal{H}^1}\cdot\|y(t)\|^2_{\mathcal{H}^{\frac{5}{4}}}. \end{aligned} \tag{3.52}$$

Combining (3.51)-(3.52), we get

$$|\mathcal{I}_2| \leq \frac{\epsilon}{24}\mathcal{E}_y(t) + C\|z_1(t)\|^2_{\mathcal{H}^1}\cdot\|y(t)\|^2_{\mathcal{H}^{\frac{5}{4}}} + C_0, \tag{3.53}$$

where C_0 may depend on $\epsilon, K_\eta, \|y_t(t)\|, \|z_1(t)\|_{\mathcal{H}^1}, \|y(t)\|_{\mathcal{H}^1}$.
Using (3.22), we have

$$|\mathcal{I}_3| \leq \frac{\epsilon}{24}\|y(t)\|^2_{\mathcal{H}^{\frac{5}{4}}} + C_{\epsilon,\|z_t(t)\|}. \tag{3.54}$$

Finally, substituting the estimates (3.47), (3.50), (3.53) and (3.54) into (3.46) and recalling (3.43), we deduce

$$\frac{d}{dt}\mathbf{E}_y(t) + \iota(t)\mathbf{E}_y(t) \leq C_2,$$

where

$$\mathbf{E}_y(t) = \mathcal{E}_y(t) + 2\langle f(u)-f(z), A^{\frac{1}{4}}y(t)\rangle + 2\langle f_1(z), A^{\frac{1}{4}}y(t)\rangle,$$

$$\iota(t) = \frac{\epsilon}{4} - C\|z_1(t)\|^2_{\mathcal{H}^1}.$$

And C_1 depends on $\epsilon, \|z_t(t)\|$, and C_2 depends on ϵ, K_η, M. Then using the Gronwall's inequality and integrating over $[T_1, t]$, we obtain that

$$\mathbf{E}_y(t) \leq e^{-\int_{T_1}^t \iota(r)dr}\mathbf{E}_y(T_1) + C_2\int_{T_1}^t e^{-\int_s^t \iota(r)dr}ds. \tag{3.55}$$

Taking η in (3.44) small enough, we obtain

$$e^{-\int_{T_1}^t \iota(r)dr} \leq e^{-\frac{\epsilon}{8}(t-T_1)}e^{C_\eta C_1}, \tag{3.56}$$

and

$$\int_{T_1}^t e^{-\int_s^t \iota(r)dr}ds \leq e^{C_\eta C_1}\int_{T_1}^t e^{-\frac{\epsilon}{8}(t-s)}ds \leq 8\epsilon^{-1}e^{C_\eta C_1}.$$

Substituting (3.56) into (3.55), we deduce that, for all $t \geq T_1$,

$$\mathbf{E}_y(t) \leq e^{C_\eta C_1}\mathbf{E}_v(T_1) + 8\epsilon^{-1}C_2e^{C_\eta C_1},$$

where C_η, C_1, and C_2 are constants independent of the initial data, and $T_1 > 0$ is fixed. This implies that $\mathbf{E}_y(t)$ is uniformly bounded for all $t \geq T_1$. Moreover, since $B \subset \mathcal{H}$ is bounded and $t \mapsto \mathbf{E}_y(t)$ is continuous, there exists a constant C_{B,T_1} such that

$$\mathbf{E}_y(t) \leq C_{B,T_1}, \quad \forall t \in [0, T_1].$$

Combining the above, we conclude that $\mathbf{E}_y(t)$ is uniformly bounded on $[0, +\infty)$,

$$\mathbf{E}_y(t) \leq C = \max\left\{C_{B,T_1},\ e^{C_\eta C_1}\mathbf{E}_y(T_1) + 8\varepsilon^{-1}C_2 e^{C_\eta C_1}\right\}, \quad \forall t \geq 0.$$

In particular, this implies

$$\|\tilde{y}(t)\|^2_{\mathcal{H}^{1/4}} \leq C, \quad \text{and} \quad \|y(t)\|^2_{\mathcal{H}^{5/4}} \leq C.$$

Since $y_t(t) = \tilde{y}(t) - \varepsilon y(t)$, we estimate

$$\|y_t(t)\|_{\mathcal{H}^{-3/4}} \leq \|\tilde{y}(t)\|_{\mathcal{H}^{1/4}} + \varepsilon\|y(t)\|_{\mathcal{H}^{1/4}} \leq C + \varepsilon C',$$

because $\|y(t)\|_{\mathcal{H}^{1/4}}$ is bounded via interpolation between $\mathcal{H}$ and $\mathcal{H}^{5/4}$, both of which are bounded. Thus, we obtain the estimate

$$\|Y(t,0)u_0\|^2_{\mathcal{H}^{1/4}} = \|y(t)\|^2_{\mathcal{H}^{1/4}} + \|y_t(t)\|^2_{\mathcal{H}^{-3/4}} \leq N_{\|B\|_{\mathcal{H}}},$$

where $N_{\|B\|_{\mathcal{H}}}$ depends only on the bound of B in $\mathcal{H}$. □

The second a priori estimate.

Lemma 3.10. *Assume B_ζ is an bounded set in $\mathcal{H}_{\frac{1}{4}}$. Then there exists a positive constant $M_{\|B_\zeta\|_{\mathcal{H}_{\frac{1}{4}}}}$ which depends only on the $\mathcal{H}_{\frac{1}{4}}$-bounds of B_ζ such that*

$$\|U(t,0)u_0\|^2_{\mathcal{H}_{\frac{1}{4}}} \leq M_{\|B_\zeta\|_{\mathcal{H}^{\frac{1}{4}}}} \quad \textit{for all } t \geq 0 \textit{ and } u_0 = B_\zeta.$$

Proof. Let $u(t)$ be the solution of equation (1.1)-(1.2) with the initial value $u_0 = (u(0), u_t(0)) \in B_\zeta$. Taking the inner product of (1.1) with $A^{\frac{1}{4}}\tilde{u}$, we obtain

$$\begin{aligned}\frac{d}{dt}\mathcal{E}_y(t) + \epsilon\|u(t)\|^2_{\mathcal{H}_{\frac{5}{4}}} + |u_t|^{q-1} - \epsilon)\|\tilde{u}(t)\|^2_{\mathcal{H}_{\frac{1}{4}}} \\ + h_u(t) + 2\langle f(u), A^{\frac{1}{4}}\tilde{u}(t)\rangle = 2\langle 0, A^{\frac{1}{4}}\tilde{u}(t)\rangle,\end{aligned}$$

where

$$\begin{aligned}\mathcal{E}_u(t) &= \|\widetilde{u}(t)\|^2_{\mathcal{H}_{\frac{1}{4}}} + \|u(t)\|^2_{\mathcal{H}_{\frac{5}{4}}},\ \widetilde{u}(t) = u_t(t) + \epsilon u(t),\\ h_u(t) &= \epsilon\|u(t)\|^2_{\mathcal{H}_{\frac{5}{4}}} + (|u_t|^{q-1} - \epsilon)\|\widetilde{u}(t)\|^2_{\mathcal{H}_{\frac{1}{4}}}\\ &\quad + 2\epsilon^2\langle u(t), A^{\frac{1}{4}}\widetilde{u}(t)\rangle - 2\epsilon|u_t|^{q-1}\langle u(t), A^{\frac{1}{4}}\widetilde{u}(t)\rangle.\end{aligned}$$

Since the other terms can be handled with a similar argument as in the Lemma 3.6, we only deal with the nonlinear term. Indeed, we have

$$\langle f(u), A^{\frac{1}{4}}\tilde{u}(t)\rangle = \frac{d}{dt}\langle f(u), A^{\frac{1}{4}}u(t)\rangle - \langle f(u), A^{\frac{1}{4}}u_t(t)\rangle - f'(u)u_t(t), A^{\frac{1}{4}}u(t)\rangle.$$

And

$$\left|\left\langle f'(u)u_t(t), A^{\frac{1}{4}}u(t)\right\rangle\right| \le C\int_\Omega |u_t(t)|\left(1 + |y_1(t)|^2 + |z_1(t)|^2\right)\cdot|A^{\frac{1}{4}}u(t)|\,dx. \tag{3.57}$$

By Lemma 3.5 and using Hölder's and Sobolev-type inequality, we have

$$\int_\Omega |u_t(t)||A^{\frac{1}{4}}u(t)|\,dx \le C\|u_t(t)\|\,\|u(t)\|_{\mathcal{H}_{\frac{5}{4}}}.$$

$$\begin{aligned}\int_\Omega |u_t(t)||y_1(t)|^2|A^{\frac{1}{4}}u(t)|\,dx &\le \|u_t(t)\|\cdot\|y_1(t)\|^2_{L^8}\cdot\|A^{\frac{1}{4}}u(t)\|_{L^4}\\ &\le C_{\|u_t(t)\|,K_\eta}\|u(t)\|_{\mathcal{H}_{\frac{5}{4}}}.\end{aligned}$$

$$\begin{aligned}&\int_\Omega |u_t(t)||z_1(t)|^2|A^{\frac{1}{4}}u(t)|\,dx\\ &\qquad\le \int_\Omega \left(|\tilde{u}(t)| + \epsilon|u(t)|\right)|z_1(t)|^2|A^{\frac{1}{4}}u(t)|\,dx\\ &\qquad\le \epsilon\|z_1(t)\|^2_{L^6}\cdot\|u(t)\|_{L^{\frac{12}{5}}}\cdot\|A^{\frac{1}{4}}u(t)\|_{L^4}\\ &\qquad\quad + \|z_1(t)\|^2_{L^6}\cdot\|A^{\frac{1}{4}}u(t)\|_{L^4}\cdot\|\tilde{u}(t)\|_{L^{\frac{12}{5}}}\\ &\qquad\le C_{\|z_1(t)\|_{\mathcal{H}^1},\|u(t)\|_{\mathcal{H}^1}}\|u(t)\|_{\mathcal{H}^{\frac{5}{4}}}\\ &\qquad\quad + C_\epsilon\|z_1(t)\|^2_{\mathcal{H}_1}\cdot\|u(t)\|^2_{\mathcal{H}_{\frac{5}{4}}} + \frac{\epsilon}{8}\|\tilde{u}(t)\|^2_{\mathcal{H}_{\frac{1}{4}}}.\end{aligned} \tag{3.58}$$

$$\begin{aligned}\left|\langle f(u), A^{\frac{1}{4}}u(t)\rangle\right| &\le C\int_\Omega \left(1 + |u(t)|^3\right)A^{\frac{1}{4}}u(t)\,dx\\ &\le C_{\|u(t)\|_{\mathcal{H}^1}}\|u(t)\|^{\frac{5}{2}}_{\mathcal{H}_{\frac{1}{4}}}.\end{aligned} \tag{3.59}$$

We combine (3.57)-(3.58) and use Young's inequality, we get

$$|\langle f'(u)u_t(t), A^{\frac{1}{4}}u(t)\rangle| \leq \frac{\epsilon}{8}\|u_t(t)\|^2_{\mathcal{H}_{\frac{5}{4}}} + C_{\|z_1(t)\|,\epsilon}\|z_1(t)\|^2_{\mathcal{H}^1}\cdot\|u(t)\|^2_{\mathcal{H}_{\frac{5}{2}}} + C_{\epsilon,K_\eta,M}.$$

Therefore, by choosing ϵ small enough and applying (3.59), we obtain

$$\frac{d}{dt}\mathcal{E}_u(t) + \zeta(t)\mathcal{E}_u(t) \leq C_{\epsilon,K_\eta,M}, \tag{3.60}$$

where

$$\mathcal{E}_u(t) = \mathcal{E}_u(t) + 2\langle f(u), A^{\frac{1}{4}}u(t)\rangle,\ \zeta(t) = \frac{\epsilon}{4} - C_{\epsilon,\|z_1(t)\|}\|z_1(t)\|^2_{\mathcal{H}_1}.$$

Multiply both sides of (3.60) by $e^{\int_0^t \zeta(r)\,dr}$

$$\frac{d}{dt}\left(e^{\int_0^t \zeta(r)\,dr}\mathcal{E}_u(t)\right) \leq C_{\epsilon,K_\eta,M}e^{\int_0^t \zeta(r)\,dr}.$$

Integrating from 0 to t, we obtain

$$\mathcal{E}_u(t) \leq e^{-\int_0^t \zeta(r)\,dr}\mathcal{E}_u(0) + C_{\epsilon,K_\eta,M}\int_0^t e^{-\int_s^t \zeta(r)\,dr}\,ds.$$

By $u_0 = (u(0), u_t(0))$ belongs to $\mathcal{H}^{\frac{1}{4}}$, we then complete the proof by the same argument as in Lemma 3.5. □

As a consequence of Lemma 3.6 and 3.7, we can state the following result.

Corollary 3.11. *Assume B_ζ is an arbitrary bounded set in $\mathcal{H}^{\frac{1}{4}}$. Then there exists a positive constant $M_{\|B_\zeta\|_{\mathcal{H}^{\frac{1}{4}}}}$ which depends only on the $\mathcal{H}^{\frac{1}{4}}$-bounds of B_ζ such that*

$$\|Z(t,0)u_0\|^2_{\mathcal{H}_{\frac{1}{4}}} \leq M_{\|B_\zeta\|_{\mathcal{H}^{\frac{1}{4}}}} \quad \text{for all} \quad t \geq 0 \text{ and } u_0 = (u(0), u_t(0)) \in B_\zeta.$$

Proof. Let $u(t) = z(t) + y(t)$ be the decomposition of the full solution associated to the system, where $Z(t,0)u_0 = z(t)$ and $Y(t,0)u_0 = y(t)$. From Lemma 3.6, we know that for any bounded set $B \subset \mathcal{H}$, there exists a constant $N_{\|B\|_{\mathcal{H}}} > 0$ such that

$$\|Y(t,0)u_0\|^2_{\mathcal{H}^{\frac{1}{4}}} \leq N_{\|B\|_{\mathcal{H}}}, \quad \forall t \geq 0.$$

From Lemma 3.7, for a bounded set $B_\zeta \subset \mathcal{H}^{\frac{1}{4}}$, there exists a constant $M_{\|B_\zeta\|_{\mathcal{H}^{\frac{1}{4}}}} > 0$ such that

$$\|U(t,0)u_0\|^2_{\mathcal{H}^{\frac{1}{4}}} \leq M_{\|B_\zeta\|_{\mathcal{H}^{\frac{1}{4}}}}, \quad \forall t \geq 0.$$

Since the total solution satisfies $u(t) = Z(t,0)u_0 + Y(t,0)u_0$, we can write

$$Z(t,0)u_0 = u(t) - Y(t,0)u_0.$$

Therefore,

$$\|Z(t,0)u_0\|_{\mathcal{H}^{\frac{1}{4}}} \leq \|u(t)\|_{\mathcal{H}^{\frac{1}{4}}} + \|Y(t,0)u_0\|_{\mathcal{H}^{\frac{1}{4}}}.$$

Applying the inequality $(a+b)^2 \leq 2a^2 + 2b^2$, we obtain

$$\|Z(t,0)u_0\|^2_{\mathcal{H}^{\frac{1}{4}}} \leq 2\|u(t)\|^2_{\mathcal{H}^{\frac{1}{4}}} + 2\|Y(t,0)u_0\|^2_{\mathcal{H}^{\frac{1}{4}}}.$$

Using the uniform bounds on $\|u(t)\|^2_{\mathcal{H}^{1/4}}$ from Lemma 3.7 and $\|Y(t,0)u_0\|^2_{\mathcal{H}^{1/4}}$ from Lemma 3.6, we conclude that there exists a constant $M_{\|B_\zeta\|_{\mathcal{H}^{1/4}}} > 0$ such that

$$\|Z(t,0)u_0\|^2_{\mathcal{H}^{\frac{1}{4}}} \leq M_{\|B_\zeta\|_{\mathcal{H}^{\frac{1}{4}}}}, \quad \forall t \geq 0. \qquad \square$$

Lemma 3.12. *For all bounded set $B_\zeta \subset \mathcal{H}^{\frac{1}{4}}$, there exists a positive constant $M_{\|B_\zeta\|_{\mathcal{H}^{\frac{1}{4}}}}$ which depends only on the $\mathcal{H}^{\frac{1}{4}}$-bounds of B_ζ such that*

$$\|Y(t,0)u_0\|^2_{\mathcal{H}^{\frac{3}{4}}} \leq M_{\|B_\zeta\|_{\mathcal{H}^{\frac{1}{4}}}} \quad \text{for all} \quad t \geq 0 \text{ and } u_0 = (u(0), u_t(0)) \in B_\zeta.$$

Proof. Taking the inner product of (3.25) with $A^{\frac{3}{4}}\tilde{y}$, we obtain

$$\begin{aligned} &\frac{d}{dt}\mathcal{E}_\nu(t) + \epsilon\|y(t)\|^2_{H^{\frac{7}{4}}} + (|u_t|^{q-1} - \epsilon)\|\tilde{y}(t)\|^2_{H^{\frac{3}{4}}} \\ &+h_\nu(t) + 2\langle f(u) - f_0(z), A^{\frac{3}{4}}\tilde{y}(t)\rangle = 2\langle 0, A^{\frac{3}{4}}u(t)\rangle, \end{aligned} \tag{3.61}$$

where

$$\begin{aligned} \mathcal{E}_\nu(t) &= \|\tilde{y}(t)\|^2_{\mathcal{H}^{\frac{3}{4}}} + \|y(t)\|^2_{\mathcal{H}^{\frac{7}{4}}}, \quad \tilde{y}(t) = y(t) + \epsilon y(t), \\ h_\nu(t) &= \epsilon\|y(t)\|^2_{\mathcal{H}^{\frac{7}{4}}} + (|u_t|^{q-1}) - \epsilon)\|\tilde{y}(t)\|^2_{\mathcal{H}^{\frac{3}{4}}} \\ &\quad + 2\epsilon^2\langle y(t), A^{\frac{3}{4}}\tilde{y}(t)\rangle - 2\epsilon|u_t|^{q-1}\langle y(t), A^{\frac{3}{4}}\tilde{y}(t)\rangle. \end{aligned}$$

The nonlinear term gives

$$\begin{aligned}\langle f(u)-f_0(z),A^{\frac{3}{4}}\tilde{y}(t)\rangle &= \frac{d}{dt}\langle f(u)-f_0(z),A^{\frac{3}{4}}y(t)\rangle\\ &\quad+\epsilon\langle f(u)-f_0(z),A^{\frac{3}{4}}y(t)\rangle\\ &\quad-\langle f'(u)u_t(t)-f_0'(z)z_t(t),A^{\frac{3}{4}}y(t)\rangle. \qquad (3.62)\end{aligned}$$

And

$$\left|\langle f'(u)u_t(t),A^{\frac{3}{4}}v(t)\rangle\right| \le C\int_\Omega |u_t(t)|(1+|u(t)|^2)|A^{\frac{3}{4}}v(t)|dx.$$

By Lemma 3.7, we can deduce

$$\begin{aligned}\int_\Omega |u_t(t)||u(t)|^2|A^{\frac{3}{4}}y(t)|dx &\le \|u_t(t)\|_{L^{\frac{12}{5}}}\cdot\|u(t)\|^2_{L^{12}}\cdot\|A^{\frac{3}{4}}y(t)\|_{L^{\frac{12}{5}}}\\ &\le C_{\|u_t(t)\|_{\mathcal{H}^{\frac{1}{4}}},\|u(t)\|_{\mathcal{H}^{\frac{5}{4}}}}\|y_t(t)\|_{\mathcal{H}^{\frac{7}{4}}}.\end{aligned}$$

Similarly, thanks to Corollary 3.11, we get

$$|\langle f_0(z)z_t(t),A^{\frac{3}{4}}y(t)\rangle| \le C_{|\Omega|,\|z_t(t)\|_{\mathcal{H}^{\frac{1}{4}}},\|z(t)\|_{\mathcal{H}^{\frac{5}{4}}}}\|y_t(t)\|_{\mathcal{H}^{\frac{7}{4}}}. \qquad (3.63)$$

Substituting (3.62) and (3.63) into (3.61), and using Young's inequality, we obtain $\frac{d}{dt}\mathcal{E}_\nu(t)+\epsilon\mathcal{E}_\nu(t)\le C(t)$, where $C(t)\le C_{\|B_\zeta\|_{\mathcal{H}^{\frac{1}{4}}}}$ is uniformly bounded for $u_0\in B_\zeta\subset\mathcal{H}^{\frac{1}{4}}$, we obtain

$$\mathcal{E}_\nu(t)\le\mathcal{E}_\nu(0)e^{-\epsilon t}+\int_0^t C(s)e^{-\epsilon(t-s)}ds.$$

Since $\mathcal{E}_\nu(0)\le C_{\|B_\zeta\|_{\mathcal{H}^{\frac{1}{4}}}}$, we conclude

$$\mathcal{E}_\nu(t)\le M_{\|B_\zeta\|_{\mathcal{H}^{\frac{1}{4}}}},\quad \text{for all } t\ge 0.$$

Hence,

$$\|Y(t,0)u_0\|^2_{\mathcal{H}^{\frac{3}{4}}}\le M_{\|B_\zeta\|_{\mathcal{H}^{\frac{1}{4}}}},\quad \text{for all } t\ge 0,\ u_0\in B_\zeta. \qquad \square$$

Lemma 3.13. *For all bounded set $B_\nu\subset\mathcal{H}^{\frac{3}{4}}$, there exists a positive constant $M_{\|B_\nu\|_{\mathcal{H}^{\frac{3}{4}}}}$ which depends only on the $\mathcal{H}^{\frac{3}{4}}$-bounds of B_ν such that*

$$\|U(t,0)u_0\|^2_{\mathcal{H}^{\frac{3}{4}}}\le M_{\|B_\nu\|_{\mathcal{H}^{\frac{3}{4}}}}\quad \forall t\ge 0 \text{ and } u_0=(u(0),u(0))\in B_\nu. \qquad (3.64)$$

Proof. We consider the solution $u(t)$ of the damped nonlinear wave equation with initial data $u_0\in B_\nu\subset\mathcal{H}^{\frac{3}{4}}$. Let us define the higher-order energy

functional

$$\mathcal{E}(t) = \|u_t(t)\|^2_{H^{\frac{3}{4}}} + \|u(t)\|^2_{H^{\frac{7}{4}}}.$$

Taking the time derivative and using the equation, we obtain

$$\frac{d}{dt}\mathcal{E}(t) + c_1\|u_t(t)\|^{q+1}_{H^{\frac{3}{4}}} + c_2\|u(t)\|^2_{H^{\frac{7}{4}}} \leq \left|\langle f(u), A^{\frac{3}{4}}u_t(t)\rangle\right|.$$

Using the growth condition on $f(u)$ and Sobolev embeddings, we estimate

$$\left|\langle f(u), A^{\frac{3}{4}}u_t\rangle\right| \leq C\int_\Omega |u_t|(1+|u|^p)|A^{\frac{3}{4}}u|\,dx \leq C_{\|u\|_{H^{\frac{7}{4}}}, \|u_t\|_{H^{\frac{3}{4}}}}.$$

Thus,

$$\frac{d}{dt}\mathcal{E}(t) + \epsilon\mathcal{E}(t) \leq C(t),$$

where $C(t)$ depends only on norms of the solution, which are uniformly bounded for $u_0 \in B_\nu$. Applying Gronwall's lemma, we get

$$\mathcal{E}(t) \leq \mathcal{E}(0)e^{-\epsilon t} + \int_0^t C(s)e^{-\epsilon(t-s)}ds \leq M_{\|B_\nu\|_{\mathcal{H}^{\frac{3}{4}}}}.$$

Hence,

$$\|U(t,0)u_0\|^2_{\mathcal{H}^{\frac{3}{4}}} \leq M_{\|B_\nu\|_{\mathcal{H}^{\frac{3}{4}}}}, \quad \text{for all } t \geq 0,\ u_0 \in B_\nu. \qquad \square$$

As an immediate consequence of the Lemma 3.9 and 3.10, we have

Corollary 3.14. *For all bounded set $B_\nu \subset \mathcal{H}^{\frac{3}{4}}$, there exists a positive constant $M_{\|B_\nu\|_{\mathcal{H}^{\frac{3}{4}}}}$ which depends only on the $\mathcal{H}^{\frac{3}{4}}$-bounds of B_ν such that*

$$\|Z(t,0)u_0\|^2_{\mathcal{H}^{\frac{3}{4}}} \leq M_{\|B_\nu\|_{\mathcal{H}^{\frac{3}{4}}}} \quad \textit{for all } t \geq 0 \textit{ and } u_0 = (u(0), u(0)) \in B_\nu.$$

Proof. From the decomposition of the evolution operator, we have

$$U(t,0)u_0 = Z(t,0)u_0 + Y(t,0)u_0.$$

Taking norms in $\mathcal{H}^{\frac{3}{4}}$, we apply the triangle inequality

$$\|Z(t,0)u_0\|_{\mathcal{H}^{\frac{3}{4}}} \leq \|U(t,0)u_0\|_{\mathcal{H}^{\frac{3}{4}}} + \|Y(t,0)u_0\|_{\mathcal{H}^{\frac{3}{4}}}.$$

Applying the inequality $(a+b)^2 \leq 2a^2 + 2b^2$, we obtain

$$\|Z(t,0)u_0\|^2_{\mathcal{H}^{\frac{3}{4}}} \leq 2\|U(t,0)u_0\|^2_{\mathcal{H}^{\frac{3}{4}}} + 2\|Y(t,0)u_0\|^2_{\mathcal{H}^{\frac{3}{4}}}.$$

Now, by Lemma 3.10 and Lemma 3.9 , there exist constants depending only on $\|B_\nu\|_{\mathcal{H}^{\frac{3}{4}}}$ such that

$$\|U(t,0)u_0\|^2_{\mathcal{H}^{\frac{3}{4}}} \leq M_1, \quad \|Y(t,0)u_0\|^2_{\mathcal{H}^{\frac{3}{4}}} \leq M_2.$$

Therefore,

$$\|Z(t,0)u_0\|^2_{\mathcal{H}^{\frac{3}{4}}} \leq 2M_1 + 2M_2 = M_{\|B_\nu\|_{\mathcal{H}^{\frac{3}{4}}}}. \qquad \square$$

Lemma 3.15. *Assume the initial data set B_ν is bounded in $\mathcal{H}^{\frac{3}{4}}$ and $\frac{3}{4} \leq \varrho \leq 1$, then the solution of (3.25) $y(t)$ satisfies that,*

$$\|Y(t,0)u_0\|^2_{\mathcal{H}^\varrho} \leq N_{\|B_\nu\|_{\mathcal{H}^{\frac{3}{4}}}} \quad \text{for all } t \geq 0 \text{ and } u_0 = (u(0), u(0)) \in B_\nu.$$

Proof. We test equation (4.38) with $A^\varrho \tilde{y}(t)$, where $\tilde{y}(t) = y_t(t) + \epsilon y(t)$, for a small parameter $\epsilon > 0$. By multiplying the equation by this test function and integrating by parts, we obtain

$$\frac{d}{dt}\mathcal{E}_\varrho(t) + \epsilon\mathcal{E}_\varrho(t) \leq 2\langle f'(u)u_t(t), A^\varrho y(t)\rangle + 2\langle f_0'(z)z_t(t), A^\varrho y(t)\rangle, \qquad (3.65)$$

where

$$\mathcal{E}_\varrho(t) = \|\tilde{y}(t)\|^2_{\mathcal{H}^\varrho} + \|y(t)\|^2_{\mathcal{H}^{\varrho+1}} + 2\langle f(u) - f_0(z), A^\varrho y(t)\rangle. \tilde{y}(t) = y(t) + \epsilon y(t).$$

We now estimate the nonlinear terms on the right-hand side of (3.65). Using Lemma 3.10 and the continuous embedding $\mathcal{H}^{\frac{7}{4}} \hookrightarrow L^\infty$, we apply Hölder's inequality to get

$$\begin{aligned} 2\left|\langle f'(u)u_t(t), A^\varrho y(t)\rangle\right| &\leq C_{\|u(t)\|_{\mathcal{H}^{\frac{7}{4}}}} \|u_t(t)\|_{L^2}\|y(t)\|_{\mathcal{H}^{\varrho+1}}, \\ 2\left|\langle f_0'(z)z_t(t), A^\varrho y(t)\rangle\right| &\leq C_{\|z(t)\|_{\mathcal{H}^{\frac{7}{4}}}} \|z_t(t)\|_{L^2}\|y(t)\|_{\mathcal{H}^{\varrho+1}}. \end{aligned} \qquad (3.66)$$

Applying Young's inequality , we can absorb the resulting terms into the left-hand side, leading to

$$\frac{d}{dt}\mathcal{E}_\varrho(t) + \frac{\epsilon}{2}\mathcal{E}_\varrho(t) \leq C\left(\|u(t)\|^2_{\mathcal{H}^{\frac{7}{4}}} + \|z(t)\|^2_{\mathcal{H}^{\frac{7}{4}}} + \|u_t(t)\|^2_{L^2} + \|z_t(t)\|^2_{L^2}\right).$$

From previous lemmas (3.7), (3.9) and (3.10), the quantities on the right-hand side are uniformly bounded for $u_0 \in B_\nu \subset \mathcal{H}^{\frac{3}{4}}$. Therefore, there exists a constant $C_0 > 0$ such that:

$$\frac{d}{dt}\mathcal{E}_\varrho(t) + \frac{\epsilon}{2}\mathcal{E}_\varrho(t) \leq C_0.$$

Applying Gronwall's inequality gives

$$\mathcal{E}_\varrho(t) \leq \mathcal{E}_\varrho(0)e^{-\frac{\epsilon}{2}t} + \frac{2C_0}{\epsilon} \leq C_{\|B_\nu\|_{\mathcal{H}^{\frac{3}{4}}}}, \quad \forall t \geq 0.$$

Finally, since $\mathcal{E}_\varrho(t)$ controls $\|Y(t,0)u_0\|^2_{\mathcal{H}^\varrho}$, we conclude

$$\|Y(t,0)u_0\|^2_{\mathcal{H}^\varrho} \leq C_{\|B_\nu\|_{\mathcal{H}^{\frac{3}{4}}}}, \quad \forall t \geq 0. \qquad \square$$

Lemma 3.16. *Let $\frac{3}{4} \leq \varrho \leq 1$, assume the initial data set B_ϱ is bounded , then there exists a positive constant $J_{\|B_\varrho\|_{\mathcal{H}^\varrho}}$ such that*

$$\|U(t,0)u_0\|^2_{\mathcal{H}^\varrho} \leq J_{\|B_\varrho\|_{\mathcal{H}^\varrho}} \quad \textit{for all } t \geq 0 \textit{ and } u_0 = (u(0), u_t(0)) \in B_\varrho.$$

Proof. Consider the solution $u(t)$ corresponding to the initial data $u_0 = (u(0), u_t(0)) \in B_\varrho$, with norm

$$\|u_0\|^2_{\mathcal{H}^\varrho} = \|u(0)\|^2_{\mathcal{H}^{\varrho+1}} + \|u_t(0)\|^2_{\mathcal{H}^\varrho}.$$

For a sufficiently small $\epsilon > 0$, define $\tilde{u}(t) = u_t(t) + \epsilon u(t)$. Define the modified energy

$$\mathcal{E}_\varrho(t) = \|\tilde{u}(t)\|^2_{\mathcal{H}^\varrho} + \|u(t)\|^2_{\mathcal{H}^{\varrho+1}} + 2\langle F(u(t)), A^\varrho u(t)\rangle.$$

Under the assumptions on f and coercivity, there exist constants $c_1, c_2 > 0$ such that

$$c_1\|U(t,0)u_0\|^2_{\mathcal{H}^\varrho} \leq \mathcal{E}_\varrho(t) \leq c_2\|U(t,0)u_0\|^2_{\mathcal{H}^\varrho}.$$

Differentiating $\mathcal{E}_\varrho(t)$ and using the equation satisfied by u, we obtain an inequality of the form

$$\frac{d}{dt}\mathcal{E}_\varrho(t) + \epsilon\mathcal{E}_\varrho(t) \leq 2\langle f'(u)u_t, A^\varrho u\rangle + 2\langle g(u_t), A^\varrho u\rangle,$$

where $g(u_t)$ represents damping term. Thanks to Sobolev embeddings and the boundedness of $\|u(t)\|_{\mathcal{H}^{\varrho+1}}$ in a sufficiently high order space, we have

$$|\langle f'(u)u_t, A^\varrho u\rangle| \leq C_{\|u\|_{\mathcal{H}^{\varrho+1}}}\|u_t\|_{\mathcal{H}^\varrho}\|u\|_{\mathcal{H}^{\varrho+1}},$$

and similarly,

$$|\langle g(u_t), A^\varrho u\rangle| \leq C\|g(u_t)\|_{\mathcal{H}^\varrho}\|u\|_{\mathcal{H}^{\varrho+1}}.$$

For any $\delta > 0$, applying Young's inequality, we get

$$2|\langle f'(u)u_t, A^\varrho u\rangle| \leq \frac{C^2_{\|u\|_{\mathcal{H}^{\varrho+1}}}}{2\delta}\|u_t\|^2_{\mathcal{H}^\varrho} + \frac{\delta}{2}\|u\|^2_{\mathcal{H}^{\varrho+1}},$$

and similarly for the terms involving $g(u_t)$. Choosing $\delta > 0$ small enough, we obtain

$$\frac{d}{dt}\mathcal{E}_\varrho(t) + \frac{\epsilon}{2}\mathcal{E}_\varrho(t) \leq C_0,$$

where the constant $C_0 > 0$ depends only on the bound of B_ϱ. By integrating and applying Gronwall's inequality,

$$\mathcal{E}_\varrho(t) \leq \mathcal{E}_\varrho(0)e^{-\frac{\epsilon}{2}t} + \frac{2C_0}{\epsilon} \leq J_{\|B_\varrho\|_{\mathcal{H}^\varrho}},$$

with $J_{\|B_\varrho\|_{\mathcal{H}^\varrho}} = \mathcal{E}_\varrho(0) + \frac{2C_0}{\epsilon}$. By equivalence of norms, this implies

$$\|U(t,0)u_0\|^2_{\mathcal{H}^\varrho} \leq \frac{1}{c_1}\mathcal{E}_\varrho(t) \leq \frac{J_{\|B_\varrho\|_{\mathcal{H}^\varrho}}}{c_1}. \qquad \square$$

Lemma 3.17. [10] *Let $(\mathcal{M}, d)$ be an abstract metric space and let $U(t,0)$ be a Lipschitz continuous dynamical process in $\mathcal{M}$, i.e*

$$d(U(t,0)m_1, U(t,0)m_2) \leq Ce^{Kt}d(m_1, m_2),$$

for appropriate constants C and K which are independent of m_i, t. We further assume that there exist three subsets $M_1, M_2, M_3 \subset \mathcal{M}$ such that

$$\begin{cases} \operatorname{dist}_{\mathcal{M}}(U(t,0)M_1, M_2) \leq C_1e^{-\alpha_1 t}, \\ \operatorname{dist}_{\mathcal{M}}(U(t,0)M_2, M_3) \leq C_2e^{-\alpha_2 t}. \end{cases}$$

Then

$$\operatorname{dist}_{\mathcal{M}}(U(t,0)M_1, M_3) \leq C'e^{-\alpha' t},$$

where $C' = CC_1 + C_2$, $\alpha' = \frac{\alpha_1\alpha_2}{K+\alpha_1+\alpha_2}$.

Proof of Theorem 3.4.

Proof. We are now ready to prove our main results of this section, and the proof based on the a priori estimates and attraction transitivity result of [2]. From Theorem 3.1, we have

$$\operatorname{dist}_{\mathbb{E}}(S(t)\mathcal{B}, \mathfrak{B}_0) \leq \mathcal{K}_0e^{\alpha_1 T_{\mathcal{B}}}e^{-\alpha_1 t}, \tag{3.67}$$

where $\alpha_1 > 0$ and $\mathfrak{B}_0 = \widehat{\mathcal{B}} \times \mathbb{T}^n \subset \mathbb{E}$, and $\mathfrak{B}_\zeta = \mathcal{M}_\zeta \times \mathbb{T}^n \subset \mathbb{E}^1$, with $\mathcal{M}_\zeta \subset \mathcal{H}^{1/4}$ a bounded set. Let $u(t)$ denote the solution corresponding to initial data $u_0 \in \widehat{\mathcal{B}} \subset \mathcal{H}$. Using Lemma 3.2.6, the solution admits a decomposition $Z(t,0)u_0 = v(t) + w(t)$, where

- $v(t) \in \mathcal{H}$ decays exponentially: $\|v(t)\|_{\mathcal{H}} \leq \mathcal{Q}_1(\|u_0\|_{\mathcal{H}})e^{-\alpha_3 t}$,
- $w(t) \in \mathcal{H}^{1/4}$ is uniformly bounded: $\sup_{t\geq 0}\|w(t)\|_{\mathcal{H}^{1/4}} \leq C(\|u_0\|_{\mathcal{H}})$.

Let us define the following set of regular components: $\mathcal{M}_\zeta = \{w(t) \mid u_0 \in \widehat{\mathcal{B}},\ t \geq 0\} \subset \mathcal{H}^{1/4}$, which is bounded in $\mathcal{H}^{1/4}$ due to the uniform bound above. Then define $\mathfrak{B}_\zeta = \mathcal{M}_\zeta \times \mathbb{T}^n \subset \mathbb{E}^1$. We now estimate the distance

between the full trajectories and the regular set:

$$\operatorname{dist}_{\mathbb{E}}(S(t)\mathfrak{B}_0, \mathfrak{B}_\zeta) = \sup_{u_0 \in \widehat{\mathcal{B}}} \inf_{z \in \mathcal{M}_\zeta} \|v(t) + w(t) - z\|_{\mathcal{H}}.$$

Since $w(t) \in \mathcal{M}_\zeta$, we can choose $z = w(t)$, yielding

$$\|v(t) + w(t) - w(t)\|_{\mathcal{H}} = \|v(t)\|_{\mathcal{H}} \leq \mathcal{Q}_1(\|u_0\|_{\mathcal{H}})e^{-\alpha_3 t}.$$

Therefore, we conclude

$$\operatorname{dist}_{\mathbb{E}}(S(t)\mathfrak{B}_\zeta, \mathfrak{B}_y) \leq \mathcal{Q}_1(\|\mathcal{M}_v\|_{\mathcal{H}})e^{-\alpha_3 t},$$

where $\mathfrak{B}_v = \mathcal{M}_v \times \mathbb{T}^n$, with $\mathcal{M}_v \subset \mathcal{H}^{3/4}$, and $\mathfrak{B}_\varrho = \mathcal{M}_\varrho \times \mathbb{T}^n$, with $\mathcal{M}_\varrho \subset \mathcal{H}^\varrho$, $\varrho \in \left[\frac{3}{4}, 1\right]$. We proceed by decomposing the evolution of the semigroup $S(t)\mathfrak{B}_v$ as

$$S(t)\mathfrak{B}_v = v(t) + y(t),$$

where

- $v(t)$ decays exponentially in $\mathcal{H}$ as a consequence of Lemma 3.2.3

$$\|v(t)\|_{\mathcal{H}} \leq \mathcal{Q}_1\left(\|u_0\|_{\mathcal{H}}\right) e^{-k_0 t},$$

- $y(t)$ belongs to a bounded subset of $\mathcal{H}^\varrho$ due to Lemma 3.2.12

$$\|y(t)\|_{\mathcal{H}^\varrho} \leq C, \quad \forall t \geq 0.$$

Now define the attracting set $\mathfrak{B}_\varrho = \mathcal{M}_\varrho \times \mathbb{T}^n$, where $\mathcal{M}_\varrho = \{y(t) \mid u_0 \in \mathcal{M}_v, t \geq 0\}$, which is bounded in $\mathcal{H}^\varrho$ by Lemma 3.2.12. Since $y(t) \in \mathcal{M}_\varrho$, for each $u_0 \in \mathcal{M}_v$ we choose $z = y(t) \in \mathfrak{B}_\varrho$, and thus

$$\operatorname{dist}_{\mathbb{E}}(S(t)u_0, \mathfrak{B}_\varrho) \leq \|v(t)\|_{\mathcal{H}} \leq \mathcal{Q}_1(\|u_0\|_{\mathcal{H}})e^{-\alpha_4 t}.$$

Taking $u_0 \in \mathcal{M}_v$, we obtain

$$\operatorname{dist}_{\mathbb{E}}(S(t)\mathfrak{B}_v, \mathfrak{B}_\varrho) \leq \mathcal{Q}_1(\|\mathcal{M}_\varrho\|_{\mathcal{H}})e^{-\alpha_4 t}, \tag{3.68}$$

where $\mathfrak{B}_\varrho = \mathcal{M}_\varrho \times \mathbb{T}^n$, $\mathcal{M}_\varrho$ is bounded in $\mathcal{H}^\varrho$ and α_4 depends on the $\mathcal{H}$-bounds of $\mathcal{M}_\varrho$. Combining now (3.63)-(3.64), recalling Lemma 3.3, and applying Lemma 3.13, we deduce

$$\operatorname{dist}_{\mathbb{E}}(S(t)\mathcal{B}, \mathcal{M} \times \mathbb{T}^n) \leq \mathcal{Q}(\|\mathcal{B}\|_{\mathbb{E}})e^{-\alpha t}, \quad \text{for all } t \geq 0,$$

where $\mathcal{M} = \mathcal{M}_\varrho$ and α only depends on α_i $(i = 0, 1, 2, 3, 4)$. □

Corollary 3.18. *Let the assumptions of Theorem 3.2.1 hold. Then the solution semigroup* $\{S(t)\}_{t\geq 0}$ *associated with problem* (1.1) – (1.2) *possesses*

a global attractor $\mathcal{A}$ in the extended phase space $\mathbb{E}$. Moreover, the attractor $\mathcal{A}$ is bounded.

Proof. The exponential attraction estimate above ensures asymptotic compactness of the semigroup. Indeed, given any bounded sequence $\{u_n\} \subset \mathbb{E}$ and sequence of times $t_n \to \infty$, the trajectories $S(t_n)u_n$ are eventually exponentially close to the compact set $\mathcal{M} \times \mathbb{R}^n \subset \mathbb{E}^1$, which is compactly embedded into $\mathbb{E}$. Therefore, up to a subsequence, $S(t_n)u_n$ converges in $\mathbb{E}$. Since the semigroup $\{S(t)\}_{t\geq 0}$ is continuous, has a bounded absorbing set, and is asymptotically compact in $\mathbb{E}$, it follows by standard attractor theory that there exists a unique global attractor $\mathcal{A} \subset \mathbb{E}$, which is compact, invariant, and attracts all bounded subsets of $\mathbb{E}$. Since $\mathcal{A} \subset \omega(\mathcal{M} \times \mathbb{R}^n)$, and this set is bounded in $\mathbb{E}^1$, it follows that $\mathcal{A} \subset \mathbb{E}^1$ and is therefore more regular than the original phase space. We conclude that the semigroup $\{S(t)\}_{t\geq 0}$ possesses a global attractor $\mathcal{A} \subset \mathbb{E}$, which is bounded in $\mathbb{E}^1$. □

4. Uniform exponential attractor

Theorem 4.1. *Let the assumptions of Theorem 3.2.1 be satisfied, and assume further that $f \in C^2(\mathbb{R})$. For all $0 < \theta < \frac{1}{2}$, we define $t^* = \frac{1}{\lambda} \ln \frac{1}{\theta}$, where $\lambda = \frac{\alpha}{2}$. Then, for all $\kappa \in (0, \frac{1}{2} - \theta)$, the semigroup $\{S(t)\}_{t\geq 0}$ induced by solutions of $(1.1)-(1.2)$ on $\mathbb{E}$ has a time-dependent exponential attractor $\mathbf{M} = \{\mathbf{M}^\kappa(s) \mid s \in \mathbb{R}\}$ in $\mathbb{E}^1$ which satisfies the following properties:*
i. There exists $\varpi > 0$ such that $\mathbf{M}^\kappa(s) = \mathbf{M}^\kappa(\varpi + s), \quad \forall s \in \mathbb{R}$.
ii. The family is positive semi-variant, that is

$$S(t)\mathbf{M}^\kappa(s) \subset \mathbf{M}^\kappa(t + s), \quad \forall t \geq 0, \quad \forall s \in \mathbb{R}.$$

iii. There exists a positive constant β such that, for every bounded subset $\mathcal{B}$ of $\mathbb{E}$,

$$\sup_{s\in[0,\varpi]} \operatorname{dist}_{\mathbb{E}}(S(t)\mathcal{B}, \mathbf{M}^\kappa(s)) \leq \mathcal{Q}(\|\mathcal{B}\|_{\mathbb{E}})e^{-\beta t}, \quad \forall t \geq 0;$$

iv. $\mathbf{M}^\kappa(s)$ is compact in $\mathbb{E}^1$ and its fractal dimension in $\mathbb{E}^1$ is uniformly bounded,

$$\sup_{s\in\mathbb{R}} \dim_{\mathfrak{F}}^{\mathbb{E}^1}(\mathbf{M}^\kappa(s)) \leq \log \frac{1}{2(\kappa + \theta)} \left(N_{\frac{k}{l^*}}^{nz}(B_1^z(0))\right), \quad \text{for all } s \in \mathbb{R},$$

where Z, nZ and l^* will be difined.

Proof. Using Lemma 3.13, there exists a bounded set $\mathfrak{B}_{\mathcal{S}} \subset \mathbb{E}^1$, such that for all bounded set $\mathcal{B} \subset \mathbb{E}^1$, the following estimates hold

$$\operatorname{dist}_{\mathbb{E}}(S(t)\mathcal{B}, \mathfrak{B}_{\mathcal{S}}) \leq \mathcal{K}_1 e^{\alpha_5 T_B} e^{-\alpha_5 t},$$

where $\alpha_5 > 0$ and $\mathcal{K}_1 = \sup\{\|S(t)\mathcal{B}\|_{\mathbb{E}^1}, 0 \leq t \leq T_B\} < \infty$. Assume that

$$S(t)\mathfrak{B}_{\mathcal{S}} \subset \mathfrak{B}_{\mathcal{S}}. \tag{3.69}$$

Now, we show the Lipschitz continuity and quasi-stability of the semigroup $S(t)$ in the extended space $\mathbb{E}^1 = \mathcal{H}^1 \times \mathbb{T}^n$. Obviously, $w = u^1 - u^2$ solves

$$w_{tt} + \ell_1(t)w_t + \ell_2(t)\big(\|u_t^1\|^2 - \|u_t^2\|^2\big)(u_t^1 + u_t^2) + Aw + f(u^1) - f(u^2) = 0,$$

where

$$\begin{aligned} \ell_1(t) &= \frac{1}{2}\Big(|u_t^1|^{q-1}) + |u_t^2|^{q-1})\Big) \geq |u_0|^{q-1}, \\ \ell_2(t) &= \frac{1}{2}\int_0^1 \Big(s|u_t^1|^{q-1} + (1-s)|u_t^2|^{q-1}\Big)' ds \geq 0. \end{aligned} \tag{3.70}$$

Using the multiplier $A\tilde{w} = A(w_t + \epsilon w)$ and using (3.70), we have

$$\begin{aligned} \frac{d}{dt}\Big(\|A^{\frac{1}{2}}\tilde{w}(t)\|^2 + \epsilon\|Aw(t)\|^2\Big) + 2(\alpha - \epsilon)\|A^{\frac{1}{2}}\tilde{w}(t)\|^2 + 2\epsilon\|Aw(t)\|^2 \\ = \sum_{i=1}^{4} \mathcal{J}_i, \end{aligned} \tag{3.71}$$

where

$$\begin{aligned} \mathcal{J}_1 &= 2\epsilon\left(\ell_1(t) - \epsilon\right)\langle A^{\frac{1}{2}}w(t), A^{\frac{1}{2}}\tilde{w}(t)\rangle, \\ \mathcal{J}_2 &= -2\ell_2(t)(\|u_t^1(t)\|^2 - \|u_t^2\|^2)\langle A^{\frac{1}{2}}u_t^1(t) + u_t^2(t), A^{\frac{1}{2}}\tilde{w}(t)\rangle, \\ \mathcal{J}_3 &= -2\langle f(u^1(t)) - f(u^2(t)), A\tilde{w}(t)\rangle, \\ \mathcal{J}_4 &= 2\langle 0, A\tilde{w}(t)\rangle. \end{aligned}$$

We now estimate $\mathcal{J}_i$ $(i = 1, 2, 3, 4)$ one by one. Using Young's inequality, we obtain

$$\begin{aligned} |\mathcal{J}_1| &= \Big|2\epsilon(\ell_1(t) - \epsilon)\langle A^{\frac{1}{2}}w(t), A^{\frac{1}{2}}\tilde{w}(t)\rangle\Big| \\ &\leq 2\epsilon(\beta - \epsilon)\Big|\langle A^{\frac{1}{2}}w(t), A^{\frac{1}{2}}\tilde{w}(t)\rangle\Big| \\ &\leq \frac{1}{8}\alpha\|A^{\frac{1}{2}}\tilde{w}(t)\|^2 + C_{\epsilon,\alpha,\beta}\|A^{\frac{1}{2}}w(t)\|^2. \end{aligned} \tag{3.72}$$

Using Lemma 3.13 and Cauchy-Schwarz inequality, and Young's inequality, we deduce

$$\begin{aligned}
|\mathcal{J}_2| &= \left|2\ell_2(t)(\|u_t^1(t)\|^2 - \|u_t^2(t)\|^2)\langle A^{\frac{1}{2}}(u_t^1+u_t^2), A^{\frac{1}{2}}\tilde{w}(t)\rangle\right| \\
&\leq (|u_{Max}|^{q-1})'\|u_t^1(t)+u_t^2(t)\|\cdot\|u_t^1(t)-u_t^2(t)\| \\
&\qquad \cdot \|A^{\frac{1}{2}}(u_t^1(t)+u_t^2(t))\|\cdot\|A^{\frac{1}{2}}\tilde{w}(t)\| \\
&\leq C_{M,(|u_{Max}|^{q-1})',J_{\|B_1\|_{\mathbb{X}^1}}}\|u_t^1(t)-u_t^2(t)\|\cdot\|A^{\frac{1}{2}}\tilde{w}(t)\| \\
&\leq \frac{1}{8}\alpha\|A^{\frac{1}{2}}\tilde{w}(t)\|^2 + C_{|u_0|^{q-1},M,(|u_{Max}|^{q-1})',J_{\|\Pi_{\mathbb{X}}\mathfrak{B}_S\|_{\mathcal{H}^1}}}\|\partial_t w(t)\|^2,
\end{aligned}$$

where $(|u_{Max}|^{q-1})' = \max\limits_{r\in[0,M]}|(|u(r)|^{q-1})'|$ and $\Pi_{\mathcal{H}}$ denotes the bounded projection from $\mathbb{E}$ to $\mathcal{H}$.

$$\begin{aligned}
|\mathcal{J}_3| &= 2\left|\langle f'(u^1(t))\nabla u^1(t) - f'(u^2(t))\nabla u^2(t), A^{\frac{1}{2}}\tilde{w}(t)\rangle\right| \\
&= \left|\langle \ell_3(t)A^{\frac{1}{2}}w(t), A^{\frac{1}{2}}\tilde{w}(t)\rangle\right| + \left|\langle \ell_4(t)w(t)A^{\frac{1}{2}}(u^1(t)+u^2(t)), A^{\frac{1}{2}}\tilde{w}(t)\rangle\right|
\end{aligned}$$

Where $\ell_3(t) = f'(u^1(t))+f'(u^2(t))$ and $\ell_4(t) = \displaystyle\int_0^1 f''(su^1(t)+(1-s)u^2(t))ds$. By Assumption (3.1)-(3.4) and Lemma 3.13, and using $\mathcal{H}^2 \hookrightarrow L^\infty$, we get

$$|\mathcal{J}_3| \leq \frac{1}{8}\alpha\|A^{\frac{1}{2}}\tilde{w}(t)\|^2 + C_{q,\alpha,J_{\|\Pi_{\mathcal{H}}\mathfrak{B}_S\|_{\mathcal{H}^1}}}(\|A^{\frac{1}{2}}w(t)\|^2 + \|w(t)\|^2).$$

And, we have

$$|\mathcal{J}_4| \leq \frac{1}{8}\alpha\|A^{\frac{1}{2}}\tilde{w}(t)\|^2 + C_\alpha\|\omega_1-\omega_2\|_{\mathbb{T}^n}^2. \tag{3.73}$$

Now substituting (3.72)-(3.73) into (3.71) and choosing ϵ small enough, we find

$$\begin{aligned}
&\frac{d}{dt}\left(\|\tilde{w}(t)\|_{\mathcal{H}^1}^2 + \|w(t)\|_{\mathcal{H}^2}^2\right) + \lambda\left(\|\tilde{w}(t)\|_{\mathcal{H}^1}^2 + \|w(t)\|_{\mathcal{H}^2}^2\right) \\
&\qquad \leq \mathcal{K}_2\left(\|w(t)\|_{\mathcal{H}^1}^2 + \|w(t)\|^2 + \|\omega_1-\omega_2\|_{\mathbb{T}^n}^2\right),
\end{aligned} \tag{3.74}$$

where $\lambda = \frac{\alpha}{2}$ and $\mathcal{K}_2$ depend on $q, \epsilon, \alpha, \beta, (|u_{Max}|^{q-1})'$ and $J_{\|\Pi_{\mathcal{H}}\mathfrak{B}_S\|_{\mathcal{H}^1}}$. Applying Gronwall's inequality, we obtain

$$\|S(t)(u^1,\omega_1) - S(t)(u^2,\omega_2)\|_{\mathbb{E}_1} \leq e^{\alpha_6 t}\|(u^1,\omega_1)-(u^2,\omega_2)\|_{\mathbb{E}^1}, \tag{3.75}$$

for all $t \geq 0$ and $(u^i) \in \mathfrak{B}_S (i=1,2)$, and α_6 depend on $\mathcal{K}_2$ and λ_1. So the Lipschitz continuity holds. Define the space

$$\mathbb{Z} = \mathcal{W}_T \times \mathbb{T}^n, \tag{3.76}$$

where

$$\mathcal{W}_T = \{(w, w_t) : \|(w, w_t)\|^2_{\mathcal{W}_T} = \int_0^T (\|w(t)\|^2_{\mathcal{H}^2} + \|w_t(t)\|^2_{\mathcal{H}^1})dt < \infty\}$$

with an appropriate T. The norm in $\mathbb{Z}$ is given by

$$\|Z\|^2_{\mathbb{Z}} = \|(w, w_t)\|^2_{\mathcal{W}_T} + \|\omega\|^2_{\mathbb{T}^n}, \quad Z = (w, w_t, \omega) \in \mathbb{Z}.$$

Let

$$\mathbf{n}^2_{\mathbb{Z}}(Z) = \int_0^T \|w(t)\|^2_{\mathcal{H}^1} dt + \|\omega\|^2_{\mathbb{T}^n}, \quad Z = (w, w_t, \omega) \in \mathbb{Z}. \tag{3.77}$$

Obviously, $\mathbf{n}_{\mathbb{Z}}(\cdot)$ is a compact seminorm on $\mathbb{Z}$. Now, we can apply to (3.74) the Gronwall's

$$\begin{aligned}
&\|S(t)(u^1, \omega_1) - S(t)(u^2, \omega_2)\|_{\mathbb{E}^1} \\
&\quad \leq 4e^{-\lambda t}\|(u^1(0), u^1_t(0), \omega_1) - (u^2(0), u^2_t(0), \omega_2)\|_{\mathbb{E}^1} \\
&\quad + 2\mathcal{K}_2 \int_0^t e^{-\lambda(t-r)}\Big((1 + \lambda_1^{-1})\|w(r)\|^2_{\mathcal{H}^1} + \|\omega_1 - \omega_2\|^2_{\mathbb{T}^n}\Big)dr.
\end{aligned} \tag{3.78}$$

Taking T large enough such that $0 < \theta = 4e^{-\lambda T} < \frac{1}{2}$, we infer from estimate (3.78) that

$$\begin{aligned}
\|S(T)(u^1, \omega_1) - S(T)(u^2, \omega_2)\|_{\mathbb{E}^1} \leq \theta\|(u^1(0), u^1_t(0), \omega_1) \\
- (u^2(0), u^2_t(0), \omega_2)\|_{\mathbb{E}^1} + \mathcal{K}_3 \cdot \mathbf{n}_{\mathbb{Z}}(S(t)(u^1, \omega_1) \\
- S(t)(u^2, \omega_2)),
\end{aligned} \tag{3.79}$$

for some positive constant $\mathcal{K}_3$ only depend on $\mathcal{K}_2$, T, λ and λ_1. By (3.75), we deduce

$$\begin{aligned}
\|(S(t)(u^1, \omega_1) - S(t)(u^2, \omega_2))\|^2_{\mathbb{Z}} = \int_0^T \big(\|w(t)\|^2_{\mathcal{H}^2} + \|w_t(t)\|^2_{\mathcal{H}^1}\big)dt \\
+ \|\omega\|^2_{\mathbb{T}^n} \leq \ell^*\|(u^1, \omega_1) - (u^2, \omega_2)\|^2_{\mathbb{E}^1}.
\end{aligned} \tag{4.80}$$

where $\ell^* = \frac{e^{2\lambda_6\alpha_6 T}}{2\alpha_6}$. Therefore, combining (3.69), (3.75), (3.79) and (3.80), then using Theorem 3.4, we conclude that the semigroup $\{S(t)\}_{t\geq 0}$ has a time-dependent exponential attractor $\mathbf{M} = \{\mathbf{M}(s)\}_{s\in\mathbb{R}}$ in $\mathbb{E}_1$, satisfying (i), (ii) and (iv). Moreover, there exists a positive constant α_7, such that for any bounded in $\mathbb{E}^1$ set $\mathcal{B}^1 \subset \mathbb{E}^1$, it holds

$$\sup_{s\in[0,\varpi]} \operatorname{dist}_{\mathbb{E}^1}(S(t)\mathcal{B}^1, \mathbf{M}(s)) \leq Ce^{-\alpha_7 t}, \quad \forall t \geq 0.$$

Using attraction transitivity result of [2], we have

$$\sup_{s\in[0,\varpi]} \operatorname{dist}_{\mathbb{E}}(S(t)\mathcal{B}, \mathbf{M}(s)) \leq \mathcal{Q}(\|\mathcal{B}\|_{\mathbb{E}})e^{-\alpha t}, \quad \forall t \geq 0,$$

for any bounded set $\mathcal{B} \subset \mathbb{E}$, and $\beta = \frac{\alpha\alpha_7}{\alpha+\alpha_6+\alpha_7}$. □

Acknowledgement: We are grateful to the referee for the comments which improved the quality of the paper.

References

[1] Abdallah, A. Y., Uniform exponential attractors for first order non-autonomous lattice dynamical systems, *J. Differential Equations* **251** (2011), 1489–1504.

[2] Babin, A., Nicolaenko, B., Exponential attractors of reaction-diffusion systems in an unbounded domain, *J. Dyn. Diff. Eqns.* **7** (1995), 567–590.

[3] Caraballo, T., Lukaszewicz, G., Real, J., Pullback attractors for asymptotically compact non-autonomous dynamical systems, *Nonlinear Anal.* **64** (2006), 484–498.

[4] Carvalho, A. N., Langa, J. A., Robinson, J. C., *Attractors for Infinite-dimensional Non-autonomous Dynamical Systems*, Applied Mathematical Sciences, vol. 182, Springer, New York, 2013.

[5] Carvalho, A. N., Sonner, S., Pullback exponential attractors for evolution processes in Banach spaces: theoretical results, *Commun. Pure Appl. Anal.* **12** (2013), 3047–3071.

[6] Cavalcanti, M. M., Domingos Cavalcanti, V. N., Jorge Silva, M. A., Webler, C. M., Exponential stability for the wave equation with degenerate nonlocal weak damping, *Israel J. Math.* **219** (2017), 189–213.

[7] Chueshov, I., Long-time dynamics of Kirchhoff wave models with strong nonlinear damping, *J. Differential Equations* **252** (2012), 1229–1262.

[8] Chueshov, I., *Dynamics of Quasi-Stable Dissipative Systems*, Springer, New York, 2015.

[9] Chueshov, I., Lasiecka, I., *Long-time Behavior of Second Order Evolution Equations with Nonlinear Damping*, Mem. Amer. Math. Soc., vol. 195, 2008.

[10] Hale, J. K., *Asymptotic Behaviour of Dissipative Dynamical Systems*, Mathematical Surveys and Monographs 25, Amer. Math. Soc., Providence, 1988.

[11] Haraux, A., *Recent results on semilinear wave equations with dissipation*, in *Semigroups, Theory and Applications*, Pitman Res. Notes Math. Ser. 141, Longman, 1986, 150–157.

[12] Jiang, H. T., Zhao, C. D., *Trajectory statistical solutions and Liouville type theorem for nonlinear wave equations with polynomial growth*, *Adv. Differential Equations* **3–4** (2021), 107–132.

[13] Kirchhoff, G., *Vorlesungen uber mathematische Physik-Mechanik*, Teubner, Leipzig, 1883.

[14] Lange, H., Perla Menzala, G., *Rates of decay of a nonlocal beam equation*, *Differential Integral Equations* **10** (1997), 1075–1092.

[15] Medeiros, L. A., Ferrel, J. L., Menezes, S. B., *Vibration of elastic strings: mathematical aspects I*, *J. Comput. Anal. Appl.* **4** (2002), 91–127.

[16] Medeiros, L. A., Ferrel, J. L., Menezes, S. B., *Vibration of elastic strings: mathematical aspects II*, *J. Comput. Anal. Appl.* **4** (2002), 211–263.

[17] Pucci, P., Saldi, S., *Asymptotic stability for nonlinear damped Kirchhoff systems involving the fractional p-Laplacian operator*, *J. Differ. Equ.* **263** (2017), 2375–2418.

[18] Robinson, J. C., *Infinite-Dimensional Dynamical Systems*, Cambridge University Press, Cambridge, 2001.

[19] Sun, C. Y., Cao, D. M., Duan, J. Q., *Non-autonomous dynamics of wave equations with nonlinear damping and critical nonlinearity*, *Nonlinearity* **19** (2006), 2645–2665.

[20] Sun, C. Y., Yang, L., Duan, J. Q., *Asymptotic behavior for a semilinear second order evolution equation*, *Trans. Amer. Math. Soc.* **363** (2011), 6085–6109.

[21] Sun, C. Y., Yang, M. H., *Dynamics of the nonclassical diffusion equations*, *Asymptotic Anal.* **59** (2008), 51–81.

[22] Sun, C. Y., Yang, M. H., Zhong, C. K., *Global attractors for the wave equation with nonlinear damping*, *J. Differential Equations* **227** (2006), 427–443.

Melouka Remil, Fatima Zohra Mahdi
Department of Mathematics
Laboratory of Analysis and Geometry and its Applications (AGA)
Relizane University, Algeria.
E-mail: melouka.remil@univ.dz
E-mail: mahdifatimazohra@yahoo.fr

Abed Yfrah
Department of Mathematics
Relizane University, Algeria
Laboratory of Fundamental and applied mathematics (LMFAO),
University of Oran 1, Algeria.
E-mail: boucennaabed@gmail.com

Kouider Daouadji Djamel
Department of Mathematics
Relizane University, Algeria.
E-mail: djamelkouiderdaouadji1@gmail.com

The Mathematics Student ISSN: 0025-5742
Vol. 95, Nos. 3-4, July - December (2026), 109–125

INVESTIGATIONS ON WEIGHTED VALUE SHARING AND ITS IMPACT ON UNIQUENESS OF MEROMORPHIC FUNCTIONS

SUDIP SAHA, AMIT KUMAR PAL AND SOUMON ROY
(Received : 18 - 07 - 2025 ; Accepted : 15 - 07 - 2026)

Abstract. In the uniqueness theory of meromorphic functions, two of the fundamental and finest results are due to R. Nevanlinna, popularly known as Nevanlinna's Five Value and Four Value theorems. In those theorems, uniqueness relations were obtained by considering the sharing of five and four values by two functions respectively. In recent years, many researchers tried to refine these results by considering less number of shared values and also by weakening the sharing condition. One of these results was due to H. X. Yi in (Complex Variable Theory Appl., 28, 1995, 1-11), where uniqueness relations were obtained by considering only one shared value and with CM sharing. In this manuscript, we have achieved different relationships between two meromorphic functions, when they share a value but under weighted sharing of weight 2.

1. Introduction

Unless otherwise stated, meromorphic function in this paper refers to a meromorphic function in the extended complex plane $\mathbb{C}_\infty$. We denote by E an arbitrary set of positive real numbers having finite linear measure of $0 < r < \infty$ and use I to denote any set of infinite linear measure of $0 < r < \infty$ ([9]). If $h(z)$ is a non-constant meromorphic function, $S(r,h)$ represents any quantity fulfilling the criteria

$$S(r,h) = o(T(r,h)),$$

as $r \to \infty,\ r \in I$.

Our notation follows the conventions of Nevanlinna theory for meromorphic functions as described in ([2]). For completeness, we also review some key

2020 Mathematics Subject Classification: 30D35, 30D30, 30D20
Key words and phrases: Meromorphic function , Value Distribution , Nevanlinna theory

© *Indian Mathematical Society, 2026*.

definitions and notations that will be used subsequently.
Let f and g be two non-constant meromorphic functions and a be a finite complex number.
We say that f and g share the value a CM (Counting Multiplicities) if $f-a$ and $g-a$ have the same set of zeros with identical multiplicities.
We say that f and g share the value a IM (Ignoring Multiplicities) if $f-a$ and $g-a$ have the same set of zeros, disregarding the multiplicities.
Furthermore, f and g are said to share ∞ CM (resp. IM) if $\frac{1}{f}$ and $\frac{1}{g}$ share 0 CM (resp. IM).

2. Necessary Definitions and Notations

Definition 2.1. We denote by $N(r,a;f|=1)$ the counting function of simple a-points of f.

