Ferdowsi University of Mashhad

Iranian Journal of Numerical Analysis and Optimization

Volume 14, Number 3

September 2024

Serial Number: 30

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 14, Number 3, September 2024

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 14, Number 3, September 2024

Ferdowsi University of Mashhad - Iran

Iranian Journal of Numerical Analysis and Optimization

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

* Full Professor

** Associate Professor

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

Nonpolynomial B-spline collocation method for solving singularly perturbed quasilinear Sobolev equation 638
F. Edosa Merga and G. File Duressa

Differential-integral Euler–Lagrange equations 662
M. Shehata

An improved imperialist competitive algorithm for solving an inverse form of the Huxley equation 681
H. Dana Mazraeh, K. Parand, H. Farahani and S.R. Kheradpisheh

Stability analysis and optimal strategies for controlling a boycotting behavior of a commercial product 708
O. Aarabate, S. Belhdid and O. Balatif

Highly accurate collocation methodology for solving the generalized Burgers–Fisher's equation 736
S. Shallu and V.K. Kukrej

Uniformly convergent numerical solution for caputo fractional order singularly perturbed delay differential equation using extended cubic B-spline collocation scheme 762
N.A. Endrie and G.F. Duressa

Finite element analysis for microscale heat equation with Neumann boundary conditions 796
M.H. Hashim and A.J. Harfash

Numerical method for the solution of high order Fredholm integro-differential difference equations using Legendre polynomials 833
P.T. Pantuvo, G. Ajileye, R. Taparki and O.O. Aduroja

A pseudo−operational collocation method for optimal control problems of fractal−fractional nonlinear Ginzburg−Landau equation 875
T. Shojaeizadeh, E. Golpar-Rabok and P. Rahimkhani

A numerical computation for solving delay and neutral differential equations based on a new modification to the Legendre wavelet method 900
N.M. El-Shazly and M.A. Ramadan

Extending quasi-GMRES method to solve generalized Sylvester tensor equations via the Einstein product 938

M.M. Izadkhah

A stabilized simulated annealing-based Barzilai–Borwein method for the solution of unconstrained optimization problems 970
H. Sharma and R. Kumar Nayak

Iranian Journal of Numerical Analysis and Optimization
Vol. 14, No. 3, 2024, pp 638–661
https://doi.org/10.22067/ijnao.2024.85929.1363
https://ijnao.um.ac.ir/

Research Article

Nonpolynomial B-spline collocation method for solving singularly perturbed quasilinear Sobolev equation

F. Edosa Merga*, and G. File Duressa

Abstract

In this paper, a singularly perturbed one-dimensional initial boundary value problem of a quasilinear Sobolev-type equation is presented. The nonlinear term of the problem is linearized by Newton's linearization method. Time derivatives are discretized by implicit Euler's method on nonuniform step size. A uniform trigonometric B-spline collocation method is used to treat the spatial variable. The convergence analysis of the scheme is proved, and the accuracy of the method is of order two in space and order one in time direction, respectively. To test the efficiency of the method, a model example is demonstrated. Results of the scheme are presented in

*Corresponding author

Received 17 December 2023; revised 24 February 2024; accepted 21 March 2024

Feyisa Edosa Merga
Department of Mathematics, Jimma University, Jimma, Oromia, Ethiopia. e-mail: feyisae.2014@gmail.com

Gemechis File Duressa
Department of Mathematics, Jimma University, Jimma, Oromia, Ethiopia. e-mail: gammeef@gmail.com

How to cite this article
Edosa Merga, F. and File Duressa, G., Nonpolynomial B-spline collocation method for solving singularly perturbed quasilinear Sobolev equation. *Iran. J. Numer. Anal. Optim.*, 2024; 14(3): 638-661. https://doi.org/10.22067/ijnao.2024.85929.1363

tabular, and the figure indicates the scheme is uniformly convergent and has an initial layer at $t = 0$.

AMS subject classifications (2020): 65M06, 65M12, 65M22, 65M50.

Keywords: Singularly perturbed; Quasilinear; Sobolev; Trigonometric B-spline.

1 Introduction

A singularly perturbed differential equation is a differential equation in which the highest order derivative is multiplied by a small positive parameter ε that is recognized as a perturbation parameter. While solving these types of problems, the use of classical numerical methods on a uniform mesh may cause large oscillations as the perturbation parameter approaches zero in the entire domain of interest due to the boundary layer behavior. Therefore, to ignore this oscillation, several researchers constructed suitable numerical methods for these problems, whose accuracy does not depend on the perturbation parameter [7, 13, 12, 14, 15, 16].

This study deals also with the singularly perturbed initial boundary value problem of quasilinear Sobolev equation in the domain $\bar{Q} = \bar{\Omega} \times [0, T], \bar{\Omega} = [0, l], Q = (0, l) \times (0, T], \Omega = (0, l)$ of the form:

$$\begin{cases} Lu + f(x,t,u) = \varepsilon \left[\frac{\partial u}{\partial t}\right] - \alpha \frac{\partial^2 u}{\partial x^2} + \beta u \frac{\partial u}{\partial x} + f(x,t,u) = 0, & (x,t) \in Q, \\ u(x,0) = \varphi(x), & x \in \bar{\Omega}, \\ u(0,t) = u(l,t) = 0, & t \in (0,T], \end{cases} \tag{1}$$

where $\left[\frac{\partial u}{\partial t}\right] = -\frac{\partial^3 u}{\partial t \partial x^2} + \frac{\partial u}{\partial t}$, ε is a small perturbation parameter $0 < \varepsilon < 1$, and $\alpha > 0$ and β are given constants. Moreover, $\varphi(x)$ and $f(x,t,u)$ are assumed to be sufficiently continuously differentiable functions in $\bar{\Omega}$ and $\bar{Q} \times \mathbb{R}$, respectively.

Sobolev types equation arises in several mathematical problems, such as homogeneous fluid flow in fissured rocks [4], thermodynamics and propagation of long waves of small amplitude [20], quasi-stationary processes in

semiconductors [5], shear in second-order fluid [11], application of control theory [21], and other physical models. The analysis, development, and implementation of numerical methods for the solution of singularly perturbed pseudo-parabolic/Sobolev types of problems have received wide attention and developed in [3, 2, 1, 6, 8, 9, 10, 17].

The numerical method of (1) has been studied in the difference schemes for the singularly perturbed one-dimensional initial boundary value problem of Sobolev equations with initial jump [1]. Finite elements with piece-wise linear functions in space and exponential functions in time variables are applied.

Trigonometric B-spline is a nonpolynomial B-spline with a sine function, which was introduced by Schoenberg in 1964 [19]. Even though the trigonometric B-spline function is used to approximate several types of differential equations, it is not applied to quasilinear Sobolev types of equations. Motivated by this, we present the cubic trigonometric B-spline collocation method for solving the one-dimensional initial boundary value problem of singularly perturbed quasilinear Sobolev types of the equation. Implicit Euler and cubic trigonometric B-spline collocation methods are used to control the time and space variables, respectively.

