Iranian Journal of Numerical Analysis and Optimization

Volume 13, Number 4

December 2023

Serial Number: 27

Ferdowsi University of Mashhad, Iran

In the Name of God

Iranian Journal of Numerical Analysis and Optimization (IJNAO)

This journal is authorized under the registration No. 174/853 dated 1386/2/26 (2007/05/16), by the Ministry of Culture and Islamic Guidance.

Volume 13, Number 4, December 2023

ISSN-Print: 2423-6977, **ISSN-Online:** 2423-6969

Publisher: Faculty of Mathematical Sciences, Ferdowsi University of Mashhad

Published by: Ferdowsi University of Mashhad Press

Printing Method: Electronic

Address: Iranian Journal of Numerical Analysis and Optimization

Faculty of Mathematical Sciences, Ferdowsi University of Mashhad

P.O. Box 1159, Mashhad 91775, Iran.

Tel. : +98-51-38806222 , **Fax:** +98-51-38807358

E-mail: ijnao@um.ac.ir

Website: http://ijnao.um.ac.ir

This journal is indexed by:

- SCOPUS
- ZbMATH Open
- ISC
- DOAJ
- SID
- Civilica
- Magiran
- Mendeley
- Academia.edu
- Linkedin

- The Journal granted the International degree by the Iranian Ministry of Science, Research, and Technology.

Iranian Journal of Numerical Analysis and Optimization

Volume 13, Number 4, December 2023

Ferdowsi University of Mashhad - Iran

Iranian Journal of Numerical Analysis and Optimization

Gachpazan, Mortaza**
(Numerical Analysis)
Ferdowsi University of Mashhad, Iran.
e-mail: gachpazan@um.ac.ir

Ghanbari, Reza**
(Operations Research)
Ferdowsi University of Mashhad, Iran.
e-mail: rghanbari@um.ac.ir

Hadizadeh Yazdi, Mahmoud**
(Numerical Analysis)
Khaje-Nassir-Toosi University of Technology, Iran.
e-mail: hadizadeh@kntu.ac.ir

Hojjati, Gholamreza*
(Numerical Analysis)
University of Tabriz, Iran.
e-mail: ghojjati@tabrizu.ac.ir

Hong, Jialin*
(Scientific Computing)
Chinese Academy of Sciences (CAS), China.
e-mail: hjl@lsec.cc.ac.cn

Karimi, Hamid Reza*
(Control)
Politecnico di Milano, Italy.
e-mail: hamidreza.karimi@polimi.it

Khojasteh Salkuyeh, Davod*
(Numerical Analysis)
University of Guilan, Iran.
e-mail: khojasteh@guilan.ac.ir

Lohmander, Peter*
(Optimization)
Swedish University of Agricultural Sciences, Sweden.
e-mail: Peter@Lohmander.com

Lopez-Ruiz, Ricardo**
(Complexity, nonlinear models)
University of Zaragoza, Spain.
e-mail: rilopez@unizar.es

Mahdavi-Amiri, Nezam*
(Optimization)
Sharif University of Technology, Iran.
e-mail: nezamm@sina.sharif.edu

Mirzaei, Davoud*
(Numerical Analysis)
University of Uppsala, Sweden.
e-mail: davoud.mirzaei@it.uu.se

Omrani, Khaled*
(Numerical Analysis)
University of Tunis El Manar, Tunisia.
khaled.omrani@issatso.rnu.tn

Salehi Fathabadi, Hasan*
(Operations Research)
University of Tehran, Iran.
e-mail: hsalehi@ut.ac.ir

Soheili, Ali Reza*
(Numerical Analysis)
Ferdowsi University of Mashhad, Iran.
e-mail: soheili@um.ac.ir

Soleimani Damaneh, Majid*
(Operations Research and Optimization, Finance, and Machine Learning)
University of Tehran, Iran.
e-mail: m.soleimani.d@ut.ac.ir

Toutounian, Faezeh*
(Numerical Analysis)
Ferdowsi University of Mashhad, Iran.
e-mail: toutouni@um.ac.ir

This journal is published under the auspices of Ferdowsi University of Mashhad

We would like to acknowledge the help of Miss Narjes khatoon Zohorian in the preparation of this issue.

Letter from the Editor–in–Chief

I would like to welcome you to the Iranian Journal of Numerical Analysis and Optimization (IJNAO). This journal has been published two issues per year and supported by the Faculty of Mathematical Sciences at the Ferdowsi University of Mashhad. The faculty of Mathematical Sciences with the centers of excellence and the research centers is well-known in mathematical communities in Iran.

The main aim of the journal is to facilitate discussions and collaborations between specialists in applied mathematics, especially in the fields of numerical analysis and optimization, in the region and worldwide. Our vision is that scholars from different applied mathematical research disciplines pool their insight, knowledge, and efforts by communicating via this international journal. In order to assure the high quality of the journal, each article is reviewed by subject-qualified referees. Our expectations for IJNAO are as high as any well-known applied mathematical journal in the world. We trust that by publishing quality research and creative work, the possibility of more collaborations between researchers would be provided. We invite all applied mathematicians especially in the fields of numerical analysis and optimization to join us by submitting their original work to the Iranian Journal of Numerical Analysis and Optimization.

We would like to inform all readers that the Iranian Journal of Numerical Analysis and Optimization (IJNAO), has changed its publishing frequency from "Semiannual" to a "Quarterly" journal since January 2023. The four journal issues per year will be published in the months of March, June, September, and December. One of our goals is to continue to improve the speed of both the review and publication processes, while try continuing to publish the best available international research in numerical analysis and optimization, with the high scientific and publication standards that the journal is known for.

Ali R. Soheili

Editor-in-Chief

Contents

Iranian Journal of Numerical Analysis and Optimization
Vol. 13, No. 4, 2023, pp 578–603
https://doi.org/10.22067/ijnao.2023.79288.1192
https://ijnao.um.ac.ir/

Research Article

A generalized form of the parametric spline methods of degree $(2k+1)$ for solving a variety of two-point boundary value problems

Z. Sarvari

Abstract

In this paper, a high order accuracy method is developed for finding the approximate solution of two-point boundary value problems. The present approach is based on a special algorithm, taken from Pascal's triangle, for obtaining a generalized form of the parametric splines of degree $(2k+1)$, $k = 1, 2, \ldots$, which has a lower computational cost and gives the better approximation. Some appropriate band matrices are used to obtain a matrix form for this algorithm.

The approximate solution converges to the exact solution of order $O(h^{4k})$, where k is a quantity related to the degree of parametric splines and the number of matrix bands that are applied in this paper. Some examples are given to illustrate the applicability of the method, and we compare the computed results with other existing known methods. It is observed that our approach produced better results.

AMS subject classifications (2020): Primary 45D05, Secondary 42C10, 65G99.

Keywords: Boundary value problems; Parametric spline; Band matrices; Pascal's triangle.

1 Introduction

Spline function is a piecewise polynomial satisfying certain conditions of the continuity of the function and its derivatives. In other words, a spline function

Received 20 October 2022; revised 25 March 2023; accepted 27 March 2023

Zahra Sarvari
Department of Applied Mathematics, Azarbaijan Shahid Madani University, Tabriz, Iran. e-mail: zsarvari8@gmail.com

$S(x)$ of degree d is defined in a region $[a, b]$ such that there exists a mesh $\Delta = \{a = x_0 < x_1 < x_2 < \cdots < x_{n-1} < x_n = b\}$ with $h_i = x_i - x_{i-1}$ for $i = 1, 2, \ldots, n$. This function satisfies the following conditions:

(i) In each subinterval $[x_i, x_{i+1}]$, $i = 0, 1, \ldots, n-1$, $S(x)$ is a polynomial of degree d.

(ii) $S(x)$ and its first $(d-1)$ derivatives are continuous on $[a, b]$.

The spline's theory and application were thoroughly discussed by Ahlberg Nilson, and Walsh [1] and Greville [9]. So far, different types of spline methods, such as approximating, interpolating, and curve fitting functions, have been developed and used to solve a wide variety of differential equations; see, for example, [2, 3, 14, 15, 17, 20, 22, 26, 25, 24, 31] and references therein. One type of splines, considered in this paper, is the parametric splines developed to address some shortcomings of ordinary spline methods. These splines, depending on a parameter $\tau > 0$, are defined through the solution of a differential equation in each subinterval. The arbitrary constants are chosen to satisfy certain smoothness conditions at the joints. These splines reduce to polynomial splines (ordinary splines) as $\tau \to 0$. The exact form depends upon the manner in which the parameter is introduced. Therefore, different types of parametric splines with distinct convergence orders can be generated. Although, these methods obtained significant results, due to the lengthy calculations, no attempt was made to extend parametric splines of higher degrees. Note that using the word "degree", in this paper, for parametric splines is only for numbering and ordering them. It does not have the common meaning that is used for polynomials.

For the first time, this paper presents a general form of parametric splines with the degree $(2k+1)$, $k = 1, 2, \ldots$, which has a lower computational cost and a higher-order of convergence than the usual methods using parametric splines. Before going into details about the method, it seems necessary to review some of the fundamental properties and definitions of these parametric splines in the following subsection.

1.1 Parametric spline methods with the degree $(2k+1)$

For simplicity, it is assumed that the subintervals are of equal length, so $h = h_i = h_{i+1}$. The interval $[a, b]$ is divided into n equal subintervals using knots x_i and the partition $\Delta = \{a = x_0, x_1, \ldots, x_n = b\}$, where $x_i = x_0 + ih$, $h = \frac{b-a}{n}$ and n is a positive integer. The parametric spline function $S(x)$, with the degree $(2k+1)$, $k = 1, 2, \ldots$, is obtained in the subinterval $[x_{i-1}, x_i]$ by solving the following differential equation and determining the constants of integration:

$$S^{(2k)}(x) + \tau^2 S^{(2k-2)}(x) = (S^{(2k)}(x_i) + \tau^2 S^{(2k-2)}(x_i))(\frac{x - x_{i-1}}{h})$$

$$+(S^{(2k)}(x_{i-1}) + \tau^2 S^{(2k-2)}(x_{i-1}))(\frac{x_i - x}{h}).$$

This function of class $C^{2k}[a,b]$ depends on a parameter τ and reduces to an ordinary spline function with the degree $(2k+1)$, as $\tau \to 0$. The continuity of its derivatives at the grid points, that is, $S_{i-1}^{(\nu)}(x_i) = S_i^{(\nu)}(x_i)$, $\nu = 1, 3, \ldots, 2k-1$, yields spline relations. Note that S_i is the spline function in the subintervals $[x_i, x_{i+1}]$. Using algebraic manipulation on these relations, a differential relation, called "consistency relation", is obtained in terms of u and its derivatives at knots. In the parametric spline methods, the approximate solution of a given boundary value problem (BVP) is determined by solving the system defined by this consistency relation.

Now, to further explain how parametric splines are used to solve equations, we consider a simple second order BVP as follows:

$$\begin{cases} u''(x) = f(x) + g(x)u(x), & x \in [a,b], \\ u(a) = \lambda, \\ u(b) = \gamma, \end{cases} \tag{1}$$

where λ and γ are finite real constants and the functions $f(x)$ and $g(x)$ are continuous on $[a,b]$. Such problems arise in the theory that describes the deflection of plates and a number of other scientific applications [10].

The consistency relation associated with (1) in spline methods, is in terms of u_i and u_i''. Note that $u_i = u(x_i)$ and $u_i'' = u''(x_i)$. A system of linear algebraic equations is generated by substituting discretized (1) in the mentioned consistency relation. Finally, by solving this system, the approximate solution of (1) is obtained. One can observe that, for $k > 1$, the number of equations in this system is less than the number of unknowns; see, for example, [2, 3, 4, 8, 7, 11, 13, 14, 16, 18, 20, 19, 21, 26, 25, 24, 30, 31] To obtain the unique solution of the system, more equations, called "end conditions or boundary formulas", are needed.

When k is a large number, two problems are encountered. First, the number of additional equations that must be defined to complete the aforementioned system increases. Second, as the value of k increases, so does the number of relations resulting from the derivative continuities of splines, and consequently, the combination of them becomes more difficult.

In this paper, we present a method that allows us to obtain a general form for the consistency relations of parametric spline methods of degree $(2k+1)$ without going through a lengthy and complex calculation process. For large values of k, it does not face the drawbacks mentioned above, because, in the proposed method, we do not need to obtain the spline function and use its continuity properties and derivatives directly. This means that we do not solve any differential equation. In fact, by providing a general pattern, both the consistency relation and the required additional equations are obtained without a need to generate many spline relations. Moreover, we transform the desired algorithm into a matrix form by defining proper band matrices,

which gives us more insight into the method and facilitates convergence study. Furthermore, the convergence order of splines is improved by this general formulation.

We apply our method to (1), which is in terms of u and u''. However, the method can be applied to more complex models of (1), such as the nonlinear form or the system of these equations. It will be demonstrated in the numerical results section. Our method along with Newton–Raphson method is used to solve Bratu's problem (Example (2)) as a nonlinear equation. Also, our method is applied to solve problems such as Perturbed (Example (3)) and two-dimensional problems in the calculus of variations (Example (4)).

It should be noted that, if an equation other than (1) is considered, then the execution process of the method, such as generating a consistency relation and additional equations, will be changed. Because they are produced and defined according to the type of equation. In addition, while this paper focuses on parametric splines with the degree $(2k+1)$, the extending of our method can be probed for other types of splines as well, that is, nonpolynomials, ordinary splines, and parametric splines with the degree of $(2k)$, $k = 2, 3, 4, \ldots$.

The outline of this paper is as follows. In section 2, a comprehensive description of the method is given. To demonstrate the efficiency and superiority of the presented method, we solve examples of linear, nonlinear, perturb, and system of two-point BVPs and compare the obtained results with the other quoted methods in section 3. Finally, some important concluding remarks are given in section 4.

2 Derivation of the method

In this section, we describe our method in detail. As mentioned previously, in the common form of the spline method, we need to generate a consistency relation proportional to the type of BVP that we are going to solve numerically. This can be time-consuming and even complicated due to the numerous calculations required, such as developing the spline function criterion, determining its coefficients, and computing successive derivatives. We provide a generalized form for the consistency relation of all parametric spline methods of degree $(2k+1)$, while solving (1), without the need for long calculations. To produce this relation, we find a specific pattern and then convert it to a matrix form by using only the properties of band matrices, especially, the following widely used a matrix C, which is $(n-1)\times(n-1)$-dimensional and evident in the majority of spline-based papers (see [8, 7, 13, 14, 18, 20, 19, 21, 31]):

$$(C)_{i,j} = \begin{cases} 2, & i = j, \\ -1, & |i-j| = 1, \\ 0, & otherwise. \end{cases} \tag{2}$$

To shed light on the above-mentioned issues, we begin subsection 2.1 by evaluating some samples of the consistency relations associated with the spline methods previously used by researchers and then identifying a general pattern for our desired consistency relation. In subsection 2.2, we define its matrix form. Subsections 2.3 and 2.4 are also dedicated to solving (1) and developing boundary formulas according to the contents of the previous two subsections.

2.1 The consistency relation

There are two types of coefficients in the consistency relations of spline methods: the coefficients of u and its derivatives. By studying the spline papers (see the references on page 3), we find that just the coefficients of u follow a certain pattern. The coefficients of u are of two kinds: known and unknown. The first type of coefficients exists in the consistency relations of splines with the degree $(2k+1)$ in solving the BVPs of order $(2k)$, while the second ones can be seen in the consistency relations of the same splines in solving the BVPs of order $2, 4, 6, \ldots, 2k-2$, for $k = 2, 3, 4, \ldots$ (i.e., for $k = 2$, we have spline with the degree 5 in solving BVP of order 2, for $k = 3$, we have spline with the degree 7 in solving BVP of order 2 and 4, for $k = 4$, we have spline with the degree 9 in solving BVP of order 2,4 and 6, etc.) We first establish a pattern for known coefficients and then extend this to unknown ones.

In the following, we highlight them in the sample format. For convenience, we consider the coefficients of u to be on the left side of the consistency relation and assume that the coefficients of derivatives of u be on the right side.

Sample 1 (k=1): *The left side of the consistency relation of spline method with the degree 3 in solving a BVP of order 2*:

For $i = 1, 2, \ldots, n-1$:

$$\mathbf{1}u_{i-1} - \mathbf{2}u_i + \mathbf{1}u_{i+1} = \cdots .$$

One can see this relation in [15, 29].

Sample 2 (k=2): *The left side of the consistency relation of spline method with the degree 5 in solving a BVP of order 4*:

For $i = 2, 3, \ldots, n-2$,

$$\mathbf{1}u_{i-2} - \mathbf{4}u_{i-1} + \mathbf{6}u_i - \mathbf{4}u_{i+1} + \mathbf{1}u_{i+2} = \cdots .$$

One can see this relation in [18].

Sample 3 (k=3): *The left side of the consistency relation of spline method with the degree 7 in solving a BVP of order 6*:

For $i = 3, 4, \ldots, n-3$,

$$\mathbf{1}u_{i-3} - \mathbf{6}u_{i-2} + \mathbf{15}u_{i-1} - \mathbf{20}u_i + \mathbf{15}u_{i+1} - \mathbf{6}u_{i+2} + \mathbf{1}u_{i+3} = \cdots .$$

One can see this relation in [3, 11, 26].

Sample 4 (k=4): *The left side of the consistency relation of spline method with the degree 9 in solving a BVP of order 8*:

For $i = 4, 5, \ldots, n-4$,

$$\begin{aligned}\mathbf{1}u_{i-4} - \mathbf{8}u_{i-3} + \mathbf{28}u_{i-2} - \mathbf{56}u_{i-1} + \mathbf{70}u_i \\ - \mathbf{56}u_{i+1} + \mathbf{28}u_{i+2} - \mathbf{8}u_{i+3} + \mathbf{1}u_{i+4} = \cdots .\end{aligned}$$

One can see this relation in references [2, 19].

Sample 5 (k=5): *The left side of the consistency relation of spline method with the degree 11 in solving a BVP of order 10*:

For $i = 5, 6, \ldots, n-5$,

$$\begin{aligned}\mathbf{1}u_{i-5} - \mathbf{10}u_{i-4} + \mathbf{45}u_{i-3} - \mathbf{120}u_{i-2} + \mathbf{210}u_{i-1} - \mathbf{252}u_i \\ + \mathbf{210}u_{i+1} - \mathbf{120}u_{i+2} + \mathbf{45}u_{i+3} - \mathbf{10}u_{i+4} + \mathbf{1}u_{i+5} = \cdots .\end{aligned}$$

One can see this relation in [20, 24].

Sample 6 (k=6): *The left side of the consistency relation of spline method with the degree 13 in solving a BVP of order 12*:

For $i = 6, 7, \ldots, n-6$,

$$\begin{aligned}\mathbf{1}u_{i-6} - \mathbf{12}u_{i-5} + \mathbf{66}u_{i-4} - \mathbf{220}u_{i-3} + \mathbf{495}u_{i-2} - \mathbf{792}u_{i-1} + \mathbf{924}u_i \\ - \mathbf{792}u_{i+1} + \mathbf{495}u_{i+2} - \mathbf{220}u_{i+3} + \mathbf{66}u_{i+4} - \mathbf{12}u_{i+5} + \mathbf{1}u_{i+6} = \cdots .\end{aligned}$$

One can see this relation in [25].

By considering the above coefficients, we find that, regardless of their sign, they are the same as the binomial coefficients or the entries in the rows of Pascal's triangle:

$$\begin{array}{c}
1 \\
1 \quad 1 \\
\mathbf{1} \quad \mathbf{2} \quad \mathbf{1} \\
1 \quad 3 \quad 3 \quad 1 \\
\mathbf{1} \quad \mathbf{4} \quad \mathbf{6} \quad \mathbf{4} \quad \mathbf{1} \\
1 \quad 5 \quad 10 \quad 10 \quad 5 \quad 1 \\
\mathbf{1} \quad \mathbf{6} \quad \mathbf{15} \quad \mathbf{20} \quad \mathbf{15} \quad \mathbf{6} \quad \mathbf{1} \\
1 \quad 7 \quad 21 \quad 35 \quad 35 \quad 21 \quad 7 \quad 1 \\
\mathbf{1} \quad \mathbf{8} \quad \mathbf{28} \quad \mathbf{56} \quad \mathbf{70} \quad \mathbf{56} \quad \mathbf{28} \quad \mathbf{8} \quad \mathbf{1} \\
1 \quad 9 \quad 36 \quad 84 \quad 126 \quad 126 \quad 84 \quad 36 \quad 9 \quad 1 \\
\mathbf{1} \quad \mathbf{10} \quad \mathbf{45} \quad \mathbf{120} \quad \mathbf{210} \quad \mathbf{252} \quad \mathbf{210} \quad \mathbf{120} \quad \mathbf{45} \quad \mathbf{10} \quad \mathbf{1}
\end{array}$$

	1	11	55	165	330	462	462	330	165	55	11	1
1	**12**	**66**	**220**	**495**	**792**	**924**	**792**	**495**	**220**	**66**	**12**	**1**

...

The correlation between the above-known coefficients of u and Pascal's triangle motivates us to find a similar correlation for the unknown ones that we will deal with in this paper.

After studying the references such as [8, 14, 15, 16, 21, 22, 29, 30, 31] (Previous studies have only investigated 3rd-, 5th-, 7th-, and 9th-degree spline methods. Higher degree splines have not been used yet), we find out to consider the initial form for the consistency relations of spline methods with the degree $(2k+1)$, $k = 2, 3, 4, \ldots$ in solving a BVP of order two as follows:

$$\begin{aligned}&*(u_{i-k} + u_{i+k}) + *(u_{i-k+1} + u_{i+k-1}) + \cdots + *(u_{i-1} + u_{i+1})\\&+ * u_i = -h^2 \left(\beta_0 u''_i + \beta_1(u''_{i-1} + u''_{i+1}) + \cdots + \beta_k(u''_{i-k} + u''_{i+k})\right),\end{aligned}$$

where β_j's are the coefficients which will be determined numerically during the process of the method. We have displayed the vacancy of the unknown coefficients of u with $*$. We intend to find a pattern for them, inspired by Pascal's triangle. For this purpose, we first consider the parameters as $\alpha_0, \alpha_1, \alpha_2, \ldots, \alpha_{k-1}$, for each k, and then we implement Pascal's algorithm for them. It should be mentioned that the numerical value of these coefficients for each k is independent of the values for other k, so it is preferable to write β_j's and α_j's with the exponent (k) as $\beta_j^{(k)}$ and $\alpha_j^{(k)}$. However, to reduce the complexity of the text, the exponent (k) could be removed from the coefficients without disturbing the whole. Indeed, the number of coefficients, which is indicated by an index in them, is affected by k.

Hence, we have the following Pascal's algorithm for $\alpha_0, \alpha_1, \alpha_2, \ldots, \alpha_{k-1}$:

$$\begin{aligned}&\alpha_{k-1}, \alpha_{k-2}, \ldots, \alpha_2, \alpha_1, \alpha_0, \alpha_1, \alpha_2, \ldots, \alpha_{k-2}, \alpha_{k-1}\\&\alpha_{k-1}, \alpha_{k-1} + \alpha_{k-2}, \ldots, \alpha_1 + \alpha_0, \alpha_0 + \alpha_1, \ldots, \alpha_{k-2} + \alpha_{k-1}, \alpha_{k-1}\\&\alpha_{k-1}, 2\alpha_{k-1} + \alpha_{k-2}, \ldots, \alpha_2 + 2\alpha_1 + \alpha_0, 2(\alpha_1 + \alpha_0), \alpha_2 + 2\alpha_1 + \alpha_0, \ldots, 2\alpha_{k-1} + \alpha_{k-2}, \alpha_{k-1}\\&\cdots.\end{aligned}$$

According to the consistency relations of the mentioned references, the third row of the above triangle, regardless of the signs, is the same as the coefficients of u in the consistency relation of the spline method of degree $(2k+1)$ for solving a BVP of second order as (1). We will demonstrate that the next rows of this triangle are the coefficients of u in the consistency relations of the spline methods of degree $(2k+1)$ for solving BVPs of higher orders (bigger than 2) in future research.

Thus, the following relation with the sign $(-1)^{q+p}$ for each phrase $\alpha_p u_{i\pm q}$, can be defined as the desired consistency relation for $k = 2, 3, 4, \ldots$:

$$\begin{aligned}
&-\alpha_{k-1}(u_{i-k}+u_{i+k})+(2\alpha_{k-1}-\alpha_{k-2})(u_{i-k+1}+u_{i+k-1})\\
&+(-\alpha_{k-1}+2\alpha_{k-2}-\alpha_{k-3})(u_{i-k+2}+u_{i+k-2})+\cdots+(2\alpha_0-2\alpha_1)u_i\\
&\quad=-h^2\left(\beta_0 u_i''+\sum_{j=1}^{k}\beta_j(u_{i-j}''+u_{i+j}'')\right),\quad i=k,k+1,\ldots,n-k. \qquad (3)
\end{aligned}$$

In the following, to verify the correctness of (3), we compare it to the consistency relations developed in related papers. A quick review shows that although the appearance of the coefficients in the consistency relations of available references slightly differs from what we propose, they are identical in content. In fact, the other authors have defined these coefficients in terms of parameter τ:

Equation (3) for $k=2$, is the consistency relation of parametric spline with degree 5 (quintic spline):

$$\begin{aligned}
&-\alpha_1(u_{i-2}+u_{i+2})+(2\alpha_1-\alpha_0)(u_{i-1}+u_{i+1})+(-2\alpha_1+2\alpha_0)u_i\\
&\quad=-h^2\left[\beta_2(u_{i-2}''+u_{i+2}'')+\beta_1(u_{i-1}''+u_{i+1}'')+\beta_0 u_i''\right],\quad i=2,3,\ldots,n-2.
\end{aligned}$$

One can compare it to the consistency relations in [16, 21, 30, 31].

Equation (3) for $k=3$, is the consistency relation of parametric spline with degree 7 (septic spline):

$$\begin{aligned}
&-\alpha_2(u_{i-3}+u_{i+3})+(2\alpha_2-\alpha_1)(u_{i-2}+u_{i+2})\\
&+(-\alpha_0+2\alpha_1-\alpha_2)(u_{i-1}+u_{i+1})+(-2\alpha_1+2\alpha_0)u_i\\
&\quad=-h^2[\beta_3(u_{i-3}''+u_{i+3}'')+\beta_2(u_{i-2}''+u_{i+2}'')+\beta_1(u_{i-1}''+u_{i+1}'')+\beta_0 u_i''],\\
&\qquad\qquad i=3,4,\ldots,n-3.
\end{aligned}$$

One can compare it to the consistency relations in [14].

Equation (3) for $k=4$, is the consistency relation of parametric spline with degree 9 (nonic spline):

$$\begin{aligned}
&-\alpha_3(u_{i-4}+u_{i+4})+(2\alpha_3-\alpha_2)(u_{i-3}+u_{i+3})+(-\alpha_3+2\alpha_2-\alpha_1)(u_{i-2}+u_{i+2})\\
&+(-\alpha_2+2\alpha_1-\alpha_0)(u_{i-1}+u_{i+1})+(-2\alpha_1+2\alpha_0)u_i\\
&\quad=-h^2[\beta_4(u_{i-4}''+u_{i+4}'')+\beta_3(u_{i-3}''+u_{i+3}'')+\beta_2(u_{i-2}''+u_{i+2}''),\\
&\qquad+\beta_1(u_{i-1}''+u_{i+1}'')+\beta_0 u_i''] \qquad\qquad i=4,5,\ldots,n-4.
\end{aligned}$$

One can compare it to the consistency relations in [8, 7].

2.2 The matrix form

To demonstrate the accuracy of the above cases, namely, the validity of our claim about the correlation between the consistency relation of the parametric spline of degree $(2k+1)$ and Pascal's algorithm, we need to obtain a matrix form for (3), dependent on k. For this purpose, we use a matrix C and the following band matrices, which are $(n-1)\times(n-1)$-dimensional and play a major role in our method:

$$(A)_{i,j} = \begin{cases} \alpha_0, & i=j, \\ \alpha_1, & |i-j|=1, \\ \vdots & \\ \alpha_{k-1}, & |i-j|=k-1, \\ 0, & otherwise, \end{cases} \tag{4}$$

$$(B)_{i,j} = \begin{cases} \beta_0, & i=j, \\ \beta_1, & |i-j|=1, \\ \vdots & \\ \beta_k, & |i-j|=k, \\ 0, & otherwise. \end{cases}$$

Since the coefficients of the mentioned samples of consistency relations can be observed in the rows of the consecutive multiplication of matrix C by itself, it is expected that the coefficients of (3) is obtained from multiplying matrix A by the matrix C. Therefore, for each $k = 1, 2, \ldots,$ the general matrix form can be defined for (3) as follows:

$$(AC)_{k+1,*}V + h^2(B)_{k+1,*}V'' = 0, \quad i = k, k+1, \ldots, n-k. \tag{5}$$

where $_{k+1,*}$ denotes $(k+1)$th row of the above matrices. Proportional to k, we also define the column vectors V and V'' as follows:

$$V_j = \begin{cases} u_{i-k+(j-1)}, & 1 \le j \le 2k+1, \\ 0, & 2k+1 < j \le n-1, \end{cases}$$

and

$$V''_j = \begin{cases} u''_{i-k+(j-1)}, & 1 \le j \le 2k+1, \\ 0, & 2k+1 < j \le n-1, \end{cases}$$

where V_j and V''_j are the jth element of vectors V and V'', respectively.

Now, according to (5), we can obtain a set of parametric spline methods by changing k, without need to know the nature of the elements of matrices A and B. The numerical values of these elements are determined by expanding

(5) in Taylor's series around x_i. For this purpose, we first rewrite (5) in the following form:

$$\begin{aligned}
&(AC)_{k+1,1}(u_{i-k}+u_{i+k})+(AC)_{k+1,2}(u_{i-k+1}+u_{i+k-1})+\dots\\
&+(AC)_{k+1,k+1}u_i+h^2((B)_{k+1,1}(u''_{i-k}+u''_{i+k})+(B)_{k+1,2}(u''_{i-k+1}+u''_{i+k-1})\\
&+\dots+(B)_{k+1,k+1}u''_i)\\
&\quad=0, \qquad i=k,k+1,\dots,n-k.
\end{aligned}$$

Note that in the above relation, we used the equalities $(AC)_{k+1,j}=(AC)_{k+1,2k+(2-j)}$ and $(B)_{k+1,j}=(B)_{k+1,2k+(2-j)}$, $1\le j\le k$. These are directly obtained from the definition of band matrices A, B and C.

Now the local truncation error T_i, corresponding to Taylor's series of (5) can be obtained as

$$\begin{aligned}
T_i =&(AC)_{k+1,1}(u_i-khu'_i+\frac{(-kh)^2}{2!}u''_i+\dots+u_i+khu'_i+\frac{(kh)^2}{2!}u''_i+\dots)\\
&+(AC)_{k+1,2}(u_i+(-k+1)hu'_i+\frac{((-k+1)h)^2}{2!}u''_i+\dots\\
&+u_i+(k-1)hu'_i+\frac{((k-1)h)^2}{2!}u''_i+\dots)+\dots+(AC)_{k+1,k+1}u_i\\
&+h^2((B)_{k+1,1}(u''_i-khu^{(3)}_i+\frac{(-kh)^2}{2!}u^{(4)}_i+\dots+u''_i+khu^{(3)}_i\\
&+\frac{(kh)^2}{2!}u^{(4)}_i+\dots)+(B)_{k+1,2}(u''_i+(-k+1)hu^{(3)}_i+\frac{((-k+1)h)^2}{2!}u^{(4)}_i\\
&+\dots+u''_i+(k-1)hu^{(3)}_i+\frac{((k-1)h)^2}{2!}u^{(4)}_i+\dots)+\dots\\
&+(B)_{k+1,k+1}u''_i).
\end{aligned}$$

On simplifying, we get

$$\begin{aligned}
T_i =&(2(AC)_{k+1,1}+2(AC)_{k+1,2}+\dots+2(AC)_{k+1,k}+(AC)_{k+1,k+1})u_i\\
&+(k^2(AC)_{k+1,1}+(k-1)^2(AC)_{k+1,2}+\dots+(k-(k-1))^2(AC)_{k+1,k}\\
&+2(B)_{k+1,1}+2(B)_{k+1,2}+\dots+2(B)_{k+1,k}+(B)_{k+1,k+1})h^2u''_i\\
&+\dots.
\end{aligned}$$

Equations (2) and (4) give us $2\sum_{j=1}^{k}(AC)_{k+1,j}+(AC)_{k+1,k+1}=0$. Therefore, the first term of the above truncation error, that is, the term with coefficient u_i, is removed. We can obtain classes of the method, namely several orders of convergence, by utilizing the above truncation error and eliminating the coefficients of the various powers of h for different choices of α_j's and β_j's. However, since our goal is to obtain the highest order of convergence, so we choose α_j's and β_j's that have the following conditions:

(i) They satisfy the following relation

$$\sum_{j=1}^{k}(k-(j-1))^2(AC)_{k+1,j}+2\sum_{j=1}^{k}(B)_{k+1,j}+(B)_{k+1,k+1}=0,$$

which eliminates the second term of the truncation error, that is, the term with coefficient h^2u_i''. The above relation can be written as follows:

$$\alpha_0+2\sum_{j=1}^{k-1}\alpha_j=\beta_0+2\sum_{j=1}^{k}\beta_j,$$

because from the definition of matrices A and B, we have:

$$\sum_{j=1}^{k}(k-(j-1))^2(AC)_{k+1,j}=-\alpha_0-2\sum_{j=1}^{k-1}\alpha_j,$$
$$2\sum_{j=1}^{k}(B)_{k+1,j}+(B)_{k+1,k+1}=\beta_0+2\sum_{j=1}^{k}\beta_j.$$

In accordance with the papers related to splines, the following relation is provided for α_0 and β_0:

$$\alpha_0+2\sum_{j=1}^{k-1}\alpha_j=1=\beta_0+2\sum_{j=1}^{k}\beta_j.$$

In other words,

$$\alpha_0=1-2\sum_{j=1}^{k-1}\alpha_j, \tag{6}$$

and

$$\beta_0=1-2\sum_{j=1}^{k}\beta_j. \tag{7}$$

For more details, the interested readers are advised to see [11, 16, 20, 19, 21, 22, 25] and other related papers.

(ii) The remaining unknown elements, namely the following $(2k-1)$ coefficients are chosen in such a way that the terms with coefficient $h^4u_i^{(4)}$ to $h^{4k}u_i^{(4k)}$, in the truncation error, are eliminated:

$$\alpha_1,\alpha_2,\ldots,\alpha_{k-1},\beta_1,\beta_2,\ldots,\beta_k.$$

Therefore the local truncation error associated with (5) is $O(h^{4k+2})$, $k=1,2,\ldots$. Consequently, the proposed method is convergent of order

$O(h^{4k})$, $k = 1, 2, \ldots$.

2.3 Spline solution

Now we discretize (1) as $u_i'' = f_i + g_i u_i,\ i = 1, 2, \ldots, n-1$, at the grid points, where $f_i = f(x_i),\ g_i = g(x_i)$. As a result, the vector V'' can be rewritten as

$$V_j'' = \begin{cases} f_{i-k+(j-1)} + g_{i-k+(j-1)} u_{i-k+(j-1)}, & 1 \le j \le 2k+1, \\ 0, & 2k+1 < j \le n-1. \end{cases}$$

If we substitute V'' in (5) for $i = k, k+1, \ldots, n-k$, then a system with $(n-2k+1)$ linear algebraic equations and $(n-1)$ unknowns as $u_1, u_2, \ldots, u_{n-1}$ is obtained. Note that $u_0 = \lambda$ and $u_n = \gamma$. It can be represented in the matrix form as follows:

$$\widetilde{D}U = \widetilde{R}, \tag{8}$$

where $U = [u_1, u_2, \ldots, u_{n-1}]^T$ is a column vector with $(n-1)$ elements. Moreover, $\widetilde{D}$ is a matrix of order $(n-2k+1) \times (n-1)$, indicated as follows:

$$\widetilde{D} = \overline{(AC)} + h^2 \overline{B} G,$$

with $G = diag(g_1, g_2, \ldots, g_{n-1})$. Matrices $\overline{AC}$ and $\overline{B}$ are also defined by omitting the $(k-1)$ first and last rows of matrices AC and B, respectively. In other words, for $1 \le i \le n-2k+1$ and $1 \le j \le n-1$, we have

$$(\overline{AC})_{i,j} = (AC)_{k+(i-1),j}, \quad (\overline{B})_{i,j} = (B)_{k+(i-1),j}.$$

Finally, the vector $\widetilde{R}$ is given by

$$\begin{aligned} \widetilde{R}_i = & -h^2 \sum_{j=1}^{2k+1} (B)_{k+1,j} f_{j+i-2} \\ & + \begin{cases} -u_0 \left((AC)_{k+1,1} + h^2 g_0 (B)_{k+1,1} \right), & i = 1, \\ 0, & 2 \le i \le n-2k, \\ -u_n \left((AC)_{k+1,1} + h^2 g_n (B)_{k+1,1} \right), & i = n-2k+1. \end{cases} \end{aligned}$$

2.4 Development of the boundary formulas

To obtain a unique solution for the system (8), we need $(2(k-1))$ more equations; thus, we define them in the following form:

$$\sum_{j=0}^{k+i+1} \widehat{\alpha}_j^i u_j + h^2 \sum_{j=0}^{4k-1} \widehat{\beta}_j^i u_j'' = 0, \qquad i = 1, 2, \ldots, k-1,$$
$$\sum_{j=0}^{k+i+1} \widehat{\alpha}_{n-j}^i u_{n-j} + h^2 \sum_{j=0}^{4k-1} \widehat{\beta}_{n-j}^i u_{n-j}'' = 0, \quad i = n-(k-1), \ldots, n-2, n-1.$$

In order to use the band matrices in the new system, that is, system (8) along with the above equations, we use the following replacements for $j = 1, 2, \ldots, i+k$:

$$\widehat{\alpha}_j^i = (AC)_{i,j}, \qquad i = 1, 2, \ldots, k-1,$$
$$\widehat{\alpha}_{n-j}^i = (AC)_{i,n-j}, \qquad i = n-(k-1), \ldots, n-2, n-1.$$

These replacements simplify the convergence analysis of the method. The other unknown coefficients, $\widehat{\beta}_j^i$'s and $\widehat{\beta}_{n-j}^i$'s, are determined by considering the local truncation error of order $O(h^{4k+2})$ for the added equations and using Taylor's expansion of these equations for $i = 1, 2, \ldots, k-1$ around x_0 (or for $i = n-(k-1), \ldots, n-2, n-1$ around x_n).

On the other hand, from (2) and (4), we have $(AC)_{i,j} = (AC)_{n-i,n-j}$. Consequently, $\widehat{\alpha}_j^i = \widehat{\alpha}_{n-j}^{n-i}$ and $\widehat{\beta}_j^i = \widehat{\beta}_{n-j}^{n-i}$. Considering these justifications, system (8) is converted to the following system:

$$DU = R, \tag{9}$$

with

$$D = AC + h^2 \widehat{B} G, \tag{10}$$

where the matrix $\widehat{B}$ is $(n-1) \times (n-1)$-dimensional as the following form:

$$(\widehat{B})_{i,j} = \begin{cases} (\widehat{B})_{n-i,n-j} = \widehat{\beta}_j^i, & 1 \le i \le k-1, \quad 1 \le j \le 4k-1 \\ (B)_{i,j}, & for\ other\ i, j. \end{cases} \tag{11}$$

For the column vector R with $(n-1)$ elements, we have

$$R_i = \begin{cases} -u_0(\widehat{\alpha}_0^i + h^2 g_0 \widehat{\beta}_0^i) - h^2(\widehat{\beta}_0^i f_0 + \sum_{j=1}^{4k-1} (\widehat{B})_{i,j} f_j), & 1 \le i \le k-1, \\ \widetilde{R}_{i-(k-1)}, & k \le i \le n-k, \\ -u_n(\widehat{\alpha}_n^i + h^2 g_n \widehat{\beta}_n^i) - h^2(\widehat{\beta}_n^i f_n + \sum_{j=1}^{4k-1} (\widehat{B})_{i,j} f_{n-j}), & n-k+1 \le i \le n-1. \end{cases}$$

Finally, by solving the system (9), we obtain the solution vector U, the elements of which are approximately equal to the solution of (1) at nodes $x_1, x_2, \ldots, x_{n-1}$.

3 Numerical results

In order to test the viability of the proposed method and to demonstrate its convergence computationally, some BVPs including the cases of linear, nonlinear, perturbed, and system are considered. We measure the accuracy in the discrete maximum norm

$$\|E\| = \|U - U_{exact}\| = \max_{1\leq i\leq n-1} |U_i - (U_{exact})_i|,$$

and the convergence rate for linear and perturbed cases

$$CR = \log_2(\frac{\|E^n\|}{\|E^{2n}\|}),$$

where $\|E^n\|$ and $\|E^{2n}\|$ are the maximum absolute errors on n and $2n$ grid points, respectively. The results are listed in tables for different choices of n and k. From the tables, we see that the quantity CR is close to $4k$ for each k. In other words, by reducing the step size from h to $\frac{h}{l}$, the observed errors are approximately reduced by a factor $(\frac{1}{l})^{4k}$ verifying the convergence order of the presented method, that is, $O(h^{4k})$, $k = 1, 2, \ldots$. For example, in the rows related to $n = 16$ and $n = 64$ from Table (1), it is observed that the maximum absolute error is decreased by a factor $(\frac{1}{4})^{4k}$ when $n = 16$ is varied to $n = 64$. Namely, we have

$$for\ k = 1: \quad (6.72 * (10^{-9})) * (\frac{1}{4})^{4*1} \simeq 2.63 * (10^{-11}),$$
$$for\ k = 2: \quad (3.12 * (10^{-15})) * (\frac{1}{4})^{4*2} \simeq 8.09 * (10^{-21}),$$

The outcomes indicate that our presented method produces more accurate results in comparison with those obtained by other methods. It should be mentioned that the computations associated with the examples in this paper were performed using Mathematica 8.0. Run applications were done in just a few minutes. The numerical results in tables were written just for some values of k, but we could solve the examples for other values and the results are quite satisfactory as was already expected.

In addition, we have used some plots to illustrate the behavior of the numerical solutions. Furthermore, since $\widehat{B}$ is $(n-1)\times(n-1)$-dimensional, in (11) we should have $4k - 1 \leq n - 1$ and $k - 1 \leq n - 1$, which results in $4k \leq n$. This can be seen in the results tables.

Example 1. We consider the following linear two-point BVP:

$$u''(x) - u(x) = x^2 - 2,$$

$$u(0) = 0, \quad u(1) = 1.$$

The exact solution is

$$u(x) = 2\left(\frac{\sinh(x)}{\sinh(1)}\right) - x^2.$$

The corresponding maximum absolute errors and convergence rates in our computed solutions are listed in Tables (1) and (2), respectively. Rashidinia, Jalilian, and Mohammadi [21] solved this problem by using the nonpolynomial quintic spline. Although their method is similar to ours for $k = 2$, the only difference is that they used a lower order of convergence of the method, see Table (3).

The graph of the exact and approximate solutions of Example (1) for $n = 20$ and $k = 1, 2, 3, 4, 5$ is depicted in Figure (1).

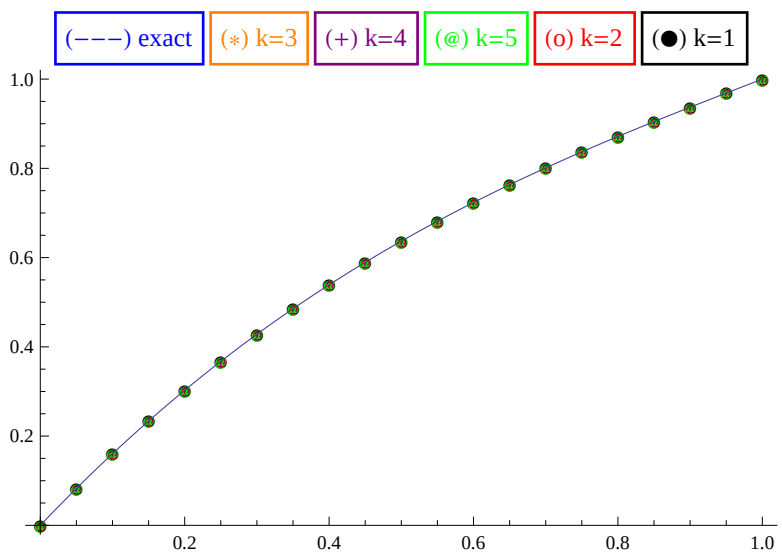


Figure 1: Plot of the exact and numerical solutions of Example (1) for $k = 1, 2, 3, 4, 5$ and $n = 20$

.

Example 2. We consider the following nonlinear BVP, the classical Bratu's problem:

$$u''(x) + \eta e^{u(x)} = 0,$$
$$u(0) = u(1) = 0,$$

where $\eta > 0$. The exact solution is

$$u(x) = -2\ln\left(\frac{\cosh((x-\frac{1}{2})\frac{\theta}{2})}{\cosh(\frac{\theta}{4})}\right),$$

where $\theta = \sqrt{2\eta}\cosh(\frac{\theta}{4})$. The Bratu's problem has zero, one, or two solutions when $\eta > \eta_c$, $\eta = \eta_c$, and $\eta < \eta_c$, respectively, where the critical value η_c satisfies the equation $1 = \frac{1}{4}\sqrt{2\eta_c}\sinh(\frac{\theta}{4})$ and it was evaluated in [5, 12] that the critical value η_c is given by $\eta_c = 3.513830719$.

We have solved this example for $\eta = 1, 2$, and 3.51 using our method with different values of k and tabulated the results in Tables (4), (5), and (6). Note that, in this example, we have used the Newton–Raphson algorithm, just with two iterations. Thus, there are errors related to the initial conjecture and the number of iterations, in addition to the error of our method. Tables (7) and (8) contain the comparison of our results and the results in [6, 13, 31]. The method in [31] is the same as our method for $k = 2$ with a lower order of convergence. Note that, the mentioned references have presented the outputs of their methods only for $n = 10$, so for the sake of comparison, in Table (7), we have to show the results of our method only for this value of n. We can use both $k = 1$ and $k = 2$ (according to condition $4k \leq n$), but $k = 2$ provides better results. Thus, we display its maximum absolute error.

Figure (2) plots the graphs of analytic and approximate solutions of Example (2) for $n = 20$, $\eta = 1$, and $k = 1, 2, 3, 4, 5$.

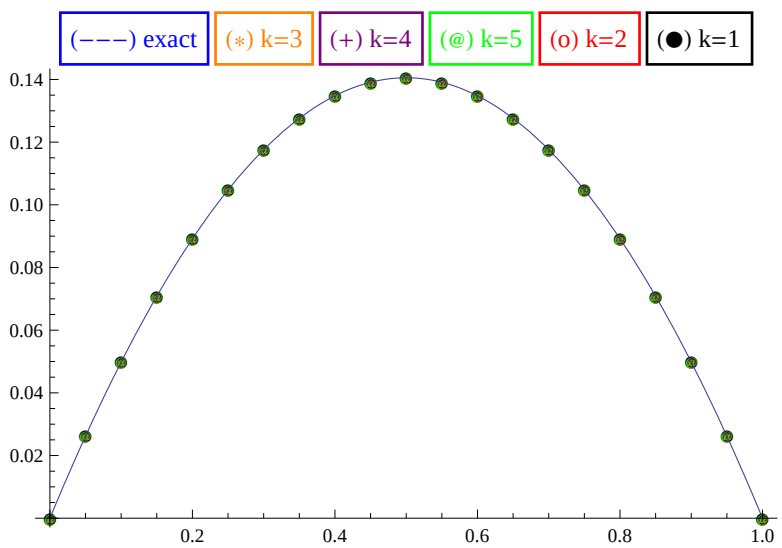


Figure 2: Plot of the exact and numerical solutions of Example (2) for $k = 1, 2, 3, 4, 5$ and $n = 20$, $\eta = 1$

.

Example 3. We consider the following singularly Perturbed BVP:

$$\epsilon u''(x) = u(x) + \cos^2(\pi x) + 2\epsilon\pi^2 \cos(2\pi x),$$
$$u(0) = u(1) = 0.$$

The exact solution is given by

$$u(x) = \frac{\exp(\frac{-(1-x)}{\sqrt{\epsilon}}) + \exp(\frac{-x}{\sqrt{\epsilon}})}{1 + \exp(\frac{-1}{\sqrt{\epsilon}})} - \cos^2(\pi x).$$

The maximum absolute errors and convergence rates for $\epsilon = \frac{1}{16}$ are tabulated in Tables (9) and (10), respectively. The results for this example from [4, 8, 17, 22, 27] are listed in Table (11). Note that the method used in [4, 22] is the same as our method for $k = 2$, but with a lower order of convergence. The results of [8] are also the same as the results of our method for $k = 4$.

We observe from Figure (3) that the graphic of the approximate solution of Example (3) for $n = 20$ and $k = 1, 2, 3, 4, 5$ coincides with the graphic of the exact solution.

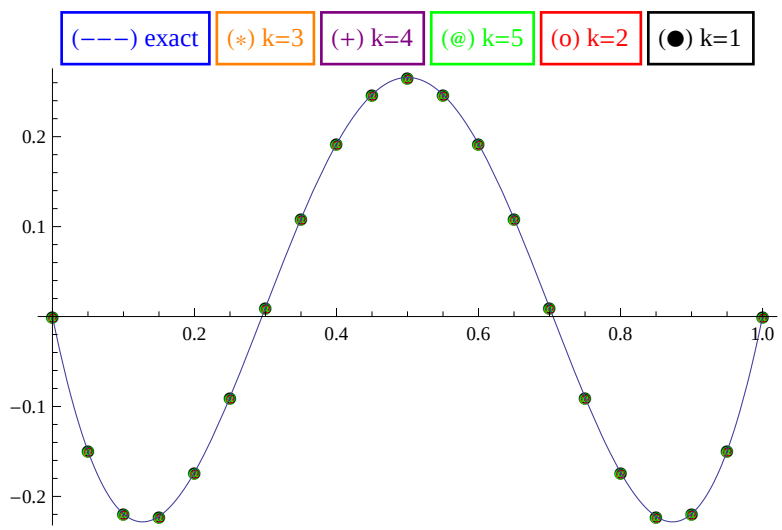


Figure 3: Plot of the exact and numerical solutions of Example (3) for $k = 1, 2, 3, 4, 5$ and $n = 20$

.

Example 4. We consider a BVP in calculus of variations, that is, the problem of finding the extremal of the functional [23]:

$$J[u_I(x), u_{II}(x)] = \int_0^{\frac{\pi}{2}} \left(u_I'^2(x) + u_{II}'^2(x) + 2u_I(x)u_{II}(x)\right) dx,$$

with boundary conditions

$$\begin{cases} u_I(0) = 0, & u_I(\frac{\pi}{2}) = 1, \\ u_{II}(0) = 0, & u_{II}(\frac{\pi}{2}) = -1. \end{cases}$$

The exact solution is given by $u_I(x) = -u_{II}(x) = \sin(x)$. For this problem, the corresponding Euler–Lagrange equations are

$$\begin{cases} u_I''(x) = u_{II}(x), \\ u_{II}''(x) = u_I(x), \end{cases}$$

that is, a system of equations such as (1). It should be mentioned that in this example, we compute $\|E_{u_I}\|$ and $\|E_{u_{II}}\|$, but one of them is displayed in Table (12), because, the value of both norms is the same. This example has already been solved by using cubic [29] and quintic [30] parametric spline methods, namely, the same as our method for $k = 1$ and $k = 2$ (with a lower convergence order), respectively. The sinc-Galerkin method [28] is also the other method that has been used for solving the above problem. The mentioned references have provided the numerical results only for $n = 5, 10, 20, \ldots, 50$. To make a proper comparison with these methods, we have shown our results only for these values, in Table (13). We have selected k's that apply to condition $4k \leq n$ and give the best outputs. For instance, for $n = 50$, we could display the numerical results of our method for $k = 1, 2, 3, \ldots, 12$, but since $k = 12$ gives the best result, we display its maximum absolute error.

The numerical results of Example (4) for $n = 20$ and $k = 1, 2, 3, 4, 5$ are plotted in Figure (4). Note that we have displayed this graph just for $u_I(x)$. Similarly, it can be shown for $u_{II}(x)$.

4 Conclusion

A long process is needed to obtain the differential relations of spline-based methods. Therefore, it is important to use a method that has considerably less computational effort with high accuracy and improves the spline methods. In this paper, for the first time, a generalized form of methods based on parametric splines of degree $(2k+1)$, $k = 1, 2, \ldots,$ was introduced that has all the mentioned properties. A very good accuracy of this method was demonstrated for solving some linear, nonlinear, perturbed, and system of BVPs. We mention some advantages of our method in the following remarks.

Remark 1. It is necessary to obtain the criterion of spline function in the spline methods. For instance, in the parametric spline method, this criterion

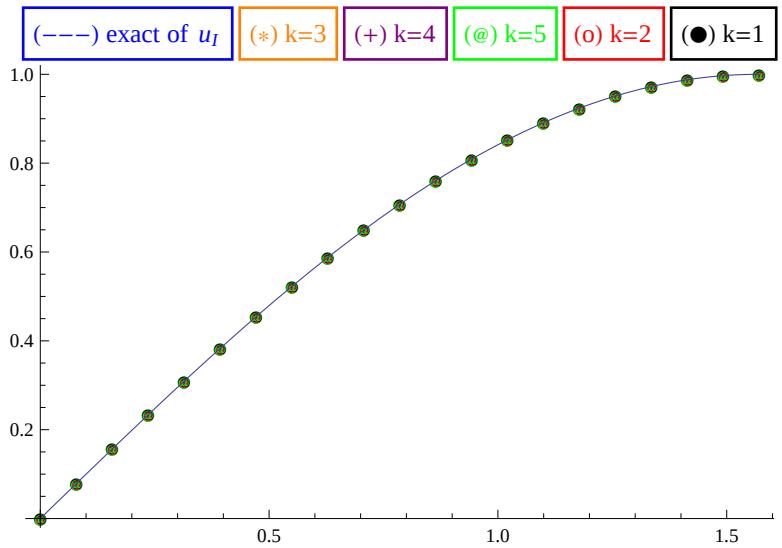


Figure 4: Plot of the exact and numerical solutions of Example (4) for $k = 1, 2, 3, 4, 5$ and $n = 20$

.

Table 1: Maximum absolute errors for Example (1).

n	$k=1$	$k=2$	$k=3$	$k=4$	$k=5$	$k=6$	$k=7\ldots$
4	1.66E−6	–	–	–	–	–	–
8	1.07E−7	2.24E−12	–	–	–	–	–
12	2.13E−8	4.92E−14	1.13E−18	–	–	–	–
16	6.72E−9	3.12E−15	2.43E−20	2.10E−25	–	–	–
20	2.76E−9	3.62E−16	1.19E−21	4.44E−27	1.74E−32	–	–
24	1.33E−9	6.17E−17	9.97E−23	1.84E−28	3.64E−34	7.42E−40	–
28	7.18E−10	1.37E−17	1.21E−23	1.23E−29	1.34E−35	1.52E−41	1.77E−47
⋮							
64	2.63E−11	8.09E−21	1.34E−28	5.35E−36	2.28E−43	1.02E−50	4.75E−58
128	1.64E−12	2.61E−23	8.75E−33	2.22E−41	6.07E−50	1.74E−58	5.16E−67
256	1.02E−13	9.71E−26	5.52E−37	8.84E−47	1.52E−56	2.76E−66	5.18E−76
512	6.43E−15	3.74E−28	4.03E−41	3.44E−52	3.73E−63	4.25E−74	5.01E−85
1024	4.02E−16	1.45E−30	6.27E−45	1.32E−57	9.02E−70	6.43E−82	4.75E−94
⋮							

Table 2: Convergence Rates, Example (1).

n	$k=1$	$k=2$	$k=3$	$k=4$	$k=5\ldots$
64	4.00	8.27	13.90	17.87	21.84
128	4.00	8.07	13.95	17.93	21.92
256	3.98	8.02	13.74	17.97	22.10
512	3.99	8.01	12.65	17.99	21.83

Table 3: Maximum absolute errors in [21] for Example (1).

n	Second-order [21]	Fourth-order[21]	Sixth-order[21]
8	1.09E−4	5.22E−8	8.75E−11
16	3.06E−5	2.31E−9	5.74E−13
32	8.11E−6	1.34E−10	2.30E−14
64	2.09E−6	8.42E−12	3.68E−14

Table 4: Maximum absolute errors for Example (2), $\eta = 1$.

n	$k=1$	$k=2$	$k=3$	$k=4$	$k=5\ldots$
4	1.17E−5	–	–	–	–
8	7.23E−7	1.78E−9	–	–	–
12	1.42E−7	1.72E−11	8.37E−13	–	–
16	4.50E−8	5.50E−13	6.32E−15	5.89E−16	–
20	1.84E−8	4.20E−14	2.94E−16	3.08E−16	3.98E−16
24	8.89E−9	6.63E−15	2.22E−16	2.77E−16	2.91E−16
28	4.79E−9	1.38E−15	4.16E−16	4.99E−16	2.49E−16
32	2.81E−9	1.41E−15	5.82E−16	4.30E−16	4.44E−16
36	1.75E−9	3.33E−16	1.05E−15	6.10E−16	1.38E−16
$\vdots$					

Table 5: Maximum absolute errors for Example (2), $\eta = 2$.

n	$k=1$	$k=2$	$k=3$	$k=4$	$k=5\ldots$
4	1.58E−4	–	–	–	–
8	9.55E−6	1.22E−7	–	–	–
12	1.87E−6	4.09E−10	3.32E−10	–	–
16	5.92E−7	8.14E−11	2.38E−12	1.07E−12	–
20	2.42E−7	1.09E−11	3.06E−13	1.78E−14	3.88E−15
$\vdots$					

Table 6: Maximum absolute errors for Example (2), $\eta = 3.51$.

n	$k=1$	$k=2$	$k=3$	$k=4$	$k=5\ldots$
4	5.55E−1	–	–	–	–
8	2.88E−3	2.29E−4	–	–	–
12	5.51E−4	2.46E−5	9.43E−6	–	–
16	1.73E−4	6.92E−7	9.35E−7	4.08E−7	–
20	7.08E−5	1.18E−8	4.85E−9	4.00E−8	1.74E−8
⋮					

Table 7: Comparison of $\|E\|$ for Example (2), $\eta = 1$ $n = 10$.

n	Our method for $k=2$	Method[31]	Method[6]
10	1.44E−10	5.87E−10	8.89E−6

Table 8: Comparison of $\|E\|$ for Example (2) with $\eta = 1, 2$ and 3.51.

	Our method			Method[13]		
n	$\eta=1$	$\eta=2$	$\eta=3.51$	$\eta=1$	$\eta=2$	$\eta=3.51$
8	1.78E−9(k= 2)	1.22E−7(k= 2)	2.29E−4(k= 2)	5.64E−9	4.53E−8	3.51E−5
16	5.89E−16(k= 4)	1.07E−12(k= 4)	4.08E−7(k= 4)	4.66E−11	1.76E−9	1.45E−7
32	4.30E−16(k= 4)	6.71E−15(k= 3)	9.32E−10(k= 2)	8.33E−13	2.13E−11	1.02E−9
64	4.71E−16(k= 6)	3.60E−15(k= 3)	1.37E−9(k= 2)	9.21E−15	2.87E−13	1.48E−11
128	2.22E−15(k= 6)	4.99E−16(k= 6)	1.37E−9(k= 2)	–	2.47E−14	1.58E−13

Table 9: Maximum absolute errors for Example (3), $\epsilon = \frac{1}{16}$.

n	$k=1$	$k=2$	$k=3$	$k=4$	$k=5$	$k=6$	$k=7\ldots$
4	1.15E−2	–	–	–	–	–	–
8	6.65E−4	4.45E−5	–	–	–	–	–
12	1.29E−4	7.49E−8	5.02E−8	–	–	–	–
16	4.07E−5	2.73E−8	4.75E−10	1.72E−11	–	–	–
20	1.66E−5	5.10E−9	1.50E−12	2.17E−13	2.45E−15	–	–
24	8.01E−6	1.05E−9	1.03E−12	3.93E−15	3.53E−17	1.73E−19	–
28	4.32E−6	2.58E−10	2.17E−13	1.51E−17	7.65E−19	2.67E−21	6.76E−24
⋮							
64	1.58E−7	9.10E−14	4.72E−18	2.66E−22	1.51E−26	8.18E−31	3.78E−35
128	9.87E−9	2.26E−16	3.28E−22	1.28E−27	5.36E−33	2.31E−38	1.02E−43
256	6.17E−10	9.47E−19	2.08E−26	5.21E−33	1.40E−39	3.95E−46	1.14E−52
512	3.86E−11	3.76E−21	1.29E−30	2.02E−38	3.43E−46	6.10E−54	1.12E−61
1024	2.41E−12	1.47E−23	7.92E−35	7.78E−44	8.25E−53	9.18E−62	1.05E−70
⋮							

Table 10: Convergence Rates, Example (3).

n	$k=1$	$k=2$	$k=3$	$k=4$	$k=5\ldots$
64	4.00	8.65	13.81	17.66	21.42
128	3.99	7.89	13.94	17.90	21.86
256	3.99	7.97	13.97	17.97	21.96
512	4.00	7.99	13.99	17.98	21.98

Table 11: Comparison of $\|E\|$ for Example (3), $\epsilon = \frac{1}{16}$.

n	Method[8]	Method[17]	Method[4]	Method[22]	Method[27]
16	1.72E−11	1.22E−6	1.57E−5	4.07E−5	1.20E−4
32	1.52E−17	6.45E−9	8.79E−7	2.53E−6	7.47E−6
64	2.66E−22	3.40E−11	5.32E−8	1.58E−7	4.67E−7
128	1.28E−27	1.03E−12	3.30E−9	9.87E−9	2.90E−8

Table 12: Maximum absolute errors for Example (4).

n	$k=1$	$k=2$	$k=3$	$k=4$	$k=5$	$k=6$	$k=7\ldots$
4	2.76E−5	–	–	–	–	–	–
8	1.72E−6	1.80E−10	–	–	–	–	–
12	3.41E−7	3.37E−12	5.75E−16	–	–	–	–
16	1.08E−7	1.97E−13	1.03E−17	6.79E−22	–	–	–
20	4.44E−8	2.32E−14	4.72E−19	1.22E−23	3.53E−28	–	–
24	2.14E−8	4.16E−15	3.74E−20	4.60E−25	6.41E−30	9.33E−35	–
28	1.15E−8	1.00E−15	4.37E−21	2.92E−26	2.16E−31	1.69E−36	1.37E−41
⋮							
512	1.03E−13	3.66E−26	1.97E−38	5.96E−49	3.95E−59	2.75E−69	1.98E−79
⋮							

Table 13: Comparison of $\|E\|$ for Example (4).

n	Our method	Method[30]	Method[29]	Method[28]
5	1.12E−5(k= 1)	–	1.12E−5	–
10	2.02E−11(k= 2)	6.70E−10	7.05E−7	2.72E−4
20	3.53E−28(k= 5)	7.07E−12	4.44E−8	8.69E−6
30	1.74E−42(k= 7)	8.10E−13	8.77E−9	5.65E−7
40	2.61E−63(k= 10)	1.55E−13	2.78E−9	5.47E−8
50	9.59E−67(k= 12)	4.21E−14	–	6.93E−9

is obtained by solving a special ordinary differential equation. Or nonpolynomial spline is a function with unknown coefficients that should be determined accordingly. In all these cases, some time-consuming calculations are needed, while the criterion and coefficients of no function are required in our method.

