DISCRETE AND CONTINUOUS MODELS AND APPLIED COMPUTATIONAL SCIENCE

Volume 34 Number 1 (2026)

Founded in 1993

Founder: PEOPLES' FRIENDSHIP UNIVERSITY OF RUSSIA NAMED AFTER PATRICE LUMUMBA

DOI: 10.22363/2658-4670-2026-34-1

Edition registered by the Federal Service for Supervision of Communications, Information Technology and Mass Media
Registration Certificate: ПИ № ФС 77-76317, 19.07.2019

ISSN 2658-7149 (Online); 2658-4670 (Print)
4 issues per year.
Language: English.

Publisher

Peoples' Friendship University of Russia named after Patrice Lumumba (RUDN University).

Indexed by

- Scopus (https://www.scopus.com),
- Ulrich's Periodicals Directory (http://www.ulrichsweb.com),
- Directory of Open Access Journals (DOAJ) (https://doaj.org),
- Russian Index of Science Citation (https://elibrary.ru),
- CyberLeninka (https://cyberleninka.ru).

Aim and Scope

Discrete and Continuous Models and Applied Computational Science arose in 2019 as a continuation of RUDN Journal of Mathematics, Information Sciences and Physics. RUDN Journal of Mathematics, Information Sciences and Physics arose in 2006 as a merger and continuation of the series "Physics", "Mathematics", "Applied Mathematics and Computer Science", "Applied Mathematics and Computer Mathematics".

Discussed issues affecting modern problems of physics, mathematics, queuing theory, the Teletraffic theory, computer science, software and databases development.

It's an international journal regarding both the editorial board and contributing authors as well as research and topics of publications. Its authors are leading researchers possessing PhD and PhDr degrees, and PhD and MA students from Russia and abroad. Articles are indexed in the Russian and foreign databases. Each paper is reviewed by at least two reviewers, the composition of which includes PhDs, are well known in their circles. Author's part of the magazine includes both young scientists, graduate students and talented students, who publish their works, and famous giants of world science.

The Journal is published in accordance with the policies of COPE (Committee on Publication Ethics).

The editors are open to thematic issue initiatives with guest editors. Further information regarding notes for contributors, subscription, and back volumes is available at http://journals.rudn.ru/miph

E-mail: miphj@rudn.ru, dcm@rudn.su

Editorial board

Computer Design: *Anna V. Korolkova*, *Dmitry S. Kulyabov*
English Text Editors: *Nikolay E. Nikolaev*, *Ivan S. Zaryadov*, *Konstantin P. Lovetskiy*
Address of editorial board: 3 Ordzhonikidze St, 115419, Moscow, Russia, +7 (495) 955-07-16, e-mail: publishing@rudn.ru
Editorial office: +7 (495) 952-02-50, e-mail: miphj@rudn.ru, dcm@rudn.su, site: https://journals.rudn.ru/miph

---

Approved for printing: 23.03.2026. Published: 30.03.2026.
Paper size 70×100/16. Offset paper. Offset printing. Typeface “Adobe Source.”
Conventional printed sheets: 12.01. Printing run 500 copies. Open price. The order: 2.
PEOPLES' FRIENDSHIP UNIVERSITY OF RUSSIA NAMED AFTER PATRICE LUMUMBA
6 Miklukho-Maklaya St, Moscow, 117198, Russian Federation
Printed at RUDN Publishing House:
3 Ordzhonikidze St, Moscow, 115419, Russian Federation,
+7 (495) 955-08-61; e-mail: publishing@rudn.ru

Discrete & Continuous Models
& Applied Computational Science
ISSN 2658-7149 (Online), 2658-4670 (Print)

2026, 34 (1)
http://journals.rudn.ru/miph

Contents

Editorial

Computer science

Modeling and simulation

Physics and astronomy

Letters

Discrete & Continuous Models & Applied Computational Science
ISSN 2658-7149 (Online), 2658-4670 (Print)

2026, 34 (1) 5–11
http://journals.rudn.ru/miph

Editorial

DOI: 10.22363/2658-4670-2026-34-1-5-11 EDN: VDNGIA

Physical dimensional quantities typesetting

Dmitry S. Kulyabov[1,2], Anna V. Korolkova[1], Leonid A. Sevastianov[1,2], Yuri P. Rybakov[1]

[1] RUDN University, 6 Miklukho-Maklaya St, Moscow, 117198, Russian Federation

[2] Joint Institute for Nuclear Research, 6 Joliot-Curie St, Dubna, 141980, Russian Federation

Abstract. The `siunitx` package is designed for typographically correct and consistent typesetting of physical quantities (numbers with units of measurement) in LaTeX documents. It automates formatting according to the rules of the International System of Units (SI), eliminating the need to manually manage spaces, fonts, and separators.

Key words and phrases: unit typesetting, LaTeX

For citation: Kulyabov, D. S., Korolkova, A. V., Sevastianov, L. A., Rybakov, Y. P. Physical dimensional quantities typesetting. *Discrete and Continuous Models and Applied Computational Science* **34** (1), 5–11. doi: 10.22363/2658-4670-2026-34-1-5-11. edn: VDNGIA (2026).

1. Introduction

We use version 3 of the `siunitx` package. In version 3, the core commands of the `siunitx` package have changed:

- Old syntax (v2): **`\SI`**`{number}{unit}` and **`\si`**`{unit}`.
- New syntax (v3): **`\qty`**`{number}{unit}` and **`\unit`**`{unit}`.

Global formatting settings for the entire document are set with the **`\sisetup`** command in the preamble.

For example, Russian-language texts often require a comma as a decimal separator:

```
\sisetup{
locale = DE, % or output-decimal-marker = {,}
}
```

Here, the `locale = DE` option will automatically adjust the format for German (and many other European languages), using a comma as a decimal separator and a period to separate groups of thousands.

2. Basic commands

The package provides three basic commands.

- **`\num`**`{<number>}` — for formatting numbers only:

Formatting Numbers

```
\num{12345.67890}\\
\num{1.34e{-12}}
```

12 345.678 90
1.34×10^{-12}

The command automatically inserts spaces between digit groups, handles scientific notation, and replaces decimal points with commas, if configured.

– **\unit**{<unit>} — for formatting units only.

Formatting Units of Measurement

```
\unit{\kilogram\metre\per\second}\\
\unit{kg.m/s^{2}}
```

$\mathrm{kg\,m\,s^{-1}}$
$\mathrm{kg\,m/s^{2}}$

Units can be entered using special macros (**\kilogram**, **\metre**) or plain text. The package will automatically convert the font to roman (not italic) and insert the correct spacing. To multiply units, use a period (.), and to divide units, use **\per** or the / symbol.

– **\qty**{<number>}{<unit>} is the basic command for outputting a quantity (number + unit). It combines the actions of **\num** and **\unit**.

Formatting Units

```
\qty{9.81}{\metre\per\second\squared}\\
\qty{100}{\kilo\metre\per\hour}
```

$\mathrm{9.81\,m\,s^{-2}}$
$\mathrm{100\,km\,h^{-1}}$

3. Additional commands

For more complex cases, separate commands are provided.

– Angles (degrees, minutes, seconds):

Angles

```
\ang{47.99}\\
\ang{47;59;43.373}
```

47.99°
47°59′43.373″

– Lists of Values:

Lists of Values

```
\qtylist{10;20;30}{\milli\metre}\\
\numlist{10;20;30}
```

10 mm, 20 mm and 30 mm
10, 20 and 30

– Ranges of Values:

Ranges of Values

```
\qtyrange{10}{100}{\metre}\\
\numrange{5}{15}
```

10 m to 100 m
5 to 15

– Product of quantities:

Product of quantities

```
\qtyproduct{2x3x4}{\centi\metre}
```

2 cm × 3 cm × 4 cm

– Commands for complex numbers:

Complex numbers

```
\complexnum{1+2i}\\
\complexqty{1+2i}{\metre}
```

1 + 2i
(1 + 2i) m

4. Tables

One of the features of `siunitx` is the ability to align numbers in tables by decimal separators.

For this, a special column type, `S`, is used.

Working with tables

```
\begin{tabular}{S S}
  {Speed (m/s)} & {Time (s)} \\
  \midrule
  1.23 & 45.6 \\
  0.012 & 789.1 \\
  123 & 0.001 \\
\end{tabular}
```

Speed (m/s)	Time (s)
1.23	45.6
0.012	789.1
123	0.001

The numbers in both columns will be aligned so that the decimal separators (periods or commas) are located below each other. Text in column headings should be enclosed in curly braces to prevent it from trying to align as a number.

Can be used in conjunction with the `tabularray` [1] package.

For this to work, you must include `tabularray` and explicitly load its `siunitx` library. This is done with the **`\UseTblrLibrary`**`{siunitx}` command. The `siunitx` package will be loaded automatically. To pass global options to `siunitx`, the **`\PassOptionsToPackage`** command must be used before loading the library.

```
\documentclass{article}
\PassOptionsToPackage{locale=DE}{siunitx}
\usepackage{tabularray}
\UseTblrLibrary{siunitx} % This command will load siunitx with the options
↪ specified above.
```

```
\begin{document}
\end{document}
```

The main strength of `siunitx` in tables is the ability to align numbers by decimal separators using a column of type `S`. In `tabularray`, this column becomes available after including the library. In the `colspec` argument, you can use `S` just like in regular LaTeX tables. The header must be enclosed in curly braces to prevent it from being aligned as a number.

Tabularray Table

```
\begin{tblr}{ colspec = { l S } }
  Element & {Value} \\
  A & 123.456 \\
  B & 2.34 \\
  C & 5678.9 \\
\end{tblr}
```

Element	Value
A	123.456
B	2.34
C	5678.9

Thanks to built-in support, you can use almost all the capabilities of `siunitx` within `tabularray`.

You can set column-specific `siunitx` options directly in `colspec`, for example, to control the number of characters.

Tabularray Table

```
\begin{tblr}{ colspec = { S[table-format=3.2] } }
  12.3 \\
  456.78 \\
  9.01 \\
\end{tblr}
```

12.3
456.78
9.01

Since `siunitx` is loaded, you can use its commands in table cells. For example, you can separate units of measurement into a separate column and format them using **`\unit`**.

Tabularray Table

```
\begin{tblr}{
    colspec = { S[table-format=1.3] l }, % The first column is numbers, the second is
     ↪ for \unit commands
    column{2} = { cmd = \unit }, % The contents of the second column will be the
     ↪ argument for \unit
  }
  \SetCell[r=1,c=2]{c} {Physical~quantities} \\ % Combined header
  1.234 & \metre                  \\
  0.835 & \candela                \\
  4.23  & \joule\per\mole         \\
\end{tblr}
```

Physical quantities	
1.234	m
0.835	cd
4.23	$\mathrm{J\,mol^{-1}}$

5. Global settings

Here are some useful options for **\sisetup**.

- `locale = DE` — quick setup for European standards (comma, spaces).
- `output-decimal-marker = {,}` — explicitly set the comma as the decimal separator.
- `group-separator = {\,}` — set the character to separate groups of digits (e.g., thin space).
- `range-phrase = {\,--\,}` — set the text for the range (default: “to”, can be replaced with a dash).
- `per-mode = symbol` — how to display a symbol to the minus first power: $/s$ instead of s^{-1}. You can also use `fraction` to display it as a fraction.
- `exponent-mode = scientific` — force all numbers to scientific notation.

Author Contributions: The contributions of the authors are equal. All authors have read and agreed to the published version of the manuscript.

Funding: This research received no external funding.

Data Availability Statement: No new data were created or analysed during this study. Data sharing is not applicable.

Conflicts of Interest: The authors declare no conflict of interest.

Declaration on Generative AI: The authors have not employed any Generative AI tools.

References

1. Kulyabov, D. S., Korolkova, A. V., Sevastianov, L. A. & Rybakov, Y. P. Typesetting tables. *Discrete and Continuous Models and Applied Computational Science* **33,** 235–241. doi:10.22363/2658-4670-2025-33-3-235-241 (Oct. 2025).

Information about the authors

Kulyabov, Dmitry S.—Professor, Doctor of Sciences in Physics and Mathematics, Professor of Department of Probability Theory and Cyber Security of RUDN University; Senior Researcher of Laboratory of Information Technologies, Joint Institute for Nuclear Research (e-mail: kulyabov-ds@rudn.ru, ORCID: 0000-0002-0877-7063, ResearcherID: I-3183-2013, Scopus Author ID: 35194130800)

Korolkova, Anna V.—Docent, Candidate of Sciences in Physics and Mathematics, Associate Professor of Department of Probability Theory and Cyber Security of RUDN University (e-mail: korolkova-av@rudn.ru, ORCID: 0000-0001-7141-7610, ResearcherID: I-3191-2013, Scopus Author ID: 36968057600)

Sevastianov, Leonid A.—Professor, Doctor of Sciences in Physics and Mathematics, Professor of Department of Computational Mathematics and Artificial Intelligence of RUDN University (e-mail: sevastianov-la@rudn.ru, ORCID: 0000-0002-1856-4643, ResearcherID: B-8497-2016, Scopus Author ID: 8783969400)

Rybakov, Yuri P.—Professor, Doctor of Sciences in Physics and Mathematics, Professor of the Institute of Physical Research and Technologies of RUDN University (e-mail: rybakov-yup@rudn.ru, ORCID: 0000-0002-7744-9725, ResearcherID: S-4813-2018, Scopus Author ID: 16454766600)

DOI: 10.22363/2658-4670-2026-34-1-5-11 EDN: VDNGIA

Набор физических размерных величин

Д. С. Кулябов[1,2], А. В. Королькова[1], Л. А. Севастьянов[1,2], Ю. П. Рыбаков[1]

[1] Российский университет дружбы народов, ул. Миклухо-Маклая, д. 6, Москва, 117198, Российская Федерация

[2] Объединённый институт ядерных исследований, ул. Жолио-Кюри, д. 6, Дубна, 141980, Российская Федерация

Аннотация. Пакет `siunitx` предназначен для типографски правильного и согласованного набора физических величин (чисел с единицами измерения) в документах LaTeX. Он автоматизирует форматирование в соответствии с правилами Международной системы единиц (СИ), что избавляет от необходимости вручную следить за пробелами, шрифтами и разделителями.

Ключевые слова: набор физических единиц, LaTeX

Discrete & Continuous Models
& Applied Computational Science
ISSN 2658-7149 (Online), 2658-4670 (Print)

2026, 34 (1) 12–23
http://journals.rudn.ru/miph

Research article

UDC 519.7
PACS 07.05.Tp
DOI: 10.22363/2658-4670-2026-34-1-12-23 EDN: URRPMT

Usage of polynomial representation of numbers for approximate homomorphic encryption

Andrey E. Krouk

Saint-Petersburg State University of Aerospace Instrumentation, 67 B. Morskaia St, 190000, Saint-Petersburg, Russian Federation

(received: December 16, 2025; revised: January 25, 2026; accepted: January 30, 2026)

Abstract. *Introduction* In the modern world of computers and networks the idea of expanding of personal computer resources with the help of cloud storages and computation looks more and more lucrative. However, usage of these resources may endanger data being processed. In last twenty years several algorithms of homomorphic encryption were developed allowing solving of this problem among other applications. However such algorithms are usually constructed as public key systems for long term storage and processing of data. In this article two algorithms of homomorphic encryption optimized for single data processing are proposed. *Purpose* The target of research is development of data coding system which allows safe data processing in public clouds. *Results* Two homomorphic coding systems had been developed, first is based on representation of numbers in the form of polynomials, second based on further representation of polynomials in the form of sets of values. Developed systems allow approximate calculations of coded data without decryption allowing processing of real numbers. System has high level of protection and provides high precision of calculations, comparable with standard personal computer calculation precision. Structure of coded data allows parallel computing. Proposed system allows safe data processing in public networks. Question of finding of optimal parameters for the system stands open both for high precision calculation of limited sets of operations and repeatedly good precision for big sets of operations.

Key words and phrases: homomorphic encryption, cloud calculations

For citation: Krouk, A. E. Usage of polynomial representation of numbers for approximate homomorphic encryption. *Discrete and Continuous Models and Applied Computational Science* **34** (1), 12–23. doi: 10.22363/2658-4670-2026-34-1-12-23. edn: URRPMT (2026).

1. Introduction

People have long been interested in the idea of hidden computing. The need to decrypt data for subsequent processing made it vulnerable. Now, with the introduction of public networks and the ability to process data on external, often public, resources, the task of performing hidden computations on encrypted data has become extremely relevant. To address this challenge, in the mid-20th century, the idea of homomorphic encryption was proposed, a method of transforming (encrypting) numerical data that allows operations to be performed on encrypted data without decryption [1, 2].

First, so-called partially homomorphic encryption systems were developed. These systems typically allowed for one type of arithmetic operation without decryption. For example, the RSA and ElGamal systems [3] enable multiplication without decryption, while the Paillier and Benaloh systems enable addition.

For a long time, attempts to create a fully homomorphic encryption system or combine several partially homomorphic encryption systems into one have been unsuccessful. The turning point was the introduction of a system developed by Craig Gentry in 2009 [4]. The system is based on ideal lattices. Numerical noise is added to the data, making it impossible to decrypt without knowing the key. The system allows for any arithmetic operations without decrypting the data. As the number of operations increases, the noise level increases as well, but the system includes a procedure to reduce the noise level. This allows for an arbitrary number of operations without decrypting the data, but at the cost of increased computational complexity.

Development of homomorphic encryption systems has significantly expanded the possibilities for conducting calculations in cloud and collaborative systems [1, 5].

In 2010, a revised version of Gentry's scheme based on integers was introduced [6]. Despite using a different mathematical foundation, this scheme employs Gentry's method of encryption through the addition of noise and a noise reduction procedure. However, many fully homomorphic encryption systems are based on lattices [7].

Along with systems that perform precise arithmetic operations, usually on integers, systems that perform approximate arithmetic operations have begun to appear, which are applicable to performing operations on real numbers. In 2016, the CKKS system [8] was released, the first system that performs approximate calculations and is designed to work with real and complex numbers. In this system, data is first represented as circular polynomials, and then the problem of learning with errors in the ring is solved to create a homomorphic encryption system that has a public key for encrypting data. However, this system is designed to perform a finite number of operations, as increasing the number of operations increases the complexity of the calculations and negatively affects the system's security. The re-encryption procedure proposed a little later [9] only partially solves the problem, as its use leads to an increase in the complexity of calculations, which does not allow it to be used too often [10]. A detailed comparison of the CKKS system [8, 11] with systems that perform exact calculations, such as the BFV system [12–14], is provided in [15]. It should also be noted that although the presence of a public key opens up additional possibilities for using the system, it is not necessary for solving certain problems and negatively affects the system's security [16, 17].

This article presents several variants of approximate fully homomorphic encryption systems that use a fundamentally new approach to creating systems: representing numbers as polynomials and, additionally, replacing the represented polynomials with a set of their values at points. Although this approach does not allow for the use of a public key, it provides near-absolute security by using session-specific keys that are not transmitted outside of the trusted PC.

This method can be effectively used to process large amounts of data on external (unsecure) computing resources (so-called public cloud computing).

2. Usage of polynomials for construction of partially homomorphic encryption system

Simple homomorphic encryption system can be built using residue number system based on Chinese Reminder Theorem (CRT) [18, 19].

This system can be further developed by replacement of residual number system with a polynomial representation. Let's discuss this approach in more detail.

Let's suppose that a sequence of arithmetic operations has to be performed on a set of numbers. We will encode the data for calculations in two stages:

1. We assign to each number a polynomial such that the value of this polynomial at some point x_s is equal to the number. The value of x_s is the same for all pairs of number-polynomial and is the secret key of the system.
2. We will choose a set of points. Values of polynomials in these points we will use to define these polynomials. (It is preferable to choose points where the polynomial values can be calculated relatively easily.) We will assign a set of values in the selected points to each polynomial. This set of values will be the result of the encoding process.

Let's combine into sets the values of all selected polynomials at each of the points (each set consists of the values of all polynomials in one of the points). Now, to perform any sequence of arithmetic operations on the original numbers, it is enough to perform the same sequence of operations within each set. Indeed, since the result of adding (multiplying) two polynomials will be a polynomial whose values at any point will be the addition (multiplication) of the values of the polynomial terms (multipliers) at these points, the resulting values match the values of the polynomial that would result from performing the given sequence of arithmetic operations on the polynomials corresponding to the original data. Let us call this polynomial the result polynomial.

The degree of this result polynomial (m) can be easily estimated by the degrees of the original polynomials. Indeed, when adding polynomials, the degree of the result can be estimated by the highest degree of the terms, and when multiplying, it can be estimated by the sum of the degrees of multipliers.

Thus, when we receive $m + 1$ calculated values back, we can perform polynomial interpolation and obtain the result polynomial.

By calculating the value of the result polynomial in the secret point x_s, we can find the final result of the calculation.

Let's evaluate the advantages and disadvantages of the proposed method. This method retains both the many advantages of the original method (simplicity of calculations, ease of parallelization of calculations) and the main disadvantage (the absence of a division operation). At the same time, the presence of a two-step encoding operation and a secret key allows for high data security without the use of false requests or other additional techniques for data protection.

3. Approximate calculations

Integer homomorphic encryption systems do not allow calculations with non-integer numbers. When working with real numbers, in particular when performing division, it is necessary to perform approximate calculations.

Approximate calculations are widely used, especially in technical and physical problems. This is because many data can only be measured with finite precision, and many elements can also be produced with finite precision. Increasing the precision of calculations leads to an excessive increase in the complexity [8, 20, 21]. In the same time there is no reason to increase the precision of calculations beyond the precision of measurements or production, as the precision of the final result is determined by the precision of the most inaccurate value.

When a computer works with real variables, it also limits the precision of the calculations, and these calculations are strictly speaking approximate.

4. Representation of numbers as polynomials

Let's improve the encoding procedure so that we can implement an approximate division procedure. This time, we will encode the data in stages and evaluate the resulting system at each stage.

Let's replace operations on numbers with operations on polynomials. To do this, we assign to each number a polynomial so that the value of this polynomial at some point x_s is equal to the number. The value of x_s must be the same for all pairs of number–polynomial and is the secret key of the system (we will call it the "secret point").

In this representation, it is easy to implement addition and multiplication operations, but there are problems with implementing division. Indeed, in most cases, it is impossible to divide polynomials completely, meaning that there will be a remainder when dividing.

$$\frac{f(x)}{g(x)} = q(x) + \frac{r(x)}{g(x)}.$$

However, this problem can be resolved by considering division as an approximate operation and choosing x to be sufficiently big. Indeed, since the degree of the remainder $r(x)$ is less than the degree of the divisor $g(x)$, the following formula holds

$$\frac{r(x)}{g(x)} \underset{x\to\infty}{\longrightarrow} 0.$$

However, this imposes restrictions on the encoding procedure. In order to use this formula, it is necessary that x be sufficiently big, meaning that the value (weight) of the leading term of the polynomial at the secret point would be significantly greater than the sum of the values of the other terms.

In addition, it is necessary to complicate the system somewhat in order to be able to divide polynomials in the case when the degree of the dividend is less than the degree of the divisor. We will use not polynomials to represent numbers, but constructions $f(x)/x^k$, i.e. polynomials divided by the degree of x. In this representation, when performing the division operation, we get the opportunity to multiply and divide the dividend by x^k (division will be carried out after calculating the result of the main operation, by simply adding the corresponding degree of x to the denominator) so that its degree exceeds the degree of the divisor. Choosing bigger values of k allows to improve the operation precision. In addition, this method can be used to improve the precision of division even when the degree of the dividend is greater than the degree of the divisor.

5. Homomorphic encryption system based on representation of numbers as polynomials

Let's build a homomorphic encryption system that satisfies the principles described in the previous paragraph. We will calculate multiplication, addition, and subtraction in the traditional exact way, and division in an approximate way. To do this, we will assign to each number in our system a polynomial divided by the power of x ($f(x)/x^k$) such that the values of the polynomials at a secret point (the secret point is the same for all polynomials) are equal to the encoded numbers, and replace operations on numbers with operations on the coefficients of the polynomials. It is impossible to restore the values of the numbers in the system without knowing the secret point. The absence of the need to transmit this secret value and the ability to choose a new secret value for each session ensure high security of the system.

Operations on polynomial coefficients are implemented as follows.

– Division. An approximate operation. First, the degree of the dividend is increased by multiplying it by x^k, and then division with remainder is performed. The incomplete quotient is used as the result of the division, and the remainder is neglected. Increasing the degree of x^k allows for greater precision in division, at the cost of increase in the complexity of the calculations and potentially the complexity of decryption. Finally, x^k is included in the weight factor (denominator) of the element.

  For example, when dividing $x + 1$ by x^3, you can multiply and divide $x + 1$ by x^2

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

  If x^5 is used instead of x^2, the result will be

$$\frac{x^6 + x^5}{x^3}\frac{1}{x^5} = \left(x^3 + x^2\right)\frac{1}{x^5},$$

  this provides greater accuracy, as the polynomials are completely divided.
– Multiplication. Polynomials are multiplied classically, and the weight factors (denominators) x^k are also multiplied. For example:

$$(x + 1)x^{-1}(x + 2)x^{-2} = (x^2 + 3x + 2)x^{-3}.$$

– Addition and subtraction. To add and subtract elements, they should first be brought to a common weight factor (denominator) (will be x with the highest absolute value of the degree), and then the polynomials are added or subtracted using the classical method. For example:

$$(x + 1)x^{-1} + (x + 2)x^{-2} = (x^2 + x)x^{-2} + (x + 2)x^{-2} = (x^2 + 2x + 2)x^{-2}.$$

It is necessary to choose an encoding algorithm to implement the system within the restrictions described in the previous paragraph. That is, it is necessary to choose the value of the secret point to be sufficiently big, and the coefficients of the polynomials to be chosen so that the leading term of the polynomials is sufficiently heavy (that is, so that the leading term contains most of the value of the number; with 70% of the value, the error is approximately 10 times greater than with 90%).

The following algorithm was used during the analysis of the system.

– The degree of the polynomial and the weight of the leading term as a percentage are selected. These are the parameters of the algorithm.
– The coefficient at the highest power is calculated so that the leading term has a value equal to the specified part of the encoded value.
– The next coefficient is chosen so that the value of the second term is equal to the specified part of the difference between the required value and the value of the leading term. And so on.
– The constant term is selected as the difference between the required value and the values of all other terms, ensuring that the required value is accurately matched.

Let's consider the algorithm's operation using the following example. Let's encrypt the number 100 using a polynomial of the second degree at the point 10 with a weight of 80% for the leading term.

– With the selected parameters, the value of the leading term should be $100 \times 0.8 = 80$. Since the degree of the polynomial is 2, the coefficient for the x^2 should be $80/10^2 = 0.8$.
– The value of the second term should be $(100 - 80) \times 0.8 = 16$. Since the degree of the second term is 1, the coefficient for the x should be $16/10 = 1.6$.
– Finally, the value of the constant term is obtained as $100 - 80 - 16 = 4$.

Thus, the number 100 is associated with a polynomial $0.8x^2 + 1.6x + 4$.

When encoding using this algorithm, the weight factors x^{-k} of the polynomials obtained during the encoding phase are always assumed to be equal to 1 ($k = 0$). However, during the division operation, the weight factor may change and participate in further calculations.

Additional measures can be taken to improve security of the system:

- Variable polynomial degrees: the polynomials describing the different points must have different degrees, but at least 2.
- Variable weight of the leading term: the weight of the leading term for each polynomial (or even for each term in each polynomial) is chosen as a random number within a specified range that ensures acceptable precision, such as between 70% and 90%.

To decrypt, it is enough to calculate the value of the polynomial at the secret point (taking into account the weight coefficient x^{-k}).

Let's evaluate the advantages and disadvantages of the proposed system. The main advantages of this system are its completeness (it supports all four arithmetic operations) and its high security due to the use of one-time, non-transmittable secret keys. The main disadvantage of this system is its computational complexity, as it requires operations on sets of numbers instead of individual numbers, and it does not support easy parallelization of calculations, which is very convenient for external networks.

6. Usage of interpolation for operations with polynomials

To increase the security of the system, as well as to facilitate parallel calculations, each polynomial can be represented as a set of values at points. After that, operations are performed not on the coefficients of the polynomial, but on these values. As the operations on the values at different points are independent of each other, they can be performed in parallel or, for example, as separate tasks for cloud computing (including in the public cloud, due to high security of the system).

There is no point in describing these arithmetic operations in detail, as they are literally operations on values (numbers) — addition, subtraction, multiplication, or division of values.

In principle, to speed up calculations, it is possible to use the values of polynomials at small points, but research has shown that this leads to a noticeable loss of division operation precision. However, this allows for the construction of an effective partially homomorphic encription system for three operations (addition, subtraction, and multiplication), with a high computational speed that can compete with modular arithmetic (such system was described above). At the same time, as long as the values used in the calculations (or a part of them) are comparable to the secret value, the precision of the operations is sufficient, allowing for the freedom to choose specific values or patterns for their generation. For example, to simplify calculations, values of the form 2^k can be used as long as at least one of them is greater than the secret value.

Encoding process follows these steps:

- Encoding of values with polynomials (as described in the previous paragraph).
- Selecting of a set of points and calculating the values of the polynomials at those points.

The values of the points in the selected set are also a secret, and are also not transmitted anywhere, which contributes to the high security of the system. The number of points in the set is selected based on the estimation of the degree of the polynomial that should be obtained as a result of the calculations.

Decryption process follows these steps:

- Estimating of the x^{-k} weight coefficient on your own computing base (this is easy to do because all operations process it uniquely) and the degree of the resulting polynomial (to determine the number of points required for decryption)
- Calculating of the values of the polynomial at all points, taking into consideration the weight coefficient x^{-k} (that is, multiply the obtained values by x^k).
- Restoring of the polynomial using the interpolation method.
- Calculating the value at the secret point (again, taking into consideration the x^{-k} weight coefficient).

To further improve security and reduce waiting times, an excessive set of points can be used (decoding can be performed as soon as sufficient number of values is obtained). This reduces the impact of various delays that occur both on data-processing servers and during packet transition through the network (so-called transport coding) [22].

It should be noted that in order to improve the precision of the system, when evaluating the weight coefficient, it is possible to include a multiplication-division operation by a weight coefficient of a relatively high degree in division operation (as in description of division in the previous section), which allows for a more precise division operation. Let's refer to the degree of x used for the multiplication-division operation as the correction and use it as a parameter for evaluating the precision of the operations. Using higher values of the correction can result in an increase in the degree of the final polynomial, which can lead to increase in complexity of the decryption process and number of points required for decryption.

When using the interpolation representation of polynomials, the coefficients of the polynomials and their degrees are hidden. This makes it unnecessary to use the additional security measures mentioned in the previous paragraph, as they can negatively impact precision. Moreover, using the interpolation representation allows for the selection of a specific structure of the polynomials to facilitate subsequent calculations, such as encoding numbers with monomials of a given degree (i.e., using 100% weight in the leading term) without compromising security.

Let's evaluate the advantages and disadvantages of the proposed system. Usage of interpolation for polynomial operations further enhances the system's security while reducing overall computational complexity by representing data as independent sets that can be processed in parallel.

7. Results of experiments

The system was tested both with and without the interpolation representation of polynomials.

The system's functionality was tested without the use of interpolation representation. Proposed methods for increasing the system's security were also tested. As a result of these tests, the division precision sufficient for engineering calculations was proved. As this method is inferior to the interpolation representation, the purpose of these tests was to verify the feasibility of the idea, and multiple tests were not conducted to verify the precision.

Multiple tests were conducted for the interpolation representation. The interpolation point sets were selected according to the principle $(i + 1) \bullet 25$, where $i = 0, 1, ...$, as they allowed for more precise calculations, and 2^i, as using this set allows for faster calculations. The secret point was not part of either set. The results were obtained for different values of the correction parameter (the correction parameter was introduced in the section describing the use of interpolation for polynomial operations) and different numbers of interpolation points.

Table 1

Single division error with leading monomial weight of 95%

Correction	Number of points	Average error rate	Maximum error
Selecting points based on the principle $(i+1) \bullet 25$			
1	2	$2.6 \cdot 10^{-16}$	$1.5 \cdot 10^{-15}$
1	3	$1.5 \cdot 10^{-15}$	$7.1 \cdot 10^{-15}$
2	5	$1.5 \cdot 10^{-14}$	$7.8 \cdot 10^{-14}$
Selecting points based on the principle 2^i			
4	5	$3.1 \cdot 10^{-16}$	$1.6 \cdot 10^{-15}$

Table 2

Multiple operation (8 divisions, 7 multiplications, 2 additions) error with leading monomial weight of 95%

Correction	Number of points	Average error rate	Maximum error
Selecting points based on the principle $(i+1) \bullet 25$			
1	2	$5.8 \cdot 10^{-16}$	$3.2 \cdot 10^{-15}$
1	3	$2.1 \cdot 10^{-15}$	$1.4 \cdot 10^{-14}$
2	5	$1.6 \cdot 10^{-14}$	$7.6 \cdot 10^{-14}$
Selecting points based on the principle 2^i			
4	5	$4.9 \cdot 10^{-16}$	$2.4 \cdot 10^{-15}$

During the tests, two important results were achieved. First, by using a weight of 95% for the leading term in the encoding algorithm, the division precision was achieved at the level of the PC's precision in performing precise (addition and multiplication) arithmetic operations (due to the PC's inaccuracy in handling real-number variables) for real numbers (with a maximum deviation of 10^{-14} and an average deviation of 10^{-15}).

The results of the most interesting test runs are presented in Tables 1 and 2.

Secondly, a research was conducted on the growth of the error as a function of the number of operations performed. In this research, significantly less convenient encoding parameters were used, with the weight of the leading term for each point randomly selected between 80% and 90%. A large number of operations was achieved through consecutive divisions and multiplications (with one more division to ensure that the degree of the result polynomial was 0, before taking correction in consideration) of various randomly selected five-digit numbers. As a result, it was possible to find parameter values (albeit not optimal) under which the increase in error with increase in operations number is almost non-existent.

Examples of errors for such parameter values (with different numbers of operations) are given in Tables 3 and 4.

Table 3

Error for different number of consecutive operations

Number of operations	Average error rate	Maximum error
3	$2.4 \cdot 10^{-10}$	$1.3 \cdot 10^{-9}$
5	$2.4 \cdot 10^{-10}$	$1.4 \cdot 10^{-9}$
17	$2.4 \cdot 10^{-10}$	$1.3 \cdot 10^{-9}$

Encoding is used with a leading term weight of 80%–90%, and the values of the points (where the calculations are performed) are selected using the formula $(i + 1) \bullet 25$.

11 points are used for correction of 8.

Table 4

Error for different number of consecutive operations

Number of operations	Average error rate	Maximum error
3	$9.1 \cdot 10^{-8}$	$4.4 \cdot 10^{-7}$
5	$9.1 \cdot 10^{-8}$	$5.1 \cdot 10^{-7}$
17	$9.1 \cdot 10^{-8}$	$5.33 \cdot 10^{-7}$

Encoding is used with a leading term weight of 80%–90%, and the values of the points (where the calculations are performed) are selected using the formula 2^i.

14 points are used for correction of 12.

8. System security evaluation

The system decryption is performed in two stages. Let's try to assess the security of these stages.

1. At the first stage, a polynomial is determined based on the values. The main problem for unauthorized decryption at this stage is that the attacker does not have information about the degree of the polynomial (this information is calculated on the base computer and is not transmitted anywhere) and about the points where the polynomial values are calculated (these values are also not transmitted anywhere). The attacker may attempt to obtain some information by using points in the set that are close to the secret point, but he does not know which points in the set to use for evaluation, and the accuracy of the evaluation remains questionable.
2. In the second step, the value at the secret point is determined using the polynomials. To determine the value at the secret point, secret key is needed. This key is unique for each calculation and is not shared with anyone, so the attacker does not know its value. If there are several values in the result, the attacker can try to exploit the higher weight of the leading term and find the relationship between them by dividing one polynomial by another. The information obtained in this way depends on the specific polynomials used for the encoding. For example, if this attack is performed on a system used during research, the result will be inaccurate because the weight of the leading term is chosen with considerable random error during the encoding process.

9. Conclusions

- The paper proposes a method of approximate homomorphic encription that provides sufficient precision for engineering calculations. The conducted research allows to hope for the existence of encoding algorithms that ensure the precision of the system's operation at the precision level of PC operations with real numbers.
- The high security of the method and the ease of dividing calculations into parallel processes make it suitable for use in public cloud networks.
- The ability to increase the number of points and the division operation parameters allows to adjust the precision and speed of calculations.
- The conducted research shows that there are system parameters under which the errors of multiple operations do not accumulate significantly.

Author Contributions: Conceptualization, Krouk A. E.; methodology, Krouk A. E.; writing—review and editing Krouk A. E.; supervision, Krouk A. E.; project administration, Krouk A. E. All authors have read and agreed to the published version of the manuscript.

Funding: This research received no external funding.

Data Availability Statement: Data sharing is not applicable.

Conflicts of Interest: The author declares no conflict of interest.

Declaration on Generative AI: The author has not employed any generative AI tools.

References

1. Rivest, R. L., Adleman, L. & Dertouzos, M. L. *On Data Banks and Privacy Homomorphisms* in *Foundations of Secure Computation* (1978).
2. Marcolla, C., Sucasas, V., Manzano, M., Bassoli, R., Fitzek, F. H. P. & Aaraj, N. Survey on Fully Homomorphic Encryption, Theory, and Applications. *Proceedings of the IEEE* **110,** 1572–1609. doi:10.1109/JPROC.2022.3205665 (2022).
3. Van Tilborg, H. C. A. *Fundamentals of Cryptology* [in Russian] (Mir, Moscow, 2006).
4. Gentry, C. *A Fully Homomorphic Encryption Scheme* PhD thesis (Stanford University, 2009).
5. Brakerski, Z., Gentry, C. & Vaikuntanathan, V. *(Leveled) fully homomorphic encryption without bootstrapping* in *Proceedings of the 3rd Innovations in Theoretical Computer Science Conference (ITCS)* (Association for Computing Machinery, Cambridge, Massachusetts, 2012), 309–325. doi: 10.1145/2090236.2090262.
6. van Dijk, M., Gentry, C., Halevi, S. & Vaikuntanathan, V. Fully Homomorphic Encryption over the Integers. Cryptology ePrint Archive, Report 2009/616 (2010).
7. Albrecht, M. *et al. Homomorphic Encryption Standard* in *Protecting Privacy through Homomorphic Encryption* (eds Lauter, K., Dai, W. & Laine, K.) (Springer International Publishing, Cham, 2021). doi:10.1007/978-3-030-77287-1_2.
8. Cheon, J. H., Kim, A., Kim, M. & Song, Y. *Homomorphic Encryption for Arithmetic of Approximate Numbers* in *Advances in Cryptology – ASIACRYPT 2017* (eds Takagi, T. & Peyrin, T.) (Springer International Publishing, Cham, 2017), 409–437. doi:10.1007/978-3-319-70694-8_15.
9. Brakerski, Z. & Vaikuntanathan, V. *Fully Homomorphic Encryption from Ring-LWE and Security for Key Dependent Messages* in *Advances in Cryptology – CRYPTO 2011* (ed Rogaway, P.) **6841** (Springer Berlin Heidelberg, Berlin, Heidelberg, 2011), 505–524. doi:10.1007/978-3-642-22792-9_29.

10. Cheon, J. H., Han, K., Kim, A., Kim, M. & Song, Y. *Bootstrapping for Approximate Homomorphic Encryption* in *Advances in Cryptology – EUROCRYPT 2018* (eds Nielsen, J. B. & Rijmen, V.) (Springer International Publishing, Cham, 2018), 360–384. doi:10.1007/978-3-319-78381-9_14.
11. Chen, H., Laine, K. & Player, R. *Simple Encrypted Arithmetic Library – SEAL (v2.1)* in *Financial Cryptography and Data Security* (eds Brenner, M., Rohloff, K., Bonneau, J., Miller, A., Ryan, P. Y., Teague, V., Bracciali, A., Sala, M., Pintore, F. & Jakobsson, M.) **10323** (Springer International Publishing, Cham, 2017), 3–18. doi:10.1007/978-3-319-70278-0_1.
12. Lepoint, T. & Naehrig, M. *A Comparison of the Homomorphic Encryption Schemes FV and YASHE* in *Progress in Cryptology – AFRICACRYPT 2014* **8469** (2014), 318–335. doi:10.1007/978-3-319-06734-6_20.
13. Bajard, J.-C., Eynard, J., Hasan, M. A. & Zucca, V. *A Full RNS Variant of FV Like Somewhat Homomorphic Encryption Schemes* in *Cryptographic Hardware and Embedded Systems – CHES 2016* (eds Avanzi, R. & Heys, H.) **10532** (2016), 423–442. doi:10.1007/978-3-319-69453-5_23.
14. Halevi, S., Polyakov, Y. & Shoup, V. *An Improved RNS Variant of the BFV Homomorphic Encryption Scheme* in *Topics in Cryptology – CT-RSA 2019* (ed Matsui, M.) **11405** (2019), 83–105. doi:10.1007/978-3-030-12612-4_5.
15. Babenko, M. G., Golimblevskaia, E. I. & Shiriaev, E. M. Comparative Analysis of Homomorphic Encryption Algorithms Based on Learning with Errors. *Proceedings of the Institute for System Programming of the RAS (Proceedings of ISP RAS)* **32,** 37–51. doi:10.15514/ISPRAS-2020-32(2)-4 (2020).
16. Gentry, C. *Fully Homomorphic Encryption Using Ideal Lattices* in *Proceedings of the 41st Annual ACM Symposium on Theory of Computing (STOC)* (2009), 169–178. doi:10.1145/1536414.1536440.
17. Pulido-Gaytan, B., Tchernykh, A., Cortés-Mendoza, J. M., Babenko, M., Radchenko, G., Avetisyan, A. & Drozdov, A. Y. Privacy-preserving Neural Networks with Homomorphic Encryption: Challenges and Opportunities. *Peer-to-Peer Networking and Applications* **14,** 1666–1691. doi:10.1007/s12083-021-01076-8 (2021).
18. Akritas, A. *Elements of Computer Algebra with Applications* 425 pp. (John Wiley & Sons, Inc., United States, 1989).
19. Krouk, A. E. & Fedorenko, S. V. *Construction of the solution of the Chinese Remainder Theorem for polynomials using the method of undetermined coefficients* in *2019 XVI International Symposium "Problems of Redundancy in Information and Control Systems" (REDUNDANCY)* (Moscow, Russia, 2019), 115–116. doi:10.1109/REDUNDANCY48165.2019.9003344.
20. Halevi, S. & Shoup, V. *Bootstrapping for HElib* in *Advances in Cryptology – EUROCRYPT 2015* (eds Oswald, E. & Fischlin, M.) **9056** (2015), 641–670. doi:10.1007/978-3-662-46800-5_25.
21. Li, B. & Micciancio, D. *On the Security of Homomorphic Encryption on Approximate Numbers* in *Advances in Cryptology – EUROCRYPT 2021* **12696** (2021), 648–677. doi:10.1007/978-3-030-77870-5_23.
22. Kabatiansky, G., Krouk, E. & Semenov, S. *Error Correcting Coding and Security for Data Networks. Analysis of the Super Channel Concept* doi:10.1002/0470867574 (John Wiley & Sons, Ltd., 2005).

Information about the authors

Krouk, Andrey E.—Candidate of Technical Sciences, Associate Professor (e-mail: svinenka@mail.ru, ORCID: 0009-0008-1162-2020)

УДК 519.7
PACS 07.05.Tp
DOI: 10.22363/2658-4670-2026-34-1-12-23 EDN: URRPMT

Использование представления чисел в виде многочленов для реализации скрытых приближённых вычислений

А. Е. Крук

Санкт-Петербургский государственный университет аэрокосмического приборостроения,
ул. Б. Морская, д. 67, Санкт-Петербург, 190000, Российская Федерация

Аннотация. *Введение* В современном мире компьютеров и сетей всё более привлекательной выглядит возможность расширения ресурсов персонального компьютера за счет облачных хранилищ и вычислений, однако, использование таких ресурсов может поставить под угрозу безопасность обрабатываемых данных. В последние двадцать лет появилось множество алгоритмов гомоморфного шифрования, позволяющих в частности решить эту задачу. Однако эти алгоритмы проектируются в основном как системы с открытым ключом, предназначенные для долгосрочного хранения и обработки данных. В данной статье предлагается два алгоритма гомоморфного шифрования, оптимизированных для однократной обработки данных. *Цель* Целью работы является разработка системы кодирования информации, позволяющей проводить безопасную обработку данных в публичных облаках. *Результаты* Разработаны две системы скрытых вычислений: первая, основанная на представлении чисел в виде многочленов и вторая, основанная на дальнейшем представлении многочленов в виде набора значений. Разработанные системы позволяют проводить приближённые вычисления над зашифрованными данными без их расшифровки, что позволяет проводить обработку вещественных чисел. Система отличается высоким уровнем защиты и обеспечивает высокую точность вычислений, сравнимую с точностью обеспечиваемой стандартными вычислениями компьютера. Структура зашифрованных данных позволяет проведение параллельных вычислений. Предложенная система позволяет безопасную обработку данных в публичных облачных сетях. Остаётся открытым вопрос оптимальных параметров системы защиты информации, обеспечивающих более высокую точность для ограниченного набора операций, либо постоянную точность для больших наборов операций.

Ключевые слова: облачные вычисления, гомоморфное шифрование

Discrete & Continuous Models & Applied Computational Science
ISSN 2658-7149 (Online), 2658-4670 (Print)

2026, 34 (1) 24–39
http://journals.rudn.ru/miph

Research article

UDC 519.872
PACS 02.50.−r
DOI: 10.22363/2658-4670-2026-34-1-24-39 EDN: VCZSIW

The waiting time extremal index in GI/G/1 system

Irina V. Peshkova[1, 2]

[1] Petrozavodsk State University, 33 Lenina Pr, Petrozavodsk, 185910, Russian Federation

[2] Institute of Applied Mathematical Research of the KarRC RAS, 11 Pushkinskaya St, Petrozavodsk, 185910, Russian Federation

(received: February 3, 2026; revised: February 15, 2026; accepted: February 16, 2026)

Abstract. In this paper the conditions to compare the extremal index of the stationary waiting time in the $M/G/1$ and $GI/M/1$ systems are obtained. These conditions include exponential asymptotic behaviour of waiting time tail and the order in failure rates for the interarrival intervals and for the service times in the systems to be compared. For $M/G/1$ system the obtained result is extended to the mixed service times with ordered components. If, in a $GI/G/1$ system, the service time is determined by a finite mixture whose dominant component of the equilibrium distribution belongs to the class of subexponential distributions then the tail of the limiting distribution of the stationary waiting time is equivalent to the tail of this distribution up to a constant obtained explicitly. Furthermore, the limiting distribution of the maximum of the stationary waiting time belongs to the maximum domain of attraction of the distribution of extreme values of the same type as the maximum of the random variables defined by the dominant component.

Key words and phrases: extremal index, queueing system, order in failure rate

For citation: Peshkova, I. V. The waiting time extremal index in GI/G/1 system. *Discrete and Continuous Models and Applied Computational Science* **34** (1), 24–39. doi: 10.22363/2658-4670-2026-34-1-24-39. edn: VCZSIW (2026).

1. Introduction

Understanding whether extreme events happen independently or in groups is crucial for forecasting them and minimizing their impact. Extreme value theory can accommodate clustering via so-called *extremal index* θ which, measures the size of the cluster and thus has an appealing physical meaning [1–4].

When $\theta = 1$, extremes behave like a Poisson process, occurring in isolation. However, when $\theta < 1$, extremes occur in groups (form clusters), following a compound Poisson process. In this case, $1/\theta$ estimates the mean cluster size or, equivalently, the mean time spent above the threshold. The limiting distribution of the maximum value in a stationary sequence is directly shaped by θ, revealing the local dependence within the data.The extremal index therefore gives a measure of the fraction of extremes that are approximately independent and identically distributed (i.i.d.) [5]. The extreme case of $\theta = 0$ indicates total dependence, where exceedances form very wide clusters. In practice, this means a sufficiently high threshold may never be crossed. Conversely, independent sequences always have $\theta = 1$, with high thresholds exceeded only by isolated events.

It is important to mention that there are other definitions of clustering and extremal index in the literature [6, 7].

From a practical point of view, the extremal index is useful for estimating the size of clusters or the average length of intervals between the exceedances. In the telecommunications the interest is the estimation of the risk to lose customers with maximum waiting times (deadlines) exceeding the threshold. The study of extremal metrics in queueing theory relies on applying extreme value theory to regenerative processes. A central objective is determining the limiting distribution for the maxima of waiting times, virtual waiting times, or queue lengths, see for example, [8–11]. The extreme value theory for independent and *one-dependent* regeneration processes is developed in [11]. The algorithm of computing the extremal index of the stationary waiting time of a stable $G/G/1$ system with distribution belonging to the domain of attraction of Gumbel distribution is given in [9]. The study of the distribution of the cluster and inter-cluster sizes is also an actual problem (see, for example, results for the Lindley process in $GI/G/1$ in [12]).

For a stable $GI/G/1$ system, if a non-zero solution $\gamma > 0$ exists for the equation $e^{\gamma(S-\tau)} = 1$, the maximum waiting time converges to a Gumbel distribution. The extremal index θ in this case is often amenable to explicit or numerical computation.

For the case of subexponential distributions of service times distributions, the parameter $\gamma = 0$, and the asymptotics of extreme values are studied using alternative methods based on the tail behavior of the waiting times themselves.

The main contribution of this paper is to establish conditions under which the extremal indexes of two queuing systems can be compared. Another objective is to determine which monotonicity properties in terms of the extremal index can be established for the systems with mixed service times.

The paper is organized as follows. In Section 2, we present known results on the extremal index for $GI/M/1$ and $M/G/1$ systems. In Section 3 Theorems 1 and 2 were proved. They establish the conditions for comparing the extremal indexes of stationary waiting times in $M/G/1$ and $GI/M/1$, respectively. For $M/G/1$ system the obtained result is extended to the case of a system with mixed service times with ordered components (Section 3.1). In Section 3.2 we extend results obtained for $GI/M/1$ system to multiserver $GI/M/c$. In Section 4 we investigate the class of limiting distributions of the stationary waiting time in $GI/G/1$ system with service time determined by a finite mixture whose dominant component of the equilibrium distribution belongs to the subexponential distributions.

2. The extremal index in GI/G/1 system

Let $GI/G/1$ system have i.i.d. service times, $\{S_i, i \geqslant 1\}$ and i.i.d. interarrivals, $\{\tau_i, i \geqslant 1\}$. Consider a reflected random walk (Lindley process), given by the recursion

$$W_{i+1} = (W_i + X_i)^+, \quad i \geqslant 1, \tag{1}$$

where $(x)^+ = \max(0, x)$ and $X_i = S_i - \tau_i$ — i.i.d. non-lattice r.v.s with common distribution function (d.f.) F_X, $\mathrm{E}X < 0$. Also assume that there is a γ such that

$$\mathrm{E}e^{\gamma X} = 1. \tag{2}$$

Let $Y_n = X_1 + \cdots + X_{n-1}$, $n \geqslant 1$ ($Y_0 = 0$). The actual waiting time W_n (or just waiting time) of customer n is the time from arrival t_n to the system until service starts. This process $\{W_n, n \geqslant 1\}$ is a Lindley process (1) generated by $\{Y_n, n \geqslant 0\}$ [8]. In particular, $W_n \stackrel{d}{=} \max\limits_{0 \leqslant k \leqslant n} Y_k$ and, if $\rho = \mathrm{E}S/\mathrm{E}\tau < 1$ (corresponds to $\mathrm{E}X < 0$), then a limiting steady-state distribution exists. By W we denote a random variable having the steady-state distribution of actual waiting time process $\{W_n\}$,

$$W_n \Rightarrow W, \quad n \to \infty.$$

It is obvious that the behaviour of the tail of $\mathrm{P}(\max Y_n > x)$ is determined by the components of X_i. Next we consider two cases: the real positive root of the equation (2) $\gamma > 0$. This case refers to light-tailed distributions of X_i and exponential asymptotic behaviour of waiting time tail. In contrast, for subexponential distributions of X_i we get $\mathrm{E}e^{\gamma X} = \infty$ for any $\gamma > 0$ [8].

Let $\overline{\mathrm{F}_S}(x) = 1 - \mathrm{F}_S(x)$ be the tail of d.f. $\mathrm{F}_S(x)$ of r.v. S. A d.f. F_S is called *subexponential* if

$$\lim_{x \to \infty} \frac{\overline{\mathrm{F}_S^{*n}}(x)}{n\overline{\mathrm{F}_S}(x)} = 1 \quad \text{for all } n \geqslant 2,$$

where $\overline{\mathrm{F}_S^{*n}}(x)$ is the tail of the n-fold convolution of the distributions $\mathrm{F}_S(x)$ with itself, i.e., $\overline{\mathrm{F}_S^{*n}}(x) = \mathrm{P}(S^1 + \cdots + S^n > x)$, where S^i is the stochastic copy of S, $i = 1, \ldots, n$.

We denote the class of subexponential distributions by $\mathcal{S}$. We also denote by S_e the stationary residual service time given by the probability density function $\overline{\mathrm{F}_S}(x)/\mathrm{E}S$, and let $\mathrm{F}_{S_e}(x)$ be the d.f. of S_e.

It is known [8, 10] that if $\rho < 1$ the system is stationary, and equation (2) has a real positive solution $\gamma > 0$ [13], then:

1) if $\mathrm{E}[Xe^{\gamma X}] < \infty$, then the tail d.f. of $\max\limits_{n \geqslant 0} Y_n$ is asymptotically (up to certain constant $K > 0$) equivalent to an exponential function, namely:

$$\lim_{x \to \infty} \frac{\mathrm{P}(\max\limits_{n \geqslant 0} Y_n > x)}{e^{-\gamma x}} = K.$$

2) if $\mathrm{E}[Xe^{\gamma X}] = \infty$, then $\lim\limits_{x \to \infty} P(\max\limits_{n \geqslant 0} Y_n > x) = \mathrm{o}(e^{-\gamma x})$, as $x \to \infty$, where $b = \mathrm{o}(a)$ means $\lim b/a = 0$.

If $S_e \in \mathcal{S}$, then the the stationary waiting time W is also subexponential, $W \in \mathcal{S}$, and [13]

$$\lim_{x \to \infty} \frac{\mathrm{P}(\max\limits_{n \geqslant 0} Y_n > x)}{\mathrm{P}(S_e > x)} = \frac{\rho}{1 - \rho}. \tag{3}$$

Denote by $M_n = \max(W_1, \ldots, W_n)$ the largest waiting time among customers $1, \ldots, n$. If $\rho \leqslant 1$ and there exists a real positive solution to equation (2), then [14–16]

$$\lim_{n \to \infty} \mathrm{P}(\gamma M_n - \log(b\theta n) \leqslant x) = \Lambda(x), \tag{4}$$

where $\Lambda(x) = \exp(-e^{-x})$ is the Gumbel distribution and b is a constant. Moreover, $M_n/\log n$ converges to $1/\gamma$ whenever possible [15] for all $\epsilon > 0$:

$$\mathrm{P}\left(\left|\frac{M_n}{\log n} - \frac{1}{\gamma}\right| > \frac{\epsilon}{\gamma}\right) \to 0 \quad \text{as } n \to \infty.$$

For the $GI/M/1$ system with $\rho < 1$, equation (2) gives a unique positive solution γ [8, Theorem 5.8]. Moreover, the distribution of the actual waiting time W is a mixture of an atom at zero and an exponential distribution with parameter γ and mixture proportions γ/μ and $1 - \gamma/\mu$, respectively, [8, Theorem 5.1]:

$$\mathrm{F}_W(x) = 1 - (1 - \gamma/\mu)e^{-\gamma x}, \quad x > 0.$$

Theorem 7.5 in [15] implies that the extremal index of the process $\{W_n, n \geqslant 1\}$ for a $GI/M/1$ system is calculated by the formula:

$$\theta = \gamma\left(\frac{d}{d\gamma}\psi_\tau(\gamma) + \frac{1}{\mu}\right), \tag{5}$$

where $\psi_\tau(\gamma) = \mathrm{E}e^{-\gamma\tau}$ is the Laplace-Stieltjes transform of interarrival times, τ, and, in addition, $b = 1 - \gamma/\mu$ in (4).

For a stationary $M/G/1$ system, equation (2) has the following form [8]:

$$\mathrm{E}e^{\gamma S} = 1 + \frac{\gamma}{\lambda}. \tag{6}$$

Furthermore, if equation (6) has a real positive solution γ, then the formula for the extremal index θ_W becomes [9]:

$$\theta = \mathrm{P}(W = 0)(1 - \psi_\tau(\gamma)) = \frac{\gamma(1-\rho)}{\gamma + \lambda}. \tag{7}$$

3. Comparison of extremal indexes in $GI/G/1$ systems

Consider two queuing systems $\Sigma^{(1)}$ and $\Sigma^{(2)}$ of type $GI/G/1$ (we assign index i to quantities related to the i-th system). Let $S_n^{(i)}$ be the service time of the n-th customer and $\tau_n^{(i)}$ be the interval between the arrivals of the n-th and $n+1$-th requests in the i-th system, $\mathrm{E}\tau^{(i)} = 1/\lambda$, $i = 1, 2$. Let $W_n^{(i)}$ be the actual waiting time of n-th customer, $i = 1, 2$. Let us denote (if they exist) the distribution limits

$$W_n^{(i)} \Rightarrow W^{(i)}, \quad n \to \infty, \quad i = 1, 2.$$

These limits exist, in particular, if the interarrival times $\tau^{(i)}$, $i = 1, 2$ are non-lattice and $\rho_i = \lambda_i \mathrm{E}S^{(i)} < 1$, $i = 1, 2$ [8].

Let us compare the extremal indexes $\theta^{(1)}$ and $\theta^{(2)}$ of stationary waiting time processes $\{W_n^{(1)}\}$ and $\{W_n^{(2)}\}$ in the systems $\Sigma^{(1)}$ and $\Sigma^{(2)}$, respectively. Further, to compare the r.v.s, we will need the stochastic order and the order in failure rate. We say that the r.v. Z_1 is less that r.v. Z_2 in stochastic order, $Z_1 \underset{st}{\leqslant} Z_2$ if

$$\overline{\mathrm{F}_{Z_1}}(x) \leqslant \overline{\mathrm{F}_{Z_2}}(x), \quad x \in \mathrm{R}.$$

Let $r_Z(x) := f_Z(x)/\overline{\mathrm{F}_Z}(x)$ be failure rate function of r.v. Z, where $f_Z(x)$ is density function. We say that r.v. Z_1 is less than r.v. Z_2 in failure rate order, $Z_1 \underset{r}{\leqslant} Z_2$, if

$$r_{Z_1}(x) \geqslant r_{Z_2}(x), \quad x \in \mathrm{R}.$$

The following theorem allows to compare the extremal indexes of stationary waiting times in two different $M/G/1$ systems for which the real positive roots of equation (2) exist.