Definition 2.2. If k is a positive integer, we denote by $N(r,a;f|\geq k)$ the counting function of those a-points of f, whose multiplicity is greater than or equal to k. Also $\overline{N}(r,a;f|\geq k)$ is the reduced counting function of the above mentioned points.

Definition 2.3. For any complex number a, by $N_2(r,a;f)$, we want to denote the sum $\overline{N}(r,a;f)+\overline{N}(r,a;f|\geq 2)$. Likewise, $N_2(r,f)=\overline{N}(r,f)+\overline{N}(r,f|\geq 2)$.

Definition 2.4. Let a be a complex number and f, g be two non-constant meromorphic functions such that f and g share the value a IM. Let z_0 be an a-point of f with multiplicity p, an a-point of g with multiplicity q. We denote by $\overline{N}_L(r,a;f)$ the reduced counting function of those a-points of f and g, where $p>q$.
We also denote by $N_E^{1)}(r,a;f)$ the counting function of those a-points of f and g where $p=q=1$ and by $\overline{N}_E^{(2}(r,a;f)$ the reduced counting function of those a-points of f and g where $p=q\geq 2$. In the same way we can define $\overline{N}_L(r,a;g), N_E^{1)}(r,a;g), \overline{N}_E^{(2}(r,a;g)$.

Definition 2.5. Let, f,g share a value a IM, we denote by $\overline{N}_*(r,a;f,g)$ the reduced counting function of those a-points of f whose multiplicities are not equal to the multiplicities of the corresponding a-points of g.
Clearly $\overline{N}_*(r,a;f,g)=\overline{N}_*(r,a;g,f)$ and $\overline{N}_*(r,a;f,g)=\overline{N}_L(r,a;f)+\overline{N}_L(r,a;g)$.

Definition 2.6. Let k be a non-negative integer or infinity. For $a \in \mathbb{C}_\infty$ we denote by $E_k(a;f)$ the set of all a-points of f where an a-point of multiplicity m is counted m times if $m \leq k$ and $k+1$ times if $m > k$.

If $E_k(a;f) = E_k(a;g)$, then we say that f and g share the value a with weight k. We sometimes also write f, g share (a,k) to mean that f, g share the value a with weight k.

Note 2.1. It is evident that, if two functions f and g share a value a ($a \in \mathbb{C}_\infty$) with weight p, then they also share the value a with weight d, where $0 \leq d < p$.

Definition 2.7. If f and g be two non-constant meromorphic functions under consideration then for any finite complex number a, the quantity $\max\{T(r,a;f),\ T(r,a;g)\}$ will be denoted by $T(r,a)$. Likewise, $T(r) = \max\{T(r,f),\ T(r,g)\}$.

3. Uniqueness of two functions

In 1920, R. Nevanlinna ([5]) established the following two results that are famously celebrated as Five Value theorem and Four Value theorem respectively, which are the fundamental results towards the theory of uniqueness of two functions:

Theorem A. If two non-constant meromorphic functions f and g share five values IM, then $f \equiv g$.

Theorem B. Let f and g be two non-constant meromorphic functions. If f and g share four values CM, then f is a Mobius transformation of g.

Numerous researchers ([1, 3, 6, 7]) have attempted to refine these results, both, by decreasing the number of shared values between two functions and by relaxing the sharing conditions.

In 1988, H. X. Yi ([8]), proved the following result:

Theorem C. Let f and g be two meromorphic functions such that f and g share $0, 1, \infty$ CM. If

$$\limsup_{r\to\infty,\ r\in I} \frac{\overline{N}(r,\frac{1}{f}) + \overline{N}(r,f)}{T(r)} < \frac{1}{2},$$

then $f \equiv g$ or $f \cdot g \equiv 1$.

Later, the same author ([9]) improved the above result by considering only two values and proved that:

Theorem D. Let f and g be two meromorphic functions such that f and g share $1, \infty$, CM. If

$$\limsup_{r\to\infty,\ r\in I} \frac{N(r,\frac{1}{f}) + N(r,\frac{1}{g}) + 2\overline{N}(r,f)}{T(r)} < 1,$$

then $f \equiv g$ or $f \cdot g \equiv 1$.

Later on the same author ([10]), went further and consider only one value in the sharing and considerably improved the previous result by establishing the following:

Theorem E. Let f and g be two meromorphic functions such that f and g share 1, CM. If

$$\limsup_{r\to\infty,\ r\in I} \frac{N_2(r,\frac{1}{f}) + N_2(r,f) + N_2(r,\frac{1}{g}) + N_2(r,g)}{T(r)} < 1,$$

then $f \equiv g$ or $f \cdot g \equiv 1$.

Now coming to the motivation behind this manuscript, in this short note, we seek to improve Theorem E by preserving the shared value 1 between two functions, while lowering the sharing environment to the weighted sharing with weight two. Through this analysis, we have uncovered various relationships between the two value-sharing functions. Below, we present the principal findings of this work:

Theorem 3.1. *Consider two distinct non-zero finite complex numbers a and b that are different from 1. Also suppose that f and g are two non-constant meromorphic functions sharing 1 with the weight 2 and fulfilling the condition:*

$$\limsup_{r\to\infty, r\in I} \frac{N_2(r,f) + N_2(r,g) + N_2(r,\frac{1}{f}) + N_2(r,\frac{1}{g}) + N_2(r,a) + N_2(r,b)}{T(r)} < 3,$$

where $N_2(r,a) = \max\{N_2(r,a;f),\ N_2(r,a;g)\}$ and similar definition for $N_2(r,b)$.

Then one of the following conclusions holds:

(i) $g(a-f) \equiv a^2$ *or* $g(a-f) \equiv ab$

(ii) $g(b-f) \equiv b^2$ *or* $g(b-f) \equiv ab$

(iii) $f(a-g) \equiv a^2$ *or* $f(a-g) \equiv ab$

(iv) $f(b-g) \equiv b^2$ *or* $f(b-g) \equiv ab$
(v) $\left(\frac{a}{f}+\frac{b}{g}\right) \equiv 1$ *or* $\left(\frac{a}{g}+\frac{b}{f}\right) \equiv 1$
(vi) $\left(\frac{f}{a}+\frac{g}{b}\right) \equiv 1$ *or* $\left(\frac{f}{b}+\frac{g}{a}\right) \equiv 1$
(vii) $(f-a)(g-a) \equiv a(a-b)$ *or* $(f-a)(g-b) \equiv b(a-b)$
(viii) $(f-b)(g-a) \equiv a(b-a)$ *or* $(f-b)(g-b) \equiv b(b-a)$
(ix) $f+g \equiv 2$ *or* $\frac{1}{f}+\frac{1}{g} \equiv 2$
(x) $fg \equiv 1$ *or* $f \equiv g$.

4. Auxiliary Lemmas

Assuming that f and g are two non-constant meromorphic functions. Let us define a new function V as

$$V = \left(\frac{f''}{f'} - \frac{g''}{g'}\right) - \left(\frac{2f'}{f-1} - \frac{2g'}{g-1}\right).$$

Lemma 4.1 ([4]). If two meromorphic functions f, g share 1 with weight 1 and V is not identically equal to 0, then

(a) $\overline{N}(r,1;f|=1) \leq N(r,V) + S(r,f) + S(r,g)$
(b) $\overline{N}(r,1;g|=1) \leq N(r,V) + S(r,f) + S(r,g)$.

Lemma 4.2. Suppose that two non-constant meromorphic functions f and g share 1, IM and suppose that a, b are two distinct complex numbers that are different from $0, 1, \infty$. Then in the case when $V \not\equiv 0$, the following estimate holds:

$$\begin{aligned} & N(r,V) \\ \leq\ & \overline{N}(r,\infty;f|\geq 2) + \overline{N}(r,0;f|\geq 2) + \overline{N}(r,\infty;g|\geq 2) + \overline{N}(r,0;g|\geq 2) \\ & + \overline{N}(r,a;f|\geq 2) + \overline{N}(r,b;f|\geq 2) + \overline{N}(r,a;g|\geq 2) + \overline{N}(r,b;g|\geq 2) \\ & + \overline{N}_*(r,1;f,g) + \overline{N}_1(r,0;f') + \overline{N}_1(r,0;g'), \end{aligned}$$

where $\overline{N}_1(r,0;f')$ is the reduced counting function of the zeros of f', that do not come from the zeros of $f(f-1)(f-a)(f-b)$. Similarly $\overline{N}_1(r,0;g')$ is defined.

Proof. Given

$$V = \left(\frac{f''}{f'} - \frac{g''}{g'}\right) - \left(\frac{2f'}{f-1} - \frac{2g'}{g-1}\right).$$

A small simplification leads us to the conclusion that the poles of V can be from

- poles of the function f and g,
- zeros of the function f' and g' and
- zeros of the function $f-1$ and $g-1$.

We now categorize the zeros of f' into two parts. Firstly the zeros of f' that are basically the zeros, or 1 points, or a-points, or b-points of f. Also in the second category there are those zeros of f' that are not the above mentioned points of f, i.e. not the zeros of $f(f-1)(f-a)(f-b)$. As per the statement, we are denoting the reduced counting function for this second category by $\overline{N}_1(r,0;f')$. Same categorization is done for the function g.
By taking V under consideration, a simple calculation infers that simple zeros, simple 1 points, simple a-points, simple b-points and simple poles of f can not contribute to the pole of V. The same assertion goes for g. Also, since f and g are sharing the value 1 IM, in the case when the multiplicities of zeros of $(f-1)$ and $(g-1)$ are equal, then that can not contribute to the poles of V. Therefore, considering all these, we can say that poles of V can only come from

- multiple poles, multiple zeros, multiple a,b-points of f and g,
- zeros of $f-1$ and $g-1$, where the multiplicities of them are different and
- zeros of f' not coming from the $0,1,a,b$-points of f and the same of g'.

Therefore,

$$\begin{aligned} & N(r,V) \\ \leq\ & \overline{N}(r,\infty;f|\geq 2)+\overline{N}(r,0;f|\geq 2)+\overline{N}(r,\infty;g|\geq 2)+\overline{N}(r,0;g|\geq 2) \\ & +\overline{N}(r,a;f|\geq 2)+\overline{N}(r,b;f|\geq 2)+\overline{N}(r,a;g|\geq 2)+\overline{N}(r,b;g|\geq 2) \\ & +\overline{N}_*(r,1;f,g)+\overline{N}_1(r,0;f')+\overline{N}_1(r,0;g'). \end{aligned}$$

This completes the proof. □

5. Proof of the main theorem

Proof of Theorem 3.1. We now assess the following two scenarios, that are arising here: $V\not\equiv 0$ and $V\equiv 0$.
Firstly let us assume that, $V\not\equiv 0$.
Since, the functions f and g share the value 1 with weight 2, then

$$\overline{N}_*(r,1;f,g)\leq \overline{N}(r,1;g|\geq 3).$$

Therefore, with the help of Lemma (4.1) and Lemma (4.2), we can deduce that

$$\begin{aligned}
& N(r,1;f|=1) \\
\leq\ & N(r,V)+S(r,f)+S(r,g) \qquad (5.1)\\
\leq\ & \overline{N}(r,\infty;f|\geq 2)+\overline{N}(r,0;f|\geq 2)+\overline{N}(r,\infty;g|\geq 2)+\overline{N}(r,0;g|\geq 2)\\
& +\overline{N}(r,a;f|\geq 2)+\overline{N}(r,b;f|\geq 2)+\overline{N}(r,a;g|\geq 2)+\overline{N}(r,b;g|\geq 2)\\
& +\overline{N}_{*}(r,1;f,g)+\overline{N}_{1}(r,0;f')+\overline{N}_{1}(r,0;g')+S(r,f)+S(r,g).\\
\leq\ & \overline{N}(r,\infty;f|\geq 2)+\overline{N}(r,0;f|\geq 2)+\overline{N}(r,\infty;g|\geq 2)+\overline{N}(r,0;g|\geq 2)\\
& +\overline{N}(r,a;f|\geq 2)+\overline{N}(r,b;f|\geq 2)+\overline{N}(r,a;g|\geq 2)+\overline{N}(r,b;g|\geq 2)\\
& +\overline{N}(r,1;g|\geq 3)+\overline{N}_{1}(r,0;f')+\overline{N}_{1}(r,0;g')+S(r,f)+S(r,g). \qquad (5.2)\\
=\ & \overline{N}(r,\infty;f|\geq 2)+\overline{N}(r,0;f|\geq 2)+\overline{N}(r,\infty;g|\geq 2)+\overline{N}(r,0;g|\geq 2)\\
& +\overline{N}(r,a;f|\geq 2)+\overline{N}(r,b;f|\geq 2)+\overline{N}_{1}(r,0;f')\\
& +\big[\ \overline{N}(r,a;g|\geq 2)+\overline{N}(r,b;g|\geq 2)+\overline{N}(r,1;g|\geq 3)\\
& +\overline{N}_{1}(r,0;g')+\overline{N}(r,1;g|\geq 2)+N(r,0;g)-\overline{N}(r,0;g)\ \big]\\
& -\overline{N}(r,1;g|\geq 2)-N(r,0;g)+\overline{N}(r,0;g)+S(r,f)+S(r,g).\\
\leq\ & \overline{N}(r,\infty;f|\geq 2)+\overline{N}(r,0;f|\geq 2)+\overline{N}(r,\infty;g|\geq 2)+\overline{N}(r,0;g|\geq 2)\\
& +\overline{N}(r,a;f|\geq 2)+\overline{N}(r,b;f|\geq 2)+\overline{N}_{1}(r,0;f')+N(r,0;g')\\
& -\overline{N}(r,1;g|\geq 2)-N(r,0;g)+\overline{N}(r,0;g)+S(r,f)+S(r,g). \qquad (5.3)
\end{aligned}$$

Also we see that, by means of Milloux Theory for meromorphic function and First Fundamental theorem, we can deduce that

$$\begin{aligned}
N(r,0;g') &\leq N\left(r,0;\frac{g'}{g}\right)+N(r,0;g)-\overline{N}(r,0;g)\\
&\leq N\left(r,\frac{g'}{g}\right)+N(r,0;g)-\overline{N}(r,0;g)+S(r,g)\\
&\leq \overline{N}(r,0;g)+\overline{N}(r,g)+N(r,0;g)-\overline{N}(r,0;g)+S(r,g)\\
&= \overline{N}(r,g)+N(r,0;g)+S(r,g). \qquad (5.4)
\end{aligned}$$

Consequently, using equation (5.4), from (5.1) we can yield that

$$\begin{aligned}
& N(r,1;f|=1) \qquad (5.5)\\
\leq\ & \overline{N}(r,\infty;f|\geq 2)+\overline{N}(r,0;f|\geq 2)+\overline{N}(r,\infty;g|\geq 2)+\overline{N}(r,0;g|\geq 2)\\
& +\overline{N}(r,a;f|\geq 2)+\overline{N}(r,b;f|\geq 2)+\overline{N}_{1}(r,0;f')+\overline{N}(r,g)+\overline{N}(r,0;g)
\end{aligned}$$

$$-\overline{N}(r,1;g|\geq 2)+S(r,f)+S(r,g). \tag{5.6}$$

Also since the functions f and g share the value 1 with weight 2, then the following assertions are obvious:

$$\begin{aligned}\overline{N}(r,1;f) &= \overline{N}(r,1;f|=1)+\overline{N}(r,1;f|\geq 2)\\ &= \overline{N}(r,1;f|=1)+\overline{N}(r,1;g|\geq 2)\\ &= N(r,1;f|=1)+\overline{N}(r,1;g|\geq 2).\end{aligned} \tag{5.7}$$

Now by the use of the Second Fundamental theorem for f, we get

$$\begin{aligned}3T(r,f) \leq\ & \overline{N}(r,f)+\overline{N}(r,1;f)+\overline{N}(r,0;f)+\overline{N}(r,a;f)+\overline{N}(r,b;f)\\ & -N_1(r,0;f')+S(r,f)\end{aligned} \tag{5.8}$$

Therefore with the help of (5.5), (5.7) and (5.8) we obtain:

$$\begin{aligned}3T(r,f) \leq\ & N_2(r,\infty;f)+N_2(r,\infty;g)+N_2(r,0;f)+N_2(r,0;g)\\ & +N_2(r,a;f)+N_2(r,b;f)+S(r,f)+S(r,g).\end{aligned} \tag{5.9}$$

Similarly, for the function g, we can obtain

$$\begin{aligned}3T(r,g) \leq\ & N_2(r,\infty;f)+N_2(r,\infty;g)+N_2(r,0;f)+N_2(r,0;g)\\ & +N_2(r,a;g)+N_2(r,b;g)+S(r,f)+S(r,g).\end{aligned} \tag{5.10}$$

Hence using (5.9) and (5.10), we obtain that

$$\begin{aligned}3\max\{T(r,f),T(r,g)\} \leq\ & N_2(r,\infty;f)+N_2(r,\infty;g)+N_2(r,0;f)\\ & +N_2(r,0;g)+N_2(r,a)+N_2(r,b)\\ & +S(r,f)+S(r,g).\end{aligned}$$

This is contradicting the condition that is provided in the statement since $T(r)=\max\{T(r,f),\ T(r,g)\}$. So this case is excluded from our consideration. The only option we are left with is $V\equiv 0$.
Then, an easy calculation deduces that

$$f=\frac{Ag+B}{Cg+D}, \tag{5.11}$$

where A,B,C,D are constants with the property $(AD-BC)\neq 0$.
From (5.11) it is obvious that

$$T(r,f)=T(r,g)+O(1). \tag{5.12}$$

If 1 be a Picard's exceptional value of f and consequently of g, then using Second Fundamental theorem we get

$$\begin{aligned} & 3T(r,f) \qquad (5.13)\\ \leq\ & \overline{N}(r,f)+\overline{N}(r,0;f)+\overline{N}(r,a;f)+\overline{N}(r,b;f)+\overline{N}(r,1;f)+S(r,f).\\ \leq\ & \overline{N}(r,f)+\overline{N}(r,0;f)+\overline{N}(r,a;f)+\overline{N}(r,b;f)+S(r,f)+S(r,g), \end{aligned}$$

Using (5.12) we can have that

$$\begin{aligned} & 3T(r,g) \qquad (5.14)\\ \leq\ & \overline{N}(r,f)+\overline{N}(r,0;f)+\overline{N}(r,a;f)+\overline{N}(r,b;f)+S(r,f)+S(r,g). \end{aligned}$$

So from (5.13) and (5.14) we can assert that

$$\begin{aligned} & 3\max\{T(r,f),T(r,g)\}\\ \leq\ & \overline{N}(r,f)+\overline{N}(r,0;f)+\overline{N}(r,a;f)+\overline{N}(r,b;f)+S(r,f)+S(r,g),\\ \leq\ & N_2(r,\infty;f)+N_2(r,\infty;g)+N_2(r,0;f)+N_2(r,0;g)\\ & +N_2(r,a)+N_2(r,b)+S(r,f)+S(r,g), \end{aligned}$$

which is contradicting the condition that is provided in the statement since $T(r)=\max\{T(r,f),\ T(r,g)\}$. Therefore, from now on we shall use the fact that 1 is not a Picard's exceptional value of f and g.

We now assess two following possible outcomes; firstly we will see the possibility of $AC\neq 0$ and subsequently we shall look at the case when $AC=0$.

Case - 1: (AC $\neq$ 0). Using the equation (5.11) we see that

$$f-\frac{A}{C}=\frac{B-\frac{AD}{C}}{Cg+D}. \qquad (5.15)$$

Subcase:1.1 Let, $\frac{A}{C}\neq a,b$. Using Second Fundamental theorem we get,

$$\begin{aligned} & 3T(r,f)\\ \leq\ & \overline{N}(r,f)+\overline{N}(r,0;f)+\overline{N}(r,a;f)+\overline{N}(r,b;f)+\overline{N}(r,\frac{A}{C};f)+S(r,f).\\ \leq\ & \overline{N}(r,f)+\overline{N}(r,0;f)+\overline{N}(r,a;f)+\overline{N}(r,b;f)+\overline{N}(r,g)\\ & +S(r,f)+S(r,g). \qquad (5.16) \end{aligned}$$

Use of equation (5.12) gives that

$$\begin{aligned} & 3T(r,g)\\ \leq\ & \overline{N}(r,f)+\overline{N}(r,0;f)+\overline{N}(r,a;f)+\overline{N}(r,b;f)+\overline{N}(r,g) \end{aligned}$$

$$+S(r,f)+S(r,g). \tag{5.17}$$

Using (5.16) and (5.17) we obtain that

$$\begin{aligned} & 3\max\{T(r,f),T(r,g)\} \\ \leq\ & \overline{N}(r,f)+\overline{N}(r,0;f)+\overline{N}(r,a;f)+\overline{N}(r,b;f) \\ & +\overline{N}(r,g)+S(r,f)+S(r,g) \\ \leq\ & N_2(r,\infty;f)+N_2(r,\infty;g)+N_2(r,0;f)+N_2(r,0;g) \\ & +N_2(r,a)+N_2(r,b)+S(r,f)+S(r,g), \end{aligned}$$

which is contradicting the condition that is provided in the statement since $T(r)=\max\{T(r,f),\ T(r,g)\}$.

Subcase:1.2 Let, us now consider the next possibility that $\frac{A}{C}=a$ and $\frac{A}{C}\neq b$. Under this, we now focus on the case when $BD=0$. Since $(AD-BC)\neq 0$, we either have $B\neq 0, D=0$ or $B=0, D\neq 0$.
Firstly, if $B\neq 0, D=0$. Then since 1 is not a Picard's exceptional value of f and g, from (5.15), we obtain,

$$f=a+\frac{1-a}{g}. \tag{5.18}$$

If $a\neq 1-\frac{1}{a}$ and $b\neq 1-\frac{1}{a}$, then Second Fundamental theorem and equation (5.18) yield the following deduction:

$$\begin{aligned} & 3T(r,g) \\ \leq\ & \overline{N}(r,g)+\overline{N}(r,0;g)+\overline{N}(r,a;g)+\overline{N}(r,b;g)+\overline{N}(r,1-\frac{1}{a};g)+S(r,g) \\ \leq\ & \overline{N}(r,g)+\overline{N}(r,0;g)+\overline{N}(r,a;g)+\overline{N}(r,b;g)+\overline{N}(r,0;f)+S(r,g), \end{aligned}$$

which is again contradicting the condition that is provided in the statement, by a similar calculation that we have done in the **Subcase 1.1.** So we are also discarding this scenario.
Also if $a=1-\frac{1}{a}$ and $b\neq 1-\frac{1}{a}$, then it seems that a is a root of the equation $x^2-x+1=0$, i.e., $a=-\omega,-\omega^2$; that eventually concludes that $g(a-f)\equiv a^2$. Similarly if $a\neq 1-\frac{1}{a}$ and $b=1-\frac{1}{a}$, it leads us to $g(a-f)\equiv ab$.
Subsequently if, $B=0, D\neq 0$. Then as 1 is not a Picard's exceptional value of f and g, with the use of equation (5.15) we reach to,

$$f=\frac{ag}{(a-1)+g}. \tag{5.19}$$

If $a \neq \frac{1}{2}$ and $a+b \neq 1$, use of Second Fundamental theorem and equation (5.19) produces the followings:

$$\begin{aligned} & 3T(r,g) \\ \leq\ & \overline{N}(r,g)+\overline{N}(r,0;g)+\overline{N}(r,a;g)+\overline{N}(r,b;g)+\overline{N}(r,1-a;g)+S(r,g) \\ =\ & \overline{N}(r,g)+\overline{N}(r,0;g)+\overline{N}(r,a;g)+\overline{N}(r,b;g)+\overline{N}(r,f)+S(r,g). \end{aligned}$$

From here, by a similar calculation that we have done in the **Subcase 1.1**, leads us to a contradiction to the condition given in the hypothesis of the statement of the result.
Also if $a = \frac{1}{2}$ and $a+b \neq 1$, using (5.19) we can obtain that $\frac{1}{f}+\frac{1}{g} \equiv 2$.
Likewise if $a \neq \frac{1}{2}$ and $a+b=1$, (5.19) induces $\frac{a}{f}+\frac{b}{g} \equiv 1$.
Let us now turn our attention to the scenario, $BD \neq 0$.
Then obviously $\frac{B}{D} \neq a$, unless since $\frac{A}{C} = a$, then we will end up with $(AD-BC)=0$, which is a contradiction as per the construction. Nevertheless, if $\frac{B}{D} \neq b$, then equation (5.11) and use of Second Fundamental theorem reciprocate that

$$\begin{aligned} & 3T(r,f) \\ \leq\ & \overline{N}(r,f)+\overline{N}(r,0;f)+\overline{N}(r,a;f)+\overline{N}(r,b;f)+\overline{N}(r,\frac{B}{D};f)+S(r,f) \\ =\ & \overline{N}(r,f)+\overline{N}(r,0;f)+\overline{N}(r,a;f)+\overline{N}(r,b;f)+\overline{N}(r,0;g)+S(r,f). \end{aligned}$$

This, with a simple calculation as we have done in **Subcase 1.1**, contradicts the hypothesis of the main result.
Thus we are left with the option $\frac{B}{D} = b$. Since 1 is not a Picard's exceptional value of f and g, then with the help of (5.15), we obtain

$$f = a + \frac{(a-b)(a-1)}{(b-1)}\left\{\frac{1}{g+(\frac{a-1}{1-b})}\right\}. \tag{5.20}$$

If $(\frac{a-1}{b-1}) \neq a, b$, then from Second Fundamental theorem, using equation (5.20) we obtain

$$\begin{aligned} & 3T(r,g) \\ \leq\ & \overline{N}(r,g)+\overline{N}(r,0;g)+\overline{N}(r,a;g)+\overline{N}(r,b;g)+\overline{N}(r,(\frac{a-1}{b-1});g) \\ & +S(r,f) \\ \leq\ & \overline{N}(r,g)+\overline{N}(r,0;g)+\overline{N}(r,a;g)+\overline{N}(r,b;g)+\overline{N}(r,f)+S(r,f), \end{aligned}$$

that is again contradicting the hypothesis of statement by a similar calculation done in **Subcase 1.1**
Also if $(\frac{a-1}{b-1}) = a$, then $(f-a)(g-a) \equiv a(a-b)$.
Similarly if $(\frac{a-1}{b-1}) = b$, then we obtain $(f-a)(g-b) \equiv b(a-b)$.

Subcase:1.3 Lets now assess and examine the possible scenario that $\frac{A}{C} = b$ and $\frac{A}{C} \neq a$.
Then by similar kind of argument as in the previous case, since$(AD-BC) \neq 0$, we can readily conclude that $\frac{B}{D} \neq b$.
Now initially we are focusing on the possibility of the scenario $BD = 0$. Then we will end up having either $B = 0$ and $D \neq 0$ or $B \neq 0$ and $D = 0$.

Firstly if $B \neq 0, D = 0$, as 1 is not a Picard's exceptional value of f and g, use of equation (5.15) induces that

$$f = b + \frac{(1-b)}{g}. \tag{5.21}$$

If $a \neq 1 - \frac{1}{b}$ and $b \neq 1 - \frac{1}{b}$, then use of (5.21) and Second Fundamental theorem gets us to,

$$\begin{aligned} & 3T(r,g) \\ \leq\ & \overline{N}(r,g) + \overline{N}(r,0;g) + \overline{N}(r,a;g) + \overline{N}(r,b;g) + \overline{N}(r,1-\frac{1}{b};g) + S(r,g) \\ =\ & \overline{N}(r,g) + \overline{N}(r,0;g) + \overline{N}(r,a;g) + \overline{N}(r,b;g) + \overline{N}(r,0;f) + S(r,g). \end{aligned}$$

This is again contradicting the given hypothesis, by a simple deduction from here similar to that of done in.
Also if $a = 1 - \frac{1}{b}$ and $b \neq 1 - \frac{1}{b}$, then from (5.21) we get $g(b-f) \equiv ab$.
Likewise if $a \neq 1 - \frac{1}{b}$ and $b = 1 - \frac{1}{b}$, then $g(b-f) \equiv b^2$.
Subsequently, if $B = 0, D \neq 0$, then similarly use of the equation (5.15) follows that,

$$f = \frac{bg}{g + (b-1)}. \tag{5.22}$$

So if $b \neq \frac{1}{2}$ and $a + b \neq 1$, Second Fundamental theorem by means of (5.22) provides that,

$$\begin{aligned} & 3T(r,g) \\ \leq\ & \overline{N}(r,g) + \overline{N}(r,0;g) + \overline{N}(r,a;g) + \overline{N}(r,b;g) + \overline{N}(r,1-b;g) + S(r,g) \\ =\ & \overline{N}(r,g) + \overline{N}(r,0;g) + \overline{N}(r,a;g) + \overline{N}(r,b;g) + \overline{N}(r,f) + S(r,g). \end{aligned}$$

This is also contradicting the given hypothesis, by a simple calculation from here on, similar to that of done in **Subcase 1.1**.
Also if $b = \frac{1}{2}$ and $a + b \neq 1$, then using (5.22), it gives that $\frac{1}{f} + \frac{1}{g} \equiv 2$.
Similarly if $b \neq \frac{1}{2}, a + b = 1$, then we get $\frac{a}{g} + \frac{b}{f} \equiv 1$.
Lets now move forward and focus on the scenario when $BD \neq 0$. Since $(AD - BC) \neq 0$, then obviously $\frac{B}{D} \neq b$.
Now if $\frac{B}{D} \neq a$, then from Second Fundamental theorem and (5.11), it follows that

$$\begin{aligned} & 3T(r,f) \\ \leq\ & \overline{N}(r,f) + \overline{N}(r,0;f) + \overline{N}(r,a;f) + \overline{N}(r,b;f) + \overline{N}(r,\frac{B}{D};f) + S(r,f) \\ =\ & \overline{N}(r,f) + \overline{N}(r,0;f) + \overline{N}(r,a;f) + \overline{N}(r,b;f) + \overline{N}(r,0;g) + S(r,f), \end{aligned}$$

making a contradiction to the given hypothesis with the help of a simple calculation, similar to that done in the **Subcase 1.1**.
So, the only possibility we have is $\frac{B}{D} = a$. Therefore as 1 is not a Picard's exceptional value of f and g, from (5.15) we deduce that

$$f = b + \frac{(b-a)(b-1)}{(a-1)} \left\{ \frac{1}{g + (\frac{b-1}{1-a})} \right\}. \tag{5.23}$$

If $(\frac{b-1}{a-1}) \neq a, b$, then (5.23) and Second Fundamental theorem reciprocates that

$$\begin{aligned} & 3T(r,g) \\ \leq\ & \overline{N}(r,g) + \overline{N}(r,0;g) + \overline{N}(r,a;g) + \overline{N}(r,b;g) + \overline{N}(r,(\frac{b-1}{a-1});g) \\ & + S(r,f) \\ \leq\ & \overline{N}(r,g) + \overline{N}(r,0;g) + \overline{N}(r,a;g) + \overline{N}(r,b;g) + \overline{N}(r,f) + S(r,f), \end{aligned}$$

which is yielding a contradiction by an argument similar to that in **Subcase 1.1**.
Also if $(\frac{b-1}{a-1}) = a$, then (5.23) implies that $(f - b)(g - a) \equiv a(b - a)$.
Likewise if $(\frac{b-1}{a-1}) = b$, then $(f - b)(g - b) \equiv b(b - a)$.

Case - 2: (AC = 0) Non-constantness of f is readily implying that A and C can not be simultaneously zero. Then we can have two possible scenarios $A = 0,\ C \neq 0$ or $A \neq 0, C = 0$.
Subcase:2.1 Let us first consider the possibility of $A = 0, C \neq 0$. Then as

1 is not a Picard's exceptional value of f and g, equation (5.11) asserts that

$$\alpha \cdot g = \frac{1}{f} - (1-\alpha), \text{ where } \alpha = \frac{C}{B}. \tag{5.24}$$

In the context of the above observation, we first suppose that $\alpha \neq 1, \alpha \neq 1-\frac{1}{a}, \alpha \neq 1-\frac{1}{b}$. Then with the help of Second Fundamental theorem and (5.24) we get

$$\begin{aligned} & 3T(r,f) \\ \leq\ & \overline{N}(r,f)+\overline{N}(r,0;f)+\overline{N}(r,a;f)+\overline{N}(r,b;f)+\overline{N}(r,\frac{1}{1-\alpha};f) \\ & +S(r,f) \\ \leq\ & \overline{N}(r,f)+\overline{N}(r,0;f)+\overline{N}(r,a;f)+\overline{N}(r,b;f)+\overline{N}(r,0;g)+S(r,f), \end{aligned}$$

which again leads us to a contradiction to the given hypothesis, similar to the earlier scenarios.

Next we shift our focus to the scenario when $\alpha \neq 1,\ \alpha \neq 1-\frac{1}{a},\ \alpha = 1-\frac{1}{b}$. From (5.24), we get

$$f = \frac{b}{(b-1)g+1}. \tag{5.25}$$

Now if $a \neq \frac{1}{1-b}, b \neq \frac{1}{1-b}$, then using Second Fundamental theorem we obtain

$$\begin{aligned} & 3T(r,g) \\ \leq\ & \overline{N}(r,g)+\overline{N}(r,0;g)+\overline{N}(r,a;g)+\overline{N}(r,b;g)+\overline{N}(r,\frac{1}{1-b};g)+S(r,g) \\ \leq\ & \overline{N}(r,g)+\overline{N}(r,0;g)+\overline{N}(r,a;g)+\overline{N}(r,b;g)+\overline{N}(r,f)+S(r,g), \end{aligned}$$

which is again making a contradiction through a simple calculation, similar to that we have done in **Subcase 1.1**

Also if $a = \frac{1}{1-b},\ b \neq \frac{1}{1-b}$, then from (5.25) we get $f(a-g) \equiv ab$.

Similarly if $a \neq \frac{1}{1-b}, b = \frac{1}{1-b}$, then $f(b-g) \equiv b^2$.

Next we assess the case when $\alpha \neq 1,\ \alpha = 1-\frac{1}{a},\ \alpha \neq 1-\frac{1}{b}$. Then use of (5.24) yields that

$$f = \frac{a}{1+(a-1)g}. \tag{5.26}$$

Then if $a \neq \frac{1}{1-a},\ b \neq \frac{1}{1-a}$, then Second Fundamental theorem implies that

$$\begin{aligned} & 3T(r,g) \\ \leq\ & \overline{N}(r,g)+\overline{N}(r,0;g)+\overline{N}(r,a;g)+\overline{N}(r,b;g)+\overline{N}(r,\frac{1}{1-a};g) \\ & +S(r,g) \end{aligned}$$

$$\leq \overline{N}(r,g)+\overline{N}(r,0;g)+\overline{N}(r,a;g)+\overline{N}(r,b;g)+\overline{N}(r,f)+S(r,g),$$

which is again making a contradiction with the help of similar calculation done in **Subcase 1.1**.
Also if $a=\frac{1}{1-a},\ b\neq\frac{1}{1-a}$, then from (5.26) we get $f(a-g)\equiv a^2$.
Likewise if $a\neq\frac{1}{1-a},\ b=\frac{1}{1-a}$, then $f(b-g)\equiv ab$.
Lastly if $\alpha=1$, then obviously $\alpha\neq 1-\frac{1}{a},\alpha\neq 1-\frac{1}{b}$, then with the help of (5.24) we have $fg\equiv 1$.

Subcase:2.2 Subsequently we now consider the case $A\neq 0,\ C=0$. Then with the fact that 1 is not a Picard's exceptional value of f and g, from (5.11) we deduce

$$f=\beta\cdot g+(1-\beta),\ \text{where}\ \beta=\frac{A}{D}. \tag{5.27}$$

If $\beta\neq 1,\beta\neq 1-a,\beta\neq 1-b$, then Second Fundamental theorem gives

$$\begin{aligned}
&3T(r,f)\\
\leq\ &\overline{N}(r,f)+\overline{N}(r,0;f)+\overline{N}(r,a;f)+\overline{N}(r,b;f)+\overline{N}(r,1-\beta;f)+S(r,f)\\
\leq\ &\overline{N}(r,f)+\overline{N}(r,0;f)+\overline{N}(r,a;f)+\overline{N}(r,b;f)+\overline{N}(r,0;g)+S(r,f),
\end{aligned}$$

which is a contradiction by means of similar deductions in **Subcase 1.1.**.
Next if $\beta\neq 1,\ \beta=1-a,\ \beta\neq 1-b$, then from equation (5.27) we get

$$f=(1-a)g+a. \tag{5.28}$$

Therefore, if $a\neq\frac{a}{a-1},b\neq\frac{a}{a-1}$, then use of Second Fundamental theorem implies that,

$$\begin{aligned}
&3T(r,g)\\
\leq\ &\overline{N}(r,g)+\overline{N}(r,0;g)+\overline{N}(r,a;g)+\overline{N}(r,b;g)+\overline{N}(r,\frac{a}{a-1};g)\\
&+S(r,g)\\
\leq\ &\overline{N}(r,g)+\overline{N}(r,0;g)+\overline{N}(r,a;g)+\overline{N}(r,b;g)+\overline{N}(r,0;f)+S(r,g),
\end{aligned}$$

from which we can again come up with a contradiction, in a similar manner that we did in the **Subcase 1.1**
Also if $a=\frac{a}{a-1},b\neq\frac{a}{a-1}$. then (5.28) induces that $f+g\equiv 2$.
Likewise if $a\neq\frac{a}{a-1},b=\frac{a}{a-1}$, then $\frac{f}{a}+\frac{g}{b}\equiv 1$.
Further if $\beta\neq 1,\ \beta\neq 1-a,\ \beta=1-b$, (5.27) gives that

$$f=(1-b)g+b. \tag{5.29}$$

So if $a \neq \frac{b}{b-1}, b \neq \frac{b}{b-1}$, then using Second Fundamental theorem we get,

$$\begin{aligned} & 3T(r,g) \\ \leq\ & \overline{N}(r,g) + \overline{N}(r,0;g) + \overline{N}(r,a;g) + \overline{N}(r,b;g) + \overline{N}(r,\frac{b}{b-1};g) \\ & + S(r,g) \\ \leq\ & \overline{N}(r,g) + \overline{N}(r,0;g) + \overline{N}(r,a;g) + \overline{N}(r,b;g) + \overline{N}(r,0;f) + S(r,g), \end{aligned}$$

which is again deducing a contradiction, by an argument similar to that in **Subcase 1.1**.

Also if $a = \frac{b}{b-1}, b \neq \frac{b}{b-1}$, then from equation (5.29) we get $\frac{f}{b} + \frac{g}{a} \equiv 1$.

Similarly when $a \neq \frac{b}{b-1}, b = \frac{b}{b-1}$, then $f + g \equiv 2$.

At the end, if $\beta = 1$, then obviously $\beta \neq 1 - a,\ \beta \neq 1 - b$. Hence using (5.27), we obtain $f \equiv g$. □

Acknowledgment: We are grateful to the referee for the comments which improved the quality of the paper. This manuscript was prepared during the Ph.D. tenure of the first author and the contents of this manuscript is included in his thesis. We declare that all the authors have equally contributed to the process of conceptualization, finding a proper methodology, analysis of the result, writing, editing and reviewing of the manuscript.

References

[1] Brosch, G., Eindeutigkeitssatze fur meromorphe Functionen, thesis, Technical University of Aachen, 1989.

[2] Hayman, W. K., Meromorphic Functions, The Clarendon Press, Oxford, 1964.

[3] Hua, X. H., Sharing values and a question of C. C. Yang, *Pacific J. Math.*, **175(1)**, 1996, 71-81.

[4] Lahiri, I., Weighted Value Sharing and Uniqueness of Meromorphic Functions, *Complex Variables*, **46**, 241-253, 2001.

[5] Nevanlinna, R., Les th'eor'emes de Picard–Borel et la theorie des fonctions meromorphes, -Gauthier-Villars, Paris, 1929.

[6] Ozawa, M., Unicity theorems for entire functions, *J. D'Anal. Math.*, **30**, 1976, 411-420.

[7] Ueda, H., Unicity theorems for meromorphic or entire functions II, *Kodai Math J*, **6**, 1983, 26-36.

[8] Yi, H. X., Meromorphic functions that share three values, *Chin. Ann. Math.*, **9A**, 1988, 434-440.

[9] Yi, H. X., Meromorphic functions that share two or three values, *Kodai. Math. J.*, **13**, 1990, 363-372.

[10] Yi, H. X., Meromorphic function that share one or two values, *Complex Variable Theory Appl.*, **28**, 1995, 1-11.

Sudip Saha(Corresponding Author)
Department of Applied Science
Haldia Institute of Technology
West Bengal, India
E-mail: sudipsaha814@gmail.com

Amit Kumar Pal
Department of Mathematics
Rajiv Gandhi National Institute of Youth Development
Tamil Nadu, India
E-mail: mail4amitpal@gmail.com

Soumon Roy
Department of Mathematics
Dr. Sudhir Chandra Sur Institute of Technology & sports complex
Kolkata, India
E-mail: rsoumon@gmail.com

The Mathematics Student ISSN: 0025-5742
Vol. 95, Nos. 3-4, July - December (2026), 126–133

A SIMPLE PROOF OF AN IDENTITY INVOLVING SKEW-HARMONIC NUMBER AND A TAIL OF $\zeta(3)$

NANDAN SAI DASIREDDY

(Received : 26 - 07 - 2025 ; Accepted : 18 - 08 - 2026)

Abstract. We present a different proof of the following identity due to Nguyen, originally posed as an open problem in the monograph [10, Chapt. 3, p. 101], which gives closed-form expression for a curious skew-harmonic number series.

$$\sum_{n=1}^{\infty}\frac{\overline{H}_n}{n}\left(\zeta(3)-1-\frac{1}{2^3}-\cdots-\frac{1}{n^3}\right)=-2\,\mathrm{Li}_5\left(\frac{1}{2}\right)-2\,\mathrm{Li}_4\left(\frac{1}{2}\right)\ln 2$$

$$+\frac{193}{64}\zeta(5)-\frac{1}{15}\ln^5 2+\frac{1}{3}\zeta(2)\ln^3 2-\frac{15}{16}\zeta(4)\ln 2-\frac{5}{16}\zeta(2)\zeta(3).$$

Nguyen's proof was based on an ordinary generating function involving a tail of $\zeta(3)$. In contrast to Nguyen's generating-function proof, our approach uses certain some linear and nonlinear harmonic sums obtained by Olaikhan. This note may be viewed as a sequel to the author's work [12], which provided a new proof of a nonlinear harmonic-sum identity.

1. Introduction

Recently, Vincent Nguyen [13] proved the following identity:

$$\sum_{n=1}^{\infty}\frac{\overline{H}_n}{n}\left(\zeta(3)-H_n^{(3)}\right)=-2\,\mathrm{Li}_5\left(\frac{1}{2}\right)-2\,\mathrm{Li}_4\left(\frac{1}{2}\right)\ln 2+\frac{193}{64}\zeta(5)$$
$$-\frac{1}{15}\ln^5 2+\frac{1}{3}\zeta(2)\ln^3 2-\frac{15}{16}\zeta(4)\ln 2-\frac{5}{16}\zeta(2)\zeta(3). \tag{1.1}$$

2020 Mathematics Subject Classification: Primary 40A25, 11M35, 33B15, 33E20; Secondary 11M06
Key words and phrases: Skew-harmonic numbers, Hurwitz zeta-function, Euler's gamma function, linear harmonic sums, generalized harmonic numbers, Riemann zeta function, Polylogarithm function

© *Indian Mathematical Society, 2026.*

Our main aim in this note is to provide a different proof of (1). We believe that the method used here can also be applied to other harmonic-series identities.

We first recall several standard formulas needed in the proof. Euler's gamma function $\Gamma(v) = \int_0^\infty t^{v-1}e^{-t}dt$ $(\Re(v) > 0)$, is a natural extension of the factorial function, since $\Gamma(p+1) = p!$ for $p = 1, 2, 3, ...$ The Hurwitz zeta-function $\zeta(z,r) = \sum_{i=0}^{\infty} \frac{1}{(i+r)^z}$ $(\Re(z) > 1; r \in \mathbb{C} \setminus \{0, -1, -2, \ldots\})$ satisfy the following relations (see, for example, [4, Eq. 2.17.1, p. 37], [6, Eq. 1.19, p. 134])

$$\zeta(z,r) = \frac{1}{\Gamma(z)} \int_0^\infty \frac{u^{z-1}e^{-ru}}{1-e^{-u}} du \qquad (\Re(z) > 1; \Re(r) > 0),$$

and for any positive integer p,

$$\zeta(z, p+1) = \zeta(z) - H_p^{(z)} \qquad (\Re(z) > 1),$$

where $H_p^{(z)} = 1 + \frac{1}{2^z} + \frac{1}{3^z} + \cdots + \frac{1}{p^z}$ is the *p-th* generalized harmonic number, and $\zeta(z,1) = \zeta(z) = \sum_{i=1}^{\infty} \frac{1}{i^z}$ is the Riemann zeta-function. In particular, we have $H_p^{(1)} := H_p = \sum_{k=1}^{p} \frac{1}{k}$ is the classical harmonic number.

The *p-th* skew-harmonic number $\overline{H}_p$ is defined by $\sum_{k=1}^{p} \frac{(-1)^{k-1}}{k}$ and the polylogarithm function is defined for $|x| \leq 1$ by $\mathrm{Li}_s(x) = \sum_{k=1}^{\infty} \frac{x^k}{k^s}, \quad s \geq 2$. The value $\mathrm{Li}_2\left(\frac{1}{2}\right)$ is given in [3, Eq. (1.16)]:

$$\mathrm{Li}_2\left(\frac{1}{2}\right) = \frac{1}{2}\zeta(2) - \frac{1}{2}\ln^2(2).$$

2. A simple proof of Nguyen's identity

We now prove Nguyen's identity (1) using several logarithmic integral formulas.

Theorem 2.1. *The following identity holds:*

$$\sum_{n=1}^{\infty} \frac{\overline{H}_n}{n}\left(\zeta(3) - 1 - \frac{1}{2^3} - \cdots - \frac{1}{n^3}\right) = -2\,\mathrm{Li}_5\left(\frac{1}{2}\right) - 2\,\mathrm{Li}_4\left(\frac{1}{2}\right)\ln 2$$

$$+\frac{193}{64}\zeta(5) - \frac{1}{15}\ln^5 2 + \frac{1}{3}\zeta(2)\ln^3 2 - \frac{15}{16}\zeta(4)\ln 2 - \frac{5}{16}\zeta(2)\zeta(3).$$

Proof. We begin with the following well-known definite integral formula:

$$\int_0^1 x^{2\alpha} \ln^2(x)\, dx = \frac{2}{(2\alpha+1)^3}, \qquad \alpha > -\frac{1}{2}. \tag{2.1}$$

we begin with the following identity.

$$\sum_{n=1}^{\infty} \frac{\overline{H}_n}{n} \left(\zeta(3) - 1 - \frac{1}{2^3} - \cdots - \frac{1}{n^3} \right) = \sum_{n=1}^{\infty} \frac{\overline{H}_n}{n} \left(\sum_{j=n}^{\infty} \frac{1}{(j+1)^3} \right). \tag{2.2}$$

We replace the inner sum on the right-hand side of (2.2) by its integral representation from (2.1). After rearranging, we obtain

$$\frac{1}{2} \left[\sum_{n=1}^{\infty} \frac{\overline{H}_n}{n} \left(\int_0^1 \ln^2(x) \left(x^n + x^{n+1} + x^{n+2} + \cdots \right) dx \right) \right].$$

Summing the geometric series in the numerator gives

$$\sum_{n=1}^{\infty} \frac{\overline{H}_n}{n} \left(\zeta(3) - 1 - \frac{1}{2^3} - \cdots - \frac{1}{n^3} \right) = \frac{1}{2} \left[\sum_{n=1}^{\infty} \frac{\overline{H}_n}{n} \left(\int_0^1 \frac{x^n \ln^2(x)}{1-x} dx \right) \right]. \tag{2.3}$$

Interchanging the order of summation and integration in (2.3), we obtain

$$\frac{1}{2} \int_0^1 \frac{\ln^2(x)}{1-x} \left(\sum_{n=1}^{\infty} \frac{\overline{H}_n x^n}{n} \right) dx.$$

Next, applying Boyadzhiev's generating function [8, Equation. (5), p. 95] to the $\sum_{n=1}^{\infty} \frac{\overline{H}_n x^n}{n}$ series gives

$$\begin{aligned}
\sum_{n=1}^{\infty} \frac{\overline{H}_n}{n} \left(\zeta(3) - 1 - \frac{1}{2^3} - \cdots - \frac{1}{n^3} \right) &= \frac{1}{2} \int_0^1 \frac{\ln^2(x)}{1-x} \operatorname{Li}_2\left(\frac{1-x}{2} \right) dx \\
&\quad + \frac{1}{2} \int_0^1 \frac{\ln^2(x)}{1-x} \left(-\operatorname{Li}_2\left(\frac{1}{2} \right) - \operatorname{Li}_2(-x) - \ln(1-x) \ln 2 \right) dx. \\
&= \frac{1}{2} \left[\int_0^1 \frac{\ln^2(x)}{1-x} \operatorname{Li}_2\left(\frac{1-x}{2} \right) dx - \operatorname{Li}_2\left(\frac{1}{2} \right) \int_0^1 \frac{\ln^2(x)}{1-x} dx \right] \\
&\quad + \frac{1}{2} \left[-\int_0^1 \frac{\ln^2(x) \operatorname{Li}_2(-x)}{1-x} dx - \ln 2 \int_0^1 \frac{\ln^2(x) \ln(1-x)}{1-x} dx \right]
\end{aligned}$$

$$= \frac{1}{2}\left[I - \mathrm{Li}_2\left(\frac{1}{2}\right)(K) - L - \ln 2\,(J)\right].$$

To evaluate I, we use Bernstein's theorem [5, Theorem. 9.30, p. 243] together with the moment formula of Vălean and Furdui [9, Lemma. 2, p. 114]

$$\int_0^1 kx^{k-1}\ln^2(1-x)\,dx = H_k^2 + H_k^{(2)}, \tag{2.4}$$

in the following manner. Making the substitution $x \mapsto 1-x$ in I, we obtain

$$I = \int_0^1 \frac{\ln^2(1-x)\,\mathrm{Li}_2\left(\frac{x}{2}\right)}{x}dx = \int_0^1 \frac{\ln^2(1-x)}{x}\left(\sum_{k=1}^{\infty}\frac{\left(\frac{x}{2}\right)^k}{k^2}\right)dx,$$

where the definition of the polylogarithm has been used.

We may interchange the order of summation and integration and obtain

$$\sum_{k=1}^{\infty}\frac{1}{2^k k^2}\left(\int_0^1 \ln^2(1-x)\,x^{k-1}dx\right).$$

Using formula (2.4), we obtain

$$\int_0^1 \frac{\ln^2(x)}{1-x}\,\mathrm{Li}_2\left(\frac{1-x}{2}\right)dx = \sum_{k=1}^{\infty}\frac{H_k^2}{2^k k^3} + \sum_{k=1}^{\infty}\frac{H_k^{(2)}}{2^k k^3}$$

holds true. The following two series evaluations are known:

$$\sum_{k=1}^{\infty}\frac{H_k^2}{2^k k^3} = -2\,\mathrm{Li}_5\left(\frac{1}{2}\right) + \frac{279}{64}\zeta(5) - \mathrm{Li}_4\left(\frac{1}{2}\right)\ln 2 - \frac{37}{16}\zeta(4)\ln 2 - \frac{9}{16}\zeta(2)\zeta(3) + \frac{7}{16}\zeta(3)\ln^2 2 + \frac{1}{12}\zeta(2)\ln^3 2 - \frac{1}{40}\ln^5 2;$$

and

$$\sum_{k=1}^{\infty}\frac{H_k^{(2)}}{2^k k^3} = -2\,\mathrm{Li}_5\left(\frac{1}{2}\right) + \frac{23}{64}\zeta(5) - 3\,\mathrm{Li}_4\left(\frac{1}{2}\right)\ln 2 - \frac{1}{16}\zeta(4)\ln 2 + \frac{23}{16}\zeta(2)\zeta(3)$$

$$-\frac{23}{16}\zeta(3)\ln^2 2+\frac{7}{12}\zeta(2)\ln^3 2-\frac{13}{120}\ln^5 2.$$

These evaluations may be found in Olaikhan [11, p. 260, Eq. (4.175); p. 258, Eq. (4.171)].

It follows that

$$\begin{aligned}\int_0^1 \frac{\ln^2(x)\operatorname{Li}_2\left(\frac{1-x}{2}\right)}{1-x}dx &= -2\operatorname{Li}_5\left(\frac{1}{2}\right)+\frac{279}{64}\zeta(5)-\operatorname{Li}_4\left(\frac{1}{2}\right)\ln 2\\ &-\frac{37}{16}\zeta(4)\ln 2-\frac{9}{16}\zeta(2)\zeta(3)+\frac{7}{16}\zeta(3)\ln^2 2+\frac{1}{12}\zeta(2)\ln^3 2-\frac{1}{40}\ln^5 2\\ &-2\operatorname{Li}_5\left(\frac{1}{2}\right)+\frac{23}{64}\zeta(5)-3\operatorname{Li}_4\left(\frac{1}{2}\right)\ln 2\\ &-\frac{1}{16}\zeta(4)\ln 2+\frac{23}{16}\zeta(2)\zeta(3)-\frac{23}{16}\zeta(3)\ln^2 2+\frac{7}{12}\zeta(2)\ln^3 2-\frac{13}{120}\ln^5 2\\ &=-4\operatorname{Li}_5\left(\frac{1}{2}\right)+\frac{151}{32}\zeta(5)-4\operatorname{Li}_4\left(\frac{1}{2}\right)\ln 2\\ &-\frac{19}{8}\zeta(4)\ln 2+\frac{7}{8}\zeta(2)\zeta(3)-\zeta(3)\ln^2 2+\frac{2}{3}\zeta(2)\ln^3 2-\frac{2}{15}\ln^5 2.\end{aligned}$$

In addition to the preceding series evaluations, we record two further integral formulas due to Olaikhan [11, p. 128, Eq. (3.3); p. 149, Eq. (3.87)] (after applying a change of variables to the first formula)

$$\int_0^1 \frac{\ln^2(1-x)}{x}dx=2\zeta(3)\quad\text{and}$$

$$\int_0^1 \frac{\ln^a(x)\ln(1-x)}{1-x}dx=(-1)^a a!\left(\zeta(a+2)-\sum_{n=1}^{\infty}\frac{H_n}{n^{a+1}}\right),\quad a\in\mathbb{Z}^+,$$

where the first integral gives value of K, while for $a=2$ the second integral reduces to $2\left(\zeta(4)-\sum_{n=1}^{\infty}\frac{H_n}{n^3}\right)$.

In the February 1955 issue of the American Mathematical Monthly [1, pp. 129–130], Whittaker and Klamkin derived the following linear harmonic sum:

$$\sum_{n=1}^{\infty} \frac{H_n}{n^3} = \frac{5}{4}\zeta(4).$$

Thus, we obtain the value of the integral $J = 2\left(\zeta(4) - \frac{5}{4}\zeta(4)\right)$

$$= -\frac{1}{2}\zeta(4).$$

The remaining integral L can be evaluated as follows.

$$L = \int_0^1 \frac{\ln^2(x)\,\mathrm{Li}_2(-x)}{1-x}dx = \int_0^1 \frac{\ln^2(x)}{1-x}\left(\sum_{k=1}^{\infty}\frac{(-x)^k}{k^2}\right)dx$$
$$= \sum_{k=1}^{\infty}\frac{(-1)^k}{k^2}\left(\int_0^1 \frac{\ln^2(x)\,x^k}{1-x}dx\right).$$

The interchange that has been made here between the integration sign and the summation is permissible due to the dominated convergence theorem. Enforcing a substitution of $x = e^{-u}$ we find

$$L = \sum_{k=1}^{\infty}\frac{(-1)^k}{k^2}\left(-\int_{\infty}^0 \frac{\ln^2(e^{-u})\left(e^{-k}\right)^u(e^{-u})}{1-e^{-u}}du\right)$$
$$= \sum_{k=1}^{\infty}\frac{(-1)^k}{k^2}\left(\int_0^{\infty}\frac{u^2e^{-(k+1)u}}{1-e^{-u}}du\right)$$
$$= \sum_{k=1}^{\infty}\frac{(-1)^k}{k^2}\left(\Gamma(3)\cdot\zeta(3,k+1)\right)$$
$$= \Gamma(3)\sum_{k=1}^{\infty}\frac{(-1)^k}{k^2}\left(\zeta(3) - H_k^{(3)}\right)$$
$$= 2\left(\zeta(3)\sum_{k=1}^{\infty}\frac{(-1)^k}{k^2} - \sum_{k=1}^{\infty}\frac{(-1)^k H_k^{(3)}}{k^2}\right)$$
$$= -\zeta(2)\zeta(3) - \frac{21}{16}\zeta(5) + \frac{3}{2}\zeta(2)\zeta(3)$$
$$= \frac{1}{2}\zeta(2)\zeta(3) - \frac{21}{16}\zeta(5),$$

where the sixth equality uses the well-known result for the alternating Basel problem of $\sum_{k=1}^{\infty}\frac{(-1)^k}{k^2}=-\frac{\zeta(2)}{2}$ [2] and alternating linear harmonic sum $\sum_{k=1}^{\infty}\frac{(-1)^k H_k^{(3)}}{k^2}=\frac{21}{32}\zeta(5)-\frac{3}{4}\zeta(2)\zeta(3)$ [7, p. 703, Equation. (4.2d)].

Substituting the preceding evaluations into the original series gives

$$\begin{aligned}
2\sum_{n=1}^{\infty}\frac{\overline{H}_n}{n}\left(\zeta(3)-1-\frac{1}{2^3}-\cdots-\frac{1}{n^3}\right) &= -4\,\mathrm{Li}_5\left(\frac{1}{2}\right)+\frac{151}{32}\zeta(5)\\
&-4\,\mathrm{Li}_4\left(\frac{1}{2}\right)\ln 2-\frac{19}{8}\zeta(4)\ln 2+\frac{7}{8}\zeta(2)\zeta(3)-\zeta(3)\ln^2 2+\frac{2}{3}\zeta(2)\ln^3 2\\
&-\frac{2}{15}\ln^5 2-2\zeta(3)\,\mathrm{Li}_2\left(\frac{1}{2}\right)-\frac{1}{2}\zeta(2)\zeta(3)+\frac{21}{16}\zeta(5)+\frac{1}{2}\zeta(4)\ln 2\\
&= -4\,\mathrm{Li}_5\left(\frac{1}{2}\right)-4\,\mathrm{Li}_4\left(\frac{1}{2}\right)\ln 2\\
&+\frac{193}{32}\zeta(5)-\frac{2}{15}\ln^5 2+\frac{2}{3}\zeta(2)\ln^3 2-\frac{15}{8}\zeta(4)\ln 2-\zeta(3)\ln^2 2+\frac{3}{8}\zeta(2)\zeta(3)\\
&-2\zeta(3)\left(\frac{1}{2}\zeta(2)-\frac{1}{2}\ln^2(2)\right)\\
&= -4\,\mathrm{Li}_5\left(\frac{1}{2}\right)-4\,\mathrm{Li}_4\left(\frac{1}{2}\right)\ln 2\\
&+\frac{193}{32}\zeta(5)-\frac{2}{15}\ln^5 2+\frac{2}{3}\zeta(2)\ln^3 2-\frac{15}{8}\zeta(4)\ln 2-\frac{5}{8}\zeta(2)\zeta(3).
\end{aligned}$$

This is Nguyen's identity and completes the proof.

□

Acknowledgement: The author is appreciative of the referee's thoughtful and detailed feedback, which improved the final version.

References

[1] Whittaker, J. V., Klamkin, M. S., Another Summation, Solution to Problem 4564, *Amer. Math. Monthly* **62**(2) (1955), 126-130.

[2] Stark, E. L., Another proof of the formula $\sum_{k=1}^{\infty}\frac{1}{k^2}=\frac{\pi^2}{6}$, *Amer. Math. Monthly* **76**(5) (1969), 552-553.

[3] Lewin, L., Polylogarithms and associated functions, *North-Holland Publishing Co.*, New York, 1981.

[4] Titchmarsh, E. C., The theory of the Riemann zeta-function, (*Oxford University (Clarendon) Press,* 1951); 2nd edition (revised by D. R. Heath-Brown), 1986.

[5] Apostol, T. M., Mathematical Analysis, 2nd edn., Addison-Wesley Publishing Co., Reading, Mass.-London-Don Mills, Ont., 1974.

[6] Adamchik, V. S., Srivastava, H. M., Some series of the zeta and related functions, *Analysis* **18**(2) (1998), 131-144.

[7] Zheng, D. -Y., Further summation formulae related to generalized harmonic numbers, *J. Math. Anal. Appl.* **335**(1) (2007), 692-706.

[8] Boyadzhiev, K. N., Power series with skew-harmonic numbers, dilogarithms, and double integrals, *Tatra Mt. Math. Publ.* **56**(1) (2013), 93-108.

[9] Vălean, C. I., Furdui, O., Reviving the quadratic series of Au-Yeung, *J. Class. Anal.* **6**(2) (2015), 113-118.

[10] Furdui, O., Sîntămărian, A., Sharpening mathematical analysis skills, Problem Books in Mathematics, Springer, Cham, 1st ed., 2021.

[11] Olaikhan, A. S., An Introduction to the Harmonic Series and Logarithmic Integrals: For High School Students up to Researchers, Ali Shadhar Olaikhan, *Phoenix*, AZ, 2021.

[12] Dasireddy, N. S., On an Identity for classical harmonic numbers, *Math. Student* **94**(3-4) (2025), 78-84.

[13] Nguyen, V., On some series involving harmonic and skew-harmonic numbers, *J. Class. Anal.* **22**(1) (2023), 25-43.

Nandan Sai Dasireddy
House No. 3-11-363, Pavani Plaza, Flat No:401,
Road No. 2, Shivaganga Colony, Siris Road,
L. B. Nagar, Hyderabad, Telangana, India.
E-mail: dasireddy.1818@gmail.com

The Mathematics Student ISSN: 0025-5742
Vol. 95, Nos. 3-4, July - December (2026), 134–139

COMPARISON OF SOME SERIES CONTAINING EULER NUMBER VIA APERY'S CONSTANT

ILIR DEMIRI AND SHPETIM REXHEPI
(Received : 28 - 07 - 2025 ; Accepted : 08 - 07 - 2026)

Abstract. This paper is devoted to the comparison of some sums of series, most of them involving the number e – *Euler number*, with the Apery's constant, respectively $\zeta(3)$, where ζ represents the Riemann zeta function, but firstly is given an inequality, and then we have achieved our desired results from here.

1. Introduction

In 1734, Euler solved Basel's problem, so he found the sum of reciprocals of the squares of natural numbers, and he proved that the sum of the series [6], [7]

$$\sum_{n=1}^{\infty} \frac{1}{n^2} = \frac{\pi^2}{6}, \text{ (see [3])}.$$

Naturally, at that time, the question was posed: "What will happen with the sum of reciprocals of the cubes of natural numbers?"
In this question answered, Roger Apery, who has proved the irrationality of $\zeta(3)$, where the standard definition is [cite: 10]

$$\zeta(3) = \sum_{n=1}^{\infty} \frac{1}{n^3}$$

but this expression converges very slowly, for that reason Apery used another form (definition) to prove the irrationality of $\zeta(3)$, which is

$$\zeta(3) = \frac{5}{2} \sum_{k=1}^{\infty} (-1)^{k-1} \frac{1}{k^3 \binom{2k}{k}}$$

Then, for this constant, many formulas were discovered. (see [4])

2020 Mathematics Subject Classification: 11A41, 16N20
Key words and phrases: Apery's constant, Euler Number, Jensen's inequality, infinite series

© *Indian Mathematical Society, 2026.*

2. Auxiliary facts

Next, we give some known results that we have used to prove our results.

Definition 2.1. Let $f : (a, b) \to \mathbb{R}$ and for any $x \in (a, b)$. Suppose there exists a second derivative $f''(x)$. The function f is convex on (a, b) if and only if for each $x \in (a, b)$ we have $f''(x) \geq 0$. If $f''(x) > 0$ for each $x \in (a, b)$, then f is strictly convex on (a, b).

Definition 2.2. Let $a_1, a_2, ..., a_n$ be positive real numbers. The numbers

i)
$$AM = \frac{a_1 + a_2 + \cdots + a_n}{n}$$
is called the arithmetic mean.

ii)
$$GM = \sqrt[n]{a_1 a_2 \cdot \cdots \cdot a_n}$$
is called the geometric mean.

Lemma 2.3 (Jensen's inequality)**.** *Let $f : (a, b) \to \mathbb{R}$ be a convex function on the interval (a, b). Let $n \in \mathbb{N}$ and $\alpha_1, \alpha_2, \ldots, \alpha_n \in (0, 1)$ be real numbers such that $\alpha_1 + \alpha_2 + \cdots + \alpha_n = 1$. Then for any $x_1, x_2, \ldots, x_n \in (a, b)$ we have*
$$f\left(\sum_{i=1}^{n} \alpha_i x_i\right) \leq \sum_{i=1}^{n} \alpha_i f(x_i)$$

Lemma 2.4. *The inequality*
$$AM \geq GM$$
holds for every set $a_1, a_2, \cdots, a_n$ of positive real numbers. Equality occurs if and only if $a_1 = a_2 = a_3 = \cdots = a_n$.