The outline of this study is the following sequences. A linearization of the numerical scheme is presented in Section 2. In Section 3, the properties of the continuous solution are discussed. Numerical formulation of the problem is presented in Section 4. Convergence analysis and numerical results are considered in Sections 5 and 6, receptively. Finally, the conclusion of the study is given in Section 7.

2 Linearization of the problem

The one-dimensional singularly perturbed initial boundary value problem of Sobolev equation (1) can be rewritten as

$$-\varepsilon\frac{\partial^3 u}{\partial t\partial x^2}+\varepsilon\frac{\partial u}{\partial t}-\alpha\frac{\partial^2 u}{\partial x^2}+F(x,t,u,\frac{\partial u}{\partial x})=0,\quad (x,t)\in Q, \tag{2}$$

where $F(x,t,u,\frac{\partial u}{\partial x})=\beta u\frac{\partial u}{\partial x}+f(x,t,u)$.

To linearize (2), we consider an initial guess for the function $u(x,t)$ by denoting $u^{(0)}(x,t)$ that satisfies both initial and boundary conditions:

$$u^{(0)}(x,t)=\frac{1}{2}\varphi(x)\left(1+e^{\frac{-2t}{\varepsilon}}\right). \tag{3}$$

Applying Newton's linearization method on $F(x,t,u,\frac{\partial u}{\partial x}))$ for the function $u^{(0)}(x,t)$, we obtain an $(n+1)$th iteration:

$$\begin{aligned}F(x,t,u^{(n+1)},\frac{\partial u^{(n+1)}}{\partial x})&=F(x,t,u^{(n)},\frac{\partial u^{(n)}}{\partial x})\\&+\left(u^{(n+1)}-u^{(n)}\right)\frac{\partial F}{\partial u}|_{\left(u^{(n)},\frac{\partial u^{(n)}}{\partial x}\right)}\\&+\left(\frac{\partial u^{(n+1)}}{\partial x}-\frac{\partial u^{(n)}}{\partial x}\right)\frac{\partial F}{\partial(\frac{\partial u}{\partial x})}|_{\left(u^{(n)},\frac{\partial u^{(n)}}{\partial x}\right)}.\end{aligned} \tag{4}$$

Substituting (4) into (2) and after some rearrangements we obtain

$$\begin{cases}-\varepsilon\frac{\partial^3 u^{(n+1)}}{\partial t\partial x^2}+\varepsilon\frac{\partial u^{(n+1)}}{\partial t}-\alpha\frac{\partial^2 u^{(n+1)}}{\partial x^2}\\+a(x,t)\frac{\partial u^{(n+1)}}{\partial x}+b(x,t)u^{(n+1)}=g(x,t),\\u(x,0)=\varphi(x), & x\in\bar{\Omega},\\u(0,t)=u(l,t)=0, & t\in(0,T],\end{cases} \tag{5}$$

where

$$\begin{aligned}a(x,t)&=\frac{\partial F}{\partial(\frac{\partial u}{\partial x})}|_{\left(u^{(n)},\frac{\partial u^{(n)}}{\partial x}\right)},\qquad b(x,t)=\frac{\partial F}{\partial u}|_{\left(u^{(n)},\frac{\partial u^{(n)}}{\partial x}\right)},\\g(x,t)&=u^{(n)}\frac{\partial F}{\partial u}|_{\left(u^{(n)},\frac{\partial u^{(n)}}{\partial x}\right)}+\left(\frac{\partial u^{(n)}}{\partial x}\right)\frac{\partial F}{\partial(\frac{\partial u}{\partial x})}|_{\left(u^{(n)},\frac{\partial u^{(n)}}{\partial x}\right)}\\&\quad-F(x,t,u^{(n)},\frac{\partial u^{(n)}}{\partial x}).\end{aligned}$$

3 Properties of continuous solution

Lemma 1. Let $\varphi(x)\in C^2[0,l]$ and the derivatives $\frac{\partial^s f}{\partial x^s},\frac{\partial^s f}{\partial u^s}(s=1,2),\frac{\partial f}{\partial t}\in C(\bar{Q})$. Then, for the solution $u(x,t)$ of (1), the following estimate holds:

$$\left|\frac{\partial^{r+s}u(x,t)}{\partial t^r\partial x^s}\right|\leq C\varepsilon^{-r},\quad \text{for all } (x,t)\in\left(\overline{Q}\right), r=0,1,\ s=0,1,2 \tag{6}$$

for any fixed l and T, and provided that

$$M_0\alpha^{-1}\frac{|\beta|\,l}{2\pi}<1,$$

where

$$\alpha_0=\left(\alpha-M_0\alpha^{-1}\frac{|\beta|\,l}{2\pi}\frac{\pi^2}{e^2+\pi^2}\right), \tag{7}$$

$$M_0=\frac{\sqrt{l}}{2}\left(\|\varphi\|_1+\frac{l^2+\pi^2}{\alpha\pi^2}\max_{0\le t\le T}\|f(\cdot,t,0)\|_0\right), \tag{8}$$

and C is a generic positive constant,which is independent of ε and mesh parameters.

Proof. Consider the integral identity

$$(Lu,u)_0+(f,u)_0=0,$$

and taking into account that $\left(u\frac{\partial u}{\partial x},u\right)_0=0$. Then,

$$\left(\varepsilon\frac{\partial u}{\partial t},u\right)_0-\left(\varepsilon\frac{\partial^3 u}{\partial t\partial x^2},u\right)_0-\left(\alpha\frac{\partial^2 u}{\partial x^2},u\right)_0+\left(\beta u\frac{\partial u}{\partial x},u\right)_0+(f(x,t,u),u)_0=0.$$

Estimating these inner products on an interval 0 to l, we obtain

$$\varepsilon\int_0^l\frac{\partial u}{\partial t}udx-\varepsilon\int_0^l\frac{\partial^3 u}{\partial t\partial x^2}udx-\alpha\int_0^l\frac{\partial^2 u}{\partial x^2}udx+\int_0^l f(x,t,u)udx=0.$$

From the linearization by assuming $f(x,t,u)\approx f(x,t,u^{(0)})+\frac{\partial f}{\partial u}u$ and $F=f(x,t,u^{(0)})$, we have

$$\frac{\varepsilon}{2}\frac{d}{dt}(u,u)_0-\frac{\varepsilon}{2}\frac{d}{dt}\left(\frac{\partial u}{\partial x},\frac{\partial u}{\partial x}\right)_0-\alpha\left(\frac{\partial u}{\partial x},\frac{\partial u}{\partial x}\right)_0+(F,u)_0+\left(\frac{\partial f}{\partial u}u,u\right)_0=0.$$