Theorem 1. *Suppose that in two $M/G/1$ systems $\Sigma^{(1)}$ and $\Sigma^{(2)}$, $\rho_i < 1$, $\mathrm{E}[S^{(i)}e^{\gamma S^{(i)}}] < \infty, i = 1, 2$ for and any $\gamma \geqslant 0$ and relations*

$$W_1^{(1)} = W_1^{(2)} = 0, \quad \tau^{(1)} \underset{r}{\geqslant} \tau^{(2)}, \quad S^{(1)} \underset{r}{\leqslant} S^{(2)}, \tag{8}$$

are satisfied. Additionally, suppose that there exist real positive roots γ_1 and γ_2 of equation (2) for both systems. Then the extremal indices of the stationary waiting times are ordered,

$$\theta^{(1)} \geqslant \theta^{(2)}. \tag{9}$$

Proof. First of all, note that the ordering in failure rate implies the ordering of the exponential moments [17]. Consequently, from the relations (8) it follows that

$$\mathrm{E}e^{\gamma S^{(1)}} \leqslant \mathrm{E}e^{\gamma S^{(2)}}, \quad \mathrm{E}e^{-\gamma\tau^{(1)}} \leqslant \mathrm{E}e^{-\gamma\tau^{(2)}}, \quad \text{for any } \gamma > 0.$$

Let γ_i denote the positive real root of equation (2) for the system $\Sigma^{(i)}$, $i = 1, 2$. If there are several such roots, then, according to Theorem 7.2 in [15], the smallest of them is chosen. Then the equalities

$$\mathrm{E}e^{\gamma_1(S^{(1)}-\tau^{(1)})} = \mathrm{E}e^{\gamma_2(S^{(2)}-\tau^{(2)})} = 1$$

are satisfied if

$$\gamma_1 \geqslant \gamma_2.$$

Since the intervals between arrivals are ordered in failure rate, $\tau^{(1)} \underset{r}{\geqslant} \tau^{(2)}$, then $\lambda_1 < \lambda_2$. Consequently, $\lambda_1/\gamma_1 \leqslant \lambda_2/\gamma_2$. Moreover, by $S^{(1)} \underset{r}{\leqslant} S^{(2)}$, we also have $\mathrm{E}S^{(1)} \leqslant \mathrm{E}S^{(2)}$ and $1 - \rho_1 \geqslant 1 - \rho_2$. Substituting these inequalities into the expression for the extremal index (7), we obtain

$$\theta^{(1)} = \frac{1-\rho_1}{1+\lambda_1/\gamma_1} \geqslant \frac{1-\rho_2}{1+\lambda_2/\gamma_2} = \theta^{(2)}.$$

To illustrate the statement of Theorem 1 we consider the simple example. It's easy to show that, for the $M/M/1$ system, the extremal index of the stationary waiting time has the form

$$\theta = (1-\rho)^2.$$

Consider two $M/M/1$ systems in which the service times are exponential with parameters μ_1 and μ_2, respectively, and $\mu_1 \geqslant \mu_2 > 0$. Assume that $\lambda_1 \leqslant \lambda_2$ and $\rho_i = \lambda_i/\mu_i < 1$, $i = 1, 2$. In this case, the conditions (8) are satisfied and

$$\gamma_1 = \mu_1 - \lambda_1 \geqslant \gamma_2 = \mu_2 - \lambda_2, \theta^{(1)} = (1-\rho_1)^2 \geqslant (1-\rho_2)^2 = \theta^{(2)},$$

i.e., the inequality (9) is satisfied.

Now consider $M/We/1$ system with Weibull service time d.f.

$$\mathrm{F}_S(x) = 1 - e^{-(x/\alpha)^\beta}, \quad \alpha,\ \beta > 0, \quad x \geqslant 0.$$

The exponential moments $\mathrm{E}e^{\gamma S}$ for the Weibull distribution exist only for $\beta \geqslant 1$, so in this case the equation (6) takes form

$$\sum_{k=0}^{\infty} \frac{(\alpha\gamma)^k}{k!}\Gamma\left(\frac{k}{\beta}+1\right) = 1 + \frac{\gamma}{\lambda}, \quad \text{for } \beta \geqslant 1.$$

Now we compare the extremal indexes of waiting times in two $M/We/1$ systems. For example, let $\lambda = 2$, $\alpha_1 = 0.25$, $\beta_1 = 1.5$ in the first system and $\lambda_2 = 2$, $\alpha_2 = 0.4$, $\beta_2 = 1.2$. With these parameters we have $S^{(1)} \underset{r}{<} S^{(2)}$. By numerical calculation we obtain a single roots $\gamma_1 = 3.7$ and $\gamma_2 = 0.79$, respectively, therefore, $\theta^{(1)} = 0.35 > \theta^{(2)} = 0.07$.

Now we prove a statement similar to the Theorem 1 for $GI/M/1$ systems.

Theorem 2. *Let the stationarity conditions $\rho_i < 1$, $\mathrm{E}[(S^{(i)}e^{\gamma S^{(i)}}] < \infty$ for any $\gamma \geqslant 0$, $i = 1, 2$, and the relations* (8) *be satisfied for two $GI/M/1$-type systems $\Sigma^{(1)}$ and $\Sigma^{(2)}$. Let there exist real positive roots γ_1 and γ_2 of the equation* (2) *for these systems and the following inequality holds:*

$$\mathrm{E}[\tau^{(1)}e^{-\gamma_1\tau^{(1)}}] \leqslant \mathrm{E}[\tau^{(2)}e^{-\gamma_2\tau^{(2)}}]. \tag{10}$$

Then the extremal indexes of the stationary waiting times are ordered as

$$\theta^{(1)} \geqslant \theta^{(2)}.$$

Proof. In the proof of Theorem 1 it is shown that the roots γ_i of equation (2) for systems $\Sigma^{(i)}$ are related by the inequality

$$\gamma_1 \geqslant \gamma_2,$$

and by (8) the exponential moments are also ordered in the same way. Therefore,

$$\psi_{\tau^{(1)}}(\gamma_1) \leq \psi_{\tau^{(2)}}(\gamma_2).$$

Note, that $\mathrm{E}\tau e^{-\gamma\tau} < \infty$ for any r.v. $\tau \geq 0$ and any $\gamma \geq 0$ and that

$$\frac{d}{d\gamma_i}\psi_{\tau^{(i)}}(\gamma_i) = -\mathrm{E}(\tau^{(i)}e^{-\gamma_i\tau^{(i)}}), \quad i = 1, 2.$$

Thus, it follows from condition (10), that

$$\frac{d}{d\gamma_1}\psi_{\tau^{(1)}}(\gamma_1) \geqslant \frac{d}{d\gamma_2}\psi_{\tau^{(2)}}(\gamma_2).$$

The equation (2) for the systems under consideration is equivalent to

$$\psi_{\tau^{(i)}}(\gamma_i) + \gamma_i/\mu_i = 1, \quad i = 1, 2.$$

Therefore, expression (5) can be rewritten as

$$\theta^{(i)} = \gamma_i\left(\frac{d}{d\gamma_i}\psi_{\tau^{(i)}}(\gamma_i) + 1/\mu_i\right) = (1 - \psi_{\tau^{(i)}}(\gamma_i))\left(\mu_i\frac{d}{d\gamma_i}\psi_{\tau^{(i)}}(\gamma_i) + 1\right), \quad i = 1, 2.$$

From the ordering in failure rate of service times, it follows that $\mu_1 \geqslant \mu_2$. Substituting the obtained inequalities into the expression for the extremal index (5), we obtain the required inequality

$$\theta^{(1)} = (1 - \psi_{\tau^{(1)}}(\gamma_1))\left(\mu_1\frac{d}{d\gamma_1}\psi_{\tau^{(1)}}(\gamma_1) + 1\right) \geqslant (1 - \psi_{\tau^{(2)}}(\gamma_2))\left(\mu_2\frac{d}{d\gamma_2}\psi_{\tau^{(2)}}(\gamma_2) + 1\right) = \theta^{(2)}.$$

It is worth noting that the set of queueing systems satisfying the inequality (10) is not empty. In particular it holds for the popular $M/M/1$ system because

$$\frac{d}{d\gamma_i}\psi_{\tau^{(i)}}(\gamma_i) = -\frac{\lambda_i}{(\lambda_i + \gamma_i)^2} = -\frac{\lambda_i}{\mu_i^2},$$

and if $\lambda_1 \leqslant \lambda_2, \mu_1 \geqslant \mu_2$, then condition (10) holds. As another example, consider two systems $\Sigma^{(1)}$ and $\Sigma^{(2)}$, with two-component hyperexponential interarrival times, where d.f. tail has form

$$\overline{\mathrm{F}_{\tau^{(i)}}}(x) = pe^{-\lambda_1^{(i)}x} + (1 - p)e^{-\lambda_2^{(i)}x}, \quad \lambda_j^{(i)} > 0, \quad 0 < p < 1, \quad x \geqslant 0, \quad i,\ j = 1, 2.$$

Let $p = 0.5, \lambda_1^{(1)} = 1$, $\lambda_2^{(1)} = 3$, $\lambda_1^{(2)} = 2$, $\lambda_2^{(2)} = 3$. Suppose that the service time is exponential with parameter $\mu_1 = 4$ in $\Sigma^{(1)}$ and with parameter $\mu_2 = 3$ in $\Sigma^{(2)}$. Then numerical analysis shows that only positive (numerical) solution of (2) is $\gamma_1 = 2.24$, for which $\theta_{W^{(1)}} = 0.33$. Similarly, the unique positive solution $\gamma_2 = 0.58$ can be obtained by solving numerically the equation (2), for which $\theta_{W^{(2)}} = 0.0385$. With these parameters, the conditions (8) and (10) are satisfied, since

$$r_{\tau^{(1)}}(x) = 1 < r_{\tau^{(2)}}(x) = 2; \qquad r_{S^{(1)}}(x) = 4 > r_{S^{(2)}}(x) = 3,$$

and

$$\gamma_1 > \gamma_2, \quad \theta^{(1)} > \theta^{(2)}.$$

3.1. $M/G/1$ system with mixed service times

Consider a single-server $M/G/1$-type system Σ with service time S given by an m-component mixture d.f.

$$\mathrm{F}_S(x) = \sum_{i=1}^{m} p_i \mathrm{F}_{S^{(i)}}(x), \quad \sum_{i=1}^{m} p_i = 1, \quad p_i \geqslant 0, \quad i = 1, \dots, m. \tag{11}$$

Assume that the r.v.'s $S^{(1)}, \dots, S^{(m)}$ are independent and $S^{(i)}$ has d.f. $\mathrm{F}_{S^{(i)}}(x)$, $i = 1, \dots, m$. Denote by

$$\rho = \lambda \mathrm{E}S = \sum_{i=1}^{m} \frac{\lambda p_i}{\mu_i} = \sum_{i=1}^{m} p_i \rho_i$$

the traffic intensity of the system Σ where $\rho_i = \lambda/\mu_i$, $\mu_i = 1/\mathrm{E}S^{(i)}$, $i = 1, \dots, m$. Assume that the components $S^{(i)}$ of the service time S are ordered in failure rate

$$S^{(1)} \underset{r}{\leqslant} \cdots \underset{r}{\leqslant} S^{(m)}.$$

Consider two queuing systems $\Sigma^{(1)}$ and $\Sigma^{(m)}$ with inputs $\tau^{(1)}$, $\tau^{(m)}$, respectively. Let τ be the input process in the original system Σ, and $\mathrm{E}\tau^{(i)} = 1/\lambda_i$, $\mathrm{E}\tau = 1/\lambda$, $i = 1, m$. (As usual, the index i relates to the i-th system.) The service time $S^{(i)}$ is given by the d.f. $\mathrm{F}_{S^{(i)}}(x)$, $i = 1, m$.

Theorem 3. *Assume the stationarity conditions $\rho_i < 1$, $i = 1, m$; $\rho < 1$ hold and the following relations are satisfied:*

$$W_1^{(1)} = W_1^{(m)} = W_1 = 0.$$

$$\tau^{(1)} \underset{r}{\geqslant} \tau \underset{r}{\geqslant} \tau^{(m)}.$$

Suppose, that the components of the service time mixture are ordered

$$S^{(1)} \underset{r}{\leqslant} \cdots \underset{r}{\leqslant} S^{(m)}.$$

Assume that real positive roots of equation (2) *exist for all three systems. Then*

$$\theta^{(1)} \geqslant \theta \geqslant \theta^{(m)}.$$

The proof follows from Theorem 1 and the monotonicity property of waiting times (see Theorems 5 and 6 in [18]).

To illustrate the statement of Theorem 3, consider an $M/H_m/1$ system as the original system where service times have m-component mixture d.f.

$$\overline{\mathrm{F}_S}(x) = \sum_{i=1}^{m} p_i e^{-\mu_i x}, \quad \mu_i > 0, \quad \sum_{i=1}^{m} p_i = 1, p_i \geqslant 0, \quad x \geqslant 0.$$

We consider two systems $\Sigma^{(1)}$ and $\Sigma^{(m)}$ (of type $M/M/1$), in which the service times $S^{(i)}$ have exponential distribution, $\overline{\mathrm{F}_{S^{(i)}}}(x) = e^{-\mu_i x}$, $i = 1, m$. The interarrival times in all three systems have exponential distribution with parameter λ. Assume that the stationarity conditions are satisfied in all systems:

$$\rho_i = \lambda/\mu_i < 1, \quad i = 1, m, \quad \rho = \lambda \sum_{i=1}^{m} p_i/\mu_i < 1.$$

The service time failure rate function r_S in the original system Σ is equal to

$$r_S(x) := \frac{\sum\limits_{i=1}^{m} p_i\mu_i e^{-\mu_i x}}{\sum\limits_{i=1}^{m} p_i e^{-\mu_i x}}, \quad x \geqslant 0.$$

The service time failure rate functions $S^{(1)}$ and $S^{(m)}$ in the systems $\Sigma^{(1)}$ and $\Sigma^{(m)}$ are, respectively, equal to $r_{S^{(1)}}(x) = \mu_1$, $r_{S^{(m)}}(x) = \mu_m$. It is easy to verify that for

$$\mu_1 \geqslant \cdots \geqslant \mu_m \tag{12}$$

the failure rate functions are ordered as follows:

$$r_{S^{(1)}}(x) \geqslant r_S(x) \geqslant r_{S^{(m)}}(x), \quad x \geqslant 0,$$

and, therefore, the service times in these systems are ordered in the failure rate as

$$S^{(1)} \underset{r}{\leqslant} S \underset{r}{\leqslant} S^{(m)}.$$

The extremal index in the original system Σ can be calculated by the formula (7). Moreover, from the condition for the parameters (12) and the equation (2) it follows that

$$\frac{\lambda+\gamma}{\lambda} = \sum_{i=1}^{m} \frac{p_i\mu_i}{\mu_i - \gamma} \geqslant \sum_{i=1}^{m} \frac{p_i}{1-\gamma/\mu_1} = \frac{\mu_1}{\mu_1 - \gamma},$$

and, therefore,

$$\gamma(\mu_1 - \lambda - \gamma) \geqslant 0$$

and $\gamma \leqslant \mu_1 - \lambda = \gamma_1$. Further, since $\rho_1 \leqslant \rho$, then

$$\theta = \frac{1-\rho}{1+\lambda/\gamma} \leqslant \frac{1-\rho_1}{1+\lambda/\gamma_1} = (1-\rho_1)^2 = \theta^{(1)}.$$

Similarly, it can be shown that

$$\frac{\lambda+\gamma}{\lambda} = \sum_{i=1}^{m} \frac{p_i\mu_i}{\mu_i - \gamma} \leqslant \frac{\mu_m}{\mu_m - \gamma},$$

and $\gamma \leqslant \mu_m - \lambda = \gamma_m$, and therefore,

$$\theta = \frac{1-\rho}{1+\lambda/\gamma} \geqslant \frac{1-\rho_m}{1+\lambda/\gamma_m} = (1-\rho_m)^2 = \theta^{(m)}.$$

3.2. Extension of Theorem 2 to multiserver systems

The extremal behaviour of multiserver system $GI/M/c$ with c working in parallel servers and service intensity μ is the same as the in $GI/M/1$ system with service intensity $c\mu$ with identical interarrival distribution $\mathrm{F}_\tau(x)$ [9]. Moreover, if $\rho = (\mu \mathrm{E}\tau)^{-1} < c$, then waiting time process has (total variation) limit which is a mixture of an atom at zero and exponential distribution with intensity γ, where γ is unique solution of the equation [8][Theorem 3.2]

$$\int_0^\infty e^{-\gamma x} \mathrm{F}_\tau(dx) = 1 - \frac{\gamma}{c\mu}.$$

From (5) one can also easily obtain the relation for extremal index of stationary waiting time in $GI/M/c$ system by substituting $c\mu$ instead μ, namely,

$$\theta = \gamma\left(\frac{d}{d\gamma}\psi_\tau(\gamma) + \frac{1}{c\mu}\right).$$

The last formula allows us to compare the extremal indexes of waiting times in two $GI/M/c_i$ systems by adding the condition $c_1 \geqslant c_2$ in Theorem 2. We formulate the obtained result as the following statement.

Theorem 4. *Let the stationarity conditions, $\rho_i = \lambda_1 \mathrm{E}S^{(i)} < c_i$, $i = 1, 2$, and the relations* (8)*, be satisfied for two systems $\Sigma^{(1)}$ and $\Sigma^{(2)}$ of type $GI/M/c_i$. Let there exist real positive roots of the equation* (2) *for these systems and*

$$c_1 \geqslant c_2.$$

Then the extremal indexes of the stationary waiting times are ordered as

$$\theta^{(1)} \geqslant \theta^{(2)}.$$

In particular, it is easy to check that for $M/M/c$ system

$$\gamma = c\mu - \lambda, \theta = \left(1 - \frac{\lambda}{c\mu}\right)^2.$$

Therefore, if $\lambda_1 \leqslant \lambda_2$, $\mu_1 \geqslant \mu_2$, $c_1 \geqslant c_2$, then $\theta^{(1)} \geqslant \theta^{(2)}$.

4. GI/G/1 system with subexponential service times

In this section we consider the systems with the subexponential service times. In contrast to the light-tailed case, the stationary waiting time W is also subexponential and the relation (3) holds. Moreover, the maximum stationary waiting time $M_n = \max(W_0, \dots, W_n)$ has the same asymptotics as $\max(X_0, \dots, X_n)$, as $n \to \infty$, with $X_i = S_i - \tau_i$ [19]. Moreover, if $S \in \mathcal{S}$, then the extremal index of the stationary waiting time for systems with subexponential service time is zero [19], i.e.,

$$\theta = 0.$$

Now consider $GI/G/1$ system with m-component mixture service times with d.f. given by (11) with a dominant component.

We say that a component $\mathrm{F}^{(j)}$, $j \in \{1, \dots, m\}$ is asymptotically r-dominant for an m-component mixture of distributions

$$\mathrm{F}(x) = p_1\mathrm{F}^{(1)}(x) + \cdots + p_m\mathrm{F}^{(m)}(x), \qquad \sum_{i=1}^{m} p_i = 1,$$

if

$$\lim_{x \to x_R} \frac{\overline{\mathrm{F}^{(i)}}(x)}{\overline{\mathrm{F}^{(j)}}(x)} = r_i, \quad i \in \{1, \dots, m\}, \quad i \neq j,$$

where $r = (r_1, \dots, r_m)$, $r_j = 1$, and $0 \leqslant r_i < 1$, for $i \neq j$, x_R is the right endpoint of $\mathrm{F}(x)$.

The following theorem states that, if in a $GI/G/1$ system, the service time is determined by a finite mixture (11) whose r-dominant component of the equilibrium distribution belongs to the class of subexponential distributions, $\mathrm{F}_{S_e^{(j)}} \in \mathcal{S}$, then the tail of the limit distribution of the stationary waiting time is equivalent to the tail of this distribution up to a constant,

$$\lim_{x\to\infty} \frac{\mathrm{P}(W > x)}{\overline{\mathrm{F}_{S_e^{(j)}}}(x)} =: \delta.$$

Furthermore, the limit distribution of the maximum of the stationary waiting time belongs to the maximum domain of attraction (MDA) of the distribution of extreme values of the same type as the maximum of the r.v. defined by the d.f. $\mathrm{F}_{S_e^{(j)}}$, and the extremal index of the stationary waiting time is obviously 0.

Theorem 5. *Let the original system Σ be stationary, $\rho < 1$. Let the service time be defined by an m-component mixture of distributions* (11) *and let there exist a set of numbers*

$$r = (r_1, \dots, r_m), \quad 0 \leqslant r_i < 1, \quad i \neq j, \quad r_j = 1,$$

such that the equilibrium distribution of the j-th component is r-dominant in the mixture and belongs to the class of subexponential distributions, $\mathrm{F}_{S_e^{(j)}} \in \mathcal{S}$. Then

1) The equilibrium distribution of service time belongs to the class of subexponential distributions, $\mathrm{F}_{S_e} \in \mathcal{S}$.

2) The tails of the equilibrium distributions F_{S_e} and $\mathrm{F}_{S_e^{(j)}}$ are equivalent up to a constant,

$$\lim_{x\to\infty} \frac{\overline{\mathrm{F}_{S_e}}(x)}{\overline{\mathrm{F}_e^{(j)}}(x)} = \sum_{i=1}^{m} q_i r_i$$

where

$$r_i := \lim_{x\to\infty} \frac{\overline{\mathrm{F}_{S_e^{(i)}}}(x)}{\overline{\mathrm{F}_{S_e^{(j)}}}(x)}, \quad 0 \leqslant r_i < 1, \quad i = 1, \dots, m, \quad i \neq j, \quad r_j = 1;$$

$q_i = \dfrac{p_i \mathrm{ES}^{(i)}}{\mathrm{ES}}$.

3) The tails of the distribution of the stationary waiting time, $\mathrm{P}(W > x)$, and the equilibrium distribution $\mathrm{F}_{S_e^{(j)}}$ are equivalent up to a constant δ,

$$\lim_{x\to\infty} \frac{\mathrm{P}(W > x)}{\overline{\mathrm{F}_{S_e^{(j)}}}(x)} = \frac{\lambda \sum\limits_{i=1}^{m} p_i r_i \mathrm{ES}^{(i)}}{1 - \lambda \sum\limits_{i=1}^{m} p_i \mathrm{ES}^{(i)}} := \delta.$$

4) If $\mathrm{F}_{S_e^{(j)}} \in MDA(G)$, then d.f. stationary waiting time in the original system $\mathrm{F}_W \in MDA(G^\delta)$.

5) The extremal index of the stationary waiting time in the original system is zero, $\theta_W = 0$.

Proof. 1) Find the equilibrium distribution of service time in the original system

$$\overline{\mathrm{F}_{S_e}}(x) = \frac{1}{\mathrm{ES}} \int_x^\infty (p_1 \overline{\mathrm{F}_{S^{(1)}}}(y) + \dots p_m \overline{\mathrm{F}_{S^{(m)}}}(y)) dy = \sum_{i=1}^{m} \frac{p_i}{\mathrm{ES}} \int_x^\infty \overline{\mathrm{F}_{S^{(i)}}}(y) dy =$$

$$= \sum_{i=1}^{m} \frac{p_i \mathrm{ES}^{(i)}}{\mathrm{ES}} \overline{\mathrm{F}_{S_e^{(i)}}}(x) = \sum_{i=1}^{m} q_i \overline{\mathrm{F}_{S_e^{(i)}}}(x).$$

Obviously, $\sum\limits_{i=1}^{m} q_i = 1$. Thus, the equilibrium service time distribution in the original system is a mixture of the equilibrium distributions of the components with mixture proportions q_i,

$$S_e = J_1 S_e^{(1)} + \cdots + J_m S_e^{(m)}, \quad \sum_{i=1}^{m} J_i = 1$$

where the indicator J_i takes the value 1 with probability q_i, $i = 1, \dots, m$. It is known that if at least one component of the final mixture has a subexponential distribution, then the mixture of distributions belongs to the class of subexponential distributions [20]. Therefore, point 1) of the theorem is proved.

2) Since F_{S_e} is an m-component mixture with proportionality coefficients q_i, where the j-th component is r-dominant, then F_{S_e} and $\mathrm{F}_{S_e^{(j)}}$ have equivalent tails up to the constant $\sum\limits_{i=1}^{m} q_i r_i$ [21].

3) By relation 3 and point 2) of this theorem,

$$\mathrm{P}(W > x) \sim \frac{\rho}{1-\rho}\overline{\mathrm{F}_{S_e}}(x) \sim \frac{\rho}{1-\rho}\sum_{i=1}^{m} q_i r_i \overline{\mathrm{F}_{S_e^{(j)}}}(x) = \delta\overline{\mathrm{F}_{S_e^{(j)}}}(x), \quad \text{as } x \to \infty,$$

where $a \sim b$ means $a/b \to 1$ and

$$\delta = \frac{\rho}{1-\rho}\sum_{i=1}^{m} q_i r_i = \frac{\lambda\sum\limits_{i=1}^{m} p_i r_i \mathrm{E}S^{(i)}}{1-\lambda\sum\limits_{i=1}^{m} p_i \mathrm{E}S^{(i)}}.$$

The point 4) follows from the point 3) of this theorem and [21, Theorem 8].

5) Since $\mathrm{F}_{S_e} \in \mathcal{S}$ then, for $GI/G/1$ system with subexponential service, the extremal index of the stationary waiting time is zero [19]. □

Corollary 1. Assume that the system Σ is stationary, $\rho < 1$. Let the service time be given by an m-component mixture of distributions (11) with components ordered by the failure rate

$$S^{(1)} \underset{r}{\leqslant} \cdots \underset{r}{\leqslant} S^{(m)}. \tag{13}$$

Suppose that $S_e^{(m)} \in \mathcal{S}$. Then all statements of Theorem 5 are true.

Proof. It suffices to show that the distribution $\mathrm{F}_e^{(m)}$ is r-dominant for F_{S_e}. Since the ordering by failure rate of service times (13) implies the stochastic ordering of the r.v. $S_e^{(i)}$,

$$S_e^{(1)} \underset{st}{\leqslant} \cdots \underset{st}{\leqslant} S_e^{(m)},$$

that

$$\frac{\overline{\mathrm{F}_{S_e^{(i)}}}(x)}{\overline{\mathrm{F}_{S_e^{(m)}}}(x)} \leqslant 1, \quad \text{for all } x.$$

Obviously, there exist r_i such that

$$\lim_{x\to x_R} \frac{\overline{\mathrm{F}_{S_e^{(i)}}}(x)}{\overline{\mathrm{F}_{S_e^{(m)}}}(x)} = r_i, \quad 0 \leqslant r_i \leqslant 1, \quad i = 1, \dots, m-1, \quad r_m = 1.$$

In this case, the distribution $\mathrm{F}_e^{(m)}$ is asymptotically r-dominant for the mixture distribution F_{S_e}, $r = (r_1, \dots, r_m)$, $r_m = 1$, and the conditions of Theorem 5 are satisfied for $j = m$. □

.

As an example, we consider a stationary $GI/G/1$ system (i.e., $\rho = \lambda \mathrm{E}S < 1$) with the service times having an exponential-Pareto distribution [22] with parameter $\alpha > 1$

$$\mathrm{F}_S(x) = 1 - pe^{-\lambda_S x} - (1-p)\left(\frac{x_0}{x_0+x}\right)^{\alpha}, \quad \lambda_S > 0, \quad \alpha > 1, \quad x_0 > 0, \quad x \geqslant 0.$$

In this case the equilibrium distribution function of S_e has the form

$$\mathrm{F}_{S_e}(x) = \mu \int_0^x \overline{\mathrm{F}_S}(t)dt = 1 - \mu\left(\frac{pe^{-\lambda_S x}}{\lambda_S} + \frac{(1-p)x_0^{\alpha}}{(\alpha-1)(x_0+x)^{\alpha-1}}\right),$$

where $\mu = 1/\mathrm{E}S$. Note that

$$\overline{\mathrm{F}_{S_e}}(x) = q_1\overline{\mathrm{F}_{S_e^{(1)}}}(x) + q_2\overline{\mathrm{F}_{S_e^{(2)}}}(x)$$

and

$$q_1 = \mu p/\lambda_S, \quad q_2 = \mu(1-p)x_0/(\alpha-1).$$

By Theorem 5 the limiting distribution of the maximum stationary waiting time M_n is a Frechet distribution of the form

$$\lim_{n\to\infty} \mathrm{P}(M_n \leqslant u_n(x)) = e^{-\frac{(1-p)\mu\lambda x_0}{(\alpha-1)(\mu-\lambda)}x^{1-\alpha}}$$

with the normalizing sequence (for $x > 0$)

$$u_n(x) = a_n x + b_n = x_0 x n^{1/(\alpha-1)} - x_0, \quad n \geqslant 1.$$

Indeed, it is obvious that the second component of the distribution is r-dominant, with $r_1 = 0, r_2 = 1$,

$$\lim_{x\to\infty} \frac{\overline{\mathrm{F}_{S_e^{(1)}}}(x)}{\overline{\mathrm{F}_{S_e^{(2)}}}(x)} = \lim_{x\to\infty} \frac{e^{-\lambda_S x}}{x_0^{\alpha-1}/(x_0+x)^{\alpha-1}} = 0 = r_1.$$

Now we find maximum domain of attraction of dominant component $S_e^{(2)}$.

$$\mathrm{F}_{S_e^{(2)}}(x) = \frac{\alpha-1}{x_0}\int_0^x \left(\frac{x_0}{x_0+y}\right)^{\alpha} dy = 1 - \left(\frac{x_0}{x_0+x}\right)^{\alpha-1}.$$

Obviously, $S_e^{(2)}$ has a Pareto distribution with parameters $\alpha - 1$, x_0 and therefore belongs to the class of subexponential distributions [23]. Let $\upsilon_n(x) = x_0 n^{1/(\alpha-1)}x - x_0$. Then for $n \to \infty$,

$$n\overline{\mathrm{F}_{S_e^{(2)}}}(\upsilon_n(x)) = \left(\frac{x_0}{x_0 + x_0 n^{1/(\alpha-1)}x - x_0}\right)^{\alpha-1} \to x^{-\alpha+1}, \quad \text{as } n \to \infty,$$

which implies $\mathrm{F}_{S_e^{(2)}} \in MDA(\Phi_{\alpha-1})$ [1] where $\Phi_\alpha(x) = e^{-x^{-\alpha}}$, $x \geqslant 0$, is Frechet distribution.

Now we can calculate δ

$$\delta = \frac{\lambda(pr_1\mathrm{E}S^{(1)} + (1-p)r_2\mathrm{E}S^{(2)})}{1 - \lambda(p\mathrm{E}S^{(i)} + (1-p)\mathrm{E}S^{(2)})} = \frac{\mu\lambda}{\mu-\lambda}\frac{(1-p)x_0}{\alpha-1}.$$

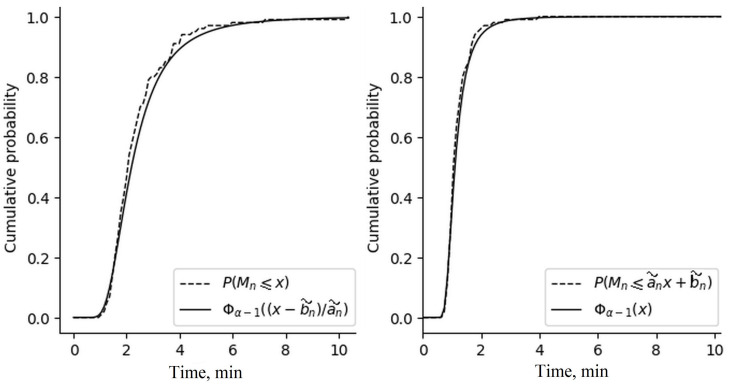

Figure 1. left figure: $\mathrm{P}(M_n \leqslant x)$ and $\Phi_{\alpha-1}(x-\tilde{b}_n)/\tilde{a}_n$, right figure: $\mathrm{P}(M_n \leqslant \tilde{a}_n x + \tilde{b}_n)$ and $\Phi_{\alpha-1}(x)$ for group size $n = 1000$

Since $\mathrm{F}_{S_e^{(2)}} \in MDA(\Phi_{\alpha-1})$, then by 4) of Theorem 5, $\mathrm{F}_W \in MDA(\Phi^{\delta}_{\alpha-1})$.

Thus, the asymptotic behaviour of the maximum stationary waiting time is determined by the second component (the Pareto distribution).

Note also, that since $F_w \in MDA(\Phi^{\delta}_{\alpha-1})$ with $u_n(x) = a_n x + b_n$, then $F_w \in MDA(\Phi_{\alpha-1})$ with $\tilde{u}_n(x) = \tilde{a}_n x + \tilde{b}_n$ where $\tilde{a}_n = a_n \delta^{1/(\alpha-1)}$, $\tilde{b}_n = b_n$.

To illustrate the conclusion of Theorem 5, we carried out a numerical simulation of the system for 100 replications. We compare the estimate of $\mathrm{P}(M_n \leqslant x)$ with $\Phi_{\alpha-1}(x-\tilde{b}_n)/\tilde{a}_n$ and the estimate of $\mathrm{P}(M_n \leqslant \tilde{a}_n x + \tilde{b}_n)$ with $\Phi_{\alpha-1}(x)$. We have run a Kolmogorov–Smirnov (K-S) test for goodness of fit. We find the empirical distribution function for M_n from observed maximum waiting times for $k = 100$ groups of customers (each group of size n). Figure 1 demonstrates results for $x_0 = 1$, $\alpha = 5$, $p = 0,5$, $\lambda_s = 4$, $\lambda = 0.5$ and group size $n = 1000$. K-S test statistic is 0.81 for left figure and 0.74 for right figure. Therefore, our hypothesis that the Frechet distribution with the normalizing constants, a_n, b_n describes nicely the maxima of waiting times (at level 0.05) is confirmed.

5. Results

The sufficient conditions for comparing the extremal indexes of stationary waiting time in the $M/G/1$ and $GI/M/1$ systems are obtained. We have proven that if in both systems the equation $ee^{\gamma(S-\tau)} = 1$ has a real positive roots and the interarrival intervals and the service times are ordered in failure rate then the extremal indexes are ordered. For $M/G/1$ the obtained result is extended to the case of a system with mixed service times with ordered components. In the case of multiserver $GI/M/c$ system we have shown that it is also possible to establish a comparison of the extremal indexes. We illustrate this results on examples with some special distributions. For $GI/G/1$ system with service time determined by a finite mixture whose dominant component of the equilibrium distribution belongs to the class of subexponential distributions,we have proven that the tail of the limiting distribution of the stationary waiting time is equivalent to the tail of this distribution up to a constant, the form of which is obtained. Furthermore, the limiting distribution of the maximum of the stationary waiting time belongs to the maximum domain of attraction of the distribution of extreme values of the same type as the

maximum of the random variables defined by the dominant component, while the extremal index of waiting time is zero.

6. Discussion

The numerical examples given in Sections 2 and 3 of the article demonstrate the correctness of the obtained statements. Note that applying the order in failure rate makes it possible to compare systems with different interarrival distributions and different service time distributions (not just with identical distributions but different parameters).

To demonstrate the assertion of Theorem 5, we considered a system with an exponential-Pareto service-time distribution. The results of our numerical experiments show that the asymptotic distribution of maximum waiting time works well when the traffic is light. Continuing the present work, we plan to extend the numerical experiments with different distributions and investigate the sensitivity of the approximation scheme to the group size n and the traffic intensity ρ.

7. Conclusion

The research examined the conditions under which the extremal indexes of two queuing systems can be compared. This can be used to select parameters for systems that guarantee a given value of the extremal index. The extreme behaviour of the stationary waiting time is also considered.

Author Contributions: Conceptualization, methodology, software, validation, formal analysis, investigation, resources, data curation, writing—original draft preparation, writing—review and editing, visualization, supervision, project administration, funding acquisition, I. V. Peshkova. All authors have read and agreed to the published version of the manuscript.

Funding: This research received no external funding.

Data Availability Statement: Data can be sent by the authors on reasonable request.

Conflicts of Interest: The authors declare no conflict of interest. The funders had no role in the design of the study; in the collection, analyses, or interpretation of data; in the writing of the manuscript; or in the decision to publish the results.

Declaration on Generative AI: The author has not employed any generative AI tools.

References

1. Leadbetter, M., Lindgren, G. & H., R. *Extremes and Related Properties of Random Sequences and Processe* (Springer Series in Statistics, 1983).
2. Resnick, S. I. *Extreme Values, Regular Variation and Point Processes* doi:10.1007/978-0-387-75953-1 (Springer New York, 1987).
3. Smith, R. L. The extremal index for a Markov chain. *Journal of Applied Probability* **29,** 37–45. doi:10.2307/3214789 (Mar. 1992).
4. Weissman, I. & Cohen, U. The extremal index and clustering of high values for derived stationary sequences. *Journal of Applied Probability* **32,** 972–981. doi:10.2307/3215211 (Dec. 1995).
5. Moloney, N. R., Faranda, D. & Sato, Y. An overview of the extremal index. *Chaos: An Interdisciplinary Journal of Nonlinear Science* **29.** doi:10.1063/1.5079656 (Feb. 2019).
6. Beirlant, J., Goegebeur, Y., Teugels, J. & Segers, J. *Statistics of Extremes: Theory and Applications* doi:10.1002/0470012382 (Wiley, Aug. 2004).
7. Ferro, C. A. T. & Segers, J. Inference for Clusters of Extreme Values. *Journal of the Royal Statistical Society. Series B (Statistical Methodology)* **65,** 545–556 (2003).

8. Asmussen, S. *Applied Probability and Queues* doi:10.1007/b97236 (Springer New York, 2003).
9. Hooghiemstra, G. & Meester, L. E. Computing the extremal index of special Markov chains and queues. *Stochastic Processes and their Applications* **65,** 171–185. doi:10.1016/s0304-4149(96)00111-1 (Dec. 1996).
10. Iglehart, D. L. Extreme Values in the GI/G/1 Queue. *The Annals of Mathematical Statistics* **43,** 627–635. doi:10.1214/aoms/1177692642 (Apr. 1972).
11. Rootzén, H. Maxima and exceedances of stationary Markov chains. *Advances in Applied Probability* **20,** 371–390. doi:10.2307/1427395 (June 1988).
12. Markovich, N. & Razumchik, R. *Cluster Modeling of Lindley Process with Application to Queuing* in *Distributed Computer and Communication Networks* (Springer International Publishing, 2019). doi:10.1007/978-3-030-36614-8_25.
13. Veraverbeke, N. Asymptotic behaviour of Wiener-Hopf factors of a random walk. *Stochastic Processes and their Applications* **5,** 27–37. doi:10.1016/0304-4149(77)90047-3 (Feb. 1977).
14. Berger, A. W. & Whitt, W. Maximum Values in Queueing Processes. *Probability in the Engineering and Informational Sciences* **9,** 375–409. doi:10.1017/s0269964800003934 (July 1995).
15. Cohen, J. W. *The Single Server Queu* (North-Holland, 1992).
16. Cohen, J. W. Asymptotic relations in queueing theory. *Advances in Applied Probability* **5,** 8–9. doi:10.2307/1425945 (Apr. 1973).
17. Marshall, A. W. & Olkin, I. *Distributions: Structure of Nonparametric, Semiparametric, and Parametric Familie* doi:10.1007/978-0-387-68477-2 (Springer New York, 2007).
18. Peshkova, I. V. & Morozov, E. V. On Comparison of Multiserver Systems with Multicomponent Mixture Distributions. *Journal of Mathematical Sciences* **267,** 260–272. doi:10.1007/s10958-022-06132-z (Oct. 2022).
19. Asmussen, S., Klüppelberg, C. & Sigman, K. Sampling at subexponential times, with queueing applications. *Stochastic Processes and their Applications* **79,** 265–286. doi:10.1016/s0304-4149(98)00064-7 (Feb. 1999).
20. Foss, S., Korshunov, D. & Zachary, S. *An Introduction to Heavy-Tailed and Subexponential Distributions* doi:10.1007/978-1-4614-7101-1 (Springer New York, 2013).
21. Simon, B. & Foley, R. D. Some Results on Sojourn Times in Acyclic Jackson Networks. *Management Science* **25,** 1027–1034 (1979).
22. Peshkova, I. EXPONENTIAL-PARETO MIXTURE DISTRIBUTION. doi:10.24412/1932-2321-2023-476-632-645 (2023).
23. Klüppelberg, C. Subexponential distributions and integrated tails. *Journal of Applied Probability* **25,** 132–141. doi:10.2307/3214240 (Mar. 1988).

Information about the authors

Peshkova, Irina V.—Candidate of Physical and Mathematical Sciences, Senior researcher, SMITS Lab, Institute of Applied Mathematical Research of the Karelian Research Center of Russian Academy of Sciences; head of the Applied Mathematics and Cybernetics department, Petrozavodsk State University (e-mail: iaminova@petrsu.ru, ORCID: 0000-0002-1461-2425, ResearcherID: P-4375-2015, Scopus Author ID: 57190062292)

УДК 519.872
PACS 02.50.−r
DOI: 10.22363/2658-4670-2026-34-1-24-39 EDN: VCZSIW

Экстремальный индекс времени ожидания в системе GI/G/1

И. В. Пешкова[1, 2]

[1] Петрозаводский государственный университет, пр. Ленина, д. 33, Петрозаводск, 185910, Российская Федерация

[2] Институт прикладных математических исследований КарНЦ РАН, ул. Пушкинская, д. 11, Петрозаводск, 185910, Российская Федерация

Аннотация. В данной работе получены условия сравнения экстремального индекса стационарного времени ожидания в системах $M/G/1$ и $GI/M/1$. Эти условия включают экспоненциальное асимптотическое поведение хвоста времени ожидания и порядок по интенсивности отказов для интервалов между приходами заявок и для времени обслуживания в сравниваемых системах. Для системы $M/G/1$ полученный результат распространяется на смешанные времена обслуживания с упорядоченными компонентами. Если в системе $GI/G/1$ время обслуживания определяется конечной смесью, доминирующая компонента равновесного распределения которой принадлежит классу субэкспоненциальных распределений, то хвост предельного распределения стационарного времени ожидания эквивалентен хвосту этого распределения с точностью до константы, вычисленной в явном виде. Кроме того, предельное распределение максимума стационарного времени ожидания принадлежит области максимального притяжения распределения экстремальных значений того же типа, что и максимум случайных величин, определяемых доминирующей компонентой.

Ключевые слова: экстремальный индекс, система обслуживания, упорядоченность по интенсивности отказа

Discrete & Continuous Models & Applied Computational Science
ISSN 2658-7149 (Online), 2658-4670 (Print)

2026, 34 (1) 40–54
http://journals.rudn.ru/miph

Research article

UDC 519.872, 519.217
PACS 07.05.Tp, 02.60.Pn, 02.70.Bf
DOI: 10.22363/2658-4670-2026-34-1-40-54 EDN: VEJQIO

Derivative-free iterations in R^n with point-wise operations for solving systems of nonlinear equations

Tugal Zhanlav[1,2], Khuder Otgondorj[2], Vandandoo Ulziibayar[2], Khangai Enkhbayar[2]

[1] Institute of Mathematics and Digital Technology, Mongolian Academy of Sciences, Ulaanbator, 13330, Mongolia

[2] Mongolian University of Science and Technology, Ulaanbator, 14191, Mongolia

(received: May 12, 2025; revised: September 30, 2025; accepted: October 30, 2025)

Abstract. In this paper, we develop a new family of high-order derivative-free iterative methods for solving systems of nonlinear equations. Specifically, we propose four two-step derivative-free schemes with convergence orders four and five, together with twelve three-step derivative-free schemes achieving convergence orders six, seven, and eight. The main specific of these iterations is that they include a vector or even a scalar iteration parameter instead of the matrix parameter inherent to other existing iterative methods. This structural simplification significantly reduces computational cost, storage requirements, and matrix operations, thereby improving overall computational efficiency. A convergence analysis is presented, establishing the theoretical order of convergence of the proposed methods. The efficiency indices of the proposed schemes are derived and compared with those of several well-known derivative-free iterative methods. The numerical experiments on standard academic problems confirm the theoretical results and demonstrate that the proposed methods are competitive and, in many cases, superior in terms of efficiency and robustness.

Key words and phrases: nonlinear systems, derivative-free iterations, efficiency index, order of convergence

For citation: Zhanlav, T., Otgondorj, K., Ulziibayar, V., Enkhbayar, K. Derivative-free iterations in R^n with point-wise operations for solving systems of nonlinear equations. *Discrete and Continuous Models and Applied Computational Science* **34** (1), 40–54. doi: 10.22363/2658-4670-2026-34-1-40-54. edn: VEJQIO (2026).

1. Introduction

We consider the following nonlinear system of equations:

$$F(x) = 0, \quad x = (x_1, x_2, \cdots, x_n)^T \in R^n, \tag{1}$$

where $F : D \subseteq R^n \to R^n$ is a nonlinear and sufficiently Fréchet differentiable function in an open convex set D. Additionally, $F'(x)$ is continuous and nonsingular at α, where α is the simple and isolated solution of equation (1). Most physical systems are inherently nonlinear nature and described by nonlinear systems. The nonlinear systems (1) also appear in many fields of applied sciences and engineering [1–12]. The solution of equation (1) cannot be computed exactly and is often approximated using iterative methods with different orders of convergence. A quite recently have been appeared some papers devoted to the constructing high efficient iterative methods containing vector and even scalar parameter coefficients [1, 3, 4, 6, 13–15]. For obtaining the numerical solution of the system (1) often used the following two-step and three-step iterative methods:

$$\begin{aligned} y_k &= x_k - F'(x_k)^{-1}F(x_k), \\ x_{k+1} &= y_k - \bar{\tau}_k F'(x_k)^{-1}F(y_k), \end{aligned} \tag{2}$$

and

$$\begin{aligned} y_k &= x_k - F'(x_k)^{-1}F(x_k), \\ z_k &= y_k - \bar{\tau}_k F'(x_k)^{-1}F(y_k), \\ x_{k+1} &= z_k - \alpha_k F'(x_k)^{-1}F(z_k), \end{aligned} \tag{3}$$

where $\bar{\tau}_k$ and α_k are iteration parameters to be determined properly.

The aim of this work is to develop derivative-free version of the iterations (2) and (3) with vector and scalar coefficients. In Section 2, we introduce new derivative-free two-step iterations of orders four and five. In Section 3, we present new derivative-free three-step iterations of order ρ ($\rho = 6, 7, 8$) and an analysis of the efficiency of the proposed iterative methods. Section 4 devoted to analysis of efficiency of proposed methods compared with other methods. In Section 5, we present the results of our experiments and compare them with known methods of the same order. The article concludes with some conclusions and references used in it.

2. The construction of two-step derivative-free iterations

First, we employ R^n with point-wise multiplication and division of vectors. Let $a = (a_1, a_2, \dots, a_n)^T \in R^n$ and $b = (b_1, b_2, \dots, b_n)^T \in R^n$. The point-wise multiplication and division of two vectors are defined by

$$a \cdot b = (a_1 b_1, a_2 b_2, \dots, a_n b_n)^T \in R^n, \tag{4a}$$

$$\frac{a}{b} = \left(\frac{a_1}{b_1}, \frac{a_2}{b_2}, \dots, \frac{a_n}{b_n}\right)^T \in R^n. \tag{4b}$$

The direct consequence of (4a) and (4b) is

$$a^2 := (a \cdot a) = (a_1^2, a_2^2, \dots, a_n^2)^T \in R^n,$$
$$\mathbf{1} = \Big(1, 1, \dots, 1\Big)^T \in R^n.$$

In [6] the following theorems were proven:

Theorem 6. *[6] The two-step iteration* (2) *has a third, fourth and fifth-order convergence if and only if the parameter $\bar{\tau}_k$ satisfies*

$$\bar{\tau}_k = \mathbf{1} + O(h),$$

$$\begin{aligned} \bar{\tau}_k &= \mathbf{1} + 2\Theta_k + O(h^2), \\ \bar{\tau}_k F'(x_k)^{-1} F(y_k) &= (\mathbf{1} + \Theta_k^2) F'(y_k)^{-1} F(y_k) + O(h^3), \end{aligned} \tag{6a}$$

Theorem 7. *[6] The three-step iteration* (3) *have order of convergence* $\rho + 1$, $\rho + 2$, $\rho + 3$ *if and only if the parameter* α_k *satisfies*

$$\begin{aligned} \alpha_k &= \mathbf{1} + O(h), \\ \alpha_k &= \mathbf{1} + 2\Theta_k + O(h^2), \\ \alpha_k F'(x_k)^{-1} F(z_k) &= (\mathbf{1} + 2\Theta_k^2) F'(y_k)^{-1} F(z_k) + O(h^3), \end{aligned} \tag{7a}$$

where

$$\Theta_k = \frac{F(y_k)}{F(x_k)}, \quad \Theta_k^2 = (\Theta_k \cdot \Theta_k),$$

and ρ *is the order of convergence of iteration* (2).

We note that the conditions (6a) and (7a) can be replaced by

$$\bar{\tau}_k = \alpha_k = \frac{\mathbf{1} + a\Theta_k + b\Theta_k^2}{\mathbf{1} + (a-2)\Theta_k + d\Theta_k^2}, \quad a, b, d \in R,$$

and in this case the convergence order maintained. We now proceed with the construction of a derivative-free analog of (2) as follows:

$$\begin{aligned} y_k &= x_k - [w_k, s_k; F]^{-1} F(x_k), \\ x_{k+1} &= y_k - T_k [w_k, s_k; F]^{-1} F(y_k), \end{aligned} \tag{8}$$

where $[w_k, s_k; F]$ is first order divided difference with

$$w_k = x_k + \gamma_1 F(x_k), \quad s_k = x_k - \gamma_1 F(x_k), \quad \gamma_1 \neq 0, \quad \gamma_1 \in R.$$

It is easy to show that

$$T_k = \bar{\tau}_k F'(x_k)^{-1} [w_k, s_k; F], \tag{9}$$

or

$$\bar{\tau}_k = T_k [w_k, s_k; F]^{-1} F'(x_k). \tag{10}$$

The passing of (2) to (8) is realized by (9). The converse is realized by (10). It is easy to show that

$$F'(x_k)^{-1} [w_k, s_k; F] = I + B_k + O(h^4), \tag{11}$$

where

$$B_k = \frac{1}{6} F'(x_k)^{-1} F'''(x_k) \gamma_1^2 F(x_k)^2 = O(h^2). \tag{12}$$

If we take (12) into account, then from (11) it follows that

$$F'(x_k)^{-1} = [w_k, s_k; F]^{-1} + O(h^2). \tag{13}$$

Analogously, using (11) and the Taylor expansion of $F(y_k)$ at point x_k, we easily obtain

$$F(y_k) = O(h^2), \quad F'(y_k) = [u_k, \varpi_k; F] + O(h^4), \tag{14}$$

where

$$u_k = y_k + \beta_1 F(x_k), \quad \varpi_k = y_k - \beta_1 F(x_k), \quad \beta_1 \neq 0, \quad \beta_1 \in R.$$

Using (9), (11), (13) and (14) it is easy to show that the ρ-order conditions (6) can be rewritten in term of T_k as:

$$T_k = \mathbf{1} + O(h),$$

$$T_k = \mathbf{1} + 2\Theta_k + O(h^2) = \frac{\mathbf{1} + 2\Theta_k + b\Theta_k^2}{1 + d\Theta_k^2} + O(h^2), \tag{15a}$$

$$T_k\,[w_k, s_k; F]^{-1} F(y_k) = (\mathbf{1} + \Theta_k^2)\,[u_k, \varpi_k; F]^{-1} F(y_k). \tag{15b}$$

Using (15a) in (8) we obtain the following family of fourth order iterations (M_1^4)

$$\begin{aligned} y_k &= x_k - [w_k, s_k; F]^{-1} F(x_k), \\ x_{k+1} &= y_k - \frac{1}{\mathbf{1} + d\Theta_k^2}\,[w_k, s_k; F]^{-1}\,[(\mathbf{1} + b\Theta_k^2)F(y_k) + 2\Theta_k^2 F(x_k)\,],\ d, b \in R. \end{aligned} \tag{16}$$

Analogously, using (15b) in (8) we obtain the following fifth order iteration (M_2^5)

$$\begin{aligned} y_k &= x_k - [w_k, s_k; F]^{-1} F(x_k), \\ x_{k+1} &= y_k - (\mathbf{1} + \Theta_k^2)\,[u_k, \varpi_k; F]^{-1} F(y_k). \end{aligned} \tag{17}$$

If $\gamma_1 \to 0$ and $\beta_1 \to 0$ then (16) and (17) lead to the iteration with derivative, considered in [6] and in [4]. The scalar coefficients versions of (16) and (17) are [1]

$$\begin{aligned} y_k &= x_k - [w_k, s_k; F]^{-1} F(x_k), \\ x_{k+1} &= y_k - \frac{1}{1 + dv_k}\,[w_k, s_k; F]^{-1}\,[(1 + bv_k)F(y_k) + 2v_k F(x_k)\,],\ d, b \in R, \end{aligned} \tag{18}$$

$$v_k = \frac{\|F(y_k)\|^2}{\|F(x_k)\|^2},$$

and

$$\begin{aligned} y_k &= x_k - [w_k, s_k; F]^{-1} F(x_k), \\ x_{k+1} &= y_k - (1 + v_k)\,[u_k, \varpi_k; F]^{-1} F(y_k), \end{aligned} \tag{19}$$

with convergence order 4 and 5 respectively. The iteration (18) completely coincides with scheme given in [13], while (19) can be considered as new scheme with fifth order of convergence. Let's denote the methods (18) and (19) as (M_3^4) and (M_4^5), respectively.

To analyze the convergence behavior of the proposed method, we first present a lemma that will be used to develop the Taylor expansion of vector functions (see [16]).

Lemma 1. *Let $F : D \subseteq R^n \to R^n$ be p-times Fréchet differentiable in a open convex set $D \subseteq R^n$, then for any $x, \hat{h} \in D$ the following expression holds:*

$$F(x + \hat{h}) = F(x) + F'(x)\hat{h} + \frac{1}{2!}F''(x)\hat{h}^2 + \cdots + \frac{1}{(p-1)!}F^{(p-1)}(x)\hat{h}^{p-1} + R_p,$$

where

$$\|R_p\| \leqslant \frac{1}{p!}\sup_{0<t<1}\|F^{(p)}(x + t\hat{h})\|\|\hat{h}\|^p \ \text{ and } \ \hat{h}^p = (\overbrace{\hat{h}, \hat{h}, \cdots, \hat{h}}^{p}),$$

and $\|\cdot\|$ denotes any norm in R^n, or a corresponding operator norm.

Definition 1. *Let $e_k = x_k - \alpha$ be the error in the k-th iteration, we call the relation*

$$e_{k+1} = L(e_k)^p + O((e_k)^{p+1}),$$

as the error equation. Here, p is the order of convergence, L is a p-linear function, i.e. $L \in \mathcal{L}(\overbrace{R^n \times \cdots \times R^n}^{p}, R^n)$.

In the following result, we establish the convergence of the family of methods given by (16) under the conditions stated in Lemma 1.

Theorem 8. *Let the function $F : D \subseteq R^n \to R^n$ be sufficiently differentiable in a convex set D containing a zero α of $F(x)$. Further, assume that $F'(x)$ is continuous and non-singular at α and the initial guess x_0 is sufficiently close to the solution. Then, the sequence generated by method* (16) *converges to the solution α with order four, for any nonzero value of parameter γ_1 and for any values of b and d.*

Proof. By applying the Taylor expansion of $F(x_k)$ around α, we obtain

$$F(x_k) = F'(\alpha)(e_k + A_2 e_k^2 + A_3 e_k^3 + A_4 e_k^4) + O(e_k^5), \tag{20}$$

$$\begin{aligned} F'(x_k) &= F'(\alpha)(I + 2A_2 e_k + 3A_3 e_k^2 + 4A_4 e_k^3) + O(e_k^4), \\ F''(x_k) &= F'(\alpha)(2A_2 + 6A_3 e_k + 12A_4 e_k^2) + O(e_k^3), \\ F'''(x_k) &= F'(\alpha)(6A_3 + 24A_4 e_k) + O(e_k^2), \end{aligned} \tag{21}$$

where

$$A_i = \frac{1}{i!}[F'(\alpha)]^{-1} F^{(i)}(\alpha), \quad i = 2, 3, \ldots .$$

Using the Genocchi–Hermite formula [17] and (20)–(21), we obtain

$$\begin{aligned} [w_k, s_k; F] &= F'(x_k) + \frac{1}{6}F'''(x_k)(\gamma_1 F(x_k))^2 + O((\gamma_1 F(x_k))^3) = \\ &= F'(\alpha)(I + 2A_2 e_k + 3A_3 e_k^2) + \frac{1}{6}F'(\alpha)6A_3\gamma_1^2 F'(\alpha)^2 (e_k)^2 + O(e_k^3) = \\ &= F'(\alpha)(I + 2A_2 e_k + A_3(3I + \gamma_1^2 F'(\alpha)^2)e_k^2) + O(e_k^3). \end{aligned}$$

Inversion of $[w_k, s_k; F]$ yields

$$[w_k, s_k; F]^{-1} = (I + C_1 e_k + C_2 e_k^2)F'(\alpha)^{-1} + O(e_k^3), \tag{22}$$

where $C_1 = -2A_2$, $C_2 = 4A_2^2 - A_3(3I + \gamma_1^2 F'(\alpha)^2)$.

Let us denote $\bar{e}_k = y_k - \alpha$. From (20) and (22), we get

$$\bar{e}_k = x_k - \alpha - [w_k, s_k; F]^{-1} F(x_k) = B_1 e_k^2 + B_2 e_k^3 + O(e_k^4),$$

where $B_1 = A_2$, $B_2 = -2A_2^2 + A_3(2I + \gamma_1^2 F'(\alpha)^2)$. We then obtain

$$F(y_k) = F'(\alpha)(\bar{e}_k + A_2 \bar{e}_k^2 + A_3 \bar{e}_k^3) + O(e_k^4) = F'(\alpha)(B_1 e_k^2 + B_2 e_k^3) + O(e_k^4). \tag{23}$$

Next, we expand the term $\Theta_k = \dfrac{F(y_k)}{F(x_k)}$, which appears in the second step of (16). From (20) and (23), we obtain

$$\Theta_k^2 = A_2^2 e_k^2 + (-6A_2^3 + 2A_2 A_3(2I + \gamma_1^2 F'(\alpha)^2))e_k^3 + O(e_k^4). \tag{24}$$

Then, from (24), we can get

$$P_k = I\frac{1+b\Theta_k^2}{1+d\Theta_k^2} = I + A_2^2(b-d)e_k^2 - 2(A_2(b-d)(3A_2^2 - A_3(2I+\gamma_1^2F'(\alpha)^2)))e_k^3 + O(e_k^4), \tag{25}$$

and

$$Q_k = \frac{2\Theta_k^2}{1+d\Theta_k^2} = 2A_2^2e_k^2 + 4A_2(-3A_2^2 + A_3(2I+\gamma_1^2F'(\alpha)^2))e_k^3 + O(e_k^4). \tag{26}$$

From (25) and (26), it follows that

$$P_kF(y_k) + Q_kF(x_k) = A_2^2e_k^2 + A_3(2I+\gamma_1^2F'(\alpha)^2))e_k^3 + \\ + A_2^3(-10+b-d) + 4A_2A_3(2I+\gamma_1^2F'(\alpha)^2)e_k^4 + O(e_k^5). \tag{27}$$

Then, using (22), and (27), the second step of the method (16) gives the error equation as

$$e_{k+1} = x_{k+1} - \alpha = 10A_2^3 - bA_2^3 - 8A_2A_3 + A_2^3d - 4A_2A_3\gamma_1^2F'(\alpha)^2 + \\ + 2A_2A_3(2I+\gamma_1^2F'(\alpha)^2) - A_2(4A_2^2 - A_3(3I+\gamma_1^2F'(\alpha)^2))e_k^4 + O(e_k^5) = \\ = -A_2(A_3 + A_2^2(-6+b-d) + A_3\gamma_1^2F'(\alpha)^2)e_k^4 + O(e_k^5).$$

This shows the fourth order convergence of the proposed family (16). □

The convergence analysis of the other proposed methods follows a similar approach to the proof of Theorem 8. Therefore, we omit it here.