Lemma 2.5. *Let f be defined on the closed and bounded interval $[a, b]$, continuous in $[a, b]$ and differentiable (at least) on (a, b). Then there is a point $x_0 \in (a, b)$ such that $f'(x_0) = \frac{f(b)-f(a)}{b-a}$.*

3. Main results

Proposition 3.1. *If $a_i \in \mathbb{R}$ and $a_i > 1, i = 1, 2, \ldots, n$ the inequality*
$$\frac{\sqrt{\ln a_1} + \sqrt{\ln a_2} + \cdots + \sqrt{\ln a_n}}{n} \leq \sqrt{\ln \frac{a_1 + a_2 + \cdots + a_n}{n}}$$

Proof. Let $f : \mathbb{R}^+ \to \mathbb{R}$ be a function is given by $f(x) = e^{x^2}$, e - Euler number, then $f''(x) = 4x^2e^{x^2} + 2$, which is greater than 0 and so it is convex. So, by Jensen's inequality, we have:

$$e^{\left(\frac{x_1+x_2+\cdots+x_n}{n}\right)^2} \leq \frac{e^{x_1^2} + e^{x_2^2} + \cdots + e^{x_n^2}}{n}, x_i \in \mathbb{R}^+, i = 1, 2, \ldots, n$$

$$\left(\frac{x_1 + x_2 + \cdots + x_n}{n}\right)^2 \leq \ln \frac{e^{x_1^2} + e^{x_2^2} + \cdots + e^{x_n^2}}{n}$$

If we take $x_i = \sqrt{\ln a_i}, a_i > 1, i = 1, 2, 3, \ldots, n$ we will achieve to

$$\left(\frac{\sqrt{\ln a_1} + \sqrt{\ln a_2} + \cdots + \sqrt{\ln a_n}}{n}\right)^2 \leq \ln \frac{a_1 + a_2 + \cdots + a_n}{n}$$

$$\frac{\sqrt{\ln a_1} + \sqrt{\ln a_2} + \cdots + \sqrt{\ln a_n}}{n} \leq \sqrt{\ln \frac{a_1 + a_2 + \cdots + a_n}{n}},$$

which is the desired result. □

Proposition 3.2. *The following relation is valid*

$$\sum_{n=1}^{\infty} \frac{e^{\left(\frac{n+1}{2}\right)^2 - \ln n^2}}{e + e^4 + e^9 + \cdots + e^{n^2}} \leq \zeta(3), e - \textit{Euler number}$$

Proof. If, in relation to Proposition 3.1, we take substitution $a_n = e^{n^2}, n \in \mathbb{N}, e - \textit{Euler number}$, we take:

$$\frac{\sqrt{\ln e^{1^2}} + \sqrt{\ln e^{2^2}} + \cdots + \sqrt{\ln e^{n^2}}}{n} \leq \sqrt{\ln \frac{e + e^4 + e^9 + \cdots + e^{n^2}}{n}}$$

$$\left(\frac{n(n+1)}{2n}\right)^2 \leq \ln \frac{e + e^4 + e^9 + \cdots + e^{n^2}}{n}$$

$$e^{\left(\frac{n+1}{2}\right)^2} \leq \frac{e + e^4 + e^9 + \cdots + e^{n^2}}{n}$$

$$\frac{e^{\left(\frac{n+1}{2}\right)^2}}{e + e^4 + e^9 + \cdots + e^{n^2}} \leq \frac{1}{n}$$

$$\frac{e^{\left(\frac{n+1}{2}\right)^2 - \ln n^2}}{e + e^4 + e^9 + \cdots + e^{n^2}} \leq \frac{1}{n^3}, n \in \mathbb{N}$$

$$\sum_{n=1}^{\infty} \frac{e^{\left(\frac{n+1}{2}\right)^2 - \ln n^2}}{e + e^4 + e^9 + \cdots + e^{n^2}} \leq \sum_{n=1}^{\infty} \frac{1}{n^3}$$

$$\sum_{n=1}^{\infty} \frac{e^{\left(\frac{n+1}{2}\right)^2 - \ln n^2}}{e + e^4 + e^9 + \cdots + e^{n^2}} \leq \zeta(3),$$

which completes the proof. □

Proposition 3.3. *The following relation is valid*

$$\sum_{n=1}^{\infty} \frac{e^{\frac{1+\sqrt{2}+\sqrt[3]{3}+\cdots+\sqrt[n]{n}}{n^2}-\ln n^2}}{e+e^{\sqrt{2}}+e^{\sqrt[3]{3}}+\cdots+e^{\sqrt[n]{n}}} \leq \zeta(3), e - \textit{Euler number.}$$

Proof. In Proposition 3.1, we take the substitution $a_n = e^{\sqrt[n]{n}}$, and will achieve the following inequalities:

$$\frac{\sqrt{\sqrt{1}+\sqrt{2}+\sqrt[3]{3}+\cdots+\sqrt[n]{n}}}{n} \leq \sqrt{\ln \frac{e+e^{\sqrt{2}}+e^{\sqrt[3]{3}}+\cdots+e^{\sqrt[n]{n}}}{n}}$$

$$\frac{1+\sqrt{2}+\sqrt[3]{3}+\cdots+\sqrt[n]{n}}{n^2} \leq \ln \frac{e+e^{\sqrt{2}}+e^{\sqrt[3]{3}}+\cdots+e^{\sqrt[n]{n}}}{n}$$

$$e^{\frac{1+\sqrt{2}+\sqrt[3]{3}+\cdots+\sqrt[n]{n}}{n^2}} \leq \frac{e+e^{\sqrt{2}}+e^{\sqrt[3]{3}}+\cdots+e^{\sqrt[n]{n}}}{n}$$

$$\frac{e^{\frac{1+\sqrt{2}+\sqrt[3]{3}+\cdots+\sqrt[n]{n}}{n^2}-\ln n^2}}{e+e^{\sqrt{2}}+e^{\sqrt[3]{3}}+\cdots+e^{\sqrt[n]{n}}} \leq \frac{1}{n^3}$$

$$\sum_{n=1}^{\infty} \frac{e^{\frac{1+\sqrt{2}+\sqrt[3]{3}+\cdots+\sqrt[n]{n}}{n^2}-\ln n^2}}{e+e^{\sqrt{2}}+e^{\sqrt[3]{3}}+\cdots+e^{\sqrt[n]{n}}} \leq \zeta(3),$$

as desired. □

Proposition 3.4. *The following relation is valid*

$$\sum_{n=1}^{\infty} \frac{e^{\sqrt[n]{n!}-\ln n^2}}{e+e^2+e^3+\cdots+e^n} \leq \zeta(3), e - \textit{Euler number}$$

Proof. From Proposition 3.1 and using the fact that $GM \leq AM$, we acquire the inequality

$$\sqrt[n]{\sqrt{\ln a_1}\sqrt{\ln a_2}\ldots\sqrt{\ln a_n}} \leq \sqrt{\ln \frac{a_1+a_2+\cdots+a_n}{n}}$$

$$\sqrt[2n]{\ln a_1 \ln a_2 \ldots \ln a_n} \leq \sqrt{\ln \frac{a_1+a_2+\cdots+a_n}{n}}$$

$$\sqrt[n]{\ln a_1 \ln a_2 \ldots \ln a_n} \leq \ln \frac{a_1+a_2+\cdots+a_n}{n}$$

Let us take $a_n = e^n, n \in \mathbb{N}, e - \textit{Euler number}$, then:

$$\sqrt[n]{1 \cdot 2 \cdot 3 \cdot \cdots \cdot n} \leq \ln \frac{e+e^2+e^3+\cdots+e^n}{n}$$

$$\frac{e^{\sqrt[n]{n!}}}{e+e^2+\cdots+e^n} \leq \frac{1}{n}$$
$$\frac{e^{\sqrt[n]{n!}-\ln n^2}}{e+e^2+\cdots+e^n} \leq \frac{1}{n^3}$$
$$\sum_{n=1}^{\infty} \frac{e^{\sqrt[n]{n!}-\ln n^2}}{e+e^2+\cdots+e^n} \leq \zeta(3),$$

which is the desired result. □

Proposition 3.5. *The following inequality*

$$\sum_{n=1}^{\infty} \frac{e^{\frac{1}{n+1}}-1}{n^2} \leq \zeta(3)$$

holds, where e-is the Euler number.

Proof. If we apply Lemma 2.3 to the function $f : \mathbb{R}^+ \to \mathbb{R}$, which is given by $f(x) = \ln x$ in interval $[n, n+1]$, where $n \in \mathbb{N}$, we acquire the inequality

$$e^{\frac{1}{n+1}}-1 < \frac{1}{n}$$
$$\frac{e^{\frac{1}{n+1}}-1}{n^2} < \frac{1}{n^3}$$
$$\sum_{n=1}^{\infty} \frac{e^{\frac{1}{n+1}}-1}{n^2} < \zeta(3),$$

which proves the result. □

References

[1] Cvetkovski, Z., Inequalities (theorems, techniques, and selected problems), *Springer*.

[2] Canuto, C. Tabacco, A., Mathematical Analysis I, *Springer*.

[3] Fairbanks, L., Notes on an approach to Apery's Contsant, *Arxiv* (2024).

[4] Raman, A., On Apery's constant and Catalan's Constant, University of California, Berkeley, 2010.

[5] Demiri, I., Shaini, B., Rexhepi, S., Iseni, E., On some new properties for the gamma functions. *Asian Research Journal of Mathematics*, **19(10)** (2023), 124-129.

[6] Demiri, I., Rexhepi, S., Some approximations of the Euler number. *The Teaching of Mathematics*, **23(1)** (2020), 51-56.

[7] Demiri, I., Rexhepi, S., Iseni, E., Comparison of some new approximations with the Leibniz formula regarding their convergence to the Archimedes' constant. *Electronic Journal of Mathematical Analysis and Applications*, **10(1)** , (2022), 122-128.

[8] Demiri, I., Rexhepi, S., SOME INEQUALITIES INVOLVING THE DIGAMMA FUNCTION. *The Mathematics Student*, **94(1-2)**,(2025), 137-141.

Ilir Demiri
Department of Mathematics
Faculty of Economics
Mother Theresa University, Skopje, North Macedonia.
E-mail: ilir.demiri@unt.edu.mk

Shpetim Rexhepi
Department of Mathematics
Faculty of Civil Engineering and Architecture
Mother Theresa University, Skopje, North Macedonia.
E-mail: shpetim.rexhepi@unt.edu.mk

The Mathematics Student ISSN: 0025-5742
Vol. 95, Nos. 3-4, July - December (2026), 140–149

A NOTE ON FIEDLER VECTOR MONOTONICITY BEYOND R-MATRICES

HARIPRASAD, M.

(Received : 30 - 07 - 2025 ; Accepted : 07 - 10 - 2026)

Abstract. We extend a classical result of Atkins, Boman, and Hendrickson, which established that the Fiedler vector of an R-matrix is monotonic and supports spectral seriation. In this note, we show that the monotonicity property persists for a broader class of matrices, specifically those whose row-integrated form is a column-R-matrix. This generalization expands the structural conditions under which spectral ordering techniques remain valid.

1. Introduction

The Fiedler vector, an eigenvector corresponding to the second-smallest eigenvalue of a graph Laplacian matrix, plays a key role in graph layout, seriation, and clustering tasks [2, 1, 3]. Seriation techniques continue to find applications in the areas such as ranking [4], data visualization [6], and the recognition of noisy structural patterns in matrices [5]. A foundational result in spectral seriation is that for R-matrices, the entries of the Fiedler vector are monotonic [1]. In this note, we extend that result to a broader class of matrices that also preserve this spectral ordering property.

1.1. **Notations and preliminary results.** Let I_n denote the identity matrix of order n. Let $\mathbf{1}_n$ be the column vector of all ones of order n. Let e_j be the jth canonical basis vector with $e_j(j) = 1$ and zero otherwise. Let $[n] = \{1, 2, 3, \cdots, n\}$. The Laplacian matrix is $L = D - A$, where

2020 Mathematics Subject Classification: 15A18, 05C50, 05C22
Key words and phrases: spectral seriation, Fiedler vector, graph Laplacian, monotonicity

© *Indian Mathematical Society, 2026.*

$D(i,i) = \sum_{j=1}^{n} A(i,j)$. The Laplacian is symmetric and positive semidefinite. The eigenvector (denoted by u) corresponding to the second smallest eigenvalue is called the Fiedler vector. The corresponding eigenvalue is known as the Fiedler value (denoted by λ_2). A symmetric matrix $A \in \mathbb{R}^{n\times n}$ is an R-matrix if: $A(i,j) \geq A(i+1,j)$ for $i > j$, and $A(i,j) \leq A(i+1,j)$ for $i < j$.

Next, we state the Perron-Frobenius theorem.

Theorem 1.1. [Perron–Frobenius] *Let M be a real nonnegative matrix, if we define $\rho(M) = \max_i |\lambda_i(M)|$, then*

- *$\rho(M)$ is an eigenvalue of M.*
- *There is a vector $x \geq 0$ such that $Mx = \rho(M)x$.*

2. Results

We begin this section with the necessary definitions and motivating examples, followed by the proof of the main theorem establishing the existence of a monotone vector in the Fiedler eigenspace.

Definition 2.1. For an $n \times n$ matrix A, we define the **row-integrated adjacency matrix** B as follows:

$$B(i,j) = \begin{cases} \sum_{q=1}^{j} A(i,q) \text{ if } i > j, \\ \sum_{q=j}^{n} A(i,q) \text{ if } i < j. \end{cases}$$

Note that we do not define the diagonal entries of the matrix B, since the Laplacian of A does not depend on the diagonal entries of A.

Definition 2.2. We say a row-integrated matrix B is a **column R-matrix** if $B(i,j) \geq B(i+1,j)$ for all $i > j$ and $B(i-1,j) \leq B(i,j)$ for all $i < j$.

Example 2.3. Consider the matrix A and its row integrated matrix B given by

$$A = \begin{bmatrix} 0 & 1 & 1 & 0 \\ 1 & 0 & 0 & 1 \\ 1 & 0 & 0 & 1 \\ 0 & 1 & 1 & 0 \end{bmatrix}, \qquad B = \begin{bmatrix} * & 2 & 1 & 0 \\ 1 & * & 1 & 1 \\ 1 & 1 & * & 1 \\ 0 & 1 & 2 & * \end{bmatrix}.$$

Note that A is not an R-matrix whereas B is a column R-matrix.
The corresponding Laplacian has Fiedler eigenvalue $\lambda_2 = 2$, with a two-dimensional Fiedler eigenspace

$$\text{span}\left\{\begin{bmatrix}-0.4081\\ 0.5775\\ -0.5775\\ 0.4081\end{bmatrix}, \begin{bmatrix}0.5775\\ 0.4081\\ -0.4081\\ -0.5775\end{bmatrix}\right\}.$$

Observe that the second basis vector $[0.5775, 0.4081, -0.4081, -0.5775]^T$ is monotone. Although the Fiedler eigenspace is two-dimensional, there is a monotonic vector in this eigenspace.

Example 2.4. Consider the graph with adjacency matrix A and the corresponding row-integrated matrix B given by

$$A = \begin{bmatrix}0&1&1&0&0&0&0\\1&0&0&1&0&0&0\\1&0&0&1&0&0&0\\0&1&1&0&1&0&0\\0&0&0&1&0&1&0\\0&0&0&0&1&0&1\\0&0&0&0&0&1&0\end{bmatrix}, \qquad B = \begin{bmatrix}0&2&1&0&0&0&0\\1&0&1&1&0&0&0\\1&1&0&1&0&0&0\\0&1&2&0&1&0&0\\0&0&0&1&0&1&0\\0&0&0&0&1&0&1\\0&0&0&0&0&1&0\end{bmatrix}.$$

Its Laplacian matrix is

$$L = D - A = \begin{bmatrix}2&-1&-1&0&0&0&0\\-1&2&0&-1&0&0&0\\-1&0&2&-1&0&0&0\\0&-1&-1&3&-1&0&0\\0&0&0&-1&2&-1&0\\0&0&0&0&-1&2&-1\\0&0&0&0&0&-1&1\end{bmatrix}.$$

The eigenvalues of L are $0, 0.2765, 1.3323, 2, 2.5219, 3.2920$, and 4.5772. Hence the Fiedler value is $\lambda_2 = 0.2765$, with Fiedler vector

$$v_F = \begin{bmatrix}-0.3801, & -0.3275, & -0.3275, & -0.1844, & 0.1528, & 0.4478, & 0.6190\end{bmatrix}^T.$$

Observe that $-0.3801 < -0.3275 = -0.3275 < -0.1844 < 0.1528 < 0.4478 < 0.6190$, so the Fiedler vector is monotone up to the repeated entries corresponding to the symmetric vertices 2 and 3. Consequently, it

induces the natural ordering $1 \prec \{2,3\} \prec 4 \prec 5 \prec 6 \prec 7$, which agrees with the geometric structure of the graph.

The examples above show that the class of matrices considered here strictly contains the class of R-matrices. In each example, the adjacency matrix is not an R-matrix (equivalently, it violates the Robinson monotonicity conditions in its given ordering), so the classical monotonicity theorem of Atkins–Boman–Hendrickson does not apply. However, the corresponding row-integrated matrix is a column-R-matrix. Hence, our result generalizes the result in [1]. These examples show that the column-R property of the row-integrated matrix ensures Fiedler monotonicity. The following theorem formalizes this observation.

Theorem 2.5. *If for the adjacency matrix A, the row-integrated adjacency matrix B is a column R-matrix, then the Fiedler eigenspace of the Laplacian corresponding to A contains a monotone vector.*

Proof. To prove that the Fiedler vector is nondecreasing, we show that the differences between consecutive entries are nonnegative. Consider the row-integrated adjacency matrix B (by Definition 2.1), we assume it is also a column R-matrix (according to Definition 2.2).

Following the argument of Atkins, Boman, and Hendrickson [1], we consider the matrices

$$S = \begin{bmatrix} -1 & 1 & 0 & \cdots & 0 \\ 0 & -1 & 1 & \ddots & \vdots \\ \vdots & \ddots & \ddots & \ddots & 0 \\ 0 & \cdots & 0 & -1 & 1 \end{bmatrix}, \qquad T = \begin{bmatrix} 0 & 0 & \cdots & 0 \\ 1 & 0 & \cdots & 0 \\ 1 & 1 & \ddots & \vdots \\ \vdots & \vdots & \ddots & 0 \\ 1 & 1 & \cdots & 1 \end{bmatrix}.$$

We have $ST = I_{n-1}$ and $TS = I_n - \mathbf{1}_n e_1^T$. Let u be the Fiedler vector with Fiedler value λ, such that $Lu = \lambda u$. Then,

$$\begin{aligned} Lu = \lambda u, \quad SLTSu &= SL(I_n - \mathbf{1}_n e_1^T)u \text{ (using } TS = I_n - \mathbf{1}_n e_1^T), \\ &= \lambda Su \text{ (using } Lu = \lambda u \text{ and } L\mathbf{1}_n = \mathbf{0}). \end{aligned}$$

Therefore, Su is an eigenvector of SLT. Let $M = SLT$. Then dimension of M is one less than that of L. We have

$$M(Su) = \lambda Su. \tag{2.1}$$

In general, every eigenvector u_j for $j > 1$ with $Lu_j = \lambda_j u_j$, we have $M(Su_j) = \lambda_j Su_j$. Also note that $\mathbf{1}_n^T u_j = 0$ for $j > 1$. Next we prove the non-diagonal entries of M are non-positive, we examine the entries of the matrix, $(SL)(i,j) = L(i+1,j) - L(i,j)$.

$$\text{This gives,} \qquad M(i,j) = \sum_{q=j+1}^{n} L(i+1,q) - L(i,q),$$

$$M(i,j) = \sum_{q=1}^{j} L(i,q) - L(i+1,q) \qquad (\text{using } \sum_j L(i,j) = 0).$$

For $i < j$ we have

$$\begin{aligned} M(i,j) &= \sum_{q=j+1}^{n} (A(i,q) - A(i+1,q)) \qquad (\text{using } L(i,j) = -A(i,j) \text{ for } i \neq j), \\ &= B(i,j+1) - B(i+1,j+1) \qquad (\text{by definition of } B). \end{aligned}$$

For $i > j$ we have

$$\begin{aligned} M(i,j) &= \sum_{q=1}^{j} (A(i+1,q) - A(i,q)) \quad (\text{using } L(i,j) = -A(i,j) \text{ for } i \neq j), \\ &= B(i+1,j) - B(i,j) \qquad (\text{by definition of } B). \end{aligned}$$

Thus we get the condition $B(i,j) \geq B(i+1,j)$ for all $i > j$ and $B(i-1,j) \leq B(i,j)$ for all $i < j$, this condition is satisfied if B is a column R-matrix.

Note that the non diagonal entries of M are non positive. Let $\alpha \geq \max_{i,j}\{M(j,j), \lambda_i\}$. Then every entry of $\alpha I_{n-1} - M$ is non-negative. Its eigenvalues are $\alpha - \lambda_i > 0$. The set $\{\alpha - \lambda_i\}$ has a unique maximum value, let us denote it by β. From the Perron-Frobenius theorem 1.1, the largest eigenvalue β has an eigenvector $y \geq 0$. Then we have $(\alpha I - M)y = \beta y$. Thus $My = (\alpha - \beta)y$. Since α is fixed and β is largest, we have $\alpha - \beta$ as the smallest possible eigenvalue of M. Note that, eigenvalues of M are $\lambda_2, \lambda_3, \cdots, \lambda_n$, we have $\lambda = \alpha - \beta$ as the Fiedler value. The corresponding eigenvector $y = Su$ is nonnegative, therefore u is nondecreasing. □

If the Fiedler value is simple, then the Fiedler eigenspace is one-dimensional, and hence the Fiedler vector is monotone (up to sign) by Theorem 2.5.

3. Application to a geometric graph with gaps

Random geometric graphs with community structure have been widely studied for their relevance in modeling spatial networks and community detection tasks [7, 8]. It is shown that the spectral clustering works for dense random geometric graphs [10]. In the sparse regime, attempts have been made to recover communities using spectral seriation [9]. Here, we consider a special type of such graph. We consider a *geometric graph* consisting of k *communities*, where nodes are placed based on spatial intervals.

The intra-community connection threshold is denoted by r_{in}, and the inter-community connection threshold by r_{out}. Two nodes of the same community are connected if their distance is less than the intra-community threshold r_{in}.

Two nodes from different communities are connected if their distance is less than the inter-community threshold r_{out}, where typically $r_{\text{out}} < r_{\text{in}}$. In this example we consider, $r_{\text{in}} = 3r_{\text{out}}$.

Nodes of each community are placed in intervals of the form $((2t)r_{\text{out}}, (2t+1)r_{\text{out}})$ for $t \geq 0$, while the alternating intervals $[(2t+1)r_{\text{out}}, (2t+2)r_{\text{out}}]$ are kept empty, introducing *gaps* in the placement.

First, we consider a single node of each community placed in the intervals. Then we extend the model to m_i nodes placed randomly in specific intervals.

3.1. **Single nodes of each community in even intervals:** Let the interval [0,1] be divided into $2J$ intervals, and let $r_{\text{out}} = \frac{1}{2J}$. Starting from the zeroth interval, let each even indexed interval consist of k nodes, one node from each community. Let $\boldsymbol{J}_k = \mathbf{1}_k \mathbf{1}_k^T$. Then the adjacency matrix is given by,

$$A_b = \begin{bmatrix} \boldsymbol{J}_k & I_k & 0 & \cdots & 0 \\ I_k & \boldsymbol{J}_k & I_k & \ddots & \vdots \\ 0 & I_k & \boldsymbol{J}_k & \ddots & 0 \\ \vdots & \ddots & \ddots & \ddots & I_k \\ 0 & \cdots & 0 & I_k & \boldsymbol{J}_k \end{bmatrix} - I_{kJ}.$$

We denote the row-integrated matrix for $\mathbf{1}\mathbf{1}_{k\times k}^T$ as T where $T(i,j) = k-j+1$ for $j > i$ and $T(i,j) = j$ for $i > j$. Let V be a lower triangular matrix with $V(i,j) = 1$ for $i \geq j$. Then the row-integrated matrix B_b of the adjacency

A_b is given by

$$B_b = \begin{bmatrix} T+\boldsymbol{J}_k & V & 0 & \cdots & 0 \\ V^T & T+\boldsymbol{J}_k & V & \ddots & \vdots \\ 0 & V^T & T+\boldsymbol{J}_k & \ddots & 0 \\ \vdots & \ddots & \ddots & \ddots & V \\ 0 & \cdots & 0 & V^T & T+\boldsymbol{J}_k \end{bmatrix}.$$

Remark 3.1. B_b is a column R-matrix, hence the Laplacian corresponding to A_b has a monotonic Fiedler vector.

Proof. Let $\mathrm{supp}(B)$ denote the set of indices corresponding to the nonzero entries of the block B in the block matrix B_b. Let $i < j$, we have four cases,

- $(i,j),(i-1,j) \in \mathrm{supp}(T + \mathbf{1}\mathbf{1}^T_{k\times k})$ then $B_b(i,j) - B_b(i-1,j) = 0$.
- $(i,j) \in \mathrm{supp}(T + \mathbf{1}\mathbf{1}^T_{k\times k})$ and $(i-1,j) \in \mathrm{supp}(V)$, then $B_b(i,j) - B_b(i-1,j) > 0$.
- $\{(i,j),(i-1,j)\} \in \mathrm{supp}(V)$ then $B_b(i,j) - B_b(i-1,j) = 0$.
- $\{(i,j)\} \in \mathrm{supp}(V)$ and $\{(i-1,j)\} \notin \mathrm{supp}(V)$. then $B_b(i,j) - B_b(i-1,j) = 1 > 0$.

Similarly when $i > j$ it can be verified that $B_b(i,j) - B_b(i+1,j) \geq 0$ implying B_b is a column R-matrix. Then, by Theorem 2.5, we have the monotonic Fiedler vector. □

3.2. **Multiple nodes per interval.** Let m_j nodes of each community be placed in the interval $2j$, then j^{th} even interval has km_j nodes in total (Illustrated for two communities in Figure 1).

The resulting adjacency matrix of the graph is given by $A = W(A_b + I_{kJ})W^T - I_n$, where $A_b + I_{kJ}$ is a block-level adjacency matrix (as defined in the previous section) describing inter-block connections, and W is a matrix encoding the placement of m_j nodes in each block. The matrix W maps the smaller block-level model into the full adjacency matrix of the graph.

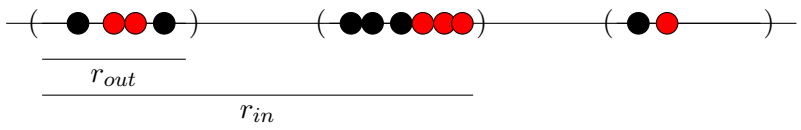

FIGURE 1. Node placement for the model.

The matrix W is given by

$$W = \left[e_1^T \otimes \mathbf{1}_{m_1\times 1}, \quad e_2^T \otimes \mathbf{1}_{m_1\times 1}, \quad \cdots, \quad e_{kJ-1}^T \otimes \mathbf{1}_{m_{kJ-1}\times 1}, \quad e_{kJ}^T \otimes \mathbf{1}_{m_{kJ}\times 1}\right]^T.$$

The Laplacian matrix is given by $L = D - A$, where D is the diagonal degree matrix with entries $D(i,i) = \sum_j A(i,j)$. Similarly, D_m is the degree matrix corresponding to the smaller Laplacian $L_m = D_m - (A_b + I_{kJ})(W^\top W) + I_{kJ}$, constructed from $(A_b + I_{kJ})(W^\top W) - I_{kJ}$.

Lemma 3.2. *We have $DW = WD_m$.*

Proof. Let d_{sj} denote the degree of a node in the jth block from the adjacency matrix A. Then we have,

$$d_{sj} = \begin{cases} km_j + m_{j-1} + m_{j+1} - 1 \text{ for } 1 < j < J, \\ km_j + m_{j+1} - 1 \text{ for } j = 1, \\ km_j + m_{j-1} - 1 \text{ for } j = J. \end{cases}$$

Let $\text{diag}(v)$ denote the diagonal matrix formed by the vector v, then we have

$$W^T W = \text{diag}(\begin{bmatrix} m_1 & m_2 & \cdots & m_J \end{bmatrix} \otimes \mathbf{1}_k^T).$$

So the degree of the matrix $(A_b + I_{kJ})(W^\top W) - I_{kJ}$ is given by,

$$d_{sj} = \begin{cases} km_j + m_{j-1} + m_{j+1} - 1 \text{ for } 1 < j < J, \\ km_j + m_{j+1} - 1 \text{ for } j = 1, \\ km_j + m_{j-1} - 1 \text{ for } j = J. \end{cases}$$

Every row corresponding to j^{th} block is multiplied by d_{sj} in DW which is equivalent to multiplying by d_{sj} to every column corresponding to block j, this leads to WD_m. □

Lemma 3.3. *If u, λ is an eigenpair of L_m, then Wu, λ is eigenpair of L.*

Proof. Consider,

$$\begin{aligned} LWu &= (DW - AW)u \\ LWu &= (WD_m - (W(A_b + I_{kJ})(W^T - I_n)W))u \\ LWu &= W(D_m - (A_b + I_{kJ})W^T W - I_{kJ})u \\ &= WL_m u = \lambda Wu. \end{aligned}$$

□

All the nodes of the same community c, in a particular interval, (corresponding to block j) have the same neighbourhood. Let their degree be d_{cj}. Then any vector u of length m_j, which is perpendicular to the all one vector $\mathbf{1}_{m_j}$ corresponds to an eigenvector v of L, with $v(i) = u(i)$ at the indices corresponding to block j and community c, and $v(i) = 0$ at other indices. Then we have $Lv = (d_{cj} + 1)v$. Let F denote the space of all such eigenvectors. With $v^T W = 0$. Note that $dim(F) + dim(range(W)) = n$.

Theorem 3.4. *If u_m is the Fiedler vector of L_m, then Wu_m is the Fiedler vector of L.*

Proof. The Fiedler value is less than or equal to the degree of every node of the graph. But eigenvectors from the space F have eigenvalues $d_{cj} + 1$ (which is 1 greater than all the individual degrees). Let δ be the minimum degree of the graph, then we have $\lambda_2(L) \leq \delta < d_{cj} + 1$. So the Fiedler value of L must be the same as the Fiedler value of L_m. □

Theorem 3.5. *Fiedler eigenspace of L_m has a monotonic vector.*

Proof. The proof follows by verifying that $(A_b + I_{kJ})(W^T W) - I_{kJ}$ is such that its row-integrated matrix is a column R-matrix.

Let T be the row-integrated matrix corresponding to $\mathbf{11^T}$, then it is easy to see that T is a column R-matrix. Then the row-integrated adjacency matrix corresponding to $(A_b + I_{kJ})W^\top W$ is given by

$$B_b = \begin{bmatrix} m_1T + m_2V^T & m_2V & 0 & 0 & \cdots & 0 \\ m_1V^T & m_2T + m_3V^T + m_1V & m_3V & 0 & \cdots & 0 \\ 0 & m_2V^T & \ddots & \ddots & \ddots & \vdots \\ \vdots & \ddots & \ddots & \ddots & \ddots & \vdots \\ 0 & \cdots & 0 & m_{J-2}V^T & m_{J-1}T + m_JV^T + m_{J-2}V & m_JV \\ 0 & 0 & \cdots & 0 & m_{J-1}V^T & m_JT_n + m_{J-1}V \end{bmatrix}.$$

By looking at the structure, similar to the proof of the Remark 3.1 we can verify the column-R property. Therefore Fiedler eigenspace of L_m has a monotonic vector. □

As a concluding remark, note that given the adjacency matrix of this graph in an arbitrary node order, one can compute the Fiedler vector of the corresponding Laplacian and sort its entries. This sorting recovers the underlying structure of the graph (reconstructing the adjacency matrix up to the interval-wise block ordering) effectively revealing the spatial layout in which the graph is embedded.

Concluding comments

We have extended the classical result on the monotonicity of Fiedler vectors from R-matrices to a broader class of matrices whose row-integrated forms are column-R-matrices. This generalization is useful for spectral seriation tasks on a broader class of graphs admitting such adjacency matrices. An application is illustrated using a geometric graph on a line with gaps.

References

[1] Atkins, J. E., Boman, E. G., and Hendrickson, B., A spectral algorithm for seriation and the consecutive ones problem, *SIAM Journal on Computing* **28** (1998), 297–310.

[2] Fiedler, M., Algebraic connectivity of graphs, *Czechoslovak Mathematical Journal* **23** (1973), 298–305.

[3] Spielman, D. A. and Teng, S.-H., Spectral partitioning works: Planar graphs and finite element meshes, *Linear Algebra and its Applications* **421** (2007), 284–305.

[4] Fogel, F., Jenatton, R., Bach, F., and d'Aspremont, A., Convex relaxations for permutation problems, *SIAM Journal on Matrix Analysis and Applications* **36** (2015), 1465–1488.

[5] Laurent, M., Seminaroti, M., and Tanigawa, S.-I., A structural characterization for certifying Robinsonian matrices, *Electronic Journal of Combinatorics* **24** (2017), P2.21.

[6] Behrisch, M., Bach, B., Riche, N. H., Schreck, T., and Fekete, J.-D., Matrix reordering methods for table and network visualization, *Computer Graphics Forum* **35** (2016), 693–716.

[7] Sankararaman, A. and Baccelli, F., Community detection on Euclidean random graphs, in *Proceedings of the Twenty-Ninth Annual ACM-SIAM Symposium on Discrete Algorithms* (2018), 2181–2200.

[8] Galhotra, S., Mazumdar, A., Pal, S., and Saha, B., The geometric block model, in *Proceedings of the Thirty-Second AAAI Conference on Artificial Intelligence* **32** (2018), 2215–2222.

[9] Hariprasad, M., A spectral seriation heuristic for community detection in one-dimensional geometric block models, *ResearchGate preprint* (2025).

[10] Allem, L. E., Avrachenkov, K., Hoppen, C., Hariprasad, M., and Sibemberg, L. S., Multi-community spectral clustering for geometric graphs, *arXiv preprint arXiv:2508.00893* (2025).

Hariprasad, M.
School of Mathematics and Natural Sciences
Chanakya University Bengaluru
Bengaluru–562110, India.
E-mail: mhariprasadkansur@gmail.com

The Mathematics Student ISSN: 0025-5742
Vol. 95, Nos. 3-4, July - December (2026), 150–157

GENERATING INFINITE SERIES FOR A UNIT FRACTION

RAMESH BHANDARI AND VIVEK BHANDARI
(Received : 20 - 08 - 2025 ; Accepted : 06 - 10 - 2026)

Abstract. In this paper, we explore infinite series for a unit fraction in terms of rational fractions, a problem which is essentially the converse of the Egyptian fraction problem. We employ a telescopic (difference-of-successive terms) technique and invoke the familiar finite sequences like the arithmetical progression (AP) and geometrical progression (GP) to illustrate our method and generate the series. We also point out some variants of the AP and the GP sequences as other possible generators of the series.

1. Introduction

In this paper, we will consider summations of the following kind:

$$\frac{3}{10}+\frac{4}{45}+\frac{5}{126}+\frac{3}{140}+\frac{7}{540}+..... \tag{1.1}$$

Although neither the limiting value of the series nor the pattern underlying its terms is immediately apparent, we shall see that both emerge in a quite natural manner. If we begin summing the terms of the series, we find that the partial sums approach 1/2 with each successive addition. In fact, we shall see later that, when continued indefinitely, the above series yields an infinite representation of the unit fraction 1/2 as a sum of rational fractions.

A unit fraction is a fraction $1/X$, where $X \in \mathbb{N}$. Any rational fraction can be expressed as a series in distinct unit fractions; e.g., 3/4= 1/3+1/4+1/6, and for historical reasons, because the ancient Egyptians first performed such representations, they are called the Egyptian fractions [1, 2]. A unit fraction $1/X$ is always representable as the sum of two unit

2020 Mathematics Subject Classification: 11D68, 40A05
Key words and phrases: unit fraction, infinite series, rational fractions, Egyption fractions, telescopic series, finite sequences

© *Indian Mathematical Society, 2026.*

fractions: $1/X = 1/(X+1)+1/(X(X+1))$. Therefore, a repeated splitting of a unit fraction into two unit fractions leads to an infinite series for a unit fraction, expressed wholly in terms of unit fractions. In this paper, out of investigative curiosity and general interest [3], we will remove this restriction and derive instead an infinite series for the unit fraction in terms of the general rational fractions, a problem which may be deemed the converse of the ancient Egyptian problem.

A straightforward way to express $1/X$ as an infinite sum in rational fractions would be to use a binomial expansion: $1/X = ((1-r)/X)(1/(1-r)) = ((1-r)/X)(1+r+r^2+r^3+....)$, which is a geometric series with the common ratio r greater than zero but less than one in magnitude; changing r produces a new infinite series for $1/X$. A common feature of this class of series is a common multiplicative factor: $(1-r)/X$.

Further, an infinite series of the type

$$\frac{1}{X} = \sum_{i=1}^{\infty} \left(\frac{1}{X+i-1} - \frac{1}{X+i} \right) = \sum_{i=1}^{\infty} \frac{1}{(X+i-1)(X+i)} \tag{1.2}$$

is often discussed in mathematical literature (especially with $X = 1$) as an example of telescoping series; see, for example, [4]. However, this infinite series also expresses a given unit fraction in terms of unit fractions.

In this paper, we will employ an approach which is reminiscent of Ramanujan's derivation of the famous nested infinite radicals [5]:

$$3 = \sqrt{1+2\sqrt{1+3\sqrt{1+4\sqrt{1+5\sqrt{1+.....}}}}} \tag{1.3}$$

Ramanujan applied a square-root identity recursively in a clever way to obtain a parametrized infinitely nested radical expression with a known limit, and then assigned values to the parameters to obtain identities such as the one in (1.3).

In a similar spirit we will derive infinite series for the unit fraction in parametrized forms and then obtain summations of the type shown in (1.1)

after assignment of values to the parameters. The methods we will use are described below.

2. Unit Fraction as a Generic Infinite Series

Let

$$S_j = \sum_{i=1}^{j} y_i, \tag{2.1}$$

where $y_i \in \mathbb{N}$. Then we can write

$$\frac{1}{S_N} = \frac{1}{S_1} - \sum_{j=1}^{N-1} \Big(\frac{1}{S_j} - \frac{1}{S_{j+1}} \Big). \tag{2.2}$$

As $N \to \infty$, $S_N \to \infty$, and (2.2) reduces to

$$\frac{1}{y_1} = \sum_{j=1}^{\infty} \Big(\frac{1}{S_j} - \frac{1}{S_{j+1}} \Big); \tag{2.3}$$

$S_1 = y_1$, and the summation on the right is a telescoping series. For fixed y_1, we can have an arbitrary (in fact, uncountable) number of infinite series for $1/y_1$, since we have complete freedom in assignment of values to y_i for which $i > 1$. In what follows, we will consider an orderly assignment of values to y_i in (2.1) using the familiar arithmetic progression (AP) and the geometric progression (GP). This will then facilitate analytic expressions for the RHS of (2.3) and thus for the infinite series for the unit fraction.

3. Infinite Series with AP as the Generator

Denoting the first term of AP by $X \in \mathbb{N}$ and the common difference by $d \in \mathbb{N}$

$$y_i = X + d(i-1), \ i \geq 1. \tag{3.1}$$

Then

$$S_j = \sum_{i=1}^{j} y_i = (X + d(j-1)/2)j \tag{3.2}$$

follows from the standard formula for the sum of a finite AP. Inserting (3.2) into (2.3), and noting $y_1 = X$, we obtain after simplification

$$\frac{1}{X} = \sum_{j=1}^{\infty} \frac{4(X + dj)}{(2X + dj)(2X + d(j-1))j(j+1)}. \tag{3.3}$$

3.1. **Properties of the AP-Generated Series.** We note the following properties of the series:

1) In addition to X, the series in (3.3) is also defined and determined by the free parameter d. For fixed X, changing d generates a new series.

2) There is no explicit common multiplicative constant determining the series, as in the example geometric series in Section 1; the products $(2X + dj)(2X + d(j-1))$ and $j(j+1)$ are each even, canceling out the numerical factor of four in the numerator.

3) As $j \to \infty$, the contributions of the terms fall off as $1/(dj^3)$; for larger values of d, the series converges faster.

4) The contribution of the first term ($j = 1$) is $(1/X)(X + d)/(2X + d)$. Consequently, for fixed X, the first term, which can be reexpressed as $(1/X)(1 + X/d)(1 + 2X/d)^{-1}$, gradually becomes closer to $1/X$ as d/X increases, reaching the limit of $1/X$ as $d \to \infty$. If we similarly rewrite the expression for the first term as $(1/(2X))(1 + d/X)(1 + d/(2X))^{-1}$, we see that for fixed d, as X (or X/d) increases, the first term gets closer to the limit of $1/(2X)$ reached as $X \to \infty$.

3.2. **Examples.** i) $X = 1, d = 2$

$$1 = \sum_{j=1}^{\infty} \frac{1+2j}{j^2(j+1)^2} = \frac{3}{4} + \frac{5}{36} + \frac{7}{144} + \frac{9}{400} + \frac{11}{900} + \ldots\ldots \tag{3.4}$$

ii) $X = 2, d = 1$

$$\frac{1}{2} = \sum_{j=1}^{\infty} \frac{4(2+j)}{(4+j)(3+j)j(j+1)} = \frac{3}{10} + \frac{4}{45} + \frac{5}{126} + \frac{3}{140} + \frac{7}{540} + \ldots\ldots \tag{3.5}$$

iii) $X = 2, d = 7$

$$\frac{1}{2} = \sum_{j=1}^{\infty} \frac{4(2+7j)}{(4+7j)(7j-3)j(j+1)} = \frac{9}{22} + \frac{16}{297} + \frac{23}{1350} + \frac{3}{400} + \frac{37}{9360} + \ldots\ldots \tag{3.6}$$

iv) $X = 6, d = 1$

$$\frac{1}{6} = \sum_{j=1}^{\infty} \frac{4(6+j)}{(12+j)(j+11)j(j+1)} = \frac{7}{78} + \frac{8}{273} + \frac{1}{70} + \frac{1}{120} + \frac{11}{2040} + \ldots\ldots \quad (3.7)$$

It is interesting to note that the analytic expression in (3.5) reproduces the series in (1.1). We further observe that the first term of the series in (3.6) is closer to $1/X$ than in (3.5) because d/X is larger, while in (3.7) it is closer to $1/(2X)$ than in (3.5) because X/d is larger (Property 4 in Section 3.1); the closeness to the limiting value is defined as |(limiting value-actual value)/limiting value|; smaller this value, closer the first term.

4. Infinite Series using GP as the Generator

Here we write

$$y_i = aX^{i-1},\ i \geq 1, \quad (4.1)$$

where $a \in \mathbb{N}$ is the first term of the GP and $X \in \mathbb{N}$ is the common ratio, with $X > 1$.

$$S_j = \sum_{i=1}^{j} aX^{i-1} = a\Big(\frac{X^j - 1}{X - 1}\Big). \quad (4.2)$$

Inserting in (2.3), we obtain after simplification

$$1 = \sum_{j=1}^{\infty} \frac{X^j (X-1)^2}{(X^j - 1)(X^{j+1} - 1)}. \quad (4.3)$$

The factor of $(X-1)^2$ in the numerator cancels the factor of $X - 1$ originating in each of the expressions, $X^j - 1$ and $X^{j+1} - 1$, in the denominator; furthermore, one of these expressions has an even exponent of X, which may be j or $j + 1$, meaning it also has $X + 1$ as a common factor. Therefore, we must multiply both sides of (4.3) by $X + 1$ to cancel this factor on the right-hand-side (RHS):

$$X + 1 = \sum_{j=1}^{\infty} \frac{X^j (X-1)^2 (X+1)}{(X^j - 1)(X^{j+1} - 1)}. \quad (4.4)$$

Noting that the first term ($j = 1$) of the series on the RHS equals X, we can then write

$$1 = \sum_{j=2}^{\infty} \frac{X^j (X-1)^2 (X+1)}{(X^j - 1)(X^{j+1} - 1)}. \quad (4.5)$$

The sum in the above equation has a common factor of X^2 arising from X^j in the numerator. Dividing both sides by X^2 then yields (after a change of index $j \to j+1$)

$$\frac{1}{X^2} = \sum_{j=1}^{\infty} \frac{X^{j-1}(X-1)^2(X+1)}{(X^{j+1}-1)(X^{j+2}-1)}. \tag{4.6}$$

4.1. **Properties of the GP-Generated Series.** Some notable properties are:

1) The assigned value of X solely determines the infinite series in (4.6); there is no free parameter as in the AP-generated series of Section 3; the series generated is for a unit fraction given by $1/X^2$, $X > 1$.

2) The series has no common factor; all common factors of numerator and denominator cancel out.

3) The terms of the series decline exponentially as $(X-1)^2(X+1)X^{-(j+4)}$ as j becomes large; larger the value of X, faster the convergence.

4.2. **Examples.** i) X=2

$$\frac{1}{4} = \frac{1}{7} + \frac{2}{35} + \frac{4}{155} + \frac{8}{651} + ... \tag{4.7}$$

ii) X=3

$$\frac{1}{9} = \frac{1}{13} + \frac{3}{130} + \frac{9}{1210} + \frac{27}{11011} + ... \tag{4.8}$$

iiii) X=4

$$\frac{1}{16} = \frac{1}{21} + \frac{4}{357} + \frac{16}{5797} + \frac{64}{93093} + ... \tag{4.9}$$

5. Discussion and Further Exploration

We have investigated infinite series for a unit fraction as a fun exercise. Specifically, we have provided a method that creates infinite series in terms of the general rational fractions and is devoid of a common multiplying factor that, for example, is present in the simple infinite geometric series representation of $1/X$ in Section 1. To our knowledge, such series for a unit fraction have not been discussed before. The method employed is based on the concept of a telescoping series (difference of two successive terms) that utilize sequences with analytic expressions for the sequence sum to provide

a well-defined structure to the infinite series. To illustrate our methodology, we used the well-known arithmetic progression (AP) and the geometric progression (GP).

The work above, however, can be explored further by considering variants of the AP and the GP sequences. The AP sequence in (3.1), which is characterized by a linear dependence on index i , can easily be generalized to $y_i = X + d(i-1)^n$, where exponent $n \geq 1$. This is possible because analytic expression for S_j (2.1) can be obtained from knowledge of the analytic expressions for summations of the kind, $\sum_{i=1}^{k} i^p$, $p \geq 1$ [6]. For example, the expression, $\sum_{i=1}^{k} i^2 = k(k+1)(2k+1)/6$ [6], enables analytic expression for S_j for the $n = 2$ case. In general, S_j for $n \geq 2$ and the corresponding infinite series can be obtained. In all such studies, it would be interesting to examine and understand the different characteristics such as the structure of the series pattern, its properties, and the convergence rate, in relation to the results of Section 3.

Likewise, variants of the GP like the polynomially-weighted GP series: $y_i = ai^n X^{i-1}$, where the ith term of the GP is multiplied by i^n $(n > 0)$, and the shifted GP series $(aX^{i-1} + b)$, where parameter b is a fixed integer, may be further considered for new investigations. In the first case, for exponent $n = 1$, writing $y_i = (\partial/\partial X)(aX^i)$, where X is treated as a continuous differentiable variable, immediately facilitates an analytic expression for $S_j = \sum_{i=1}^{j} y_i$, and thus an infinite series. In the shifted GP case, analytic expressions are straightforward, but there are two free parameters, a and b to play with. In the standard GP case of Section 4, the parameter a canceled out. Furthermore, the unit fraction was limited in solution space to $1/X^2$, i.e., not all possible unit fractions $1/X$ were covered by the standard GP sequence. It would be interesting to investigate the solution space for the above variants and compare with the standard GP case. Many more sequences yielding infinite series may still be possible, including, for example, a hybrid AP-GP sequence like $y_i = (X + (i-1)d)(aX^{i-1})$, where the AP and GP sequences are enmeshed through the product of their ith terms; we invite the reader to look into those as well.

Acknowledgment: We thank the referee for the valuable comments and suggestions, which helped us improve the presentation of the material.

References

[1] Bloom, T.F. and Elsholtz C., Egyptian Fractions, https://arXiv.org/pdf/2210.04496v1.pdf

[2] Kosheleva, O and Kreinovich, V., Why ancient Egyptians preferred some sum-of-inverses representations of fractions?, *Applied Mathematical Sciences* **14** (2020), 859-865.

[3] Varadarajan, V.S., Euler and his work on infinite series, *Bulletin of the American Mathematical Society (N.S.)* **44** (2007), 515–539.

[4] Knopp, K., Theory and Applications of Infinite Series, Dover Publications, 1990.

[5] Borwein, J. M. and de Barra, G., Nested Radicals, *The American Mathematical Monthly* **98** (1991), 735-739.

[6] Graham, R. L., Knuth, D. E., and Patashnik, O., Concrete Mathematics, 2nd ed., Addison–Wesley, 1994.

Ramesh Bhandari
1811 Metzerott Road, Apt 301
Adelphi
Maryland 20783, USA.
E-mail: rbhandari617@gmail.com

Vivek Bhandari
13 Arion Road, Takanini 2112
Auckland, NEW ZEALAND.
E-mail: vivekbhandari.singapore@gmail.com

The Mathematics Student ISSN: 0025-5742
Vol. 95, Nos. 3-4, July - December (2026), 158–169

SOME EXTENSIONS OF A PROBLEM OF PRASOLOV AND SHARYGIN IN SOLID GEOMETRY

QUANG HUNG TRAN

(Received : 24 - 03 - 2024 ; Accepted : 22 - 07 - 2026)

Abstract. The aim of this paper is to generalize certain theorems arising from a problem in the book of Prasolov and Sharygin concerning a trirectangular tetrahedron. We present some extensions and generalizations of the original three-dimensional geometric problem. The method adopted here is to prove a general result and then to deduce the original problem as a particular case.

1. Introduction

Solid geometry stands as one of the most pivotal branches of mathematics, crucial for developing not only visual thinking and spatial imagination but also rigorous logical reasoning. Unlike plane geometry, solid geometry immerses students in three-dimensional concepts, enabling them to observe, analyze, and resolve practical challenges effectively.

A cornerstone of this field is found in Book XI of Euclid's *Elements* (see [9]), a seminal work that laid the foundation for the systematic study of three-dimensional space. Through the exploration of Book XI, students are introduced to the essential definitions, postulates, and propositions that have shaped the understanding of spatial structures for centuries. This historical text not only enriches one's perspective on geometry but also provides a timeless framework applicable to modern disciplines such as architecture, mechanics, and computer science. The historical development of modern geometry has also been presented by Blumenthal [4].

2020 Mathematics Subject Classification: 51M04, 51M20, 51-01, 51-02
Key words and phrases: solid geometry, trirectangular tetrahedron, plane angle, Pythagorean theorem, congruent triangles, surface area

© *Indian Mathematical Society, 2026.*

In secondary education, teaching solid geometry goes beyond merely transmitting knowledge. It equips students with a robust foundational understanding and cultivates abstract thinking, an indispensable skill in advanced mathematical studies and research. By engaging with complex geometric problems, learners develop precision, disciplined logical reasoning, and the capacity to express sophisticated ideas clearly. The International Commission on Mathematical Instruction has provided a detailed account of the teaching of geometry for the twenty-first century [13].

Euler's formula for a convex polyhedron is significant and is stated as $V-E+F=2$, where V, E, and F are respectively the numbers of vertices, edges, and faces. This formula is of particular significance in the context of a tetrahedron. An interesting area of research is the relationship between geometric structures and graph theory. In this context, tetrahedral graphs have been studied by Aigner [1] and Bose and Laskar [5]. Another important field of research is tetrahedral molecular geometry; see, for example, Ramachandran, Deepa, and Namboori [16].

The following references are excellent resources for teaching solid geometry [2, 3, 6, 7, 8, 10, 11, 14, 15, 17, 18].

Alongside these classical sources, several recent contributions illustrate the continuing vitality of tetrahedral and solid geometry. Grace's inequality for right tetrahedra admits a short elementary proof in [20]; tetrahedra with congruent facet pairs are studied through a generalized volume formula in [22]; and a vector generalization of de Gua's theorem is developed in [19]. Further examples include an analogue of Thébault's theorem connecting the radical center of four spheres with the insphere and the Monge point of a tetrahedron [12], as well as an extension of Morley's trisector theorem to isosceles tetrahedra [21]. Together, these works show that classical tetrahedral geometry continues to generate new metric, vectorial, and incidence-geometric results.

In summary, the study of solid geometry, anchored by the insights of Euclid's Book XI, serves as a dynamic tool for nurturing critical cognitive

abilities and fostering the essential skills required for both academic excellence and lifelong success.

In the book of Prasolov and Sharygin [15], the following Problem No. 6.21 is specified.

Problem 6.21. Let $SABC$ be a trirectangular tetrahedron at S such that $SA = SB + SC$. Prove that the sum of the plane angles at vertex A is $90°$.

This intriguing problem involving a trirectangular tetrahedron can be developed into certain theorems. We restate it as the following theorem.

Theorem 1.1. *If $SABC$ is a tetrahedron that is trirectangular at S with $SA = SB + SC$, then the sum of the plane angles at vertex A is $90°$.*

The above theorem has the following converse.

Theorem 1.2. *If $SABC$ is a tetrahedron that is trirectangular at S such that the sum of the plane angles at vertex A is $90°$, then $SA = SB + SC$.*

We also establish an extension of Theorem 1.1 as follows.

Theorem 1.3. *Let $SABC$ be a tetrahedron that is trirectangular at S and satisfies $SA = SB + SC$. Denote by Δ the area of triangle SBC and by r the radius of the inscribed sphere of the tetrahedron. Then*

$$\Delta = r \cdot SA.$$

The aim of this paper is to generalize Theorems 1.1, 1.2, and 1.3, and to provide proofs that synthesize various trigonometric methods. Throughout the paper, we denote the area of triangle XYZ by $[XYZ]$.

We begin with a generalization that encompasses both Theorem 1.1 and Theorem 1.2.

Theorem 1.4 (Generalization of Theorems 1.1 and 1.2)**.** *Let $SABC$ be a tetrahedron that is trirectangular at S, and suppose that the sum of the plane angles at vertex A is α. Then the following relation holds:*

$$SA = SB + SC + \cos\alpha \left(\delta SA + \frac{2[ABC]}{SA} \right),$$

where

$$\delta = \frac{\cos\frac{\alpha}{2} - \sin\frac{\alpha}{2}}{\cos\frac{\alpha}{2} + \sin\frac{\alpha}{2}}.$$

Remark 1.5. In the subsequent proof, we shall demonstrate that the expression $\delta SA + \frac{2[ABC]}{SA}$ is always positive. Consequently, from Theorem 1.4, it follows immediately that $SA = SB + SC$ if and only if $\cos\alpha = 0$, that is, $\alpha = \frac{\pi}{2}$.

Theorem 1.6 (Generalization of Theorem 1.2)**.** *Let $SABC$ be a tetrahedron that is trirectangular at S, and suppose that the sum of the plane angles at vertex A is α. Then the following relation holds:*

$$\sin\alpha\, SA = SB + SC + 2\cos\alpha\, \frac{[ABC]}{SA}.$$

Remark 1.7. An immediate consequence of Theorem 1.6 is that when $\alpha = \frac{\pi}{2}$, we have $SA = SB + SC$. Thus, Theorem 1.2 is a particular case of Theorem 1.6.

Theorem 1.8 (Generalization of Theorem 1.3)**.** *Let $SABC$ be a tetrahedron satisfying the following two conditions:*

i) $\angle ASB = \angle ASC = 180^\circ - \angle BSC$;
ii) $SA = SB + SC$.

Denote by Δ the area of triangle SBC, by r the radius of the inscribed sphere of the tetrahedron, and by h the altitude from A to the base plane of the tetrahedron. Then $\Delta = rh$ if and only if the sum of the plane angles at vertex A is 90°.

2. Proofs of the Theorems

In this section, we present detailed proofs of Theorems 1.4, 1.6, and 1.8. Naturally, these proofs also encompass Theorems 1.1, 1.2, and 1.3.

First, we approach Theorem 1.4 using trigonometric techniques and the law of cosines.

Proof of Theorem 1.4. See Figure 1. First, by applying the law of cosines together with the Pythagorean theorem, we obtain

$$\cos\angle BAC = \frac{AB^2 + AC^2 - BC^2}{2AB\cdot AC} = \\ = \frac{SA^2 + SB^2 + SA^2 + SC^2 - (SB^2 + SC^2)}{2AB\cdot AC} = \frac{SA^2}{AB\cdot AC}. \quad (2.1)$$

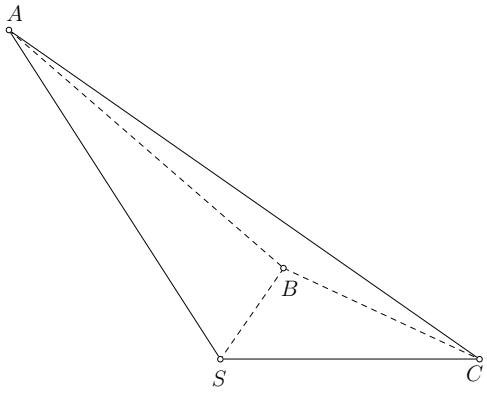

FIGURE 1. Illustration for the proof of Theorem 1.4.

Using the sine addition formula, we have

$$\begin{aligned}\sin(\angle SAC + \angle SAB) &= \sin\angle SAC \cdot \cos\angle SAB + \cos\angle SAC \cdot \sin\angle SAB \\ &= \frac{SC}{AC} \cdot \frac{SA}{AB} + \frac{SB}{AB} \cdot \frac{SA}{AC} \\ &= \frac{SA(SB+SC)}{AB \cdot AC}.\end{aligned} \tag{2.2}$$

Thus, from (2.1) and (2.2), we deduce

$$\frac{\sin(\angle SAC + \angle SAB)}{\cos\angle BAC} = \frac{SB+SC}{SA}. \tag{2.3}$$

Rearranging equation (2.3) yields

$$\frac{SB+SC-SA}{SA}\cos\angle BAC = \sin(\angle SAC + \angle SAB) - \cos\angle BAC. \tag{2.4}$$

Using (2.1) and the assumption that

$$\angle BAC + \angle SAC + \angle SAB = \alpha,$$

equation (2.4) becomes equivalent to

$$\frac{SB+SC-SA}{SA} \cdot \frac{SA^2}{AB \cdot AC} = \sin(\alpha - \angle BAC) - \sin\left(\frac{\pi}{2} - \angle BAC\right). \tag{2.5}$$

Equation (2.5) is equivalent to

$$SA(SB+SC-SA) = 2AB \cdot AC \cos\left(\frac{\alpha}{2} + \frac{\pi}{4} - \angle BAC\right)\sin\left(\frac{\alpha}{2} - \frac{\pi}{4}\right). \tag{2.6}$$

By transforming the right-hand side of equation (2.6), we obtain

$$\begin{aligned}
&2AB \cdot AC \cos\left(\frac{\alpha}{2}+\frac{\pi}{4}-\angle BAC\right)\sin\left(\frac{\alpha}{2}-\frac{\pi}{4}\right)\\
&= 2AB \cdot AC\Big(\cos\left(\frac{\alpha}{2}+\frac{\pi}{4}\right)\cos\angle BAC+\sin\left(\frac{\alpha}{2}+\frac{\pi}{4}\right)\sin\angle BAC\Big)\\
&\quad\cdot\frac{\sin\left(\alpha-\frac{\pi}{2}\right)}{2\cos\left(\frac{\alpha}{2}-\frac{\pi}{4}\right)} \qquad (2.7)\\
&= -\cos\alpha\left(\frac{\cos\left(\frac{\alpha}{2}+\frac{\pi}{4}\right)}{\cos\left(\frac{\alpha}{2}-\frac{\pi}{4}\right)}SA^2+\frac{\sin\left(\frac{\alpha}{2}+\frac{\pi}{4}\right)}{\cos\left(\frac{\alpha}{2}-\frac{\pi}{4}\right)}\cdot 2[ABC]\right)\\
&= -\cos\alpha\left(\delta SA^2+2[ABC]\right),
\end{aligned}$$

where

$$\delta=\frac{\cos\left(\frac{\alpha}{2}+\frac{\pi}{4}\right)}{\cos\left(\frac{\alpha}{2}-\frac{\pi}{4}\right)}=\frac{\cos\frac{\alpha}{2}-\sin\frac{\alpha}{2}}{\cos\frac{\alpha}{2}+\sin\frac{\alpha}{2}}.$$

From equations (2.6) and (2.7), we deduce that

$$SA=SB+SC+\cos\alpha\left(\delta SA+\frac{2[ABC]}{SA}\right).$$

This completes the proof. □

Remark 2.1. Note that from (2.1),

$$\cos\angle BAC=\frac{SA^2}{AB\cdot AC}=\cos\angle SAC\cdot\cos\angle SAB<\cos\angle SAC.$$

Hence, $\angle BAC>\max\{\angle SAC,\angle SAB\}$. Moreover, since the angles $\angle SAC$, $\angle SAB$, and $\angle BAC$ are all acute, it follows that

$$\frac{\alpha}{2}+\frac{\pi}{4}-\angle BAC=\frac{\angle SAC+\angle SAB-\angle BAC}{2}+\frac{\pi}{4}<\frac{\angle SAC}{2}+\frac{\pi}{4}<\frac{\pi}{2}.$$

Consequently, $\cos\left(\frac{\alpha}{2}+\frac{\pi}{4}-\angle BAC\right)$ is always positive, which means that $\delta SA^2+2[ABC]$ is always positive. This observation concludes the remark following Theorem 1.4.

Using some of the formulas from the previous proof, we now prove Theorem 1.6.

Proof of Theorem 1.6. See Figure 1. From the hypothesis

$$\angle BAC+\angle SAC+\angle SAB=\alpha,$$

we deduce that

$$\angle SAB+\angle SAC=\alpha-\angle BAC.$$

Hence,

$$\sin(\alpha - \angle BAC) = \sin(\angle SAB + \angle SAC). \tag{2.8}$$

Transforming the left-hand side of (2.8) gives

$$\begin{aligned}\sin(\alpha - \angle BAC) &= \sin\alpha\cos\angle BAC - \cos\alpha\sin\angle BAC \\ &= \sin\alpha\frac{SA^2}{AB\cdot AC} - \cos\alpha\sin\angle BAC \quad (\text{by } (2.1)) \\ &= \frac{\sin\alpha SA^2 - \cos\alpha\, AB\cdot AC\sin\angle BAC}{AB\cdot AC} \\ &= \frac{\sin\alpha SA^2 - 2\cos\alpha\,[ABC]}{AB\cdot AC}.\end{aligned} \tag{2.9}$$

On the other hand, combining (2.2), (2.8), and (2.9), we deduce that

$$\frac{\sin\alpha SA^2 - 2\cos\alpha\,[ABC]}{AB\cdot AC} = \frac{SA(SB+SC)}{AB\cdot AC}.$$

This equation is equivalent to

$$\sin\alpha\, SA = SB + SC + 2\cos\alpha\,\frac{[ABC]}{SA}.$$

This is exactly what needed to be proved. □

Finally, we employ the method of unfolding the tetrahedron to prove Theorem 1.8. Although this approach is primarily synthetic, it ultimately incorporates a trigonometric equation.

Proof of Theorem 1.8. See Figure 2.

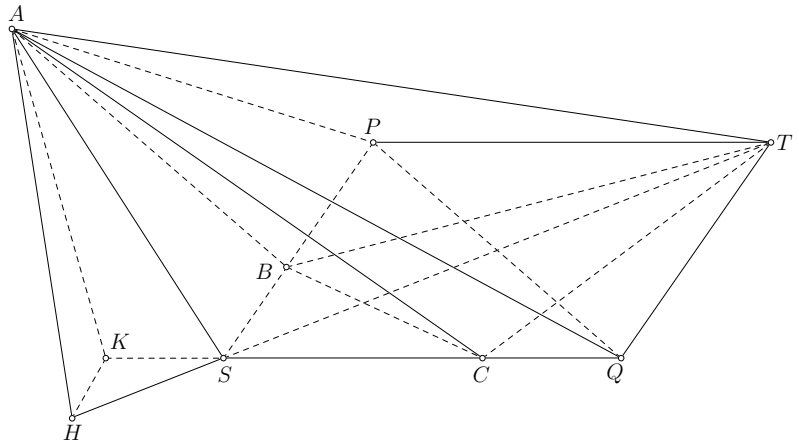

FIGURE 2. Illustration for the proof of Theorem 1.8.

On the rays SB and SC, choose points P and Q, respectively, so that $SP = SQ = SA$. Then, using the hypothesis $SA = SB + SC$, it is easy to see that

$$PB = SP - SB = SA - SB = SC$$

and

$$QC = SQ - SC = SA - SC = SB.$$

Reflect S across the line PQ to obtain the point T. Since triangle SPQ is isosceles with vertex S, the quadrilateral $SPTQ$ is a rhombus. Furthermore, by the hypothesis

$$\angle ASB = \angle ASC = 180^\circ - \angle BSC,$$

we deduce that

$$\angle BPT = \angle CQT = 180^\circ - \angle BSC = \angle ASB = \angle ASC.$$

Thus, triangles SAC and PTB, as well as triangles SAB and QTC, are congruent by SAS. Also, triangles ABC and TBC are congruent by SSS. Consequently, the corresponding angles are equal:

$$\angle SAC = \angle PTB, \quad \angle SAB = \angle QTC, \quad \angle BAC = \angle BTC.$$

It follows that the sum of the plane angles at vertex A is

$$\angle SAB + \angle BAC + \angle CAS = \angle QTC + \angle CTB + \angle BTP = \angle QTP = \angle BSC.$$

Denote $\angle BSC = \theta$. Then the sum of the plane angles at A equals θ and, simultaneously,

$$\angle ASB = \angle ASC = 180^\circ - \theta.$$

We now show that $\theta = 90^\circ$ if and only if $\Delta = rh$. Let H be the foot of the perpendicular from A onto the plane (SBC). Since $\angle ASB = \angle ASC$ and $SPTQ$ is a rhombus, the point H lies on the line ST. Denote by K the projection of A onto the line SQ.