This become

$$\frac{\varepsilon}{2}\frac{d}{dt}\left(\left\|\frac{\partial u}{\partial x}\right\|_0^2+\|u\|_0^2\right)+\alpha\left\|\frac{\partial u}{\partial x}\right\|_0^2+(F,u)_0+\left(\frac{\partial f}{\partial u}u,u\right)_0=0. \tag{9}$$

Applying an inequality $(F,u)_0\ge-\|F\|_0\|u\|_0$, $\gamma=\frac{l^2}{l^2+\pi^2}$ for $0<\gamma<1$ into (9), and after rearrangement, we get

$$\varepsilon \frac{d}{dt}\left(\left\|\frac{\partial u}{\partial x}\right\|_0^2 + \|u\|_0^2\right) + 2\alpha\left(\frac{\pi^2}{l^2+\pi^2}\right)\left(\left\|\frac{\partial u}{\partial x}\right\|_0^2 + \|u\|_0^2\right) \leq 2\|F\|_0 \|u\|_0 . \tag{10}$$

Choosing $C_1 = \alpha\frac{\pi^2}{l^2+\pi^2}$ and $\delta = \left\|\frac{\partial u}{\partial x}\right\|_0^2 + \|u\|_0^2$, the inequality (10) is written as

$$\varepsilon\delta'(t) + 2C_1\delta(t) \leq 2\|F\|_0 \|u\|_0 . \tag{11}$$

Solving the differential inequality (11), we obtain

$$\delta(t) \leq \delta_0 e^{\frac{-C_1 t}{\varepsilon}} + \left(\frac{1}{C_1^2}\max_{0\leq\tau\leq t}\|f(\cdot,t,0)\|_0^2\left(1 - e^{\frac{-C_1 t}{\varepsilon}}\right)\right) \tag{12}$$

with $\delta_0 = \|\varphi\|_1^2 = \|\varphi\|_0^2 + \|\varphi'\|_0^2$.

Using by the virtue of embedding inequality $\frac{l}{4}\left\|\frac{\partial u}{\partial x}\right\|_0^2 \geq \|u\|_{\infty,\Omega}^2$ into (12) and after some mathematical manipulation, we obtain

$$|u(x,t)| \leq \frac{\sqrt{l}}{2}\left(\delta_0^{\frac{1}{2}} e^{\frac{-C_1 t}{2\varepsilon}} + \frac{1}{C_1}\max_{0\leq\tau\leq t}\|f(\cdot,t,0)\|_0\right). \tag{13}$$

Using an identity

$$\left(Lu, \frac{\partial^2 u}{\partial x^2}\right)_0 = \left(f, \frac{\partial^2 u}{\partial x^2}\right)_0, \tag{14}$$

$$\frac{\varepsilon}{2}\frac{d}{dt}\left(\frac{\partial u}{\partial x}, \frac{\partial u}{\partial x}\right)_0 + \frac{\varepsilon}{2}\frac{d}{dt}\left(\frac{\partial^2 u}{\partial x^2}, \frac{\partial u}{\partial x}\right)_0 + \alpha\left(\frac{\partial^2 u}{\partial x^2}, \frac{\partial u}{\partial x}\right)_0 + \left(F, \frac{\partial^2 u}{\partial x^2}\right)_0 + \left(\frac{\partial f}{\partial u}u, \frac{\partial^2 u}{\partial x^2}\right)_0 = 0. \tag{15}$$

Using of Friedrich's inequality (15) then, after some rearrangement, we obtain

$$\varepsilon\frac{d}{dt}\left(\left\|\frac{\partial^2 u}{\partial x^2}\right\|_0^2 + \left\|\frac{\partial u}{\partial x}\right\|_0^2\right)_0 + 2\alpha\left(\frac{\pi^2}{l^2+\pi^2}\right)\left(\left\|\frac{\partial^2 u}{\partial x^2}\right\|_0^2 + \left\|\frac{\partial u}{\partial x}\right\|_0^2\right) \leq 2\|F\|_0\left\|\frac{\partial^2 u}{\partial x^2}\right\|_0, \tag{16}$$

which is written as

$$\varepsilon\delta'(t) + 2C_1\delta(t) \leq 2\|F\|_0\left\|\frac{\partial^2 u}{\partial x^2}\right\|_0, \tag{17}$$

where $\delta(t) = \left\|\frac{\partial^2 u}{\partial x^2}\right\|_0^2 + \left\|\frac{\partial u}{\partial x}\right\|_0^2$ and $C_1 = \alpha\frac{\pi^2}{l^2+\pi^2}$.

By solving the differential inequality (17) and applying embedding inequality, we get

$$|\frac{\partial^2 u}{\partial x^2}| \leq C. \tag{18}$$

With the same process from an identity

$$\left(Lu, \frac{\partial^3 u}{\partial t \partial x^2}\right)_0 = \left(f, \frac{\partial^3 u}{\partial t \partial x^2}\right)_0, \tag{19}$$

we obtain

$$\left\|\frac{\partial^3 u}{\partial t \partial x^2}\right\|_0^2 + \left\|\frac{\partial^2 u}{\partial t \partial x}\right\|_0^2 \leq \left(\|u\|_0^2 + \left\|\frac{\partial u}{\partial x}\right\|_0^2 + \left\|\frac{\partial^2 u}{\partial x^2}\right\|_0^2 + \frac{C}{\varepsilon^2}\|F\|_0^2\right), \tag{20}$$

which leads to (6) for $r = 1$, $s = 1, 2$, by using the proved estimate for $\left\|\frac{\partial^s u}{\partial x^2}\right\|_0, s = 0, 1, 2$.

Finally, we can write (1) as a form

$$\varepsilon \frac{\partial^3 u}{\partial t \partial x^2} + \alpha \frac{\partial^2 u}{\partial x^2} = \phi(x, t), \tag{21}$$

where $\phi(x, t) = \varepsilon \frac{\partial u}{\partial t} + \beta u \frac{\partial u}{\partial x} + f(t, x, u)$ and $|\phi(x, t)| \leq C$ with the estimate of (6) immediately for $r = 0, 1, s = 2$. □

Lemma 2. Under the assumption of Lemma 1, the following inequality holds:

$$\left\|\frac{\partial u}{\partial t}\right\|_1 \leq C\{1 + \varepsilon^{-1} e^{-\frac{\varpi_0 t}{\varepsilon}}\}, \qquad t \in [0, T], \tag{22}$$

where $\varpi_0 = \frac{C_1}{2}$ with $C_1 = \alpha \frac{\pi^2}{l^2 + \pi^2}$, which is given in the above.