3. The construction of three-step derivative-free iterations

The derivative-free analogy of iteration (3) obtained as:

$$y_k = x_k - [w_k, s_k; F]^{-1}F(x_k),$$
$$z_k = y_k - T_k[w_k, s_k; F]^{-1}F(y_k),$$
$$x_{k+1} = z_k - H_k[w_k, s_k; F]^{-1}F(z_k),$$

where T_k is given by (15) and H_k determined as:

$$H_k = \alpha_kF(x_k)^{-1}[w_k, s_k; F].$$

As before, the condition (7) can be rewritten in term of H_k as:

$$H_k = \mathbf{1} + O(h),$$
$$H_k = \mathbf{1} + 2\Theta_k + O(h^2) = \frac{\mathbf{1}+2\Theta_k+b\Theta_k^2}{1+d\Theta_k^2} + O(h^2),\ b,d \in R,$$
$$H_k[w_k, s_k; F]^{-1}F(z_k) = (\mathbf{1}+2\Theta_k^2)[u_k, \varpi_k; F]^{-1}F(z_k) + O(h^3).$$

Theorem 7 and the combination of choices (15) and (28) yields different derivative-free three-step iterations and we list some of these methods below.

Sixth-order iterations:

$$\begin{aligned} y_k &= x_k - [w_k, s_k; F]^{-1} F(x_k), \\ z_k &= y_k - \frac{1}{\mathbf{1} + d\Theta_k^2} [w_k, s_k; F]^{-1} [(\mathbf{1} + b\Theta_k)F(y_k) + 2\Theta_k^2 F(x_k)], \\ x_{k+1} &= z_k - (\mathbf{1} + 2\Theta_k) [w_k, s_k; F]^{-1} F(z_k), \end{aligned} \tag{29}$$

and

$$\begin{aligned} y_k &= x_k - [w_k, s_k; F]^{-1} F(x_k), \\ z_k &= y_k - (\mathbf{1} + \Theta_k^2) [u_k, \varpi_k; F]^{-1} F(y_k), \\ x_{k+1} &= z_k - [w_k, s_k; F]^{-1} F(z_k), \end{aligned}$$

and

$$\begin{aligned} y_k &= x_k - [w_k, s_k; F]^{-1} F(x_k), \\ z_k &= y_k - [w_k, s_k; F]^{-1} F(y_k), \\ x_{k+1} &= z_k - (\mathbf{1} + 2\Theta_k^2) [u_k, \varpi_k; F]^{-1} F(z_k). \end{aligned}$$

Seventh-order iterations

$$\begin{aligned} y_k &= x_k - [w_k, s_k; F]^{-1} F(x_k), \\ z_k &= y_k - \frac{1}{\mathbf{1} + d\Theta_k^2} [w_k, s_k; F]^{-1} [(\mathbf{1} + b\Theta_k)F(y_k) + 2\Theta_k^2 F(x_k)], \\ x_{k+1} &= z_k - (\mathbf{1} + 2\Theta_k^2) [u_k, \varpi_k; F]^{-1} F(z_k), \end{aligned}$$

and

$$\begin{aligned} y_k &= x_k - [w_k, s_k; F]^{-1} F(x_k), \\ z_k &= y_k - (\mathbf{1} + \Theta_k^2) [u_k, \varpi_k; F]^{-1} F(y_k), \\ x_{k+1} &= z_k - (\mathbf{1} + 2\Theta_k) [w_k, s_k; F]^{-1} F(z_k). \end{aligned}$$

Eighth-order iterations

$$\begin{aligned} y_k &= x_k - [w_k, s_k; F]^{-1} F(x_k), \\ z_k &= y_k - (\mathbf{1} + \Theta_k^2) [u_k, \varpi_k; F]^{-1} F(y_k), \\ x_{k+1} &= z_k - (\mathbf{1} + 2\Theta_k^2) [u_k, \varpi_k; F]^{-1} F(z_k). \end{aligned} \tag{30}$$

In the remainder of the paper, the methods (29)–(30) will be denoted by M_5^6, M_6^6, M_7^6, M_8^7, M_9^7 and M_{10}^8, respectively. If we take the transition rule that established in [14] into account, then we easily obtain from (29)–(30) its scalar coefficients variants:
Sixth-order iterations

$$\begin{aligned} y_k &= x_k - [w_k, s_k; F]^{-1} F(x_k), \\ z_k &= y_k - \frac{1}{1 + dv_k} [w_k, s_k; F]^{-1} [(1 + bv_k)F(y_k) + 2v_k F(x_k)], \\ x_{k+1} &= z_k - [u_k, \varpi_k; F]^{-1} F(z_k), \end{aligned} \tag{31}$$

where

$$\upsilon_k = \frac{\|F(y_k)\|^2}{\|F(x_k)\|^2},$$

and

$$\begin{aligned} y_k &= x_k - [w_k, s_k; F]^{-1} F(x_k), \\ z_k &= y_k - (1+\upsilon_k)\,[u_k, \varpi_k; F]^{-1} F(y_k), \\ x_{k+1} &= z_k - [w_k, s_k; F]^{-1} F(z_k). \end{aligned}$$

and

$$\begin{aligned} y_k &= x_k - [w_k, s_k; F]^{-1} F(x_k), \\ z_k &= y_k - [w_k, s_k; F]^{-1} F(y_k), \\ x_{k+1} &= z_k - (1+2\upsilon_k)\,[u_k, \varpi_k; F]^{-1} F(z_k). \end{aligned}$$

Seventh-order methods:

$$\begin{aligned} y_k &= x_k - [w_k, s_k; F]^{-1} F(x_k), \\ z_k &= y_k - \frac{1}{1+d\upsilon_k}\,[w_k, s_k; F]^{-1}\,[(1+b\upsilon_k)F(y_k) + 2\upsilon_k F(x_k)\,], \\ x_{k+1} &= z_k - (1+2\upsilon_k)\,[u_k, \varpi_k; F]^{-1} F(z_k), \end{aligned}$$

and

$$\begin{aligned} y_k &= x_k - [w_k, s_k; F]^{-1} F(x_k), \\ z_k &= y_k - (1+\upsilon_k)\,[u_k, \varpi_k; F]^{-1} F(y_k), \\ x_{k+1} &= z_k - [u_k, \varpi_k; F]^{-1}\,(F(z_k) - \beta_k F(x_k)), \quad \beta_k = \frac{\|F(z_k)\|^2}{\|F(y_k)\|^2}. \end{aligned}$$

Eight-order method:

$$\begin{aligned} y_k &= x_k - [w_k, s_k; F]^{-1} F(x_k), \\ z_k &= y_k - (1+\upsilon_k)\,[u_k, \varpi_k; F]^{-1} F(y_k), \\ x_{k+1} &= z_k - (1+2\upsilon_k)\,[u_k, \varpi_k; F]^{-1} F(z_k). \end{aligned} \tag{32}$$

In the rest of the paper, the methods (31)–(32) will be denoted by M_{11}^6, M_{12}^6, M_{13}^6, M_{14}^7, M_{15}^7 and M_{16}^8, respectively.

4. Computational efficiency

The computational efficiency index of an iterative method for solving a nonlinear system is defined by $CI = \rho^{\frac{1}{C}}$, where ρ is the order of convergence and C is the computational cost of each method. We will study the computational efficiency of the presented methods and compare it with that of other methods presented in the literature, namely M_3^4 [13], $M_{6,2}$ [18], NM7 [19] and $PM1$ [20]. To compute F in any iterative method we evaluate n scalar functions, whereas the number of scalar evaluations is $n(n-1)$ scalar functions for any divided difference $[\cdot, \cdot; F]$. In addition, we must include the number of operations shown in Table 1.

As we can see in Figs. 1, 2 and in Table 2, in terms of computational efficiency the proposed method M_5^6 is significantly superior to other considered methods. Additionally, fourth-order M_3^4 and eighth-order M_{16}^8 also have high computational efficiency.

Table 1

Computational cost of different operations

	Computational cost
LU decomposition	$\frac{1}{3}(n^3 - n)$
Solution of two triangular systems	n^2
Quotients in divided difference operator	n^2
Matrix-vector multiplication	n^2
Scalar-vector multiplication	n
Component-wise multiplication (division) of vectors	n

Table 2

Comparison of computational efficiency

№	methods	ρ	C_i	CI
1	M_5^6	6	$C_1 = \frac{1}{3}n^3 + 5n^2 + \frac{44}{3}n$	$6^{1/C_1}$
2	M_6^6	6	$C_2 = \frac{2}{3}n^3 + 5n^2 + \frac{23}{3}n$	$6^{1/C_2}$
3	M_7^6	6	$C_3 = \frac{2}{3}n^3 + 6n^2 + \frac{28}{3}n$	$6^{1/C_3}$
4	M_8^7	7	$C_4 = \frac{2}{3}n^3 + 6n^2 + \frac{46}{3}n$	$7^{1/C_4}$
5	M_9^7	7	$C_5 = \frac{2}{3}n^3 + 6n^2 + \frac{31}{3}n$	$7^{1/C_5}$
6	M_{10}^8	8	$C_6 = \frac{2}{3}n^3 + 6n^2 + \frac{31}{3}n$	$8^{1/C_6}$
7	M_{11}^6	6	$C_7 = \frac{2}{3}n^3 + 6n^2 + \frac{28}{3}n$	$6^{1/C_7}$
8	M_{12}^6	6	$C_8 = \frac{2}{3}n^3 + 6n^2 + \frac{22}{3}n$	$6^{1/C_8}$
9	M_{13}^6	6	$C_9 = \frac{2}{3}n^3 + 6n^2 + \frac{25}{3}n$	$6^{1/C_9}$
10	M_{14}^7	7	$C_{10} = \frac{2}{3}n^3 + 6n^2 + \frac{31}{3}n$	$7^{1/C_{10}}$
11	M_{15}^7	7	$C_{11} = \frac{2}{3}n^3 + 6n^2 + \frac{28}{3}n$	$7^{1/C_{11}}$
12	M_{16}^8	8	$C_{12} = \frac{2}{3}n^3 + 6n^2 + \frac{25}{3}n$	$8^{1/C_{12}}$

5. Numerical results and discussion

To evaluate the effectiveness of the new method and provide a comparison with existing methods, numerical experiments have been conducted and the results are presented in this section. To achieve this goal, we consider the following nonlinear problems, most of which are the same as in [13, 19, 21].

Example 1. Considering the following system of 20 equations:

$$y_i - \cos\left(2y_i - \sum_{j=1}^{20} y_j\right) = 0, \quad i = 1, 2, \dots, 20.$$

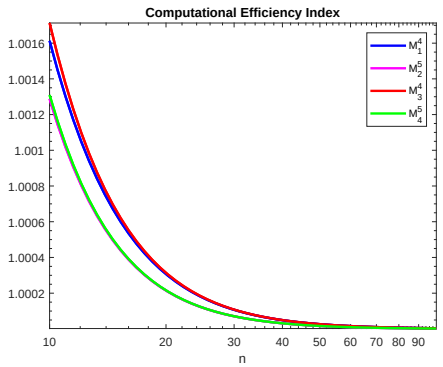

Figure 1. Computational Efficiency Index for $n = 10$ to 100 (logarithmic scale)

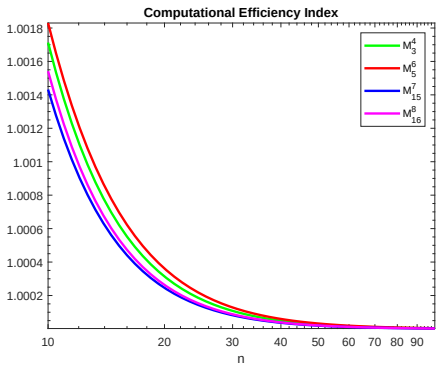

Figure 2. Computational Efficiency Index for $n = 10$ to 100 (logarithmic scale)

The solution is $y^* = \{-0.89, -0.89, \dots, -0.89\}^T$. For this solution, we choose the starting vector $x_0 = \{-0.9, -0.9, \dots, -0.9\}^T$.

Example 2. Consider the system of twenty equations

$$-y_i - 3 + \sum_{j=1}^{20} y_j - e^{y_i} + 4\cos(2\ln(|1 + y_i|)) = 0, \quad i = 1, 2, \dots, 20.$$

The exact solution $y^* = \{0, 0, \dots, 0\}^T$ of above system. For this solution, we choose the starting vector $x_0 = \{0.01, 0.01, \dots, 0.01\}^T$.

In Tables 3 and 4, we present the residual error of the example function $\|F(x_{k+1})\|$, the error between two consecutive iterations $\|x_{k+1} - x_k\|$ and the computational order of convergence ρ_{co}. The computational order of convergence (p_{co}) is calculated using the formula [4]

$$\rho_{co} = \frac{\ln(\|x_{k+1} - x_k\|/\|x_k - x_{k-1}\|)}{\ln(\|x_k - x_{k-1}\|/\|x_{k-1} - x_{k-2}\|)}.$$

The following stopping criterion is used in these experiments:

$$\|x_{k+1} - x_k\| + \|F(x_k)\| \leqslant 10^{-60}.$$

Tables 3 and 4 report the numerical performance of the considered derivative-free iterative methods. The first column lists the names of the methods under comparison. The second column shows the total CPU time (in seconds) required by each method to reach the prescribed stopping criterion. The third column indicates the number of iterations (Iter) needed for convergence.

The fourth column presents the absolute error measured by the norm $\|x_{k+1} - x_k\|$, while the fifth column reports the residual norm $\|F(x_{k+1})\|$ at the final iteration, which reflects the accuracy of the computed solution. The last column displays the approximate computational order of convergence (ACOC), confirming the theoretical convergence order of each method.

Table 3

Comparison numerical results on Example 1

Methods	CPUTime	Iter	$\|x_{k+1} - x_k\|$	$\|F(x_{k+1})\|$	ACOC
M_1^4	0.356	4	1.6016×10^{-76}	4.4188×10^{-300}	4.00
M_3^4	0.344	4	1.6016×10^{-76}	4.4188×10^{-300}	4.00
M_2^5	0.344	4	5.8123×10^{-166}	1.9572×10^{-822}	5.00
M_4^5	0.625	4	5.8123×10^{-166}	1.9572×10^{-822}	5.00
M_5^6	0.343	4	3.4780×10^{-228}	3.2447×10^{-1359}	6.00
M_6^6	0.640	4	9.7496×10^{-287}	1.8481×10^{-1711}	6.00
M_7^6	0.703	4	1.2693×10^{-274}	1.7032×10^{-1638}	6.00
M_{11}^6	0.672	4	7.8812×10^{-283}	4.2801×10^{-1688}	6.00
M_{12}^6	0.735	4	9.7496×10^{-287}	1.8481×10^{-1711}	6.00
M_{13}^6	0.782	4	1.2693×10^{-274}	1.7032×10^{-1638}	6.00
M_8^7	0.656	4	9.7872×10^{-388}	3.0845×10^{-2523}	7.00
M_9^7	0.640	4	2.2881×10^{-402}	2.9211×10^{-2524}	7.00
M_{14}^7	0.734	4	9.7872×10^{-388}	9.3521×10^{-2524}	7.00
M_{15}^7	0.732	4	1.5722×10^{-400}	8.5164×10^{-2524}	7.00
M_{10}^8	0.585	3	1.2259×10^{-79}	2.7310×10^{-624}	8.00
M_{16}^8	0.532	3	1.2259×10^{-79}	2.7310×10^{-624}	8.00
$M_{6,2}$ [18]	1.614	4	1.4692×10^{-107}	6.3590×10^{-633}	6.00
NM7 [19]	2.750	3	3.8545×10^{-71}	1.2384×10^{-487}	7.00
PM1 [20]	1.984	4	2.9442×10^{-271}	1.0079×10^{-2152}	8.00

From Tables 3 and 4, we observe that the M_3^4 iterative method is faster than the considered fourth- and fifth-order methods. Furthermore, Tables 3 and 4 indicate that the proposed M_5^6 method is the fastest among the considered methods with orders $\rho = 6, 7$ and 8. This finding is consistent with the results presented in Section 4. From these tables, it follows that M_{16}^8 is not only faster but also more accurate than the considered seventh- and eighth-order methods. Thus, the eighth-order method M_{16}^8 can be highly useful in practical applications that require high accuracy. In conclusion, the numerical results clearly demonstrate that the proposed derivative-free methods with vector and scalar coefficients are superior to those employing matrix coefficients, both in terms of computational time and overall computational cost.

Conclusions

We obtain family of two-step derivative-free iterations of order 4 and 5 and three-step derivative-free iterations of order 6, 7 and 8 with vector and scalar parameter. The specific of these iterations is that they include vector or even scalar parameter of iteration instead of matrix parameter that inherent to other existing iterative methods. The theoretical conclusions are confirmed by numerical

Table 4

Comparison numerical results on Example 2

Methods	CPUTime	Iter	$\|x_{k+1} - x_k\|$	$\|F(x_{k+1})\|$	ACOC
M_1^4	20.672	4	7.0707×10^{-76}	1.6644×10^{-300}	4.00
M_3^4	20.594	4	7.0707×10^{-76}	1.6644×10^{-300}	4.00
M_2^5	35.572	4	5.9230×10^{-172}	4.1262×10^{-858}	5.00
M_4^5	37.563	4	5.9230×10^{-172}	4.1262×10^{-858}	5.00
M_5^6	20.282	4	2.4149×10^{-219}	1.0347×10^{-1310}	6.00
M_6^6	37.782	4	1.1978×10^{-315}	2.8416×10^{-1893}	6.00
M_7^6	37.532	4	4.4919×10^{-302}	1.5809×10^{-1811}	6.00
M_{11}^6	29.656	3	7.6581×10^{-66}	1.9409×10^{-394}	6.00
M_{12}^6	37.469	4	1.1978×10^{-315}	2.8416×10^{-1893}	6.00
M_{13}^6	37.375	4	4.4919×10^{-302}	1.5809×10^{-1811}	6.00
M_8^7	42.219	4	9.1502×10^{-380}	4.7631×10^{-2654}	7.00
M_9^7	38.344	4	1.7236×10^{-399}	2.0037×10^{-2792}	7.00
M_{14}^7	37.922	4	9.1502×10^{-380}	4.7631×10^{-2654}	7.00
M_{15}^7	37.641	4	2.5776×10^{-400}	3.3521×10^{-2798}	7.00
M_{10}^8	29.906	3	6.0681×10^{-78}	1.3860×10^{-620}	8.00
M_{16}^8	29.391	3	6.0681×10^{-78}	1.3860×10^{-620}	8.00
$M_{6,2}$ [18]	56.801	4	7.8477×10^{-204}	1.4880×10^{-1216}	6.00
NM7 [19]	186.703	3	3.4339×10^{-83}	5.5892×10^{-288}	7.00
PM1 [20]	57.985	3	5.8566×10^{-78}	9.7060×10^{-617}	8.00

experiments. Based on numerical examples, one can conclude that our proposed iterations are the most efficient and faster than the existing ones of similar nature.

Author Contributions: Development and original draft preparation, T. Zhanlav; writing-original draft preparation, Kh. Otgondorj; manuscript review and editing, V. Ulziibayar; writing-review and numerical experiments, Kh. Enkhbayar. All authors have reviewed and approved the published version of the manuscript.

Funding: This research received no external funding.

Data Availability Statement: Data sharing is not applicable.

Acknowledgments: The authors wish to thank the editor and the anonymous referees for their valuable suggestions and comments on the first version of this paper.

Conflicts of Interest: The authors declare no conflict of interest.

References

1. Cordero, A., Rojas-Hiciano, R. V., Torregrosa, J. R. & Vassileva, M. P. A highly efficient class of optimal fourth-order methods for solving nonlinear systems. *Numerical Algorithms* **95,** 1879–1904. doi:10.1007/s11075-023-01631-9 (2024).
2. Changbum, C. & Neta, B. Three-step iterative methods for numerical solution of systems of nonlinear equations. *Engineering with Computers* **38,** 1015–1028. doi:10.1007/s00366-020-01072-1 (2020).
3. Erfanifar, R., Hajarian, M. & Sayevand, K. A family of iterative methods to solve nonlinear problems with applications in fractional differential equations. *Mathematical Methods in the Applied Sciences* **47,** 1–21. doi:10.1002/mma.9736 (2023).
4. Singh, H., Sharma, J. R. & Kumar, S. A simple yet efficient two-step fifth-order weighted-Newton method for nonlinear models. *Numerical Algorithms* **93,** 203–225. doi:10.1007/s11075-022-01412-w (2023).
5. Su, Q. A unified model for solving a system of nonlinear equations. *Applied Mathematics and Computation* **290,** 46–55. doi:10.1016/j.amc.2016.05.047 (2016).
6. Zhanlav, T. & Otgondorj, K. Development and adaptation of higher-order iterative methods in R^n with specific rules. *Discrete and Continuous Models and Applied Computational Science* **32,** 425–444. doi:10.22363/2658-4670-2024-32-4-425-444 (2024).
7. Zhanlav, T. & Chuluunbaatar, O. *New development of Newton-type iterations for solving nonlinear problems* 281 pp. doi:10.1007/978-3-031-63361-4 (Switzerland, Springer Nature, 2024).
8. Zhanlav, T., Chuluunbaatar, O. & Ulziibayar, V. Necessary and sufficient conditions for two and three-point iterative method of Newton's type iterations. *Computational Mathematics and Mathematical Physics* **57.** doi:10.1134/S0965542517070120 (2017).
9. Zhanlav, T., Chuluunbaatar, O. & Ulziibayar, V. Generating functionsmethod for construction new iterations. *Applied Mathematics and Computation* **315.** doi:10.1016/j.amc.2017.07.078 (2017).
10. Zhanlav, T., Mijiddorj, R. & Otgondorj, K. A family of Newton-type methods with seventh and eighth-order of convergence for solving systems of nonlinear equations. *Applied Mathematics and Computation* **54.** doi:10.15672/hujms.1061471 (2023).
11. Zhanlav, T., Otgondorj, K., Saruul, L. & Mijiddorj, R. A unified approach to the construction of higher–order derivative–free iterative methods for solving systems of nonlinear equations. *Proceedings of the Mongolian Academy of Sciences* **64.** doi:10.5564/pmas.v64i02.3649 (2023).
12. Zhanlav, T. & Otgondorj, K. Optimal eight-order three-step iterative methods for solving systems of nonlinear equations. *Discrete and Continuous Models and Applied Computational Science* **33.** doi:10.22363/2658-4670-2025-33-4-389-403 (2025).
13. Cordero, A., Rojas-Hiciano, R. V., Torregrosa, J. R. & Triguero-Navarro, P. Efficient parametric family of fourth-order Jacobian free iterative vectorial schemes. *Numerical Algorithms* **97,** 2011–2029. doi:10.1007/s11075-024-01776-1 (2024).
14. Zhanlav, T. & Otgondorj, K. High efficient iterative methods with scalar parameter coefficients for systems of nonlinear equations. *Journal of Mathematical Sciences* **279,** 866–875. doi:10.1007/s10958-024-07066-4 (2024).
15. Cordero, A., Rojas-Hiciano, R. V., Torregrosa, J. R. & Triguero-Navarro, P. Efficient parametric family of fourth-order Jacobian-free iterative vectorial schemes. *Numerical Algorithms* **97.** doi:10.1007/s11075-024-01776-1 (2024).
16. Ortega, J. M. & Rheinboldt, W. C. *Iterative Solution of Nonlinear Equations in Several Variables* 572 pp. (Academic Press, New York, 1970).
17. Kung, H. T. & Traub, J. F. Optimal order of one-point and multipoint iteration. *Association for Computing Machinery* **21,** 643–651. doi:10.1145/321850.321860 (1973).

18. Sharma, J. R. & Arora, H. Efficient higher order derivative-free multipoint methods with and without memory for systems of nonlinear equations. *International Journal of Computer Mathematics* **95,** 920–938. doi:10.1080/00207160.2017.1298747 (2018).
19. Narang, M., Bhatia, S. & Kanwar, V. New efficient derivative free family of seventh-order methods for solving systems of nonlinear equations. *Numerical Algorithms* **76,** 283–307. doi:10.1007/s11075-016-0254-0 (2017).
20. Ahmad, F., Soleymani, F., Khaksar Haghani, F. & Serra-Capizzano, S. Higher order derivative-free iterative methods with and without memory for systems of nonlinear equations. *Applied Mathematics and Computation* **314,** 199–211. doi:10.1016/j.amc.2017.07.012 (2017).
21. Zhanlav, T., Chun, C., Otgondorj, K. & Ulziibayar, V. High-order iterations for systems of nonlinear equations. *International Journal of Computer Mathematics* **97,** 1704–1724. doi:10.1080/00207160.2019.1652739 (2020).

Information about the authors

Zhanlav, Tugal—Academician, Professor, Doctor of Sciences in Physics and Mathematics, (e-mail: tzhanlav@yahoo.com, ORCID: 0000-0003-0743-5587, Scopus Author ID: 24484328800)

Otgondorj, Khuder—Associate Professor of Department of Mathematics at School of Applied Sciences, Mongolian University of Science and Technology (e-mail: otgondorj@gmail.com, ORCID: 0000-0003-1635-7971, Scopus Author ID: 57209734799)

Ulziibayar, Vandandoo—Professor of Department of Mathematics at School of Applied Sciences, Mongolian University of Science and Technology (e-mail: v_ulzii@must.edu.mn, phone: +(976)99071795, ORCID: 0000-0003-2279-0755)

Enkhbayar, Khangai—Senior Lecturer of Department of Mathematics at School of Applied Sciences, Mongolian University of Science and Technology (e-mail: eegii33@must.edu.mn, ORCID: 0000-0002-1259-5502)

УДК 519.872, 519.217
PACS 07.05.Tp, 02.60.Pn, 02.70.Bf
DOI: 10.22363/2658-4670-2026-34-1-40-54 EDN: VEJQIO

Итерации без производных в R^n с поточечными операциями для решения систем нелинейных уравнений

Т. Жанлав[1, 2], Х. Отгондорж[2], В. Улзийбаяр[2], Х. Энхбаяр[2]

[1] Институт математики и информационной технологии, Монгольская Академия Наук, Улан-батор, 13330, Монголия

[2] Монгольский Государственный Университет Науки и Технологии, Улан-батор, 14191, Монголия

Аннотация. В данной работе мы разрабатываем новое семейство итерационных методов высокого порядка без использования производных для решения систем нелинейных уравнений. В частности, мы предлагаем четыре двухшаговые схемы без использования производных с порядками сходимости четыре и пять, а также двенадцать трёхшаговых схем без использования производных, достигающих порядков сходимости шесть, семь и восемь. Главная особенность этих итераций заключается в том, что они включают векторный или даже скалярный параметр итерации вместо матричного параметра, присущего другим существующим итерационным методам. Это структурное упрощение значительно снижает вычислительные затраты, требования к хранению данных и матричные операции, тем самым повышая общую вычислительную эффективность. Представлен анализ сходимости, устанавливающий теоретический порядок сходимости предлагаемых методов. Выведены показатели эффективности предложенных схем и проведено их сравнение с показателями нескольких известных итерационных методов без использования производных. Численные эксперименты на стандартных академических задачах подтверждают теоретические результаты и демонстрируют, что предложенные методы являются конкурентоспособными и во многих случаях превосходят другие методы с точки зрения эффективности и устойчивости.

Ключевые слова: нелинейные системы, итерации без производных, индекс эффективности, порядок сходимости

Discrete & Continuous Models
& Applied Computational Science
ISSN 2658-7149 (Online), 2658-4670 (Print)

2026, 34 (1) 55–69
http://journals.rudn.ru/miph

Research article

UDC 004.942
PACS 07.05.Tp,
DOI: 10.22363/2658-4670-2026-34-1-55-69

EDN: UNFRNH

Simulation of the evacuation of passengers and crew from aircraft during a fire on the ground

Aleksandr S. Baklashov, Leonid Yu. Filimonyuk

V. A. Trapeznikov Institute of Control Sciences of Russian Academy of Sciences, 65 Profsoyuznaya St, Moscow, 117997, Russian Federation

(received: January 12, 2026; revised: January 25, 2026; accepted: January 30, 2026)

Abstract. *Background* Currently, incidents, including fires on board of aircrafts during takeoff and landing, are becoming more frequent. To address this issue, we introduce new models of fire propagation dynamics and the evacuation process for aircraft passengers, accounting for their physical interactions, along with an integrated model combining such processes as the spread of fire, smoke, and temperature. Nowadays aviation incidents involving onboard fires occur regularly, often resulting in traumas among passengers, as well as material damage. *Purpose* The main purpose of this study is to create integrated models that enable analysis of aircraft evacuation under various fire hazard scenarios. Much attention is given to using these models to analyze the process of leaving the aircraft, taking into account various scenarios of the spread of damaging fire factors, which will allow us to develop an optimal sequence of actions for each particular situation. *Method* It uses mathematical apparatus of the multi-dimensional cellular automata to describe fire spread, dividing the aircraft into cubic cells with 4 states: burning, burned, consisting of combustible, and non-combustible materials. Calculation of the probabilities of combustion is based on the influence of the neighboring cells, while evacuation models incorporate multi-agent approaches considering passengers' movements, physical contacts, and hazardous factor distributions. The model was created, and graphs were obtained using Python 3.12. *Results* The results indicate that the integrated model accurately simulates fire dynamics and evacuation interactions, allowing us to analyze different scenarios to make scenario-based predictions of optimal post-accident exit routes. The model was implemented for two scenarios: a fire in the left engine of the Embraer E-190 and Airbus A320-100 aircraft. *Conclusions* Based on the findings, it can be concluded that this approach facilitates decision support systems for enhancing safety during ground-based aircraft fires, providing the model for analyzing and minimizing risks in sudden emergencies.

Key words and phrases: multi-agent model, passenger evacuation, fire, aircraft

For citation: Baklashov, A. S., Filimonyuk, L. Y. Simulation of the evacuation of passengers and crew from aircraft during a fire on the ground. *Discrete and Continuous Models and Applied Computational Science* **34** (1), 55–69. doi: 10.22363/2658-4670-2026-34-1-55-69. edn: UNFRNH (2026).

1. Introduction

Aviation incidents involving fires on board of aircrafts are regularly recorded in Russia and worldwide [1–3]. Such incidents often result in injuries and even victims among passengers and crew members. Nevertheless, in nearly all cases, they cause substantial material damage [4].

Strict guidelines exist [5] that regulate the actions of the crew and passengers during aircraft evacuation. However, these guidelines are typically applicable in certain key details only to specific aircraft types and cannot be effectively implemented in a general context [6].

At the same time, it is important to note that analysis of statistical data reveals a notable trend driven by the increasing adoption of onboard decision support systems [7]. The relative proportion of "sudden" accidents, where there is virtually no time for decision-making, has been rising over the years compared to "anticipated" accidents, which allow sufficient time for preparing the aircraft for landing and subsequent evacuation [8].

In this field of research, we cannot ignore works about crowds in confined spaces [9, 10] and their behavior dictated by people's psychological features [11–13]. However, these works don't take into account the features of enclosed spaces such as the cabin of the aircraft.

Both domestic and international researchers conducted studies on modeling fire dynamics in confined spaces [14–17]. However, these works are primarily useful for analyzing individual hazardous factors. Some of the models considering people's interaction neglect the effects of physical connection and inertial forces between agents [18, 19]. This limitation can be addressed through the application of integrated model research findings [20–23], although these models, in turn, do not account for the unique characteristics of fires occurring on board of aircrafts.

This determines the relevance of research focused on developing integrated models that allow simultaneous consideration of diverse hazardous factors during emergency landings (e.g., the spread of carbon monoxide, high combustion temperatures, flooding in water landings, and others). Such models can be employed to analyze the aircraft evacuation process under various scenarios of hazardous factor spreading, thereby allowing the development of optimal action sequences for each specific situation.

1.1. Structure of the paper

The article includes several sections, each addressing a specific aspect of the research:

The section "Theoretical Basis" defines models of spreading fire and carbon monoxide, and mathematical model of people evacuation from the plane.

The section "Results" presents the graphs obtained by application of the model in specific conditions.

The section "Discussion" summarizes the experimental findings.

The section "Conclusion" outlines the main outcomes and discusses directions for future research.

2. Theoretical Basis

2.1. Problem statement

The specific problem addressed in this study is the development of new models for the dynamics of fire propagation and the evacuation process of aircraft crew and passengers, taking into account their physical interactions. It also includes development of the integrated model that combines the processes of fire, smoke, and temperature spread with evacuation.

2.2. Proposed mathematical models of the spread of fire hazards on board an aircraft

2.2.1. Fire propagation model

We propose to use the mathematical framework of multidimensional cellular automata for describing the fire propagation process.

The aircraft, including the volume of the passenger cabin and the cockpit, is discretized into cubic cells. This yields a three-dimensional grid $A = \{a_{ijk} \mid 0 \le i < n,\ 0 \le j < m,\ 0 \le k < l\}$, where n, m, l denote the number of cubes along the horizontal, vertical, and height axes, respectively; c_{ijk} represents the elementary cube (EC) at coordinates (i, j, k).

Let

- F_t be the set of ECs where combustion is occurring at the given time instant.
- V_t be the set of ECs where the material has fully burned out, preventing any further combustion.
- M_t be the set of ECs where combustion is fundamentally possible due to the presence of combustible material but has not yet initiated at time t.
- N be the set of ECs where fire is impossible (non-combustible material).

The direction of fire propagation at time $t + 1$ is governed by the ignition probability P^t_{ijk} of an EC in state I_t (a state that subsequently assumes values from the aforementioned sets). The ignition probability of an EC can be determined via the expression:

$$P^t_{ijk} = \upsilon_{ijk} f^t_{ijk}\, dt/8l, \tag{1}$$

where υ_{ijk} is the fire propagation velocity of the material composing EC c_{ijk}; f^t_{ijk} is a parameter characterizing the state of neighboring ECs; and dt is the model time step.

The parameter f^t_{ijk} is calculated using the formula:

$$f^t_{ijk} = 3q^t_{ijk1} + 2q^t_{ijk2} + q^t_{ijk3},$$

where q^t_{ijk1} is the number of ECs sharing a face with the considered EC, where combustion is occurring at time t. Their coordinates (see Table 1): $(i, j, k+1)$, $(i-1, j, k)$, $(i, j-1, k)$, $(i, j+1, k)$, $(i+1, j, k)$, $(i, j, k-1)$;

q^t_{ijk2} is the number of ECs sharing an edge but not a face with the considered EC, where combustion is occurring at time t. Their coordinates (see Table 1): $(i-1, j, k+1)$, $(i, j-1, k+1)$, $(i, j+1, k+1)$, $(i+1, j, k+1)$, $(i-1, j-1, k)$, $(i-1, j+1, k)$, $(i+1, j-1, k)$, $(i+1, j+1, k)$, $(i-1, j, k-1)$, $(i, j-1, k-1)$, $(i, j+1, k-1)$, $(i+1, j, k-1)$;

q^t_{ijk3} is the number of ECs sharing a vertex but neither an edge nor a face with the considered EC, where combustion is occurring at time t. Their coordinates (see Table 1): $(i-1, j-1, k+1)$, $(i-1, j+1, k+1)$, $(i+1, j-1, k+1)$, $(i+1, j+1, k+1)$, $(i-1, j-1, k-1)$, $(i-1, j+1, k-1)$, $(i+1, j-1, k-1)$, $(i+1, j+1, k-1)$.

The parameter f^t_{ijk} can take one of the values 0, 1, 2, ..., 50, determined according to the following principle (see Fig. 1 and Table 2).

Each non-boundary elementary cube has:

- 6 ECs sharing a common *face* with the considered EC, and each of these six takes the value 3.
- 12 ECs sharing a common *edge but not a face* with the considered EC, and each of these 12 takes the value 2.
- 8 ECs sharing a common *vertex but not a face or edge* with the considered EC, and each of these 8 takes the value 1.

Table 1

Coordinates of neighboring elementary cubes to the considered one

Upper row of EC			**Middle row of EC**			**Lower row of EC**		
(i-1, j+1, k+1)	(i, j+1, k+1)	(i+1, j+1, k+1)	(i-1, j+1, k)	(i, j+1, k)	(i+1, j+1, k)	(i-1, j+1, k-1)	(i, j+1, k-1)	(i+1, j+1, k-1)
(i-1, j, k+1)	(i, j, k+1)	(i+1, j, k+1)	(i-1, j, k)	(i, j, k)	(i+1, j, k)	(i-1, j, k-1)	(i, j, k-1)	(i+1, j, k-1)
(i-1, j-1, k+1)	(i, j-1, k+1)	(i+1, j-1, k+1)	(i-1, j-1, k)	(i, j-1, k)	(i+1, j-1, k)	(i-1, j-1, k-1)	(i, j-1, k-1)	(i+1, j-1, k-1)

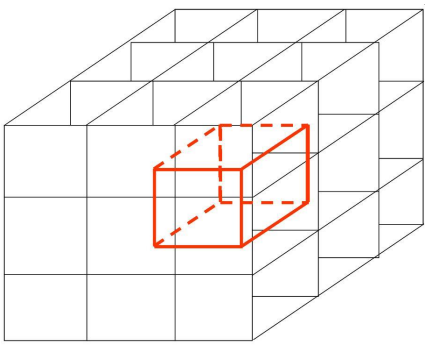

Figure 1. Graphical representation of the EC (highlighted in red) and all its neighboring elementary cubes

f_{ijk}^t will be equal to the sum of the values of neighboring burning ECs.

This value is chosen based on the fact that the distance between the centers of ECs sharing a common *face* with the considered EC is smaller than the distance between the centers of ECs sharing a common *edge*. At the same time, it is still smaller than between the centers of ECs sharing a common *vertex* with the considered EC.

For edge ECs, the parameter f_{ijk}^t can take a value from the set $\{0, 1, 2, \dots, 35\}$, and for corner ones — from $\{0, 1, 2, \dots, 24\}$.

The elementary cube transitions from the burning state at moment t to the burned state at $t + 1$, if no combustible mass remains in it. For each EC a_{ijk}, the mass of combustible substance m_{ijk} and burning speed v_{ijk} are specified.

The mass of the combustible substance of the combustible material m_{ijk} is determined as follows:

$$m_{ijk}^{t+1} = m_{ijk}^t - v_{ijk}\, dt.$$

Figure 2 shows an example of fire propagation on board a twin-engine aircraft.

2.2.2. Carbon monoxide propagation model

In the proposed model, the mass of combustible material in each burning EC changes according to the relation

$$m_{ijk}^{t+1} = m_{ijk}^t - v_{ijk}\, dt.$$

Table 2

Values of the terms in the parameter f_{ijk}^{t} for neighboring elementary cubes

Upper row of EC			**Middle row of EC**			**Lower row of EC**		
1	2	1	2	3	2	1	2	1
2	3	2	3	EC	3	2	3	2
1	2	1	2	3	2	1	2	1

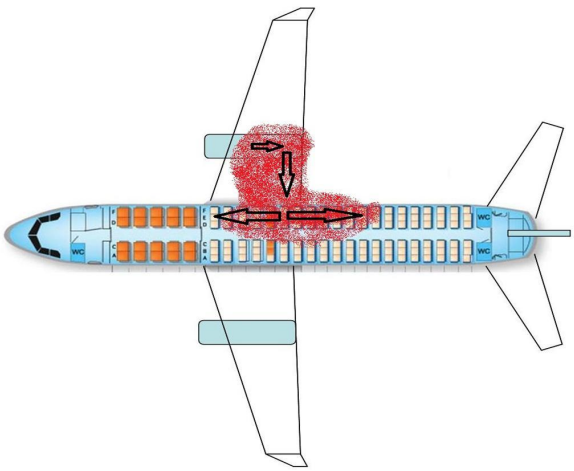

Figure 2. Dynamics of fire propagation on board an aircraft during an engine fire

During combustion, carbon monoxide is emitted. The volume of this emission depends on the type of material being burned. The density of carbon monoxide μ_{ijk} in the elementary cube a_{ijk} increases according to the relation:

$$\mu_{ijk}^{t+1} = \mu_{ijk}^{t} + U_{ijk}\upsilon_{ijk}\, dt/h^3,$$

where U_{ijk} is the CO emission coefficient of the burning material, h^3 is the cell volume.

The propagation of carbon monoxide is determined by the expression:

$$\mu_{ijk}^{t+1} = \mu_{ijk}^{t} + d'_{ijk} \sum_{s \in S'_{ijk}} d'_s(\mu_s^t - \mu_{ijk}^t),$$

where d'_{ijk}, d'_s are coefficients regulating the rate of carbon monoxide propagation; μ_s^t is the density of carbon monoxide for EC.

2.3. Mathematical model of the evacuation process of people from the aircraft

In the proposed mathematical model of the evacuation process, the following objects will be considered:

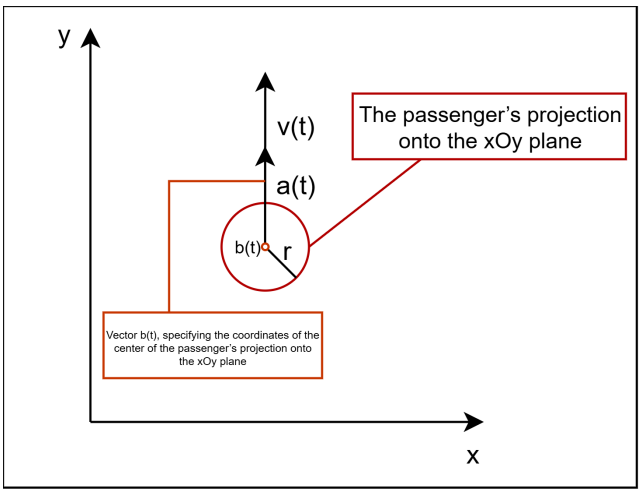

Figure 3. Model of the passenger projected onto the xOy plane

1. A finite set of obstacles in the aircraft cabin, each represented by parameters: (x_W, y_W) are the coordinates of the lower left corner, x_{WW} and y_{WW} — their length and width, respectively.
2. A finite set of exits from the aircraft, specified by the coordinates of the lower left corner (x_E, y_E), width and length of the opening x_{WE} and y_{WE}. The exit zone outside of the aircraft should be located at some distance from the door opening, as after exiting the aircraft, the passenger in some cases still influences the evacuation process.
3. A finite set of evacuating passengers, which we will consider as a set of their projections onto the xOy plane in the form of circles (see Figure 3). The coordinates of these passengers are the centers of the corresponding circles.
4. A finite set of evacuation zones, each specified by parameters: (x_Z, y_Z) are the coordinates of the lower left corner of the evacuation zone, x_{WZ} and y_{WZ} — its length and width, respectively. It is assumed that passengers are located within these zones at the initial moment of evacuation.

The parameters considered in the model for passengers (or agents in terms of multi-agent systems theory):

- vector $b(t)$, specifying the coordinates of the center of the passenger's projection onto the xOy plane.
- r — projection radius.
- m — human mass.
- $v(t)$ — velocity vector of movement.
- $a(t)$ — acceleration vector of movement.
- $v_{\max}$ — maximum possible speed.
- $a_{\max}$ — maximum possible acceleration.

Figure 3 shows the passenger's projection onto the xOy plane.
In the proposed model, the law of motion for passengers is determined by the relations:

$$b(t + \Delta t) = b(t) + v(t)\Delta t,$$
$$v(t + \Delta t) = v(t) + a(t)\Delta t,$$

where Δt is the model time step.

Passenger speeds change during collisions. We will consider collisions as partially elastic impacts. To describe a partially elastic collision, we introduce a restitution coefficient $0 \le R \le 1$. We calculate the normal components of the motion velocities to the common tangent plane to the surfaces of the colliding bodies at their contact point (collision plane) after the impact in a partially elastic impact using the formulas:

$$p_{1n} = -Rv_{1n} + (1+R)\frac{m_1 v_{1n} + m_2 v_{2n}}{m_1 + m_2}, \qquad p_{2n} = -Rv_{2n} + (1+R)\frac{m_1 v_{1n} + m_2 v_{2n}}{m_1 + m_2}. \tag{2}$$

Here, v_{1n} and v_{2n} are the normal projections of the passengers' motion velocities to the collision plane before the impact, and p_{1n} and p_{2n} are the normal projections of the passengers' motion velocities to the collision plane after the impact; m_1 and m_2 are the masses of the colliding agents. We assume that the tangential projections of the velocities to the collision plane do not change during the impact.

When a passenger collides with an obstacle, the projection of their motion velocity that is in parallel to the obstacle does not change. The other projection changes its sign to the opposite and decreases its value depending on R.

Let us determine the direction of the acceleration vector $a(t)$ for each passenger, assuming that each strives to reach the nearest aircraft exit as quickly as possible. For this, let us introduce the concept of the optimal (from the passenger's perspective) velocity vector $v_{\text{opt}}(t)$, approximating the passenger to the chosen exit and allowing avoidance of collisions with obstacles such as the other passengers and obstacles. Also, the actual speed and direction of the agent's movement $v(t)$ may change due to these collisions. Then, it can be considered that passengers move with acceleration $a(t)$, whose absolute magnitude depends on physical capabilities: $|a(t)| = a_{\max}$, and the direction matches with the direction of $v_{\text{opt}}(t) - v(t)$. If $v_{\text{opt}}(t) = v(t)$, it is considered that the agent moves without acceleration. The absolute magnitude of acceleration is always at its maximum value, except in the cases when the passenger is already moving with maximum speed. It can be explained by the *stressful* situation and the passenger's desire to leave the aircraft as quickly as possible.

The absolute magnitude of the optimal velocity $v_{\text{opt}}(t)$ is bounded above by the agent's physical capabilities, namely the speed $v_{\max}$. If there are no obstacles ahead along the route, then $|v_{\text{opt}}(t)| = v_{\max}$. To avoid collision with other agents along the route, the absolute magnitude $|v_{\text{opt}}(t)|$ may decrease.

Let us consider the approach to choosing the vector $v_{\text{opt}}(t)$. Let w be the vector specifying the direction to the nearest exit, accounting for the cabin layout. The vector w is an attribute of the space cell where the agent is located and is determined by the positions of obstacles and exit zones. Let h_γ be the distance from the agent's center to the nearest obstacle in the motion direction at angle γ to w, B is the given critical distance, r is the radius of the passenger's projection onto the xOy plane.

Consequently, the absolute magnitude of the optimal speed $v_{\gamma\text{opt}}(\gamma)$ of the passenger in the motion direction at angle γ to the vector w can be calculated by formula:

$$v_{\gamma\text{opt}}(\gamma) = \begin{cases} v_{\max} & \text{if } h_\gamma \ge B + r, \\ \frac{v_{\max}(h_\gamma - r)}{B} & \text{if } r \le h_\gamma \le B + r, \\ 0 & \text{if } h_\gamma \le r. \end{cases}$$

Passengers' physical capabilities vary, so each passenger must have individual values of $v_{\max}$ and $a_{\max}$.

Let us introduce the function $g(\gamma)$:

$$g(\gamma) = v_{\gamma\text{opt}}(\gamma)\cos(\gamma), \quad \gamma \in [-\pi/2, \pi/2]. \tag{3}$$

It can be inferred from formula (3) that the angle α between the vectors v_{opt} and w is defined as the angle at which the function $g(\gamma)$ is maximal.

The absolute magnitude of v_{opt} determined as:

$$|v_{\text{opt}}| = v_{\gamma\text{opt}}(\alpha).$$

The presented method for choosing v_{opt} allows the agent to maneuver between other agents, aiming to avoid collisions and approach the nearest exit faster (assuming the passenger knows where the exit is).

Thus, the choice of the optimal velocity vector v_{opt} accounts for the cabin layout and the positions of other agents. This vector must be recalculated for each passenger at each model time step Δt. Agents also exhibit additional properties such as heterogeneity due to differences in their mass and projection radius, as well as goal-directedness, since all presented passengers strive to exit the aircraft cabin.

Furthermore, we have to consider the impact of carbon monoxide spreading on the velocity of agents. This impact can be calculated by formula:

$$v'_{\max} = \begin{cases} v_{\max}, & \text{if } \mu_{ijk} \le 0.1, \\ v_{\max} \cdot \max\left(0.6,\, 1 - 1.1 \cdot \mu_{ijk}\right), & \text{if } \mu_{ijk} > 0.1. \end{cases}$$

It should be noted that the presented model also includes known attributes such as the angle and range of the agent's view of the surrounding space, moment of inertia, and head rotation angle. The direction leading the agent to the exit (considering obstacles) is determined by the vector w. The agent's view angle can be considered equal to 180°, since $\gamma \in [-\pi/2, \pi/2]$, i.e., the agent analyzes all possible alternatives for movement in the plane ahead.

Since the agent calculates the optimal speed values based on analyzing the situation within the critical zone, the agent's view range is at least B.

Part of the energy during impact transfers to rotational motion. In the model, these energy losses are accounted for by the restitution coefficient.

2.4. Algorithm for modeling the evacuation process

We created an algorithm for modeling emergency situations (an event of fire on the ground in our case) consisting of such steps:

1. Input initial data: cabin dimensions, coordinates of obstacles, exits from the cabin, number of agents and their parameters.
2. Generate agents inside evacuation zones according to seating positions.
3. Calculate v_{opt} for each agent.
4. Calculate a for each agent.
5. Calculate v for each agent.
6. Calculate b for each agent.
7. Check for collisions for each pair of agents. If a collision occurs, recalculate velocities.
8. Check collisions with obstacles for each agent. If a collision occurs, recalculate velocities.
9. Check the condition for reaching the emergency exit for each agent. If the agent has reached the exit, exclude them from the list.

10. Collect and store statistics. If no passengers remain in the cabin, proceed to step 12.
11. Proceed to the next model time step. Go to step 3.
12. Display the results of the experiments (graphs of functions).
13. Go to step 1 (upon user request).

The model is created according to this algorithm.

3. Results

The model was built using Python 3 according to the proposed theoretical basis, the application of the mathematical apparatus of multidimensional cellular automata, and the rules of mathematical models of fire and carbon monoxide propagation. Also, a mathematical model of the evacuation of passengers was implemented.

All of the mentioned models were integrated into the program according to the algorithm suggested in Section 2.4.

The results include graphs of the dependence between the time and passengers evacuated, as well as of the dependence between the time and burned cubes.

The scenario of the model is the combustion of the left engine of an aircraft. One of the model conventions is the fact that the fire begins its propagation from the cabin (exactly from the part of the cabin in front of the left engine).

The graphs introduced for the aircraft models, such as the Embraer E-190 with 100 seats and a full load of passengers and the Airbus A320-100 with 150 seats and also a full load of passengers.

Let us introduce the common parameters of the model for Airbus A320-100 and Embraer E-190. (see Table 3).

As we can observe, the parameters in question are primarily relevant to passengers and the general model configuration. A uniform distribution is set for the parameters that are calculated at each step of the algorithm.

Also, there are differing parameters to consider (see Table 4).

Figures 4 , 5 display the dynamic of evacuation from the aircraft and fire propagation via displaying the cubic cells that have been burned.

In comparing the scenarios for the Embraer E-190 and Airbus A320-100, it is noted that the bigger A320, accommodating 50% more passengers and featuring a broader cabin, requires approximately 50 seconds longer for total evacuation (140 seconds against 90 seconds). Meanwhile, the fire extends to 43 additional burned cells, and the time increases to 140 seconds, illustrating how larger cabin volume and passengers' capacity affect the model. The uniform distribution applied for parameters like v_{max}, a_{max}, r and m makes this model and the result more realistic. These findings highlight the vital role of right exits and evacuation zone locations.

4. Discussion

The results obtained from the integrated model show that combining cellular automata for fire and carbon monoxide propagation with the dynamics of agent evacuation provides a method to predict the evacuation process from aircraft in the event of a fire on the ground.

For example, in the Embraer E-190 case (Fig. 4), 100 passengers could evacuate in about 90 seconds when only 94 cubic cells was burned. As the maximum count of burning EC is 528, and for this time only 94 of them were burned, this time is enough to evacuate all of the passengers without injuries.

Table 3

Common parameters of the E190 and A320 models

Parameter	**Value**	**Description**
Cell size	0.5 m	Size of elementary cube (EC) in meters
Time step Δt	0.05 s	Model time step
Restitution coefficient R	0.2	Coefficient for collision handling
Critical distance B	2.0 m	Distance for optimal velocity calculation
Initial temperature	20°C	Initial EC temperature
Maximum agent velocity v_{max}	1.2–1.8 m/s (uniform distribution)	Maximum speed of agents
Maximum acceleration a_{max}	2.0–4.0 m/s^2 (uniform distribution)	Maximum acceleration of agents
Reaction time	0–8 s (uniform distribution)	Delay before agent starts moving
Agent radius r	0.20–0.28 m (uniform distribution)	Radius of projection for agents
Agent mass m	50–90 kg (uniform distribution)	Mass of agents

Table 4

Differing parameters of the E190 and A320 models

Parameter	**E190 Value**	**A320 Value**
Aircraft model	Embraer E190	Airbus A320-100
Number of passengers	100	150
Cabin length	25.0 m	27.5 m
Cabin width	2.74 m	3.70 m
Cabin height	2.0 m	2.22 m
Grid size	52 x 7 x 6	57 x 9 x 6
Engine fire position	(26, 0, 0)	(28, 0, 0)
Max. burning EC	528	682
Number of exits	4	6
Seat configuration	2+2	3+3

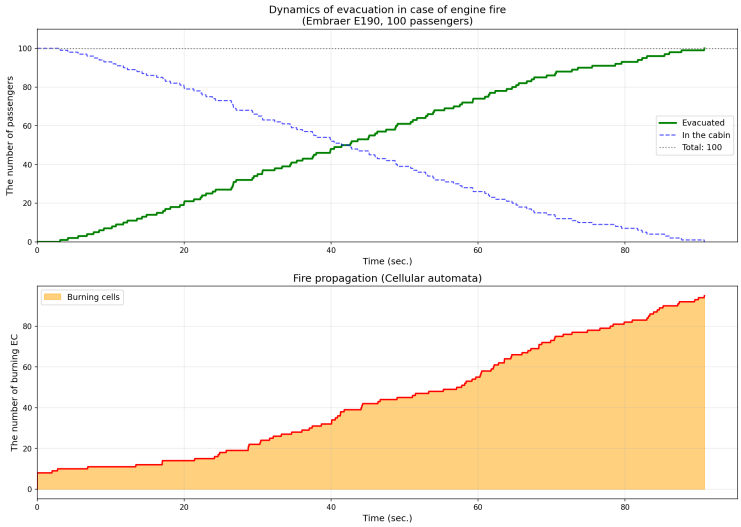

Figure 4. Dynamic of evacuation from the aircraft / fire propagation for Embraer E-190

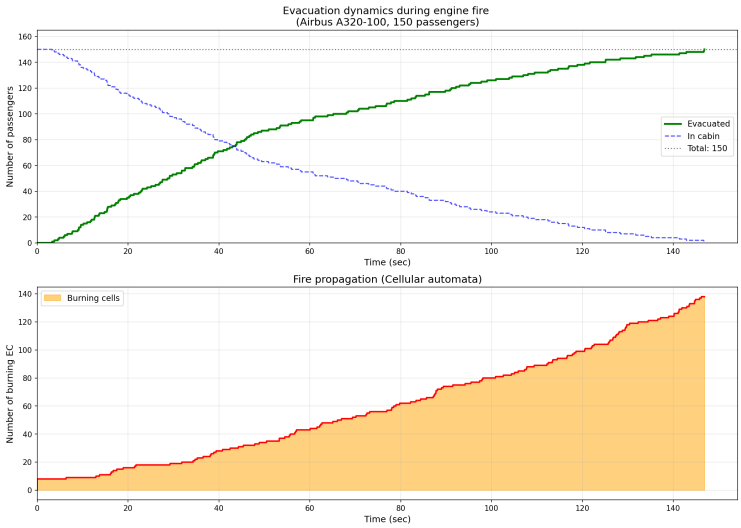

Figure 5. Dynamic of evacuation from the aircraft / fire propagation for Airbus A320-100

At the same time, the Airbus A320-100 scenario (Fig. 5) took longer, around 140 seconds for 150 passengers with 137 cells affected. These results are obtained because of the larger cabin and more people involved.

The fire spread follows a smoothly growing curve due to the mechanism of neighbor influences in the automata, as does the curve of evacuation in both cases. It follows from this that physical interactions, like passenger collisions handled through restitution coefficients (Eq. (2)), and hazard progression (Eq. (1)) lead to accurate and realistic results of the model.

There are no major deviations in the graphs. The variability of the random starting conditions (e.g., reaction times 0–8 s) confirming the model's stability. The suggested approach stands out by its simple adaptation to aircraft specifics and could feed into real decision-support tools to cut down on casualties in the "sudden" incidents.

There is a perspective for the future work, as could be added more data for the output in the model, such as detailed temperature maps, a map of cabin and burned cubes etc. Overall, this research enhances the importance of comprehensive models for aviation safety. This research could reduce sudden accidents by generating results for different scenarios and considering them when designing aircraft.

5. Conclusion

A complex of mathematical models and algorithms has been developed that enables the simulation of situations arising during aircraft fires, the calculation of the dynamics of hazardous fire factors, and passenger evacuation under the spread of these factors.

The integrated model of the dynamics of hazardous fire factors development and the process of evacuating passengers and crew from the aircraft is built on the basis of the mathematical apparatus of cellular automata and multi-agent systems.

It is proposed to use the integrated model for analyzing the process of leaving the aircraft, taking into account various scenarios of the spread of fire's hazardous factors, which will allow developing optimal lists of actions for each specific situation.