Let $\mathcal{A}$ be the total surface area of the tetrahedron $SABC$. From the congruent triangles mentioned above, $\mathcal{A}$ equals the area of the rhombus $SPTQ$; hence,

$$\mathcal{A} = SP \cdot SQ \sin\theta = SA^2 \sin\theta = \frac{AK^2}{\sin\theta}. \tag{2.10}$$

On the other hand, since r is the radius of the inscribed sphere of the tetrahedron $SABC$ and V is its volume, we have

$$\Delta h = 3V = \mathcal{A}r. \tag{2.11}$$

Thus, from (2.10) and (2.11), we deduce

$$\Delta h = \frac{AK^2}{\sin\theta}\, r. \tag{2.12}$$

Since H is the projection of A onto the plane (SBC), we have $AH = h$. Notice that triangle AHK is right-angled at H. By the Pythagorean theorem,

$$h^2 = AK^2 - KH^2 = AK^2 - SK^2 \tan^2\frac{\theta}{2} = AK^2 - AK^2\frac{\tan^2\frac{\theta}{2}}{\tan^2\theta}.$$

Thus,

$$\frac{h^2}{AK^2} = 1 - \frac{\tan^2\frac{\theta}{2}}{\tan^2\theta}. \tag{2.13}$$

Combining (2.12) and (2.13), we obtain

$$r = \frac{\Delta h}{h^2}\sin\theta\left(1 - \frac{\tan^2\frac{\theta}{2}}{\tan^2\theta}\right),$$

or equivalently,

$$\frac{hr}{\Delta} = \sin\theta\left(1 - \frac{\tan^2\frac{\theta}{2}}{\tan^2\theta}\right). \tag{2.14}$$

Set $\tan\frac{\theta}{2} = t$ with $t > 0$. From the standard trigonometric formulas, we have

$$\sin\theta = \frac{2t}{1+t^2}, \quad \tan\theta = \frac{2t}{1-t^2}.$$

Substituting these into (2.14), we deduce

$$1 - \frac{hr}{\Delta} = 1 - \frac{2t}{1+t^2}\left(1 - \frac{t^2}{\frac{4t^2}{(1-t^2)^2}}\right) = \frac{1}{2}(t-1)^2(t+2). \tag{2.15}$$

Therefore,

$$\Delta = rh \quad \text{if and only if} \quad \frac{1}{2}(t-1)^2(t+2) = 0.$$

Since $t > 0$, this is equivalent to $t = 1$. Hence, $\tan\frac{\theta}{2} = 1$ if and only if $\theta = 90^\circ$. In other words, $\Delta = rh$ if and only if the sum of the plane angles at vertex A of the tetrahedron is 90°. This completes the proof. □

Remark 2.2. When $\theta = 90°$, the tetrahedron is also trirectangular at S. Moreover, it is evident that

$$1 - \frac{hr}{\Delta} = \frac{1}{2}(t-1)^2(t+2) \geq 0.$$

Hence, we always have the inequality

$$\Delta \geq hr,$$

with equality if and only if the tetrahedron is trirectangular at S.

3. Conclusion

In this work, we have demonstrated the effectiveness of combining trigonometric methods with synthetic approaches to solve intricate geometric problems. By integrating classical trigonometry with the method of unfolding the tetrahedron, we provided proofs of several nontrivial theorems. This approach not only simplifies complex three-dimensional relationships by converting them into more manageable two-dimensional problems, but also highlights a versatile strategy applicable to a broad spectrum of similar challenges.

The unfolding of the tetrahedron, in particular, stands out as a creative and insightful method. This technique transforms the spatial problem into a planar configuration, allowing the use of well-known trigonometric identities and geometric properties. Such a method reveals hidden symmetries and connections that might otherwise remain obscured in a purely spatial analysis.

Furthermore, the progression from the initial problem presented in Theorem 1.1 to its extensions and generalizations has significant educational implications. Starting from a foundational problem and expanding it to encompass more complex situations enriches the theoretical landscape and offers a useful framework for teaching. This layered development helps students build confidence as they tackle increasingly challenging problems and illustrates the interconnectedness of different mathematical ideas.

In summary, the combination of trigonometric and synthetic methods, especially through tetrahedral unfolding, provides a powerful strategy for

solving and understanding complex geometric problems. This study exemplifies how such methods can lead to elegant solutions while also serving as an effective pedagogical tool in mathematics education.

Acknowledgements: The author is grateful to the referee for the helpful suggestion concerning the title and for valuable comments.

The author would also like to express his sincere gratitude to Professor Youvaraj, Editor-in-Chief of *The Mathematics Student*, for his kind support and guidance during the editorial process.

The author is also grateful to the students of the two specialized mathematics classes A1 and A2 at the High School for Gifted Students, VNU University of Science, cohort 2023–2026. The time and dedication that he devoted to teaching solid geometry in these classes inspired the writing of this article.

The author also thanks ChatGPT 5.2 for its assistance in enhancing the English language of this manuscript.

References

[1] Aigner, M., On the tetrahedral graph, *Pacific Journal of Mathematics* **25** (1968), no. 2, 219–227.

[2] Altshiller-Court, N., *Modern Pure Solid Geometry*, Chelsea, New York, 1979.

[3] Balliccioni, A., *Coordonnées barycentriques et géométrie*, Claude Hermant, 1964.

[4] Blumenthal, L. M., *A Modern View of Geometry*, W. H. Freeman and Company, San Francisco, 1961; Dover Publications, Mineola, New York, 1980, 2017.

[5] Bose, R. C. and Laskar, R., A characterization of tetrahedral graphs, *Journal of Combinatorial Theory* **3** (1967), no. 4, 366–385.

[6] Couderc, P. and Balliccioni, A., *Premier livre du tétraèdre à l'usage des èlèves de première, de mathématiques, des candidats aux grandes ècoles et à l'agrégation*, Gauthier-Villars, Paris, 1935.

[7] Crelle, A. L., Einige Bemerkungen über die dreiseitige Pyramide, *Sammlung mathematischer Aufsätze u. Bemerkungen* **1** (1821).

[8] Dostor, G., *Éléments de la théorie des déterminants, avec application à l'algèbre, la trigonométrie et la géométrie analytique dans le plan et l'espace*, 2nd ed., Gauthier-Villars, Paris, 1905.

[9] Euclid, Book XI of the *Elements*, `http://aleph0.clarku.edu/~djoyce/java/elements/elements.html`.

[10] Gardner, M., Tetrahedrons, Chapter 19 in *The Sixth Book of Mathematical Games from Scientific American*, University of Chicago Press, Chicago, IL, 1984.

[11] Hopf, H., *Selected Chapters of Geometry*, ETH Zürich lecture notes, 1940.

[12] Herrera, B. and Tran, Q. H., An analogue of Thébault's theorem linking the radical center of four spheres with the insphere and the Monge point of a tetrahedron, *International Electronic Journal of Geometry* **15** (2022), no. 1, 75–78.

[13] The International Commission on Mathematical Instruction, Perspectives on the teaching of geometry for the 21st century, *Educational Studies in Mathematics* **28** (1995), no. 1, 91–98.

[14] Lee, J. R., The law of cosines in a tetrahedron, *Journal of the Korea Society of Mathematical Education, Series B: Pure and Applied Mathematics* **4** (1997).

[15] Prasolov, V. V. and Sharygin, I. F., *Stereometry Problems*, Nauka, Moscow, 1991.

[16] Ramachandran, K. I., Deepa, G. and Namboori, P. K., *Computational Chemistry and Molecular Modeling: Principles and Applications*, Springer-Verlag, Berlin Heidelberg, 2008.

[17] Rouche, E. and de Comberousse, C., *Traité de géométrie, nouv. ed., vol. 2: Géométrie dans l'espace*, Gauthier-Villars, Paris, 1922.

[18] von Staudt, K. G. C., Ueber einige geometrische Sätze, *Journal für die reine und angewandte Mathematik* **57** (1860).

[19] Tran, Q. H., A generalization of de Gua's theorem with a vector proof, *The Mathematical Intelligencer* **46** (2024), no. 3, 236–238.

[20] Tran, Q. H., A simple proof of Grace's inequality in a right tetrahedron, *Journal of Geometry* **117** (2026), Art. 17.

[21] Tran, Q. H., Morley's trisector theorem for isosceles tetrahedron, *Acta Mathematica Hungarica* **165** (2021), no. 2, 308–315.

[22] Tran, Q. H., A generalized volume formula for tetrahedra with congruent facet pairs, *The Mathematical Intelligencer* **47** (2025), no. 2, 136–139.

Quang Hung Tran
High School for Gifted Students
VNU University of Science
Vietnam.
E-mail: tranquanghung@hus.edu.vn

The Mathematics Student ISSN: 0025-5742
Vol. 95, Nos. 3-4, July - December (2026), 170–176

APPROXIMATING THE SUMS OF POWERS OF RECIPROCALS

GEOFFREY W. BROWN, ADAM C. BROWN

(Received : 14 - 10 - 2025 ; Accepted : 17 - 07 - 2026)

Abstract. This paper derives a formula to approximate summing powers of reciprocals for any given power, and it accomplishes this by integrating a certain transform of a discrete random variable in two different ways.

The problem of finding the sums of powers of reciprocals of natural numbers has a long and illustrious history. The simplest case of this problem is to find simply the sum of the reciprocals of the natural numbers. If we let S be the value of the infinite sum, then we have

$$S = 1 + \frac{1}{2} + \frac{1}{3} + \cdots + \frac{1}{k} + \cdots = \sum_{k=1}^{\infty} \frac{1}{k}.$$

The series for S is known as the Harmonic series. It is also known that this sum diverges, and so it evaluates to positive infinity. This fact was established by Jakob Bernoulli in 1689 in his work *Tractatus de seriebus infinitis* (Treatise on Infinite Series), although Jakob attributes the proof appearing there to his brother Johann. Many proofs of the divergence of the Harmonic series are now known.

The next case of this problem is to find the sum of the reciprocals of the squares of the natural numbers. If we again let S be the value of this infinite sum, then the problem is to determine the value of

$$S = 1 + \frac{1}{2^2} + \frac{1}{3^2} + \cdots + \frac{1}{k^2} + \cdots = \sum_{k=1}^{\infty} \frac{1}{k^2}.$$

2020 Mathematics Subject Classification: 40-02, 60-02
Key words and phrases: probability theory, random variables, number theory

© *Indian Mathematical Society, 2026.*

This problem was unsolved in 1689, but the sum was finally determined in 1734 by Leonhard Euler, who had studied mathematics under Johann Bernoulli himself. Euler discovered an ingenious argument to show that

$$\sum_{k=1}^{\infty} \frac{1}{k^2} = \frac{\pi^2}{6},$$

thus solving the so-called Basel problem. See [1] *Journey through Genius*, Chapter 9, for a wonderful account of Euler's original proof. More generally, one can ask for the sum of the following series S, where $n \geq 1$ is a natural number:

$$S = 1 + \frac{1}{2^n} + \frac{1}{3^n} + \cdots + \frac{1}{k^n} + \cdots = \sum_{k=1}^{\infty} \frac{1}{k^n}.$$

This sum S amounts to determining the value of the Riemann zeta function at a positive integral value of n. The Riemann zeta function is defined by the following series as a function of s:

$$\zeta(s) = \sum_{k \geq 1} \frac{1}{k^s}.$$

When $s = 1$, it is clear that $\zeta(1)$ gives the divergent Harmonic series. When $s = 2$, we have just seen that

$$\zeta(2) = \sum_{k \geq 1} \frac{1}{k^2} = \frac{\pi^2}{6},$$

the famous result due to Euler. The formulas for $\zeta(s)$ when s is an even positive integer are all known, and in fact one has

$$\zeta(2n) = \sum_{k \geq 1} \frac{1}{k^{2n}} = (-1)^{n-1} \frac{2^{2n-1} \pi^{2n} B_{2n}}{(2n)!},$$

where B_{2n} is the $2n$-th Bernoulli number. See [2] *Concrete Mathematics*, Section 6.5, for a derivation of these remarkable formulas. For odd positive integers, however, no such formulas for $\zeta(2n+1)$ are known. Thus determining, for example, the exact value of $\zeta(3) = \sum_{k \geq 1} \frac{1}{k^3}$ is an unsolved problem. This paper uses an entirely different approach to derive approximations for the finite sums

$$\sum_{l=1}^{k} \frac{1}{l^{m+1}},$$

where $m \geq 0$ is a non-negative integer and k is a fixed positive integer. We will derive a formula to approximate the finite sum of powers of reciprocals $\sum_{l=1}^{k} \frac{1}{l^{m+1}}$, and we will accomplish this by integrating a certain transform of a discrete random variable in two different ways.

Given the discrete random variable X, we define the following transform of X, which is analogous to the Laplace transform:

$$M_X(t) = E[e^{-tX} t^m], \tag{0.1}$$

where $m \geq 0$ is a non-negative integer. We now specify X to be the discrete uniform distribution, so that X takes the values $1, 2, 3, \ldots, k$ each with probability $\frac{1}{k}$. We calculate $M_X(t)$ for X the discrete uniform random variable, so that (0.1) becomes

$$\begin{aligned} M_X(t) &= E[e^{-tX} t^m] \\ &= e^{-t} \cdot t^m \cdot \frac{1}{k} + e^{-2t} \cdot t^m \cdot \frac{1}{k} + \cdots + e^{-kt} \cdot t^m \cdot \frac{1}{k} \\ &= \frac{1}{k}[e^{-t} t^m + e^{-2t} t^m + \cdots + e^{-kt} t^m]. \end{aligned} \tag{0.2}$$

The next step of the argument is to integrate $M_X(t)$ from zero to infinity, and thus to calculate

$$\int_0^\infty M_X(t) dt,$$

where X is the discrete uniform random variable. Thus (0.2) becomes:

$$\int_0^\infty M_X(t) dt = \frac{1}{k} \left[\int_0^\infty e^{-t} t^m dt + \int_0^\infty e^{-2t} t^m dt + \cdots + \int_0^\infty e^{-kt} t^m dt \right]. \tag{0.3}$$

We now calculate

$$\int_0^\infty e^{-lt} t^m dt, \tag{0.4}$$

where $l \in \{1, 2, \ldots, k\}$. In order to simplify (0.4), we make the change of variables $lt = x$. Thus $t = x/l$ and $dt = dx/l$. Thus (0.4) becomes

$$\int_0^\infty e^{-x} \left(\frac{x}{l}\right)^m \frac{dx}{l} = \frac{1}{l^{m+1}} \int_0^\infty e^{-x} x^m dx. \tag{0.5}$$

At this point it is instructive to recall the definition of the Gamma function. See [3] *Mathematical Statistics with Applications*, p. 185. We have the definition of the Gamma function as

$$\Gamma(\alpha) = \int_0^\infty e^{-x} x^{\alpha-1} dx, \text{ where } \alpha \geq 1. \tag{0.6}$$

It is easy to check that $\Gamma(1) = 1$, and integration by parts will verify that $\Gamma(\alpha) = (\alpha - 1)\Gamma(\alpha - 1)$ for any $\alpha > 1$, and that $\Gamma(n) = (n-1)!$, provided that n is a positive integer. Putting $\alpha = m + 1$ in (0.6), we see that

$$\Gamma(m+1) = \int_0^\infty e^{-x} x^m dx = m!. \tag{0.7}$$

Using (0.7) in (0.5), we see that

$$\int_0^\infty e^{-lt} t^m dt = \frac{1}{l^{m+1}} \cdot m!. \tag{0.8}$$

Using (0.8) in (0.3), we find that

$$\int_0^\infty M_X(t) dt = \frac{1}{k} \cdot \sum_{l=1}^{k} \frac{1}{l^{m+1}} \cdot m!. \tag{0.9}$$

Thus (0.9) shows that all the improper integrals appearing in (0.3) converge, so next we integrate (0.2) using finite limits of integration instead. Let N be a positive real number. Then imitating (0.3), we see that

$$\int_0^N M_X(t) dt = \frac{1}{k} \left[\int_0^N e^{-t} t^m dt + \int_0^N e^{-2t} t^m dt + \cdots + \int_0^N e^{-kt} t^m dt \right]. \tag{0.10}$$

We now calculate

$$\int_0^N e^{-lt} t^m dt,$$

where $l \in \{1, 2, \ldots, k\}$. Using the same change of variables $lt = x$, we find that

$$\int_0^N e^{-lt} t^m dt = \int_0^{lN} e^{-x} \left(\frac{x}{l}\right)^m \frac{dx}{l} = \frac{1}{l^{m+1}} \int_0^{lN} e^{-x} x^m dx.$$

From (0.7) we know that $\int_0^\infty e^{-x} x^m dx = m!$. The basic idea in what follows is that we want to choose the positive real number N large enough so

that $\int_0^N e^{-x}x^m dx$ is very close to $\int_0^\infty e^{-x}x^m dx$, which we know from (0.7) has the value of $m!$. To make this more precise, we let $\epsilon_1 > 0$ be given, and we choose N such that $\int_0^N e^{-x}x^m dx = m! - \epsilon_1$. Then because $2N > N$, $\int_0^{2N} e^{-x}x^m dx$ will be even closer to the value of $m!$. In particular, we have $\int_0^{2N} e^{-x}x^m dx = m! - \epsilon_2$ where $\epsilon_2 < \epsilon_1$, and continuing this argument we see that

$$\int_0^{lN} e^{-x}x^m dx = m! - \epsilon_l, \text{ where } \epsilon_l < \epsilon_{l-1} < \cdots < \epsilon_2 < \epsilon_1.$$

Using these results in (0.10), we see that

$$\int_0^N M_X(t)dt = \frac{1}{k}\left[\sum_{l=1}^{k} \frac{1}{l^{m+1}}(m! - \epsilon_l)\right], \text{ where } \epsilon_k < \epsilon_{k-1} < \cdots < \epsilon_2 < \epsilon_1. \tag{0.11}$$

It is illuminating to compare (0.9) to (0.11).

This completes the first way to calculate $\int_0^N M_X(t)dt$. As for the second way, we may find $\int_0^N M_X(t)dt$ by using the Taylor series expansion for e^{-lt}. We have that

$$e^{-lt} = \sum_{n=0}^{\infty} \frac{(-lt)^n}{n!} \text{ where } l \in \{1, 2, \ldots, k\}. \tag{0.12}$$

Using (0.12) in (0.2), we find that

$$\begin{aligned} M_X(t) &= \frac{1}{k}\left[\sum_{l=1}^{k} e^{-lt}t^m\right] \\ &= \frac{1}{k}\left[\sum_{l=1}^{k}\left(\sum_{n=0}^{\infty}\frac{(-lt)^n}{n!}\right)t^m\right] \\ &= \frac{1}{k}\left[\sum_{n=0}^{\infty}\left(\sum_{l=1}^{k}(-l)^n\right)\frac{t^{n+m}}{n!}\right]. \end{aligned} \tag{0.13}$$

Letting $S_n(k) = 1^n + 2^n + \cdots + k^n$, we note that

$$\sum_{l=1}^{k}(-l)^n = (-1)^n\sum_{l=1}^{k} l^n = (-1)^n S_n(k),$$

so that (0.13) becomes

$$M_X(t) = \frac{1}{k}\left[\sum_{n=0}^{\infty}(-1)^n S_n(k)\frac{t^{n+m}}{n!}\right]. \tag{0.14}$$

Thus we directly calculate from (0.14) that

$$\begin{aligned}\int_0^N M_X(t)dt &= \frac{1}{k}\left[\sum_{n=0}^{\infty}(-1)^n S_n(k)\int_0^N \frac{t^{n+m}}{n!}dt\right]\\ &= \frac{1}{k}\left[\sum_{n=0}^{\infty}(-1)^n S_n(k)\frac{t^{n+m+1}}{(n+m+1)n!}\Bigg|_0^N\right]\\ &= \frac{1}{k}\left[\sum_{n=0}^{\infty}(-1)^n S_n(k)\frac{N^{n+m+1}}{(n+m+1)n!}\right]. \end{aligned}\tag{0.15}$$

Now equating (0.11) to (0.15) and cancelling the factor of $\frac{1}{k}$ appearing in each, we have

$$\sum_{l=1}^{k}\frac{1}{l^{m+1}}(m! - \epsilon_l) = \sum_{n=0}^{\infty}(-1)^n S_n(k)\frac{N^{n+m+1}}{(n+m+1)n!}.$$

What follows now is a bit of algebraic manipulation. Thus:

$$\sum_{l=1}^{k}\frac{1}{l^{m+1}}\cdot m! - \sum_{n=0}^{\infty}(-1)^n S_n(k)\frac{N^{n+m+1}}{(n+m+1)n!} = \sum_{l=1}^{k}\frac{1}{l^{m+1}}\epsilon_l \tag{0.16}$$

$$\sum_{l=1}^{k}\frac{1}{l^{m+1}} - \frac{1}{m!}\sum_{n=0}^{\infty}(-1)^n S_n(k)\frac{N^{n+m+1}}{(n+m+1)n!} = \frac{1}{m!}\sum_{l=1}^{k}\frac{1}{l^{m+1}}\epsilon_l. \tag{0.17}$$

Looking at the right hand side of (0.16), we see that

$$\sum_{l=1}^{k}\frac{1}{l^{m+1}}\epsilon_l < \sum_{l=1}^{k}\frac{1}{l^{m+1}}\epsilon_1 \leq \left(\sum_{l=1}^{k}\frac{1}{l}\right)\epsilon_1. \tag{0.18}$$

Using (0.18) in (0.17), we find that

$$\sum_{l=1}^{k}\frac{1}{l^{m+1}} - \frac{1}{m!}\sum_{n=0}^{\infty}(-1)^n S_n(k)\frac{N^{n+m+1}}{(n+m+1)n!} < \frac{\left(\sum_{l=1}^{k}\frac{1}{l}\right)}{m!}\epsilon_1. \tag{0.19}$$

Now $\sum_{l=1}^{k}\frac{1}{l}$ is a bounded quantity, so we may choose

$$\epsilon = \frac{\left(\sum_{l=1}^{k} \frac{1}{l}\right)}{m!}\epsilon_1$$

in (0.19). Thus we have arrived at the following theorem, which is indeed the main result of the paper.

Theorem: Let $\epsilon > 0$ be given. Then there exists an $N(\epsilon)$ such that for all $N > N(\epsilon)$,

$$\left| \sum_{l=1}^{k} \frac{1}{l^{m+1}} - \frac{1}{m!} \sum_{n=0}^{\infty} (-1)^n S_n(k) \frac{N^{n+m+1}}{(n+m+1)n!} \right| < \epsilon.$$

Now the sums $S_n(k) = 1^n + 2^n + \cdots + k^n$ appearing in these equations are well known in combinatorics. We find that $S_0(k) = k$, $S_1(k) = \frac{k(k+1)}{2} \approx \frac{k^2}{2}$, $S_2(k) = \frac{k(k+1)(2k+1)}{6} \approx \frac{k^3}{3}$, and it is known that $S_n(k) \approx \frac{k^{n+1}}{n+1}$, implying that (0.15) is an absolutely convergent series for a fixed k. It is known that the sequence $\{S_n(k)\}_{n=0}^{\infty}$ is related to the Bernoulli numbers, and also that the same sequence $\{S_n(k)\}_{n=0}^{\infty}$ has a simple *exponential generating series*. See [2] *Concrete Mathematics*, Sections 6.5 and 7.6, for details.

References

[1] Dunham W., *Journey through Genius*, Penguin Books, 1991.

[2] Graham R.L., Knuth D.E., Patashnik O., *Concrete Mathematics*, 2nd Edition, Addison-Wesley, 1994.

[3] Wackerly D.D., Mendenhall W. III, Scheaffer R.L., *Mathematical Statistics with Applications*, 7th Edition, Brooks/Cole Cengage Learning, 2008.

Geoffrey W. Brown
Toronto Metropolitan University
350 Victoria Street
Toronto, ON, M5B 2K3, CANADA.
E-mail: geoffrey.brown@torontomu.ca

Adam C. Brown
University of Toronto Schools
371 Bloor St W
Toronto, ON, M5S 2R7, CANADA.
E-mail: abrown@utschools.ca

The Mathematics Student ISSN: 0025-5742
Vol. 95, Nos. 3-4, July - December (2026), 177–194

ON BLOCK-TRIDIAGONAL MATRIX FORM OF COMPACT OPERATORS

SASMITA PATNAIK AND RAHUL SETHI

(Received : 11 - 11 - 2025 ; Accepted : 13 - 07 - 2026)

Abstract. This article is a preliminary step to relate the study of compact operators to finite matrix theory through the block-tridiagonal matrix form associated with a compact operator. We obtain a matrix representation of a compact operator from its block-tridiagonal matrix form via blockwise triangularization, which is the sum of a strictly lower triangular matrix and an upper triangular matrix, where each row and each column has finitely supported complex scalar entries. We also offer a matricial approach to investigate questions on the quasinilpotency of commutators of compact operators.

1. Introduction

This paper initiates the study of the structure of compact operators and its commutators via block-tridiagonal matrix form of a compact operator acting on an infinite-dimensional separable complex Hilbert space. The block-tridiagonal matrix form associated with bounded operators continues to be an invaluable tool in operator theory. In [11, Theorem 20.4], it is proved that every bounded operator on an infinite-dimensional separable complex Hilbert space admits a block-tridiagonal matrix form whose diagonal blocks are growing exponentially in size. The most recent work on applications of block-tridiagonal matrix forms can be found in the joint work of the first author of this paper with J. Loreaux, S. Petrovic, and G. Weiss in [10, Section 3, Theorems 3.1-3.4].

In this paper, we provide an alternate matricial framework to study compact operators and related problems. The results obtained here can

2020 Mathematics Subject Classification:
Primary: 47B02, 47B47, 47A08, 47A10, Secondary: 47A65, 47B36
Key words and phrases: Block-tridiagonal matrix form of bounded operator, Compact operator Commutator Spectrum Quasinilpotent operator

© *Indian Mathematical Society, 2026.*

be proved in much shorter ways, but here we aim at emphasizing the role of the block-tridiagonal matrix form in obtaining some results. The key point is that through this matricial approach, one can directly connect the problem with the problem in finite matrix theory. Our goal in this paper is twofold: link the quasinilpotency of commutators of compact operators with the simultaneous triangularization of finite matrices and initiate a study of the impact of the block-tridiagonal matrix form on the structure of a compact operator. We established a structural decomposition of a compact operator via its block-tridiagonal matrix form, which we believe offers a more informative model as discussed in the concluding remarks mentioned at the end of this paper.

In Section 2, we offer a matricial view of investigating questions on the quasinilpotency of commutators of compact operators. We do this by connecting the finite matrices that appear in the block-tridiagonal matrix form of a compact operator with the simultaneous triangularization of pairs of finite matrices. The advantage of this approach is that there are several efficient and tractable ways of determining the simultaneous triangularizability of a pair of *finite* matrices, for instance, see [13, Chapters 1-4] that includes many classical theorems such as Engel, McCoy, Levitzki, Kolchin, and Kaplansky theorems, to name a few. We believe this matrix approach has the potential to offer a better understanding of the structure of commutators, in general.

In Section 3, we investigate the structure of a compact operator by slicing its block-tridiagonal matrix form appropriately and discuss the consequences of it. In the finite-dimensional case, it is well-known that every matrix is unitarily equivalent to a lower triangular matrix, which can also be viewed as the sum of a diagonal matrix (an upper triangular matrix) and a nilpotent matrix (a strictly lower triangular matrix). In this paper, we establish an analog of this for a compact operator via its block-tridiagonal matrix form: the matrix of every compact operator is unitarily similar to the sum of a *strictly lower triangular matrix* and an *upper triangular matrix*, where each row and each column have at most finitely many nonzero complex scalar entries. At the end, we make some remarks on the block-tridiagonal matrix form of a compact operator allowing one to dissect the compact operator into some nice pieces, which exhibit certain universal phenomena that occur in certain parts of it, if not fully.

Though the results and observations in this paper are pretty straightforward, the purpose is to illuminate the matricial approach in the study of compact operators.

We shall use matrix and bounded operator interchangeably depending on the context.

2. On quasinilpotency of a commutator of compact operators

The commutator structure of bounded operators is very well understood, that is, a bounded operator is a commutator of bounded operators $AB-BA$ if and only if it is not of the form $\lambda I+K$, where λ is a nonzero complex number and K is a compact operator [3]. Unlike the well-understood commutator structure of bounded operators, the question on the commutator structure of compact operators posed by Pearcy and Topping in 1971 remains open [12, Problem 1]: is every compact operator a commutator of compact operators? The first author of this paper together with Jireh Loreaux, John Petrovic, and Gary Weiss have made some progress on this problem, see [10].

In this section, we provide an application of the simultaneous triangularization of finite matrices in the study of commutator of compact operators via the block-tridiagonal matrix form associated with a compact operator. We determine sufficient conditions on compact operators C and Z for which the commutator $CZ-ZC$ is quasinilpotent via the simultaneous triangularization of finite matrices that appear in the block-tridiagonal matrix form for C and Z (see Theorem 2.6 below).

We recall here the definition of simultaneous triangularization of a family of finite matrices.

Definition 2.1. *We say that a family of finite matrices is simultaneously triangularizable if there is an orthonormal basis of a Hilbert space such that all matrices in the family are upper triangular matrix with scalar entries (or lower triangular matrix with scalar entries) with respect to that basis.*

We start by first recalling the well-known result of A. Albert and B. Muckenhoupt on commutator of finite matrices, namely, $C \in M_n(\mathbb{C})$ [1, Theorem]: if $\mathrm{tr}(C) = 0$, then there exist matrices A and B such that $AB - BA = C$.

An immediate consequence is that nilpotent matrices are commutator of matrices.

For operators on an infinite-dimensional Hilbert space, compact operators are the direct generalization of finite matrices, so a natural question for the compact counterpart is as follows: *is every nilpotent compact operator a commutator of compact operators?* An affirmative answer to this question stems from a more than half-a-century old open question of Pearcy and Topping posed in 1971: is every compact operator a commutator of compact operators? In 2017, Ken Dykema and Amudhan Krishnaswamy Usha proved that every nilpotent compact operator is a commutator of compact operators [4, Theorem 3.2]. Their work involved the following structure theorem for nilpotents, which they modified for entries to act on the same space [4, Lemma 2.2]: for every nilpotent $T \in B(\mathcal{H})$, i.e., $T^n = 0$ for some $n > 1$, there exists a decomposition $\mathcal{H} = \oplus_{i=1}^{n} \mathcal{H}_i$ of the Hilbert space $\mathcal{H}$ into the direct sum of an orthogonal family of subspaces $\mathcal{H}_1, \mathcal{H}_2, \ldots, \mathcal{H}_n$ such that the matrix of T corresponding to this decomposition has the upper triangular form [7, Theorem 1]. This finite block matrix model allowed them to deal with only finitely many operator equations to construct appropriate compact operators C and Z for which $CZ - ZC = T$, though the computations are quite intricate.

In the case of operators acting on finite-dimensional Hilbert space, the spectrum of an operator being $\{0\}$ implies that the operator must be nilpotent. In contrast, there are operators acting on infinite-dimensional Hilbert spaces whose spectrum is $\{0\}$, but they are not nilpotent. Such operators are called quasinilpotent operators.

Our study is centered around the following question for the class of compact quasinilpotent operators that are not nilpotent.

Question. *Is every compact quasinilpotent operator a commutator of compact operators?*

Unlike the case of nilpotent compact operators, one does not have a finite matrix structure for compact quasinilpotent operators, so addressing such a question in full generality seems intractable! In the literature, some sufficient conditions on compact operators C and Z are provided for the quasinilpotency of $CZ - ZC$. For instance, if the pair of compact operators $\{C, Z\}$ is simultaneously triangularizable, then $[C, Z]$ is quasinilpotent [13, Theorem 7.3.3]; if every word in CZ and ZC is quasinilpotent, then $[C, Z]$ is quasinilpotent (a consequence of Turovskii's theorem [13, Corollary 8.1.14] applied to CZ and ZC). However, showing the existence of a maximal

chain of subspaces invariant under both C and Z for the pair $\{C, Z\}$ to be simultaneously triangularizable or determining if every word in CZ and ZC is quasinilpotent seems to us a difficult task at hand.

In the case of finite matrices, note that if C and Z are simultaneously triangularizable, then $CZ - ZC$ is unitarily similar to a strictly upper triangular matrix and hence is nilpotent. We combined this observation and the continuity of the spectrum on the set of compact operators to determine a sufficient condition involving finite matrices that yields quasinilpotency of the commutator $[C, Z]$ in Theorem 2.6. To do this, we reduced the problem from operator setup to matrix setup, by invoking Theorem 2.4 mentioned below and using the fact that for any $C, Z \in K(\mathcal{H})$, the commutator $[C, Z]$ is quasinilpotent if and only if $[U^*CU, U^*ZU]$ is quasinilpotent for each unitary $U \in \mathcal{U}(\mathcal{H})$.

A fundamental tool that we use throughout in this paper is the block-tridiagonal matrix form associated with a bounded operator and so, we briefly mention the matrix form here for a single bounded operator in Theorem 2.3 and for the set of two bounded operators in Theorem 2.4 ([11, Theorem 20.4], see also [9, Remark 2.2]). (For details, see [10, Section 4, Theorems 4.1-4.4, Remark 4.5]).

The following theorems give rise to block-tridiagonal matrix form for a single bounded operator and for a collection of N bounded operators.

Theorem 2.2 ([14, Theorem 2–Corollary 3, combined])**.** *Let $C \in \mathcal{B}(\mathcal{H})$ be arbitrary and let (e_n) be any orthonormal basis of $\mathcal{H}$. There is an orthonormal basis (f_n) is derived from (e_n) such that $f_1 = e_1$ and $e_n \in \vee_{k=1}^{3n} f_k$ for all $n \in \mathbb{N}$, and moreover, the matrix of C with respect to (f_n) takes the staircase form*

$$C = \begin{pmatrix} * & * & * & 0 & \cdots & & & \\ * & * & * & * & * & * & 0 & \cdots \\ * & * & \cdots & & & & & \\ 0 & * & & & & & & \\ 0 & * & & & & & & \\ 0 & * & & & & & & \\ \vdots & 0 & & & & & & \end{pmatrix}, \tag{2.1}$$

where this matrix has row and column support lengths $(3n)_{n=1}^{\infty}$.

If $S_1, \ldots, S_N$ is any finite set of selfadjoint operators, then there is an orthonormal basis (f_n) of $\mathcal{H}$, such that $f_1 = e_1$ and $e_n \in \vee_{k=1}^{1+(n-1)(N+1)} f_k$, and with respect to which S_k takes the form (2.1) with $(3n)_{n=1}^{\infty}$ replaced by $((N+1)n)_{n=1}^{\infty}$. When $S_1, \ldots, S_N$ are not necessarily selfadjoint, the row and column support lengths are given by $((2N+1)n)_{n=1}^{\infty}$, and the inclusion relationships are $e_n \in \vee_{k=1}^{1+(n-1)(2N+1)} f_k$.

The staircase form (2.1) allows us to represent a matrix in universal block tridiagonal form as in (2.2) below:

$$C = \begin{pmatrix} C_1 & A_1 & 0 & \ldots \\ B_1 & C_2 & A_2 & \ddots \\ 0 & B_2 & C_3 & \ddots \\ \vdots & \ddots & \ddots & \ddots \end{pmatrix}. \tag{2.2}$$

Theorem 2.3. [11, Theorem 20.4] *For $C \in \mathcal{B}(\mathcal{H})$, the block-tridiagonal partition of the matrix of C induced by (2.1) (i.e., in the derived basis (f_n) from Theorem 2.2) is given by (2.2), where the central blocks C_n have block size sequence (k_n) given by $k_1 = 1$ and $k_n = 2(3^{n-2})$ for $n > 1$.*

More generally, the necessary and sufficient condition that (2.2) cover the (2.1) staircase support entries is: choose k_1 arbitrarily and for $n \geq 1$, $k_{n+1} \geq 2(k_1 + k_2 + \cdots + k_n)$.

For any N operators in $\mathcal{B}(\mathcal{H})$, the block tridiagonal partition of their matrices induced by the analog of (2.1) for N (i.e., in the new basis (f_n)) is given by (2.2). For each of them simultaneously the central blocks have block size sequence (k_n) given by $k_1 = 1$ and $k_n = 2N((2N+1)^{n-2})$ for $n > 1$.

In particular, for 2 operators $k_n = 4(5^{n-2})$ for $n > 1$, and one has

Theorem 2.4. [10, Theorem 4.4] *For $A, B \in \mathcal{B}(\mathcal{H})$, there is an orthonormal basis $\{f_n\}$ with respect to which both A and B have matrices in the block-tridiagonal form. That is, there is a unitary operator U such that for each of them U^*AU and U^*BU simultaneously the central blocks have block size sequence $\langle k_n \rangle$ with $k_1 = 1$ and $k_n = 4(5^{n-2})$ for $n > 1$. That is, with respect to $\{f_n\}$, the block-tridiagonal matrix forms of $C := U^*AU$ and*

$Z := U^*BU$ are given by

$$C = \begin{pmatrix} C_1 & A_1 & \mathbf{0} & \dots \\ B_1 & C_2 & A_2 & \ddots \\ \mathbf{0} & B_2 & C_3 & \ddots \\ \vdots & \ddots & \ddots & \ddots \end{pmatrix} \quad \textit{and} \quad Z = \begin{pmatrix} Z_1 & X_1 & \mathbf{0} & \dots \\ Y_1 & Z_2 & X_2 & \ddots \\ \mathbf{0} & Y_2 & Z_3 & \ddots \\ \vdots & \ddots & \ddots & \ddots \end{pmatrix}. \tag{2.3}$$

The central blocks C_n and Z_n are of size $k_n \times k_n$, where $k_1 = 1$, $k_n = 4(5^{n-2})$ for $n \geq 2$. The blocks $\{A_n\}$ and $\{X_n\}$ are of size $k_n \times 5k_n$ and the blocks $\{B_n\}$ and $\{Y_n\}$ are of size $5k_n \times k_n$.

Remark 2.5. *It is known that for any compact operators A and B, there exist sequences of finite-rank operators F_n and E_n such that $F_n \to A$ and $E_n \to B$ in the norm as $n \to \infty$. But, one may not be able to make sense of simultaneous triangularizability of these F_n, E_n's. The advantage of the simultaneous block-tridiagonalization of compact operators C and Z in its matrix form allowed us to employ the simultaneous triangularization of its finite-rank operators in its finite matrix form that converges to C and Z in the norm.*

Since $[C, Z]$ is quasinilpotent if and only if $[U^*CU, U^*ZU]$ is quasinilpotent for each unitary $U \in \mathcal{U}(\mathcal{H})$ and that by Theorem 2.4, there is a unitary U for which both U^*CU and U^*ZU are simultaneously in block-tridiagonal matrix form, so we consider both C and Z with their block-tridiagonal matrix form with respect to a fixed orthonormal basis $\{f_n\}$.

We next introduce a few notations that are required to do the computations. Let C and Z be compact operators on $\mathcal{H}$. We denote by $\{f_n\}$ the orthonormal basis with respect to which the matrix forms of both C and Z are given by Equation (2.3). Let P_n be the projection onto the subspace spanned by $\{f_1, f_2, \dots, f_{1+4+\cdots+4(5^{n-2})}\}$ for $n > 1$ and P_1 the projection onto the span of $\{f_1\}$. Then, for each $n \geq 1$, the matrix forms for $C'_n := P_nCP_n$ and $Z'_n := P_nZP_n$ are respectively given by

$$C'_n = \begin{pmatrix} C_1 & A_1 & \mathbf{0} & \mathbf{0} & \dots & \dots & \dots \\ B_1 & C_2 & A_2 & \mathbf{0} & \dots & \dots & \dots \\ \mathbf{0} & \ddots & \ddots & \ddots & \ddots & \ddots & \ddots \\ \mathbf{0} & \ddots & B_{n-2} & C_{n-1} & A_{n-1} & \mathbf{0} & \dots \\ \vdots & \mathbf{0} & \ddots & B_{n-1} & C_n & \mathbf{0} & \dots \\ \vdots & \mathbf{0} & \mathbf{0} & \mathbf{0} & \mathbf{0} & \mathbf{0} & \dots \\ \vdots & \vdots & \vdots & \vdots & \ddots & \ddots & \ddots \end{pmatrix} \tag{2.4}$$

$$Z'_n = \begin{pmatrix} Z_1 & X_1 & \mathbf{0} & \mathbf{0} & \dots & \dots & \dots \\ Y_1 & Z_2 & X_2 & \mathbf{0} & \dots & \dots & \dots \\ \mathbf{0} & \ddots & \ddots & \ddots & \ddots & \ddots & \ddots \\ \mathbf{0} & \ddots & Y_{n-2} & Z_{n-1} & X_{n-1} & \mathbf{0} & \dots \\ \vdots & \mathbf{0} & \ddots & Y_{n-1} & Z_n & \mathbf{0} & \dots \\ \vdots & \mathbf{0} & \mathbf{0} & \mathbf{0} & \mathbf{0} & \mathbf{0} & \dots \\ \vdots & \vdots & \vdots & \vdots & \ddots & \ddots & \ddots \end{pmatrix}$$

Since C and Z are compact operators, it follows from [2, Theorem 2] that

$$\|C_n\|, \|B_n\|, \|A_n\|, \|X_n\|, \|Y_n\|, \|Z_n\| \to 0 \text{ as } n \to \infty,$$

which further implies that $\|C - C'_n\| \to 0$ and $\|Z - Z'_n\| \to 0$ as $n \to \infty$. It immediately follows that $\|[C, Z] - [C'_n, Z'_n]\| \to 0$ as $n \to \infty$. We re-write C'_n and Z'_n as 2×2 block-matrices with respect to the Hilbert space decomposition $\mathcal{H} = \mathcal{H}_{5^{n-1}} \oplus (\mathcal{H}_{5^{n-1}})^{\perp}$, where the upper left corner of each block-matrix is a square matrix of size 5^{n-1} for $n > 1$ and the other entries of the matrix are zero matrices. So,

$$C'_n = \begin{pmatrix} C''_n & \mathbf{0} \\ \mathbf{0} & \mathbf{0} \end{pmatrix} \quad \text{and} \quad Z'_n = \begin{pmatrix} Z''_n & \mathbf{0} \\ \mathbf{0} & \mathbf{0} \end{pmatrix},$$

where

$$C''_n = \begin{pmatrix} C_1 & A_1 & \mathbf{0} & \dots & \mathbf{0} \\ B_1 & C_2 & A_2 & \mathbf{0} & \vdots \\ \mathbf{0} & \ddots & \ddots & \ddots & \mathbf{0} \\ \vdots & \ddots & B_{n-2} & C_{n-1} & A_{n-1} \\ \mathbf{0} & \mathbf{0} & \ddots & B_{n-1} & C_n \end{pmatrix} \text{ and } Z''_n = \begin{pmatrix} Z_1 & X_1 & \mathbf{0} & \dots & \mathbf{0} \\ Y_1 & Z_2 & X_2 & \mathbf{0} & \vdots \\ \mathbf{0} & \ddots & \ddots & \ddots & \mathbf{0} \\ \vdots & \ddots & Y_{n-2} & Z_{n-1} & X_{n-1} \\ \mathbf{0} & \mathbf{0} & \ddots & Y_{n-1} & Z_n \end{pmatrix}. \tag{2.5}$$

Since each pair $\{C''_n, Z''_n\}$ is simultaneously triangularizable for $n \geq 1$, there exist unitary matrix $U_n \in M_{5^{n-1}}(\mathbb{C})$ such that $U_n^* C''_n U_n$ and $U_n^* Z''_n U_n$

are both in upper triangular form. This further implies that

$$(U_n^*\oplus I)C_n'(U_n\oplus I) = \begin{pmatrix} U_n^*C_n''U_n & \mathbf{0} \\ \mathbf{0} & \mathbf{0} \end{pmatrix} \text{ and } (U_n^*\oplus I)Z_n'(U_n\oplus I) = \begin{pmatrix} U_n^*Z_n''U_n & \mathbf{0} \\ \mathbf{0} & \mathbf{0} \end{pmatrix},$$

where I is the identity operator on the orthogonal complement of the finite-dimensional subspace $\mathcal{H}_n$ of $\mathcal{H}$ of dimension 5^{n-1}. Therefore,

$$C_n' = (U_n\oplus I)\begin{pmatrix} U_n^*C_n''U_n & \mathbf{0} \\ \mathbf{0} & \mathbf{0} \end{pmatrix}(U_n^*\oplus I) \text{ and } Z_n' = (U_n\oplus I)\begin{pmatrix} U_n^*Z_n''U_n & \mathbf{0} \\ \mathbf{0} & \mathbf{0} \end{pmatrix}(U_n^*\oplus I).$$

This implies that the commutator

$$[C_n', Z_n'] = C_n'Z_n' - Z_n'C_n' = (U_n \oplus I)\begin{pmatrix} U_n^*(C_n''Z_n'' - Z_n''C_n'')U_n & \mathbf{0} \\ \mathbf{0} & \mathbf{0} \end{pmatrix}(U_n^* \oplus I)$$

is nilpotent because $U_n^*(C_n''Z_n'' - Z_n''C_n'')U_n$ is a strictly upper triangular finite matrix and so the matrix on the left-hand side of the above latter equality is nilpotent. Moreover, since $\|C_n' - C\| \to 0$ and $\|Z_n' - Z\| \to 0$ as $n \to \infty$, we have $\|[C_n', Z_n'] - [C, Z]\| \to 0$ as $n \to \infty$. By the continuity of the spectrum on compact operators, $\sigma([C_n', Z_n']) \to \sigma([C, Z])$. However, $\sigma([C_n', Z_n']) = \{0\}$ for each $n \geq 1$, and so $\sigma([C, Z]) = \{0\}$, i.e., $[C, Z]$ is quasinilpotent. (Here, spectrum is considered as a set-valued function on the Banach space of all bounded linear operators acting on a Hilbert space mapping each operator to its spectrum. The spectrum is continuous at an operator A, if $\sigma(A_n) \to \sigma(A)$ whenever $A_n \to A$ in the norm. The notions of semicontinuity and continuity of the spectrum are explained in [6, Problems 103-105].)

We now summarize the above discussion in the theorem below.

Theorem 2.6. *Let C and Z be compact operators, where both simultaneously have block-tridiagonal matrix form given in Equation (2.3). If each pair $\{C_n'', Z_n''\}$ is simultaneously triangularizable for $n \geq 1$, then $CZ - ZC$ is quasinilpotent.*

As one would anticipate, the converse of Theorem 2.6 does not hold in general. Towards this, we construct $C, Z \in \mathcal{K}(\mathcal{H})$, which are in block-tridiagonal matrix forms given in Equation (2.3) such that $CZ - ZC$ is quasinilpotent, but the pair $\{C_n'', Z_n''\}$ is not triangularizable for some $n \in \mathbb{N}$. To do so, we first observe the following: for each $n \geq 3$, one can construct matrices $A_n, B_n \in M_n(\mathbb{C})$ and a polynomial $p_n(x, y) \in \mathbb{C}[x, y]$ such that

$[A_n, B_n]$ is nilpotent, but $p_n(A_n, B_n)[A_n, B_n]$ is not nilpotent. Indeed, take

$$A_n = \begin{pmatrix} 0 & 1 & 0 & \cdots & 0 \\ 0 & 0 & 1 & \cdots & 0 \\ \vdots & \vdots & \vdots & \ddots & \vdots \\ 0 & 0 & 0 & \cdots & 1 \\ 0 & 0 & 0 & \cdots & 0 \end{pmatrix}_{n\times n} \qquad B_n = \begin{pmatrix} 0 & 0 & 0 & \cdots & 0 \\ 0 & 0 & 0 & \cdots & 0 \\ \vdots & \vdots & \vdots & \ddots & \vdots \\ 0 & 0 & 0 & \cdots & 0 \\ 1 & 0 & 0 & \cdots & 0 \end{pmatrix}_{n\times n}.$$

Then,

$$A_nB_n - B_nA_n = \begin{pmatrix} 0 & 0 & 0 & \cdots & 0 \\ \vdots & \vdots & \vdots & \ddots & \vdots \\ 0 & 0 & 0 & \cdots & 0 \\ 0 & 0 & 0 & \cdots & 0 \\ 1 & 0 & 0 & \cdots & 0 \\ 0 & -1 & 0 & \cdots & 0 \end{pmatrix}_{n\times n}$$

is nilpotent, but

$$A_n^{n-2}(A_nB_n - B_nA_n) = \begin{pmatrix} 1 & 0 & 0 & \cdots & 0 \\ 0 & -1 & 0 & \cdots & 0 \\ 0 & 0 & 0 & \cdots & 0 \\ \vdots & \vdots & \vdots & \ddots & \vdots \\ 0 & 0 & 0 & \cdots & 0 \end{pmatrix}_{n\times n}$$

is clearly not nilpotent. It is easy to see that $\|A_n\| = 1 = \|B_n\|$ for all $n \geq 3$. Define

$$P_n = \frac{1}{n}A_n \quad \text{and} \quad Q_n = \frac{1}{n}B_n \tag{2.6}$$

for all $n \geq 3$. So, $\|P_n\| = \frac{1}{n} = \|Q_n\|$, and hence $\|P_n\|, \|Q_n\| \to 0$ as $n \to \infty$. Observe that

$$P_nQ_n - Q_nP_n = \begin{pmatrix} 0 & 0 & 0 & \cdots & 0 \\ \vdots & \vdots & \vdots & \ddots & \vdots \\ 0 & 0 & 0 & \cdots & 0 \\ 0 & 0 & 0 & \cdots & 0 \\ \frac{1}{n^2} & 0 & 0 & \cdots & 0 \\ 0 & \frac{-1}{n^2} & 0 & \cdots & 0 \end{pmatrix}_{n\times n},$$

which is nilpotent, but for $p_n(x, y) = x^{n-2}$,

$$p_n(P_n,Q_n)[P_n,Q_n]=P_n^{n-2}(P_nQ_n-Q_nP_n)=\begin{pmatrix} n^{-n} & 0 & 0 & \cdots & 0\\ 0 & -n^{-n} & 0 & \cdots & 0\\ 0 & 0 & 0 & \cdots & 0\\ \vdots & \vdots & \vdots & \ddots & \vdots\\ 0 & 0 & 0 & \cdots & 0\end{pmatrix}_{n\times n} \tag{2.7}$$

is not nilpotent.

We are now ready to construct C and Z for which $CZ-ZC$ is quasinilpotent, but the pair $\{C_n'', Z_n''\}$ is not triangularizable for some $n\in\mathbb{N}$.

Example 2.7. *For $k_1=1$ and $k_n=4(5^{n-2})$ for each $n\geq 2$, let*

$$C=\begin{pmatrix} C_1 & 0 & 0 & \cdots\\ 0 & C_2 & 0 & \cdots\\ 0 & 0 & C_3 & \cdots\\ \vdots & \vdots & \vdots & \ddots\end{pmatrix} \quad \textit{and} \quad Z=\begin{pmatrix} Z_1 & 0 & 0 & \cdots\\ 0 & Z_2 & 0 & \cdots\\ 0 & 0 & Z_3 & \cdots\\ \vdots & \vdots & \vdots & \ddots\end{pmatrix},$$

where $C_n:=P_{k_n}$ and $Z_n:=Q_{k_n}$ for $n\geq 1$.

We first prove that $[C,Z]$ is a compact quasinilpotent operator. Clearly, as $\|P_{k_n}\|,\|Q_{k_n}\|\to 0$ as $n\to\infty$, we have $\|C_n\|,\|Z_n\|\to 0$ as $n\to\infty$. Therefore, it follows from [2, Theorem 2] *that C and Z are compact operators.*

By construction, $[C_n,Z_n]$ is nilpotent for each $n\geq 1$ and $[C_n,Z_n]\to 0$ as $n\to\infty$. The commutator $CZ-ZC$ is given by

$$[C,Z]=\begin{pmatrix} [C_1,Z_1] & 0 & 0 & \cdots\\ 0 & [C_2,Z_2] & 0 & \cdots\\ 0 & 0 & [C_3,Z_3] & \cdots\\ \vdots & \vdots & \vdots & \ddots\end{pmatrix}.$$

For each $n\geq 1$, the truncated sequence of nilpotent operators constructed from the above matrix form of $[C,Z]$ is defined as

$$[C,Z]_n:=\begin{pmatrix} [C_1,Z_1] & 0 & 0 & \cdots & 0 & 0 & \ldots\\ 0 & [C_2,Z_2] & 0 & \cdots & 0 & 0 & \ldots\\ \vdots & \vdots & \vdots & \ddots & \vdots & \vdots & \ldots\\ 0 & 0 & 0 & \cdots & [C_n,Z_n] & 0 & \ldots\\ 0 & 0 & 0 & \ldots & 0 & 0 & \ddots\\ \vdots & \vdots & \vdots & \vdots & \vdots & \ddots & \ddots\end{pmatrix}.$$

A straightforward computation shows that $\|[C,Z]_n - [C,Z]\| \to 0$ *as* $n \to \infty$. *Moreover, since the norm limits of compact quasinilpotents (here in particular,* $[C,Z]_n$ *are nilpotents) is again a compact quasinilpotent* [13, Corollary 7.2.11]*, so is the commutator* $[C,Z]$. *Following the matrix notation from Equation (2.5), we have*

$$C'_n = \begin{pmatrix} C_1 & 0 & 0 & \cdots & 0 & 0 & \dots \\ 0 & C_2 & 0 & \cdots & 0 & 0 & \dots \\ 0 & 0 & C_3 & \cdots & 0 & 0 & \dots \\ \vdots & \vdots & \vdots & \ddots & \vdots & \vdots & \\ 0 & 0 & 0 & \cdots & C_n & 0 & \dots \\ & & & & & & \ddots \\ 0 & 0 & 0 & \dots & 0 & 0 & \ddots \\ \vdots & \vdots & \vdots & \vdots & \vdots & \ddots & \ddots \end{pmatrix}, Z'_n = \begin{pmatrix} Z_1 & 0 & 0 & \cdots & 0 & 0 & \dots \\ 0 & Z_2 & 0 & \cdots & 0 & 0 & \dots \\ 0 & 0 & Z_3 & \cdots & 0 & 0 & \dots \\ \vdots & \vdots & \vdots & \ddots & \vdots & \vdots & \\ 0 & 0 & 0 & \cdots & Z_n & 0 & \dots \\ & & & & & & \ddots \\ 0 & 0 & 0 & \dots & 0 & 0 & \ddots \\ \vdots & \vdots & \vdots & \vdots & \vdots & \ddots & \ddots \end{pmatrix},$$

where $C_n, Z_n \in M_{k_n}(\mathbb{C})$ *for each* $n \geq 1$. *We next show that the pair* $\{C''_n, Z''_n\}$ *is not triangularizable for each* $n \geq 2$. *(The target was to find some* $n \in \mathbb{N}$ *such that* $\{C''_n, Z''_n\}$ *is not triangularizable. Interestingly, this example yields more than we need - we shall prove that* $\{C''_n, Z''_n\}$ *is not triangularizable for any* $n \geq 2$.*) The proof is by contradiction. Indeed, assume that* $\{C''_n, Z''_n\}$ *is triangularizable for some* $n \geq 2$. *Then, by McCoy's Theorem for finite matrices (*[13, Theorem 1.3.4]*),* $p(C''_n, Z''_n)[C''_n, Z''_n]$ *is nilpotent for every* $p(x,y) \in \mathbb{C}[x,y]$. *Choose* $p(x,y) = p_{k_n}(x,y) = x^{k_n - 2}$. *We recall that* $k_1 = 1$ *and* $k_n = 4(5^{n-2})$ *for* $n > 1$. *We have* $p(C'_n, Z'_n)[C'_n, Z'_n]$ *equal to*

$$\begin{pmatrix} p(C_1,Z_1)[C_1,Z_1] & 0 & 0 & \cdots & 0 & 0 & \dots \\ 0 & p(C_2,Z_2)[C_2,Z_2] & 0 & \cdots & 0 & 0 & \dots \\ & & \ddots & & & & \\ 0 & 0 & & \cdots & 0 & 0 & \dots \\ \vdots & \vdots & \vdots & \ddots & \vdots & \vdots & \\ 0 & 0 & 0 & \cdots & p(C_n,Z_n)[C_n,Z_n] & 0 & \dots \\ & & & & & & \ddots \\ 0 & 0 & 0 & \dots & 0 & 0 & \ddots \\ \vdots & \vdots & \vdots & \vdots & \vdots & \ddots & \ddots \end{pmatrix}.$$

And so,

$$p(C'_n, Z'_n)[C'_n, Z'_n] = \begin{pmatrix} p(C''_n, Z''_n)[C''_n, Z''_n] & \mathbf{0} \\ \mathbf{0} & \mathbf{0} \end{pmatrix}.$$

Note that for $k_1 = 1$, *the* 1×1 *matrix* $p(C_1,Z_1)[C_1,Z_1]$ *is always zero and for each* $n > 1$, *the matrix* $p(C_n,Z_n)[C_n,Z_n]$ *is of size greater than* 3. *For convenience, let* $T_n := p(C'_n, Z'_n)[C'_n, Z'_n]$ *and* $S_j := p(C_j,Z_j)[C_j,Z_j]$ *for each* $1 \leq j \leq n$. *Recall that by hypothesis* $C_j := P_{k_j}, Z_j := Q_{k_j}$, *where* P_{k_j}, Q_{k_j} *are given by Equation (2.6). Since the matrix of* T_n *has finitely*

many nonzero blocks, it follows from [6, Chapter 11, §98] *that for* $n > 1$,

$$\sigma(T_n) = \bigcup_{j=1}^{n} \sigma(S_j).$$

Since $p(C_n'', Z_n'')[C_n'', Z_n'']$ *is nilpotent,* T_n *is also nilpotent, and thus we have* $\sigma(T_n) = \{0\}$. *Hence* $\bigcup_{j=1}^{n} \sigma(S_j) = \{0\}$. *This implies that* $\sigma(S_j) = \{0\}$ *for all* $1 \leq j \leq n$. *However, for* $j > 1$, $\sigma(S_j) \neq \{0\}$ *as* S_j *is not nilpotent by construction in Equation (2.7) (in there, replace* j *with* $k_j = 4(5^{j-2}) > 3$*), and hence a contradiction. Therefore,* $\{C_n'', Z_n''\}$ *is not triangularizable for all* $n \geq 2$.

Remark 2.8. *The construction above works more generally when* $\{k_n\}_{n=1}^{\infty}$ *is any strictly increasing sequence of natural numbers. In fact, if* $k_1 = 1$, *then* $\{C_n'', Z_n''\}$ *is not triangularizable for all* $n \geq 2$. *If* $k_1 > 1$, *then* $\{C_n'', Z_n''\}$ *is not triangularizable for all* $n \geq 1$.

3. On the structure of a compact operator via its block-tridiagonal matrix form

In [5, Theorem 2], Fong and Sourour showed that every bounded operator on a Hilbert space is the sum of two quasinilpotent operators if and only if it is not a nonzero scalar plus a compact operator. As a consequence, they showed in [5, Corollary 2] that every compact operator is a sum of two compact quasinilpotent operators.

In this section, we establish an alternate structure theorem for a compact operator through the direct use of the block-tridiagonal matrix form associated with it. In the finite-dimensional case, it is well-known that every matrix is unitarily equivalent to an upper triangular matrix, which is the sum of a diagonal matrix (upper triangular matrix) and a nilpotent matrix (strictly lower triangular matrix). Here, we prove via the block-tridiagonal matrix form of a compact operator that every compact operator is unitarily similar to the sum of an *upper triangularizable compact operator* and a *strictly lower triangularizable compact quasinilpotent operator*. Moreover, in this matrix representation, each row and each column has at most finitely many nonzero complex scalar entries. This is discussed below and summarized in Theorem 3.1.

For any compact operator T (more generally, for any bounded operator), by [11, Theorem 20.4], there is an orthonormal basis $\{f_n\}$ such that

with respect to that basis the matrix representation of $K := W^*TW$ has a block-tridiagonal matrix form, where the central blocks C_n have block size sequence $\{k_n\}$ given by $k_1 = 1$ and $k_n = 2(3^{n-2})$ for $n > 1$. (Here W is the unitary associated with the change of basis.) That is, with respect to the orthonormal basis $\{f_n\}$ obtained in [11, Theorem 20.4], one has

$$K = \begin{pmatrix} C_1 & A_1 & 0 & \dots \\ B_1 & C_2 & A_2 & \ddots \\ 0 & B_2 & C_3 & \ddots \\ \vdots & \ddots & \ddots & \ddots \end{pmatrix}. \tag{3.1}$$

We then split this matrix in two pieces as follows.

$$K = \begin{pmatrix} C_1 & A_1 & 0 & \dots \\ B_1 & C_2 & A_2 & \ddots \\ 0 & B_2 & C_3 & \ddots \\ \vdots & \ddots & \ddots & \ddots \end{pmatrix} = \underbrace{\begin{pmatrix} C_1 & A_1 & 0 & \dots \\ 0 & C_2 & A_2 & \ddots \\ 0 & 0 & C_3 & \ddots \\ \vdots & \ddots & \ddots & \ddots \end{pmatrix}}_{S} + \underbrace{\begin{pmatrix} 0 & 0 & 0 & \dots \\ B_1 & 0 & 0 & \ddots \\ 0 & B_2 & 0 & \ddots \\ \vdots & \ddots & \ddots & \ddots \end{pmatrix}}_{Q}, \tag{3.2}$$

where the"first piece" of K denoted by S is an upper triangularizable compact operator and the "second piece" of K denoted by Q is a compact quasinilpotent operator. Indeed, since K is a compact operator, it follows from [2, Theorem 2] that $\|A_n\|, \|B_n\|, \|C_n\| \to 0$ as $n \to \infty$. Since $\|B_n\| \to 0$ as $n \to \infty$, so Q is a compact operator. For each $n \geq 1$, let

$$F_n := \begin{pmatrix} 0 & 0 & 0 & \dots & \dots & \dots & \dots \\ B_1 & 0 & 0 & \ddots & \ddots & \ddots & \ddots \\ 0 & B_2 & 0 & \ddots & \ddots & \ddots & \ddots \\ \vdots & \ddots & \ddots & \ddots & \ddots & \ddots & \ddots \\ \vdots & \ddots & \ddots & B_n & \ddots & \ddots & \ddots \\ \vdots & \ddots & \ddots & \ddots & 0 & \ddots & \ddots \\ \vdots & \ddots & \ddots & \ddots & \ddots & \ddots & \ddots \end{pmatrix}.$$

Note that $\|Q - F_n\| \to 0$ as $n \to \infty$, where F_n's are finite-rank nilpotent operators. And as the norm limit of nilpotent operators is quasinilpotent, therefore Q is quasinilpotent.

Consider the compact operator S in Equation (3.2). Since C_n is a square matrix for each $n \geq 1$, C_n is upper triangularizable. That is, for each $n \geq 1$,

there exists a unitary matrix U_n such that $U_n^* C_n U_n = \Delta_n$, where Δ_n is an upper triangular matrix. Consider the unitary operator $U := \oplus U_n$ on the Hilbert space $\mathcal{H}$. Then, a straightforward computation shows that U^*SU is an upper triangular matrix with complex scalar entries. Therefore, S represents a compact operator whose matrix is upper triangularizable. Let $\Delta := U^*SU$. Then the matrix form associated with Δ is given by

$$\Delta = \begin{pmatrix} \Delta_1 & A_1' & 0 & \dots \\ 0 & \Delta_2 & A_2' & \ddots \\ 0 & 0 & \Delta_3 & \ddots \\ \vdots & \ddots & \ddots & \ddots \end{pmatrix}. \tag{3.3}$$

Consider the same unitary operator U constructed above using the U_n's. Since $K = S + Q$ (Equation (3.2)),

$$U^*KU = \begin{pmatrix} \Delta_1 & A_1' & 0 & \dots \\ 0 & \Delta_2 & A_2' & \ddots \\ 0 & 0 & \Delta_3 & \ddots \\ \vdots & \ddots & \ddots & \ddots \end{pmatrix} + \begin{pmatrix} 0 & 0 & 0 & \dots \\ U_2^* B_1 U_1 & 0 & 0 & \ddots \\ 0 & U_3^* B_2 U_2 & 0 & \ddots \\ \vdots & \ddots & \ddots & \ddots \end{pmatrix} = \Delta + U^*QU. \tag{3.4}$$

Since $K = W^*TW$, one has $U^*W^*TWU = \Delta + U^*QU$. Let $U_0 := WU$, which is a unitary operator. Therefore, for a given compact operator T, we have

$$U_0^* T U_0 = \Delta + U^*QU, \tag{3.5}$$

where Δ is an upper triangular matrix representing a compact operator and U^*QU is a quasinilpotent matrix representing a compact quasinilpotent operator.

Recall that in the finite-dimensional case, a finite matrix A is unitarily equivalent to an upper triangular matrix, i.e., $V^*AV = D + Q$, where D is a diagonal matrix (upper triangular matrix) and Q is a nilpotent matrix (strictly lower triangular matrix) for some unitary matrix V. In the language of the spectrum of a finite matrix, $V^*AV = D + Q$ implies that

$$\sigma(V^*AV - D) = \sigma(Q) = \{0\}. \tag{3.6}$$

One can view Equation (3.5) as an infinite-dimensional analog of the above mentioned finite-dimensional matrix decomposition in the following theorem.

To summarize:

Theorem 3.1. *(A structure theorem for compact operators: an infinite-dimensional analog of finite matrices) Every compact operator T is unitarily equivalent to the sum of an upper triangularizable compact operator and a compact quasinilpotent operator. That is, following the notation from Equation (3.5), one has*

$$U_0^* T U_0 = \Delta + U^* Q U.$$

Moreover, $\sigma(U_0^* T U_0 - \Delta) = \{0\}$, which is an infinite-dimensional analog of Equation (3.6). In particular, if $B_n = 0$ for $n \geq 1$, then W^*TW is upper triangularizable (i.e., there is an orthonormal basis with respect to which the matrix of W^*TW is upper triangular with complex scalar entries).