Proof. Differentiating (1) with respect t and proceeded with $\frac{\partial u}{\partial t}$, we have

$$\frac{d}{dt}\left(\varepsilon \frac{\partial u}{\partial t} - \varepsilon \frac{\partial^3 u}{\partial t \partial x^2} - \alpha \frac{\partial^2 u}{\partial x^2} + \beta u \frac{\partial u}{\partial x} + f(x, t, u) = 0\right). \tag{23}$$

With an assumption $\frac{d}{dt}\beta u \frac{\partial u}{\partial x} = 0$, we can have

$$\left(\varepsilon \frac{\partial^2 u}{\partial t^2}, \frac{\partial u}{\partial t}\right)_0 - \left(\varepsilon \frac{\partial^4 u}{\partial t^2 \partial x^2}, \frac{\partial u}{\partial t}\right)_0 - \left(\alpha \frac{\partial^3 u}{\partial t \partial x^2}, \frac{\partial u}{\partial t}\right)_0 + \left(\frac{\partial f}{\partial t}, \frac{\partial u}{\partial t}\right)_0 = 0.$$

For $\left(\frac{\partial f}{\partial t}, \frac{\partial u}{\partial t}\right)_0 \geq -\left\|\frac{\partial f}{\partial t}\right\|_0 \left\|\frac{\partial u}{\partial t}\right\|_0$, we have

$$\frac{\varepsilon}{2}\frac{d}{dt}\left(\left\|\frac{\partial u}{\partial t}\right\|_0^2+\left\|\frac{\partial^2 u}{\partial t\partial x}\right\|_0^2\right)+\alpha\left\|\frac{\partial^2 u}{\partial t\partial x}\right\|_0^2\leq\left\|\frac{\partial f}{\partial t}\right\|_0\left\|\frac{\partial u}{\partial t}\right\|_0.$$

Applying an inequality relation $\left\|\frac{\partial f}{\partial t}\right\|_0\left\|\frac{\partial u}{\partial t}\right\|_0\leq\frac{1}{C_1}\left\|\frac{\partial f}{\partial t}\right\|_0^2+C_1\left\|\frac{\partial u}{\partial t}\right\|_0^2$, it gives

$$\varepsilon\frac{d}{dt}\left(\left\|\frac{\partial u}{\partial t}\right\|_0^2+\left\|\frac{\partial^2 u}{\partial t\partial x}\right\|_0^2\right)+C_1\left(\left\|\frac{\partial^2 u}{\partial t\partial x}\right\|_0^2+\left\|\frac{\partial u}{\partial t}\right\|_0^2\right)\leq C, \tag{24}$$

where $C=(C_1-2\alpha)\left\|\frac{\partial^2 u}{\partial t\partial x}\right\|_0^2+3C_1\left\|\frac{\partial u}{\partial t}\right\|_0^2+\frac{2}{C_1}\left\|\frac{\partial f}{\partial t}\right\|_0^2$.

Inequality (24) is written as

$$\varepsilon\delta'(t)+C_1\delta(t)\leq C, \tag{25}$$

where $\delta(t)=\left\|\frac{\partial u}{\partial t}\right\|_1^2=\left\|\frac{\partial u}{\partial t}\right\|_0^2+\left\|\frac{\partial^2 u}{\partial t\partial x}\right\|_0^2$.

This is similar to $\delta(t)=\|v\|_0^2+\|v'\|_0^2$ for $v=\left\|\frac{\partial u}{\partial t}\right\|_0^2$ and $C_1=\alpha\frac{\pi^2}{l^2+\pi^2}$.

Solving inequality (25), we get

$$\delta(t)\leq C\left(1+\frac{1}{\varepsilon^2}e^{\frac{-C_1 t}{\varepsilon}}\right), \tag{26}$$

which yields

$$\left\|\frac{\partial u}{\partial t}\right\|_1\leq C\left(1+\frac{1}{\varepsilon}e^{\frac{-C_1 t}{2\varepsilon}}\right).$$

With $\varpi_0=\frac{C_1}{2}$, it gives

$$\left\|\frac{\partial u}{\partial t}\right\|_1\leq C\left(1+\varepsilon^{-1}e^{\frac{-\varpi_0 t}{\varepsilon}}\right).$$

□

4 Numerical scheme formulation

4.1 Temporal discretization

Mesh generation

Based on (22), there is an initial layer in the neighborhood of $t=0$ of order $\varpi_0^{-1}\varepsilon|\ln\varepsilon|$ thickness, where ϖ_0 is given by (7). We divide two nonoverlapping subintervals $[0,\varrho]$ and $[\varrho,T]$, with the transition parameter

$$\varrho = \min\{\frac{T}{2}, \varpi_0^{-1}\varepsilon|\ln\varepsilon|\}.$$

Let $\overline{\Omega}_t^N = \{t_j\}_j^N$ be the set of mesh points. Now, we define piece-wise uniform mesh points as

$$t_j = \begin{cases} -\varpi_0^{-1}\varepsilon \ln N \left[1-(1-\varepsilon)\frac{2j}{N}\right], & j=0,\ldots,\frac{N}{2}, \quad if\, \varrho = \frac{T}{2}, \\ -\varpi_0^{-1}\varepsilon \ln N \left[1-\left(1-e^{\frac{-\varpi_0 T}{2\varepsilon}}\right)\frac{2j}{N}\right], & j=0,\ldots,\frac{N}{2}, \quad if\, \varrho < \frac{T}{2}, \\ \varrho + \left(1-\frac{N}{2}\right)\tau, & j=\frac{N}{2},\ldots,N, \quad \tau = 2\frac{(T-\varrho)}{N}. \end{cases}$$

To discretize time derivative of (5), we use the implicit Euler method on nonuniform step size on the domain: $\Omega_t^N = 0 = t_0 < t_1 < \cdots < t_j < t_{j+1} < \cdots < t_M = T, \quad j = 0.1,\ldots,M-1, \quad \tau(j) = t(j+1) - t(j)$ at the point (x, t_j). Then, (5) becomes

$$-\left(\frac{\varepsilon}{\tau(j)}+\frac{\alpha}{2}\right)\frac{\partial^2 u^{j+1}}{\partial x^2}(x) + a\,(x,t_{j+1})\,\frac{\partial u^{j+1}}{\partial x}(x) + \left(\frac{\varepsilon}{\tau(j)} + b\,(x,t_{j+1})\right)u^{j+1}(x) = \frac{-\varepsilon}{\tau(j)}\frac{\partial^2 u^j}{\partial x^2}(x) + \frac{\varepsilon}{\tau(j)}u^j(x) + g(x,t_{j+1}). \tag{27}$$

Lemma 3 (Semi-discrete maximum principle)**.** For each $j = 1, 2, \ldots, N-1$, let Z_{j+1} be a sufficiently smooth function on domain $\bar{\Omega}$. If $Z_{j+1}(0) \geq 0$, $Z_{j+1}(1) \geq 0$, and $L^{\tau(j)}u_{j+1}(x) \geq 0$, $x \in \Omega$, then $Z_{j+1} \geq 0$, for all $x \in \bar{\Omega}$.