Author Contributions: For research articles with several authors, a short paragraph specifying their individual contributions must be provided. Conceptualization, Aleksandr S. Baklashov and Leonid Yu. Filimonyuk; methodology, Leonid Yu. Filimonyuk; software, Aleksandr S. Baklashov; validation, Aleksandr S. Baklashov and Leonid Yu. Filimonyuk; formal analysis, Leonid Yu. Filimonyuk; investigation, Aleksandr S. Baklashov and Leonid Yu. Filimonyuk; resources, Aleksandr S. Baklashov; data curation, Aleksandr S. Baklashov; writing—original draft preparation, Leonid Yu. Filimonyuk; writing—review and editing, Aleksandr S. Baklashov; visualization, Aleksandr S. Baklashov; supervision, Leonid Yu. Filimonyuk.; project administration, Leonid Yu. Filimonyuk. All authors have read and agreed to the published version of the manuscript. All authors have read and agreed to the published version of the manuscript.

Funding: This research received no external funding.

Data Availability Statement: Data sharing is not applicable.

Conflicts of Interest: The authors declare no conflict of interest.

Declaration on Generative AI: The authors have not employed any Generative AI tools.

References

1. *Report of the Interstate Aviation Committee on the Results of the Investigation of the Aviation Accident. Airbus A-310 Catastrophe* tech. rep. (Interstate Aviation Committee, 2006).
2. Suharev, A., Šestakovs, V. & Vinogradov, L. Estimation of evacuation time of passengers in aircraft accidents with fire in airfield areas. *Aviation* **24,** 72–79. doi:10.3846/aviation.2020.12653 (July 2020).
3. *Report of the Interstate Aviation Committee on the Results of the Investigation of the Aviation Accident. An-24RV Catastrophe* Russian. Tech. rep. (Interstate Aviation Committee, 2011).
4. Wang, J., Tao, Z., Yang, R., Gao, Z., Shan, D. & Wang, W. A review of aircraft fire accident investigation techniques: Research, process, and cases. **153.** doi:10.1016/j.engfailanal.2023.107558 (2023).

5. Butcher, N., Barnett, J. & Buckland, T. *Emergency evacuation of commercial passenger aeroplanes* June 2020.
6. Choochart, P. & Thipyopas, C. Study of Passenger Evacuation from the Airbus A330-300 Aircraft. *Proceedings* **39.** doi:10.3390/proceedings2019039025 (2019).
7. Lv, W., Xing, L., Li, J., Zhao, C. & Yang, Y. Evaluating personnel evacuation risks under fire scenario of Airbus wide-body aircraft: A simulation study. English. *Frontiers in Public Health* **10.** doi:10.3389/fpubh.2022.994031 (Sept. 2022).
8. Boyd, D. D. & Howell, C. Accident Rates, Causes, and Occupant Injury Involving High-Performance General Aviation Aircraft. *Aerospace Medicine and Human Performance* **91,** 387–393. doi:10.3357/AMHP.5509.2020 (May 2020).
9. Winter, H. Modelling Crowd Dynamics During Evacuation Situations Using Simulation, 20 (2012).
10. Beklaryan, A. Exit Front in the Model of Crowd Behavior in Emergency Situations. Russian. *Bulletin of Tambov University. Series: Natural and Technical Sciences* **20,** 851–856 (2015).
11. Cherif, F. & Chighoub, R. Crowd simulation influenced by agent's socio-psychological state. *Journal of computing* **2,** 48–54 (2010).
12. Henderson, L. F. The statistics of crowd fluids. *Nature* **229,** 381–383 (1971).
13. Helbing, D., Farkas, I. & Vicsek, T. Simulating dynamical features of escape panic. *Nature* **407,** 487–490 (2000).
14. Fedosov, S., Ibragimov, A., Solov'ev, R., *et al.* Mathematical Model of Fire Development in a System of Premises. Russian. *Vestnik MGSU,* 121–128 (2013).
15. Drysdale, D. *Introduction to Fire Dynamics* Russian. 424 pp. (Stroyizdat, 1990).
16. Apiecionek, L., Zarzycki, H., Czerniak, J., *et al.* The Cellular Automata Theory with Fuzzy Numbers in Simulation of Real Fires in Buildings. *Advances in Intelligent Systems and Computing* **559,** 169–182 (2018).
17. Zhang, Y., Yang, Z. & Sun, Z. A dynamic estimation method for aircraft emergency evacuation based on cellular automata. *Advances in Mechanical Engineering* **11,** 12. doi:10 . 1177 / 1687814019825702 (2019).
18. Akopov, A. S. & Beklaryan, L. A. An agent model of crowd behavior in emergencies. **76,** 1817–1827. doi:10.1134/S0005117915100094 (Oct. 2015).
19. Korepanov, V. O. Simulation models of agents' tactical behavior. *Management of large systems: collection of works,* 145–157 (2009).
20. Samartsev, A. *et al.* Mathematical Model of the Dynamics of Fire Development in Premises. Russian. *Large-Scale Systems Control,* 42–62 (2018).
21. Samartsev, A. *et al.* Fire and Heat Spreading Model Based on Cellular Automata Theory. *Journal of Physics: Conference Series* **1015,** 5 (2018).
22. Göttlich, S., Kühn, S., Ohst, J. P., Ruzika, S. & Thiemann, M. Evacuation dynamics influenced by spreading hazardous material. *Networks and Heterogeneous Media* **6,** 443–464. doi:10.3934/nhm.2011.6.443 (2011).
23. Korolchenko D.A., P. S. Introduction of a flame suppression pattern into integrated and zone models used to analyze the dynamics of hazardous factors of indoor fires. Russian. *Fire and Explosion Safety* **30,** 78–87. doi:10.22227/PVB.2021.30.02.78-87 (May 2021).

Information about the authors

Baklashov, Aleksandr S.—Postgraduate Student of V.A. Trapeznikov Institute of Control Sciences of RAS; Junior Researcher of Laboratory of Technical Diagnostics and Fault Tolerance, V. A. Trapeznikov Institute of Control Sciences of Russian Academy

of Sciences (e-mail: baklashov2001@mail.ru, ORCID: 0009-0000-9046-3225, ResearcherID: KLZ-4503-2024)

Filimonyuk, Leonid Yu.—Doctor of Technical Sciences, Leading Researcher of of Laboratory of Technical Diagnostics and Fault Tolerance, V. A. Trapeznikov Institute of Control Sciences of Russian Academy of Sciences (e-mail: filimonyukleonid@mail.ru, ORCID: 0000-0002-0007-3969, ResearcherID: F-2611-2017, Scopus Author ID: 57189024654)

УДК 004.942
PACS 07.05.Tp,
DOI: 10.22363/2658-4670-2026-34-1-55-69 EDN: UNFRNH

Моделирование эвакуации пассажиров и экипажа из воздушных судов при пожаре на земле

А. С. Баклашов, Л. Ю. Филимонюк

Институт проблем управления им. В. А. Трапезникова Российской академии наук,
ул. Профсоюзная, д. 65, Москва, 117997, Российская Федерация

Аннотация. *Предпосылки* В настоящее время происшествия, в том числе пожары на борту самолёта при взлете/посадке, возникают всё чаще. Для исследования этой проблемы в статье представлены новые модели динамики распространения пожара и процесса эвакуации пассажиров воздушного судна с учётом их физического взаимодействия, а также интегрированная модель, объединяющая процессы, такие как распространение огня, дыма и температуры. *Цель* Основная цель данного исследования заключается в создании комплексных моделей, позволяющих анализировать процесс эвакуации из воздушного судна при различных сценариях. Особое внимание уделяется использованию этих моделей для анализа процесса покидания самолёта с учётом различных сценариев распространения поражающих факторов пожара, что позволит разработать оптимальную последовательность действий для каждой конкретной ситуации. *Методы* Для описания распространения огня используется математический аппарат многомерных клеточных автоматов, разделяющих воздушное судно на кубические ячейки, которым присваиваются 4 различных состояния: горения, выгоревшего, состоящего из горючего и негорючего материалов. Вероятности возгорания рассчитываются на основе влияния соседних ячеек, а модели эвакуации объединяют в себе мультиагентные подходы, учитывающие движения пассажиров, физические контакты и распределения опасных факторов. Модель была создана и графики получены с использованием языка программирования Python 3.12. *Результаты* Результаты показывают, что интегрированная модель точно симулирует динамику пожара и действия пассажиров при эвакуации. Также она позволяет анализировать различные сценарии для прогнозирования оптимальных путей эвакуации после аварии. Модель была реализована для двух сценариев: возгорания в левом двигателе самолётов Embraer E-190 и Airbus A320-100. *Заключение* В заключение можно констатировать, что предложенный подход способствует разработке систем поддержки принятия решений, необходимых для повышения безопасности при пожарах на воздушных судах на земле, предлагая модель для анализа и минимизации рисков в условиях внезапных чрезвычайных ситуаций.

Ключевые слова: мультиагентная модель, эвакуация пассажиров, пожар, воздушное судно

Discrete & Continuous Models & Applied Computational Science
ISSN 2658-7149 (Online), 2658-4670 (Print)

2026, 34 (1) 70–97
http://journals.rudn.ru/miph

Research article

UDC 511.84:512.523.282.2:519.711
DOI: 10.22363/2658-4670-2026-34-1-70-97 EDN: UOBPEG

Dual quaternion representation of geometrical motion in 3D space

Olesya M. Abakumova[1], Migran N. Gevorkyan[1], Anna V. Korolkova[1], Dmitry S. Kulyabov[1,2]

[1] RUDN University, 6 Miklukho-Maklaya St, Moscow, 117198, Russian Federation

[2] Joint Institute for Nuclear Research, 6 Joliot-Curie St, Dubna, 141980, Russian Federation

(received: December 4, 2025; revised: December 28, 2025; accepted: January 10, 2025)

Abstract. *Background* In a previous article we discussed the use of dual quaternions for modeling points, lines and planes and solving standard geometric problems. This article is a logical continuation and reveals the use of dual quaternions to describe isometries of three-dimensional space. *Purpose* The derivation of all necessary formulas for the screw motion of points, straight lines and planes, as well as reflection relative to the plane. Refinement of notation and formalism. *Method* The algebra of dual numbers, quaternions and dual quaternions is used, as well as elements of the theory of screws and sliding vectors. *Results* Formulas for rotation, translation, reflection, helical motion, and mirror rotation are obtained and systematized. *Conclusions* Dual quaternions can serve as a full-fledged tool for describing helical motion in space. Due to the possibility of expressing dual quaternion operations in terms of standard vector and scalar products, the formulas obtained allow for effective software implementation.

Key words and phrases: natural modeling, reproducible research, research as code

For citation: Abakumova, O. M., Gevorkyan, M. N., Korolkova, A. V., Kulyabov, D. S. Dual quaternion representation of geometrical motion in 3D space. *Discrete and Continuous Models and Applied Computational Science* **34** (1), 70–97. doi: 10.22363/2658-4670-2026-34-1-70-97. edn: UOBPEG (2026).

1. Introduction

In the previous article [1], we considered parabolic biquaternions (dual quaternions), the discovery of which is attributed to W. Clifford and systematic study began later in the works of E. Study [2, 3] and A. P. Kotelnikov [4] including under the guise of the theory of screws [5–9].

This article logically continues the previous one [1] and focuses on the application of dual quaternions to the description of isometries (proper and improper motions) of three-dimensional space. There are two main objectives.

- Output the necessary formulas for calculating all possible movements of three-dimensional space for points, lines and planes. These movements include rotations, translations (parallel transfers), and mirror symmetries.
- In the process of deducing formulas, illustrate the work of mathematical formalism, for which the conclusion is given in great detail, with all the details. A number of quaternion formulas have also been preliminarily obtained, also for the purpose of illustrating the notation.

As a novelty of the paper we can mention the description of not only the movement of lines (or screws, as it is done in the fundamental monograph [10]), but also points and planes, as well as consideration of reflections, which is rare in the literature. In the process, we relied on books [11–13], however, we also provide a number of new formulas. We also base our conclusion on the principle of Kotelnikov–Study transference, which is apparently unknown in foreign papers.

Dual quaternions have significant applied importance, as evidenced by publications on the topic [14–19]. In this article, we do not provide any software implementations, since adding this material would make it necessary to shorten the calculations. Examples of software implementation and an illustration of the operation of all the formulas obtained are planned in further publications by the authors.

1.1. Structure of the paper

The article consists of an introduction, 2 parts and a conclusion listing the results.

In the first part of the article, the quaternionic rotation formula is derived. The conclusion is based on the Rodrigues formula, which can be obtained from relatively elementary geometric constructions. The relation of quaternions to rotation matrices is shown below. Much attention is paid to the algorithm for calculating the rotation quaternion according to a given matrix. This algorithm takes into account the problem of rounding when working with floating-point numbers. In the last paragraph, a formula is derived for reflection relative to a plane, including one that does not pass through the origin.

The second part is the main one. It begins with the formulation of the Kotelnikov–Rudy transference principle. Next, this principle is applied to the quaternionic formula of rotation, which allows you to immediately obtain a dual quaternionic formula of helical motion for a straight line. The dual quaternion of helical motion is written out explicitly and a number of properties are proved for it. Next, it is divided into translational and rotational parts, and their actions on the direct line are studied separately. By creating compositions from these parts, it is possible to describe more complex cases, for example, the mismatch of the axis of rotation with the axis of translation.

Further, using the duality principle allows us to generalize the formula of helical motion to the cases of points and planes. For these cases, explicit formulas for pure translation and pure rotation are also given in detail.

This part ends with the derivation of dual quaternion formulas for reflecting straight lines, points, and planes relative to an arbitrary plane.

1.2. Notations and conventions

The following naming conventions are accepted in this article

- Quaternions are indicated by lowercase Latin letters from the end of the alphabet: p, q, r. The components of the quaternions are indicated by the same letters, but with the indexes p_0, p_1, etc.
- Dual quaternions are indicated by uppercase Latin letters from the end of the alphabet: P, Q, R. The components of the quaternions are indicated by the same letters, but with the indexes P_0, P_1, etc.
- Vectors and pure quaternions are indicated by lowercase bold Latin letters: $\mathbf{q}$, $\mathbf{v}$, etc.
- Pure dual quaternions are indicated by uppercase bold Latin letters: $\mathbf{Q}$, $\mathbf{V}$, etc.
- Individual scalars (real numbers) are denoted by the Greek letters α, β, etc.

To avoid ambiguity in the notation system, we do not use multiple quaternions and dual quaternions designated by the same letter, but distinguished by an index. The only exceptions are dual quaternions of points, lines, and planes, the components of which are denoted by letters other than the letters denoting these dual quaternions.

1.3. Description of geometric motion in three-dimensional space using quaternions

By motion in Euclidean space we mean an affine transformation that preserves the scalar product (metric). An affine transformation can be written as:

Motion in three-dimensional space is reduced to three:

- translation (parallel translation);
- rotation;
- reflection.

The linear part of the affine transformation is responsible for the rotation.

1.4. Rotations using quaternions

1.4.1. Sandwich formula for quaternion rotation of a point

The quaternion formula for rotating a point about an axis passing through the origin is widely known [20, 21]. Let a unit quaternion be given

$$\lambda = \lambda_0 o + \ = \lambda_0 o + \lambda_1 \mathrm{i} + \lambda_2 \mathrm{j} + \lambda_3 \mathrm{k}, \quad \lambda_0 = \cos\frac{\theta}{2}, \quad \lambda_i = \sin\frac{\theta}{2} a_i,$$

where $\mathbf{a} = (a_1, a_2, a_3)^T$ is the direction vector of the rotation axis passing through the origin, and θ is the magnitude of the rotation angle. Then, for a point P with homogeneous coordinates $\vec{\mathrm{p}} = (x, y, z \mid w)$, given in quaternion form as $p = wo + x\mathrm{i} + y\mathrm{j} + z\mathrm{k}$, the following sandwich formula holds:

$$p' = \lambda p \lambda^*,$$

where p' is a quaternion that specifies the homogeneous coordinates of the point P', into which the original point P has passed after the rotation.

Note that the interpretation of the scalar component of the quaternion p as a homogeneous coordinate w is not often encountered in the literature, and here we rely on the sources [12].

Let us prove the above formula purely algebraically using the Rodrigues formula in the Rodrigues–Hamilton form. Recall that this formula in the space $\mathbb{R}^3$ is written as

$$\mathbf{p}' = \mathbf{p} + 2\lambda_0 \times \mathbf{p} + 2 \times (\times \mathbf{p}),$$

where $\mathbf{p}$ is the radius vector defining the initial position of the point P, and λ_0, λ_1, λ_2, λ_3 are the Rodrigues–Hamilton parameters that completely coincide with the components of the quaternion λ and obey the condition $\lambda_0^2 + \|\|^2 = 1$. The proof of the Rodrigues–Hamilton formula does not use the concept of quaternions in any way, so if we transform it so as to replace vector multiplication with quaternion multiplication and the components λ_i with the quaternion λ, then we will also prove the quaternion sandwich formula.

We transform the Rodrigues–Hamilton formula using the normalization $\lambda_0^2 + \|\|^2 = 1$

$$\mathbf{p}' = (\lambda_0^2 + \|\|^2)\mathbf{p} + 2\lambda_0 \times \mathbf{p} + 2 \times (\times \mathbf{p}) = \lambda_0^2 \mathbf{p} + \|\|^2 \mathbf{p} + 2\lambda_0 \times \mathbf{p} + 2 \times (\times \mathbf{p}).$$

Using the formula $\mathbf{a} \times (\mathbf{b} \times \mathbf{c}) = \mathbf{b}(\mathbf{a}, \mathbf{c}) - \mathbf{c}(\mathbf{a}, \mathbf{b})$ we write

$$\times (\times \mathbf{p}) = (, \mathbf{p}) - \mathbf{p}(,) = (, \mathbf{p}) - \mathbf{p}\|\|^2,$$

and substitute into the Rodrigues–Hamilton formula

$$\mathbf{p}' = \lambda_0^2\mathbf{p} + \|\|^2\mathbf{p} + 2\lambda_0 \times \mathbf{p} + (, \mathbf{p}) - \mathbf{p}\|\|^2 + \times (\times \mathbf{p}) = \lambda_0^2\mathbf{p} + 2\lambda_0 \times \mathbf{p} + (, \mathbf{p}) + \times (\times \mathbf{p}).$$

We make the following transformation: $\times (\times \mathbf{p}) = -(\times \mathbf{p}) \times$ and add a dummy term $(\times \mathbf{p},) = 0$. Equality to zero is true due to $\times \mathbf{p} \perp$, then

$$\mathbf{p}' = \lambda_0^2\mathbf{p} + 2\lambda_0 \times \mathbf{p} + (, \mathbf{p}) - \big(- (\times \mathbf{p},) + (\times \mathbf{p}) \times \big).$$

Let us now temporarily, within the limits of this derivation, denote quaternion multiplication by the symbol $\circ$ and interpret all vectors in the formula as pure quaternions. We write

$$-(\times \mathbf{p},) + (\times \mathbf{p}) \times = (\times \mathbf{p}) \circ \quad \text{и} \quad (, \mathbf{p}) = (, \mathbf{p}) \circ .$$

and also replace $2 \times \mathbf{p} = \circ \mathbf{p} - \mathbf{p} \circ$ and continue transforming the Rodrigues–Hamilton formula:

$$\begin{aligned}
\mathbf{p}' &= \lambda_0^2\mathbf{p} + 2\lambda_0 \times \mathbf{p} + (, \mathbf{p}) \circ + (\times \mathbf{p}) \circ = \lambda_0^2\mathbf{p} + \lambda_0(\circ \mathbf{p} - \mathbf{p} \circ) - \big(- (, \mathbf{p}) + \times \mathbf{p}\big) \circ = \\
&= \lambda_0^2\mathbf{p} + \lambda_0(\circ \mathbf{p} - \mathbf{p} \circ) - (\circ \mathbf{p}) \circ = \lambda_0^2\mathbf{p} - \lambda_0\mathbf{p} \circ + \lambda_0 \circ \mathbf{p} - (\circ \mathbf{p}) \circ = \lambda_0(\lambda_0\mathbf{p} - \mathbf{p} \circ) + \lambda_0 \circ \mathbf{p} - (\circ \mathbf{p}) \circ = \\
&= \lambda_0\mathbf{p}_0(\lambda_0 -) + \circ \mathbf{p} \circ (\lambda_0 -) = (\lambda_0\mathbf{p} + \circ \mathbf{p}) \circ (\lambda_0 -) = (\lambda_0 +) \circ \mathbf{p} \circ (\lambda_0 -)
\end{aligned}$$

As a result, we obtained a sandwich formula:

$$\mathbf{p}' = (\lambda_0 +) \circ \mathbf{p} \circ (\lambda_0 -) = \lambda \circ \mathbf{p} \circ \lambda^*.$$

Due to $\lambda\lambda^* = 1$ we can finally write:

$$p' = w + \mathbf{p}' = w\lambda\lambda^* + \lambda \circ \mathbf{p} \circ \lambda^* = \lambda \circ (w + \mathbf{p}) \circ \lambda^* \Rightarrow p' = \lambda \circ p \circ \lambda^*.$$

1.4.2. Calculating a rotation matrix given a rotation quaternion

Let's use the Rodrigues formula, written in terms of the Rodrigues–Hamilton coefficients

$$\mathbf{p}' = \mathbf{p} + 2\lambda_0 \times \mathbf{p} + 2 \times \times \mathbf{p},$$

but let us represent the vector product in matrix form as follows:

$$\times \mathbf{p} = \Lambda\mathbf{p} = \begin{bmatrix} 0 & -\lambda_3 & \lambda_2 \\ \lambda_3 & 0 & -\lambda_1 \\ -\lambda_2 & \lambda_1 & 0 \end{bmatrix} \begin{bmatrix} p_1 \\ p_2 \\ p_3 \end{bmatrix}.$$

$$\times \times \mathbf{p} = \Lambda(\Lambda\mathbf{p}) = \Lambda^2\mathbf{p} = \begin{bmatrix} 0 & -\lambda_3 & \lambda_2 \\ \lambda_3 & 0 & -\lambda_1 \\ -\lambda_2 & \lambda_1 & 0 \end{bmatrix} \begin{bmatrix} 0 & -\lambda_3 & \lambda_2 \\ \lambda_3 & 0 & -\lambda_1 \\ -\lambda_2 & \lambda_1 & 0 \end{bmatrix} \begin{bmatrix} p_1 \\ p_2 \\ p_3 \end{bmatrix} =$$

$$= \begin{bmatrix} -(\lambda_2^2+\lambda_3^2) & \lambda_1\lambda_2 & \lambda_1\lambda_3 \\ \lambda_1\lambda_2 & -(\lambda_1^2+\lambda_3^2) & \lambda_2\lambda_3 \\ \lambda_1\lambda_3 & \lambda_2\lambda_3 & -(\lambda_1^2+\lambda_2^2) \end{bmatrix} \begin{bmatrix} p_1 \\ p_2 \\ p_3 \end{bmatrix},$$

where Λ is a matrix composed of the components of the vector part of the unit quaternion λ. Now Rodrigues' formula can be rewritten in matrix form as follows:

$$\mathbf{p}' = I\mathbf{p} + 2\lambda_0\Lambda\mathbf{p} + 2\Lambda^2\mathbf{p} = (I + 2\lambda_0\Lambda + 2\Lambda^2)\mathbf{p},$$

$$I + 2\lambda_0\Lambda + 2\Lambda^2 = \begin{bmatrix} 1 & 0 & 0 \\ 0 & 1 & 0 \\ 0 & 0 & 1 \end{bmatrix} + \begin{bmatrix} 0 & -2\lambda_0\lambda_3 & 2\lambda_0\lambda_2 \\ 2\lambda_0\lambda_3 & 0 & -2\lambda_0\lambda_1 \\ -2\lambda_0\lambda_2 & 2\lambda_0\lambda_1 & 0 \end{bmatrix} +$$

$$= \begin{bmatrix} -2(\lambda_2^2+\lambda_3^2) & 2\lambda_1\lambda_2 & 2\lambda_1\lambda_3 \\ 2\lambda_1\lambda_2 & -2(\lambda_1^2+\lambda_3^2) & 2\lambda_2\lambda_3 \\ 2\lambda_1\lambda_3 & 2\lambda_2\lambda_3 & -2(\lambda_1^2+\lambda_2^2) \end{bmatrix} = \begin{bmatrix} 1-2(\lambda_2^2+\lambda_3^2) & 2(\lambda_1\lambda_2-\lambda_0\lambda_3) & 2(\lambda_0\lambda_2+\lambda_1\lambda_3) \\ 2(\lambda_0\lambda_3+\lambda_1\lambda_2) & 1-2(\lambda_1^2+\lambda_3^2) & 2(\lambda_2\lambda_3-\lambda_0\lambda_1) \\ 2(\lambda_1\lambda_3-\lambda_0\lambda_2) & 2(\lambda_0\lambda_1+\lambda_2\lambda_3) & 1-2(\lambda_1^2+\lambda_2^2) \end{bmatrix} =$$

$$= \begin{bmatrix} \lambda_0^2+\lambda_1^2-\lambda_2^2-\lambda_3^2 & 2(\lambda_1\lambda_2-\lambda_0\lambda_3) & 2(\lambda_0\lambda_2+\lambda_1\lambda_3) \\ 2(\lambda_0\lambda_3+\lambda_1\lambda_2) & \lambda_0^2-\lambda_1^2+\lambda_2^2-\lambda_3^2 & 2(\lambda_2\lambda_3-\lambda_0\lambda_1) \\ 2(\lambda_1\lambda_3-\lambda_0\lambda_2) & 2(\lambda_0\lambda_1+\lambda_2\lambda_3) & \lambda_0^2-\lambda_1^2-\lambda_2^2+\lambda_3^2 \end{bmatrix}.$$

The last matrix was obtained by replacing $1 = \lambda_0^2 + \|\mathbf{q}\| = \lambda_0^2 + \lambda_1^2 + \lambda_2^2\lambda_3^2$.

1.4.3. Calculating a rotation quaternion from a given rotation matrix

Here we present an algorithm for calculating the coefficients of a quaternion λ given a rotation matrix R. The algorithm follows the calculation method described in [22, pp. 90–94], which allows for some compensation for the loss of precision when working with floating-point numbers.

Let us write the rotation matrix in quaternion form

$$R = \begin{bmatrix} 1-2(\lambda_2^2+\lambda_3^2) & 2(\lambda_1\lambda_2-\lambda_0\lambda_3) & 2(\lambda_1\lambda_3-\lambda_0\lambda_2) \\ 2(\lambda_0\lambda_3+\lambda_1\lambda_2) & 1-2(\lambda_1^2+\lambda_3^2) & 2(\lambda_2\lambda_3-\lambda_0\lambda_1) \\ 2(\lambda_1\lambda_3-\lambda_0\lambda_2) & 2(\lambda_2\lambda_3+\lambda_0\lambda_1) & 1-2(\lambda_1^2+\lambda_2^2) \end{bmatrix}$$

where $\lambda = \cos\dfrac{\theta}{2} + \sin\dfrac{\theta}{2}, \ = \lambda_1\mathrm{i} + \lambda_2\mathrm{j} + \lambda_3\mathrm{k}, \lambda_0 = \cos\dfrac{\theta}{2}$ и $|\lambda| = 1$.

Note that λ and $-\lambda$ define the same rotation, since the minus sign is neutralized in the sandwich formula

$$(-\lambda)p(-\lambda)^* = \lambda p\lambda^*.$$

This property allows us to choose the sign before the quaternion so that the scalar part is always positive. If $\lambda_0 > 0$, then we can store only three vector components $\lambda_1, \lambda_2, \lambda_3$, and calculate the scalar part from the unity condition $\lambda_0 = \sqrt{1 - \|\|} = \sqrt{1-\lambda_1^2-\lambda_2^2-\lambda_3^2}$ and choose a positive sign before the root.

Let us represent the matrix R as coefficients r^i_j, where $i = 1, 2, 3$ is the column index, and j is the row index

$$R = \begin{bmatrix} r^1_1 & r^1_2 & r^1_3 \\ r^2_1 & r^2_2 & r^2_3 \\ r^3_1 & r^3_2 & r^3_3 \end{bmatrix}$$

Let us show that the scalar element λ_0 can be calculated through the trace of the matrix.

$$r^1_1 + r^2_2 + r^3_3 = 3 - 2(\lambda^2_2 + \lambda^2_3 + \lambda^2_1 + \lambda^2_3 + \lambda^2_1 + \lambda^2_2) = 3 - 4(\lambda^2_1 + \lambda^2_2 + \lambda^2_3) = 3 - 4\|\lambda\|^2 = 3 - 4(1 - \lambda^2_0) = 4\lambda^2_0 - 1.$$

Next we write $4\lambda^2_0 = 1 + r^1_1 + r^2_2 + r^3_3 = 1 + \operatorname{Tr} R$, which will finally give an expression for λ_0

$$\lambda_0 = \pm\frac{1}{2}\sqrt{1 + \operatorname{Tr} R}.$$

Note that the expression under the root is always positive, since $1 + \operatorname{Tr} R = 4\lambda^2_0 \geqslant 0$.

To calculate $\lambda_1, \lambda_2, \lambda_3$ we write

$$\begin{aligned} r^3_2 - r^2_3 &= 2(\lambda_2\lambda_3 + \lambda_0\lambda_1 - \lambda_2\lambda_3 + \lambda_0\lambda_1) = 4\lambda_0\lambda_1, \\ r^1_3 - r^3_1 &= 2(\lambda_1\lambda_3 + \lambda_0\lambda_2 - \lambda_1\lambda_3 + \lambda_0\lambda_2) = 4\lambda_0\lambda_2, \\ r^2_1 - r^1_2 &= 2(\lambda_0\lambda_3 + \lambda_1\lambda_2 - \lambda_1\lambda_2 + \lambda_0\lambda_3) = 4\lambda_0\lambda_3. \end{aligned}$$

As a result, we obtain a set of formulas

$$\begin{aligned} &4\lambda_0\lambda_1 = r^3_2 - r^2_3, \\ &4\lambda_0\lambda_2 = r^1_3 - r^3_1, \quad \Longleftrightarrow \quad 4\lambda_0\lambda_i = r^{i-1}_{i+1} - r^{i+1}_{i-1}, \quad i = 1, 2, 3. \\ &4\lambda_0\lambda_3 = r^2_1 - r^1_2, \end{aligned} \tag{1}$$

If $|\lambda_0|$ is large enough, say $|\lambda_0| > 1/2$, then the quaternion coefficients can be calculated using the following formulas:

$$\lambda_0 = \pm\sqrt{1 + \operatorname{Tr} R}, \quad \lambda_1 = \frac{r^3_2 - r^2_3}{4\lambda_0}, \quad \lambda_2 = \frac{r^1_3 - r^3_1}{4\lambda_0}, \quad \lambda_3 = \frac{r^2_1 - r^1_2}{4\lambda_0}.$$

If $|\lambda_0|$ is small, say $|\lambda_0| \leqslant 1/2$, then a more sophisticated calculation scheme will have to be used.

We will need a set of three groups of formulas. To obtain the first group, we write

$$\begin{aligned} r^1_1 - r^2_2 - r^3_3 + 1 &= 1 - 2\lambda^2_2 - 2\lambda^2_3 - 1 + 2\lambda^2_1 + 2\lambda^2_3 - 1 + 2\lambda^2_1 + 2\lambda^2_2 + 1 = 4\lambda^2_1, \\ -r^1_1 + r^2_2 - r^3_3 + 1 &= -1 + 2\lambda^2_2 + 2\lambda^2_3 + 1 - 2\lambda^2_1 - 2\lambda^2_3 - 1 + 2\lambda^2_1 + \lambda^2_2 + 1 = 4\lambda^2_2, \\ -r^1_1 - r^2_2 + r^3_3 + 1 &= -1 + 2\lambda^2_2 + 2\lambda^2_3 - 1 + 2\lambda^2_1 + 2\lambda^2_3 + 1 - 2\lambda^2_1 - 2\lambda^2_2 + 1 = 4\lambda^2_3, \end{aligned}$$

obtained the following set of formulas:

$$4\lambda^2_1 = 1 + r^1_1 - r^2_2 - r^3_3, \quad 4\lambda^2_2 = 1 + r^2_2 - r^1_1 - r^3_3, \quad 4\lambda^2_3 = 1 + r^3_3 - r^1_1 - r^2_2 \tag{2}$$

Let us further consider the diagonal elements of the matrix R and using the unity condition of the quaternion λ we write

$$\begin{aligned} r^1_1 &= 1 - 2(\lambda^2_2 + \lambda^2_3) = 1 - 2(\lambda^2_1 + \lambda^2_2 + \lambda^2_3) + 2\lambda^2_1 = 1 - 2(1 - \lambda^2_0) + 2\lambda^2_1 = -1 + 2\lambda^2_0 + 2\lambda^2_1, \\ r^2_2 &= 1 - 2(\lambda^2_1 + \lambda^2_3) = 1 - 2(\lambda^2_1 + \lambda^2_2 + \lambda^2_3) + 2\lambda^2_2 = 1 - 2(1 - \lambda^2_0) + 2\lambda^2_2 = -1 + 2\lambda^2_0 + 2\lambda^2_2, \\ r^3_3 &= 1 - 2(\lambda^2_1 + \lambda^2_2) = 1 - 2(\lambda^2_1 + \lambda^2_2 + \lambda^2_3) + 2\lambda^2_3 = 1 - 2(1 - \lambda^2_0) + 2\lambda^2_3 = -1 + 2\lambda^2_0 + 2\lambda^2_3. \end{aligned}$$

The second group of formulas we need is obtained:

$$2\lambda_1^2 = r_1^1 - 2\lambda_0^2 + 1, \quad 2\lambda_2^2 = r_2^2 - 2\lambda_0^2 + 1, \quad 2\lambda_3^2 = r_3^3 - 2\lambda_0^2 + 1. \tag{3}$$

Formulas (3) allow us to find the largest absolute value component λ_1, λ_2, λ_3 of the vector part of a quaternion using the elements of the matrix and λ_0.

We obtain the last set of formulas by summing the symmetric elements of the matrix R:

$$\begin{aligned} r_1^2 + r_2^1 &= 2\lambda_0\lambda_3 + 2\lambda_1\lambda_2 + 2\lambda_1\lambda_2 - 2\lambda_0\lambda_3 = 4\lambda_1\lambda_2, \\ r_3^1 + r_1^3 &= 2\lambda_1\lambda_3 + 2\lambda_0\lambda_2 + 2\lambda_1\lambda_3 - 2\lambda_0\lambda_2 = 4\lambda_1\lambda_3, \\ r_2^3 + r_3^2 &= 2\lambda_2\lambda_3 - 2\lambda_0\lambda_1 + 2\lambda_2\lambda_3 + 2\lambda_0\lambda_1 = 4\lambda_2\lambda_3, \end{aligned}$$

we obtain the third, final group of necessary formulas:

$$4\lambda_1\lambda_2 = r_1^2 + r_2^1, 4\lambda_1\lambda_3 = r_3^1 + r_1^3, 4\lambda_2\lambda_3 = r_2^3 + r_3^1. \tag{4}$$

Now we have the entire necessary set of formulas at our disposal and we can move on to the presentation of the algorithm itself.

- If $|\lambda_0| \leqslant 1/2$, then dividing by λ_0 when calculating λ_1, λ_2, λ_3 using the set of formulas (1) can lead to an accumulation of errors, so it is more correct to start calculating from the largest component of the vector part.
- To find out which of the components λ_1, λ_2, λ_3 is larger, use the formulas (3). From the formulas it is clear that there is no need to calculate the entire expression; it is sufficient to compare the components of the matrix r_1^1, r_2^2, and r_3^3.
- If $r_1^1 > r_2^2$ and $r_1^1 > r_3^3$, then the component λ_1 is the largest. We calculate it using the formula from (2), λ_2 and λ_3 using the formulas (4), and λ_0 using the formula from (1), i.e.

$$\lambda_1 = \pm\frac{1}{2}\sqrt{1 + r_1^1 - r_2^2 - r_3^3}, \quad \lambda_2 = \frac{r_1^2 + r_2^1}{4\lambda_1}, \quad \lambda_3 = \frac{r_3^1 + r_1^3}{4\lambda_1}, \quad \lambda_0 = \frac{r_2^3 - r_3^2}{4\lambda_1}.$$

- Otherwise, if $r_2^2 > r_3^3$, then λ_2 is the largest, then

$$\lambda_2 = \pm\frac{1}{2}\sqrt{1 + r_2^2 - r_1^1 - r_3^3}, \quad \lambda_1 = \frac{r_1^2 + r_2^1}{4\lambda_2}, \quad \lambda_3 = \frac{r_2^3 + r_3^1}{4\lambda_2}, \quad \lambda_0 = \frac{r_3^1 - r_1^3}{4\lambda_2}.$$

- Otherwise, the only option left is when the largest is λ_3, then

$$\lambda_3 = \pm\frac{1}{2}\sqrt{1 + r_3^3 - r_1^1 - r_2^2}, \quad \lambda_1 = \frac{r_3^1 + r_1^3}{4\lambda_3}, \quad \lambda_2 = \frac{r_2^3 + r_3^1}{4\lambda_3}, \quad \lambda_0 = \frac{r_1^2 - r_2^1}{4\lambda_3}.$$

1.5. Reflections using quaternions

1.5.1. Reflection about a plane passing through the origin

Vector formula Consider a point P with radius vector $\mathbf{p}$ and a plane π passing through the origin O with unit normal vector $\mathbf{n}$, where $\mathbf{n} \perp \pi$. Relative to the vector $\mathbf{n}$, the vector $\mathbf{p}$ splits into two components $\mathbf{p}_{\parallel\mathbf{n}}$ and $\mathbf{p}_{\perp\mathbf{n}}$:

$$\mathbf{p} = \mathbf{p}_{\parallel\mathbf{n}} + \mathbf{p}_{\perp\mathbf{n}}.$$

Reflection with respect to the π plane affects only the vector $\mathbf{p}_{\parallel\mathbf{n}}$, while leaving the vector $\mathbf{p}_{\perp\mathbf{n}}$ unchanged. However, the vector $\mathbf{p}_{\parallel\mathbf{n}}$ changes sign upon reflection. As a result, the reflected vector $\mathbf{p}'$ is calculated as follows:

$$\mathbf{p}' = \mathbf{p}_{\perp\mathbf{n}} - \mathbf{p}_{\parallel\mathbf{n}},$$

where $\mathbf{p}_{\|\mathbf{n}} = (\mathbf{p}, \mathbf{n})\mathbf{n}$ and $\mathbf{p}_{\perp\mathbf{n}} = \mathbf{p} - \mathbf{p}_{\|\mathbf{n}} = \mathbf{p} - (\mathbf{p}, \mathbf{n})\mathbf{n}$, therefore

$$\mathbf{p}' = \underbrace{\mathbf{p} - (\mathbf{p}, \mathbf{n})\mathbf{n}}_{\mathbf{p}_{\perp\mathbf{n}}} - \underbrace{(\mathbf{p}, \mathbf{n})\mathbf{n}}_{\|\mathbf{n}} = \mathbf{p} - 2(\mathbf{p}, \mathbf{n})\mathbf{n}. \tag{5}$$

The same formula can be written in matrix form:

$$\mathbf{p}' = (\mathrm{I} - 2\mathbf{n}\mathbf{n}^T)\mathbf{p}.$$

Quaternion formula If we associate with the radius vector $\mathbf{p}$ a pure quaternion $p = 0 + \mathbf{p}$, then the quaternion formula for reflection can be obtained from the equalityby setting $\mathbf{q} = \mathbf{n}$ and applying it to (5) from right to left:

$$\mathbf{npn} = \|\mathbf{n}\|^2\mathbf{p} - 2(\mathbf{p}, \mathbf{n})\mathbf{n} = \mathbf{p} - 2(\mathbf{p}, \mathbf{n})\mathbf{n} = \mathbf{p}'.$$

Above, we associated a point with a quaternion of the general form $p = p_0 + p_x\mathrm{i} + p_y\mathrm{j} + p_z\mathrm{k}$, so in the quaternion formula for reflecting a point, we would like to see a quaternion p with a nonzero scalar part. However, the formula $\mathbf{n}p\mathbf{n}$ will give an incorrect result:

$$\mathbf{n}p\mathbf{n} = \mathbf{n}(p_0 + \mathbf{p})\mathbf{n} = \mathbf{nn}p_0 + \mathbf{npn} = -p_0 + \mathbf{p}',$$

where the scalar part becomes negative, although it should remain positive.

To eliminate this discrepancy, we change the formula as follows:

$$p' = \mathbf{n}(p - 2p_0)\mathbf{n} = \mathbf{n}(p_0 + \mathbf{p} - 2p_0)\mathbf{n} = \mathbf{n}(\mathbf{p} - p_0)\mathbf{n} = -\mathbf{nn}p_0 + \mathbf{npn} = p_0 + \mathbf{p}'.$$

If we additionally note that $p - 2p_0 = -p_0 + \mathbf{p} = -(p_0 - \mathbf{p}) = -p^*$, then the quaternion reflection formula can be written in its final form as follows:

$$p' = -\mathbf{n}p^*\mathbf{n}$$

1.5.2. Reflection relative to an arbitrary plane

Vector formula Consider an arbitrary plane π, with a unit normal vector $\mathbf{n}$, located at a distance $-d$ from the origin O. Such a plane is given by the equation:

$$(\mathbf{q}, \mathbf{n}) + d = 0,$$

where $\mathbf{q}$ is the radius vector of an arbitrary point Q belonging to the π plane. The radius vector of the projection of the origin onto the plane is calculated as $\mathbf{OO}_\perp = -d\mathbf{n}$.

The reflection of a certain point P with a radius vector $\mathbf{p}$ is carried out in three stages:

1. transferring the origin to a point on the plane by subtracting the vector $\mathbf{OO}_\perp$ from $\mathbf{p}$;
2. reflection using the formula (5);
3. returning the origin by adding the resulting vector to $\mathbf{OO}_\perp$.

Combining all three actions into one formula, we write:

$$\mathbf{p}' = \mathbf{p} + d\mathbf{n} - 2(\mathbf{p} + d\mathbf{n}, \mathbf{n})\mathbf{n} - d\mathbf{n} = \mathbf{p} - 2(\mathbf{p}, \mathbf{n})\mathbf{n} - 2d(\mathbf{n}, \mathbf{n})\mathbf{n} = \mathbf{p} - 2(\mathbf{p}, \mathbf{n})\mathbf{n} - 2d\mathbf{n}$$

and as a result we get:

$$\mathbf{p}' = \mathbf{p} - 2(\mathbf{p}, \mathbf{n})\mathbf{n} - 2d\mathbf{n}. \tag{6}$$

It is worth paying attention to the minus sign in the formula $\mathbf{OO}_\perp = -d\mathbf{n}$, due to which, when subtracting the vector $\mathbf{OO}_\perp$ in the calculations, addition with $d\mathbf{n}$ occurs, and when adding, on the contrary, subtraction occurs.

Table 1

Kotelnikov–Study Transfer Principle

Radius vector $\mathbf{p}$	Screw $\mathbf{L}$
Angle θ	Dual angle Θ
Real number λ	Dual number Λ

Quaternion formula Similarly, it can be written in quaternion form:

$$\mathbf{p}' = \mathbf{p} - 2(\mathbf{p}, \mathbf{n})\mathbf{n} - 2d\mathbf{n}.$$

To prove the formula, we perform some transformations

$$p' = -\mathbf{n}(p + d\mathbf{n})^*\mathbf{n} - d\mathbf{n} = -\mathbf{n}p^*\mathbf{n} - \mathbf{n}d\mathbf{n}^*\mathbf{n} - d\mathbf{n} = -\mathbf{n}p^*\mathbf{n} - d\mathbf{n} - d\mathbf{n} =$$
$$= -\mathbf{n}(1 - \mathbf{p})\mathbf{n} - 2d\mathbf{n} = -\mathbf{n}\mathbf{n} + \mathbf{n}\mathbf{p}\mathbf{n} = 1 + \mathbf{n}\mathbf{p}\mathbf{n} - 2d\mathbf{n}.$$

Since $\mathbf{npn} = \|\mathbf{n}\|^2\mathbf{p} - 2(\mathbf{p}, \mathbf{n})\mathbf{n} = \mathbf{p} - 2(\mathbf{p}, \mathbf{n})\mathbf{n}$, we can finally write

$$p' = 1 + \mathbf{p} - 2(\mathbf{p}, \mathbf{n})\mathbf{n} - 2d\mathbf{n} = 1 + \mathbf{p}',$$

from which it is clear that the vector part completely coincides with the formula (6).

2. Description of screw motion and reflection using dual quaternions

2.1. Kotelnikov–Study's transfer principle

Transfer principle. All formulas of the theory of finite rotations and the kinematics of motion of a rigid body with one fixed point, when replacing real quantities with dual analogs, are transformed into formulas of the theory of finite displacements and the kinematics of motion of a free rigid body [10, p. 67].

The principle was formulated by Alexander Petrovich Kotelnikov and Eduard Study (Eduard Study) [5, pp. 12–13].

In other words, if in the formulas for the rotation of a point in space we replace real numbers, vectors, angles and quaternions with dual numbers, screws, dual angles and dual quaternions, then we obtain the correct formulas for screw motion (table 1).

If the formulas for rotations in space are applied to **affine points** (radius vectors), then the formulas obtained by the principle of transfer should be applied to **screws**, that is, to **lines** in space.

2.2. Application of the Kotelnikov–Study transfer principle

2.2.1. Obtaining a dual quaternion of screw motion

Let's apply the Kotelnikov–Study transfer principle to derive dual quaternion formulas for screw motion. First, we write down the necessary quaternion formulas. Rotational motion around an axis passing through the origin defines a unit quaternion, which is most conveniently written in trigonometric form:

$$\lambda = \cos\frac{\theta}{2} + \sin\frac{\theta}{2}\mathbf{a}, \quad \mathbf{a} = a_x\mathrm{i} + a_y\mathrm{j} + a_z\mathrm{k},$$

where θ is the angle of rotation around the axis with unit direction vector $\mathbf{a} = (a_x, a_y, a_z)^T$, $\|\mathbf{a}\| = 1$. The rotation of an affine point P, represented in homogeneous coordinates by the quaternion $p = 1 + x\mathrm{i} + y\mathrm{j} + z\mathrm{k}$, is given by the sandwich formula:

$$p' = \lambda p \lambda^*,$$

where the new position of the point is expressed by the quaternion $p' = 1 + x'\mathrm{i} + y'\mathrm{j} + z'\mathrm{k}$.

Note also that the quaternion p need not have a unit scalar part and can define a projective point of any form, since the sandwich formula leaves the coordinate w unchanged due to the unity of $\lambda\lambda^* = 1$:

$$\lambda(w + \mathbf{p})\lambda^* = \lambda w \lambda^* + \lambda \mathbf{p} \lambda^* = w\lambda\lambda^* + \lambda \mathbf{p} \lambda^* = w + \lambda \mathbf{p} \lambda^*.$$

The scalar part w can also be equal to zero, in which case the formula defines the rotation of a point at infinity in projective space or an equivalent free vector in Cartesian space.

According to the translation principle, dual quaternion defining screw motion (translation + rotation) is obtained from a rotational quaternion by the following substitution:

- $\theta \longrightarrow \Theta = \theta + \theta^o\varepsilon$ — the angle is replaced by its dual angle;
- $\mathbf{a} \longrightarrow \mathbf{A} = \mathbf{a} + \mathbf{a}^o\varepsilon$ — the vector is replaced by a pure dual quaternion (a screw);
- $* \longrightarrow \dagger$ — the quaternion conjugate $*$ is replaced by the dual quaternion conjugate $\dagger$.

As a result of such a replacement, the dual quaternion of screw motion will be written as follows:

$$\Lambda = \cos\frac{\Theta}{2} + \sin\frac{\Theta}{2}\mathbf{A}, \quad \Theta = \theta + \theta^o\varepsilon, \quad \mathbf{A} = \mathbf{a} + \mathbf{a}^o\varepsilon. \tag{7}$$

A pure dual quaternion $\mathbf{A}$ defines an arbitrary axis of rotation with a direction unit vector $\mathbf{a}$ and a moment $\mathbf{a}^o$. The set of vectors $\{\mathbf{a} \mid \mathbf{a}^o\}$ are the Plücker coordinates, for which th Plücker condition $(\mathbf{a}, \mathbf{a}^o) = 0$ must be satisfied. The Plücker condition and the condition $\|\mathbf{a}\| = 1$ guarantee the unity of the pure dual quaternion $\mathbf{A}$, since $|\mathbf{A}| = \mathbf{A}\mathbf{A}^* = (\mathbf{a} + \mathbf{a}^o\varepsilon)(-\mathbf{a} - \mathbf{a}^o\varepsilon) = \|\mathbf{a}^2 + 2(\mathbf{a}, \mathbf{a}^o)\varepsilon\| = \|\mathbf{a}\|^2 = 1$.

The dual angle Θ specifies both the magnitude of the rotation angle θ around the axis $\mathbf{A}$ and the translation distance θ^o along the same axis $\mathbf{A}$. Trigonometric functions of the dual angle are calculated using the formulas

$$\sin\Theta = \sin(\theta + \varepsilon\theta^o) = \sin\theta + \theta^o\cos\theta\varepsilon, \quad \cos\Theta = \cos(\theta + \theta^o\varepsilon) = \cos\theta - \theta^o\sin\theta\varepsilon.$$

Using these formulas, we replace the dual number Θ in the formula (7) with the real numbers θ and θ^o and transform the dual quaternion Λ as follows:

$$\begin{aligned}\Lambda = \cos\frac{\Theta}{2} + \sin\frac{\Theta}{2}\mathbf{A} &= \cos\frac{\theta}{2} - \frac{\theta^o}{2}\sin\frac{\theta}{2}\varepsilon + \left(\sin\frac{\theta}{2} + \frac{\theta^2}{2}\cos\frac{\theta}{2}\varepsilon\right)(\mathbf{a} + \mathbf{a}^o\varepsilon) = \\ &= \cos\frac{\theta}{2} - \frac{\theta^o}{2}\sin\frac{\theta}{2}\varepsilon + \sin\frac{\theta}{2}\mathbf{a} + \frac{\theta^o}{2}\cos\frac{\theta}{2}\varepsilon\mathbf{a} + \sin\frac{\theta}{2}\mathbf{a}^o\varepsilon + \frac{\theta^o}{2}\cos\frac{\theta}{2}\mathbf{a}^o\varepsilon^2 = \\ &= \cos\frac{\theta}{2} - \frac{\theta^o}{2}\sin\frac{\theta}{2}\varepsilon + \sin\frac{\theta}{2}\mathbf{a} + \frac{\theta^o}{2}\cos\frac{\theta}{2}\mathbf{a}\varepsilon + \sin\frac{\theta}{2}\mathbf{a}^o\varepsilon.\end{aligned}$$

Having grouped separately the terms with θ^o and without θ^o, we write the formula in the following form:

$$\Lambda = \cos\frac{\theta}{2} + \sin\frac{\theta}{2}(\mathbf{a} + \mathbf{a}^o\varepsilon) + \left(\cos\frac{\theta}{2}\mathbf{a} - \sin\frac{\theta}{2}\right)\frac{\theta^o}{2}\varepsilon. \tag{8}$$

This notation has the advantage of allowing us to distinguish between pure rotation and pure translation:

- for $\theta^o = 0$, we obtain pure rotation around an arbitrary axis $\mathbf{A}$ defined by the dual quaternion $R = \cos\frac{\theta}{2} + \sin\frac{\theta}{2}(\mathbf{a} + \mathbf{a}^o\varepsilon)$;
- for $\theta = 0$, the trigonometric functions take the values $\sin 0 = 0$ and $\cos 0 = 1$, and we obtain pure translation along the axis $\mathbf{A}$ defined by the dual quaternion $T = 1 + \frac{\theta^o}{2}\mathbf{a}\varepsilon$.

2.2.2. Conjugate dual quaternion of screw motion

Now let's find the conjugate dual quaternion Λ. Recall that dual quaternion conjugation is introduced in three different ways:

- $Q^* = (q + q^o\varepsilon)^* = q^* + q^{o*}\varepsilon$ is the quaternion conjugation, which can also be called complex conjugation.
- $\overline{Q} = \overline{q + q^o\varepsilon} = q - q^o\varepsilon$ is the dual conjugation.
- $Q^\dagger = (\overline{q + q^o\varepsilon})^* = q^* - q^{o*}\varepsilon$ is the dual quaternion conjugation.

We need to calculate the dual conjugation of the dual quaternion Λ:

$$\Lambda^\dagger = \overline{\cos\Theta/2} + \overline{\sin\Theta/2}\mathbf{A}^\dagger, \quad \mathbf{A}^\dagger = \mathbf{a}^* - \mathbf{a}^{o*}\varepsilon = -\mathbf{a} + \mathbf{a}^o\varepsilon = -(\mathbf{a} - \mathbf{a}^o\varepsilon).$$

Next, we can use formulas for trigonometric functions of dual numbers:

$$\overline{\cos\Theta/2} = \cos\frac{\theta}{2} + \frac{\theta^o}{2}\sin\frac{\theta}{2}\varepsilon, \quad \overline{\sin\Theta/2} = \sin\frac{\theta}{2} - \frac{\theta^o}{2}\cos\frac{\theta}{2}\varepsilon.$$

We can, however, obtain this formula differently by writing Λ in quaternion form:

$$\Lambda = \cos\frac{\theta}{2} + \sin\frac{\theta}{2}\mathbf{a} + \left(-\frac{\theta^o}{2}\sin\frac{\theta}{2} + \frac{\theta^o}{2}\cos\frac{\theta}{2}\mathbf{a} + \sin\frac{\theta}{2}\mathbf{a}^o\right)\varepsilon = \lambda + \lambda^o\varepsilon,$$

where $\lambda = \cos\frac{\theta}{2} + \sin\frac{\theta}{2}\mathbf{a}$ and $\lambda^o = -\frac{\theta^o}{2}\sin\frac{\theta}{2} + \frac{\theta^o}{2}\cos\frac{\theta}{2}\mathbf{a} + \sin\frac{\theta}{2}\mathbf{a}^o$ are quaternions. Then we calculate:

$$\begin{aligned}\Lambda^\dagger = \lambda^* - \lambda^{o*}\varepsilon &= \left(\cos\frac{\theta}{2} - \sin\frac{\theta}{2}\mathbf{a}\right) - \left(-\frac{\theta^o}{2}\sin\frac{\theta}{2} - \frac{\theta^o}{2}\cos\frac{\theta}{2}\mathbf{a} - \sin\frac{\theta}{2}\mathbf{a}^o\right)\varepsilon = \\ &= \left(\cos\frac{\theta}{2} - \sin\frac{\theta}{2}\mathbf{a}\right) + \left(\frac{\theta^o}{2}\sin\frac{\theta}{2} + \frac{\theta^o}{2}\cos\frac{\theta}{2}\mathbf{a} + \sin\frac{\theta}{2}\mathbf{a}^o\right)\varepsilon = \\ &= \left(\cos\frac{\theta}{2} - \sin\frac{\theta}{2}(\mathbf{a} - \mathbf{a}^o\varepsilon)\right) + \left(\sin\frac{\theta}{2} + \cos\frac{\theta}{2}\mathbf{a}\right)\frac{\theta^o}{2}\varepsilon.\end{aligned}$$

Accordingly, the final formula for the conjugate dual quaternion will look like this:

$$\Lambda^\dagger = \cos\frac{\theta}{2} - \sin\frac{\theta}{2}(\mathbf{a} - \mathbf{a}^o\varepsilon) + \left(\sin\frac{\theta}{2} + \cos\frac{\theta}{2}\mathbf{a}\right)\frac{\theta^o}{2}\varepsilon.$$

2.2.3. Proof of the unity of the dual quaternion of screw motion

Let us also recall the definition of a unit dual quaternion: a dual quaternion Q is called unit dual quaternion if its modulus is equal to 1, that is, $|Q|^2 = QQ^* = |q|^2 + 2(q, q^o)\varepsilon = 1$. Let us check that Λ is unit dual quaternion, for which it is necessary to calculate the quaternion conjugate. Let us do this and at the same time note that the expression for Λ^* will differ from $\Lambda^\dagger$.

$$\Lambda^* = \lambda^* + \lambda^{o*}\varepsilon = \cos\frac{\theta}{2} - \sin\frac{\theta}{2}\mathbf{a} + \left(-\frac{\theta^o}{2}\sin\frac{\theta}{2} - \frac{\theta^o}{2}\cos\frac{\theta}{2}\mathbf{a} - \sin\frac{\theta}{2}\mathbf{a}^o\right)\varepsilon =$$

$$= \cos\frac{\theta}{2} - \frac{\theta^o}{2}\sin\frac{\theta}{2}\varepsilon - \sin\frac{\theta}{2}\mathbf{a} - \frac{\theta^o}{2}\cos\frac{\theta}{2}\mathbf{a}\varepsilon - \sin\frac{\theta}{2}\mathbf{a}^o\varepsilon - \frac{\theta^o}{2}\cos\frac{\theta}{2}\mathbf{a}^o\varepsilon^2 =$$
$$= \left(\cos\frac{\theta}{2} - \frac{\theta^o}{2}\sin\frac{\theta}{2}\varepsilon\right) - \left[\left(\sin\frac{\theta}{2} + \frac{\theta^o}{2}\cos\frac{\theta}{2}\varepsilon\right)\mathbf{a} + \left(\sin\frac{\theta}{2} + \frac{\theta^o}{2}\cos\frac{\theta}{2}\varepsilon\right)\mathbf{a}^o\varepsilon\right] =$$
$$= \left(\cos\frac{\theta}{2} - \frac{\theta^o}{2}\sin\frac{\theta}{2}\varepsilon\right) - \left(\sin\frac{\theta}{2} + \frac{\theta^o}{2}\cos\frac{\theta}{2}\varepsilon\right)(\mathbf{a} + \mathbf{a}^o\varepsilon) = \cos\frac{\Theta}{2} - \sin\frac{\Theta}{2}\mathbf{A},$$

where $\mathbf{A} = \mathbf{a} + \mathbf{a}^o\varepsilon$. Finally:

$$\Lambda^* = \left(\cos\frac{\Theta}{2} + \sin\frac{\Theta}{2}\mathbf{A}\right)^* = \cos\frac{\Theta}{2} - \sin\frac{\Theta}{2}\mathbf{A},$$

since $\mathbf{A}^* = -\mathbf{a} - \mathbf{a}^o\varepsilon = -\mathbf{A}$.