Remark 2. The continuity property of spline and its derivatives in grid points plays a major role in all of the spline methods. One can use this property to obtain the required spline relations. In this paper, instead of using the properties of spline directly, to save time and reduce calculations, we derived the consistency relations from a special algorithm and then obtained its matrix form by defining some band matrices.

Remark 3. The approximate solution converges to the exact solution of order $O(h^{4k})$. It follows that $\|E\| \to 0$ as $h \to 0$. The convergence occurs more quickly when k is a larger number. Indeed, the order of error is not fixed and decreases by increasing the value of k. It is regarded as one of our method's advantages. In addition, since we have $h = \frac{b-a}{n}$, it concludes that $\|E\| = O((\frac{b-a}{n})^{4k})$. This indicates that an increasing k is more effective than that n in reducing error. It can be seen in the tables containing numerical results.

Remark 4. We claim that the proposed method can be applied to solve other similar differential equations in particular as $u^{(2m)}(x) = f(x) + g(x)u(x)$, where m indicates a positive integer. This will be considered in our future research. Moreover, because of the adequate flexibility and expandability of this method, there is a possibility of achieving the generalized form of the methods based on splines with the degree $(2k)$, $k = 1, 2, \ldots$.

Acknowledgements

The authors are grateful to the anonymous referees for their time, effort, and extensive comments which improve the quality of the presentation of the paper.

References

[1] Ahlberg, J.H., Nilson, E.N. and Walsh, J.L. *The Theory of Splines and Their Applications*, Academic Press, New York. (1967),

[2] Akram, G. and Siddiqi, S.S. *Nonic spline solutions of eighth-order boundary value problems*, Appl. Math. Comput. 182 (2006), 829–845.

[3] Akram, G. and Siddiqi, S.S. *Solution of sixth order boundary value problems using non-polynomial spline technique*, Appl. Math. Comput. 181 (2006), 708–720.

[4] Aziz, T. and Khan, A. *Quintic spline approach to the solution of a singularly-perturbed boundary-value problem*, J. Optim. Theory Appl. 112 (2002), 517–527.

[5] Buckmire, R. *Application of a Mickens finite-difference scheme to the cylindrical Bratu-Gelfand problem*, Numer. Meth. Part. Differ. Equat. 20 (2004), 327–337.

[6] Caglar, H., Caglar, N., Ozer, M., Valaristos, A. and Anagnostopoulos, A.N. *B-spline method for solving Bratu's problem*, Int. J. Comput. Math. 87 (2010), 1885–1891.

[7] Farajeyan, K., Rashidinia, J. and Jalilian, R. *Classes of high-order numerical methods for solution of certain problem in calculus of variations*, Cogent. Math. Stat. 4 (2017), 1–15.

[8] Farajeyan, K., Rashidinia, J., Jalilian, R. and Maleki, N.R. *Application of spline to approximate the solution of singularly perturbed boundary-value problems*, Comput. Methods Differ. Equ. 8 (2020), 373–388.

[9] Greville, T.N.E. *Introduction to spline functions, in: Theory and Application of Spline Functions*, Academic Press, New York. (1969),

[10] Henrici, P. *Discrete variable methods in ordinary differential equations*, New York, Wiley. (1961),

[11] Islam, S. U., Tirmizi, I.A. , Haq, F. and Khan, M.A. *Non-polynomial splines approach to the solution of sixth-order boundary-value problems*, Appl. Math. Comput. 195 (2008), 270–284.

[12] Jacobsen, J. and Schmitt, K. *The Liouville-Bratu-Gelfand problem for radial operators*, J. Differ. Equat. 184 (2002), 283-298.

[13] Jalilian, R. *Non-polynomial spline method for solving Bratu's problem*, Comput. Phys. Commun. 181 (2010), 1868–1872.

[14] Jalilian, R., Rashidinia, J., Farajyan, K. and Jalilian, H. *Non-Polynomial Spline for the Numerical Solution of Problems in Calculus of Variations*, Int. J. Math. Comput. 5 (2015), 1–14.

[15] Khan, A. *Parametric cubic spline solution of two point boundary value problems*, Appl. Math. Comput. 154 (2004), 175–182.

[16] Khan, A., Khan, I. and Aziz, T. *A survey on parametric spline function approximation*, Appl. Math. Comput. 171 (2005), 983–1003.

[17] Khan, A., Khan, I. and Aziz, T. *Sextic spline solution of a singularly perturbed boundary-value problems*, Appl. Math. Comput. 181 (2006), 432–439.

[18] Rashidinia, J. and Golbabaee, A. *Convergence of numerical solution of a fourth-order boundary value problem*, Appl. Math. Comput. 171 (2005), 1296–1305.

[19] Rashidinia, J., Jalilian, R. and Farajeyan, K. *Spline approximate solution of eighth-order boundary-value problems*, Int. J. Comput. Math. 86 (2009), 1319–1333.

[20] Rashidinia, J., Jalilian, R. and Farajeyan, K. *Non polynomial spline solutions for special linear tenth-order boundary value problems*, World J. Model. Simul. 7 (2011), 40–51.

[21] Rashidinia, J., Jalilian, R. and Mohammadi, R. *Convergence analysis of spline solution of certain two-point boundary value problems*, Computer Science and Engineering and Electrical Engineering. 16 (2009), 128–136.

[22] Rashidinia, J. and Mahmoodi, Z. *Non-polynomial spline solution of a singularly perturbed boundary-value problems*, Int. J. Contemp. Math. Sciences. 2 (2007), 1581–1586.

[23] Razzaghi, M. and Yousefi, S. *Legendre wavelets direct method for variational problems*, Math. Comput. Simulat. 53 (2000), 185–192.

[24] Siddiqi, S.S. and Akram, G. *Solution of tenth-order boundary value problems using eleventh degree spline*, Appl. Math. Comput. 185 (2007), 115–127.

[25] Siddiqi, S.S. and Akram, G. *Solutions of 12th order boundary value problems using non-polynomial spline technique*, Appl. Math. Comput. 199 (2008), 559–571.

[26] Siddiqi, S.S. and Akram, G. *Septic spline solutions of sixth-order boundary value problems*, J. Comput. Appl. Math. 215 (2008), 288–301.

[27] Surla, K. and Vukoslavcević, V. *A spline difference scheme for boundary value problems with a small parameter*, Zb. Rad. Prirod.-Mat. Fak. Ser. Mat. 25 (1995), 159–168.

[28] Zarebnia, M. and Aliniya, N. *Sinc-Galerkin method for the solution of problems in calculus of variations*, Int. J. Nat. Eng. Sci. 5 (2011), 140–145.

[29] Zarebnia, M. and Sarvari, Z. *Numerical solution of the boundary value problems in calculus of variations using parametric cubic spline method*, J. Inform. Comput. Sci. 8 (2013), 275–282.

[30] Zarebnia, M. and Sarvari, Z. *Numerical solution of Variational Problems via parametric quintic spline method*, J. Hyperstruct. 3 (2014), 40–52.

[31] Zarebnia, M. and Sarvari, Z. *Parametric spline method for solving Bratu's problem*, Int. J. Nonlinear Sci. 14 (2012), 3–10.

How to cite this article
Sarvari. Z, A generalized form of the parametric spline methods of degree $(2k+1)$ for solving a variety of two-point boundary value problems. *Iran. J. Numer. Anal. Optim.*, 2023; 13(4): 578-603. https://doi.org/10.22067/ijnao.2023.79288.1192

Iranian Journal of Numerical Analysis and Optimization
Vol. 13, No. 4, 2023, pp 604–626
https://doi.org/10.22067/ijnao.2023.81586.1232
https://ijnao.um.ac.ir/

Research Article

Collection-based numerical method for multi-order fractional integro-differential equations

G. Ajileye*, T. Oyedepo, L. Adiku and J. Sabo

Abstract

In this paper, the standard collocation approach is used to solve multi-order fractional integro-differential equations using Caputo sense. We obtain the integral form of the problem and transform it into a system of linear algebraic equations using standard collocation points. The algebraic equations are then solved using the matrix inversion method. By substituting the algebraic equation solutions into the approximate solution, the numerical result is obtained. We establish the method's uniqueness as well as the convergence of the method. Numerical examples show that the developed method is efficient in problem-solving and competes favorably with the existing method.

AMS subject classifications (2020): 65C30, 65L06, 65C03.

*Corresponding author

Received 12 March 2023; revised 30 April 2023; accepted 15 May 2023

Ganiyu Ajileye
Department of Mathematics and Statistics, Federal University Wukari, Taraba State, Nigeria.
e-mail: ajileye@fuwukari.edu.ng

Taiye Oyedepo
Federal College of Dental Technology and Therapy, Enugu, Nigeria.
e-mail: oyedepotaiye@yahoo.com

Lydia Adiku
Department of Mathematics and Statistics, Federal University Wukari, Taraba State, Nigeria.
e-mail: adiku@fuwukari.edu.ng

John Sabo
Department of Mathematics, Adamawa State University, Mubi, Nigeria.
e-mail: sabojohn21@gmail.com

Keywords: Integro-differential equations; Collocation method; Fredholm-Volterra equations; Multi- order.

1 Introduction

Fractional calculus is one of the subfields of mathematics that looks at the characteristics of the derivatives and integrals of noninteger orders. This discipline examines the notion and method of solving differential equations with fractional derivatives of unknown functions. In recent years, a significant amount of interest in fractional calculus has emerged as a result of the fact that it may be used in a wide variety of fields of scientific interest; see [9]. Some of the numerical methods for the solution of fractional integro-differential equations developed in the literature include: Multi-order fractional by [12, 5, 16], Collocation method by [1, 3], Least square method by [13], Adomian decomposition method by [10], Chebyshev cardinal functions by [8], Laplace decomposition method by [11, 18, 14], Taylor expansion method by [7, 19], Haar wavelets by [4], Legendre Wavelets Method [6], and variational iteration method by [17]. Collocation approach to first-order Volterra integro-differential equations. The class of integro-differential equations was reformulated to assume an approximate solution in terms of the constructed polynomial. After solving for the unknown, we obtained a system of linear algebraic equations by collocating the resulting equation at various places within the range $[0, 1]$ [2]. The Laplace Adomian decomposition technique based on the Bernstein polynomial is employed to obtain an approximate solution for solving Volterra integral and integro-differential equations. Rani and Mistra [15] concluded that only orthogonal polynomials such as Legendre, Chebyshev, or Jacobi polynomials can improve the Adomian decomposition method.

In this research, we present efficient method for solving multi-order fractional integro-differential equations with fractional derivatives of the form

$$D^{\beta}y(x) = \sum_{j=0}^{N} q_j(x)D^{\alpha_j}y(x)+h(x)+\int_0^b k_1(x,t)y(t)dt+\int_0^x k_2(x,t)y(t)dt \tag{1}$$

subject to the initial condition

$$y^{(j)}(a_j) = \lambda_j, \quad j = 0, 1, \ldots, n-1, \ n \in \mathbb{N}, \quad \beta > \alpha_N, \tag{2}$$

where $y(x)$ is the unknown function to be determined, D^{α_j} and D^{β} are Caputo's derivative, and $h(x)$ is the force known prior. Moreover, $k_1(x,t)$ and $k_2(x,t)$ are the Fredholm and Volterra integral kernel functions, respectively. Also, $q_j(x)$ is the known function and a_j and λ_j are known constants.

2 Basic definitions

In this section, we present certain definitions and fundamental ideas of fractional calculus for the purpose of the formulation of the problem that has been presented.

Definition 1. The Caputo derivative with order $\alpha > 0$ of the given function $f(x),\ x \in (a,b)$ is defined as [11]

$$ {}_{x}^{C}D_{a}^{\alpha} y(x) = \frac{1}{\Gamma(m-\alpha)} \int_{a}^{x} (x-s)^{m-\alpha-1} y^{(m)}(s) ds, \tag{3} $$

where $m-1 \leq \alpha \leq m,\ m \in N,\ x > 0$.

Definition 2. Let $(a_n), n \geq 0$ be a sequence of real numbers. The power series in x with coefficients a_n is an expression [11]

$$ y(x) = a_0 + a_1 x + a_2 x^2 + a_3 x^3 + \cdots a_N x^N = \sum_{n=0}^{N} a_n x^n = \ \phi(x)\ \mathbf{A}, \tag{4} $$

where $\phi(x) = [1 \quad x \quad x^2 \quad \cdots \quad x^N]$, $\mathbf{A} = [a_0 \quad a_1 \quad \cdots \quad a_N]^T$. Then $y(x,n) = x^n \mathbf{A},\ n = 0(1)N,\ n \in Z^+$.

Definition 3 (Standard Collocation Method (SCM))**.** This method is used to determine the desired collocation points within an interval, $[a,b]$ and is given by [1]

$$ x_i = a + \frac{(b-a)i}{N}, \quad i = 1, 2, 3, \ldots, N. \tag{5} $$

Definition 4. Let $y(x)$ be a continuous function. Then [3]

$$ {}_{0}I_{x}^{\beta} \left({}_{0}^{C}D_{x}^{\beta} y(x) \right) = y(x) - \sum_{k=0}^{N} \frac{y^{(k)}(0)}{k!} x^k, \tag{6} $$

where $m-1 < \beta < 1$.

Definition 5. Let $p(s)$ be an integrable function. Then [3]

$$ {}_{0}I_{x}^{\beta} \left(p(s) \right) = \frac{1}{\Gamma(\beta)} \int_{0}^{x} (x-s)^{\beta-1} p(s) ds. \tag{7} $$

Definition 6. The Riemann–Liouville derivative of order $\alpha > 0$ with $n-1 < \alpha < n$ of the power function $f(t) = t^{p-\alpha}$ is given by [11]

$$ D^{\alpha} t^{p} = \frac{\Gamma(p+1)}{\Gamma(p-\alpha+1)} t^{p-\alpha}. \tag{8} $$

Definition 7. A metric on a set M is a function $d : M \times M \longrightarrow \mathbb{R}$ with the following properties, for all $x, y \in M$ [3],

(a) $d(x,y) \geq 0$,

(b) $d(x,y) = 0 \Longleftrightarrow x = y$,

(c) $d(x,y) = d(y,x)$,

(d) $d(x,y) \leq d(x,z) + d(x,y)$.

If d is a metric on M, then the pair (M,d) is called a metric space.

Definition 8. Let (X,d) be a metric space. A mapping $T : X \longrightarrow X$ is Lipschitzian if $\exists$ a constant $L > 0$ such that $d(Tx,Ty) \leq Ld(x,y)$ for all $x,y \in X$ [3].

3 Mathematical background

In this section, we develop an enhanced method for the numerical solution of multi-order fractional integro-differential equations. This method is based on the collocation approach and also considered power series polynomials as our basic function.

Theorem 1 (Banach's fixed point theorem)**.** Let (X,d) be a complete metric space. It follows that each contraction mapping $T : X \longrightarrow X$ has a unique fixed point x of T in X, such that $T(x) = x$.

Lemma 1 (Integral form)**.** Let $y(x)$ be a solution to (1) subject to (2). Then the integral form is

$$\begin{aligned} y(x) = W(x) &+ \sum_{j=0}^{N} \frac{1}{\Gamma(m_j - \alpha_j)} \frac{1}{\Gamma(\beta)} \\ &\times \int_0^x (x-s)^{\beta-1} q_j(s) \left[\int_0^s (s-t)^{m_j - \alpha_j - 1} y^{(m_j)}(t) dt \right] ds \\ &+ \frac{1}{\Gamma(\beta)} \int_0^x (x-s)^{\beta-1} \left(\int_0^b k_1(s,t) y(t) dt \right) ds \\ &+ \frac{1}{\Gamma(\beta)} \int_0^x (x-s)^{\beta-1} \left(\int_0^s k_2(s,t) y(t) dt \right) ds, \end{aligned} \tag{9}$$

where

$$W(x) = \sum_{k=0}^{N} \frac{y^{(k)}(0)}{k!} x^k + \frac{1}{\Gamma(\beta)} \int_0^x (x-s)^{\beta-1} h(s) ds.$$

Proof. Multiplying (1) by ${}_0I_x^{\beta}(\cdot)$ gives

$${}_0I_x^{\beta} \left(D^{\beta} y(x) \right) = {}_0I_x^{\beta} \left(\sum_{j=0}^{N} q_j(x) D^{\alpha_j} y(x) \right)$$

$$+ \ {}_0I_x^{\beta}\left(h(x)\right) + \ {}_0I_x^{\beta}\left(\int_0^b k_1(x,t)y(t)dt\right)$$
$$+ \ {}_0I_x^{\beta}\left(\int_0^s k_2(s,t)y(t)dt\right). \tag{10}$$

Using (6) on (9) gives

$$y(x) = \sum_{k=0}^{N} \frac{y^{(k)}(0)}{k!}x^k + \ {}_0I_x^{\beta}\left(\sum_{j=0}^{N} q_j(x)D^{\alpha_j}y(x)\right)$$
$$+ \ {}_0I_x^{\beta}\left(h(x)\right) + \ {}_0I_x^{\beta}\left(\int_0^b k_1(x,t)y(t)dt\right)$$
$$+ \ {}_0I_x^{\beta}\left(\int_0^s k_2(s,t)y(t)dt\right). \tag{11}$$

Applying (3) and (7) to (11) gives

$$y(x) = \sum_{k=0}^{N} \frac{y^{(k)}(0)}{k!}x^k + \ \frac{1}{\Gamma(\beta)}\int_0^x (x-s)^{\beta-1}$$
$$\times \left(\sum_{j=0}^{N} q_j(x)\frac{1}{\Gamma(m_j-\alpha_j)}\int_0^s (s-t)^{m_j-\alpha_j-1}\, y^{(m_j)}(t)dt\right) ds$$
$$+ \frac{1}{\Gamma(\beta)}\int_0^x (x-s)^{\beta-1}\, h(s)ds$$
$$+ \frac{1}{\Gamma(\beta)}\int_0^x (x-s)^{\beta-1}\left(\int_0^b k_1(x,t)y(t)dt\right) ds$$
$$+ \frac{1}{\Gamma(\beta)}\int_0^x (x-s)^{\beta-1}\left(\int_0^s k_2(s,t)y(t)dt\right) ds. \tag{12}$$

Substituting (4) into (12) gives

$$y(x) = \sum_{k=0}^{N} \frac{y^{(k)}(0)}{k!}x^k + \ \frac{1}{\Gamma(\beta)}\int_0^x (x-s)^{\beta-1}$$
$$\times \left(\sum_{j=0}^{N} q_j(x)\frac{1}{\Gamma(m_j-\alpha_j)}\int_0^s (s-t)^{m_j-\alpha_j-1}\, \frac{d^{m_j}}{dt^{m_j}}\left(\phi(t)\right)\ dt\mathbf{A}\right) ds$$
$$+ \frac{1}{\Gamma(\beta)}\int_0^x (x-s)^{\beta-1}\, h(s)ds + \frac{1}{\Gamma(\beta)}\int_0^x (x-s)^{\beta-1}$$
$$\times \left(\int_0^b k_1(x,t)\phi(t)dt\right) ds\mathbf{A} + \frac{1}{\Gamma(\beta)}\int_0^x (x-s)^{\beta-1}$$

$$\times \left(\int_0^s k_2(s,t)\phi(t)dt \right) ds\mathbf{A}. \tag{13}$$

□

3.1 Method of solution

Collocating at x_i in (13) gives

$$\begin{aligned} y(x_i) = W(x_i) &+ \sum_{j=0}^{N} \frac{1}{\Gamma(m_j-\alpha_j)} \frac{1}{\Gamma(\beta)} \int_0^{x_i} (x_i-s)^{\beta-1} q_j(s) \\ &\times \left(\int_0^s (s-t)^{m_j-\alpha_j-1} \frac{d^{m_j}}{dt^{m_j}} \left(\phi(t)\right) \, dt \right) ds \ \mathbf{A} \\ &+ \frac{1}{\Gamma(\beta)} \int_0^{x_i} (x_i-s)^{\beta-1} \left(\int_0^b k_1(s,t)\phi(t)dt \right) ds\mathbf{A} \\ &+ \frac{1}{\Gamma(\beta)} \int_0^{x_i} (x_i-s)^{\beta-1} \left(\int_0^s k_2(s,t)\phi(t)dt \right) ds \ \mathbf{A}, \end{aligned} \tag{14}$$

where

$$W(x_i) = \sum_{k=0}^{N} \frac{y^{(k)}(0)}{k!} x^k + \frac{1}{\Gamma(\beta)} \int_0^x (x-s)^{\beta-1} h(s)ds.$$

Simplifying (14) gives

$$\phi(x_i)\mathbf{A} = W(x_i) + \begin{bmatrix} \sum_{j=0}^{N} \frac{1}{\Gamma(m_j-\alpha_j)} \frac{1}{\Gamma(\beta)} \int_0^{x_i} (x_i-s)^{\beta-1} q_j(s) \\ \times \left(\int_0^s (s-t)^{m_j-\alpha_j-1} \frac{d^{m_j}}{dt^{m_j}} \left(\phi(t)\right) dt \right) ds \\ + \frac{1}{\Gamma(\beta)} \int_0^{x_i} (x_i-s)^{\beta-1} \\ \times \left(\int_0^b k_1(s,t) \left(\phi(t)\right) dt + \int_0^s k_2(s,t) \left(\phi(t)\right) dt \right) ds \end{bmatrix} \mathbf{A}. \tag{15}$$

Factorizing the values of $\mathbf{A}$ from (15) gives

$$\begin{bmatrix} \phi(x_i) - \sum_{j=0}^{N} \frac{1}{\Gamma(m_j-\alpha_j)} \frac{1}{\Gamma(\beta)} \int_0^{x_i} (x_i-s)^{\beta-1} q_j(s) \\ \times \left(\int_0^s (s-t)^{m_j-\alpha_j-1} \frac{d^{m_j}}{dt^{m_j}} \left(\phi(t)\right) dt \right) ds - \\ \frac{1}{\Gamma(\beta)} \int_0^{x_i} (x_i-s)^{\beta-1} \\ \times \left(\int_0^b k_1(s,t) \left(\phi(t)\right) dt + \int_0^s k_2(s,t) \left(\phi(t)\right) dt \right) ds \end{bmatrix} \mathbf{A} = W(x_i). \tag{16}$$

Equation (16) can be in the form

$$V(x_i)\mathbf{A} = W(x_i), \tag{17}$$

where

$$\begin{aligned} V(x_i) = \phi(x_i) - \sum_{j=0}^{N} \frac{1}{\Gamma(m_j-\alpha_j)} \frac{1}{\Gamma(\beta)} \int_0^{x_i} (x_i-s)^{\beta-1} q_j(s) \\ \left(\int_0^s (s-t)^{m_j-\alpha_j-1} \frac{d^{m_j}}{dt^{m_j}} (\phi(t)) \ dt \right) ds \ - \frac{1}{\Gamma(\beta)} \int_0^{x_i} (x_i-s)^{\beta-1} \\ \left(\int_0^b k_1(s,t) (\phi(t)) \, dt + \int_0^s k_2(s,t) (\phi(t)) \, dt \right) ds \end{aligned} \tag{18}$$

and

$$\mathbf{A} = [a_0 \quad a_1 \quad \cdots \quad a_N]^T$$

multiply both sides of (17) by $V^{-1}(x_i)$ gives

$$\mathbf{A} = V^{-1}(x_i) W(x_i). \tag{19}$$

Lemma 2. Let $y(t)$ be approximated by (10) and let

$$L(x) = {}_0I_x^{\beta} \left(\sum_{j=0}^{N} q_j(x) D^{\alpha_j} y(x) \right). \tag{20}$$

If $q_j(s) = s^{p_j}$, then

$$\mathbf{L}(x;n) = \frac{\Gamma(n+1)\Gamma(n-\alpha_j+p_j+1)}{\Gamma(n-\alpha_j+1)\Gamma(\beta+n-\alpha_j+p_j+1)} x_i^{\beta+n-\alpha_j+p_j} \mathbf{A}. \tag{21}$$

Proof. Applying (3) and (7) into (20) gives

$$\begin{aligned} {}_0I_x^{\beta} \left(\sum_{j=0}^{N} q_j(x) D^{\alpha_j} y(x) \right) = \sum_{j=0}^{N} \frac{1}{\Gamma(m_j-\alpha_j)} \frac{1}{\Gamma(\beta)} \int_0^{x_i} (x-s)^{\beta-1} q_j(s) \\ \left[\int_0^s (s-t)^{m_j-\alpha_j-1} y^{(m_j)}(t) dt \right] ds. \end{aligned} \tag{22}$$

Substituting (8) into (22) gives

$${}_0I_x^{\beta} \left(\sum_{j=0}^{N} q_j(x) D^{\alpha_j} y(x) \right)$$

$$= \sum_{j=0}^{N} \frac{1}{\Gamma(m_j - \alpha_j)} \frac{1}{\Gamma(\beta)} \int_0^x (x-s)^{\beta-1} S^{p_j}$$
$$\left[\int_0^s (s-t)^{m_j-\alpha_j-1} \left(\frac{\Gamma(n+1)}{\Gamma(n-m_j+1)} t^{n-m_j} \right) dt \right] ds\ \mathbf{A}. \tag{23}$$

Let $s - t = (1-v)s$. Then $t = vs \Longrightarrow \dfrac{dt}{dv} = s \Longrightarrow dt = sdv$. Substituting them into (23) gives

$$ {}_0I_x^{\beta} \left(\sum_{j=0}^{N} q_j(x) D^{\alpha_j} y(x) \right)$$
$$= \sum_{j=0}^{N} \frac{\Gamma(n+1)}{\Gamma(m_j-\alpha_j)\Gamma(n-m_j+1)} \frac{1}{\Gamma(\beta)} \int_0^x (x-s)^{\beta-1} S^{p_j}$$
$$\left[S^{n-\alpha_j} \int_0^1 (1-v)^{m_j-\alpha_j-1} V^{n-m_j} dt \right] ds\ \mathbf{A}. \tag{24}$$

Simplifying (24), we get

$$\mathbf{L}(x;n) = \frac{\Gamma(n+1)\Gamma(n-\alpha_j+p_j+1)}{\Gamma(n-\alpha_j+1)\Gamma(\beta+n-\alpha_j+p_j+1)} x^{\beta+n-\alpha_j+p_j} \mathbf{A}. \tag{25}$$

□

Lemma 3. Let $y(t)$ be approximated by (10) and let

$$E(x) = {}_0I_x^{\beta} \left[\int_0^b k_1(s,t) y(t) dt + \int_0^s k_2(s,t) y(t) dt \right]. \tag{26}$$

If $k_1(s,t) = s^r t^{\sigma}$ $k_2(s,t) = s^g t^v$, then

$$\mathbf{E}(x;n) = \left(\begin{array}{c} \dfrac{b^{r+n+1}\Gamma(r+1)}{(\sigma+n+1)\,\Gamma(\beta+r+1)} x^{\beta+r} + \\ \dfrac{\Gamma(g+v+n+2)}{(v+n+1)\,\Gamma(\beta+g+v+n+2)} x^{\beta+g+v+n+1} \end{array} \right) \mathbf{A}. \tag{27}$$

Proof. Applying (10) to (26) gives

$$ {}_0I_x^{\beta} \left[\int_0^b k_1(s,t) y(t) dt + \int_0^s k_2(s,t) y(t) dt \right]$$
$$= \frac{1}{\Gamma(\beta)} \int_0^x (x-s)_0^{\beta-1} \left[\int_0^b k_1(s,t) y(t) dt + \int_0^s k_2(s,t) y(t) dt \right] ds.$$

Substituting $k_1(s,t) = s^r t^\sigma$ $k_2(s,t) = s^g t^v$ gives

$$\begin{aligned} &{}_0I_x^\beta \left[\int_0^b k_1(s,t)y(t)dt + \int_0^s k_2(s,t)y(t)dt \right] \\ &= \frac{1}{\Gamma(\beta)} \int_0^x (x-s)^{\beta-1} \left(\int_0^b s^r t^\sigma y(t)dt \right) ds \\ &\quad + \frac{1}{\Gamma(\beta)} \int_0^x (x-s)^{\beta-1} \left(\int_0^s s^g t^v y(t)dt \right) ds. \end{aligned} \tag{28}$$

Applying (4) to (28) and simplifying give

$$\begin{aligned} &{}_0I_x^\beta \left[\int_0^b k_1(s,t)y(t)dt + \int_0^s k_2(s,t)y(t)dt \right] \\ &= \frac{1}{\Gamma(\beta)} \int_0^x (x-s)^{\beta-1} \left[s^r \frac{b^{\sigma+n+1}}{\sigma+n+1} \right] \mathbf{A}\ ds \\ &\quad + \frac{1}{\Gamma(\beta)} \int_0^x (x-s)^{\beta-1} \left[s^g \frac{s^{v+n+1}}{v+n+1} \right] \mathbf{A}\ ds. \end{aligned} \tag{29}$$

Let $x - s = (1-u)x$. Then $s = ux \Longrightarrow ds = xdu$. Substituting them into (29) gives

$$\begin{aligned} &{}_0I_x^\beta \left[\int_0^b k_1(s,t)y(t)dt + \int_0^s k_2(s,t)y(t)dt \right] \\ &= \left(\begin{array}{c} \dfrac{b^{\sigma+n+1}}{\Gamma(\beta)\ (\sigma+n+1)} \int_0^1 \left((1-u)\,x\right)^{\beta-1} (ux)^r x\ du + \\ \dfrac{1}{\Gamma(\beta)\ (v+n+1)} \int_0^1 \left((1-u)\,x\right)^{\beta-1} (ux)^{g+v+n+1}\ x\ du \end{array} \right) \mathbf{A}. \end{aligned} \tag{30}$$

Solving (30) gives

$$\mathbf{E}(x;n) = \left(\begin{array}{c} \dfrac{b^{\sigma+n+1}\Gamma(r+1)}{(\sigma+n+1)\,\Gamma(\beta+r+1)} x^{\beta+r} + \\ \dfrac{\Gamma(g+v+n+2)}{(v+n+1)\,\Gamma(\beta+g+v+n+2)} x^{\beta+g+v+n+1} \end{array} \right) \mathbf{A}.$$

□

Lemma 4. Let $y(t)$ be approximated by (10) and let

$$C(x) = {}_0I_x^\beta \left(h(x)\right). \tag{31}$$

If $h(s) = s^m$, then

$$C(x) = \frac{\Gamma(m+1)}{\Gamma(\beta+m+1)} x^{\beta+m}.$$

Proof. Applying (7) to (31) gives

$$ {}_0I_x^{\beta}\left(h(x)\right) = \frac{1}{\Gamma(\beta)}\int_0^x (x-s)^{\beta-1}h(s)\ ds. $$

Substituting for $h(s)$ gives

$$ {}_0I_x^{\beta}\left(h(x)\right) = \frac{1}{\Gamma(\beta)}\int_0^x (x-s)^{\beta-1}s^m ds. $$

Let $x-s=(1-u)x,\ s=ux \Longrightarrow \dfrac{ds}{du}=x \Longrightarrow ds=xdu$. Then

$$ C(x) = \frac{\Gamma(m+1)}{\Gamma(\beta+m+1)}x^{\beta+m}. \tag{32} $$

□

Lemma 5. Let $y(x)$ be the solution of (1) and (2). Then the numerical result gives

$$ y(x) = \phi(x_i)V^{-1}(x_i)\ W(x_i), \tag{33} $$

where

$$ \begin{aligned} V(x_i) = &\frac{\Gamma(n+1)\Gamma(n-\alpha_j+p_j+1)}{\Gamma(n-\alpha_j+1)\Gamma(\beta+n-\alpha_j+p_j+1)}x_i^{\beta+n-\alpha_j+p_j} \\ &+\frac{b^{r+n+1}\Gamma(r+1)}{(\sigma+n+1)\,\Gamma(\beta+r+1)}x^{\beta+r} \\ &+\frac{\Gamma(r+\sigma+n+2)}{(\sigma+n+1)\,\Gamma(\beta+r+\sigma+n+2)}x^{\beta+r+\sigma+n+1} \end{aligned} $$

and

$$ W(x_i) = -\sum_{k=0}^{N}\frac{y^{(k)}(0)}{k!}x_i^k + \frac{\Gamma(m+1)}{\Gamma(\beta+m+1)}x_i^{\beta+m}. $$

Proof. Approximate solution of (11) is

$$ y(x) = \phi(x)\ \mathbf{A}. $$

From (19) we have $\mathbf{A} = V^{-1}(x_i)\ W(x_i)$ where

$$ \begin{aligned} V(x_i) = &\frac{\Gamma(n+1)\Gamma(n-\alpha_j+p_j+1)}{\Gamma(n-\alpha_j+1)\Gamma(\beta+n-\alpha_j+p_j+1)}x_i^{\beta+n-\alpha_j+p_j} \\ &+\frac{b^{r+n+1}\Gamma(r+1)}{(\sigma+n+1)\,\Gamma(\beta+r+1)}x^{\beta+r} \\ &+\frac{\Gamma(r+\sigma+n+2)}{(\sigma+n+1)\,\Gamma(\beta+r+\sigma+n+2)}x^{\beta+r+\sigma+n+1}. \end{aligned} $$

Substituting for **A** in the approximate solution gives the numerical result

$$y(x) = \phi(x_i)V^{-1}(x_i)\ W(x_i).$$

□

4 Uniqueness of the solution

In this section, we establish the uniqueness of the method by introducing the following hypothesis:

$$\begin{aligned}
&H_1 : q^* = \max_{x\in[0,1]} |q(x)|\,,\\
&H_2 : k_1^* = \max_{x\in[0,1]} \int_0^b |k_1(x,t)|\,dt,\\
&H_3 : k_2^* = \max_{x\in[0,1]} \int_0^x |k_2(x,t)|\,dt,\\
&H_4 : \left|y_N^{(m_j)} - y^{(m_j)}\right| \le L_{m_j}\,|y_N - y|\,,\\
&H_5 : u = \max_{x\epsilon j} \sum_{x\epsilon j}^{N} \frac{L_{m_j}}{\Gamma(m_j-\alpha+1)}.
\end{aligned}$$

Lemma 6. [q-contraction] Let $T : X \longrightarrow X$ be a mapping defined by Theorem 1 for $y_1,\ y_2\ \in X$. Then T is q-contraction if and only if

$$\frac{1}{\Gamma(\beta+1)}\left[\frac{uq_j^*}{\Gamma(m_j-\alpha_j+1)} + K_1^* + K_2^*\right] < 1.$$

Moreover, there exist a unique solution of T.

Proof. We have

$$\begin{aligned}
(Ty_1)(x) = W(x) &+ \sum_{j=0}^{N} \frac{1}{\Gamma(m_j-\alpha_j)}\frac{1}{\Gamma(\beta)}\int_0^x (x-s)^{\beta-1} q_j(s)\\
&\times\left[\int_0^s (s-t)^{m_j-\alpha_j-1} y_1^{(m_j)}(t)dt\right] ds + \frac{1}{\Gamma(\beta)}\int_0^x (x-s)^{\beta-1}\\
&\times\left[\int_0^b k_1(s,t)y_1(t)dt + \int_0^s k_2(s,t)y_1(t)dt\right] ds
\end{aligned}$$

and

$$(Ty_2)(x) = W(x) + \sum_{j=0}^{N} \frac{1}{\Gamma(m_j-\alpha_j)}\frac{1}{\Gamma(\beta)}\int_0^x (x-s)^{\beta-1} q_j(s)$$

$$\times \left[\int_0^s (s-t)^{m_j-\alpha_j-1} y_2^{(m_j)}(t)dt\right] ds + \frac{1}{\Gamma(\beta)} \int_0^x (x-s)^{\beta-1}$$
$$\times \left[\int_0^b k_1(s,t)y_2(t)dt + \int_0^s k_2(s,t)y_2(t)dt\right] ds$$

$$\begin{aligned}
|(Ty_1)(x) - (Ty_2)(x)| = \sum_{j=0}^{N} \frac{1}{\Gamma(m_j-\alpha_j)} \frac{1}{\Gamma(\beta)} \int_0^x (x-s)^{\beta-1} |q_j(s)| \\
\times \left[\int_0^s (s-t)^{m_j-\alpha_j-1} \left|y_1^{(m_j)}(t) - y_2^{(m_j)}(t)\right| dt\right] ds \\
+ \frac{1}{\Gamma(\beta)} \int_0^x (x-s)^{\beta-1} \\
\times \left[\begin{array}{c} \int_0^b |k_1(s,t)| \, |y_1(t) - y_2(t)| \, dt \\ + \int_0^s |k_2(s,t)| \, |(y_1(t) - y_2(t))| \, dt \end{array}\right] ds.
\end{aligned}$$

Taking maximum of both sides and using H_1 *to* H_5 give

$$d\left(Ty_1(x), Ty_2(x)\right) \leq \frac{1}{\Gamma(\beta+1)} \left[\frac{uq_j^*}{\Gamma(m_j-\alpha_j+1)} + K_1^* + K_2^*\right] d(y_N, y).$$

Since T is a contraction,

$$\frac{1}{\Gamma(\beta+1)} \left[\frac{uq_j^*}{\Gamma(m_j-\alpha_j+1)} + K_1^* + K_2^*\right] < 1.$$

□

5 Convergence analysis

In this section, we establish the convergence of the method by substituting the approximate solution into (3.0). We have

$$\begin{aligned}
y_N(x) = W(x) + \sum_{j=0}^{N} \frac{1}{\Gamma(m_j-\alpha_j)} \frac{1}{\Gamma(\beta)} \\
\times \int_0^x (x-s)^{\beta-1} q_j(s) \left[\int_0^s (s-t)^{m_j-\alpha_j-1} y_N^{(m_j)}(t)dt\right] ds \\
+ \frac{1}{\Gamma(\beta)} \int_0^x (x-s)^{\beta-1} \left(\int_0^b k_1(s,t)y_N(t)dt\right) ds \\
+ \frac{1}{\Gamma(\beta)} \int_0^x (x-s)^{\beta-1} \left(\int_0^s k_2(s,t)y_N(t)dt\right) ds. \qquad (34)
\end{aligned}$$

Subtracting (9) from (34) gives

$$E_N(x) = y_N(x) - y(x).$$

Hence

$$\begin{aligned}
&|E_N(x)| \\
&\leq \frac{1}{\Gamma(\beta)} \int_0^x (x-s)^{\beta-1} \sum_{j=0}^{N} \frac{1}{\Gamma(m_j-\alpha_j)} q_j(s) \left| \left[\int_0^s (s-t)^{m_j-\alpha_j-1} E_N(t) dt \right] ds \right| \\
&\quad + \frac{1}{\Gamma(\beta)} \int_0^x (x-s)^{\beta-1} \left[\left| \int_0^b k_1(s,t) E_N(t) dt \right| + \left| \int_0^s k_2(s,t) E_N(t) dt \right| \right] ds.
\end{aligned}$$

Therefore

$$\begin{aligned}
&\frac{\|E_N(x_i)\|_\infty}{\|E_N(t)\|_\infty} \\
&\leq \frac{1}{\Gamma(\beta)} \int_0^{x_i} (x-s)^{\beta-1} \left| \begin{bmatrix} \sum_{j=0}^{N} \frac{1}{\Gamma(m_j-\alpha_j)} q_j(s) \left[\int_0^s (s-t)^{m_j-\alpha_j-1} dt \right] \\ + \left[\int_0^b k_1(s,t)\, dt + \int_0^s k_2(s,t)\, dt \right] \end{bmatrix} \right| ds.
\end{aligned}$$

The method of solution converges.

6 Numerical examples

In this section, we present numerical examples to evaluate the effectiveness and clarity of the method. A MAPLE 18 program is used to perform the computations. Let $y_n(x)$ and $y(x)$ be the approximate and exact solutions, respectively. Error $_N = |y_n(x) - y(x)|$.

Example 1. [6] Consider the following multi-order Fractional integro-differential equation:

$$D^{1.7}y(x) = x^2 D^{1.5}y(x) + xD^{0.5}y(x) - \int_0^x (x-t)\, y(t) dt - \int_0^1 (x+t)\, y(t) dt + f(x)$$

with this condition $y^{'}(0) = y(0) = 0$ and exact solution $y(x) = x^2 + x^3$, and $f(x) = \left(\frac{\Gamma(3)}{\Gamma(1.5)} + \frac{\Gamma(3)}{\Gamma(2.5)}\right) x^{2.5} + \left(\frac{\Gamma(4)}{\Gamma(2.5)} + \frac{\Gamma(4)}{\Gamma(3.5)}\right) x^{3.5} - \frac{\Gamma(3)}{\Gamma(1.3)} x^{0.3} - \frac{\Gamma(4)}{\Gamma(2.3)} x^{1.3} - \frac{x^4}{12} - \frac{x^5}{20} - \frac{7x}{12} - \frac{9}{20}$.

Solution 1. Comparing with (1) and (2), $\beta = 1.7, \alpha_1 = 1.5, \alpha_2 = 0.5, k_1(x,t) = (x+t)\,, k_2(x,t) = (x-t)$.

Using $N = 3$ for illustration. Applying (6) gives

$$
\begin{aligned}
y(x) = W(x) &+ \frac{1}{\Gamma(2-1.5)}\frac{1}{\Gamma(1.7)}\int_0^x (x-s)^{1.7-1}s^2 \\
&\left[\int_0^s (s-t)^{2-1.5-1}\frac{\Gamma(n+1)}{\Gamma(n-2+1)}t^{n-2}dt\right] ds\mathbf{A} \\
&+ \frac{1}{\Gamma(1-0.5)}\frac{1}{\Gamma(1.7)}\int_0^x (x-s)^{1.7-1} \\
&s\left[\int_0^s (s-t)^{1-0.5-1}\frac{\Gamma(n+1)}{\Gamma(n-1+1)}t^{n-1}dt\right] ds\ \mathbf{A} \\
&- \frac{1}{\Gamma(1.7)}\int_0^x (x-s)^{1.7-1} \\
&\left[\begin{array}{l} \left(x\frac{\Gamma(n+1)}{\Gamma(n+m+1)}1^{m+n} + \frac{\Gamma(n+1)}{\Gamma(n+m+1)}1^{m+n}\right) \\ +\left(x\frac{\Gamma(n+1)}{\Gamma(n+m+1)}x^{m+n} - \frac{\Gamma(n+1)}{\Gamma(n+m+1)}x^{m+n}\right)\end{array}\right] ds\ \mathbf{A}, \quad (35)
\end{aligned}
$$

where

$$
W(x) = \frac{1}{\Gamma(\beta)}\int_0^x (x-s)^{\beta-1}f(s)ds. \quad (36)
$$

Substituting $f(s)$ into (36) gives

$$
\begin{aligned}
W(x) = &\frac{1}{\Gamma(1.7)}\int_0^x (x-s)^{1.7-1} \\
&\left[\begin{array}{c} \left(\frac{\Gamma(3)}{\Gamma(1.5)} + \frac{\Gamma(3)}{\Gamma(2.5)}\right)s^{2.5} + \left(\frac{\Gamma(4)}{\Gamma(2.5)} + \frac{\Gamma(4)}{\Gamma(3.5)}\right)s^{3.5} \\ -\frac{\Gamma(3)}{\Gamma(1.3)}s^{0.3} - \frac{\Gamma(4)}{\Gamma(2.3)}s^{1.3} - \frac{s^4}{12} - \frac{s^5}{20} - \frac{7s}{12} - \frac{9}{20}\end{array}\right] ds. \quad (37)
\end{aligned}
$$

Simplify further

$$
\begin{aligned}
W(x) = &\left(\frac{\Gamma(3)}{\Gamma(1.5)} + \frac{\Gamma(3)}{\Gamma(2.5)}\right)\frac{1}{\Gamma(1.7)}\int_0^x (x-s)^{1.7-1}x^{2.5}ds \\
&+ \left(\frac{\Gamma(4)}{\Gamma(2.5)} + \frac{\Gamma(4)}{\Gamma(3.5)}\right)\frac{1}{\Gamma(1.7)}\int_0^x (x-s)^{1.7-1}x^{3.5}ds \\
&- \frac{\Gamma(3)}{\Gamma(1.3)}\frac{1}{\Gamma(1.7)}\int_0^x (x-s)^{1.7-1}x^{0.3}ds \\
&- \frac{\Gamma(4)}{\Gamma(2.3)}\frac{1}{\Gamma(1.7)}\int_0^x (x-s)^{1.7-1}x^{1.3}ds \\
&- \frac{1}{12\Gamma(1.7)}\int_0^x (x-s)^{1.7-1}x^4ds - \frac{1}{20\Gamma(1.7)}\int_0^x (x-s)^{1.7-1}x^5ds \\
&- \frac{7}{12\Gamma(1.7)}\int_0^x (x-s)^{1.7-1}xds - \frac{9}{20}\frac{1}{\Gamma(1.7)}\int_0^x (x-s)^{1.7-1}x^0ds
\end{aligned}
$$

$$
\begin{aligned}
W(x) = & \left(\frac{\Gamma(3)}{\Gamma(1.5)} + \frac{\Gamma(3)}{\Gamma(2.5)}\right) \frac{\Gamma(2.5+1)}{\Gamma(1.7+2.5+1)} x^{1.7+2.5} \\
& \left(\frac{\Gamma(4)}{\Gamma(2.5)} + \frac{\Gamma(4)}{\Gamma(3.5)}\right) \frac{\Gamma(3.5+1)}{\Gamma(1.7+3.5+1)} x^{1.7+3.5} \\
& - \frac{\Gamma(3)}{\Gamma(1.3)} \frac{\Gamma(0.3+1)}{\Gamma(1.7+0.3+1)} x^{1.7+0.3} \\
& - \frac{\Gamma(4)}{\Gamma(2.3)} \frac{\Gamma(1.3+1)}{\Gamma(1.7+1.3+1)} x^{1.7+1.3} \\
& - 12 \frac{\Gamma(4+1)}{\Gamma(1.7+4+1)} x^{1.7+4} - 20 \frac{\Gamma(5+1)}{\Gamma(1.7+5+1)} x^{1.7+5} \\
& - \frac{7\Gamma(1+1)}{12\Gamma(1.7+1+1)} x^{1.7+1} - \frac{9}{20} \frac{\Gamma(0+1)}{\Gamma(1.7+0+1)} x^{1.7+0}. \qquad (38)
\end{aligned}
$$

Substituting (38) into (35) gives

$$y(x) = \phi(x_i) V^{-1}(x_i)\ W(x_i).$$

We obtain the result

$$y_3 = \begin{pmatrix} 1.8956614056 \times 10^{-10} + 1.4273998650 \times 10^{-12} x \\ +0.9999999968 x^2 + 1.0000000024 x^3 \end{pmatrix}.$$

Table 1: Exact and approximate values of Example 1

x	**Exact**	**N=3**	**N=4**	**N=6**
0.25	0.0781250000	0.0781249983	0.0781249999	0.0781250000
0.5	0.3750000000	0.3749999996	0.3749999999	0.3750000000
0.75	0.9843750000	0.9843749992	0.9843749995	0.9843749999
1.0	2.000000000	1.9999999990	2.0000000000	2.0000000000

Table 2: Absolute Error for Example 1

x	$\mathbf{ERR}_3$	$\mathbf{ERR}_4$	$\mathbf{ERR}_6$	$[12]_{64}$	$[6]_7$
0.25	1.7e-9	1.0e-10	0.0	2,45e-4	1.000e-5
0.5	4.0e-10	1.0e-10	0.0	1.375e-3	1.200e-5
0.75	8.0e-10	5.0e-10	1.0e-10	5.387e-3	2.13e-4
1.0	1.0e-9	0.0	0.0	4.166e-3	8.970e-4

Example 2. [5] Consider multi-order Fractional integro-differential equation of the form

$$D^2y(x) = -D^{1.5}y(x) - y(x) + \int_0^1 y(t)dt + x - \frac{1}{2}$$

with the condition $y(0) = y^{'}(0) = 1$, and the exact solution is $y(x) = x + 1$.

Solution 2. Comparing with (1) and (2), $\beta = 2, \sim \beta = 1.5, h(x) = x - \frac{1}{2}$.

Use $N = 3$ for illustration.
Write in the integral form

$$\begin{aligned} y(x) = W(x) - \frac{1}{\Gamma(2-1.5)}\frac{1}{\Gamma(2)}\int_0^x (x-s)^{2-1} \\ \left[\int_0^s (s-t)^{2-1.5-1}\frac{\Gamma(n+1)}{\Gamma(n-2+1)}t^{n-2}dt\right] ds\ \mathbf{A} \\ - \frac{1}{\Gamma(2)}\int_0^x (x-s)^{2-1}s^n ds\mathbf{A} + \frac{1}{\Gamma(2)}\int_0^x (x-s)^{2-1} \\ \left[\int_0^1 t^n dt\right] ds\ \mathbf{A} \end{aligned} \tag{39}$$

where

$$W(x) = \frac{1}{\Gamma(2)}\int_0^x (x-s)^{2-1}\left(s - \frac{1}{2}\right) ds \tag{40}$$

$$\begin{aligned} y(x) = W(x) + \frac{1}{\Gamma(2)}\int_0^x (x-s)^{2-1} \\ \left[\frac{\Gamma(n+1)}{\Gamma(n-0.5)}s^{n+0.5}\right] ds\ \mathbf{A} \\ - \frac{1}{\Gamma(2)}\int_0^x (x-s)^{2-1}s^n ds\mathbf{A} + \frac{1}{\Gamma(2)}\int_0^x (x-s)^{2-1} \\ \left[\int_0^1 t^n dt\right] ds\ \mathbf{A} \end{aligned}$$

$$\begin{aligned} y(x) = W(x) + \frac{\Gamma(n+1)\Gamma(n+1.5)}{\Gamma(n-0.5)\Gamma(n+3.5)}x^{n+4.5}ds\ \mathbf{A} \\ - \frac{\Gamma(n+1)}{\Gamma(n+3)}x^{n+4}\mathbf{A} + \frac{\Gamma(n+1)}{\Gamma(n+4)}1^{n+5}\ \mathbf{A} \end{aligned} \tag{41}$$

$$W(x) = \frac{\Gamma(2)}{\Gamma(4)}x^5 - \frac{\Gamma(1)}{2\Gamma(3)}x^4 \tag{42}$$

for $n = 0(1)N$. Applying (41) and (42) gives

$$y(x) = \phi(x_i)V^{-1}(x_i)\ W(x_i).$$

we obtain the result

$$y_3(x) = \begin{pmatrix} 1.0000000000 + 1.0000000000x + \\ 8.8817841970 \times 10^{-16}x^2 + 2.2204460493 \times 10^{-16}x^3 \end{pmatrix}.$$

Table 3: Exact and approximate values of Example 2

x	Exact	N=3	Absolute Error
0.2	1.2000000000	1.2000000000	0.00
0.4	1.4000000000	1.4000000000	0.00
0.6	1.6000000000	1.6000000000	0.00
0.8	1.8000000000	1.8000000000	0.00
1.0	2.0000000000	2.0000000000	0.00

Example 3. [5] consider fractional fredholm integro-differential equations of the form

$$D^{1.5}y(x) = D^{0.5}y(x) + \int_0^1 e^x y(t)dt + f(x),$$

where $f(x) = e^x - e^{x+1}$ with the condition $y(0) = 0$ and exact solution $y(x) = e^x$.

Solution 3. Comparing with (1) and (2), we have $\beta = 1.5, \alpha = 0.5, \sim k(x,t) = e^x, \sim f(x) = e^x - e^{x+1}$.

Write in the integral form

$$\begin{aligned} y(x) = W(x) + \sum_{j=0}^{N} \frac{1}{\Gamma(m_j - \alpha_j)} \frac{1}{\Gamma(\beta)} \int_0^x (x-s)^{\beta-1} \\ \left[\int_0^s (s-t)^{m_j-\alpha_j-1} \frac{\Gamma(n+1)}{\Gamma(n-m_j+1)} t^{n-m_j} dt\right] ds\ \mathbf{A} \\ - \frac{1}{\Gamma(\beta)} \int_0^x (x-s)^{\beta-1} \left[\int_0^1 e^s\ t^n dt\right] ds\ \mathbf{A}, \end{aligned} \tag{43}$$

where

$$W(x) = \frac{1}{\Gamma(\beta)} \int_0^x (x-s)^{\beta-1} f(s)ds. \tag{44}$$

Use $N = 3$ for illustration.

Substituting for $\beta = 1.5, \alpha = 0.5, f(x) = e^x - e^{x+1}$ in (43) and (44) gives

$$\begin{aligned} y(x) = W(x) + \frac{1}{\Gamma(1-0.5)} \frac{1}{\Gamma(1.5)} \int_0^x (x-s)^{1.5-1} \\ \left[\int_0^s (s-t)^{1-0.5-1} \frac{\Gamma(n+1)}{\Gamma(n-1+1)} t^{n-1} dt\right] ds\ \mathbf{A} \end{aligned}$$

$$+\frac{1}{\Gamma(1.5)}\int_0^x (x-s)^{1.5-1}\left[\int_0^1 e^s\ t^n dt\right] ds\ \mathbf{A}$$

$$W(x)=\frac{1}{\Gamma(1.5)}\int_0^x (x-s)^{1.5-1}\left(e^s-e^{s+1}\right) ds$$

for $n=0(1)N$. Applying Lemma 4 gives

$$y(x)=W(x)+\frac{\Gamma(n+1)}{\Gamma(n+0.5)}\frac{1}{\Gamma(1.5)}\int_0^x (x-s)^{1.5-1}s^{n+1.5}ds\ \mathbf{A}$$
$$+\frac{1}{\Gamma(1.5)}\int_0^x (x-s)^{1.5-1}\left[\int_0^1 e^s\ t^n dt\right] ds\ \mathbf{A}, \tag{45}$$

where

$$W(x)=\frac{1}{\Gamma(1.5)}\int_0^x (x-s)^{1.5-1}\frac{s^n}{\Gamma(n+1)}ds$$
$$-\frac{1}{\Gamma(1.5)}\int_0^x (x-s)^{1.5-1}\frac{(s+1)^n}{\Gamma(n+1)}ds. \tag{46}$$

Using (45) and (46) gives

$$y(x)=\phi(x_i)V^{-1}(x_i)\ W(x_i).$$

We obtain the result

$$y_3=\begin{pmatrix} 0.9990233401+1.0116982759x+ \\ 0.4050677749x^2+0.3003042742x^3 \end{pmatrix}.$$

Table 4: Exact and approximate values of Example 3

x	Exact	N=3	N=5	N=6
0.2	1.2214027580	1.2199681400	1.2213960720	1.2214031090
0.4	1.4918246980	1.4877329680	1.4918178290	1.4918249590
0.6	1.8221188000	1.8167324280	1.8221150110	1.8221190520
0.8	2.2255409280	2.2213811250	2.2255348760	2.2255409420
1.0	2.7182818280	2.7160936650	2.7182828500	2.7182817820

Example 4. [16] consider the initial value problem of equation

$$D^2y(x)=x^2D^{1.5}y(x)+x^{\frac{1}{2}}D^{1.5}y(x)+x^{\frac{1}{3}}y(x)+f(x),\quad 0<x\leq 1,$$

$$y(0)=0,\ y^{'}(0)=0$$

Table 5: Absolute error for Example 3

x	$\mathbf{ERR}_3$	$\mathbf{ERR}_5$	$\mathbf{ERR}_6$	$[5]_{\mathbf{N=18}}$
0.2	1.4346180e-3	6.686e-6	3.51e-7	0.21823e-7
0.4	4.09173e-3	6.869e-6	2.61e-7	0.21586e-7
0.6	5.386372e-3	3.789e-6	2.52e-7	0.86325e-7
0.8	4159803e-3	6.052e-6	1.40e-8	0.12423-5
1.0	2.1881634e-3	1.022e-6	4.60e-8	0.83792e-5

$$f(x) = 6\pi^{1/2} - 8x^{7/2} - \frac{16}{5}x^3 - x^{10/3}\pi^{1/2}.$$

The exact solution is

$$y(x) = \pi^{1/2}x^3.$$

Solution 4. Comparing with (1) and (2), we have $\beta = 2, \alpha_1 = 1.5, \alpha_2 = 0.5, h(x) = 6\pi^{1/2} - 8x^{7/2} - \frac{16}{5}x^3 - x^{10/3}\pi^{1/2} \sim$.

Use $N = 3$ for illustration. Applying (6) gives

$$\begin{aligned} y(x) = W(x) &+ \frac{1}{\Gamma(2-1.5)}\frac{1}{\Gamma(2)}\int_0^x (x-s)^{2-1}\ s^2 \\ &\left[\int_0^s (s-t)^{2-1.5-1}\frac{\Gamma(n+1)}{\Gamma(n-2+1)}t^{n-2}dt\right] ds\ \mathbf{A} \\ &+\frac{1}{\Gamma(1-0.5)}\frac{1}{\Gamma(2)}\int_0^x (x-s)^{2-1}\ s^{1/2} \\ &\left[\int_0^s (s-t)^{1-0.5-1}\frac{\Gamma(n+1)}{\Gamma(n-1+1)}t^{n-1}dt\right] ds\ \mathbf{A} \\ &+ \frac{1}{\Gamma(2)}\int_0^x (x-s)^{2-1}\ s^{1/3}\left[seq(s^n, n=0,\ldots,N)\right] ds\ \mathbf{A}, \end{aligned} \tag{47}$$

for $n = 0(1)N$, where

$$W(x) = \frac{1}{\Gamma(2)}\int_0^x (x-s)^{2-1}\left(6\pi^{1/2}s - 8s^{7/2} - \frac{16}{5}s^3 - s^{10/3}\pi^{1/2}\right) ds.$$

Simplifying (47) gives

$$\begin{aligned} y(x) = W(x) &+ \frac{\Gamma(n+1)\Gamma(n-1)\Gamma(3.5+n)}{\Gamma(n-3)\Gamma(n-0.5)\Gamma(5.5+n)}x^{6.5+n}\mathbf{A} \\ &+ \frac{\Gamma(n-1)}{\Gamma(n+0.5)}x^{n+2}\ \mathbf{A} + \frac{\Gamma(n+0.8)}{\Gamma(n+2.8)}x^{n+3.8}\ \mathbf{A}, \end{aligned} \tag{48}$$

$$W(x) = \frac{6\pi^{1/2}\Gamma(2)}{\Gamma(4)}x^5 - \frac{8\Gamma(3.5)}{\Gamma(5.5)}x^{6.5} - \frac{16}{5}\frac{\Gamma(3)}{\Gamma(5)}x^6 - \frac{\pi^{1/2}\Gamma(3.3)}{\Gamma(5.3)}x^{6.3} \tag{49}$$

Substituting (49) into (48) gives

$$y(x) = \phi(x_i)V^{-1}(x_i)\ W(x_i).$$

We obtain the result

$$y_3 = \begin{pmatrix} -0.778452e-4x^0 - 0.212739e-4x+ \\ 0.12350293e-2x^2 + 1.7678381513x^3 \end{pmatrix}.$$

Table 6: Exact and approximate values of Example 4

x	Exact	N=3	N=5	N=7	N=10
0.1	0.0017725688	0.0017002159	0.0017710059	0.0017729787	0.0017721932
0.3	0.0478593564	0.0477585554	0.0478450493	0.0478592986	0.0478593029
0.5	0.2215710946	0.2212000441	0.2215300075	0.2215712791	0.2215710855
0.7	0.6079910837	0.6068809132	0.6079068083	0.6079913542	0.6079910989
0.9	1.2922026240	1.2896573940	1.2920556370	1.2922025970	1.2922026170

Table 7: Absolute error for Example 4

x	$\mathbf{ERR}_3$	$\mathbf{ERR}_5$	$\mathbf{ERR}_7$	$\mathbf{ERR}_{10}$	$[16]_{10}$
0.1	7.23529e-5	1.5629e-6	4.099e-7	3.756e-7	7.45873e-7
0.3	1.00801e-4	1.43072e-5	4.422e-6	5.35e-7	1.4833e-6
0.5	3.7105e-4	4.1087e-5	1.845e-7	9.1e-9	1.74701e-6
0.7	1.1101e-3	8.4275e-5	1.705e-7	1.52e-9	5.5116e-7
0.9	2.54523e-3	1.46987e-4	2.7e-8	7.0e-9	2.47276e-6

7 Discussion of results

In this section, we discuss the numerical results obtained from the solved examples using the derived numerical method.

In Example 1, the approximate solution obtained as $N = 3$ gives $y_3 = 1.8956614056 \times 10^{-10} + 1.4273998650 \times 10^{-12}x + 0.9999999968x^2 + 1.0000000024x^3$. Solving for $N = 4$ and $N = 6$, we obtained Table 1, which shows the results obtained from solving Example 1. Table 2 shows the absolute error of Example 1, and it indicates that as the values of N increase, the error becomes smaller and more consistent across all values of x. For instance, the least error of [12] at $N = 64$ is $2.45e-4$ while the least error in our method is 0.00 at $N = 6$. This confirmed that our method performed better.

In Example 2, the approximate solution obtained at $N = 3$ gives $y_3(x) = 1.00000000 + 1.0000000000x + 8.8817841970 \times 10^{-16}x^2 + 2.2204460493 \times 10^{-16}x^3$, which shows that the result converges to the exact solution as displayed in Table 3.

In Example 3, the approximate solution at $N = 3$ gives $y_3(x) = 0.9990233401 + 1.0116982759x + 0.4050677749x^2 + 0.3003042742x^3$. Solving $N = 5$ and 7, we obtained Table 4, which displays the results obtained at $x = 0.2$ *to* 1.0 for various values of N and the exact solution. The absolute error of Example 3 as shown in Table 5 indicates that as the values of N increase, the error becomes smaller. For instance, the least error in [5] at $N = 18$ is $0.21823e - 7$ while the least error in our method at $N = 6$ is $1.40e - 8$. This shows that the numerical method developed is consistent and converges faster.

In Example 4, the approximate solution at $N = 3$ gives $y_3(x) = -2.5324187192 \times 10^{-13} + 6.5978333907 \times 10^{-12}x - 1.0000000002x^2 + 1.0000000000x^3$. Solving at $N = 5$, $N = 7$, and $N = 10$, we obtained Table 6, which shows the results obtained at $x = 0.1$ *to* 0.9 for various values of N and the exact solution. Table 7 shows the absolute error of problem 1, and it indicates that as the value of N increases, the error becomes smaller. We also compare our results with [16]. For instance, the least error in [16]at $N = 10$ is $5.5116e - 7$ while the least error in our method is $2.7e - 8$ at $N = 7$. This clearly shows that our method performs better.