Proof. Assume that there is (x^*) such that

$$Z_{j+1}(x^*) = \min_{x\in\overline{\Omega}_x} Z_{j+1}(x) \geq 0.$$

From the assumption it indicates that $x^* \notin \{1, 2\}$, which implies $x^* \in (0, 1)$. Applying the property of extreme values in calculus gives $\frac{d}{dx}Z_{j+1}(x^*) = 0$, and $\frac{d^2}{dx^2}Z_{j+1}(x^*) \geq 0$, given that $L^{\tau(j)}u_{j+1}(x^*) < 0$, which contradicts to $L^{\tau(j)}u_{j+1}(x^*) \geq 0, x \in \Omega$. Therefore, we conclude that $Z_{j+1} \geq 0$, for all $x \in \Omega$. Hence, the operator $L^{\tau(j)}$ satisfies a semi-discrete maximum principle. □

Lemma 4 (Local truncation error)**.** Consider the bound on the derivatives of $u(x,t)$ with respect to $t$given by $\left|\frac{\partial^k u(x,t)}{\partial x^k}\right| \leq C, \quad$ for all $(x,t) \in (\overline{\Omega})$. Then the local error estimate in the temporal direction is given by

$$\|e_{j+1}\| \leq C\,(\tau)^2,$$

where $e_{j+1} = u^{j+1}(x) - U^{j+1}(x)$ is the local error estimate in the temporal direction at $(j+1)th$ time level.

Proof. From (27), we have

$$\begin{aligned}
&\varepsilon\left(\frac{u^{j+1}(x)-u^{j}(x)}{\tau(j)}-\frac{\tau(j)}{2}\frac{\partial^2 u^{j}}{\partial x^2}(x)\right)-\varepsilon\frac{\partial^2}{\partial x^2}\left(\frac{u^{j+1}(x)-u^{j}(x)}{\tau(j)}-\frac{\tau(j)}{2}\frac{\partial^2 u^{j}}{\partial x^2}(x)\right)\\
&\quad-\alpha\frac{\partial^2 u^{j+1}}{\partial x^2}+a(x,t_{j+1})\frac{\partial u^{j+1}}{\partial x}+b(x,t_{j+1})u\\
&\qquad=g(x,t_{j+1})+O(\tau^2(j))^2,\quad (x,t)\in Q.
\end{aligned}\tag{28}$$

Multiplying (28) by $\tau(j)$, it gives

$$\begin{aligned}
&\varepsilon\left(u^{j+1}(x)-u^{j}(x)-\frac{\tau(j)^2}{2}\frac{\partial^2 u^{j}}{\partial t^2}(x)\right)-\varepsilon\frac{\partial^2}{\partial x^2}\left(u^{j+1}(x)-u^{j}(x)-\frac{\tau(j)^2}{2}\frac{\partial^2 u^{j}}{\partial t^2}(x)\right)\\
&\quad-\alpha\tau(j)\frac{\partial^2 u^{j+1}}{\partial x^2}+\tau(j)a(x,t_{j+1})\frac{\partial u^{j+1}}{\partial x}+\tau(j)b(x,t_{j+1})u\\
&\qquad=\tau(j)g(x,t_{j+1})+O(\tau(j))^3,\quad (x,t)\in Q.
\end{aligned}\tag{29}$$

By rearranging this, we obtain

$$\begin{aligned}
&\left(\varepsilon-\varepsilon\frac{\partial^2}{\partial x^2}-\frac{\tau(j)}{2}\left(\alpha\frac{\partial^2}{\partial x^2}-a(x,t_{j+1})\frac{\partial}{\partial x}-b(x,t_{j+1})\right)\right)u^{j+1}(x)\\
&\quad=\left(\varepsilon-\varepsilon\frac{\partial^2}{\partial x^2}\right)u^{j}(x)+\tau(j)\left(g(x,t_{j+1})\right)\\
&\qquad+(\tau(j))^2\left(\frac{\varepsilon}{2}\frac{\partial^2}{\partial x^2}-\frac{\varepsilon}{2}\frac{\partial^4}{\partial x^4}\right)u^{j}(x)+O\left(\Delta t(j)\right)^3,
\end{aligned}\tag{30}$$

which is written as

$$\mathcal{L}_{\varepsilon}^{\tau}u^{j+1}(x)=\Gamma(x,t_{j+1})+O\left(\tau(j)\right)^2,\tag{31}$$

where

$$\begin{aligned}
\mathcal{L}_{\varepsilon}^{\tau}&=\left(\varepsilon-\varepsilon\frac{\partial^2}{\partial x^2}-\frac{\tau(j)}{2}\left(\alpha\frac{\partial^2}{\partial x^2}-a(x,t_{j+1})\frac{\partial}{\partial x}-b(x,t_{j+1})\right)\right)\\
\Gamma(x,t_{j+1})&=\left(\varepsilon-\varepsilon\frac{\partial^2}{\partial x^2}\right)u^{j}(x,t)+\tau(j)\left(g(x,t_{j+1})\right).
\end{aligned}$$

From the boundedness of the solution, we have

$$\mathcal{L}_{\varepsilon}^{\tau} U^{j+1}(x) = \Gamma(x, t_{j+1}) \text{ for all } x \in \bar{\Omega}. \tag{32}$$

Now, from the desired mesh, we consider two case:

Case 1: For $\tau(j) \in [0, \sigma]$, let us consider $\max\{\tau(j) = \tau_1\}$.
Now, from (31) and (32), it yields

$$\left\| \mathcal{L}_{\varepsilon}^{\tau} \left(u^{j+1}(x) - U^{j+1}(x) \right) \right\| = \| \mathcal{L}_{\varepsilon}^{\tau} e_{j+1} \| \leq C \left(\tau_1 \right)^2.$$

Case 2: For $\tau(j) \in [\sigma, 1]$, let us consider $\max\{\tau(j) = \tau_2\}$.
Again from (31) and (32), we have

$$\left\| \mathcal{L}_{\varepsilon}^{\tau} \left(u^{j+1}(x) - U^{j+1}(x) \right) \right\| = \| \mathcal{L}_{\varepsilon}^{\tau} e_{j+1} \| \leq C \left(\tau_2 \right)^2$$

with the boundary conditions $u^{j+1}(0) - U^{j+1}(0) = e_{j+1}(0) = 0$ and $u^{j+1}(1) - U^{j+1}(1) = e_{j+1}(1) = 0$. Hence applying the maximum principles and choosing that $\tau = \max\{\tau_1, \tau_2\}$ give

$$\| e_{j+1} \| \leq C \left(\tau \right)^2.$$

□

Lemma 5. [Global error estimate] Under the hypothesis of the Lemma 4, the global error estimate in the temporal direction is given by

$$\| E_{j+1} \|_{\infty} \leq C(\tau)^2, \quad \text{for all } \ j \leq T/\Delta t,$$

where E_{j+1} is the global error estimate in the temporal direction at $(j+1)$th time level.