Now we can find the modulus of the dual quaternion $|\Lambda|^2 = \Lambda\Lambda^*$:

$$\Lambda\Lambda^* = \left(\cos\frac{\Theta}{2} + \sin\frac{\Theta}{2}\mathbf{A}\right)\left(\cos\frac{\Theta}{2} - \sin\frac{\Theta}{2}\mathbf{A}\right) =$$
$$= \cos^2\frac{\Theta}{2} - \cos\frac{\Theta}{2}\sin\frac{\Theta}{2}\mathbf{A} + \sin\frac{\Theta}{2}\mathbf{A}\cos\frac{\Theta}{2} - \sin\frac{\Theta}{2}\mathbf{A}\sin\frac{\Theta}{2}\mathbf{A}.$$

Recall that the multiplication of dual numbers is commutative $(a + b\varepsilon)(c + d\varepsilon) = (c + d\varepsilon)(a + b\varepsilon)$, and the multiplication of a dual number by a pure dual quaternion is also commutative: $(a + b\varepsilon)(\mathbf{a} + \mathbf{b}\varepsilon) = (\mathbf{a} + \mathbf{b}\varepsilon)(a + b\varepsilon)$. This allows us to write the expression $\Lambda\Lambda^*$ in the following form:

$$\Lambda\Lambda^* = \cos^2\frac{\Theta}{2} - \cos\frac{\Theta}{2}\sin\frac{\Theta}{2}\mathbf{A} + \sin\frac{\Theta}{2}\cos\frac{\Theta}{2}\mathbf{A} - \sin^2\frac{\Theta}{2}\mathbf{A}\mathbf{A} = \cos^2\frac{\Theta}{2} - \sin^2\frac{\Theta}{2}\mathbf{A}\mathbf{A}.$$

Let us calculate the product of pure dual quaternions $\mathbf{AA}$:

$$\mathbf{AA} = (\mathbf{a} + \mathbf{a}^o\varepsilon)(\mathbf{a} + \mathbf{a}^o\varepsilon) = \mathbf{aa} + \mathbf{a}\mathbf{a}^o\varepsilon + \mathbf{a}^o\mathbf{a}\varepsilon + \mathbf{a}^o\mathbf{a}^o\varepsilon^2 = \mathbf{aa} + (\mathbf{a}\mathbf{a}^o + \mathbf{a}^o\mathbf{a})\varepsilon.$$

For further simplification, we use the rule of multiplication of pure quaternions $\mathbf{pq} = -(\mathbf{p}, \mathbf{q}) + \mathbf{p} \times \mathbf{q}$, then $\mathbf{aa} = -(\mathbf{a}, \mathbf{a}) = -\|\mathbf{a}\|^2 = -1$ since $\|\mathbf{a}\| = 1$ by condition. In view of the fact that $\mathbf{a}\mathbf{a}^o = -(\mathbf{a}, \mathbf{a}^o) + \mathbf{a} \times \mathbf{a}^o$ and $\mathbf{a}^o\mathbf{a} = -(\mathbf{a}^o, \mathbf{a}) + \mathbf{a}^o \times \mathbf{a}$ the dual part of $\mathbf{AA}$ is simplified:

$$\mathbf{a}\mathbf{a}^o + \mathbf{a}^o\mathbf{a} = -2(\mathbf{a}, \mathbf{a}^o) + \mathbf{a} \times \mathbf{a}^o - \mathbf{a} \times \mathbf{a}^o = -2(\mathbf{a}, \mathbf{a}^o).$$

In addition, by the Plücker condition $(\mathbf{a}, \mathbf{a}^o) = 0$, because $\mathbf{a} \perp \mathbf{a}^o$, therefore $\mathbf{AA} = -1$ subject to $||\mathbf{a}|| = 1$. We obtain that $\Lambda\Lambda^* = \cos^2\frac{\Theta}{2} + \sin^2\frac{\Theta}{2}$. We use the formula $(a + b\varepsilon)^2 = a^2 + 2ab\varepsilon$ to calculate $\cos^2\Theta/2$ and $\sin^2\Theta/2$, for them the following will be true:

$$\cos^2\frac{\Theta}{2} = \left(\cos\frac{\theta}{2} - \frac{\theta^o}{2}\sin\frac{\theta}{2}\varepsilon\right)^2 = \cos^2\frac{\theta}{2} - 2\cos\frac{\theta}{2}\sin\frac{\theta}{2}\frac{\theta^o}{2}\varepsilon,$$
$$\sin^2\frac{\Theta}{2} = \left(\sin\frac{\theta}{2} + \frac{\theta^o}{2}\cos\frac{\theta}{2}\varepsilon\right)^2 = \sin^2\frac{\theta}{2} + 2\sin\frac{\theta}{2}\cos\frac{\theta}{2}\frac{\theta^o}{2}\varepsilon.$$

Hence:

$$\cos^2\frac{\Theta}{2} + \sin^2\frac{\Theta}{2} = \cos^2\frac{\theta}{2} + \sin^2\frac{\theta}{2} - 2\cos\frac{\theta}{2}\sin\frac{\theta}{2}\frac{\theta^o}{2}\varepsilon + 2\sin\frac{\theta}{2}\cos\frac{\theta}{2}\frac{\theta^o}{2}\varepsilon =$$
$$= \cos^2\frac{\theta}{2} + \sin^2\frac{\theta}{2} = 1 \Rightarrow \cos^2\frac{\Theta}{2} + \sin^2\frac{\Theta}{2} = 1,$$

where Θ is the dual angle.

As a result, we have proved the unity of the dual quaternion $\Lambda = \cos\frac{\Theta}{2} + \sin\frac{\Theta}{2}\mathbf{A}$. It should be especially noted that an important condition is the unity of the pure quaternion $\mathbf{a}$, i.e. $\|\mathbf{a}\| = 1$ and the fulfillment of the Plücker condition $(\mathbf{a}, \mathbf{a}^o) = 0$. Without these two conditions, Λ will not be unity.

2.3. Screw motion of a point and a vector

2.3.1. Rotation around an arbitrary axis without translation

Let us consider the dual quaternion representation of an affine point P:

$$P = 1 + \mathbf{p}^o\varepsilon = p + p^o,$$

where $p^o = \mathbf{p}^o$ is a pure quaternion (radius vector), and $p = 1$ is a scalar quaternion (point O). Let's now construct a sandwich operator from the dual quaternion R:

$$R = \cos\frac{\theta}{2} + \sin\frac{\theta}{2}(\mathbf{a} + \mathbf{a}^o\varepsilon).$$

where $\mathbf{A} = \mathbf{a} + \mathbf{a}^o\varepsilon$ is a pure dual quaternion defining the axis of rotation, $\mathbf{a}$ is the direction vector of the axis of rotation, $\mathbf{a}^o$ is the moment of the axis of rotation, and θ is the actual angle of rotation about the axis.

Once again, we require two important conditions to be satisfied:

1. vector $\mathbf{a}$ must be the unit vector $\|\mathbf{a}\| = 1$;
2. vectors $\mathbf{a}$ and $\mathbf{a}^o$ satisfy the Plücker condition $(\mathbf{a}, \mathbf{a}^o) = 0$.

Let's find the conjugate dual quaternion $R^\dagger = (\bar{R})^* = \overline{(R^*)}$:

$$R^* = \cos\frac{\theta}{2} - \sin\frac{\theta}{2}\mathbf{a} - \sin\frac{\theta}{2}\mathbf{a}^o\varepsilon \quad R^\dagger = \overline{(R^*)} = \cos\frac{\theta}{2} - \sin\frac{\theta}{2}\mathbf{a} + \sin\frac{\theta}{2}\mathbf{a}^o\varepsilon = \cos\frac{\theta}{2} - \sin\frac{\theta}{2}(\mathbf{a} - \mathbf{a}^o\varepsilon).$$

The final formula for the sandwich operator will look like this:

$$P' = RPR^\dagger = R(1 + \mathbf{p}^o\varepsilon)R^\dagger = RR^\dagger + R\mathbf{p}^o\varepsilon R^\dagger.$$

Let's first find the dual quaternion product $RR^\dagger$:

$$\begin{aligned}
RR^\dagger &= \left(\cos\frac{\theta}{2} + \sin\frac{\theta}{2}(\mathbf{a} + \mathbf{a}^o\varepsilon)\right)\left(\cos\frac{\theta}{2} - \sin\frac{\theta}{2}(\mathbf{a} - \mathbf{a}^o\varepsilon)\right) = \\
&= \cos^2\frac{\theta}{2} - \sin\frac{\theta}{2}\cos\frac{\theta}{2}(\mathbf{a} - \mathbf{a}^o\varepsilon) + \sin\frac{\theta}{2}\cos\frac{\theta}{2}(\mathbf{a} + \mathbf{a}^o\varepsilon) - \sin^2\frac{\theta}{2}(\mathbf{a} + \mathbf{a}^o\varepsilon)(\mathbf{a} - \mathbf{a}^o\varepsilon) = \\
&= \cos^2\frac{\theta}{2} + \sin\frac{\theta}{2}\cos\frac{\theta}{2}(\mathbf{a} + \mathbf{a}^o\varepsilon - \mathbf{a} + \mathbf{a}^o\varepsilon) + \sin^2\frac{\theta}{2}(1 + 2\mathbf{a} \times \mathbf{a}^o\varepsilon) = \\
&= \cos^2\frac{\theta}{2} + \sin^2\frac{\theta}{2} + 2\sin\frac{\theta}{2}\cos\frac{\theta}{2}\mathbf{a}^o\varepsilon + 2\sin^2\frac{\theta}{2}\mathbf{a} \times \mathbf{a}^o\varepsilon = \\
&= 1 + \sin\theta\mathbf{a}^o\varepsilon + (1 - \cos\theta)\mathbf{a} \times \mathbf{a}^o\varepsilon = 1 + (\sin\theta\mathbf{a}^o + (1 - \cos\theta)\mathbf{a} \times \mathbf{a}^o)\varepsilon.
\end{aligned}$$

Note that to simplify the product of vectors $(\mathbf{a} + \mathbf{a}^o\varepsilon)(\mathbf{a} - \mathbf{a}^o\varepsilon)$, the following calculations were performed:

$$(\mathbf{a} + \mathbf{a}^o\varepsilon)(\mathbf{a} - \mathbf{a}^o\varepsilon) = \mathbf{a}\mathbf{a} - \mathbf{a}\mathbf{a}^o\varepsilon + \mathbf{a}^o\mathbf{a}\varepsilon - \mathbf{a}^o\mathbf{a}^o\varepsilon\varepsilon = \mathbf{a}\mathbf{a} + (\mathbf{a}^o\mathbf{a} - \mathbf{a}\mathbf{a}^o)\varepsilon,$$

where $\mathbf{a}\mathbf{a} = -(\mathbf{a}, \mathbf{a}) + \mathbf{a} \times \mathbf{a} = -\|\mathbf{a}\|^2 = -1$, therefore $\mathbf{a}\mathbf{a} = -1$. The subtraction was simplified as follows:

$$\mathbf{a}^o\mathbf{a} - \mathbf{a}\mathbf{a}^o = -(\mathbf{a}^o, \mathbf{a}) + \mathbf{a}^o \times \mathbf{a} + (\mathbf{a}, \mathbf{a}^o) - \mathbf{a} \times \mathbf{a}^o = -2\mathbf{a} \times \mathbf{a}^o.$$

The result of simplifications is an expression of the form:

$$(\mathbf{a} + \mathbf{a}^o\varepsilon)(\mathbf{a} - \mathbf{a}^o\varepsilon) = (-1 - 2\mathbf{a} \times \mathbf{a}^o\varepsilon).$$

As a result, we obtain the formula for the dual quaternion product $RR^\dagger$:

$$RR^\dagger = 1 + (\sin\theta\mathbf{a}^o + (1-\cos\theta)\mathbf{a}\times\mathbf{a}^o)\varepsilon, \tag{9}$$

where $\mathbf{a}\times\mathbf{a}^o$ has the geometric meaning of the projection of the origin onto the **A** axis or, in other words, the point of the axis closest to the origin.

Now we calculate the second term $R\mathbf{p}^o\varepsilon R^\dagger$:

$$\begin{aligned}
R\mathbf{p}^o\varepsilon R^\dagger &= \left(\cos\frac{\theta}{2} + \sin\frac{\theta}{2}(\mathbf{a}+\mathbf{a}^o\varepsilon)\right)\mathbf{p}^o\varepsilon\left(\cos\frac{\theta}{2} - \sin\frac{\theta}{2}(\mathbf{a}-\mathbf{a}^o\varepsilon)\right) = \\
&= \left(\cos\frac{\theta}{2} + \sin\frac{\theta}{2}(\mathbf{a}+\mathbf{a}^o\varepsilon)\right)\left(\cos\frac{\theta}{2}\mathbf{p}^o\varepsilon - \sin\frac{\theta}{2}\mathbf{p}^o\mathbf{a}\varepsilon + \sin\frac{\theta}{2}\mathbf{p}^o\mathbf{a}^o\varepsilon^2\right) = \\
&= \cos^2\frac{\theta}{2}\mathbf{p}^o\varepsilon - \sin\frac{\theta}{2}\cos\frac{\theta}{2}\mathbf{p}^o\mathbf{a}\varepsilon + \sin\frac{\theta}{2}\cos\frac{\theta}{2}(\mathbf{a}+\mathbf{a}^o\varepsilon)\mathbf{p}^o\varepsilon - \sin^2\frac{\theta}{2}(\mathbf{a}+\mathbf{a}^o\varepsilon)\mathbf{p}^o\mathbf{a}\varepsilon = \\
&= \cos^2\frac{\theta}{2}\mathbf{p}^o\varepsilon - \sin\frac{\theta}{2}\cos\frac{\theta}{2}\mathbf{p}^o\mathbf{a}\varepsilon + \sin\frac{\theta}{2}\cos\frac{\theta}{2}\mathbf{a}\mathbf{p}^o\varepsilon - \sin^2\frac{\theta}{2}\mathbf{a}\mathbf{p}^o\mathbf{a}\varepsilon = \\
&= \cos^2\frac{\theta}{2}\mathbf{p}^o\varepsilon + \sin\frac{\theta}{2}\cos\frac{\theta}{2}(\mathbf{a}\mathbf{p}^o - \mathbf{p}^o\mathbf{a})\varepsilon - \sin^2\frac{\theta}{2}\mathbf{a}\mathbf{p}^o\mathbf{a}\varepsilon = \\
&= \cos^2\frac{\theta}{2}\mathbf{p}^o\varepsilon + 2\sin\frac{\theta}{2}\cos\frac{\theta}{2}\mathbf{a}\times\mathbf{p}^o\varepsilon - \sin^2\frac{\theta}{2}(\mathbf{p}^o - 2(\mathbf{a},\mathbf{p}^o)\mathbf{a})\varepsilon = \\
&= \left(\cos^2\frac{\theta}{2} - \sin^2\frac{\theta}{2}\right)\mathbf{p}^o\varepsilon + \sin\theta\mathbf{a}\times\mathbf{p}^o\varepsilon + \sin^2\frac{\theta}{2}\cdot 2(\mathbf{a},\mathbf{p}^o)\mathbf{a}\varepsilon = \\
&= \cos\theta\mathbf{p}^o\varepsilon + \sin\theta\mathbf{a}\times\mathbf{p}^o\varepsilon + (1-\cos\theta)(\mathbf{a},\mathbf{p}^o)\mathbf{a}\varepsilon.
\end{aligned}$$

The following simplifications were used in calculating the main part:

$$(\mathbf{a}+\mathbf{a}^o\varepsilon)\mathbf{p}^o\varepsilon = \mathbf{a}\mathbf{p}^o\varepsilon + \mathbf{a}^o\mathbf{p}^o\varepsilon^2 = \mathbf{a}\mathbf{p}^o\varepsilon,$$

$$(\mathbf{a}+\mathbf{a}^o\varepsilon)\mathbf{p}^o\mathbf{a}\varepsilon = \mathbf{a}\mathbf{p}^o\mathbf{a}\varepsilon + \mathbf{a}^o\mathbf{p}^o\mathbf{a}\varepsilon^2 = \mathbf{a}\mathbf{p}^o\mathbf{a}\varepsilon,$$

$$\mathbf{a}\mathbf{p}^o - \mathbf{p}^o\mathbf{a} = -(\mathbf{a},\mathbf{p}^o) + \mathbf{a}\times\mathbf{p}^o + (\mathbf{p}^o,\mathbf{a}) - \mathbf{p}^o\times\mathbf{a} = 2\mathbf{a}\times\mathbf{p}^o,$$

$$\begin{aligned}
\mathbf{a}\mathbf{p}^o\mathbf{a} &= \mathbf{a}(-(\mathbf{p}^o,\mathbf{a}) + \mathbf{p}^o\times\mathbf{a}) = -(\mathbf{p}^o,\mathbf{a})\mathbf{a} + \mathbf{a}(\mathbf{p}^o\times\mathbf{a}) = \\
&= -(\mathbf{p}^o,\mathbf{a})\mathbf{a} - (\mathbf{a},\mathbf{p}^o\times\mathbf{a}) + \mathbf{a}\times\mathbf{p}^o\times\mathbf{a} = -(\mathbf{p}^o,\mathbf{a})\mathbf{a} + \mathbf{a}\times\mathbf{p}^o\times\mathbf{a} = \\
&= -(\mathbf{p}^o,\mathbf{a})\mathbf{a} + \mathbf{p}^o||\mathbf{a}||^2 - (\mathbf{a},\mathbf{p}^o)\mathbf{a} = \mathbf{p}^o - 2(\mathbf{a},\mathbf{p}^o)\mathbf{a}.
\end{aligned}$$

In the end we got:

$$R\mathbf{p}^o\varepsilon R^\dagger = (\cos\theta\mathbf{p}^o + \sin\theta\mathbf{a}\times\mathbf{p}^o + (1-\cos\theta)(\mathbf{a},\mathbf{p}^o)\mathbf{a})\,\varepsilon.$$

Note that in this expression the moment $\mathbf{a}^o$ is missing and the expression in brackets at the imaginary unit ε exactly repeats Rodrigues' formula for the rotation of the radius vector around the axis with the direction vector **a** passing through the origin.

Let's now write down the complete formula:

$$RPR^\dagger = RR^\dagger + R\mathbf{p}^o\varepsilon R^\dagger = 1 + [\cos\theta\mathbf{p}^o + \sin\theta\mathbf{a}\times\mathbf{p}^o + (1-\cos\theta)(\mathbf{a},\mathbf{p}^o)\mathbf{a} + \sin\theta\mathbf{a}^o + (1-\cos\theta)\mathbf{a}\times\mathbf{a}^o]\varepsilon, \tag{10}$$

where $\mathbf{a}\times\mathbf{a}^o$ corresponds to a point on the axis of rotation. The part responsible for rotation around an arbitrary axis in this case is hidden in the term $RR^\dagger$.

It is also worth noting the importance of the scalar part in the dual quaternion representation of the point $P = 1 + \mathbf{p}^o\varepsilon$. Without this part, there would be no $RR^\dagger$ term in the final formula.

A direction vector (free vector) can be represented by a pure dual quaternion:

$$\mathbf{V} = 0 + \mathbf{v}^o\varepsilon$$

and the «translational» part of $RR^\dagger$ does not act on such a dual quaternion:

$$R\mathbf{V}R^\dagger = R0R^\dagger + R\mathbf{v}^o\varepsilon R^\dagger = R\mathbf{v}^o\varepsilon R^\dagger = (\cos\theta\mathbf{v}^o + \sin\theta\mathbf{a}\times\mathbf{v}^o + (1-\cos\theta)(\mathbf{a},\mathbf{v}^o)\mathbf{a})\,\varepsilon. \tag{11}$$

Similarly, a point in projective space with homogeneous coordinates $(x, y, z : w) = (\mathbf{p}^o \mid w)$ is represented by the dual quaternion $P_w = w + \mathbf{p}^o\varepsilon$ and the same formula can be used:

$$\begin{aligned} RP_wR^\dagger &= RwR^\dagger + R\mathbf{p}^o\varepsilon R^\dagger = wRR^\dagger + R\mathbf{p}^o\varepsilon R^\dagger = \\ &= w + \left[\cos\theta\mathbf{p}^o + \sin\theta\mathbf{a}\times\mathbf{p}^o + (1-\cos\theta)(\mathbf{a},\mathbf{p}^o)\mathbf{a} + w\sin\theta\mathbf{a}^o + w(1-\cos\theta)\mathbf{a}\times\mathbf{a}^o\right]\varepsilon. \end{aligned}$$

2.3.2. Rotation around an arbitrary axis using Rodrigues' formula

Let us show that the formulas (10) and (11) can be obtained from the usual Rodrigues vector formula. We will consider the rotation of a point using the usual vector notation. We will associate a point P with a vector $\mathbf{p}$, where the superscript o is removed since we have moved to the vector formalism.

We perform the rotation around an axis passing through point P_0 in the direction of the radius vector $\mathbf{a}$. We associate point P_0 with the radius vector $\mathbf{p}_0$. We perform the rotation using Rodrigues' formula in three steps:

1. Subtract the vector $\mathbf{p}_0$ from $\mathbf{p}$, thereby moving the origin to point P_0 or, alternatively, moving the axis to the origin of the new coordinate system.
2. We use Rodrigues' formula to perform the rotation by applying it to the vector $\mathbf{p} - \mathbf{p_0}$.
3. Add the vector p_0 to the result of the rotation, returning the coordinate system to its original position.

If $R_{\theta,\mathbf{a}}(\mathbf{p}) = \cos\theta\mathbf{p} + \sin\theta\mathbf{a}\times\mathbf{p} + (1-\cos\theta)(\mathbf{a},\mathbf{p})\mathbf{a}$, then the rotation around the axis passing through the point P_0 with the direction vector $\mathbf{a}$ will be given by the formula:

$$\mathbf{p}' = R_{\theta,\mathbf{a}}(\mathbf{p} - \mathbf{p}_0) + \mathbf{p}_0.$$

Let's expand on this formula:

$$\begin{aligned} R_{\theta,\mathbf{a}}(\mathbf{p}-\mathbf{p_0}) + \mathbf{p}_0 &= \cos\theta(\mathbf{p}-\mathbf{p}_0) + \sin\theta\mathbf{a}\times(\mathbf{p}-\mathbf{p}_0) + (1-\cos\theta)(\mathbf{a},\mathbf{p}-\mathbf{p}_0)\mathbf{a} + \mathbf{p}_0 = \\ &= \cos\theta\mathbf{p} - \cos\theta\mathbf{p}_0 + \sin\theta\mathbf{a}\times\mathbf{p} - \sin\theta\mathbf{a}\times\mathbf{p}_0 + (1-\cos\theta)(\mathbf{a},\mathbf{p})\mathbf{a} - (1-\cos\theta)(\mathbf{a},\mathbf{p}_0)\mathbf{a} + \mathbf{p}_0 = \\ &= \cos\theta\mathbf{p} + \sin\theta\mathbf{a}\times\mathbf{p} + (1-\cos\theta)(\mathbf{a},\mathbf{p})\mathbf{a} - \cos\theta\mathbf{p}_0 - \sin\theta\mathbf{a}\times\mathbf{p}_0 - (1-\cos\theta)(\mathbf{a},\mathbf{p}_0)\mathbf{a} + \mathbf{p}_0. \end{aligned}$$

The choice of the point P_0 on the straight axis of rotation is generally arbitrary, however, if we define the axis by Plücker coordinates using the screw $\mathbf{A} = \mathbf{a} + \mathbf{a}^o\varepsilon$, then we can choose $\mathbf{p}_0 = \mathbf{a}\times\mathbf{a}^o$ — the projection of the point O onto the straight line. With this choice of point, the expressions in the tail will be simplified:

$$\mathbf{a}\times\mathbf{p}_0 = \mathbf{a}\times\mathbf{a}\times\mathbf{a}^o = \mathbf{a}(\mathbf{a},\mathbf{a}^o) - \mathbf{a}^o(\mathbf{a},\mathbf{a}) = -\mathbf{a}^o$$

due to the Plücker condition $(\mathbf{a},\mathbf{a}^o) = 0$ and the normalization of $\mathbf{a}$ $\|\mathbf{a}\| = 1$

$$(\mathbf{a},\mathbf{p}_0)\mathbf{a} = (\mathbf{a},\mathbf{a}\times\mathbf{a}^o)\mathbf{a} = 0,$$

because $(\mathbf{a},\mathbf{a}\times\mathbf{a}^o) = 0$.

$$-\cos\theta\mathbf{p}_0 + \sin\theta\mathbf{a} + \mathbf{p}_0 = (1-\cos\theta)\mathbf{p}_0 + \sin\theta\mathbf{a}^o = (1-\cos\theta)\mathbf{a}\times\mathbf{a}^o + \sin\theta\mathbf{a}^o.$$

As a result, given that $\mathbf{p}_0 = \mathbf{a} \times \mathbf{a}^o$ we get:

$$R_{\theta,\mathbf{a}}(\mathbf{p} - \mathbf{p}_0) + \mathbf{p}_0 = \cos\theta\mathbf{p} + \sin\theta\mathbf{a} \times \mathbf{p} + (1 - \cos\theta)(\mathbf{a}, \mathbf{p})\mathbf{a} + \sin\theta\mathbf{a}^o + (1 - \cos\theta)\mathbf{a} \times \mathbf{a}^o,$$

which completely coincides with the formula (10) obtained by the dual quaternion method up to the notation $\mathbf{p} \to \mathbf{p}^o$.

2.3.3. Translation along an axis without rotation

Let's write the dual quaternion for translation along the axis $\mathbf{A} = \mathbf{a} + \mathbf{a}^o\varepsilon$:

$$T = 1 + \frac{\theta^o}{2}\mathbf{a}\varepsilon,$$

where θ^o is the dual part of the dual angle Θ.

$$T^\dagger = \overline{(T^*)} = \overline{(1 + \frac{\theta^o}{2}\mathbf{a}^*\varepsilon)} = \overline{(1 - \frac{\theta^o}{2}\mathbf{a}\varepsilon)} = 1 + \frac{\theta^o}{2}\mathbf{a}\varepsilon = T.$$

We get that $T^\dagger = T$.

Let's apply the sandwich product to the point $P = 1 + \mathbf{p}^o\varepsilon$:

$$\begin{aligned} P' = TPT^\dagger = TPT &= \left(1 + \frac{\theta^o}{2}\mathbf{a}\varepsilon\right)(1 + \mathbf{p}^o\varepsilon)\left(1 + \frac{\theta^o}{2}\mathbf{a}\varepsilon\right) = \\ &= \left(1 + \frac{\theta^o}{2}\mathbf{a}\varepsilon\right)\left(1 + \frac{\theta^o}{2}\mathbf{a}\varepsilon + \mathbf{p}^o\varepsilon\right) = 1 + \frac{\theta^o}{2}\mathbf{a}\varepsilon + \mathbf{p}^o\varepsilon + \frac{\theta^o}{2}\mathbf{a}\varepsilon + \frac{\theta^o}{2}\frac{\theta^o}{2}\mathbf{a}\mathbf{a}\varepsilon^2 + \frac{\theta^o}{2}\mathbf{a}\mathbf{p}^o\varepsilon^2 = \\ &= 1 + \frac{\theta^o}{2}\mathbf{a}\varepsilon + \mathbf{p}^o\varepsilon + \frac{\theta^o}{2}\mathbf{a}\varepsilon = 1 + \theta^o\mathbf{a}\varepsilon + \mathbf{p}^o\varepsilon = 1 + (\theta^o\mathbf{a} + \mathbf{p}^o)\varepsilon = 1 + (\mathbf{p}^o + \theta^o\mathbf{a})\varepsilon. \end{aligned}$$

$$P' = TPT^\dagger = 1 + (\mathbf{p}^o + \theta^o\mathbf{a})\varepsilon.$$

2.3.4. Composition of rotations and translations

Above, we used the translation principle to obtain the screw motion dual quaternion and wrote it in terms of the dual (7) and real (8) angles. However, interestingly, it can be obtained by dual quaternion multiplication of only the rotational and only the translational dual quaternions.

Let us designate:

$$R = \cos\frac{\theta}{2} + \sin\frac{\theta}{2}(\mathbf{a} + \mathbf{a}^o\varepsilon),$$

$$T = 1 + \frac{\theta^o}{2}n\varepsilon,$$

$$\begin{aligned} RT &= \left(\cos\frac{\theta}{2} + \sin\frac{\theta}{2}(\mathbf{a} + \mathbf{a}^o\varepsilon)\right)\left(1 + \frac{\theta^o}{2}\mathbf{a}\varepsilon\right) = \\ &= \cos\frac{\theta}{2} + \sin\frac{\theta}{2}(\mathbf{a} + \mathbf{a}^o\varepsilon) + \frac{\theta^o}{2}\cos\frac{\theta}{2}\mathbf{a}\varepsilon + \sin\frac{\theta}{2}\mathbf{a}\frac{\theta^o}{2}\mathbf{a}\varepsilon + \sin\frac{\theta}{2}\frac{\theta^o}{2}\mathbf{a}^o\mathbf{a}\varepsilon^2 = \\ &= \cos\frac{\theta}{2} + \sin\frac{\theta}{2}(\mathbf{a} + \mathbf{a}^o\varepsilon) + \cos\frac{\theta}{2}\frac{\theta^o}{2}\mathbf{a}\varepsilon - \sin\frac{\theta}{2}\frac{\theta^o}{2}\varepsilon = \\ &= \cos\frac{\theta}{2} + \sin\frac{\theta}{2}(\mathbf{a} + \mathbf{a}^o\varepsilon) + \left(\cos\frac{\theta}{2}\mathbf{a} - \sin\frac{\theta}{2}\right)\frac{\theta^o}{2}\varepsilon = \Lambda. \end{aligned}$$

Taking into account the simplification $\mathbf{aa} = -(\mathbf{a},\mathbf{a}) + \mathbf{a}\times\mathbf{a} = -\|\mathbf{a}\|^2 = -1$, we obtain the final formula:

$$\Lambda = RT,$$

which completely coincides with the full formula:

$$TR = \left(1 + \frac{\theta^o}{2}\mathbf{a}\varepsilon\right)\left(\cos\frac{\theta}{2} + \sin\frac{\theta}{2}(\mathbf{a} + \mathbf{a}^o\varepsilon)\right) = \cos\frac{\theta}{2} + \frac{\theta^o}{2}\mathbf{a}\cos\frac{\theta}{2}\varepsilon + \sin\frac{\theta}{2}(\mathbf{a} + \mathbf{a}^o\varepsilon) + \sin\frac{\theta}{2}\frac{\theta^o}{2}\mathbf{aa}\varepsilon =$$
$$= \cos\frac{\theta}{2} + \sin\frac{\theta}{2}(\mathbf{a} + \mathbf{a}^o\varepsilon) + \left(\cos\frac{\theta}{2}\mathbf{a} - \sin\frac{\theta}{2}\right)\frac{\theta^o}{2}\varepsilon = \Lambda.$$

It turns out that translation and rotation along the screw axis commute:

$$\Lambda = RT = TR.$$

2.4. Screw motion of a straight line

An arbitrary line with a specified direction $\mathbf{a}$ is defined by a pure dual quaternion $\mathbf{L} = \mathbf{v} + \mathbf{m}\varepsilon$. The components of the vectors $\{\mathbf{v} \mid \mathbf{m}\}$ are Plücker coordinates, and the Plücker condition $(\mathbf{v},\mathbf{m}) = 0$ is satisfied. Screw motion is defined by the same dual quaternion Λ, but the sandwich formula looks somewhat different:

$$\mathbf{L}' = \Lambda\mathbf{L}\Lambda^*.$$

Just as for a point, we will first consider the translation of a straight line, then rotations, and then find their composition.

2.4.1. Translation of a straight line along an axis

Consider the line $\mathbf{L} = \mathbf{v} + \mathbf{m}\varepsilon$ and apply T to it as a sandwich operator:

$$T = 1 + \frac{\theta^o}{2}\mathbf{a}\varepsilon, \quad T^* = 1 - \frac{\theta^o}{2}\mathbf{a}\varepsilon,$$

$$T(\mathbf{v} + \mathbf{m}\varepsilon)T^* = \left(1 + \frac{\theta^o}{2}\mathbf{a}\varepsilon\right)(\mathbf{v} + \mathbf{m}\varepsilon)\left(1 - \frac{\theta^o}{2}\mathbf{a}\varepsilon\right).$$

It can be shown that $TT^* = 1$, since $T^* = 1 - \frac{\theta^o}{2}\mathbf{a}\varepsilon$, then

$$\left(1 + \frac{\theta^o}{2}\mathbf{a}\varepsilon\right)\left(1 - \frac{\theta^o}{2}\mathbf{a}\varepsilon\right) = 1 - \frac{\theta^o}{2}\mathbf{a}\varepsilon + \frac{\theta^o}{2}\mathbf{a}\varepsilon = 1 \Rightarrow |T| = TT^* = 1.$$

Let's now reveal the formula:

$$T(\mathbf{v} + \mathbf{m}\varepsilon)T^* = T\mathbf{v}T^* + T\mathbf{m}T^*\varepsilon.$$

Because

$$T\mathbf{v}T^* = \left(1 + \frac{\theta^o}{2}\mathbf{a}\varepsilon\right)\left(\mathbf{v} - \frac{\theta^o}{2}\mathbf{v}\mathbf{a}\varepsilon\right) = \mathbf{v} + \frac{\theta^o}{2}\mathbf{a}\mathbf{v}\varepsilon - \frac{\theta^o}{2}\mathbf{v}\mathbf{a}\varepsilon$$

and

$$T\mathbf{m}T^* = \mathbf{m} + \frac{\theta^o}{2}\mathbf{a}\mathbf{m}\varepsilon - \frac{\theta^o}{2}\mathbf{m}\mathbf{a}\varepsilon,$$

that

$$T\mathbf{m}T^*\varepsilon = \mathbf{m}\varepsilon + \frac{\theta^o}{2}\mathbf{a}\mathbf{m}\varepsilon^2 - \frac{\theta^o}{2}\mathbf{m}\mathbf{a}\varepsilon^2 = \mathbf{m}\varepsilon.$$

$$T\mathbf{v}T^* + T\mathbf{m}T^*\varepsilon = \mathbf{v} + \frac{\theta^o}{2}(\mathbf{av} - \mathbf{va})\varepsilon + \mathbf{m}\varepsilon.$$

$$\mathbf{av} - \mathbf{va} = -(\mathbf{a}, \mathbf{v}) + \mathbf{a} \times \mathbf{v} + (\mathbf{v}, \mathbf{a}) - \mathbf{v} \times \mathbf{a} = 2\mathbf{a} \times \mathbf{v}.$$

Then it turns out that:

$$L' = TLT^* = \mathbf{v} + \frac{\theta^o}{2} 2\mathbf{a} \times \mathbf{v}\varepsilon + \mathbf{m}\varepsilon = \mathbf{v} + (\mathbf{m} + \theta^o \mathbf{a} \times \mathbf{v})\varepsilon,$$

where $\mathbf{m} = \mathbf{p}_0 \times \mathbf{v} \Rightarrow \mathbf{m} + \theta^o \mathbf{a} \times \mathbf{v} = \mathbf{p}_0 \times \mathbf{v} + \theta^o \mathbf{a} \times \mathbf{v} = (\mathbf{p}_0 + \theta^o \mathbf{a}) \times \mathbf{v}$, therefore we get:

$$L' = \mathbf{v} + (\mathbf{p}_0 + \theta^o \mathbf{a}) \times \mathbf{v}\varepsilon.$$

As a result, as expected, the direction vector $\mathbf{v}$ does not change during parallel translation (translation), but the moment of the line is transformed. The line was moved from the point P_0 towards the $\mathbf{a}$ axis by the amount θ^o. Note that we can map T^{-1} onto T^*, since $|T| = 1$ and $T^* = T^{-1}$. Therefore, the sandwich formula for translation can also be written as follows [13, p. 51]:

$$L' = TLT^* = TLT^{-1}.$$

2.4.2. Rotation of a line around an axis

Let's consider the rotating dual quaternion R and find its conjugate dual quaternion R^*:

$$R = \cos\frac{\theta}{2} + \sin\frac{\theta}{2}(\mathbf{a} + \mathbf{a}^o\varepsilon), \quad R^* = \cos\frac{\theta}{2} - \sin\frac{\theta}{2}(\mathbf{a} + \mathbf{a}^o\varepsilon).$$

It was shown earlier that $\mathbf{AA} = -1$, where $\mathbf{A} = \mathbf{a} + \mathbf{a}^o\varepsilon$, so:

$$RR^* = \cos^2\frac{\theta}{2} - \sin\frac{\theta}{2}\cos\frac{\theta}{2}\mathbf{A} + \sin\frac{\theta}{2}\cos\frac{\theta}{2}\mathbf{A} - \sin^2\frac{\theta}{2}\mathbf{AA} = \cos^2\frac{\theta}{2} + \sin^2\frac{\theta}{2} = 1 \Rightarrow RR^* = 1,$$

from which it follows that $R^{-1} = R^*$.

The rotation of the line $\mathbf{L} = \mathbf{v} + \mathbf{m}\varepsilon$ is carried out by the sandwich operator:

$$\mathbf{L}' = R\mathbf{L}R^* = R\mathbf{v}R^* + R\mathbf{m}R^*\varepsilon,$$

Let's calculate the first term in this formula:

$$\begin{aligned} R\mathbf{v}R &= \left(\cos\frac{\theta}{2} + \sin\frac{\theta}{2}\mathbf{A}\right)\mathbf{v}\left(\cos\frac{\theta}{2} - \sin\frac{\theta}{2}\mathbf{A}\right) = \\ &= \cos^2\frac{\theta}{2}\mathbf{v} - \sin\frac{\theta}{2}\cos\frac{\theta}{2}\mathbf{vA} + \sin\frac{\theta}{2}\cos\frac{\theta}{2}\mathbf{Av} - \sin^2\frac{\theta}{2}\mathbf{AvA} = \\ &= \cos^2\frac{\theta}{2}\mathbf{v} + \sin\frac{\theta}{2}\cos\frac{\theta}{2}(\mathbf{Av} - \mathbf{vA}) - \sin^2\frac{\theta}{2}\mathbf{AvA}. \end{aligned}$$

Because

$$\mathbf{Av} = (\mathbf{a} + \mathbf{a}^o\varepsilon)\mathbf{v} = \mathbf{av} + \mathbf{a}^o\mathbf{v}\varepsilon = -(\mathbf{a}, \mathbf{v}) - (\mathbf{a}^o, \mathbf{v})\varepsilon + \mathbf{a} \times \mathbf{v} + \mathbf{a}^o \times \mathbf{v}\varepsilon$$

and

$$\mathbf{vA} = \mathbf{v}(\mathbf{a} + \mathbf{a}^o\varepsilon) = \mathbf{va} + \mathbf{va}^o\varepsilon = -(\mathbf{v}, \mathbf{a}) - (\mathbf{v}, \mathbf{a}^o)\varepsilon + \mathbf{v} \times \mathbf{a} + \mathbf{v} \times \mathbf{a}^o\varepsilon,$$

that

$$\mathbf{Av} - \mathbf{vA} = 2\mathbf{a} \times \mathbf{v} + 2\mathbf{a}^o \times \mathbf{v}\varepsilon.$$

Next, we calculate the dual quaternion product **AvA**:

$$\mathbf{AvA} = (\mathbf{a}+\mathbf{a}^o\varepsilon)\mathbf{v}(\mathbf{a}+\mathbf{a}^o\varepsilon) = (\mathbf{a}+\mathbf{a}^o\varepsilon)(\mathbf{va}+\mathbf{va}^o\varepsilon) = \mathbf{ava}+\mathbf{ava}^o\varepsilon+\mathbf{a}^o\mathbf{va}\varepsilon+\mathbf{a}^o\mathbf{va}^o\varepsilon^2 = \mathbf{ava}+(\mathbf{ava}^o+\mathbf{a}^o\mathbf{va})\varepsilon,$$

where to simplify the dual part $\mathbf{ava}^o + \mathbf{a}^o\mathbf{va}$ we take into account that

$$\mathbf{ava}^o = -\mathbf{a}(\mathbf{v},\mathbf{a}^o) + \mathbf{av}\times\mathbf{a}^o = -(\mathbf{v},\mathbf{a}^o)\mathbf{a} - (\mathbf{a},\mathbf{v}\times\mathbf{a}^o) + \mathbf{a}\times\mathbf{v}\times\mathbf{a}^o = \\ = -(\mathbf{v},\mathbf{a}^o)\mathbf{a} - (\mathbf{a},\mathbf{v}\times\mathbf{a}^o) + \mathbf{v}(\mathbf{a},\mathbf{a}^o) - \mathbf{a}^o(\mathbf{a},\mathbf{v}),$$

$$\mathbf{a}^o\mathbf{va} = -\mathbf{a}^o(\mathbf{v},\mathbf{a}) + \mathbf{a}^o\mathbf{v}\times\mathbf{a} = -(\mathbf{v},\mathbf{a})\mathbf{a}^o - (\mathbf{a}^o,\mathbf{v}\times\mathbf{a}) + \mathbf{a}^o\times\mathbf{v}\times\mathbf{a} = \\ = -(\mathbf{v},\mathbf{a})\mathbf{a}^o - (\mathbf{a}^o,\mathbf{v}\times\mathbf{a}) + \mathbf{v}(\mathbf{a}^o,\mathbf{a}) - \mathbf{a}(\mathbf{a}^o,\mathbf{v}),$$

after which the expression in brackets is simplified:

$$\mathbf{ava}^o + \mathbf{a}^o\mathbf{va} = -2(\mathbf{v},\mathbf{a}^o)\mathbf{a} - 2(\mathbf{v},\mathbf{a})\mathbf{a}^o + 2(\mathbf{a},\mathbf{a}^o)\mathbf{v} = -2\big((\mathbf{v},\mathbf{a}^o)\mathbf{a} + (\mathbf{v},\mathbf{a})\mathbf{a}^o\big).$$

We used the Plücker condition $(\mathbf{a}^o,\mathbf{a}) = 0$ and the mixed product property $(\mathbf{a}^o,\mathbf{v}\times\mathbf{a}) = -(\mathbf{a},\mathbf{v}\times\mathbf{a}^o)$. Given that

$$\mathbf{ava} = -\mathbf{a}(\mathbf{v},\mathbf{a}) + \mathbf{av}\times\mathbf{a} = -\mathbf{a}(\mathbf{v},\mathbf{a}) - (\mathbf{a},\mathbf{v}\times\mathbf{a}) + \mathbf{a}\times\mathbf{v}\times\mathbf{a} = -\mathbf{a}(\mathbf{v},\mathbf{a}) + \mathbf{v}(\mathbf{a},\mathbf{a}) - \mathbf{a}(\mathbf{a},\mathbf{v}) = \mathbf{v} - 2\mathbf{a}(\mathbf{v},\mathbf{a}),$$

we obtain the final expression:

$$\mathbf{AvA} = \mathbf{v} - 2\mathbf{a}(\mathbf{v},\mathbf{a}) - 2\big((\mathbf{v},\mathbf{a}^o)\mathbf{a} + (\mathbf{v},\mathbf{a})\mathbf{a}^o\big)\varepsilon.$$

Substituting into the formula for $R\mathbf{v}R$ we get:

$$\cos^2\frac{\theta}{2}\mathbf{v} + 2\sin\frac{\theta}{2}\cos\frac{\theta}{2}(\mathbf{a}\times\mathbf{v} + \mathbf{a}^o\times\mathbf{v}\varepsilon) - \sin^2\frac{\theta}{2}\big(\mathbf{v} - 2\mathbf{a}(\mathbf{v},\mathbf{a}) - 2\big((\mathbf{v},\mathbf{a}^o)\mathbf{a} + (\mathbf{v},\mathbf{a})\mathbf{a}^o\big)\varepsilon\big) = \\ = \left(\cos^2\frac{\theta}{2} - \sin^2\frac{\theta}{2}\right)\mathbf{v} + \sin\theta(\mathbf{a}\times\mathbf{v} + \mathbf{a}^o\times\mathbf{v}\varepsilon) + 2\sin^2\frac{\theta}{2}\big(\mathbf{a}(\mathbf{v},\mathbf{a}) + \big((\mathbf{v},\mathbf{a}^o)\mathbf{a} + (\mathbf{v},\mathbf{a})\mathbf{a}^o\big)\varepsilon\big) = \\ = \cos\theta\mathbf{v} + \sin\theta\mathbf{a}\times\mathbf{v} + (1-\cos\theta)\mathbf{a}(\mathbf{v},\mathbf{a}) + \big[\sin\theta\mathbf{a}^o\times\mathbf{v} + (1-\cos\theta)\big((\mathbf{v},\mathbf{a}^o)\mathbf{a} + (\mathbf{v},\mathbf{a})\mathbf{a}^o\big)\big]\varepsilon.$$

The expression $R\mathbf{m}R^*\varepsilon$ is calculated in a completely similar way and is significantly simplified by multiplying by the dual imaginary unit ε:

$$R\mathbf{m}R^*\varepsilon = (\cos\theta\mathbf{m} + \sin\theta\mathbf{a}\times\mathbf{m} + (1-\cos\theta)\mathbf{a}(\mathbf{m},\mathbf{a}))\,\varepsilon.$$

Now we can write a sandwich formula for the rotation of a line around an axis:

$$\mathbf{L}' = R\mathbf{L}R^* = \mathbf{v}' + \mathbf{m}'\varepsilon,$$

where:

$$\mathbf{v}' = \cos\theta\mathbf{v} + \sin\theta\mathbf{a}\times\mathbf{v} + (1-\cos\theta)\mathbf{a}(\mathbf{v},\mathbf{a}),$$

$$\mathbf{m}' = \cos\theta\mathbf{m} + \sin\theta\mathbf{a}\times\mathbf{m} + (1-\cos\theta)\mathbf{a}(\mathbf{m},\mathbf{a}) + \sin\theta\mathbf{a}^o\times\mathbf{v} + (1-\cos\theta)\big((\mathbf{v},\mathbf{a}^o)\mathbf{a} + (\mathbf{v},\mathbf{a})\mathbf{a}^o\big).$$

To check this equality, we can take Rodrigues' formula and use the principle of transfer to write:

$$\mathbf{L}' = \cos\theta\mathbf{L} + \sin\theta\mathbf{A}\times\mathbf{L} + (1-\cos\theta)(\mathbf{A},\mathbf{L})\mathbf{A},$$

where

$$\mathbf{L} = \mathbf{v} + \mathbf{m}\varepsilon,$$
$$\mathbf{A} = \mathbf{a} + \mathbf{a}^o\varepsilon.$$

The arguments of the functions cos and sin are real, since we have pure rotation without translation:

$$\mathbf{A} \times \mathbf{L} = \mathbf{a} \times \mathbf{v} + (\mathbf{a} \times \mathbf{m} + \mathbf{a}^o \times \mathbf{v})\varepsilon,$$
$$(\mathbf{A}, \mathbf{L})\mathbf{A} = (\mathbf{a}, \mathbf{v})\mathbf{a} + [(\mathbf{a}, \mathbf{m})\mathbf{a} + (\mathbf{a}^o, \mathbf{v})\mathbf{a}]\varepsilon + (\mathbf{a}, \mathbf{v})\mathbf{a}^o\varepsilon.$$

Combining the expressions we obtain:

$$\begin{aligned}\mathbf{L}' = \cos\theta\mathbf{v} + \sin\theta\mathbf{a} \times \mathbf{v} + (1 - \cos\theta)\mathbf{a}(\mathbf{v}, \mathbf{a}) + [\cos\theta\mathbf{m} + \sin\theta\mathbf{a} \times \mathbf{m} + (1 - \cos\theta)(\mathbf{a}, \mathbf{m})\mathbf{a}]\,\varepsilon + \\ + [\sin\theta\mathbf{a}^o \times \mathbf{v} + (1 - \cos\theta)\,((\mathbf{a}^o, \mathbf{v})\mathbf{a} + (\mathbf{a}, \mathbf{v})\mathbf{a}^o)]\,\varepsilon = \\ = R_{\theta,\mathbf{a}}(\mathbf{v}) + R_{\theta,\mathbf{a}}(\mathbf{m})\varepsilon + \left[\sin\theta\mathbf{a}^o \times \mathbf{v} + (1 - \cos\theta)\big((\mathbf{v}, \mathbf{a}^o)\mathbf{a} + (\mathbf{v}, \mathbf{a})\mathbf{a}^o\big)\right]\varepsilon,\end{aligned}$$

which exactly repeats the result obtained from the dual quaternion formula.

2.5. Screw motion of a plane

Let us recall that the dual quaternion representation of a plane has the following form:

$$\Pi = ai + bj + ck + d\varepsilon = \mathbf{n} + d\varepsilon.$$

where $\mathbf{n}$ is the unit direction normal vector, and d is the distance from the origin to the plane. The sandwich formula for the screw motion of the plane will have the form [13, pp. 49–50]:

$$\Pi' = \Lambda\Pi\Lambda^\dagger,$$

where the dual quaternion of screw motion is already known to us:

$$\Lambda = \cos\frac{\theta}{2} + \sin\frac{\theta}{2}(\mathbf{a} + \mathbf{a}^o\varepsilon) + \left(\cos\frac{\theta}{2}\mathbf{a} - \sin\frac{\theta}{2}\right)\frac{\theta^o}{2}\varepsilon.$$

2.5.1. Plane Translation

Consider the translation $T = 1 + \frac{\theta^o}{2}\mathbf{a}\varepsilon$. Note that:

$$TT = \left(1 + \frac{\theta^o}{2}\mathbf{a}\varepsilon\right)^2 = 1 + 2\frac{\theta^o}{2}\mathbf{a}\varepsilon = 1 + \theta^o\mathbf{a}\varepsilon,$$
$$T\Pi T^\dagger = T\Pi T = Td\varepsilon T + T\mathbf{n}T.$$

Let's carry out the calculations separately:

$$Td\varepsilon T = d\varepsilon TT = d\varepsilon + 2\theta^o\mathbf{a}\varepsilon d\varepsilon = d\varepsilon,$$

$$\begin{aligned}T\mathbf{n}T = \left(1 + \frac{\theta^o}{2}\mathbf{a}\varepsilon\right)\mathbf{n}\left(1 + \frac{\theta^o}{2}\mathbf{a}\varepsilon\right) = \left(1 + \frac{\theta^o}{2}\mathbf{a}\varepsilon\right)\left(\mathbf{n} + \frac{\theta^o}{2}\mathbf{n}\mathbf{a}\varepsilon\right) = \\ = \mathbf{n} + \frac{\theta^o}{2}\mathbf{n}\mathbf{a}\varepsilon + \frac{\theta^o}{2}\mathbf{a}\mathbf{n}\varepsilon = \mathbf{n} + \frac{\theta^o}{2}(\mathbf{n}\mathbf{a} + \mathbf{a}\mathbf{n})\varepsilon = \mathbf{n} - \theta^o(\mathbf{a}, \mathbf{n})\varepsilon\end{aligned}$$

because the product of vectors:

$$\mathbf{n}\mathbf{a} + \mathbf{a}\mathbf{n} = -(\mathbf{n}, \mathbf{a}) + \mathbf{n} \times \mathbf{a} - (\mathbf{a}, \mathbf{n}) + \mathbf{a} \times \mathbf{n} = -2(\mathbf{a}, \mathbf{n}).$$

As a result we get:

$$\Pi' = T\Pi T = \mathbf{n} + (d - \theta^o(\mathbf{a}, \mathbf{n}))\varepsilon$$

2.5.2. Rotation of a plane

Consider now the pure rotation $R = \cos\frac{\theta}{2} + \sin\frac{\theta}{2}(\mathbf{a} + \mathbf{a}^o\varepsilon)$ and $R^\dagger = \cos\frac{\theta}{2} - \sin\frac{\theta}{2}(\mathbf{a} + \mathbf{a}^o\varepsilon)$. Using the formula (9), we write:

$$R(\mathbf{n} + d\varepsilon)R^\dagger = R\mathbf{n}R^\dagger + d\varepsilon RR^\dagger = R\mathbf{n}R^\dagger + d\varepsilon,$$

The calculation of $R\mathbf{n}R^\dagger$ is similar to the calculation we already did for a straight line, but it differs from it in details, so we will also present all the calculations in detail:

$$\begin{aligned}
&\left(\cos\frac{\theta}{2} + \sin\frac{\theta}{2}(\mathbf{a} + \mathbf{a}^o\varepsilon)\right)\left(\cos\frac{\theta}{2}\mathbf{n} - \sin\frac{\theta}{2}(\mathbf{n}\mathbf{a} - \mathbf{n}\mathbf{a}^o\varepsilon)\right) = \\
&= \cos^2\frac{\theta}{2}\mathbf{n} - \cos\frac{\theta}{2}\sin\frac{\theta}{2}(\mathbf{n}\mathbf{a} - \mathbf{n}\mathbf{a}^o\varepsilon) + \sin\frac{\theta}{2}\cos\frac{\theta}{2}(\mathbf{a}\mathbf{n} + \mathbf{a}^o\mathbf{n}\varepsilon) - \sin^2\frac{\theta}{2}(\mathbf{a}\mathbf{n}\mathbf{a} - \mathbf{a}\mathbf{n}\mathbf{a}^o\varepsilon + \mathbf{a}^o\mathbf{n}\mathbf{a}\varepsilon) = \\
&= \cos^2\frac{\theta}{2}\mathbf{n} + 2\cos\frac{\theta}{2}\sin\frac{\theta}{2}(\mathbf{a}\times\mathbf{n} - (\mathbf{a}^o, \mathbf{n})\varepsilon) - \sin^2\frac{\theta}{2}(\mathbf{n} - 2(\mathbf{n}, \mathbf{a})\mathbf{a} + 2(\mathbf{a}, \mathbf{n}, \mathbf{a}^o)\varepsilon) = \\
&= \cos\theta\mathbf{n} + \sin\theta(\mathbf{a}\times\mathbf{n} - (\mathbf{a}^o, \mathbf{n})\varepsilon) + (1 - \cos\theta)[(\mathbf{n}, \mathbf{a})\mathbf{a} - (\mathbf{a}, \mathbf{n}, \mathbf{a}^o)\varepsilon] = \\
&= \cos\theta\mathbf{n} + \sin\theta\mathbf{a}\times\mathbf{n} + (1 - \cos\theta)(\mathbf{n}, \mathbf{a})\mathbf{a} - [\sin\theta(\mathbf{a}^o, \mathbf{n}) + (\mathbf{a}, \mathbf{n}, \mathbf{a}^o)(1 - \cos\theta)]\varepsilon.
\end{aligned}$$

The following simplifications were used during the calculation:

$$-\mathbf{n}\mathbf{a} + \mathbf{n}\mathbf{a}^o\varepsilon + \mathbf{a}\mathbf{n} + \mathbf{a}^o\mathbf{n}\varepsilon = (\mathbf{a}\mathbf{n} - \mathbf{n}\mathbf{a}) + (\mathbf{a}^o\mathbf{n} + \mathbf{n}\mathbf{a}^o)\varepsilon = 2\mathbf{a}\times\mathbf{n} - 2(\mathbf{a}^o, \mathbf{n})\varepsilon,$$

$$\mathbf{a}\mathbf{n}\mathbf{a} = \|\mathbf{a}\|^2\mathbf{n} - 2(\mathbf{n}, \mathbf{a})\mathbf{a} = \mathbf{n} - 2(\mathbf{n}, \mathbf{a})\mathbf{a},$$

$$\begin{aligned}
\mathbf{a}^o\mathbf{n}\mathbf{a} = -\mathbf{a}^o(\mathbf{n}, \mathbf{a}) + \mathbf{a}^o\mathbf{n}\times\mathbf{a} = -\mathbf{a}^o(\mathbf{n}, \mathbf{a}) - (\mathbf{a}^o, \mathbf{n}\times\mathbf{a}) + \mathbf{a}^o\times\mathbf{n}\times\mathbf{a} = \\
= -\mathbf{a}^o(\mathbf{n}, \mathbf{a}) - (\mathbf{a}^o, \mathbf{n}\times\mathbf{a}) + \mathbf{n}(\mathbf{a}^o, \mathbf{a}) - \mathbf{a}(\mathbf{a}^o, \mathbf{n}),
\end{aligned}$$

$$\begin{aligned}
\mathbf{a}\mathbf{n}\mathbf{a}^o = -\mathbf{a}(\mathbf{n}, \mathbf{a}^o) + \mathbf{a}\mathbf{n}\times\mathbf{a}^o = -\mathbf{a}(\mathbf{n}, \mathbf{a}^o) - (\mathbf{a}, \mathbf{n}\times\mathbf{a}^o) + \mathbf{a}\times\mathbf{n}\times\mathbf{a}^o = \\
= -\mathbf{a}(\mathbf{n}, \mathbf{a}^o) - (\mathbf{a}, \mathbf{n}\times\mathbf{a}^o) + \mathbf{n}(\mathbf{a}, \mathbf{a}^o) - \mathbf{a}^o(\mathbf{a}, \mathbf{n}),
\end{aligned}$$

$$\mathbf{a}^o\mathbf{n}\mathbf{a} - \mathbf{a}\mathbf{n}\mathbf{a}^o = -(\mathbf{a}^o, \mathbf{n}\times\mathbf{a}) - (\mathbf{a}, \mathbf{n}\times\mathbf{a}^o) = +(\mathbf{a}, \mathbf{n}, \mathbf{a}^o) + (\mathbf{a}, \mathbf{n}, \mathbf{a}^o) = 2(\mathbf{a}, \mathbf{n}, \mathbf{a}^o).$$

The result is that:

$$\Pi' = R\Pi R^\dagger = \cos\theta\mathbf{n} + \sin\theta\mathbf{a}\times\mathbf{n} + (1 - \cos\theta)(\mathbf{n}, \mathbf{a})\mathbf{a} + [d - \sin\theta(\mathbf{a}^o, \mathbf{n}) - (1 - \cos\theta)(\mathbf{a}, \mathbf{n}, \mathbf{a}^o)]\varepsilon$$

In [13] there is only a formula for rotation around an axis passing through the origin, but there is no formula for an arbitrary axis of rotation.

2.6. Relationship of dual quaternions to projective transformation matrices in $\mathbb{R}\mathrm{P}^3$

Let us write the dual quaternion that defines the screw motion in quaternion form:

$$\Lambda = \lambda + \lambda^o\varepsilon = \cos\frac{\theta}{2} + \sin\frac{\theta}{2}\mathbf{a} + \left(\sin\frac{\theta}{2}\mathbf{a}^o + \frac{\theta^o}{2}\cos\frac{\theta}{2}\mathbf{a} - \frac{\theta^o}{2}\sin\frac{\theta}{2}\right)\varepsilon,$$

$$\lambda = \lambda_0 + \lambda_1\mathrm{i} + \lambda_2\mathrm{j} + \lambda_3\mathrm{k} = \cos\frac{\theta}{2} + \sin\frac{\theta}{2}\mathbf{a}, \quad \lambda^o = \lambda_0^o + \lambda_1^o\mathrm{i} + \lambda_2^o\mathrm{j} + \lambda_3^o\mathrm{k} = \sin\frac{\theta}{2}\mathbf{a}^o + \frac{\theta^o}{2}\cos\frac{\theta}{2}\mathbf{a} - \frac{\theta^o}{2}\sin\frac{\theta}{2}.$$

The quaternion λ specifies the rotation, and the quaternion λ^o is responsible for translation operations.