Hence, from the numerical results obtained, we can conclude that the numerical method derived is efficient, consistent, and computationally reliable.

8 Conclusion

An enhanced numerical method was developed for the solution of multi-order fractional integro-differential equations with initial conditions using the collocation method. The numerical method derived is consistent, efficient, and reliable. Maple code was used to implement the developed method. Solved numerical examples showed that the method is reliable and suitable for such kinds of problems.

References

[1] Agbolade, A.O. and Anake, T.A. *Solution of first order Volterra linear integro-differential equations by collocation method*, J. Appl. Math. (2017), Article ID, 1510267.

[2] Ajileye, G. and Aminu, F.A. *Approximate solution to first-order integro-differential equations Using polynomial collocation approach*, J. Appl. Computat. Math. 11 (2022), 486.

[3] Ajileye, G., James, A., Abdullahi, A. and Oyedepo, T. *Collocation approach for the computational solution of Fredholm-Volterra fractional order of integro-differential equations*, J. Niger. Soc. Phys. Sci. (2022), 834–834.

[4] Ghafoor, A., Haq, S., Rasool, A. and Baleanu, D. *An efficient numerical algorithm for the study of time fractional Tricomi and Keldysh type equations*, Engineering with Computers 38(4) (2022), 3185–3195.

[5] Gülsu, M., Öztürk, Y. and Anapalı, A. *Numerical approach for solving fractional Fredholm integro-differential equation*, Int. J. Comput. Math. 90(7) (2013), 1413–1434.

[6] Guo, N. and Ma, Y. *Numerical algorithm to solve fractional integro-differential equations based on Legendre wavelets method*, IAENG Int. J. Appl. Math. 48(2) (2018), 140–145.

[7] Huang, L., Li, X.,-F., Zhao, Y. and Duan, X.-Y. *Approximate solution of fractional integro-differential equations by Taylor expansion method*, Comput. Math. Appl. 62(3) (2011), 1127–1134.

[8] Irandoust-pakchin, S., Kheiri, H. and Abdi-mazraeh, S. *Chebyshev cardinal functions: an effective tool for solving nonlinear Volterra and Fredholm integro-differential equations of fractional order*, Iran. J. Sci. Technol. Trans. A Sci. 37 (2013), 53–62.

[9] Khan, R.H. and Bakodah, H.O. *Adomian decomposition method and its modification for nonlinear Abel's integral equations*, Int. J. Math. Anal. (Ruse) 7 (45-48) (2013), 2349–2358.

[10] Li, C. and Wang, Y. *Numerical algorithm based on Adomian decomposition for fractional differential equations*, Comput. Math. Appl. 57(10) (2009), 1672–1681.

[11] Lotfi, A. Dehghan, M. and Yousefi, S.A. *A numerical technique for solving fractional optimal control problems*, Comput. Math. Appl. 62(3) (2011), 1055–1067.

[12] Ma, Y., Wang, L. and Meng, Z. *Numerical algorithm to solve fractional integro-differential equations based on operational matrix of generalized block pulse functions*, CMES - Comput. Model. Eng. Sci. 96(1) (2013), 31–47.

[13] Mohammed, D.Sh. *Numerical solution of fractional integro-differential equations by least squares method and shifted Chebyshev polynomial*, Math. Probl. Eng. (2014), Art. ID 431965, 5 pp.

[14] Nawaz, Y. *Variational iteration method and homotopy perturbation method for fourth-order fractional integro-differential equations*, Comput. Math. Appl. 61(8) (2011), 2330–2341.

[15] Rani, D. and Mishra, V. *Solutions of Volterra integral and integro-differential equations using modified Laplace Adomian decomposition method*, J. Appl. Math. Stat. Inform. 15(1) (2019), 5–18.

[16] Rostamy, D., Alipour, M., Jafari, H. and Baleanu, D. *Solving multi-term orders fractional differential equations by operational matrices of BPs with convergence analysis*, Rom. Rep. Phys. 65(2) (2013), 334–349.

[17] Thabet, H., Kendre, S. and Unhale, S. *Numerical analysis of iterative fractional partial integro-differential equations*, J. Math. (2022), Art. ID 8781186, 14 pp.

[18] Yang, C. and Hou, J. *Numerical solution of Volterra integro-differential equations of fractional order by Laplace decomposition method*, International Journal of Mathematical and Computational Sciences 7(5) (2013), 863–867.

[19] Zhou, Y. *Basic theory of fractional differential equations*, World Scientific Publishing Co. Pte. Ltd., Hackensack, NJ, 2014.

How to cite this article
Ajileye, G., Oyedepo, T., Adiku, L. and Sabo, J., Collection-based numerical method for multi-order fractional integro-differential equations. *Iran. J. Numer. Anal. Optim.*, 2023; 13(4): 604-626. https://doi.org/10.22067/ijnao.2023.81586.1232

Iranian Journal of Numerical Analysis and Optimization
Vol. 13, No. 4, 2023, pp 627–645
https://doi.org/10.22067/ijnao.2023.80721.1214
https://ijnao.um.ac.ir/

A robust uniformly convergent scheme for two parameters singularly perturbed parabolic problems with time delay

N.T. Negero

Abstract

A singularly perturbed time delay parabolic problem with two small parameters is considered. The paper develops a finite difference scheme that is exponentially fitted on a uniform mesh in the spatial direction and uses the implicit-Euler method to discretize the time derivative in the temporal direction in order to obtain a better numerical approximation to the solutions of this class of problems. We establish the parameter-uniform error estimate and discuss the stability of the suggested approach. In order to demonstrate the improvement in terms of accuracy, numerical results are also shown to validate the theoretical conclusions and are contrasted with the current hybrid scheme.

AMS subject classifications (2020): Primary 65M06; Secondary 65M12, 65L11.

Keywords: Singular perturbation, Two parameters parabolic convection-diffusion problem, Time delay, Fitted operator scheme, the Error estimate

1 Introduction

We deal with the following class of singularly perturbed parabolic initial-boundary-value problems (IBVPs) on the domain $D = \Omega_x \times (0, T]$, $\Omega_x = (0, 1)$:

$$\begin{cases} \$_{\varepsilon,\mu} u(x,t) \equiv u_t - \varepsilon u_{xx} - \mu a(x,t) u_x + b(x,t) u = w(x,t), \\ u(x,t) = \phi_b(x,t), \qquad (x,t) \in \Gamma_b = [0,1] \times [-\tau, 0], \\ u(0,t) = \phi_l(t), \qquad \Gamma_l = \{(0,t) : 0 \le t \le T\}, \\ u(1,t) = \phi_r(t), \qquad \Gamma_r = \{(1,t) : 0 \le t \le T\}, \end{cases} \tag{1}$$

Received 21 January 2023; revised 20 April 2023; accepted 7 June 2023

Naol Tufa Negero
Department of Mathematics, Wollega University, Nekemte, Ethiopia. e-mail: natitfa@gmail.com

where $w(x,t) = -c(x,t)u(x,t-\tau)+f(x,t), (x,t) \in D$. Here in this paper, $\Gamma = \Gamma_b \cup \Gamma_l \cup \Gamma_r$ with the parameters ε and μ such that $0 < \varepsilon \leq 1, 0 \leq \mu \leq 1$, and $\tau > 0$ represents the delay parameter and the functions $a(x,t), b(x,t), c(x,t), f(x,t)$, $\phi_b(x,t), \phi_l(t)$, and $\phi_r(t)$ are sufficiently smooth, bounded functions independent of ε and μ with

$$a(x,t) \geq \alpha > 0, \quad b(x,t) \geq \beta > 0, \quad c(x,t) \geq \vartheta > 0, \qquad (x,t) \in \overline{D} = [0,T] \times [0,1]\,.$$

The type of singularly perturbed two-parameter problems changes depending on the values of the perturbation parameters ε and μ; for $\mu = 0$, the problem is a reaction-diffusion problem, whereas, for $\mu = 1$, it is a convection-diffusion problem. It is well known that due to the presence of layers, classical numerical methods using a uniform mesh cannot properly approximate the exact solution when the parameter decreases unless a large number of mesh-intervals are utilized. However, even for lesser values of the perturbation parameters, one can overcome this difficulty by employing the fitted operator technique, which works without the prior location of the boundary layer. Time delay parabolic differential equations have recently attracted increasing amounts of attention due to their widespread use in many diverse application fields, including material science, biosciences, medicine, control theory, economics, and so on; see [20, 1, 10, 21, 23]. Many researchers have discussed the numerical results of the solutions of one-parameter singularly perturbed parabolic differential equations with time delay. For instance, one can refer to the articles by Das and Natesan [3], Gowrisankar and Natesan [6], Kumar [7], Woldaregay et al. [22], and Negero and Duressa [13, 14, 15, 16, 17].

In recent years, the development of a fitted numerical scheme for solving singularly perturbed time-delay parabolic problems having two parameters has received significant attention from a few authors. One such efficient fitted numerical scheme is an upwind difference scheme, which is proposed for solving singularly perturbed time-delay parabolic problems having two parameters in [5] by Govindarao, Mohapatra, and Sahu. They constructed a method on the Shishkin type meshes (standard Shishkin mesh, Bakhvalov-Shishkin mesh) and proved that the method is first-order accurate. Negero [12] considered the same problem in [5] and produced a second-order convergent scheme using an exponentially fitted cubic spline scheme. Prior to Negero's [12] strategy, there were no developed numerical techniques for addressing two-parameter singularly perturbed time-delayed parabolic problems based on fitted operators. Kumar et al. [8] devised and analyzed a hybrid monotone finite difference scheme for singularly perturbed IBVPs of the form (1). In [8], a first-order uniformly convergent method is given for problem (1) using a hybrid monotone finite difference scheme on a rectangular mesh, which is a combination of a uniform mesh in time and a layer-adapted Shishkin mesh in space. There were no established numerical methods for dealing with two-parameter singularly perturbed time-delay parabolic problems based on

fitted operators prior to Negero's [12] strategy. Thus, the main aim of the present study is to provide robust parameter uniform convergent numerical methods based on exponentially fitted for the solution of problem (1).

Organization of the paper: In Section 2, the properties of the continuous solution are given. In Section 3, we describe the construction of an exponentially fitted finite difference discretization of problem (1). The stability and uniform convergence analysis of the suggested technique are given in Section 4. Some numerical results that validate our theory are reported in Section 5. Lastly, in Section 6, we present the conclusion of the paper.

Notations: The norm $\|\cdot\|$ is used to denote the maximum norm over the domain $\bar{D}$, defined as $\|g\| = \max_{\bar{D}} |g(x,t)|$ for a function g defined on some domain $\bar{D}$. In addition, C and its subscripts stand for positive constants independent of the perturbation parameters ε, μ, and mesh sizes.

2 Properties of the continuous solution

The required compatibility conditions at the corner points are

$$\begin{cases} \phi_b(0,0) = \phi_l(0), \\ \phi_b(1,0) = \phi_r(0), \end{cases} \tag{2}$$

$$\begin{cases} \dfrac{\partial \phi_l(0)}{\partial t} - \varepsilon \dfrac{\partial^2 \phi_b(0,0)}{\partial x^2} - \mu a(0,0) \dfrac{\partial \phi_b(0,0)}{\partial x} + b(0,0)\phi_b(0,0) \\ = -c(0,0)\phi_b(0,-\tau) + f(0,0), \\ \dfrac{\partial \phi_r(0)}{\partial t} - \varepsilon \dfrac{\partial^2 \phi_b(1,0)}{\partial x^2} - \mu a(1,0) \dfrac{\partial \phi_b(1,0)}{\partial x} + b(1,0)\phi_b(1,0) \\ = -c(1,0)\phi_b(0,-\tau) + f(1,0), \end{cases} \tag{3}$$

so that the data matches at the two corners $(0,0)$ and $(1,0)$. Let a, b, c, and f be continuous on a domain D. Then (1) has a unique solution $u \in C^2(D)$ [9].

Lemma 1 (Continuous maximum principle)**.** Let $\Phi(x,t) \in C^2(D) \cap C^0(\bar{D})$ and $\Phi(x,t) \geq 0$ for all $(x,t) \in \Gamma = \Gamma_l \cup \Gamma_b \cup \Gamma_r$. Then $\mathcal{S}_{\varepsilon,\mu}\pi(x,t) \geq 0$ in D gives $\Phi(x,t) \geq 0$, for all $(x,t) \in \bar{D}$.

Proof. Assume $(\theta^*, \zeta^*) \in D$ such that $\Phi(\theta^*, \zeta^*) = \min_{(x,t)\in\bar{D}} \Phi(x,t)$ and $\Phi(\theta^*, \zeta^*) < 0$. Then, it is easy to verify that $\mathcal{S}_{\varepsilon,\mu}\Phi(\theta^*, \zeta^*) < 0$, which is a contradiction. Thus, we have $\Phi(x,t) \geq 0$ for all $(x,t) \in \bar{D}$. □

Lemma 2. The solution $u(x,t)$ of the continuous problem (1) is bounded as

$$|u(x,t) - \phi_b(x,0)| \leq Ct.$$

Proof. Refer to [8]. □

Lemma 3. The bound on the solution $u(x,t)$ of the continuous problem (1) is given by

$$|u(x,t)| \leq C, \qquad (x,t) \in \bar{D}.$$

Proof. Refer to [8]. □

Lemma 4 (Uniform stability estimate)**.** Let $u(x,t)$ be the solution of the continuous problem in (1). Then we have the bound

$$\|u(x,t)\| \leq \beta^{-1} \|w\| + \max\left\{|\phi_b| + \max\left(|\phi_l|, |\phi_r|\right)\right\}.$$

Proof. An application of Lemma 1 to the comparison function

$$\chi^{\pm}(x,t) = \beta^{-1} \|g\| + \max\left(|\phi_b|, \left(|\phi_l| + |\phi_r|\right)\right) \pm u(x,t), \qquad (x,t) \in \bar{D},$$

yields the required estimate. □

Lemma 5. Let $u(x,t)$ be the solution of problem (1), satisfying $0 \leq i+2j \leq 4$. Then $u(x,t)$ satisfies the following bound:

$$\left\| \frac{\partial^{i+j} u}{\partial x^i \partial t^j} \right\| \leq C \begin{cases} \frac{1}{\left(\sqrt{\varepsilon}\right)^i} & \text{when} \quad \alpha\mu^2 \leq \varepsilon\eta, \\ \left(\frac{\mu}{\varepsilon}\right)^i \left(\frac{\mu^2}{\varepsilon}\right)^j & \text{when} \quad \alpha\mu^2 \geq \varepsilon\eta, \end{cases}$$

where $\eta \approx \min_{(x,t)\in\bar{D}} \frac{b(x,t)}{a(x,t)}$.

Proof. Refer to [8]. □

3 Numerical scheme formulation

3.1 Temporal discretization

The time interval $[0,T]$ is partitioned into a uniform step size as follows:

$$\Omega_t^M = \left\{t_m = m\Delta t, m = 0, 1, \ldots, M, \Delta t = T/M\right\}, \quad T = ks, \quad s = m_s \Delta t,$$

where k is a positive constant, m_s is a positive integer, Δt is the time step size, and M is the number of mesh intervals.

Hence, the problem (1) is discretized by using the implicit Euler method as follows:

$$\begin{cases} \dfrac{U^{m+1}(x)-U^{m}(x)}{\Delta t}-\varepsilon\left(U_{xx}\right)^{m+1}(x)-\mu a^{m+1}(x) \\ \left(U_{x}\right)^{m+1}(x)+b^{m+1}(x)\,U^{m+1}(x)=w^{m+1}(x), \\ U^{m}(0)=\phi_{l}\left(t_{m}\right), \quad 0\leq m\leq M, x\in\Omega_{x}, \\ U^{m}(1)=\phi_{r}\left(t_{m}\right), \quad 0\leq m\leq M, x\in\Omega_{x}, \\ U^{m+1}(x)=\phi_{b}\left(x,t_{m+1}\right), \quad -(s+1)\leq m\leq -1, \quad x\in\Omega_{x}, \end{cases} \tag{4}$$

where $w^{m+1}(x)=-c^{m+1}(x)U^{m+1-s}(x)+f^{m+1}(x)$, $0\leq m\leq M, x\in\Omega_x$ and $U^{m+1}(x)$ is the semidiscrete approximation to the exact solution $u(x,t_{m+1})$ of (1) at the $(m+1)th$ time level. Then, let us rewrite (4) in the following operator form:

$$\begin{cases} \$_{\varepsilon,\mu}^{M}U^{m+1}(x)=H\left(x,t_{m+1}\right), \\ U^{m+1}(0)=\phi_{l}\left(t_{m+1}\right), \qquad 0\leq m\leq M, \\ U^{m+1}(1)=\phi_{r}\left(t_{m+1}\right), \qquad 0\leq m\leq M, \quad x\in\Omega_{x}, \\ U^{m+1}(x)=\phi_{b}\left(x,t_{m+1}\right), \qquad -(s+1)\leq m\leq -1, \quad x\in\Omega_{x}, \end{cases} \tag{5}$$

where

$$\$_{\varepsilon,\mu}^{M}U^{m+1}(x)=-\varepsilon\left(U_{xx}\right)^{m+1}(x)-\mu a^{m+1}(x)\left(U_{x}\right)^{m+1}(x)+q^{m+1}(x)\,U^{m+1}(x)$$

and

$$H\left(x,t_{m+1}\right)=\frac{1}{\Delta t}U^{m}(x)-c^{m+1}(x)U^{m-s+1}(x)+f^{m+1}(x), \qquad 1\leq m\leq M, \quad x\in\Omega_{x},$$

for $q^{m+1}(x)=\dfrac{1}{\Delta t}+b^{m+1}(x)$.

Lemma 6 (Semidiscrete maximum principle)**.** Let $\varphi^{m+1}(x)\in C^{2}(D)\cap C^{0}\left(\bar{D}\right)$. If $\varphi^{m+1}(0)\geq 0$, $\varphi^{m+1}(1)\geq 0$, and $\$_{\varepsilon,\mu}^{M}\varphi^{m+1}(x)\geq 0$ for all $x\in D$, then $\varphi^{m+1}(x)\geq 0$ for all $x\in\bar{D}$.

Proof. One can prove this lemma by the same procedure as the proof of Lemma 1. □

Lemma 7 (Local error estimate)**.** Suppose $\dfrac{\partial^{i}u(x,t)}{\partial t^{i}}\leq C, (x,t)\in\bar{D}\times(0,T]$, $0\leq i\leq 2$. The local truncation error defined as $e_{m+1}=u(x,t_m)-U^{m}(x)$, associated to scheme (5) satisfies

$$\|e_{m+1}\|\leq C\left(\Delta t\right)^{2}, \qquad m=1,2,\ldots,M.$$

Proof. See [2]. □

Lemma 8 (Global error estimate.)**.** The global error E_{m+1} is estimated as

$$\|E_{m+1}\| \leq C\left(\Delta t\right).$$

Proof. See [3]. □

At the $(n+1)th$ time level, the characteristics equation of the homogeneous part of the differential equation (5) can be

$$\varepsilon\lambda^2\left(x\right) + \mu a^{m+1}\left(x\right)\lambda\left(x\right) - \left(b^{m+1}\left(x\right) + \frac{1}{\Delta t}\right) = 0. \tag{6}$$

Then, the roots of (5) are

$$\lambda_1\left(x\right) = \frac{-\mu a^{m+1}(x)}{2\varepsilon} + \sqrt{\left(\frac{-\mu a^{m+1}(x)}{2\varepsilon}\right)^2 + \frac{\varrho^*}{\varepsilon}} > 0,$$
$$\lambda_2\left(x\right) = \frac{-\mu a^{m+1}(x)}{2\varepsilon} - \sqrt{\left(\frac{-\mu a^{m+1}(x)}{2\varepsilon}\right)^2 + \frac{\varrho^*}{\varepsilon}} < 0,$$

where $\varrho^* = b^{m+1}\left(x\right) + \dfrac{1}{\Delta t}$. From these roots, it is possible to see the boundary layer behavior of the solution in the neighborhood of $x = 0$ and $x = 1$. Let $\varrho_0 = -\max_{x\in[0,1]}\lambda_1\left(x\right)$ and $\varrho_1 = \min_{x\in[0,1]}\lambda_2\left(x\right)$. Then we have two cases
i) When $\frac{\mu^2}{\varepsilon} \to 0$, as $\varepsilon \to 0$, $\varrho_0 \approx \varrho_1 = \sqrt{\frac{\varrho_1^*}{\varepsilon}}$, where $0 < \varrho_1^* < \varrho^*$.
ii) When $\frac{\varepsilon}{\mu^2} \to 0$, as $\mu \to 0$, $= \frac{\mu}{\varepsilon}\varrho_2^*$ and $\varrho_1 = 0$, where $0 < \varrho_2^* < \mu a^{m+1(x)}$.
Next, we give the semidiscrete bound of the solution $U^{m+1}(x)$ of the problems in (6).

Lemma 9. [8]For a fixed number $0 < p < 1$ and for a certain order k, the solution $U^m(x)$ of (5) satisfies the following derivative bound

$$\left|\frac{d^iU^m(x)}{dx^i}\right| \leq C\left(1 + \varrho_0^{-i}e^{-p\varrho_0 x} + \varrho_1^{-i}e^{-p\varrho_1(1-x)}\right), \qquad \text{for} \;\; 0 \leq i \leq k.$$

3.2 Fully discrete problem

In this section, we fully discretize the problem under consideration via an exponentially fitted finite difference scheme for space derivative discretization. On the space domain $[0,1]$, we introduce the equidistant meshes with uniform mesh length h such that

$$\Omega_x^N = \left\{x_n = nh,\; n = 1, 2, \ldots, N, x_0 = 0, x_N = 1, h = 1/N\right\},$$

where h is the step size, and N is the number of mesh points in the space direction. Using the theory applied in the asymptotic method developed

in [18], we develop an exponentially fitted numerical scheme to solve the singularly perturbed BVPs in (6). In the considered case, the boundary layer is on the left side of the domain, so for the singularly perturbed problem of (6), the zero-order approximation asymptotic solution is given as

$$\begin{aligned} U^{m+1}(x) =& U_0^{m+1}(x) + \left(\phi_l\left(t_{m+1}\right) - U_0^{m+1}(0)\right) \exp\left\{-\int_0^x \left(\frac{\mu a^{m+1}\left(x\right)}{\varepsilon}\right) dx\right\} \\ &+ O\left(\varepsilon\right), \end{aligned} \tag{7}$$

where $U_0^{m+1}(x)$ is the solution of the reduced problem in (6) obtained by setting $\varepsilon = 0$ written as

$$\begin{cases} \mu a^{m+1}\left(x\right) \dfrac{d}{dx} U_0^{m+1}\left(x\right) - q^{m+1}\left(x\right) U_0^{m+1}\left(x\right) = G^{m+1}\left(x\right), \\ U_0^{m+1}\left(0\right) = \phi_l\left(t_{m+1}\right). \end{cases} \tag{8}$$

Taking Taylor's series expansion for $a(x, t_m)$ about $x = 0$ and taking their first terms, (7) gives

$$\begin{aligned} U^{m+1}(x) =& U_0^{m+1}(x) + \left(\phi_l\left(t_{m+1}\right) - U_0^{m+1}(0)\right) \exp\left\{-\left(\frac{\mu a^{m+1}\left(0\right)}{\varepsilon}\right) x\right\} \\ &+ O\left(\varepsilon\right). \end{aligned} \tag{9}$$

At the mesh $x_n = nh$, (9) becomes

$$\begin{aligned} U^{m+1}(nh) =& U_0^{m+1}(nh) + \left(\phi_l\left(t_{m+1}\right) - U_0^{m+1}(0)\right) \exp\left\{-\left(\frac{\mu a^{m+1}\left(x\right)}{\varepsilon}\right)(nh)\right\} \\ &+ O\left(\varepsilon\right). \end{aligned}$$

Therefore,

$$\lim_{h\to 0} U^{m+1}(nh) = U_0^{m+1}(0) + \left(\phi_l\left(t_{m+1}\right) - U_0^{m+1}(0)\right) \exp\left\{-\mu a^{m+1}\left(0\right) n\rho\right\}, \tag{10}$$

where $\rho = \frac{\mu h}{\varepsilon}$.

Now, we consider the derivative approximation of the problem in (1) and (2) as

$$D^- U_n = \frac{U_n - U_{n-1}}{h},\ D^+ U_n = \frac{U_{n+1} - U_n}{h},\ D^0 U_n = \frac{U_{n+1} - U_{n-1}}{2h}, \text{ and}$$
$$D^+ D^- U_n = \frac{U_{n+1} - 2U_n + U_{n-1}}{h^2},$$

and

$$\begin{aligned}\varepsilon\sigma\left(\rho,\varepsilon,\mu\right)D^{+}D^{-}U^{m+1}\left(x_n\right)+\mu a^{m+1}\left(x_n\right)D^{0}U^{m+1}\left(x_n\right)\\ -q^{m+1}\left(x_n\right)U^{m+1}\left(x_n\right)=G\left(x_n,t_{m+1}\right),\end{aligned}\tag{11}$$

where $\sigma\left(\rho,\varepsilon,\mu\right)$ is a fitting factor.
Multiplying (11) by h and evaluating the limit as $h\to 0$ give

$$\begin{aligned}&\lim_{h\to 0}\left[\frac{\sigma\left(\rho,\varepsilon,\mu\right)}{\rho}\left(U_{n+1}^{m+1}-2U_{n}^{m+1}+U_{n-1}^{m+1}\right)\right]\\&+\frac{1}{2}a^{m+1}\left(nh\right)\left(U_{n+1}^{m+1}-U_{n-1}^{m+1}\right)=0.\end{aligned}\tag{12}$$

Substituting (10) into (12) and taking $a(x,t)=a$ constant with some manipulation give the fitting factor as

$$\sigma\left(\rho,\varepsilon,\mu\right)=a^{m+1}\left(0\right)\tfrac{\rho}{2}\coth\left(\tfrac{\rho a^{m+1}(0)}{2}\right).$$

For the variable fitting factor, we define as

$$\sigma_n\left(\rho,\varepsilon,\mu\right)=a^{m+1}\left(x_n\right)\frac{\rho}{2}\coth\left(\frac{\rho a^{m+1}\left(x_n\right)}{2}\right).\tag{13}$$

Hence, using (12) ,the resulting finite difference scheme can be given as

$$\begin{aligned}\$_{\varepsilon,\mu}^{N,M}U_m^{n+1}\equiv&\left(\frac{\varepsilon\sigma_n\left(\rho,\varepsilon,\mu\right)}{h^2}-\frac{1}{2}\mu a_n^{m+1}\right)U_{n-1}^{m+1}\\&+\left(\frac{-2\varepsilon\sigma_n\left(\rho,\varepsilon,\mu\right)}{h^2}-q_n^{m+1}\right)U_n^{m+1}\\&+\left(\frac{\varepsilon\sigma_n\left(\rho,\varepsilon,\mu\right)}{h^2}+\frac{1}{2}\mu a_n^{m+1}\right)U_{n+1}^{m+1}\\=&H_n^{m+1}\end{aligned}\tag{14}$$

subject to the following conditions:

$$\begin{cases}U_0^{m+1}=\phi_l\left(t_{m+1}\right), & 0\le m\le M,\\ U_N^{m+1}=\phi_r\left(t_{m+1}\right), & 0\le m\le M,\\ U\left(x_n,t_{m+1}\right)=\phi_b\left(x_n,t_{m+1}\right), x_n\in\bar{\Omega}^N, & -(\wp+1)\le m\le -1,\end{cases}\tag{15}$$

where

$$\begin{aligned}H_n^{m+1}&=H\left(x_n,t_{m+1}\right)\\&=-\frac{1}{\Delta t}U^m\left(x_n\right)+c^{m+1}(x_n)U^{m-\wp+1}\left(x_n\right)-f^{m+1}(x_n).\end{aligned}$$

The schemes in (14) and (15) can be rewritten as

$$\$_{\varepsilon,\mu}^{N,M}U_n^{m+1}\equiv E_n^{m+1}U_{n-1}^{m+1}-F_n^{m+1}U_n^{m+1}+G_n^{m+1}U_{n-1}^{m+1}=H_n^{m+1},\tag{16}$$

where

$$
\begin{aligned}
E_n^{m+1} &= \frac{\varepsilon\sigma_n\left(\rho,\varepsilon,\mu\right)}{h^2} - \frac{1}{2}\mu a_n^{m+1},\\
F_n^{m+1} &= \frac{2\varepsilon\sigma_n\left(\rho,\varepsilon,\mu\right)}{h^2} + q_n^{m+1},\\
G_n^{m+1} &= \frac{\varepsilon\sigma_n\left(\rho,\varepsilon,\mu\right)}{h^2} + \frac{1}{2}\mu a_n^{m+1},\\
H_n^{m+1} &= -\frac{1}{\Delta t}U^m\left(x_n\right) + c^{m+1}(x_n)U^{m-\wp+1}\left(x_n\right) - f^{m+1}(x_n).
\end{aligned}
$$

From the entries $E_n^{m+1}, F_n^{m+1}, G_n^{m+1}$ of tridiagonal system of (16), it is evident that $E_n^{m+1} < 0, G_n^{m+1} < 0$ and $E_n^{m+1} + F_n^{m+1} + G_n^{m+1} > 0$. Thus the system is an M-matrix, and therefore its inverse exists, and it is positive. Hence, the tridiagonal system in (16) can be easily solved by any existing methods.

4 Stability and uniform convergence analysis

Lemma 10 (Discrete maximum principle)**.** Assume that ψ_n^{m+1} is any mesh function that satisfies $\psi_0^{m+1} \geq 0$, $\psi_N^{m+1} \geq 0$, and that $\$_{\varepsilon,\mu}^{N,M}$ is the discrete operator of (16). Then $\$_{\varepsilon,\mu}^{N,M}\psi_n^{m+1} \geq 0$, for $1 \leq n \leq N-1$, implies that $\psi_n^{m+1} \geq 0$, for $0 \leq n \leq N$.

Proof. Refer to [12]. □

Lemma 11 (Uniform stability estimate for discrete problem)**.** Let U_n^{m+1} be any mesh function such that $U_0^{m+1} = 0, U_N^{m+1} = 0$ on $0 \leq n \leq N$. Then

$$
\left|U_n^{m+1}\right| \leq \frac{\max\left|\$_{\varepsilon,\mu}^{N,M}U_n^{m+1}\right|}{q^*} + C\max\left\{\left|\phi_l(t_{m+1})\right|, \left|\phi_r(t_{m+1})\right|\right\},
$$

where $q_n^{m+1} = \dfrac{1}{\Delta t} + b\left(x_n, t_{m+1}\right) \geq q^* > 0$.

Proof. Refer to [12]. □

Theorem 1. Let $U\left(x_n, t_{m+1}\right)$ be the continuous solution of (1) and (2) and let U_n^{m+1} be the approximate solution of (16). Then, for sufficiently large N, the following error bound holds:

$$
\left|\$_{\varepsilon,\mu}^{N,M}\left(U\left(x_n, t_{m+1}\right) - U_n^{m+1}\right)\right| \leq CN^{-2}.
$$

Proof. Consider the error bound in the spatial direction as

$$
\begin{aligned}
&\left|\$_{\varepsilon,\mu}^{N,M}\left(U\left(x_n,t_{m+1}\right)-U_n^{m+1}\right)\right| \\
&\quad=\left|\$_{\varepsilon,\mu}^{N,M}U\left(x_n,t_{m+1}\right)-\$_{\varepsilon,\mu}^{N,M}U_n^{m+1}\right| \\
&\quad=\Bigg|\varepsilon\left(U_{xx}\right)^{m+1}\left(x_n\right)+\mu a^{m+1}\left(x_n\right)\left(U_x\right)^{m+1}\left(x_n\right) \\
&\qquad-\left\{\varepsilon\sigma\left(\rho,\varepsilon,\mu\right)D^+D^-U_n^{m+1}+\mu a_n^{m+1}D^0U_n^{m+1}\right\}\Bigg| \\
&\quad\leq\left|\varepsilon\sigma\left(\rho,\varepsilon,\mu\right)\left(\frac{d^2}{dx^2}-D^+D^-\right)U_n^{m+1}+\mu a_n^{m+1}\left(\frac{d}{dx}-D^0\right)U_n^{m+1}\right| \\
&\quad\leq\left|\varepsilon\left[a^{m+1}\left(x_n\right)\frac{\rho\mu}{2}\coth\left(\frac{\rho\mu a^{m+1}\left(x_n\right)}{2}\right)-1\right]D^+D^-U_n^{m+1}\right| \\
&\qquad+\left|\varepsilon\left(\frac{d^2}{dx^2}-D^+D^-\right)U_n^{m+1}\right|+\left|\mu a_n^{m+1}\left(\frac{d}{dx}-D^0\right)U_n^{m+1}\right|.
\end{aligned}
\tag{17}
$$

Now, (17) becomes

$$
\begin{aligned}
&\left|\$_{\varepsilon,\mu}^{N,M}\left(U\left(x_n,t_{m+1}\right)-U_n^{m+1}\right)\right| \\
&\qquad\leq C\mu h^2\frac{d^2U_n^{m+1}}{dx^2}+C\varepsilon h^2\frac{d^4U_n^{m+1}}{dx^4}+C\mu h^2\frac{d^3U_n^{m+1}}{dx^3}.
\end{aligned}
$$

Using Lemma 9, we have

$$
\begin{aligned}
&\left|\$_{\varepsilon,\mu}^{N,M}\left(U\left(x_n,t_{m+1}\right)-U_n^{m+1}\right)\right| \\
&\qquad\leq C\mu h^2\left(1+\omega_1^{-2}e^{-\nu\omega_1x}+\omega_2^{-2}e^{-\nu\omega_2(1-x)}\right) \\
&\qquad\quad+Ch^2\Bigg[\varepsilon\left(1+\omega_1^{-4}e^{-\nu\omega_1x}+\omega_2^{-4}e^{-\nu\omega_2(1-x)}\right) \\
&\qquad\quad+\mu\left(1+\omega_1^{-3}e^{-\nu\omega_1x}+\omega_2^{-3}e^{-\nu\omega_2(1-x)}\right)\Bigg].
\end{aligned}
$$

As $\varepsilon\to 0$ both $\omega_1^{-i}e^{-\nu\omega_1x_m}$ and $\omega_2^{-i}e^{-\nu\omega_2(1-x_m)}$ approach zero for $0\leq i\leq 4$. Thus, we obtain the following error bound:

$$
\left|\$_{\varepsilon}^{N,M}\left(U\left(x_n,t_{m+1}\right)-U_n^{m+1}\right)\right|\leq CN^{-2},
$$

since $h=N^{-1}$. □

Under the hypothesis of Lemmas 11 and 10, the following error estimate holds:

$$
\max_{0\leq n<N}\left|U\left(x_n,t_{m+1}\right)-U_n^{m+1}\right|\leq Ch=CN^{-2}. \tag{18}
$$

Theorem 2. Let $u(x,t)$ be the exact solution of (1) and (2) and let U_n^{m+1} be the numerical solution of (16). For the discrete scheme, there exist a constant C independent of ε, h and Δt such that

$$\max_{0\le n\le N, 0\le m\le M} \left|u\left(x_n, t_{m+1}\right) - U_n^{m+1}\right| \le C\left(N^{-2} + (\Delta t)\right).$$

for sufficiently large N.

Proof. The result follows from the error estimate given in Lemma 8 and Theorem 1. □

5 Numerical results

In this section, we illustrate the proposed scheme using two numerical examples of the form given in (1). We investigate the theoretical results in this paper by performing experiments using the proposed scheme. The exact solution of these two examples is not known. Thus, we use the double mesh principle to evaluate maximum absolute errors $E_{\varepsilon,\mu}^{N,M}$ and the corresponding order of convergence $p_{\varepsilon,\mu}^{N,M}$ as

$$E_{\varepsilon,\mu}^{N,M} = \max_{0\le n\le N, 0\le m\le M} \left|U_n^{m+1} - U_{2n}^{2m+1}\right|, \quad p_{\varepsilon,\mu}^{N,M} = \log_2\left(\frac{E_{\varepsilon,\mu}^{N,M}}{E_{\varepsilon,\mu}^{2N,2M}}\right).$$

From these values, we obtain the ε-uniform error $E^{N,M}$ and the corresponding ε-uniform order of convergence $p^{N,M}$ as

$$E^{N,M} = \max_{0\le n\le N, 0\le m\le M} E_{\varepsilon}^{N,M} \text{ and } p^{N,M} = \log_2\left(\frac{E^{N,M}}{E^{2N,2M}}\right),$$

where U_m^{n+1} is the numerical solutions obtained by using $N \times M$ mesh intervals in space and time direction with mesh size h and Δt, respectively.

Example 1. Consider the problem

$$\frac{\partial u}{\partial t} - \varepsilon\frac{\partial^2 u}{\partial x^2} - \mu(1+x)\frac{\partial u}{\partial x} + u(x,t) = -u(x,t-\tau) + 16x^2(1-x)^2, \quad (x,t) \in (0,1)\times(0,2],$$

with

$$\begin{cases} u(0,t) = 0, & u(1,t) = 0, \quad t \in (0,2], \\ u(x,t) = 0, & (x,t) \in [0,1]\times[-\tau,0]. \end{cases}$$

Example 2. Consider the problem

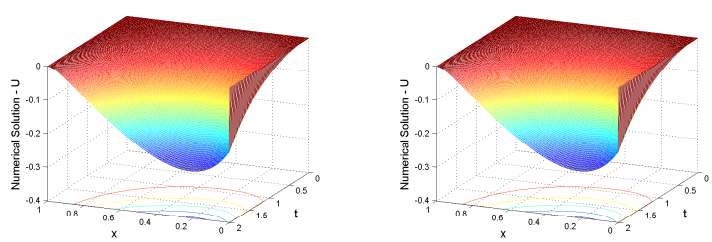
$$\frac{\partial u}{\partial t}-\varepsilon\frac{\partial^2 u}{\partial x^2}-\mu\left(1+x\left(1-x\right)+t^2\right)\frac{\partial u}{\partial x}+\left(1+5xt\right)u(x,t)$$
$$=-u(x,t-\tau)+x(1-x)\left(e^t-1\right),\qquad (x,t)\in(0,1)\times(0,2],$$

with

$$\begin{cases} u(0,t)=0, & u(1,t)=0, \qquad t\in(0,2]\,, \\ u(x,t)=0, & (x,t)\in[0,1]\times[-\tau,0]\,. \end{cases}$$

Table 1: Maximum pointwise errors ($E^{N,M}_{\varepsilon,\mu}$) and rate of convergence ($p^{N,M}_{\varepsilon,\mu}$) for Example 1.

$\mu=10^{-4}$	N=32	N=64	N=128	N=256	N=512
$\varepsilon\downarrow$	M=16	M=32	M=64	M=128	M=256
10^{-0}	5.7516e-03	3.6382e-03	2.1286e-03	1.1627e-03	6.0947e-04
	0.66074	0.77332	0.87243	0.93185	-
10^{-2}	1.0422e-02	5.4491e-03	2.7875e-03	1.4101e-03	7.0919e-04
	0.93554	0.96705	0.98317	0.99155	-
10^{-4}	1.0658e-02	5.5312e-03	2.8189e-03	1.4231e-03	7.1502e-04
	0.94627	0.97246	0.98610	0.99298	-
10^{-6}	1.0663e-02	5.5328e-03	2.8193e-03	1.4232e-03	7.1504e-04
	0.94653	0.97267	0.98620	0.99304	-
10^{-8}	1.0664e-02	5.5339e-03	2.8202e-03	1.4237e-03	7.1533e-04
	0.94638	0.97250	0.98615	0.99296	-
10^{-10}	1.0664e-02	5.5339e-03	2.8202e-03	1.4237e-03	7.1533e-04
	0.94638	0.97250	0.98615	0.99296	-
10^{-12}	1.0664e-02	5.5339e-03	2.8202e-03	1.4237e-03	7.1533e-04
	0.94638	0.97250	0.98615	0.99296	-
$E^{N,M}$	**1.0664e-02**	**5.5339e-03**	**2.8202e-03**	**1.4237e-03**	**7.1533e-04**
$p^{N,M}$	**0.94638**	**0.97250**	**0.98615**	**0.99296**	-



(a) (b)

Figure 1: Surface plot of the numerical solution for Example 2 with $N=256, M=128$, **a** $\varepsilon=10^{-1}$,$\mu=10^{-12}$ **b** $\varepsilon=10^{-12}$,$\mu=10^{-1}$

Table 2: Maximum pointwise errors ($E_{\varepsilon,\mu}^{N,M}$) and rate of convergence ($p_{\varepsilon,\mu}^{N,M}$) for Example 1.

$\mu = 10^{-12}$	Number of mesh intervals N=M				
$\varepsilon \downarrow$	32	64	128	256	512
10^{-0}	3.6218e-03	2.1253e-03	1.1619e-03	6.0925e-04	3.1225e-04
	0.57919	0.87118	0.93138	0.96433	-
10^{-2}	5.4110e-03	2.7780e-03	1.4077e-03	7.0860e-04	3.5549e-04
	0.96185	0.98071	0.99030	0.99516	-
10^{-4}	5.5307e-03	2.8187e-03	1.4231e-03	7.1501e-04	3.5838e-04
	0.97243	0.98599	0.99300	0.99647	-
10^{-6}	5.5315e-03	2.8189e-03	1.4231e-03	7.1502e-04	3.5839e-04
	0.97254	0.98610	0.99298	0.99645	-
10^{-8}	5.5315e-03	2.8189e-03	1.4231e-03	7.1502e-04	3.5838e-04
	0.97254	0.98610	0.99298	0.99645	-
10^{-10}	5.5315e-03	2.8189e-03	1.4231e-03	7.1502e-04	3.5838e-04
	0.97254	0.98610	0.99298	0.99645	-
10^{-12}	5.5315e-03	2.8189e-03	1.4231e-03	7.1502e-04	3.5838e-04
	0.97254	0.98610	0.99298	0.99645	-
$E^{N,M}$	**5.5315e-03**	**2.8189e-03**	**1.4231e-03**	**7.1502e-04**	**3.5839e-04**
$p^{N,M}$	**0.97254**	**0.98610**	**0.99298**	**0.99645**	-

Table 3: Maximum pointwise errors ($E_{\varepsilon,\mu}^{N,M}$) and rate of convergence ($p_{\varepsilon,\mu}^{N,M}$) for Example 2.

$\mu = 10^{-4}$	N=32	N=64	N=128	N=256	N=512
$\varepsilon \downarrow$	M=16	M=32	M=64	M=128	M=256
10^{-0}	2.1475e-04	1.0912e-04	5.4962e-05	2.7578e-05	1.3813e-05
	0.97674	0.98941	0.99492	0.99749	-
10^{-2}	2.1465e-03	1.1561e-03	6.0053e-04	3.0592e-04	1.5440e-04
	0.89272	0.94496	0.97308	0.98648	-
10^{-4}	2.6764e-03	1.4488e-03	7.5345e-04	3.8424e-04	1.9401e-04
	0.88544	0.94327	0.97150	0.98588	-
10^{-6}	2.6771e-03	1.4491e-03	7.5407e-04	3.8484e-04	1.9445e-04
	0.88551	0.94239	0.97044	0.98486	-
10^{-8}	2.6771e-03	1.4490e-03	7.5351e-04	3.8423e-04	1.9401e-04
	0.88561	0.94336	0.97166	0. 98584	-
10^{-10}	2.6771e-03	1.4490e-03	7.5351e-04	3.8423e-04	1.9401e-04
	0.88561	0.94336	0.97166	0. 98584	-
10^{-12}	2.6771e-03	1.4490e-03	7.5351e-04	3.8423e-04	1.9401e-04
	0.88561	0.94336	0.97166	0. 98584	-
$E^{N,M}$	**2.6771e-03**	**1.4491e-03**	**7.5407e-04**	**3.8424e-04**	**1.9445e-04**
$p^{N,M}$	**0.88551**	**0.94239**	**0.97269**	**0.98261**	-

Table 4: Maximum pointwise errors ($E^{N,M}_{\varepsilon,\mu}$) and rate of convergence ($p^{N,M}_{\varepsilon,\mu}$) for Example 2.

$\mu = 10^{-12}$	Number of mesh intervals N=M				
$\varepsilon \downarrow$	32	64	128	256	512
10^{-0}	1.3372e-04	6.1251e-05	2.9166e-05	1.4212e-05	7.0123e-06
	1.1264	1.0704	1.0372	1.0191	-
10^{-2}	1.1701e-03	6.0326e-04	3.0645e-04	1.5447e-04	7.7560e-05
	0.95578	0.97713	0.98833	0.99394	-
10^{-4}	1.4466e-03	7.5262e-04	3.8382e-04	1.9386e-04	9.7408e-05
	0.94267	0.97149	0.98541	0.99290	-
10^{-6}	1.4522e-03	7.5525e-04	3.8509e-04	1.9444e-04	9.7702e-05
	0.94321	0.97176	0.98587	0.99287	-
10^{-8}	1.4522e-03	7.5527e-04	3.8510e-04	1.9445e-04	9.7705e-05
	0.94318	0.97176	0.98583	0.99289	-
10^{-10}	1.4523e-03	7.5527e-04	3.8510e-04	1.9445e-04	9.7705e-05
	0.94328	0.97176	0.98583	0.99289	-
10^{-12}	1.4523e-03	7.5527e-04	3.8510e-04	1.9445e-04	9.7705e-05
	0.94328	0.97176	0.98583	0.99289	-
$E^{N,M}$	**1.4523e-03**	**7.5527e-04**	**3.8510e-04**	**1.9445e-04**	**9.7705e-05**
$p^{N,M}$	**0.94328**	**0.97176**	**0.98583**	**0.99289**	-

Table 5: Comparison of uniform error ($E^{N,M}$) for Example 1.

$\mu = 10^{-3}$	N=32	N=64	N=128	N=256	N=512
$\varepsilon \downarrow$	M=8	M=16	M=32	M=64	M=128
Proposed method					
10^{-4}	1.9859e-02	1.0660e-02	5.5318e-03	2.8190e-03	1.4232e-03
10^{-6}	1.9905e-02	1.0684e-02	5.5439e-03	2.8245e-03	1.4251e-03
10^{-8}	1.9905e-02	1.0684e-02	5.5440e-03	2.8252e-03	1.4262e-03
10^{-10}	1.9905e-02	1.0684e-02	5.5440e-03	2.8252e-03	1.4262e-03
10^{-12}	1.9905e-02	1.0684e-02	5.5440e-03	2.8252e-03	1.4262e-03
Method in [8]					
10^{-4}	4.3705e-2	1.6704e-2	7.3802e-3	3.7406e-3	1.8967e-3
10^{-6}	4.3471e-2	1.6596e-2	7.3290e-3	3.7218e-3	1.8873e-3
10^{-8}	4.3429e-2	1.6573e-2	7.3303e-3	3.7211e-3	1.8870e-3
10^{-10}	4.4343e-2	1.6572e-2	7.3303e-3	3.7211e-3	1.8870e-3
10^{-12}	4.4343e2	1.6572e-2	7.3303e-3	3.7211e-3	1.8870e-3

Table 6: Comparison of uniform error ($E^{N,M}$) for Example 2.

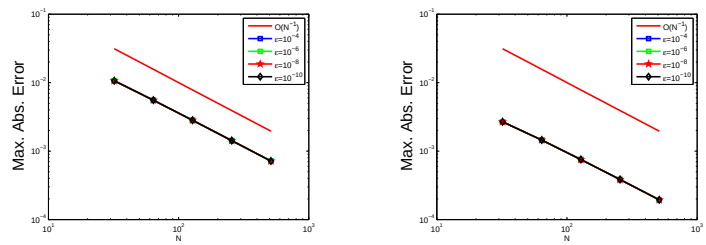
$\mu = 10^{-3}$	N=32	N=64	N=128	N=256	N=512
$\varepsilon \downarrow$	M=8	M=16	M=32	M=64	M=128
Proposed method					
10^{-4}	4.5627e-03	2.6876e-03	1.4564e-03	7.5785e-04	3.8653e-04
10^{-6}	4.5254e-03	2.6603e-03	1.4402e-03	7.4904e-04	3.8203e-04
10^{-8}	4.5254e-03	2.6603e-03	1.4402e-03	7.4904e-04	3.8195e-04
10^{-10}	4.5254e-03	2.6603e-03	1.4402e-03	7.4904e-04	3.8195e-04
10^{-12}	4.5254e-03	2.6603e-03	1.4402e-03	7.4904e-04	3.8195e-04
Method in [8]					
10^{-4}	1.1161e-2	5.1087e-3	2.4749e-3	1.2214e-3	6.0706e-4
10^{-6}	1.1008e-2	5.0450e-3	2.4437e-3	1.2073e-3	6.0036e-4
10^{-8}	1.0941e-2	5.0426e-3	2.4442e-3	1.2071e-3	6.0016e-4
10^{-10}	1.0940e-2	5.0428e-3	2.4442e-3	1.2071e-3	6.0016e-4
10^{-12}	1.0940e-2	5.0428e-3	2.4442e-3	1.2071e-3	6.0016e-4

Table 7: Comparison of uniform error ($E^{N,M}$) for Example 1.

$\mu = 10^{-9}$	N=32	N=64	N=128	N=256	N=512
$\varepsilon \downarrow$	M=8	M=16	M=32	M=64	M=128
Proposed method					
10^{-4}	1.9853e-02	1.0658e-02	5.5313e-03	2.8189e-03	1.4231e-03
10^{-6}	1.9856e-02	1.0659e-02	5.5315e-03	2.8189e-03	1.4231e-03
10^{-8}	1.9856e-02	1.0659e-02	5.5315e-03	2.8189e-03	1.4231e-03
10^{-10}	1.9856e-02	1.0659e-02	5.5315e-03	2.8189e-03	1.4231e-03
10^{-12}	1.9856e-02	1.0659e-02	5.5315e-03	2.8189e-03	1.4231e-03
Method in [8]					
10^{-4}	4.3708e-2	1.6705e-2	7.3807e-3	3.7407e-3	1.8967e-3
10^{-6}	4.3816e-2	1.6749e-2	7.4017e-3	3.7489e-3	1.9008e-3
10^{-8}	4.3817e-2	1.6750e-2	7.4019e-3	3.7490e-3	1.9008e-3
10^{-10}	4.3817e-2	1.6750e-2	7.4019e-3	3.7490e-3	1.9008e-3
10^{-12}	4.3817e-2	1.6750e-2	7.4019e-3	3.7490e-3	1.9008e-3

Table 8: Comparison of uniform error ($E^{N,M}$) for Example 2.

$\mu = 10^{-9}$	N=32	N=64	N=128	N=256	N=512
$\varepsilon \downarrow$	M=8	M=16	M=32	M=64	M=128
Proposed method					
10^{-4}	4.5499e-03	2.6744e-03	1.4477e-03	7.5307e-04	3.8398e-04
10^{-6}	4.5651e-03	2.6830e-03	1.4523e-03	7.5527e-04	3.8513e-04
10^{-8}	4.5652e-03	2.6831e-03	1.4523e-03	7.5529e-04	3.8514e-04
10^{-10}	4.5652e-03	2.6831e-03	1.4523e-03	7.5529e-04	3.8514e-04
10^{-12}	4.5652e-03	2.6831e-03	1.4523e-03	7.5529e-04	3.8514e-04
Method in [8]					
10^{-4}	1.1053e-2	5.0755e-3	2.4578e-3	1.2132e-3	6.0309e-4
10^{-6}	1.1046e-2	5.0765e-3	2.4625e-3	1.2161e-3	6.0456e-4
10^{-8}	1.1100e-2	5.0838e-3	2.4627e-3	1.2161e-3	6.0457e-4
10^{-10}	1.1093e-2	5.0782e-3	2.4639e-3	1.2162e-3	6.0457e-4
10^{-12}	1.1092e-2	5.0775e-3	2.4640e-3	1.2162e-3	6.0457e-4



(a) (b)

Figure 2: Log-Log plot of the maximum error on left (a) for Example 1 with $\mu = 10^{-4}$ and on right (b) for Example 2 with $\mu = 10^{-4}$.

We have demonstrated maximum pointwise errors ($E_{\varepsilon,\mu}^{N,M}$) and the rate of convergence ($p_{\varepsilon,\mu}^{N,M}$) for Example 1 using scheme (16) by fixing $\mu = 10^{-4}$ in Table 1 and $\mu = 10^{-12}$ in Table 2 with various values of ε. Similarly, Tables 3 and 4 have presented the result obtained for Example 2. The results given in Tables 1–4 clearly indicate that the proposed numerical method is accurate of order $O\left(N^{-2} + \Delta t\right)$, which approves the hypothetical result predicted in the theory. Numerical solutions obtained by the presented numerical scheme (16) for Example 2 have been shown in Figure 1 (a), (b), and it shows the effects of the two parameters ε and μ on the steepness of the layer of the solutions. From Figure 1 (a), we confirm the nonoccurrence of both left and right boundary layers near $x = 0$ and $x = 1$ for $\mu \to 0$ as ε becomes large. Similarly, from Figure 1 (b), we confirm the occurrence of left boundary layers near $x = 0$ for $\mu = 1$ as ε becomes small. The graphs between N and maximum pointwise errors of Examples 1 and 2 are plotted as the log-log scale, respectively, in Figure 2 (a) and (b). From these two graphs, one can observe that the numerical scheme converges ε-uniformly as the perturbation parameter goes very small. The comparison of our numerical results with that of [8] is presented in Tables 5–8. From these tables, we can confirm the improved accuracy of our proposed numerical method.

6 Conclusion

A singularly perturbed parabolic differential equation exhibiting boundary layers was considered. The considered problem contains two small perturbation parameters multiplied by the highest order derivative a term of the equation and a large delay parameter on the time variable. An exponentially fitted operator numerical scheme was proposed for solving the problem. First, the equation was approximated by equivalent singularly perturbed parabolic partial differential equations using the implicit Euler method in the time direction. Inducing an exponential fitting factor for a term with the per-

turbation parameter ε and determining its value, a fully discrete numerical scheme was developed using implicit Euler in temporal discretization and the central finite difference method for spatial discretization. The uniform stability and uniform convergence of the scheme were established. It was shown that the scheme is accurate and converges uniformly with the order of convergence $O\left(N^{-2} + (\Delta t)\right)$.

Acknowledgements

The author is grateful to his anonymous referees and editor for their constructive comments.

References

[1] Asl, F.M., and Ulsoy, A.G. *Analysis of a system of linear delay differential equations*, J. Dyn. Sys., Meas., Control, 125 (2) (2003), 215–223.

[2] Clavero, C., Jorge, J. and Lisbona, F. *A uniformly convergent scheme on a nonuniform mesh for convection-diffusion parabolic problems.* J. Comput. Appl. Math. 154(2) (2003), 415–429.

[3] Das, A. and Natesan, S. *Uniformly convergent hybrid numerical scheme for singularly perturbed delay parabolic convection-diffusion problems on Shishkin mesh*, Appl. Math. Comput. 271 (2015), 168–186.

[4] Epstein, I.R. *Delay effects and differential delay equations in chemical kinetics*, Int. Rev. Phys. Chem. 11 (1) (1992), 135–160.

[5] Govindarao, L., Sahu, S.R. and Mohapatra, J. *Uniformly convergent numerical method for singularly perturbed time delay parabolic problem with two small parameters*, Iran. J. Sci. Technol. Trans. A Sci. 43(5) (2019), 2373–2383.

[6] Gowrisankar, S. and Natesan, S. *ε- uniformly convergent numerical scheme for singularly perturbed delay parabolic partial differential equations*, Int. J. Comput. Math. 94 (2017), 902–921.

[7] Kumar, D. *A parameter-uniform scheme for the parabolic singularly perturbed problem with a delay in time*, Numer. Methods Partial Differ. Equ. 37 (1) (2021), 626–642.

[8] Kumar, S. and Kumar, M. *A robust numerical method for a two-parameter singularly perturbed time delay parabolic problem*, Comput. Appl. Math. 39(3) (2020), 1–25.

[9] Ladyzhenskaia, O.A., Solonnikov, V.A. and Ural'tseva, N.N. *Linear and quasilinear equations of parabolic type.*(Russian) Translated from the Russian by S. Smith Translations of Mathematical Monographs, Vol. 23 American Mathematical Society, Providence, R.I. 1968.

[10] McCartin, B.J. *Discretization of the semiconductor device equations,*"New problems and new solutions for device and process modeling", Boole, (1985), 72–82.

[11] Miller, J., O'Riordan, E., Shishkin, G. and Shishkina, L. *Fitted mesh methods for problems with parabolic boundary layers*, Math. Proc. R. Ir. Acad. 98A (1998), no. 2, 173–190.

[12] Negero, N.T. *A uniformly convergent numerical scheme for two parameters singularly perturbed parabolic convection-diffusion problems with a large temporal lag*, Results Appl. Math. 16 (2022), Paper No. 100338, 15 pp.

[13] Negero, N.T. and Duressa, G.F. *A method of line with improved accuracy for singularly perturbed parabolic convection-diffusion problems with large temporal lag*, Results Appl. Math. 11 (2021), 100174, 13 pp.

[14] Negero, N.T. and Duressa, G.F. *An efficient numerical approach for singularly perturbed parabolic convection-diffusion problems with large time-lag*, J. Math. Model. 10(2) (2022), 173–110.

[15] Negero, N.T. and Duressa, G.F. *Uniform convergent solution of singularly perturbed parabolic differential equations with general temporal-lag*, Iran. J. Sci. Technol. Trans. A Sci. 46(2) (2022), 507–524.

[16] Negero, N.T. and Duressa, G.F. *An exponentially fitted spline method for singularly perturbed parabolic convection-diffusion problems with large time delay*, Tamkang J. Math. (2022).

[17] Negero, N.T. and Duressa, G.F. *Parameter-uniform robust scheme for singularly perturbed parabolic convection-diffusion problems with large time-lag*, Comput. Methods Differ. Equ. 10 (4) (2022), 954–968.

[18] O'Malley Jr, R.E. *Introduction to singular perturbations,* Applied Mathematics and Mechanics, Vol. 14. Academic Press [Harcourt Brace Jovanovich, Publishers], New York-London, 1974.

[19] Roos, H.G. and Uzelac, Z. *The SDFEM for a convection-diffusion problem with two small parameters*, Dedicated to John J. H. Miller on the occasion of his 65th birthday. Comput. Methods Appl. Math. 3(3) (2003), 443–458.

[20] Tikhonov, A.N. and Samarskii, A.A. *Equations of mathematical physics*, Courier Corporation, 2013.

[21] Van Harten, A. and Schumacher, J. *On a class of partial functional differential equations arising in feed-back control theory*, Differential equations and applications (Proc. Third Scheveningen Conf., Scheveningen, (1977), pp. 161–179, North-Holland Math. Stud., 31, North-Holland, Amsterdam-New York, 1978.

[22] Woldaregay, M. M. Aniley, W. T. and Duressa,G.F. *Novel numerical scheme for singularly perturbed time delay convection-diffusion equation* Adv. Math. Phys. (2021), Art. ID 6641236, 13 pp.

[23] Wu, J. *Theory and applications of partial functional differential equations*, New York, Springer, 119, (2012).

How to cite this article
Negero, N.T., A robust uniformly convergent scheme for two parameters singularly perturbed parabolic problems with time delay. *Iran. J. Numer. Anal. Optim.*, 2023; 13(4): 627-645. https://doi.org/10.22067/ijnao.2023.80721.1214

Iranian Journal of Numerical Analysis and Optimization
Vol. 13, No. 4, 2023, pp 646–671
https://doi.org/10.22067/ijnao.2023.79648.1198
https://ijnao.um.ac.ir/

Research Article

Numerical nonlinear model solutions for the hepatitis C transmission between people and medical equipment using Jacobi wavelets method

N. Hamidat, S.M. Bahri and N. Abbassa*

Abstract

In this work, we present a new mathematical model for the spread of hepatitis C disease in two populations: human population and medical equipment population. Then, we apply the Jacobi wavelets method combined with the decoupling and quasi-linearization technique to solve this set of nonlinear differential equations for numerical simulation.

AMS subject classifications (2020): Primary 92C60; Secondary 65L10, 65T60.

Keywords: Hepatitis C; Sterilization; Jacobi wavelets; Operational matrix of derivative; Simulation.

*Corresponding author

Received 17 November 2022; revised 12 January 2023; accepted 25 January 2023

Nadjat Hamidat
Laboratory of pure and applied mathematics, Faculty of exact science and computer science, University of Abdelhamid Ibn Badis, Mostaganem -Algeria. e-mail: nadjat.hamidat.etu@univ-mosta.dz

Sidi Mohamed Bahri
Laboratory of pure and applied mathematics, Faculty of exact science and computer science, University of Abdelhamid Ibn Badis, Mostaganem -Algeria. e-mail: sidimohamed.bahri@univ-mosta.dz

Nadira Abbassa
Laboratory of pure and applied mathematics, Faculty of exact science and computer science, University of Abdelhamid Ibn Badis, Mostaganem -Algeria. e-mail: abbassanadira91@gmail.com

1 Introduction

Viral hepatitis is a major health problem worldwide, comparable to that posed by other major communicable diseases, such as human immunodeficiency virus (HIV), tuberculosis, malaria, or, more recently, coronavirus disease 2019 (COVID-19). In this work, we are interested in viral hepatitis C (HCV).
Hepatitis C is an inflammation of the liver caused by the hepatitis C virus. The virus can cause both acute and chronic hepatitis. According to the fact sheet of the World Health Organization (WHO) updated on October 2017 for hepatitis C, 71 million people have been estimated for chronic hepatitis C infection in the whole world, and approximately 399,000 people die each year from hepatitis C [28]. Until today, researchers could not develop a vaccine or effective treatment that heals hepatitis C at 100% [25].

The hepatitis C virus is transmitted by exposure to contaminated blood resulting from bringing the blood of an infected person into contact with that of a person likely to be contaminated directly (transfusion) or indirectly (equipment of contaminated injection for example). In 2016, WHO introduced global targets, for the care and management of HCV, a 90% reduction in new cases of chronic hepatitis C, a 65% reduction in hepatitis C deaths, and treatment of 80% of eligible people with chronic hepatitis C infections [30]. In Algeria, the president of the "SOS hepatitis association", spoke in an interview about the need to draw up a national plan against viral hepatitis, which will aim to improve prevention, care, and the availability of drugs. He also mentioned that the prevention of viral hepatitis poses a problem in Algeria because there is no real prevention against these viral infections, especially at the dentist. It is obvious to know that the majority of contaminations by these viruses are done during dental care. Therefore, raising awareness against viral hepatitis "B" and "C" is very important to detect these diseases, especially since they are silent. Indeed, better prevention requires better knowledge of the modes of transmission and the populations at risk in order to improve education and teach the appropriate protective measures. The last century has seen the emergence and rapid development of mathematical modeling, which plays an important role in assessing and anticipate the impact of Public Health programs.