Proof. Using local error estimates up to $(j+1)$th time step given in Lemma 2, we get the following global error estimates at $(j+1)$th time step

$$\begin{aligned}
\| E_{j+1} \|_{\infty} &= \left\| \sum_{k=1}^{j} e_k \right\|_{\infty}, \quad j+1 \leq T/\tau(j) \\
&\leq \| e_1 \|_{\infty} + \| e_2 \|_{\infty} + \| e_3 \|_{\infty} + \cdots + \| e_+ 1 \|_{\infty} \\
&\leq c_1 j + 1 \left(\tau(j) \right)^2 \qquad \text{(by Lemma 4)} \\
&\leq c_1 \left((j+1)\tau(j) \right) \left(\tau(j) \right) \\
&\leq c_1 T \left(\tau(j) \right), \qquad ((j+1)\tau(j) \leq T)
\end{aligned}$$

$$\leq C(\tau(j)) \quad \textit{choosing} \quad \tau = \max\{\tau(j)\}$$
$$\leq C(\tau).$$

□

4.2 Spatial discretization

Descretizing the interval equally by knots x_i into N subintervals $[x_i, x_{i+1}]$, $i = 0, 1, \ldots, N-1$, such that $0 = x_0 < x_1 < \cdots < x_N = l$ as a uniform partition of the solution domain $0 \leq x \leq l$ with the step length $h = x_{i+1} - x_i = \frac{l}{N},\ i = 0, 1, \ldots, N-1$.

The piece-wise cubic trigonometric B-spline basis function $TB_i(x)$ over the uniform mesh is defined as [22]:

$$TB_i(x) = \frac{1}{\omega(h)} \begin{cases} (\sin)^3\left(\frac{x-x_{i-2}}{2}\right), & x \leq x_{i-1}, \\ \sin\left(\frac{x-x_{i-2}}{2}\right)\left[\sin\left(\frac{x-x_{i-2}}{2}\right)\sin\left(\frac{x_i-x}{2}\right) + \sin\left(\frac{x_{i+1}-x}{2}\right)\sin\left(\frac{x-x_{i-1}}{2}\right)\right] & \\ +\sin\left(\frac{x-x_{i-2}}{2}\right)\sin^2\left(\frac{x_i-x}{2}\right), & x_{i-1} \leq x \leq x_i, \\ \sin\left(\frac{x_{i+2}-x}{2}\right)\left[\sin\left(\frac{x-x_{i-1}}{2}\right)\sin\left(\frac{x_{i+1}-x}{2}\right) + \sin\left(\frac{x_{i+2}-x}{2}\right)\sin\left(\frac{x-x_i}{2}\right)\right] & \quad (33) \\ +\sin\left(\frac{x-x_{i-2}}{2}\right)\sin^2\left(\frac{x_{i+1}-x}{2}\right), & x_i \leq x \leq x_{i+1}, \\ \sin^3\left(\frac{x_{i+2}-x}{2}\right), & x \leq x_{i+2}, \\ 0, & \text{otherwise}, \end{cases}$$

where $\omega(h = \sin\left(\frac{h}{2}\right)\sin(h)\sin\left(\frac{3h}{2}\right)$ and $\{TB_{-1}(x), TB_0(x), \ldots, TB_N(x), TB_{N+1}(x)\}$ form a basis over the region $0 \leq x \leq l$. The coefficients of the approximate function $TB_i(x)$ and its derivatives are given in Table 1.

Table 1: Coefficients of cubic B-splines and its derivatives at knots

x	x_{i-2}	x_{i-1}	x_i	x_{i+1}	x_{i+2}
$TB_i(x)$	0	η_1	η_2	η_1	0
$TB_i^{'}(x)$	0	$-\eta_3$	0	η_3	0
$TB_i^{"}(x)$	0	η_4	η_5	η_4	0

We have

$$\eta_1 = \frac{\sin^2\frac{h}{2}}{\sin(h)\sin(\frac{3h}{2})}, \qquad \eta_2 = \frac{2}{1+2\cos(h)}, \quad \eta_3 = \frac{3}{4\sin(\frac{3h}{2})},$$
$$\eta_4 = \frac{3(1+3\cos(h))}{16\sin^2(\frac{h}{2})\left(2\cos(\frac{h}{2})+\cos(\frac{3h}{2})\right)}, \quad \eta_5 = -\frac{3\cos^2(\frac{h}{2})}{2\sin^2(\frac{h}{2})\left(1+2\cos(h)\right)}.$$

Let $U(x)$ be the cubic trigonometric B-spline collocation to approximate (1) and given as

$$U(x) \approx \sum_{i=-1}^{M+1} \alpha_i(t) TB_i(x), \tag{34}$$

where $\alpha_i(t)$ is an unknown time-dependent parameter to be determined from the collocation method together with using the boundary and initial conditions. Using (34) and Table 1, an approximate values of $U(x,t)$ and its first and second derivatives at the knots are

$$\begin{cases} U_i = \eta_1\alpha_{i-1} + \eta_2\alpha_i + \eta_1\alpha_{i+1}, \\ U_i^{'} = -\eta_3\alpha_{i-1} + \eta_3\alpha_{i+1}, \\ U_i^{"} = \eta_4\alpha_{i-1} + \eta_5\alpha_i + \eta_4\alpha_{i+1}. \end{cases} \tag{35}$$

By substituting (35) into (27), we obtain

$$\begin{aligned} &\left(-\eta_4\left(\frac{\varepsilon}{\tau(j)}+\alpha\right) - \eta_3 a_i^{j+1} + \eta_1 b_i^{j+1}\right)\alpha_{i-1}^{j+1} + \left(-\eta_5\left(\frac{\varepsilon}{\tau(j)}+\alpha\right) + \eta_2 b_i^{j+1}\right)\alpha_i^{j+1} \\ &+ \left(-\eta_4\left(\frac{\varepsilon}{\tau(j)}+\alpha\right) + \eta_3 a_i^{j+1} + \eta_1 b_i^{j+1}\right)\alpha_{i+1}^{j+1} \\ &= \left(-\frac{\varepsilon}{\tau(j)}(\eta_4-\eta_1)\right)\alpha_{i-1}^{j} + \left(-\frac{\varepsilon}{\tau(j)}(\eta_5-\eta_2)\right)\alpha_i^{j} \\ &+ \left(-\frac{\varepsilon}{\tau(j)}(\eta_4-\eta_1)\right)\alpha_{i+1}^{j} + g_i^{j+1}. \end{aligned} \tag{36}$$