Table 2

Relationship of dual quaternions to projective transformation matrices

Comparison criterion	**Matrixes**	**Dual quaternions**
Number of scalar coefficients	$4 \times 4 = 16$	$4 + 4 = 8$
Multiplications (compositions)	48 scalar multiplications	48 scalar multiplications
Point motion	12 scalar multiplications	96 scalar multiplications

The projective transformation matrix that defines rotation and translation looks like this:

$$M = \left[\begin{array}{c|c} R & \mathbf{t} \\ \hline \mathbf{0}^T & 1 \end{array}\right] = \left[\begin{array}{ccc|c} r_1^1 & r_2^1 & r_3^1 & t_x \\ r_1^2 & r_2^2 & r_3^2 & t_y \\ r_1^3 & r_2^3 & r_3^3 & t_z \\ \hline 0 & 0 & 0 & 1 \end{array}\right]$$

where the matrix R is written in terms of the quaternion coefficients $\lambda_0, \lambda_1, \lambda_2, \lambda_3$ as

$$R = \begin{bmatrix} \lambda_0^2 + \lambda_1^2 - \lambda_2^2 - \lambda_3^2 & 2(\lambda_1\lambda_2 + \lambda_0\lambda_3) & 2(\lambda_1\lambda_3 + \lambda_0\lambda_2) \\ 2(\lambda_1\lambda_2 + \lambda_0\lambda_3) & \lambda_0^2 - \lambda_1^2 + \lambda_2^2 - \lambda_3^2 & 2(\lambda_2\lambda_3 - \lambda_0\lambda_1) \\ 2(\lambda_1\lambda_3 - \lambda_0\lambda_2) & 2(\lambda_2\lambda_3 + \lambda_0\lambda_1) & \lambda_0^2 - \lambda_1^2 - \lambda_2^2 + \lambda_3^2 \end{bmatrix}.$$

Let us show that the column vector $\mathbf{t}$ can be calculated using the formula

$$\mathbf{t} = 2\lambda^o\lambda^*.$$

Using the Plücker condition $(\mathbf{a}, \mathbf{a}^o) = 0$ and the unity of the vector $\mathbf{a}$, we can prove:

$$2\lambda^o\lambda^* = \theta^o\mathbf{a} + \sin\theta\mathbf{a}^o + 2\sin^2\frac{\theta}{2}\mathbf{a} \times \mathbf{a}^o = \theta^o\mathbf{a} + \sin\theta\mathbf{a}^o + (1 - \cos\theta)\mathbf{a} \times \mathbf{a}^o.$$

This formula has two terms:

- $\sin\theta\mathbf{a}^o + (1 - \cos\theta)\mathbf{a} \times \mathbf{a}^o$ — coincides with the part of the formula for rotation around an arbitrary axis (translation and return to the origin);
- $\theta^o\mathbf{a}$ — specifies a translation along the $\mathbf{a}$ axis by a distance θ.

If the matrix M is given, then the principal part λ of the dual quaternion Λ can be calculated using a special algorithm, described in detail in section 1.4.3, and the calculation of λ^o can be carried out using the formula:

$$\lambda^o = \frac{1}{2}\mathbf{t}\lambda = \frac{1}{2}t\lambda, \quad t = 0 + \mathbf{t}.$$

The formula is valid due to (see table 2):

$$\mathbf{t} = 2\lambda^o\lambda^* \quad \Rightarrow \quad \frac{1}{2}\mathbf{t}\lambda = \frac{1}{2}2\lambda^o\underbrace{\lambda^*\lambda}_{=1} = \lambda^o.$$

- Dual quaternions are computationally less efficient than matrices.
- Dual quaternions are more convenient for defining an arbitrary axis.
- Dual quaternions are easier to use for making heuristic inferences.

2.7. Reflection about a plane using dual quaternions

For a point P represented by a dual quaternion $P = w + \mathbf{p}\varepsilon = w + (x\mathrm{i} + y\mathrm{j} + z\mathrm{k})\varepsilon$, the reflection formula with respect to the plane represented by the dual quaternion $N = \mathbf{n} + d\varepsilon$, where $\|\mathbf{n}\| = 1$, has the following form:

$$P' = -NP^*N.$$

A similar reflection formula is valid for an arbitrary plane π, represented by the dual quaternion $\Pi = \ + \delta\varepsilon$:

$$\Pi' = -N\Pi^*N.$$

In the case of reflection of a straight line represented by the dual quaternion $L = \mathbf{v} + \mathbf{m}\varepsilon$, a different formula is valid:

$$L' = NL^\dagger N^{-1}.$$

Since usually $\|\mathbf{n}\| = 1$, then $|N|^2 = \|\mathbf{n}\|^2 = 1$ and $N^{-1} = N^*/|N|^2 = N^*$.

Let us prove the formulas.

2.7.1. Point reflection

$$\begin{aligned} P' = -NP^*N &= -(\mathbf{n} + d\varepsilon)(w - \mathbf{p}\varepsilon)(\mathbf{n} + d\varepsilon) = -(\mathbf{n} + d\varepsilon)(w\mathbf{n} + dw\varepsilon - \mathbf{pn}\varepsilon - d\mathbf{p}\varepsilon^2) = \\ &= -(\mathbf{n} + d\varepsilon)(w\mathbf{n} + dw\varepsilon - \mathbf{pn}\varepsilon) = -(w\mathbf{nn} + dw\mathbf{n}\varepsilon - \mathbf{npn}\varepsilon + dw\mathbf{n}\varepsilon + d^2w\varepsilon^2 - \mathbf{pn}d\varepsilon^2) = \\ &= -(-w + 2dw\mathbf{n}\varepsilon - \mathbf{npn}\varepsilon) = w + \mathbf{npn}\varepsilon - 2dw\mathbf{n}\varepsilon = \\ &= w + (\mathbf{npn} - 2dw\mathbf{n})\varepsilon = w + (\mathbf{p} - 2(\mathbf{p}, \mathbf{n})\mathbf{n} - 2dw\mathbf{n})\varepsilon. \end{aligned}$$

It can be seen that in the dual part of the dual quaternion P' we obtain exactly the same expression as in the formula (6), if the point is affine and $w = 1$.

For an affine point the expression is somewhat simplified:

$$P' = -NP^*N = 1 + \Big(\mathbf{p} - 2(\mathbf{p}, \mathbf{n})\mathbf{n} - 2d\mathbf{n}\Big)\varepsilon = 1 + (\mathbf{npn} - 2d\mathbf{n})\varepsilon.$$

2.7.2. Reflection of a plane

The reflected plane is defined by the dual quaternion $\Pi = \ + \delta\varepsilon$, and the plane relative to which the reflection occurs is represented by the dual quaternion $N = \mathbf{n} + d\varepsilon$, $\|\mathbf{n}\| = 1$.

Note that $\Pi^* = -\nu + \delta\varepsilon$ and write

$$\begin{aligned} -N\Pi^*N &= -(\mathbf{n} + d\varepsilon)(- + \delta\varepsilon)(\mathbf{n} + d\varepsilon) = \\ &= -(\mathbf{n} + d\varepsilon)(-\mathbf{n} - d\varepsilon + \delta\mathbf{n}\varepsilon + \delta d\varepsilon^2) = -(\mathbf{n} + d\varepsilon)(-\mathbf{n} - d\varepsilon + \delta\mathbf{n}\varepsilon) = \\ &= -(-\mathbf{nn} - d\mathbf{n}\varepsilon + \delta\mathbf{nn}\varepsilon - d\mathbf{n}\varepsilon - d^2\varepsilon^2 + \delta d\mathbf{n}\varepsilon^2) = \mathbf{nn} + d\mathbf{n}\varepsilon + \delta\varepsilon + d\mathbf{n}\varepsilon = \\ &= \mathbf{nn} + d(\mathbf{n} + \mathbf{n})\varepsilon + \delta\varepsilon = \ - 2(, \mathbf{n})\mathbf{n} - \big(2d(, \mathbf{n}) - \delta\big)\varepsilon. \end{aligned}$$

As a result, we obtain the formula:

$$-N\Pi^*N = \ - 2(, \mathbf{n})\mathbf{n} - (2d(, \mathbf{n}) - \delta)\varepsilon.$$

This formula can be rewritten in a different form if we consider the case when the planes N and Π are not parallel and have a common point P, then

$$\delta\varepsilon = -(P,) = -(1 + \mathbf{p}\varepsilon,) = -(1,) - (\mathbf{p},)\varepsilon = -(\mathbf{p},)\varepsilon, \quad d\varepsilon = -(P, \mathbf{n}).$$

Using these two relations, the plane reflection formula can be written as follows:

$$-N\Pi^*N = -2(,\mathbf{n})\mathbf{n} - (-2(P,\mathbf{n})(,\mathbf{n}) + (P,)) = -2(,\mathbf{n})\mathbf{n} - ((P,-2(,\mathbf{n})\mathbf{n}) + (P,)) =$$
$$= -2(,\mathbf{n})\mathbf{n} - (P,\ -2(,\mathbf{n})\mathbf{n}) = -2(,\mathbf{n})\mathbf{n} - (P,\ -2(,\mathbf{n})\mathbf{n}) = \mathbf{nn} - (P,\mathbf{nn}).$$

2.7.3. Reflection of a line

We will assume that the normal vector of the plane relative to which the reflection occurs is unitary. Then $N^{-1} = N^*/|N| = N^*/\|\mathbf{n}\| = N^*$. As a result, $N^{-1} = N^* = \mathbf{n}^* + d\varepsilon = -\mathbf{n} + d\varepsilon$.

Let us now find the dual conjugation of the dual quaternion of the line L

$$L^\dagger = (\mathbf{v} + \mathbf{m}\varepsilon)^\dagger = \mathbf{v}^* - \mathbf{m}\varepsilon = -\mathbf{v} + \mathbf{m}\varepsilon.$$

Now we transform the expression

$$L' = NL^\dagger N^{-1} = (\mathbf{n} + d\varepsilon)(-\mathbf{v} + \mathbf{m}\varepsilon)(-\mathbf{n} + d\varepsilon) = (\mathbf{n} + d\varepsilon)(\mathbf{vn} - d\mathbf{v}\varepsilon - \mathbf{mn}\varepsilon + d\mathbf{m}\varepsilon^2) =$$
$$= (\mathbf{n} + d\varepsilon)(\mathbf{vn} - d\mathbf{v}\varepsilon - \mathbf{mn}\varepsilon) = \mathbf{nvn} - d\mathbf{nv}\varepsilon - \mathbf{nmn}\varepsilon + d\mathbf{vn}\varepsilon - d^2\mathbf{v}\varepsilon^2 - d\mathbf{mn}\varepsilon^2 =$$
$$= \mathbf{nvn} - \mathbf{nmn}\varepsilon + d(\mathbf{vn} - \mathbf{nv})\varepsilon = \mathbf{nvn} - \mathbf{nmn}\varepsilon + 2d\mathbf{v} \times \mathbf{n}\varepsilon =$$
$$= \|\mathbf{n}\|^2\mathbf{v} - 2(\mathbf{v},\mathbf{n})\mathbf{n} - (\|\mathbf{n}\|^2\mathbf{m} - 2(\mathbf{m},\mathbf{n})\mathbf{n})\varepsilon + 2d\mathbf{v} \times \mathbf{n}\varepsilon = \mathbf{v} - 2(\mathbf{v},\mathbf{n})\mathbf{n} - (\mathbf{m} - 2(\mathbf{m},\mathbf{n})\mathbf{n} - 2d\mathbf{v} \times \mathbf{n})\varepsilon.$$

Let us express the moment $\mathbf{m} = \mathbf{p} \times \mathbf{v}$ through the point P on the line and the vector $\mathbf{v}$. In this case, we take into account that $\mathbf{v} - 2(\mathbf{v},\mathbf{n})\mathbf{n} = \mathbf{nvn}$ and $\mathbf{m} - 2(\mathbf{m},\mathbf{n})\mathbf{n} = \mathbf{nmn}$, then

$$NL^\dagger N^{-1} = \mathbf{nvn} - \mathbf{n}(\mathbf{p} \times \mathbf{v})\mathbf{n}\varepsilon + 2d\mathbf{v} \times \mathbf{n}\varepsilon = \mathbf{nvn} + (\mathbf{npn}) \times (\mathbf{nvn})\varepsilon + 2d\mathbf{v} \times \mathbf{n}\varepsilon.$$

Consider the expression $2d\mathbf{v} \times \mathbf{n}$ and transform it as follows:

$$\mathbf{nvn} = \mathbf{v} - 2(\mathbf{v},\mathbf{n})\mathbf{n} \Rightarrow \mathbf{v} = \mathbf{nvn} + 2(\mathbf{v},\mathbf{n})\mathbf{n},$$
$$2d\mathbf{v} \times \mathbf{n} = -2d\mathbf{n} \times \mathbf{v} = -2d\mathbf{n} \times (\mathbf{nvn} + 2(\mathbf{v},\mathbf{n})\mathbf{n}) =$$
$$= -2d\mathbf{n} \times (\mathbf{nvn}) - 2d(\mathbf{v},\mathbf{n})\mathbf{n} \times \mathbf{n} = -2d\mathbf{n} \times (\mathbf{nvn}),$$
$$2d\mathbf{v} \times \mathbf{n} = -2d\mathbf{n} \times (\mathbf{nvn}).$$

Let's substitute into the basic formula for reflection:

$$NL^\dagger N^{-1} = \mathbf{nvn} + (\mathbf{npn}) \times (\mathbf{nvn})\varepsilon - 2d\mathbf{n} \times (\mathbf{nvn}) = \mathbf{nvn} + (\mathbf{npn} - 2d\mathbf{n}) \times (\mathbf{nvn})\varepsilon.$$

Above we obtained the formula $P' = -NP^*N = 1 + (\mathbf{npn} - 2d\mathbf{n})\varepsilon$, with the help of which we can write

$$(\mathbf{npn} - 2d\mathbf{n})\varepsilon = (P' - 1)$$
$$NL^\dagger N^{-1} = \mathbf{nvn} + (P' - 1) \times (\mathbf{nvn}) = \mathbf{nvn} + (-NP^*N - 1) \times (\mathbf{nvn}).$$

In this form, this formula is written in [13, p. 59]. As a result, we have the formula in three versions:

$$L' = NL^\dagger N^{-1} = \mathbf{nvn} + (2d\mathbf{v} \times \mathbf{n} - \mathbf{nmn})\varepsilon = \mathbf{v} - 2(\mathbf{v},\mathbf{n})\mathbf{n} - (\mathbf{m} - 2(\mathbf{m},\mathbf{n})\mathbf{n} - 2d\mathbf{v} \times \mathbf{n})\varepsilon =$$
$$= \mathbf{nvn} + (-NP^*N - 1) \times (\mathbf{nvn}).$$

3. Results

As a result of the work, dual quaternionic formulas were obtained for

- of screw motion of a straight line, point, and plane;
- for pure rotation and pure parallel transfer (translation) of straight lines, points and planes;
- formula for calculating the projective matrix of screw motion according to a given dual quaternion;
- formulas for reflecting points, straight lines, and planes relative to randomly positioned planes.

We dare to attribute to one of the results the very fact of detailed derivation of these formulas, as this will make it easier for readers of the article to master working with dual quaternions.

4. Discussion

The material presented in this article shows that dual quaternions can be used to comprehensively describe both proper and improper movements in three-dimensional space. Currently, there are many alternative formalisms, the most popular and actively promoted of which is the geometric algebra formalism.

We will not go into a comparison of dual quaternions and geometric algebra here, since without a detailed presentation of the latter in terms consistent with dual quaternions, this comparison is unproductive. However, we note the following.

- In our opinion, when applying the Cayley–Dickson procedure, dual quaternions cannot be called some kind of artificial formation and they are quite logical.
- All dual quaternion formulas, when disclosed, are reduced to standard calculations with three-dimensional vectors, which makes it possible to efficiently implement calculations using, for example, shader languages where support for these vectors is implemented at the hardware level.

5. Conclusion

All the formulas obtained are of practical interest, as they can be used to calculate finite complex movements, as well as to construct surfaces (for example, linear ones). In future publications, we plan to use them to visualize various examples of movement, focusing on software implementation. The detailed material in this article will allow you not to be distracted by mathematical details and focus on the subtleties of software implementation and the algorithmic side of the issue.

Author Contributions: Conceptualization, Migran N. Gevorkyan, Dmitry S. Kulyabov; methodology, Migran N. Gevorkyan; writing—original draft preparation, Migran N. Gevorkyan, Olesya M. Abakumova; writing—review and editing, Anna V. Korolkova. All authors have read and agreed to the published version of the manuscript.

Funding: This research received no external funding.

Data Availability Statement: Data sharing is not applicable.

Conflicts of Interest: The authors declare no conflict of interest.

Declaration on Generative AI: The authors have not employed any Generative AI tools.

References

1. Gevorkyan, M. N., Vishnevskiy, N. A., Didus, K. V., Korolkova, A. V. & Kulyabov, D. S. Dual quaternion representation of points, lines and planes. *Discrete and Continuous Models and Applied Computational Science* **33,** 411–439. doi:10.22363/2658-4670-2025-33-4-411-439 (Dec. 2025).
2. Blaschke, W. J. E. *Anwendung dualer Quaternionen auf Kinematik* German (Suomalainen tiedeakatemia, Helsinki, 1958).
3. Blaschke, W. J. E. *Kinematics and Quaternions* trans. from the German by Delphenich, D. H. Berlin, 1960. doi:10.1002/zamm.19620420724.
4. Kotelnikov, A. P. *The Screw Calculus and Some of Its Applications to Geometry and Mechanics* 222 pp. (Annals of the Imperial University of Kazan, Kazan, 1895).
5. Dimentberg, F. M. *The Screw Calculus and Its Applications in Mechanics* 162 pp. (Foreign Technology Division, Springfield, 1969).
6. Fischer, I. *Dual-Number Methods in Kinematics, Statics and Dynamics* 240 pp. (CRC Press, 1998).
7. Huang, Z., Li, Q. & Ding, H. *Basics of Screw Theory* in *Theory of Parallel Mechanisms* (Springer Netherlands, Dordrecht, 2013). doi:10.1007/978-94-007-4201-7_1.
8. Featherstone, R. A Beginner's Guide to 6-D Vectors (Part 1). *IEEE Robotics and Automation Magazine* **17,** 83–94. doi:10.1109/MRA.2010.937853 (2010).
9. Featherstone, R. A Beginner's Guide to 6-D Vectors (Part 2) [Tutorial]. *IEEE Robotics and Automation Magazine* **17,** 88–99. doi:10.1109/MRA.2010.939560 (2010).
10. Chelnokov, Y. N. *Quaternionic and biquaternionic models and methods of solid mechanics and their applications. Geometry and kinematics of motion* 512 pp. (FIZMATLIT, Moscow, 2006).
11. Goldman, R. *An integrated introduction to computer graphics and geometric modeling* 592 pp. (CRC Press Taylor & Francis Group, Boca Raton, London, New York, 2009).
12. Goldman, R. *Rethinking Quaternions. Theory and Computation* doi:10 . 2200 / S00292ED1V01Y201008CGR013 (Morgan & Claypool, 2010).
13. Goldman, R. *Dual Quaternions and Their Associated Clifford Algebras* 279 pp. (CRC Press Taylor & Francis Group, Boca Raton, London, New York, 2024).
14. Kenwright, B. *A Survey on Dual-Quaternions* 2023. arXiv: 2303.14765 [math.OC].
15. Thomas, F. Approaching Dual Quaternions From Matrix Algebra. *IEEE Transactions on Robotics* **30,** 1–12. doi:10.1109/TRO.2014.2341312 (Aug. 2014).
16. Bruno, V. A. *Robot Kinematic Modeling and Control Based on Dual Quaternion Algebra. Part I: Fundamentals* working paper or preprint. https://hal.science/hal-01478225.
17. Wang, X., Han, D., Yu, C. & Zheng, Z. The geometric structure of unit dual quaternion with application in kinematic control. *Journal of Mathematical Analysis and Applications* **389,** 1352–1364. doi:10.1016/j.jmaa.2012.01.016 (2012).
18. Bekar, M. & Yayli, Y. Kinematics of Dual Quaternion Involution Matrices. *SDU Journal of Science* **11,** 121–132 (2016).
19. Dantam, N. T. *Practical Exponential Coordinates using Implicit Dual Quaternions* (Workshop on the Algorithmic Foundations of Robotics, 2018).
20. Kuipers, J. B. *Quaternions and rotation sequences* a primer with applications to orbits, aerospace and virtual reality. 371 pp. (Princeton University Press, 41 William Street, Princeton, New Jersey 08540, 1999).
21. Vince, J. *Rotation transforms for computer graphics* 1st ed. 232 pp. (Springer-Verlag London, 2011).
22. Lengyel, E. *Foundations of Game Engine Development.* 1: *Mathematics* 4 vols. 195 pp. (Terathon Software LLC, Lincoln, California, 2016).

Information about the authors

Abakumova, Olesya M.—Student of Department of Probability Theory and Cyber Security of RUDN University (e-mail: 1132220832@rudn.ru, ORCID: 0009-0002-5236-0027)

Gevorkyan, Migran N.—Docent, Candidate of Sciences in Physics and Mathematics, Associate Professor of Department of Probability Theory and Cyber Security of RUDN University (e-mail: gevorkyan-mn@rudn.ru, ORCID: 0000-0002-4834-4895, ResearcherID: E-9214-2016, Scopus Author ID: 57190004380)

Korolkova, Anna V.—Docent, Candidate of Sciences in Physics and Mathematics, Associate Professor of Department of Probability Theory and Cyber Security of RUDN University (e-mail: korolkova-av@rudn.ru, ORCID: 0000-0001-7141-7610, ResearcherID: I-3191-2013, Scopus Author ID: 36968057600)

Kulyabov, Dmitry S.—Professor, Doctor of Sciences in Physics and Mathematics, Professor of Department of Probability Theory and Cyber Security of RUDN University; Senior Researcher of Laboratory of Information Technologies, Joint Institute for Nuclear Research (e-mail: kulyabov-ds@rudn.ru, ORCID: 0000-0002-0877-7063, ResearcherID: I-3183-2013, Scopus Author ID: 35194130800)

УДК 511.84:512.523.282.2:519.711
DOI: 10.22363/2658-4670-2026-34-1-70-97 EDN: UOBPEG

Бикватернионное представление движения в трёхмерном пространстве

О. М. Абакумова[1], М. Н. Геворкян[1], А. В. Королькова[1], Д. С. Кулябов[1, 2]

[1] Российский университет дружбы народов, ул. Миклухо-Маклая, д. 6, Москва, 117198, Российская Федерация

[2] Объединённый институт ядерных исследований, ул. Жолио-Кюри, д. 6, Дубна, 141980, Российская Федерация

Аннотация. *Предпосылки* В предыдущей статье авторов был подробно рассмотрен вопрос использования бикватернионов для задания точек, прямых и плоскостей и решения стандартных геометрических задач. Данная статья является логическим продолжением и раскрывает применение бикватернионов для описания изометрий трёхмерного пространства. *Цель* Вывод всех необходимых формул для винтового движения точек, прямых и плоскостей, а также зеркальной симметрии (отражения) относительно плоскости. Доработка обозначений и формализма. *Методы* Используется алгебра дуальных чисел, кватернионов и бикватернионов, а также элементы теории винтов и скользящих векторов. *Результаты* Получены и систематизированы формулы для вращения, трансляции, отражения, винтового движения и зеркального вращения. *Выводы* Бикватернионы могут служить полноценным инструментом для описания винтового движения в пространстве. Благодаря возможности выражения бикватернионных операций через стандартное векторное и скалярное произведения, полученные формулы допускают эффективную программную реализацию.

Ключевые слова: натурное моделирование, воспроизводимое исследование, исследование как код

Discrete & Continuous Models & Applied Computational Science
ISSN 2658-7149 (Online), 2658-4670 (Print)

2026, 34 (1) 98–112
http://journals.rudn.ru/miph

Research article

UDC 519.872, 519.217
PACS 07.05.Tp, 02.70.Bf
DOI: 10.22363/2658-4670-2026-34-1-98-112
EDN: URTKJP

On a finite-difference scheme defining a birational non-quadratic map between time layers

Lyubov O. Lapshenkova[1], Kseniya S. Mashkovtseva[1], Alina A. Trusova[1], Mikhail D. Malykh[1,2]

[1] RUDN University, 6 Miklukho-Maklaya St, Moscow, 117198, Russian Federation

[2] Joint Institute for Nuclear Research, 6 Joliot-Curie St, Dubna, 141980, Russian Federation

(received: January 21, 2026; revised: February 1, 2026; accepted: February 10, 2026)

Abstract. The article considers reversible difference schemes for dynamical systems based on the system doubling method proposed by V.N. Abrashin and S.N. Sytova. The method duplicates the original variables, leading to an extended system whose finite-difference approximation defines a birational map between time layers. The preservation of algebraic integrals in such schemes is investigated. It is proved that if the original system admits a homogeneous quadratic first integral, the corresponding bilinear form is exactly preserved by the discrete scheme. This property is demonstrated on the Jacobi oscillator, where the geometric mean of the duplicated variables ensures exact conservation of the quadratic integral. A more detailed analysis is performed on the non-trivial Vanhaecke system, an integrable Hamiltonian system with two degrees of freedom and higher-degree polynomial integrals. Numerical experiments carried out in the computer algebra system Sage using the package fdm.sage confirm that the two copies oscillate synchronously around the exact values of the first integrals, and averaging reduces the oscillation amplitude. For separable Hamiltonian systems, the scheme is shown to be symplectic. The results obtained allow recommending the doubling method for constructing stable and structure-preserving numerical integrators for a wide class of dynamical systems with polynomial right-hand sides, including high-dimensional systems.

Key words and phrases: dynamical system, finite difference scheme, Kahan's method, integrable system, Vanhaecke system

For citation: Lapshenkova, L. O., Mashkovtseva, K. S., Trusova, A. A., Malykh, M. D. On a finite-difference scheme defining a birational non-quadratic map between time layers. *Discrete and Continuous Models and Applied Computational Science* **34** (1), 98–112. doi: 10.22363/2658-4670-2026-34-1-98-112. edn: URTKJP (2026).

1. Introduction

Dynamical systems with quadratic right–hand sides can be approximated by finite-difference schemes. Timestep of these schemes is a birational map [1–5]. Appelroth's quadratization allows any dynamical system with a polynomial right-hand side to be transformed into a system with a quadratic right-hand side to which the Kahan scheme can be applied [6]. Consequently, a very broad class of continuous dynamical systems admits approximation by t-symmetric finite-difference schemes that define a birational map between time layers. Following [5], we refer to such schemes as reversible. This property is of fundamental physical relevance. The time evolution of a dynamical system should, in general, be reversible: every initial state corresponds to exactly one final state, and every final state comes from exactly one initial state. Surprisingly, although the discrete-time methods often preserve one-to-one correspondence, it is not always true for the original continuous systems over the different time intervals [5].

It is worth mentioning that in the papers above the time step was implemented using a quadratic Cremona transformation. However, in the 1980s, V. N. Abrashin proposed a multicomponent alternating-direction method, which was successfully applied by S. N. Sytova. This method was used for modeling the nonlinear dynamics of electromagnetic radiation generated by charged particle beams in multidimensional spatially periodic structures [7, n. 4]. Like Appelroth's quadratization, this approach also introduces extra variables. However, instead of using a complicated substitution, it does so in a very simple and clear way: by duplicating every original variable. Because of this straightforward construction, we call it the system doubling method. A closer look shows that this method can be used almost for any dynamical system to build a finite-difference scheme where each time step is described by a Cremona transformation. In this paper, we give a general description of the doubling method and explain its key algebraic properties.

2. Methods

2.1. System doubling method

Let $\mathfrak{x} = (x_1, \dots, x_n)$ and there is a dynamical system

$$\frac{d\mathfrak{x}}{dt} = \mathfrak{f}(\mathfrak{x}). \tag{1}$$

For simplicity, assume that the right-hand sides in (1) are polynomials.

Introduce a new set of variables $\mathfrak{x}'$, which duplicate the original set of variables $\mathfrak{x}$, and form the doubled system:

$$\begin{cases} \dfrac{d\mathfrak{x}}{dt} = \mathfrak{f}(\mathfrak{x}'), \\ \dfrac{d\mathfrak{x}'}{dt} = \mathfrak{f}(\mathfrak{x}). \end{cases} \tag{2}$$

We formulate the equivalence of these systems as a theorem.

Theorem 9. *Every solution $\mathfrak{x} = \mathfrak{h}(t)$ of system* (1) *extends to a solution $\mathfrak{x} = \mathfrak{x}' = \mathfrak{h}(t)$ of the doubled system* (2). *Conversely, any solution of system* (2) *that satisfies the initial condition $\mathfrak{x} = \mathfrak{x}'$ at some time t_0 satisfies the equation $\mathfrak{x}(t) = \mathfrak{x}'(t)$ for all t, and $\mathfrak{x} = \mathfrak{x}(t)$ is a solution of the original system* (1).

Proof. (i) Suppose $\mathfrak{x} = \mathfrak{h}(t)$ is a solution of system (1). Then

$$\frac{d\mathfrak{h}}{dt} = \mathfrak{f}(\mathfrak{h}).$$

Substituting $\mathfrak{x}(t) = \mathfrak{x}'(t) = \mathfrak{h}(t)$ into system (2), both equations are satisfied identically, since

$$\frac{d\mathfrak{x}}{dt} = \frac{d\mathfrak{h}}{dt} = \mathfrak{f}(\mathfrak{h}) = \mathfrak{f}(\mathfrak{x}'), \qquad \frac{d\mathfrak{x}'}{dt} = \frac{d\mathfrak{h}}{dt} = \mathfrak{f}(\mathfrak{h}) = \mathfrak{f}(\mathfrak{x}).$$

Thus, $(\mathfrak{h}(t), \mathfrak{h}(t))$ is a solution of the doubled system (2).

(ii) Conversely, let $(\mathfrak{x}(t), \mathfrak{x}'(t))$ be a solution of (2) satisfying the initial condition $\mathfrak{x}(0) = \mathfrak{x}'(0)$. Define the differences

$$u_i(t) = x_i'(t) - x_i(t), \qquad i = 1, \dots, n,$$

and collect them into the vector $\mathfrak{u}(t) = \mathfrak{x}'(t) - \mathfrak{x}(t)$. Differentiating gives

$$\frac{du_i}{dt} = \frac{dx_i'}{dt} - \frac{dx_i}{dt} = f_i(\mathfrak{x}) - f_i(\mathfrak{x}').$$

Since each component f_i is a polynomial (by assumption), the difference $f_i(\mathfrak{x}) - f_i(\mathfrak{x}')$ can be expressed as a linear combination of the differences $x_j' - x_j = u_j$ with polynomial coefficients. More precisely, there exist polynomials $g_{ij}(\mathfrak{x}, \mathfrak{x}')$ such that

$$f_i(\mathfrak{x}) - f_i(\mathfrak{x}') = \sum_{j=1}^{n} g_{ij}(\mathfrak{x}, \mathfrak{x}')\, u_j.$$

Hence, $\mathfrak{u}(t)$ satisfies the linear homogeneous system

$$\frac{du_i}{dt} = \sum_{j=1}^{n} g_{ij}(\mathfrak{x}(t), \mathfrak{x}'(t))\, u_j, \qquad i = 1, \dots, n.$$

This system has continuous (in fact, smooth) coefficients along the trajectory $(\mathfrak{x}(t), \mathfrak{x}'(t))$, and the initial condition $\mathfrak{u}(0) = 0$ holds by hypothesis. By uniqueness of solutions to linear ODEs, it follows that $\mathfrak{u}(t) \equiv 0$ for all t in the interval of existence. Therefore,

$$\mathfrak{x}(t) = \mathfrak{x}'(t) \quad \text{for all } t.$$

Substituting this identity into the first equation of (2) results in

$$\frac{d\mathfrak{x}}{dt} = \mathfrak{f}(\mathfrak{x}),$$

so $\mathfrak{x}(t)$ is indeed a solution of the original system (1). □

Remark 1. The polynomials u_i arising in the proof are a natural generalization of Darboux polynomials [8–10].

System (2) is more complex than system (1), yet it admits a straightforward finite-difference discretization:

$$\begin{cases} \hat{\mathfrak{x}} - \mathfrak{x} = \mathfrak{f}(\mathfrak{x}')\,\Delta t, \\ \hat{\mathfrak{x}}' - \mathfrak{x}' = \mathfrak{f}(\hat{\mathfrak{x}})\,\Delta t. \end{cases} \tag{3}$$

This scheme defines a birational map between the points $(\mathfrak{x}, \mathfrak{x}')$ and $(\hat{\mathfrak{x}}, \hat{\mathfrak{x}}')$.

In contrast to Kahan's method, the scheme (3) is not t-symmetric [5]. However, it possesses a generalized analogue of t-symmetry: the equations (3) are invariant under the transformation

$$\Delta t \mapsto -\Delta t, \qquad \mathfrak{x} \mapsto \hat{\mathfrak{x}}', \qquad \mathfrak{x}' \mapsto \hat{\mathfrak{x}}.$$

We shall refer to this property as generalized t-symmetry.

We have implemented the doubling-based scheme (3) in our `fdm` package for SAGE. All computations presented below were made using this package.

2.2. Quadratic integrals

Theorem 10. *Suppose that system* (1) *admits a homogeneous quadratic first integral*

$$v = \sum_{i,j} v_{ij}\, x_i x_j.$$

Then the doubled system (2) *possesses the first integral*

$$w = \sum_{i,j} v_{ij}\, x_i x'_j,$$

which is exactly preserved by the discrete scheme (3)*, i.e.,*

$$w(\hat{\mathfrak{x}}, \hat{\mathfrak{x}}') = w(\mathfrak{x}, \mathfrak{x}').$$

Proof. We first observe that

$$v = \mathfrak{x}^{\mathsf{T}} V \mathfrak{x},$$

where V is the symmetric matrix representing the quadratic form v. By assumption,

$$\frac{dv}{dt} = 0,$$

along solutions of (1), i.e.,

$$\mathfrak{f}(\mathfrak{x})^{\mathsf{T}} V \mathfrak{x} + \mathfrak{x}^{\mathsf{T}} V \mathfrak{f}(\mathfrak{x}) = 0.$$

Since V is symmetric, this simplifies to

$$\mathfrak{x}^{\mathsf{T}} V \mathfrak{f}(\mathfrak{x}) = 0. \tag{4}$$

This identity holds for all $\mathfrak{x} \in \mathbb{R}_n$, as it is satisfied along every solution of (1).

The quantity w is the bilinear form associated with the quadratic form v, so we can write

$$w = \mathfrak{x}^{\mathsf{T}} V \mathfrak{x}'.$$

Differentiating w along solutions of the doubled system (2) gives

$$\frac{dw}{dt} = \mathfrak{f}(\mathfrak{x}')^{\mathsf{T}} V \mathfrak{x}' + \mathfrak{x}^{\mathsf{T}} V \mathfrak{f}(\mathfrak{x}).$$

Each term on the right-hand side vanishes by identity (4). Consequently, w is a first integral of system (2).

We now turn to the finite-difference setting. Consider the increment

$$\hat{w} - w = w(\hat{\mathfrak{x}}, \hat{\mathfrak{x}}') - w(\mathfrak{x}, \mathfrak{x}') = \hat{\mathfrak{x}}^{\mathsf{T}} V \hat{\mathfrak{x}}' - \mathfrak{x}^{\mathsf{T}} V \mathfrak{x}'.$$

Adding and subtracting the term $\hat{\mathfrak{x}}^{\mathsf{T}} V \mathfrak{x}'$ gives

$$\hat{w} - w = \big(\hat{\mathfrak{x}}^{\mathsf{T}} V \hat{\mathfrak{x}}' - \hat{\mathfrak{x}}^{\mathsf{T}} V \mathfrak{x}'\big) + \big(\hat{\mathfrak{x}}^{\mathsf{T}} V \mathfrak{x}' - \mathfrak{x}^{\mathsf{T}} V \mathfrak{x}'\big).$$

Using the discrete scheme (3), we have

$$\hat{\mathfrak{x}}' - \mathfrak{x}' = \mathfrak{f}(\hat{\mathfrak{x}})\,\Delta t, \qquad \hat{\mathfrak{x}} - \mathfrak{x} = \mathfrak{f}(\mathfrak{x}')\,\Delta t,$$

and therefore

$$\hat{\mathfrak{x}}^{\mathsf{T}} V \hat{\mathfrak{x}}' - \hat{\mathfrak{x}}^{\mathsf{T}} V \mathfrak{x}' = \hat{\mathfrak{x}}^{\mathsf{T}} V\big(\hat{\mathfrak{x}}' - \mathfrak{x}'\big) = \Delta t\, \hat{\mathfrak{x}}^{\mathsf{T}} V \mathfrak{f}(\hat{\mathfrak{x}}),$$

$$\hat{\mathfrak{x}}^{\mathsf{T}} V \mathfrak{x}' - \mathfrak{x}^{\mathsf{T}} V \mathfrak{x}' = \big(\hat{\mathfrak{x}} - \mathfrak{x}\big)^{\mathsf{T}} V \mathfrak{x}' = \Delta t\, \mathfrak{f}(\mathfrak{x}')^{\mathsf{T}} V \mathfrak{x}'.$$

By identity (4), both terms vanish. Hence $\hat{w} = w$, which shows that w is exactly preserved by the discrete map (3). □

2.3. Hamiltonian formulation

Suppose the original system (1) is Hamiltonian, i.e., it can be written as

$$\frac{dp_i}{dt} = -\frac{\partial H}{\partial q_i}, \quad \frac{dq_i}{dt} = \frac{\partial H}{\partial p_i}, \quad i = 1, \dots, n.$$

Then the doubled system takes the form

$$\begin{cases} \dfrac{dp_i}{dt} = -\dfrac{\partial H'}{\partial q'_i}, \quad \dfrac{dq_i}{dt} = \dfrac{\partial H'}{\partial p'_i}, \quad i = 1, \dots, n, \\ \dfrac{dp'_i}{dt} = -\dfrac{\partial H}{\partial q_i}, \quad \dfrac{dq'_i}{dt} = \dfrac{\partial H}{\partial p_i}, \quad i = 1, \dots, n. \end{cases}$$

where $H = H(p, q)$ and $H' = H(p', q')$. We introduce

$$H_2 = H(p, q) + H(p', q') \tag{5}$$

and rewrite the system in the following way

$$\begin{cases} \dfrac{dp_i}{dt} = -\dfrac{\partial H_2}{\partial q'_i}, \quad \dfrac{dq'_i}{dt} = \dfrac{\partial H_2}{\partial p_i}, \quad i = 1, \dots n, \\ \dfrac{dp'_i}{dt} = -\dfrac{\partial H_2}{\partial q_i}, \quad \dfrac{dq_i}{dt} = \dfrac{\partial H_2}{\partial p'_i}, \quad i = 1, \dots n. \end{cases} \tag{6}$$

This combined Hamiltonian allows us to write the doubled system in a compact Hamiltonian form. It is now clear that this system is Hamiltonian, with the same Hamiltonian as in (5). Its symplectic structure is defined by the 1-form

$$\omega = \sum_{i=1}^{n} (p_i \, dq'_i + p'_i \, dq_i).$$

Note the “cross” pairing of coordinates and momenta — each momentum is paired with the other set of coordinates.

The finite-difference scheme obtained by the doubling method reads

$$\begin{cases} \dfrac{\hat{p}_i - p_i}{\Delta t} = -\dfrac{\partial H'}{\partial q'_i}(p', q'), \qquad \dfrac{\hat{q}_i - q_i}{\Delta t} = \dfrac{\partial H'}{\partial p'_i}(p', q'), \quad i = 1, \dots, n, \\ \dfrac{\hat{p}'_i - p'_i}{\Delta t} = -\dfrac{\partial H}{\partial q_i}(\hat{p}, \hat{q}), \qquad \dfrac{\hat{q}'_i - q'_i}{\Delta t} = \dfrac{\partial H}{\partial p_i}(\hat{p}, \hat{q}), \quad i = 1, \dots, n. \end{cases}$$

Or, using (6), the scheme can be written as

$$\begin{cases} \dfrac{\hat{p}_i - p_i}{\Delta t} = -\dfrac{\partial H_2}{\partial q'_i}, \qquad \dfrac{\hat{q}'_i - q'_i}{\Delta t} = \left.\dfrac{\partial H_2}{\partial p_i}\right|_{p=\hat{p},\, q=\hat{q}}, \quad i = 1, \dots, n, \\ \dfrac{\hat{p}'_i - p'_i}{\Delta t} = -\left.\dfrac{\partial H_2}{\partial q_i}\right|_{p=\hat{p},\, q=\hat{q}}, \qquad \dfrac{\hat{q}_i - q_i}{\Delta t} = \dfrac{\partial H_2}{\partial p'_i}, \quad i = 1, \dots, n. \end{cases} \tag{7}$$

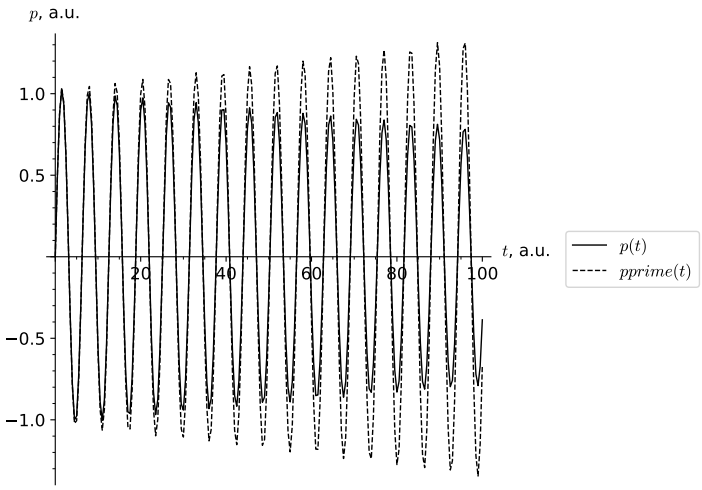

Figure 1. Time evolution of the variables $p(t)$ and $p'(t)$ for the Jacobi oscillator

3. Results

3.1. Jacobi oscillator

To understand the purpose of doubling the variables, consider a simple and well-studied example — the Jacobi oscillator:

$$\begin{cases} \dfrac{dp}{dt} = qr, \\ \dfrac{dq}{dt} = -pr, \\ \dfrac{dr}{dt} = -k^2 pq, \\ p(0) = 0, \\ q(0) = 1, \\ r(0) = 1. \end{cases}$$

It is well known that the solution of this system is periodic and can be expressed in terms of elliptic functions. Figure 1 shows the time dependence of $p(t)$ and $p'(t)$: the amplitude of oscillations of p grows over time, while the amplitude of p' decreases. The exact solution lies somewhere between these two trajectories. The behaviour shown in Fig. 1 closely resembles the classical idea of upper and lower solutions for differential equations, first introduced by Chaplygin [11].

Looking at this figure, one can expect that the average of p and p' approximates the exact solution, while half the absolute difference $\frac{1}{2}|p - p'|$ provides an estimate of the error incurred by the method. The graph of this error estimate is shown in Fig. 2.

Nevertheless, taking arithmetic averages does not produce the behavior typically expected from a geometric integrator. This is particularly clear from the phase portrait in the (p, q)-plane. Figure 3 displays the trajectories of the points (p, q) and (p', q'). As anticipated, (p, q) spirals outward to infinity,

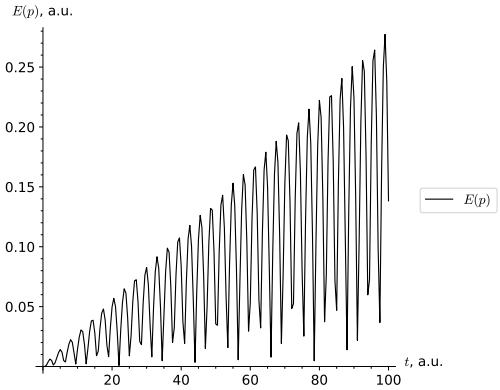

Figure 2. Error function $E(p)$ as a function of time t

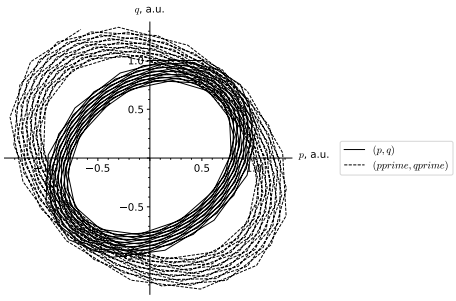

Figure 3. Trajectories of the points (p, q) and (p', q')

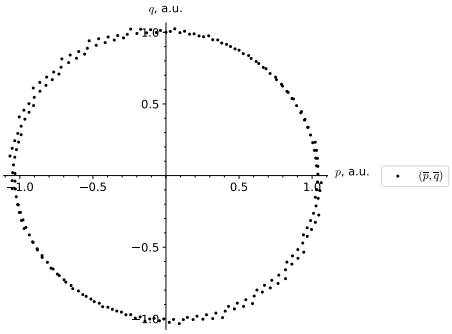

Figure 4. Trajectory of the point $(\overline{p}, \overline{q})$

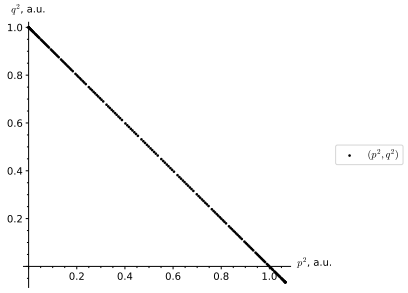

Figure 5. Phase portrait of the Jacobi oscillator in the (q, p)-plane obtained using the geometric mean

whereas (p', q') spirals inward toward the origin. Forming the arithmetic mean — namely, the point

$$(\overline{p}, \overline{q}) = \left(\frac{p + p'}{2}, \frac{q + q'}{2}\right),$$

partially corrects this behavior (see Fig. 4), but the improvement is incomplete. The averaged trajectory still fails to lie on a closed curve, in contrast to the exact periodic orbits or those obtained with Kahan's method. This indicates that the discretization introduces a small amount of numerical dissipation into the system.

However, if we use geometric means in our calculations, the situation improves dramatically. As clearly seen in Fig. 5, the point

$$(\overline{\overline{p}}, \overline{\overline{q}}) = \left(\sqrt{pp'}, \sqrt{qq'}\right)$$

lies exactly on the circle

$$\overline{\overline{p}}^2 + \overline{\overline{q}}^2 = 1,$$

which is precisely the first integral of the Jacobi oscillator,

$$p^2 + q^2 = 1.$$

Thus, in this case, using geometric averages exactly preserves the original system's motion integral.

As an example of the quadratic integral preservation, the Jacobi oscillator possesses the quadratic first integral

$$v = p^2 + q^2.$$

Consequently, the doubled system admits the invariant

$$w = pp' + qq',$$

which is exactly conserved under the numerical scheme (3). When expressed in terms of geometric means,

$$\overline{\overline{p}} = \sqrt{pp'}, \qquad \overline{\overline{q}} = \sqrt{qq'},$$

this invariant takes the form

$$w = \overline{\overline{p}}^2 + \overline{\overline{q}}^2,$$

which explains the exact preservation of the circular phase trajectories observed in Fig. 5.

3.2. Vanhaecke system

To see how integrals of degree higher than two behave under the discrete scheme (3), we look at a well-known example of a completely integrable Hamiltonian system — the Vanhaecke system [12–14], first introduced by P. Vanhaecke in the 1990s. This system has two degrees of freedom and Hamiltonian

$$H = \frac{1}{2}(p_1^2 + p_2^2) + \frac{1}{2}(q_1^2 + q_2^2)^2 + \alpha q_1^2 + \beta q_2^2 - \alpha - \beta,$$

where α and β are parameters of the system. This system has two polynomial first integrals:

$$F = (p_2 q_1 - p_1 q_2)^2 - \left(2q_1^4 + 2q_1^2 q_2^2 + 2\alpha q_1^2 + p_1^2\right)(\alpha - \beta),$$

and

$$G = (p_2 q_1 - p_1 q_2)^2 + \left(2q_1^2 q_2^2 + 2q_2^4 + 2\beta q_2^2 + p_2^2\right)(\alpha - \beta),$$

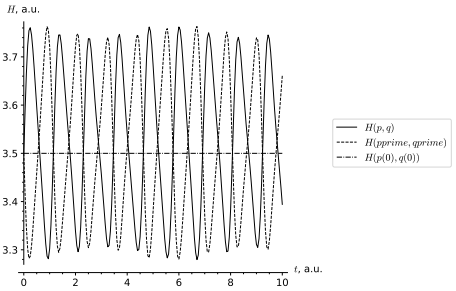

Figure 6. Plots of H and H'

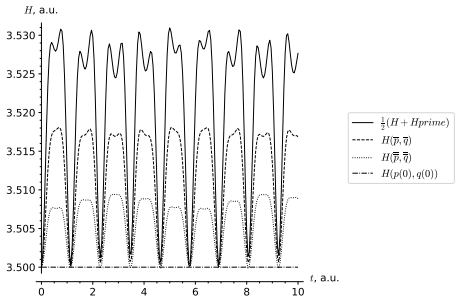

Figure 7. Plot of the averaged Hamiltonian

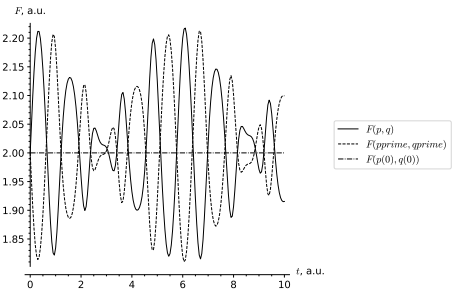

Figure 8. Plots of F and F'

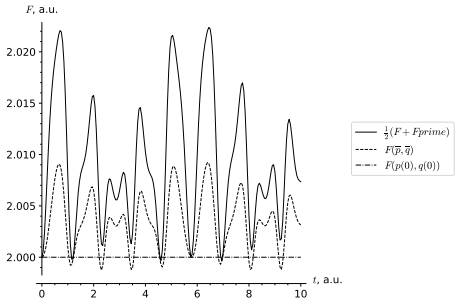

Figure 9. Plots of averaged values of F

which are in involution. Moreover, their difference is proportional to the Hamiltonian H.

We integrated the system using the scheme (3) with the initial conditions

$$\begin{cases} q_1(0) &= 0, \\ q_2(0) &= 1, \\ p_1(0) &= 1, \\ p_2(0) &= 0, \end{cases}$$

parameters $\alpha = 1$, $\beta = 2$, and time step $\Delta t = 0.05$.

Figure 6 shows the values of $H = H(p, q)$ and $H' = H(p', q')$. Even with such a small time step, the energy changes noticeably — already in the first decimal place. However, H and H' oscillate out of phase around a value close to the exact one, and — most importantly — neither shows any long-term drift upward or downward.

Averaging in different ways reduces the size of these oscillations. We tried three approaches:

- the simple average $\frac{1}{2}(H + H')$ — shown as the solid line in Fig. 7,
- $H(\overline{p}, \overline{q})$, where $\overline{p} = \frac{p+p'}{2}$ and $\overline{q} = \frac{q+q'}{2}$ — dashed line,
- $H(\overline{\overline{p}}, \overline{\overline{q}})$, where $\overline{\overline{p}} = \sqrt{pp'}$ and $\overline{\overline{q}} = \sqrt{qq'}$ — doted line.

This keeps the energy with an error of order 10^{-3}, and, most importantly, the energy shows no long-term drift upward or downward.

Figure 8 shows how the integral F behaves over time. When computing the averages (Fig. 9), we used only the first two of the three methods described above. The geometric mean was not used here because the expression for F involves more than just squares of coordinates and momenta, which caused technical difficulties.

The same pattern is clearly visible as in the energy plot: the two copies F and F' oscillate out of phase around the true value, and averaging reduces the oscillation amplitude without introducing any drift.

3.3. Symplecticity

The scheme (7) is a partitioned (or split) finite-difference method [15, 16]. The first $2n$ variables (p, q') and the last $2n$ variables (p', q) are updated using different formulas, which is typical for such methods. Because of this structure, the scheme is expected to be symplectic.

Theorem 11. *If the Hamiltonian splits as*

$$H = T(p) + U(q),$$

then the discrete scheme (7) *is symplectic, i.e.,*

$$\hat{\omega} - \omega = dS$$

for some function S.

Proof. Let us denote the partial derivatives of T and U by

$$T_i = \frac{\partial T}{\partial p_i}, \qquad U_i = \frac{\partial U}{\partial q_i}.$$

Recall that the discrete symplectic form is

$$\hat{\omega} = \sum_{i=1}^{n} \big(\hat{p}_i \, d\hat{q}'_i + \hat{p}'_i \, d\hat{q}_i\big).$$

We treat the two parts separately. First, using the update rule from (7),

$$\hat{q}'_i = q'_i + \Delta t \, T_i(\hat{p}),$$

so

$$\sum \hat{p}_i \, d\hat{q}'_i = \sum \hat{p}_i \, d\big(q'_i + \Delta t \, T_i(\hat{p})\big) = \sum \hat{p}_i \, dq'_i + \Delta t \sum \hat{p}_i \, dT_i(\hat{p}).$$

The second term is an exact differential. Indeed,

$$\sum \hat{p}_i \, dT_i(\hat{p}) = d\Big(\sum \hat{p}_i T_i(\hat{p}) - T(\hat{p})\Big),$$

so we can write

$$\sum \hat{p}_i \, d\hat{q}'_i = \sum \hat{p}_i \, dq'_i + dS_1(\hat{p}),$$

where $S_1(\hat{p}) = \Delta t\big(\sum \hat{p}_i T_i(\hat{p}) - T(\hat{p})\big)$.

Now use the other half of the scheme: $\hat{p}_i = p_i - \Delta t \, U_i(q')$, so

$$\sum \hat{p}_i \, dq'_i = \sum p_i \, dq'_i - \Delta t \sum U_i(q') \, dq'_i = \sum p_i \, dq'_i - dU(q').$$

Putting everything together,

$$\sum \hat{p}_i \, d\hat{q}'_i = \sum p_i \, dq'_i - dU(q') + dS_1(\hat{p}).$$

A similar calculation for the second part gives

$$\sum \hat{p}'_i \, d\hat{q}_i = \sum p'_i \, d\hat{q}_i + dU(\hat{q}) = \sum p'_i \, dq_i + dS_2(p') + dU(\hat{q}),$$

where $S_2(p') = \Delta t\big(\sum p'_i T_i(p') - T(p')\big)$.

Adding both contributions and subtracting the original form

$$\omega = \sum (p_i \, dq'_i + p'_i \, dq_i),$$

we find that all non-exact terms cancel each other out, and

$$\hat{\omega} - \omega = dS$$

with $S = S_1(\hat{p}) + S_2(p') - U(q') + U(\hat{q})$. Hence the scheme is symplectic. □

4. Discussion

In general, it is highly convenient when a numerical method inherently includes a built-in error estimator. In contrast, methods like Runge–Kutta require such error estimation techniques to be added separately [17].

Nevertheless, taking arithmetic averages does not produce the behavior typically expected from a geometric integrator. This is particularly clear from the phase portrait in the (p, q)-plane. Figure 3 displays the trajectories of the points (p, q) and (p', q'). As anticipated, (p, q) spirals outward to infinity, whereas (p', q') spirals inward toward the origin.

Forming the arithmetic mean — namely, the point

$$(\overline{p}, \overline{q}) = \left(\frac{p + p'}{2}, \frac{q + q'}{2} \right),$$

partially corrects this behavior (see Fig. 4), but the improvement is incomplete. The averaged trajectory still fails to lie on a closed curve, in contrast to the exact periodic orbits or those obtained with Kahan's method. This indicates that the discretization introduces a small amount of numerical dissipation into the system.

However, if we use geometric means in our calculations, the situation improves dramatically. As clearly seen in Fig. 5, the point

$$(\overline{\overline{p}}, \overline{\overline{q}}) = \left(\sqrt{p p'}, \sqrt{q q'} \right)$$

lies exactly on the circle

$$\overline{\overline{p}}^{\,2} + \overline{\overline{q}}^{\,2} = 1,$$

which is precisely the first integral of the Jacobi oscillator,

$$p^2 + q^2 = 1.$$

Thus, in this case, using geometric averages exactly preserves the original system's motion integral.

In this sense, scheme (3) actually works better than Kahan's method. Unlike the doubling scheme, Kahan's method does not preserve the original integral $p^2 + q^2$. Instead, it conserves a modified expression dependent on Δt. Finding this expression for the Kahan scheme was the main result of the work [18], in [5] this expression was found according to the Lagutinsky method.

Exact preservation is important in many cases. For example, one conservation law for the spinning top simply says that the sum of the squares of the direction cosines is always 1 [19]. Kahan's method preserves a modified expression [20], i.e., adds tiny corrections that don't have a clear geometric meaning. Because scheme (3) preserves this integral exactly, it looks very promising for building geometric integrators — methods that respect the underlying geometry of the problem.

Overall, we can say that the scheme (3) keeps all integrals of the system close to their exact values allowing only small oscillations around them. This is a very good property and suggests that the scheme may be reliable for long-time simulations. At the same time, it is important to note that the energy integral is not preserved exactly, no matter how we average it.

It should also be mentioned that the Vanhaecke system has cubic right-hand sides, so it cannot be directly discretized by Kahan's method. Even if it could, there are no general results guaranteeing that Kahan's method preserves arbitrary polynomial integrals. The only case that is well understood is the preservation of the energy (Hamiltonian) itself [2].

Energy conservation in symplectic integrators has been studied extensively. It is well known that the original Hamiltonian is not preserved exactly. However, one can construct a modified Hamiltonian $H^{(m)}$ that is conserved up to any desired order of accuracy [15, 21]. Because of this, numerical experiments typically show the energy oscillating around a constant value that is close to — but not exactly equal to — the true energy. This is precisely what we observed in the Vanhaecke system. Looking at Fig. 7, one might even guess that the different averaging strategies we used correspond to modified Hamiltonians of different orders.

5. Conclusion

In this work, we studied reversible finite-difference schemes for dynamical systems based on the doubling method introduced by V. N. Abrashin and S. N. Sytova. We carried out a detailed analysis of algebraic integral preservation for two benchmark systems: the classical Jacobi oscillator and the Vanhaecke system.

First of all, quadratic integrals are proven to preserve exactly (Th. 9). In particular, the quadratic integrals for the Jacobi oscillator are preserved exactly if we replace the squares of the quantities with the geometric mean. For comparison's sake, Kahan's method does not preserve the original quadratic integral $p^2 + q^2$, but conserves a modified expression dependent on Δt. Exact preservation is important for geometrical interpretation of expressions like the sum of the squares of the direction cosines.

For Hamiltonian systems, this method is symplectic in the sense of the theorem 11. Thus the original Hamiltonian is not preserved exactly, but one can construct a modified Hamiltonian $H^{(m)}$ that is conserved up to any desired order. It seems that the different averaging strategies we used correspond to modified Hamiltonians of different orders. In any way, the scheme keeps all integrals of the Vanhaecke system close to their exact values allowing only small oscillations around them. Thus the scheme may be reliable for long-time simulations.

Author Contributions: Conceptualization, M.Malykh; methodology, M.Malykh, L.Lapshenkova; software, M.Malykh, L.Lapshenkova; validation, A.Trusova, K.Mashkovtseva; writing—original draft preparation, L.Lapshenkova; writing—review and editing, M.Malykh. All authors have read and agreed to the published version of the manuscript.