Alternatively, every compact operator is unitarily equivalent to a compact operator K, which has a normal part N such that $K - N$ is a sum of two quasinilpotent compact operators. One can see this by splitting U^*SU in Equation (3.3) into the sum of a diagonal matrix and a strictly upper triangular matrix. (This strictly upper triangular matrix represents a compact quasinilpotent operator.)

Concluding remarks

We filter out some observations from Sections 2-3 that are perhaps worth mentioning when we consider the block-tridiagonal matrix form associated with a compact operator or a pair of compact operators. In particular, observations (i)-(iii) below reveal that a compact operator is always triangularizable with a quasinilpotent remainder.

(1) In the finite-dimensional case, every matrix is upper triangularizable. Since every compact operator is block-tridiagonalizable (see Equation (3.1), there is a piece of it, namely, S in Equation (3.2), which is always upper triangularizable with complex scalar entries and each row has at most finitely many nonzero entries. That is, if T is a compact operator, then $W^*TW - Q$ is upper triangularizable, see Equation (3.2).

(2) A pair of compact operators $\{K_1, K_2\}$ is simultaneously block-tridiagonalizable (Theorem 2.4, Section 2). If K_1 and K_2 both are simultaneously in block-tridiagonal matrix forms, then by extracting its upper triangularizing pieces, the pair $\{K_1 - S_1, K_2 - S_2\}$ is

always strictly lower triangular with complex scalar entries. Indeed, the set $\{K_1 - S_1, K_2 - S_2\} = \{Q_1, Q_2\}$, where $K_1 = S_1 + Q_1$ and $K_2 = S_2 + Q_2$ is given by Equation (3.2).

(3) The matrix splitting given in Equation (3.2) of the block-tridiagonal matrix form associated with a compact operator (or a pair of compact operators) touches upon another question on commutators, namely, on commutators that are trace class with trace zero. In particular, it is of interest to know if $C, Z \in B(\mathcal{H})$ with C or Z a special operator and that $[C, Z] = CZ - ZC$ is trace class, then is $\mathrm{tr}(CZ - ZC) = 0$? (See [8] and references therein.) Because the trace is unitarily invariant and that T is trace class if and only if U^*TU is trace class for each unitary operator U, with no loss of generality, the above question can be restated for when both C and Z are in block-tridiagonal matrix forms: *if $C, Z \in B(\mathcal{H})$ with both C and Z in block-tridiagonal matrix form and that $[C, Z] = CZ - ZC$ is trace class, then is $\mathrm{tr}(CZ - ZC) = 0$?*

We have the following partial answer to it: for C and Z compact operators each with block-tridiagonal matrix form, if the commutator $[C-S_1, Z-S_2]$ is trace class, then $\mathrm{tr}([C-S_1, Z-S_2]) = 0$, where $C = S_1 + Q_1$ and $Z = S_2 + Q_2$ given in Equation (3.2). Indeed, $\mathrm{tr}([C - S_1, Z - S_2]) = \mathrm{tr}(Q_1Q_2 - Q_2Q_1)$ and note that the diagonal entries of $Q_1Q_2 - Q_2Q_1$ are all equal to zero, so the trace is zero. (In [8, Theorems 4-5], F. Kittaneh proved that $\mathrm{tr}(CZ - ZC) = 0$ under certain special sufficient conditions on C and Z.)

(4) For a compact operator K with matrix decomposition given in Equation (3.2), if for each $n \geq 1$, C_n is a nilpotent matrix, then K is *unitarily* similar to a *zero-diagonal* operator. In this case, K is a sum of two quasinilpotent operators.

Acknowledgement: We are grateful for the support by the Science and Engineering Research Board, Core Research Grant 002514.

References

[1] Albert, A. A., Muckenhoupt, B., On matrices of trace zeros, *Michigan Mathematics Journal*, **4** (1957), 1-3.

[2] Damir, B., Boris, G., A note on compact operators and operator matrices, *Mathematical Communications*, **4** (1999), 159-165

[3] Brown, A., Pearcy, C., Structure of commutators of operators, *Annals of Mathematics. Second Series*, **82** (1965), 112-127.

[4] Dykema, K., Krishnaswamy-Usha, A., Nilpotent elements of operator ideals as single commutators, *Proceedings of the American Mathematical Society*, **146** (2018), 3031-3037.

[5] Fong, C. K., Sourour, A. R., Sums and products of quasinilpotent operators, *Proceedings of the Royal Society of Edinburgh Sect. A*, **99** (1984), 193-200.

[6] Halmos, P. R., A Hilbert space problem book, *Springer-Verlag, New York-Berlin*, (1982), xvii+369.

[7] Halmos, P. R., Capacity in Banach algebras, *Indiana University Mathematics Journal*, **20** (1970), 855-863.

[8] Kittaneh, F., Some trace class commutators of trace zero, *Proceedings of the American Mathematical Society*, **113** (1991), 655-661.

[9] Marcoux, L. W., Quasidiagonality and a generalized version of Berg's technique, *Journal of Operator Theory*, **35** (1996), 133-145.

[10] Loreaux, J., Patnaik, S., Petrovic, S., Weiss, G., On commutators of compact operators: generalizations and limitations of Anderson's approach, *Integral Equations Operator Theory*, **96** (2024), Paper No. 16, 28.

[11] Patnaik, S., Petrovic, S., Weiss, G., Universal block tridiagonalization in $\mathcal{B}(\mathcal{H})$ and beyond, *The mathematical legacy of Victor Lomonosov—operator theory*, **2** (2020), 317-326.

[12] Pearcy, C., Topping, D., On commutators in ideals of compact operators, *Michigan Mathematical Journal*, **18** (1971), 247-252.

[13] Radjavi, H., Rosenthal, P., Simultaneous triangularization, *Springer-Verlag, New York* (2000), xii+318.

[14] Weiss, G., Commutators of Hilbert-Schmidt operators II, *Integral Equations Operator Theory*, **3** (1980), 574-600.

Sasmita Patnaik
Department of Mathematics and Statistics,
Indian Institute of Technology,
Kanpur 208016, INDIA
E-mail: sasmita@iitk.ac.in

Rahul Sethi
School of Mathematics,
Georgia Institute of Technology,
GA 30332-0160, USA.
E-mail: rsethi43@gatech.edu

The Mathematics Student ISSN: 0025-5742
Vol. 95, Nos. 3-4, July - December (2026), 195–211

FINSLER SPACES WITH P-POWER (α, β)-METRIC

SRUTHY ASHA BABY

(Received : 31 - 07 - 2025 ; Accepted : 04 - 10 - 2026)

Abstract. In this paper, we study Finsler spaces with a p-power (α, β)-metric. The purpose of this paper is first to determine the two Finsler deformations of the p-power Finsler metric $F = \frac{(\alpha+\beta)^p}{\alpha^{p-1}}$ ($p \neq 0$). Consequently, we obtain the nonholonomic frame for Finsler space with p-power Finsler metrics, which can be considered as the generalisation of the Randers metric. Later, we study the $\mathbb{R}$-complex Finsler spaces with p-power (α, β)-metric. The fundamental metric tensor fields g_{ij} and $g_{i\bar{j}}$ and their determinants and inverse tensor fields are obtained. Some examples of non-Hermitian $\mathbb{R}$-complex Finsler spaces with the p-power (α, β)-metric are given.

1. Introduction

Randers metric was introduced by G. Randers [37] and named after him for the first time by R. S. Ingarden. The idea of the Randers Manifold was generated from the theory of electron optics. In this, the Randers metric was applied to describe the motion of relativistic electrons, and in many physical areas, its application [11] was studied. Randers metrics were seen in a more general class of metrics, which emerged in the study of metrics, which we call as (α, β) metrics. The more recent application of Randers metrics was studied in papers relating to [27] Zermelo's Navigation Problem.

In [41], G. Vranceanu introduces the concept of a nonholonomic space, which is more general than a Riemannian space and generalises the parallelism of Levi-Civita and geodesic curves in that space. From another standpoint, Z. Horak [22] considers a nonholonomic region as a space with a nonholonomic dynamical system. In [19], T. Hosokawa discusses the nonholonomic system in a space of line elements with an affine connection. In

2020 Mathematics Subject Classification: 53B40, 53C60
Key words and phrases: Finsler space with (α, β)-metric, Generalized Lagrange metric, Nonholonomic Finsler frame, p-power Finsler metric, $\mathbb{R}$-Complex Finsler spaces

© *Indian Mathematical Society, 2026.*

[25], K. Yoshie introduces the theory of nonholonomic systems in Finsler space.

In ([20], [21]), P. R. Holland studies a unified formalism that uses a nonholonomic frame on a space-time arising from consideration of a charged particle moving in an external electromagnetic field. R. S. Ingarden [24] was the first to point out that the Lorentz force law could be written as geodesic equations on a Finsler space, called Randers space. In ([13], [14]), R. G. Beil studies gauge transformation as a nonholonomic frame on the tangent bundle of a four-dimensional base manifold. In these papers, the common Finsler idea used by these physicists is the existence of a nonholonomic frame on the vertical subbundle VTM of the tangent bundle of a base manifold M. In ([8], [9]), P. L. Antonelli finds such a nonholonomic frame for two important classes that are dual in the sense of Randers and Kropina spaces. Though the study of nonholonomic frames for Finsler spaces with (α, β)-metric is quite old and so many results have been obtained by authors from many countries, it is still an important topic of research in Finsler geometry.

In this paper, we are studying a special class of Finsler spaces with p-power (α, β)-metric given by $F = \frac{(\alpha+\beta)^p}{\alpha^{p-1}}$, $(p \neq 0)$, where α is a Riemannian metric and β is a 1-form. This metric can be considered as a generalisation of the Randers metric. The main purpose of the current paper is to first determine a nonholonomic frame for a Finsler space with a p-power Finsler metric and study the more generalised $\mathbb{R}$-Complex Finsler space with a p-power (α, β)-metric. The paper is organised as follows:

Starting with a literature survey in section one, we recall some basic definitions and results in section two. In section three, initially we determine the two Finsler deformations of the p-power (α, β)-metric $F = \frac{(\alpha+\beta)^p}{\alpha^{p-1}}$ ($p \neq 0$) and obtain a corresponding frame for each of these Finsler deformations. Then, a nonholonomic frame for Finsler space with the aforesaid metric is determined as the product of these Finsler frames (see Theorem 3.3). In section four, we study on the notion of $\mathbb{R}$-complex Finsler spaces with a p-power (α, β)-metric $\frac{(\alpha+\beta)^p}{\alpha^{p-1}}$, $p \neq 0$ (see Theorem 4.8).

2. Preliminaries

An important class of Finsler spaces is the class of Finsler spaces with (α, β)-metric [31]. The first Finsler space with (α, β)-metric was introduced

in forties by the physicist G. Randers [37], and it is called Randers space. The other notable Finsler spaces with (α,β)-metric are Kropina space [28], Generalized Kropina Space and Matsumoto space [29].

Definition 2.1. A Finsler space $F^n=(M,F(x,y))$ is called the (α,β)-metric if there exists a 2-homogenous function L of two variables such that the Finsler metric F: $TM\longrightarrow\Re$ is given by:

$$F^2(x,y)=L(\alpha(x,y),\beta(x,y)), \tag{2.1}$$

where $\alpha^2(x,y)=a_{ij}(x)y^iy^j$, α is Riemannian metric on M, and $\beta(x,y)=b_i(x)y^i$ is a 1-form on M.

Example 1.1: If $L(\alpha,\beta)=(\alpha+\beta)^2$, then the Finsler space with metric $F(x,y)=\sqrt{a_{ij}(x)y^iy^j}+b_i(x)y^i$ is called a Randers space.

Example 1.2: If $L(\alpha,\beta)=\frac{\alpha^4}{\beta^2}$, then the Finsler space with metric $F(x,y)=\frac{a_{ij}(x)y^iy^j}{b_i(x)y^i}$ is called a Kropina space.

For a Finsler space with (α,β)-metric $F^2(x,y)=L(\alpha,\beta)$, the following Finsler invariants are well known [30]:

$$\rho=\frac{1}{2\alpha}\frac{\partial L}{\partial\alpha},\quad \rho_0=\frac{1}{2}\frac{\partial^2L}{\partial\beta^2},\quad \rho_{-1}=\frac{1}{2}\frac{\partial^2L}{\partial\alpha\partial\beta},\quad \rho_{-2}=\frac{1}{2\alpha^2}\Big(\frac{\partial^2L}{\partial\alpha^2}-\frac{1}{\alpha}\frac{\partial L}{\partial\alpha}\Big). \tag{2.2}$$

For a Finsler space with (α,β)-metric, we have

$$\rho_{-1}\beta+\rho_{-2}\alpha^2=0. \tag{2.3}$$

With these Finsler invariants, the metric tensor g_{ij} of a Finsler space with (α,β)-metric is given by ([7], [31])

$$g_{ij}(x,y)=\rho a_{ij}(x)+\rho_0b_i(x)b_j(x)+\rho_{-1}(b_i(x)y_j+b_j(x)y_i)+\rho_{-2}y_iy_j. \tag{2.4}$$

The metric tensor g_{ij} of a Lagrange space with (α,β)-metric can be arranged in the form

$$\begin{aligned} g_{ij}(x,y) &= \rho a_{ij}(x)+\frac{1}{\rho_{-2}}(\rho_{-1}b_i+\rho_{-2}y_i)(\rho_{-1}b_j+\rho_{-2}y_j)\\ &\quad+\frac{1}{\rho_{-2}}(\rho_0\rho_{-2}-\rho_{-1}^2)b_ib_j. \end{aligned} \tag{2.5}$$

From (2.5), we can see that g_{ij} is the result of two Finsler deformations

$$a_{ij}\mapsto h_{ij} = \rho a_{ij}+\frac{1}{\rho_{-2}}(\rho_{-1}b_i+\rho_{-2}y_i)(\rho_{-1}b_j+\rho_{-2}y_j), \tag{2.6}$$

$$h_{ij} \mapsto g_{ij} \quad = \quad h_{ij} + \frac{1}{\rho_{-2}}(\rho_0\rho_{-2} - \rho_{-1}^2)b_ib_j. \tag{2.7}$$

The nonholonomic Finsler frame that corresponds to the first deformation (2.6) is, according to Theorem (7.9.1) in [24], given by

$$\begin{aligned} X_j^i \quad = \quad & \sqrt{\rho}\delta_j^i - \frac{1}{B^2}\Big(\sqrt{\rho} \pm \sqrt{\rho + \frac{B^2}{\rho_{-2}}}\Big)(\rho_{-1}b^i + \rho_{-2}y^i) \\ & (\rho_{-1}b_j + \rho_{-2}y_j), \end{aligned} \tag{2.8}$$

where $B^2 = a_{ij}(\rho_{-1}b^i + \rho_{-2}y^i)(\rho_{-1}b^j + \rho_{-2}y^j) = \rho_{-1}^2b^2 + \beta\rho_{-1}\rho_{-2}$. The metric tensors a_{ij} and h_{ij} are related by the formula

$$h_{ij} = X_i^kX_j^la_{kl}. \tag{2.9}$$

According to Theorem (7.9.1) in [24], the nonholonomic Finsler frame that corresponds to the second deformation (2.7) is given by

$$Y_j^i = \delta_j^i - \frac{1}{C^2}\Big(1 \pm \sqrt{1 + \frac{\rho_{-2}C^2}{\rho_0\rho_{-2} - \rho_{-1}^2}}\Big)b^ib_j, \tag{2.10}$$

where $C^2 = h_{ij}b^ib^j = \rho b^2 + \frac{1}{\rho_{-2}}(\rho_{-1}b^2 + \rho_{-2}\beta)^2$. The metric tensors h_{ij} and g_{ij} are related by the formula

$$g_{mn} = Y_m^iY_n^jh_{ij}. \tag{2.11}$$

From (2.9) and (2.11), we have that $V_m^k = X_i^kY_m^i$ with X_i^k given by (2.8).

Theorem 2.2. *([15]) Let $F^2(x,y) = L(\alpha(x,y), \beta(x,y))$ be a metric function of a Finsler space with (α,β)-metric for which the condition (2.3) is true. Then*

$$V_j^i = X_k^iY_j^k \tag{2.12}$$

is a nonholonomic Finsler frame where X_k^i and Y_j^k are given by (2.8) and (2.10) respectively.

3. Nonholonomic Frame for Finsler space with p-power (α,β)-metric

We consider a Finsler space with p-power (α,β)-metric given by

$$F = \frac{(\alpha+\beta)^p}{\alpha^{p-1}}, (p \neq 0)$$

Proposition 3.1. *A p-power (α,β)-metric $F = \frac{(\alpha+\beta)^p}{\alpha^{p-1}}$, $(p \neq 0)$ is positive definite if and only if one of the following cases holds:[18]*

(1) $(p > 2$ *or* $p < 0)$ b^2 *satisfies* $b^2 < \frac{1}{(p-1)^2}$.
(2) $(\frac{1}{2} \leq p \leq 2)$ b^2 *satisfies* $b^2 < 1$.
(3) $(0 < p < \frac{1}{2})$ b^2 *satisfies* $b^2 < \frac{4-5p}{4(1-p)^2(1+p)}$.

Proposition 3.2. *An n-dimensional p-power metric* $F = \frac{(\alpha+\beta)^p}{\alpha^{p-1}}$, $(p \neq 1, 2, -1, \frac{1}{2})$ *is Einstein reversible [18] if and only if* F *is Ricci flat parallel.*

For the fundamental function $L = \frac{(\alpha+\beta)^{2p}}{\alpha^{2(p-1)}}$, the Finsler invariants (2.2) are given by

$$\begin{aligned}
\rho &= p\alpha^{-2p+1}(\alpha+\beta)^{2p-1} - (p-1)\alpha^{-2p}(\alpha+\beta)^{2p}\alpha^{-2p},\\
\rho_0 &= \frac{p(2p-1)(\alpha+\beta)^{2p-2}}{\alpha^{2(p-1)}},\\
\rho_{-1} &= p(2p-1)\alpha^{-2p+1}(\alpha+\beta)^{2p-2} - 2p(p-1)\alpha^{-2p}(\alpha+\beta)^{2p-1},\\
\rho_{-2} &= (-4p^2+3p)(\alpha+\beta)^{2p-1}\alpha^{-2p-1} + 2(p-1)^2(\alpha+\beta)^{2p}\alpha^{-2p-2},\\
B^2 &= (\alpha+\beta)^{4p}\alpha^{-4p}\Big[\big\{(2p-1)p^2\alpha(\alpha+\beta)^{-2} - 2(p-1)(\alpha+\beta)^{-1}\big\}^2\\
&\quad + \beta\big[(-4p^2+3p)(\alpha+\beta)^{-1}\alpha^{-1} + 2(p-1)^2\alpha^{-2}\big]\big[(2p-1)\alpha(\alpha+\beta)^{-2}\\
&\quad - 2(p-1)(\alpha+\beta)^{-1}\big]\Big].
\end{aligned} \tag{3.1}$$

(3.2)

Using (3.1) in (2.8), we have

$$\begin{aligned}
X_k^i &= \sqrt{p\alpha^{-2p+1}(\alpha+\beta)^{2p-1} - (p-1)\alpha^{-2p}(\alpha+\beta)^{2p}}\delta_j^i - \frac{1}{B^2}\\
&\quad \Big[\sqrt{p\alpha^{-2p+1}(\alpha+\beta)^{2p-1} - (p-1)\alpha^{-2p}(\alpha+\beta)^{2p}\alpha^{-2p}} \pm\\
&\quad \Big\{p\alpha^{-2p+1}(\alpha+\beta)^{2p-1} - (p-1)\alpha^{-2p}(\alpha+\beta)^{2p}\alpha^{-2p}\\
&\quad + \frac{B^2}{(-4p^2+3p)(\alpha+\beta)^{2p-1}\alpha^{-2p-1} + 2(p-1)^2(\alpha+\beta)^{2p}\alpha^{-2p-2}}\Big\}^{\frac{1}{2}}\Big]\\
&\quad \Big[\{p(2p-1)\alpha^{-2p+1}(\alpha+\beta)^{2p-2} - 2p(p-1)\alpha^{-2p}(\alpha+\beta)^{2p-1}\}b^i +\\
&\quad \{(-4p^2+3p)(\alpha+\beta)^{2p-1}\alpha^{-2p-1} + 2(p-1)^2(\alpha+\beta)^{2p}\alpha^{-2p-2}\}y^i\Big]\\
&\quad \Big[\{p(2p-1)\alpha^{-2p+1}(\alpha+\beta)^{2p-2} - 2p(p-1)\alpha^{-2p}(\alpha+\beta)^{2p-1}\}b_j + \{\\
&\quad (-4p^2+3p)(\alpha+\beta)^{2p-1}\alpha^{-2p-1} + 2(p-1)^2(\alpha+\beta)^{2p}\alpha^{-2p-2}\}y_j\Big].
\end{aligned} \tag{3.3}$$

Again, using (3.1) in (2.10), we have

$$
\begin{aligned}
Y_j^i &= \delta_j^i - \frac{1}{C^2}\Big[1 \pm \Big\{1 + \big\{\alpha^{2(p-1)}[(-4p^2+3p)(\alpha+\beta)^{2p-1}\alpha^{-2p-1} + \\
&\quad 2(p-1)^2(\alpha+\beta)^{2p}\alpha^{-2p-2}]C^2\big\}/\big\{[p(2p-1)(\alpha+\beta)^{2p-2}] \\
&\quad [(-4p^2+3p)(\alpha+\beta)^{2p-1}\alpha^{-2p-1} + 2(p-1)^2(\alpha+\beta)^{2p}\alpha^{-2p-2}] - \\
&\quad \big(p(2p-1)\alpha^{-2p+1}(\alpha+\beta)^{2p-2} - 2p(p-1)\alpha^{-2p}(\alpha+\beta)^{2p-1}\big)^2\big\}\Big\}^{\frac{1}{2}}\Big] b^i b_j, \qquad (3.4)
\end{aligned}
$$

where

$$
\begin{aligned}
C^2 &= [p\alpha^{-2p+1}(\alpha+\beta)^{2p-1} - (p-1)\alpha^{-2p}(\alpha+\beta)^{2p}]b^2 + \\
&\quad \frac{1}{[(-4p^2+3p)(\alpha+\beta)^{2p-1}\alpha^{-2p-1} + 2(p-1)^2(\alpha+\beta)^{2p}\alpha^{-2p-2}]} \\
&\quad \Big[\Big(p(2p-1)\alpha^{-2p+1}(\alpha+\beta)^{2p-2} - 2p(p-1)\alpha^{-2p}(\alpha+\beta)^{2p-1}\Big)b^2 + \\
&\quad \Big((-4p^2+3p)(\alpha+\beta)^{2p-1}\alpha^{-2p-1} + 2(p-1)^2(\alpha+\beta)^{2p}\alpha^{-2p-2}\Big)\beta\Big]^2 \qquad (3.5)
\end{aligned}
$$

Hence we have the following:

Theorem 3.3. *Consider a Finsler space $F^n = (M, F)$ with $L = \frac{(\alpha+\beta)^{2p}}{\alpha^{2(p-1)}}$ $(p \neq 0)$, for which the condition (2.3) is true, then*

$$V_j^i = X_k^i Y_j^k, \qquad (3.6)$$

is a nonholonomic Finsler frame, where X_k^i and Y_j^k are given by (3.3) and (3.4) respectively.

4. $\mathbb{R}$- complex Finsler space with (α, β)- metric

G. Munteanu and R. Purcaru [36] have introduced the concept of R complex Finsler space ([1], [3], [35], [38], [39]) extending complex Finsler space by reducing the scalars $\lambda \in \mathbb{C}$ to $\mathbb{R}$. Further, it was studied by the authors in [11], [31], [32], [34], [38].

In the present section, we introduce a similar notion on $\mathbb{R}$-complex Finsler spaces with a p-power (α, β)-metric $\frac{(\alpha+\beta)^p}{\alpha^{p-1}}, p \neq 0$. In this section, we keep the general setting from ([35], [36]) and subsequently we recall only some needed notions.

Let M be a complex manifold with $dim_{\mathbb{C}}M = n$, let T_zM be the tangent space at z, let (z^k) be local complex coordinates in a chart (U, ϕ), and let $T'M$ be its holomorphic tangent bundle. It has a natural structure of a complex manifold, $dim_{\mathbb{C}}T'M = dim(MXT_zM') = dimM + dimT_zM' = n + n = 2n$ and the induced coordinates in a local chart on u $\in T'M$ are

denoted by u $= (z^k, \eta^k)$. The changes of local coordinates in u are given by the rules:

$$z^{ik} = z^{ik}(z); \quad \eta'^k = \frac{\partial z'^k}{\partial z^j}\eta^j. \tag{4.1}$$

The natural frame $\left\{\frac{\partial}{\partial z^k}, \frac{\partial}{\partial \eta^k}\right\}$ of $T'_u(T'M)$ transforms with the Jacobi matrix of (4.1) changes, $\frac{\partial}{\partial z^k} = \frac{\partial z'^k}{\partial z^k}\frac{\partial}{\partial z'^j} + \frac{\partial^2 z'^j}{\partial z^k \partial z^h}\eta^h \frac{\partial}{\partial \eta'^j}$; $\frac{\partial}{\partial \eta^k} = \frac{\partial z'^j}{\partial z^k}\frac{\partial}{\partial \eta'^j}$.

A complex nonlinear connection, briefly (c. n. c.), is a supplementary distribution $H(T'M)$ to the vertical distribution spanned by $\left\{\frac{\partial}{\partial \eta^k}\right\}$ and an adapted frame in $H(T'M)$ is $\frac{\delta}{\delta z^k} = \frac{\partial}{\partial z^k} - N^j_k \frac{\partial}{\partial \eta^j}$, where N^j_k are the coefficients of c. n. c. Next we have some notations $\partial_k = \frac{\partial}{\partial z^k}, \delta_k = \frac{\delta}{\delta z^k}, \dot\partial_k = \frac{\partial}{\partial \eta^k}$ and $\partial_{\bar k}$, $\dot\partial_{\bar k}$ $\delta_{\bar k}$ for their conjugates. The dual adapted basis of $\{\delta_k, \dot\partial_k\}$ are $\{dz^k, \delta\eta^k = d\eta^k + N^k_j dz^j\}$ and $\{d\bar z^k, d\delta\eta^k\}$ theirs conjugates.

The homogeneity of the metric function of a complex Finsler space ([1], [3], [35], [38]) is with respect to complex scalars and the metric tensor of the space, which is a Hermitian form.

A positive real valued continuous function F on $T'M$ is an $\mathbb{R}$-complex Finsler metric on M, if
(i) $L = F^2$ is smooth on $T'M$ (except the 0 sections);
(ii) $F(z,\eta) \geq 0$, $F(z,\eta) = 0$ if and only if $\eta = 0$;
(iii) $F(z, \lambda\eta, \bar z, \lambda\bar\eta) = |\lambda| F(z,\eta,\bar z,\bar\eta)$;
It follows that L is (2, 0) homogeneous with respect to the real scalars λ, and in [36] the following identities are proved:

$$\frac{\partial L}{\partial \eta^i}\eta^i + \frac{\partial L}{\partial \bar\eta^i}\bar\eta^i = 2L; \quad g_{ij}\eta^i + g_{\bar j i}\bar\eta^i = \frac{\partial L}{\partial \eta^j}; \tag{4.2}$$

$$\frac{\partial g_{ik}}{\partial \eta^j}\eta^j + \frac{\partial g_{ik}}{\partial \bar\eta^j}\bar\eta^j = 0; \frac{\partial g_{i\bar k}}{\partial \eta^j}\eta^j + \frac{\partial g_{i\bar k}}{\partial \bar\eta^j}\bar\eta^j = 0; \tag{4.3}$$

$$2L = g_{ij}\eta^i\eta^j + g_{\bar i\bar j}\bar\eta^i\bar\eta^j + 2g_{i\bar j}\eta^i\bar\eta^j; \tag{4.4}$$

where

$$g_{ij} = \frac{\partial^2 L}{\partial \eta^i \partial \eta^j}; \quad g_{i\bar j} = \frac{\partial^2 L}{\partial \eta^i \partial \bar\eta^j}; \quad g_{\bar i\bar j} = \frac{\partial^2 L}{\partial \bar\eta^i \partial \bar\eta^j} \tag{4.5}$$

are the metric tensors of the space.

4.1. **$\mathbb{R}$- complex Finsler space with a p-power (α,β)- metric $\frac{(\alpha+\beta)^p}{\alpha^{p-1}}, p \neq 0$.** An $\mathbb{R}$-complex Finsler space [36] produces two fundamental tensor fields g_{ij} and $g_{i\bar j}$. The invertibility of $g_{i\bar j}$ is a compulsory requirement, for a proper Hermitian geometry.

These problems led to the study of Hermitian $\mathbb{R}$-complex Finsler spaces (i. e., $\det(g_{i\bar{j}}) \neq 0$). In the present section, we have results on $\mathbb{R}$-complex Finsler spaces with a p-power (α, β)- metric $\frac{(\alpha+\beta)^p}{\alpha^{p-1}}, p \neq 0$.

Definition 4.1 ($\mathbb{R}$-complex Finsler space)**.** An $\mathbb{R}$-complex Finsler space with (α, β)-metric is a pair (M, F), where the fundamental function $F(z, \eta, \bar{z}, \bar{\eta})$ is $\mathbb{R}$-homogenous by means of functions $\alpha(z, \eta, \bar{z}, \bar{\eta})$ and $\beta(z, \eta, \bar{z}, \bar{\eta})$ which are dependent on $z^i, \eta^i, \bar{z}^i$ and $\bar{\eta}^i$, (i = 1, 2,..n), i. e.,

$$F(z, \eta, \bar{z}, \bar{\eta}) = F(\alpha(z, \eta, \bar{z}, \bar{\eta}), \beta(z, \eta, \bar{z}, \bar{\eta})) \tag{4.6}$$

where

$$\alpha^2(z, \eta, \bar{z}, \bar{\eta}) = \frac{1}{2}(a_{ij}\eta^i\eta^j + a_{\bar{i}\bar{j}}\bar{\eta}^i\bar{\eta}^j + 2a_{i\bar{j}}\eta^i\bar{\eta}^j) = Re\{a_{ij}\eta^i\eta^j + a_{i\bar{j}}\eta^i\bar{\eta}^j\},$$
$$\beta(z, \eta, \bar{z}, \bar{\eta}) = \frac{1}{2}(b_i\eta^i + b_{\bar{i}}\bar{\eta}^i) = Re\{b_i\eta^i\},$$

with:

$$a_{ij} = a_{ij}(z), a_{i\bar{j}} = a_{i\bar{j}}(z), b_i = b_i(z), \tag{4.7}$$

where $b_i(z)dz^i$ is a 1-form on the complex manifold M.

We denote

$$L(\alpha(z, \eta, \bar{z}, \bar{\eta}), \beta(z, \eta, \bar{z}, \bar{\eta})) = F^2(\alpha(z, \eta, \bar{z}, \bar{\eta}), \beta(z, \eta, \bar{z}, \bar{\eta})). \tag{4.8}$$

Remark 4.2. $L = F^2$ is $\mathbb{R}$-complex Finsler space with (α, β)-metric.

Definition 4.3. An $\mathbb{R}$-complex Finsler space with a p-power (α, β)-metric $\frac{(\alpha+\beta)^p}{\alpha^{p-1}}$, $p \neq 0$ is given by

$$L(\alpha, \beta) = \frac{(\alpha + \beta)^{2p}}{\alpha^{2p-2}}. \tag{4.9}$$

Taking into account the 2-homogeneity condition of L

$$L(\alpha(z, \lambda\eta, \bar{z}, \lambda\bar{\eta}), \beta(z, \lambda\eta, \bar{z}, \lambda\bar{\eta})) = \lambda^2 L(\alpha(z, \eta, \bar{z}, \bar{\eta}), \beta(z, \eta, \bar{z}, \bar{\eta})), \lambda \in \mathbb{R}_+. \tag{4.10}$$

We know the following proposition:

Proposition 4.4. *([6]) In an $\mathbb{R}$-complex Finsler space with (α, β)-metric the following equalities hold:*

$$\alpha L_\alpha + \beta L_\beta = 2L, \quad \alpha L_{\alpha\alpha} + \beta L_{\alpha\beta} = L_\alpha, \tag{4.11}$$
$$\alpha L_{\alpha\beta} + \beta L_{\beta\beta} = L_\beta, \alpha^2 L_{\alpha\alpha} + 2\alpha\beta L_{\alpha\beta} + \beta^2 L_{\beta\beta} = 2L, \tag{4.12}$$

where

$$L_\alpha = \frac{\partial L}{\partial \alpha},\ L_\beta = \frac{\partial L}{\partial \beta},\ L_{\alpha\beta} = \frac{\partial^2 L}{\partial \alpha^2},\ L_{\beta\beta} = \frac{\partial^2 L}{\partial \beta^2}. \tag{4.13}$$

Particular case: For an $\mathbb{R}$-complex Finsler space with special (α,β)-metric $L(\alpha,\beta) = \frac{(\alpha+\beta)^{2p}}{\alpha^{2p-2}}$, we have

$$L_\alpha = 2\frac{(\alpha+\beta)^{2p-1}}{\alpha^{2p-1}}[\alpha + (1+p)\beta], \tag{4.14}$$

$$L_\beta = \frac{2p(\alpha+\beta)^{2p-1}}{\alpha^{2p-2}}, \tag{4.15}$$

$$\alpha L_\alpha + \beta L_\beta = 2L \tag{4.16}$$

$$L_{\alpha\alpha} = \frac{2(\alpha+\beta)^{2p-1}}{\alpha^{2p-1}}\Big[\{1 + (2p-1)(\alpha+(1+p)\beta)(\alpha+\beta)^{-1}\} + \alpha^{-1}(\alpha+(1+p)\beta)\Big], \tag{4.17}$$

$$L_{\alpha\beta} = \frac{2p(\alpha+\beta)^{2p-2}}{\alpha^{2p-1}}\Big[(3-2p)\alpha - (2p-2)\beta\Big], \tag{4.18}$$

$$L_{\beta\beta} = \frac{2p(2p-1)(\alpha+\beta)^{2p-2}}{\alpha^{2p-2}}. \tag{4.19}$$

We consider,

$$\frac{\partial \alpha}{\partial \eta^i} = \frac{1}{2\alpha}(a_{ij}\eta^j + a_{i\bar{j}}\bar{\eta}^j) = \frac{1}{2\alpha}l_i, \frac{\partial \beta}{\partial \eta^i} = \frac{1}{2}b_i, \tag{4.20}$$

$$\frac{\partial \alpha}{\partial \bar{\eta}^i} = \frac{1}{2\alpha}(a_{i\bar{j}}\bar{\eta}^j + a_{i\bar{j}}\eta^j) = \frac{1}{2\alpha}l_i, \frac{\partial \beta}{\partial \bar{\eta}^i} = \frac{1}{2}b_{\bar{i}}, \tag{4.21}$$

where

$$l_i = a_{ij}\eta^j + a_{i\bar{j}}\bar{\eta}^j,\quad l_{\bar{j}} = a_{\bar{i}\bar{j}}\bar{\eta}^i + a_{i\bar{j}}\bar{\eta}^i. \tag{4.22}$$

We find immediately

$$l_i\eta^i + l_{\bar{j}}\eta^j = 2\alpha^2. \tag{4.23}$$

We denote $\eta_i = \frac{\partial l}{\partial \eta^i}$. Consequently, we obtain

$$\eta_i = \rho_0 l_i + \rho_1 b_i, \tag{4.24}$$

where

$$\rho_0 = \frac{1}{2}\frac{L_\alpha}{\alpha}(0 - homogeneity),\quad \rho_1 = \frac{1}{2}L_\alpha\ \ (1 - homogeneity), \tag{4.25}$$

Differentiating (4.25) by η^j and $\bar{\eta}^j$ respectively, we obtain

$$\frac{\partial \rho_0}{\partial \eta^j} = \rho_{-2}l_j + \rho_{-1}b_j,\quad \frac{\partial \rho_0}{\partial \bar{\eta}^j} = \rho_{-2}l_{\bar{j}} + \rho_{-1}b_{\bar{j}}, \tag{4.26}$$

$$\frac{\partial \rho_1}{\partial \eta^i} = \rho_{-1}l_j + \mu_0 b_i,\quad \frac{\partial \rho_1}{\partial \bar{\eta}^i} = \rho_{-2}l_{\bar{i}} + \mu_0 b_{\bar{i}}, \tag{4.27}$$

where

$$\rho_{-2} = \frac{\alpha L_{\alpha\alpha} - L_\alpha}{4\alpha^3}, \quad \rho_{-1} = \frac{L_{\alpha\beta}}{4\alpha}, \quad \mu_0 = \frac{L_{\beta\beta}}{4}. \tag{4.28}$$

Proposition 4.5. *The invariants of $\mathbb{R}$-complex Finsler space with special (α,β)- metric $\frac{(\alpha+\beta)^p}{\alpha^{p-1}}$ are given by*

$$\begin{aligned}
\rho_0 &= \frac{(\alpha+\beta)^{2p-1}}{\alpha^{2p}}[\alpha + (1+p)\beta], \\
\rho_1 &= \frac{(\alpha+\beta)^{2p-1}}{\alpha^{2p-1}}[\alpha + (1+p)\beta], \\
\rho_{-2} &= \frac{(\alpha+\beta)^{2p-1}}{2\alpha^{2p+1}}[1 + (2p-1)(\alpha + (1+p)\beta)(\alpha+\beta)^{-1}], \\
\rho_{-1} &= \frac{p(\alpha+\beta)^{2p-2}}{2\alpha^{2p}}[(3-2p)\alpha - (2p-2)\beta], \\
\mu_0 &= \frac{p(2p-1)(\alpha+\beta)^{2p-2}}{2\alpha^{2p-2}}.
\end{aligned} \tag{4.29}$$

We have immediately

Proposition 4.6. *For an $\mathbb{R}$-complex Finsler space with p-power (α,β)-metric $\frac{(\alpha+\beta)^p}{\alpha^{p-1}}$, the fundamental tensor fields are given by*

$$\begin{aligned}
g_{ij} &= \frac{(\alpha+\beta)^{2p-1}}{\alpha^{2p}}[\alpha + (1+p)\beta]a_{ij} + \frac{(\alpha+\beta)^{2p-1}}{2\alpha^{2p+1}}[1 + (2p-1) \\
&\quad (\alpha + (1+p)\beta)(\alpha+\beta)^{-1}]l_i l_j + \frac{p(2p-1)(\alpha+\beta)^{2p-2}}{2\alpha^{2p-2}} b_i b_j \\
&\quad + \frac{p(\alpha+\beta)^{2p-2}}{2\alpha^{2p+3}}[(3-2p)\alpha - (2p-2)\beta](b_j l_i + b_i l_j), \\
g_{i\bar{j}} &= \frac{(\alpha^4 + \epsilon\alpha^3\beta - \lambda^2\beta^4 - \epsilon\delta\alpha\beta^3)}{\alpha^4} a_{i\bar{j}} + \frac{\beta}{2\alpha^6}(4\delta^2\beta^3 - \epsilon\alpha^3 - \epsilon\delta\alpha\beta^2) l_i l_{\bar{j}} \\
&\quad \frac{1}{2\alpha^2}\Big[(2\lambda + \epsilon^2)\alpha^2 + 6\epsilon\delta\alpha\beta + 6\lambda^2\beta^2\Big] b_i b_{\bar{j}} \\
&\quad + \frac{1}{2\alpha^4}(\epsilon\alpha^3 - 3\epsilon\delta\alpha\beta^2 - 4\delta^2\beta^3)\,(b_{\bar{j}} l_i + b_i l_{\bar{j}}),
\end{aligned}$$

or, in the equivalent form

$$g_{ij} = \rho_0[a_{ij} - p_1 l_i l_j + p_2 b_i b_j + p_3 \eta_i \eta_j], \tag{4.30}$$

$$g_{i\bar{j}} = \rho_0[a_{i\bar{j}} - p_1 l_i l_{\bar{j}} + p_2 b_i b_{\bar{j}} + p_3 \eta_i \eta_{\bar{j}}]. \tag{4.31}$$

where

$$\begin{aligned} p_1 &= \frac{-(\alpha+\beta)^{-1}}{2\alpha^{2p+1}}\Big[\big(1+(2p-1)(\alpha+(1+p)\beta)(\alpha+\beta)^{-1}\big)- \\ &\quad (\alpha+\beta)^{-1}\big((3-2p)\alpha-(2p-2)\beta\big)\Big], \\ p_2 &= \frac{p(\alpha+\beta)^{2p-2}}{2\alpha^{2p-1}}\Big[(2p-1)\alpha-(\{3-2p\}\alpha-\{2p-2\}\beta)\Big], \\ p_3 &= \frac{p\alpha^{2p-1}(\{3-2p\}\alpha-\{2p-2\}\beta)}{(\alpha+\beta)^{2p}(\alpha+(1+p)\beta)^2}. \end{aligned}$$

Proof. From the equation (4.29) in Theorem 2.1 [6] by direct calculus, we have the results. □

Now, we find the determinant and the inverse of the tensor field g_{ij}. We obtain the solution from Proposition 2.2 [5],for an arbitrary non singular non- Hermitian matrix (Q_{ij}). The result is

Proposition 4.7. *Suppose*

(1) (Q_{ij}) is a non-singular $n\times n$ complex matrix with inverse (Q^{ji});
(2) C_i and $C_{\bar{i}}=\bar{C}_i$, i = 1, 2, .., n are complex numbers;
(3) $C^i=Q^{ji}C_j$ and its conjugates $C^2=C^iC_i=\bar{C}^i C_{\bar{i}}$; $H_{ij}=Q_{ij}\pm C_iC_j$
Then $det(H_{ij})=(1\pm C^2)det(Q_{ij})$.
And whenever $1\pm C^2\neq 0$, the matrix (H_{ij}) is invertible and in this case its inverse is $H^{ji}=Q^{ji}\mp\frac{1}{1\pm C^2}C^iC^j$.

Theorem 4.8. *For a non-Hermitian $\mathbb{R}$-complex Finsler space with p-power (α,β)-metric, we have*

(1) The contravariant tensor g^{ij} of the fundamental tensor field g_{ij} is

$$\begin{aligned} g^{ij} &= \frac{1}{\rho_0}\Big[a^{ji}+\Big(\frac{p_1}{1-\gamma p_1}-\frac{\epsilon^2p_1^2p_2}{\tau(1-\gamma p_1)^2}\Big)\eta^i\eta^j-\frac{p_2b^ib^j}{\tau} \\ &\quad -\frac{\epsilon p_1p_2(b^i\eta^j+b^j\eta^i)}{\tau(1-\gamma p_1)} \\ &\quad -\frac{P^2\eta^i\eta^j+PQ(\eta^ib^j+\eta^jb^i)+Q^2b^ib^j}{1+(P\gamma+Q\epsilon)\sqrt{p_3}}\Big] \end{aligned} \tag{4.32}$$

where

$$l_i=a_{ij}\eta^j,\ \gamma=a_{ij}\eta^j\eta^k=l_i\eta^k,\ \epsilon=b_j\eta^j,\ \omega=b_jb^j,$$
$$b^k=a^{jk}b_j,\ b_l=b^ka_{kl},\delta=a_{jk}\eta^jb^k=l_kb^k,l^j=a^{ji}l_i=\eta^j.$$

(2) $det(g_{ij}) = (\rho_0)^n\Big[1+(P\gamma+Q\epsilon)\sqrt{p_3}\Big]\Big[1+\omega+\frac{p_1\epsilon^2}{1-p_1\gamma}\Big](1-p_1\gamma)det(a_{ij})$, *where*

$$\begin{aligned}
\rho_0 &= \frac{(\alpha+\beta)^{2p-1}}{\alpha^{2p}}[\alpha+(1+p)\beta],\\
P &= 1+\Big[\frac{p_1}{1-\gamma p_1}-\frac{\epsilon^2p_1^2p_2}{\tau(1-\gamma p_1)^2}\Big]\gamma-\Big[\frac{p_1p_2}{\tau(1-\gamma p_1)}\Big]\epsilon^2,\\
Q &= -\frac{p_2}{\tau}\epsilon+\Big[\frac{p_1p_2\epsilon}{\tau(1-\gamma p_1)}\Big]\gamma,\\
p_1 &= \frac{-(\alpha+\beta)^{-1}}{2\alpha^{2p+1}}\Big[\big(1+(2p-1)(\alpha+(1+p)\beta)(\alpha+\beta)^{-1}\big)-\\
&\quad (\alpha+\beta)^{-1}\big((3-2p)\alpha-(2p-2)\beta\big)\Big],\\
p_2 &= \frac{p(\alpha+\beta)^{2p-2}}{2\alpha^{2p-1}}\Big[(2p-1)\alpha-(\{3-2p\}\alpha-\{2p-2\}\beta)\Big],\\
p_3 &= \frac{p\alpha^{2p-1}(\{3-2p\}\alpha-\{2p-2\}\beta)}{(\alpha+\beta)^{2p}(\alpha+(1+p)\beta)^2}.
\end{aligned}$$

Proof. To prove, we use the previous proposition in a recursive algorithm in three steps. From (4.30) we have

$$g_{ij} = \rho_0[a_{ij}-p_1l_il_j+p_2b_ib_j+p_3\eta_i\eta_j]. \tag{4.33}$$

1. In the first step, we set $Q_{ij}=a_{ij}$ and $c_i=\sqrt{p_1}l_i$. From proposition 4.7, we get $Q^{ji}=a^{ji}$, $c^2=\gamma p_1$. Then, the matrix $H_{ij}=a_{ij}-p_1l_il_j$ is invertible with have

$$H^{ji} = a^{ji}+\frac{p_1\eta^i\eta^j}{1-p_1\gamma} \tag{4.34}$$

and $det(a_{ij}-p_1l_il_j)=(1-p_1\gamma)det(a_{ij})$.

2. Next, consider $Q_{ij}=a_{ij}-p_1l_il_j$ and $c_i=\sqrt{p_2}b_i$. From proposition 4.7, we have : $Q^{ji}=a^{ji}+\frac{p_1\eta^i\eta^j}{1-p_1\gamma}$, $c^2=p_2\Big(\omega+\frac{p_1\epsilon^2}{1-p_1\gamma}\Big)$, $1+c^2=1+\omega+\frac{p_1\epsilon^2}{1-p_1\gamma}\neq 0$ and $c^i=\sqrt{p_2}\Big(b^i+\frac{\epsilon p_1\eta^i}{1-p_1\gamma}\Big)$.

Then, the matrix $H_{ij}=a_{ij}-p_1l_il_j+p_2b_ib_j$ is invertible with

$$H^{ji} = a^{ji}+\Big[\frac{p_1}{1-\gamma p_1}-\frac{\epsilon^2p_1^2p_2}{\tau(1-\gamma p_1)^2}\Big]\eta^i\eta^j-\frac{p_2b^ib^j}{\tau}-\frac{\epsilon p_1p_2(b^i\eta^j+b^j\eta^i)}{\tau(1-\gamma p_1)} \tag{4.35}$$

where $\tau=1+\omega+\frac{p_1\epsilon^2}{1-p_1\gamma}$ and $det(a_{ij}-p_1l_il_j+p_2b_ib_j)=\Big[1+\omega+\frac{p_1\epsilon^2}{1-p_1\gamma}\Big](1-p_1\gamma)det(a_{ij})$.

3. In this step, we have $Q_{ij} = a_{ij} - p_1 l_i l_j + p_2 b_i b_j$ and $c_i = \sqrt{p_3}\eta_i$. By applying the proposition 4.7, we obtain this time

$$Q^{ji} = a^{ji} + \Big[\frac{p_1}{1-\gamma p_1} - \frac{\epsilon^2 p_1^2 p_2}{\tau(1-\gamma p_1)^2}\Big]\eta^i\eta^j - \frac{p_2 b^i b^j}{\tau} - \frac{\epsilon p_1 p_2(b^i\eta^j + b^j\eta^i)}{\tau(1-\gamma p_1)}, \tag{4.36}$$

$$\begin{aligned} c^i &= \Big[\eta^i + \Big(\frac{p_1}{1-\gamma p_1} - \frac{\epsilon^2 p_1^2 p_2}{\tau(1-\gamma p_1)^2}\Big)\eta^i\gamma - \frac{p_2 b^i\epsilon}{\tau} - \frac{\epsilon p_1 p_2(b^i\gamma + \epsilon\eta^i)}{\tau(1-\gamma p_1)}\Big]\sqrt{p_3} \\ &= P\eta^i + Qb^i, \end{aligned} \tag{4.37}$$

where

$$P = 1 + \Big[\frac{p_1}{1-\gamma p_1} - \frac{\epsilon^2 p_1^2 p_2}{\tau(1-\gamma p_1)^2}\Big]\gamma - \Big[\frac{p_1 p_2}{\tau(1-\gamma p_1)}\Big]\epsilon^2, \tag{4.38}$$

$$Q = -\frac{p_2}{\tau}\epsilon + \Big[\frac{p_1 p_2\epsilon}{\tau(1-\gamma p_1)}\Big]\gamma \tag{4.39}$$

and

$$c^2 = (P\gamma + Q\epsilon)\sqrt{p_3}, 1 + c^2 = 1 + (P\gamma + Q\epsilon)\sqrt{p_3} \neq 0, \tag{4.40}$$

where

$c^i c^j = P^2\eta^i\eta^j + PQ(\eta^i b^j + \eta^j b^i) + Q^2 b^i b^j$.

Then, the matrix $H_{ij} = a_{ij} - p_1 l_i l_j + p_2 b_i b_j + p_3\eta_i\eta_j$ is invertible with

$$\begin{aligned} H^{ji} &= a^{ji} + \Big[\frac{p_1}{1-\gamma p_1} - \frac{\epsilon^2 p_1^2 p_2}{\tau(1-\gamma p_1)^2}\Big]\eta^i\eta^j - \frac{p_2 b^i b^j}{\tau} - \frac{\epsilon p_1 p_2(b^i\eta^j + b^j\eta^i)}{\tau(1-\gamma p_1)} \\ &\quad - \frac{P^2\eta^i\eta^j + PQ(\eta^i b^j + \eta^j b^i) + Q^2 b^i b^j}{1 + (P\gamma + Q\epsilon)\sqrt{p_3}} \end{aligned} \tag{4.41}$$

and $det(a_{ij} - p_1 l_i l_j + p_2 b_i b_j + p_3\eta_i\eta_j) = \Big[1 + (P\gamma + Q\epsilon)\sqrt{p_3}\Big]\Big[1 + \omega + \frac{p_1\epsilon^2}{1-p_1\gamma}\Big](1 - p_1\gamma)det(a_{ij})$. But $g_{ij} = \rho_0 H_{ij}$, with H_{ij} from step 3. Thus, $g^{ij} = \frac{1}{\rho_0}H^{ji}$ and $det(g_{ij}) = (\rho_0)^n\Big[1 + (P\gamma + Q\epsilon)\sqrt{p_3}\Big]\Big[1 + \omega + \frac{p_1\epsilon^2}{1-p_1\gamma}\Big](1 - p_1\gamma)det(a_{ij})$. Hence the statement holds. □

Proposition 4.9. *In a non-Hermitian $\mathbb{R}$-complex Finsler space with p-power (α,β)- metric $\frac{(\alpha+\beta)^p}{\alpha^{p-1}}, p \neq 0$, we have the following :*

$$\gamma + \bar{\gamma} = l_i\eta^i + l_{\bar{j}}\eta^{\bar{j}} = a_{ij}\eta^j\eta^i + a_{\bar{j}\bar{k}}\eta^{\bar{k}}\eta^{\bar{j}} = 2\alpha^2, \tag{4.42}$$

$$\theta + \bar{\theta} = b_j\eta^j + b_{\bar{j}}\eta^{\bar{j}} = 2\beta, \delta = \theta, \tag{4.43}$$

where:

$$l_i = a_{ij}\eta^j,\ \gamma = a_{ij}\eta^j\eta^k = l_i\eta^k,\ \theta = b_j\eta^j,\ \omega = b_jb^j,$$
$$b^k = a^{jk}b_j,\ b_l = b^ka_{kl}, \delta = a_{jk}\eta^jb^k = l_kb^k, l^j = a^{ji}l_i = \eta^j.$$

Example 4.10. We set α as

$$\alpha^2 = \frac{(1+\theta|z|^)\sum_{k=1}^{n} Re(\eta^j)^2 - \theta Re < z,\eta >^2}{(1+\theta|z|^2)^2}, \tag{4.44}$$

where $|z|^2 = \sum_{k=1}^{n} z^k\bar{z}^k$, $< z,\eta >= \sum_{k=1}^{n} z^k\bar{\eta}^k$, defined over the disk $\Delta_r^n = \{z \in \mathbb{C}^n, |z| < r, r = \sqrt{\frac{1}{|\theta|}}\}$ if $\theta = 0$ and on the complex projective space $P^n(C)$ if $\theta > 0$. By computation, we obtain $a_{ij} = \frac{1}{1+\theta|z|^2}\Big(\delta_{ij} - \frac{\theta\bar{z}^i\bar{z}^j}{1+\theta|z|^2}\Big)$ and $a_{i\bar{j}} = 0$ and so, $\alpha^2(z,\eta) = \frac{1}{2}(a_{ij}\eta^j\eta^i + a_{\bar{j}\bar{k}}\eta^{\bar{i}}\eta^{\bar{j}})$. Now, taking $\beta(z,\eta) = Re\frac{<z,\eta>}{1+\theta|z|^2}$, where $b_i = \frac{\bar{z}^1}{1+\theta|z|^2}$, we obtain some examples of non- Hermitian $\mathbb{R}$- complex p-power (α,β)- metric $\frac{(\alpha+\beta)^p}{\alpha^{p-1}}, p \neq 0$

$$F_\theta = \frac{\Big[\sqrt{\frac{(1+\theta|z|^)\sum_{k=1}^{n} Re(\eta^j)^2-\theta Re<z,\eta>^2}{(1+\theta|z|^2)^2}} + Re\frac{<z,\eta>}{1+\theta|z|^2}\Big]^p}{\Big[\frac{(1+\theta|z|^)\sum_{k=1}^{n} Re(\eta^j)^2-\theta Re<z,\eta>^2}{(1+\theta|z|^2)^2}\Big]^{\frac{p-1}{2}}}. \tag{4.45}$$

Similarly, one can obtain the inverse for $g_{i\bar{j}}$.

5. Applications

Nonholonomic frames in Finsler spaces have broad applications in fields that involve systems with velocity-dependent constraints, complex geometries, and optimisation. These areas include robotics, vehicle dynamics, nonholonomic mechanics, theoretical physics, control theory, mathematical biology, and computer science ([11], [33]). Nonholonomic Finsler frames provide a robust geometric tool to address practical problems in constrained systems, facilitating the design of more efficient algorithms and providing deeper insights into the behaviour of systems with complex constraints.

The study of $\mathbb{R}$-Complex Finsler spaces provides a powerful generalisation of traditional Riemannian geometry, allowing for a richer understanding of complex structures, especially when combined with Finsler metrics

[5] and further developed by Abate and Patrizio [1]. Its applications are broad, extending from theoretical physics and quantum mechanics to control theory, robotics, machine learning, complex analysis, and algebraic geometry [2]. These spaces provide a natural setting to model systems with complex-valued data, constraints, and geometries, leading to new insights and solutions in fields as diverse as quantum information, nonlinear control systems, and the study of moduli spaces in algebraic geometry [1].

Acknowledgement: We thank the referee for the comments, which improved the quality of the paper.

References

[1] Abate M., Patrizio G., *Finsler metrics- A Global Approach,* Lecture Notes in Math.,**1591**, Springer-Verlag, (1994).

[2] Agrachev A., Barilari D.,Boscain U., *A Comprehensive Introduction to Sub-Riemannian Geometry,* Cambridge University Press, 2020.

[3] Aikou T., *Projective flatness of complex Finsler metrics,* Publ. Math. Dereccan, **63/3**, (2003), 343-362.

[4] Aldea N., *Complex Finsler space of constant holomorphic curvature,* Diff. Geom. and its applications, Proc. Conf. Prague 2004, Charles Univ., Prague (Czech Republic), (2005), 179-190.

[5] Aldea N., Munteanu G., *On complex Finsler spaces with Randers metric,* J. Korean Math. Soc. **46** (5), (2009), 949-966.

[6] Aldea N., Purcaru M., $\mathbb{R}$-*Complex Finsler space with* (α, β)-*metric,* Novi Saj J. Math. Math., **38** (1), (2008), 1-9.

[7] Anastasiei, M. and Shimada, H., *The Beil Metrics Associated to a Finsler Space,* BJGA, **3**, (2) (1998), 1-16.

[8] Antonelli, P. L. and Bucataru, I., *On Holland's frame for Randers space and its applications in Physics,* Steps in differential geometry- Proceedings of the colloquium on differential geometry, Debrecen, Hungary, (2000) 25-30.

[9] Antonelli, P. L. and Bucataru, I., *Finsler connections in anholonomic geometry of Kropina space,* Nonlinear Studies, **8** (1), (2001), 171–184.

[10] Antonelli, P. L., Ingarden, R. S. and Matsumato. M., *The Theory of Sprays and Finsler Spaces with Applications in Physics and Biology,* Kluwer Acad. publ., FTPH, no, 1993.

[11] Bao D., Chern S. S., Shen Z., *An introduction to Riemann- Finsler geometry,* Graduate Texts in Math., **200**, Springer- Verlag, 2000.

[12] Bao D., Robles C., *On Randers spaces of constant flag curvature,* Rep. Math. Phys., **51** (2003), 39-42.

[13] Beil, R. G., *Finsler Gauge Transformations and General Relativity,* Intern. J. Theor. Phys., **31** (6), (1992), 1025-1044.

[14] Beil, R. G., *Finsler and Kaluza-Klein Gauge Theories,* Intern. J. Theor. Phys., **32** (6), (1993), 1021-1031.

[15] Bucataru, I., *Nonholonomic Frames in Finsler Geometry,* Balkan Journal of Geometry and Its Applications, 7 (1), (2002), 13-27.

[16] Bucataru, I., *Nonholonomic frames for Finsler spaces with (α, β)-metrics,* Proceedings of the conference on Finsler and Lagrange Geometries, Iasi, August 26-31, 2001, Kluwer Academic Publishers, (2003), 69-78.

[17] Bucataru, I., Miron, R., *Finsler-Lagrange Geometry applications to dynamical systems,* CEEX ET 3174/2005-2007 & CEEX M III 12595/2007 (2007).

[18] Guojun, Y., *On a class of Einstein reversible Finsler metrics,* Differential Geometry Appl. **60**, (2018), 80-103.

[19] Hosokawa, T., *Ueber nicht-holonome Uebertragung in allgemeiner Mannigfaltigkeit T_n,* Jour. Fac. Sci. Hokkaido Imp. Univ., **I**, (2) (1934), 1-11.

[20] Holland, P. R., *Electromagnetism, particles and anholonomy,* Phys. Lett., **91 A**, (1982), 275-278.

[21] Holland, P. R. and Philippidis, *Anholonomic deformations in the ether: a significance for the electrodynamics potentials,* in the volume *Quantum Implications,* B.J. Hiley and F. D. Peat (eds.), Routledge and Kegan Paul, London and New York(1987), 295-311.

[22] Horak, Z., *Sur une generalisation de la notion de variete,* Publ. Fac. Sc. Univ. Masaryk, Brno, 86 (1927), 1-20.

[23] Hrimiuc, D. and Shimada, H., *On the L-duality between Lagrange and Hamilton manifolds,* Nonlinear World, **3**, (1996), 613-641.

[24] Ingarden, R. S., *Differential Geometry and physics,* Tensor N. S., **30**, (1976), 201-209.

[25] Katsurada, Y., *On the Theory of Non-holonomic system in Finsler Spaces,* Tohoku Math. J., (2) **3** (2), (1951), 140-148.

[26] Kobayashi S., Horst C., Wu K. H., *Complex Differential geometry,* Birkhauser Verlag, (1983).

[27] Kopacz P., *On the generaization of Zermelo Navigation problem on Riemannian Manifolds,* International Journal of Geometric methods in modern physics, (16) **4**, (2019).

[28] Kropina V. K., *On projective two-dimensional Finsler spaces with a special metric,* Trudy Sem. Vektor. Tenzor. Anal., **11**, (1961), 277-292.

[29] Matsumoto, M., *A slope of a mountain is a Finsler surface with respect to a time measure,* J. Math. Kyoto Univ. **29**(1), (1989), 17-25.

[30] Matsumoto, M., *Foundations of Finsler Geometry and Special Finsler Spaces,* Kaisheisha Press, Otsu, Japan (1986).

[31] Matsumoto, M., *Theory of Finsler spaces with (α, β)-metrics,* Rep. Math. Phys. **31**, (1991), 43-83.

[32] Miron R., *General Randers spaces,* Kluwer Acad. Publ., FTPH **76**, (1996), 126-140.

[33] R. Miron, Anastasiei M., *The Geometry of Lagrange Spaces: Theory and Applications,* Kluwer Acad. Publ., FTPH (1993).

[34] Miron R., Hassan B. T., *Variational problem in Finsler spaces with* (α,β)*- metric,* Algebras Groups and Geometries, **20**, (2003), 285-300.

[35] Munteanu G., *Complex Spaces in Finsler, Lagrange and Hamilton Geometries,* Kluwer Acad. Publ., FTPH **141**, (2004).

[36] Munteanu G., Purcaru M., *On* $\mathbb{R}$*- complex Finsler spaces,* Balkan J. Geom. Appl., **14** (1), (2009), 52-59.

[37] Randers, G., *On an asymmetric metric in the four-spaces of general relativity,* Phys. Rev., **59**, (1941), 195-199.

[38] Sabau V. S., Shimada H., *Remarkable classes of* (α,β)*-metric spaces,* Rep. on Math. Phys., **47** (1), (2001), 31-48.

[39] Shanker G., Baby S. A., *On* $\mathbb{R}$*-complex Finsler spaces with special* (α,β)*-metric,* International Journal of Applied mathematics, **29** (5), (2016), 609-619.

[40] Shanker, G., Baby, S. A., *On the Nonholonomic Frames for Finsler spaces with general* $(\alpha,\ \beta)$*-metrics,* **107** (4), (2016), 1013-1023.

[41] Vranceanu, G., *Sur les espaces non holonomes, Sur le calcul differential absolu pour les varietes non holonomes,* Comptes rendus hebdomadaires des séances de l'Académie des sciences, **183**, (1926), 852-1083.

Sruthy Asha Baby
Post Graduate Department of Mathematics
Devaswom Board College, Thalayolaparambu,
Kottayam, Kerala, INDIA.
E-mail: sruthymuthu123@gmail.com, sruthyasha@dbct.ac.in

The Mathematics Student ISSN: 0025-5742
Vol. 95, Nos. 3-4, July - December (2026), 212–241

LOCAL CONVERGENCE ANALYSIS OF INITIAL VALUE PROBLEMS OF NONLINEAR RIEMANN-LIOUVILLE FRACTIONAL HYBRID DIFFERENTIAL EQUATIONS

JANHAVI B. DHAGE AND BAPURAO C. DHAGE
(Received : 27 - 01 - 2026 ; Accepted : 28 - 06 - 2026)

Abstract. In this paper we prove local approximation, existence and uniqueness results for the initial value problems of nonlinear Riemann-Liouville fractional ordinary hybrid differential equations by using the Dhage monotone iteration method based on the recent hybrid fixed point principles. Next, an approximation result for Ulam-Hyers stability and monotonic dependence of the local solution on initial data is also discussed. A couple of examples are indicated to illustrate the hypotheses and abstract results of this paper. Finally, some concluding remarks are given at the end of this paper.

1. Introduction

The current trend of several mathematicians all over the world working in the field of nonlinear analysis is changed from ordinary to fractional differential equations because of their appearance in several dynamic phenomena of natural and physical sciences. See for example, Diethelm [20], Podlubny [28], Miller and Ross [27], Hadmard [2], Hilfer [23] and references therein. Therefore, several fractional derivatives such as Riemann-Liouville, Caputo, Hadmard and Hilfer etc., have been employed in the analysis of such problems. Of the above fractional Riemann-Liouville and Caputo derivatives are frequently used for the analysis of fractional order linear as well a nonlinear differential equations for the analysis of different qualitative properties of the solution. Convergence analysis of Picard's successive approximations for existence, uniqueness of solution for such problems is available in the

2020 Mathematics Subject Classification: 34A34, 34A08, 34A12, 47H10
Key words and phrases: Fractional differential equation, hybrid fixed point principle, Dhage iteration method, Local approximation theorem, Ulam-Hyers stability.

© *Indian Mathematical Society, 2026.*

literature, but the monotonic local convergence analysis is new to the theory fractional calculus which is the main motivation of this paper. In the present work we discuss some approximate results for local existence, uniqueness and stability of the solution of the nonlinear fractional Riemann-Liouville differential equations via monotonic convergence of successive approximation based on Dhage iteration method.

Given a closed and bounded interval $J = [t_0, t_0 + a]$ of the real line $\mathbb{R}$, for some $t_0, a \in \mathbb{R}$ with $t_0 \geq 0$ and $a > 0$, we denote $J' = (t_0, t_0 + a]$. Now we consider the initial value problem (IVP) of nonlinear Riemann-Liouville hybrid fractional differential equation (HFRDE)

$$^{RL}D^{q}_{t_0}x(t) = f(t, x(t)), \;\; t \in J', \tag{1.1}$$

together with the fractional integral initial condition (IC)

$$\lim_{t \to t_{0+}} I^{1-q}_{t_0+}x(t) = b_0 \in \mathbb{R}, \tag{1.2}$$

where $^{RL}D^{q}$ is a Riemann-Liouville fractional derivative of order q with $0 < q < 1$ and the function $f : J' \times \mathbb{R} \to \mathbb{R}$ is continuous and satisfies certain hybrid, that is, mixed hypotheses from algebra, analysis and topology to be specified later. Here, the initial condition (1.2) does not indicate the instantaneous observable physical state of the system but represents the past history or memory of the solution of anomalous dynamic systems modeled on the HFDE (1.1).