This can be written as

$$r_i^-\alpha_{i-1}^{j+1} + r_i^c\alpha_i^{j+1} + r_i^+\alpha_{i+1}^{j+1} = q_i^-\alpha_{i-1}^{j} + q_i^c\alpha_i^{j} + q_i^+\alpha_{i+1}^{j} + g_i^{j+1}, \tag{37}$$

where

$$r_i^- = -\eta_4\left(\frac{\varepsilon}{\tau(j)}+\alpha\right) - \eta_3 a_i^{j+1} + \eta_1 b_i^{j+1},$$
$$r_i^c = -\eta_5\left(\frac{\varepsilon}{\tau(j)}+\alpha\right) + \eta_2 b_i^{j+1},$$

$$
\begin{aligned}
r_i^+ &= -\eta_4\left(\frac{\varepsilon}{\tau(j)}+\alpha\right)+\eta_3 a_i^{j+1}+\eta_1 b_i^{j+1},\\
q_i^- &= -\frac{\varepsilon}{\tau(j)}\left(\eta_4-\eta_1\right),\\
q_i^c &= -\frac{\varepsilon}{\tau(j)}\left(\eta_5-\eta_2\right),\\
q_i^+ &= -\frac{\varepsilon}{\tau(j)}\left(\eta_4-\eta_1\right).
\end{aligned}
$$

Imposing the boundary condition

Using the boundary conditions into (35), we have for $i=0$,

$$
\eta_1\alpha_{-1}^j+\eta_2\alpha_0^j+\eta_1\alpha_1^j=\phi_0\Rightarrow\alpha_{-1}^j=\frac{1}{\eta_1}\phi_0-\frac{\eta_2}{\eta_1}\alpha_0^j-\alpha_1^j, \tag{38}
$$

for $i=N$,

$$
\eta_1\alpha_{N-1}^j+\eta_2\alpha_N^j+\eta_1\alpha_{N+1}^j=\phi_N\Rightarrow\alpha_{N+1}^j=\frac{1}{\eta_1}\phi_N-\alpha_{N-1}^j-\frac{\eta_2}{\eta_1}\alpha_N^j. \tag{39}
$$

Substituting (38) and (39) into (37), we obtain

$$
\begin{cases}
\left(r_0^c-\frac{\eta_2}{\eta_1}r_0^-\right)\alpha_0^{j+1}+\left(r_0^+-r_0^-\right)\alpha_0^{j+1}\\
=\left(q_0^c-\frac{\eta_2}{\eta_1}q_0^-\right)\alpha_0^j+\left(q_0^+-q_0^-\right)\alpha_0^j+\left(\frac{q_0^-}{\eta_1}\phi_0^j-\frac{r_0^-}{\eta_1}\phi_0^{j+1}\right)+g_0^{j+1},\\
r_i^-\alpha_{i-1}^{j+1}+r_i^c\alpha_i^{j+1}+r_i^+\alpha_{i+1}^{j+1}=q_i^-\alpha_{i-1}^j+q_i^c\alpha_i^j+q_i^+\alpha_{i+1}^j+g_i^{j+1},\\
\left(r_N^--r_N^+\right)\alpha_{N-1}^{j+1}+\left(r_N^c-\frac{\eta_2}{\eta_1}r_N^+\right)\alpha_N^{j+1}\\
=\left(q_N^+-q_N^+\right)\alpha_{N-1}^j+\left(q_N^c-\frac{\eta_2}{\eta_1}q_N^+\right)\alpha_N^j+\left(\frac{q_N^+}{\eta_1}\phi_N^j-\frac{r_N^+}{\eta_1}\phi_N^{j+1}\right)+g_N^{j+1},\\
u(x_i,0)=\varphi(x_i), & x_i\in\bar{\Omega},\\
u(0,t_{j+1})=u(l,t_{j+1})=0, & t_{j+1}\in(0,T].
\end{cases} \tag{40}
$$

Equation (40) is an $(N+1)\times(N+1)$ system of linear equations.

Determination of the initial vector α_i^0

An initial vector can be calculated from the initial condition and first space derivative of the initial conditions at the boundaries. At the knots x_i, the following relations are used:

$$
\begin{aligned}
U^0(x_0,0)=\phi_0^0&=\eta_1\alpha_{-1}^0+\eta_2\alpha_0^0+\eta_1\alpha_1^0\\
U^0(i,0)=\phi_0^i&=\eta_1\alpha_{i-1}^0+\eta_2\alpha_i^0+\eta_1\alpha_{i+1}^0, \qquad i=1,2,\ldots,N-1,\\
U^0(1,0)=\phi_N^0&=\eta_1\alpha_{N-1}^0+\eta_2\alpha_N^0+\eta_1\alpha_{N+1}^0.
\end{aligned} \tag{41}
$$

From first derivative of (35), we also have

$$\eta_3\alpha_1^0 - \eta_3\alpha_{-1}^0 = \left(\phi_0^0\right)' \Rightarrow \alpha_{-1}^0 = \alpha_1^0 - \frac{1}{\eta_3}\left(\phi_0^0\right)',$$
$$\eta_3\alpha_{N+1}^0 - \eta_3\alpha_{N-1}^0 = \left(\phi_N^0\right)' \Rightarrow \alpha_{N+1}^0 = \alpha_{N-1}^0 + \frac{1}{\eta_3}\left(\phi_N^0\right)'. \tag{42}$$

Substituting (42) into (41), we obtain an $(N+1)\times(N+1)$ system of linear equations:

$$\begin{aligned}
&\eta_2\alpha_0^0 + 2\eta_1\alpha_1^0 = \phi_0^0 + \frac{\eta_1}{\eta_3}\left(\phi_0^0\right)' \\
&\eta_1\alpha_{i-1}^0 + \eta_2\alpha_i^0 + \eta_1\alpha_{i+1}^0 = U^0(i,0), \qquad i = 1,2,\ldots,N-1, \\
&2\eta_1\alpha_{N-1}^0 + \eta_2\alpha_N^0 = \phi_N^0 - \frac{\eta_1}{\eta_3}\left(\phi_N^0\right)'.
\end{aligned} \tag{43}$$

5 Convergence analysis

Lemma 6. The trigonometric B-spline collocation $\{TB_{-1}(x), TB_0(x), \ldots, TB_N(x), TB_{N+1}(x)\}$ defined in (33) satisfies the inequality

$$\sum_{i=-1}^{N+1} |TB_i(x)| \le 6, \quad x \in [0,1]. \tag{44}$$

Proof. From the triangular inequality, we have

$$\left|\sum_{i=-1}^{N+1} TB_i(x)\right| \le \sum_{i=-1}^{N+1} |TB_i(x)|.$$

At any node x_i, we have

$$\begin{aligned}
\sum_{i=-1}^{N+1} |TB_i(x)| &= |TB_{i-1}(x)| + |TB_i(x)| + |TB_{i+1}(x)| \\
&= |\eta_1| + |\eta_2| + |\eta_1| \le 4.
\end{aligned}$$