Over the last decade, a large number of mathematical models have been developed to simulate, analyze, and understand the dynamics of a population of hepatitis C. In a related research work, Martcheva and Castillo-Chavez [19] proposed a model to study the role of a chronic infectious stage on the dynamics of HCV over the long term. Incorporating the immune class in [10], in [32], the latency period was merged. In [4], the authors showed both the effect of processing and immigration. Another model describes the effect of isolating chronically infected people [15]. Several studies have been carried out in [11, 20, 23, 21, 22, 33] showing the impact of HCV treatment in drug users on the prevalence of the disease. The optimal control theory has been

used to understand the efforts made to prevent the spread of the disease by different measures and strategies [1, 31, 34].

Our aim in this article is to understand how hepatitis C disease can evolve, by highlighting the role of sterilization of infected material, modeled by ordinary differential equation (ODE), unlike the model of Miller et al. [24], which targets the population of drug users. Therefore a single mode of contamination which plays the role of a vector of the disease and a single host. Our new model SIR-MI consists of taking into account other causes of contamination, such as dental equipment, toilet equipment, needles, tattooing, and piercing equipment in interaction with a mixed human population, and then we resolve this model.

On the other hand, wavelet theory plays an important role in many areas of mathematics and applied sciences, for instance, signal analysis in medicine, image processing, signal processing, data compression, statistics, and numerical methods [7, 18]. In recent years, wavelets based on orthogonal polynomials have been used in many researches to solve different problems such as ODE, partial differential equations, fractional differential equations, optimal control, and variational calculus [2, 9, 8, 26, 27], and this is due to orthogonality property. We propose the Jacobi wavelets method with general indices (α, β) in this work in order to obtain computational solutions. This method generalized other methods like Legendre wavelets and Chebyshev wavelets. The Jacobi wavelets method reduces an ODE to a system of algebraic equations by using the operational matrix of the derivative of Jacobi wavelets. In our numerical simulations, we have found that using the operational matrix of derivative simplifies the implementation of the method compared to using the operational matrix of integration [3]. Then, we apply the decoupling and quasi-linearization technique (DQLT) combined with the Jacobi wavelets method to solve the underlying problem.

In this paper, we propose, in section 2, a mathematical model SIR-MI that describes the dynamics of a population of hepatitis C. Section 3 will concern the mathematical analysis of the proposed model. Section 4 is devoted to explaining the different steps that lead to the implementation of the Jacobi wavelet method combined with DQLT. In Section 5, we apply the proposed method to simulate the model SIR-MI. Finally, Section 6 presents our conclusions.

2 Model formulation

In order to understand the effect of sterilization of the material on the transmission and dynamics of hepatitis C, we propose a mathematical model SIR-MI developed by Miller et al. [24] with five compartments. That is, let N_H be the total population of humans, which is subdivided into three subclasses: S_H (susceptible), I_H (infected), R_H (recovered). The total population of N_M material is divided into two subclasses: M_U (uninfected), M_I (infected).

The graphical representation of the proposed model is shown in Figure 1.

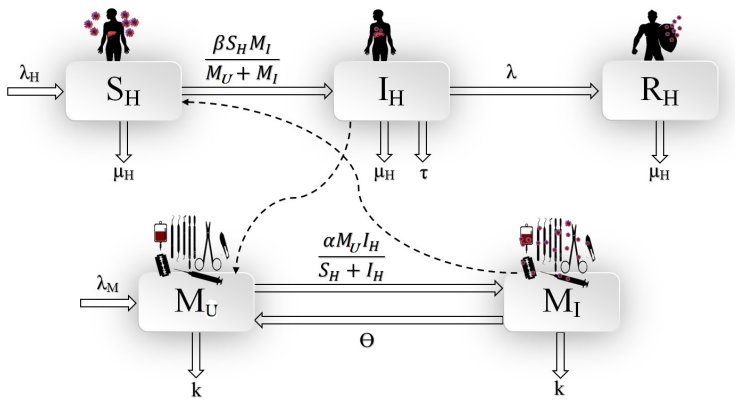


Figure 1: The compartmental model diagram.

The model SIR-MI is given by the following system of ODEs:

$$\begin{cases} \dfrac{dS_H}{dt}(t) = \lambda_H - \dfrac{\beta S_H(t) M_I(t)}{M_I(t) + M_U(t)} - \mu_H S_H(t), \\ \dfrac{dI_H}{dt}(t) = \dfrac{\beta S_H(t) M_I(t)}{M_I(t) + M_U(t)} - \lambda I_H(t) - (\tau + \mu_H) I_H(t), \\ \dfrac{dR_H}{dt}(t) = \lambda I_H(t) - \mu_H R_H(t), \\ \dfrac{dM_U}{dt}(t) = \lambda_M - \dfrac{\alpha M_U(t) I_H(t)}{S_H(t) + I_H(t)} + \theta M_I(t) - k M_U(t), \\ \dfrac{dM_I}{dt}(t) = \dfrac{\alpha M_U(t) I_H(t)}{S_H(t) + I_H(t)} - \theta M_I(t) - k M_I(t), \\ S_H(0) = S_{H_0}, I_H(0) = I_{H_0}, R_H(0) = R_{H_0}, M_U(0) = M_{S_0}, M_I(0) = M_{I_0}, \\ S_{H_0}, I_{H_0}, R_{H_0}, M_{U_0}, M_{I_0} > 0, \end{cases} \tag{1}$$

with

$$\lambda_M = k N_M \neq 0.$$

The parameters used in our model are defined in Table 1.

Table 1: Definitions of parameters used in model.

Parameters	Description
λ_H	birth rate of susceptible
μ_H	natural mortality rate of the human population
β	rate of interaction between susceptible humans and infected material
τ	mortality rate due to the disease
λ	disease cure rate
λ_M	birth rate of uninfected material
α	interaction rate between infected humans and uninfected material
k	rejection rate of infected or non-infected material
θ	sterilization rate of infected material

3 Mathematical Analysis of the Model

In this section and in the first moment, we will apply the theorem of Cauchy–Lipschitz to demonstrate the existence and the uniqueness of the solution of the system (1). Then we will study the behavior of this solution by going through the calculation of the points of equilibrium as well as the stability of these points. However, first, we note that the population of material $N_M(t)$ is constant.
Indeed,

$$N_M(t) = M_I(t) + M_U(t) \Longleftrightarrow \frac{dN_M}{dt}(t) = \frac{dM_U}{dt}(t) + \frac{dM_I}{dt}(t).$$

So

$$\begin{aligned} \frac{dM_U}{dt}(t) + \frac{dM_I}{dt}(t) &= \lambda_M - k\left(M_I(t) + M_U(t)\right), \\ \frac{dN_M}{dt}(t) &= 0. \end{aligned}$$

We show that the human population is not constant. So we have

$$N_H(t) = S_H(t) + I_H(t) + R_H(t) \Longleftrightarrow \frac{dN_H}{dt}(t) = \frac{dS_H}{dt}(t) + \frac{dI_H}{dt}(t) + \frac{dR_H}{dt}(t).$$

Then

$$\begin{aligned} \frac{dS_H}{dt}(t) + \frac{dI_H}{dt}(t) + \frac{dR_H}{dt}(t) &= \lambda_H - \mu_H\left(S_H(t) + I_H(t) + R_H(t)\right) - \tau I_H(t), \\ \frac{dN_H}{dt}(t) &= -\tau I_H(t). \end{aligned}$$

3.1 Existence and uniqueness of a positive solution

To study the existence and uniqueness of the solution of problem (1), we need to apply the Cauchy–Lipschitz theorem.
Our model (1) is a system of nonlinear, autonomous first-order differential equations that can be written as the following Cauchy problem:

$$\begin{cases} X'(t) = F(X(t)), t \in [0,T], \\ X(0) = X_0, \end{cases} \tag{2}$$

with

$$X(t) = \begin{pmatrix} S_H(t) \\ I_H(t) \\ R_H(t) \\ M_U(t) \\ M_I(t) \end{pmatrix} \text{ and } F(X(t)) = \begin{pmatrix} f_1(X(t)) \\ f_2(X(t)) \\ f_3(X(t)) \\ f_4(X(t)) \\ f_5(X(t)) \end{pmatrix},$$

where

$$f_1(X(t)) = \lambda_H - \frac{\beta S_H(t) M_I(t)}{N_M} - \mu_H S_H(t), \tag{3}$$

$$f_2(X(t)) = \frac{\beta S_H(t) M_I(t)}{N_M} - \lambda I_H(t) - (\tau + \mu_H) I_H(t), \tag{4}$$

$$f_3(X(t)) = \lambda I_H(t) - \mu_H R_H(t), \tag{5}$$

$$f_4(X(t)) = \lambda_M - \frac{\alpha M_U(t) I_H(t)}{S_H(t) + I_H(t)} + \theta M_I(t) - k(t), \tag{6}$$

$$f_5(X(t)) = \frac{\alpha M_U(t) I_H(t)}{S_H(t) + I_H(t)} - \theta M_I(t) - k M_I(t). \tag{7}$$

We recall that the norm $Norm(\cdot)$ in the space of continuous functions from I to $\mathbb{R}^5$ (denoted by $C\left(I, \mathbb{R}^5\right)$) is defined by

$$Norm(F) = \max_{t \in I} \|F(t)\|_2,$$

with $\|\cdot\|_2$ is the usual Euclidean norm in $\mathbb{R}^5$.
We are now able to state the following result.

Theorem 1. The differential problem (1) admits a unique solution $(S_H(t), I_H(t), R_H(t), M_U(t), M_I(t))^T \in \mathbb{R}^5$ for all $t \in [0,T]$.

Proof. To demonstrate that the Cauchy problem (1) admits a unique solution, it suffices to show that the vector function F of the equivalent problem (2), is Liptschizian.
Let $t \in [0,T]$, $X_1, X_2 \in \mathbb{R}^5$. Then

$$\|F(X_1(t)) - F(X_2(t))\| = \max \begin{cases} |f_1(X_1(t)) - f_1(X_2(t))|, \\ |f_2(X_1(t)) - f_2(X_2(t))|, \\ |f_3(X_1(t)) - f_3(X_2(t))|, \\ |f_4(X_1(t)) - f_4(X_2(t))|, \\ |f_5(X_1(t)) - f_5(X_2(t))|. \end{cases}$$

We assume that at any instant $t \in [0, T]$, the human population $N_H(t) = S_H(t) + I_H(t)$ is between two real numbers strictly positive $N_{\min}$ and $N_{\max}$.

We will examine each of the components $|f_i(X_1(t)) - f_i(X_2(t))|$, $i = 1, \ldots, 5$. Therefore

$$\begin{aligned} &|f_1(X_1(t)) - f_1(X_2(t))| \\ &= \left| -\frac{\beta S_{H_1}(t) M_{I_1}(t)}{N_M(t)} - \mu_H S_{H_1}(t) + \frac{\beta S_{H_2}(t) M_{I_2}(t)}{N_M(t)} + \mu_H S_{H_2}(t) \right| \\ &\leq \frac{\beta}{N_M(t)} |-S_{H_1}(t) M_{I_1}(t) + S_{H_2}(t) M_{I_2}(t)| + \mu_H |-S_{H_1}(t) + S_{H_2}(t)|. \end{aligned}$$

By adding and subtracting the term $S_{H_1} M_{I_2}$, we have

$$\begin{aligned} &|f_1(X_1(t)) - f_1(X_2(t))| \\ &\quad \leq \frac{\beta}{N_M(t)} S_{H_1}(t) |-M_{I_1}(t) + M_{I_2}(t)| + \beta \frac{M_{I_2}(t)}{N_M(t)} |-S_{H_1}(t) + S_{H_2}(t)| \\ &\quad\quad + \mu_H |-S_{H_1}(t) + S_{H_2}(t)|. \end{aligned}$$

Since $S_{H_1} \leq N_{\max}$ and $\dfrac{M_{I_2}}{N_M} \leq 1$, then

$$|f_1(X_1(t)) - f_1(X_2(t))| \leq \left(\frac{\beta}{N_M(t)} N_{\max} + \beta + \mu_H \right) \|X_1(t) - X_2(t)\|.$$

For (4) and following the same reasoning, we find

$$\begin{aligned} &|f_2(X_1(t)) - f_2(X_2(t))| \\ &\leq \left(\frac{\beta}{N_M(t)} N_{\max} + \beta + \lambda + \tau + \mu_H \right) \|X_1(t) - X_2(t)\|. \end{aligned}$$

The linearity of terms in (5) leads to

$$|f_3(X_1(t)) - f_3(X_2(t))| \leq (\lambda + \mu_H) \|X_1(t) - X_2(t)\|.$$

From (6), we have

$$\begin{aligned}|f_4(X_1(t)) - f_4(X_2(t))| \leq & \frac{\alpha}{N_H(t)} |-M_{U_1}(t) I_{H_1}(t) + M_{U_2}(t) I_{H_2}(t)| \\ & +\theta |M_{I_1}(t) - M_{I_2}(t)| + k |-M_{U_1}(t) + M_{U_2}(t)|.\end{aligned}$$

By adding and subtracting the term $M_{U_1} I_{H_2}$, we have

$$\begin{aligned}|f_4(X_1(t)) - f_4(X_2(t))| \leq & \frac{\alpha}{N_H(t)} M_{U_1}(t) |-I_{H_1}(t) + I_{H_2}(t)| \\ & +\frac{\alpha}{N_H(t)} I_{H_2}(t) |-M_{U_1}(t) + M_{U_2}(t)| \\ & +\theta |M_{I_1}(t) - M_{I_2}(t)| + k |-M_{U_1}(t) + M_{U_2}(t)|.\end{aligned}$$

Knowing $M_{U_1} \leq N_M$, $N_H(t) \geq N_{\min}$ and $\dfrac{I_{H_2}}{N_H(t)} \leq 1$, we arrive at

$$|f_4(X_1(t)) - f_4(X_2(t))| \leq \left(\frac{\alpha N_M(t)}{N_{\min}} + \alpha + \theta + k \right) \|X_1(t) - X_2(t)\|.$$

Finally, from (7) and following the previous steps, we have

$$|f_5(X_1(t)) - f_5(X_2(t))| \leq \left(\frac{\alpha N_M(t)}{N_{\min}} + \alpha + \theta + k \right) \|X_1(t) - X_2(t)\|.$$

Therefore, we have

$$\|F(X_1(t)) - F(X_2(t))\| \leq C \|X_1(t) - X_2(t)\|,$$

with

$$\begin{aligned}C = \max \Bigg(& \frac{\beta}{N_M(t)} N_{\max} + \beta + \mu_H, \ \frac{\beta}{N_M(t)} N_{\max} + \beta + \lambda + \tau + \mu_H, \ \lambda + \mu_H, \\ & \frac{\alpha N_M(t)}{N_{\min}} + \alpha + \theta + k \Bigg).\end{aligned}$$

□

3.2 Equilibrium points

In this subsection, we will look for points of equilibrium E_0 and E_1 (Theorem 2) and study their stabilities. We limit ourselves to the stability of the point E_0. The stability of the point E_1 will be made numerically.

The basic reproduction rate R_0

Understanding how an epidemic develops once it has appeared is crucial if we are to hope to control it. To do this, various models have been developed, which highlight the crucial role played by the R_0 parameter, describing the average number of new infections due to a sick individual. As one can imagine, if this number is less than 1, then the epidemic will tend to die out, while it may persist or even spread to the entire population if $R_0 > 1$ ([12]).

We recall, for a given matrix A, that $Sp(A)$ represents the spectrum of A and that the spectral radius of the matrix A, denoted $\rho(A)$, is defined by

$$\rho(A) = \max\left\{|\lambda|, \lambda \in Sp(A)\right\}.$$

The disease-free point is

$$(S_H = \frac{\lambda_H}{\mu_H}, I_H = 0, R_H = 0, M_U = N_M, M_I = 0).$$

We consider different infected populations of the model. That is,

$$\frac{dI_H}{dt}(t) = \frac{\beta S_H(t) M_I(t)}{N_M} - \lambda I_H(t) - (\tau + \mu_H) I_H(t)$$

and

$$\frac{dM_I}{dt}(t) = \frac{\alpha M_U(t) I_H(t)}{S_H(t) + I_H(t)} - \theta M_I(t) - k M_I(t).$$

To be able to calculate R_0, we use two matrices F and V, where the matrix F represents the appearance of new infected; that is, what comes from other compartments and which enters the infected compartment following an infection,

$$F(I_H, M_I) = \begin{pmatrix} 0 & \frac{\beta S_H}{N_M} \\ \frac{\alpha M_U}{N_H} & 0 \end{pmatrix}.$$

The matrix V represents all those who leave the compartments of the infected and those who come there for any other reason,

$$V(I_H, M_I) = \begin{pmatrix} -\lambda - (\tau + \mu_H) & 0 \\ 0 & -\theta - k \end{pmatrix}.$$

We have

$$-FV^{-1} = \begin{pmatrix} 0 & \dfrac{\beta\lambda_H}{N_M\,(k+\theta)\,\mu_H} \\ \dfrac{\alpha N_M \mu_H}{(\lambda+\tau+\mu_H)\,\lambda_H} & 0 \end{pmatrix}.$$

The matrix $-FV^{-1}$ represents the next generation matrix. The basic reproduction rate is given by

$$R_0 = \rho(-FV^{-1}).$$

After calculating the eigenvalues of the matrix $-FV^{-1}$, we find

$$\lambda_1 = \sqrt{\tfrac{\beta\alpha}{(\lambda+\tau+\mu_H)(k+\theta)}} \text{ and } \lambda_2 = -\sqrt{\tfrac{\beta\alpha}{(\lambda+\tau+\mu_H)(k+\theta)}}.$$

We then conclude

$$R_0 = \sqrt{\frac{\beta\alpha}{(\lambda+\tau+\mu_H)\,(k+\theta)}}. \tag{8}$$

The calculation of equilibrium points

Theorem 2. The system (1) admits two equilibrium points E_0 and E_1, for strictly positive parameters. They are given indeed as follows.

1. If $R_0 < 1$, then the point E_0 exists and it is given by

$$E_0 = \left(\frac{\lambda_H}{\mu_H}, 0, 0, N_M, 0\right).$$

2. If

$$R_0 > 1 \quad \text{and} \quad \alpha\beta + \alpha\mu_H > (k+\theta)\,(\tau+\lambda)\,,$$

then the endemic point E_1 exists and it is given by

$$E_1 = (S_H^*, I_H^*, R_H^*, N_M - M_I^*, M_I^*),$$

with

$$\begin{cases} S_H^* = \dfrac{\lambda_H - (\tau+\mu_H+\lambda)\,I_H^*}{\mu_H}, \\ I_H^* = \dfrac{\beta\alpha\lambda_H - \lambda_H\,(k+\theta)\,(\lambda+\tau+\mu_H)\,N_M}{\alpha\beta\,(\tau+\mu_H+\lambda) + N_M\,(\lambda+\tau+\mu_H)\,(\alpha\mu_H - (k+\theta)\,(\tau+\lambda))}, \\ R_H^* = \dfrac{\lambda}{\mu_H} I_H^*, \\ M_U^* = N_M - M_I^*, \\ M_I^* = \dfrac{\alpha\mu_H N_M I_H^*}{((\alpha\mu_H - (k+\theta)\,(\tau+\lambda))\,I_H^* + \lambda_H\,(k+\theta))}. \end{cases}$$

Proof. The equilibriums of the system (1) are given by the solutions of the following system of algebraic equations:

$$\begin{cases} \lambda_H - \dfrac{\beta S_H^* M_I^*}{N_M} - \mu_H S_H^* = 0, \\ \dfrac{\beta S_H^* M_I^*}{N_M} - \lambda I_H^* - (\tau + \mu_H) I_H^* = 0, \\ \lambda I_H^* - \mu_H R_H^* = 0, \\ \lambda_M - \dfrac{\alpha M_U^* I_H^*}{S_H^* + I_H^*} + \theta M_I^* - k M_U^* = 0, \\ \dfrac{\alpha}{S_H + I_H} M_U^* I_H^* - (k + \theta) M_I^* = 0. \end{cases} \tag{9}$$

As the population of the material is constant $N_M = M_U + M_I$, then the system (9) is reduced to the following four equations:

$$\lambda_H - \frac{\beta S_H^* M_I^*}{N_M} - \mu_H S_H^* = 0, \tag{10}$$

$$\frac{\beta S_H^* M_I^*}{N_M} - \lambda I_H^* - (\tau + \mu_H) I_H^* = 0, \tag{11}$$

$$\lambda I_H^* - \mu_H R_H^* = 0, \tag{12}$$

and

$$\frac{\alpha (N_M - M_I^*) I_H^*}{S_H^* + I_H^*} - (k + \theta) M_I^* = 0. \tag{13}$$

The sum of (10) and (11) gives

$$S_H^* = \frac{\lambda_H - (\tau + \mu_H + \lambda) I_H^*}{\mu_H}. \tag{14}$$

From (12), it is clear that

$$R_H^* = \frac{\lambda}{\mu_H} I_H^*. \tag{15}$$

To determine M_I^*, we replace (14) in (13) and obtain

$$M_I^* = \frac{\alpha \mu_H N_M I_H^*}{((\alpha \mu_H - (k + \theta)(\tau + \lambda)) I_H^* + \lambda_H (k + \theta))}. \tag{16}$$

Substituting (14) and (16) into (11), we find

$$\left(\frac{\beta \alpha (\lambda_H - (\tau + \mu_H + \lambda) I_H^*)}{((\alpha \mu_H - (k + \theta)(\tau + \lambda)) I_H^* + \lambda_H (k + \theta))} - (\lambda + \tau + \mu_H) N_M \right) I_H^* = 0. \tag{17}$$

According to (17), we distinguish two cases:

Case I $I_H^* = 0$, we then find

$$S_H = \frac{\lambda_H}{\mu_H}, R_H = M_I = 0 \text{ et } M_U = N_M.$$

Hence the existence of the first equilibrium point E_0 is as follows:

$$E_0 = \left(\frac{\lambda_H}{\mu_H}, 0, 0, N_M, 0\right).$$

Case II $I_H^* \neq 0$, we have

$$\frac{\beta\alpha\left(\lambda_H - (\tau + \mu_H + \lambda) I_H^*\right)}{\left((\alpha\mu_H - (k+\theta)(\tau+\lambda)) I_H^* + \lambda_H (k+\theta)\right)} - (\lambda + \tau + \mu_H) N_M = 0.$$

We can easily write the last equation in the form

$$BI_H^{*2} + AI_H^* = 0$$

with

$$\begin{cases} A = \beta\alpha(\tau + \mu_H + \lambda) + (\lambda + \tau + \mu_H) N_M (\alpha\mu_H - (k+\theta)(\tau+\lambda)), \\ B = -\lambda_H (k+\theta)(\lambda + \tau + \mu_H) N_M + \beta\alpha\lambda_H. \end{cases}$$

Hence,

$$I_H^* = \frac{\beta\alpha\lambda_H - \lambda_H (k+\theta)(\lambda + \tau + \mu_H) N_M}{\alpha\beta(\tau + \mu_H + \lambda) + (\lambda + \tau + \mu_H) N_M (\alpha\mu_H - (k+\theta)(\tau+\lambda))}. \tag{18}$$

Then, the existence of the endemic point E_1 is given by

$$E_1 = (S_H^*, I_H^*, R_H^*, N_M - M_I^*, M_I^*),$$

with

$$\begin{cases} S_H^* = \dfrac{\lambda_H - (\tau + \mu_H + \lambda) I_H^*}{\mu_H}, \\ I_H^* = \dfrac{\beta\alpha\lambda_H - \lambda_H (k+\theta)(\lambda + \tau + \mu_H) N_M}{\alpha\beta(\tau + \mu_H + \lambda) + (\lambda + \tau + \mu_H) N_M (\alpha\mu_H - (k+\theta)(\tau+\lambda))}, \\ R_H^* = \dfrac{\lambda}{\mu_H} I_H^*, \\ M_U^* = N_M - M_I^*, \\ M_I^* = \dfrac{\alpha\mu_H N_M I_H^*}{\left((\alpha\mu_H - (k+\theta)(\tau+\lambda)) I_H^* + \lambda_H (k+\theta)\right)}. \end{cases}$$

□

The positivity of the equilibrium points

It is clear that the point E_0 is positive (belong to the positive orthant) without condition. It then remains to show that the point E_1 is positive. This amounts to showing that $S_H^*, I_H^*, R_H^*, M_I^*$ are positive.

1. The positivity of S_H^* is according to (14).

   It is clear that the denominator of S_H^* is positive, so it suffices to study the positivity of the numerator

$$\begin{aligned}\lambda_H - (\tau + \mu_H + \lambda) I_H^* &= \frac{\lambda_H \alpha \mu_H + \lambda_H (k+\theta) \mu_H}{\alpha\beta + \alpha\mu_H - (k+\theta)(\tau+\lambda)} \\ &= \frac{S_1}{S_2}.\end{aligned}$$

   As the numerator S_1 is positive, then it remains to show that S_2 is positive; that is,

$$\alpha\beta + \alpha\mu_H > (k+\theta)(\tau+\lambda).$$

2. The positivity of I_H^* is according to (18). We let

$$I_H^* = \frac{A}{B}.$$

   We write A as a function of R_0 defined in (8),

$$A = \alpha\lambda_H\beta\frac{R_0^2 - 1}{R_0^2}.$$

   Therefore A is positive if $R_0 > 1$.

   We write B as a function of R_0,

$$B \geq \alpha\beta(\lambda+\tau)\frac{R_0^2 - 1}{R_0^2}.$$

   We note that B is positive if $R_0 > 1$.

3. The positivity of R_H^* is according to (15). We note that R_H^* is positive if I_H^* is positive.

4. The positivity of M_I^* is according to (16). We let

$$M_I^* = \frac{M_1}{M_2}.$$

It is clear that M_1 is positive if I_H^* is positive. Then, M_I^* is positive if M_2 is positive. Indeed

$$M_2 > 0 \Longrightarrow \frac{C_1}{C_2} > 0,$$

with

$$\begin{cases} C_1 = \alpha\mu_H\left(\lambda+\tau+\mu_H\right)\left(\alpha\beta+\alpha\mu_H-(k+\theta)\left(\tau+\lambda\right)\right) \\ \qquad +\lambda_H\left(k+\theta\right)\left(\alpha\beta\mu_H+\left(\tau+\mu_H+\lambda\right)\alpha\mu_H\right) > 0, \\ C_2 = \alpha\beta\left(\tau+\mu_H+\lambda\right)+\left(\lambda+\tau+\mu_H\right)\left(\alpha\mu_H-(k+\theta)\left(\tau+\lambda\right)\right) > 0. \end{cases}$$

We then deduce that M_2 is positive if

$$\alpha\beta+\alpha\mu_H > (k+\theta)\left(\tau+\lambda\right),$$

and therefore M_I^* is positive if I_H^* is positive and if

$$\alpha\beta+\alpha\mu_H > (k+\theta)\left(\tau+\lambda\right).$$

3.3 Stability

The stability of the equilibrium point [5] results from the stability of the Jacobian matrix of the system (10), (11), (12), (13) (i.e., its eigenvalues must be negative), which is given by

$$J(S_H, I_H, R_H, M_I) = \begin{pmatrix} -\dfrac{\beta M_I}{N_M}-\mu_H & 0 & 0 & -\dfrac{\beta S_H}{N_M} \\ \dfrac{\beta M_I}{N_M} & -\lambda-(\tau+\mu_H) & 0 & \dfrac{\beta S_H}{N_M} \\ 0 & \lambda & -\mu_H & 0 \\ \dfrac{-\alpha(N_M-M_I)I_H}{(S_H+I_H)^2} & \dfrac{\alpha(N_M-M_I)S_H}{(S_H+I_H)^2} & 0 & \dfrac{-\alpha I_H}{S_H+I_H}-(k+\theta) \end{pmatrix}$$

Theorem 3. It holds that E_0 is locally asymptotically stable (the solutions must approach an equilibrium point under initial conditions close to the equilibrium point) if and only if

$$R_0 < 1.$$

Proof. The Jacobian matrix at point E_0 is given by

$$J(E_0) = \begin{pmatrix} -\mu_H & 0 & 0 & -\dfrac{\beta\lambda_H}{N_M\mu_H} \\ 0 & -(\lambda+\tau+\mu_H) & 0 & \dfrac{\beta\lambda_H}{N_M\mu_H} \\ 0 & \lambda & -\mu_H & 0 \\ 0 & \dfrac{\alpha N_M\mu_H}{\lambda_H} & 0 & -(k+\theta) \end{pmatrix}$$
$$= \begin{pmatrix} -A & 0 & 0 & -B \\ 0 & -C & 0 & B \\ 0 & D & -A & 0 \\ 0 & E & 0 & -F \end{pmatrix}.$$

We calculate the characteristic polynomial of $J(E_0)$,

$$\begin{aligned} \det(J(E_0) - XI_3) &= -(X+\mu_H)^2 (k\lambda + k\tau + k\mu_H + \theta\lambda \\ &\quad + \theta\tau - \alpha\beta + \theta\mu_H + (\theta+\lambda+\tau+\mu_H+k)X + X^2) \\ &= -(X+\mu_H)^2 P(X). \end{aligned}$$

We have the first eigenvalues

$$X_1 = X_2 = -\mu_H < 0,$$

and

$$P(X) = A + BX + CX^2,$$

with

$$\begin{aligned} A &= k\lambda + k\tau + k\mu_H + \theta\lambda + \theta\tau + \theta\mu_H - \alpha\beta, \\ B &= (\theta+\lambda+\tau+\mu_H+k), \\ C &= 1. \end{aligned}$$

Let us use Descartes' rule [16] to show that the coefficients of the polynomial P do not change signs.

It is clear that B and C are positive. It only remains to show that A is positive or equivalently

$$1 - R_0 > 0.$$

So, A is positive if $R_0 < 1$.

According to Descartes' rule, the polynomial does not admit any positive root. Hence, the stability of the point E_0. □

4 Method of resolution

4.1 Jacobi wavelets

The Jacobi polynomials $J_m^{(\alpha,\beta)}$ $(\alpha > -1, \beta > -1)$ are orthogonal polynomials on the interval $[-1,1]$ ([13, 29]) with the weight function

$$\omega(x) = (1-x)^{\alpha}(1+x)^{\beta}, \tag{19}$$

where m is a positive integer, which represents the degree of the polynomial. These polynomials belong to the weight space $L^2_\omega([-1,1])$. The Jacobi polynomials can be represented by the recursive formula given by

$$\begin{aligned} &J_m^{(\alpha,\beta)}(x) \\ &= \frac{(\alpha+\beta+2m-1)\left[\alpha^2-\beta^2+x(\alpha+\beta+2m)(\alpha+\beta+2m-2)\right]}{2m(\alpha+\beta+2m-2)(\alpha+\beta+m)} J_{m-1}^{(\alpha,\beta)}(x) \\ &\quad -\frac{(\alpha+m-1)(\beta+m-1)(\alpha+\beta+2m)}{m(\alpha+\beta+2m-2)(\alpha+\beta+m)} J_{m-2}^{(\alpha,\beta)}(x), \end{aligned} \tag{20}$$

where

$$J_0^{(\alpha,\beta)}(x) = 1,\ J_1^{(\alpha,\beta)}(x) = \frac{\alpha+\beta+2}{2}x + \frac{\alpha-\beta}{2}. \tag{21}$$

As the Jacobi polynomials are orthogonal with respect to the weight function ω, then

$$\left\langle J_n^{(\alpha,\beta)}, J_m^{(\alpha,\beta)} \right\rangle_{L^2_\omega} = h_m^{(\alpha,\beta)}\delta_{n,m}, \qquad \text{for all } n,m \in \mathbb{N}, \tag{22}$$

where

$$h_m^{(\alpha,\beta)} = \left\| J_m^{(\alpha,\beta)} \right\|^2 = \frac{2^{\alpha+\beta+1}\Gamma(\alpha+m+1)\Gamma(\beta+m+1)}{(2m+1+\alpha+\beta)\,m!\,\Gamma(\alpha+\beta+m+1)}, \tag{23}$$

$\delta_{n,m}$ represents the Kronecker symbol, Γ is the Euler gamma function, and $\langle\cdot,\cdot\rangle_{L^2_\omega}$ denotes the inner product of $L^2_\omega([-1,1])$.

The Jacobi wavelets are defined by

$$\psi_{n,m}^{(\alpha,\beta)}(x) = \begin{cases} \frac{2^{\frac{k+1}{2}}}{\sqrt{h_m^{(\alpha,\beta)}}} J_m^{(\alpha,\beta)}\left(2^{k+1}x-2n+1\right), & \frac{n-1}{2^k} \le x < \frac{n}{2^k} \\ 0, & \text{otherwise}, \end{cases} \tag{24}$$

where $k \in \mathbb{N}$, $n = 1,\ldots,2^k$ represents the number of decomposition levels, $m = 0,1,\ldots,M$ is the degree of the Jacobi polynomials ($M \in \mathbb{N}^*$). The coefficient $\frac{2^{\frac{k+1}{2}}}{\sqrt{h_m^{(\alpha,\beta)}}}$ is for normality.

4.2 Decomposition in Jacobi wavelets basis

Since the Jacobi wavelets family $\left\{\psi_{n,m}^{(\alpha,\beta)}\right\}_{\substack{n=1,\ldots,2^k \\ m\geq 0}}$ forms an orthonormal basis in $L_\omega^2([0,1])$, we can express all functions f in $L_\omega^2([0,1])$ as a unique linear combination of elements of this basis:

$$f(x)=\sum_{n=1}^{2^k}\sum_{m=0}^{\infty}c_{n,m}\psi_{n,m}^{(\alpha,\beta)}(x)\,, \tag{25}$$

where $c_{n,m}=\left\langle f,\psi_{n,m}^{(\alpha,\beta)}\right\rangle_{L_\omega^2([0,1])}$. From the point of view of the numerical analysis, we take the truncated sum (its projection on finite space)

$$f(x)=\sum_{n=1}^{2^k}\sum_{m=0}^{M}c_{n,m}\psi_{n,m}^{(\alpha,\beta)}(x). \tag{26}$$

Let

$$C=\left[c_{1,0},\ldots,c_{1,M},c_{2,0},\ldots,c_{2,M},\ldots,c_{2^k,0},\ldots,c_{2^k,M}\right]^T,$$

and let

$$\Psi^{(\alpha,\beta)}=\left[\psi_{1,0}^{(\alpha,\beta)},\ldots,\psi_{1,M}^{(\alpha,\beta)},\psi_{2,0}^{(\alpha,\beta)},\ldots,\psi_{2,M}^{(\alpha,\beta)},\ldots,\psi_{2^k,0}^{(\alpha,\beta)},\ldots,\psi_{2^k,M}^{(\alpha,\beta)}\right]^T. \tag{27}$$

We can find the following matrix notation:

$$f(x)=C^T\Psi^{(\alpha,\beta)}(x)\,. \tag{28}$$

In this case, the $\Psi^{(\alpha,\beta)}$ are called the $2^k(M+1)$ Jacobi wavelets vector and C is a $2^k(M+1)$ vector.

The operational matrix of derivative

The derivative of the Jacobi wavelets vector $\Psi^{(\alpha,\beta)}$ from (27) can be expressed by [17]

$$\frac{d\Psi^{(\alpha,\beta)}(x)}{dx}=D^{(\alpha,\beta)}\Psi^{(\alpha,\beta)}(x),$$

where $D^{(\alpha,\beta)}$ denotes the $2^k(M+1)\times 2^k(M+1)$ operational matrix given by

$$D^{(\alpha,\beta)} = \begin{pmatrix} F^{(\alpha,\beta)} & 0 & \dots & 0 \\ 0 & F^{(\alpha,\beta)} & \ddots & \vdots \\ \vdots & \ddots & \ddots & 0 \\ 0 & \dots & 0 & F^{(\alpha,\beta)} \end{pmatrix},$$

$F^{(\alpha,\beta)}$ is $(M+1)\times(M+1)$ matrix, where its (i,j)th element is given by

$$F_{i,j}^{(\alpha,\beta)} = \begin{cases} 2^{k+1}\frac{\sqrt{h_{j-1}^{(\alpha,\beta)}}}{\sqrt{h_{i-1}^{(\alpha,\beta)}}}\gamma_{i-1,j-1}^{(\alpha,\beta)}, & \text{if } i > j \\ 0, & \text{otherwise,} \end{cases} \tag{29}$$

in which $h_{i-1}^{(\alpha,\beta)}$ and $h_{j-1}^{(\alpha,\beta)}$ are defined from (23), and $\gamma_{i-1,j-1}^{(\alpha,\beta)}$ are given by

$$\gamma_{i-1,j-1}^{(\alpha,\beta)} = \frac{\Gamma(i+\beta)}{2\Gamma(i+\alpha+\beta)}\frac{(2(j-1)+\alpha+\beta+1)\Gamma(\alpha+\beta+j)}{\Gamma(\alpha+j)} \times \left[\sum_{d=j-1}^{i-1}(-1)^{d-j-1}\frac{(2(d+1)+\alpha+\beta)\Gamma(\alpha+d+1)}{\Gamma(\beta+d+2)}\right]. \tag{30}$$

4.3 Description of the solution method

In this subsection, we describe how to apply the Jacobi wavelets to solve ODEs. Then, we use the DQLT with Jacobi wavelets method to solve a set of nonlinear differential equations. In the end, we present the formula of errors calculation.

Linear first order differential equation

Consider the linear first order differential equation with initial condition

$$\begin{cases} f'(x) + a(x)f(x) = g(x), & x \in]0,1], \\ f(0) = f_0, \end{cases} \tag{31}$$

where f_0 is arbitrary constant. To solve the problem (31), we decompose $f(x)$ in the Jacobi wavelets basis $\left\{\psi_{n,m}^{(\alpha,\beta)}\right\}_{\substack{n=1,\dots,2^k \\ m=0,\dots,M}}$ by estimating (28),

$$f(x) = C^T\Psi^{(\alpha,\beta)}(x), \tag{32}$$

where C denotes the solution vector of the problem. Then, we have

$$f'(x) = C^T D^{(\alpha,\beta)}\Psi^{(\alpha,\beta)}(x). \tag{33}$$

Now, by substituting (32)–(33) into problem (29), we get the following algebraic system:

$$C^T(D^{(\alpha,\beta)} + a(x_i)I)\Psi^{(\alpha,\beta)}(x_i) = g(x_i), \quad i = 1, \ldots, nc, \tag{34}$$

where I is the identity matrix and nc is the number of collocation points. We have to insert the initial condition

$$f_0 = C^T\Psi^{(\alpha,\beta)}(0). \tag{35}$$

Equations (34) and (35) generate $2^k(M+1)$ set of linear algebraic equations, which can easily be solved for the unknown C by using one of the method of resolution an algebraic system. Consequently, $f(x)$ given in (32) will be easily calculated.

Set of nonlinear differential equation

To solve a set of nonlinear differential equations, we will use the DQLT to transform this problem by iterative steps into a set of decoupled and linearized differential equations, where each equation can be written as the problem (31). Then we use the Jacobi wavelets method described in the previous subsection. Let us consider a set of p nonlinear differential equations. This iterative technique can be defined by

$$\begin{cases} \text{Given initial profile} f_1^{(0)}, f_2^{(0)}, \ldots, f_p^{(0)}, \\ (f_1'(x))^{(l+1)} + a_1(x)f_1^{(l+1)} = g_1\left(x, f_1^{(l)}, f_2^{(l)}, \ldots, f_p^{(l)}\right), \\ (f_2'(x))^{(l+1)} + a_2(x)f_2^{(l+1)} = g_2\left(x, f_1^{(l+1)}, f_2^{(l)}, \ldots, f_p^{(l)}\right), \\ \vdots \\ \left(f_p'(x)\right)^{(l+1)} + a_p(x)f_p^{(l+1)} = g_p\left(x, f_1^{(l+1)}, f_2^{(l+1)}, \ldots, f_p^{(l)}\right), \end{cases} \tag{36}$$

where $f_i^{(l+1)}$ and $f_i^{(l)}$ are the approximations of the solution f_i at the current and the precedent iteration, respectively. At each iteration, we apply the Jacobi wavelets method to solve p linear differential equation. Then, for $(l+1)$th iteration, we can calculate the decoupling error using the following formula:

$$E_{DQLT} = \max\left(\|f_1^l - f_1^{l+1}\|_2, \|f_2^l - f_2^{l+1}\|_2, \ldots, \|f_p^l - f_p^{l+1}\|_2\right). \tag{37}$$

The procedure is terminated when the error of decoupling is sufficiently small.

Error estimation

Since the ODEs solutions are only known at collocation points, the most appropriate norm is the euclidean norm if the exact solution is given. The accuracy of the proposed method is estimated by

$$error = \|f(x) - f_{ex}(x)\|_2 = \sqrt{\sum_{i=1}^{nc} |f(x_i) - f_{ex}(x_i)|^2}, \tag{38}$$

where f_{ex} is the analytic solution, f is the approximate solution, and nc the number of collocation points.

5 Numerical simulations of model SIR-MI

In this section, we will study the stability of the point E_1 numerically. Then, we simulate our model to see the importance of studying the effect of the sterilization parameter infected material and management on the evolution of the human population. We apply the Jacobi wavelets with DQLT, which makes it possible to numerically evaluate the solutions of the ODEs and to build their graphs. We conclude our section with a discussion of the results obtained.

5.1 The study of the stability of the second equilibrium point E_1

The following table gives us the biological parameters that verify the conditions of existence and stability of the second point of equilibrium E_1:

Table 2: The parameters verifying the stability of E_1.

Equilibrium point	$E_1 = (1622, 9920, 1668, 18172, 11828)$								
Parameter	λ_H	β	μ_H	λ	λ_M	τ	α	k	θ
Value	230	0.23	0.05116	0.086	1500	0.011	0.6	0.05	0.3

1. For the conditions of existence, we have

$$(R_0 = 1.6313) > 1, \tag{39}$$

and

$$(\alpha\beta + \alpha\mu_H = 0.1687) > ((k+\theta)(\tau+\lambda) = 0.0339). \tag{40}$$

Hence, we have the existence of the point E_1.

2. For stability, the Jacobian matrix of system (34) at point E_1 after substitution of the parameters given in Table 2 is given by

$$J\left(E_1\right)=\begin{pmatrix} -0.1418 & 0 & 0 & -0.0124 \\ 0.0907 & -0.1482 & 0 & 0.0124 \\ 0 & 0.0860 & -0.0512 & 0 \\ -1.5836 & 2.5875 & 0 & -0.5778 \end{pmatrix}.$$

The eigenvalues of $J(E_1)$ are

$$\begin{cases} vp_1=-0.0512, \\ vp_2=-0.6842, \\ vp_3=-0.0918+0.0515i, \\ vp_4=-0.0918-0.0515i. \end{cases}$$

The eigenvalues of $J(E_1)$ have a negative real part, hence, the asymptotic stability of the second equilibrium point E_1.

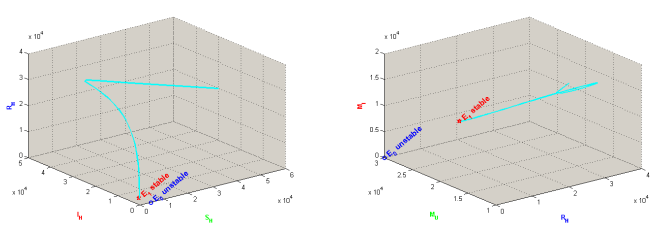


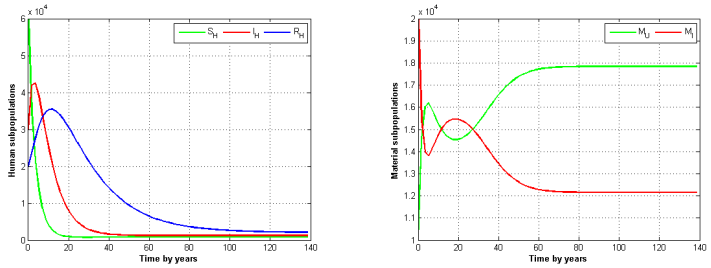
Figure 2: The convergence of the system toward E_1.

Figure 2: The convergence of the system toward E_1.

We note that the solutions obtained in Figure 2. All converge towards the equilibrium point E_1 when $t \to +\infty$.

Figure 3 shows that the five subpopulations converge after a fairly large time to the second equilibrium point E_1.

In what follows, we carried out simulation experiments with the parameters illustrated by Table 3.



(A) (B)

Figure 3: Evolution of human and material subpopulations: (A) represents the human subpopulations, (B) represents the material subpopulations.

Table 3: Variations in estimated values of biological data.

The time	40 years								
I.C	S_H	I_H		R_H		M_U		M_I	
Value	60000	30000		20000		10000		20000	
Parameter	λ_H	β	μ_H	λ	λ_M	τ	α	k	θ
Value	230	0.072-0.6	0.01-0.05116	0.006-0.235	1500	0.011	0.1-0.6	0.05	0.17-0.4

5.2 The impact of equipment sterilization on disease progression

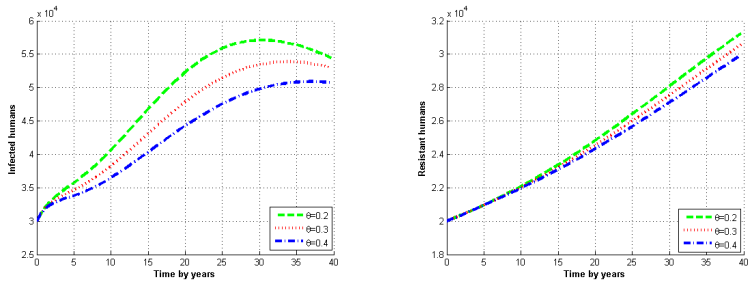


Figure 4: Evolution of human subpopulations for different values of θ.

For different values of $\theta = 0.2, 0.3, 0.4$, we see, in Figure 4, the positive effect played by the sterilization parameter to reduce the number of infections. This shows that better compliance with universal hygiene rules and recommendations for disinfection of nondisposable medical equipment and the development of equipment for use single should allow in the long term a quasi-disappearance of infections.

5.3 The impact of the transition rate from I_H to R_H

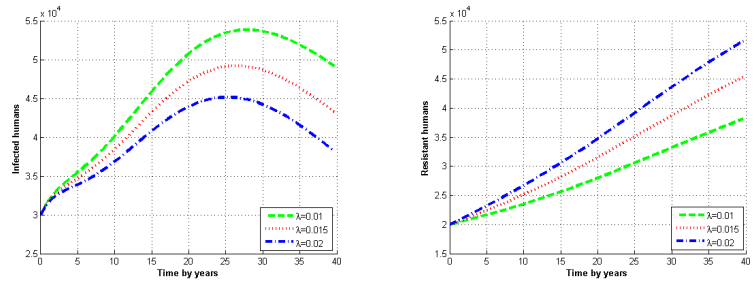


Figure 5: Evolution of human subpopulations for different values of λ.

For different values of $\lambda = 0.01, 0.015, 0.02$, the curves obtained in Figure 5 have made it possible to understand the important role of good care for infected people. Being infected with HCV does not protect against the risk of a new infection, which could worsen the medical situation. The development of a better therapeutic strategy can significantly improve the quality of life of people infected with hepatitis C.

6 Discussions and Conclusion

World Health Organization recommends that countries develop national strategies to reduce the burden of disease associated with hepatitis C hampered by weak or lacking national surveillance systems and unreliable estimates of the burden of hepatitis C morbidity.

In this work, we described and mathematically analyzed the dynamics of hepatitis C. The different numerical simulations were presented to see the behavior of the model at infinity, and the results obtained showed that the trends related to the prevention and management of infection considerably influence the subpopulations. We also applied the Jacobi wavelets method associated with the DQLT to obtain a numerical solution, which gave a very satisfactory results.

Due to the lack of data, our model has not been validated for the case of Algeria. Nevertheless the results of this modest work constitute the bases of work to be continued and improved for a much more in-depth study.

References

[1] Abdelrazec, A., Bélair, J., Shan, C. and Zhu, H. *Modeling the spread and control of dengue with limited public health resources*, Mathematical biosciences, 217 (2016), 136–145.

[2] Ablaoui-Lahmar, N., Belhamiti, O. and Bahri, S. M. *A new Legendre wavelets decomposition method for solving PDEs*, Malaya Journal of Matematik (MJM), 2 (1) (2014), 136–145.

[3] Abualrub, T. and Sadek, I. *Legendre wavelet operational matrix of derivative for optimal control in a convective–diffusive fluid problem*, J. Frankl. Inst. 351 (2) (2014), 682–693.

[4] Ainea, N., Massawe, E. S. and Makinde, O. D. *Modelling the effect of treatment and infected immigrants on the spread of hepatitis c virus disease with acute and chronic stages*, Am. J. Comput. Math. 2(1)(2012), 10–20.

[5] Ak Gümüs, O. *Global and local stability analysis in a nonlinear discrete-time population model* Adv. Difference Equ. 2014, 2014:299, 9 pp.

[6] Ali Merina, H and Belhamiti, O.*Simulation study of nonlinear reverse osmosis desalination system using third and fourth chebyshev wavelet methods*, MATCH Commun. Math. Comut. Chem. 75 (3)(2016), 629–652.

[7] Antoniadis, A. *Wavelet methods in statistics: Some recent developments and their applications*, Stat. Surv. 1 (2007), 16–55.

[8] Azodi, H. D. *Numerical solution of fractional-order sir epidemic model via Jacobi wavelets*, J. Int. Math. Virtual Inst. 10 (1) (2020), 183–197.

[9] Bokhari, A., Amir, A., Bahri, S. M. and Belgacem, F. B. M. *A generalized Bernoulli wavelet operational matrix of derivative applications to optimal control problems*, Nonlinear Stud. 24 (4) (2017), 75–90.

[10] Das, P., Mukherjee, D. and Sarkar, A. K. *Analysis of a disease transmission model of hepatitis C*, J. Biol. Syst. 13 (4) (2005), 331–339.

[11] Echevarria, D., Gutfraind, A., Boodram, B., Major, M., Del Valle, S., Cotler, S. J. and Dahari, H. *Mathematical modeling of hepatitis C prevalence reduction with antiviral treatment scale-up in persons who inject drugs in metropolitan Chicago*, PloS one, 10 (8) (2015) e0135901.

[12] Falconet, H., Jego, A., Veber, A. and Calvez, V. *Modéliser la propagation d'une épidémie*, thése sous la direction d'Amandine Veber et Vincent Calvez, (2015).

[13] Jackson, D. *Fourier series and orthogonal polynomials*, Courier Corporation, (2012).

[14] Hamou Maamar, M. and Belhamiti, O. *New (0,2) Jacobi multi-wavelets adaptive method for numerical simulation of gas separations using hollow fiber membranes*, Commun. Appl. Nonlinear Anal. 22 (3) (2015), 61–81.

[15] Khan, A., Sial, S. and Imran, M. *Transmission dynamics of hepatitis C with control strategies*, J. Comput. Med. (2014) 2014.

[16] Laguerre, E. N. *Sur la règle des signes de Descartes*, Nouvelles Annales de Mathématiques, 2e série, 18 (1879), 67–71.

[17] Mahmoud, A., Ameen, I. G. and Mohamed, A. *A new operational matrix based on Jacobi wavelets for a class of variable-order fractional differential equations*, Proc. Rom. Acad. - Math. Phys. Tech. Sci. Inf. Sci. 18 (4) (2017), 315–322.

[18] Mallat, S. *A wavelet tour of signal processing*, The sparse way. Third edition. With contributions from Gabriel Peyré. Elsevier/Academic Press, Amsterdam, 2009.

[19] Martcheva, M. and Castillo-Chavez, C. *Diseases with chronic stage in a population with varying size*, Math. Biosci. Elsevier, 182 (1) (2003), 1–25.

[20] Martin, N. K., Vickerman, P., Foster, G. R., Hutchinson, S. J., Goldberg, D. J. and Hickman, M. *Can antiviral therapy for hepatitis C reduce the prevalence of HCV among injecting drug user populations? A modeling analysis of its prevention utility*, J. Hepatol. 54 (6) (2011), 1137–1144.

[21] Martin, N. K., Vickerman, P., Grebely, J., Hellard, M., Hutchinson, S. J., Lima, V. D., Foster, G. R., Dillon, J. F., Goldberg, D. J., Dore, G. J. and Hickman, M., *Hepatitis C virus treatment for prevention among people who inject drugs: modeling treatment scale-up in the age of direct-acting antivirals*, Hepatology, 58 (5) (2013), 1598–1609.

[22] Martin, N. K., Vickerman, P., Hickman, M. *Mathematical modelling of hepatitis C treatment for injecting drug users*, J. Theoret. Biol. 274 (1) (2013), 58–66.

[23] Martin, N. K., Vickerman, P., Miners, A., Foster, G. R., Hutchinson, S. J., Goldberg, D. J. and Hickman, M. *Cost-effectiveness of hepatitis C virus antiviral treatment for injection drug user populations*, Hepatology, 55 (1) (2012), 49–57.

[24] Miller-Dickson, M. D., Meszaros, V. A., Almagro-Moreno, S. and Brandon Ogbunugafor, C. *Hepatitis C virus modelled as an indirectly transmitted infection highlights the centrality of injection drug equipment in disease dynamics*, J. R. Soc. Interface. 16 (158) (2019), 58–66.

[25] Organisation mondiale de la Santé. *Organisation mondiale de la Santé* Prévention des maladies chroniques: un investissement vital [Internet]. Geneva (2005).

[26] Razzaghi, M. and Yousefi, S. *Legendre wavelets direct method for variational problems*,Math. Comput. Simulation 53 (3) (2000), 185–192.

[27] Rong, L. J. and Chang, P. *Jacobi wavelet operational matrix of fractional integration for solving fractional integro-differential equation*, In Journal of Physics: Conference Series, vol. 693, no. 1, p. 012002. IOP Publishing, 2016.

[28] Shukla, N., Angelopoulou, A and Hodhod, R. *Non-Invasive Diagnosis of Liver Fibrosis in Chronic Hepatitis C using Mathematical Modeling and Simulation*, Electronics, MDPI, 11 (8) (2022) 1260.

[29] Szegö, G. *Orthogonal polynomials*, American Mathematical Society Colloquium Publications, Vol. 23 American Mathematical Society, New York, 1939.

[30] World Health Organization *Global hepatitis report 2017: executive summary*, No. WHO/HIV/2017.06. World Health Organization, 2017.

[31] Yang, Y., Tang, S., Ren, X., Zhao, H. and Guo, C. *Global stability and optimal control for a tuberculosis model with vaccination and treatment*, Discrete Contin. Dyn. Syst. Ser. B, 21 (3) (2016).

[32] Yuan, J. and Yang, Z. *Global dynamics of an SEI model with acute and chronic stages*, J. Comput. Appl. Math. 213 (2) (2008), 465–476.

[33] Zeiler, I., Langlands, T., Murray, J. M. and Ritter, A. *Optimal targeting of Hepatitis C virus treatment among injecting drug users to those not enrolled in methadone maintenance programs*, Drug and alcohol dependence, 110 (3) (2010), 228–233.

[34] Zhang, S. and Xu, X. *Dynamic analysis and optimal control for a model of hepatitis C with treatment*, Commun. Nonlinear Sci. Numer. Simul. 46 (2017), 14–25.

How to cite this article
Hamidat, N., Bahri, S.M. and Abbassa, N., Numerical nonlinear model solutions for the hepatitis C transmission between people and medical equipment using Jacobi wavelets method. *Iran. J. Numer. Anal. Optim.*, 2023; 13(4): 646-671. https://doi.org/10.22067/ijnao.2023.79648.1198

Iranian Journal of Numerical Analysis and Optimization
Vol. 13, No. 4, 2023, pp 672–694
https://doi.org/10.22067/ijnao.2023.81716.1238
https://ijnao.um.ac.ir/

Research Article

A shifted fractional-order Hahn functions Tau method for time-fractional PDE with nonsmooth solution

N. Mollahasani*,

Abstract

In this paper, a new orthogonal system of nonpolynomial basis functions is introduced and used to solve a class of time-fractional partial differential equations that have nonsmooth solutions. In fact, unlike polynomial bases, such basis functions have singularity and are constructed with a fractional variable change on Hahn polynomials. This feature leads to obtaining more accurate spectral approximations than polynomial bases. The introduced method is a spectral method that uses the operational matrix of fractional order integral of fractional-order shifted Hahn functions and finally converts the equation into a matrix equation system. In the introduced method, no collocation method has been used, and initial and boundary conditions are applied during the execution of the method. Error and convergence analysis of the numerical method has been investigated in a Sobolev space. Finally, some numerical experiments are considered in the form of tables and figures to demonstrate the accuracy and capability of the proposed method.

AMS subject classifications (2020): 65M70, 65M22, 65M15, 35R11, 26A33.

Keywords: Fractional-order shifted Hahn functions; Fractional-time partial differential equations; Spectral method, Error analysis; Convergence analysis.

*Corresponding author

Received 23 March 2023; revised 4 May 2023; accepted 4 May 2023

Nasibeh Mollahasani
Department of applied mathematics, Graduate university of advanced technology, Kerman, Iran.
e-mails: n.mollahasani@math.uk.ac.ir, nasibmo62@gmail.com

1 Introduction

In recent works, science and engineering researchers found that the use of fractional calculus in modeling gives a more realistic description of various complex phenomena with long-range temporal cumulative memory. Fractional order operators have nonlocal and memory features. Therefore, these two important properties simulate and describe a variety of engineering and scientific problems with memory characteristics and inheritance more appropriately than integer order differential equations, such as finance [30], physics [36], and hydrology [3], by using fractional differential equations. Many analytical methods have been used to solve fractional differential equations, such as the Green function method, Fourier, Laplace, and Mellin transform methods [24]. The complexity of integral and fractional differential operators and also the nonobservance of many properties expected in classic calculus encouraged researchers to study effective and reliable numerical methods for solving fractional differential equations. These numerical methods mainly include finite element and finite difference methods, spectral methods, and so on [31, 9, 8, 29, 34, 11, 1, 13, 22, 5, 23, 19]. In solving fractional order differential equations, two basic features that make classical methods not efficient and accurate are that fractional order operators have nonlocal properties and the other is the singularity of the solutions of fractional equations. Therefore, spectral methods based on ordinary polynomials, which have high accuracy for solving problems with smooth solutions (see, for example, [33, 12]), are not suitable for solving fractional differential equations with nonsmooth solutions since they do not have high expected accuracy. Concerning the numerical solution of partial differential equations dependent on time, one of the most common approaches is to use the finite difference approximation together with the spectral approximation for time and spatial derivatives, respectively. One of the main drawbacks of this approach is that the temporal discretization error may overcome the spatial discretization error, and the unknowns have to be solved simultaneously at all times [20]. As emphasized above, fractional differential equations mostly have nonsmooth solutions. It is also possible to encounter coefficients in terms of the given fractional equation in a nonsmooth case. On the other hand, in most of the spectral-introduced solving methods, in order to achieve high accuracy, they raise unrealistic assumptions. For instance, one of the assumptions in most of them is the smoothness of the unique regular solution of the fractional differential equation at the initial time $t = 0$ [32, 21, 35, 16]. So far, very few works have been done to solve fractional differential and integral equations with nonsmooth solutions, numerically, some of which can be seen in [15, 25, 26]. Due to their high accuracy, spectral methods have become one of the first choices researchers study to solve fractional differential equations with nonsmooth solutions. Among these techniques, we can refer to the methods available in [37, 38, 7]. Analytical and numerical studies indicate the exponential convergence of these methods for nonsmooth solutions

in certain situations, and by using specific techniques, though, the exact solution of fractional time differential equations does not generally follow the mentioned form [27, 14].

This paper is organized as follows: fractional Hahn functions and their properties are defined in section 2, and also, function approximation and the operational matrix of fractional integration are introduced. In section 3, our proposed method is described. An error analysis is presented in section 4, and finally, some numerical examples are depicted in section 5.

2 Fractional-order shifted Hahn functions approximation

The main goal of this section is to introduce a new class of fractional basis functions, which are defined using shifted Hahn polynomials (SHPs) and applied to calculating their operational matrix of fractional integration.

Definition 1. For given constants $\sigma_1, \sigma_2 > -1$, and $M \in \mathbb{N}$, Hahn polynomials on $[0, M]$ are defined as [17]

$$h_k(x; \sigma_1, \sigma_2, M) = \sum_{i=0}^{k} \frac{(-k)_i (k+\sigma_1+\sigma_2+1)_i (-x)_i}{(\sigma_1+1)_i (-M)_i i!}, \quad k = 0, 1, 2, \ldots, M, \tag{1}$$

where $(\cdot)_i$ is the Pochhammer notation, which is defined as

$$\begin{cases} (\zeta)_0 = 1, \\ (\zeta)_i = \zeta(\zeta+1)\cdots(\zeta+i-1), \quad i \in \mathbb{N}, \quad \text{for } \zeta \in \mathbb{R}^+. \end{cases} \tag{2}$$

Remark 1. The relationship between Stirling numbers and Pochhammer notation is as follows:

$$(-k)_i = (-1)^i \sum_{l=0}^{i} S_i^{(l)} k^l, \tag{3}$$

where $S_i^{(l)}$ are Stirling numbers of the first kind defined as

$$S_i^{(l)} = \sum_{r=0}^{i-l} (-1)^r \binom{i-1+r}{i-l+r} \binom{2r-l}{i-l-r} s_{i-l+r}^{(r)},$$

in which $s_i^{(l)}$ are Stirling numbers of the second kind in the form

$$s_i^{(l)} = \frac{1}{l!} \sum_{r=0}^{l} (-1)^{l-r} \binom{r}{l} r^i.$$

Now, by using (3) in (1) and the changing of variables as $x = \frac{Mt}{L}$, we can achieve the following standard polynomial form of SHPs on $[0, L]$ as

$$\begin{aligned}\overline{h_k}(t;\sigma_1,\sigma_2,M,L) &= h_k(\frac{Mt}{L};\sigma_1,\sigma_2,M)\\ &= \sum_{i=0}^{k}\sum_{l=0}^{i}(-1)^i\frac{(-k)_i(k+\sigma_1+\sigma_2+1)_i}{(\sigma_1+1)_i(-M)_i i!}\times S_i^{(l)}(\frac{M}{L})^l t^l\\ &= \sum_{i=0}^{k}\sum_{l=0}^{i}\Delta_{i,k,l}t^l,\end{aligned}$$

for $k = 0, 1, 2, \ldots, M$, where $\Delta_{i,k,l} = (-1)^i\frac{(-k)_i(k+\sigma_1+\sigma_2+1)_i}{(\sigma_1+1)_i(-M)_i i!}\times S_i^{(l)}(\frac{M}{L})^l$.