Funding: This research received no external funding.

Data Availability Statement: Data sharing is not applicable.

Acknowledgments: Non profit acknowledgments.

Conflicts of Interest: The authors declare no conflict of interest.

Declaration on Generative AI: The author has not employed any generative AI tools.

References

1. Petrera, M. & Suris, Y. B. On the Hamiltonian structure of Hirota-Kimura discretization of the Euler top. *Math. Nachr.* **283,** 1654–1663. doi:10.1002/mana.200711162 (2010).
2. Celledoni, E., McLachlan, R. I., Owren, B. & Quispel, G. R. W. Geometric properties of Kahan's method. *J. Phys. A: Math. Theor.* **46,** 025201. doi:10.1088/1751-8113/46/2/025201 (2013).
3. Petrera, M., Pfadler, A., Suris, Y. B. & Fedorov, Y. N. On the Construction of Elliptic Solutions of Integrable Birational Maps. *Experimental Mathematics* **26,** 324–341. doi:10.1080/10586458.2016.1166354 (2017).
4. Petrera, M., Smirin, J. & Suris, Y. B. Geometry of the Kahan discretizations of planar quadratic Hamiltonian systems. *Proc. R. Soc. A* **475,** 20180761. doi:10.1098/rspa.2018.0761 (2019).
5. Malykh, M., Gambaryan, M., Kroytor, O. & Zorin, A. Finite Difference Models of Dynamical Systems with Quadratic Right-Hand Side. *Mathematics* **12,** 167. doi:10.3390/math12010167 (2024).
6. Malykh, M., Ayryan, E., Lapshenkova, L. & Sevastianov, L. Difference Schemes for Differential Equations with a Polynomial Right-Hand Side, Defining Birational Correspondences. *Mathematics* **12,** 2725. doi:10.3390/math12172725 (2024).
7. Sytova, S. N. Finite-Difference Methods in Problems Modeling Volume Free Electron Lasers. *Differ. Equ.* **37,** 1026–1031 (2001).
8. Goriely, A. *Integrability and nonintegrability of dynamical systems* (World Scientific, 2001).
9. Gorbuzov, V. N. *Integrals of differential systems* In Russian (Grodno State University, Grodno, 2006).
10. Pranevich, A. F. On Poisson's Theorem of Building First Integrals for Ordinary Differential Systems. *Rus. J. Nonlin. Dyn.* **15,** 87–96 (2019).
11. Nefedov, N. N. *Differential equations – Lectures* Russian. URL: https://teach-in.ru/file/synopsis/pdf/differential-equation-M1.pdf (date: 14.01.2025). 2019.
12. Vanhaecke, P. A special case of the Garnier system, $(1, 4)$-polarized abelian surfaces and their moduli. *Compositio Mathematica* **92,** 157–203 (1994).
13. Vanhaecke, P. *Integrable Systems in the Realm of Algebraic Geometry* 2nd (Springer, 2001).
14. Shiwei, W., Malykh, M. D., Sevastianov, L. A. & Zorin, A. V. On the behavior of orbits of Vanhaecke system on integral surfaces. *Discrete and Continuous Models and Applied Computational Science* **34,** xx–xx (2026).
15. Sanz-Serna, J. M. & Calvo, M. P. *Numerical Hamiltonian Problems* (CHAPMAN & HALL, London, Glasgow, New York, Tokyo, Melbourne, Madras, 1994).
16. Gevorkyan, M. N. Specific implementations of symplectic numerical methods. *Bulletin of the Peoples Friendship University of Russia. Series: Mathematics informatics physics.,* 77–89 (2013).
17. Sarafyan, D., Outlaw, C. & Derr, L. An investigation of Runge-Kutta processes, and equivalence of scalar and vector cases. *Journal of Mathematical Analysis and Applications* **104,** 568–588. doi:10.1016/0022-247X(84)90021-0 (1984).
18. Hirota, R. & Kimura, K. Discretization of the Euler Top. *Journal of the Physical Society of Japan* **69,** 627–630 (2000).
19. Golubev, V. V. *Lectures on integration of the equations of motion of a rigid body about a fixed point* (Israel Program for Scientific Translations, Jerusalem, 1960).

20. Hirota, R. & Kimura, K. Discretization of the Lagrange Top. *Journal of the Physical Society of Japan* **69,** 3193–3199 (2000).
21. Malykh, M. D. & Konyaeva, M. A. Calculation of modified Hamiltonian in Sage. *Discrete and Continuous Models and Applied Computational Science* **34,** xx–xx (2026).

Information about the authors

Lapshenkova, Lyubov O.—Assistant of Department of Computational Mathematics and Artificial Intelligence of RUDN University (e-mail: lapshenkova-lo@rudn.ru, ORCID: 0000-0002-1053-4925, ResearcherID: OXC-7984-2025)

Mashkovtseva, Kseniia S.—Student of Department of Computational Mathematics and Artificial Intelligence of RUDN University (e-mail: 1132226438@rudn.ru, ORCID: 0009-0002-6927-4467)

Trusova, Alina A.—Student of Department of Computational Mathematics and Artificial Intelligence of RUDN University (e-mail: 1132246715@rudn.ru, ORCID: 0009-0008-5140-5629)

Malykh, Mikhail D.—DSc., Head of Department of Computational Mathematics and Artificial Intelligence of RUDN University; Senior Researcher of Joint Institute for Nuclear Research (e-mail: malykh-md@rudn.ru, ORCID: 0000-0001-6541-6603, ResearcherID: P-8123-20168, Scopus Author ID: 6602318510)

УДК 519.872, 519.217
PACS 07.05.Tp, 02.70.Bf
DOI: 10.22363/2658-4670-2026-34-1-98-112 EDN: URTKJP

Об одной разностной схеме, задающей бирациональное, но не квадратичное соответствие между слоями

Л. О. Лапшенкова[1], К. С. Машковцева[1], А. А. Трусова[1], М. Д. Малых[1, 2]

[1] Российский университет дружбы народов, ул. Миклухо-Маклая, д. 6, Москва, 117198, Российская Федерация

[2] Объединённый институт ядерных исследований, ул. Жолио-Кюри, д. 6, Дубна, 141980, Российская Федерация

Аннотация. В статье рассматриваются обратимые разностные схемы для динамических систем, основанные на методе удвоения системы, предложенном В. Н. Абрашиным и С. Н. Сытовой. Метод заключается в дублировании исходного набора переменных, что позволяет перейти к расширенной системе, для которой строится конечно-разностная аппроксимация, задающая бирациональное отображение между соседними временными слоями. Исследуется сохранение алгебраических интегралов таких схем. Доказывается, что если исходная система допускает однородный квадратичный первый интеграл, то соответствующая билинейная форма является точным интегралом дискретной схемы. Это свойство демонстрируется на классическом примере осциллятора Якоби, где схема сохраняет точную величину, выраженную через среднее геометрическое дублированных переменных, воспроизводя корректную геометрию фазовых траекторий. Более глубокий анализ проводится на примере нетривиальной системы Ванхаеке — интегрируемой гамильтоновой системы с двумя степенями свободы, обладающей полиномиальными интегралами высших степеней, интегрируемость которой выражается через абелевы функции. Численные эксперименты, реализованные в системе компьютерной алгебры Sage с использованием специализированного пакета fdm.sage, подтверждают, что при дискретизации методом удвоения две копии системы синхронно колеблются около точных значений первых интегралов, а применение усреднения снижает амплитуду колебаний. Для сепарабельных гамильтоновых систем показана симплектичность схемы. Полученные результаты позволяют рекомендовать метод удвоения для построения устойчивых и структуросохраняющих численных интеграторов для широкого класса динамических систем с полиномиальными правыми частями, включая системы высокой размерности.

Ключевые слова: динамические системы, конечные разности, схема Кагана, интегрируемые системы, система Ванхаеке

Discrete & Continuous Models & Applied Computational Science
ISSN 2658-7149 (Online), 2658-4670 (Print)

2026, 34 (1) 113–124
http://journals.rudn.ru/miph

Review

UDC 537.533.7
PACS 52.25.Os, 52.30.–q, 52.35.–g, 52.38.–r, 52.40.Db, 52.50.Jm, 52.57.–z
DOI: 10.22363/2658-4670-2026-34-1-113-124 EDN: UPXGCS

Interaction of relativistic electrons with intense electromagnetic fields: ponderomotive effect, acceleration, refraction, reflection, dependence on initial conditions

Alejandro J. Castillo[1, 2, 3], Yuriy Gr. Rudoy[1]

[1] RUDN University, 6 Miklukho-Maklaya St, Moscow, 117198, Russian Federation

[2] P. N. Lebedev Physical Institute of the Russian Academy of Sciences, 53 Leninskiy Prosp, Moscow, 119991, Russian Federation

[3] N. I. Pirogov Russian National Research Medical University, 1 Ostrovityanova St, Moscow, 117513, Russian Federation

(received: July 17, 2025; revised: December 25, 2025; accepted: January 10, 2026)

Abstract. The rigorous theory and characterization of charged-particle dynamics in high-intensity electromagnetic fields are fundamental for the development of advanced plasma-based applications. Accurate analytical models must bridge the gap between smoothed trajectories and exact particle motion to predetermine injection and energy gain. The main objective of this review is to establish a rigorous framework for the averaged relativistic motion of electrons, focusing on the strict derivation of ponderomotive forces and the impact of fast-oscillating periodic additions on dynamical variables. By making use of the Krylov–Bogoliubov–Mitropolsky averaging method to obtain the equations of motion, the study analyzes relativistic effects in laser beams and waveguides. These theoretical predictions are substantiated through numerical validation, including test-particle simulations and three-dimensional particle-in-cell simulations (PIC) of relativistic self-trapping regimes such as "laser bullet" and "bubble" structures. The review details the independence of the results on the formulation framework, the strict dependence on wave polarization, and the non-strict potential character of the relativistic ponderomotive force. The analysis demonstrates that periodic fast-oscillating additions are essential for a complete description, accurately setting initial conditions in averaged equations and enabling precise predictions of electron reflection and refraction. Simulations confirm that these fast-oscillating corrections determine electron injection and beam charge in realistic laser–plasma acceleration scenarios. The present review clearly shows that the dual framework of test-particle and PIC models is vital for probing the limits of averaged motion theory. The findings are of direct practical relevance for the optimization of radiation sources and guide the development of future theories incorporating non-adiabatic and field topology dependent effects.

Key words and phrases: averaged motion, relativistic ponderomotive forces, laser radiation, Gaussian beam, waveguides, beat wave

For citation: Castillo, A. J., Rudoy, Y. G. Interaction of relativistic electrons with intense electromagnetic fields: ponderomotive effect, acceleration, refraction, reflection, dependence on initial conditions. *Discrete and Continuous Models and Applied Computational Science* **34** (1), 113–124. doi: 10.22363/2658-4670-2026-34-1-113-124. edn: UPXGCS (2026).

1. Introduction

The dynamics of charged particles in intense electromagnetic (EM) fields is a fundamental area of research with significant implications for fields such as laser–plasma interaction, particle acceleration [1], and plasma heating [2]. The complexity of particle motion increases significantly when the EM field is inhomogeneous in space and time [3] or the relativistic effects become considerable [4]. To address this, various approximate methods, particularly averaging techniques [5], have been developed to derive simplified equations of motion for the "guiding center" or "smoothed" variables of the particle, while accounting for the "fast oscillations" or "periodic additions", as mentioned in [6–9], is treated as an intrinsic intermediate procedure.

A central concept derived from the averaging analysis is the ponderomotive force, which describes the averaged driving of a high-frequency EM field on a charged particle [10–13]. Initially derived by Gaponov and Miller [10] for nonrelativistic particles in a weak monochromatic field, this force is potential and points towards regions of lower field intensity [11, 14]. Its relativistic generalization, initially presented by Kibble [15], introduced the concept of an "effective mass" of the particle, $m^* = m\sqrt{1 + A^2}$, where $A^2 = (e/mc\omega)^2\langle E^2\rangle$ is the normalized vector potential, redefined as a function of the averaged (slow) coordinates. This initial relativistic treatment also introduced the concept of mutual refraction of electrons and photons, an appealing name that captures the physical analogy of an electron's deflection in a wave field mirroring light refraction in plasmas. This powerful, physical similarity offers considerable insight into the problem, much of which has unfortunately vanished from modern discourse.

Further analyses of applicability were conducted, yielding various generalizations, particularly considering the influence of the relativistic nature of particle motion [16, 17], the superposition of multiple waves [18], and external magnetic fields. Such diversity of scenarios, field configurations, and formalisms employed leads to a wide range of significant results. Notably, the average wave-particle interaction can depend on propagation direction, polarization, spatiotemporal amplitude profile, and wave intensity, and it may not necessarily be potential. However, the absence of a unified rigorous approach to these problems often results in contradictions across studies [13, 19, 20]. Furthermore, while averaging particle motion simplifies the description of dynamics in non-uniform, rapidly oscillating fields, the rapidly oscillating components of dynamic variables, along with their initial conditions, are frequently overlooked. This omission prevents a complete and unambiguous understanding of particle motion.

The collective works [6–9] offer a rigorous and consistent application of the Krylov–Bogoliubov–Mitropolsky (KBM) [21] averaging over fast phase [22] method to analyze the relativistic motion of charged particles in a variety of intense and inhomogeneous EM fields. These studies go beyond typical derivations of ponderomotive forces by meticulously calculating oscillating additions and highlighting their crucial role, particularly in defining initial conditions for averaged equations. They refine theoretical models and uncover previously overlooked complexities in particle-wave interactions [2]. In this review, we assess the results, validity, and significance of these researches. Our goal is to provide a coherent framework that connects foundational principles with modern applications.

2. Relativistic averaged motion theory: model and methods

Classical averaging techniques, mainly applied to finding the ponderomotive force, often prove insufficient when examining the motion of charged particles in relativistic regimes. Early attempts introduced more rigorous derivations using multiple-scale perturbation theory [12], Hamiltonian

[23] and Lagrangian [24] averaging, and covariant formulations [17, 25] . These efforts primarily revealed that the relativistic ponderomotive force is not only amplitude-dependent but also sensitive to polarization and space–time structure of the EM field. The KBM averaging over the fast phase theory suggests that a correct averaging procedure first requires establishing a standard form of the equation of motion. This allows one to prove the existence of a specific change of variables that excludes time from the right-hand sides of these equations with a prescribed degree of accuracy in terms of a small parameter [26]. Here, we present a synthesis of results from theoretical manuscripts [6–9], which make use of the KBM method. These works consider a full coupling between field geometry [27], relativistic particle momentum [28], and relativistic kinematics [29], while scrutinizing the meaning of the fast oscillating terms explicitly obtained by the KBM method, thus including their influence, especially in the setting of initial conditions.

2.1. Polarization effects

The paper [6] extensively examines the relativistic motion of a charged particle in intense linearly and circularly polarized EM radiation within the geometrical optics approximation [30]. Whereas it is usually accepted that ponderomotive forces are independent of the polarization of the wave [31], research as [13] has shown that the averaged equations of motion of the particle (and consequently the expressions for the ponderomotive force) are different for circularly and linearly polarized waves. A key aspect of the work [6] is the consistent derivation of averaged relativistic equations of motion, confirming that the expressions for the ponderomotive force differ for circularly and linearly polarized waves. This directly addresses and clarifies previous contradictions between [13] and [31].

The analysis introduces two dimensionless parameters: $g = eE/mc\omega$, representing the ratio of the particle's oscillating velocity amplitude to the speed of light, and μ, associated with the space-time variations of the amplitude. For intense radiation fields, g can be comparable to or greater than unity (e.g., $g \approx 1$ for a wavelength of 1 μm and intensity of 10^{18} W/cm^2). This means that g is not generally small, making averaging via this parameter unfeasible according to the Bogoliubov-Krylov theorem.Thus averaging must be performed by expanding in the small parameter μ, which is small in the geometrical optics approximation, $\mu \approx 1/kL \approx 1/\omega T$. Let us note that all references to the parameter g in cited works [6–9] correspond to the amplitude of the normalized vector potential in the averaged description and should be identified with the standard symbol a_0 for consistency [32].

The derived expressions for the averaged relativistic force for both polarizations contain new additional small terms weakening its module. While these terms are small, their effect can be noticeable at small radiation field gradients. A critical distinction arises for linearly polarized waves: rapidly oscillating terms in the relativistic factor, specifically those oscillating with a doubled phase whose amplitude depends only on wave intensity, are not associated with the small expansion parameter. This implies that for a linearly polarized wave, the field cannot be excessively strong, as the usual binomial expansion for the relativistic factor might become invalid, complicating the averaging procedure significantly. The averaged action of a linearly polarized wave on a particle is shown to be more weakened than that of a circularly polarized wave.

2.2. The impact of spatiotemporal wave structure

The manuscript [7] focuses on ponderomotive forces in intense laser radiation fields described as the superposition of Gaussian beams of arbitrary modes in the quasi-optical approximation. This approximation is often considered more adequate for laser radiation [33] and provides a physically realistic description of the field's spatial structure analysis [30]. The small parameter here is $\mu = 2/ka$, where a is the beam waist. The ponderomotive force for circularly polarized Gaussian beams is

derived. The method involves representing the laser field via a vector potential $\boldsymbol{A}(\boldsymbol{r}, t)$ and expanding its complex amplitude $A_0(r)$ in even powers of μ. The longitudinal component of the vector potential, $A_{zm}^{(1)}$, emerges in odd powers of μ from the Coulomb gauge condition $\nabla \boldsymbol{A} = 0$. Similar to the previous work [6], a generalized momentum vector, $\boldsymbol{\pi} = \boldsymbol{p} + (e/c)\boldsymbol{A}$ is employed to handle large, fast-oscillating terms. A significant finding is that relativistic effects lead to a weakening (attenuation) of the averaged force of high-power laser radiation on the particle. The ponderomotive force is proportional to the gradients of the Gaussian radiation intensity, aligning with experimental data [34]. The difficulties in averaging the longitudinal motion equations, particularly due to large rapidly oscillating components proportional to the wave field, are addressed by assuming weakly relativistic motion in the transverse plane, which is often a natural assumption for acceleration problems. The mean energy of the particle-radiation system is conserved to second-order terms when the amplitude is time-independent. Furthermore, particle acceleration or deceleration is shown to depend on its injection into a divergent or convergent Gaussian beam, respectively.

2.3. Beat waves approach to modulated wavefronts

The study in [8] extends the analysis to the relativistic motion of a charged particle in the field of a laser beat wave, formed by the superposition of two circularly polarized Gaussian beams in the fundamental mode propagating in the same direction. This is crucial for understanding mechanisms of particle acceleration [35, 36] and plasma heating [37]. A unique and significant result in this context is the demonstration that, although a relativistic generalization of the ponderomotive potential is defined, the averaged force in the field of a beat wave is not completely potential. This contrasts sharply with the potential nature of the standard form of the Gaponov-Miller force. The force is found to depend significantly on the slowly varying combination (beat wave) phase, which evolves according to its own nonlinear equation. This implies a more complex interaction than for single-wave fields. The averaging procedure here has a distinct feature: while partial phases of the constituent waves are considered “fast” and averaged over, the combination phase is “slow” and is not averaged. Relativistic effects and the diffractive spreading of the beams further weaken the averaged action on the particle. It is also shown that the transverse components of the ponderomotive force are first-order effects in their expansion, while the longitudinal component is a second-order effect. This implies that ponderomotive expulsion of particles in the radial direction (towards weaker fields) occurs faster than acceleration in the direction of wave propagation. The particle’s trajectory in the transverse plane can be approximated as a circle whose radius slowly decreases as the beat wave propagates, unlike the constant radius for a plane wave.

2.4. Oscillating additions and initial conditions

Reference [9] delves into the relativistic motion of a single electron entering a rectangular waveguide supporting an arbitrary H_{mn}-mode wave. Here, the small parameter for expansion is $g = eE/m_e c\omega$, which is typically small for waveguide fields. A salient finding is that the averaged (ponderomotive) force along the longitudinal axis of the waveguide is absent, regardless of the wave mode, meaning that no non-gradient forces are generated in this direction. As a standard effect, the transverse components of the ponderomotive force expel the charged particle from regions of high field intensity. The constant of motion, $\gamma - p_z/mc = C$, which behaves analogously to the refractive index for plane waves in a dielectric medium, is confirmed. Owing to the influence of the explicitly presented periodic additions terms, the analytical and numerical results from [9] are crucial: together, these results address and reconcile the contradiction between [19] and [38].

The crucial role of oscillating additions and Initial Conditions is studied deeper in this work. Across all the studies [6–9], a consistent and highly emphasized theme is the meticulous calculation and application of "periodic additions" to the smoothed (averaged) dynamical variables. While previous works often used these oscillating parts only for deriving averaged equations and then disregarded them, the aforesaid works rigorously derive the terms usually up to second-order expansions. The numerical simulations repeatedly demonstrate that an excellent agreement between the exact equations of motion and the averaged solutions is achieved only when the initial conditions for the averaged equations are correctly defined by incorporating the periodic additions across the entire time evolution, including the initial moment. For the initial moment onward This "initial leap" between the exact and averaged momenta, determined by the periodic additions, is critical. For example, in the waveguide case, the longitudinal averaged momentum at the initial instant may differ from zero and even be negative, which is correctly predicted by their model and verified numerically. This detailed treatment of initial conditions allows for an accurate description of phenomena like electron refraction and reflection by the waveguide field, consistent with Kibble's [15] earlier work on mutual refraction of electrons and photons. As remarked and shown in [9] and collectively supported by [6–9], depending on injection conditions and initial phase, an electron may either penetrate or be reflected by the waveguide field, with the critical momentum being defined by the periodic additions.

2.5. Methodological constraints and physical limitations

A rigorous application of the Krylov–Bogoliubov–Mitropolskiy (KBM) averaging method to the relativistic equations of motion reveals fundamental constraints on the resulting analytical models for the ponderomotive forces. These constraints establish the foundational bounds within which the KBM-derived formalism remains both mathematically correct and physically adequate. A primary limitation arises because the Lorentz force equation in an intense field ($a_0 = eE/mc\omega \sim 1$) is not initially in the "standard form" required by the KBM formalism, which presupposes a clear separation between slow dynamics and small, fast oscillations. The right-hand side of the equation of motion contains large-amplitude terms, proportional to the field strength a_0, that oscillate at the optical frequency. Direct averaging is therefore impossible [6]. As demonstrated in the works [8], is a preliminary transformation to the particle's canonical momentum, $\boldsymbol{\pi} = \mathbf{p} + (e/c)\mathbf{A}$ automatically absorbs and eliminates the dominant oscillatory force terms. However, this necessary step complicates the subsequent averaging of the relativistic factor $\gamma = \sqrt{1 + (\mathbf{p}/mc)^2}$. Expressed in terms of the canonical momentum $\boldsymbol{\pi}$ and the field momentum $\mathbf{p}_E = (e/c)\mathbf{A}$, γ becomes depending on the $\boldsymbol{\pi} - \mathbf{p}_E$. A consistent KBM expansion of γ is only straightforward if the oscillatory part $\mathbf{p}_E$ is the dominant momentum scale. This leads to a critical, often implicit, assumption: the smoothed transverse canonical momentum must satisfy $|\boldsymbol{\pi}_\perp| < |\mathbf{p}_E|$. When this condition holds, a binomial expansion of γ in powers of $(\boldsymbol{\pi} \cdot \mathbf{p}_E)$ is justified and allows for systematic averaging [6, 7]. Violation of this condition signals a regime where the particle's quiver motion is no longer the primary relativistic effect, and the standard ponderomotive expansion fails.

A separate and stringent limitation concerns the longitudinal motion. For the wave phase $\theta = kz - \omega t$ to be a "fast" variable suitable for averaging, its derivative $d\theta/dt = -\omega\,(1 - v_z/c)$ must remain large. This requires that the quantity $G = \gamma - p_z/mc = \gamma(1 - v_z/c)$ is not too small [6]. Physically, this condition $|1 - v_z/c| \gg \mu$ (where μ is the slow-variation parameter) ensures that the Doppler-shifted frequency experienced by the particle remains high [39]. As the particle's longitudinal velocity approaches c, this condition breaks down; the phase evolution becomes slow, the separation of time scales vanishes, and the averaging procedure is invalidated.

Finally, accounting for the finite extent in time of the EM fields is essential, particularly in laser EM fields, where attaining higher radiation intensities unavoidably entails shorter pulse durations. We note that the entire KBM approach and the adopted aprroximations(e.g. quasioptical or geometrical) rest on adiabatic assumptions, and consequently the laser pulse envelope (characterized by scales L, T, or waist a) must vary slowly compared to the optical period ($\mu \ll 1$). This assumption underpins the definition of the small expansion parameter $\mu = 1/(kL) \approx 2/(ka)$ [31]. For few-cycle or sub-cycle laser pulses, this clear scale separation collapses. In such ultra-short pulse regimes, the particle dynamics is intrinsically and strongly phase-dependent, the concept of a time-averaged ponderomotive force becomes ill-defined, and analysis must revert to fully time-resolved models or direct numerical integration of the Lorentz equations.

Let us remark that within the regime of laser intensity and particle energy considered in the articles which are the scope of this work, radiation reaction and quantum electrodynamical (QED) effects are negligible [40]. Collective plasma behavior is treated qualitatively within a phenomenological framework.

3. Numerical and experimental verification

The theory of averaged relativistic motion for single charged particles, developed in [6–9], requires systematic benchmarking against both numerical simulations and experimental data. In particular, the works [1, 41] have been instrumental — not only in validating the foundational framework — but also in extending it to more complex, less idealized scenarios of laser–plasma interaction. The theoretical predictions derived therein show strong agreement with both single-particle simulations and full particle-in-cell (PIC) results for canonical configurations, such as Gaussian laser pulses interacting with preformed or self-generated plasma channels, as well as in the relativistic self-trapping regime. Key validation metrics include trajectory fidelity over multiple laser cycles, long-term energy gain or loss, the spatial distribution of electrons expelled by the laser field, and the subsequent dynamics of these particles under quasi-static fields.

However it is crucial to recognize the boundaries of this averaged description. Beside the limitations described in section the formalism based in KBM method breaks down in stochastic regimes. Specifically, when stochasticity develops, for instance in complex field configurations where a laser field is assisted by large-amplitude plasma waves [42] the foundational KBM averaging method loses its strict applicability [43]. In such cases, the concept of a well-defined ponderomotive force is partially employed, serving primarily for qualitative estimates or as a guiding approximation and the particle dynamics must be analyzed using tools for chaotic systems, such as Lyapunov exponents, Poincaré maps, and other indicators of non-integrability and phase-space mixing.

Experimental [14, 44, 45] benchmarking remains more challenging due to the difficulty of isolating pure ponderomotive effects from competing processes such as collisional heating, space-charge fields, and instabilities. Nevertheless, combined diagnostics can enable semi-quantitative validation of predicted density modulations and EM field structures consistent with relativistic ponderomotive theory [46, 47].

3.1. Test-particle models

For controlled investigations of single-particle dynamics, test-particle models offer a complementary and highly flexible approach. In these models, prescribed EM fields — such as Gaussian laser beams, beat waves, or idealized waveguide modes — are used to integrate the relativistic equations of motion for ensembles of particles. This setup allows direct comparison with analytical expressions for the

ponderomotive force and enables the isolation of effects due to field geometry [44], polarization [48], and carrier-envelope phase [49]. Future studies will employ hybrid approaches that combine test-particle trajectories with envelope-averaged field models, thereby bridging the gap between rapid quiver motion and slow drift dynamics.

Accurate modeling of the ponderomotive force, augmented by information from the periodic additions to the smoothed motion, has proven essential for predicting injection thresholds, beam quality, and energy spread. These parameters are critical for applications ranging from compact accelerators and bright radiation sources to medical therapies.

A related study of direct laser electron acceleration in plasma channels formed by ultrashort, relativistically intense pulses, both linearly and circularly polarized, examines post-injection electron dynamics governed by self-generated quasi-static fields [41]. Using test-particle simulations, this work incorporates the radial electric field, azimuthal magnetic field, and, in the case of circular polarization — an axial quasi-static magnetic field component, the latter having been first identified in PIC simulations. Through consistent application and numerical testing of the KBM averaging method, this research reveals distinct mechanisms of drift, diffusion, and acceleration, alongside detailed analyses of trajectory stability and chaotic motion.

3.2. Particle-in-Cell (PIC) simulations

Relativistic ponderomotive forces lie at the heart of modern laser-driven plasma acceleration schemes. The leading edge of an intense laser pulse expels background electrons via the ponderomotive force, creating a trailing, charge-separated cavity that sustains enormous longitudinal electric fields [50]. Particle-in-cell (PIC) simulations serve as the primary computational tool for self-consistently modeling these interactions at relativistic intensities; crucially, they capture ponderomotive effects naturally by resolving the fast oscillatory motion of particles without invoking explicit cycle averaging.

Among mechanisms associated with the formation of an ion cloud in material that moves together with the laser pulse by the effects of ponderomotive force, Relativistic self-trapping (RST) of an intense laser pulse represents one of the most efficient mechanisms for laser-driven electron acceleration, delivering extreme charge ($>$ 10 nC) with high energy conversion efficiency ($\sim$ 40–50 %). This mechanism operates in two characteristic regimes: the "laser bullet", where the pulse length $c\tau$ is comparable to the cavity diameter D ($c\tau \sim D$), and the "bubble" regime, where $c\tau \ll D$. The distinction depends on the fraction of the cavity volume filled by the laser field. A quantitative comparison of these regimes, supported by recent work [1], relies on a physical interpretation grounded in the averaged dynamics of test particles. This approach has clarified distinct electron injection mechanisms and enabled predictive scaling laws for injection efficiency.

The equilibrium size of the plasma channel (or cavity) is governed by a balance between the outward radial ponderomotive force and the inward Coulomb force of the ion core at the channel boundary [51]. The standard scaling law for the channel radius, $R \propto I_0^{1/4}$, emerges from this force balance. However, its derivation traditionally assumes a uniform transverse laser profile, where the dimensionless field amplitude a_0 is constant. In a more realistic, non-uniform transverse profile, the field amplitude at the boundary, $a_0^{(\mathrm{b})}$, is lower than the peak axial value, $a_0^{(\mathrm{axis})}$. Consequently, the proportionality constant α in the scaling law must be adjusted when the boundary field is used, resolving an ambiguity present in the literature [52].

Although full-scale particle-in-cell (PIC) simulations naturally capture collective, non-adiabatic, and stochastic effects in ultra-short laser–plasma interactions, the underlying ponderomotive dynamics can still be identified through tailored diagnostic procedures. In the analysis of [1], signatures of the averaged ponderomotive force will be extracted by examining phase-averaged particle momenta, correlating particle expulsion with local intensity gradients, and reconstructing effective force fields

from simulation data. These methods allow for the assessment of the relative contribution of ponderomotive mechanisms even in strongly nonlinear regimes, providing a self-consistent bridge between single-particle averaged theory and collective plasma behavior.

It should be noted that high-resolution simulations of modern relativistic laser–plasma interactions often uncover phenomena that lie beyond conventional averaged descriptions, such as phase trapping, stochastic heating in chaotic field regions, and transient momentum kicks during the pulse rise time. These effects underscore the limitations of adiabatic approximations and motivate the development of refined, non-perturbative theories of relativistic ponderomotive dynamics [51].

4. Conclusion

The body of work exemplified by the articles [6–9] provides a comprehensive and rigorous framework for understanding the relativistic motion of charged particles in intense and inhomogeneous EM fields. Through the consistent application of the KBM averaging method, these studies yield refined expressions for the ponderomotive force, revealing its dependence on wave polarization, its non-potential character in beat-wave fields, and its complete absence along the propagation axis in certain waveguide modes.

Crucially, these works highlight the often-overlooked importance of the fast-oscillating "periodic additions" to the smoothed dynamical variables. It has been both analytically derived and numerically validated that these terms are essential for accurately setting initial conditions in the averaged equations, thereby bridging the gap between exact and approximate solutions and enabling precise predictions of phenomena such as electron reflection and refraction.

Collectively, these contributions significantly advance the theoretical understanding of charged-particle dynamics in strong fields, offering robust analytical tools for analyzing complex interactions relevant to high-intensity laser–matter physics, plasma-based accelerators, and radiation-source development. Their systematic methodology and thorough validation render them particularly valuable for future research.

Moreover, the complementary use of test-particle models and full particle-in-cell (PIC) simulations provides a powerful dual framework for probing the validity and limitations of the theory of averaged relativistic motion. While PIC simulations capture collective plasma effects and self-consistent field evolution, test-particle models isolate the fundamental single-particle physics underlying ponderomotive acceleration, injection, and transport. Together, they not only validate existing analytical models but also guide the development of next-generation theories that incorporate non-adiabatic, phase-resolved, and field-topology-dependent effects — capabilities essential for the design and optimization of next-generation laser–plasma accelerators and compact radiation sources.

Author Contributions: Conceptualization, Castillo A. J.; methodology, Castillo A. J. and Rudoy Yu. Gr.; formal analysis, Castillo A. J. and Rudoy Yu. Gr.; investigation, Castillo A. J. and Rudoy Yu. Gr.; writing original draft preparation, Castillo A. J.; writing review and editing, Castillo A. J. and Rudoy Yu. Gr.; supervision, Rudoy Yu. Gr.; project administration, Castillo A. J. All authors have read and agreed to the published version of the manuscript.

Funding: This research received no external funding.

Data Availability Statement: Due to the nature of the research, data sharing is not applicable to this article.

Acknowledgments: To the memory of Prof. V. P. Milant'ev.

Conflicts of Interest: The authors declare no conflict of interest.

Declaration on Generative AI: The authors have not employed any Generative AI tools.

References

1. Bychenkov, V. Y., Castillo, A. J., Bochkarev, S. G. & Lobok, M. G. Laser Acceleration of Electrons:“Laser Buller” or “Bubble”? *JETP Letters* **121,** 512–519 (2025).
2. Mulser, P. *Hot Matter from High-Power Lasers* (Springer, 2020).
3. Bolotovski, B. M. & Serov, A. V. Special features of motion of particles in an electromagnetic wave. *Physics-Uspekhi* **46,** 645 (2003).
4. Andreev, S. N., Makarov, V. P. & Rukhadze, A. A. On the motion of a charged particle in a plane monochromatic electromagnetic wave. *Quantum Electronics* **39,** 68 (2009).
5. Morozov, A. I. & Solov'Ev, L. S. *Problems of plasma theory, second release* (Atomizdat Moscow, 1963).
6. Milant'ev, V. P. & Castillo, A. J. On the theory of the relativistic motion of a charged particle in the field of intense electromagnetic radiation. *Journal of Experimental and Theoretical Physics* **116,** 558–566 (2013).
7. Castillo, A. J. & Milant'ev, V. P. Relativistic ponderomotive forces in the field of intense laser radiation. *Technical Physics* **59,** 1261–1266 (2014).
8. Castillo, A. J. & Milant'ev, V. P. On the averaged relativistic forces in the field of laser beat wave. *Inzheniernaya Fizika* **4,** 16–22 (2014).
9. Castillo, A. J. & Milant'ev, V. P. Features of the relativistic motion of a single electron entering a waveguide. *Physics of Plasmas* **28** (2021).
10. Gaponov, A. V. & Miller, M. A. Potential Wells For Charged Particles In A High-Frequency Electro-Magnetic Field. *Journal of Experimental and Theoretical Physics* **34,** 242–243 (1958).
11. Kibble, T. W. B. Refraction of electron beams by intense electromagnetic waves. *Physical Review Letters* **16,** 1054 (1966).
12. Startsev, E. A. & McKinstrie, C. J. Multiple scale derivation of the relativistic ponderomotive force. *Physical Review E* **55,** 7527 (1997).
13. Taranukhin, V. D. Structure of ponderomotive forces interacting with an electron in the laser fields of relativistic intensity. *Zhurnal Ehksperimental'noj i Teoreticheskoj Fiziki* **117** (2000).
14. Malka, G., Lefebvre, E. & Miquel, J. L. Experimental observation of electrons accelerated in vacuum to relativistic energies by a high-intensity laser. *Physical review letters* **78,** 3314 (1997).
15. Kibble, T. W. B. Mutual refraction of electrons and photons. *Physical Review* **150,** 1060 (1966).
16. Lindman, E. L. & Stroscio, M. A. On the relativistic corrections to the ponderomotive force. *Nuclear Fusion* **17,** 619 (1977).
17. Bauer, D., Mulser, P. & Steeb, W.-H. Relativistic ponderomotive force, uphill acceleration, and transition to chaos. *Physical review letters* **75,** 4622 (1995).
18. Ruiz, D. E. & Dodin, I. Y. Ponderomotive dynamics of waves in quasiperiodically modulated media. *Physical Review A* **95,** 032114 (2017).
19. Bituk, D. R. & Fedorov, M. V. Relativistic ponderomotive forces. *Journal of Experimental and Theoretical Physics* **89,** 640–646 (1999).
20. Smorenburg, P. W., Kanters, J. H., Lassise, A., Brussaard, G. J., Kamp, L. P. & Luiten, O. J. Polarization-dependent ponderomotive gradient force in a standing wave. *Physical Review A—Atomic, Molecular, and Optical Physics* **83,** 063810 (2011).
21. Bogolyubov, N. N. & Mitropolskii, Y. A. *Asymptotic Methods in Oscillation Theory* 1974.
22. Shiryaev, O. B. Asymptotic theory of ponderomotive dynamics of an electron in the field of a focused relativistically intense electromagnetic envelope. *Quantum Electronics* **49,** 936 (2019).
23. Kaplan, A. E. & Pokrovsky, A. L. Fully relativistic theory of the ponderomotive force in an ultraintense standing wave. *Physical review letters* **95,** 053601 (2005).
24. Dodin, I. Y. Ponderomotive forces and wave dispersion: two sides of the same coin. *arXiv preprint arXiv:1107.2852* (2011).

25. Manheimer, W. M. A covariant derivation of the ponderomotive force. *The Physics of Fluids* **28,** 1569–1571 (1985).
26. Esmailzadeh, E., Younesian, D. & Askari, H. Analytical methods in nonlinear oscillations. *Netherlands: Springer* (2018).
27. Dodin, I. Y. & Fisch, N. J. Axiomatic geometrical optics, Abraham-Minkowski controversy, and photon properties derived classically. *Physical Review A—Atomic, Molecular, and Optical Physics* **86,** 053834 (2012).
28. Kentwell, G. W. & Jones, D. A. The time-dependent ponderomotive force. *Physics Reports* **145,** 319–403 (1987).
29. Yang, J. H., Craxton, R. S. & Haines, M. G. Explicit general solutions to relativistic electron dynamics in plane-wave electromagnetic fields and simulations of ponderomotive acceleration. *Plasma Physics and Controlled Fusion* **53,** 125006 (2011).
30. Vinogradova, M. B., Rudenko, O. V. & Sukhorukov, A. P. *Theory of Waves* (Nauka Moscow, 1979).
31. Quesnel, B. & Mora, P. Theory and simulation of the interaction of ultraintense laser pulses with electrons in vacuum. *Physical Review E* **58,** 3719 (1998).
32. Gonoskov, A., Blackburn, T. G., Marklund, M. & Bulanov, S. S. Charged particle motion and radiation in strong electromagnetic fields. *Reviews of Modern Physics* **94,** 045001 (2022).
33. D'ippolito, D. A. & Myra, J. R. Quasilinear theory of the ponderomotive force: Induced stability and transport in axisymmetric mirrors. *The Physics of fluids* **28,** 1895–1905 (1985).
34. Aseyev, S. A., Mironov, B. N., Minogin, V. G. & Chekalin, S. V. Measurement of the Gaponov-Miller force produced in vacuum by tightly focused intense femtosecond laser radiation. *Journal of Experimental and Theoretical Physics* **112,** 780–783 (2011).
35. Tajima, T. & Dawson, J. M. Laser electron accelerator. *Physical review letters* **43,** 267 (1979).
36. Sprangle, P., Esarey, E., Krall, J. & Ting, A. *Vacuum Laser Acceleration* tech. rep. (1995).
37. Litvak, A. G. & Trakhtengerts, V. Y. Induced scattering of waves and plasma heating by coherent radiation. *Sov. Phys. JETP* **33,** 921 (1971).
38. Serov, A. V. Ponderomotive nongradient force acting on a relativistic particle crossing an inhomogeneous electromagnetic wave. *Journal of Experimental and Theoretical Physics* **92,** 20–27 (2001).
39. Milant'ev, V. P. On the possibility of averaging the equations of an electron motion in the intense laser radiation. *Discrete and Continuous Models and Applied Computational Science* **29,** 105–113 (2021).
40. Popruzhenko, S. V. & Fedotov, A. M. Dynamics and radiation of charged particles in ultra-intense laser fields. *Uspekhi Fizicheskikh Nauk* **193,** 491–527 (2023).
41. Castillo, A. J., Bochkarev, S. G. & Bychenkov, V. Y. *Particle drift, diffusion, and acceleration in quasi-static fields generated by ultrashort relativistically intense laser pulse channeling in near-critical density targets* in *2024 International Conference Laser Optics (ICLO)* (2024), 226–226.
42. Bochkarev, S. G., Brantov, A. V., Bychenkov, V. Y., Torshin, D. V., Kovalev, V. F., Baidin, G. V. & Lykov, V. A. Stochastic electron acceleration in plasma waves driven by a high-power subpicosecond laser pulse. *Plasma Physics Reports* **40,** 202–214 (2014).
43. Zhang, Y. & Krasheninnikov, S. I. Electron heating in the laser and static electric and magnetic fields. *Physics of Plasmas* **25** (2018).
44. Burton, D. A., Cairns, R. A., Ersfeld, B., Noble, A., Yoffe, S. & Jaroszynski, D. A. *Observations on the ponderomotive force* in *Relativistic Plasma Waves and Particle Beams as Coherent and Incoherent Radiation Sources II* **10234** (2017), 17–22.
45. Malka, G. & Miquel, J. L. Experimental confirmation of ponderomotive-force electrons produced by an ultrarelativistic laser pulse on a solid target. *Physical review letters* **77,** 75 (1996).

46. Roso, L., Pérez-Hernández, J. A., Lera, R. & Fedosejevs, R. *The Role of the Ponderomotive Force in High Field Experiments* in *Progress in Ultrafast Intense Laser Science XVI* 149–177 (Springer, 2021).
47. Hegelich, B. M., Labun, L. & Labun, O. Z. Revisiting experimental signatures of the ponderomotive force. *Photonics* **10,** 226 (2023).
48. Galkin, A. L., Korobkin, V. V., Romanovskii, M. Y. & Shiryaev, O. B. Relativistic motion and radiation of an electron in the field of an intense laser pulse. *Quantum Electronics* **37,** 903 (2007).
49. Wang, P. X., Ho, Y. K., Yuan, X. Q., Kong, Q., Cao, N., Sessler, A. M., Esarey, E. & Nishida, Y. Vacuum electron acceleration by an intense laser. *Applied Physics Letters* **78,** 2253–2255 (2001).
50. Gibbon, P. *Short pulse laser interactions with matter: an introduction* (World Scientific, 2005).
51. Galkin, A. L., Korobkin, V. V., Romanovsky, M. Y. & Shiryaev, O. B. Electron acceleration in quasi-stationary electromagnetic fields during the self-channeling of intense light pulses. *Journal of Experimental and Theoretical Physics* **100,** 1050–1060 (2005).
52. Bychenkov, V. Y. & Kovalev, V. F. Self-Trapping of a Laser Beam of Ultrarelativistic Intensities. *JETP Letters* **120,** 334–340 (2024).

Information about the authors

Castillo, Alejandro J.—PhD Student of Institute of Physical Research and Technology, RUDN University; Junior Research Fellow of P.N. Lebedev Physical Institute of the Russian Academy of Sciences (LPI); Lecturer of Department of Higher Mathematics of State Autonomous Educational Institution of Higher Education "N.I. Pirogov Russian National Research Medical University" of Ministry of Health of Russian Federation (e-mail: 114222068@rudn.ru, ORCID: 0000-0003-0001-8764)

Rudoy, Yuriy Gr.—Professor of Institute of Physical Research and Technology of RUDN University (e-mail: rudikar@mail.ru, ORCID: 0000-0002-7130-4859)

УДК 537.533.7
PACS 52.25.Os, 52.30.−q, 52.35.−g, 52.38.−r, 52.40.Db, 52.50.Jm, 52.57.−z
DOI: 10.22363/2658-4670-2026-34-1-113-124 EDN: UPXGCS

Взаимодействие релятивистских электронов с интенсивными электромагнитными полями: пондеромоторные эффекты, ускорение, преломление, отражение и зависимость от начальных условий

А. Х. Кастильо[1, 2, 3], Ю. Г. Рудой[1]

[1] Российский университет дружбы народов, ул. Миклухо-Маклая, д. 6, Москва, 117198, Российская Федерация

[2] Физический институт имени П. Н. Лебедева РАН, Ленинский проспект, д. 53, Москва, 119991, Российская Федерация

[3] РНИМУ имени Н. И. Пирогова, ул. Островитянова, д. 1, стр. 6, Москва, 117513, Российская Федерация

Аннотация. Строгая теория и описание динамики заряженных частиц в высокоинтенсивных электромагнитных полях имеют фундаментальное значение для разработки перспективных плазменных приложений. Точные аналитические модели должны устранять разрыв между усредненными траекториями и истинным движением частиц для предварительного определения параметров инжекции и набора энергии. Основная цель данного обзора заключается в создании строгого аналитического описания усредненного релятивистского движения электронов с упором на строгий вывод пондеромоторных сил и влияние быстро осциллирующих периодических добавок на динамические переменные. С помощью метода усреднения Крылова–Боголюбова–Митропольского для получения уравнений движения в работе анализируются релятивистские эффекты в лазерных пучках и волноводах. Теоретические результаты подтверждаются в ходе численной валидации, включающей моделирование тестовых частиц в заданных полях и трехмерное моделирование плазмы в ячейках (PIC) для режимов релятивистского самозахвата, таких как структуры «лазерная пуля» и «пузырь». В обзоре подробно описаны независимость результатов от способа описания (приближение), строгая зависимость от поляризации волны и нестрого потенциальный характер релятивистской пондеромоторной силы. Анализ показывает, что периодические быстро осциллирующие добавки необходимы для полного описания, точного задания начальных условий в усреднённых уравнениях и обеспечения достоверного прогнозирования явлений отражения и преломления электронов. Моделирование подтверждает, что быстро осциллирующие добавки определяют инжекцию электронов и заряд пучка в реалистичных сценариях лазерно-плазменного ускорения. Данный обзор демонстрирует, что комбинированное использование моделей тестовых частиц и PIC-моделирования является крайне важным для исследования пределов теории усреднённого движения. Полученные результаты имеют прямую практическую значимость для оптимизации источников излучения и служат ориентиром для развития будущих теорий, учитывающих неадиабатические эффекты и эффекты, зависящие от топологии поля.

Ключевые слова: усреднённое движение, релятивистские пондеромоторные силы, лазерное излучение, гауссов пучок, волноводы, биение

Discrete & Continuous Models & Applied Computational Science
ISSN 2658-7149 (Online), 2658-4670 (Print)

2026, 34 (1) 125–138
http://journals.rudn.ru/miph

Research article

UDC 537.527,533.9.03
PACS 52.80 Pi, 52.80.Sm, 52.50.Sw, 52.40.Db
DOI: 10.22363/2658-4670-2026-34-1-125-138 EDN: UOSEFX

Mathematical models of low-pressure discharge in a magnetic field supported by UHF electromagnetic field

Sergey A. Dvinin[1, 2], Denis V. Chuprov[2], Konstantin N. Kornev[1, 2], Zafari A. Qodirzoda[3], Davlat K. Solikhzoda[3]

[1] RUDN University, 6 Miklukho-Maklaya St, Moscow, 117198, Russian Federation

[2] Lomonosov Moscow State University 1 build 2 Leninskiye gory, Moscow, 119991, Russian Federation

[3] Tajik National University, 17 Rudaki Av, Dushanbe, 973402, Tajikistan

(received: February 6, 2026; revised: February 22, 2026; accepted: February 25, 2026)

Abstract. Electron cyclotron resonance (ECR) discharges are an effective way to generate plasma at low working gas pressure. The aim of this work is to develop a mathematical model of the ECR discharge implemented at the RAPIRA facility (RUDN University), which is used for a wide range of scientific research. The evolution of plasma particles is described within the framework of the hydrodynamic approximation (a two-dimensional model with cylindrical symmetry). A three-dimensional model of cold plasma is used to calculate the spatial distribution of the electromagnetic field. Calculations have shown that in the operating mode of the facility (gas pressures from $4 \cdot 10^{-4}$ to 10^{-2} Torr, magnetic field up to 2500 G), the electron temperature is equalized along the magnetic field lines, and at the same time, the magnetic field ensures a decrease in energy losses to the side walls of the facility. The spatial distributions of the electron density and temperature and the electromagnetic field in the plasma are calculated. The implemented model can serve as a basis for developing a more advanced set of software codes that take into account the non-Maxwellian nature of the electron velocity distribution function, caused by the non-adiabatic nature of their heating in a non-uniform magnetic field.

Key words and phrases: ECR discharge, discharge in a resonator, discharge in a magnetic trap, drift-diffusion model

For citation: Dvinin, S. A., Chuprov, D. V., Kornev, K. N., Qodirzoda, Z. A., Solikhzoda, D. K. Mathematical models of low-pressure discharge in a magnetic field supported by UHF electromagnetic field. *Discrete and Continuous Models and Applied Computational Science* **34** (1), 125–138. doi: 10.22363/2658-4670-2026-34-1-125-138. edn: UOSEFX (2026).

1. Introduction

Electron cyclotron resonance (ECR) discharge is currently used in various fields of science and technology: plasma-chemistry installations for material processing [1–7], sources for multiply charged ions (MCI) [8–10], sources of hydrogen ions for proton accelerators [11, 12], and microwave plasma thruster [13]. The multitude of possible applications has led to a variety of discharge installation geometries in which ECR interaction is realized, differing both in the spatial configuration of the constant magnetic field and in the method of exciting the electromagnetic field and its frequencies. On the other hand, the diversity of installation options determines different approaches to constructing mathematical models of the processes occurring in these installations.

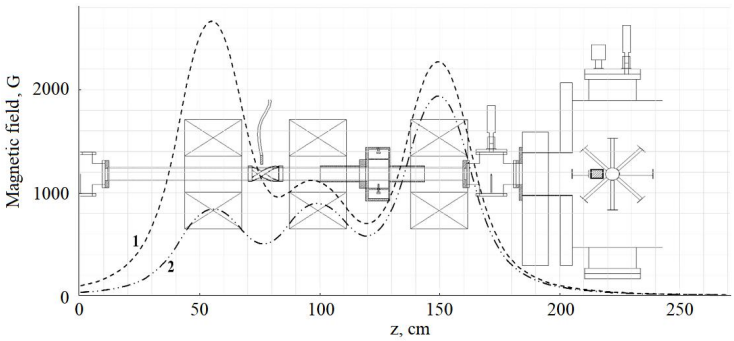

Figure 1. Setup diagram: A—processing chamber, B—quartz plasma pipeline, C—gas inlet connection point, D—helicon antenna, E—microwave resonator, F1, F2, F3—magnetic coils. Axial distribution of magnetic field induction along plasma pipeline under microwave (1) and HF (2) discharges

The purpose of this work is to formulate an approach for developing a model of microwave discharge implemented on the multifunctional installation—RAPIRA (Resonant Accelerated Plasma Installation Research & Application, RUDN University), used to study the absorption of microwave power by a magnetized plasma filling a cavity, the processes of plasma transport along a cylindrical quartz discharge tube (plasma pipeline) from the source to the processing chamber and processing of various chemical and biological objects by plasma created.

2. Experimental setup and computer modeling tools

The schematic view of the RAPIRA installation is shown in Figure 1. First of all, we list the elements and systems that are important and relevant for the numerical model being developed. The installation contains (A)—a processing chamber in which the processed samples are placed, (B)—a quartz plasma pipeline, (C)—a gas inlet system, (D)—helicon antenna for generation of RF (13.56 MHz) discharge, (F)—coils (1, 2, 3) for generating a magnetic field, and (E)—a microwave cavity. The magnetic field configuration is controlled by currents through coils (F.1–F.3).

The experimental setup was developed to use RF and microwave plasma discharges to create plasma flows to study their interaction with various substrates. The RF discharge is generated using a half-wave helicon antenna, the microwave discharge is initiated in a cylindrical resonator with the fundamental oscillation mode H_{111}. The curves of two longitudinal magnetic field distributions, providing resonant conditions during the operation of microwave 1 and RF 2 plasma sources, are also shown in figure 1. The range of possible pressures of the plasma-forming gases and mixtures of the installation is quite wide, but in this paper, we will consider the option of generating microwave plasma at 0.01–0.04 Pa. The microwave resonator is excited by two rod antennas inserted into the resonator perpendicular to the side wall. Each rod is 6 mm in diameter and inserted 32.1 mm deep of the cavity. An electrodynamic model for a microwave discharge is considered as an example for calculation. To prevent the loss of microwave radiation through the holes in the end walls of the cavity, the axisymmetric quartz pipeline was shielded with cylindrical evanescent waveguides.

The paper shows that in the specified pressure range (with the possible exception of the lowest pressures), the discharge can be described within the drift-diffusion model, including the particle balance equations, the energy balance equation, and Maxwell's equations. This approach is standard for most gas discharge models [14]. The specificity of this work is that this system of equations is

used to describe the discharge in a non-uniform magnetic field. The system of equations obtained below was solved using the Comsol Multiphysics software package [15]. The RF module of Comsol was used to solve the electrodynamic equations, and the diffusion and heat transfer equations were solved using the main module. The magnetic field of each coil was approximated as the field of the current flowing along a ring of radius R. The radius R was chosen in such a way as to approximate the experimentally obtained dependence of the magnetic field of each coil along the plasma guide axis as accurately as possible. The zero coordinate of the calculation problem corresponded to the position of the resonator exciters.

The model of a microwave discharge in the specified pressure range (with the possible exception of the lowest pressures) can be described in the framework of the drift-diffusion approach, which includes particle balance equations, energy balance equation, and Maxwell's equations. This approach is standard for most gas discharge models [14]. The specificity of this paper is that the above-mentioned system of equations is used to describe discharge in a non-uniform magnetic field. The system of equations was solved using the Comsol Multiphysics software package [15]. The specificity of this paper is that the system of equations is used to describe the discharge in a non-uniform magnetic field. The RF module of Comsol was used to solve the electrodynamic equations, and the drift-diffusion model equations were solved using the main module. The magnetic field of each coil was approximated as the field of the current flowing along a ring of radius R. The radius R was chosen in such a way as to approximate the experimentally obtained dependence of the magnetic field of each coil along the plasma guide axis as accurately as possible. The zero coordinate of the calculation problem corresponded to the position of the resonator exciters.

3. Diffusion and loss of particles in the discharge

Estimates show that the longitudinal dimensions of the plasma conduit in the pressure range of 0.01–0.04 Pa are greater than the wavelength, with the possible exception of the lowest pressures in this range. In the transverse direction, the magnetization conditions are satisfied: $|\Omega_\alpha|\tau_\alpha > 1$, where $\Omega_\alpha = e_\alpha B/m_\alpha c$, $\tau_\alpha^{-1} = \nu_\alpha$ is the cyclotron frequency and the collision frequency of type α particles ($\alpha = e$ for electrons and $\alpha = +$ for ions). In this case, the transverse discharge dimensions also exceed the Larmor radius, so the latter can be considered as the mean free path when considering the radial motion of charged particles. Therefore, in this case, the discharge can be described within the framework of the drift-diffusion (hydrodynamic) model.

In this case, the diffusion and thermal conductivity coefficients become anisotropic [16]. In a uniform magnetic field, the diffusion equations have the form:

$$n_e\mathbf{V}_e = -\frac{n_e}{1+(\Omega_e/\nu_{en})^2}\left\{\left(\mu_e\mathbf{E} + D_e\frac{\nabla n_e}{n_e}\right) + \left[\frac{e}{\nu_{en}}\times\left(-\mu_e\mathbf{E} - D_e\frac{\nabla n_e}{n_e}\right)\right]\right\} - \\ - \mu_e n_e\frac{e(\mathbf{E}_e)}{\Omega_e^2} - D_e\frac{e(\nabla n_{ee})}{\Omega_e^2}, \quad (1)$$

$$n_+\mathbf{V}_+ = \frac{n_+}{1+(\Omega_+/\nu_{+n})^2}\left\{\left(\mu_+\mathbf{E} - D_+\frac{\nabla n_+}{n_+}\right) + \left[\frac{+}{\nu_{+n}}\times\left(\mu_+\mathbf{E} - D_+\frac{\nabla n_+}{n_+}\right)\right]\right\} + \\ + \mu_+ n_+\frac{+(\mathbf{E}_+)}{\Omega_+^2} - D_+\frac{+(\nabla n_{++})}{\Omega_+^2}. \quad (2)$$

Here n_e, n_+, Ω_e, Ω_+ and ν_{en}, ν_{+n} are the densities, cyclotron frequencies and effective collision frequencies for electrons and ions, μ_e, μ_+, D_e, D_+ are the mobilities and diffusion coefficients of electrons and ions along the magnetic field, **E** is the ambipolar field. Thus, the magnetic field does not affect the motion of particles along the magnetic field lines. In addition, from equations (1) and (2) it follows that particles participate in drift motion in the direction perpendicular to both the electric and magnetic fields, with negative and positive particles drifting in different directions. Finally, there is drift and diffusion of particles in the direction parallel to the electric field and the density gradient of charged particles. The value of diffusion coefficients in the direction across the magnetic field are significantly smaller than the value, when particles moves along a magnetic field.

$$D_{e\perp} = \frac{D_e}{1+(\Omega_e/\nu_{en})^2}, \qquad D_{+\perp} = \frac{D_+}{1+(\Omega_+/\nu_{+n})^2},$$
$$\mu_{e\perp} = \frac{\mu_e}{1+(\Omega_e/\nu_{en})^2}, \qquad \mu_{+\perp} = \frac{\mu_+}{1+(\Omega_+/\nu_{+n})^2}.$$

The complete system of equations in the drift-diffusion model for a homogeneous magnetic field includes equations for the electron and ion currents (1), (2), and the electron and ion balance equations. The Poisson equation, which should close the system of equations, is replaced by the quasi-neutrality condition, whereby the equation for the electron density is excluded from consideration, and instead, the equation for the electric current is used, which is also a consequence of the quasi-neutrality condition: $n_+ = n_e = n$, $\left(\nabla \cdot n(\mathbf{V}_e - \mathbf{V}_+)\right) = 0$. Using equations (1) and (2), we also eliminate the equations for the electron and ion currents. Thus, the complete system of equations takes the form:

$$-\frac{\partial}{\partial z}\left(D_{+zz}\frac{\partial n}{\partial z} + n\mu_{+zz}\frac{\partial \varphi}{\partial z}\right) - \frac{\partial}{\partial x}\left(D_{+xx}\frac{\partial n}{\partial x} + n\mu_{+xx}\frac{\partial \varphi}{\partial x}\right) = \nu_i n, \tag{3}$$

$$\frac{\partial}{\partial z}\left((D_{ezz} - D_{+zz})\frac{\partial n}{\partial z} - n(\mu_{ezz} + \mu_{+zz})\frac{\partial \varphi}{\partial z}\right) + \frac{\partial}{\partial x}\left((D_{exx} - D_{+xx})\frac{\partial n}{\partial x} - n(\mu_{exx} + \mu_{+xx})\frac{\partial \varphi}{\partial x}\right) = 0. \tag{4}$$

The final system of equations for a system with one type of ion includes equations (3), (4). The boundary conditions are usually set in the form ($\boldsymbol{\eta}$ is the normal to the wall surface, Λ_i is the mean free path of the ion).

$$\left(\boldsymbol{\eta} \cdot \nabla n\right) = n/\Lambda_i, \qquad \left(\boldsymbol{\eta} \cdot (-\mathbf{j}_e + \mathbf{j}_+)\right) = 0. \tag{5}$$

Equations (5) are valid in the case when ion mean free path is less than the size of the sheath between the plasma and the quartz tube. Otherwise, the Bohm criterion is used, which states that the plasma flow velocity at the boundary is equal to the ion-sound velocity. The partial differential equations were solved using the Comsol Multiphysics mathematical package [15].