Definition 1.1. A function $\xi_{b_0} \in C_{1-q}(J, \mathbb{R}) = \big\{x : J \to \mathbb{R} : x \in C(J', \mathbb{R})$ and $(t - t_0)^{1-q}x \in C(J, \mathbb{R})\big\}$ is said to be a classical *solution* of the HFRDE (1.1 if it satisfies the equations in (1.1)-(1.2) on J, where $C(J, \mathbb{R})$ is the space of continuous real-valued functions defined on J. If a solution ξ_{b_0} lies in a neighborhood $\mathcal{N}(x_0)$ of some point $x_0 \in C_{1-q}(J, \mathbb{R})$, then we say it is a *local solution* or *neighborhood solution* (in short *nbhd solution*) of the HFDE (1.1)-(1.2) defined on J.

Remark 1.2. It is well-known that a closed ball $B[x, r]$ in $C_{1-q}(J, \mathbb{R})$ centered at a point $x \in C_{1-q}(J, \mathbb{R})$ of radius $r > 0$ is a neighborhood of the point x, so if a solution ξ_{b_0} of the HFRDE (1.1)-(1.2) lies in a closed ball $B[x, r] = \overline{B(x, r)}$ in $C_{1-q}(J, \mathbb{R})$, then it is a *local solution* or *nbhd solution*) of the given HFRDE (1.1)-(1.2). Note that the present notion of the local solution is different that given in Coddington [4].

It is known that if the nonlinearity or nonlinearities involved in a fractional differential equation (FDE) are continuous or discontinuous, then it is called a continuous or discontinuous FDE. Similarly, if the nonlinear function or functions involved in a FDE satisfy certain mixed conditions form different branches of mathematics such as algebra, analysis and topology etc., then it is called a HFDE (see Dhage *et al.* [12]). For example if $f(t,x) = x^2$, then the FDE (1.1) is the continuous DE and not a HFDE because the map $x \mapsto x^2$ is only continuous on $\mathbb{R}$. But if $f(t,x) = \tan^{-1} x$, then the FDE (1.1) becomes a HFDE, since the function $x \mapsto \tan^{-1} x$ is monotonic, Lipschitz and continuous and bounded on $\mathbb{R}$ which are the characteristics of algebra, analysis and topology. The HFDEs are generally handled with the help of hybrid fixed point theory to get more information about the solution.

The nonlinear HFRDE (1.1)-(1.2) is is quite well-known and discussed for different aspects of the solution by using various techniques. The existence result for the local solution of the IVP (1.1)-(1.2) can be proved by using the Schauder fixed point principle, see for example, Lakshmikantham *et al.* [25], Granas and Dugundji [21], Agarwal *el.* [1], Dhage *et al.* [18] and references therein. Similarly, the approximation result for the uniqueness of solution can be proved by using well-known Banach fixed point theorem under a Lipschitz condition which is considered to be very strong in the topic of nonlinear analysis. But to the knowledge of the present authors, the approximation result for local existence and uniqueness theorems without using the Lipschitz condition is not discussed so far in the theory of nonlinear fractional differential and integral equations. In this paper, we undertake the study of approximation results for local existence and uniqueness of solution under weaker form of one sided or partial Lipschitz condition but via construction of the algorithms based on the Dhage monotone iteration method and a hybrid fixed point theorem given in [8] on the lines of Dhage and Dhage [14, 15, 16] which is generalization of the hybrid fixed point theorems proved in [7].

The rest of the paper is organized as follows. Section 2 deals with the auxiliary results and main hybrid fixed point theorems involved in the Dhage iteration method. The hypotheses and main approximation results for the

local existence and uniqueness of solution are given in Section 3. The approximation of the Ulam-Hyers stability is discussed in Section 4 and monotonic dependence of local solution on initial data is discussed in section 5. Finally a couple of illustrative examples are presented in Section 6 and some conclusive remarks are given in Section 7.

2. Auxiliary Results

We need the following two basic definitions of Riemann-Liouville integration and Riemann-Liouville derivative from fractional calculus in what follows.

Definition 2.1. The Riemann-Liouville fractional integration of a function $x : J \to \mathbb{R}$ of fractional order q is defined as

$$I_{t_0}^q x(t) = \frac{1}{\Gamma(q)} \int_{t_0}^t (t-s)^{q-1} x(s)\, ds, \quad t \in J, \tag{2.1}$$

provided the integral on the right hand side exists on the interval $(t_0, t_0+a]$, where Γ is the Euler's gamma function defined by $\Gamma(q) = \int_0^\infty e^{-t} t^{q-1}\, dt$.

Remark 2.2. The integral on the right hand side of (2.1) exists in particular if $(t-s)^{q-1}x \in L^1((t_0, t_0+a], \mathbb{R})$ or $(t-s)^{q-1}x \in C((t_0, t_0+a], \mathbb{R})$ for all $t, s \in (t_0, t_0+a]$, $t \geq s$, where $L^1((t_0, t_0+a], \mathbb{R})$ and $C((t_0, t_0+a], \mathbb{R})$ denote receptively the spaces of Lebesgue integrable and continuous real valued functions on $(t_0, t_0+a]$.

Definition 2.3. If $x \in L^1(J, \mathbb{R})$, then the Riemann-Liouville fractional derivative ${}^{RL}D_{t_0}^q x$ of x of order q is defined as

$${}^{RL}D_{t_0}^q x(t) = \frac{1}{\Gamma(n-q)} \left(\frac{d}{dt}\right)^n \int_{t_0}^t (t-s)^{n-q-1} x(s)\, ds, n = [q]+1, \tag{2.2}$$

provided the right hand side exists, where $[q]$ denotes the integer part of the real number q.

Then we have the following fundamental results which are needed in what follows.

Lemma 2.4 (Podlubny [28]). *For any $q \in (0,1)$ and $\beta > 0$, we have*

(i) $I_{t_0}^q (t-t_0)^{\beta-1} = \dfrac{\Gamma(\beta)}{\Gamma(\beta+q)} (t-t_0)^{\beta+q-1}$,

(ii) ${}^{RL}D_{t_0}^q (t-t_0)^{\beta-1} = \dfrac{\Gamma(\beta)}{\Gamma(\beta-q)} (t-t_0)^{\beta-q-1}$, *and*

(iii) ${}^{RL}D^{q}_{t_0}(t-t_0)^{q-1}=0,$

for all $t\in J$.

Lemma 2.5 (Podlubny [28]). *If* $x\in C(J,\mathbb{R})$ *and* $q\in(0,1)$, *then*

(a) ${}^{RL}D^{q}_{t_0}I^{q}_{t_0}x(t)=x(t)$, *and*

(b) $I^{q}_{t_0}\left({}^{RL}D^{q}_{t_0}x(t)\right)=x(t)-\dfrac{I^{1-q}_{t_0}x(t)}{\Gamma q}\Big|_{t=t_0}(t-t_0)^{q-1},$

for all $t\in J$, *where* $I^{1-q}_{t_0}x\in C(J,\mathbb{R})$.

We place the problem of HFRDE (1.1)-(1.2) in the function space $C_{1-q}(J,\mathbb{R})$ of continuous, real-valued functions defined on J. We introduce a supremum norms $\|\cdot\|$ and $\|\cdot\|_{1-q}$ in the spaces $C(J,\mathbb{R})$ and $C_{1-q}(J,\mathbb{R})$ defined respectively by

$$\|x\|=\sup_{t\in J}|x(t)|, \tag{2.3}$$

and

$$\|x\|_{1-q}=\sup_{t\in J}\big|(t-t_0)^{1-q}x(t)\big|. \tag{2.4}$$

Define cones K and K_{1-q} in the Banach spaces $C(J,\mathbb{R})$ and $C_{1-q}(J,\mathbb{R})$ by

$$K=\{x\in C(J,\mathbb{R})\mid x(t)\geq 0 \text{ for all } t\in J\} \tag{2.5}$$

and

$$\begin{aligned}K_{1-q}=\big\{x\in C_{1-q}(J,\mathbb{R})\mid x(t)\geq 0 \text{ for all } t\in J'\\ \text{and } (t-t_0)^{1-q}x(t)\geq 0 \text{ for all } t\in J\big\}\end{aligned} \tag{2.6}$$

respectively.

Next, we introduce a partial order relation $\preceq$ in the Banach space $C_{1-q}(J,\overline{\mathbb{R}})$ by the order cone K_{1-q}. Thus,

$$x\preceq y \iff y-x\in K_{1-q},$$

or equivalently,

$$\begin{aligned}x\preceq y \iff & x(t)\leq y(t) \text{ for all } t\in J'\\ & \text{and } (t-t_0)^{1-q}x(t)\leq (t-t_0)^{1-q}y(t) \text{ for all } t\in J.\end{aligned} \tag{2.7}$$

The details of cones an their properties along with some nice applications may be found in Granas and Dugundji [21] and Guo and Lakshmikantham [22]. It is known that the Banach space $C_{1-q}(J,\mathbb{R})$ together with the above partial order relation $\preceq$ becomes an ordered Banach space which we denote

for convenience, by $\big(C_{1-q}(J,\mathbb{R}), K_{1-q}\big)$. We denote the open and closed spheres centered at $x_0 \in C_{1-q}(J,\mathbb{R})$ of radius r, for some $r > 0$, by

$$B_r(x_0) = \{x \in C_{1-q}(J,\mathbb{R}) \mid \|x - x_0\|_{1-q} < r\} = B(x_0, r)$$

and

$$B_r[x_0] = \{x \in C_{1-q}(J,\mathbb{R}) \mid \|x - x_0\|_{1-q} \leq r\} = \overline{B(x_0, r)}$$

respectively. It is clear that $B_r(x_0) \subset B_r[x_0] \subset B_{r+\epsilon}(x_0)$ for every $\epsilon > 0$.

It is well-known that the fixed point theoretic technique is very much useful in the subject of nonlinear analysis for dealing with the nonlinear equations. See Granas and Dugundji [21] and the references therein. Here, we employ the Dhage monotone iteration method or simply *Dhage iteration method* based on two hybrid fixed point theorems proved in [7]. Before stating them, we give some preliminaries:

Let $\big(E, d, \preceq\big)$ be a partially ordered metric space and let $S \subset E$. E is called **regular** if a monotone nondecreasing (resp. monotone nonincreasing) sequence $\{x_n\}$ in E converges to x_*, then $x_n \preceq x_*$ (resp. $x_* \preceq x_n$) for all $n \in \mathbb{N}$. The metric d and the order relation $\preceq$ are said to be **$\mathcal{D}$-compatible** in S if a monotone sequence $\{x_n\}$ in S has a convergent subsequence, then the original sequence $\{x_n\}$ is convergent and converges to the same limit point. S is called a **Janhavi set** if d and $\preceq$ are $\mathcal{D}$-compatible in it. S is called **partial bounded** (resp. partially closed, partially compact) if every chain C in S is bounded (resp. closed, compact). A few details of the above notions along with their utility in hybrid fixed point theorems appear in Dhage [5]. The following results are frequently used in the theory of hybrid fixed point theory and applications to nonlinear differential and integral equations.

Lemma 2.6 (Dhage [8,9]). *Every ordered Banach space $\big(E, \|\cdot\|, K\big)$ is regular.*

Lemma 2.7 (Dhage [8,9]). *Every non-empty partially compact subset S of an ordered Banach space $\big(E, \|\cdot\|, K\big)$ is a Janhavi set.*

A mapping $\mathcal{T} : S \to S$ is called **monotone nondecreasing** (resp. monotone nonincreasing) if $x \preceq y$ implies $\mathcal{T}x \preceq \mathcal{T}y$ (resp. $x \preceq y$ implies $\mathcal{T}x \succeq \mathcal{T}y$). $\mathcal{T}$ is **monotone** or **monotonic** if it is either monotone nondecreasing or monotone nonincreasing on S. $\mathcal{T}$ is called **partial bounded**

(resp. partially totally bounded or partially precompact) if $\mathcal{T}(S)$ is partially bounded (resp. partially totally bounded or partially precompact for partially bounded S). T is **partially continuous** if and only if $\{x_n\} \subset S$ converges to x_* with either $x_n \preceq x_*$ or $x_* \preceq x_n$ for each $n \in \mathbb{N}$, then $\mathcal{T}x_n \to \mathcal{T}x$. T is called **partial completely continuous** if it is partially continuous and partially totally bounded. The details of the above concepts along with applications to nonlinear equations appear in Dhage [5, 6, 7, 8], Dhage and Dhage [9], Dhage *et al.* [19] and references therein. The following lemma is obvious and frequently used in the applications of Dhage iteration method.

Lemma 2.8 (Dhage and Dhage [14, 15])**.** *Every mapping $\mathcal{T}$ on a subset S of an ordered metric space E into itself is partially continuous if and only if it is continuous on every chain C of S.*

Now we are equipped with all the necessary details to state our required hybrid fixed point theorems needed in what follows.

Theorem 2.9. *Let S be a non-empty, partial closed and partial bounded subset of a regular partially ordered Banach space $\big(E, \|\cdot\|, \preceq\big)$. Suppose that $\mathcal{T} : S \to S$ is a partial completely continuous and monotone nondecreasing operator and every chain C in $\mathcal{T}(S)$ is Janhavi set. If there exists an element $x_0 \in S$ such that $x_0 \preceq \mathcal{T}x_0$ or $x_0 \succeq \mathcal{T}x_0$, then $\mathcal{T}$ has a fixed point x^* in S and the sequence $\{\mathcal{T}^n x_0\}_{n=0}^{\infty}$ of successive iterations converges monotonically to x^*.*

Proof. The proof is similar to a hybrid fixed point theorem proved in Dhage [8] with obvious modifications, however the details appear in Dhage *et al.* [12, 13]. □

Remark 2.10. Note that the statement "every chain C in $\mathcal{T}(S)$ is Janhavi set" holds if every chain C in the subset S of the partially ordered Banach space $\big(E, \|\cdot\|, \preceq\big)$ is a Janavi set and the later statement holds in particular if S is a partially compact subset of $\big(E, \|\cdot\|, K\big)$.

Definition 2.11 ([6], [29])**.** A mapping $\mathcal{T} : \big(E, d, \preceq\big) \to \big(E, d, \preceq\big)$ is called a partial contraction if

$$d(\mathcal{T}x, \mathcal{T}y) \leq k\, d(x, y) \tag{2.8}$$

for all comparable elements $x, y \in E$, where k is a real number satisfying $0 \leq k < 1$. The number k is called the contraction constant of $\mathcal{T}$.

Theorem 2.12 (Dhage [10]). *Let $\left(E, \|\cdot\|, \preceq, \right)$ denote a regular partially ordered Banach space and let $\mathcal{T} : E \to E$ be a monotone nondecreasing and partial contraction operator with contraction constant k. If there exists an element $x_0 \in X$ such that $x_0 \preceq \mathcal{T}x_0$ or $x_0 \succeq \mathcal{T}x_0$ satisfying*

$$\|x_0 - \mathcal{T}x_0\| \leq (1-k)r, \tag{2.9}$$

for some real number $r > 0$, then $\mathcal{T}$ has a unique comparable fixed point x^* in $B_r[x_0]$ and the sequence $\{x_n\}_{n=0}^{\infty}$ of successive iterations converges geometrically and monotonically to x^*. Furthermore, if every pair of elements in X has a lower or upper bound, then x^* is unique.

Remark 2.13. Note that every every pair of elements in a partially ordered set (in short poset) $(E, \preceq)$ has a lower or upper bound if $(E, \preceq)$ is a lattice, that is, $\preceq$ is a lattice order in E. In this case the poset $(E, \|\cdot\|, \preceq)$ is called a **partially lattice ordered Banach space**. There do exist several lattice partially ordered Banach spaces which are useful for applications in nonlinear analysis. For example, every Banach lattice is a partially lattice ordered Banach space. Clearly, an ordered Banach space $\big(C(J, \mathbb{R}), K\big)$ is a lattice ordered Banach space. The details of the lattice structure of a Banach space appear in the monograph Birkhoff [3].

As a consequence of Remark 2.13, we obtain

Theorem 2.14. *Let $\left(E, \|\cdot\|, \preceq, \right)$ denote a regular partially lattice ordered Banach space and let $\mathcal{T} : E \to E$ be a monotone nondecreasing and partial contraction operator with contraction constant q. If there exists an element $x_0 \in X$ such that $x_0 \preceq \mathcal{T}x_0$ or $x_0 \succeq \mathcal{T}x_0$ satisfying (2.7), then $\mathcal{T}$ has a unique fixed point x^* in $B_r[x_0]$ and the sequence $\{\mathcal{T}^n x_0\}_{n=0}^{\infty}$ of successive iterations converges geometrically and monotonically to x^*.*

In the following sections we prove the main approximation results of this paper concerning the existence, uniqueness, stability and monotonic dependence of solution under suitable natural conditions.

3. Local Approximation Results

We consider the following set of hypotheses in what follows.

- (H$_1$) The function f is continuous and bounded on $J \times \mathbb{R}$ with bound M_f.
- (H$_2$) $f(t, x)$ is nondecreasing in x for each $t \in J$.

(H$_3$) $f(t, \alpha_0(t)) \geq 0$ for all $t \in J$, where $\alpha_0(t) = \frac{b_0}{\Gamma(q)}(t-t_0)^{q-1}$, $t \in J'$.

Then we have the following useful lemma.

Lemma 3.1 (Kilbas *et al.* [26]). *For any function $h \in L^1(J, \mathbb{R})$, the function $x \in C(J, \mathbb{R})$ is a solution of the FRDE*

$$ {}^{RL}D^q_{t_0} x(t) = h(t), \ t \in J', \tag{3.1}$$

satisfying the initial condition

$$\lim_{t \to t_{0+}} I^{1-q}_{t_0+} x(t) = b_0, \tag{3.2}$$

if and only if x satisfies the nonlinear fractional integral equation (FRIE)

$$x(t) = \frac{b_0}{\Gamma(q)}(t-t_0)^{q-1} + \frac{1}{\Gamma(q)} \int_{t_0}^{t} (t-s)^{q-1} h(s)\, ds \tag{3.3}$$

for all $t \in J'$.

Proof. The proof of the lemma follows by the applications of Lemmas 2.4 and 2.5 to the FRDE (3.1). We omit the details. □

Theorem 3.2. *Suppose that all the hypotheses (H$_1$) through (H$_3$) hold. If $\frac{M_f\, a}{\Gamma(q+1)} \leq r$, then the HFRDE (1.1)-(1.2) has a local solution ξ_{b_0} in $B_r[x_0]$, where $x_0(t) = \frac{b_0}{\Gamma(q)}(t-t_0)^{q-1}$, $t \in J'$, and the sequence $\{x_n\}_{n=0}^{\infty}$ of successive approximations defined by*

$$\left.\begin{aligned} x_0(t) &= \frac{b_0}{\Gamma(q)}(t-t_0)^{q-1}, \ t \in J', \\ x_{n+1}(t) &= \frac{b_0}{\Gamma(q)}(t-t_0)^{q-1} + \frac{1}{\Gamma(q)} \int_{t_0}^{t} (t-s)^{q-1} f\big(s, x_n(s)\big)\, ds, \ t \in J', \end{aligned}\right\} \tag{3.4}$$

for $n = 1, 2, \ldots$; is monotone nondecreasing and converges to ξ_{b_0}.

Proof. Set $X = C_{1-q}(J, \mathbb{R})$. Clearly, (X, K_{1-q}) is a regular partially ordered Banach space in view of Lemma 2.6. Let x_0 be a real-valued function on J given by $x_0(t) = \frac{b_0}{\Gamma(q)}(t-t_0)^{q-1}$ for all $t \in J'$ and define a closed ball $B_r[x_0]$ in X defined by (2.4). By Lemma 3.1, the HFRDE (1.1)-(1.2) is equivalent to the nonlinear hybrid fractional integral equation (HFIE)

$$x(t) = \frac{b_0}{\Gamma(q)}(t-t_0)^{q-1} + \frac{1}{\Gamma(q)} \int_{t_0}^{t} (t-s)^{q-1} f\big(s, x(s)\big)\, ds, \ t \in J'. \tag{3.5}$$

Now, define an operator $\mathcal{T}$ on $B_r[x_0]$ into X by

$$\mathcal{T}x(t) = \frac{b_0}{\Gamma(q)}(t-t_0)^{q-1} + \frac{1}{\Gamma(q)}\int_{t_0}^{t}(t-s)^{q-1}f\big(s,x(s)\big)\,ds,\ \ t\in J', \quad (3.6)$$

so that

$$(t-t_0)^{1-q}\mathcal{T}x(t) = \frac{b_0}{\Gamma(q)} + \frac{(t-t_0)^{1-q}}{\Gamma(q)}\int_{t_0}^{t}(t-s)^{q-1}f\big(s,x(s)\big)\,ds \in C(J,\mathbb{R})$$

for all $t\in J$. Therefore, the fractional integral equation (3.5) is transformed into the operator equation as

$$\mathcal{T}x = x \ \text{ on } J. \quad (3.7)$$

We shall show that the operator $\mathcal{T}$ satisfies all the conditions of Theorem 2.9 on $B_r[x_0]$ in the following series of steps.

Claim I: *The operator $\mathcal{T}$ maps $B_r[x_0]$ into itself.*

Firstly, we show that $\mathcal{T}$ maps $B_r[x_0]$ into itself. Let $x\in B_r[x_0]$ be arbitrary element. Then,

$$\begin{aligned}|\mathcal{T}x(t)-x_0(t)| &= \left|\frac{1}{\Gamma(q)}\int_{t_0}^{t}(t-s)^{q-1}f\big(s,x(s)\big)\,ds\right| \\ &\le \frac{1}{\Gamma(q)}\int_{t_0}^{t}(t-s)^{q-1}\big|f\big(s,x(s)\big)\big|\,ds \\ &\le M_f\int_{t_0}^{t}(t-s)^{q-1}\,ds = \frac{M_f\,a^q}{\Gamma(q+1)}.\end{aligned}$$

Therefore,

$$\big|(t-t_0)^{1-q}\big(\mathcal{T}x(t)-x_0(t)\big)\big| \le \frac{M_f\,a}{\Gamma(q+1)}$$

for all $t\in J$. Taking the supremum over t in the above inequality yields

$$\|\mathcal{T}x-x_0\|_{1-q} \le \frac{M_f\,a}{\Gamma(q+1)} \le r,$$

which implies that $\mathcal{T}x\in B_r[x_0]$ for all $x\in B_r[x_0]$. As a result, we have $\mathcal{T}(B_r[x_0])\subset B_r[x_0]$.

Claim II: *$\mathcal{T}$ is a monotone nondecreasing operator on $B_r[x_0]$.*

Let $x,y\in B_r[x_0]$ be any two elements such that $x\succeq y$ on J. Then, since (H_2) and (H_3) hold, we have that

$$\mathcal{T}x(t) = \frac{b_0}{\Gamma(q)}(t-t_0)^{q-1} + \frac{1}{\Gamma(q)}\int_{t_0}^{t}(t-s)^{q-1}f\big(s,x(s)\big)\,ds$$

$$\geq \frac{b_0}{\Gamma(q)}(t-t_0)^{q-1} + \frac{1}{\Gamma(q)}\int_{t_0}^{t}(t-s)^{q-1}f\big(s,y(s)\big)\,ds$$

$$= \mathcal{T}y(t)$$

for all $t \in J'$. Since $(t-t_0)^{1-q} \geq 0$ for all $t \in J$, one has $(t-t_0)^{1-q}\mathcal{T}x(t) \geq (t-t_0)^{1-q}\mathcal{T}y(t)$ for all $t \in J$. Therefore, $\mathcal{T}x \succeq \mathcal{T}y$, that is, $\mathcal{T}$ is monotone nondecreasing on $B_r[x_0]$.

Claim III: *$\mathcal{T}$ is partially continuous operator on $B_r[x_0]$.*

Let C be a chain in $B_r[x_0]$ and let $\{x_n\}$ be a monotone nondecreaing sequence in C converging to a point $x \in C$. Then, from Lemma 2.6 it follows that $x_n \preceq x$ for each $n \in \mathbb{N}$. By hypothesis (H_1), the function f is bounded on $J \times \mathbb{R}$. Therefore, by dominated convergence theorem,we obtain

$$\lim_{n\to\infty}(t-t_0)^{1-q}\mathcal{T}x_n(t)$$

$$= \lim_{n\to\infty}\left[\frac{b_0}{\Gamma(q)} + \frac{(t-t_0)^{1-q}}{\Gamma(q)}\int_{t_0}^{t}(t-s)^{q-1}f\big(s,x_n(s)\big)\,ds\right]$$

$$= \frac{b_0}{\Gamma(q)} + \lim_{n\to\infty}\frac{(t-t_0)^{1-q}}{\Gamma(q)}\int_{t_0}^{t}(t-s)^{q-1}f\big(s,x_n(s)\big)\,ds$$

$$= \frac{b_0}{\Gamma(q)} + \frac{(t-t_0)^{1-q}}{\Gamma(q)}\int_{t_0}^{t}(t-s)^{q-1}\left[\lim_{n\to\infty}f\big(s,x_n(s)\big)\right]ds$$

$$= \frac{b_0}{\Gamma(q)}(t-t_0)^{q-1} + \frac{1}{\Gamma(q)}\int_{t_0}^{t}(t-s)^{q-1}f\big(s,x(s)\big)\,ds$$

$$= (t-t_0)^{1-q}\mathcal{T}x(t)$$

for all $t \in J$. Therefore, $\mathcal{T}x_n \to \mathcal{T}x$ pointwise on J.

Next we show that $\{\mathcal{T}x_n\}$ is an equicontinuous sequence of points in X. Let $t_1, t_2 \in J$ be arbitrary. Then, for any $n \in \mathbb{N}$, by hypothesis (H_1), we obtain

$$\Big|(t_1-t_0)^{1-q}\mathcal{T}x_n(t_1) - (t_2-t_0)^{1-q}\mathcal{T}x_n(t_2)\Big|$$

$$= \left|\frac{(t_1-t_0)^{1-q}}{\Gamma(q)}\int_{t_0}^{t_1}(t_1-t_0)^{q-1}f(s,x(s))\,ds\right.$$

$$\left. -\frac{(t_2-t_0)^{q-1}}{\Gamma(q)}\int_{t_0}^{t_2}(t_2-t_0)^{q-1}f(s,x(s))\,ds\right|$$

$$\leq \frac{a}{\Gamma(q)} \left| \int_{t_0}^{t_1} (t_1 - s)^{q-1} f(s, x_n(s))\, ds - \int_{t_0}^{t_2} (t_2 - s)^{q-1} f(s, x_n(s))\, ds \right|$$

$$+ \frac{1}{\Gamma(q)} \left| (t_1 - t_0)^{1-q} - (t_2 - t_0)^{1-q} \right| \left| \int_{t_0}^{t_2} (t_2 - t_0)^{q-1} f(s, x(s))\, ds \right|.$$

Now,

$$\frac{a}{\Gamma(q)} \left| \int_{t_0}^{t_1} (t_1 - s)^{q-1} f(s, x_n(s))\, ds - \int_{t_0}^{t_2} (t_2 - s)^{q-1} f(s, x_n(s))\, ds \right|$$

$$\leq \frac{a}{\Gamma(q)} \left| \int_{t_0}^{t_1} (t_1 - s)^{q-1} f(s, x_n(s))\, ds - \int_{t_0}^{t_1} (t_2 - s)^{q-1} f(s, x_n(s))\, ds \right|$$

$$+ \frac{a}{\Gamma(q)} \left| \int_{t_0}^{t_1} (t_2 - s)^{q-1} f(s, x(s))\, ds - \int_{t_0}^{t_2} (t_2 - s)^{q-1} f(s, x_n(s))\, ds \right|$$

$$\leq \frac{a}{\Gamma(q)} \int_{t_0}^{t_1} \left| (t_1 - s)^{q-1} - (t_2 - s)^{q-1} \right| |f(s, x_n(s))|\, ds$$

$$+ \frac{a}{\Gamma(q)} \left| \int_{t_2}^{t_1} \left| (t_2 - s)^{q-1} \right| |f(s, x_n(s))|\, ds \right|$$

$$\leq \frac{M_f a}{\Gamma(q)} \int_{t_0}^{t_1} \left| (t_1 - s)^{q-1} - (t_2 - s)^{q-1} \right| ds$$

$$+ \frac{M_f a}{\Gamma(q)} \left| \int_{t_2}^{t_1} \left| (t_2 - s)^{q-1} \right| ds \right|$$

$$\leq \frac{M_f a}{\Gamma(q)} \int_{t_0}^{t_{t_0+a}} \left| (t_1 - s)^{q-1} - (t_2 - s)^{q-1} \right| ds + \frac{M_f a^q}{\Gamma(q)} |t_1 - t_2|$$

$$\to 0 \quad \text{as} \quad t_1 \to t_2,$$

uniformly for each $n \in \mathbb{N}$. Similarly, we have

$$\frac{1}{\Gamma(q)} \left| (t_1 - t_0)^{1-q} - (t_2 - t_0)^{1-q} \right| \left| \int_{t_0}^{t_2} (t_2 - t_0)^{q-1} f(s, x(s))\, ds \right|$$

$$\leq \frac{M_f}{\Gamma(q)} \left| (t_1 - t_0)^{1-q} - (t_2 - t_0)^{1-q} \right| \left| \int_{t_0}^{t_2} (t_2 - t_0)^{q-1} |f(s, x(s))|\, ds \right|$$

$$\leq \frac{M_f}{\Gamma(q)} \left| (t_1 - t_0)^{1-q} - (t_2 - t_0)^{1-q} \right| \left| \int_{t_0}^{t_2} (t_2 - t_0)^{q-1}\, ds \right|$$

$$\leq \frac{M_f a^q}{\Gamma(q)} \left| (t_1 - t_0)^{1-q} - (t_2 - t_0)^{1-q} \right|$$

$$\to 0 \quad \text{as} \quad t_1 \to t_2,$$

uniformly for each $n \in \mathbb{N}$. As a result, $\{\mathcal{T}x_n\}$ is an equicontinuous sequence of points in X. Therefore, we have that $\mathcal{T}x_n \to \mathcal{T}x$ uniformly on J. Hence $\mathcal{T}$ is a partially continuous operator on $B_r[x_0]$.

Claim IV: *$\mathcal{T}$ is partially compact operator on $B_r[x_0]$ into itself.*

To show $\mathcal{T}$ is partially comact operator, it is enough to prove that $\mathcal{T}(B_r[x_0])$ is a partially compact subset of $B_r[x_0]$. Let C be an arbitrary chain in $B_r[x_0]$. We show that $\mathcal{T}(C)$ is a relatvely compact subset of $B_r[x_0]$. We apply the Arzelá-Ascoli theorem for compactness of a set in $C(J, \mathbb{R})$. Firstly, let $y \in \mathcal{T}(C)$ be any element. Then there is an element $x \in C$ such that $y = \mathcal{T}x$. Now, by hypothesis (H_1), we get

$$\begin{aligned} \|y\|_{1-q} = \|\mathcal{T}x\|_{1-q} &\leq \left|\frac{b_0}{\Gamma(q)}\right| + \frac{M_f a^q}{\Gamma(q+1)} \sup_{t\in J}(t-t_0)^{1-q} \\ &\leq \left|\frac{b_0}{\Gamma(q)}\right| + \frac{M_f a}{\Gamma(q+1)} \end{aligned}$$

for all $y \in \mathcal{T}(C)$. This shows that $\mathcal{T}(C)$ is uniformly bounded subset of $B_r[x_0]$. Next, let t_1, t_2 be arbitrary. Then following the arguments given in Claim III, it can be proved that

$$|y(t_1) - y(t_2)| = \big|\mathcal{T}x(t_1) - \mathcal{T}x(t_2)\big| \to 0 \quad \text{as} \quad t_1 \to t_2,$$

uniformly for $y \in \mathcal{T}(C)$. Now, by an application of Arzelá-Ascoli theorem, we conclude that $\mathcal{T}(C)$ is a relatiely compact subset of $B_r[x_0]$. Consequently $\mathcal{T}$ is a partially compact operator on $B_r[x_0]$ into itelf.

Claim V: *The element $x_0 \in B_r[x_0]$ satisfies the order relation $x_0 \preceq \mathcal{T}x_0$.*

Since (H_3) holds, one has

$$\begin{aligned} x_0(t) &= \frac{b_0}{\Gamma(q)}(t-t_0)^{q-1} \\ &\leq x_0(t) + \frac{1}{\Gamma(q)} \int_{t_0}^{t} (t-s)^{q-1} f\left(s, \frac{b_0}{\Gamma(q)}(t-t_0)^{q-1}\right) ds \\ &= \frac{b_0}{\Gamma(q)}(t-t_0)^{q-1} + \frac{1}{\Gamma(q)} \int_{t_0}^{t} (t-s)^{q-1} f\big(s, x_0(s)\big)\, ds \\ &= \mathcal{T}x_0(t) \end{aligned}$$

for all $t \in J'$. Since $(t-t_0)^{1-q} \geq 0$ for all $t \in J$, one has $(t-t_0)^{1-q} x_0(t) \leq (t-t_0)^{1-q} \mathcal{T}x_0(t)$ for all $t \in J$. As a result, we have that $x_0 \preceq \mathcal{T}x_0$.

This shows that the function x_0 in $B_r[x_0]$ serves as to satisfy the operator inequality $x_0 \preceq \mathcal{T}x_0$.

Now, $\mathcal{T}$ is a partially completely continuous operator on $B_r[x_0]$ and so every chain C in $\mathcal{T}(B_r[x_0])$ is a Janahavi set in view of Lemma 2.7. Thus, the operator $\mathcal{T}$ satisfies all the conditions of Theorem 2.9, and so $\mathcal{T}$ has a fixed point ξ_{b_0} in $B_r[x_0]$ and the sequence $\{\mathcal{T}^n x_0\}_{n=0}^{\infty}$ of successive iterations converges monotone nondecreasingly to ξ_{b_0}. Hence, we infer that the HFIE (3.5) and consequently the HFRDE (1.1)-(1.2) has a local solution ξ_{b_0} and the sequence $\{x_n\}_{n=0}^{\infty}$ of successive approximations defined by (3.4) converges monotone nondecreasingly to ξ_{b_0}. This completes the proof. □

Next, we prove an approximation result concerning the existence and uniqueness of solution simultaneously under weaker form of partial Lipschitz or one sided Lipschitz condition. We need the following hypotheses in what follows. Se for example,

(H$_4$) There exists a constant $\ell > 0$ such that

$$0 \leq f(t,x) - f(t,y) \leq \ell(x-y)$$

for all $t \in J$ and $x, y \in \mathbb{R}$ with $x \geq y$. Moreover, $\dfrac{\ell\, a^q\, \Gamma(q)}{\Gamma(2q)} < 1$.

This conditionis natural and has been used in earlier several papers for nonlinear fractional differential equations. Se for eample Ran and Reurings [29], Dhage and Dhage [9] and Dhage [5, 6] and references therein.

Theorem 3.3. *Suppose that the hypotheses (H$_1$), (H$_3$) and (H$_4$) hold. Furthermore, if the condition*

$$\frac{M_f\, a}{\Gamma(q+1)} \leq \left[1 - \frac{\ell\, a^q\, \Gamma(q)}{\Gamma(2q)}\right] r, \tag{3.8}$$

holds for some real number $r > 0$, then the HFRDE (1.1)-(1.2) has a unique solution ξ_{b_0} in $B_r[x_0]$ defined on J, where $x_0 \equiv \alpha_0$, and the sequence $\{x_n\}_{n=0}^{\infty}$ of successive approximations defined by (3.4) is monotone nondecreasing and converges to ξ_{b_0}.

Proof. Set $(X, K_{1-q}) = \big(C_{1-q}(J,\mathbb{R}), \preceq\big)$ which is a lattice ordered Banach space w.r.t. to the lattice operations $x \vee y = \max\{x,y\}$ and $x \wedge y = \min\{x,y\}$, and so every pair of elements of X has a lower and an upper bound. Let $r > 0$ be a fixed real number and consider the closed ball $B_r[x_0]$

centered at $x_0 \in C_{1-q}(J,\mathbb{R})$ of radius r in the partially ordered Banach space (X, K_{1-q}).

Define an operator $\mathcal{T}$ on X into X by (3.6). Clearly, $\mathcal{T}$ is monotone nondecreasing on X. To see this, let $x, y \in X$ be any two elements such that $x \succeq y$. Then, by hypotheses (H_1) and (H_4), we obtain

$$\mathcal{T}x(t) - \mathcal{T}y(t) = \frac{1}{\Gamma(q)} \int_{t_0}^{t} (t-s)^{q-1} \left[f\big(s, x(s)\big) - f\big(s, y(s)\big) \right] ds \geq 0,$$

for all $t \in J'$. Since $(t-t_0)^{1-q} \geq 0$ for all $t \in J$, we have

$$(t-t_0)^{1-q}\mathcal{T}x(t) - (t-t_0)^{1-q}\mathcal{T}y(t) \geq (t-t_0)^{1-q}\big[\mathcal{T}x(t) - \mathcal{T}y(t)\big] \geq 0,$$

for all $t \in J$. Therefore, $\mathcal{T}x \succeq \mathcal{T}y$ and consequently $\mathcal{T}$ is monotone nondecreasing on X.

Next, we show that $\mathcal{T}$ is a partial contraction on X. Let $x, y \in X$ be such that $x \succeq y$. Then, by hypotheses (H_7) and (H_6), we obtain

$$\begin{aligned}
&\Big|(t-t_0)^{1-q}\big(\mathcal{T}x(t) - \mathcal{T}y(t)\big)\Big| \\
&= \left| \frac{(t-t_0)^{1-q}}{\Gamma(q)} \int_{t_0}^{t} (t-s)^{q-1} \left[f\big(s, x(s)\big) - f\big(s, y(s)\big) \right] ds \right| \\
&\leq \left| \frac{(t-t_0)^{1-q}}{\Gamma(q)} \int_{t_0}^{t} (t-s)^{q-1} \ell\big(x(s) - y(s)\big)\, ds \right| \\
&= \left| \frac{(t-t_0)^{1-q}}{\Gamma(q)} \int_{t_0}^{t} (t-s)^{q-1} (s-t_0)^{q-1} \ell\big[(s-t_0)^{1-q}\big(\big(x(s) - y(s)\big)\big]\, ds \right| \\
&\leq \left| \frac{(t-t_0)^{1-q}}{\Gamma(q)} \int_{t_0}^{t} (t-s)^{q-1} (s-t_0)^{q-1} \ell\big|(s-t_0)^{1-q}\big(\big(x(s) - y(s)\big)\big|\, ds \right| \\
&\leq \left| \frac{\ell(t-t_0)^{1-q}}{\Gamma(q)} \int_{t_0}^{t} (t-s)^{q-1} (s-t_0)^{q-1} \|x-y\|_{1-q}\, ds \right| \\
&\leq \frac{\ell(t-t_0)^{q}}{\Gamma(q)} \int_{0}^{1} u^{q-1} (1-u)^{q-1} \|x-y\|_{1-q}\, du \quad \big(\because s = t_0 + (t-t_0)u\big) \\
&\leq \left(\frac{\ell a^q}{\Gamma(q)} \int_{0}^{1} u^{q-1} (1-u)^{q-1} du \right) \|x-y\|_{1-q} \\
&\leq \frac{\ell a^q}{\Gamma(q)}\, B(q,q) \|x-y\|_{1-q}
\end{aligned}$$

$$\leq \frac{\ell\, a^q\, \Gamma(q)}{\Gamma(2q)} \|x - y\|_{1-q}$$

$$= \lambda \|x - y\|_{1-q}$$

for all $t \in J$. Therefore, taking the supremum over t in the above inequality yields the estimate,

$$\|\mathcal{T}x - \mathcal{T}y\|_{1-q} \leq \lambda \|x - y\|_{1-q}$$

for all comparable elements $x, y \in X$, where $\lambda = \dfrac{\ell\, a^q\, \Gamma(q)}{\Gamma(2q)} < 1$. This shows that $\mathcal{T}$ is a partial contraction on X with contraction constant λ. Furthermore, it can be shown as in the proof of Theorem 3.2 that the element $x_0 \in B_r[x_0]$ satisfies the relation $x_0 \preceq \mathcal{T}x_0$ in view of hypothesis (H_4) and (H_5). Finally, by hypothesis (H_1) and condition (3.6), one has

$$\begin{aligned}\|x_0 - \mathcal{T}x_0\|_{1-q} &= \sup_{t\in J} \left| \frac{(t-t_0)^{1-q}}{\Gamma(q)} \int_{t_0}^{t} (t-s)^{q-1} f\big(s, x_0(s)\big)\, ds \right| \\ &\leq \frac{M_f\, a}{\Gamma(q+1)} \\ &\leq \left[1 - \frac{\ell\, a^q\, \Gamma(q)}{\Gamma(2q)}\right] r\end{aligned}$$

which shows that the condition (3.8) of Theorem 2.12 is satisfied. Hence $\mathcal{T}$ has a unique fixed point ξ_{b_0} in $B_r[x_0]$ and the sequence $\{\mathcal{T}^n x_0\}_{n=0}^{\infty}$ of successive iterations converges monotone nondecreasingly to ξ_{b_0}. This further implies that the HFIE (3.5) and consequently the HFRDE (1.1)-(1.2) has a unique local solution ξ_{b_0} defined on J and the sequence $\{x_n\}_{n=0}^{\infty}$ of successive approximations defined by (3.4) converges monotone nondecreasingly to ξ_{b_0}. This completes the proof. □

Next we prove the approximation results if the nonlinearity f satisfies the opposite condition to that of hypothesis (H_3). We consider ehe following hypothesis in what follows.

(H_5) The function f satisfies the functional inequality $f(t, \alpha_0(t)) \leq 0$ for all $t \in J$, where $\alpha_0(t) = \frac{b_0}{\Gamma(q)}(t - t_0)^{q-1}, \quad t \in J'$.

Theorem 3.4. *Suppose that the hypotheses (H_1)-(H_2) and (H_5) hold. Furthermore, if there is a real number $r > 0$ such that $\frac{M_f\, a}{\Gamma(q+1)} \leq r$, then the HFRDE (1.1)-(1.2) has a local solution ξ_{b_0} in $B_r[x_0]$, where $x_0(t) =$*

$\alpha_0(t)$, $t \in J'$, and the sequence $\{x_n\}_{n=0}^{\infty}$ of successive approximations defined by (3.4) is monotone nonincreasing and converges to ξ_{b_0}.

Proof. Now by hypothesis (H_5), the initial function $x_0(t) = \alpha_0(t)$, $t \in J$, satisfies the order relation $x_0 \succeq \mathcal{T}x_0$, so by an application of Theorem 2.9, the HFRDE (1.1)-(1.2) has a local solution ξ_{b_0} defined on J and the sequence $\{x_n\}_{n=0}^{\infty}$ of successive approximations defined by (3.4) is monotone nonincreasing and converges to the solution ξ_{b_0}. □

Theorem 3.5. *Suppose that the hypotheses (H_1), (H_4) and (H_5) hold. Furthermore, if there is a number $r > 0$ such that the condition (3.8) holds, then the HFRDE (1.1)-(1.2) has a unique local solution ξ_{b_0} in $B_r[x_0]$ defined on J, where $x_0 \equiv \alpha_0$, and the sequence $\{x_n\}_{n=0}^{\infty}$ of successive approximations defined by (3.4) is monotone nonincreasing and converges to ξ_{b_0}.*

Proof. The proof is similar to Theorem 2.12 with appropriate modifications. We omit the details. □

Remark 3.6. If the initial condition in the equation (1.1) is such that $b_0 > 0$, then under the conditions of Theorem 3.2, the HFRDE (1.1)-(1.2) has a local positive solution ξ_{b_0} defined on J in view of the fact that $\dfrac{b_0}{\Gamma(q)}(t - t_0)^{q-1} > 0$ for all $t \in J'$ and the sequence $\{x_n\}_{n=0}^{\infty}$ of successive approximations defined by (3.4) converges monotone nondecreasingly to the positive solution ξ_{b_0}. Similarly, under the conditions of Theorem 3.3, the HFRDE (1.1)-(1.2) has a unique local positive solution ξ_{b_0} defined on J and the sequence $\{x_n\}_{n=0}^{\infty}$ of successive approximations given by (3.4) converges monotone nondecreasingly to the unique positive solution ξ_{b_0} on entire interval J.

Remark 3.7. The convergence of the sequence in Theorems 3.2 and 3.4 is not geometrical but monotonic and so the rate of convergence of the sequence of successive approximations is not known, however the convergence of the sequence in Theorems 3.3 and 3.5 is geometrical and hence we have the idea of rate of convergence. The more the value of one sised Lipschitz constant ℓ, we have lesser rhe rate of conergence and smaller the value of ℓ one has more rate of convergene of sequence of sucessive approximations to the soution. Thus the rate of convergence of the sequence is inversely proportional to the value of the partial Lipschitz constant ℓ.

4. Approximation of Ulam-Hyers Stability of local solution

The Ulam-Hyers stability for various dynamic systems has already been discussed by several authors under the conditions of integral inequalities and classical Banach fixed point theorem on nonlinearity f (see Tripathy [30], Huang *et al.* [24] and references therein). Here, in the present paper, we discuss the approximation of the Ulam-Hyers stability of local solution of the HFRDE (1.1)-(1.2) under the conditions of hybrid fixed point principle of partial contraction mappings stated in Theorem 2.12. We need the following definition in what follows.

Definition 4.1. The HFRDE (1.1)-(1.2) is said to be locally Ulam-Hyers stable if for $\epsilon > 0$ and for each local solution $y \in B_r[x_0] \subset C_{1-q}(J, \mathbb{R})$ of the inequality

$$\left.\begin{aligned} \left|{}^{RL}D^q_{t_0} y(t) - f(t, y(t))\right| \leq \epsilon, \;\; t \in J', \\ \lim_{t \to t_{0+}} I^{1-q}_{t_{0+}} x(t) = b_0, \end{aligned}\right\} \tag{4.1}$$

there exists a constant $K_f > 0$ such that

$$\left|(t - t_0)^{1-q}\big(y(t) - \xi_{b_0}(t)\big)\right| \leq K_f \epsilon \tag{4.2}$$

for all $t \in J$, where $\xi_{b_0} \in B_r[x_0]$ is a local solution of the HFRDE (1.1)-(1.2) defined on J. The solution ξ_{b_0} of the HFRDE (1.1)-(1.2) is called weighed Ulam-Hyers stable local solution on J.

Note that the Ulam-hyers stability condition (4.2) is different from usual one beause the solution belongs to the Banach space $C_{1-q}(J, \mathbb{R})$. Now we are equipped with all necessary details to prove our main Ulam-Hyers stability result of this section.

Theorem 4.2. *Assume that all the hypotheses of Theorem 3.3 hold. Then the HFRDE (1.1)-(1.2) has a unique Ulam-Hyers stable local solution $\xi_{b_0} \in B_r[x_0]$, where $x_0 \equiv \alpha_0$, and the sequence $\{x_n\}_{n=0}^{\infty}$ of successive approximations given by (3.4) converges monotone nondecreasingly to ξ_{b_0}.*

Proof. Let $\epsilon > 0$ be given and let $y \in B_r[x_0]$ be a solution of the functional inequality (*) on J, that is, we have

$$\left.\begin{aligned} \left|{}^{RL}D^q_{t_0} y(t) - f(t, y(t))\right| \leq \epsilon, \;\; t \in J', \\ \lim_{t \to t_{0+}} I^{1-q}_{t_{0+}} y(t) = b_0. \end{aligned}\right\} \tag{4.3}$$

By Theorem 3.3, the HFRDE (1.1)-(1.2) has a unique local solution $\xi_{b_0} \in B_r[x_0]$. Then by Lemma 3.1, one has

$$\xi_{b_0}(t) = \frac{b_0}{\Gamma(q)}(t-t_0)^{q-1} + \frac{1}{\Gamma(q)} \int_{t_0}^{t} (t-s)^{q-1} f\big(s, \xi_{b_0}(s)\big)\, ds, \quad t \in J'. \tag{4.4}$$

Now, applying Riemann-Louville integration on both sides of (4.1) yields the estimate:

$$\left| y(t) - \frac{b_0}{\Gamma(q)}(t-t_0)^{q-1} - \frac{1}{\Gamma(q)} \int_{t_0}^{t} (t-s)^{q-1} f\big(s, y(s)\big)\, ds \right| \leq \frac{\epsilon\, a^q}{\Gamma(q+1)} \tag{4.5}$$

for all $t \in J'$. Next, from (4.2) and (4.3) we obtain

$$\begin{aligned}
&\Big|(t-t_0)^{1-q}\big(y(t) - \xi_{b_0}(t)\big)\Big| \\
&= \left| (t-t_0)^{1-q} y(t) - \frac{b_0}{\Gamma(q)} - \frac{(t-t_0)^{1-q}}{\Gamma(q)} \int_{t_0}^{t} (t-s)^{q-1} f\big(s, \xi_{b_0}(s)\big)\, ds \right| \\
&\leq \left| (t-t_0)^{1-q} y(t) - \frac{b_0}{\Gamma(q)} - \frac{(t-t_0)^{1-q}}{\Gamma(q)} \int_{t_0}^{t} (t-s)^{q-1} f\big(s, y(s)\big)\, ds \right| \\
&\quad + \left| \frac{(t-t_0)^{1-q}}{\Gamma(q)} \int_{t_0}^{t} (t-s)^{q-1} f\big(s, y(s)\big)\, ds \right. \\
&\qquad \left. - \frac{(t-t_0)^{1-q}}{\Gamma(q)} \int_{t_0}^{t} (t-s)^{q-1} f\big(s, \xi_{b_0}(s)\big)\, ds \right| \\
&\leq \frac{\epsilon\, a^{q+1}}{\Gamma(q+1)} + \left| \frac{(t-t_0)^{1-q}}{\Gamma(q)} \int_{t_0}^{t} (t-s)^{q-1} \big[f\big(s, y(s)\big) - f\big(s, \xi_{b_0}(s)\big) \big]\, ds \right| \\
&\leq \frac{\epsilon\, a^{q+1}}{\Gamma(q+1)} + \left| \frac{(t-t_0)^{1-q}}{\Gamma(q)} \int_{t_0}^{t} \frac{(t-s)^{q-1}}{(t-t_0)^{1-q}}\, \ell \left[(t-t_0)^{1-q} \big(y(s) - \xi_{b_0}(s) \big) \right] ds \right| \\
&= \frac{\epsilon\, a^{q+1}}{\Gamma(q+1)} + \frac{\ell\, a^q\, \Gamma(q)}{\Gamma(2q)}\, \|y - \xi_{b_0}\|_{1-q}.
\end{aligned}$$

for all $t \in J$. Taking the supremum over t in the above inequality, we obtain

$$\|y - \xi_{b_0}\|_{1-q} \leq \frac{\epsilon\, a^{q+1}}{\Gamma(q+1)} + \frac{\ell\, a^q\, \Gamma(q)}{\Gamma(2q)}\, \|y - \xi_{b_0}\|_{1-q}.$$

or

$$\|y - \xi_{b_0}\|_{1-q} \leq \left[\frac{\dfrac{a^{q+1}}{\Gamma(q+1)}}{1 - \frac{\ell\, a^q\, \Gamma(q)}{\Gamma(2q)}} \right] \epsilon, \quad \big(\because \lambda = \frac{\ell\, a^q\, \Gamma(q)}{\Gamma(2q)} < 1 \big).$$

Letting $K_f = \left[\frac{a^{q+1}}{\Gamma(q+1)}\Big/1 - \frac{\ell\, a^q\, \Gamma(q)}{\Gamma(2q)}\right] > 0$, we obtain the estimate

$$\Big|(t-t_0)^{1-q}\big(y(t)-\xi_{b_0}(t)\big)\Big| \leq K_f\epsilon$$

for all $t \in J$. As a result, ξ is a Ulam-Hyers stable local solution of the HFRDE (1.1)-(1.2) on J and the sequence $\{x_n\}_{n=0}^{\infty}$ of successive approximations defined by (3.4) converges monotone nondecreasingly to ξ_{b_0}. Consequently the HFRDE (1.1)-(1.2) is a locally Ulam-Hyers stable on J. This completes the proof. □

Remark 4.3. If the initial condition in the equation (1.1)-(1.2) is such that $b_0 > 0$, then under the conditions of Theorem 4.2, the HFRDE (1.1)-(1.2) has a unique Ulam-Hyers stable local positive solution ξ_{b_0} defined on J in view of the fact that $\dfrac{b_0}{\Gamma(q)}(t-t_0)^{q-1} > 0$ for all $t \in J'$ and the sequence $\{x_n\}_{n=0}^{\infty}$ of successive approximations defined by (3.4) converges monotone nondecreasingly to ξ_{b_0} on the entire interval J.

In the following section we prove some comparison results concerning monotonic dependence of the local solution of HFRDE (1.1)-(1.2) on the initial condition. In particular, it is shown that values of the local solution of the HFRDE like (1.1)-(1.2) are directly proportional to the values of initial data of past history or memory of related dynamic system.

5. Monotonic Dependence on Initial Data

The qualitative analysis concerning the monotonic dependence of the local solution on initial data is initiated by Dhage and Dhage [16] who proved that if the initial conditions of two nonlinear ordinary differential equations with integer order are comparable, then in the same order the corresponding local solutions are also comparable. Here, we obtain similar results for the non-integer, that is, fractional order HFRDE (1.1)-(1.2) on the entire interval J.

Now consider the IVP of HFRDE

$$^{RL}D_{t_0}^{q}x(t) = f(t,x(t)),\ \ t \in J', \tag{5.1}$$

satisfying the fractional integral initial condition (IC)

$$\lim_{t\to t_{0+}} I_{t_0+}^{1-q}x(t) = b_1 \in \mathbb{R}, \tag{5.2}$$

where the function $f : J \times \mathbb{R} \to \mathbb{R}$ is given as in HFRDE (1.1)-(1.2). We consider the following hypothesis in the sequel.

(H$_6$) $f(t, \alpha_1(t)) \geq 0$ for each $t \in J$, where $\alpha_1(t) = \frac{b_1}{\Gamma(q)}(t-t_0)^{q-1}$, $t \in J'$.

Our first monotonic dependence result is as follows.

Theorem 5.1. *Suppose that the hypotheses (H_1), (H_2), (H_3) and (H_6) hold. If there is a real number $r > 0$ such that $\frac{M_f\, a}{\Gamma(q+1)} \leq r$ and $b_0 \leq b_1$, then the HFRDEs (1.1)-(1.2) and (5.1)-(5.2) have local solutions respectively ξ_{b_0} and ξ_{b_1} satisfying $\xi_{b_0} \preceq \xi_{b_1}$ on J. Moreover, if $\|\alpha_0 - \alpha_1\|_{1-q} > 2r$, then $\xi_{b_0} \preceq \xi_{b_1}$, $\xi_{b_0} \neq \xi_{b_1}$ on J.*

Proof. By Theorem 3.2, the HFRDE (5.1)-(5.2) has a local solution $\xi_{b_1} \in B_r[\alpha_1]$ and the sequence $\{y_n\}_{n=0}^{\infty}$ of successive approximations defined by

$$\left.\begin{aligned} y_0(t) &= \frac{b_1}{\Gamma(q)}(t-t_0)^{q-1}, \quad t \in J', \\ y_{n+1}(t) &= \frac{b_1}{\Gamma(q)}(t-t_0)^{q-1} + \frac{1}{\Gamma(q)} \int_{t_0}^{t} (t-s)^{q-1} f\big(s, y_n(s)\big)\, ds, \quad t \in J', \end{aligned}\right\} \tag{5.3}$$

for $n = 0, 1, \ldots$; is monotone nondecreasing and converges to ξ_{b_1}. We show that $\xi_{b_0} \preceq \xi_{b_1}$ on J. Since $b_0 \leq b_1$ one has

$$x_0(t) = \alpha_0(t) = \frac{b_0}{\Gamma(q)}(t-t_0)^{q-1} \leq \frac{b_1}{\Gamma(q)}(t-t_0)^{q-1} = \alpha_1(t) = y_0(t)$$

and

$$\begin{aligned} x_1(t) &= \frac{b_0}{\Gamma(q)}(t-t_0)^{q-1} + \frac{1}{\Gamma(q)} \int_{t_0}^{t} (t-s)^{q-1} f(s, x_0(s))\, ds \\ &\leq \frac{b_1}{\Gamma(q)}(t-t_0)^{q-1} + \frac{1}{\Gamma(q)} \int_{t_0}^{t} (t-s)^{q-1} f(s, y_0(s))\, ds \\ &= y_1(t) \end{aligned}$$

for all $t \in J'$. Since $(t-t_0)^{1-q} \geq 0$ for all $t \in J$, one has $(t-t_0)^{1-q} x_1(t) \leq (t-t_0)^{1-q} y_1(t)$ for all $t \in J$. Proceeding in this way, by induction, we obtain $x_n \preceq y_n$ on J. By definition of the partial order $\preceq$, we obtain $x_n \preceq y_n$ for all $n \in \mathbb{N}$ and $x_n \to \xi_{b_0}$, $y_n \to \xi_{b_1}$. Now from Lemma 2.4 it follows that $\xi_{b_0} \preceq \xi_{b_1}$ on J, that is, $\xi_{b_0}(t) \leq \xi_{b_1}(t)$ for all $t \in J'$.

Next, suppose that $\|\alpha_0 - \alpha_1\|_{1-q} > 2r$. Then $B_r[\alpha] \bigcap B_r[\alpha_1] = \emptyset$ and ξ_{b_0} and ξ_{b_1} completely lie in the closed balls $B_r[\alpha_0]$ and $B_r[\alpha_1]$ respectively.

Therefore, $\xi_{b_0}(t) \neq \xi_{b_1}(t)$ for all $t \in J'$. As a result, we obtain $\xi_{b_0} \preceq \xi_{b_1}$, $\xi_{b_0} \neq \xi_{b_1}$ on J. This completes the proof. □

Theorem 5.2. *Suppose that the hypotheses (H_1), (H_3), (H_4) and (H_6) hold. If there is a real number $r > 0$ such that the condition (3.8) holds and $b_0 \leq b_1$, then the HFRDEs (1.1)-(1.2) and (5.1)-(5.2) have unique local solutions respectively ξ_{b_0} and ξ_{b_1} satisfying $\xi_{b_0} \preceq \xi_{b_1}$ on J. Moreover, if $\|\alpha_0 - \alpha_1\|_{1-q} > 2r$, then $\xi_{b_0} \preceq \xi_{b_1}$, $\xi_{b_0} \neq \xi_{b_1}$ on J.*

Proof. The proof is similar to Theorem 5.1 and the conclusion follows by an application of Theorem 2.12. □

Next we consider the following hypothesis in what follows.

(H_7) $f(t, \alpha_1(t)) \leq 0$ for each $t \in J$, where $\alpha_1(t) = \frac{b_1}{\Gamma(q)}(t - t_0)^{q-1}$, $t \in J'$.

The following approximate existence and uniqueness theorems similar to Theorems 3.4 and 3.5 are obvious and may be stated without proof.

Theorem 5.3. *Assume that the hypotheses (H_1), (H_2) and (H_7) hold. If there is a real number $r > 0$ such that $\frac{M_f\, a^{q+1}}{\Gamma(q+1)} \leq r$, then the HFRDE (5.1)-(5.2) has a local solution $\xi_{b_1} \in B_r[\alpha_1]$ and the sequence $\{y_n\}_{n=0}^{\infty}$ of successive approximations defined by (5.3) is monotone nonincreasing and converges to ξ_{b_1}.*

Theorem 5.4. *Assume that the hypotheses (H_1), (H_4) and (H_7) hold. If there is a real number $r > 0$ such that the condition (3.6) holds, then the HFRDE (5.1)-(5.2) has a unique local solution $\xi_{b_1} \in B_r[\alpha_1]$ and the sequence $\{y_n\}_{n=0}^{\infty}$ of successive approximations defined by (5.3) is monotone nonincreasing and converges to ξ_{b_1}.*

Our next monotonic dependence result of the local solution on initial data is obtained between the given vector segment of the initial history representation.

Theorem 5.5. *Suppose that the hypotheses (H_1), (H_2), (H_3) and $H_6)$ hold. If there is a real number $r > 0$ such that $\frac{M_f\, a}{\Gamma(q+1)} \leq r$ and $b_0 \leq b_1$, then the HFRDEs (1.1)-(1.2) and (5.1)-(5.2) have local solutions respectively ξ_{b_0} and ξ_{b_1} in the vector segment $[\alpha_0, \alpha_1]$ satisfying $\xi_{b_0} \preceq \xi_{b_1}$ on J. Furthermore, if $\|\alpha_0 - \alpha_1\|_{1-q} > 2r$, then $\xi_{b_0} \preceq \xi_{b_1}$ and $\xi_{b_0} \neq \xi_{b_1}$ on J.*

Proof. Since all the conditions of Theorem 3.2 are satisfied, we have a sequence $\{x_n\}_{n=0}^{\infty}$ of successive approximations defined by (3.3) satisfying

$$\alpha_0 = x_0 \preceq x_1 \preceq \cdots \preceq x_n \cdots ; \quad x_n \to \xi_{b_0},$$

and ξ_{b_0} is a local solution of the HFRDE (1.1)-(1.2) on J. Similarly, since all the conditions of Theorem 5.3 are satisfied, we have a sequence $\{y_n\}_{n=0}^{\infty}$ of successive approximations defined by (5.3) satisfying

$$\alpha_1 = y_0 \succeq y_1 \succeq \cdots \succeq y_n \cdots ; \quad y_n \to \xi_{b_1},$$

and ξ_{b_1} is a local solution of the HFDE (5.1)-(5.2) on J. Now, given that $b_0 \leq b_1$, so we have

$$\alpha_0(t) = x_0(t) \leq y_0(t) = \alpha_1(t)$$

for all $t \in J'$. Next,

$$\begin{aligned} x_1(t) &= \frac{b_0}{\Gamma(q)}(t-t_0)^{q-1} + \frac{1}{\Gamma(q)} \int_{t_0}^{t} (t-s)^{q-1} f(s, x_0(s))\, ds \\ &\leq \frac{b_1}{\Gamma(q)}(t-t_0)^{q-1} + \frac{1}{\Gamma(q)} \int_{t_0}^{t} (t-s)^{q-1} f(s, y_0(s))\, ds \\ &= y_1(t) \end{aligned}$$

for all $t \in J'$. Since $(t-t_0)^{1-q} \geq 0$ for all $t \in J$, one has $(t-t_0)^{1-q} x_1(t) \leq (t-t_0)^{1-q} y_1(t)$ for all $t \in J$. Thus, we have

$$\alpha_0 \preceq x_1 \preceq y_1 \preceq \alpha_1$$

on J. Proceeding in this way, by induction, we obtain

$$\alpha_0 \preceq x_n \preceq y_n \preceq \alpha_1$$

for all $n = 1, \ldots,$ on J. Passing the limit as $n \to \infty$ in the above inequality, we get

$$\alpha_0 \preceq \xi_{b_0} \preceq \xi_{b_1} \preceq \alpha_1$$

on J. Consequently the local solutions $\xi_{b_0} \in B_r[\alpha_0]$ and $\xi_{b_1} \in B_r[\alpha_1]$ of the fractional differential equations (1.1)-(1.2) and (5.1)-(5.2) lie in the vector segment $[\alpha_0, \alpha_1]$ satisfying $\xi_{b_0} \preceq \xi_{b_1}$ for all $t \in J$. Next, suppose that $\|\alpha_0 - \alpha_1\|_{1-q} > 2r$, then $B_r[\alpha_0] \bigcap B_r[\alpha_1] = \emptyset$, so the graphs of the local solutions ξ_{b_0} and ξ_{b_1} do not intersect at any point of J, whence $\xi_{b_0} \preceq \xi_{b_1}$ and $\xi_{b_0} \neq \xi_{b_1}$ on J. This completes the proof. □

Theorem 5.6. *Suppose that the hypotheses (H_1), (H_3), (H_4) and (H_6) hold. If there is a real number $r > 0$ such that the condition (3.8) holds, and $b_0 \leq b_1$, then the HFRDEs (1.1)-(1.2) and (5.1)-(5.2) have unique local solutions ξ_{b_0} and ξ_{b_1} respectively in the vector segment $[\alpha_0, \alpha_1]$ satisfying $\xi_{b_0} \preceq \xi_{b_1}$ on J. Furthermore, if $\|\alpha_0 - \alpha_1\|_{1-q} > 2r$, then $\xi_{b_0} \preceq \xi_{b_1}$ and $\xi_{b_0} \neq \xi_{b_1}$ on J.*

Proof. The proof is similar to Theorem 5.5 with appropriate modifications and now the desired conclusion follows by the applications of Theorem 2.12 and 3.5. We omit the details. □

Remark 5.7. If the real numbers b_0 and b_1 given respectively in the initial conditions (1.2) and (5.2) of the HFRDEs (1.1) and (5.1) satisfy the inequality $0 \leq b_0 \leq b_1$, then the local solutions ξ_{b_0} and ξ_{b_1} of the HFRDEs (1.1)-(1.2) and (5.1)-(5.2) mentioned in Theorems 5.1 through 5.6 are both positive and satisfy the stated inequalities on the entire interval J.

6. Examples

In the following we present a couple of numerical examples to illustrate the hypotheses and abstract results involved in the approximation theorems of this paper concerning existence, uniqueness, Ulam-Hyers stability and monotonic dependence of the solution.