SHPs are orthogonal on $[0, L]$ via the inner product in the following form [28]:

$$\langle f, g\rangle_{\tilde{\omega}} := \sum_{r=0}^{M} f(\frac{L}{M}r)g(\frac{L}{M}r)\tilde{\omega}(r), \tag{4}$$

where $\tilde{\omega}(r)$ is a real nonnegative weight function defined by

$$\tilde{\omega}(x;\sigma_1,\sigma_2,M) = \binom{\sigma_1+x}{x}\binom{\sigma_2+M-x}{M-x}. \tag{5}$$

The orthogonal relationship of SHPs is as follows:

$$\langle\overline{h_k},\overline{h_j}\rangle_{\tilde{\omega}} := \begin{cases}\sum_{r=0}^{M}\overline{h_k}^2(\frac{L}{M}r,\sigma_1,\sigma_2,M,L)\tilde{\omega}(r), & k=j,\\ 0, & k\neq j.\end{cases} \tag{6}$$

To define fractional-order shifted Hahn functions (FOSHFs), t is substituted by t^α in SHPs such that α is a positive real number. Therefore, FOSHFs can be defined in the following form:

$$\begin{aligned}\overline{h_k^\alpha}(t;\sigma_1,\sigma_2,M,L) &= \sum_{i=0}^{k}\sum_{l=0}^{i}(-1)^i\frac{(-k)_i(k+\sigma_1+\sigma_2+1)_i}{(\sigma_1+1)_i(-M)_i i!}\times S_i^{(l)}(\frac{M}{L})^l t^{\alpha l}\\ &= \sum_{i=0}^{k}\sum_{l=0}^{i}\Delta_{i,k,l}t^{\alpha l}, \qquad k=0,1,2,\ldots,M.\end{aligned} \tag{7}$$

Proposition 1. FOSHFs are orthogonal on $[0, L]$ via the inner product in the following form:

$$\langle f, g\rangle_{\tilde{\omega}}^{\alpha} := \sum_{r=0}^{M} f((\frac{L}{M}r)^{\frac{1}{\alpha}})g((\frac{L}{M}r)^{\frac{1}{\alpha}})\tilde{\omega}(r), \tag{8}$$

where $\tilde{\omega}(r)$ is defined in (5).

Proof. Replacing f, g by $\overline{h_k^\alpha}, \overline{h_j^\alpha}$ in (8) and then using the orthogonality property (6), the assertion is available. □

Definition 2. Associated with the FOSHFs, the orthonormal FOSHFs can be defined as

$$\begin{aligned}&\overline{\mathcal{H}_k^\alpha}(t;\sigma_1,\sigma_2,M,L)\\&\quad=\frac{1}{\sqrt{\langle\overline{h_k^\alpha}(t;\sigma_1,\sigma_2,M,L),\overline{h_k^\alpha}(t;\sigma_1,\sigma_2,M,L)\rangle_{\tilde{\omega}}^\alpha}}\overline{h_k^\alpha}(t;\sigma_1,\sigma_2,M,L).\end{aligned}\tag{9}$$

2.1 Function approximation

For an integer $m \geq 0$, the Sobolev space $H_{\tilde{\omega}}^m\ [a,b]$ is

$$H_{\tilde{\omega}}^m[a,b]=\{u\in L_{\tilde{\omega}}^2[a,b]:\ 0\leq j\leq m,\ u^{(j)}(x)\in L_{\tilde{\omega}}^2[a,b]\},$$

where $L_{\tilde{\omega}}^2$ is the space of all square-integrable functions with respect to the weight function $\tilde{\omega}$. Indeed, $H_{\tilde{\omega}}^m[a,b]$ is defined as the vector space of functions $u \in L_{\tilde{\omega}}^2\ [a,b]$ such that all derivatives of u of order up to m can be represented by functions in $L_{\tilde{\omega}}^2\ [a,b]$.
Goertz and Öffner described the expansion of a function by Hahn polynomials and concluded that the series expansion of a function by Hahn polynomials converges pointwise under some assumptions (for more details, see [10]). Therefore, any function $u(t) \in L_{\tilde{\omega}}^2[0,L]$ can be expanded in terms of FOSHFs basis. In practice, only the first $(M+1)$ terms of FOSHFs are considered. Hence

$$u(t)\simeq\sum_{i=0}^{M}u_i\overline{\mathcal{H}_i^\alpha}(t;\sigma_1,\sigma_2,M,L)=u_M(t)=\mathbf{U}^T\mathcal{H}^{(\alpha)}(t;\sigma_1,\sigma_2,M,L),\tag{10}$$

where $\mathbf{U}^T=[u_0,u_1,\ldots,u_M]$ is the vector of FOSHFs coefficients, which can be derived as

$$\begin{aligned}u_i&=\left\langle u(t),\overline{\mathcal{H}_i^\alpha}(t;\sigma_1,\sigma_2,M,L)\right\rangle_{\tilde{\omega}}^\alpha\\&:=\sum_{r=0}^{M}u((\frac{L}{M}r)^{\frac{1}{\alpha}})\overline{\mathcal{H}_i^\alpha}((\frac{L}{M}r)^{\frac{1}{\alpha}};\sigma_1,\sigma_2,M,L)\tilde{\omega}(r),\quad i=0,1,\ldots,M,\end{aligned}\tag{11}$$

and $\mathcal{H}^{(\alpha)}(t;\sigma_1,\sigma_2,M,L)$ is the vector of FOSHFs defined as follows:

$$\mathcal{H}^{(\alpha)}(t;\sigma_1,\sigma_2,M,L):=[\overline{\mathcal{H}_0^\alpha}(t;\sigma_1,\sigma_2,M,L),\overline{\mathcal{H}_1^\alpha}(t;\sigma_1,\sigma_2,M,L),$$

$$\ldots, \overline{\mathcal{H}_M^{\alpha}}(t;\sigma_1,\sigma_2,M,L)]^T. \tag{12}$$

For simplicity, from now on, $\mathcal{H}^{(\alpha)}(t;\sigma_1,\sigma_2,M,L)$ is presented by $\mathcal{H}_M^{(\alpha)}(t)$.

Similarly, any two variables function $f(x,t) \in L^2_{\tilde{\omega}}([0,L]\times[0,T])$ can be approximated by the FOSHFs as follows:

$$\begin{aligned} f(x,t) &\simeq \sum_{i=0}^{M}\sum_{j=0}^{N} f_{i,j}\overline{\mathcal{H}_i^{\alpha}}(x;\sigma_1,\sigma_2,N,L)\overline{\mathcal{H}_j^{\beta}}(t;\sigma_1,\sigma_2,N,L) \\ &=: f_{M,N}(x,t) = (\mathcal{H}_M^{(\alpha)}(x))^T F \mathcal{H}_N^{(\beta)}(t), \end{aligned} \tag{13}$$

where $F=[f_{i,j}]$ is an $(M+1)\times(N+1)$ matrix that its entries are

$$\begin{aligned} f_{i,j} = \sum_{r_1=0}^{M}\sum_{r_2=0}^{N} & u((\frac{L}{M}r_1)^{\frac{1}{\alpha}},(\frac{T}{N}r_2)^{\frac{1}{\beta}})\overline{\mathcal{H}_i^{\alpha}}((\frac{L}{M}r_1)^{\frac{1}{\alpha}};\sigma_1,\sigma_2,M,L) \\ & \overline{\mathcal{H}_j^{\beta}}((\frac{T}{N}r_2)^{\frac{1}{\beta}};\sigma_1,\sigma_2,N,T)\tilde{\omega}(r_1)\tilde{\omega}(r_2), \end{aligned} \tag{14}$$

for $i=0,1,\ldots,M$ and $j=0,1,\ldots,N$.

Theorem 1. Let $M,N\in\mathbb{N}$, $\Lambda=[0,L]\times[0,T]$ and let $f\in H^2_{\tilde{\omega}}(\Lambda)$. Suppose that $f_{M,N}(x,t)=(\mathcal{H}_M^{(\alpha)}(x))^T F\mathcal{H}_N^{(\beta)}(t)$ is the best approximation of f in $\Omega = span\{\overline{\mathcal{H}_i^{\alpha}}(x;\sigma_1,\sigma_2,M,L)\overline{\mathcal{H}_j^{\beta}}(t;\sigma_1,\sigma_2,N,T)|\ i=0,1,\ldots,M,\ j=0,1,\ldots,N\}$. We will have

$$\|f(x,t)-f_{M,N}(x,t)\|^2_{L^2_{\tilde{\omega}}(\Lambda)} \le \frac{L^{M+2}T^{M+2}}{2^{2(M+N+1)}(M+1)!(N+1)!}\tilde{F},$$

where $\tilde{F}=\max_{(x,t)\in\Lambda}|\frac{\partial^{M+N}g(x,t)}{\partial x^M\partial t^N}|$ such that $g(x,t)=f(x^{\frac{1}{\alpha}},t^{\frac{1}{\alpha}})$.

Proof. Let $\phi_{M,N}(\eta,\xi)$ be the interpolation polynomial of $g(\eta,\xi)=f(\eta^{\frac{1}{\alpha}},\xi^{\frac{1}{\alpha}})$ at $(M+1)(N+1)$ shifted Chebyshev points in Λ. Then

$$|g(\eta,\xi)-\phi_{M,N}(\eta,\xi)| \le \frac{1}{2^{M+N}(M+1)!}(\frac{L}{2})^{M+1}(\frac{T}{2})^{N+1}\max_{(\eta,\xi)\in\Lambda}|\frac{\partial^{M+N}g(\eta,\xi)}{\partial\eta^M\partial\xi^N}|.$$

If $\tilde{F}=\max_{(\eta,\xi)\in\Lambda}|\frac{\partial^{M+N}g(\eta,\xi)}{\partial\eta^M\partial\xi^N}|$ and $\eta=x^{\alpha},\ \xi=t^{\beta}$ are sets, then we get

$$|g(x^{\alpha},t^{\beta})-\phi_{M,N}(x^{\alpha},t^{\beta})| \le \frac{1}{2^{M+N}(M+1)!}(\frac{L}{2})^{M+1}(\frac{T}{2})^{N+1}\tilde{F}. \tag{15}$$

It is obvious that $\phi_{M,N}(x^{\alpha},t^{\beta})\in\Omega$. So, since $f_{M,N}(x,t)$ is the best approximation of f concerning L^2-norm, we have

$$\|f(x,t)-f_{M,N}(x,t)\|_2 \le \|f(x,t)-\phi_{M,N}(x,t)\|_2$$

$$= (\int_0^L \int_0^T (f(x,t) - \phi_{M,N}(x,t))^2 dtdx)^{\frac{1}{2}}.$$

Thus, from (15) the assertion is derived. □

2.2 FOSHFs operational matrix of fractional integration

Here, an operational matrix of fractional integration for FOSHFs is going to be obtained. Note that the Riemann–Liouville fractional integration of order β for a function f is defined as

$$I^{\vartheta} f(x) = \frac{1}{\Gamma(\vartheta)} \int_a^x (x-t)^{\vartheta-1} f(t)dt, \qquad x > a, \quad \vartheta \geq 0. \tag{16}$$

For this special type of fractional integration, there are some particular properties. The most useful of which is

$$I^{\vartheta} x^{\gamma} = \frac{\Gamma(\gamma+1)}{\Gamma(\gamma+\vartheta+1)} x^{\vartheta+\gamma}. \tag{17}$$

Using the above concepts, the following lemma states the FOSHFs operational matrix of fractional integration.

Lemma 1. The fractional integration of order β of the vector $\mathbf{H}_M^{(\alpha)}(t)$ can be expanded by itself as follows:

$$I^{\vartheta} \mathcal{H}_M^{(\alpha)}(t) \simeq \mathfrak{P}_{\vartheta} \mathcal{H}_M^{(\alpha)}(t), \tag{18}$$

where $\mathfrak{P}_{\vartheta} = [\mathfrak{p}_{kj}]_{(M+1)\times(M+1)}$, which is called the FOSHFs operational matrix of fractional integration with

$$\mathfrak{p}_{kj} = \sum_{r=0}^{M} \sum_{i=0}^{k} \sum_{l=0}^{i} \sum_{i_1=0}^{j} \sum_{l_1=0}^{i_1} \tilde{\omega}(r) \overline{\Delta}_{i,k,l} \Delta_{i_1,j,l_1} \frac{\Gamma(\alpha l+1)}{\Gamma(\alpha l+\vartheta+1)} (\frac{L}{M} r)^{\frac{\vartheta+l\alpha}{\alpha}+l_1}.$$

Proof. According to (12), we have

$$I^{\vartheta} \mathcal{H}_M^{(\alpha)}(t) = \begin{bmatrix} I^{\vartheta} \overline{\mathcal{H}_0^{\alpha}}(t;\sigma_1,\sigma_2,M,L) \\ I^{\vartheta} \overline{\mathcal{H}_1^{\alpha}}(t;\sigma_1,\sigma_2,M,L) \\ \vdots \\ I^{\vartheta} \overline{\mathcal{H}_k^{\alpha}}(t;\sigma_1,\sigma_2,M,L) \\ \vdots \\ I^{\vartheta} \overline{\mathcal{H}_M^{\alpha}}(t;\sigma_1,\sigma_2,M,L) \end{bmatrix}. \tag{19}$$

By using (7), Proposition 1, the linear property of operator I, and (17) for each entry in (19), we will have

$$\begin{aligned} I^{\vartheta}\overline{\mathcal{H}_k^{\alpha}}(t;\sigma_1,\sigma_2,M,L) &= \sum_{i=0}^{k}\sum_{l=0}^{i}\overline{\Delta}_{i,k,l}I^{\beta}t^{\alpha l} \\ &= \sum_{i=0}^{k}\sum_{l=0}^{i}\overline{\Delta}_{i,k,l}\frac{\Gamma(\alpha l+1)}{\Gamma(\alpha l+\vartheta+1)}t^{\vartheta+l\alpha} \\ &\simeq \sum_{j=0}^{M}\mathfrak{p}_{kj}\overline{\mathcal{H}_j^{\alpha}}(t;\sigma_1,\sigma_2,M,L), \quad k=0,1,\ldots,M, \end{aligned}$$

where

$$\begin{aligned} \mathfrak{p}_{kj} &= \langle\sum_{i=0}^{k}\sum_{l=0}^{i}\overline{\Delta}_{i,k,l}\frac{\Gamma(\alpha l+1)}{\Gamma(\alpha l+\vartheta+1)}t^{\vartheta+l\alpha},\overline{\mathcal{H}_j^{\alpha}}(t;\sigma_1,\sigma_2,M,L)\rangle_{\tilde{\omega}}^{\alpha} \qquad (20)\\ &= \sum_{r=0}^{M}\tilde{\omega}(r)\sum_{i=0}^{k}\sum_{l=0}^{i}\overline{\Delta}_{i,k,l}\frac{\Gamma(\alpha l+1)}{\Gamma(\alpha l+\vartheta+1)}(\frac{L}{M}r)^{\frac{\vartheta+l\alpha}{\alpha}}\overline{\mathcal{H}_j^{\alpha}}((\frac{L}{M}r)^{\frac{1}{\alpha}};\sigma_1,\sigma_2,M,L) \\ &= \sum_{r=0}^{M}\sum_{i=0}^{k}\sum_{l=0}^{i}\tilde{\omega}(r)\overline{\Delta}_{i,k,l}\frac{\Gamma(\alpha l+1)}{\Gamma(\alpha l+\vartheta+1)}(\frac{L}{M}r)^{\frac{\vartheta+l\alpha}{\alpha}}\overline{\mathcal{H}_j^{\alpha}}((\frac{L}{M}r)^{\frac{1}{\alpha}};\sigma_1,\sigma_2,M,L), \end{aligned}$$

where $\overline{\Delta}_{i,k,l} = \frac{\Delta_{i,k,l}}{\sqrt{\langle\overline{h_k^{\alpha}}(t;\sigma_1,\sigma_2,M,L),\overline{h_k^{\alpha}}(t;\sigma_1,\sigma_2,M,L)\rangle_{\tilde{\omega}}^{\alpha}}}$. substituting (7) in (20) instead of $\overline{\mathcal{H}_j^{\alpha}}((\frac{L}{M}r)^{\frac{1}{\alpha}};\sigma_1,\sigma_2,M,L)$ finishes the proof. □

3 Description of method

The main aim of this section is to approximate the solution of the following equation:

$$D_t^{\vartheta}u(x,t) = -au(x,t) + b\frac{\partial u(x,t)}{\partial x} + c\frac{\partial^2 u(x,t)}{\partial x^2} + f(x,t), \qquad (21)$$

subject to the initial and boundary conditions

$$u(x,0)=g(x), \quad u(0,t)=\lambda(t), \quad u(L,t)=\eta(t) \quad \text{for } 0\le x\le L,\ 0\le t\le T.$$

Let

$$\frac{\partial^2 u(x,t)}{\partial x^2} \simeq (\mathcal{H}_M^{\alpha}(x))^T\mathbf{U}\mathcal{H}_N^{\beta}(t). \qquad (22)$$

Applying the integration operator I^{ϑ} on both sides of (22) and using the operational matrix of integration (18), for $\vartheta=1$ and $\vartheta=2$, respectively, yield

$$\frac{\partial u(x,t)}{\partial x} \simeq (\mathcal{H}_M^\alpha(x))^T(\mathfrak{P})^T\mathbf{U}\mathcal{H}_N^\beta(t) + xh(t), \tag{23}$$

$$u(x,t) \simeq (\mathcal{H}_M^\alpha(x))^T(\mathfrak{P}^2)^T\mathbf{U}\mathcal{H}_N^\beta(t) + xh(t) + \lambda(t), \tag{24}$$

in which the function $h(t)$ is calculated by putting $x = L$ in (24) and then applying the final condition $u(L,t) = \eta(t)$ as follows:

$$h(t) = \frac{1}{L}(\eta(t) - \lambda(t) - (\mathcal{H}_M^\alpha(L))^T(\mathfrak{P}^2)^T\mathbf{U}\mathcal{H}_N^\beta(t)).$$

Therefore (25) and (24) can be rewritten as follows:

$$\begin{aligned}\frac{\partial u(x,t)}{\partial x} &\simeq (\mathcal{H}_M^\alpha(x))^T(\mathfrak{P}^2)^T\mathbf{U}\mathcal{H}_N^\beta(t)\\ &\quad + \frac{1}{L}(\eta(t) - \lambda(t) - (\mathcal{H}_M^\alpha(L))^T(\mathfrak{P}^2)^T\mathbf{U}\mathcal{H}_N^\beta(t)), \qquad (25)\\ u(x,t) &\simeq (\mathcal{H}_M^\alpha(x))^T(\mathfrak{P}^2)^T\mathbf{U}\mathcal{H}_N^\beta(t)\\ &\quad + \frac{x}{L}(\eta(t) - \lambda(t) - (\mathcal{H}_M^\alpha(L))^T(\mathfrak{P}^2)^T\mathbf{U}\mathcal{H}_N^\beta(t)) + \lambda(t). \qquad (26)\end{aligned}$$

By substituting (22), (25), and (26) in (21), we get

$$\begin{aligned}D_t^\alpha u(x,t) &\simeq -a[(\mathcal{H}_M^\alpha(x))^T(\mathfrak{P}^2)^T\mathbf{U}\mathcal{H}_N^\beta(t)\\ &\quad + \frac{x}{L}(\eta(t) - \lambda(t) - (\mathcal{H}_M^\alpha(L))^T(\mathfrak{P}^2)^T\mathbf{U}\mathcal{H}_N^\beta(t)) + \lambda(t)]\\ &\quad + b[(\mathcal{H}_M^\alpha(x))^T(\mathfrak{P}^2)^T\mathbf{U}\mathcal{H}_N^\beta(t)\\ &\quad + \frac{1}{L}(\eta(t) - \lambda(t) - (\mathcal{H}_M^\alpha(L))^T(\mathfrak{P}^2)^T\mathbf{U}\mathcal{H}_N^\beta(t))]\\ &\quad + c(\mathcal{H}_M^\alpha(x))^T\mathbf{U}\mathcal{H}_N^\beta(t) + f(x,t)\\ &= (\mathcal{H}_M^\alpha(x))^T\mathfrak{A}\mathcal{H}_N^\beta(t), \qquad (27)\end{aligned}$$

where

$$\begin{aligned}\mathfrak{A} &= -a(\mathfrak{P}^2)^T\mathbf{U} + a\frac{\mathcal{X}}{L}(\mathcal{H}_M^\alpha(L))^T(\mathfrak{P}^2)^T\mathbf{U}\\ &\quad + b(\mathfrak{P})^T\mathbf{U} - b\frac{\infty}{L}(\mathcal{H}_M^\alpha(L))^T(\mathfrak{P}^2)^T\mathbf{U} + c\mathbf{U} + \mathbf{K}_1,\end{aligned}$$

and $\mathbf{K}_1$, $\mathbf{X}$, and $\mathbf{1}$ are the matrix and vector coefficient of FOSHF-approximation related to the following relations:

$$\begin{aligned}k_1(x,t) &= f(x,t) - a(\frac{x}{L}(\eta(t) - \lambda(t)) + \lambda(t)) + \frac{b}{L}(\eta(t) - \lambda(t))\\ &\simeq (\mathcal{H}_M^\alpha(x))^T\mathbf{K}_1\mathcal{H}_N^\beta(t),\\ x &\simeq (\mathcal{H}_M^\alpha(x))^T\mathbf{X},\end{aligned}$$

$$1 \simeq (\mathcal{H}_M^\alpha(x))^T \mathbf{1}.$$

Again, by applying the integration operator I_t^ϑ on both sides of (27) and using the operational matrix of integration $\mathfrak{P}^\vartheta$, we will have

$$u(x,t) \simeq (\mathcal{H}_M^\alpha(x))^T \mathfrak{A}\mathfrak{P}^\vartheta \mathcal{H}_N^\beta(t) + g(x). \tag{28}$$

Equalizing the right sides of (26) and (28), we get

$$(\mathcal{H}_M^\alpha(x))^T[(\mathfrak{P}^2)^T\mathbf{U} - \frac{\mathbf{X}}{L}(\mathcal{H}_M^\alpha(L))^T(\mathfrak{P}^2)^T\mathbf{U} - \mathfrak{A}\mathfrak{P}^\vartheta]\mathcal{H}_N^\beta(t) = (\mathcal{H}_M^\alpha(x))^T\mathbf{K}_2\mathcal{H}_N^\beta(t),$$

where $\mathbf{K}_2$ is the matrix coefficient of FOSHF-approximation related to the following relation:

$$k_2(x,t) = g(x) - \frac{x}{L}(\eta(t) - \lambda(t)) - \lambda(t) \simeq (\mathcal{H}_M^\alpha(x))^T\mathbf{K}_2\mathcal{H}_N^\beta(t).$$

Thus

$$(\mathfrak{P}^2)^T\mathbf{U} - \frac{\mathbf{X}}{L}(\mathcal{H}_M^\alpha(L))^T(\mathfrak{P}^2)^T\mathbf{U} - \mathfrak{A}\mathfrak{P}^\vartheta(t) = \mathbf{K}_2,$$

which can be rewritten as

$$\mathfrak{B}\mathbf{U} + \mathfrak{C}\mathbf{U}\mathfrak{D} = \mathfrak{E}, \tag{29}$$

where $\mathfrak{B} = (I - \frac{\mathbf{X}}{L}(\mathcal{H}_M^\alpha(L))^T)(\mathfrak{P}^2)^T$,
$\mathfrak{C} = [aI - \frac{a}{L}\mathbf{X}(\mathcal{H}_M^\alpha(L))^T + \frac{b}{L}\mathbf{1}(\mathcal{H}_M^\alpha(L))^T](\mathfrak{P}^2)^T - b\mathfrak{P}^T - cI$, $\mathfrak{D} = \mathfrak{P}^\vartheta$, and $\mathfrak{E} = \mathbf{K}_1\mathfrak{P}^\vartheta + \mathbf{K}_2$. Equation (29) is a matrix equation with the unknown matrix U. It can be solved by the global GMRES method. After solving the equation, by placing the obtained matrix U in (24), the approximate solution of the problem is obtained.

4 Error analysis

In this section, the convergence of the introduced method in a Sobolev space is considered. An upper bound is derived for the absolute error of the proposed method. To this end, some bounds are obtained for the approximations of different parts of the mentioned equation. First, the basic definitions and concepts related to Sobolev spaces are from the books [4, 18], with a slight change in symbols.

Let Λ be an open subset of $\mathbb{R}^n$ and let $L^2_{\tilde{\omega}}(\Lambda)$ be the space of all square-integrable functions concerning the weight function $\tilde{\omega}$. For an integer $m \geq 0$, the Sobolev space $H^m_{\tilde{\omega}}(\Lambda)$ is

$$H^m_{\tilde{\omega}}(\Lambda) = \{u|\ u \in L^2_{\tilde{\omega}}(\Lambda),\ \partial^\nu u \in L^2_{\tilde{\omega}}(\Lambda) \text{ for all } |\nu| \leq m\},$$

where ∂^ν is called the distributional derivatives and defined as the following form:

$$\partial^\nu u = \frac{\partial^{|\nu|} u}{\partial x_1^{\nu_1} \partial x_2^{\nu_2} \cdots \partial x_n^{\nu_n}}, \quad |\nu| = \nu_1 + \nu_2 + \cdots + \nu_n.$$

For all $u,\ \nu \in H^m_\omega(\Lambda)$, the inner product is given as

$$\langle u, \nu \rangle_{H^m_\omega(\Lambda)} = \langle u, \nu \rangle_{L^2_\omega(\Lambda)} + \sum_{1 \le |\nu| \le m} \langle \partial^\nu u, \partial^\nu \nu \rangle_{L^2_\omega(\Lambda)}.$$

The corresponding norm and seminorm are defined as

$$\|u\|_{H^m_\omega(\Lambda)} = (\|u\|^2_{L^2_\omega(\Lambda)} + \sum_{1 \le |\nu| \le m} \|\partial^\nu u\|^2_{L^2_\omega(\Lambda)})^{\frac{1}{2}},$$
$$|u|_{H^m_\omega(\Lambda)} = (\sum_{|\nu| = m} \|\partial^\nu u\|^2_{L^2_\omega(\Lambda)})^{\frac{1}{2}}.$$

It is obvious to see that if $m \geqslant 0$, then $\|u\|_{L^2_\omega(\Lambda)} \leqslant \|u\|_{H^m_\omega(\Lambda)}$. In a special case, for $m = 0$, it yields $\|u\|_{H^m_\omega(\Lambda)} = \|u\|_{L^2_\omega(\Lambda)}$. Also, for $m = 0$, we have $|u|_{H^m_\omega(\Lambda)} = \|u\|_{L^2_\omega(\Lambda)}$.

Suppose that $u \in H^m_\omega(\Lambda)$ and $\mathcal{P}^{\alpha,\beta}_{M,N}$ are the orthogonal projection operator, where $\Lambda = [0, L] \times [0, T]$ and

$$\mathcal{P}^{\alpha,\beta}_{M,N} u := \sum_{i=0}^{M} \sum_{j=0}^{N} u_{i,j} \overline{\mathcal{H}^\alpha_i}(x; \sigma_1, \sigma_2, N, L) \overline{\mathcal{H}^\beta_j}(t; \sigma_1, \sigma_2, N, L).$$

In other words, $\mathcal{P}^{\alpha,\beta}_{M,N} u = u_{M,N}(x,t) = (\mathcal{H}^{(\alpha)}_M(x))^T U \mathcal{H}^{(\beta)}_N(t)$.

In the following, for simplicity and brevity, $M = N$, $\alpha = \beta$, and $\mathcal{P}_M := \mathcal{P}^{\alpha,\beta}_{M,N}$ are stated. According to [6], for all $u \in H^m_\omega(\Lambda)$, we have

$$\|u - \mathcal{P}_M u\|_{H^j_\omega(\Lambda)} \le C M^{\rho(j) - m} |u|_{H^{m;M}_\omega(\Lambda)}, \quad 0 \le j \le m, \tag{30}$$

where C is a constant independent of M and only depends on m,

$$\rho(j) = \begin{cases} 0, & j = 0, \\ 2j - \frac{1}{2}, & j > 0, \end{cases}$$

and

$$|u|_{H^{m;M}_\omega(\Lambda)} = (\sum_{k = \min(m, M+1)}^{m} \sum_{i=1}^{2} \|D^k_i u\|^2_{L^2_\omega})^{\frac{1}{2}}.$$

Theorem 2. Suppose that $u(x,t) \in H^m_\omega(\Lambda)$, $m \geqslant 0$ and that $u_{N,M}(x,t)$ is the best approximation of u. Then

$$\|u(x,t) - u_{N,M}(x,t)\|_{L^2_\omega(\Lambda)} \leqslant \|u(x,t) - u_{N,M}(x,t)\|_{H^j_\omega(\Lambda)}(\Lambda)$$

$$\leqslant CM^{\rho(j)-m}|u|_{H^{m;M}_{\tilde{\omega}}(\Lambda)}, \quad 0 \leq j \leq m. \quad (31)$$

Proof. Considering this fact that $\|\cdot\|_{L^2_{\tilde{\omega}}(\Lambda)} \leqslant \|\cdot\|_{H^m_{\tilde{\omega}}(\Lambda)}$, inequality (30), and the uniqueness of the best approximation, the proof of the theorem is easily done. □

Lemma 2. Suppose that the assumptions of Theorem 2 are true, that $u(x,t) \simeq u_{M,N}(x,t) = (\mathcal{H}^{(\alpha)}_M(x))^T U \mathcal{H}^{(\beta)}_N(t)$, and that $\mathfrak{P}_\vartheta$ is the FOSHF-operational matrix of fractional integration. Then

$$\begin{aligned}&\|I^\vartheta_x u(x,t) - (\mathcal{H}^{(\alpha)}_M(x))^T \mathfrak{P}^T_\vartheta U \mathcal{H}^{(\beta)}_N(t)\|_{L_2(I)}\\ &\quad\leq \frac{L^\vartheta}{\Gamma(\vartheta+1)} CM^{\rho(j)-m}|u|_{H^{m;M}_{\tilde{\omega}}(\Lambda)}, \quad 0 \leq j \leq m.\end{aligned}$$

Proof. According to (16), we have

$$\begin{aligned}&\|I^\vartheta_x u(x,t) - (\mathcal{H}^{(\alpha)}_M(x))^T \mathfrak{P}^T_\vartheta U \mathcal{H}^{(\beta)}_N(t)\|_{L^2_{\tilde{\omega}}(\Lambda)}\\ &\quad= \|I^\vartheta_x u(x,t) - I^\vartheta_x u_{M,N}(t)\|_{L^2_{\tilde{\omega}}(\Lambda)}\\ &\quad= \|I^\vartheta_x (u(x,t) - u_{M,N}(x,t))\|_{L^2_{\tilde{\omega}}(\Lambda)}\\ &\quad= \|\frac{1}{\Gamma(\vartheta)}\int_0^x (x-\xi)^{\vartheta-1}(u(\xi,t) - u_{M,N}(\xi,t))d\vartheta\|_{L^2_{\tilde{\omega}}(\Lambda)}\\ &\quad= \frac{1}{\Gamma(\vartheta)}\|x^{\vartheta-1} * (u(x,t) - u_{M,N}(x,t))\|_{L^2_{\tilde{\omega}}(\Lambda)}. \qquad (32)\end{aligned}$$

Now, by using this fact that $\|f * g\|_\rho \leq \|f\|_1.\|g\|_\rho$, and Theorem 2, respectively, we get

$$\begin{aligned}&\|I^\vartheta_x u(x,t) - (\mathcal{H}^{(\alpha)}_M(x))^T \mathfrak{P}^T_\vartheta U \mathcal{H}^{(\beta)}_N(t)\|_{L^2_{\tilde{\omega}}(\Lambda)}\\ &\quad\leq \frac{L^\vartheta}{\vartheta\Gamma(\vartheta)}\|u(x,t) - u_M(x,t)\|_{L^2_{\tilde{\omega}}(\Lambda)}\\ &\quad\leq \frac{L^\vartheta}{\Gamma(\vartheta+1)} CM^{\rho(j)-m}|u|_{H^{m;M}_{\tilde{\omega}}(\Lambda)}, \quad 0 \leq j \leq m. \qquad (33)\end{aligned}$$

□

To get an error bound for derived approximation in the proposed method, which has been introduced in section 3, without losing the generality, we suppose that

$$\frac{\partial^2 u(x,t)}{\partial x^2} \simeq (\mathcal{H}^{(\alpha)}_M(x))^T U \mathcal{H}^{(\beta)}_N(t) =: \phi_{M,N}(x,t), \qquad (34)$$

$$\frac{\partial u(x,t)}{\partial x} \simeq (\mathcal{H}^{(\alpha)}_M(x))^T W \mathcal{H}^{(\beta)}_N(t) = \varphi_{M,N}(x,t), \qquad (35)$$

$$u(x,t) \simeq (\mathcal{H}_M^{(\alpha)}(x))^T V \mathcal{H}_N^{(\beta)}(t) =: \psi_{M,N}(x,t). \tag{36}$$

As it can be seen from the process of the presented method in section 3 that relation (26) has appeared in applying the operator I_x^2 on the sides of (34) and (36) can be derived by expanding (26) in terms of FOSHFs basis. It is easy to see that

$$\|u(x,t) - \psi_{M,N}(x,t)\|_{L^2_{\tilde{\omega}}(\Lambda)} = \|I_x^2 \frac{\partial^2 u(x,t)}{\partial x^2} - I_x^2 \phi_{M,N}(x,t)\|_{L^2_{\tilde{\omega}}(\Lambda)}. \tag{37}$$

So, considering (37) and applying Lemma 2, the following corollary is obtained.

Corollary 1. If relation (34) is true, then

$$\|u(x,t) - \psi_{M,N}(x,t)\|_{L^2_{\tilde{\omega}}(\Lambda)} \leq \frac{L^2}{\Gamma(3)} C M^{\rho(j)-m} |\frac{\partial^2 u(x,t)}{\partial x^2}|_{H^{m;M}_{\tilde{\omega}}(\Lambda)}, \quad 0 \leq j \leq m. \tag{38}$$

Consider the main equation (21) and the presented method in section 3, by substituting (34)–(36) on the right side of (21) and applying the operator I_t^ϑ on it, we get

$$u(x,t) \simeq -aI_t^\vartheta \psi(x,t) + bI_t^\vartheta \varphi(x,t) + cI_t^\vartheta \phi(x,t) + I_t^\vartheta f(x,t) + g(x). \tag{39}$$

On the other hand, we have

$$u(x,t) = -aI_t^\vartheta u(x,t) + bI_t^\vartheta \frac{\partial u(x,t)}{\partial x} + cI_t^\vartheta \frac{\partial^2 u(x,t)}{\partial x^2} + I_t^\vartheta f(x,t) + g(x). \tag{40}$$

Putting the right side of (39) and (40) as equivalent, we define perturbation term as follows:

$$\begin{aligned} \mathfrak{R}_{M,N}(x,t) := -aI_t^\vartheta (u(x,t) - \psi_{M,N}(x,t)) + bI_t^\vartheta (\frac{\partial u(x,t)}{\partial x} \\ - \varphi_{M,N}(x,t)) + cI_t^\vartheta (\frac{\partial^2 u(x,t)}{\partial x^2} - \phi_{M,N}(x,t)). \end{aligned} \tag{41}$$

Theorem 3. Suppose that, $u(x,t) \in H^m_{\tilde{\omega}}(\Lambda)$ for $m \geqslant 0$ is the exact solution of (21). If $\psi_{M,N}(x,t)$ is the approximate solution, obtained by applying the presented method, then $\mathfrak{R}_{M,N}(x,t) \longrightarrow 0$ as $M,\ N \longrightarrow \infty$.

Proof. According to (41), we have

$$\begin{aligned} \|\mathfrak{R}_{M,N}(x,t)\|_{L^2_{\tilde{\omega}}(\Lambda)} \leq |a| \|I_t^\vartheta (u(x,t) - \psi_{M,N}(x,t))\|_{L^2_{\tilde{\omega}}(\Lambda)} \\ + |b| \|I_t^\vartheta (\frac{\partial u(x,t)}{\partial x} - \varphi_{M,N}(x,t))\|_{L^2_{\tilde{\omega}}(\Lambda)} \\ + |c| \|I_t^\vartheta (\frac{\partial^2 u(x,t)}{\partial x^2} - \phi_{M,N}(x,t))\|_{L^2_{\tilde{\omega}}(\Lambda)}. \end{aligned} \tag{42}$$

Now, by applying Lemma 2 in approximations (34)–(36), respectively, we get

$$\begin{aligned}
&\|I_t^{\vartheta}(u(x,t)-\psi_{M,N}(x,t))\|_{L^2_{\tilde{\omega}}(\Lambda)}\\
&\quad\leq \frac{T^{\vartheta}}{\Gamma(\vartheta+1)}CM^{\rho(j)-m}|u|_{H^{m;M}_{\tilde{\omega}}(\Lambda)}, \quad 0\leq j\leq m, \qquad (43)\\
&\|I_t^{\vartheta}(\frac{\partial u(x,t)}{\partial x}-\phi_{M,N}(x,t))\|_{L^2_{\tilde{\omega}}(\Lambda)}\\
&\quad\leq \frac{T^{\vartheta}}{\Gamma(\vartheta+1)}CM^{\rho(j)-m}|\frac{\partial u}{\partial x}|_{H^{m;M}_{\tilde{\omega}}(\Lambda)}, \quad 0\leq j\leq m, \qquad (44)\\
&\|I_t^{\vartheta}(\frac{\partial^2 u(x,t)}{\partial x^2}-\phi_{M,N}(x,t))\|_{L^2_{\tilde{\omega}}(\Lambda)}\\
&\quad\leq \frac{T^{\vartheta}}{\Gamma(\vartheta+1)}CM^{\rho(j)-m}|\frac{\partial^2 u}{\partial x^2}|_{H^{m;M}_{\tilde{\omega}}(\Lambda)}, \quad 0\leq j\leq m. \qquad (45)
\end{aligned}$$

So, by using (43)–(45) in (42), it yields

$$\begin{aligned}
&\|\mathfrak{R}_{M,N}(x,t)\|_{L^2_{\tilde{\omega}}(\Lambda)} \qquad (46)\\
&\quad\leq \frac{T^{\vartheta}}{\Gamma(\vartheta+1)}CM^{\rho(j)-m}(|a||u|_{H^{m;M}_{\tilde{\omega}}(\Lambda)}+|b||\frac{\partial u}{\partial x}|_{H^{m;M}_{\tilde{\omega}}(\Lambda)}+|c||\frac{\partial^2 u}{\partial x^2}|_{H^{m;M}_{\tilde{\omega}}(\Lambda)}).
\end{aligned}$$

Hence, it is concluded that $\mathfrak{R}_{M,N}(x,t)\longrightarrow 0$ as $M,\ N\longrightarrow\infty$. □

5 Numerical results

In this section, the introduced method in section 3 is utilized to approximate the solutions to problems. It should be mentioned that the maximum of absolute error is the infinity norm of the error function and

$$L_\infty=\max_{1\leq j\leq N}|e(x_j,T)|.$$

All numerical experiments have been performed using MATLAB R2017a on a Core(TM)2 laptop with 4GB RAM and a speed of 2.00 GHz.

Example 1. Consider the following time-fractional diffusion differential equation:

$$D_t^{\vartheta}u(x,t)=-u(x,t)+\frac{\partial^2 u(x,t)}{\partial x^2}+f(x,t), \quad 0<\vartheta<1,\ (x,t)\in[0,1]\times[0,1], \qquad (47)$$

where $f(x,t)=sin(\pi x)(1+\frac{t^{\vartheta}}{\Gamma(\vartheta+1)})+\frac{\pi^2 t^{\vartheta}}{\Gamma(\vartheta+1)})$, subject to the initial and boundary conditions:

$$u(x,0)=0, \quad u(0,t)=0, \quad u(1,t)=0.$$

The exact solution is $u(x,t) = \frac{t^{\vartheta}}{\Gamma(\vartheta+1)} sin(\pi x)$. Table 1 shows the L_∞-norm of absolute error for fixed $N = 1$ and some M and β in comparison to [2]. In Figure 1, the L_∞-norm of absolute error for fixed $N = 1,\ \vartheta = 0.9$, and some $M = 4, 5, \ldots, 10$ is shown, which demonstrates that the approximate solution converges to the exact solution as M increases. Finally, Figure 2 shows the absolute error functions for fixed $N = 1,\ \vartheta = 0.9$, and some $M = 6,\ 8,\ 10$.

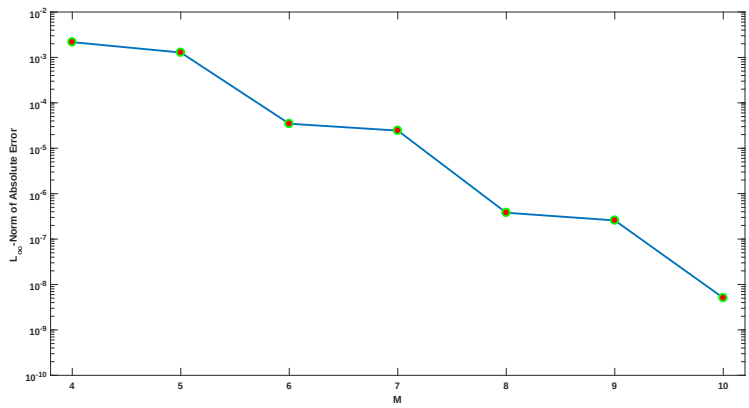


Figure 1: L_∞-norm of the absolute error function for fixed $N = 1,\ \vartheta = 0.9$, and some $M = 4, 5, \ldots, 10$ (Example 1)

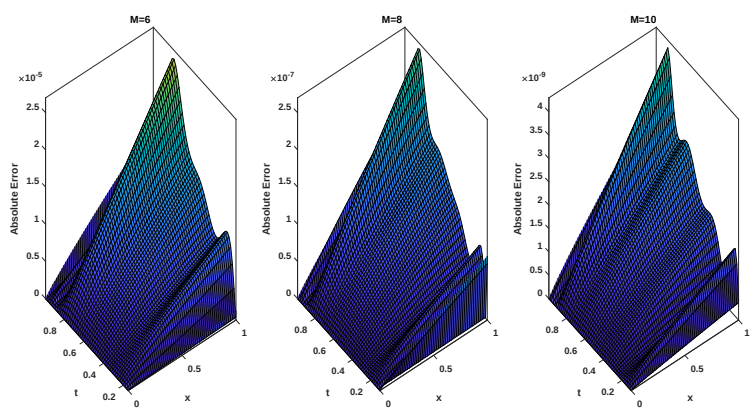


Figure 2: Absolute error functions for fixed $N = 1,\ \vartheta = 0.9$, and $M = 6,\ 8,\ 10$ (Example 1)

Example 2. Consider the following inhomogeneous fractional-order Burger's equation:

Table 1: L_∞-norm of absolute error for fixed $N = 1$ and some M and β in comparison to [2] (Example 1)

M	$\vartheta = 0.25$	[2]	$\vartheta = 0.5$	[2]	$\vartheta = 0.75$	[2]
4	2.301e-3	1.690e-3	2.389e-4	4.979e-3	2.287e-4	2.918e-3
6	3.638e-5	5.764e-4	3.721e-6	3.331e-5	3.589e-6	2.752e-5
8	4.517e-7	1.761e-6	4.621e-8	1.754e-7	4.455e-8	1798e-7
10	7.101e-10	3.127e-9	7.263e-10	8.428e-10	7.003e-10	8.116e-10

$$D_t^{\vartheta} u(x,t) = \frac{\partial^2 u(x,t)}{\partial x^2} - \frac{\partial u(x,t)}{\partial x} + f(x,t), \quad 0 < \vartheta \leq 1, \ (x,t) \in [0,1] \times [0,1], \tag{48}$$

where $f(x,t) = \frac{2t^{2-\vartheta}}{\Gamma(3-\vartheta)} + 2x - 2$, subject to the initial and boundary conditions:

$$u(x,0) = x^2, \quad u(0,t) = t^2, \quad u(1,t) = 1 + t^2.$$

The exact solution is $u(x,t) = x^2 + t^2$. Figures 3 and 4 show the absolute error functions after solving the problem by using the presented method with $M = 2,\ N = 4,\ \alpha = 1,\ \beta = 0.5$ for the fractional-order derivative $\vartheta = 0.5$ and $M = 2,\ N = 2,\ \alpha = 1$, $\beta = 1$, and $\vartheta = 1$, respectively.

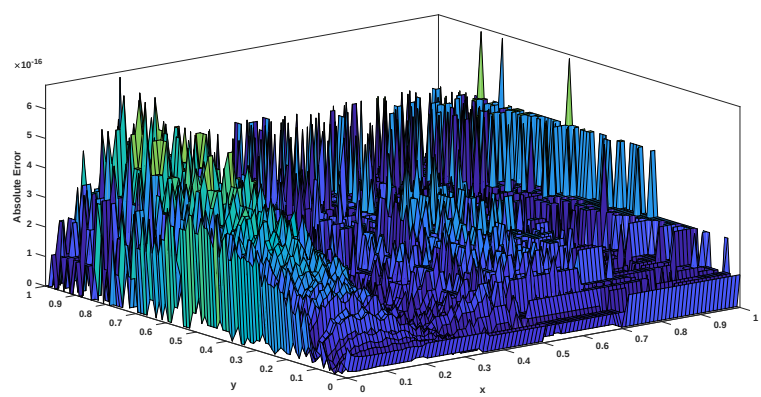


Figure 3: Absolute error function for $M = 2,\ N = 4,\ \alpha = 1,\ \beta = 0.5$, and $\vartheta = 0.5$ (Example 2)

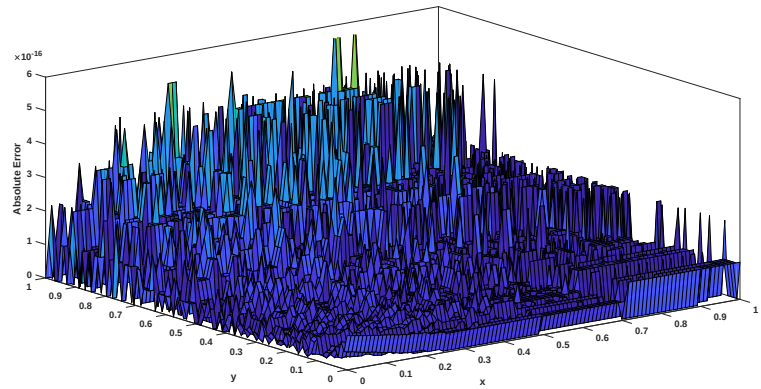


Figure 4: Absolute error function for $M = 2,\ N = 2,\ \alpha = 1,\ \beta = 1$, and $\vartheta = 1$ (Example 2)

Example 3. Consider the following transformed time-fractional Black–Scholes model with homogeneous boundary conditions:

$$D_t^{\vartheta} u(x,t) - \frac{\sigma^2}{2}\frac{\partial^2 u(x,t)}{\partial x^2} - (r - \frac{\sigma^2}{2})\frac{\partial u(x,t)}{\partial x} + ru(x,t) = f(x,t),$$
$$0 < \vartheta \leq 1,\ (x,t) \in (0,1) \times (0,1], \tag{49}$$

where

$$f(x,t) = \frac{6t^{3-\vartheta}}{\Gamma(4-\vartheta)}(x^5 - x^4) - (t^3+1)[\frac{\sigma^2}{2}(20x^3 - 12x^2)$$
$$+ (r - \frac{\sigma^2}{2})(5x^4 - 4x^3) - r(x^5 - x^4)],$$

subject to the initial and boundary conditions:

$$u(x,0) = x^5 - x^4, \quad u(0,t) = 0, \quad u(1,t) = 0.$$

The exact solution is $u(x,t) = (t^3+1)(x^5 - x^4)$. Let $r = 0.02$ and let $\sigma = 0.8$. Figure 5 shows the absolute error function obtained by applying the presented method for $\vartheta = 0.5$ $\alpha = 1,\ \beta = 0.5,\ M = 5$, and $N = 6$. Also, Figure 6 shows the absolute error after solving the problem by using the presented method with $M = 5,\ N = 3,\ \alpha = 1$, and $\beta = 1$ for $\vartheta = 1$.

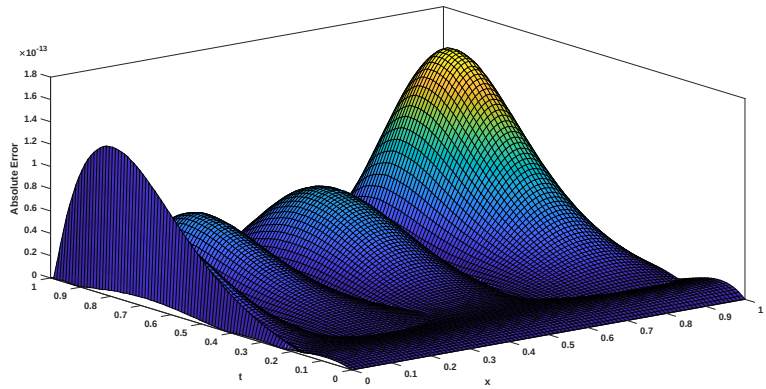


Figure 5: Absolute error function for $\vartheta = 0.5$ $\alpha = 1$, $\beta = 0.5$, $M = 5$, and $N = 6$ (Example 3)

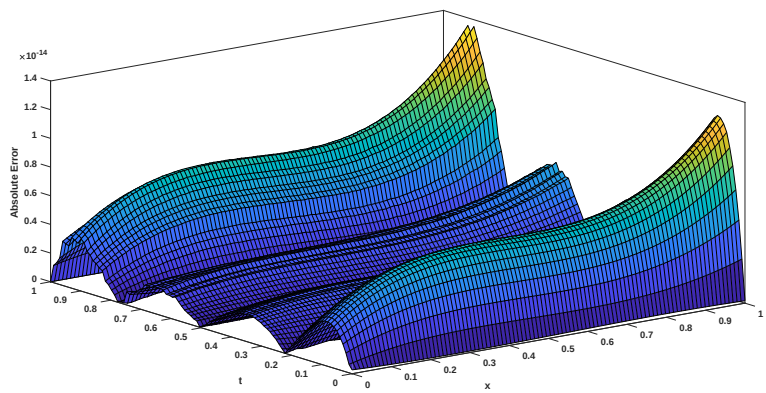


Figure 6: Absolute error function for $\vartheta = 0.5$ $\alpha = 1$, $\beta = 1$, $M = 5$, and $N = 3$ (Example 3).

Example 4. Consider the following time-fractional equation:

$$D_t^{\vartheta} u(x,t) = -au(x,t) + b\frac{\partial u(x,t)}{\partial x} + c\frac{\partial^2 u(x,t)}{\partial x^2} = f(x,t),$$
$$0 < \vartheta \leq 1, \ (x,t) \in (0,L) \times (0,T], \quad (50)$$

subject to the initial and boundary conditions:

$$u(x,0) = x^{\frac{5}{2}}, \quad u(0,t) = 0, \quad u(L,t) = L^{\frac{5}{2}} e^{-t},$$

where in the case of $\vartheta = 1$ and the function f is chosen as $f(x,t) = -e^{-t}\frac{5}{2}(x^{\frac{3}{2}} + \frac{3}{2}\sqrt{x})$, the exact solution is $u(x,t) = e^{-t}x^{\frac{5}{2}}$. It is notable that in other cases of $0 < \vartheta < 1$, the exact solution is unknown. Figure 7 shows the absolute error functions obtained by applying the presented method for $\vartheta = 0.5,\ \alpha = 0.5,\ \beta = 1,\ M = 5$, and $N = 4,\ 6,\ 8$. Also, Figure 8 shows the L_∞-norm of the absolute error function for fixed $M = 5,\ \vartheta = 0.5$, and some $N = 2, 3, \ldots, 8$, which demonstrates that the L_∞-norm of the absolute error function converges to zero as N increases. Finally, Figure 9 depicts approximate solutions for different $0 < \vartheta \leq 1,\ M = 5,\ N = 7$, which shows that as $\vartheta \to 1$, the approximate solution converges to the exact solution when $\vartheta = 1$.

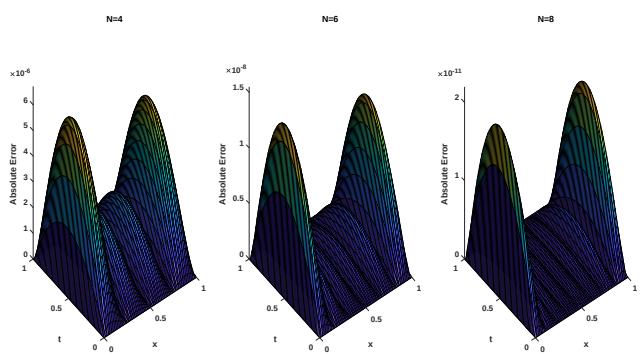


Figure 7: Absolute error functions for $\vartheta = 0.5,\ \alpha = 0.5,\ \beta = 1,\ M = 5$, and $N = 4,\ 6,\ 8$. (Example 4)

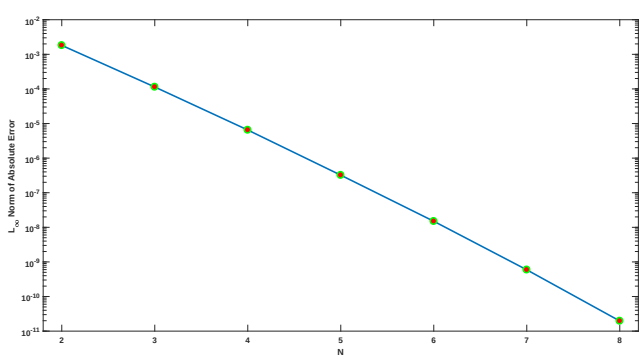


Figure 8: L_∞-norm of absolute error function for fixed $M = 5,\ \vartheta = 0.5$, and $N = 2, 3, \ldots, 8$. (Example 4)

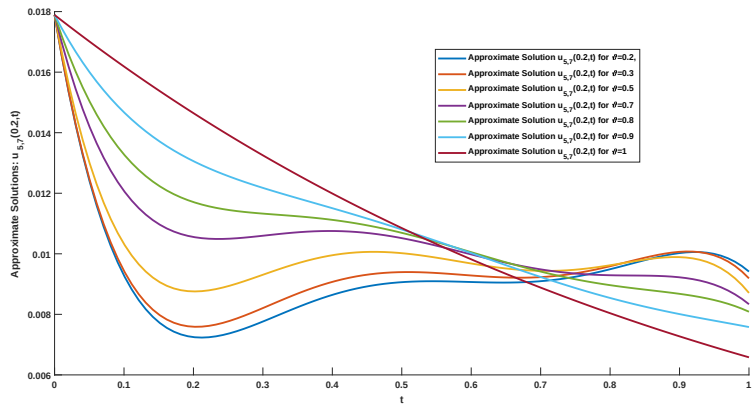


Figure 9: Approximate solutions for different $0 < \vartheta \leq 1,\ M = 5,\ N = 7$. (Example 4)

6 Conclusion

In this paper, a new orthogonal system of nonpolynomial basis functions, named FOSHFs, has been introduced and used to solve a class of fractional-time partial differential equations with nonsmooth solutions. For introducing the method, an operational matrix of fractional order integral of Hahn functions has been used for the first time as basis functions here. Furthermore, the convergence of FOSHFs approximation has been proved. In numerical examples, the obtained results have demonstrated the efficiency and convergence of the proposed method for the cases of nonsmooth solutions.

7 Declarations

The author declares that there is no conflict of interest.

References

[1] Abbaszadeh, M. and Dehghan, M. *Numerical investigation of reproducing kernel particle Galerkin method for solving fractional modified distributed-order anomalous sub-diffusion equation with error estimation*, Appl. Math. Comput. 392 (2021), 125–718.

[2] Abo-Gabal, H., Zaky, M.A. and Doha, E.H. *Fractional Romanovski–Jacobi tau method for time-fractional partial differential equations with nonsmooth solutions*, Appl. Numer. Math. 182 (2022), 214–234.

[3] Benson, D.A., Wheatcraft, S.W. and Meerschaert, M.M. *The fractional-order governing equation of Levy motion*, Water Resour. Res. 36 (6) (2000), 1413–1423.

[4] Bhattacharyya, P.K. *Distributions Generalized Functions with Applications in Sobolev Spaces*, Distributions. de Gruyter, 2012.

[5] Bhrawy, A. and Zaky, M. *An improved collocation method for multi-dimensional space-time variable-order fractional Schroedinger equations*, Appl. Numer. Math. 111 (2017), 197–218.

[6] Canuto, C., Quarteroni, A., Hussaini, M.Y. and Zang, T.A. *Spectral methods; Fundamentals in single domains*, Springer Science & Business Media, 2007.

[7] Chen, S., Shen, J. and Wang, L. *Generalized Jacobi functions and their applications to fractional differential equations*, Math. Comput. 85 (300) (2016), 1603–1638.

[8] Dehghan, M., Abbaszadeh, M. and Mohebbi, A. *Legendre spectral element method for solving time fractional modified anomalous subdiffusion equation*, Appl. Math. Model. 40 (5-6) (2016), 3635–3654.

[9] Deng, W. *Finite element method for the space and time fractional Fokker–Planck equation,* SINUM. 47 (1) (2009), 204–226.

[10] Goertz, R. and Öffner, P. *Spectral accuracy for the Hahn polynomials*, ArXiv e-prints: arXiv:1609.07291, 2016.

[11] Hendy, A.S. and Zaky, M.A. *Global consistency analysis of L1-Galerkin spectral schemes for coupled nonlinear space-time fractional Schrödinger equations*, Appl. Numer. Math. 156 (2020), 276–302.

[12] Hesthaven, J.S., Gottlieb, S. and Gottlieb, D. *Spectral methods for time-dependent problems*, Cambridge Monographs on Applied and Computational Mathematics, 21. Cambridge University Press, Cambridge, 2007.

[13] Heydari, M., Avazzadeh, Z. and Atangana, A. *Orthonormal shifted discrete Legendre polynomials for solving a coupled system of nonlinear variable-order time fractional reaction-advection-diffusion equations*, Appl. Numer. Math. 161 (2021), 425–436.

[14] Hou, D., Hasan, M.T. and Xu, C. *Muntz spectral methods for the time-fractional diffusion equation*, Comput. Methods Appl. Math. 18 (1) (2018), 43–62.

[15] Jin, B., Lazarov, R. and Zhou, Z. *Error estimates for a semi-discrete finite element method for fractional order parabolic equations*, SIAM Journal on Numerical Analysis 51 (1) (2013), 445–466.

[16] Jin, B., Lazarov, R. and Zhou, Z. *Numerical methods for time-fractional evolution equations with nonsmooth data: a concise overview*, Comput. Methods Appl. Mech. Eng. 346 (2019), 332–358.

[17] Karlin, S. and McGregor, J.L. *The Hahn polynomials, formulas and an application*, Scripta Math. 26 (1961), 33–46.

[18] Kreyszig, E. *Introductory functional analysis with applications*, John Wiley & Sons, New York-London-Sydney, 1978.

[19] Latifi, S. and Delkhosh, M. *Generalized Lagrange Jacobi-Gauss-Lobatto vs Jacobi-Gauss-Lobatto collocation approximations for solving (2+ 1)-dimensional sine-Gordon equations*, Math. Methods Appl. Sci. 43(4) (2020), 2001–2019.

[20] Lui, S. and Nataj, S. *Spectral collocation in space and time for linear PDEs*, J. Comput. Phys. 424 (2020), 109–843.

[21] Lyu, P. and Vong, S. *A nonuniform L2 formula of Caputo derivative and its application to a fractional Benjamin–Bona–Mahony-type equation with nonsmooth solutions*, Numer. Methods Partial. Differ. Equ. 36 (3) (2020), 579–600.

[22] Nikan, O., Avazzadeh, Z. and Machado, J.T. *Numerical investigation of fractional nonlinear sine-Gordon and Klein-Gordon models arising in relativistic quantum mechanics*, Eng. Anal. Bound. Elem. 120 (2020), 223–237.

[23] Parand, K. and Delkhosh, M. *Operational matrices to solve nonlinear Riccati differential equations of arbitrary order*, St. Petersburg Polytechnical University Journal: Physics and Mathematics 3 (3) (2017), 242–254.

[24] Podlubny, I. *Fractional differential equations: an introduction to fractional derivatives, fractional differential equations, to methods of their solution and some of their applications*, Elsevier, 1998.

[25] Saeedi, H. *A fractional-order operational method for numerical treatment of multi-order fractional partial differential equation with variable coefficients*, SeMA J. 75(3) (2018), 421–433.

[26] Saeedi, H. and Chuev, G.N. *Triangular functions for operational matrix of nonlinear fractional Volterra integral equations*, J. Appl. Math. Comput. 49(1-2) (2015), 213–232.

[27] Sakamoto, K. and Yamamoto, M. *Initial value/boundary value problems for fractional diffusion-wave equations and applications to some inverse problems*, J. Math. Anal. Appl. 382 (1) (2011), 426–447.

[28] Salehi, F., Saeedi, H. and Mohseni Moghadam, M. *Discrete Hahn polynomials for numerical solution of two-dimensional variable-order fractional Rayleigh–Stokes problem*, Comput. Appl. Math. 37 (4) (2018), 5274–5292.

[29] Sheng, C., Shen, J., Tang, T., Wang, L. and Yuan, H. *Fast Fourier-like mapped Chebyshev spectral-Galerkin methods for PDEs with integral fractional Laplacian in unbounded domains*, SIAM J. Numer. Anal. 58 (5) (2020), 2435–2464.

[30] Tarasov, V.E. *Mathematical economics: Application of fractional calculus*, Mathematics 8(5) (2020), 660.

[31] Yang, Z., Liu, F., Nie, Y. and Turner, I. *An unstructured mesh finite difference/finite element method for the three-dimensional time-space fractional Bloch-Torrey equations on irregular domains*, J. Comput. Phys. 408 (2020), 109–284.

[32] Zaky, M.A. *Recovery of high order accuracy in Jacobi spectral collocation methods for fractional terminal value problems with nonsmooth solutions*, J. Comput. Appl. Math. 357 (2019), 103–122.

[33] Zaky, M.A. and Ameen, I.G. *A priori error estimates of a Jacobi spectral method for nonlinear systems of fractional boundary value problems and related Volterra-Fredholm integral equations with smooth solutions*, Numer. Algorithms 84(1) (2020), 63–89.

[34] Zaky, M.A. and Hendy, A.S. *Convergence analysis of an L1-continuous Galerkin method for nonlinear time-space fractional Schrödinger equations*, Int. J. Comput. Math. 98(7) (2021), 1420–1437.

[35] Zaky, M.A., Hendy, A.S. and Macías-Díaz, J.E. *Semi-implicit Galerkin–Legendre spectral schemes for nonlinear time-space fractional diffusion–reaction equations with smooth and nonsmooth solutions*, J. Sci. Comput. 82 (1) No. 13 (2020), 1–27.