In a nonuniform magnetic field, the induction is not directed along the 0Z axis. Therefore, the diffusion and mobility tensors of charged particles will no longer be diagonal. Below we write the ion balance equation and the equations for the currents, which replace equations (3) and (4). Further in the formulas we replace n_e and n_+ by n.

1. Equation of charged particle densities (The upper sign + corresponds to ions, the lower sign − to electrons):

$$\frac{\partial n}{\partial t} - \frac{\partial}{\partial z}\Big[\left(D_{e,i\perp} - D_{e,i\parallel}\right)\sin\theta\cos\theta\frac{\partial n}{\partial r} - \left(D_{e,i\perp}\sin^2\theta + D_{e,i\parallel}\cos^2\theta\right)\frac{\partial n}{\partial z} \pm$$
$$\pm\, n\left(\mu_{e,i\perp} - \mu_{e,i\parallel}\right)\sin\theta\cos\theta\frac{\partial \varphi}{\partial r} \pm n\left(\mu_{e,i\perp}\sin^2\theta + \mu_{e,i\parallel}\cos^2\theta\right)\frac{\partial \varphi}{\partial z}\Big] -$$
$$-\frac{\partial}{r\partial r}\Big[\left(D_{e,i\perp}\cos^2\theta + D_{e,i\parallel}\sin^2\theta\right)r\frac{\partial n}{\partial r} - \left(D_{e,i\parallel} - D_{e,i\perp}\right)r\frac{\partial n}{\partial z}\sin\theta\cos\theta \pm$$

$$\pm n\left(\mu_{e,i\perp}\cos^2\theta + \mu_{e,i\parallel}\sin^2\theta\right)\frac{\partial\varphi}{\partial r} \pm n\left(\mu_{e,i\perp} - \mu_{e,i\parallel}\right)\frac{\partial\varphi}{\partial z}\sin\theta\cos\theta\Big] = \nu_i n.$$

2. Equations for the ambipolar field potential

$$\begin{aligned}(\nabla\cdot\mathbf{J}) = \frac{\partial}{\partial z}\Big\{&\Big(\left(D_{e\perp} - D_{i\perp}\right) - \left(D_{e\parallel} - D_{i\parallel}\right)\Big)\sin\theta\cos\theta\frac{\partial n}{\partial r} - \\ &- \left(\left(D_{e\perp} - D_{i\perp}\right)\sin^2\theta + \left(D_{e\parallel} - D_{i\parallel}\right)\cos^2\theta\right)\frac{\partial n}{\partial z} - \\ &- \Big(\left(\mu_{e\perp} - \mu_{i\perp}\right) - \left(\mu_{e\parallel} - \mu_{i\parallel}\right)\Big)\sin\theta\cos\theta\frac{\partial\varphi}{\partial r} - \\ &- \left(\left(\mu_{e\perp} - \mu_{i\perp}\right)\sin^2\theta + \left(\mu_{e\parallel} - \mu_{i\parallel}\right)\cos^2\theta\right)\frac{\partial\varphi}{\partial z}\Big\} - \\ -\frac{1}{r}\frac{\partial}{\partial r}r\Big\{&\left(\left(D_{e\perp} - D_{i\perp}\right)\cos^2\theta + \left(D_{e\parallel} - D_{i\parallel}\right)\sin^2\theta\right)\frac{\partial n}{\partial r} - \\ &- \Big(\left(D_{e\parallel} - D_{i\parallel}\right) - \left(D_{e\perp} - D_{i\perp}\right)\Big)\sin\theta\cos\theta\frac{\partial n}{\partial z} - \\ &- n\Big(\left(\mu_{e\perp} + \mu_{i\perp}\right)\cos^2\theta + \left(\mu_{e\parallel} + \mu_{i\parallel}\right)\sin^2\theta\Big)\frac{\partial\varphi}{\partial r} + \\ &+ n\Big(\left(\mu_{e\parallel} + \mu_{i\parallel}\right) - \left(\mu_{e\perp} + \mu_{i\perp}\right)\Big)\sin\theta\cos\theta\frac{\partial\varphi}{\partial z}\Big\} = 0. \quad (6)\end{aligned}$$

In these equations, θ is the angle between the direction of the constant magnetic field and the 0Z axis: $\theta = \arctan\left(H_r(r,z)/H_z(r,z)\right)$. The boundary conditions coincide with the conditions in a uniform magnetic field.

4. Heat transfer in the discharge. Heating and energy loss of electrons in the plasma

The charged particle balance equations include the electron and ion production rate, which depends on the chemical reactions occurring in the plasma. The rate of these reactions, in turn, depends on the electron temperature and the temperatures of the heavy particles. Since a low-pressure discharge is being considered, it can be expected that no temperature change along the field line should occur. In the transverse direction, where particle transport is suppressed by the magnetic field, energy transfer may be insufficient and the temperature may vary. Ideally, to calculate the frequencies of chemical processes, it is necessary to solve the heat conduction equation for electrons, which has the form

$$\begin{aligned}\frac{3}{2}nk\frac{\partial T_e}{\partial t} - \frac{\partial}{\partial z}\Big[&\left(\chi_{e\perp} - \chi_{e\parallel}\right)\sin\theta\cos\theta\frac{\partial T_e}{\partial r} - \left(\chi_{e\perp}\sin^2\theta + \chi_{e\parallel}\cos^2\theta\right)\frac{\partial n}{\partial z}\Big] - \\ &- \frac{\partial}{r\partial r}\Big[\left(\chi_{e\perp}\cos^2\theta + \chi_{e\parallel}\sin^2\theta\right)r\frac{\partial T_e}{\partial r} - \left(\chi_{e\perp} - \chi_{e\parallel}\right)r\frac{\partial T_e}{\partial z}\sin\theta\cos\theta\Big] = \\ &= \sum_{j=1}^{3}\sum_{i=1}^{3}\sigma_{ij}E_iE_j^* - Q. \quad (7)\end{aligned}$$

Here χ is the thermal conductivity coefficient of the plasma, which was calculated in accordance with [16, 17], k is Boltzmann constant. The role of the magnetic field was taken into account in accordance with [16]. Here $Q = nw_1$ is the energy transferred by electrons to other particles in elastic and inelastic collision. The calculation of w_1 will be discussed below. It can be expected that under the experimental conditions, due to the high thermal conductivity along the magnetic field lines, the electron temperature in this direction should equalize. Here, σ_{ij} is the plasma conductivity, accounting for its anisotropy and the high-frequency nature of the field. The field absorption is calculated using the effective collision frequency, which takes into consideration both collisional and collisionless energy gain by electrons.

In the direction perpendicular to the magnetic surface, the thermal conductivity of the electron gas is significantly lower, so radial temperature non-uniformity can be expected in cases where heating across the plasma cross-section is non-uniform. Therefore, the process of establishing the spatial distribution of electron temperature should be investigated using mathematical modeling.

According to models of a steady-state low-pressure discharge, ionization balances losses. Losses are determined by the discharge geometry (i.e., the position of the boundaries and the magnetic field strength profile), the chemical properties, and the pressure of the working gas. If the spatial distribution of electron temperature is uniform, then particle losses determine the ionization required in the discharge and, consequently, the electron temperature.

Then the value of this temperature should ensure particle balance, i.e., as is usually the case in a stationary discharge, the required temperature is determined by the particle balance. If the ionization cross-section is known, the temperature is determined from the relation

$$\nu_{i,s} = 4\pi N \int_0^\infty V q_{i,s}(V) f_e(V, T_e) V^2 \, dV, \tag{8}$$

where $q_{i,s}$ is the ionization or excitation cross-section, N is the density of neutral atoms, $f_e(V)$ is the electron energy distribution function, which is assumed to be isotropic. If the function is Maxwellian and a linear approximation is used for the process cross-section (ε_i is ionization threshold),

$$q_i = a(\varepsilon - \varepsilon_i), \tag{9}$$

or Fabrikant's approximation (ε_m is the energy at which the ionization cross-section is maximum and equal to q_m)

$$q_s = q_{ms} \frac{\varepsilon - \varepsilon_s}{\varepsilon_{ms} - \varepsilon_s} \exp\left(\frac{\varepsilon_{ms} - \varepsilon}{\varepsilon_{ms} - \varepsilon_s}\right),$$

then the following expressions can be obtained for the frequencies (e, m are electron charge and mass):

$$\nu_i = \frac{4}{\sqrt{\pi}} \left(aN \frac{kT_e}{e}\right) \left(\frac{2kT_e}{m}\right)^{1/2} \left(1 + \frac{\varepsilon_i}{2kT_e}\right) \exp\left(-\frac{\varepsilon_i}{kT_e}\right),$$

$$\nu_s = \frac{4}{\sqrt{\pi}} q_{ms} N \left(\frac{2kT_e}{m}\right)^{1/2} \frac{kT}{\varepsilon_{ms} - \varepsilon_s} \left[1 + \frac{\varepsilon_s}{2kT_e}\left(1 + \frac{kT_e}{\varepsilon_{ms} - \varepsilon_s}\right)\right] \times \exp\left(1 - \frac{\varepsilon_s}{kT_e}\right) \Big/ \left(1 + \frac{kT_e}{\varepsilon_{ms} - \varepsilon_s}\right)^3.$$

If more accurate results are required, approximations [18] and numerical integration (8) can be used. The energy losses of electrons are determined by the relation

$$w_1(T_e) = \frac{2m}{M} \nu_{en} \frac{3}{2} k(T_e - T_g) + \sum_s \nu_s(T_e) \varepsilon_s + \nu_i(T_e) \left(\varepsilon_i + 2kT_e + \varepsilon_{ist}\right).$$

The first term accounts for elastic energy losses, the second for excitation losses, and the third for ionization losses. The last term accounts for the energy carried away by ions toward the wall. Ion acceleration occurs due to the ambipolar field. The temperature distribution in the discharge is given by the thermal conductivity equation (7).

By integrating equation (7) over the entire discharge volume, we obtain the energy required to maintain a discharge with a given electron density. Note also that using cross-section (9) for the ionization frequency yields a temperature of 4.5 eV in the setup's operating modes, but the formula itself overestimates the ionization frequency. The energy required to create an electron under these conditions was 10^9 eV/s/Torr.

5. Spatial distribution of the electromagnetic field in the discharge

The particle balance equation (6) allows us to determine the ionization frequency (averaged over the volume) required to maintain the discharge at steady state. The particle balance equation (7) quantifies the energy required to create the required number of electrons in 1 second. Knowing this energy, we can determine the power required to maintain a plasma with a given average density n_0 by integrating the solution of equation (7) over the entire plasma volume:

$$W = \iiint\limits_V dx\, dy\, dz\, n_e(x, y, z) w_1(x, y, z).$$

The final step required to complete the mathematical model is to solve Maxwell's equations. Knowing the microwave power required to maintain the discharge allows us to determine the amplitude, spatial distribution of the microwave density, and discharge impedance, and select an appropriate method for matching the discharge to the generator. The model construction procedure described above is not self-consistent, since the solution of the particle balance equation assumed the electron temperature to be uniform throughout the volume, etc. Nevertheless, it usually allows for a fairly accurate determination of the averaged discharge parameters as functions of given conditions (geometry, chemical nature of the gas, etc.). The necessary refinement of the model can be made at later stages, possibly using well-developed perturbation theory or other methods.

Let us now turn to the presentation of the electrodynamic part of the problem. Electrodynamically, the discharge was described using the cold plasma model [19, 20]. Maxwell's equations were solved using the "Comsol Multiphysics" software package. The permittivity is written as:

$$(\varepsilon_{ij}) = \begin{pmatrix} \varepsilon_\perp & ig & 0 \\ -ig & \varepsilon_\perp & 0 \\ 0 & 0 & \varepsilon_\parallel \end{pmatrix},$$

where

$$\varepsilon_\perp = 1 - \frac{n_e}{n_c} \frac{(1 + i\nu_{en}/\omega)}{n_c(1 + i\nu_{en}/\omega)^2 - \Omega_e^2/\omega^2},$$
$$g = -\frac{n_e}{n_c} \frac{\Omega_e/\omega}{(1 + i\nu_{en}/\omega)^2 - \Omega_e^2/\omega^2},$$
$$\varepsilon_\parallel = 1 - \frac{n_e}{n_c} \frac{1}{1 + i\nu_{en}/\omega}.$$

Here $\omega_{Le} = \sqrt{4\pi ne^2/m}$ is the Langmuir frequency, n_e is the electron density, e and m are their charge and mass, ν_{en} is the effective electron collision frequency, and $_e = e\mathbf{B}_z/mc$ is the cyclotron frequency. In an inhomogeneous medium [21]

$$\hat{\varepsilon}_{ij} = \Phi^{-1}T^{-1}\begin{pmatrix} \varepsilon_\perp & ig & 0 \\ -ig & \varepsilon_\perp & 0 \\ 0 & 0 & \varepsilon_\parallel \end{pmatrix} T\Phi, \qquad T = \begin{pmatrix} \cos\theta & 0 & -\sin\theta \\ 0 & 1 & 0 \\ \sin\theta & 0 & \cos\theta \end{pmatrix},$$

$$\Phi = \begin{pmatrix} \cos\varphi & \sin\varphi & 0 \\ -\sin\varphi & \cos\varphi & 0 \\ 0 & 0 & 1 \end{pmatrix}.$$

The electrodynamic models used in the calculations differed in the geometry of the excitation system, the configuration of the magnetic field, the frequency of the microwave, and the frequency of electron-neutral collisions.

6. Simulation results and their discussion

The results of numerical modeling of the equations discussed above working gas argon, pressure $4 \cdot 10^{-4}$ Torr, spatial distribution of constant magnetic field corresponded to that measured in the experiment (Figure 1), describing the diffusion and drift of charged particles and heat transfer in plasma in cylindrical geometry (azimuth distribution was considered uniform) showed the following.

The size at which the electron temperature equalization along and across the magnetic field occurs can be estimated from the theory of dimensions $L_\parallel \approx (\chi_\parallel/n_e)/w_1$ and $L_\perp \approx (\chi_\perp/n_e)/w_1$, where $\chi_{\parallel,\perp}/n_e$ are the thermal conductivity coefficients per electron along and across the magnetic field, and w1 is the energy lost by an electron in collisions per unit time. Furthermore, energy losses at the wall play a significant role in equalizing temperatures in space. In earlier studies, these energies were neglected when calculating spatial plasma density distributions due to the fact that the bulk of the electrons are reflected at the boundary from the resulting potential barrier, equalizing the electron and ion flows to the wall. Calculations showed that using plasma thermal insulation conditions leads to a significantly more uniform electron temperature distribution in space. Examples of calculating the spatial electron temperature distribution are shown in Figure 2.

It was assumed that electron heating occurs in a region of space near the resonator (the resonator center corresponds to the coordinate $z = 0$ in figure 2), and its intensity is independent of the radial coordinate.

Figure 3 shows a similar calculation for the case where heating occurs only in the central region of the plasma with a radius of 1 cm. It is evident that temperature equalization along the radius does not occur, indicating good thermal insulation of the plasma due to the magnetic field.

Figure 4 shows the calculated spatial distribution of the plasma density under the assumption of a constant spatial distribution of the electron temperature. A leveling of the electron density in the central region and a noticeable decrease in the region where the magnetic nozzle begins to form are noticeable.

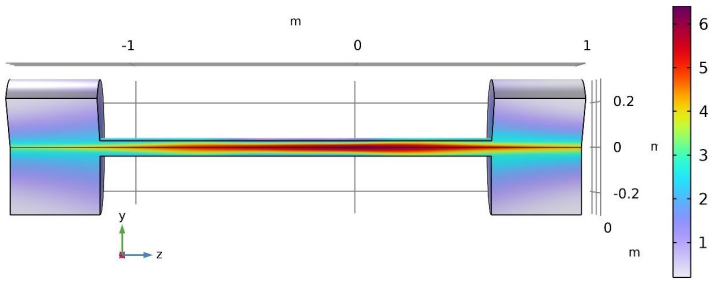

Figure 2. Electron temperature (eV) distribution in space. Energy is deposited uniformly across the cross section. All energy is deposited within a region of $|z| < 10$ cm relative to the resonator center

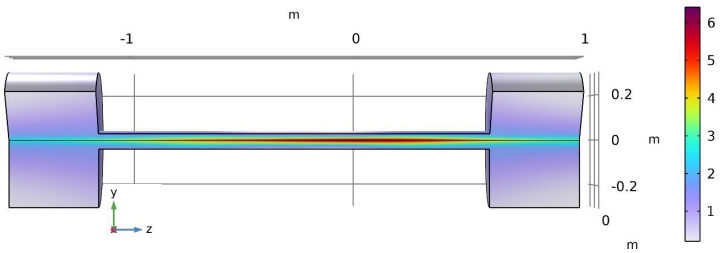

Figure 3. Electron temperature (in relative units) distribution in space. All energy is deposited within a region $|z| < 10$ cm and $|r| < 1$ cm relative to the resonator center

The spatial distribution of the electromagnetic field in the cavity was also calculated (figure 5). In [22], the cavity was excited using a slit in the side wall excited by a waveguide in the center; waves propagating in the azimuthal direction were excited, and the amplitude of the z-component of the electric field was small. In this case, the observed spatial distribution of the field has a more complex structure, with axial components of both the magnetic and electric fields present. Furthermore, various figures suggest the excitation of fields with azimuthal modes $m = 2$, 3, 4, and 5. When calculating the distribution of the electromagnetic field in the plasma near the cavity, the longitudinal distribution of the plasma density was considered constant, since the field is concentrated almost entirely in the region limited by the cavity due to the presence of cutoff waveguides surrounding the quartz tube, where the longitudinal inhomogeneity of the plasma is small.

Azimuthal non-uniformity of the magnetic field energy input may lead to the need to move from solving a two-dimensional axisymmetric problem to solving a three-dimensional one, which will take into account the more complex nature of the movement of charged particles, which is quite possible in a given range of working gas pressures and magnetic field strengths [23–27].

7. Conclusions

1. The paper formulates a simple discharge model based on the solution of the diffusion equations for charged particles, the energy balance equation for electrons, and Maxwell's equations. The solutions are not completely consistent, as the assumptions of uniform plasma heating by the microwave field inside the resonator, equalization of the electron temperature along magnetic

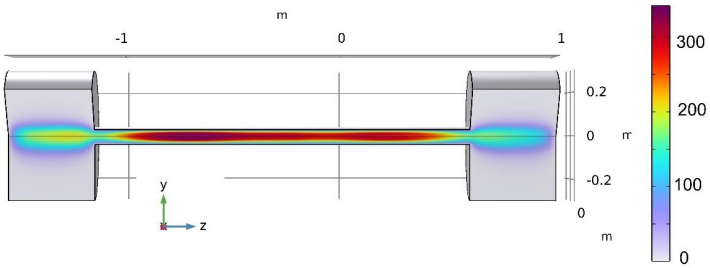

Figure 4. Distribution of electron density (in relative units) in the discharge at a constant electron temperature in space

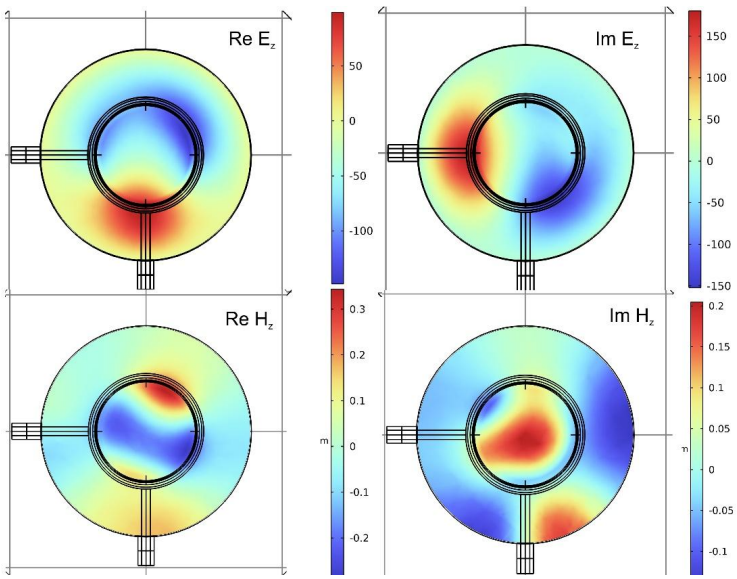

Figure 5. Distribution of the z-component of the electric (V/m) and magnetic (A/m) field in space in the excitation plane of the resonator. Electron density in the center of plasma is equal to 10^{10} cm^{-3}. The ratio of the effective frequency of electron collisions ν to the field frequency ω during calculation is 0.1. In-phase voltages with a frequency of 2.45 GHz and a voltage of 1 V are applied to the rod exciters. In the approximations used, Maxwell's equations are linear for a given electron density distribution, so the fields at other supplied powers increase or decrease proportionally to the gender of the exciting wave

lines, and uniformity of the longitudinal plasma distribution along the quarts pipeline were used to speed up the computation time.

2. Solutions to the heat conduction, diffusion, and Maxwell equations showed that the approximations used are satisfactorily fulfilled in the model under consideration, with the exception of the assumption of azimuthal heating homogeneity. Therefore, to assess the influence of this effect, it is necessary to complicate the model to a fully three-dimensional form.
3. The decrease in electron density near the working chamber may be due to the fact that parts of the field lines in the magnetic nozzle can pass through the boundaries of the quarts pipeline, which increases particle losses in this region.

Author Contributions: Conceptualization, Sergey A. Dvinin and Davlat K. Solikhzoda; methodology, Denis V. Chuprov, Konstantin N. Kornev and Zafari A. Qodirzoda; software, Sergey A. Dvinin and Zafari A. Qodirzoda; validation and visualisation, Konstantin N. Kornev, Zafari A. Qodirzoda; investigation, Sergey A. Dvinin, Zafari A. Qodirzoda, writing—original draft preparation, Sergey A. Dvinin D.V.Chuprov, Davlat K. Solikhzoda; writing—review and editing, Sergey A. Dvinin and Denis V. Chuprov. All authors have read and agreed to the published version of the manuscript.

Funding: The research was carried out with the support of the Ministry of Science and Higher Education of the Russian Federation (State Assignment No. FSSF-2026-0043) within the framework of the federal project "Development of technologies for controlled thermonuclear fusion and innovative plasma technologies".

Data Availability Statement: Data sharing is not applicable.

Conflicts of Interest: The authors declare no conflict of interest.

Declaration on Generative AI: The authors have not employed any Generative AI tools.

References

1. Alton, G. D. & Smithe, D. N. Design studies for an advanced ECR ion source. *Review of Scientific Instruments* **65,** 775–787. doi:10.1063/1.1144954. eprint: https://pubs.aip.org/aip/rsi/article-pdf/65/4/775/19216150/775_1_online.pdf (1994).
2. Asmussen, J., Grotjohn, T., Mak, P. & Perrin, M. The design and application of electron cyclotron resonance discharges. *IEEE Transactions on Plasma Science* **25,** 1196–1221. doi:10.1109/27.650896 (1997).
3. Yonesu, A., Shinohara, S., Yamashiro, Y. & Kawai, Y. Ion and neutral temperatures in an electron cyclotron resonance plasma. *Thin Solid Films* **390.** Proceedings of the 5th Asia-Pacific Conference on Plasma Science & Technology and the 13th Symposium on Plasma Science for Materials, 208–211. doi:10.1016/S0040-6090(01)00921-X (2001).
4. Muta, H., Koga, M., Itagaki, N. & Kawai, Y. Numerical investigation of a low-electron-temperature ECR plasma in Ar/N_2 mixtures. *Surface and Coatings Technology* **171.** Proceedings from the Joint International Symposia of the 6th APCPST, 15th SPSM, 4th International Conference on Open Magnetic Sytems for Plasma Confinement and 11th KAPRA, 157–161. doi:10.1016/S0257-8972(03)00261-5 (2003).
5. Koga, M., Yonesu, A. & Kawai, Y. Measurement of ion temperature in ECR Ar/N2 plasma. *Surface and Coatings Technology* **171.** Proceedings from the Joint International Symposia of the 6th APCPST, 15th SPSM, 4th International Conference on Open Magnetic Sytems for Plasma Confinement and 11th KAPRA, 216–221. doi:10.1016/S0257-8972(03)00274-3 (2003).
6. Kim, S. B., Kim, D. C., Namkung, W., Cho, M. & Yoo, S. J. Design and characterization of 2.45 GHz electron cyclotron resonance plasma source with magnetron magnetic field configuration for high flux of hyperthermal neutral beam. *Review of Scientific Instruments* **81,** 083301. doi:10.1063/1.3477998. eprint: https://pubs.aip.org/aip/rsi/article-pdf/doi/10.1063/1.3477998/15899550/083301_1_online.pdf (Aug. 2010).
7. Jauberteau, J.-L., Jauberteau, I., Cortázar, O. D. & Megía-Macías, A. Langmuir probe in magnetized plasma: Determination of the electron diffusion parameter and of the electron energy distribution function. *Contributions to Plasma Physics* **60,** e201900067. doi:10.1002/ctpp.201900067. eprint: https://onlinelibrary.wiley.com/doi/pdf/10.1002/ctpp.201900067 (2020).
8. Gammino, S. Production of High-Intensity, Highly Charged Ions, 123–164. doi:10.5170/CERN-2013-007.123. arXiv: 1410.7974 (2013).

9. Nakamura, T., Wada, H., Asaji, T. & Furuse, M. Effect of axial magnetic field on a 2.45 GHz permanent magnet ECR ion source. *Review of Scientific Instruments* **87,** 02A737. doi:10.1063/1.4937012. eprint: https://pubs.aip.org/aip/rsi/article-pdf/doi/10.1063/1.4937012/15842829/02a737_1_online.pdf (Dec. 2015).
10. Bogomolov, S. L., Bondarchenko, A. E., Efremov, A. A., *et al.* Production of High-Intensity Ion Beams from the DECRIS-PM-14 ECR Ion Source. *Physics of Particles and Nuclei Letters* **15,** 878–881. doi:10.1134/S1547477118070191 (2018).
11. Gammino, S., Celona, L., Ciavola, G., Maimone, F. & Mascali, D. Review on high current 2.45 GHz electron cyclotron resonance sources (invited)a). *Review of Scientific Instruments* **81,** 02B313. doi:10.1063/1.3266145. eprint: https://pubs.aip.org/aip/rsi/article-pdf/doi/10.1063/1.3266145/13935479/02b313_1_online.pdf (Feb. 2010).
12. Zhang, W. H. *et al.* A 2.45 GHz electron cyclotron resonance proton ion source and a dual-lens low energy beam transporta). *Review of Scientific Instruments* **83,** 02A329. doi:10.1063/1.3669802. eprint: https://pubs.aip.org/aip/rsi/article-pdf/doi/10.1063/1.3669802/15749851/02a329_1_online.pdf (Feb. 2012).
13. Fu, S., Ding, Z., Ke, Y. & Tian, L. Design Optimization and Experiment of 5-cm ECR Ion Thruster. *IEEE Transactions on Plasma Science* **PP,** 1–9. doi:10.1109/TPS.2020.2966662 (Feb. 2020).
14. Lieberman, M. A. & Lichtenberg, A. J. *Principles of Plasma Discharges and Material Processing* (Wiley, New York, 2005).
15. *Comsol Multiphysics. Reference Manual. Comsol Multyphysics. Programming Reference Manual* (2023).
16. Braginsky, S. I. *Transport equations in plasma* in *Problems of Plasma Theory* (ed Leontovich, M. A.) In Russian (1963).
17. Granovsky, V. L. *Electric current in gas, steady-state current* in *General properties of plasma* (ed V. L. Granovsky, A. K. M.) In Russian (Nauka, GRFML, Moscow, 1971).
18. Golyatina, H. V. & Mayorov, S. A. Analytical approximation of collision cross sections of electrons with atoms in inert gases. *Uspekhi Pricladnoy Fisiki* **9.** In Russian, 298–309. doi:10.51368/2307-4469-2021-9-4-298-309 (2021).
19. Alexandrov, A. F., Bogdankevich, L. S. & Rukhadze, A. A. *Principles of Plasma Electrodynamics* doi:10.1007/978-3-642-69247-5 (Springer-Verlag, Berlin, Heidelberg, New York, Tokyo, 1984).
20. *Plasma Electrodynamics* (ed Akhiezer, A. I.) In Russian (Nauka, GRFML, Moscow, 1974).
21. Mironov, V., Bogomolov, S., Bondarchenko, A., Efremov, A., Loginov, V. & Pugachev, D. Three-dimensional modelling of processes in Electron Cyclotron Resonance Ion Source. *Journal of Instrumentation* **15,** P10030. doi:10.1088/1748-0221/15/10/P10030 (2020).
22. Dvinin, S. A. & Korneeva, M. A. Numerical Simulation of the Spatial Structure of the Electromagnetic Field of a Microwave Discharge in a Magnetic Mirror Trap. *Plasma Phys. Rep.* **49,** 1448–1452. doi:10.1134/S1063780X23601438 (2023).
23. Kadomtsev, B. B. & Nedospasov, A. V. Instability of the positive column in a magnetic field and the 'anomalous' diffusion effect. *Journal of Nuclear Energy. Part C, Plasma Physics, Accelerators, Thermonuclear Research* **1,** 230. doi:10.1088/0368-3281/1/4/306 (1960).
24. Nedospasov, A. V. & Khait, V. D. *Oscillations and instabilities of low-temperature plasma* In Russian. 160 pp. (Nauka, GRFML, Moscow, 1979).
25. Nedospasov, A. V. & Khait, V. D. *Fundamentals of Physics of Processes in Devices with Low-Temperature Plasma* In Russian. 224 pp. (Energoatomizdat, Moscow, 1991).
26. Mikhailovsky, A. B. *Plasma Instabilities in Magnetic Traps* In Russian. 296 pp. (Atomizdat, Moscow, 1978).
27. Timofeev, A. V. & Shvilkin, B. N. Drift-dissipative instability of an inhomogeneous plasma in a magnetic field. *Phys. Usp.* **19,** 149–168. doi:10.1070/PU1976v019n02ABEH005134 (1976).

Information about the authors

Dvinin, Sergey A.—Doctor of Physical and Mathematical Sciences, Professor of Lomonosov Moscow State University, leading researcher of Institute of Physical Research and Technology of Peoples' Friendship University of Russia (RUDN University) (e-mail: dvininsa@phys.msu.ru, ORCID: 0000-0002-0163-9282, ResearcherID: J-6595-2012, Scopus Author ID: 6602388907)

Chuprov, Denis V.—Senior Lecturer, Research Associate of Institute of Physical Research and Technology of Peoples' Friendship University of Russia (RUDN University) (e-mail: chuprov-dv@rudn.ru, ORCID: 0000-0002-6768-6196, ResearcherID: O-3193-2013, Scopus Author ID: 6508067157)

Kornev, Konstantin N.—Lead engineer, of Lomonosov Moscow State university, research intern of Institute of Physical Research and Technology of Peoples' Friendship University of Russia (RUDN University) (e-mail: singuliarnost@yandex.ru, ORCID: 0000-0002-7574-566X, Scopus Author ID: 57213826116)

Qodirzoda, Zafari A.—Candidate of Science, Associate Professor of Tajik National University (e-mail: zafar.kodirzoda@yandex.ru, ORCID: 0009-0004-2276-3786, ResearcherID: NMK-4101-2025, Scopus Author ID: 57220783014)

Solikhzoda, Davlat K.—Doctor of Science, Professor of Tajik National University (e-mail: davlat56@mail.ru, ORCID: 0009-0006-8624-3274, Scopus Author ID: 57215526726)

УДК 537.527,533.9.03
PACS 52.80 Pi, 52.80.Sm, 52.50.Sw, 52.40.Db
DOI: 10.22363/2658-4670-2026-34-1-125-138 EDN: UOSEFX

Математические модели разряда низкого давления в магнитном поле, поддерживаемого быстропеременным электромагнитным полем

С. А. Двинин[1,2], Д. В. Чупров[2], К. Н. Корнев[1,2], З. А. Кодирзода[3], Д. К. Солихзода[3]

[1] Российский университет дружбы народов, ул. Миклухо-Маклая, д. 6, Москва, 117198, Российская Федерация

[2] Московский Государственный университет имени М. В. Ломоносова, Ленинские Горы д. 1 стр. 2, Москва, 119991, Российская Федерация

[3] Таджикский национальный университет, Проспект Рудаки, д. 17, Душанбе, 973402, Таджикистан

Аннотация. Разряды использующие электронный циклотронный резонанс (ЭЦР) для нагрева электронов, представляют собой эффективный способ создания плазмы при низком давлении рабочего газа. Цель данной работы — разработка математической модели ЭЦР разряда, реализованного на установке RAPIRA (РУДН), применяемой для реализации целого ряда научных исследований. Эволюция частиц плазма описывается в рамках гидродинамического приближения/ (двумерная модель с цилиндрической симметрией), При расчете пространственного распределения электромагнитного поля используется трехмерная модель холодной плазмы. Расчеты показали, что в рабочем режиме установки (давления газа от $4 \cdot 10^{-4}$ до 10^{-2} Торр, магнитное поле до 2500 Гс) происходит выравнивание температуры электронов вдоль силовых магнитного поля, и в то же время магнитное поле обеспечивает уменьшение потерь энергии на боковые стенки установки. Рассчитаны пространственные распределения плотности и температуры электронов и электромагнитного поля в плазме. Реализованная модель может служить основой для разработки более совершенного набора программных кодов, учитывающих немаксвелловскую природу функции распределения скоростей электронов, обусловленную неадиабатическим характером их нагрева в неоднородном магнитном поле.

Ключевые слова: электронный циклотронный резонанс, ЭЦР-разряд, разряд в резонаторе, разряд в магнитной ловушке, дрейфово-диффузионная модель

Discrete & Continuous Models
& Applied Computational Science
ISSN 2658-7149 (Online), 2658-4670 (Print)

2026, 34 (1) 139–144
http://journals.rudn.ru/miph

Research article

UDC 51-72.530.145
PACS 07.05.Tp
DOI: 10.22363/2658-4670-2026-34-1-139-144 EDN: UNJXAC

Solution of the one-dimensional Schrödinger equation for a heterostructure with a triangular potential function by the power series method

Irina N. Belyaeva[1], Nikolay A. Chekanov[1], Roman V. Korotenko[1], Natalia N. Chekanova[2]

[1] Belgorod State National Research University, 85 Pobedy St, Belgorod, 308015, Russian Federation

[2] Kharkov National University named after V.N. Karazin, 1 Mironositskaya St, Kharkov, 61001, Ukraine

(received: February 4, 2026; revised: February 16, 2026; accepted: February 20, 2026)

Abstract. In the work by the power series method the one-dimensional Schrödinger equation is solved with a triangular potential function which is applied in various modern heterostructures, in particular for GaAs and the others. By varying available parameters it is possible to obtain the desired precision of the numerical solution of the Schrödinger equation with any type of potential function for modern heterostructures. For the original Schrödinger equation are obtained wave functions in the form Airy functions and the analytical formula for the energy levels through the zeros of the Airy function. The values energy levels from this analytical formula agree with its results obtained by direct power series method with precision up to 10^{-4} percents, that is, up to 5 decimal signs. However, it is more rational and easier to use the Schrödinger equation solution, because the numerical calculations zeros of Airy function present separate complex and complicated numerical problem. But in order to achieve high numerical accuracy, it is necessary to set the Digits flag to several dozen significant digits and increasing the number of power series, that leads to an increasing in the time spent on the computer.

Key words and phrases: Schrödinger equation, triangular potential function, heterostructures, energy levels, wave functions, the Airy equation, zeros of the Airy function, power series, mathematical modeling, the Maple computer system

For citation: Belyaeva, I. N., Chekanov, N. A., Korotenko, R. V., Chekanova, N. N. Solution of the one-dimensional Schrödinger equation for a heterostructure with a triangular potential function by the power series method. *Discrete and Continuous Models and Applied Computational Science* **34** (1), 139–144. doi: 10.22363/2658-4670-2026-34-1-139-144. edn: UNJXAC (2026).

1. Introduction

In this work, the one-dimensional Schrödinger equation with a triangular potential function has been solved using the power series method [1–4], which is used in the study of semiconductor nano-dimensional structures in the field of modern advanced microelectronics for the creation of new devices and devices in various fields of technology [5–16]. However, heterostructures are complex quantum systems with many quantum features. For example, the heterostructure between the layers $GaAs$ and $Al_xGa_{1-x}As$ electrons are in the triangular potential well [13, 17, 18].

A very promising direction is their use in the field of the new generation of microelectronics for the creation of devices and devices that will become elements of large integrated circuits capable of storing huge amounts of information and processing them at high speed and will form the basis of a new generation of electronic and optoelectronic machines of small sizes [6, 12, 15]. Motion of electrons in these structures is essentially described by the laws of quantum mechanics, and various quantum models have been developed to describe them [5–10, 12, 13, 15].

2. Solving of the basic equations

In our work, we solved the Schrödinger equation with the triangular potential function as

$$V(x) = \begin{cases} \alpha x, & \alpha = \mid e \mid \cdot \mid \vec{E} \mid, x > 0, \\ \infty, & x \leq 0. \end{cases} \tag{1}$$

In the atomic system of units($m = e = \hbar = 1$), the Schrödinger equation has the form

$$\left[-\frac{1}{2} \cdot \frac{d^2}{dx^2} + V(x)\right] \psi(x) = E\psi(x), \tag{2}$$

where

$$\psi(0) = 0, \qquad \psi(\infty) \longrightarrow 0. \tag{3}$$

Is boundary condition. Here e is the elementary charge, $\vec{E}$ is the electric field strength. The integration of equation (1)–(2) is performed on the segment $[R_{\text{left}}; R_{\text{right}}]$ with the help of a developed computer program [1] in the Maple system. Our maple program have three parameters R_{left}; R_{right}, and n-number of member in power series. By variation of these parameters one can achieved desirable exactness.

The optimal cut-off values of segment select in our calculations by variation method and it are equal $R_{\text{left}} = -0.28 \cdot 10^{-26}$ *и* $R_{\text{right}} = 13.5$ and with the number of members in the power series equal to $n = 200$. As it know, the power series method first calculates two linearly independent solutions and , which depend on the total energy as a parameter. Their linear combination gives the general solution of the Schrödinger equation (2). Consideration of the boundary conditions (3) leads to a homogeneous algebraic system, the nontrivial solutions of which are given by the allowable energy levels and the corresponding wave functions. The following lower energy levels were calculated. The optimal cut-off values were the target of selection and in our calculations are equal and with the number of members in the power series equal . As know, in the power series method first calculates two linearly independent solutions $\psi_1(x, E)$ and $\psi_2(x, E)$, which depend on the total energy as a parameter. Their linear combination gives the general solution of the Schrödinger equation (2). Taking into account the boundary conditions (3) leads to a homogeneous algebraic system, the nontrivial solutions of which are given by the allowable energy levels and the corresponding wave functions. If $\alpha = 1$ the following lower energy levels were calculated:

$$E_k = 1.855575;\ 3.24446;\ 4.381671;\ 5.386613;\ 6.305263.$$

and the corresponding wave functions, which because of their bulkiness are represented in the following form:

$$\psi_1(x) = 0.61139759 \cdot 10^{-5}x - 0.142163116 \cdot 10^{-30}x^2 - 0.0000169241805x^3 + \cdots$$
$$- 0.593525574 \cdot 10^{-15}x^{25} + 0.590773946 \cdot 10^{-16}x^{26} + 0.171191325 \cdot 10^{-31}x^{27}$$
$$\psi_2(x) = 0.136823205 \cdot 10^{-6}x - 0.31814343910 \cdot 10^{-32}x^2 - 0.378742190 \cdot 10^{-6}x^3 + \cdots$$
$$- 0.132823670 \cdot 10^{-16}x^{25} + 0.132207889 \cdot 10^{-17}x^{26} + 0.383104973 \cdot 10^{-33}x^{27}$$
$$\psi_3(x) = 0.353019093 \cdot 10^{-9}x - 0.820845473 \cdot 10^{-35}x^2 - 0.977196991 \cdot 10^{-9}x^3 + \cdots$$
$$- 0.198239895 \cdot 10^{-17}x^{22} + 0.999308236 \cdot 10^{-18}x^{23} - 0.131621212 \cdot 10^{-19}x^{24} -$$
$$- 0.342699846 \cdot 10^{-19}x^{25} + \cdots$$

It is shown also, that initial problem admit the analytical solution. Indeed, rewrite equation (2) in the form

$$\psi''_{xx} - 2\alpha\left(x - \frac{E}{\alpha}\right)\psi(x) = 0$$

and do following substation:

$$z = \beta\left(x - \frac{E}{\alpha}\right), \qquad \left(x - \frac{E}{\alpha}\right) = \frac{z}{\beta}.$$

Then by $\beta^3 = 2\alpha$ initial equation (2) bring to [19–21]:

$$\psi''_{zz} - z\psi(z) = 0. \tag{4}$$

Solution of this equation (4) will be known function Airy and solution initial problem (2) - (3) will be following wave function:

$$\psi(x) = \text{const} \cdot Ai\,[z(x)] = \text{const} \cdot Ai\left[\beta\left(x - \frac{E}{\alpha}\right)\right]. \tag{5}$$

From (5) and (3) obtain equality $\beta \cdot E = -\alpha \cdot z_k$, where z_k are zeros of Airy function $Ai(z_k) = 0$. And thus we have analytical expression for energy levels in atomic units:

$$E_k = -(z_{k+1}) \cdot \sqrt[3]{\frac{\alpha^2}{2}}, \qquad k = 0, 1, 2, \ldots \tag{6}$$

The values energy levels from formula (6) agree with its results obtained by direct power series method with precision up to $10^{-4}\%$. However, it is more rational and faster to calculate the energy levels with the help of direct solution of the Schrödinger's equation by some known method [2].

3. Results

A computer program of symbolic-numerical solution of the one-dimensional Schrödinger equation is developed, and calculations of energy levels and wave functions of a perspective gallium arsenide semiconductor with a triangular potential function are carried out, which is experimentally detected for electrons at the boundary between layers of this semiconductor.

4. Discussion

It is shown that the Schrödinger equation with a triangular potential function admits analytical solutions, both for wave functions and for the energy spectrum. In particular, an analytical formula for energy levels is obtained. which uses the zeros of the Airy function. In the calculations, it was found that the energy levels obtained by direct numerical calculation of the Schrödinger equation practically coincide with their values calculated by the analytical formula (five decimal places coincide). It should be pointed out that the developed method for solving the Schrödinger equation is quite applicable for calculations with other types of potential functions in other heterostructures. It can be hoped that the results of the calculations and the developed program will be applied in the field of modern research on semiconductors.

5. Conclusions

Thus, this program finds the solution of the Schrödinger equation with high precision by variations of its three parameters: $(R_{\text{left}}, R_{\text{right}}, n)$ and the Digits commands from the Maple system provide high precision in solving the Schrödinger equation with other potential functions that are used in heterostructures. Thus, It has been shown that the developed method of solving the Schrödinger equation allows for high numerical accuracy and our program can be made available to interested parties.

Author Contributions: The contributions of the authors are equal. All authors have read and agreed to the published version of the manuscript.

Funding: This research received no external funding.

Data Availability Statement: Data sharing is not applicable.

Conflicts of Interest: The author declares no conflict of interest.

Declaration on Generative AI: The author has not employed any generative AI tools.

References

1. Belyaeva, I. N., Ukolov, Y. A. & Chekanov, N. A. Construction general solution of differential equations for Fuchs type as power series. Russian. *Registered in Fond of algorithm and program - Moscow,* 50200500089 (2005).
2. Belyaeva, I. N., Chekanov, N. A., Chekanova, N. N., Kirichenko, I. K. & Yarho, T. *The methods for solving differential equations of classical and quantum mechanics* 183 pp. (Kharkov, KNAHU, 2021).
3. Trikomi, F. *Differential Equations* Russian. 352 pp. (Moscow: Foreign Literature Publishing House, 1962).
4. Tikhonov, A. N., Vasilyeva, A. B. & Sveshnikov, A. G. *Differential Equations* Russian. 231 pp. (Moscow: Nauka, 1980).
5. Lesovik, G. *Electronic Transport in Meso- and Nano-Scale Conductors* 156 pp. (ETH Zurich, Herbstsemester, 2008).
6. Datta, S. *Quantum Transport: Atom to Transistor* 420 pp. (Cambridge University Press, 2005).
7. Harrison, P. *Quantum Wells, Wires and Dots* 502 pp. (The University of Leeds, UK, John Wiley and Sons, LTD, 2005).
8. Ledentsov, N. N., Ustinov, V. M., Schukin, V. A., Kopiev, P. S., Alferov, Z. I. & Bimberg, D. Heterostructures with quantum dots: fabrication, properties, and lasers. *Semiconductors* **32,** 343–365 (1998).

9. Datta, S. *Electronic Transport in Mesoscopic Systems* 318 pp. (Cambridge University Press, 1995).
10. Shik, A. Y., Bakueva, L. G., Musikhin, S. F. & Rykov, S. A. *Physics of low-dimensional systems* Russian. 377 pp. (St.-Petersburg: Nauka, 1969).
11. Khludkov, S. S., Tolbanov, O. P., Vilisova, M. D. & Prudaev, I. A. *Semiconductor devices based on gallium arsenide with deep impurity centers* Russian. 258 pp. (Tomsk: Publishing House of Tomsk State University, 2016).
12. Demikhovskii, V. Y. Quantum wells, wires and dots. What's this? *Soros Educational Journal* **5,** 80–86 (1997).
13. Demikhovskii, V. Y. & Vugalter, G. A. *The physics of quantum low-dimensional structures* Russian. 248 pp. (Moscow: Logos, 2000).
14. Tavger, B. A. & Demikhovskii, V. Y. Quantum size effects in semiconductor and semimetallic films. *Soviet Physics Uspekhi* **11,** 644–658. doi:10.1070/PU1969v011n05ABEH003739 (1969).
15. Neverov, V. N. & Titov, A. N. *The physics of quantum low-dimensional systems* Russian. 240 pp. (Ekaterinburg: Center "Nanotechnology and perspective material", 2008).
16. Alferov, Z. I. The History and Future of semiconductor heterostructures. *Semiconductors* **32,** 1–14 (1998).
17. Weiss, D. & Richter, K. Complex and quantized electron motion in antidot arrays. *Physica D* **83,** 290–298 (1995).
18. Stockmann, H.-J. *Quantum Chaos: An Introduction* 368 pp. (Cambridge University Press, 1999).
19. Airy, G. B. On the intensity of light in the neighbourhood of a caustic. *Transactions of the Cambridge Philosophical Society* **6,** 379–402 (1838).
20. Abramowitz, M. & Stegun, I. A. *Handbook of Mathematical Functions with Formulas, Graphs and Mathematical Tables* Russian. 832 pp. (Moscow: Nauka, 1979).
21. Davies, J. H. *The Physics of Low-Dimensional Semiconductors: An Introduction* 438 pp. (Cambridge University Press, 1998).

Information about the authors

Belyaeva, Irina N.—Docent, Candidate of Sciences in Physics and Mathematics, Associate Professor of department of Mathematics of Faculty of Mathematics and Science of Institute of Pedagogy (e-mail: ibelyaeva@bsuedu.ru, ORCID: 0000-0002-7674-1716)

Chekanov, Nikolay A.—Professor, Doctor of Sciences in Physics and Mathematics, Professor of department of Mathematics of Faculty of Mathematics and Science of Institute of Pedagogy (e-mail: nikchek137@gmail.com, ORCID: 0000 0003 1131 3195)

Korotenko, Roman V.—Student of department of Mathematics of Faculty of Mathematics and Science of Institute of Pedagogy (e-mail: 1474589@bsuedu.ru, ORCID: 0009-0002-6353-3552)

Chekanova, Natalia N.—Candidate of Sciences in Physics and Mathematics, Associate Professor of Department Information Technology and Mathematic Modeling (e-mail: natchek1976@gmail.com, ORCID: 0000-0001-9134-2951)

УДК 51-72.530.145
PACS 07.05.Tp
DOI: 10.22363/2658-4670-2026-34-1-139-144 EDN: UNJXAC

Решение одномерного уравнения Шредингера для гетероструктур с треугольной потенциальной функцией методом степенных рядов

И. Н. Беляева[1], Н. А. Чеканов[1], Р. В. Коротенко[1], Н. Н. Чеканова[2]

[1] Белгородский государственный национальный исследовательский университет, ул. Победы, д. 85, Белгород, Российская Федерация

[2] Харьковский национальный университет имени В. Н. Каразина, ул. Мироносицкая, д. 1, Харьков, 61001, Украина

Аннотация. В работе методом степенных рядов решается одномерное уравнение Шрёдингера с треугольной потенциальной функцией, которая применяется в различных современных гетероструктурах, в частности для GaAs и других. Варьируя доступные параметры, можно получить желаемую точность численного решения уравнения Шрёдингера с любым типом потенциальной функции для современных гетероструктур. Для исходного уравнения Шрёдингера получены волновые функции в виде функций Эйри и аналитическая формула для уровней энергии с помощью нулей функции Эйри. Значения энергетических уровней из этой аналитической формулы согласуются с результатами, полученными методом прямых степенных рядов, с точностью до 10^{-4} процентов, то есть до 5 десятичных знаков. Однако рациональнее и проще использовать решение уравнения Шрёдингера. Но для достижения высокой точности вычислений необходимо установить флажок Digits на несколько десятков значащих цифр и увеличить количество членов степенного ряда, что приводит увеличению времени счета на компьютере.

Ключевые слова: уравнение Шрёдингера, треугольная потенциальная функция, гетероструктуры, энергетические уровни, волновые функции, уравнение Эйри, нули функции Эйри, степенные ряды, математическое моделирование, компьютерная система Maple

Discrete & Continuous Models & Applied Computational Science
ISSN 2658-7149 (Online), 2658-4670 (Print)

2026, 34 (1) 145–149
http://journals.rudn.ru/miph

Research article

UDC 519.87
DOI: 10.22363/2658-4670-2026-34-1-145-149 EDN: UYXCCK

A model of cumulative advantage for conference dynamics

Anna M. Ermolayeva

RUDN University, 6 Miklukho-Maklaya St, Moscow, 117198, Russian Federation

(received: December 15, 2025; revised: January 10, 2026; accepted: January 16, 2026)

Abstract. This paper attempts to modify the standard Verhulst model to describe the dynamics of scientific conferences taking into account cumulative advantage.

Key words and phrases: scientometrics, conferences, the Matthew law

For citation: Ermolayeva, A. M. A model of cumulative advantage for conference dynamics. *Discrete and Continuous Models and Applied Computational Science* **34** (1), 145–149. doi: 10.22363/2658-4670-2026-34-1-145-149. edn: UYXCCK (2026).

1. Introduction

The problem of conference evaluation is currently very pressing for researchers in the field of scientometrics, as there is no universal methodology for evaluating all conferences in all fields. Several conference rankings exist, such as the Australian CORE, the Chinese CCF Conference Rankings, the Brazilian QUALIS, and the industry-specific Microsoft Academic Conference Rankings. All of these rankings are compiled for computer science conferences, due to the extremely important nature of conferences in this field, as over 60% of research results are published in conference proceedings.

A study of the development of scientific conferences showed that conferences develop unevenly, with some becoming stellar, while others quickly fade away. This led us to use the standard Verhulst model [1] for our study, but to expand it by taking into account the Matthew effect [2].

2. Basic model

Let there be n conferences in a given scientific field. Let $R^i(t) \geqslant 0$ denote the numerical measure of the ranking of the ith conference at time t. The ranking is considered as a single aggregate value.

Let the ranking dynamics of each conference be determined by the following mechanisms:

- internal growth;
- competition;
- natural decay;
- external influences.

Internal growth is associated with the desire to increase ratings through internal efforts (attracting renowned speakers, improving the quality of peer review, and improving organization). Competition is caused by the mutual inhibition of conferences, as resources (people, money) are limited. Natural decay (dissipation) is associated with obsolescence, loss of relevance, etc. External influences are caused by unpredictable factors (black swans) (e.g., changes in program committees, publication of breakthrough results, scandals).

We will use the Verhulst model for an isolated conference as a basis:

$$\frac{dR}{dt} = rR\left(1 - \frac{R}{K}\right) - \delta R,$$

where r is the maximum growth rate, K is the capacity (the maximum achievable rating in the absence of competitors), δ is the decay coefficient.

Then the equilibrium rating is:

$$R^* = K(1 - \delta/r), \qquad r > \delta.$$

We'll introduce competitive inhibition for several conferences. The growth of each conference is slowed not only by its own rankings, but also by the rankings of other conferences.

$$\frac{dR^i}{dt} = r_i R^i \left(1 - \frac{\sum_{j=1}^{n} \alpha_{ij} R^j}{K_i}\right),$$

where α_{ij} is the coefficient of influence of conference j on conference i. It is natural to assume $\alpha_{ii} = 1$. α_{ij} for $i \neq j$ shows how strongly competitors suppress the growth of the ith conference.

Let's add attenuation and external influences:

$$\frac{dR^i}{dt} = r_i R^i \left(1 - \frac{\sum_{j=1}^{n} \alpha_{ij} R^j}{K_i}\right) - \delta_i R^i + \gamma_i F_i(t),$$

where:

- $r_i > 0$ — potential growth rate,
- $K_i > 0$ — maximum possible rating in the absence of competitors,
- $\alpha_{ij} \geqslant 0$ — competition coefficients,
- $\delta_i \geqslant 0$ — natural decay rate,
- $\gamma_i \geqslant 0$ — sensitivity to external influences,
- $F_i(t) \geqslant 0$ — external impulse function.

In a more compact form, we can rewrite:

$$\frac{dR^i}{dt} = R^i \left(r_i - \frac{r_i}{K_i} \sum_{j=1}^{n} \alpha_{ij} R^j - \delta_i\right) + \gamma_i F_i(t).$$

The term $-\frac{r_i}{K_i} \alpha_{ij} R^i R^j$ describes mutual inhibition.

3. Accounting for Matthew's law

Matthew's Law (for to everyone who has, more will be given, and he will have abundance; but from him who does not have, even what he has will be taken away) is a manifestation of cumulative advantage. The higher a conference's rating, the easier it is to attract the best authors, receive more citations, and, consequently, further increase its rating.

3.1. Dependence on the current rating

Let's add the term $\beta_i R_\theta^i$, which increases the growth rate proportionally to the current rating:

$$\frac{dR^i}{dt} = r_i R^i \left(1 - \frac{\sum \alpha_{ij} R^j}{K_i}\right) + \beta_i R_\theta^i - \delta_i R^i + \gamma_i F_i(t),$$

where:

- $\beta_i \geqslant 0$ — intensity of the cumulative advantage,
- $\theta > 0$ — nonlinearity index.

3.2. Capacity dependence on rating

Let's make the capacity K_i dependent on the rating:

$$K_i = K_i^{(0)} + \varkappa_i R^i,$$

where $\varkappa_i \geqslant 0$.

Then the logistical constraint becomes less severe for the leaders, which facilitates their further growth.

3.3. Asymmetric competition

To take Matthew's law into account, we can make α_{ij} dependent on the difference in ratings:

$$\alpha_{ij} = \alpha_{ij\,\text{base}} \cdot \exp\bigl(-\lambda(R^i - R^j)\bigr), \qquad R^i > R^j,$$

where $\lambda > 0$.

A conference with a higher rating R^i experiences less inhibition from a conference with a lower R^j. This creates a positive feedback loop: leaders become less vulnerable to competitors.

3.4. Threshold effect

If the cumulative advantage is very strong, the system may exhibit bistability. The conference either becomes a leader or remains at a low level. To achieve this, we add a threshold term:

$$\frac{dR^i}{dt} = R^i \bigl(\rho_i(R^i - R_{\text{th}})\bigr) - \delta_i R^i + \ldots$$

3.5. Cumulative advantage

Cumulative advantage means that the rating increase is proportional to the current rating (or its degree) with a positive coefficient. In the simplest case ($\beta_i > 0$, $\theta = 1$) at the initial stage we obtain:

$$\frac{dR^i}{dt} \approx (r_i + \beta_i) R^i.$$

This leads to exponential growth, limited only by the capacity K_i and competition. If two conferences initially have similar parameters, but one gains a small advantage ε, this advantage will increase over time, and the system may converge to an equilibrium with a strong leader dominance. The equilibrium becomes unstable, and a "winner-takes-all" regime emerges.

Author Contributions: All authors have read and agreed to the published version of the manuscript.

Funding: This research received no external funding.

Data Availability Statement: Data can be sent by the authors on reasonable request.

Conflicts of Interest: The authors declare no conflict of interest. The funders had no role in the design of the study; in the collection, analyses, or interpretation of data; in the writing of the manuscript; or in the decision to publish the results.

Declaration on Generative AI: The author has not employed any generative AI tools.

References

1. Verhulst, P. F. *Notice sur la loi que la population suit dans son accroissement* 113–117 (1838).
2. Merton, R. K. The Matthew Effect in Science: The reward and communication systems of science are considered. *Science* **159,** 56–63. doi:10.1126/science.159.3810.56 (Jan. 1968).

Information about the authors

Ermolayeva, Anna M.—Assistent Professor of Department of Probability Theory and Cyber Security of RUDN University (e-mail: ermolaeva-am@rudn.ru, ORCID: 0000-0001-6107-6461)

УДК 519.87
DOI: 10.22363/2658-4670-2026-34-1-145-149 EDN: UYXCCK

Модель динамики конференций с учётом кумулятивного преимущества

А. М. Ермолаева

Российский университет дружбы народов, ул. Миклухо-Маклая, д. 6, Москва, 117198, Российская Федерация

Аннотация. В статье делается попытка модифицировать стандартную модель Ферхюльста для описания динамики научных конференций с учётом кумулятивного преимущества.

Ключевые слова: наукометрия, конференции, эффект Матфея