Example 6.1. Given a closed and bounded interval $J = [0, 1]$ in $\mathbb{R}$, consider the IVP of nonlinear first order HFRDE,

$$^{RL}D_0^{1/2} x(t) = (t+1)\tanh x(t), \;\; t \in (0, 1], \tag{6.1}$$

satsfying the initial conditions

$$\lim_{t \to 0} I_0^{1/2}\, x(t) = 1, \tag{6.2}$$

$$\lim_{t \to 0} I_0^{1/2}\, x(t) = 2, \tag{6.3}$$

$$\lim_{t \to 0} I_0^{1/2}\, x(t) = 17. \tag{6.4}$$

Here, $q = \frac{1}{2}$, $b_0 = 1$, $f(t, x) = (t+1)\tanh x$, $(t, x) \in (0, 1] \times \mathbb{R}$. We show that the functions f satisfies all the conditions of Theorem 3.2. Clearly, f is bounded on $[0, 1] \times \mathbb{R}$ with bound $M_f = 2$ and so the hypothesis (H_1) is satisfied. Also the function $f(t, x)$ is nondecreasing in x for each $t \in (0, 1]$.

Therefore, hypothesis (H_2) is satisfied. Moreover, we have

$$\begin{aligned} f(t,\alpha_0(t)) &= f\Big(t,\frac{t^{-1/2}}{\sqrt{\pi}}\Big) \\ &= (t+1)\tanh\Big(\frac{t^{-1/2}}{\sqrt{\pi}}\Big) \\ &\geq 0 \end{aligned}$$

for each $t \in (0,1]$, where $\alpha_0(t) = \dfrac{t^{-1/2}}{\sqrt{\pi}}$, so the hypothesis (H_3) holds. If we take $r = 4$, all the conditions of Theorem 3.2 are satisfied. Hence, the HFRDE (6.1)-(6.2) has a local solution ξ_1 in the closed ball $B_4\big[\alpha_0\big]$ of $C_{1-q}(J,\mathbb{R})$ which is positive in view of Remark 3.6. Moreover, the sequence $\{x_n\}_{n=0}^{\infty}$ of successive approximations defined by

$$x_0(t) = \frac{t^{-1/2}}{\sqrt{\pi}},\ \ t \in (0,1],$$

$$x_{n+1}(t) = \frac{t^{-1/2}}{\sqrt{\pi}} + \frac{1}{\sqrt{\pi}}\int_0^t (t-s)^{-1/2}(s+1)\tanh x_n(s)\,ds,\ \ t \in (0,1],$$

for $n = 1, 2, \ldots$; converges monotone nondecreasingly to ξ_1. Similarly, following the analogous arguments it can be proved that the HFRDEs (6.1)-(6.3) and (6.1)-(6.4) have the local positive solutions $\xi_2 \in B_4\big[\alpha_1\big]$ and $\xi_{17} \in B_4\big[\alpha_2\big]$ defined on $[0,1]$ in view of Remark 3.6, where $\alpha_1(t) = \dfrac{2t^{-1/2}}{\sqrt{\pi}}$ and $\alpha_2(t) = \dfrac{17t^{-1/2}}{\sqrt{\pi}}$. Since $b_0 = 1 \leq 2 = b_1$, we have by Theorem 5.1, $\xi_1 \preceq \xi_2$ on entire interval $[0,1]$, that is $\xi_1(t) \leq \xi_2(t)$ for all $t \in (0,1]$ and $t^{1/2}\xi_1(t) \leq t^{1/2}\xi_2(t)$ for all $t \in [0,1]$. Again, we have

$$\begin{aligned} \|\alpha_0 - \alpha_2\|_{1-q} &= \sup_{t\in[0,1]}\left|(t^{1/2}\Big(\frac{t^{-1/2}}{\sqrt{\pi}} - \frac{17t^{-1/2}}{\sqrt{\pi}}\Big)\right| \\ &= \frac{16}{\sqrt{\pi}} > 8 = 2r. \end{aligned}$$

Therefore, by Theorem 5.1, $\xi_1 \preceq \xi_{17}$, $\xi_1 \neq \xi_{17}$ on the whole interval $[0,1]$. Consequently $\xi_1(t) < \xi_{17}(t)$ for all $t \in (0,1]$ and $t^{1/2}\xi_1(t) < t^{1/2}\xi_{17}(t)$ for all $t \in [0,1]$.

The approximation concerning the existence and uniqueness of local solution is elaborated in the following example.

Example 6.2. Given a closed and bounded interval $J = [0, 1]$ in $\mathbb{R}$, consider the IVP of nonlinear first order HFRDE,

$$^{RL}D_0^{1/2}x(t) = \frac{1}{6}(t+1)\tan^{-1}x(t), \quad t \in (0, 1], \tag{6.5}$$

satisfying the intial conditions

$$\lim_{t\to 0} I_0^{1/2}\, x(t) = 1, \tag{6.6}$$

$$\lim_{t\to 0} I_0^{1/2}\, x(t) = 2, \tag{6.7}$$

$$\lim_{t\to 0} I_0^{1/2}\, x(t) = 9. \tag{6.8}$$

Here, $q = 1/2$, $b_0 = 1$, $b_1 = 2$, $b_2 = 9$, and $f(t, x) = \frac{1}{6}(t+1)\tan^{-1}x$ for $(t, x) \in (0, 1] \times \mathbb{R}$. We show that f satisfies all the conditions of Theorem 3.3. Clearly, f is bounded on $(0, 1] \times \mathbb{R}$ with bound $M_f = \frac{11}{21}$ and so, the hypothesis (H_1) is satisfied. Further, the function $f(t, x)$ is nondecreasing in x for each $t \in (0, 1]$, so the hypothesis (H_2) is satisfied. Next, let $x, y \in \mathbb{R}$ be such that $x \geq y$. Then there exists a constant ξ with $x < \xi < y$ satisfying

$$\begin{aligned} 0 \leq f(t, x) - f(t, y) &\leq \frac{1}{6} \cdot \frac{t+1}{(1+\xi^2)}(x - y) \\ &\leq \frac{1}{3}(x - y) \end{aligned}$$

for all $t \in (0, 1]$. So the hypothesis (H_7) holds with $\ell = \frac{1}{3}$. Next, we have

$$f(t, \alpha_0(t)) = f\left(t, \frac{t^{-1/2}}{\sqrt{\pi}}\right) = (t+1)\tan^{-1}\left(\frac{t^{-1/2}}{\sqrt{\pi}}\right) \geq 0$$

for each $t \in [0, 1]$, where $\alpha_0(t) = \dfrac{t^{-1/2}}{\sqrt{\pi}}$, and so the hypothesis (H_3) holds. Furthermore, we obtain

$$\frac{\ell\, a^q\, \Gamma(q)}{\Gamma(2q)} = \frac{\Gamma(1/2)}{3\Gamma(1)} \leq \frac{2}{5} < 1.$$

If we take $r = 2$, then we have

$$\frac{M_f\, a}{\Gamma(q+1)} \leq 1 \leq \left(1 - \frac{2}{5}\right) \cdot 2 \leq \left[1 - \frac{\ell\, a^q\, \Gamma(q)}{\Gamma(2q)}\right] r$$

and so, the condition (3.8) is satisfied. Thus, all the conditions of Theorem 3.3 are fulfilled. Hence, the HFRDE (6.5)-(6.6) has a unique local solution ξ_1 in the closed ball $B_2[\alpha_0]$ of $C_{1-q}(J, \mathbb{R})$ which is positive in view of Remark 3.6. Moreover, the sequence $\{x_n\}_{n=0}^{\infty}$ of successive approximations defined

by

$$x_0(t) = \frac{t^{-1/2}}{\sqrt{\pi}}, \ \ t \in (0,1],$$

$$x_{n+1}(t) = \frac{t^{-1/2}}{\sqrt{\pi}} + \frac{1}{12} \cdot \frac{1}{\sqrt{\pi}} \int_0^t (t-s)^{-1/2}(s+1) \tan^{-1} x_n(s)\, ds,$$

for all $t \in (0,1]$ and for $n = 1, 2, \ldots$; converges monotone nondecreasingly to ξ_1. Furthermore, the unique local positive solution ξ_1 is Ulam-Hyers stable on $[0,1]$ in view of Definition 4.1. Consequently the HFRDE (6.5)-(6.6) is a locally Ulam-Hyers stable on the interval $[0,1]$. Again, following the similar arguments it is proved that the HFRDEs (6.5)-(6.7) and (6.5)-(6.8) have unique local positive Ulam-Hyers stable solutions $\xi_2 \in B_2[\alpha_1]$ and $\xi_9 \in B_2[\alpha_2]$ defined on $[0,1]$, where $\alpha_1(t) = \dfrac{2t^{-1/2}}{\sqrt{\pi}}$ and $\alpha_2(t) = \dfrac{9t^{-1/2}}{\sqrt{\pi}}$. Since $b_0 = 1 \leq 2 = b_1$, we have by Theorem 5.1, $\xi_1 \preceq \xi_2$ on $[0,1]$. Again, we have

$$\begin{aligned} \|\alpha_0 - \alpha_2\|_{1-q} &= \sup_{t\in[0,1]} \left| t^{1/2} \Big(\frac{t^{-1/2}}{\sqrt{\pi}} - \frac{9t^{-1/2}}{\sqrt{\pi}} \Big) \right| \\ &= \frac{8}{\sqrt{\pi}} > 4 = 2r. \end{aligned}$$

Therefore, by an application of Theorem 5.1, $\xi_1 \preceq \xi_9$, $\xi_1 \neq \xi_9$ on the whole interval $[0,1]$. Consequently $\xi_1(t) < \xi_9(t)$ for all $t \in (0,1]$ and $t^{1/2}\xi_1(t) < t^{1/2}\xi_9(t)$ for all $t \in [0,1]$.

7. Remarks and Conclusion

The existence of local solution for the HFRDE (1.1)-(1.2) in the sense that given in Coddington [4] can be proved under the hypothesis (H_1) using a method based on Schauder fixed point principle, but in that case we do not get any sequence of iterations for approximating the local solution and the existence of solution is obtained on a very small part of the interval of J. On other hand with our present new approach, we obtain an algorithm for the solution and the solution exists on whole of the interval J. Similarly, unlike the method involving Banach contraction mapping principle, we do not need the usual Lipschitz condition on the nonlinear function involved in the equation, but instead a partial Lipschitz condition serves our purpose. Again, we obtain additional qualitative information of the monotonic

convergence of the sequence of successive approximations to the local solution. Moreover, because of our present method we get qualitative analysis concerning the monotonic dependence of the local solution on the initial data which otherwise not possible by older methods. A key advantage of the present approach lies in the fact that Dhage iteration method became instrumental in proving the monotonic convergence of the Picard's successive approximations to the local solution of the considered problem without the use of usual Lipschitz condition. Again, the hypotheses considered in this paper for eistence results are nattural and widely used in the study of nonlinear differential equations. Next, the results of this paper concern the theoretical approximation, however the numerical error analysis of solution of the problems given in (6.1)-(6.2) and (6.5) -(6.6) may be obtained via the algorithms provided there by using advanced computer techniques. Thus our Dhage iteration method used in this paper has some advantages over the older ones. However, there are certain limitations of this method. First the convergence criteria fails if the solution space is not a Janhavi Banach space and secondly, the uniqueness outcome fails if the solution space is not a lattice ordered Banach space. Finally, the present method may be extended to other sophisticated complex integer or fractional order nonlinear differential equations such as functional and impulsive fractional differential and integrodifferential equations etc., with different fractional derivatives and with appropriate modifications to prove similar qualitative aspects of the local solution. The study in this direction forms further scope for the research work and some of the results of this direction will be reported elsewhere.

Acknowledgement: The authors are thankful to the editor and referee for giving some useful suggestions for the improvement of this paper.

References

[1] Agarwal, R., Hristova, S., O'Regan, D., Basic concepts of Riemann-Liouville fractional differential equations with non-instantaneous impulses, *Symmetry* **11** (2019), 1–21.

[2] Ahmad, B., Alsaedi, A., Ntouyas, S.K., Tariboon, J., *Hadmard-type Fractional Differential Equations, Inclusions and inequalities*, Springer International Publishing AG, Delhi, India, 2017.

[3] Birkhoff, G., *Lattice Theory*, Amer. Math. Soc. Coll. Publ. New York 1967.

[4] Coddington, E. A., *An Introduction to Ordinary Differential Equations*, Dover Publications Inc. New York, 1989.

[5] Dhage, B. C., Partially condensing mappings in partially ordered normed linear spaces and applications to functional integral equations, *Tamkang J. Math.*, **45** (4) (2014), 397-427.

[6] Dhage, B. C., Coupled and mixed coupled hybrid fixed point principles in a partially ordered Banach algebra and PBVPs of nonlinear coupled quadratic differential equations, *Differ. Equ. Appl.*, **11**(1) (2019), 1-85.

[7] Dhage, B. C., A Schauder type hybrid fixed point theorem in a partially ordered metric space with applications to nonlinear functional integral equations, *Jñānābha* **52** (2) (2022), 168-181.

[8] Dhage, B. C., Nonlinear partial completely continuous operators in a partially ordered Banach space and nonlinear hyperbolic partial differential equations, *Malaya J. Mat.*, **12** (4) (2024), 330-338.

[9] Dhage, B. C., Dhage, S. B., Approximating solutions of nonlinear first order ordinary differential equations, *GJMS Special issue for Recent Advances in Mathematical Sciences and Applications-13, GJMS*, **2** (2) (2013), 25-35.

[10] Dhage, J. B., Dhage, B. C., Approximating local solution of an IVP of nonlinear first order ordinary hybrid differential equations, *Nonlinear Studies* **30** (3) (2023), 721-732.

[11] Dhage, J. B., Dhage, B. C., Approximating local solution of IVPs of nonlinear first order ordinary hybrid intgrodifferential equations, *Malaya J. Mat.*, **11** (04) (2023), 344-355.

[12] Dhage, J. B., Dhage, S. B., Dhage, B. C., An algorithmic approach to local solution of the nonlinear higher order ordinary hybrid differential equations, *Jñānābha* **54** (1) (2024), 29-40.

[13] Dhage, J. B., Dhage, S. B., Dhage, B. C., Approximation results for local solution of the initial value problems of nonlinear first order ordinary hybrid functional integrodifferential equations, *Dyn. Continuous, Descrete and Impulsive Systems, Series B*: **32** (2025), 369-386.

[14] Dhage, J. B., Dhage, B. C., Approximation, existence and uniqueness of the integrable local solution of nonlinear hybrid functional integral equations, *J. Ramanujan Soc. Math. & Math. Sci.*, **13** (1) (2025), 19-38.

[15] Dhage, J. B., Dhage, B. C., On local monotonic analysis of the IVPs of nonlinear Caputo fractional ordinary hybrid differential equations, *Carpathian J. Math.*, **42** 94) (2026), 853-869.

[16] Dhage, J. B., Dhage, B. C., On monotonic dependence of the local solution of nonlinear hybrid differential equations on initial data, *J. Appl. & Pure Math.*, **8** (1-2) (2026), 37-44.

[17] Dhage, J. B., Dhage, B. C., Local convergence analysis of initial value problems of nonlinear first order ordinary hybrid functional differential equations, *An. Stint. Univ. Al. I. Cuza Iasi, Mat.(N.S.)*, LXXII (2026), 183-203.

[18] Dhage, B. C., Graef, J. R., Dhage, S. B., Existence, stability and global attractivity results for nonlinear Riemann-Liouville fractional differential equtions, *Cubo, A Mathematical Journal*, **25** (1) (2023), 23-36.

[19] Dhage, S. B., Dhage, B. C., Graef, J. R., Dhage iteration method for initial value problem for nonlinear first order integrodifferential equations, *J. Fixed Point Theory Appl.* **18** (2016), 309-325.

[20] Diethelm, K., The Analysis of Fractional Differential Equations. An Application-Oriented Exposition Using Differential Operators of Caputo Type, *Lecture Notes in Mathematics, Springer, Berlin*, 2010.

[21] Granas, A., Dugundji, J., *Fixed Point Theory*, Springer 2003.

[22] Guo, D., Lakshmikantham, V., *Nonlinear Problems in Abstract Cones*, Academic Press, New York, London 1988.

[23] Hilfer, R., *Applications of Fractional Calculus in Physics*, World Scientific, Singapore, 2000.

[24] Huang, J., Jung, S., Li, Y., On Hyers-Ulam stability of nonlinear differential equations, *Bull. Korean Math. Soc.* **52** (2) (2015), 685-697.

[25] Lakshmikantham, V., Leela, S., Vasundhara Devi, J., *Theory of Fractional Dynamic Systems*, Cambridge Scientific Publishers, Colterham, 2009.

[26] Kilbas, A. A., Srivastava, H. M., Trujillo, J.J., *Theory and Applications of Fractional Differential Equations*, North-Holland Mathematics Studies, 204. Elsevier Science B.V., Amsterdam, 2006.

[27] Miller, K. S., Ross, B., *An Introduction to the Fractional Calculus and Fractional Differential Equations*, Wiley, New York, 1993.

[28] Podlubny, I., *Fractional differential equations*, Academic Press, (1999).

[29] Ran A. C. M., Reurings, M. C. B., A fixed point theorem in partially ordered sets and some applications to matrix equations, *Proc. Amer. Math. Soc.*, **132** (5) (2004), 1435-1443.

[30] Tripathy, A. K., *Hyers-Ulam stability of ordinary differential equations*, Chapman and Hall / CRC, London, NY, 2021.

Janhavi B. Dhage, Bapurao C. Dhage
Kasubai, Gurukul Colony,
Thodga Road, Ahmedpur,
Ahmedpur-413 515, Dist. Latur,
Maharashtra, India.
E-mail: jbdhage@gmail.com
E-mail: bcdhage@gmail.com

The Mathematics Student ISSN: 0025-5742
Vol. 95, Nos. 3-4, July - December (2026), 242–249

AN ELEMENTARY PROOF OF QUESTION 359 OF RAMANUJAN

SURAJIT RAJAGOPAL

(Received : 20 - 04 - 2026 ; Accepted : 14 - 11 - 2026)

Abstract. This paper presents a proof for Question 359 posed by Ramanujan in the Journal of the Indian Mathematical Society. The problem requires establishing a specific trigonometric identity connecting the trigonometric functions of three variables, x, y and z, given the conditions $\sin(x+y) = 2\sin(\frac{x-y}{2})$ and $\sin(y+z) = 2\sin(\frac{y-z}{2})$. We employ algebraic substitutions to transform the given trigonometric constraints into a polynomial framework. We demonstrate that $\tan\frac{x}{2}$ and $\tan\frac{z}{2}$ are roots of a specific quartic polynomial. By analyzing symmetric functions of the roots of this polynomial and deriving key relationships between the polynomial coefficients and the target trigonometric expressions, we prove the validity of Ramanujan's claim for all real cases.

1. Introduction

Between 1911 and 1919, Srinivasa Ramanujan submitted a sequence of problems to the Journal of the Indian Mathematical Society. These problems, fifty-eight in total, range across number theory, continued fractions, infinite series, and trigonometric identities, and they continue to reward study today — both for the depth of the mathematics they conceal behind deceptively simple statements and for the ingenuity they invite from those who attempt them. A comprehensive survey of these problems, together with solutions and historical commentary, is given by Berndt, Choi, and Kang [1].

The present paper is concerned with Question 359, which appears on page 78 of Volume 4 of the Journal (see [1]). Ramanujan asks the reader to

2020 Mathematics Subject Classification: 33B10, 12E10
Key words and phrases: JIMS Question 359, quartic polynomials, trigonometric identities

© *Indian Mathematical Society, 2026.*

establish the following: if

$$\sin(x+y) = 2\sin(\frac{1}{2}(x-y))$$

and

$$\sin(y+z) = 2\sin(\frac{1}{2}(y-z)),$$

prove that

$$(\frac{1}{2}\sin x\cos z)^{1/4} + (\frac{1}{2}\sin z\cos x)^{1/4} = (\sin(2y))^{1/12}.$$

In this identity, the left-hand side is a sum of fourth roots in x and z, while the right-hand side is a twelfth root in y alone. There is no visible reason, from the hypotheses, why such a precise relationship among fractional powers should hold — and yet Ramanujan asserts that it does, whenever the two constraints linking x, y and z are satisfied. Recovering the identity from the hypotheses is therefore an exercise in extracting hidden algebraic structure from the initial constraints.

Existing treatments of Question 359 in the literature tend to rely on manipulations that, while correct, obscure why the identity takes the form it does. Sometimes topics in higher mathematics such as modular equations are used to prove the above identity [2]. Our aim here is different: we give a proof that uses only the Weierstrass half-angle substitution and elementary properties of polynomials — specifically, Vieta's formulas applied to a quartic. No tool beyond a standard undergraduate curriculum is required.

The method can be summarized as follows. By setting $u = \tan(x/2)$, $a = \tan(y/2)$, and $v = \tan(z/2)$, the two trigonometric constraints transform into a single polynomial condition: both u and v are roots of the quartic $P(x) = x^4 - 4a^3x^3 + 6a^2x^2 - 4ax + a^4$. This reduction is the heart of the argument. Once it is in hand, the symmetric functions $s = u+v$ and $t = uv$ can be related to the remaining two roots of P through Vieta's formulas, and a pair of polynomial identities in s, t, and a emerges. From these, after a careful case analysis that isolates the degenerate configurations $t^2 = a^4$ and $t = 0$, the target identity follows.

The paper is organized around five claims. Claim 1 establishes a relation between $\tan(x/2)$ and $\tan(z/2)$ and a certain quartic polynomial. Claim 2 expresses $\sin(\frac{1}{2}(x+y))$, $\cos(\frac{1}{2}(x+y))$, and their analogues in y and z, as fourth roots of products of sines and cosines — the source of the fourth-root exponents in the identity. Claims 3 and 4 connect the symmetric functions

of the roots to $\sin(\frac{1}{2}(x+z)+y)$ and, through it, to $(\sin(2y))^{1/3}$. Claim 5 assembles these ingredients into Ramanujan's identity. The final three sections handle the degenerate cases that fall outside the generic argument.

2. The Proof

Let x, y and z be real numbers such that

$$\sin(x+y) = 2\sin(\frac{1}{2}(x-y))$$

and

$$\sin(y+z) = 2\sin(\frac{1}{2}(y-z)).$$

Assume $\tan(x/2)$, $\tan(y/2)$ and $\tan(z/2)$ are well-defined (we will handle the cases when x, y or z is an odd multiple of π later). Let $u = \tan(x/2)$, $a = \tan(y/2)$ and $v = \tan(z/2)$. We claim:

Claim 1. *If $P(x) = x^4 - 4a^3x^3 + 6a^2x^2 - 4ax + a^4$, then u and v are both roots of $P(x)$.*

Proof. Note that $\sin\frac{x}{2} = \frac{u}{\sqrt{1+u^2}}$ and $\sin\frac{y}{2} = \frac{a}{\sqrt{1+a^2}}$, and similarly $\cos\frac{x}{2} = \frac{1}{\sqrt{1+u^2}}$ and $\cos\frac{y}{2} = \frac{1}{\sqrt{1+a^2}}$. As $\sin(x+y) = 2\sin\frac{1}{2}(x-y)$ we see

$$\sin\frac{x+y}{2}\cos\frac{x+y}{2} = \sin\frac{x-y}{2}.$$

Writing $\sin(\frac{1}{2}(x+y))$ as $\sin\frac{x}{2}\cos\frac{y}{2} + \sin\frac{y}{2}\cos\frac{x}{2}$, substituting $\sin\frac{x}{2} = \frac{u}{\sqrt{1+u^2}}$ (and similarly for $\cos(\frac{1}{2}(x+y))$) yields

$$\frac{(u+a)(1-au)}{(1+u^2)(1+a^2)} = \frac{(u-a)}{\sqrt{(1+u^2)(1+a^2)}}.$$

We deduce

$$(u+a)(1-au) = \sqrt{(1+u^2)(1+a^2)}(u-a).$$

Squaring both sides and simplifying yields $P(u) = 0$. Similarly, one can deduce $P(v) = 0$. □

Claim 2. *We claim $(\sin(\frac{1}{2}(x+y)))^4 = \sin x \sin y$ and $(\cos(\frac{1}{2}(x+y)))^4 = \cos x \cos y$ and $(\sin(\frac{1}{2}(y+z)))^4 = \sin y \sin z$ and $(\cos(\frac{1}{2}(y+z)))^4 = \cos y \cos z$.*

Proof. Note that the equation $P(u) = 0$ can be re-written as

$$(u+a)^4 = 4au(1+u^2)(1+a^2).$$

Substituting $u = \tan\frac{x}{2} = \frac{\sin\frac{x}{2}}{\cos\frac{x}{2}}$ and $a = \tan\frac{y}{2} = \frac{\sin\frac{y}{2}}{\cos\frac{y}{2}}$ in this equation, we see

$$(\sin(\frac{1}{2}(x+y)))^4 = \sin x \sin y. \tag{2.1}$$

The equality $P(u) = 0$ can also be re-written as

$$(u-a)^4 = 4au(1-u^2)(1-a^2).$$

From this equation, substituting $u = \frac{\sin\frac{x}{2}}{\cos\frac{x}{2}}$ and $a = \frac{\sin\frac{y}{2}}{\cos\frac{y}{2}}$ we get

$$(\sin(\frac{1}{2}(x-y)))^4 = \sin x \sin y \cos x \cos y.$$

As $\sin(\frac{1}{2}(x+y))\cos(\frac{1}{2}(x+y)) = \sin(\frac{1}{2}(x-y))$ and $(\sin(\frac{1}{2}(x+y)))^4 = \sin x \sin y$ from (2.1), we deduce $(\cos(\frac{1}{2}(x+y)))^4 = \cos x \cos y$.

We can do a similar analysis for $(\sin(\frac{1}{2}(y+z)))^4$ and $(\cos(\frac{1}{2}(y+z)))^4$, which yields the results given above. □

In the following theorem, we derive some equations giving relations between s (defined as $u+v$) and t (defined as uv) which will be useful in our proof.

Theorem 2.1. *Let $s = u+v$ and $t = uv$. Then*

$$(4a^3 - s)s + t + \frac{a^4}{t} = 6a^2 \tag{2.2}$$

and

$$s(a^4 - t^2) = 4at - 4a^3t^2 \tag{2.3}$$

Proof. Consider the quartic polynomial $P(x) = x^4 - 4a^3x^3 + 6a^2x^2 - 4ax + a^4$ (defined before). By Claim 1, we know u and v are roots of $P(x)$. Labeling the other two roots as r_1 and r_2, from Vieta's formulae, we have

$$a^4 = r_1 r_2 t,$$

$$4a = t(r_1 + r_2) + r_1 r_2 s,$$

$$6a^2 = t + r_1 r_2 + s(r_1 + r_2),$$

$$4a^3 = s + r_1 + r_2.$$

From these facts we deduce (2.2) and (2.3). □

Now we split our proof into cases. When $t^2 \neq a^4$, one can prove the following.

3. Case 1: When $t^2 \neq a^4$ and $t \neq 0$

We deduce from (2.3) that

$$s = 4at\frac{(1-a^2t)}{(a^4-t^2)}. \tag{3.1}$$

Substituting this into (2.2) and simplifying, we get $\frac{Q(t)}{t(a^4-t^2)^2} = 0$, where $Q(t) = a^{12} - 16a^{10}t^3 - 6a^{10}t + 15a^8t^2 + 12a^6t^3 + 15a^4t^4 - 6a^2t^5 - 16a^2t^3 + t^6$. It follows that

$$a^{12}-16a^{10}t^3-6a^{10}t+15a^8t^2+12a^6t^3+15a^4t^4-6a^2t^5-16a^2t^3+t^6 = 0 \tag{3.2}$$

Let

$$k = (16a^2(1-a^4)^2)^{1/3}. \tag{3.3}$$

Note that $Q(t) = (t-a^2)^6 - k^3t^3$, so $Q(t) = 0$ implies

$$(t-a^2)^2 = kt. \tag{3.4}$$

We use the above equation to prove our next claim.

Claim 3. *We claim* $(\frac{s(1-a^2)+2a(1-t)}{\sqrt{s^2+(1-t)^2}})^2 = k$.

Proof. From (3.4), for establishing the claim, it is sufficient to prove

$$t(\frac{s(1-a^2)+2a(1-t)}{\sqrt{s^2+(1-t)^2}})^2 = (t-a^2)^2.$$

Considering the expression $t[s(1-a^2)+2a(1-t)]^2-(t-a^2)^2[s^2+(1-t)^2]$ and substituting our known expression for s from (3.1), that is $s = \frac{4at(1-a^2t)}{a^4-t^2}$, we see

$$t[s(1-a^2)+2a(1-t)]^2 - (t-a^2)^2[s^2+(1-t)^2] = \frac{-(t-1)^2Q(t)}{(a^4-t^2)^2}.$$

As $Q(t) = 0$ from (3.2), it follows that

$$t(\frac{s(1-a^2)+2a(1-t)}{\sqrt{s^2+(1-t)^2}})^2 = (t-a^2)^2.$$

This concludes the proof of the above claim. □

Claim 4. *We claim* $(\sin(2y))^{1/3} = \sin(\frac{1}{2}(x+z)+y)$.

Proof. Note that taking square roots on both sides of Claim 3, and using the fact that

$$\frac{(t-a^2)^2}{t} = k \tag{3.5}$$

we have

$$\frac{s(1-a^2)+2a(1-t)}{\sqrt{s^2+(1-t)^2}}=\sqrt{k}.$$

Recall $a=\tan(\frac{y}{2})$, so we see $\sin(2y)=\frac{4a(1-a^2)}{(1+a^2)^2}$, so $(\sin(2y))^{1/3}=\frac{\sqrt{k}}{1+a^2}$ and

$$\frac{1}{1+a^2}\frac{s(1-a^2)+2a(1-t)}{\sqrt{s^2+(1-t)^2}}=\sin(\frac{1}{2}(x+z)+y).$$

This proves our claim. □

Claim 5. *We claim* $(\sin(2y))^{1/12}=(\frac{1}{2}\sin x\cos z)^{1/4}+(\frac{1}{2}\sin z\cos x)^{1/4}$.

Proof. From Claim 4 above,

$$(\sin(2y))^{1/3}=\sin(\frac{1}{2}(x+y))\cos(\frac{1}{2}(y+z))+\sin(\frac{1}{2}(y+z))\cos(\frac{1}{2}(x+y)). \quad (3.6)$$

Using Claim 2,

$$\sin(\tfrac{1}{2}(x+y))=(\sin x\sin y)^{1/4},\quad \cos(\tfrac{1}{2}(x+y))=(\cos x\cos y)^{1/4},$$
$$\sin(\tfrac{1}{2}(y+z))=(\sin y\sin z)^{1/4},\quad \cos(\tfrac{1}{2}(y+z))=(\cos y\cos z)^{1/4},$$

and substituting these into (3.6), we see

$$[(\frac{1}{2}\sin x\cos z)^{1/4}+(\frac{1}{2}\sin z\cos x)^{1/4}]\cdot(2\sin y\cos y)^{1/4}=(\sin(2y))^{1/3}. \quad (3.7)$$

This implies $(\sin(2y))^{1/12}=(\frac{1}{2}\sin x\cos z)^{1/4}+(\frac{1}{2}\sin z\cos x)^{1/4}$.

Thus we have proven Ramanujan's claim, however, in the above proof, we assumed none of x, y and z are odd multiples of pi, and also assumed $t^2\neq a^4$ and $t\neq 0$. These cases will be dealt with in the next few sections of the paper. □

4. Case 2: When $t^2=a^4$ and $t\neq 0$

Case 2(a): When $t=a^2$ and $t\neq 0$. Note $t^2=a^4$ implies $t=a^2$ or $t=-a^2$. In the first case, let $t=a^2$. From (2.3),

$$s(a^4-t^2)=4at-4a^3t^2.$$

So

$$0=4at-4a^3t^2.$$

As $t\neq 0$ and $a^4=t^2$, we see $a\neq 0$; thus we have $t=\frac{1}{a^2}$.

So $a^2=\frac{1}{a^2}$, or $a^4=1$. Given a is real, we see $a=1$ or $a=-1$. If $a=1$, as $P(u)=0$ we see $u^4-4u^3+6u^2-4u+1=0$. So $(u-1)^4=0$, so $u=1$ and $v=1$ similarly. So, in this case $s=2$, $t=1$ and $a=1$.

We needed to show $\frac{s(1-a^2)+2a(1-t)}{\sqrt{s^2+(1-t)^2}} = \sqrt{k}$. But this is true as $s(1-a^2)+2a(1-t)=0$ and $\sqrt{k} = \sqrt{(16a^2(1-a^4)^2)^{1/3}} = 0$. If $a=-1$, then $u^4+4u^3+6u^2+4u+1=0$ so $u=v=-1$ (so $t=uv=1$, as before) and we have a similar analysis as above.

Case 2(b): When $t=-a^2$ and $t\neq 0$. If $t=-a^2$ then (from equation (2.3)), we see $4at-4a^3t^2=0$. Noting $t\neq 0$, this means $t=\frac{1}{a^2}$ so $a^4=-1$ and thus a is not real. This is a contradiction as we defined $a=\tan(\frac{y}{2})$, which must be a real number when y is real and not an odd multiple of π.

5. Case 3: When $t=0$

Consider $P(x)=x^4-4a^3x^3+6a^2x^2-4ax+a^4$. By Claim 1, we know u and v are roots of $P(x)$. By assumption $uv=t=0$. If r_1 and r_2 are the other roots (of P), we see $tr_1r_2=a^4$ so $a=0$, and thus $P(x)=x^4$. As u and v are roots of P, $u=v=0$ and thus $s=0$ and $t=0$. As $s=0$, note $\frac{s(1-a^2)+2a(1-t)}{\sqrt{s^2+(1-t)^2}} = \sqrt{k}$ as $k=0$ and $s(1-a^2)+2a(1-t)=0+0=0$.

So, we have proved $\frac{s(1-a^2)+2a(1-t)}{\sqrt{s^2+(1-t)^2}} = \sqrt{k}$ in the case $t=0$. This concludes the proof of Case 3.

6. Case 4: When x or y or z is an odd multiple of π

Let $x=(2n+1)\pi$, for an integer n. The proof we gave in Case 1 would not work as $u=\tan(x/2)$ is undefined. For this case, by the 2π-periodicity of sine, we may consider only the cases when $x=\pi$, or $x=3\pi$.

When $x=\pi$, the condition $\sin(x+y)=2\sin(\frac{1}{2}(x-y))$ yields

$$-\sin y = 2\cos(\frac{y}{2}). \tag{6.1}$$

So, from (6.1), $\cos(\frac{y}{2})=0$ or $\sin(\frac{y}{2})=-1$. So, in either case, y is an odd multiple of π. A similar analysis holds when $x=3\pi$.

Consider the equation

$$\sin(y+z)=2\sin(\frac{1}{2}(y-z)). \tag{6.2}$$

From the above analysis, y must be an odd multiple of π. Applying exactly the same logic as above to the equation $\sin(y+z)=2\sin(\frac{1}{2}(y-z))$, we get that z must be an odd multiple of π.

We intended to show $(\frac{1}{2}\sin x\cos z)^{1/4}+(\frac{1}{2}\sin z\cos x)^{1/4}=(\sin(2y))^{1/12}$. But this is true as the left-hand side is $0+0$ or 0, as both x and z are odd

multiples of π so $\sin x$ and $\sin z$ are 0 — and the right-hand side is 0 as $\sin(2y) = 0$ when y is an integer multiple of π. This concludes our proof.

Acknowledgement: We are grateful to the referee for the comments which improved the quality of the paper.

References

[1] Berndt, B. C., Choi, Y.-S. and Kang, S.-Y., The problems submitted by Ramanujan to the Journal of the Indian Mathematical Society, in: *Ramanujan: Essays and Surveys*, History of Mathematics **22** (2001), 215–258.

[2] Piezas III, T., On Ramanujan's Question 359, *Mathematics Stack Exchange*, https://math.stackexchange.com/questions/310026 (accessed 17-04-2026).

Surajit Rajagopal
401 Dev Arti Building
Mahim 400016
Mumbai, INDIA.
E-mail: surajitrajagopal@gmail.com

The Mathematics Student ISSN: 0025-5742
Vol. 95, Nos. 3-4, July - December (2026), 250–273

MODULARITY IN MATHEMATICS AND PHYSICS: FROM MODULAR FORMS TO RATIONAL CONFORMAL FIELD THEORY

ARPIT DAS AND JISHU DAS

(Received : 28- 08 - 2026 ; Accepted : 28 - 10 - 2026)

Abstract. Modular symmetry provides a striking bridge between number theory and two-dimensional physics. We begin with its arithmetic and geometric foundations: the modular group, complex tori, modular forms, Eisenstein series, the discriminant, the Dedekind eta function, and Hecke operators. We then explain how scale and conformal invariance, finite temperature, and torus geometry produce modular symmetry in two-dimensional conformal field theory. Finally, we show how characters of rational conformal field theories form vector-valued modular objects and how modular constraints organise spectra and support classification through modular linear differential equations. The presentation is expository and self-contained.

1. Introduction

Modularity is a rare idea whose mathematical formulation and physical meaning illuminate one another. In number theory it governs highly constrained holomorphic functions, their Fourier coefficients, and their associated arithmetic data. In two-dimensional physics it expresses the fact that a torus can be described by different choices of fundamental cycles without changing the underlying system. The same fractional linear transformations therefore appear both in the theory of modular forms and in finite-temperature conformal field theory.

This connection has consequences in both directions. On the mathematical side, modular transformation laws turn analytic functions into

2020 Mathematics Subject Classification: 11F11, 81T40
Key words and phrases: modular forms; modular invariance; complex tori; conformal field theory; rational conformal field theory.

© *Indian Mathematical Society, 2026.*

rigid arithmetic objects; Eisenstein series, the modular discriminant, the Dedekind eta function, and Hecke operators provide central examples. On the physical side, modular invariance relates low- and high-temperature regimes, constrains the spectrum of a conformal field theory, and underlies Cardy's asymptotic formula for the density of states [1, 5]. In a rational conformal field theory, the finitely many characters transform together as a vector-valued modular object, making modularity a practical tool for organising and classifying candidate theories [6, 8].

Our purpose is expository. The individual results reviewed below are standard; the aim is to provide a single, student-friendly route from the arithmetic definition of a modular form to its geometric origin and its use in conformal field theory. We keep the prerequisites modest, spell out the complex-analytic and thermal constructions, and repeatedly connect formal transformation laws with the quantities they constrain.

The order is deliberately mathematics first. Section 2 gives the historical and arithmetic background. Section 3 introduces modular forms, Hecke operators, and the principal examples needed later. Section 4 explains complex tori and the modular group geometrically. We then turn to physics: section 5 develops conformal symmetry and the thermal torus, while section 6 studies the torus partition function and the compact boson. Section 7 closes the main discussion with characters and modular constraints in rational conformal field theory, and section 8 gathers the two viewpoints into one conclusion.

2. Historical development of modular forms

The theory of modular forms is one of the most prominent meeting points of analysis, algebra, geometry, and number theory. What makes the subject especially striking is that it originates from very concrete questions about infinite series and special functions, yet leads naturally to some of the deepest results in modern mathematics.

Historically, early examples of what we now recognise as modular forms appeared in the nineteenth century through the study of elliptic functions and q-series. In his *Fundamenta Nova Theoriae Functionum Ellipticarum*

(1829), Jacobi introduced powerful identities involving infinite products and series. A celebrated example is the following triple-product identity.

Theorem 2.1 (Jacobi). *If $|q| < 1$ and $x \neq 0$, then*

$$\prod_{n=0}^{\infty}(1-q^{2n+2})(1+xq^{2n+1})\left(1+\frac{q^{2n+1}}{x}\right) = \sum_{k=-\infty}^{\infty} q^{k^2}x^k. \tag{2.1}$$

Rename the nome in the identity above as Q. Setting $Q = t^3$ and $x = -t^{-1}$, and then writing $q = t^2$, gives Euler's pentagonal number theorem:

$$\prod_{n=1}^{\infty}(1-q^n) = \sum_{k=-\infty}^{\infty}(-1)^k q^{\frac{k(3k-1)}{2}}. \tag{2.2}$$

These analytic formulae later revealed unexpected arithmetic content, including formulae for representations of integers as sums of squares.

During the nineteenth century, Eisenstein, Dedekind, and Kronecker developed these ideas further. Eisenstein series provided explicit functions with remarkable symmetry properties, while Dedekind's η-function suggested a deep connection between infinite products and modular transformations. Despite this progress, these developments remained somewhat fragmented, lacking a conceptual framework.

A decisive shift occurred with the work of Srinivasa Ramanujan. In his 1916 paper [12], he introduced the arithmetic function $\tau(n)$ through the identity

$$q\prod_{n=1}^{\infty}(1-q^n)^{24} = \sum_{n=1}^{\infty}\tau(n)q^n. \tag{2.3}$$

The coefficients $\tau(n)$ display remarkable arithmetic structure. Based on extensive numerical calculations, Ramanujan conjectured that they are multiplicative, satisfy explicit prime-power recurrences, and obey strong bounds. More precisely, he conjectured:

(1) $\tau(mn) = \tau(m)\tau(n)$ whenever $\gcd(m,n) = 1$.
(2) If p is prime and a is a non-negative integer, then

$$\tau(p^{a+2}) = \tau(p)\tau(p^{a+1}) - p^{11}\tau(p^a). \tag{2.4}$$

(3) If p is prime, then $|\tau(p)| \leq 2p^{11/2}$.

At the time, there was no clear reason why the coefficients of such an innocent-looking power series should possess such rigid arithmetic structure.

The explanation for these phenomena began to emerge in the 1930s through the work of Hecke [7]. Hecke developed a systematic theory of modular forms and introduced the operators that now bear his name. Modular forms are holomorphic functions on the upper half-plane that satisfy precise transformation laws under the action of $\mathrm{SL}_2(\mathbb{Z})$. By defining the linear operators now called Hecke operators, Hecke showed that Ramanujan's multiplicativity and prime-power recurrence arise naturally. The deepest of these, concerning bounds on the size of $\tau(p)$ for prime p, resisted proof for decades. Its eventual resolution by Deligne [4] in 1974 relied on ideas from algebraic geometry, specifically the theory of étale cohomology and the Weil conjectures. This marked a turning point in the subject, revealing that modular forms are intimately connected to geometry and Galois representations.

Following Hecke's foundational work, the theory of modular forms expanded in multiple directions. Maass discovered real-analytic analogues, now known as Maass forms, which play an important role in spectral theory. Rankin and Selberg developed methods to associate Dirichlet series and L-functions to modular forms, laying the groundwork for systematic analytic study. In parallel, Harish-Chandra and Langlands reformulated modular forms within the broader setting of automorphic representations, leading to the Langlands programme—a vast network of conjectures linking number theory, harmonic analysis, and representation theory.

One of the most dramatic applications of modular forms arose from their connection to elliptic curves. Frey's insight suggested that a hypothetical solution of the Fermat equation would produce an elliptic curve incompatible with the expected modularity properties. Serre made this link precise, and Ribet proved the required epsilon conjecture in 1986. Fermat's Last Theorem was thereby reduced to the semistable case of the Modularity Conjecture. Wiles's proof of that semistable case, completed with Taylor, established the result needed for Fermat's Last Theorem [2, 15] and confirmed the central role of modular forms in arithmetic geometry.

Today, modular forms appear throughout mathematics and theoretical physics. They play a fundamental role in the study of elliptic curves, Galois representations, L-functions, partitions, coding theory, and even string theory. Despite their sophisticated applications, the basic definitions rely on relatively elementary ingredients: complex analysis, group actions, and Fourier series.

The historical survey explains why modular forms became important. We now give their analytic definition and principal examples; section 4 will then supply the geometric reason for the modular group's fractional linear action.

3. Modular forms and arithmetic q-series

In this section, we define modular forms for $\mathrm{SL}_2(\mathbb{Z})$ and discuss the elementary facts needed later. We start by defining the upper half-plane and the modular group.

3.1. **The upper half-plane and the modular group.** Let

$$\mathbb{H} = \{\tau \in \mathbb{C} : \mathrm{Im}(\tau) > 0\} \tag{3.1}$$

denote the complex upper half-plane. This is the natural domain for the modular forms defined below. The variable $q = \mathrm{e}^{2\pi \mathrm{i}\tau}$ satisfies $|q| < 1$ for $\tau \in \mathbb{H}$. A holomorphic function on $\mathbb{H}$ that is invariant under $\tau \mapsto \tau + 1$ descends to the punctured unit disc $0 < |q| < 1$ and has a Laurent expansion in q; holomorphy at the cusp removes its negative powers. The point $q = 0$ represents the cusp at infinity.

The *modular group* is defined as

$$\mathrm{SL}_2(\mathbb{Z}) = \left\{ \begin{pmatrix} a & b \\ c & d \end{pmatrix} : a, b, c, d \in \mathbb{Z},\ ad - bc = 1 \right\}. \tag{3.2}$$

It acts on $\mathbb{H}$ by the fractional linear transformation

$$\tau \mapsto \frac{a\tau + b}{c\tau + d}. \tag{3.3}$$

The group $\mathrm{SL}_2(\mathbb{Z})$ is generated by the two elements

$$S = \begin{pmatrix} 0 & -1 \\ 1 & 0 \end{pmatrix}, \qquad T = \begin{pmatrix} 1 & 1 \\ 0 & 1 \end{pmatrix}, \tag{3.4}$$

corresponding to the transformations $\tau \mapsto -1/\tau$ and $\tau \mapsto \tau + 1$, respectively. Since I and $-I$ act identically on τ, the faithful action is by

$\mathrm{PSL}_2(\mathbb{Z}) = \mathrm{SL}_2(\mathbb{Z})/\{\pm I\}$. We shall nevertheless follow the common convention of referring to $\mathrm{SL}_2(\mathbb{Z})$ as the modular group. In the following section, we shall derive this group from the geometry of a torus; see also [14].

3.2. **Modular forms for** $\mathrm{SL}_2(\mathbb{Z})$**.**

Definition 3.1. Let k be an even integer. A *modular form of weight k for* $\mathrm{SL}_2(\mathbb{Z})$ is a holomorphic function $f : \mathbb{H} \to \mathbb{C}$ satisfying:

(1) (*Modular transformation*)

$$f\left(\frac{a\tau + b}{c\tau + d}\right) = (c\tau + d)^k f(\tau) \quad \text{for all } \begin{pmatrix} a & b \\ c & d \end{pmatrix} \in \mathrm{SL}_2(\mathbb{Z}); \tag{3.5}$$

(2) (*Holomorphy at infinity*) $f(\tau)$ remains bounded as $\mathrm{Im}(\tau) \to \infty$, equivalently as $q \to 0$.

Since S and T generate $\mathrm{SL}_2(\mathbb{Z})$, the modular transformation property is equivalent to the two conditions $f(\tau+1) = f(\tau)$ and $f(-1/\tau) = \tau^k f(\tau)$. The periodicity condition $f(\tau+1) = f(\tau)$ first gives a Laurent expansion in q on the punctured unit disc. The holomorphy-at-infinity condition removes its negative powers, yielding the Fourier expansion

$$f(\tau) = \sum_{n=0}^{\infty} a(n) q^n, \qquad q = \mathrm{e}^{2\pi \mathrm{i} \tau}. \tag{3.6}$$

If $a(0) = 0$, then f is called a *cusp form*. We now turn to standard examples.

Example 3.2 (Eisenstein series)**.** For even integers $k \geq 4$, the Eisenstein series

$$G_k(\tau) = \sum_{(m,n) \neq (0,0)} \frac{1}{(m\tau + n)^k} = 2\zeta(k) + 2 \sum_{m=1}^{\infty} \sum_{n \in \mathbb{Z}} \frac{1}{(m\tau + n)^k}, \tag{3.7}$$

defines a modular form of weight k. Its normalisation with constant term 1 has Fourier expansion

$$E_k(\tau) := \frac{G_k(\tau)}{2\zeta(k)} = 1 - \frac{2k}{B_k} \sum_{n=1}^{\infty} \left(\sum_{d \mid n} d^{k-1} \right) q^n, \tag{3.8}$$

where B_k denotes the k-th Bernoulli number. One can show that E_4 and E_6 generate the graded ring of modular forms for the full modular group.

For each fixed k, the modular forms of weight k form a finite-dimensional vector space. More precisely, we have the following theorem.

Theorem 3.3. *For an even integer $k \geq 4$, let $M_k(\mathrm{SL}_2(\mathbb{Z}))$ denote the set of modular forms of weight k for $\mathrm{SL}_2(\mathbb{Z})$. We have*

$$M_k(\mathrm{SL}_2(\mathbb{Z})) = \bigoplus_{\substack{a,b\geq 0 \\ 4a+6b=k}} \mathbb{C}\, E_4^a E_6^b, \tag{3.9}$$

where $\bigoplus$ denotes direct sum of vector spaces. In particular, every modular form for $\mathrm{SL}_2(\mathbb{Z})$ can be written uniquely as a polynomial in E_4 and E_6; see [13, Chapter 7, Corollary 2] *for a proof.*

Example 3.4. Consider

$$\Delta(\tau) := \frac{E_4(\tau)^3 - E_6(\tau)^2}{1728}. \tag{3.10}$$

The constant terms in E_4^3 and E_6^2 cancel, so Δ is a cusp form of weight 12. Moreover,

$$\Delta(\tau) = q\prod_{n=1}^{\infty}(1-q^n)^{24} = \sum_{n=1}^{\infty}\tau(n)q^n. \tag{3.11}$$

Thus the coefficients are precisely Ramanujan's arithmetic function. The identity $\Delta = \eta^{24}$ will later connect this modular form to oscillator partition functions.

3.3. **Hecke operators.** For a positive integer m, the Hecke operator T_m is a linear map from $M_k(\mathrm{SL}_2(\mathbb{Z}))$ to itself. In one standard normalisation,

$$(T_m f)(\tau) = m^{k-1}\sum_{\substack{ad=m \\ d>0}} \frac{1}{d^k} \sum_{b \,(\mathrm{mod}\, d)} f\left(\frac{a\tau+b}{d}\right). \tag{3.12}$$

This is a well-defined operator on $M_k(\mathrm{SL}_2(\mathbb{Z}))$ [9, Theorem 5.2.2]. If a normalised simultaneous Hecke eigenform has expansion $f(\tau) = \sum_{n=0}^{\infty} a(n)q^n$ and satisfies $T_n f = \lambda(n) f$, then $a(n) = a(1)\lambda(n)$ [13, Chapter 7, Theorem 7]. The following relations explain Ramanujan's first two conjectures; see [9, Theorems 7.2.1 and 7.2.2].

Theorem 3.5. *Let $m, n \in \mathbb{N}$. Then*

- $T_m T_n = T_{mn}$ *if* $\gcd(m,n) = 1$;
- $T_p T_{p^n} = T_{p^{n+1}} + p^{k-1} T_{p^{n-1}}$, *where p is prime and $n \geq 1$.*

The one-dimensionality of the space of weight-12 cusp forms implies that Δ is a simultaneous Hecke eigenform [13, Chapter 7, Proposition 14]. The two relations above then give the multiplicativity and prime-power recurrence of Ramanujan's τ-function.

3.4. **The Klein j-invariant.** A weight-zero modular function that is holomorphic both on $\mathbb{H}$ and at the cusp must be constant. Non-constant examples arise when holomorphy at the cusp is relaxed while holomorphy on $\mathbb{H}$ is retained.

The basic example is the Klein modular invariant

$$j(\tau) = \frac{E_4(\tau)^3}{\Delta(\tau)}, \tag{3.13}$$

where E_4 is the Eisenstein series of weight 4 and Δ is the weight-12 cusp form introduced above.

3.5. **The Dedekind eta function.** The Dedekind eta function transforms with weight $1/2$ and a multiplier system. It is therefore not an ordinary integral-weight modular form for the full modular group. It is defined by

$$\eta(\tau) = q^{1/24} \prod_{n=1}^{\infty} (1 - q^n), \qquad \eta(\tau)^{24} = \Delta(\tau). \tag{3.14}$$

With the standard branch of the square root, it satisfies

$$\eta(\tau + 1) = \mathrm{e}^{\pi \mathrm{i}/12} \eta(\tau), \qquad \eta(-1/\tau) = \sqrt{-\mathrm{i}\tau}\, \eta(\tau). \tag{3.15}$$

Let $p(n)$ be the number of unrestricted partitions of n into positive integers; for example, $p(1) = 1$, $p(5) = 7$, and $p(7) = 15$. Then

$$\frac{q^{1/24}}{\eta(\tau)} = \sum_{n=0}^{\infty} p(n) q^n. \tag{3.16}$$

Thus the Fourier coefficients of $q^{1/24}/\eta(\tau)$ are the partition numbers. They also imply striking congruences; for instance, $p(7n + 5) \equiv 0 \pmod 7$; see [9, Exercise 5.4.5].

4. Complex tori and the modular group

Having introduced modular forms analytically, we now explain the geometry behind their transformation law. A complex torus is obtained by quotienting the complex plane by a rank-two lattice. Its lattice basis is not unique, and changes of oriented integral basis produce precisely the modular transformations. This also prepares for the later appearance of the same torus as Euclidean spacetime at finite temperature.

4.1. **Quotient topology and complex tori.** Let X be a topological space and let $\sim$ be an equivalence relation on X. For each $x \in X$, its equivalence

class is

$$[x] = \{y \in X \mid y \sim x\}. \tag{4.1}$$

The quotient set is

$$X/\sim := \ \{[x] \mid x \in X\}. \tag{4.2}$$

The natural surjection $\pi : X \to X/\sim$ is given by $\pi(x) = [x]$. The quotient topology on $X/\sim$ is defined by

$$U \subseteq X/\sim \ \text{is open} \iff \pi^{-1}(U) \subseteq X \ \text{is open}. \tag{4.3}$$

The quotient $\mathbb{R}/\mathbb{Z}$ as the circle S^1: Define an equivalence relation on the real numbers as: $x \sim y$ if and only if $x - y \in \mathbb{Z}$. The map

$$\Phi : \mathbb{R}/\mathbb{Z} \to S^1, \qquad \Phi([x]) = \mathrm{e}^{2\pi \mathrm{i} x} \tag{4.4}$$

is well-defined because replacing x by $x + n$, with $n \in \mathbb{Z}$, leaves its value unchanged. It is continuous and bijective; since $\mathbb{R}/\mathbb{Z}$ is compact and S^1 is Hausdorff, it is a homeomorphism.

The complex torus $\mathbb{C}/\Lambda$: Choose $\omega_1, \omega_2 \in \mathbb{C}$ that are linearly independent over $\mathbb{R}$ and define the *lattice*

$$\Lambda = \mathbb{Z}\omega_1 + \mathbb{Z}\omega_2 = \{m\omega_1 + n\omega_2 \mid m, n \in \mathbb{Z}\}. \tag{4.5}$$

Define $z \sim z'$ if and only if $z - z' \in \Lambda$. Topologically, the quotient by any rank-two lattice is a torus:

$$\mathbb{C}/\Lambda \cong \mathbb{R}^2/\Lambda \ \cong \ \mathbb{R}^2/\mathbb{Z}^2 \ \cong \ (\mathbb{R}/\mathbb{Z}) \times (\mathbb{R}/\mathbb{Z}) \ \cong \ S^1 \times S^1 =: \mathbb{T}^2. \tag{4.6}$$

Fundamental parallelogram and opposite-edge gluing: A fundamental domain for the translation action of Λ on $\mathbb{C}$ is

$$P = \{a\omega_1 + b\omega_2 \mid a, b \in [0, 1)\}. \tag{4.7}$$

Every $z \in \mathbb{C}$ is equivalent to exactly one point of the half-open set P. Equivalently, gluing opposite edges of the corresponding closed parallelogram, as illustrated in figure 1, produces the torus $\mathbb{T}^2$.

Area and orientation: Write $\omega_1 = a + \mathrm{i}b$ and $\omega_2 = c + \mathrm{i}d$, where $a, b, c, d \in \mathbb{R}$, and regard them as vectors in $\mathbb{R}^2$. The signed area of the parallelogram is the determinant

$$\mathrm{Area}(\omega_1, \omega_2) = \det \begin{pmatrix} a & c \\ b & d \end{pmatrix} = ad - bc = \mathrm{Im}(\overline{\omega_1}\omega_2), \tag{4.8}$$

where a positive area corresponds to a positively oriented basis.

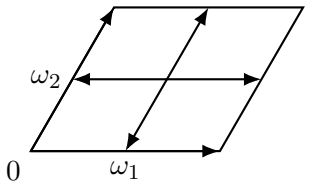

FIGURE 1. Parallelogram fundamental domain for translations by Λ, with opposite edges identified to form $\mathbb{C}/\Lambda$.

4.2. **The modular parameter and the upper half-plane.** If we scale the lattice Λ by a non-zero complex number α, setting $\omega_1' = \alpha\omega_1$ and $\omega_2' = \alpha\omega_2$, then $\Lambda' = \alpha\Lambda$. The map

$$f : \mathbb{C}/\Lambda \to \mathbb{C}/\Lambda', \qquad f([z]) = [\alpha z] \tag{4.9}$$

is well-defined, has inverse $[z] \mapsto [\alpha^{-1}z]$, and is biholomorphic. Hence the two scaled lattices define equivalent complex tori.

Define the *modular parameter* by $\tau = \omega_2/\omega_1$. It is unchanged by the common scaling:

$$\tau' = \frac{\omega_2'}{\omega_1'} = \frac{\alpha\omega_2}{\alpha\omega_1} = \tau. \tag{4.10}$$

Choosing the convenient normalisation $\omega_1 = 1$ gives $\Lambda = \mathbb{Z} + \tau\mathbb{Z}$.

The condition that ω_1, ω_2 are $\mathbb{R}$-linearly independent implies that τ is not real. We choose the orientation so that the area is positive; then $\mathrm{Im}(\tau) > 0$, and hence $\tau \in \mathbb{H}$.

4.3. **Changing a lattice basis.** The lattice does not come with a unique basis. To display the standard fractional linear action, parameterise an oriented change of basis by

$$\begin{pmatrix} \omega_2' \\ \omega_1' \end{pmatrix} = \begin{pmatrix} a & b \\ c & d \end{pmatrix} \begin{pmatrix} \omega_2 \\ \omega_1 \end{pmatrix}, \qquad a, b, c, d \in \mathbb{Z}, \qquad ad - bc = 1. \tag{4.11}$$

An integral matrix with determinant 1 has an integral inverse, so (ω_1', ω_2') generates the same lattice and preserves its orientation. Conversely, every oriented integral basis change is obtained this way. This is the geometric origin of $\mathrm{SL}_2(\mathbb{Z})$.

With $\tau = \omega_2/\omega_1$, the transformed parameter is

$$\tau' = \frac{\omega_2'}{\omega_1'} = \frac{a\omega_2 + b\omega_1}{c\omega_2 + d\omega_1} = \frac{a\tau + b}{c\tau + d}. \tag{4.12}$$

For $\tau \in \mathbb{H}$,

$$\mathrm{Im}\left(\frac{a\tau+b}{c\tau+d}\right) = \frac{\mathrm{Im}(\tau)}{|c\tau+d|^2} > 0. \tag{4.13}$$

Thus $\mathrm{SL}_2(\mathbb{Z})$ preserves the upper half-plane.

4.4. **Generators and the fundamental domain.** The lattice vectors ω_1 and ω_2 determine two oriented cycles on the torus $\mathbb{C}/\Lambda$: each cycle is obtained by projecting the straight segment from 0 to the corresponding lattice vector onto the quotient. In terms of the ordered lattice basis, the generators act as follows:

- T sends (ω_1, ω_2) to $(\omega_1, \omega_2 + \omega_1)$, thereby shearing the cycle basis;
- S sends (ω_1, ω_2) to $(\omega_2, -\omega_1)$, thereby exchanging the two cycles and reversing the orientation of one of them, while preserving the orientation of the torus.

The thermal significance of this cycle exchange will be developed in section 5.

A fundamental domain: Consider the action of $\mathrm{SL}_2(\mathbb{Z})$ on $\mathbb{H}$ by $\tau \mapsto \frac{a\tau+b}{c\tau+d}$. A fundamental domain $\mathcal{F} \subset \mathbb{H}$ contains at least one representative of each orbit. Different points in the *interior* of $\mathcal{F}$ correspond to different complex-torus shapes because they are not related by an $\mathrm{SL}_2(\mathbb{Z})$ transformation. A standard choice is

$$\mathcal{F} = \left\{ \tau \in \mathbb{H} : \ |\tau| \geq 1, \ -\frac{1}{2} \leq \mathrm{Re}(\tau) \leq \frac{1}{2} \right\}, \tag{4.14}$$

where appropriate boundary points are identified by modular transformations [14]. Thus $\mathbb{H}/\mathrm{PSL}_2(\mathbb{Z})$ classifies complex structures on oriented tori. This is a conformal or complex-analytic classification: topologically, all rank-two lattice quotients are homeomorphic to $S^1 \times S^1$.

Up to this point, τ and $q = \mathrm{e}^{2\pi \mathrm{i} \tau}$ have been mathematical variables describing a complex torus and Fourier expansions on its moduli space. We now show that they also have a direct physical meaning. For a two-dimensional quantum system at finite temperature, τ measures the relative sizes and twist of the Euclidean-time and spatial cycles, while powers of q and $\bar{q}$ keep track of the left- and right-moving contributions to the energy.

5. From criticality to the thermal torus

We now turn from the mathematical description of tori to physical systems in which the same geometry arises naturally. In a gapped phase of a

system with short-range interactions, connected correlation functions typically decay exponentially:

$$\langle \mathcal{O}(x)\mathcal{O}(0)\rangle_{\mathrm{c}} \sim \mathrm{e}^{-|x|/\xi}, \tag{5.1}$$

where ξ is the correlation length. At a continuous, or second-order, phase transition, ξ diverges. The long-distance theory then has no distinguished finite length scale, motivating scale invariance under

$$x \mapsto \lambda x, \qquad \lambda > 0. \tag{5.2}$$

Scale-invariant two-point functions therefore have power-law rather than exponential behaviour [10].

5.1. **Power laws from scale invariance.** Assume translation and rotation invariance, so the two-point function of a scalar operator depends only on the distance $r = |x|$:

$$G(r) := \langle \mathcal{O}(x)\mathcal{O}(0)\rangle. \tag{5.3}$$

If $\mathcal{O}$ has scaling dimension Δ, then

$$\mathcal{O}(\lambda x) = \lambda^{-\Delta}\,\mathcal{O}(x), \tag{5.4}$$

which implies

$$G(\lambda r) = \lambda^{-2\Delta} G(r). \tag{5.5}$$

The only solutions are power laws,

$$G(r) \propto \frac{1}{r^{2\Delta}}. \tag{5.6}$$

5.2. **From scale invariance to conformal invariance.** In many, though not all, relativistic quantum field theories, scale invariance can be enhanced to conformal invariance. Conformal transformations preserve angles locally. Scaling is one example; another important example is the logarithmic/exponential map relating the punctured complex plane to a cylinder.

Many continuous critical points are described at long distances by conformal field theories (CFTs), a viewpoint pioneered by Polyakov [10]. A relativistic CFT is invariant under translations, Lorentz transformations, dilatations, and special conformal transformations [11]. In two dimensions the local conformal symmetry is infinite-dimensional, making the resulting constraints exceptionally strong.

5.3. **Conformal maps in two dimensions and the Cauchy–Riemann equations.** Let $f : \mathbb{R}^2 \to \mathbb{R}^2$ be a differentiable map and write

$$(u, v) = f(x, y), \qquad u = u(x, y), \quad v = v(x, y). \tag{5.7}$$

Fix a point $p = (x_0, y_0)$. The key idea is to note that *any differentiable map looks linear to first order* near a point. More precisely, for a small displacement $\delta\mathbf{x} = (\delta x, \delta y)^T$, we have the first-order Taylor expansion

$$f(p + \delta\mathbf{x}) = f(p) + J(p)\,\delta\mathbf{x} + o(\|\delta\mathbf{x}\|), \tag{5.8}$$

where the *Jacobian* (derivative) matrix at p is

$$J(p) = Df(p) = \begin{pmatrix} \frac{\partial u}{\partial x} & \frac{\partial u}{\partial y} \\ \frac{\partial v}{\partial x} & \frac{\partial v}{\partial y} \end{pmatrix}\Bigg|_{(x_0,y_0)}. \tag{5.9}$$

Thus, *infinitesimal vectors* based at p transform by the linear map $J(p)$.

Angles in linear algebra: Given two non-zero vectors $a, b \in \mathbb{R}^2$, the angle $\angle(a, b)$ is determined by the inner product:

$$\cos\angle(a, b) = \frac{\langle a, b\rangle}{\|a\|\,\|b\|}, \tag{5.10}$$

where $\|\cdot\|$ denotes the norm. Assuming $J(p)$ is invertible, we say that f is conformal at p if, for all non-zero tangent vectors a, b,

$$\angle(Ja, Jb) = \angle(a, b) \qquad \text{(to first order, i.e. for infinitesimal vectors).} \tag{5.11}$$

It is orientation-preserving at p if, in addition, $\det J(p) > 0$.

The Jacobian is a similarity transformation: Basic linear algebra gives

$$\begin{aligned} J \text{ preserves angles} &\iff J^TJ = \lambda^2\mathbf{1} \quad \text{for some } \lambda > 0, \\ J \text{ preserves orientation} &\iff \det J > 0, \end{aligned} \tag{5.12}$$

where $\mathbf{1}$ is the 2×2 identity matrix. Indeed, if $J^TJ = \lambda^2\mathbf{1}$, then for any a, b,

$$\langle Ja, Jb\rangle = a^T(J^TJ)b = a^T(\lambda^2\mathbf{1})b = \lambda^2\langle a, b\rangle, \tag{5.13}$$

and also $\|Ja\|^2 = \langle Ja, Ja\rangle = \lambda^2\|a\|^2$. Therefore

$$\frac{\langle Ja, Jb\rangle}{\|Ja\|\,\|Jb\|} = \frac{\lambda^2\langle a, b\rangle}{(\lambda\|a\|)(\lambda\|b\|)} = \frac{\langle a, b\rangle}{\|a\|\,\|b\|}, \tag{5.14}$$

implying that the angles are preserved. Conversely, preservation of the angle between two orthonormal basis vectors makes their images orthogonal, while

preservation of the angle between their sum and difference makes those images equal in length. Hence $J^T J$ is a scalar multiple of the identity; orientation preservation then requires $\det J > 0$.