[36] Zaslavsky, G.M. *Chaos, fractional kinetics, and anomalous transport*, Phys. Rep. 371 (6) (2002), 461–580.

[37] Zayernouri, M., Ainsworth, M. and Karniadakis, G.E. *A unified Petrov–Galerkin spectral method for fractional PDEs*, Comput. Methods Appl. Mech. Eng. 283 (2015), 1545–1569.

[38] Zayernouri, M. and Karniadakis, G.E. *Fractional Sturm–Liouville eigenproblems: theory and numerical approximation*, J. Comput. Phys. 252 (2013), 495–517.

How to cite this article

Mollahasani, S., A shifted fractional-order Hahn functions Tau method for time-fractional PDE with nonsmooth solution. *Iran. J. Numer. Anal. Optim.*, 2023; 13(4): 672-694. https://doi.org/10.22067/ijnao.2023.81716.1238

Iranian Journal of Numerical Analysis and Optimization
Vol. 13, No. 4, 2023, pp 695–710
https://doi.org/10.22067/ijnao.2023.81548.1230
https://ijnao.um.ac.ir/

Research Article

Numerical solution of fractional Bagley–Torvik equations using Lucas polynomials

M. Askari

Abstract

The aim of this article is to present a new method based on Lucas polynomials and residual error function for a numerical solution of fractional Bagley–Torvik equations. Here, the approximate solution is expanded as a linear combination of Lucas polynomials, and by using the collocation method, the original problem is reduced to a system of linear equations. So, the approximate solution to the problem could be found by solving this system. Then, by using the residual error function and approximating the error function by utilizing the same approach, we achieve more accurate results. In addition, the convergence analysis of the method is investigated. Numerical examples demonstrate the validity and applicability of the method.

AMS subject classifications (2020): Primary 45D05; Secondary 42C10, 65G99.

Keywords: Fractional Bagley–Torvik equation; Caputo derivative; Lucas polynomials; Residual error function; Convergence analysis.

1 Introduction

Fractional differential equations have important rules in many fields of science and engineering. For example, in viscoelasticity [4, 3], economic growth model and finance [5, 16], biology [24], control theory [8, 14, 20, 21], dynamics of particle [29], electrical circuits [8], and vibration [25], some issues can be modeled as fractional differential equations.

Received 11 March 2021; revised 7 July 2022; accepted 1 August 2022

Maysam Askari
Department of Mathematics, Professor Hesabi Branch, Islamic Azad University, Tafresh, Iran.
e-mail: maysam.askari@gmail.com, Maysam.Askari@iau.ac.ir

The fractional Bagley–Torvik equation was originally introduced in 1983 to describe the motion of an immersed plate in a Newtonian fluid [30] as

$$m\frac{d^2}{dx^2}U(x) + 2A\sqrt{\eta r}\frac{d^{\frac{3}{2}}}{dx^{\frac{3}{2}}}U(x) + cU(x) = 0,$$

where m, and A are the mass and area of the plate, respectively, r is the fluid density, c is the spring of stiffness, and η is viscosity. The solution of the Bagley–Torvik equation has been studied by researchers for the past two decades. In [26] authors applied the Adomian decomposition method for the solution of Bagley–Torvik equation. El-Gamel and Abd-El-Hadi [9] presented the Legendre-collocation method to approximate the solution of fractional Bagley–Torvik equations. Zolfaghari et al. [34] studied an application of the enhanced homotopy perturbation method to find the approximate solution of Bagley–Torvik equation. In [32], an integral transform method is considered for solving Bagley–Torvik equation. Srivastava, Shah, and Abass [28] proposed a numerical method for studying Bagley–Torvik equations based on the Gegenbauer wavelet together with block pulse function. In [10], authors presented Chelyshkov–Tau as an effective tool for solving Bagley–Torvik equation. Cenesiz, Keskin, and Kurnaz [6] solved Bagley–Torvik equations by using the generalized Taylor collocation method. Authors of [15] utilized hybrid functions approximation, which consists of the block pulse function and Bernoulli polynomials, for the numerical solution of Bagley–Torvik equations. Zahra and Van Daele [33] used a discrete spline function and nonstandard Grunwald–Letnikov and weighted and shifted Grunwald–Letnikov difference operators to propose the solution to Bagley–Torvik equations. El-Gamel and Abd-El-Hadi [9], by using Legendre basis functions, reduced Bagley–Torvik equation to a system of linear equations and by solving this system presented a numerical solution to the Bagley–Torvik equation. Authors of [27] introduced the numerical solution of Bagley–Torvik based on reproducing kernel Hilbert space. In [31], generalized Bessel functions of the first kind are applied for the numerical solution of the fractional Bagley–Torvik equation.

The outlines of the article are as follows: In section 2, we briefly introduce the Caputo fractional derivative, Fibonacci, and Lucas polynomials and describe their properties. In section 3, we construct a numerical method for a solution of fractional Bagley–Torvik equations using Lucas polynomials and residual error function. In section 4, the convergence analysis of the proposed method is studied. The numerical results for some problems are given in section 5, and at the end, we have a brief conclusion.

2 Basic definitions and requirements

Definition 1. If $\alpha > 0$, then the Caputo fractional derivative operator of order α is defined as

$$D^{\alpha}f(x) = \frac{1}{\Gamma(m-\alpha)}\int_0^x (x-t)^{(m-\alpha-1)}f^{(m)}(t)d_t,$$

where $m-1 < \alpha \leq m$.

The Caputo derivative has linear property and

$$D^{\alpha}(c) = 0, \quad c \text{ is a constant.}$$

$$D^{\alpha}(x^k) = \begin{cases} \frac{\Gamma(k+1)}{\Gamma(k-\alpha+1)}x^{k-\alpha} & \text{if } k \geq \lceil \alpha \rceil,\ k \in \mathbb{N}, \\ 0 & \text{if } k < \lceil \alpha \rceil,\ k \in \mathbb{N}_0, \end{cases}$$

where $\mathbb{N}_0 = \{0, 1, 2, \ldots\}$ and $\mathbb{N}$ is the set of natural numbers.

The Fibonacci polynomials $F_n(x)$ and and Lucas polynomials $L_n(x)$ are defined by recursive relations as

$$\begin{aligned} &F_0(x) = 0, \quad F_1(x) = 1, \\ &F_n(x) = xF_{n-1}(x) + F_{n-2}(x), \quad n \geq 2, \end{aligned}$$

and

$$\begin{aligned} &L_0(x) = 2, \quad L_1(x) = x, \\ &L_n(x) = xL_{n-1}(x) + L_{n-2}(x), \quad n \geq 2, \end{aligned}$$

respectively. Here we remark that Fibonacci and Lucas polynomials are the special case of Chebyshev polynomials (see [22]). Lucas polynomials have explicit form as

$$L_n(x) = \sum_{i=0}^{\lfloor \frac{n}{2} \rfloor} \frac{n}{n-i}\binom{n-i}{i}x^{n-2i}, \quad n \geq 1,$$

where $\lfloor x \rfloor$ is the largest integer less or equal to x. According to [18], the first derivative of Lucas polynomials can be evaluated using the Fibonacci polynomials as

$$L_n'(x) = nF_n(x). \tag{1}$$

Continuing this approach by repeating derivation on both sides of (1) gives

$$L_n^{(k)}(x) = nF_n^{(k-1)}(x), \quad k \geq 2.$$

If $u(x)$ is a continuous function, then we can approximate $u(x)$ by a linear combination of Lucas polynomials as

$$u(x) \approx \sum_{j=0}^{m} a_j L_j(x) = \mathbf{L}(x)\mathbf{A},$$

where $\mathbf{L} = [L_0(x), L_1(x), \ldots, L_m(x)]$ and $\mathbf{A} = [a_0, a_1, \ldots, a_m]^T$. Moreover, for the kth derivation of $u(x)$, we have the approximation

$$u^{(k)}(x) \approx \sum_{j=0}^{m} a_j L_j^{(k)}(x).$$

Therefore, as mentioned in [11], the approximation for the kth ($k \geq 2$) derivation of $u(x)$ can be formulated as

$$u^{(k)}(x) \approx n\mathbf{F}(x)D^{k-1}\mathbf{A},$$

where

$$\mathbf{F} = [F_0(x), F_1(x), \ldots, F_m(x)],$$

$$\mathbf{D}_{(m+1)\times(m+1)} = \begin{pmatrix} 0 \ldots 0 \\ \vdots \ \ \mathbf{d} \\ 0 \end{pmatrix},$$

and $\mathbf{d}$ is an $m \times m$ matrix, which is defined as

$$\mathbf{d}_{i,j} = \begin{cases} i \ \sin \frac{(j-i)\pi}{2} & \text{if } j > i, \\ 0 & \text{if } j \leq i. \end{cases}$$

Further details about Lucas polynomials and application of Lucas polynomials for solving problems arising in engineering, such as ordinary and partial differential equations, can be found in [1, 2, 7, 12, 11, 13, 18, 19].

3 Construction of method

Consider the fractional Bagley–Torvik equation

$$A \ D^2 f(x) + B \ D^{\frac{3}{2}} f(x) + C \ f(x) = g(x), \quad x \in [0, 1], \tag{2}$$

with initial conditions

$$f(0) = f_0, \quad f'(0) = f'_0,$$

or boundary conditions

$$f(0) = f_0, \quad f(1) = f_1.$$

By using the Caputo fractional derivation, (2) can be rewritten as

$$A \ D^2 f(x) + \frac{B}{\Gamma(\frac{1}{2})} \int_0^x (x-t)^{-\frac{1}{2}} f''(t) d_t + C \ f(x) = g(x). \tag{3}$$

Let the approximate estimation for the solution of (2) have the following form:

$$f(x) \approx \sum_{j=0}^{M} \alpha_j L_j(x). \tag{4}$$

Now, by collocating at the nodes $\{x_i : i = 1, \dots, M-1\}$, where $0 < x_1 < \cdots < x_{M-1} < 1$, and utilizing (3), we get

$$A \sum_{j=0}^{M} \alpha_j L_j''(x_i) + \frac{B}{\Gamma(\frac{1}{2})} \sum_{j=0}^{M} \alpha_j \int_0^{x_i} (x_i - t)^{-\frac{1}{2}} L_j''(t) d_t + C \sum_{j=0}^{M} \alpha_j L_j(x_i) = g(x_i). \tag{5}$$

Also, initial and boundary conditions lead to

$$\sum_{j=0}^{M} \alpha_j L_j(0) = f_0, \quad \sum_{j=0}^{M} \alpha_j L_j'(0) = f_0', \tag{6}$$

and

$$\sum_{j=0}^{M} \alpha_j L_j(0) = f_0, \quad \sum_{j=0}^{M} \alpha_j L_j(1) = f_1, \tag{7}$$

respectively. Hence, the combination of (5) together with (6) or (7) gives a system of linear equations as

$$\mathbf{U}\lambda = \mathbf{b},$$

where, for initial conditions,

$$\mathbf{b} = [g(x_1), \dots, g(x_{M-1}), f_0, f_0']^T,$$

$$\mathbf{U}_{i,j} = \begin{cases} AL_{j-1}''(x_i) + \frac{B}{\Gamma(\frac{1}{2})} \int_0^{x_i} (x_i - t)^{-\frac{1}{2}} L_{j-1}''(t) d_t + CL_{j-1}(x_i) & \\ & \text{if } 1 \le i \le M-1, \\ L_{j-1}(0) & \text{if } i = M, \\ L_{j-1}'(0) & \text{if } i = M+1, \end{cases}$$

and for boundary conditions,

$$\mathbf{b} = [g(x_1), \dots, g(x_{M-1}), f_0, f_1]^T,$$

$$\mathbf{U}_{i,j} = \begin{cases} AL_{j-1}''(x_i) + \frac{B}{\Gamma(\frac{1}{2})} \int_0^{x_i} (x_i - t)^{-\frac{1}{2}} L_{j-1}''(t) d_t + CL_{j-1}(x_i) & \\ & \text{if } 1 \le i \le M-1, \\ L_{j-1}(0) & \text{if } i = M, \\ L_{j-1}(1) & \text{if } i = M+1. \end{cases}$$

For example, if $M = 2$, by using Chebyshev–Gauss–Lobatto nodes, U has the following structure:

$$U_I = \begin{bmatrix} 2C & \frac{C}{2} & 2A + 2\sqrt{2}\frac{B}{\sqrt{\pi}} + 9\frac{C}{4} \\ 2 & 0 & 2 \\ 0 & 1 & 0 \end{bmatrix}$$

and

$$U_B = \begin{bmatrix} 2C & \frac{C}{2} & 2A + 2\sqrt{2}\frac{B}{\sqrt{\pi}} + 9\frac{C}{4} \\ 2 & 0 & 2 \\ 2 & 1 & 3 \end{bmatrix}.$$

Therefore, by solving the obtained system of linear equations, the approximate solution of the fractional Bagley–Torvik equation is determined. Here, we present a more accurate method using the residual error function [7, 17] for the solution of the fractional Bagley–Torvik equation. If we display the error of approximation (4), as

$$e(x) = f(x) - \sum_{j=0}^{M} \alpha_j L_j(x),$$

then the error function satisfies the fractional differential equation

$$A\ D^2 e(x) + B\ D^{\frac{3}{2}} e(x) + C\ e(x) = R(x), \tag{8}$$

where

$$R(x) = g(x) - A \sum_{j=0}^{M} \alpha_j L_j''(x) - \frac{B}{\Gamma(\frac{1}{2})} \sum_{j=0}^{M} \alpha_j \int_0^x (x-t)^{-\frac{1}{2}} L_j''(t) d_t \tag{9}$$

$$- C \sum_{j=0}^{M} \alpha_j L_j(x). \tag{10}$$

The above fractional differential equation is accompanied with initial conditions

$$e(0) = e'(0) = 0, \tag{11}$$

or the boundary conditions

$$e(0) = e(1) = 0. \tag{12}$$

Now, we propose the approximate solution to (8)–(12) using Lucas polynomials as

$$e(x) \approx \sum_{j=0}^{N} \beta_j L_j(x), \quad N > M.$$

By using the idea described above, we can get the an approximation for error function $e(x)$. So, we obtain a better approximation

$$f(x) \approx \sum_{j=0}^{M} \alpha_j L_j(x) + \sum_{j=0}^{N} \beta_j L_j(x)$$

for the numerical solution of the fractional Bagley–Torvik equation.

4 Convergence analysis

In this section, we argue about the convergence of the proposed method. For this aim, first, some requirements are given.

Lemma 1. [1] Assume that $f(x)$ is an infinitely differentiable at $x = 0$. Then $f(x)$ can be represented by using Lucas polynomials as

$$f(x) = \sum_{i=0}^{\infty} \sum_{j=0}^{\infty} \frac{(-1)^j \delta_i f^{(i+2j)}}{j!(i+j)!} L_i(x),$$

where

$$\delta_i = \begin{cases} \frac{1}{2} & \text{if } i = 0, \\ 1 & \text{if } i \neq 0. \end{cases}$$

Lemma 2. [1] For every $i \geq 0$, the Lucas polynomials can be bounded as

$$|L_i(x)| \leq 2\sigma^i,$$

where σ is the golden ratio.

Theorem 1. [1] Let $f(x)$ be defined on $[0, 1]$, and there is a positive constant A such that $|f^{(i)}(0)| \leq A^i$, $i \geq 0$. Moreover, suppose that $f(x)$ has a representation

$$f(x) = \sum_{i=0}^{\infty} c_i L_i(x), \tag{13}$$

and that $e(x) = \sum_{i=M+1}^{\infty} c_i L_i(x)$ is an error of approximation function $f(x)$ by Lucas polynomials of degree M. Then

$$|c_i| \leq \frac{A^i \cosh(2A)}{i!}$$

and the series (13) is convergent and

$$|e(x)| < \frac{2e^{A\sigma} \cosh(2A)(A\sigma)^{M+1}}{(M+1)!}.$$

In the following theorem, we discuss the convergence of the presented method of the previous section.

Theorem 2. Let $f(x)$ be an infinitely differentiable at $x = 0$, and there is a constant $A > 0$ such that $|f^{(i)}(0)| \leq A^i$, $i > 0$. If $e(x)$ is defined as $e(x) = f(x) - \sum_{i=0}^{M} a_i L_i(x)$ and has a representation

$$e(x) = \sum_{i=0}^{\infty} b_i L_i(x),$$

then the proposed method has the error estimation

$$|E(x)| < \frac{2e^{\mathcal{A}\sigma} \cosh(2\mathcal{A})(\mathcal{A}\sigma)^{N+1}}{(N+1)!}.$$

Proof. According to the previous section, the approximation

$$f(x) \approx \sum_{i=0}^{M} a_i L_i(x) + \sum_{i=0}^{N} b_i L_i(x)$$

] has the error $E(x) = e(x) - \sum_{i=0}^{N} b_i L_i(x)$. Also,

$$e^{(i)}(x) = f^{(i)}(x) - \sum_{j=0}^{M} a_j L_j^{(i)}(x).$$

Since $L_j(x)$ is a polynomial of degree j, so $L_j^{(i)}(x)$ has the following representation:

$$L_j^{(i)}(x) = \begin{cases} \alpha_{j_0} + \alpha_{j_1} x + \cdots + \alpha_{j_{j-i}} x^{j-i} & \text{if } j \geq i, \\ 0 & \text{if } j < i. \end{cases}$$

Therefore

$$L_j^{(i)}(0) = \begin{cases} \alpha_{j_0} & \text{if } j \geq i, \\ 0 & \text{if } j < i. \end{cases}$$

If we set $P = \max\{|L_j^{(i)}(0)| : \ i, j = 0, \ldots, M\}$, then for $i = 1, \ldots, M$ by using Theorem 1, we get

$$\begin{aligned} |e^{(i)}(0)| &< A^i + \sum_{j=0}^{M} \frac{A^j \cosh(2A) P}{j!} \\ &< (A + \cosh(2A) P e^A)^i. \end{aligned}$$

Moreover, for $i \geq M + 1$, we have $|e^{(i)}(0)| \leq A^i$. If we apply Theorem 1 for the function $e(x) = f(x) - \sum_{i=0}^{M} a_i L_i(x)$, then

$$|E(x)| \leq \frac{2e^{\mathcal{A}\sigma} \cosh(2\mathcal{A})(\mathcal{A}\sigma)^{N+1}}{(N+1)!},$$

where

$$\mathcal{A} = A + \cosh(2A)Pe^{A}.$$

□

5 Numerical results

In this section, some examples are presented to show the accuracy of the proposed method. These examples consist of initial and boundary conditions. Also, to show the accuracy and validity of the proposed method, we have a comparison between our approach and a number of other methods. In computations, we utilize Chebyshev–Gauss–Lobatto nodes as collocation points, and all of the computations have been performed in MAPLE 18 software.

Example 1. Consider fractional Bagley–Torvik equation

$$D^2 f(x) + D^{\frac{3}{2}} f(x) + f(x) = x^3 + 7x + 1 + \frac{8x^{\frac{3}{2}}}{\sqrt{\pi}}$$

with the initial conditions $f(0) = 1$ and $f'(0) = 1$. This problem has the exact solution $f(x) = x^3 + x + 1$. Here we take $M = 6$ and $N = 10$. We compare the Lucas collocation method (LCM) and Lucas collocation method combined with residual error function (LCM-REF) with the Chelyshkov–Tau method [10] and Legendre collocation method [9]. Results are given in Table 1. Absolute errors of LCM and LCM-REF are listed in Table 2 and plotted in Figure 1.

Table 1: Comparisons of the presented methods for Example 1

x	Exact solution	LCM-REF	LCM	Chelyshkov–Tau [10]	Legendre collocation [9]
0.1	1.101000	1.101000	1.101000	1.101000	1.101000
0.25	1.265625	1.265625	1.265625	1.265625	1.265625
0.5	1.625000	1.625000	1.625000	1.625000	1.625000
0.75	2.171875	2.171875	2.171875	2.171875	2.171875
1	3.000000	3.000000	3.000000	3.000000	3.000002

Table 2: Absolute errors of the presented methods for Example 1

x	0.1	0.3	0.5	0.7	0.9
LCM	8.90000E-48	1.22000E-47	1.67000E-47	2.23000E-47	3.28000E-47
LCM-REF	8.72634E-48	1.06045E-47	1.19949E-47	1.30881E-47	1.39519E-47

Example 2. In this example, we study the fractional Bagley–Torvik equation

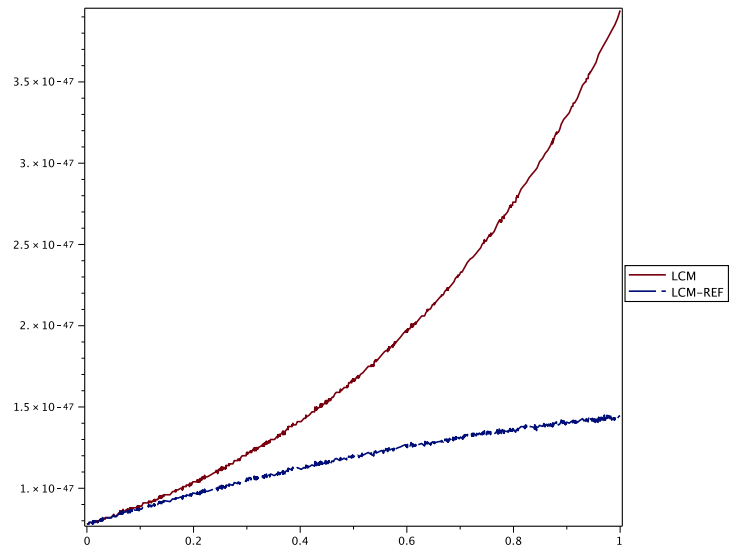


Figure 1: The plot of absolute errors of LCM and LCM-REF for Example 1

$$D^2 f(x) + \frac{8}{17} D^{\frac{3}{2}} f(x) + \frac{13}{51} f(x) = \frac{x^{-\frac{1}{2}}}{89250\sqrt{\pi}} (48p(x) + 7\sqrt{x}q(x)),$$

where

$$\begin{aligned} p(x) =& 16000x^4 - 32480x^3 + 21280x^2 - 4746x + 189, \\ q(x) =& 3250x^5 - 9425x^4 + 264880x^3 - 44, \end{aligned}$$

with the boundary conditions $f(0) = 0,\ f(1) = 0$. This problem has the exact solution

$$f(x) = x^5 - \frac{29}{10}x^4 + \frac{76}{25}x^3 - \frac{339}{250} + \frac{27}{125}x.$$

We examine the proposed method with $M = 6,\ N = 10$. In Table 3, a comparison between absolute errors of Lucas collocation method combined with residual error function (LCM-REF), Chelyshkov-Tau method [10], Harr wavelets method [23], and Bessel collocation method [31] is given. In Figure 2, the plot of the exact solution and approximate solution, which is obtained by the combination of the Lucas collocation method and residual error function, is displayed.

Example 3. Consider the fractional Bagley–Torvik equation

$$A\ D^2 f(x) + B\ D^{\frac{3}{2}} f(x) + C\ f(x) = g(x),$$

Table 3: Comparisons of LCM-REF for Example 2

x	LCM-REF	Chelyshkov–Tau [10]	Harr wavelets [23]	Bessel collocation [31]
0.1	2.69915E-48	5.92720E-14	6.49908E-7	1.0800E-2
0.2	3.21281E-48	1.18400E-13	6.35657E-7	8.9595E-3
0.3	3.72702E-48	1.77249E-13	3.71584E-7	3.7797E-3
0.4	4.23527E-48	2.35568E-13	9.48220E-7	1.4413E-7
0.5	4.71891E-48	2.17578E-13	1.59573E-6	1.0001E-3
0.6	5.20793E-48	2.92504E-13	1.05494E-6	6.6150E-8
0.7	5.69369E-48	3.82671E-13	6.34678E-7	1.2599E-3
0.8	6.20644E-48	3.82256E-13	1.88690E-6	1.2800E-3
0.9	6.54825E-48	2.90107E-13	3.13999E-6	2.0656E-8

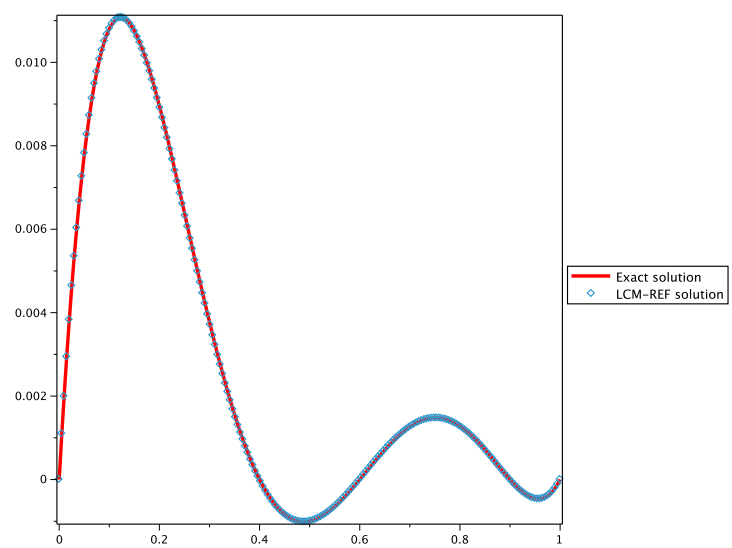


Figure 2: The plot of exact and LCM-REF solutions for Example 2

with the initial conditions $f(0) = 0,\ f'(0) = 0$. This problem has the exact solution

$$f(x) = \int_0^x G_3(x-\tau)g(\tau)d_\tau = \frac{1}{A}\sum_{k=0}^{\infty}\frac{(-1)^k}{k!}(\frac{C}{A})^k x^{2k+1}E^{(k)}_{\frac{1}{2},2+\frac{3k}{2}}(-\frac{B}{A}\sqrt{x}),$$

where $G_3(x)$ is three-term Green's function, which is defined as

$$G_3(x) = \frac{1}{A}\sum_{k=0}^{\infty}\frac{(-1)^k}{k!}(\frac{C}{A})^k x^{2k+1}E^{(k)}_{\frac{1}{2},2+\frac{3k}{2}}(-\frac{B}{A}\sqrt{x}),$$

and $E_{\lambda,\mu}$ is the Mittag-Leffler function with two parameters λ and μ, and

$$E^{(k)}_{\lambda,\mu}(y) = \sum_{j=0}^{\infty} \frac{(j+k)!y^j}{j!\Gamma(\lambda j + \lambda k + \mu)}, \quad k = 0, 1, 2, \dots \ .$$

Let $A = 1$, $B = \frac{1}{2}$, $C = \frac{1}{2}$, and $g(x) = 8$. For this case, we choose $M = 30$ and $N = 40$. Numerical comparisons of the proposed methods with Chelyshkov–Tau method [10], Legendre collocation method [9], and generalized Taylor collocation method [6] are listed in Table 4. In Table 5, absolute errors of LCM and LCM-REF are displayed. Figure 3 exhibits the comparison of analytical and LCM-REF solutions of this example. The plot of absolute errors of LCM and LCM-REF is illustrated in Figure 4.

Table 4: Comparisons of the presented methods for Example 3

x	Exact solution	LCM-REF	LCM	Chelyshkov-Tau [10]	Taylor-collocation [6]	Legendre collocation [9]
0.1	0.036487	0.036486	0.036483	0.036453	0.036485	0.036471
0.2	0.140639	0.140636	0.140632	0.140575	0.140634	0.140615
0.3	0.307484	0.307480	0.307473	0.307403	0.307476	0.307434
0.4	0.533284	0.533278	0.533269	0.533252	0.533271	0.533225
0.5	0.814756	0.814749	0.814739	0.814860	0.814735	0.814661
0.6	1.148837	1.148828	1.148816	1.149069	1.148805	1.148733
0.7	1.532565	1.532555	1.532541	1.532870	1.532521	1.532424
0.8	1.963029	1.963018	1.963002	1.963440	1.962974	1.962874
0.9	2.437334	2.437322	2.437305	2.437829	2.437455	2.437134

Table 5: Absolute errors of the presented methods for Example 3

x	0.1	0.3	0.5	0.7	0.9
LCM	4.23327E-6	1.16836E-5	1.82885E-5	2.41776E-5	2.93789E-5
LCM-REF	1.78503E-6	4.92605E-6	7.71291E-6	1.01973E-5	1.23905E-5

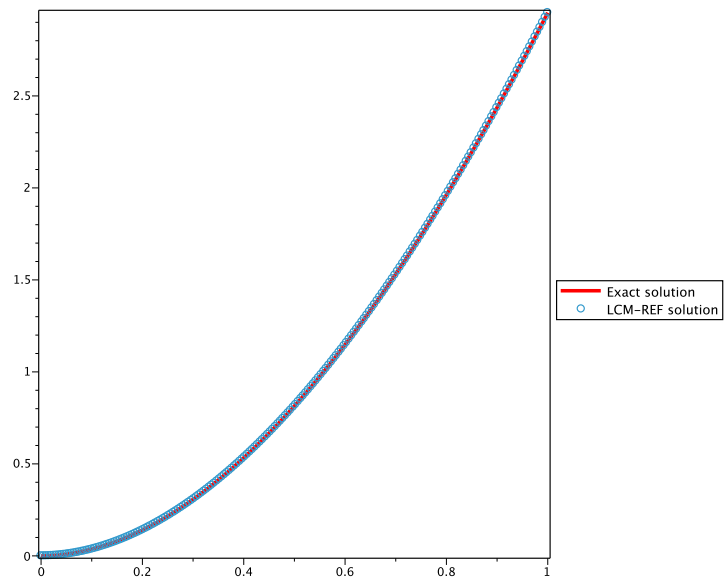


Figure 3: The plot of exact and LCM-REF solutions for Example 3

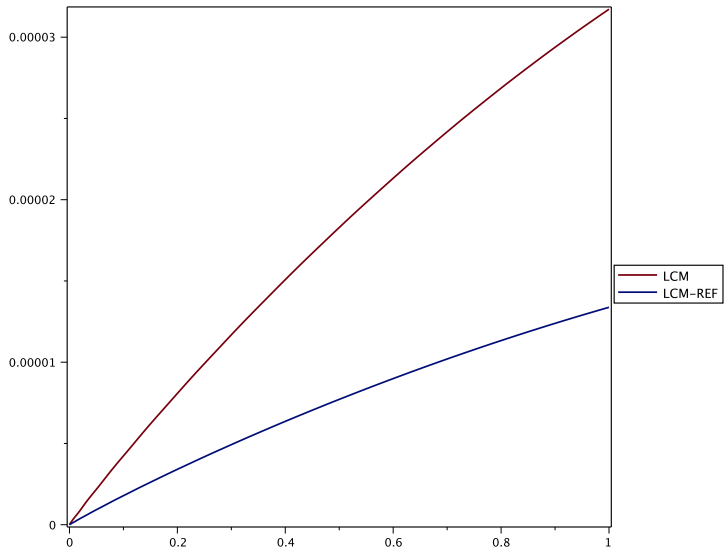


Figure 4: Absolute errors of LCM and LCM-REF for example 3

6 Conclusion

A new numerical method using Lucas polynomials was proposed to solve the fractional Bagley–Torvik equation. In this approach, we expanded the exact solution as a finite linear combination of Lucas polynomials. Then, by using Chebyshev–Gauss–Lobatto nodes as collocation points, the approximate solution was obtained. To improve the results, we applied the residual error function, and the error function was estimated by Lucas polynomials. So we can improve the results and get a more accurate approximation. Numerical tests and comparisons with other numerical methods indicated that this method is reliable and has acceptable accuracy.

References

[1] Abd-Elhameed, W.M. and Youssri, Y.H. *Spectral solutions for fractional differential equations via a novel Lucas operational matrix of fractional derivatives*, Rom. J. Phys. 61 (2016), 795–813.

[2] Ali, I., Haq, S., Nisar, K.S. and Baleanu, D. *An efficient numerical scheme based on Lucas polynomials for the study of multidimensional Burgers-type equations*, Adv. Differ. Equ. (2021), 1–24.

[3] Bagley, R.L. and Torvik, P.J. *A theoretical basis for the application of fractional calculus to viscoelasticity*, J. Rheology. 27 (1983), 201–210.

[4] Bagley, R.L. and Torvik, P.J. *Fractional calculus in the transient analysis of viscoelastically damped structures*, AIAA J. 23 (1985), 918–925.

[5] Baillie, R.T. *Long memory processes and fractional integration in econometrics*, J. Econometrics. 73 (1996), 5–59.

[6] Cenesiz, Y., Keskin, Y. and Kurnaz, A. *The solution of the Bagley–Torvik equation with the generalized Taylor collocation method*, J. Franklin Inst. 347 (2010), 452–466.

[7] Cetin, M., Sezer, M. and Guler, C. *Lucas polynomial approach for system of high-order linear differential equations and residual error estimation*, Math. Probl. Eng. (2015), Art. ID 625984, 14.

[8] Debnath, L. *Recent applications of fractional calculus to science and engineering*, Int. J. Math. Math. Sci. 54 (2003), 3413–3442.

[9] El-Gamel, M. and Abd-El-Hadi, M. *Numerical solution of the Bagley–Torvik equation by Legendre-collocation method*, SeMA J. 74 (2017), 371–383.

[10] El-Gamel, M., Abd-El-Hadi, M. and El-Azab, M. *Chelyshkov-Tau approach for solving Bagley–Torvik equation*, Appl. Math. 8 (2017), 1795–1807.

[11] Filipponi, P. and Horadam, A.F. *Derivative sequences of Fibonacci and Lucas polynomials*, Applications of Fibonacci numbers 4 (Winston-Salem, NC, 1990), 99–108, Kluwer Acad. Publ., Dordrecht, 1991.

[12] Filipponi, P. and Horadam, A.F. *Second derivative sequences of Fibonacci and Lucas polynomials*, Fibonacci Quart. 31 (1993), 194–204.

[13] Koshy, T. *Fibonacci and Lucas Numbers with Applications*, Wiley, New York, USA, 2001.

[14] Lin, W. *Global existence theory and chaos control of fractional differential equations*, J. Math. Anal. Appl. 332 (2007), 709–726.

[15] Mashayekhi, S. and Razzaghi, M. *Numerical solution of the fractional Bagley–Torvik equation by using hybrid functions approximation*, Math. Meth. Appl. Sci. 39 (2016), 353–365.

[16] Ming, H., Wang, J.R. and Feckan, M. *The application of fractional calculus in Chinese economic growth models*, Mathematics, 7(8) (2019), 665.

[17] Oliveira, F.A. *Collocation and residual correction*, Numerische Mathematik 36 (1980), 27–31.

[18] Oruc, O. *A new algorithm based on Lucas polynomials for approximate solution of 1D and 2D nonlinear generalized Benjamin-Bona-Mahony Burgers equation*, Comp. Math. Appl. (2017).

[19] Oruc, O. *A new numerical treatment based on Lucas polynomials for 1D and 2D sinh-Gordon equation*, Commun. Nonlinear Sci. Numer. Simulat. 57 (2018), 14–25.

[20] Oustaloup, A. *Systemes asservis lineaires d'ordre fractionnaire*, masson, Paris, 1983.

[21] Podlubny, I. *Fractional differential equations*, Academic Press, New York, 1999.

[22] Purav, V.N. *Fibonacci, Lucas and Chebyshev polynomials*, Int. J. Sci. Res. 7 (2018), 1693–1695.

[23] Rehman and Ali Khan, R. *A numerical method for solving boundary value problems for fractional differential equation*, Appl. Math. Model. 36 (2012), 894–907.

[24] Rihan, F.A. *Numerical modeling of fractional-order biological systems*, Abstr. Appl. Anal. 2013 (2013), 1–11.

[25] Rossikhin, Y.A. and Shitikova, M.V. *Applications of fractional calculus to dynamic problems of linear and nonlinear hereditary mechanics of solids*, Appl. Mech. Rev. 50 (1997), 15–67.

[26] Saha Ray, S. and Bera, R.K. *Analytical solution of the Bagley–Torvik equation by Adomian decomposition method*, Appl. Math. Comput. 168 (2005), 398–410.

[27] Sakar, M.G., Saldir, O. and Akgul, A. *A novel technique for fractional Bagley–Torvik equation*, Proc. Natl. Acad. Sci. India Sect. A Phys. Sci. 89 (2019), 539–545.

[28] Srivastava, H.M., Shah, F.A. and Abass, R. *An application of the Genenbauer wavelet method for the numerical solution of the fractional Bagley–Torvik equation*, Russian. J. Math. Phys. 26 (2019), 77–93.

[29] Tarasov, V.E. *Fractional dynamics: applications of fractional calculus to dynamics of particles, fields and media, nonlinear physical science*, Springer, Heidelberg, Germany, 2011.

[30] Torvik, P.J. and Bagley, R.L. *On the apearance of the Fractional derivative in the behavior of real materials*, J. Appl. Mech. 51 (1984), 294–298.

[31] Yuzbasi, S. *Numerical solution of the Bagley–Torvik equation by the Bessel collcation method*, Math. Meth. Appl. Sci. 36 (2012), 300–312.

[32] Zafar, A.A., Kudra, R.G. and Awrejcewicz, J. *An investigation of fractional Bagley–Torvik equation*, Entropy. 22 (1) (2019), 28.

[33] Zahra, W.K. and Van Daele, M. *Discrete spline methods for solving two point fractional Bagley–Torvik equation*, Appl. Math. Comput. 296 (2017), 42–56.

[34] Zolfaghari, M., Ghaderi, R., SheikholEslami, A., Ranjbar, A., Hosseinnia, S.H., Momani, S. and Sadati, J. *Application of the enhanced homotopy perturbation method to solve the fractional-order Bagley–Torvik differential*, Phys. Scr. 2009 (T136) (2009), 014032.

How to cite this article

Askari, M., Numerical solution of fractional Bagley–Torvik equations using Lucas polynomials. *Iran. J. Numer. Anal. Optim.*, 2023; 13(4): 695-710. https://doi.org/10.22067/ijnao.2023.81548.1230

Iranian Journal of Numerical Analysis and Optimization
Vol. 13, No. 4, 2023, pp 711–727
https://doi.org/10.22067/ijnao.2023.83070.1283
https://ijnao.um.ac.ir/

Research Article

Singularly perturbed two-point boundary value problem by applying exponential fitted finite difference method

N. Kumar*, R. Kumar Sinha and R. Ranjan

Abstract

The present study addresses an exponentially fitted finite difference method to obtain the solution of singularly perturbed two-point boundary value problems (BVPs) having a boundary layer at one end (left or right) point on uniform mesh. A fitting factor is introduced in the derived scheme using the theory of singular perturbations. Thomas algorithm is employed to solve the resulting tri-diagonal system of equations. The convergence of the presented method is investigated. Several model example problems are solved using the proposed method. The results are presented with terms of maximum absolute errors, which demonstrate the accuracy and efficiency of the method. It is observed that the proposed method is capable of producing highly accurate results with minimal computational effort for a fixed value of step size h, when the perturbation parameter tends to zero. From the graphs, we also observed that a numerical solution approximates the exact solution very well in the boundary layers for smaller value of ε.

AMS subject classifications (2020): Primary AMS 65L10; Secondary 65L11, 65L12, 65L20.

*Corresponding author

Received 22 June 2023; revised 12 August 2023; accepted 14 August 2023

Narendra Kumar
Department of Mathematics, National Institute of Technology Patna, Patna - 800005, India.
e-mail: narendra.ma18@nitp.ac.in

Rajesh Kumar Sinha
Department of Mathematics, National Institute of Technology Patna, Patna - 800005, India.
e-mail: rajesh@nitp.ac.in

Rakesh Ranjan
Department of Science and Technology, Bihar, Government Polytechnic, Lakhisarai, Lakhisarai- 811311, India.
e-mail: 90.ranjan@gmail.com

Keywords: Singular perturbation problem; Stability and convergence; Finite difference method.

1 Introduction

Singular perturbation problems are of mainly deal in fluid mechanics and other areas of practical/applied mathematics. The solution of the singularly perturbed boundary value problems (BVPs) has a multi-scale nature. The solution varies rapidly in some parts of the domain and varies slowly in some other parts of the domain. The numerical solution of singular perturbation problems (SPPs) is far from trivial, because of the boundary layer behavior of the solution. There are many physical situations in which the sharp changes occur inside the domain of interest, and the narrow regions across which these changes take place are usually referred as Navier–Stokes flow problems, involving high Reynolds number [4, 17, 28], mathematical models of liquid crystal materials and chemical reactions, control theory, and electrical networks [6, 7, 30]. These quick shifts can be managed by fast scales, magnified scales, or stretched scales, but not by slow scales. The domain of integration is typically divided into two subdomains, with a distinct scheme being applied to each subdomain as a common approach to solving this type of problem. In recent years, a large number of analytical methods have been proposed (see [22, 21, 2, 20, 19, 11, 16, 29]. Numerical methods based schemes with and without fitting factors on boundary value techniques and initial value techniques are given in [9, 1, 12, 13, 23, 14]. Phaneendra and Lalu[24] presented Gaussian quadrature for two-point singularly perturbed BVPs with the exponential fitting with a layer at one endpoint, dual boundary layers, and internal boundary layers. In this paper, the given BVP is reduced into an equivalent first-order differential equation with the perturbation parameter as a deviating argument. Then, the Gaussian two-point quadrature technique with exponential fitting is implemented to solve the first-order equation with deviating parameters. Mishra and Saini [18] studied the Liouville–Green transform to solve a singularly perturbed two-point BVP with a right-end boundary layer. Articles [3, 5, 8, 9, 31] proposed different numerical approaches combining fitted mesh methods and fitted operator methods employed by several researchers for solving SPPs, whereas Kadalbajoo and Kumar [10] presented a detailed outline on the numerical methods for solving SPPs. Indeed these existing numerical methods are mostly based on fitted operator techniques or use reasonable theoretical information regarding the solutions, which forms a limitation of these approaches. An efficient method of numerical integration for a class of singularly perturbed two-point BVPs at one endpoint (either left or right) has been presented in [25]. Ranjan, Prasad, and Alam [27] developed a simple method of numerical integration for a class of singularly perturbed two-point BVPs at one endpoint (either left or right). Ranjan and Prasad [26] proposed a fitted finite

difference scheme for solving singularly perturbed two-point BVPs having boundary layer at left or right endpoints. Madhu Latha, Phaneendra, and Reddy [15] presented a numerical solution of SPP using numerical integration with an exponential fitting factor.

In view of the wealth of literature on SPPs, we raise the question of whether there are other ways to attack SPPs, namely ways that are very easy to use and ready for computer implementation. In response to this need for a fresh approach to SPPs, we propose and illustrate in this paper a fitted finite difference technique for singularly perturbed two-point BVPs with a boundary layer on the left (or right) end of the underlying interval. Numerical experience with several linear examples is described.

The paper is organized as follows: Section 2 presents the description of the presented new effective method to solve a second-order singularly perturbed two-point BVP. In Section 3, the convergence of the presented method is investigated. To demonstrate the accuracy and efficiency of the presented method, numerical experiments are carried out for several model test problems, and the results are shown in tables in Section 4. Finally, the discussions and conclusions are presented in the last section 5.

2 Statement of the problems

Consider the singularly perturbed two-point BVPs of the following type:

$$\varepsilon\, v''(t) + r(t)v'(t) + s(t)v(t) = \psi(t) \text{ on } \Omega = [0,\, 1], \tag{1}$$

subject to the boundary conditions and interval conditions,

$$v(0) = \alpha, \quad v(1) = \beta, \tag{2}$$

where ε is a small positive perturbation parameter ($0 < \varepsilon << 1$). Furthermore, the functions $r(t)$, $s(t)$, and $\psi(t)$ are continuously differentiable functions in $[0, 1]$, where α and β are constant. In the scenario where we assume that $r(t) \geq M > 0$ holds true for the entire interval $[0, 1]$, with M representing a positive constant, the boundary layer is expected to occur in the vicinity of $t = 0$. On the other hand, if we consider that $r(t) \leq M < 0$ holds throughout the interval $[0, 1]$, with M being a negative constant, then the boundary layer is anticipated to be located near $t = 1$.

2.1 Description of the method for left-end boundary layer problems

In this subsection, we describe the proposed method for the solution of the problem (1)–(2) having boundary layer at left-end point of the interval considered.

The solution of (1) with (2) is of the following form (see . 22–261[22]):

$$v(t) = v_0(t) + \frac{r(0)}{r(t)}\left(\alpha_0 - v_0(0)\right) e^{-\int_0^t \left(\frac{r(t)}{\varepsilon} - \frac{s(t)}{r(t)}\right)dt} + o(\varepsilon), \tag{3}$$

where $v_0(t)$ denotes the simplified problem's solution:

$$r(t)v_0'(t) + s(t)v_0(t) \;=\; \psi(t), \qquad v_0(1) = \beta. \tag{4}$$

By considering the Taylor series expansions of $r(t)$ and $s(t)$ around the point $t = 0$ up to their respective first terms, we can simplify (3) as follows:

$$v(t) = v_0(t) + \left(\alpha_0 - v_0(0)\right) e^{-\left(\frac{r(0)}{\varepsilon} - \frac{s(0)}{r(0)}\right)t} + o(\varepsilon). \tag{5}$$

Taking the limit as $h \to 0$ and applying (3) to the point $t = t_i = ih$, $i = 0, 1, 2, \ldots, N$, we obtain

$$\lim_{h\to 0} v(ih) = v_0(0) + \left(\alpha_0 - v_0(0)\right) e^{-\left(\frac{r^2(0) - \varepsilon s(0)}{r(0)}\right)i\rho} + o(\varepsilon), \tag{6}$$

where $\rho = h/\varepsilon$, the first and second-order approximations have been used as below:

$$v_i' = \frac{3v_{i+1} - 2v_i - v_{i-1}}{4h}, \tag{7}$$

$$v_i'' = \frac{v_{i+1} - 2v_i + v_{i-1}}{h^2}. \tag{8}$$

Substituting (7) and (8) in (1), we have

$$\varepsilon\left[\frac{v_{i+1} - 2v_i + v_{i-1}}{h^2}\right] + r_i\left[\frac{3v_{i+1} - 2v_i - v_{i-1}}{4h}\right] + s_i v_i = \psi_i. \tag{9}$$

Introducing the fitting factor $\sigma(\rho)$ into the aforementioned approach, we obtain the following result:

$$\sigma\varepsilon\left[\frac{v_{i+1} - 2v_i + v_{i-1}}{h^2}\right] + r_i\left[\frac{3v_{i+1} - 2v_i - v_{i-1}}{4h}\right] + s_i v_i = \psi_i. \tag{10}$$

The determination of the fitting factor $\sigma(\rho)$ aims to ensure that the solution of the difference scheme described in (10) achieves uniform convergence towards the solution of (1) with (2).

By multiplying (10) by h and considering the limit as $h \longrightarrow 0$, the result of (10) is as follows:

$$\frac{\sigma}{\rho}\left[v_{i+1}-2v_i-v_{i-1}\right]+\frac{r(0)}{4}\left[3v_{i+1}-2v_i-v_{i-1}\right]=0. \tag{11}$$

Let $\mu=\frac{r^2(0)-\varepsilon s(0)}{r(0)}$. By using (6), we get

$$\lim_{h\to 0}\left(v(ih-h)+v(ih+h)-2v(ih)\right)=\left(\alpha_0-v_0(0)\right)e^{-\mu i\rho}\left(e^{\mu\rho}+e^{-\mu\rho}-2\right),$$
$$\lim_{h\to 0}\left(3v(ih+h)-2v(ih)-v(ih-h)\right)=\left(\alpha_0-v_0(0)\right)e^{-\mu i\rho}\left(3e^{-\mu\rho}-2-e^{\mu\rho}\right).$$

By using the above equations in (11), we get

$$\sigma\left(\rho\right)=\frac{r(0)\rho}{2}\coth\left(\frac{\left(r^2(0)-\varepsilon s(0)\right)\rho}{2r(0)}\right)-\frac{r(0)\rho}{4}, \tag{12}$$

which is a required fitting factor $\sigma\left(\rho\right)$.
Finally, from (11) with the value of $\sigma(\rho)$ given by (12), we obtain the following three-term recurrence relationship:

$$P_iv_{i-1}-Q_iv_i+R_iv_{i+1}=H_i \quad (i=1,2,3,\ldots,N-1), \tag{13}$$

where

$$\begin{aligned}
P_i &= \tfrac{\sigma\varepsilon}{h^2}-\tfrac{r_i}{4h},\\
Q_i &= \tfrac{2\sigma\varepsilon}{h^2}+\tfrac{2r_i}{4h}-s_i,\\
R_i &= \tfrac{\sigma\varepsilon}{h^2}+\tfrac{3r_i}{4h},\\
H_i &= \psi_i.
\end{aligned}$$

Equation (13) generates an $(N-1)$ equations system involving $(N-1)$ undetermined ranging from v_1 to v_{N-1}. These $(N-1)$ equations together with the boundary conditions equation (2), are sufficient to solve the obtained tri-diagonal system with the help of an efficient solver called the Thomas algorithm, commonly called as the "Discrete Invariant Imbedding algorithm".

2.2 Description of the method for right-end boundary layer problems

In this subsection, we will describe the proposed method for the solution of the problem (1)–(2) having boundary layer at right-end point of the interval considered.

The solution of (1) with (2) is of the following form . 22–261[22]):

$$v(t) = v_0(t) + \frac{r(0)}{r(t)} \left(\alpha_0 - v_0(1)\right) e^{-\int_0^t \left(\frac{r(t)}{\varepsilon} - \frac{s(t)}{r(t)}\right) dt} + o(\varepsilon), \tag{14}$$

where $y_0(t)$ denotes the simplified problem's solution:

$$r(t)v_0'(t) + s(t)v_0(t) \;=\; \psi(t), \qquad v_0(1) = \beta. \tag{15}$$

By considering the Taylor series expansions of $r(t)$ and $s(t)$ around the point $t = 0$ up to their respective first terms, we can simplify (14) as follows:

$$v(t) = v_0(t) + \left(\alpha_0 - v_0(0)\right) e^{-\left(\frac{r(1)}{\varepsilon} - \frac{s(1)}{r(1)}\right)t} + o(\varepsilon). \tag{16}$$

Taking the limit as $h \to 0$ and applying (3) to the point $t = t_i = ih$, $i = 0, 1, 2, \ldots, N$, we obtain

$$\lim_{h\to 0} v(ih) = v_0(0) + \left(\alpha_0 - y_0(0)\right) e^{-\left(\frac{r^2(1) - \varepsilon s(1)}{r(1)}\right) i\rho} + o(\varepsilon), \tag{17}$$

where $\rho = h/\varepsilon$.

After multiplying (10) by h and taking the limit as $h \to 0$, (10) converts into the following form:

$$\frac{\sigma}{\rho}\left[v_{i+1} - 2v_i - v_{i-1}\right] + \frac{r(0)}{4}\left[3v_{i+1} - 2v_i - v_{i-1}\right] = 0. \tag{18}$$

Let $\mu = \frac{r^2(0) - \varepsilon s(0)}{r(0)}$. By using (17), we get

$$\lim_{h\to 0}\left(v(ih-h) + v(ih+h) - 2v(ih)\right) = \left(\alpha_0 - v_0(1)\right) e^{-\mu i\rho}\left(e^{\mu\rho} + e^{-\mu\rho} - 2\right),$$
$$\lim_{h\to 0}\left(3v(ih+h) - 2v(ih) - v(ih-h)\right) = \left(\alpha_0 - v_0(1)\right) e^{-\mu i\rho}\left(3e^{-\mu\rho} - 2 - e^{\mu\rho}\right).$$

By substituting the aforementioned equations into (18), we get

$$\sigma(\rho) = \frac{r(0)\rho}{2} \coth\left(\frac{\left(r^2(1) - \varepsilon s(1)\right)\rho}{2r(1)}\right) - \frac{r(0)\rho}{4}, \tag{19}$$

which is a required fitting factor $\sigma(\rho)$ for right-end boundary layer problem.

3 Convergence analysis

This section focuses on the analysis of the convergence of the method.

Definition 1 (Consistency)**.** Let

$$\phi_i[v] = L_h v(t_i) - L_\phi v(t_i), \quad i = 1, 2, \ldots, N.$$

In this context, v denotes a smooth function defined on the interval $I = [0, 1]$, and L_h represents the discrete difference operator. Consequently, the difference equation (13)–(2) exhibits consistency with the corresponding differential equation (1)–(2), if

$$|\phi_i[v]| \to 0 \ as \ h \to o.$$

The quantities $\phi_i[v], i = 1, 2, 3 \ldots, N$ is called the local truncation (or local discretization) errors.

Definition 2. The differential equation (13)–(2) is said to possess local pth-order accuracy when, for suitably smooth data, there exists a positive constant C that remains independent of h and ε such that

$$\max_{1 \le i \le N} |\phi_i[v]| \le Ch^p.$$

The agreement between the differential equation (13)–(2) and (1)–(2), along with its locally second-order accuracy, is established through the lemma provided below.

Lemma 1. If $v \in C^2(I)$, then

$$|\phi_i[v]| \le \max_{t_{i-1} \le t \le t_{i+1}} \left\{ \frac{r_i h}{4} |v_i''| \right\} + O(h^2), \quad i = 1, 2, 3, \ldots, N-1.$$

Proof. By definition,

$$\begin{aligned}
&\phi_i = \sigma\varepsilon \left\{ \frac{v_{i+1} - 2v_i + v_{i-1}}{h^2} - v_i'' \right\} + \left\{ \frac{3v_{i+1} - 2v_i - v_{i-1}}{4h} \right\}, \\
&\phi_i = \sigma\varepsilon \left\{ \frac{h^2}{12} v_i^{iv} + \frac{h^4}{360} v_i^{vi} + \cdots \right\} + r_i \left\{ \frac{h}{12} v_i'' + \frac{h^2}{3!} y_i''' + \cdots \right\}, \\
&|\phi_i| = \max_{t_{i-1} \le t \le t_{i+1}} \left\{ \frac{\sigma\varepsilon h^2}{12} |v_i^{iv}| \right\} + \max_{t_{i-1} \le t \le t_{i+1}} \left\{ \frac{r_i h}{4} |v_i''| \right\}, \\
&|\phi_i| \le \max_{t_{i-1} \le t \le t_{i+1}} \left\{ \frac{r_i h}{4} |v_i''| \right\} + O(h^2), \\
&|\phi_i| \le O(h). \quad i = 1, 2, 3, \ldots, N-1.
\end{aligned}$$

As a result, the intended outcome is attained. □

We will now examine the proposed method's convergence across the entire interval range $0 \le t \le 1$. We write the tridiagonal system (13) in the matrix-vector form

$$WV = D, \tag{20}$$

where $W = (a_{ij}),\ 1 \le i, j \le N-1$ is a tridiagonal matrix of order $N-1$ with

$$a_{i,i-1} = \sigma\varepsilon - \frac{r_i h}{4},$$
$$a_{i,i} = -2\sigma\varepsilon - \frac{2hr_i}{4} + s_i h^2,$$
$$a_{i,i+1} = \sigma\varepsilon + \frac{3hr_i}{4},$$

and $D = (d_i)$ is a column vector with $d_i = h^2\phi_i$ for $i = 1, 2, 3, \ldots, N-1$ with local truncation error ϕ_i:

$$|\phi_i| \leq O(h). \tag{21}$$

We also have

$$W\bar{V} - \phi(h) = D, \tag{22}$$

where $\bar{V} = (\bar{V}_0, \bar{V}_1, \bar{V}_2, \bar{V}_3, \ldots, \bar{V}_N)^t$ and $\phi(h) = (\phi_1(h), \phi_2(h), \phi_3(h), \ldots, \phi_N(h))^t$ stands for the local truncation error and the real solution, respectively. (20) and (22) give us

$$W(\bar{V} - V) = \phi(h). \tag{23}$$

Thus the error equation is

$$WE = \phi(h), \tag{24}$$

where $E = \bar{V} - V = (e_0, e_1, e_2, \ldots, e_N)^t$. If S_i^* is the total of the components in the ith row of W, then

$$S_1^* = \sum_{j=1}^{N-1} a_{1,j} = \tfrac{-\sigma\varepsilon}{h^2} - \tfrac{r_1}{4h} + s_1,$$
$$S_{N-1}^* = \sum_{j=1}^{N-1} a_{N-1,j} = \tfrac{-\sigma\varepsilon}{h^2} - \tfrac{3r_{N-1}}{4h} + s_{N-1},$$
$$S_i^* = \sum_{j=1}^{N-1} a_{i,j} = s_i = B_{i0}.$$

Since $0 < \varepsilon << 1$, The matrix W is irreducible and monotone for sufficiently small h. As a result, W^{-1} must exist and contain nonnegative elements. Therefore, we have from (24) that

$$E = W^{-1}\phi(h), \tag{25}$$
$$\|E\| \leq \|W^{-1}\|\|\phi(h)\|. \tag{26}$$

Let $\bar{a}_{ki}$ represent the (ki)th components of W^{-1}. Since $\bar{a}_{ki} \geq 0$, we have from the operations on matrices:

$$\sum_{j=1}^{N-1} \bar{a}_{ki} S_i^* = 1, \quad k = 1, 2, \ldots, N-1. \tag{27}$$

Therefore, its follows

$$\sum_{j=1}^{N-1} \bar{a}_{ki} \leq \frac{1}{\min_{0\leq i\leq N-1} S_i^*} = \frac{1}{B_{i0}} \leq \frac{1}{|B_{i0}|}, \tag{28}$$

for some i_0 between 1 and $N-1$, and $B_{i0} = q_i$.
Therefore, from (21), (25), and (27), we get

$$e_j = \sum_{i=1}^{N-1} \bar{a}_{ki}\phi_i(h), \quad j = 1(1)N-1, \tag{29}$$

which implies

$$e_j \leq \frac{O(h)}{|q_i|}, \quad j = 1(1)N-1. \tag{30}$$

Consequently, by applying the definitions and (29), we obtain:

$$\|E\| = O(h).$$

This implies that the purposed method is the first-order rate of convergence on uniform mesh. As above, we can apply the same procedure for showing the purposed method is of first-order rate of convergence on uniform mesh for the right layer problem.

4 Numerical illustrations

The effectiveness of the purposed method has been demonstrated by implementing it on the three linear SPPs at left-end boundary layer as well as one problem involving a right-end boundary layer and presented the computational results in the tables in terms of the maximum absolute errors E_ε^N. These examples have been chosen because they have been widely discussed in literature. For various values of mesh point N and perturbation parameter ε, the E_ε^N are defined by $E_\varepsilon^N = \max\limits_{0\leq i\leq N-1}[|v(t_i) - v_i|]$, where $v(t_i)$ and v_i denote the exact and approximate solution, respectively. The double mesh principle is used to calculate the rate of convergence defined as $r_\varepsilon^N = \log_2\left(\frac{E_\varepsilon^N}{E_\varepsilon^{2N}}\right)$. The purposed method is capable of achieving uniform results, when perturbation parameter ε tends to 0 for any fixed value of the mesh size h.

Example 1. First, consider the following homogeneous SPP from [15]:

$$\varepsilon v''(t) + v'(t) - v(t) = 0, \quad t \in [0,1],$$

with boundary condition $v(0) = 1$ and $v(1) = 1$.
The exact solution is given by

$$v(t) = \frac{(e^{m_2} - 1)e^{m_1 t} + (1 - e^{m_1})e^{m_2 t}}{e^{m_2} - e^{m_1}},$$

where $m_1 = \dfrac{(-1 + \sqrt{1+4\varepsilon})}{2\varepsilon}$ and $m_2 = \dfrac{(-1 - \sqrt{1+4\varepsilon})}{2\varepsilon}$.

The maximum absolute errors for various values of N and singular perturbation parameter ε are presented in Table 1 for example 1. It can be easily observed from Table 1 that the maximum absolute errors tends uniformly, when the singular perturbation parameter ε tends to 0, for any fixed value of $N = 1/h$. Also, rates of convergence presented in Table 1 show that the purposed scheme is capable of producing almost first-order accurate uniformly convergent solution. In Figure 1, we present our solution and the exact solution for various values of ε and a fixed value of N. Clearly, as shown in the figure, the numerical solution and the exact solution are very close within the boundary layers for smaller values of ε.

Example 2. Consider the following non-homogeneous SPP involving a constant term $f(t)$ [25, 15]:

$$\varepsilon v''(t) + v'(t) = 2, \quad t \in [0, 1],$$

with boundary condition $v(0) = 1$ and $v(1) = 1$. The exact solution is given by $v(t) = 2t + \frac{1-e^{t/\varepsilon}}{e^{t/\varepsilon}-1}$.

The maximum absolute errors for various values of N and singular perturbation parameter ε are presented in Table 2 for example 2. It can be easily observed from Table 2 that the maximum absolute errors tends uniformly, when the singular perturbation parameter ε tends to 0, for any fixed value of $N = 1/h$. In Figure 2, we present our solution and the exact solution for various values of ε and a fixed value of N. Clearly, as shown in the figure, the numerical solution and the exact solution are very close within the boundary layers for smaller values of ε.

Example 3. Consider the following non-homogeneous SPP involving a variable term $f(t)$ [25, 15]:

$$\varepsilon v''(t) + v'(t) = 1 + 2t, \quad t \in [0, 1],$$

with boundary condition $v(0) = 1$ and $v(1) = 1$. The exact solution is given by $v(t) = \frac{1-e^{-t/\varepsilon}}{1-e^{1/\varepsilon}}(2\varepsilon - 1) + t(t + 1 - 2\varepsilon)$.

The maximum absolute errors for various values of N and singular perturbation parameter ε are presented in Table 3 for example3. It can be easily observed from Table 3 that the maximum absolute errors tends uniformly, when the singular perturbation parameter ε tends to 0, for any fixed value of $N = 1/h$. Also, rates of convergence presented in Table 3 show that the purposed scheme is capable of producing almost first-order accurate uniformly convergent solution. In Figure 3, we present our solution and the exact solution for various values of ε and a fixed value of N. Clearly, as shown in the figure, the numerical solution and the exact solution are very close within the boundary layers for smaller values of ε.

Example 4. Lastly, consider the following homogeneous SPP at right-end boundary layer [25, 26, 15]:

$$\varepsilon v''(t) - v'(t) - (1+\varepsilon)v(t) = 0, \quad t \in [0,1],$$

with boundary condition $v(0) = 1 + exp(-(1+\varepsilon)/\varepsilon)$ and $v(1) = 1 + 1/e$. The exact solution is given by $v(t) = e^{(1+\varepsilon)(t-1)/\varepsilon} + e^{-t}$.

The maximum absolute errors for various values of N and singular perturbation parameter ε are presented in Table 4 for Example 4. It can be easily observed from the Table 4 that the maximum absolute errors tends uniformly, when the singular perturbation parameter ε tends to 0, for any fixed value of $N = 1/h$. Also, rates of convergence presented in Table 4 show that the purposed scheme is capable of producing almost first-order accurate uniformly convergent solution. In Figure 4, we present our solution and the exact solution for various values of ε and a fixed value of N. Clearly, as shown in the figure, the numerical solution and the exact solution are very close within the boundary layers for smaller values of ε.