Equation (5.12) says that locally J is a *uniform scaling* by λ followed by an *orthogonal transformation*. In two dimensions, an orientation-preserving orthogonal map is a rotation. Hence

$$\text{Angle preservation and } \det J > 0 \iff J = \lambda R, \quad \lambda > 0, \quad R \in SO(2), \tag{5.15}$$

where $SO(2)$ is the group of 2×2 orthogonal matrices with unit determinant; it represents two-dimensional rotations.

The rotation matrix: Every $R \in SO(2)$ has the form

$$R = \begin{pmatrix} \cos\theta & -\sin\theta \\ \sin\theta & \cos\theta \end{pmatrix}, \tag{5.16}$$

for some angle $\theta \in [0, 2\pi)$. Plugging this into (5.15) gives

$$J = \begin{pmatrix} \frac{\partial u}{\partial x} & \frac{\partial u}{\partial y} \\ \frac{\partial v}{\partial x} & \frac{\partial v}{\partial y} \end{pmatrix} = \lambda \begin{pmatrix} \cos\theta & -\sin\theta \\ \sin\theta & \cos\theta \end{pmatrix}. \tag{5.17}$$

Equating matrix entries yields the relations

$$\frac{\partial u}{\partial x} = \lambda\cos\theta, \quad \frac{\partial u}{\partial y} = -\lambda\sin\theta; \qquad \frac{\partial v}{\partial x} = \lambda\sin\theta, \quad \frac{\partial v}{\partial y} = \lambda\cos\theta. \tag{5.18}$$

From these, we immediately obtain

$$\frac{\partial u}{\partial x} = \frac{\partial v}{\partial y}, \qquad \frac{\partial u}{\partial y} = -\frac{\partial v}{\partial x}, \tag{5.19}$$

which are exactly the *Cauchy–Riemann equations* at the point p.

Interpreting (5.19), identify (x, y) with $z = x + \mathrm{i}y$ and (u, v) with $w = u + \mathrm{i}v$. The derivative then sends an infinitesimal displacement δz to $\lambda \mathrm{e}^{\mathrm{i}\theta}\delta z$; locally, it scales by λ and rotates by θ. Thus, in two dimensions, orientation-preserving conformal infinitesimal behaviour is equivalent to *complex-linearity*, which is encoded by the Cauchy–Riemann equations. The orientation-reversing case is instead anti-complex-linear.

Orientation-preserving conformal maps in two dimensions are therefore locally holomorphic. This is why complex analysis is the natural language of two-dimensional CFT.

From holomorphic maps to the Virasoro algebra: From above discussion, we see that an infinitesimal local conformal transformation can be

written as

$$z \longmapsto z + \epsilon f(z), \tag{5.20}$$

where ϵ is a small parameter and f is holomorphic. The corresponding vector field is $f(z)\partial_z$, where $\partial_z = \partial/\partial z$. On an annulus centred at $z = 0$, the function f admits a Laurent expansion. This motivates introducing the modes

$$\ell_n = -z^{n+1}\partial_z, \qquad n \in \mathbb{Z}. \tag{5.21}$$

For two such vector fields, their commutator/Lie bracket gives

$$[f(z)\partial_z, g(z)\partial_z] = \big(f(z)g'(z) - g(z)f'(z)\big)\partial_z. \tag{5.22}$$

Thus, we get

$$[\ell_m, \ell_n] = (m-n)\ell_{m+n}, \tag{5.23}$$

where $[A, B] = AB - BA$ denotes the commutator. The finite linear combinations of the generators ℓ_n, with this bracket, form the *Witt algebra*. These modes describe local conformal transformations; the three modes $\ell_{-1}, \ell_0, \ell_1$ generate the globally defined Möbius transformations of the Riemann sphere.

In a quantum conformal field theory, the corresponding symmetry generators act as operators L_n on the space of states. Their commutators may acquire an additional term proportional to the identity operator. To understand why such terms are allowed, recall that multiplying a normalised quantum state vector by an overall phase leaves all measurement probabilities unchanged. Therefore, symmetry operators need only satisfy the group composition law up to a phase,

$$U(g_1)U(g_2) = \mathrm{e}^{\mathrm{i}\omega(g_1,g_2)}U(g_1g_2), \tag{5.24}$$

where g_1, g_2 are symmetry transformations and the phase depends on these transformations, not on the state. Such an action is called a *projective representation*. At the level of infinitesimal generators, this permits additional terms proportional to the identity in their commutators. These are called *central terms*.

For the Witt algebra, the standard non-trivial central extension is the *Virasoro algebra*. In a representation with central charge c, its defining

relations are

$$[L_m, L_n] = (m-n)L_{m+n} + \frac{c}{12} m(m^2-1)\delta_{m+n,0}\,\mathbf{1}, \qquad m, n \in \mathbb{Z}. \quad (5.25)$$

Here $\mathbf{1}$ is the identity operator and $\delta_{m+n,0}$ is the Kronecker delta. The additional term is called *central* because it commutes with every generator; the constant c is a characteristic of the theory. For $c = 0$, these commutation relations reduce to those of the Witt algebra [5].

The anti-holomorphic sector has an analogous Virasoro algebra with generators $\bar{L}_n$ and central charge $\bar{c}$. The two sectors commute:

$$[L_m, \bar{L}_n] = 0. \quad (5.26)$$

The eigenvalues of the zero modes L_0 and $\bar{L}_0$ are the conformal weights h and $\bar{h}$ that appear in the torus partition function below.

5.4. **Compact Euclidean time leads to a thermal theory.** We next study a two-dimensional conformal system at temperature $T > 0$. The torus structure emerges from the Euclidean description of its thermal partition function.

Finite-temperature physics is obtained by averaging over quantum states with Boltzmann weights.

We use Dirac's *bra-ket notation*. A ket $|\psi\rangle$ denotes a vector in the complex Hilbert space $\mathcal{H}$ of quantum states, while the corresponding bra $\langle\psi|$ denotes its adjoint, viewed as a linear functional. In finite-dimensional matrix notation, these are a column vector and its complex-conjugate transpose, respectively. Thus $\langle\phi|\psi\rangle$ denotes their inner product, and

$$\langle\phi|A|\psi\rangle = \langle\phi|\,\big(A|\psi\rangle\big) \quad (5.27)$$

is a matrix element of an operator A. A state is normalised when $\langle\psi|\psi\rangle = 1$; for such a state, $\langle\psi|A|\psi\rangle$ is the expectation value of an observable A.

For a system with Hamiltonian H at temperature T (we set $k_B = \hbar = 1$), the canonical ensemble assigns

$$p_n = \frac{\mathrm{e}^{-\beta E_n}}{Z(\beta)}, \qquad \beta = \frac{1}{T}, \qquad Z(\beta) = \sum_n \mathrm{e}^{-\beta E_n}, \quad (5.28)$$

where $\{|n\rangle\}$ is an orthonormal basis of energy eigenstates, so that $H|n\rangle = E_n|n\rangle$ and $\langle m|n\rangle = \delta_{mn}$, where δ_{mn} is the Kronecker delta. The quantity $Z(\beta)$ is the thermal partition function. The thermal expectation value of

an observable $\mathcal{O}$ is then

$$\langle \mathcal{O} \rangle_\beta = \sum_n p_n \langle n|\mathcal{O}|n\rangle = \frac{1}{Z(\beta)} \sum_n \langle n|\mathrm{e}^{-\beta H}\mathcal{O}|n\rangle. \tag{5.29}$$

The right-hand side is precisely a *trace*:

$$\mathrm{Tr}(A) = \sum_n \langle n|A|n\rangle, \tag{5.30}$$

$$\langle \mathcal{O} \rangle_\beta = \frac{1}{Z(\beta)} \mathrm{Tr}\left(\mathrm{e}^{-\beta H}\mathcal{O}\right), \qquad Z(\beta) = \mathrm{Tr}\left(\mathrm{e}^{-\beta H}\right). \tag{5.31}$$

It is convenient to package these averages into the (thermal) *density matrix*

$$\rho = \frac{\mathrm{e}^{-\beta H}}{Z(\beta)}, \qquad \mathrm{Tr}\ \rho = 1, \tag{5.32}$$

so that $\langle \mathcal{O} \rangle_\beta = \mathrm{Tr}(\rho\, \mathcal{O})$ and $\mathrm{e}^{-\beta H}$ is called the *Boltzmann factor*.

Imaginary time and Wick rotation: Real-time evolution is generated by $U(t) = \mathrm{e}^{-\mathrm{i}tH}$. Formally, the Boltzmann factor is the same operator evaluated at *imaginary* time:

$$\mathrm{e}^{-\beta H} = \mathrm{e}^{-\mathrm{i}H(-\mathrm{i}\beta)}. \tag{5.33}$$

This motivates the Wick rotation $t = -\mathrm{i}\tau_E$, which turns oscillatory amplitudes into decaying weights and converts Lorentzian dynamics into a Euclidean theory.

Thermal partition function as a Euclidean path integral: Let us insert a complete basis of field-configuration states $|\varphi\rangle$ to rewrite the trace:

$$Z(\beta) = \mathrm{Tr}\big(\mathrm{e}^{-\beta H}\big) = \int \mathcal{D}\varphi\, \langle \varphi|\mathrm{e}^{-\beta H}|\varphi\rangle. \tag{5.34}$$

The matrix element admits a path-integral representation in Euclidean time:

$$\langle \varphi|\mathrm{e}^{-\beta H}|\varphi\rangle = \int_{\phi(0)=\varphi}^{\phi(\beta)=\varphi} \mathcal{D}\phi(\tau_E)\, \mathrm{e}^{-S_E[\phi]}, \qquad S_E = \int_0^\beta L_E\, \mathrm{d}\tau_E. \tag{5.35}$$

Here S_E is the Euclidean action. Because the expression is a trace, the initial and final configurations are identified. Equivalently, fields are defined on a circle of circumference β, with $\tau_E \sim \tau_E + \beta$. Bosons are periodic around the thermal circle, whereas fermions are antiperiodic.

5.5. **From the thermal cylinder to the torus.** For a (1+1)-dimensional theory, Wick rotation gives a two-dimensional Euclidean spacetime. If space is also periodic, then $x \sim x+L$ and $\tau_E \sim \tau_E+\beta$, so the Euclidean spacetime

is $S^1_x \times S^1_{\tau_E} \cong \mathbb{T}^2$. Define

$$w = x + \mathrm{i}\tau_E. \tag{5.36}$$

For a rectangular torus, $w \sim w + L$ and $w \sim w + \mathrm{i}\beta$. Introducing the dimensionless coordinate $\widetilde{w} = w/L$ gives

$$\widetilde{w} \sim \widetilde{w} + 1, \qquad \widetilde{w} \sim \widetilde{w} + \tau, \qquad \tau = \mathrm{i}\frac{\beta}{L}. \tag{5.37}$$

5.6. **The physical meaning of modular transformations.** The torus in (5.37) is precisely the complex torus discussed in section 4. A different oriented basis for its two non-contractible cycles changes τ by a modular transformation without changing the underlying geometry. Therefore, the torus partition function of a consistent non-chiral bosonic theory is invariant under this change of description. More generally, chiral quantities and partition functions with a fixed spin structure may transform covariantly or may be permuted among themselves.

For the rectangular torus $\tau = \mathrm{i}\beta/L$, the generator $S : \tau \mapsto -1/\tau$ exchanges the Euclidean-time and spatial cycles:

$$\mathrm{i}\frac{\beta}{L} \longmapsto \mathrm{i}\frac{L}{\beta}. \tag{5.38}$$

It therefore relates the low- and high-temperature regimes. The nome becomes

$$q = \mathrm{e}^{2\pi\mathrm{i}\tau} = \mathrm{e}^{-2\pi\beta/L}, \tag{5.39}$$

and $\bar{q} = q$ for this rectangular torus. Thus the q- and $\bar{q}$-expansions jointly encode the thermal energy levels.

6. Modular invariance in two-dimensional conformal field theory

Equation (5.39) gives the mathematical nome a direct physical interpretation. When β/L is large, q is small and the partition function is dominated by low-energy states. The modular S transformation relates this regime to high temperature. This is the setting in which modularity constrains the spectrum and yields universal asymptotic results such as Cardy's formula [1].

6.1. **The torus partition function.** For a two-dimensional CFT on a torus, the thermal partition function introduced in (5.28) can be written as [5]

$$Z(\tau,\bar\tau) = \mathrm{Tr}_{\mathcal{H}}\Big(q^{L_0-\frac{c}{24}}\bar q^{\bar L_0-\frac{\bar c}{24}}\Big)$$
$$= \sum_{\text{states}} q^{h-\frac{c}{24}}\,\bar q^{\bar h-\frac{\bar c}{24}}, \qquad q=\mathrm{e}^{2\pi\mathrm{i}\tau}, \qquad \bar q=\mathrm{e}^{-2\pi\mathrm{i}\bar\tau}. \tag{6.1}$$

Here h and $\bar h$ are the left- and right-moving conformal weights, while c and $\bar c$ are the corresponding central charges. The cylinder Hamiltonian is proportional to $L_0+\bar L_0-(c+\bar c)/24$.

Changing the oriented basis of torus cycles is a large diffeomorphism and acts on τ through the modular group, as derived in section 4.3. For a consistent non-chiral bosonic CFT without a gravitational anomaly, the full partition function is invariant under this action. Chiral theories and fermionic theories with a specified spin structure require the corresponding modular-covariant or spin-structure-valued formulation. In every case, modular consistency strongly constrains the allowed spectrum.

6.2. **Example: the compact free boson on a torus.** Consider a compact free boson, a scalar field valued on a circle of radius R:

$$X:\mathbb{T}^2 \longrightarrow S^1_R, \qquad X\sim X+2\pi R. \tag{6.2}$$

Its torus partition function factorises into an oscillator contribution and a sum over momentum and winding sectors $(m,n)\in\mathbb{Z}^2$ [5]:

$$Z_{X\in S^1_R}(\tau,\bar\tau) = \frac{1}{|\eta(\tau)|^2}\sum_{m,n\in\mathbb{Z}} q^{\frac{\alpha'}{4}p_L^2}\,\bar q^{\frac{\alpha'}{4}p_R^2}, \tag{6.3}$$

$$p_L = \frac{m}{R}+\frac{nR}{\alpha'}, \qquad p_R = \frac{m}{R}-\frac{nR}{\alpha'}. \tag{6.4}$$

The parameter α' fixes the length convention. The eta function encodes the oscillator modes, whereas the lattice sum records momentum and winding. Thus $q^{1/24}/\eta(\tau)$, the partition generating function in (3.16), counts holomorphic oscillator excitations; after including the vacuum-energy factors and both chiralities, the oscillator contribution is $1/|\eta(\tau)|^2$. At a generic radius, this theory is not rational. With a suitable extended $U(1)$ chiral algebra, it becomes an RCFT when R^2/α' is rational (up to convention-dependent rescalings).

The preceding discussion applies to two-dimensional conformal field theories quite generally. Rational conformal field theories form a particularly rigid subclass: after choosing an appropriate chiral algebra, only finitely

many irreducible representations occur, and modular covariance becomes a finite-dimensional problem.

7. Rational conformal field theories and modular constraints

In two dimensions, conformal symmetry factorises into holomorphic (left-moving) and anti-holomorphic (right-moving) sectors, as seen in (6.1). Often, the chiral symmetry is extended beyond the Virasoro algebra, for example to an affine Lie algebra or a W-algebra [5]. The chosen extension is called the *chiral algebra*.

A theory is called *rational* with respect to this chiral algebra if only finitely many irreducible representations occur. Its torus partition function decomposes into a finite sum of characters. Each character counts descendant states built over a primary state within one conformal family; the vacuum is the primary of the vacuum family.

7.1. **Structure of states.** The Hilbert space decomposes into conformal representations. Each primary state generates a tower of descendants under the action of the chiral algebra. These towers form the conformal families that organise the spectrum, as illustrated in figure 2.

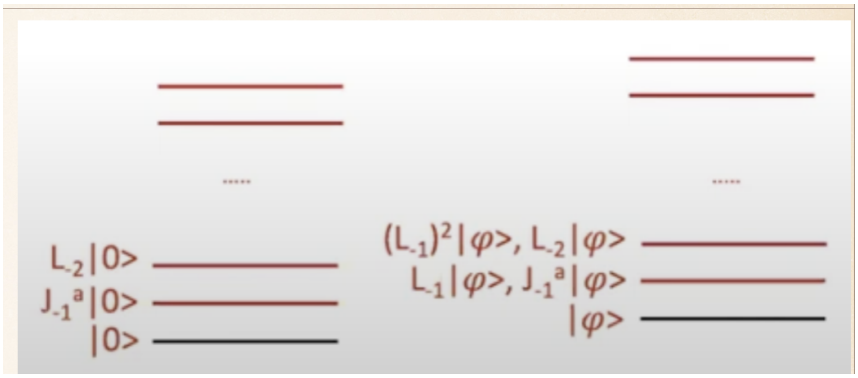

Figure 2. Primary states and their towers of holomorphic descendants. An analogous construction applies to the anti-holomorphic sector.

7.2. **Characters as q-series.** For a conformal family $\mathcal{V}_i$, the character is defined as

$$\chi_i(\tau) = \mathrm{Tr}_{\mathcal{V}_i}\left(q^{L_0 - \frac{c}{24}}\right), \qquad q = \mathrm{e}^{2\pi \mathrm{i}\tau}, \tag{7.1}$$

where L_0 is the holomorphic Virasoro zero mode. On a circle of circumference L, the full Hamiltonian is $H = (2\pi/L)\big(L_0 + \bar{L}_0 - (c+\bar{c})/24\big)$.

The character is a generating function for descendant degeneracies and has the expansion

$$\chi_i(\tau) = q^{h_i - \frac{c}{24}}\big(d_i(0) + d_i(1)q + d_i(2)q^2 + \cdots\big), \tag{7.2}$$

where h_i is the L_0 eigenvalue of the primary state. The coefficient $d_i(n)$ counts states at level n in the conformal family $\mathcal{V}_i$.

7.3. **Partition functions in terms of characters.** Assume for simplicity that the left- and right-moving sectors use the same chiral algebra and character basis, and let $\boldsymbol{\chi}(\tau)$ be the corresponding column vector of characters. Regrouping the state sum (6.1) into chiral representations gives

$$Z(\tau,\bar{\tau}) = \boldsymbol{\chi}(\tau)^\dagger M \boldsymbol{\chi}(\tau), \tag{7.3}$$

where M is a matrix of non-negative integer multiplicities.

7.4. **Modular transformations of characters.** Under the usual rationality and finiteness hypotheses, modular transformations preserve the span of the linearly independent characters, which therefore carries a finite-dimensional representation of $\mathrm{SL}_2(\mathbb{Z})$:

$$\boldsymbol{\chi}(\tau+1) = T\boldsymbol{\chi}(\tau), \qquad \boldsymbol{\chi}(-1/\tau) = S\boldsymbol{\chi}(\tau). \tag{7.4}$$

Modular invariance of the full partition function then implies

$$S^\dagger M S = M, \qquad T^\dagger M T = M. \tag{7.5}$$

For unitary modular matrices, these equations say that M commutes with both S and T.

7.5. **Classification via modular constraints.** Modular symmetry is not merely a consistency condition; it is a classification tool. If an RCFT has n linearly independent characters, their span gives an n-component, weight-zero, weakly holomorphic vector-valued modular object for a finite-dimensional representation of $\mathrm{SL}_2(\mathbb{Z})$ [6, 8]. The qualifier "weakly holomorphic" permits finitely many polar terms at the cusp, including the factor $q^{-c/24}$ in the vacuum character when $c > 0$.

Modular covariance and controlled growth at the cusp $\tau \to \mathrm{i}\infty$ impose strong restrictions on these functions. In the Mathur–Mukhi–Sen (MMS)

classification programme, one exploits the constraints by constructing modular linear differential equations (MLDEs) whose solutions furnish candidate characters. At fixed MLDE order and Wronskian data, requiring a vacuum-type leading exponent and non-negative integral Fourier coefficients can reduce the solutions to a finite list of admissible candidates [3, 8]. These are necessary consistency tests; an admissible solution need not by itself define an actual RCFT.

8. Conclusion

We have followed one modular structure through three complementary settings. The arithmetic discussion showed how a simple transformation law on the upper half-plane leads to rigid spaces of modular forms, distinguished examples such as E_4, E_6, Δ, and η, and the multiplicative information encoded by Hecke operators. The geometry of complex tori then explained why the group $\mathrm{SL}_2(\mathbb{Z})$ appears: changing an oriented lattice basis changes the parameter τ but not the complex torus it describes.

The physical discussion supplied a second interpretation of the same geometry. At a continuous critical point, the divergence of the correlation length removes any distinguished finite infrared length scale, producing universal power-law behaviour and, in many two-dimensional systems, conformal symmetry. After Wick rotation, the thermal trace makes Euclidean time periodic; periodic space then gives a torus. Modular transformations are large changes of the torus-cycle basis. For a non-chiral bosonic CFT without a gravitational anomaly, the full torus partition function is invariant under this change of description. The S transformation exchanges the two cycles and relates low- and high-temperature data.

In rational conformal field theory, this general requirement becomes finite-dimensional. Characters count descendant states within conformal families and transform among themselves under the modular group. Their Fourier expansions, positivity and integrality properties, and modular covariance make them simultaneously physical counting functions and vector-valued modular objects. Modular linear differential equations exploit this meeting point to turn consistency conditions into a concrete classification strategy.

The central lesson is therefore not merely that modular forms occur in physics. Rather, arithmetic functions, complex-torus geometry, thermal

partition functions, and conformal characters are different realisations of the same symmetry principle. Seeing the complete chain makes clear why modularity is both mathematically rigid and physically predictive.

Acknowledgements: We are grateful to the anonymous referee for their valuable comments and suggestions, which helped improve the presentation of the manuscript. Part of this material grew out of a seminar presented by AD at COEP Technological University, Pune. AD gratefully acknowledges the hospitality of COEP Technological University. AD acknowledges support from a recently concluded STFC Consolidated Grant ST/T000600/1 – "Particle Theory at the Higgs Centre". The work of AD is supported by the grant ANRF/ECRG/2024/000247/PMS from the Anusandhan National Research Foundation (ANRF), India. JD thanks COEP Technological University for an excellent working environment.

Conflict of interest: The authors report no competing financial or non-financial interests.

References

[1] Cardy, J. L., Operator content of two-dimensional conformally invariant theories, *Nuclear Physics B* **270** (1986), 186–204.

[2] Cornell, G., Silverman, J. H. and Stevens, G. (eds.), Modular forms and Fermat's last theorem, Springer-Verlag, New York, 1997.

[3] Das, A., Gowdigere, C. N. and Santara, J., Wronskian indices and rational conformal field theories, *Journal of High Energy Physics* **2021** (2021), no. 4, 294.

[4] Deligne, P., La conjecture de Weil. I, *Publications Mathématiques de l'IHÉS* **43** (1974), 273–307.

[5] Di Francesco, P., Mathieu, P. and Sénéchal, D., Conformal field theory, Graduate Texts in Contemporary Physics, Springer-Verlag, New York, 1997.

[6] Gannon, T., The theory of vector-modular forms for the modular group, Conformal Field Theory, Automorphic Forms and Related Topics (Kohnen, W. and Weissauer, R., eds.), Contributions in Mathematical and Computational Sciences, vol. 8, Springer, 2014, 247–286.

[7] Hecke, E., Über die Bestimmung Dirichletscher Reihen durch ihre Funktionalgleichung, *Mathematische Annalen* **112** (1936), no. 1, 664–699.

[8] Mathur, S. D., Mukhi, S. and Sen, A., On the classification of rational conformal field theories, *Physics Letters B* **213** (1988), 303–308.

[9] Murty, M. R., Dewar, M. and Graves, H., Problems in the theory of modular forms, Institute of Mathematical Sciences Lecture Notes, vol. 1, Hindustan Book Agency, New Delhi, 2015.

[10] Polyakov, A. M., Conformal symmetry of critical fluctuations, *JETP Letters* **12** (1970), 381–383.
[11] Qualls, J. D., Lectures on conformal field theory, arXiv:1511.04074 [hep-th], 2015.
[12] Ramanujan, S., On certain arithmetical functions, Collected Papers of Srinivasa Ramanujan, AMS Chelsea Publishing, Providence, RI, 2000, 136–162. Originally published in 1916.
[13] Serre, J.-P., A course in arithmetic, Graduate Texts in Mathematics, vol. 7, Springer-Verlag, New York, 1973.
[14] Tatitscheff, V., Monstrous moonshine: A short introduction, *Resonance* **27** (2022), no. 12, 2107–2126.
[15] Wiles, A., Modular elliptic curves and Fermat's last theorem, *Annals of Mathematics* **141** (1995), no. 3, 443–551.

Arpit Das
School of Mathematics and Maxwell Institute for Mathematical Sciences,
University of Edinburgh, Edinburgh EH9 3FD, U.K.
Higgs Centre for Theoretical Physics,
University of Edinburgh, Edinburgh EH8 9YL, U.K.

Department of Astrophysics and High Energy Physics,
S.N. Bose National Centre for Basic Sciences,
Salt Lake, Kolkata 700106, India
E-mail: adas95@bose.res.in

Jishu Das
Department of Applied Sciences and Humanities, Mathematics Section,
COEP Technological University, Wellesley Road,
Shivajinagar, Pune, Maharashtra 411005, India.
E-mail: dasj.maths@coeptech.ac.in

The Mathematics Student ISSN: 0025-5742
Vol. 95, Nos. 3-4, July - December (2026), 274–284

PROBLEM SECTION

In the volume 95(1-2) 2026 of the Mathematics Student, we had invited solutions for a set of five new problems, in addition to the problems 2, 3 and 5 from 94(3-4) 2025. We have received solutions for problems 1, 4 and 5 from 95(1-2). We shall present these solutions and the solutions for problems 2, 3 and 5 of 94(3-4) as they still remain unsolved. We shall give some more time for problems 2 and 3 from 95(1-2).

We present five new problems in this volume and invite solutions for these problems from the readers till April 30, 2027. Correct solutions received by this date will be published in volume 96 (1-2) 2027 of The Mathematics Student, which is scheduled to be published in May 2027.

We sincerely acknowledge all the cooperation and support from the people who contribute problems and solutions to the problem section. We also welcome new problems along with detailed solutions for the problem section from the readers.

New Problems

MS 95(3-4) 2026 : Problem 1 (by **Mr. Ryota Inagaki**, Department of Mathematics, Massachusetts Institute of Technology, Cambridge, Massachusetts, U.S.A).

Let $n \geq 3$ be an integer. Consider cities $1, 2, \ldots, n$. Three travelers α, β, γ begin in cities $1, 2, 3$, respectively. At each turn, all three move simultaneously. A traveler currently in city c chooses uniformly at random one of the $n-1$ cities other than c and moves there. At each turn, conditional on the travelers' current positions, the travelers choose their destinations independently and uniformly. The random choices used on different turns are also independent.

Find the expected number of turns until the three travelers occupy the same city for the first time.

© *Indian Mathematical Society, 2026.*

MS 95(3-4) 2026 : Problem 2 (by **Dr. Quang Hung Tran**, High School for Gifted Students, VNU University of Science, Vietnam)

Let $ABCD.A'B'C'D'$ be a cube, and let O be the center of the face $ABCD$. Let Ω be the sphere with center O passing through A, B, C, D. Choose a point P on the open hemisphere of Ω lying on the side of the plane $(ABCD)$ opposite to the cube. Define

$$A_0 = PA' \cap (ABCD), \quad B_0 = PB' \cap (ABCD),$$
$$C_0 = PC' \cap (ABCD), \quad D_0 = PD' \cap (ABCD).$$

For a planar polygon $X_1X_2\cdots X_k$, denote its area by $[X_1X_2\cdots X_k]$. Prove that

$$[ABCD]\left(\sqrt{[ABCD]} + \sqrt{[A_0B_0C_0D_0]}\right)^2$$
$$= 2\left([A_0D_0CB]^2 + [B_0A_0DC]^2 + [C_0B_0AD]^2 + [D_0C_0BA]^2\right).$$

MS 95(3-4) 2026 : Problem 3 (by **Dr. B. Sury**, ISI, Bengaluru, India).

(i) Show that $\sqrt{t^2-t+1}, \sqrt{t^2+t+1}, \sqrt{3+4t^2}$, $t \in \mathbb{R}$, can be the lengths of sides of a triangle.
(ii) Show that the triangle must be obtuse.
(iii) Find the area of the triangle.
(iv) Show that length of one median of the triangle is independent of t.

MS 95(3-4) 2026 : Problem 4 (by **Dr. R. Roopkumar**, Department of Mathematics, Central University of Tamil Nadu, Thiruvarur, India).

Is the Schröder-Bernstein type theorem valid for homomorphism between groups? More precisely, if G and H are groups such that there exist two injective homomorphisms $\phi : G \to H$ and $\psi : H \to G$, can we conclude that G and H are isomorphic groups?

MS 95(3-4) 2026 : Problem 5 (by **Dr. S. K. Tomar**, Department of Mathematics, Panjab University, Chandigarh, India)

Using the calculus of variations, find the curve of shortest length that splits a given triangle into two regions of equal area. Verify your result with a suitable example.

Solutions for problems 2, 3 and 5 from MS 94(3-4) 2025.

MS 94(3-4) 2025 : Problem 2 (by **Dr. B. Sury**, ISI, Bengaluru, India).

For each real number $x \geq 0$, let $N_x = \{n \in \mathbb{N} \mid nx \quad \text{is even}\}$. If $f(x) = \sum_{n \in N_x} \frac{1}{2^n}$ for $x \in [0, 1)$, prove that $\text{Inf}\{f(x) \mid x \in [0, 1)\} = \frac{4}{7}$.

Solution: (by the Proposer)

We show that when $x \to \frac{2}{3}$ from the left, then $f(x)$ approaches $4/7$. After that, we show that $f(x) \geq 4/7$ for all $x \geq 0$. Firstly, let $x = \frac{2}{3} - \epsilon$ with $\epsilon > 0$ small. Let $n \in \mathbb{N}$ be small to start with so that $n\epsilon < 1$. Now if $n = 3r + 1$, for $r \in \mathbb{N}$, then

$$[nx] = [(3r+1)(\frac{2}{3} - \epsilon)] = [2r + \frac{2}{3} - (3r+1)\epsilon] = 2r.$$

If $n = 3r + 2$, then

$$[nx] = [(3r+2)(\frac{2}{3} - \epsilon)] = [2r + \frac{4}{3} - (3r+2)\epsilon] = 2r + 1.$$

If $n = 3r + 3$, then

$$[nx] = [(3r+3)(\frac{2}{3} - \epsilon)] = [2r + 2 - (3r+3)\epsilon] = 2r + 1.$$

Thus, $[nx]$ is even when $n = 3r + 1$ and odd in the other two cases. Also, this parity continues periodically. Therefore, for $x = \frac{2}{3} - \epsilon$, we have N_x to consist of n of the form $3r + 1$. Therefore,

$$f(x) = \sum_{r \geq 0} \frac{1}{2^{3r+1}} = \frac{1/2}{1 - 1/8} = 4/7.$$

Now $f(0) \geq 1 > 4/7$ and so, we consider $0 < x < 1/2$, $1/2 \leq x < 2/3$, $2/3 \leq x < 1$. It suffices to show that $f(x) > 4/7$ in each case.
If $0 < x < 1/2$, then evidently $[x] = [2x] = 0$ which means $1, 2 \in N_x$. Therefore,

$$f(x) > \frac{1}{2^1} + \frac{1}{2^2} = \frac{3}{4} > \frac{4}{7}.$$

Let $1/2 \leq x < 2/3$. We show that among any three consecutive integers $n, n+1, n+2$ at least one of them belongs to N_x.

In deed, if $[nx] = 2r + 1$, then

$$2r + 1 \leq nx < 2r + 2$$

which, on adding the inequalities

$$1 \leq 2x < 4/3$$

gives us

$$2r + 2 \leq (n + 2)x < 2r + 10/3.$$

This gives $n + 2 \in N_x$ if $(n + 2)x < 2r + 3$ and gives $n + 1 \in N_x$ if $(n + 2)x \geq 2r + 3$. Thus, one of $n, n + 1, n + 2$ is in N_x for $x \in [1/2, 2/3)$. So, if we partition the integers $n \geq 5$ into sets $\{3r + 2, 3r + 3, 3r + 4\}$, then the contribution from these 3 terms is at least as large as the least element $\frac{1}{2^{3r+4}}$.
For such x, we also have $1, 4 \in N_x$ as $[x] = 0, [4x] = 2$ are even. Therefore, for such x,

$$f(x) > \frac{1}{2} + \frac{1}{2^4} + \sum_{r\geq 1} \frac{1}{2^{3r+4}} = \frac{1}{+\frac{1}{[}16} + \frac{1/128}{1 - 1/8} = 64/112 = 4/7.$$

Finally, when $x \in [2/3, 1)$, we see that $[x] = 0, [3x] = 2$ are even and so $1, 3 \in N_x$. Hence

$$f(x) > 1 + 1/8 = 7/8 > 4/7.$$

Thus, we have shown that over all x, the infimum of $f(x)$ equals $4/7$.

MS 94(3-4) 2025 : Problem 3 (by **Dr. B. Sury**, ISI, Bengaluru, India).
If we have $n \in \mathbb{N}$, $n > 2$, discs on the plane with the property that any three of them intersect, then prove that the intersection of all discs is non-empty.

Solution: (by the Proposer)
We prove by induction on d that if we have $n > d$ closed balls in $\mathbb{R}^d$ satisfying the property that any $d + 1$ of the n balls intersect, then all of the n balls have a common intersection. For $d = 1$, we are given n intervals $I_i = [a_i, b_i]$ with $1 \leq i \leq n$ such that $I_i \cap I_j \neq \emptyset$ for all i, j. This means clearly that $a_i < b_j$ for all i, j. Therefore,

$$\emptyset \neq [\max_i a_i, \min_j b_j] \subseteq \bigcap_{k=1}^{n} I_k.$$

Let us assume now that $d > 1$ and that the result holds for any $1 \leq r < d$. Let $D_1, \cdots, D_n$ be closed balls in $\mathbb{R}^{d+1}$ such that each set of d balls intersects nontrivially. If the result is not true for this situation, then we have $D_1, \cdots, D_k$ satisfying

$$\cap_{i=1}^{k-1} D_i \neq \emptyset = \cap_{i=1}^{k} D_i.$$

Write $E = \cap_{i=1}^{k-1} D_i$; then E is not empty but $E \cap D_k$ is empty. Consider the hyperplane H in $\mathbb{R}^{d+1}$ which separates E and D_k. As D_k intersects each of $D_1, \cdots, D_{k-1}$, the intersections $H_i := H \cap D_i$ are non-empty for $i = 1, \cdots, k-1$. By hypothesis, D_k along with any d among the $k-1$ balls $D_1, \cdots, D_{k-1}$ have a common intersection, and therefore, any d of the sets H_i have a non-empty intersection. By induction hypothesis (for dimension $d-1$), the intersection $\cap_{i=1}^{k-1} H_i \neq \emptyset$. So

$$H \cap D_1 \cap D_2 \cap \cdots \cap D_{k-1} \neq \emptyset.$$

That is, $H \cap E \neq \emptyset$, which is a contradiction of the definition of the hyperplane H. The result is proved by contradiction.

MS 94(3-4) 2025 : Problem 5 (by **Dr. Chudamani Pranesachar Anil Kumar**, KREA University, India).

Baker, Heegner and Stark in the middle of the 20th century have proved a version of the statement that if $d > 0$ is squarefree and $d \equiv -1 \mod 4$, then the ring

$$R_d = \frac{\mathbb{Z}[X]}{\langle X^2 - X + \frac{(1+d)}{4} \rangle}$$

is a principal ideal domain if and only if $d = 3, 7, 11, 19, 43, 67, 163$. This problem is about proving the following.

Let $p \in \mathbb{N}$ be a prime and $D > 0$ be a squarefree integer such that $D \equiv -1 \mod 4$ and let $E = \frac{D+1}{4}$. Suppose $0 < p < \frac{D}{4}, p \nmid D$ and the polynomial $X^2 - X + E$ modulo p factorizes into linear factors in $\mathbb{F}_p[X]$. Then:

(1) Show that $R_D = \frac{\mathbb{Z}[X]}{\langle X^2 - X + E \rangle}$ is not a principal ideal domain.
(2) Conclude using (1) that, if $D > 3, D$ squarefree and $E = \frac{D+1}{4}$ is not prime then the ring R_D is not a principal ideal domain.
(3) Conclude that the rings R_D for $D = 51, 91, 115, 123$ are not principal ideal domains. What are the choices of p in these two cases?

(4) Conclude that if $D > 0, D \equiv -1 \mod 4$, D is squarefree then R_D is a not a principal ideal domain for all $3 \leq D \leq 163$ and $D \notin \{3, 7, 11, 19, 43, 67, 163\}$.

Solution: (by the Proposer)

We prove (1). First of all we have

$$\frac{R_D}{\langle p \rangle} \cong \frac{\mathbb{Z}[X]}{\langle p, X^2 - X + E \rangle} \cong \frac{\mathbb{F}_p[X]}{\langle X^2 - X + \overline{E} \rangle}$$

where $\overline{E}$ is the residue class of $E \mod p$. We observe that the roots of $X^2 - X + E$ are $\frac{1 \pm \sqrt{-D}}{2}$. Let $\delta = \frac{1+\sqrt{-D}}{2} \in \mathbb{C}$, then $\overline{\delta} = \frac{1-\sqrt{-D}}{2}$ and $R_D = \mathbb{Z}[\delta] = \mathbb{Z}[\overline{\delta}], \overline{\delta} = 1 - \delta$.

Any element in $\mathbb{Z}[\delta]$ is of the form $a + b\delta$ where $a, b \in \mathbb{Z}$.

Suppose $X^2 - X + \overline{E}$ factorizes in $\mathbb{F}_p[X]$ and say $X^2 - X + \overline{E} = (X - \alpha)(X + \alpha - 1)$ for some $\alpha \in \mathbb{F}_p$.

Then the ideals $M_1 = \langle X - \alpha \rangle, M_2 = \langle X + \alpha - 1 \rangle$ are maximal principal ideals in $\mathbb{F}_p[X]$ containing the principal ideal $\langle X^2 - X + \overline{E} \rangle$ in $\mathbb{F}_p[X]$.

Note that $M_1 \neq M_2$ because $\alpha \not\equiv 1 - \alpha \mod p$. If $\alpha \equiv 1 - \alpha \mod p$, then we have $2\alpha \equiv 1 \mod p \Rightarrow p \neq 2$ and $\alpha^2 \equiv E \mod p \Rightarrow \frac{1}{4} \equiv \frac{D+1}{4} \mod p \Rightarrow D \equiv 0 \mod p \Rightarrow p \mid D$. But in the hypothesis we are given that $p \nmid D$. Hence M_1, M_2 are two distinct maximal ideals.

Now we get that $M_1 \cap M_2 = M_1 M_2 = \langle X^2 - X + \overline{E} \rangle$. So the distinct ideals $\overline{M}_1, \overline{M}_2$ in $\frac{\mathbb{F}_p[X]}{\langle X^2 - X + \overline{E} \rangle}$ associated to M_1, M_2 respectively are maximal ideals in $\frac{\mathbb{F}_p[X]}{\langle X^2 - X + \overline{E} \rangle} = \frac{R_D}{\langle p \rangle}$ and we also have $\overline{M}_1 \overline{M}_2 = \langle 0 \rangle$. If $\phi : R_D \longrightarrow \frac{R_D}{\langle p \rangle} = \frac{\mathbb{F}_p[X]}{\langle X^2 - X + \overline{E} \rangle}$ is the quotient homomorphism, then the ideals $N_1 = \phi^{-1}(\overline{M}_1), N_2 = \phi^{-1}(\overline{M}_2)$ are maximal ideals in R_D and N_1, N_2 are two distinct maximal ideals containing $\langle p \rangle$ such that $N_1 \cap N_2 = N_1 N_2 = \langle p \rangle$ since $\overline{M}_1 \overline{M}_2 = \langle 0 \rangle$.

We also observe that $X^2 - X + E = (X - \delta)(X - \overline{\delta})$. The conjugation automorphism $c : R_D \longrightarrow R_D$ takes the ideal $\langle p \rangle$ to itself. Hence it induces an automorphism of $\frac{R_D}{\langle p \rangle}$. Also $c(\delta) = \overline{\delta} = 1 - \delta$. Consider the following commutative diagram.

$$\begin{array}{ccc} \mathbb{Z}[X] & \xrightarrow[\text{kill } p]{\tau} & \mathbb{F}_p[X] \\ {\scriptstyle \text{kill } X^2-X+E}\big\downarrow{\scriptstyle X \atop \delta}\pi & & \psi\big\downarrow{\scriptstyle \text{kill } X^2-X+E} \\ \mathbb{Z}[\delta] = \frac{\mathbb{Z}[X]}{\langle X^2-X+E \rangle} = R_D & \xrightarrow[\text{kill } p]{\phi} & \frac{R_D}{\langle p \rangle} = \frac{\mathbb{F}_p[X]}{\langle X^2-X+E \rangle} = \frac{\mathbb{Z}[X]}{\langle p, X^2-X+E \rangle} \end{array}$$

Let $P_1 = \pi^{-1}(N_1), P_2 = \pi^{-1}(N_2)$ be the two distinct maximal ideals in $\mathbb{Z}[X]$ which contain the ideal $\langle p, X^2 - X + E\rangle$. Then we have by commutativity of the above diagram
$P_1 = \tau^{-1}(M_1) = \tau^{-1}(\psi^{-1}(\overline{M}_1)), P_2 = \tau^{-1}(M_2) = \tau^{-1}(\psi^{-1}(\overline{M}_2))$. Let $n \in \mathbb{Z}$ and $n \equiv \alpha \mod p$. Then we have $1 - n \equiv 1 - \alpha \mod p$ and we have $P_1 = \langle p, X - n\rangle, P_2 = \langle p, X + n - 1\rangle$. So we get

$$N_1 = \frac{\langle p, X - n\rangle}{\langle X^2 - X + E\rangle}, N_2 = \frac{\langle p, X + n - 1\rangle}{\langle X^2 - X + E\rangle}.$$

Now we observe that $\pi(X) = \delta \in \mathbb{Z}[\delta]$. Hence we have

$$N_1 = \langle p, \delta - n\rangle, N_2 = \langle p, \delta + n - 1\rangle \subset \mathbb{Z}[\delta].$$

Here we see that the conjugation map c takes p to p and $\delta - n$ to $\overline{\delta} - n = -(\delta + n - 1)$. We also observe that the conjugation map takes $\delta + n - 1$ to $\overline{\delta} + n - 1 = -(\delta - n)$. Hence the conjugation map c takes the maximal ideal N_1 to N_2 and vice-versa.
Now if R_D is a principal ideal domain then $N_1 = \langle x + y\delta\rangle$ and $N_2 = \langle x + y\overline{\delta}\rangle$ for some $x, y \in \mathbb{Z}$. Since $\langle p\rangle = N_1N_2 = \langle (x + y\delta)(x + y\overline{\delta})\rangle$ we have that there exists a unit $u \in \mathbb{Z}[\delta]$ such that $p = u(x + y\delta)(x + y\overline{\delta})$. Now the only units in $\mathbb{Z}[\delta]$ are ± 1 as we can assume $E > 1$ since there exists a prime p satisfying $0 < p < \frac{D}{4} < E$. Now the reason for the units to be only ± 1 is as follows. In the ring $\mathbb{Z}[\delta]$ if $(a + b\delta)(c + d\delta) = 1$ then we have $(a + b\overline{\delta})(c + d\overline{\delta}) = 1$. Hence we get that
$(a + b\delta)(a + b\overline{\delta})(c + d\overline{\delta})(c + d\delta) = 1 \Rightarrow (a^2 + ab + Eb^2)(c^2 + cd + Ed^2) = 1 \Rightarrow (a^2 + ab + Eb^2) = \pm 1 = (c^2 + cd + Ed^2)$. But then we have $(2a + b)^2 + (4E - 1)b^2 = \pm 4, (2c + d)^2 + (4E - 1)d^2 = \pm 4$. The possibility -4 cannot occur. Since $E > 1$ the only solution to the equations $(2a + b)^2 + (4E - 1)b^2 = 4, (2c + d)^2 + (4E - 1)d^2 = 4$ is $a = \pm 1, c = \pm 1, b = 0 = d$ and we have $ac = 1$. Therefore the only units are ± 1. Hence we get $p = \pm(x + y\delta)(x + y\overline{\delta}) = \pm(x^2 + xy + Ey^2)$. So $p = \pm\frac{1}{4}\big((2x + y)^2 + (4E - 1)y^2\big)$. The negative sign is impossible. The possibility that
$p = (x^2 + xy + Ey^2) = \frac{1}{4}\big((2x + y)^2 + (4E - 1)y^2\big) = \frac{1}{4}\big((2x + y)^2 + Dy^2\big)$ also cannot occur. Since $p < \frac{D}{4}$, $y = 0$ in which case $p = x^2$ is not possible. Thus we must have that $\mathbb{Z}[\delta]$ is not a principal ideal domain, which proves (1).

Next we prove (2). If $E = \frac{D+1}{4}$ is not prime, let $p \mid E$ where p is a prime. Then we have $p < E = \frac{D+1}{4} \Rightarrow p < \frac{D}{4}$. Also
$p \mid E \Rightarrow p \mid 4E \Rightarrow p \mid D+1 \Rightarrow p \nmid D$. Now we observe that $X^2 - X + E$ modulo p factorizes as $X(X-1)$. So using (1) we conclude that $\mathbb{Z}[\delta]$ is not a principal ideal domain where $\delta = \frac{1+\sqrt{-D}}{2}$. This proves (2).

We prove (3). If $D = 51$ then $E = 13$ which is a prime and if $D = 91$ then $E = 23$ which is also a prime. In both cases we choose $p = 5$. Then we have $p \nmid D$ and $p < \frac{D}{4}$. We also have in both cases,
$X^2 - X + E \equiv X^2 - X - 2 \equiv (X-2)(X+1) \mod 5$. If $D = 115$ then $E = 29$ and choose $p = 7 \Rightarrow p \nmid D, p < \frac{D}{4}$. We have
$X^2 - X + E \equiv X^2 - X - 6 \equiv (X-3)(X+2) \mod 7$. If $D = 123$ then $E = 31$ and choose $p = 11 \Rightarrow p \nmid D, p < \frac{D}{4}$. We have
$X^2 - X + E \equiv X^2 - X - 2 \equiv (X-2)(X+1) \mod 11$. Hence using (1) we conclude that $\mathbb{Z}[\delta]$ is not a principal ideal domain where $\delta = \frac{1+\sqrt{-D}}{2}$ in all cases $D \in \{51, 91, 115, 123\}$, which proves (3).

Finally we prove (4). For $3 \leq D \leq 163, D \notin \{3, 7, 11, 19, 43, 67, 163\}, D$ squarefree and $\frac{D+1}{4}$ is not a prime then using (2) we are done. Suppose $\frac{D+1}{4}$ is prime. Then we have
$\frac{D+1}{4} < \frac{163+1}{4} = 41 \Rightarrow \frac{D+1}{4} \in \{2, 3, 5, 7, 11, 13, 17, 19, 23, 29, 31, 37\}$.
Hence $D \in \{7, 11, 19, 27, 51, 67, 75, 91, 115, 123, 147\}$ which implies $D \in \{51, 91, 115, 123\}$ because $D \notin \{3, 7, 11, 19, 43, 67, 163\}$ and D is squarefree. In the cases $D \in \{51, 91, 115, 123\}$ we use (3). Hence we have for $3 \leq D \leq 163, D \notin \{3, 7, 11, 19, 43, 67, 163\}, D$ squarefree, the ring $\mathbb{Z}[\delta]$ is not a principal ideal domain where $\delta = \frac{1+\sqrt{-D}}{2}$, which proves (4) and hence the proof completes.

Solutions for the problems 1, 4 and 5 from MS 95(1-2) 2026.

MS 95(1-2) 2026 : Problem 1 (by **Dr. Nguyen Tien Lam**, High School for Gifted Students, Hanoi University of Science, Vietnam).

Find all pairs of positive integers (a, b) such that there exist infinitely many pairs of positive integers (m, n) for which both $m^3 + 3m^2 + an + b$ and $n^3 + 3n^2 + am + b$ are perfect cubes.

Solution: (by Mr. Santanu Panda, Hooghly, West Bengal, India.)
Since $m^3 + 3m^2 + an + b$ is a perfect cube and a, b, m, n are positive integers, $m^3 + 3m^2 + an + b > m^3$. So, we can say $m^3 + 3m^2 + an + b \geq (m+1)^3$,

so $m^3+3m^2+an+b=(m+1)^3 \Rightarrow an+b=3m+1$ or $m^3+3m^2+an+b \geq (m+2)^3 \Rightarrow an+b \geq 3m^2+12m+8$. Similar things happen for n^3+3n^2+am+b. Combining them, there are four possible cases. At least one of these four must be satisfied by infinitely many pairs of positive integers m, n.

Case-I : $an+b=3m+1$ and $am+b=3n+1$
This gives $an-3m=am-3n=1-b$.
$an-3m=am-3n \Rightarrow a(n-m)=3(m-n) \Rightarrow (a+3)(n-m)=0$.
As $a+3>0$, $m=n$. Then $(a-3)m=1-b$ has infinitely many solutions m only if $a=3$ and $b=1$.

Case-II : $an+b \geq 3m^2+12m+8$ and $am+b=3n+1$
This gives $9m^2+(36-a^2)m+(24-ab+a-3b) \leq 0$.
The LHS of the inequality is a quadratic function of m with a positive coefficient of m^2. So for any (a,b), the solution set of the inequality is a bounded interval. In this bounded interval, there can only be finitely many integer values of m. And for each possible integer value of m in the solution set, there can be at most one value of n. So, this case does not give infinitely many pairs of positive integers (m,n).

Case-III : $an+b=3m+1$ and $am+b \geq 3n^2+12n+8$
For a similar reason as Case-II, this case does not give infinitely many pairs of positive integers (m,n).

Case-IV : $an+b \geq 3m^2+12m+8$ and $am+b \geq 3n^2+12n+8$
Without loss of generality, we can assume $m \geq n$, since $a>0$, $am+b \geq an+b \Rightarrow am+b \geq 3m^2+12m+8 \Rightarrow 3m^2+(12-a)m+(8-b) \leq 0$.The LHS of the inequality is a quadratic function of m with positive coefficient of m^2. For a similar reason as used in Case-II, there can only be finitely many integer values of m. As $m \geq n > 0$, for each m, there exist only finitely many n. So, this case does not give infinitely many pairs of positive integers (m,n).
When $(a,b)=(3,1)$ and $m=n$, $m^3+3m^2+an+b=n^3+3n^2+am+b=(m+1)^3$, every $m=n$ gives a solution. So, $(a,b)=(3,1)$ is the only pair of positive integers for which there are infinitely many pairs of positive integers (m,n) such that both m^3+3m^2+an+b and n^3+3n^2+am+b are perfect cubes.

MS 95(1-2) 2025 : Problem 4 (by **Dr. B. Sury**, ISI, Bengaluru, India).
Let f be a monic polynomial of degree $n \in \mathbb{N}$ with real coefficients and suppose $a_0 > a_1 > \cdots > a_n$ are integers. Prove that $|f(a_i)| \geq \dfrac{n!}{2^n}$ for some $0 \leq i \leq n$.

Solution: (by Mr. Santanu Panda, Hooghly, West Bengal, India.)
Here, we prove that if $a_0 > a_1 > \cdots > a_n$ are integers and $|f(a_i)| < \frac{n!}{2^n}$ for all $0 \leq i \leq n$, then the polynomial is not monic. This is equivalent to the given statement.
Since all the coefficients of f are real, for all $0 \leq i \leq n$, $f(a_i)$ is real. Let $f(a_i) = b_i, \forall 0 \leq i \leq n$. As all a_i's are distinct, by Lagrange's interpolation method, we can say,

$$f(x) = \sum_{i=0}^{n} b_i \prod_{j=0, j\neq i}^{n} \frac{x - a_j}{a_i - a_j}.$$

Since the function has degree n as given, the leading coefficient of f is :

$$\sum_{i=0}^{n} b_i \prod_{j=0, j\neq i}^{n} \frac{1}{a_i - a_j}.$$

Now,

$$\begin{aligned}
\left| \sum_{i=0}^{n} b_i \prod_{j=0, j\neq i}^{n} \frac{1}{a_i - a_j} \right| &\leq \sum_{i=0}^{n} \left| b_i \prod_{j=0, j\neq i}^{n} \frac{1}{a_i - a_j} \right| \\
&= \sum_{i=0}^{n} |b_i| \prod_{j=0, j\neq i}^{n} \frac{1}{|a_i - a_j|} \\
&< \frac{n!}{2^n} \sum_{i=0}^{n} \prod_{j=0, j\neq i}^{n} \frac{1}{|a_i - a_j|} \quad \left[\text{as } |b_i| = |f(a_i)| < \frac{n!}{2^n} \right].
\end{aligned} \tag{1}$$

Now, since $a_0 > a_1 > \cdots > a_n$ are integers, there exist at least $|i - j| - 1$ distinct integers(a_k) between a_i and a_j (excluding a_i and a_j). So, $|a_i - a_j| \geq |i - j|$ for all $0 \leq i, j \leq n$. So, for all $0 \leq i \leq n$,

$$\begin{aligned}
\prod_{j=0, j\neq i}^{n} \frac{1}{|a_i - a_j|} &= \frac{1}{(|a_0 - a_i| \cdots |a_{i-1} - a_i|)(|a_i - a_{i+1}| \cdots |a_i - a_n|)} \\
&\leq \frac{1}{(i \cdots 1)(1 \cdots (n-i))} = \frac{1}{i!(n-i)!}.
\end{aligned}$$

From inequality (1),

$$\left|\sum_{i=0}^{n} b_i \prod_{j=0, j\neq i}^{n} \frac{1}{a_i - a_j}\right| < \frac{n!}{2^n} \sum_{i=0}^{n} \prod_{j=0, j\neq i}^{n} \frac{1}{|a_i - a_j|}$$
$$\leq \frac{n!}{2^n} \sum_{i=0}^{n} \frac{1}{i!(n-i)!} = \frac{n!}{2^n} \frac{1}{n!} \sum_{i=0}^{n} \binom{n}{i} = \frac{n!}{2^n} \frac{2^n}{n!} = 1.$$

So, the absolute value of the leading coefficient of f is less than 1. So, the function cannot be a monic polynomial.

MS 95(1-2) 2026 : Problem 5(by **Dr. B. Sury**, ISI, Bengaluru, India) For $1 \leq k \leq n$, consider the set:

$$S_k(n) := \left\{(a_1, a_2, \cdots, a_k) \in \mathbb{N}^k : \sum_{i=1}^{k} a_i = n\right\}.$$

For instance, $S_2(3) = \{(1,2),(2,1)\}$ and $S_2(4) = \{(1,3),(2,2),(3,1)\}$. Note that

$$\sum_{k=1}^{3} \left(\frac{(-1)^k}{k!} \sum_{(a_1,\cdots,a_k)\in S_k(3)} \frac{1}{a_1 \cdots a_k}\right) = -\frac{1}{1} \cdot \frac{1}{3} + \frac{1}{2!}\left(\frac{1}{1.2} + \frac{1}{2.1}\right) - \frac{1}{3!} \cdot \frac{1}{1.1.1} = 0.$$

Prove that for all $n > 1$ that $\sum_{k=1}^{n} \left(\frac{(-1)^k}{k!} \sum_{(a_1,\cdots,a_k)\in S_k(n)} \frac{1}{a_1 a_2 \cdots a_k}\right) = 0$.

This problem was also solved by Mr. Santanu Panda, Hooghly, West Bengal, India.

Solution:(by the Proposer) If $|x| < 1$, we have

$$1 - x = \exp(\ln(1-x)) = \sum_{n\geq 0} \frac{\ln(1-x)^n}{n!} = \sum_{n\geq 0} \frac{(-1)^n}{n!} \left(\sum_{r\geq 1} \frac{x^r}{r}\right)^n.$$

We consider the coefficients of various powers of x on the right hand side. The first two terms on the right hand side are $1 - x$ (which are the two terms on the left hand side). The coefficient of any x^n with $n \geq 2$ on the right hand side is

$$\sum_{r\geq 1} \frac{(-1)^r}{r!} \sum_{a_1+\cdots+a_r=n} \frac{1}{a_1 a_2 \cdots a_r} = \sum_{r=1}^{n} \frac{(-1)^r}{r!} \sum_{(a_1,\cdots,a_r)\in S_r(n)} \frac{1}{a_1 a_2 \cdots a_r}.$$

For $|x| < 1$ the coefficients of x^n with $n \geq 2$ are zero; hence, for $n \geq 2$:

$$\sum_{r\geq 1} \frac{(-1)^r}{r!} \sum_{(a_1,\cdots,a_r)\in S_r(n)} \frac{1}{a_1 a_2 \cdots a_r} = 0.$$

FORM IV

1.	Place of Publication:	PUNE
2.	Periodicity of publication:	QUARTERLY
3.	Printer's Name: Nationality: Address:	ANANT BARVE INDIAN PARASURAM PRINTERS A-7, CO-OPERATIVE INDUSTRIAL ESTATE RAILWAY STATION MIDC CHHATRAPATI SAMBHAJINAGAR-431 005, INDIA
4.	Publisher's Name: Nationality: Address:	M. M. SHIKARE INDIAN GENERAL SECRETARY THE INDIAN MATHEMATICAL SOCIETY EMERITUS PROFESSOR DEPARTMENT OF MATHEMATICS JSPM UNIVERSITY, PUNE 412207 MAHARASHTRA, INDIA
5.	Editor's Name: Nationality: Address:	G. P. YOUVARAJ INDIAN RAMANUJAN INSTITUTE FOR ADVANCED STUDY IN MATHEMATICS, UNIVERSITY OF MADRAS, CHENNAI 600005, TN, INDIA
6.	Names and addresses of individuals who own the newspaper and partners or shareholders holding more than 1% of the total capital:	THE INDIAN MATHEMATICAL SOCIETY

I, M. M. Shikare, the General Secretary of the IMS, hereby declare that the particulars given above are true to the best of my knowledge and belief.

M. M. Shikare

Dated: November 15, 2026 Signature of the Publisher

Published by Prof. M. M. Shikare for the Indian Mathematical Society, type set by Prof. G. P. Youvaraj, Former Director and Head, Ramanujan Institute for Advanced Study in Mathematics, Chepauk, Chennai - 600005, India and printed by Anant Barve at Parashuram Printers, A-17, Co-Operative Industrial Estate, Railway station MIDC, Chhatrapati Sambhajinagar-431 005, Maharashtra, India.

Printed in India

The Mathematics Student ISSN: 0025-5742
Vol. 95, Nos. 3-4, July - December (2026)

EDITORIAL BOARD

G. P. Youvaraj (Editor-in-Chief)
Ramanujan Institute for Advanced Study in Mathematics
University of Madras, Chennai- 600 005, Tamil Nadu, India
E-mail : msindianmathsociety@gmail.com

Bruce C. Berndt
Dept. of Mathematics, University of Illinois 1409 *West Green St. Urbana, IL* 61801, *USA*
E − mail : berndt@illinois.edu

George E. Andrews
Dept. of Mathematics, The Pennsylvania State University, University Park PA 16802, *USA*
E − mail : gea1@psu.edu

M. Ram Murty
Dept. of Mathematics and Statistics Jeffery Hall, Queen's University Kingston, Ontario, K7L3N6, Canada
E − mail : murty@mast.queensu.ca

B. Sury
Theoretical Stat. and Mat. Unit Indian Statistical Institute Bangalore 560059, *India*
E − mail : surybang@gmail.com

Siddharth Gadgil
Indian Institute of Science Bangalore − 560012, *India*
E − mail : siddharth.gadgil@gmail.com

Gadadhar Misra
Indian Statistical Institute Bangalore − 560059, *India*
E − mail : gm@iisc.ac.in

Sukanta Pati
IIT Guwahati Guwahati 781039, *Assam, India*
E − mail : pati@iitg.ac.in

Atul Dixit
IIT Gandhinagar Gandhinagar − 382355, *Gujarat, India*
E − mail : adixit@iitg.ac.in

S. K. Tomar
Panjab University Chandigarh − 160014, *India*
E − mail : sktomar@pu.ac.in

Aparna Dar
IITKanpur Kanpur − 208016, *U.P., India*
E − mail : adar@iitk.ac.in

Clare D′Cruz
Dept. of Mathematics, CMI IT Park Padur P.O., Siruseri Kelambakkam − 603103, *T.N., India*
E − mail : clare@cmi.ac.in

L. Sunil Chandran
Dept. of Computer Science&Automation Indian Institute of Science Bangalore − 560012, *India*
E − mail : sunil.cl@gmail.com

Subhojoy Gupta
Indian Institute of Science Bangalore − 560012, *India*
E − mail : subhojoy@iisc.ac.in

T. S. S. R. K. Rao
ManipalUniversity MJaipur 303007, *Rajasthan, India*
E − mail : tssrkrao@gmail.com

Kaushal Verma
Dept. of Mathematics Indian Institute of Science Bangalore − 560012, *India*
E − mail : kverma@iisc.ac.in

C. S. Aravinda
TIFR Centre for Applicable Mathematics P. B. No. 6503, *GKVK Post Sharadanagara Chikkabommasandra, Bangalore* − 560065, *India*
E − mail : aravinda@math.tifrbng.res.in

Indranil Biswas
School of Mathematics, Tata Institute of Fundamental Research, Homi Bhabha Rd., Mumbai − 400005, *India*
E − mail : indranil29@gmail.com

Timothy Huber
School of Mathematics and statistical Sciences University of Texas Rio Grande Valley, 1201 *West Univ. Avenue, Edinburg, TX* 78539 *USA*
E − mail : timothy.huber@utrgv.edu

Debjani Chakrabory
Department of Mathematics IIT Kharagpur, Kharagpur 721302, *India*
E − mail : debjani@maths.iitkgp.ac.in

Safique Ahmad
Department of Mathematics IIT Indore, Indore 452020, *India*
E − mail : safique@iiti.ac.in

T. Raja Sekhar
Department of Mathematics IIT Kharagpur, Kharagpur 721302, *India*
E − mail : trajasekhar@maths.iitkgp.ac.in

THE INDIAN MATHEMATICAL SOCIETY

Founded in 1907

Registered Office: Center for Advanced Study in Mathematics
Savitribai Phule Pune University, Pune - 411007

COUNCIL FOR THE SESSION 2026-2027

PRESIDENT: Amin Sofi, Formerly at Department of Mathematics, Kashmir University, Srinagar-190006, India

IMMEDIATE PAST PRESIDENT: S. S. Khare, Formerly at Department of Mathematics, Northeast Hill University, Shillong-793022, India

GENERAL SECRETARY: M. M. Shikare, Emeritus Professor, Department of Mathematics, JSPM University, Pune-412207, Maharashtra, India

ADMINISTRATIVE SECRETARY: B. N. Waphare, Center for Advanced Study in Mathematics, S. P. Pune University, Pune-411007, Maharashtra, India

ACADEMIC SECRETARY: G. P. Raja Sekhar, Department of Mathematics IIT Kharagpur, Kharagpur-721302, West Bengal, India

TREASURER: M. M. Pawar, Formerly at K. B. C. North Maharastra University, Jalgon, Maharashtra, India

EDITOR: **J. of the Indian Math. Society**: Pankaj Jain, South Asian University, Rajpur Road, Maidan Garhi, New Delhi-110068, India

EDITOR: **The Mathematics Student**: G. P. Youvaraj, Former Director, Ramanujan Inst. for Advanced Study in Mathematics, University of Madras, Chennai-600005, India

LIBRARIAN: M. Pitchaimani, Director, Ramanujan Inst. for Advanced Study in Mathematics, University of Madras, Chennai-600005, India

OTHER MEMBERS OF THE COUNCIL:

Sameer Chavan, Department of Mathematics and Statistics, IIT Kanpur, Kanpur, U. P.

Swaroop Nandan Bora, Department of Mathematics, IIT Guwahati, Guwahati, Asam

Seshadev Padhi, Department of Mathematics, BITS, Mesra, Ranchi, Jharkhand

Pallavi Manhale, Department of Mathematics, VNIT Nagpur, Maharashtra

Punita Batra, Department of Mathematics, HRI, Allahabad, U. P.

P. K. Sahoo, Department of Mathematics, BITS Pillani, Hyderabad Campus, Hyderabad, Telangana

P. P. Malavadkar, Department of Mathematics, MIT World Peace University, Pune, Maharastra

Apoorva Khare, Department of Mathematics, IISc, Bangaluru, Karnataka

Ram Prakash Sharma, Department of Mathematics, NIT Jote Papum, Arunachal Pradesh

Edited by G. P. Youvaraj and published by M. M. Shikare, the General Secretary of the IMS, for the Indian Mathematical Society.

Typeset by Prof. G. P. Youvaraj, Former Director and Head, Ramanujan Institute for Advanced Study in Mathematics, Chepauk, Chennai - 600005, India and printed by Anant Barve at Parashuram Printers, A-17, Co-Operative Industrial Estate, Railway station MIDC, Chhatrapati Sambhajinagar-431 005, Maharashtra, India. Printed in India

Back volumes of our periodicals, except for a few numbers out of stock, are available.

Copyright ©The Indian Mathematical Society, 2026.