Table 1: Computational results in terms of maximum absolute errors for various values of N and ε and the rate of convergence r_ε^N for Example 1

N	16	32	64	128	256	512
$\varepsilon = 10^{-5}$	0.0112	0.0057	0.0029	0.0014	0.0007	0.0003
r_ε^N	0.9745	0.9749	1.0506	1.0000	1.2224	
$\varepsilon = 10^{-6}$	0.0112	0.0057	0.0029	0.0014	0.0007	0.0003
r_ε^N	0.9745	0.9749	1.0506	1.0000	1.2224	
$\varepsilon = 10^{-7}$	0.0112	0.0057	0.0029	0.0014	0.0007	0.0003
r_ε^N	0.9745	0.9749	1.0506	1.0000	1.2224	
$\varepsilon = 10^{-8}$	0.0112	0.0057	0.0029	0.0014	0.0007	0.0003
r_ε^N	0.9745	0.9749	1.0506	1.0000	1.2224	
$\varepsilon = 10^{-9}$	0.0112	0.0057	0.0029	0.0014	0.0007	0.0003
r_ε^N	0.9745	0.9749	1.0506	1.0000	1.2224	

Table 2: Computational results in terms of maximum absolute errors for various values of N and ε for Example 2

N	16	32	64	128	256	512
$\varepsilon = 10^{-5}$	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000
$\varepsilon = 10^{-6}$	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000
$\varepsilon = 10^{-7}$	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000
$\varepsilon = 10^{-8}$	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000
$\varepsilon = 10^{-9}$	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000

Table 3: Computational results in terms of maximum absolute errors for various values of N and ε and the rate of convergence r_ε^N for Example 3

N	16	32	64	128	256	512
$\varepsilon = 10^{-5}$	0.0586	0.0303	0.0154	0.0079	0.0039	0.0019
r_ε^N	0.9516	0.9764	0.9630	1.0184	1.0375	
$\varepsilon = 10^{-6}$	0.0586	0.0303	0.0154	0.0077	0.0039	0.0019
r_ε^N	0.9516	0.9764	1.0000	0.9814	1.0375	
$\varepsilon = 10^{-7}$	0.0586	0.0303	0.0154	0.0078	0.0039	0.0019
r_ε^N	0.9516	0.9764	0.9814	1.0000	1.0375	
$\varepsilon = 10^{-8}$	0.0586	0.0303	0.0154	0.0078	0.0039	0.0019
r_ε^N	0.9516	0.9764	0.9814	1.0000	1.0375	
$\varepsilon = 10^{-9}$	0.0586	0.0303	0.0154	0.0078	0.0039	0.0019
r_ε^N	0.9516	0.9764	0.9814	1.0000	1.0375	

Table 4: Computational results in terms of maximum absolute errors for various values of N and ε and the rate of convergence r_ε^N for Example 1

N	16	32	64	128	256	512
$\varepsilon = 10^{-5}$	0.0112	0.0057	0.0029	0.0014	0.0007	0.0003
r_ε^N	0.9745	0.9749	1.0506	1.0000	1.2224	
$\varepsilon = 10^{-6}$	0.0112	0.0057	0.0029	0.0014	0.0007	0.0003
r_ε^N	0.9745	0.9749	1.0506	1.0000	1.2224	
$\varepsilon = 10^{-7}$	0.0112	0.0057	0.0029	0.0014	0.0007	0.0003
r_ε^N	0.9745	0.9749	1.0506	1.0000	1.2224	
$\varepsilon = 10^{-8}$	0.0112	0.0057	0.0029	0.0014	0.0007	0.0003
r_ε^N	0.9745	0.9749	1.0506	1.0000	1.2224	
$\varepsilon = 10^{-9}$	0.0112	0.0057	0.0029	0.0014	0.0007	0.0003
r_ε^N	0.9745	0.9749	1.0506	1.0000	1.2224	

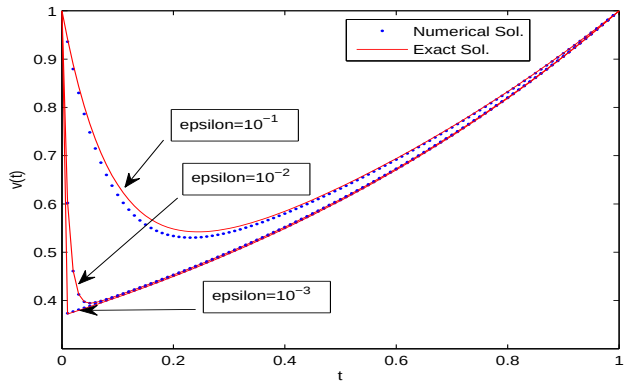


Figure 1: Computational solution of the given Example 1 for the fixed value $N = 100$ and various values of ε

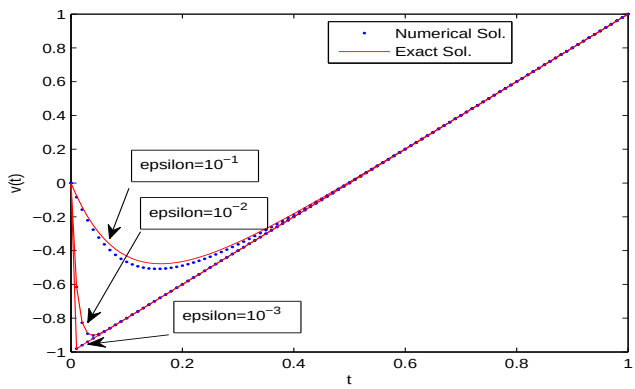


Figure 2: Computational solution of the given Example 2 for the fixed value $N = 100$ and various values of ε

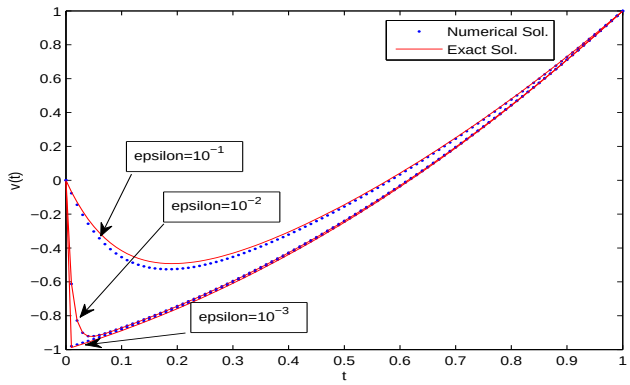


Figure 3: Computational solution of the given Example 3 for the fixed value $N = 100$ and various values of ε

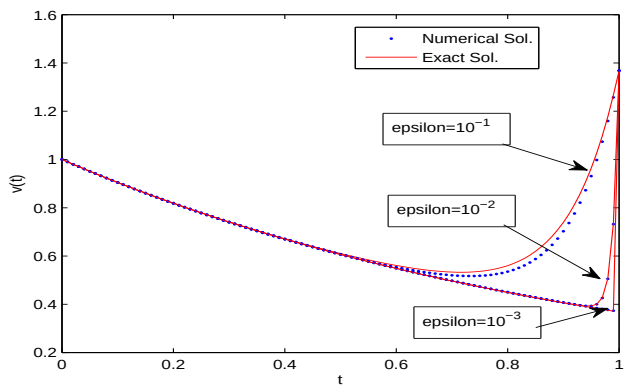


Figure 4: Computational solution of the given Example 4 for fixed value of $N = 100$ and various values of ε

5 Conclusion

We have derived an exponentially fitted finite difference tridiagonal scheme for solving singularly perturbed two-point BVPs at one endpoint (left or right). We have carried out the convergence analysis for the proposed scheme and performed the numerical experiments on four example problems (three left-end and one right-end problems) for various values of $N = 1/h$ and perturbation parameter ε, which show that the scheme is of almost first-order accurate. The computational results in terms of maximum absolute

error are presented in Tables 1–4. It is easily observed from the tables that the proposed method is capable of producing highly accurate results for a fixed value of mesh size h, when the perturbation parameter ε tends to 0. The maximum absolute errors are becoming uniform for any fixed values of N when $\varepsilon \to 0$. Furthermore, one can easily observe from Tables 1, 3, and 4 that the proposed exponential fitted finite difference scheme is capable of producing first-order accurate uniformly convergent solution for any fixed value of mesh size $h = 1/N$ when perturbation parameter ε tends to 0. In Figures 1–4, we present our solution and the exact solution for various values of ε and a fixed value of N. Clearly, as shown in the figures, the numerical solution and the exact solution are very close within the boundary layers for smaller values of ε. Notably, the novelty of our method lies in its independence from both deviating arguments and fitted meshes [15].

Conflict of interest The author declare that he has no competing interests.

References

[1] Andargie, A. and Reddy, Y. *Two initial value problems approach for solving singular perturbations problems*, Am. J. Comput. Math. 2(03) (2012), 213–216.

[2] Bender, C.M. and Orsazag, S.A. *Advanced mathematical methods for scientists and engineers*, Springer, New York, 1999.

[3] Chakravarthy, P. and Reddy, Y.N. *Exponentially fitted modified upwind scheme for singular perturbation problems*, Int. J. Fluid Mech. Res. 33 (2006), 119–136.

[4] Gold, R.R. *Magneto hydrodynamic pipe flow*, Part I. J. Fluid Mech. 13 (1962), 505–512.

[5] Habib, H.M. and El-Zahar, E.R. *An algorithm for solving singular perturbation problems with mechanization*, Appl. Math. Comput. 188 (2007), 286–302.

[6] Hinch, E.J. *Perturbation methods*, Cambridge University Press, Cambridge, 1991.

[7] Holmes, M.H. *Introduction to perturbation methods*, Springer, Berlin, 1995.

[8] Jayakumar, J. and Ramanujam, N. *A numerical method for singular perturbation problems arising in chemical reactor theory*, Comput. Math. Appl. 27 (1994,) 83–99.

[9] Kadalbajoo, M.K. and Kumar, D. *Initial value technique for singularly perturbed two point boundary value problems using an exponentially fitted finite difference scheme*, Comput. Math. Appl. 57 (2009), 1147–1156.

[10] Kadalbajoo, M.K. and Kumar, D. *A brief survey on numerical methods for solving singularly perturbed problems*, Appl. Math. Comput. 217 (2010), 3641–3716.

[11] Kevorkian, J. and Cole, J.D. *Perturbation methods in applied mathematics*, 2nd Edition, Springer-Verlag, New York, 1981.

[12] Kumar, M. and Surabhi, T. *An initial-value technique to solve third-order reaction-diffusion singularly perturbed boundary-value problems*, Int. J. Comput. Math. 89(17) (2012), 2345–2352.

[13] Kumar, M., Singh, P. and Hradyesh Kumar, M. *An initial-value technique for singularly perturbed boundary value problems via cubic spline*, Int. J. Comput. Methods Eng. Sci. Mech. 8(6) (2007), 419–427.

[14] Lorenz, J. *Combinations of initial and boundary value methods for a class of singular perturbation problems*, Numerical analysis of singular perturbation problems (Proc. Conf., Math. Inst., Catholic Univ., Nijmegen, 1978), pp. 295–315, Academic Press, London-New York, 1979.

[15] Madhu Latha, K., Phaneendra, K. and Reddy, Y.N. *Numerical integration with exponential fitting factor for singularly perturbed two point boundary value problems*, British Journal of Mathematics & Computer Science 3(3) (2013), 397–414.

[16] Miller, J.J.H. *Singular perturbation problems in chemical physics, analytic and computational methods*, XCVII Wiley, New York, 1997.

[17] Miller, J.J.H., Riordan, R.E.O. and Shishkin, G.I. *Fitted numerical methods for singular perturbation problems, error estimates in the maximum norm for linear problems in one and two dimensions*, World Scientific, 1996.

[18] Mishra, H. and Saini, S. *Numerical solution of singularly perturbed two-point boundary value problem via Liouville-Green transform*, Am. J. Comput. Math. 3(1) (2013), 1–5.

[19] Nayfeh, A.H. *Perturbation methods*, John Wiley & Sons, Inc., New York, 1979.

[20] Nayfeh, A.H. *Introduction to perturbation techniques*, Wiley-VCH, New York, 1993.

[21] O'Malley, R.E. *Introduction to singular perturbations*, Academic Press, New York, 1974.

[22] O'Malley, R.E. *Singular perturbation methods for ordinary differential equations*, Applied Mathematical Sciences, 89, Springer, Berlin, 1990.

[23] Padmaja, P., Aparna, P. and Gorla, R.S.R. *An initial-value technique for self-adjoint singularly perturbed two-point boundary value problems*, Int. J. Appl. Mech. Eng. 25(1) (2020), 106–126.

[24] Phaneendra, K. and Lalu, M. *Gaussian quadrature for two-point singularly perturbed boundary value problems with exponential fitting*, Communications in Mathematics and Applications 10(3) (2019), 447–467.

[25] Ranjan, R. and Prasad, H.S. *An efficient method of numerical integration for a class of singularly perturbed two point boundary value problems*, WSEAS Trans. Math. 17 (2018), 265–273.

[26] Ranjan, R. and Prasad, H.S., *A fitted finite difference scheme for solving singularly perturbed two point boundary value problems*, Inf. Sci. Lett. 9(2), (2020), 65–73.

[27] Ranjan, R., Prasad, H.S. and Alam, J. *A simple method of numerical integration for a class of singularly perturbed two point boundary value problems*, i-manager's Journal on Mathematics 7(1) (2018), 41.

[28] Roos, H.G., Stynes, M. and Tobiska, L. *Numerical methods for singularly perturbed differential equations*, Springer, Berlin 1996.

[29] Smith, D.R. *Singular perturbation theory: An introduction with applications*, Cambridge University Press, Cambridge, 1985.

[30] Verhulst, F. *Methods and applications of singular perturbations: Boundary layers and multiple timescale dynamics*, Springer, Berlin 2005.

[31] Vigo-Aguiar, J. and Natesan, S. *An efficient numerical method for singular perturbation problems*, J. Comput. Appl. Math., 192 (2006), 132–141.

How to cite this article
Kumar, N., Kumar Sinha, R. and Ranjan, R., Singularly perturbed two-point boundary value problem by applying exponential fitted finite difference method. *Iran. J. Numer. Anal. Optim.*, 2023; 13(4): 711-727. https://doi.org/10.22067/ijnao.2023.83070.1283

Iranian Journal of Numerical Analysis and Optimization
Vol. 13, No. 4, 2023, pp 728–746
https://doi.org/10.22067/ijnao.2023.81484.1229
https://ijnao.um.ac.ir/

Research Article

Numerical study of sine-Gordon equations using Bessel collocation method

S. Arora and I. Bala*

Abstract

The nonlinear space time dynamics have been discussed in terms of a hyperbolic equation known as a sine-Gordon equation. The proposed equation has been discretized using the Bessel collocation method with Bessel polynomials as base functions. The proposed hyperbolic equation has been transformed into a system of parabolic equations using a continuously differentiable function. The system of equations involves one linear and the other nonlinear diffusion equation. The convergence of the present technique has been discussed through absolute error, L_2-norm, and L_∞-norm. The numerical values obtained from the Bessel collocation method have been compared with the values already given in the literature. The present technique has been applied to different problems to check its applicability. Numerical values obtained from the Bessel collocation method have been presented in tabular as well as in graphical form.

AMS subject classifications (2020): 35L10, 33C10, 35L05, 65M70.

Keywords: Sine-Gordon equation; Bessel polynomials; Wave equation; Orthogonal Collocation.

*Corresponding author

Received 9 March 2023; revised 20 June 2023; accepted 20 June 2023

Shelly Arora
Department of Mathematics, Punjabi University, Patiala, 147002, Punjab, India. e-mail: aroshelly@pbi.ac.in

Indu Bala
Department of Mathematics, Punjabi University, Patiala, 147002, Punjab, India. e-mail: indu13121994@gmail.com

1 Introduction

Nonlinear partial differential equations have wide applications in different branches of science and engineering, which helps to understand the diversity of physical phenomena in a logical manner. The nonlinear wave equations such as Klein–Gordon and sine-Gordon equations have wide applications in optics, plasma physics, quantum mechanics and fluid mechanics, and so on. Finding the analytic solution to these problems is a tedious task due to the complexity of the nonlinear terms. The numerical approximation of the solution in terms of the discrete set of points is often desirable by researchers due to the simplicity of the numerical computations.

The derivation of the Klein–Gordon equation is a generalization of the Schrödinger equation. It was named after the physicists Oskar Klein and Walter Gordon. They together in 1926, proposed that relativistic electrons can be described by the Klein–Gordon equation during the research for the equation describing de Broglie waves. Schrödinger considered the Klein–Gordon equation as a quantum wave equation [18, 19]. Klein–Gordon equation plays a significant role in many scientific applications, such as nonlinear optics and quantum field theory, and solid state physics. Klein–Gordon equation in one dimension can be considered as

$$\frac{\partial^2 y}{\partial t^2} - \beta \frac{\partial^2 y}{\partial \xi^2} + f_1(y) = f_2(\xi, t), \tag{1}$$

where $(\xi, t) \in (\xi_a, \xi_b) \times (0, T)$, $f_1(y)$ is nonlinear force, and β is a constant. The sine-Gordon equation is a special case of (1) for $f_1(y) = \sin(y)$ and $f_2(\xi, t) = 0$. Interestingly, the sine-Gordon equation was discovered separately in 1939 by Frenkel and Kontorova while studying the propagation of slip in an infinite chain of elastically bound atoms lying over a fixed chain. Later it was found that sine-Gordon is a special case of Klein–Gordon equation. The sine-Gordon equation in one dimension can be described as

$$\frac{\partial^2 y}{\partial t^2} - \frac{\partial^2 y}{\partial \xi^2} + \sin(y(\xi, t)) = 0, \tag{2}$$

and the initial and boundary conditions can be described as

$$\begin{aligned} y(\xi, 0) = \phi_1(\xi) \quad \textit{and} \quad y_t(\xi, 0) = \phi_2(\xi), \\ y(\xi_a, t) = \phi_3(t) \quad \textit{and} \quad y(\xi_b, t) = \phi_4(t). \end{aligned} \tag{3}$$

The present problem of the equation is a continuum model for waves in mechanical systems, coupled-pendulum, the study of the domain wall dynamics in magnetic crystals, magnetic-flux propagation in large Josephson junctions, propagation of crystal dislocations in solids, propagation of ultra-short optical pulses in optical fibers, as a nonlinear effective field theory for strong interactions in particle physics, and so on; see [8, 16, 18, 19, 28, 37].

A variety of numerical techniques have been developed by different researchers to study the behavior of nonlinear sine-Gordon equation, such as finite difference method, inverse scattering method, auxiliary equation method, spectral method, pseudo-spectral method, tanh-sech method, Adomian decomposition method, sine-cosine method, Jacobi elliptic functions, Backlund transformation, Riccati equation expansion method, homotopy perturbation method, and variational iteration method [15, 18, 33, 32, 40].

Method of characteristics and a leapfrog finite difference scheme [3, 24] were the first two methods developed to obtain the numerical solution of sine-Gordon equation. Strauss and Vázquez [36] developed a leapfrog finite difference, an implicit, energy-conserving scheme for the Klein–Gordon equation.

Apart from these, other numerical methods have also been developed for the solution of sine-Gordon equations, which include pseudospectral methods and spectral methods. Pseudospectral methods include the split-step Fourier scheme [1, 2, 41] and spectral methods include energy-conserving, wavelet spectral method, Fourier scheme, Legendre spectral element method, and multiresolution analysis method based on Legendre wavelets [27]. Finite element methods is based on a collocation scheme using Legendre–Gauss–Lobatto points [29], cubic B-splines [31] and Petrov–Galerkin scheme [5].

In the present study, the Bessel collocation method (BCM) has been followed to study the behavior of the nonlinear sine-Gordon equation. The Bessel polynomials of degree n have been taken as base polynomials. To discretize the time direction, the temporal variable has been split by the introduction of a continuously differentiable function. It converts the wave equation into a system of equations involving one linear and the other nonlinear equation.

The present manuscript has been divided into six sections starting from introduction. The BCM has been described in section 2. The explanation of collocation points as well as the implementation of the collocation technique has been described in sections 3 and 4, respectively. Convergence analysis has been discussed in section 5, whereas the numerical application has been given in section 6.

2 Bessel collocation method (BCM)

The collocation method belongs to the general class of approximate methods, known as weighted residual methods. In this method, the residual is set orthogonal to the weight function. In an orthogonal collocation, the trial function $y(\xi, t)$ is represented in a series of known polynomials with unknown coefficients [13, 22, 38]. The residual is set equal to zero at the collocation points.

On the basis of the implementation of the trial function, the collocation technique can be classified into three categories. If the trial function satisfies

the differential equation $\ell\mathbf{V}(\mathbf{y}) = \mathbf{0}$ with volume V, then it is termed as interior collocation. If the trial function satisfies the boundary $\ell\mathbf{B}(\mathbf{y}) = \mathbf{0}$, where B is the boundary adjoining volume V, then it is termed as boundary collocation. If the trail function satisfies neither the equation nor the boundary and is adjusted to both, then it is termed as a mixed collocation.

The choice of base function is the first important step in the technique of collocation. In the present study, Bessel polynomials of order n have been chosen as a trial function, and the technique is called the BCM.
During the study of problems in dynamic astronomy to solve the Kepler's problem, a German astronomer Bessel in 1824, introduced Bessel polynomials, which are the solution to a second order boundary value problem. These polynomials can be written in terms of limit confluent hypergeometric function ${}_0F_1$. The details of these hypergeometric functions are given in [7, 25, 34, 39]:

$$J_n(\xi) = \frac{\xi^n}{2^n n!}{}_0F_1(-; n+1; -\frac{1}{4}\xi^2). \tag{4}$$

The Bessel coefficients also follow from the power series expansion for small values of ξ

$$\lim_{\xi\to 0}{}_0F_1(-; n+1; -\frac{1}{4}\xi^2) = 1,$$

$$\lim_{\xi\to 0}\xi^{-n}J_n(\xi) = \frac{1}{2^n n!},$$

which shows that as $\xi \to 0$, the Bessel coefficient $J_n(\xi)$ approaches to $\frac{1}{2^n n!}$.

The first order derivative of the Bessel function is defined as

$$\frac{d}{d\xi}(\xi^n J_n(\xi)) = \xi^n J_{n-1}(\xi),$$

$$\frac{d}{d\xi}(\xi^{-n} J_n(\xi)) = -\xi^{-n} J_{n+1}(\xi).$$

3 Collocation points

The next step is the choice of collocation points. It is an important part of the collocation technique. In this study, instead of taking the uniform points, the zeros of orthogonal polynomials, such as Jacobi polynomials, have been taken as collocation points. Legendre and Chebyshev polynomials are special cases of Jacobi polynomials, and the zeros of these orthogonal polynomials are preferably taken as collocation points. Runge's divergence formula also states that nonuniform collocation points give less error as compared to uniform collocation points.

Theorem 1. [26] If $\mathcal{Q}_n(\xi)$ form a simple set of real polynomials and $w(\xi) > 0$ on $a \le \xi \le b$, then the necessary and sufficient condition that the set $\mathcal{Q}_n(\xi)$

is orthogonal with respect to the $w(\xi)$ over the interval $a \leq \xi \leq b$ is that

$$\int_a^b w(\xi)x^k \mathcal{Q}_n(\xi)d\xi = 0, \qquad k = 0, 1, 2, 3, \ldots, (n-1).$$

Theorem 2. [26] If the simple set of real polynomials $\mathcal{Q}_n(\xi)$ is orthogonal with respect to the weight function $w(\xi) > 0$ on the interval $a \leq \xi \leq b$, then the zeros of $\mathcal{Q}_n(\xi)$ are distinct and lie in the interval $a \leq \xi \leq b$.

Since $\mathcal{Q}_n(\xi)$ is a polynomial of degree n, then it has exactly n roots, multiplicity counted, such that the roots are distinct and all lie in $a \leq \xi \leq b$.

Usually, the collocation points are selected from the Legendre or Chebyshev polynomials, and these polynomials are also particular cases of Jacobi polynomials. The details of the collocation points are given elsewhere [4, 6, 11, 12, 14, 17, 20, 35]. The zeros of Chebyshev polynomials have been taken as collocation points:

$$\xi_j = \cos\big(\frac{\pi(j-1)}{n}\big), \qquad j = 1, 2, \ldots, n+1.$$

The Chebyshev collocation points have been transformed from the interval $[-1, 1]$ to $[\xi_a, \xi_b]$ using one to one correspondence.

4 Implementation of BCM

To discretize the given problem, BCM is applied in space direction. To apply BCM, (2) has been split in time direction. For this purpose, a new continuously differentiable function $z(\xi, t)$, differentiable with respect to t has been introduced:

$$\begin{aligned} \frac{\partial y}{\partial t} &= z(\xi, t), \qquad (\xi, t) \in \ (\xi_a, \xi_b) \times (0, T), \\ \frac{\partial z}{\partial t} &= \frac{\partial^2 y}{\partial \xi^2} - \sin(y(\xi, t)), \qquad (\xi, t) \in \ (\xi_a, \xi_b) \times (0, T). \end{aligned} \tag{5}$$

To apply BCM on a system of equations defined by (5), the two functions $y(\xi, t)$ and $z(\xi, t)$ have been approximated in terms of Bessel polynomials as

$$\begin{aligned} y(\xi, t) &= \sum_{i=1}^{n+1} J_i(\xi)c_i(t), \\ z(\xi, t) &= \sum_{i=1}^{n+1} J_i(\xi)d_i(t), \end{aligned} \tag{6}$$

where $J_i(\xi)$ are ith order Bessel polynomials. To simplify (6), the Bessel polynomials can be rewritten as suggested by [10, 42, 43, 44, 45, 46, 47, 48]:

$$
\begin{aligned}
y(\xi,t) &= \sum_{i=1}^{n+1} \xi^{i-1} R c_i(t),\\
z(\xi,t) &= \sum_{i=1}^{n+1} \xi^{i-1} R d_i(t),
\end{aligned}
\tag{7}
$$

where R is a square matrix of order $(n+1)\times(n+1)$ and $c_i(t)$ and $d_i(t)$ are the unknown coefficients of t, which are to be determined.

For n being an odd integer, R is defined as

$$
R = \begin{bmatrix}
\frac{1}{0!0!2^0} & 0 & 0 & 0 & \cdots & 0 & 0\\
0 & \frac{1}{0!1!2^1} & 0 & 0 & \cdots & 0 & 0\\
\frac{-1}{1!1!2^2} & 0 & \frac{1}{0!2!2^2} & 0 & \cdots & 0 & 0\\
0 & \frac{-1}{1!2!2^3} & 0 & \frac{1}{0!3!2^3} & \cdots & 0 & 0\\
\frac{1}{2!2!2^4} & 0 & \frac{-1}{1!3!2^4} & 0 & \cdots & 0 & 0\\
\vdots & \vdots & \vdots & \vdots & \vdots & \vdots & \vdots\\
\vdots & \vdots & \vdots & \vdots & \vdots & \vdots & \vdots\\
\frac{(-1)^{\frac{n-1}{2}}}{(\frac{n-1}{2})!(\frac{n-1}{2})!2^{n-1}} & 0 & \frac{(-1)^{\frac{n-3}{2}}}{(\frac{n-3}{2})!(\frac{n+1}{2})!2^{n-1}} & 0 & \cdots & \frac{1}{0!(n-1)!2^{n-1}} & 0\\
0 & \frac{(-1)^{\frac{n-1}{2}}}{(\frac{n-1}{2})!(\frac{n+1}{2})!2^{n}} & 0 & \frac{(-1)^{\frac{n-3}{2}}}{(\frac{n-3}{2})!(\frac{n+3}{2})!2^{n}} & \cdots & 0 & \frac{1}{0!n!2^{n}}
\end{bmatrix}
$$

However, for n being an even integer, the matrix R can be written as

$$
R = \begin{bmatrix}
\frac{1}{0!0!2^0} & 0 & 0 & 0 & \cdots & 0 & 0\\
0 & \frac{1}{0!1!2^1} & 0 & 0 & \cdots & 0 & 0\\
\frac{-1}{1!1!2^2} & 0 & \frac{1}{0!2!2^2} & 0 & \cdots & 0 & 0\\
0 & \frac{-1}{1!2!2^3} & 0 & \frac{1}{0!3!2^3} & \cdots & 0 & 0\\
\frac{1}{2!2!2^4} & 0 & \frac{-1}{1!3!2^4} & 0 & \cdots & 0 & 0\\
\vdots & \vdots & \vdots & \vdots & \vdots & \vdots & \vdots\\
\vdots & \vdots & \vdots & \vdots & \vdots & \vdots & \vdots\\
0 & \frac{(-1)^{\frac{n-2}{2}}}{(\frac{n-2}{2})!(\frac{n}{2})!2^{n-1}} & 0 & \frac{(-1)^{\frac{n-4}{2}}}{(\frac{n-4}{2})!(\frac{n+2}{2})!2^{n-1}} & \cdots & \frac{1}{0!(n-1)!2^{n-1}} & 0\\
\frac{(-1)^{\frac{n}{2}}}{(\frac{n}{2})!(\frac{n}{2})!2^{n}} & 0 & \frac{(-1)^{\frac{n-2}{2}}}{(\frac{n-2}{2})!(\frac{n+2}{2})!2^{n}} & 0 & \cdots & 0 & \frac{1}{0!n!2^{n}}
\end{bmatrix}
$$

At jth collocation point, (7) can be written as

$$
\begin{aligned}
y(\xi_j,t) &= \sum_{i=1}^{n+1} \xi_j^{i-1} R c_i(t), \qquad j=1,2,\ldots,n+1,\\
z(\xi_j,t) &= \sum_{i=1}^{n+1} \xi_j^{i-1} R d_i(t), \qquad j=1,2,\ldots,n+1.
\end{aligned}
\tag{8}
$$

Now, rewrite (8) in a matrix form at jth collocation point

$$
[y_j] = [X][c_i(t)], \qquad [z_j] = [X][d_i(t)], \tag{9}
$$

where $X = [\xi_j^{i-1}]R$ and y_j represents the value of y at jth collocation point. We have

$$[X]^{-1}[y_j] = [\mathbf{c}], \qquad [X]^{-1}[z_j] = [\mathbf{d}]. \tag{10}$$

Substituting collocation coefficients from (10) in (7) results in

$$\begin{aligned} y(\xi,t) &= \sum_{i=1}^{n+1} \xi^{i-1} X^{-1} y_i, \\ z(\xi,t) &= \sum_{i=1}^{n+1} \xi^{i-1} X^{-1} z_i. \end{aligned} \tag{11}$$

The first and second order derivatives of $y(\xi,t)$ with respect to ξ can be obtained as

$$\begin{aligned} \frac{\partial y}{\partial \xi} &= \sum_{i=1}^{n+1} (i-1)\xi^{i-2} X^{-1} y_i, \\ \frac{\partial^2 y}{\partial \xi^2} &= \sum_{i=1}^{n+1} (i-1)(i-2)\xi^{i-3} X^{-1} y_i. \end{aligned} \tag{12}$$

Using the discretized forms of $y(\xi,t)$ and $z(\xi,t)$ in (5) leads to the following system of equations:

$$\begin{aligned} \frac{dy_j}{dt} &= z_j, \\ \frac{dz_j}{dt} &= \beta \sum_{i=1}^{n+1} B_{ji} y_i - \sin(y_j), \end{aligned} \tag{13}$$

where $j = 2, 3, \ldots, n$.
In the above coupled form of equations, A_{ji} and B_{ji} are first and second order discretized forms of derivatives of $y(\xi)$ with respect to ξ at jth collocation point, respectively. Boundary conditions for both $y(\xi,t)$ and $z(\xi,t)$ configurations assumed to be $y(a,t) = y_1 = y_a$, $y(b,t) = y_{n+1} = y_b$, $z(a,t) = z_1 = z_a$, and $z(b,t) = z_{n+1} = z_b$. The matrix representation of (13) of sine-Gordon can be written as:

$$\begin{bmatrix} \frac{dY}{dt} \\ \frac{dZ}{dt} \end{bmatrix} = \begin{bmatrix} I & O \\ O & B \end{bmatrix} \begin{bmatrix} Z \\ Y \end{bmatrix} - F. \tag{14}$$

In the above system of equations, there are $n-1$ collocation equations and two boundary conditions for each function $y(\xi,t)$ and $z(\xi,t)$, respectively. It results in $2(n-1)$ collocation equations in total and four boundary conditions. There is no effect of boundary conditions in the matrix representation of (14) as they are in scalar form and get merged into F. Moreover, O represents the zero matrix, and I represents the identity matrix in relation (14). Also,

$$O = \begin{bmatrix} 0\,0\,0\ldots 0 \\ 0\,0\,0\ldots 0 \\ \vdots\;\vdots\;\vdots\;\ddots\;\vdots \\ 0\,0\,0\ldots 0 \end{bmatrix}_{(n-1)\times(n-1)}, \qquad I = \begin{bmatrix} 1\,1\,1\ldots 1 \\ 1\,1\,1\ldots 1 \\ \vdots\;\vdots\;\vdots\;\ddots\;\vdots \\ 1\,1\,1\ldots 1 \end{bmatrix}_{(n-1)\times(n-1)},$$

$$Y = \begin{bmatrix} y_2 \\ y_3 \\ \vdots \\ y_n \end{bmatrix}_{(n-1)\times 1}, \qquad Z = \begin{bmatrix} z_2 \\ z_3 \\ \vdots \\ z_n \end{bmatrix}_{(n-1)\times 1},$$

$$B = \begin{bmatrix} B(2,2) & B(2,3) & B(2,4) & \ldots & B(2,n) \\ B(3,2) & B(3,3) & B(3,4) & \ldots & B(3,n) \\ \vdots & \vdots & \vdots & \ddots & \vdots \\ B(n,2) & B(n,3) & B(3,4) & \ldots & B(n,n) \end{bmatrix}_{(n-1)\times(n-1)},$$

and

$$F = \begin{bmatrix} O \\ \bar{F} \end{bmatrix}$$

for the Klein–Gordon equation $\bar{F}$ can be represented as

$$\bar{F} = \begin{bmatrix} \sin(y_2) \\ \sin(y_3) \\ \vdots \\ \sin(y_n) \end{bmatrix}_{(n-1)\times 1}$$

The matrix representation corresponds to a nonlinear system of $2(n-1)$ equations form block matrix structure (14). The left-hand side includes the column vector of time derivatives of functions y and z, respectively, and the right-hand side does not include any product with the inverse of a coefficient matrix, which helps to reduce the stiffness of the system of equations. The system of ordinary differential equations is solved using MATLAB with the ode23s module.

5 Convergence analysis

Theorem 3. [26] If $\{\mathcal{Q}_n(\xi)\}$ represents for simple set of polynomials and if $\mathcal{Y}(\xi)$ is a polynomial of degree m, then there exist constants a_k such that

$$\mathcal{Y}(\xi) = \sum_{k=0}^{m} a_k \mathcal{Q}_k(\xi).$$

The a_k's are functions of k and any parameter involved in $\mathcal{Y}(\xi)$.

Theorem 4. [25]There exists a unique polynomial $P_n(\xi)$ of degree n, which assumes prescribed values at $n+1$ distinct points $\xi_0 < \xi_1 < \cdots < \xi_n$.

Theorem 5. [25] Given any interval $a \leq \xi \leq b$, real number $\varepsilon > 0$ and any real valued continuous function $f(\xi)$ on $a \leq \xi \leq b$, there exists a polynomial $P(\xi)$ such that

$$|| \, f(\xi) - P(\xi) \, || < \varepsilon.$$

To study the convergence behavior of orthogonal collocation scheme, the definition given by [21] is quoted here:
Consider a family of mathematical problems parametrized by singular perturbation parameter ε, where ε lies in the semiopen interval $0 < \varepsilon \leq 1$. Assume that each problem in the family has a unique solution denoted by y_ε, and that each y_ε is approximated by a sequence of numerical solutions $\left\{(Y_\varepsilon, \Omega^N)\right\}_{N=1}^{\infty}$, where Y_ε is defined on the Ω^N representing the set of points in $\boldsymbol{R}$, and N is the discretization parameter. Then the numerical solutions Y_ε are said to converge to the exact solution y_ε, if there exists a positive integer N_0 and positive numbers G and p, where N_0, G, and p are all independent of N and ε, such that for all $N \geq N_0$

$$sup_{0<\varepsilon\leq 1}\left|Y_\varepsilon - y_\varepsilon\right|_{\Omega^N} \leq GN^{-p}.$$

Here p is the rate of convergence and G is the error constant. It shows that the rate of convergence in the case of the collocation technique depends upon the number of collocation points.

Let $y(\xi, t)$ be the exact solution, and let $y_h(\xi, t)$ be the approximate solution. The absolute error between the exact and approximate solution is calculated as

$$E_a = \mid y(\xi, t) - y_h(\xi, t) \mid . \tag{15}$$

The error in terms of L_2-norm and L_∞-norm has been calculated with respect to the weight function $w(\xi)$ such that

$$||y - y_h||_2^2 = \sum_{i=1}^{n+1} |w_i(\xi)(y - y_h)_i^2|, \tag{16}$$

where $y(\xi, t)$ represent analytic solutions and $y_h(\xi, t)$ represent approximate solutions [22]. The error between exact and numerical values has been shown by $e = y - y_h$.

L_2-norm is said to converge to the exact solution if $||y - y_h||_2 \to 0$ as $n \to \infty$. Thus

$$||e||_2 = ||y - y_h||_2, \tag{17}$$

Similarly, in L_∞-norm for $||y - y_h||$, it has been taken as

$$||y - y_h||_\infty = \max |(y - y_h)_i|, \qquad i = 1, 2, 3, \ldots, n+1, \tag{18}$$

$$||e||_\infty = ||y - y_h||_\infty. \tag{19}$$

6 Numerical examples

To verify the applicability of BCM, the scheme has been applied on different nonlinear hyperbolic equations.

Example 1. Consider the sine-Gordon nonlinear hyperbolic equation from the generalized (2) coupled form of two interacting configurations $y(\xi,t)$ and $\frac{\partial y}{\partial t}(\xi,t) = z(\xi,t)$:

$$\frac{\partial^2 y}{\partial t^2} = \frac{\partial^2 y}{\partial \xi^2} - \sin(y),$$

by defining

$$\frac{\partial y}{\partial t} = z,$$

$$\frac{\partial z}{\partial t} = \frac{\partial^2 y}{\partial \xi^2} - \sin(y),$$

with respect to the initial conditions

$$y(\xi,0) = 4\tan^{-1}(\exp(g\xi)),$$

$$z(\xi,0) = -4cg\frac{\exp(g\xi)}{(1+\exp(2g\xi))}.$$

Then exact solutions of the above equations have been obtained as

$$y(\xi,t) = 4\tan^{-1}(\exp(g(\xi - ct))),$$

where $\xi \in [-3,3]$, $g = \frac{1}{(1-c^2)^{\frac{1}{2}}}$, and $c = 0.5$, and boundary conditions can be extracted from the exact solution [9, 23, 30].

The values of $y(\xi,t)$ are calculated for different numbers of collocation points and compared with the exact solution in Table 1. The numerical values have been presented for a fixed value of $\xi = 0$ and different values of time in Table 1. The numerical results are found to be enough close to the exact solution. It is also observed that no particular change occurs in numerical values after 25 collocation points. A graphical representation of experimental results with respect to the time and collocation points has been presented in Figure 1.

A comparison of error in terms of L_2-norm and L_∞-norm at different numbers of collocation points has also been given in Table 2 for different values of time.

A graphical representation of error in terms of L_2-norm and L_∞-norm at 25 and 31 collocation points with respect to the time values have been presented in Figures 2 and 3, respectively. It can be analyzed from Figures 2 and 3 that error decreases with the increase in collocation points. From these figures, it can also be analyzed that the decrease in error is not so large and almost similar at 25 and 31 collocation points.

Numerical results have also been compared with [9, 23, 30] and have been discussed in Table 3. It is observed that the results obtained by BCM are better and give less error.

Example 2. The sine-Gordon nonlinear hyperbolic equation from the generalized (2) in coupled form of two interacting configurations $y(\xi, t)$ and $z(\xi, t)$ with respect to the initial conditions

$$y(\xi, 0) = 0,$$

$$z(\xi, 0) = 4sech(\xi),$$

the exact solution of equation has been obtained as

$$y(\xi, t) = 4\tan^{-1}(t.sech(\xi)),$$

$$z(\xi, t) = 4\frac{sech(\xi)}{1 + t^2 sech^2(\xi)}.$$

The boundary conditions have been taken from the exact solutions [9, 23, 30].

The numerical values calculated at 17 and 19 collocation points have been compared with the exact values and are presented in Table 4 for fixed $\xi = 0.5$ but at different values of the time. It has been observed from Table 4 that the numerical results are close enough to the exact solutions. It is also observed that no particular change in numerical values occurs after 17 collocation points. The graphical representation of numerical values with respect to the time and collocation points has been presented in Figure 4.

A comparison of error in terms of L_2-norm and L_∞-norms at 17 and 19 collocation points has been given in Table 5 at different time levels. The graphical representation of error in terms of L_2-norm and L_∞-norm at 17 and 19 collocation points with respect to the time has been presented in Figures 5 and 6, respectively. It is observed from Figures 5 and 6 that at 17 and 19 collocation points, it does not make much difference in numerical results, but if still count the difference at 19 collocation points, it shows better results than 17 collocation points.

A comparison of numerical results with [9, 23, 30] has been discussed in Table 6 and found to be close enough to be accepted.

7 Conclusion

The given nonlinear sine-Gordon equation has been solved successfully by using the BCM over Chebeshev collocation points. By the above analysis, the proposed method of BCM is proved to have some desired and popular features, such as high order accuracy and preserving energy conservation. Consistency and convergence of the computational technique have been obtained by computing the results of numerical solutions with analytic solutions. Error

analysis in terms of L_2- and L_∞-norms with respect to the weight function employed showed that the Bessel collocation approach is very stable, and the results obtained by this approach are consistent and convergent.

Table 1: Comparison of absolute error (E_a) of Example 1 at $\xi = 0$ for different numbers of collocation points

t	E_a at 8 collocation points	E_a at 16 collocation points	E_a at 25 collocation points	E_a at 31 collocation points
0.1	1.2362e-02	3.6988e-04	1.3449e-07	4.0068e-07
0.2	2.4923e-02	7.4329e-04	9.8035e-07	1.9933e-07
0.3	3.7862e-02	1.1126e-03	2.8106e-06	1.9219e-06
0.4	5.1315e-02	1.4626e-03	5.2349e-06	1.9219e-06
0.5	6.5366e-02	1.7749e-03	7.3092e-06	5.0514e-06
0.6	8.0030e-02	2.0225e-03	7.9431e-06	1.6903e-06
0.7	9.5257e-02	2.1711e-03	6.4173e-06	1.3360e-06
0.8	1.1093e-01	2.1925e-03	2.7694e-06	5.4602e-06
0.9	1.2689e-01	2.0728e-03	2.1323e-06	5.0533e-06
1.0	1.4291e-01	1.8073e-03	6.8670e-06	6.3394e-06

Table 2: Comparison of error for $y(\xi, t)$ of Example 1 at different collocation points

t	At 25 collocation points		At 31 collocation points	
	$\|\|e\|\|_\infty$	$\|\|e\|\|_2$	$\|\|e\|\|_\infty$	$\|\|e\|\|_2$
0.1	6.7963e-07	4.3658e-07	1.2869e-06	0.0000e-00
0.2	2.4566e-06	1.2844e-06	3.5131e-06	1.2717e-07
0.3	4.7272e-06	1.9502e-06	4.9638e-06	1.1174e-07
0.4	6.7848e-06	2.5348e-06	6.4823e-06	1.6430e-07
0.5	7.3091e-06	3.7359e-06	7.1190e-06	2.9715e-07
0.6	7.9430e-06	5.3058e-06	7.7593e-06	2.8624e-07
0.7	6.8875e-06	6.2351e-06	7.9344e-06	1.3433e-07
0.8	7.9192e-06	5.8165e-06	8.1705e-06	0.0000e-00
0.9	9.8062e-06	4.2143e-06	8.1533e-06	2.8093e-07

Table 3: Comparison of $||e||_2$ and $||e||_\infty$ calculated by Bessel collocation for $y(\xi, t)$ with different techniques of Example 1

t	Dehgan & Shokri [9]		Mittal & Bhatia [23]		Shukla & Tamsir [30]		Bessel collocation	
	$\|\|e\|\|_\infty$	$\|\|e\|\|_2$	$\|\|e\|\|_\infty$	$\|\|e\|\|_2$	$\|\|e\|\|_\infty$	$\|\|e\|\|_2$	$\|\|e\|\|_\infty$	$\|\|e\|\|_2$
0.25	4.95e-06	1.76e-05	4.90e-05	3.66e-05	9.61e-06	5.67e-06	4.44e-06	1.46e-07
0.50	8.42e-06	4.31e-05	7.55e-05	9.00e-05	1.10e-05	8.39e-06	7.38e-06	2.97e-07
0.75	1.65e-05	8.25e-05	1.43e-04	1.60e-04	1.26e-05	1.05e-05	7.99e-06	0.00e-00
1.00	2.51e-05	1.27e-04	2.10e-04	2.27e-04	1.44e-05	1.24e-05	1.49e-05	2.81e-07

Table 4: Comparison of absolute error (E_a) of Example 2 at $\xi = 0.5$ for different numbers of collocation points.

t	E_a at 8 collocation points	E_a at 16 collocation points	E_a at 17 collocation points	E_a at 19 collocation points
0.1	4.6603e-06	1.0754e-10	4.4239e-10	5.3072e-11
0.2	2.9466e-05	8.3158e-10	6.1702e-11	2.3970e-10
0.3	6.5426e-05	2.2862e-09	1.1462e-09	6.9454e-10
0.4	7.9905e-05	3.3772e-09	2.5359e-09	1.1188e-09
0.5	5.2043e-05	3.7229e-09	3.3682e-09	1.4447e-09
0.6	4.2174e-05	3.7595e-09	3.1963e-09	1.4496e-09
0.7	3.7627e-05	3.0807e-09	2.0244e-09	1.0735e-09
0.8	3.1964e-05	7.4804e-10	3.7364e-10	4.0962e-10
0.9	2.4734e-05	3.0871e-09	1.3676e-09	4.7658e-10
1.0	3.1170e-05	6.9712e-09	3.3311e-09	1.5792e-09

Table 5: Comparison of error for $y(\xi, t)$ of Example 2 at different collocation points

t	At 17 collocation points		At 19 collocation points	
	$\|\|e\|\|_\infty$	$\|\|e\|\|_2$	$\|\|e\|\|_\infty$	$\|\|e\|\|_2$
0.1	3.6956e-10	2.0235e-10	1.2951e-10	7.3902e-11
0.2	7.5355e-10	4.0045e-10	4.6931e-10	2.9364e-10
0.3	2.5508e-09	1.6149e-09	1.3351e-09	8.3540e-10
0.4	4.2276e-09	2.6935e-09	1.9375e-09	1.2585e-09
0.5	4.5143e-09	3.0129e-09	2.0691e-09	1.4681e-09
0.6	3.0879e-09	2.4971e-09	1.5904e-09	1.3228e-09
0.7	2.2236e-09	1.4627e-09	1.2266e-09	9.1829e-10
0.8	2.2897e-09	1.1368e-09	8.0804e-10	5.1226e-10
0.9	4.4816e-09	2.4093e-09	1.5020e-09	7.6839e-10

Table 6: Comparison of $||e||_2$ and $||e||_\infty$ calculated by Bessel collocation for $y(\xi, t)$ with different techniques of Example 2

t	Dehgan& Shokri [9]		Mittal& Bhatia [23]		Shukla& Tamsir [30]		Bessel collocation	
	$\|\|e\|\|_\infty$	$\|\|e\|\|_2$	$\|\|e\|\|_\infty$	$\|\|e\|\|_2$	$\|\|e\|\|_\infty$	$\|\|e\|\|_2$	$\|\|e\|\|_\infty$	$\|\|e\|\|_2$
0.25	5.89e-06	3.91e-05	2.32e-05	1.18e-05	5.46e-06	2.43e-06	9.06e-10	5.64e-10
0.50	2.01e-05	1.30e-04	4.11e-05	4.19e-05	7.39e-06	5.54e-06	1.88e-09	1.44e-09
0.75	3.63e-05	2.35e-04	1.02e-04	7.78e-05	7.78e-06	6.45e-06	1.02e-09	6.87e-10
1.00	5.07e-05	3.27e-04	1.64e-04	1.30e-04	8.75e-06	7.84e-06	2.54e-09	1.53e-09

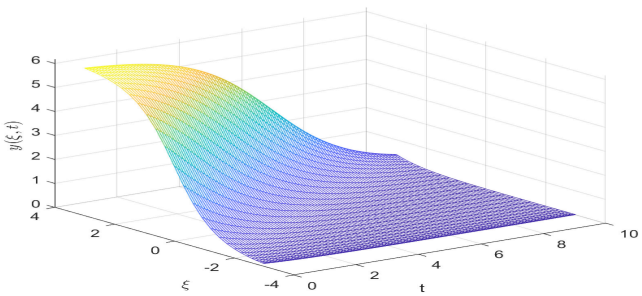


Figure 1: Graphical representation of $y(\xi, t)$ of Example 1

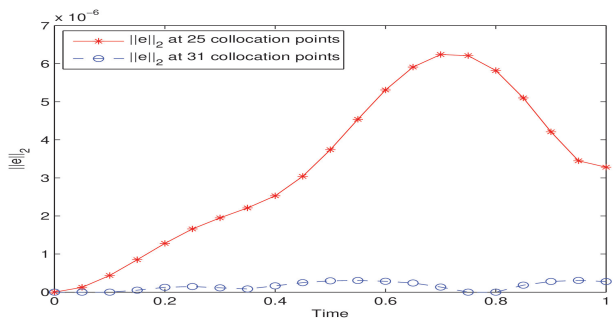


Figure 2: Graphical representation of comparison of error in form of L_2-norm with respect to time and number of collocation points of Example 1

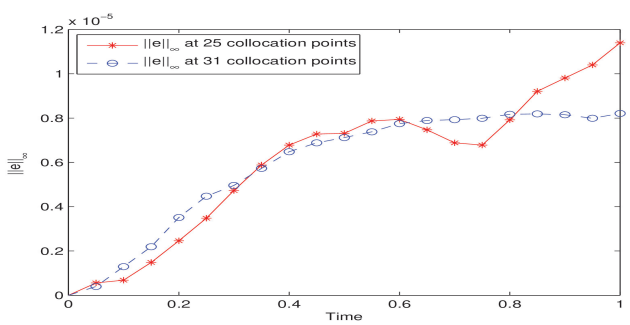


Figure 3: Graphical representation of comparison of error in form of L_∞-norm with respect to time and number of collocation points of Example 1

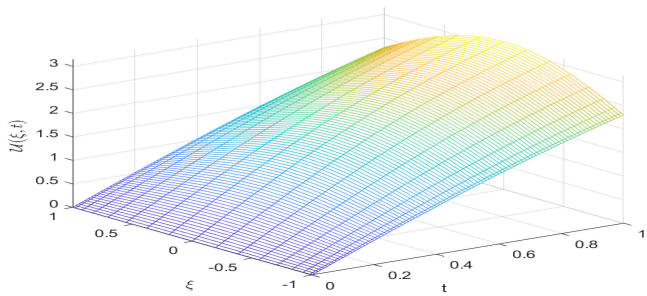


Figure 4: Graphical representation of $y(\xi, t)$ of Example 2

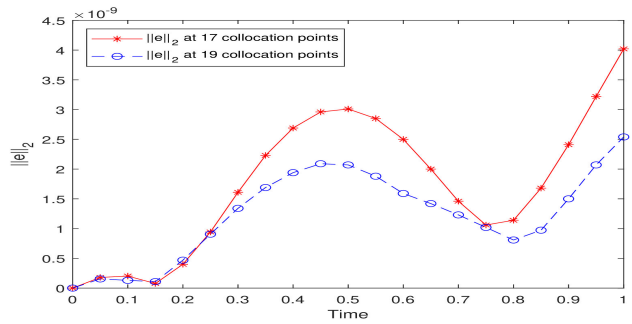


Figure 5: Graphical representation of comparison of error in form of L_2-norm with respect to time and number of collocation points of Example 2

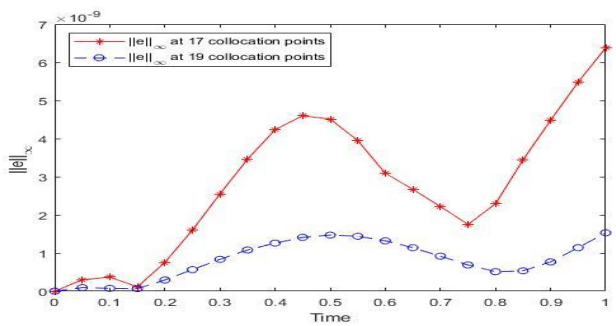


Figure 6: Graphical representation of comparison of error in form of L_∞-norm with respect to time and number of collocation points of Example 2

References

[1] Ablowitz, M.J., Herbst, B.M. and Schober, C.M. *Numerical simulation of quasi-periodic solutions of the sine-Gordon equation*, Phys. D

87 (1995), 37–47.

[2] Ablowitz, M.J., Herbst, B.M. and Schober, C.M. *On the numerical solution of the sine-Gordon equation*, J. Comput. Phys. 131 (1997), 354–367.

[3] Ablowitz, M.J., Kruskal, M.D. and Ladik, J.F. *Solitary wave collisions*, SIAM J. Appl. Math. 36 (1979), 428–443.

[4] Abramowitz, M. and Stegun, I.A. *Handbook of mathematical functions with formulas and mathematical tables*, with corrections Superintendent of Documents. National Bureau of Standards Applied Mathematics Series, No. 55 U. S. Government Printing Office, Washington, D.C., 1965,

[5] Argyris, J. and Haase, M. *An engineer's guide to soliton phenomena: application of the finite element method*, Comput. Meth. Appl. Mech. Eng. 61(1) (1987), 71–122.

[6] Arora, S., Dhaliwal, S.S. and Kukreja, V.K. *Solution of two point boundary value problems using orthogonal collocation on finite elements*, Appl. Math. Comput. 171 (2005), 358–370.

[7] Bailey, W.N. *Generalized Hypergeometric Series*, Cambridge Tracts in Mathematics and Mathematical Physics, No. 32 Stechert-Hafner, Inc., New York, 1964,

[8] Barone, A., Esposito, F., Magee, C.J., and Scott, A.C. *Theory and applications of the sine-Gordon equation*, Riv. Nuovo Cimento. 1 (1971), 227–267.

[9] Dehghan, M. and Shokri, A. *A numerical method for one-dimensional nonlinear Sine-Gordon equation using collocation and radial basis functions*, Numer. Methods. Partial. Differ. Eq. 24(2) (2008), 687–698.

[10] Evans, W.D., Everitt, W.N., Kwon, K.H. and Littlejohn, L.L. *Real orthogonalizing weights for Bessel polynomials*, J. Comput. Appl. Math. 49 (1993), 51–57.

[11] Everitt, W.N. and Markett, C. *On a generalization of Bessel functions satisfying higher-order differential equations*, J. Comput. Appl. Math. 54 (1994), 325–349.

[12] Ferguson, N.B. and Finlayson, B.A. *Transient chemical reaction analysis by orthogonal collocation*, Chem. Eng. J. 1(4) (1970), 327–336.

[13] Finlayson, B.A. *Packed bed reactor analysis by orthogonal collocation*, Chem. Eng. Sci. 26 (1971), 1081–1091.

[14] Gautschi, W. *Numerical Analysis*, Second Edition, Springer-Verlag, New York, 2012.

[15] Ilati M. and Dehghan, M. *The use of radial basis functions (RBFs) collocation and RBF-QR methods for solving the coupled nonlinear sine-Gordon equations*, Eng. Anal. Bound. Elem. 52 (2015), 99–109.

[16] Jianga, C., Sun, J., Li, H. and Wang, Y. *A fourth-order AVF method for the numerical integration of sine-Gordon equation*, Appl. Math. Comput. 313 (2017), 144–158.

[17] Koornwinder, T.H. *Orthogonal polynomial with weight function* $(1-x)^{\alpha}(1-x)^{\beta}+M\delta(x-1)+N\delta(x-1)$, Can. Math. Bull. 27(2) (1984), 205–214.

[18] Kumar, D., Singh, J., Kumar and S., Sushila *Numerical computation of Klein–Gordon equations arising in quantum field theory by using homotopy analysis transform method*, Alex. Eng. J. 53 (2014), 469–474.

[19] Martin-Vergara, F., Rus, F. and Villatoro, F.R. *Padé numerical schemes for the sine-Gordon equation*, Appl. Math. Comput. 358 (2019), 232–243.

[20] McLachlan, N.W. *Bessel functions for engineers*, University of Illinois, Oxford University Press, London, England. 1961.

[21] Miller, J.J.H., O'Riordan, E. and Shishkin, G.I. *Fitted Numerical methods for singular perturbation problems, error estimate in the maximum norm for linear problems in one and two dimensions*, World Scientific, 1996.

[22] Mishra, P., Sharma, K.K., Pani, A.K. and Fairweather, G. *Orthogonal spline collocation for singularly perturbed reaction diffusion problems in one dimension*, Int. J. Numer. Anal. Model. 16(4) (2019), 647–667.

[23] Mittal, R.C., and Bhatia, R. *Numerical solution of nonlinear Sine-Gordon equation by modified cubic B-Spline collocation method*, Int. J. Partial Differ. Equ. 2014 (2014), 1–8.

[24] Perring, J.K. and Skyrme, T.H.R. *A model unified field equation*, Nucl. Phys. 31 (1962), 550–555.

[25] Prenter, P.M. *Splines and variational methods*, Wiley-Interscience [John Wiley & Sons], New York-London-Sydney, 1975.

[26] Rainville, E.D. *Special functions*, Reprint of 1960 first edition. Chelsea Publishing Co., Bronx, N.Y., 1971.

[27] Saray, B.N., Lakestani, M. and Cattani, C. *Evaluation of mixed Crank-Nicolson scheme and tau method for the solution of Klein-Gordon equation*, App. Math. Comput. 331 (2018), 169–181.

[28] Scott, A.C., Chu, F.Y.F. and McLaughlin, D.W. *The soliton: A new concept in applied science*, Proc. IEEE 61 (1973), 1443–1483.

[29] Shan, Y., Liu, W. and Wu, B. *Space-time Legendre-Gauss-Lobatto collocation method for two-dimensional generalized sine-Gordon equation*, Appl. Numer. Math. 122 (2017), 92–107.

[30] Shukla, H.S. and Tamsir M. *Numerical solution of nonlinear Sine–Gordon equation by using the modified cubic B-spline differential quadrature method*, Beni-Seuf Univ. J. Basic Appl. Sci. 7 (2018), 359–366.

[31] Shukla, H.S., Tamsir, M. and Srivastava, V. K. *Numerical simulation of two dimensional sine-Gordon solitons using modified cubic B-spline differential quadrature method*, AIP Adv. 5(1) (2015), 017121.

[32] Sirendaoreji *A new auxiliary equation and exact travelling wave solutions of nonlinear equations*, Phys. Lett. A 356(2) (2006), 124–130.

[33] Sirendaoreji *Auxiliary equation method and new solutions of Klein–Gordon equations*, Chaos Solitons Fractals 31(4) (2007), 943–950.

[34] Sneddon, I. N. *Special functions of mathematical physics and chemistry,* Oliver and Boyd, Edinburgh-London; Interscience Publishers, Inc., New York, 1956.

[35] Stempak, K. *A weighted uniform L^p-estimate of Bessel functions: a note on a paper of K. Guo: "A uniform L^p estimate of Bessel functions and distributions supported on S^{n-1}" [Proc. Amer. Math. Soc. 125 (1997), no. 5, 1329–1340; MR1363462 (97g:46047)],* Proc. Amer. Math. Soc. 128(10) (2000), 2943–2945.

[36] Strauss, W. and Vazquez, L. *Numerical solution of a nonlinear Klein-Gordon equation*, J. Comput. Phys. 28 (1978), 271–278.

[37] Tasbozan, O., Yagmurlu, N.M.,Ucar, Y. and Esen, A. *Numerical solutions of the Sine-Gordon equation by collocation method*, Sohag J. Math. 3(1) (2016), 1–6.

[38] Villadsen, J.V. and Stewart, W.E. *Solution of boundary value problem by orthogonal collocation*, Chem. Eng. Sci. 22 (1967), 1483–1501.

[39] Watson, G.N. *A treatise on the theory of Bessel functions*, Cambridge University Press, Cambridge, England; The Macmillan Company, New York, 1944.

[40] Wazwaz, A.M. *New travelling wave solutions to the Boussinesq and the Klein–Gordon equations*, Commun. Nonlinear Sci. Numer. Simul. 13(5) (2008), 889–901.

[41] Wingate, C. *Numerical search for a ϕ^4 breather mode*, SIAM J. Appl. Math. 43 (1983), 120–140.

[42] Yüzbaşı, Ş. *A numerical approach for solving a class of the nonlinear Lane–Emden type equations arising in astrophysics*, Math. Methods Appl. Sci. 34(8) (2011), 2218–2230.

[43] Yüzbaşı, Ş. *A numerical approach for solving the high-order linear singular differential-difference equations*, Comput. Math. Appl. 62(5) (2011), 2289–2303.

[44] Yüzbaşı, Ş. *A numerical approximation based on the Bessel functions of first kind for solutions of Riccati type differential-difference equations*, Comput. Math. Appl. 64(6) (2012), 1691–1705.

[45] Yüzbaşı, Ş. *Bessel collocation approach for solving continuous population models for single and interacting species*, Appl. Math. Model. 36 (2012), 3787–3802.

[46] Yüzbaşı, Ş., Şahin, N. and Sezer, M. *A Bessel polynomial approach for solving linear neutral delay differential equations with variable coefficients*, J. Adv. Res. Appl. Math. 3(1) (2011), 81–101.

[47] Yüzbaşı, Ş., Şahin, N. and Sezer, M. *Bessel matrix method for solving high-order linear Fredholm integro-differential equations*, J. Adv. Res. Appl. Math. 3(2) (2011), 23–47.

[48] Yüzbaşı, Ş., Şahin, N. and Sezer, M. *Numerical solutions of systems of linear Fredholm integro-differential equations with Bessel polynomial bases*, Comput. Math. Appl. 61(10) (2011), 3079–3096.

How to cite this article
Arora. S and Bala. I, Numerical study of sine-Gordon equations using Bessel collocation method. *Iran. J. Numer. Anal. Optim.*, 2023; 13(4): 728-746. https://doi.org/10.22067/ijnao.2023.81484.1229

Iranian Journal of Numerical Analysis and Optimization
Vol. 13, No. 4, 2023, pp 747–762
https://doi.org/10.22067/ijnao.2023.78096.1165
https://ijnao.um.ac.ir/

Research Article

Optimal control analysis for modeling HIV transmission

K. R. Cheneke

Abstract

In this study, a modified model of HIV with therapeutic and preventive controls is developed. Moreover, a simple evaluation of the optimal control problem is investigated. We construct the Hamiltonian function by way of integrating Pontryagin's maximal principle to achieve the point-wise optimal solution. The effects obtained from the version analysis strengthen public health education to a conscious population, PrEP for early activation of HIV infection prevention, and early treatment with artwork for safe life after HIV infection. Moreover, numerical simulations are done using the MATLAB platform to illustrate the qualitative conduct of the HIV infection. In the end, we receive that adhering to ART protective prone people, the usage of PrEP along with different prevention control is safer control measures.

AMS subject classifications (2020): Primary 45D05; Secondary 42C10, 65G99.

Keywords: HIV; Optimal control problem; Basic reproduction number, Numerical simulation.

1 Introduction

Human immunodeficiency virus (HIV), the cause of HIV infection, has no curative medication until now [2]. Moreover, the long-time existence of the virus in the body leads to a serious infection called acquired immunodeficiency syndrome (AIDS) disease [6]. However, optimal controls are the effective way to combat HIV transmission and progression in the community [2, 6]. Public health education, condom, and anti-retrovirus therapy are the

Received 5 August 2022; revised 26 May 2023; accepted 16 June 2023

Kumama Regassa Cheneke
Department of Mathematics, Wollega University, Nekemte, Ethiopia. e-mail: kumamaregassa@gmail.com

major measures taken by both governmental and nongovernmental institutions to stop further progression and transmission of HIV in the populations [1, 3, 12, 4, 5, 7, 8, 9]. Moreover, effective pre-exposure prophylaxis (PrEP) is the drug used to prevent the survival of HIV in the human blood [14]. On the other hand, the abstinence of sexual practices through the activation of public health education reduces the fate of acquiring HIV from potentially infectious individuals. Mathematical models are very important tools to describe the behavior of biological events. Particularly, with the great contribution of Pontryagin's Maximum Principle (PMP) in the construction of optimal control problems, the nature of biological dynamics is studied intensively [13, 14, 15, 18, 19, 20, 10, 22, 23]. Based on the works done in [12], the motivation of this study is due to the significant contribution of public health education and prophylaxis in controlling the transmission of HIV infection among human individuals. Particularly, abstinence due to consolidated public health education builds positive awareness toward controlling oneself, whereas prophylaxis helps to prevent the progression of HIV in the human body. Mathematical models are important tools to control infections [24, 11, 27, 26, 16, 17]. In this study, we have included prophylaxis, antiretroviral therapy (ART), and prevention for controlling the transmission of HIV infection by modifying the model studied in [12].

2 Formulation of model

In this study, a mathematical model is formulated by classifying the total population into compartments of (i) Susceptible individuals (S), (ii) Individuals on Pre-exposure prophylaxis (E), (iii) HIV infected with primary stage (P), (iv) Not on treatment HIV infected individuals (J), (v) HIV Undetectable individuals (U), and (vi) On treatment HIV infected individuals (I).

Moreover, the subsequent assumptions are considered in the modeling of the infection (i) a new susceptible individuals becomes susceptible at recruitment rate of λ, (ii) individuals in S transfer to E due to taking PrEP at the rate of ρ; (iii) transmission rate of HIV infection from individuals in P to S is β_1 and transmission rate of HIV infection from I to S is β_2; (iv) individuals transfer from P to I at progression rate of ξ; (v) individuals transfer from P to J at transfer rate of η; (vi) individuals transfer from J to I at transfer rate of γ; (vii) individuals in the compartment J die due to infection at the rate ζ; (viii) individuals transfer from I to U due to adherence to ART at transferring rate of θ; (ix) individuals transfer from U to I due to default using of ART at the rate of ϕ; (x) natural induced death rate of all people is μ; (xi) AIDS induced death rate is δ; (xii) in this study, standard incidence rate is applied; (xiii) PrEP engagement effort is u_3; (xiv) Condom using effort is u_1; (xv) ART using effort is u_2.

The pictorial representation of the deterministic model with control measures is given in Figure 1.

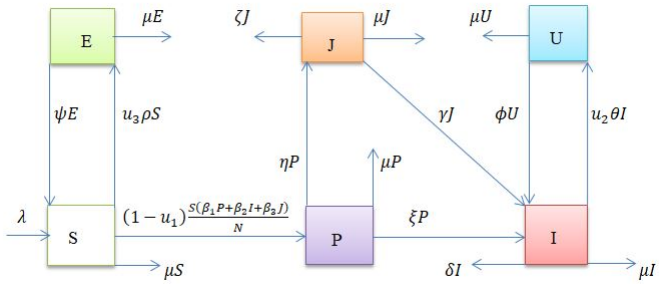


Figure 1: Schematic diagram of HIV transmission dynamics.

The deterministic model of population dynamics subject to HIV infection in the presence of control measures is given by

$$\begin{aligned}
\frac{dS}{dt} &= \lambda - \frac{(1-u_1)\,S\,(\beta_1 P + \beta_2 I + \beta_3 J)}{N} + \psi E - (u_3\rho + \mu)\,S,\\
\frac{dE}{dt} &= u_3\rho S - (\mu + \psi)\,E,\\
\frac{dP}{dt} &= \frac{(1-u_1)\,S\,(\beta_1 P + \beta_2 I + \beta_3 J)}{N} - (\xi + \eta + \mu)\,P,\\
\frac{dJ}{dt} &= \eta P - (\gamma + \zeta + \mu)J,\\
\frac{dI}{dt} &= \xi P + \gamma J + \phi U - (u_2\theta + \mu + \delta)\,I,\\
\frac{dU}{dt} &= u_2\theta I - (\phi + \mu)\,U,
\end{aligned} \tag{1}$$

with initial conditions: $S(0) \geq 0,\ E(0) \geq 0,\ P(0) \geq 0,\ J(0) \geq 0,\ I(0) \geq 0,\ U(0) \geq 0,\ 0 \leq u_i \leq 1,\ i = 1, 2, 3, 4.$

3 Analysis of the model without control

3.1 Invariant region

Theorem 1. The solution of model (1) is invariant in the region Ω proper-subset of six-dimensional space over the set of nonnegative real numbers such that

$$\Omega = \{(S, E, P, J, U, I) \in R_+^6 : N(0) \leq \frac{\lambda}{\mu}\}. \tag{2}$$

Proof. The equations of model (1) gives the subsequent equation:

$$\frac{dN}{dt} = \lambda - \mu N - \delta I - \zeta J,$$

which implies

$$\frac{dN}{dt} \leq \lambda - \mu N.$$

Applying mathematical procedures, the preceding inequality gives

$$N(t) \leq \frac{\lambda}{\mu} - \left(\frac{\lambda}{\mu} - N(0)\right) e^{-\mu t},$$

which implies, as time t varies, the total population size is bounded for all time t, with the given initial condition. □

3.2 Nonnegative property

Theorem 2. All solution variables of model (1) without control are nonnegative in the stated invariant region of the solution.

Proof. Consider the first equation of model (1) without control. Then

$$\frac{dS}{dt} = \lambda - \frac{S\left(\beta_1 P + \beta_2 I + \beta_3 J\right)}{N} + \psi E - \mu S, \tag{3}$$

which implies

$$\frac{dS}{dt} \geq -\frac{S\left(\beta_1 P + \beta_2 I + \beta_3 J\right)}{N} - \mu S. \tag{4}$$

Solving the preceding inequality, we get

$$S(t) \geq S(0) e^{-\mu t - \int_0^t \frac{(\beta_1 P(\xi) + \beta_2 I(\xi) + \beta_3 J(\xi))}{N(\xi)} d\xi}. \tag{5}$$

Hence, based on the initial condition, the susceptible population size is nonnegative for all time t. □

3.3 Basic reproduction number

The basic reproduction number R_0 of model (1) without control is the average number of infected individuals produced by typical infectious individuals in the susceptible population during the entire period of infection. Based on the techniques applied, we compute basic reproduction numbers from model (1) without control as follows. Let F and V be the Jacobian matrices obtained from model (1) as given below:

$$F = \begin{pmatrix} \beta_1 & \beta_2 & \beta_3 & 0 \\ 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 \end{pmatrix}, \quad V = \begin{pmatrix} \xi+\eta+\mu & 0 & 0 & 0 \\ -\eta & \gamma+\zeta+\mu & 0 & 0 \\ -\xi & -\gamma & \delta+\mu & -\phi \\ 0 & 0 & 0 & \phi+\mu \end{pmatrix}.$$

The spectral radius $\rho\left(FV^{-1}\right)$ computed from next-generation matrix FV^{-1} of foregoing matrices is given by

$$\begin{aligned} \rho\left(FV^{-1}\right) =& \frac{\beta_1}{\xi+\eta+\mu} + \frac{\beta_2\eta}{(\xi+\eta+\mu)(\gamma+\zeta+\mu)} \\ &+ \frac{\beta_3(\xi\mu+\xi\gamma+\xi\zeta+\eta\gamma)}{(\delta+\mu)(\xi+\eta+\mu)(\gamma+\zeta+\mu)}. \end{aligned}$$

Therefore, by the definition, we obtain

$$\begin{aligned} R_0 =& \frac{\beta_1}{\xi+\eta+\mu} + \frac{\beta_2\eta}{(\xi+\eta+\mu)(\gamma+\zeta+\mu)} \\ &+ \frac{\beta_3(\xi\mu+\xi\gamma+\xi\zeta+\eta\gamma)}{(\delta+\mu)(\xi+\eta+\mu)(\gamma+\zeta+\mu)}. \end{aligned}$$

3.4 Global stability of disease-free equilibrium

Theorem 3. The global stability of a disease-free equilibrium point is described as a steady state where the trajectory of solution shows the tendency of moving toward it for all time t.

Proof. To show the global stability of disease-free equilibrium, we incorporate the method applied in the works of [21]. Next, from the computed matrices for construction of next-generation, we obtain

$$F = \begin{pmatrix} \beta_1 & \beta_2 & \beta_3 & 0 \\ 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 \end{pmatrix}, \quad V = \begin{pmatrix} \xi+\eta+\mu & 0 & 0 & 0 \\ -\eta & \gamma+\zeta+\mu & 0 & 0 \\ -\xi & -\gamma & \delta+\mu & -\phi \\ 0 & 0 & 0 & \phi+\mu \end{pmatrix}.$$

Moreover, the rate of change of variables (P, J, I, U) at disease-free equilibrium can be written as

$$\begin{pmatrix} \frac{dP}{dt} \\ \frac{dJ}{dt} \\ \frac{dI}{dt} \\ \frac{dU}{dt} \end{pmatrix} \leq (F - V) \begin{pmatrix} P \\ J \\ I \\ U \end{pmatrix}.$$

Therefore, by the comparison method applied in [21], we justify that model (1) without control has a globally asymptotically stable disease-free equilibrium. □

4 Extension to control problem

The deterministic model of population dynamics subject to HIV infection, in the presence of control measures, is given by

$$\begin{aligned}
\frac{dS}{dt} &= \lambda - \frac{(1-u_1)\,S\,(\beta_1 P + \beta_2 I + \beta_3 J)}{N} + \psi E - (u_3\rho + \mu)\,S,\\
\frac{dE}{dt} &= u_3\rho S - (\mu+\psi)\,E,\\
\frac{dP}{dt} &= \frac{(1-u_1)\,S\,(\beta_1 P + \beta_2 I + \beta_3 J)}{N} - (\xi+\eta+\mu)\,P,\\
\frac{dJ}{dt} &= \eta P - (\gamma+\zeta+\mu)J,\\
\frac{dI}{dt} &= \xi P + \gamma J + \phi U - (u_2\theta + \delta + \mu)\,I,\\
\frac{dU}{dt} &= u_2\theta I - (\phi+\mu)\,U,
\end{aligned} \tag{6}$$

with initial conditions $S(0) \geq 0,\ E(0) \geq 0,\ P(0) \geq 0,\ J(0) \geq 0,\ I(0) \geq 0,\ U(0) \geq 0,\ 0 \leq u_i \leq 1,\ i = 1,2,3.$
To study the optimal levels of the controls, we define the Lebesgue measurable control set U as

$$U = \{(u_1, u_2, u_3) : 0 \leq u_1 \leq 1, 0 \leq u_2 \leq 1, 0 \leq u_3 \leq 1, 0 \leq t \leq t_f\}. \tag{7}$$

Our goal is to find the optimal controls that minimize objective functional J given by

$$J = \min_{(u_1,u_2,u_3)} \int_0^{t_f} c_1 P + c_2 I + c_3 J + \frac{1}{2}\left(w_1 u_1^2 + w_2 u_2^2 + w_3 u_3^2\right), \tag{8}$$

where $c_j,\ j = 1,2,3$ and $w_i\ , i = 1,2,3$ are constants. The expressions $0.5 w_i u_i^2, i = 1,2,3$ are costs associated with controls. The form of cost is quadratic because we assumed it to be nonlinear in nature [24]. Also, for four optimal controls u_1^*, u_2^*, u_3^*, we have

$$J(u_1^*, u_2^*, u_3^*) = \min\{J(u_1, u_2, u_3) : u_1, u_2, u_3 \in U\},$$

where $U = \{(u_1, u_2, u_3) : 0 \leq u_1 \leq 1, 0 \leq u_2 \leq 1, 0 \leq u_3 \leq 1\}$. Furthermore, $u_1, u_2 and u_3$ are measurable controls.

4.1 Existence of optimal control solution

Theorem 4. The optimal control solution of a control problem exists if the following Fleming's and Rishel's conditions are satisfied:

(i) The set of all solutions to optimal control problem and objective functional must be nonempty.

(ii) The state system is a linear function of controls with coefficients dependent on state variables and time.

(iii) The integrand in objective functional is convex and bounded above by $d_1(|u_1|^2+|u_2|^2+|u_3|^2)^d - d_2 \leq c_1P+c_2I+c_3J+\frac{1}{2}\left(w_1u_1^2+w_2u_2^2+w_3u_3^2\right), d_1 > 0$ and $d > 1$.

Proof. We employ the method from to demonstrate the existence of optimal control. The condition (i) is satisfied if the state system has bounded coefficients. Additionally, the state system operates in accordance with controls, satisfying requirement (ii). The integrand in the objective functional is used to demonstrate condition (iii). Moreover, $c_1P + c_2I + c_3J + \frac{1}{2}\left(w_1u_1^2 + w_2u_2^2 + w_3u_3^2\right)$ is convex on U as any constant, linear and quadratic are convex. Furthermore, assume that there are $d_1, d_2 > 0$, and $d > 1$ satisfying $d_1(|u_1|^2+|u_2|^2+|u_3|^2)^d - d_2 \leq c_1P+c_2I+c_3J+\frac{1}{2}\left(w_1u_1^2+w_2u_2^2+w_3u_3^2\right), d_1 = \min\{w_i, i = 1, 2, 3\}$, $d = 2$, and d_2 is the half of coefficient of control functions. Therefore, the optimal solution exists. □

4.2 The Hamiltonian and optimality system

The PMP stated the necessary conditions that are satisfied optimal pair. Hence, by this principle, we obtain the Hamiltonian function (H) defined as [24]

$$\begin{aligned} H(x,u,t) =& c_1P + c_2I + c_3J + \frac{1}{2}\left(w_1u_1^2 + w_2u_2^2 + w_3u_3^2\right) \\ &+ \lambda_1\frac{dS}{dt} + \lambda_2\frac{dE}{dt} + \lambda_3\frac{dP}{dt} + \lambda_4\frac{dJ}{dt} + \lambda_5\frac{dI}{dt} + \lambda_6\frac{dU}{dt}, \end{aligned}$$

where $\lambda_i, i = 1, 2, 3, 4, 5, 6$ are the adjoint variable corresponding to state variables S, E, P, J, I, and U, respectively, and to be determined using the PMP for the existence of optimal pairs.

Theorem 5. Let S, E, P, J, I, U be optimal state variables and let optimal control $u_i, i = 1, 2, 3$ be the optimal controls. Then there exist costate variables $\lambda_1, \lambda_2, \lambda_3, \lambda_4, \lambda_5, and \lambda_6$ that satisfy

$$\frac{d\lambda_1}{dt} = -\frac{\partial H}{\partial S}, \qquad \frac{d\lambda_2}{dt} = -\frac{\partial H}{\partial E},$$
$$\frac{d\lambda_3}{dt} = -\frac{\partial H}{\partial P}, \qquad \frac{d\lambda_4}{dt} = -\frac{\partial H}{\partial J},$$
$$\frac{d\lambda_5}{dt} = -\frac{\partial H}{\partial I}, \qquad \frac{d\lambda_6}{dt} = -\frac{\partial H}{\partial U},$$

with transversality or final time conditions $\lambda_1(t_f) = \lambda_2(t_f) = \lambda_3(t_f) = \lambda_4(t_f) = \lambda_5(t_f) = \lambda_6(t_f) = 0$, where H is the Hamiltonian function. Moreover, the optimal controls u_1^*, u_2^*, and u_3^* are $u_1^* = \min\{0, \max\{\frac{\beta SI(\lambda_2-\lambda_1)}{(w_1 N)}, 1\}\}$ and $u_2^* = \min\{0, \max\{\frac{\alpha I(\lambda_4-\lambda_2)}{(w_2 N)}, 1\}\}$, over the constraints $0 \le u_1 \le 1, 0 \le u_2 \le 1$.

Proof. The PMP gives the standard form of adjoint equation with transversality conditions. Now, differentiating the Hamiltonian function with respect to state variables, we have

$$\begin{aligned}
\frac{d\lambda_1}{dt} = -\frac{\partial H}{\partial S} =& (1-u_1)\frac{(\beta_1 P + \beta_2 I + \beta_3 J)\, N - S\,(\beta_1 P + \beta_2 I + \beta_3 J)}{N^2}\,(\lambda_1 - \lambda_3) \\
&+ u_3 \rho\,(\lambda_1 - \lambda_2) + \mu\lambda_1, \\
\frac{d\lambda_2}{dt} = -\frac{\partial H}{\partial E} =& \psi\,(\lambda_2 - \lambda_1) + \mu\lambda_2, \\
\frac{d\lambda_3}{dt} = -\frac{\partial H}{\partial P} =& -c_1 + \frac{(1-u_1)\,(\beta_1 SN - S\,(\beta_1 P + \beta_2 I + \beta_3 J))}{N^2}\,(\lambda_1 - \lambda_3) \\
&+ \xi\,(\lambda_3 - \lambda_5) + \eta\,(\lambda_3 - \lambda_4) + \mu\lambda_3, \\
\frac{d\lambda_4}{dt} = -\frac{\partial H}{\partial J} =& \frac{(1-u_1)\,(\beta_3 SN - S\,(\beta_1 P + \beta_2 I + \beta_3 J))}{N^2}\,(\lambda_1 - \lambda_3) \\
&+ \gamma\,(\lambda_4 - \lambda_5) + (\zeta + \mu)\lambda_4, \\
\frac{d\lambda_5}{dt} = -\frac{\partial H}{\partial I} =& -c_2 + \frac{(1-u_1)(\beta_2 SN - S(\beta_1 P + \beta_2 I + \beta_3 J))}{N^2}(\lambda_1 - \lambda_3) \\
&+ u_2\theta(\lambda_5 - \lambda_6) + (\delta + \mu)\lambda_5, \\
\frac{d\lambda_6}{dt} = -\frac{\partial H}{\partial U} =& \phi(\lambda_6 - \lambda_5) + \mu\lambda_6.
\end{aligned}$$

Furthermore, the characterization of optimal controls u_1^*, u_2^* and u_3^* shows that

$$\frac{\partial H}{\partial u_1} = \frac{\partial H}{\partial u_2} = \frac{\partial H}{\partial u_3} = 0.$$

Hence, optimal controls over $0 \le u_1 \le 1, 0 \le u_2 \le 1, 0 \le u_3 \le 1$ are given by

$$u_1^* = u_1 = \frac{S\,(\beta_1 P + \beta_2 I + \beta_3 J)\,(\lambda_3 - \lambda_1)}{w_1 N},$$

$$u_2^* = u_2 = \frac{\theta I\left(\lambda_5 - \lambda_6\right)}{w_2},$$
$$u_3^* = u_3 = \frac{\rho S\left(\lambda_1 - \lambda_2\right)}{w_3}.$$

Therefore, the bounds of the optimal control variables are given by

$$u_1^* = \begin{cases} \frac{S(\beta_1 P+\beta_2 I+\beta_3 J)(\lambda_3-\lambda_1)}{w_1 N} & \text{if } 0 < \frac{S(\beta_1 P+\beta_2 I+\beta_3 J)(\lambda_3-\lambda_1)}{w_1 N} < 1, \\ 0 & \text{if } \frac{S(\beta_1 P+\beta_2 I+\beta_3 J)(\lambda_3-\lambda_1)}{w_1 N} \leq 0, \\ 1 & \text{if } 1 \leq \frac{S(\beta_1 P+\beta_2 I+\beta_3 J)(\lambda_3-\lambda_1)}{w_1 N}, \end{cases}$$

$$u_2^* = \begin{cases} \frac{\theta I(\lambda_5-\lambda_6)}{w_2} & \text{if } 0 < \frac{\theta I(\lambda_5-\lambda_6)}{w_2} < 1, \\ 0 & \text{if } \frac{\theta I(\lambda_5-\lambda_6)}{w_2} \leq 0, \\ 1 & \text{if } 1 \leq \frac{\theta I(\lambda_5-\lambda_6)}{w_2}, \end{cases}$$

$$u_3^* = \begin{cases} \frac{\rho S(\lambda_1-\lambda_2)}{w_3} & \text{if } 0 < \frac{\rho S(\lambda_1-\lambda_2)}{w_3} < 1, \\ 0 & \text{if } \frac{\rho S(\lambda_1-\lambda_2)}{w_3} \leq 0, \\ 1 & \text{if } \quad 1 \leq \frac{\rho S(\lambda_1-\lambda_2)}{w_3}. \end{cases}$$

In a compact form, the optimal controls can be written as
$u_1^* = \min\{0, \max\{\frac{S(\beta_1 P+\beta_2 I+\beta_3 J)(\lambda_3-\lambda_1)}{w_1 N}, 1\}\}$,
$u_2^* = \min\{0, \max\{\frac{\theta I(\lambda_5-\lambda_6)}{w_2}, 1\}\}$,
$u_3^* = \min\{0, \max\{\frac{\rho S(\lambda_1-\lambda_2)}{w_3}, 1\}\}$. □

Moreover, the optimality system of the optimal control problem can be written as

$$\begin{aligned}
\frac{dS}{dt} =& \lambda - \frac{(1-u_1)\, S\left(\beta_1 P + \beta_2 I + \beta_3 J\right)}{N} + \psi E - \left(u_3 \rho + \mu\right) S, \\
\frac{dE}{dt} =& u_3 \rho S - \left(\mu + \psi\right) E, \\
\frac{dP}{dt} =& \frac{(1-u_1)\, S\left(\beta_1 P + \beta_2 I + \beta_3 J\right)}{N} - \left(\xi + \eta + \mu\right) P, \\
\frac{dJ}{dt} =& \eta P - (\gamma + \zeta + \mu) J, \\
\frac{dI}{dt} =& \xi P + \gamma J + \phi U - \left(u_2 \theta + \delta + \mu\right) I, \\
\frac{dU}{dt} =& u_2 \theta I - \left(\phi + \mu\right) U, \\
\frac{d\lambda_1}{dt} =& (1-u_1) \frac{\left(\beta_1 P + \beta_2 I + \beta_3 J\right) N - S\left(\beta_1 P + \beta_2 I + \beta_3 J\right)}{N^2} \left(\lambda_1 - \lambda_3\right) \\
& + u_3 \rho \left(\lambda_1 - \lambda_2\right) + \mu \lambda_1,
\end{aligned}$$

$$
\begin{aligned}
\frac{d\lambda_2}{dt} =& \psi\left(\lambda_2-\lambda_1\right)+\mu\lambda_2,\\
\frac{d\lambda_3}{dt} =& -c_1+\frac{(1-u_1)\left(\beta_1 SN-S\left(\beta_1 P+\beta_2 I+\beta_3 J\right)\right)}{N^2}\left(\lambda_1-\lambda_3\right)\\
&+\xi\left(\lambda_3-\lambda_5\right)+\eta\left(\lambda_3-\lambda_4\right)+\mu\lambda_3,\\
\frac{d\lambda_4}{dt} =& \frac{(1-u_1)\left(\beta_3 SN-S\left(\beta_1 P+\beta_2 I+\beta_3 J\right)\right)}{N^2}\left(\lambda_1-\lambda_3\right)\\
&+\gamma\left(\lambda_4-\lambda_5\right)+(\zeta+\mu)\lambda_4,\\
\frac{d\lambda_5}{dt} =& -c_2+\frac{(1-u_1)(\beta_2 SN-S(\beta_1 P+\beta_2 I+\beta_3 J))}{N^2}(\lambda_1-\lambda_3)\\
&+u_2\theta(\lambda_5-\lambda_6)+(\delta+\mu)\lambda_5,\\
\frac{d\lambda_6}{dt} =& \phi(\lambda_6-\lambda_5)+\mu\lambda_6,
\end{aligned}
$$

with $\lambda_1(t_f)=\lambda_2(t_f)=\lambda_3(t_f)=\lambda_4(t_f)=\lambda_5(t_f)=\lambda_6(t_f)=0, S(0)=S_0, P(0)=P_0, J(0)=J_0, I(0)=I_0, U(0)=U_0$.

4.3 Numerical simulations and discussion

4.3.1 Analysis using the numerical methods

In this study, the numerical methods are involved in simulating the general results of the analytical findings that give real meaning to both the mathematical and biological communities. Furthermore, the parameter values used in the simulation are either taken from the literature or assumed, as given in Table 1. Also, $w_1=50, w_2=20, w_3=30, c_1=5, c_2=25, T=20, S(0)=1000, H(0)=0, W(0)=300, I(0)=500, U(0)=0, A(0)=0$.
Moreover, MATLAB software is applied in the simulation process. Fractional derivatives and stochastic findings are widely applied as reviewed in this paper. Hence, we incorporate both forward and backward sweep methods of fourth-order Runge–Kutta method to simulate the results. The applied control strategies are as follows:
Strategy 1: Using together control measures u_1 and u_2.
Strategy 2: Using together control measures u_1 and u_3.
Strategy 3: Using together control measures u_2 and u_3.
Strategy 4: Using together control measures u_1, u_2, and u_3.
Moreover, we have used the parameters given in Table 1 to simulate subsequent numerical solutions.

Table 1: Parameter/constants value.

Parameter/constants	Value
λ	200
β_1	0.9915
β_2	0.75
β_3	0.9815
ξ	0.5
μ	0.02
η	0.5
ζ	0.1
ϕ	0.09
θ	0.5
δ	1
ρ	0 .1
γ	0.1
ψ	0.001

Based on the aforementioned control strategies, the following numerical simulations are performed.

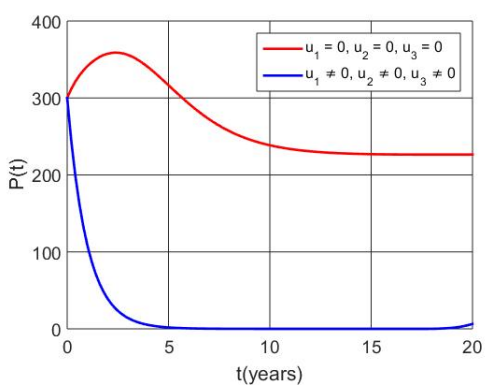


Figure 2: Primary HIV infected population.

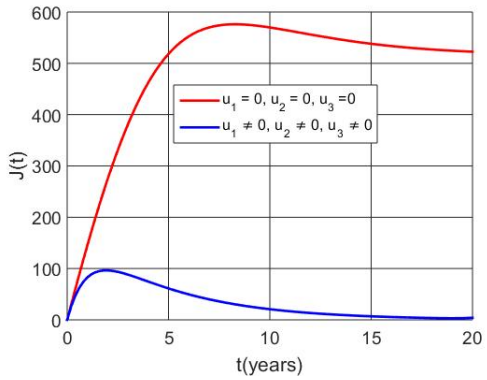


Figure 3: HIV not tested population.

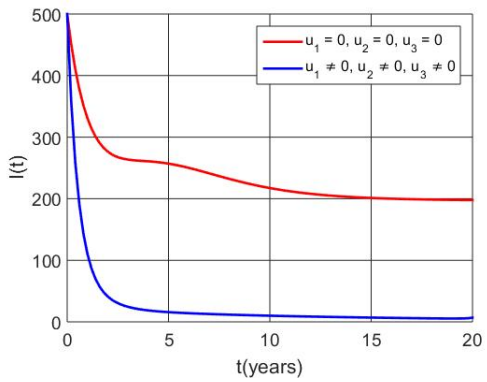


Figure 4: HIV infected and on treatment individuals.

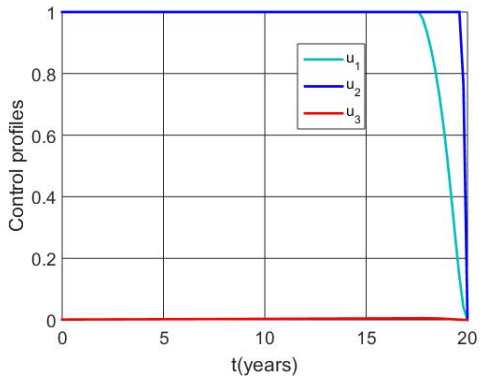


Figure 5: Control functions effect illustration.

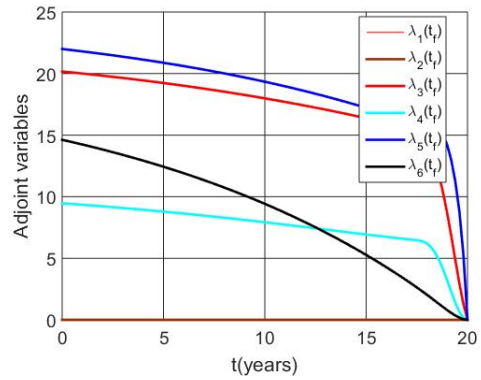


Figure 6: Adjoint variables condition descriptions.

4.3.2 Numerical results and discussion

This study develops and analyzes a mathematical model of HIV with the best possible control measures. The conceptual diagram for population dynamics is shown in Figure 1. The inclusion of the intervention with control serves to emphasize the significance of control measures in minimizing the effects of HIV infection. The numerical simulation results are shown in Figure 2 and show how a successful combination of control methods lowers the number of newly infected people. The numerical results in Figure 3 show that a reduction in the number of people who have not begun ART is shown when an intervention with control functions is present. Figure 4 shows a simulation of the number of HIV-positive people who are now receiving treatment. The results show that the intervention with three control groups dramatically lowers the number of people infected with HIV. When applied correctly, u_1 and u_2 are effective from the beginning to the end of initiation, as seen in Figure 5, where the applied control functions are simulated. Control u_3, on the other hand, makes a smaller contribution to regulating HIV infection dynamics because of its limited availability. In Figure 6, the adjoint variable is simulated to show that the transversal requirement has been satisfied.

5 Conclusion

According to the results of analytical and numerical simulations, adopting the best control measures to stop the further progression and transmission of HIV dynamics is more successful if done before the HIV infection even begins to spread. Additionally, maintaining ART and protecting those who are susceptible are considered the most crucial ways to lessen the effects of

HIV infection. Intervention with pre-exposure prophylaxis contributes less to lowering the risk of HIV infection since it is less affordable and accessible.

Acknowledgements

Authors are grateful to there anonymous referees and editor for their constructive comments.

References

[1] Ali Biswas, H. *On the evolution of AIDS/HIV treatment: an optimal control approach,* Curr. HIV Res. 12(1), (2014), 1–12.

[2] Akudibillah G., Pandey A., and Medlock J. *Optimal control for HIV treatment.* Math. Biosci. Eng. 16(1), (2018), 373–396.

[3] Arenas, A.J., González-Parra, G., Naranjo, J.J., Cogollo, M. and De La Espriella, N. *Mathematical analysis and numerical solution of a model of HIV with a discrete time delay,* Mathematics, 9(3), (2021), 257.

[4] Bakare, E.A. and Hoskova-Mayerova S.*Optimal control analysis of cholera dynamics in the presence of asymptotic transmission, Axioms* 10 (2), (2021), 60.

[5] Bórquez, A., Guanira, J.V., Revill, P., Caballero, P., Silva-Santisteban, A., Kelly, S., Salazar, X., Bracamonte, P., Minaya, P., Hallett, T.B. and Cáceres, C.F. *The impact and cost-effectiveness of combined HIV prevention scenarios among transgender women sex-workers in Lima, Peru: a mathematical modelling study,* The Lancet Public Health, 4(3), (2019), e127–e136.

[6] Boukhouima, A., Lotfi, E.M., Mahrouf, M., Rosa, S., Torres, D.F. and Yousfi, N. *Stability analysis and optimal control of a fractional HIV-AIDS epidemic model with memory and general incidence rate,* Eur. Phys. J. Plus, 136(1), (2021), 1–20.

[7] Camlin, C.S., Koss, C.A., Getahun, M., Owino, L., Itiakorit, H., Akatukwasa, C., Maeri, I., Bakanoma, R., Onyango, A., Atwine, F. and Ayieko, J. *Understanding demand for PrEP and early experiences of PrEP use among young adults in rural Kenya and Uganda: a qualitative study,* AIDS Behav. 24, (2020), 2149–2162.

[8] Campos, C., Silva, C.J. and Torres, D.F. *Numerical optimal control of HIV transmission in Octave/MATLAB,* Math. Comput. Appl. 25(1) (2020), Paper No. 1, 20 pp.

[9] Campos, N.G., Lince-Deroche, N., Chibwesha, C.J., Firnhaber, C., Smith, J.S., Michelow, P., Meyer-Rath, G., Jamieson, L., Jordaan, S., Sharma, M. and Regan, C. *Cost-effectiveness of cervical cancer screening in women living with HIV in South Africa: a mathematical modeling study.* Acquir. Immune Defic. Syndr. (1999), 79(2) (2018), 195.

[10] Cheneke, K.R. *Optimal Control and Bifurcation Analysis of HIV Model,* Comput. Math. Methods Med. (2023).

[11] Cheneke, K.R., Rao, K.P. and Edessa, G.K. *Application of a new generalized fractional derivative and rank of control measures on cholera transmission dynamics*, International Journal of Mathematics and Mathematical Sciences, 2021, (2021), 1–9.

[12] Cheneke, K.R., Rao, K.P. and Edessa, G.K. *Bifurcation and stabillity analysis of HIV transmission model with optimal control,* J. Math. 2021, (2021), 1–14.

[13] Choi, H., Suh, J., Lee, W., Kim, J.H., Kim, J.H., Seong, H., Ahn, J.Y., Jeong, S.J., Ku, N.S., Park, Y.S. and Yeom, J.S. *Cost-effectiveness analysis of pre-exposure prophylaxis for the prevention of HIV in men who have sex with men in South Korea: a mathematical modelling study,* Sci. Rep. 10(1), (2020), 14609.

[14] Đorđević, J. and Rognlien Dahl, K., 2022. *Stochastic optimal control of pre-exposure prophylaxis for HIV infection,* Math, Med. Biol. 39(3), (2022), 197–225.

[15] Ghosh, I., Tiwari, P.K., Samanta, S., Elmojtaba, I.M., Al-Salti, N. and Chattopadhyay, J. *A simple SI-type model for HIV/AIDS with media and self-imposed psychological fear,* Math. Biosci. 306 (2018), 160–169.

[16] Hattaf, K. and Yousfi, N. *Two optimal treatments of HIV infection model,* World J. Model. Simul. 8(1), (2012), 27–36.

[17] Hattaf, K. and Yousfi, N. *Optimal control of a delayed HIV infection model with immune response using an efficient numerical method,* Int. Sch. Res. Notices, (2012).

[18] Hidayat, N., R. B. E. Wibowo, et al., Marsudi, Hidayat, N. and Wibowo, R. B. E. *Optimal control and cost-effectiveness analysis of HIV model with educational campaigns and therapy,* Matematika (Johor) 35 (2019), Special issue, 123–138.

[19] Hove-Musekwa, S.D., Nyabadza, F., Mambili-Mamboundou, H., Chiyaka, C. and Mukandavire, Z. *Cost-effectiveness analysis of hospitalization and home-based care strategies for people living with HIV/AIDS: the case of Zimbabwe,* Int. Sch. Res. Notices, (2014).

[20] Huo, H.-F., Chen, R. and Wang, X.-Y. *Modelling and stability of HIV/AIDS epidemic model with treatment,* Appl. Math. Model. 40(13-14) (2016), 6550–6559.

[21] Huo, H.-F. and Li-Xiang, F. *Global stability for an HIV/AIDS epidemic model with different latent stages and treatment.* Applied Mathematical Modelling, 37(3) (2013), 1480–1489.

[22] Khajanchi, S., Bera, S. and Roy, T.K. *Mathematical analysis of the global dynamics of a HTLV-I infection model, considering the role of cytotoxic T-lymphocytes,* Math. Comput. Simul. 180 (2021), 354–378.

[23] Marsudi, M., Hidayat, N. and Wibowo, R.B.E. *Application of Optimal Control Strategies for the Spread of HIV in a Population,* J. Life Sci. Res. 4(1) (2017), 1–9.

[24] Marsudi, T., Suryanto, A. and Darti, I. *Global stability and optimal control of an HIV/AIDS epidemic model with behavioral change and treatment.* Eng. Lett. 29(2) (2021).

[25] Naik, P.A., Yavuz, M., Qureshi, S., Zu, J. and Townley, S. *Modeling and analysis of COVID-19 epidemics with treatment in fractional derivatives using real data from Pakistan,* Eur. Phys. J. Plus, 135 (2020), 1–42.

[26] Olaniyi, S., Obabiyi, O.S., Okosun, K.O., Oladipo, A.T. and Adewale, S.O. *Mathematical modelling and optimal cost-effective control of COVID-19 transmission dynamics,* Eur. Phys. J. Plus, 135(11) (2020), 938.

[27] Silva, C.J. and Torres, D.F. *Modeling and optimal control of HIV/AIDS prevention through PrEP.* arXiv preprint arXiv:1703.06446 (2017).

How to cite this article
Cheneke, K. R., Optimal control analysis for modeling HIV transmission. *Iran. J. Numer. Anal. Optim.*, 2023; 13(4): 747-762. https://doi.org/10.22067/ijnao.2023.78096.1165

Iranian Journal of Numerical Analysis and Optimization
Vol. 13, No. 4, 2023, pp 763–774
https://doi.org/10.22067/ijnao.2023.82361.1260
https://ijnao.um.ac.ir/

Research Article

Improving the performance of the FCM algorithm in clustering using the DBSCAN algorithm†

S. Barkhordari Firozabadi, S.A. Shahzadeh Fazeli*, J. Zarepour Ahmadabadi and S.M. Karbassi

Abstract

The fuzzy-C-means (FCM) algorithm is one of the most famous fuzzy clustering algorithms, but it gets stuck in local optima. In addition, this algorithm requires the number of clusters. Also, the density-based spatial of the application with noise (DBSCAN) algorithm, which is a density-based clustering algorithm, unlike the FCM algorithm, should not be pre-numbered. If the clusters are specific and depend on the number of clusters, then it can determine the number of clusters. Another advantage of the DBSCAN clustering algorithm over FCM is its ability to cluster data of different shapes. In this paper, in order to overcome these limitations, a hybrid approach for clustering is proposed, which uses FCM and DBSCAN algorithms. In

*Corresponding author

Received 11 May 2023; revised 9 July 2023; accepted 27 July 2023

Saeideh Barkhordari Firozabadi
PhD candidate, Department of Computer Science, Yazd University , Yazd, Iran.
e-mail: s.barkhordari@stu.yazd.ac.ir

Seyed Abolfazl Shahzadeh Fazeli
Parallel Processing Lab, Department of Computer Science, Yazd University, Yazd, Iran.
e-mail: fazeli@yazd.ac.ir

Jamal Zarepour Ahmadabadi
Department of Computer Science, Yazd University, Yazd, Iran.
e-mail: zarepourjamal@yazd.ac.ir

Seyed Mehdi Karbassi
Department of Applied Mathematics, Faculty of Mathematical Sciences, Yazd University, Yazd, Iran.
e-mail: smkarbassi@yazd.ac.ir

† This article was suggested by the scientific committee of the 5th national seminar on control and optimization for publication in IJNAO, which was accepted after independent review.

this method, the optimal number of clusters and the optimal location for the centers of the clusters are determined based on the changes that take place according to the data set in three phases by predicting the possibility of the problems stated in the FCM algorithm. With this improvement, the values of none of the initial parameters of the FCM algorithm are random, and in the first phase, it has been tried to replace these random values to the optimal in the FCM algorithm, which has a significant effect on the convergence of the algorithm because it helps to reduce iterations. The proposed method has been examined on the Iris flower and compared the results with basic FCM algorithm and another algorithm. Results shows the better performance of the proposed method.

AMS subject classifications (2020): 68T10, 62H30.

Keywords: Clustering; Fuzzy clustering; DBSCAN.

1 Introduction

Clustering is one of the important techniques of knowledge discovery in databases. Density-based clustering algorithms are one of the main methods for clustering in data mining. The density-based spatial of application with noise (DBSCAN) algorithm is a clustering method that is based on density. This algorithm has the ability to discover clusters of different sizes and shapes from a large amount of data and performs well against noise [3, 6]. Another method that has received a lot of attention is the fuzzy method. In these methods, unlike deterministic clustering that, any data belongs to exactly one cluster; the data can belong to several clusters [7]. Although the approach adopted by both algorithms is widely accepted to deal with clustering problems, due to the weakness in each and in order to achieve a better method for data clustering, various methods have been used. In all methods, it has been tried to find values that are as close as possible to an exact answer.

2 Related works

Wei and Xie [10], after a better analysis of the slower convergence speed, introduced a new competitive learning-based rival checked fuzzy c-means clustering algorithm. In the method proposed by Xue, Shang, and Feng [12], a fuzzy rough semi-supervised outlier detection is used, which is able to minimize the sum squared errors of the clustering. Maraziotis [4], for gene expression profile clustering, proposed a novel semi-supervised fuzzy clustering algorithm (SSFCA). Abdellahoum et al. [1] presented a new version of fuzzy clustering based on the ABC algorithm, namely ABC − SFCM. For detecting the malicious behavior in wireless sensor networks, Shamshirband

et al. [8] presented a hybrid clustering method, namely a density-based fuzzy imperialist competitive clustering algorithm. Mekhmoukh and Mokrani [5] introduced an improved fuzzy-C-means (FCM) using particle swarm optimization based on the outlier rejection and level set. The results of this method were compared with related works, which showed more effectiveness. Zhang et al. [13] proposed a variant of FCM for image segmentation, which has reduced the complexity of the algorithm compared to similar types. Alomoush et al. [2] proposed a method for choosing cluster centers to avoid getting stuck in local optimum.

3 Preliminaries and definitions

Here we present some necessary algorithms.

3.1 Clustering

Clustering is a process by which a set of objects can be separated into distinct groups. Each release is called a cluster. Members of each cluster, according to characteristics which, are very similar to each other, but instead, the degree of similarity between the clusters is the lowest [9]. Although most clustering algorithms or methods have the same basis, there are differences in the method of measuring similarity or distance, as well as choosing labels for objects in each cluster. There are methods: for example, discriminative clustering, hierarchical clustering, model-based clustering, fuzzy clustering and density-based clustering. Here We deal with the last two methods: fuzzy clustering and density-based clustering. By combining these methods, we try to provide a new method for clustering.

3.2 FCM clustering algorithm

As mentioned in fuzzy clustering, unlike classical clustering, where each input sample belongs to only one cluster, one sample can belong to more than one cluster. Actually the basic idea in fuzzy clustering is to assume that each element can be placed in several clusters with different degrees of membership [7]. As a result, we can have clusters that are more consistent with reality.

One of the basic fuzzy clustering algorithms is FCM. In the FCM algorithm, we try to optimize the following objective function [11]:

$$J_m = \sum_{i=1}^{c} \sum_{k=1}^{n} {u_{ik}}^m {d_{ik}}^2 = \sum_{i=1}^{c} \sum_{k=1}^{n} {u_{ik}}^m \|x_k - v_i\|^2,$$

where m is a real number greater than one. Moreover, u_{ik} is the degree of membership of the kth data in the ith cluster, d_{ik} is the measure of similarity in the next n space, x_k represents the kth data, and v_i is the center of the ith cluster. The complete procedure of the algorithm is as follows:

Algorithm 1 FCM algorithm

Input: Data set, number of clusters, $max - iter$, $threshold$ (minimum objective function improvement value), and m (the value for exponentiation of matrix U)
Output: Cluster centers, objective function values and matrix U

1. Initialization: Randomly determine the value of each data belonging to the desired cluster, put it in the matrix U, set the value of the $iter$, and the value of the objective function to zero.
2. Calculate new centers for each cluster.
3. Calculate the distance from the data to the cluster centers.
4. Calculate the value of the objective function in terms of distance values.
5. Calculate the matrix U in terms of the values obtained from the previous steps.
6. Calculate imp (the difference between the value of the objective function in the new step and in the previous step) and set $iter = iter + 1$, if $imp \geq threshold$ and $iter \leq max - iter$, then repeat step 2; otherwise, the algorithm terminates.

As mentioned, FCM clustering performs well when working with overlapping data and performs well with noise-free data. Since they cannot distinguish between data points and noise, it leads to the center, which may gravitate toward the outliers. Also, it may be located at a local optimum. To improve the algorithm, we use the DBSCAN algorithm. In which follows, the DBSCAN algorithm is presented.

3.3 DBSCAN algorithm

In density-based clustering algorithms, points with high density are identified and placed in a cluster. One of the famous algorithms cited in this field to DBSCAN, which was presented by Ester and colleagues in 1996. This algorithm has the ability to identify remote points [3]. In the DBSCAN algorithm, there are two parameters, the radius (Epsilon) and the minimum points in a cluster (MinPoints). Each data point has a distance from other

points. Any point whose distance to an assumed point is less than Epsilon is considered a neighbor of that point. Any given point that has MinPoints of neighbors is the center of the cluster.
The way that the algorithm works is that the algorithm first selects a sample (which is a point in the vector space) and according to the radius Epsilon, the neighbor looks for this point in space. If the algorithm is able to find at least as many points as MinPoints within the specified Epsilon radius, then all those points together belong to a cluster. The algorithm then looks for one of the points adjacent to the current point to look again at that point with the Epsilon radius. The other neighbor points are searched, and if the number of serious new neighbor points is found again, then this algorithm again places all those new points in the same cluster with the previous points. If it does not find a new point in the neighborhood, then this cluster is complete. To find other clusters at other points, it randomly selects another point and starts finding neighbors and forming a new cluster for that point. If the algorithm is within the desired radius of a point but does not find enough samples, then the DBSCAN algorithm identifies this point as outlier data and does not assign it to any cluster. It should make all the clusters and check all the points to be able to identify whether it is an outlier or not. The algorithm continues in the same way to find other clusters that have at least as much as $MinPoints$ in their radius and are clustered. Finally, those that are not assigned to any cluster are identified as outlier data. This continues until all points have been checked [3, 6].

4 Proposed method

To improve the fuzzy clustering method, changes are made in three phases:

1. In the FCM clustering method, as stated, the value of each data belonging to the desired cluster is randomly determined and placed in the matrix U. In the proposed method, first, the data set is clustered through the DBSCAN algorithm. Since the number of clusters must be given to the FCM algorithm as input, the initial cluster number is determined in this way. Then, the distance between each data to the centers of the clusters obtained from the DBSCAN algorithm is calculated. In the next step, these distances are reversed and normalized. To help improve the convergence of the algorithm, the above values are placed in the matrix U to determine the value belonging of each data to the clusters. The points that are closer to the cluster centers get more value. As a result, better convergence is achieved, and the number of iterations also becomes less.

2. Similar to the idea of the simulated annealing method, changes are applied to the number of clusters. In this way, in the range of $+k/2$ and $-k/2$, a value is randomly selected and added to the number of

clusters. If the number of clusters increases, then centers are randomly selected from the data, and if the number of clusters decreases, then some centers are randomly deleted. In the event that the objective function is improved by changing the number of clusters, results will be updated.

3. The cluster centers are moved to find the optimal centers. In the proposed method, the primary centers are obtained with the DBSCAN algorithm. Considering the criteria of the DBSCAN algorithm, in data density clustering, several data may be close to each other, but according to the changes in the value of data dimensions from the first to the last data, it is more appropriate that this data should not be placed in a cluster. For this reason, in the second and third phases of the proposed algorithm, the number and location of the cluster centers are changed to reach the optimal centers. These changes increase in the first iteration and decrease in subsequent iterations. The process is as follows: in the first iteration, based on the data diameter and the angle that is randomly determined, the transfer value is determined. In the next iterations, a coefficient from the diameter of the data determines that the amount of displacement is based on this coefficient takes place, and this displacement will be reduced. At each stage, based on the new centers, the matrix U and the objective function are calculated, and if improved, results will be replaced.

In the proposed method, different aspects of clustering and different ways of improving these methods were studied and investigated. Then, according to the weaknesses of the FCM algorithm, based on the changes made in three phases in the proposed method, the algorithm was improved with new methods from three points of view. In each point of view, different aspects of clustering are considered:

1. In the first phase of improvement, combine the algorithm with DBSCAN algorithm. In addition to solving the basic problem of the algorithm in determining the number of clusters, it is tried to make the initialization of the matrix U in a completely intelligent and accurate way by making changes. Because by conducting tests, we found that the initial values have a significant effect on the convergence and accuracy of the algorithm result, and if this value is done with the random method used in the FCM algorithm, then the number of repetitions will increase.

2. In the second phase, it was tried to find the optimal value for the number of clusters with a creative method. In the FCM algorithm, due to the unknown number of clusters, this value should be given as an initial parameter to the algorithm.

3. In this method, the initial value for the location of the centers is done according to the criteria of the DBSCAN algorithm. According to this

fact, in the third phase of the proposed method, it is tried to find the best place for the centers of the clusters with a new method, which has a significant effect on reaching the optimal solution.

The proposed method is summarized in Figure 1. Here, IMP is the improvement value of the objective function, NC is the number of clusters, and C is the centers of the clusters.

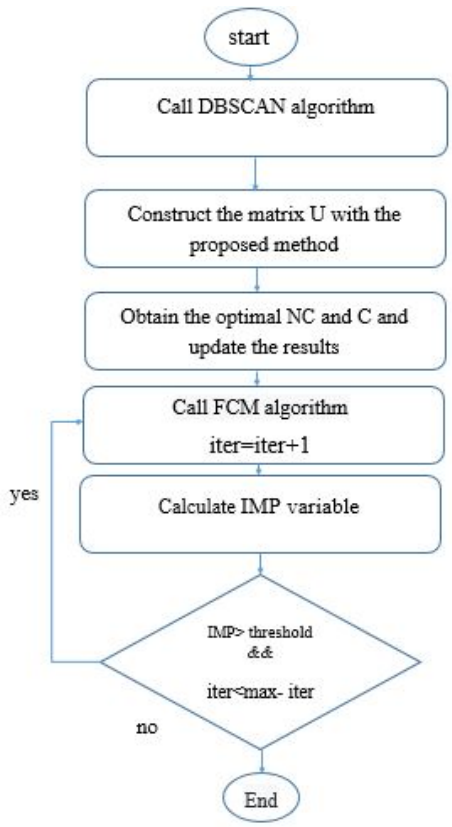


Figure 1: The proposed method

The general routine of the algorithm is given in Algorithm 2 as follows:

Algorithm 2 Proposed algorithm MODFCM

Input: Data set, number of clusters, $max - iter$, threshold (minimum objective function improvement value), and m (the value for exponentiation of matrix U)
Output: Cluster centers, objective function values, and matrix U

1. Initialization:
   (a) Call the DBSCAN algorithm and determine the number of clusters in the data set.
   (b) Calculate the distance of each data to the centers obtained from the DBSCAN algorithm and construct the matrix U with the proposed method.
   (c) Set $iter$ and the value of the objective function to zero.
2. Obtain the optimal number of clusters using the second phase of the proposed method, update the new results and go to the next step.
3. Initialize f by calculating the diameter of the data set.
4. Initialize s by randomly choosing an angle.
5. Set $d = f * \cos(s)$ and displace all centers by d.
6. Calculate the distance of the data to the centers of the new clusters and the value of the objective function in terms of the distance values.
7. Calculate the values of the matrix U according to the values obtained from the previous steps.
8. If the objective function is improved, then update new results and go to the next step; otherwise, go to step 10.
9. Set $f = .9f$. If $f \geq 5$, then go to step 4; otherwise, go to the next step.
10. Calculate the new centers for each cluster, the distance of the data to the centers of the clusters, and the value of the objective function in terms of the distance values.
11. Calculate the values of the matrix U according to the values obtained from the previous steps.
12. Calculate imp (the difference between the value of the objective function in the new step and in the previous step), and set $iter = iter + 1$. If $imp \geq threshold$ and $iter \leq max - iter$, then repeat step 10; otherwise, the algorithm terminates.

5 Experimental results

Two sets of tests have been performed on the FCM algorithm and the proposed MODFCM algorithm on the Iris flower with four features. Different similarity measures in the solution clustering problems are used. Here, the Euclidean distance criterion is used due to its high efficiency. Also, the evaluation of the results obtained from the clustering of the data set with the DBSCAN algorithm and the direct transfer of the results to the FCM algorithm was performed. The algorithm was named DBSCAN − FCM. We analyze the convergence and the iterative process of algorithms. The convergence and the iterative process for these algorithms are shown in Table 1. As we can see from Table 1, the convergence speed of the proposed MODFCM algorithm is faster than the FCM algorithm and DBSCAN − FCM algorithm. This shows that the proposed MODFCM algorithm improves the convergence speed. Further analysis reveals that the proposed MODFCM algorithm can reduce the required clustering time effectively and improve the efficiency of the data processing.

Table 1: Comparison table of algorithms

Algorithm	Objective function	Number of iterations
FCM	12.469286	15
DBSCAN − FCM	14.169376	19
MODFCM	9.589957	19
MODFCM	9.589961	15

In the second experiment, the GENETIC algorithm and RAND index were used, and the performance of two algorithms was evaluated. The results show that although both algorithms reach the final solution, the speed of the proposed algorithm is increased because the algorithm is converged in fewer iterations. In addition, the RAND index in the first population generated was evaluated for both algorithms. It was about 0.67 for the proposed algorithm in most iterations, but the same amount for FCM was about 0.41. The evaluation diagram of two algorithms, FCM and MODFCM, are given in Figures 2 and 3, respectively. As a result shows, the proposed algorithm MODFCM has a better performance in achieving the desired clustering.

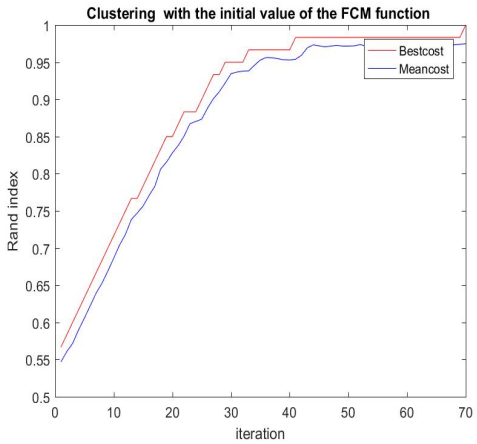


Figure 2: Evaluation of the performance of the FCM algorithm with RAND index

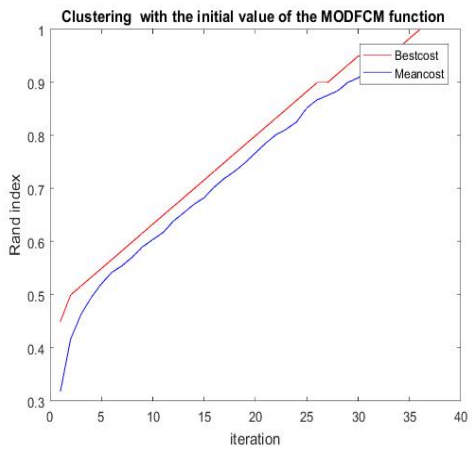


Figure 3: Evaluation of the performance of the MODFCM algorithm with RAND index

6 Conclusion

Today, there are many methods for data clustering, each of which has advantages and disadvantages. Methods can be achieved by combining algorithms to improve the results by covering each other's weaknesses. In this article, the FCM algorithm and the DBSCAN algorithm were combined. One of the advantages of the proposed algorithm in all the experiments compared

to FCM is that we do not face the problem of determining the number of clusters. The algorithm was improved by determining the optimal number of clusters. In addition, to increase the quality of clustering, optimal centers were also obtained. In total, by making these changes in the proposed method, it was found that by evaluating the objective function in both algorithms, the improvement of the objective function in the proposed algorithm with the same number of iterations has better performance than the FCM algorithm and the speed of convergence increases.

References

[1] Abdellahoum, H., Mokhtari, N., Brahimi, A. and Boukra, A. *CSFCM: An improved fuzzy C-Means image segmentation algorithm using a cooperative approach*, Expert Syst. Appl. 166 (2021), 114063.

[2] Alomoush, W., Khashan, O.A., Alrosan, A., Houssein, E.H., Attar, H., Alweshah, M. and Alhosban, F.*Fuzzy clustering algorithm based on improved global best-guided artificial bee colony with new search probability model for image segmentation*, Sensors 22(22) (2022), 8956.

[3] Ester, M., Kriegel, H. P. and Sander, J. *A density-based algorithm for discovering clusters in large spatial databases with noise*, kdd 96(34) (1996), 226–231.

[4] Maraziotis, I.A. *A semi-supervised fuzzy clustering algorithm applied to gene expression data*, Pattern Recognit. 45(1) (2012), 637–648.

[5] Mekhmoukh, A. and Mokrani, K. *Improved fuzzy C-Means based particle swarm optimization (PSO) initialization and outlier rejection with level set methods for MR brain image segmentation*, Comput. Methods Programs Biomed. 122(2) (2015), 266–281.

[6] Parimala, M., Lopez, D. and Senthilkumar, N.C. *A survey on density based clustering algorithms for mining large spatial databases*, Int. J. Adv. Sci. Technol. 31(1) (2011), 59–66.

[7] Ruspini, E.H., Bezdek, J.C. and Keller, J.M. *Fuzzy clustering: A historical perspective*, IEEE Comput. Intell. Mag. 14(1) (2019), 45–55.

[8] Shamshirband, S., Amini, A., Anuar, N.B., Kiah, M.L.M., Teh, Y.W. and Furnell, S. *D-FICCA: A density-based fuzzy imperialist competitive clustering algorithm for intrusion detection in wireless sensor networks*, Measurement 55 (2014), 212–226.

[9] Singh, T., Saxena, N., Khurana, M., Singh, D., Abdalla, M. and Alshazly, H. *Data clustering using moth-flame optimization algorithm*, Sensors 21(12) (2021), 4086.

[10] Wei, L.M. and Xie, W.X. *Rival checked fuzzy c-means algorithm*, ACTA ELECTONICA SINICA 28(7) (2000), 79.

[11] Xu, R. and Wunsch, D. *Survey of clustering algorithms*, IEEE Trans. Neural Netw. 16(3) (2005), 645–678.

[12] Xue, Z., Shang, Y. and Feng, A. *Semi-supervised outlier detection based on fuzzy rough C-means clustering*, Math. Comput. Simul. 80(9) (2010), 1911–1921.

[13] Zhang, H., Li, H., Chen, N., Chen, S. and Liu, J. *Novel fuzzy clustering algorithm with variable multi-pixel fitting spatial information for image segmentation*, Pattern Recognit. 121 (2022), 108201.

How to cite this article
Barkhordari Firozabadi, S., Shahzadeh Fazeli, S.A., Zarepour Ahmadabadi, J. and Karbassi, S.M., Improving the performance of the FCM algorithm in clustering using the DBSCAN algorithm. *Iran. J. Numer. Anal. Optim.*, 2023; 13(4): 763-774. https://doi.org/10.22067/ijnao.2023.82361.1260

Aims and scope

Iranian Journal of Numerical Analysis and Optimization (IJNAO) is published twice a year by the Department of Applied Mathematics, Faculty of Mathematical Sciences, Ferdowsi University of Mashhad. Papers dealing with different aspects of numerical analysis and optimization, theories and their applications in engineering and industry are considered for publication.

Journal Policy

All submissions to IJNAO are first evaluated by the journal's Editor-in-Chief or one of the journal's Associate Editors for their appropriateness to the scope and objectives of IJNAO. If deemed appropriate, the paper is sent out for review using a single blind process. Manuscripts are reviewed simultaneously by reviewers who are experts in their respective fields. The first review of every manuscript is performed by at least two anonymous referees. Upon the receipt of the referee's reports, the paper is accepted, rejected, or sent back to the author(s) for revision. Revised papers are assigned to an Associate Editor who makes an evaluation of the acceptability of the revision. Based upon the Associate Editor's evaluation, the paper is accepted, rejected, or returned to the author(s) for another revision. The second revision is then evaluated by the Editor-in-Chief, possibly in consultation with the Associate Editor who handled the original paper and the first revision, for a usually final resolution.

The authors can track their submissions and the process of peer review via: http://ijnao.um.ac.ir

All manuscripts submitted to IJNAO are tracked by using "iThenticate" for possible plagiarism before acceptance.

Instruction for Authors

The Journal publishes all papers in the fields of numerical analysis and optimization. Articles must be written in English.
All submitted papers will be refereed and the authors may be asked to revise their manuscripts according to the referee's reports. The Editorial Board of the Journal keeps the right to accept or reject the papers for publication.
The papers with more than one authors, should determine the corresponding author. The e-mail address of the corresponding author must appear at the end of the manuscript or as a footnote of the first page.
It is strongly recommended to set up the manuscript by Latex or Tex, using the template provided in the web site of the Journal. Manuscripts should be typed double-spaced with wide margins to provide enough room for editorial remarks.

References should be arranged in alphabetical order by the surname of the first author as examples below:

[1] Brunner, H. *A survey of recent advances in the numerical treatment of Volterra integral and integro-differential equations*, J. Comput. Appl. Math. 8 (1982), 213-229.

[2] Stoer, J. and Bulirsch, R. *Introduction to Numerical Analysis*, Springer-verlag, New York, 2002.

Iranian Journal of Numerical Analysis and Optimization

CONTENTS Vol. 13, No. 4, pp 578-774

web site: https://ijnao.um.ac.ir
Email: ijnao@um.ac.ir
ISSN-Print: 2423-6977
ISSN-Online: 2423-6969