At any point in each subinterval $x_{i-1} \le x \le x_i$, we also have

$$\sum_{i=-1}^{N+1} |TB_i(x_i)| \le 2 \quad \textit{and} \quad \sum_{i=-1}^{N+1} |TB_{i-1}(x_{i-1})| \le 2,$$

and similarly for $x \in [x_{i-1}, x_i]$, we have

$$\sum_{i=-1}^{N+1} |TB_{i+1}(x_i)| \leq 1 \quad and \quad \sum_{i=-1}^{N+1} |TB_{i-2}(x_{i-1})| \leq 1.$$

Therefore

$$\sum_{i=-1}^{N+1} |TB_i(x)| = |TB_{i-2}(x)| + |TB_{i-1}(x)| + |TB_i(x)| + |TB_{i+1}(x)| \leq 6. \quad (45)$$

□

Lemma 7. Let $u(x)$ be the exact solution of the boundary value problem (40) and let $U(x) = \sum_{i=-1}^{N+1} \alpha_i(t) TB_i(x)$ be the trigonometric B-spline collocation approximation of $u(x)$. Then

$$\|u(x) - U(x)\|_\infty \leq C\left(h^2\right) \quad (46)$$

for sufficiently small h, and C is a positive constant.

Proof. Let $\bar{U}(x) = \sum_{i=-1}^{N+1} \bar{\alpha}_i TB_i(x)$ be a unique spline interpolate to be computed B-spline approximation to $u(x)$, where $\bar{\alpha}_i = (\bar{\alpha}_0, \bar{\alpha}_1, \ldots, \bar{\alpha}_N)^T$.

To estimate $\|u(x) - U(x)\|$, we must estimate the errors $\left\|u(x) - \bar{U}(x)\right\|$ and $\left\|\bar{U}(x) - U(x)\right\|$, respectively. Now, (40) is written as

$$A\alpha_i^{j+1} = H, \quad (47)$$

where $H = B\alpha_i^j + D$.

Following (47) for $\bar{U}(x)$, we get

$$A\bar{\alpha}_i^{j+1} = \bar{H}, \quad (48)$$

where $\bar{\alpha}_i^{j+1} = \left(\bar{\alpha}_0^{j+1}, \bar{\alpha}_1^{j+1}, \ldots, \bar{\alpha}_N^{j+1}\right)^T$.

Now, from (47) and (48), we obtain

$$A\left(\alpha_i^{j+1} - \bar{\alpha}_i^{j+1}\right) = \left(H - \bar{H}\right). \quad (49)$$

To proceed this, we consider the following theorem.

Theorem 1. Suppose that $H \in C^2[0,l]$ that $u(x) \in C^4[0,l]$, and that $\bar{\Omega}_x = \{0 = x_0 < x_1 < \cdots < x_N = l\}$ is a uniform partition of $[0,l]$ with the step size h. If $U(x)$ is the unique trigonometric B-spline approximation for

$u(x)$ at the knots $x_0, \ldots, x_N$, then

$$\begin{aligned} &|U(x) - u(x)| \leq O(h^3), \\ &\left|U^{(k)}(x) - u^{(k)}(x)\right| \leq O(h^2), \quad k = 1, 2, \\ &\left|U^{(k)}(x) - u^{(k)}(x)\right| \leq O(h), \quad k = 3, \end{aligned}$$

Proof. See [18]. □

Setting the right-hand side of (27) by H_i and using Theorem 1, we obtain the bound $\left\|H - \breve{H}\right\|$ as

$$\begin{aligned} \left|H_i - \bar{H}_i\right| &= \left|cu_i^{''} + au_i^{'} + bu_i - c\bar{u}_i^{''} + a\bar{u}_i^{'} + b\bar{u}_i\right| \\ &\leq |c| \left|u_i^{''} - \bar{u}_i^{''}\right| + |b| \left|u_i^{'} - \bar{u}_i^{'}\right| + |b| \left|u_i - \bar{u}_i\right|, \end{aligned}$$

where
$c = -\left(\frac{\varepsilon}{\tau(j)} + \frac{\alpha}{2}\right), a = a\left(x, t_{j+1}\right), b = \frac{\varepsilon}{\tau(j)} + b\left(x, t_{j+1}\right),$
which is

$$\left|H_i - \bar{H}_i\right| \leq |c|\, O(h^2) + |b|\, O(h^2) + |b|\, O(h^3) \leq K(h^2), \tag{50}$$

where $K = |c| + |b| + |b|\, O(h)$. Now from (49) and (50), we have

$$\|A\| \left\|\alpha_i^{j+1} - \bar{\alpha}_i^{j+1}\right\| = \left\|H - \bar{H}\right\| \leq Kh^2.$$

This yields

$$\left\|\alpha_i^{j+1} - \bar{\alpha}_i^{j+1}\right\| \leq Kh^2 \left\|A^{-1}\right\|. \tag{51}$$

Moreover, $TB_i(x)$ and its derivative up to the second order have nonvanishing values at the mesh points $[x_{i=2}, x_{i+2}]$ and at other mesh points it is zero. Using these facts, the matrix $\|A\|$ is a tridiagonal and diagonally dominant matrix. Hence, the matrix is nonsingular, and A^{-1} is bounded. Then, we get

$$\left\|\alpha_i^{j+1} - \bar{\alpha}_i^{j+1}\right\| \leq K_1 h^2, \tag{52}$$

where $K_1 = K\left\|A^{-1}\right\|$. Again from $U(x) - \bar{U}(x)$ and Lemma 7, we get

$$U(x) - \bar{U}(x) = \sum_{i=-1}^{N+1} (\alpha_i - \bar{\alpha}_i)\, TB_i(x) \leq \bar{k}h^2, \quad \bar{k} = 6k_1. \tag{53}$$

Therefore, from Theorem 1 and (53), we get

$$\begin{aligned}\|u(x)-U(x)\|_\infty &= \left\|u(x)-U(x)+\bar{U}(x)-\bar{U}(x)\right\|_\infty \\ &\leq \left\|u(x)-\bar{U}(x)\right\| + \left\|\bar{U}(x)-U(x)\right\| \\ &\leq O(h^3)+\bar{K}(h^2) \\ &\leq Ch^2, \quad C=\bar{K}+O(h).\end{aligned}$$

□

Theorem 2. Let $u(x,t)$ be the solution of (1) and let $U(x_i,t_{j+1})$ be the solution of the total discretized equation. Under the hypothesis of Lemmas 5 and 7, then the $\varepsilon-$uniform estimate holds

$$\sup_{1\leq i\leq N-1} = \max_{1\leq i\leq N-1, 0<j<M} |u(x_i,t_{j+1})-U(x_i,t_{j+1})| \leq C\left(h^2+\tau\right), \quad (54)$$

where C is the constant independent of ε, h, and τ.

Proof. The proof is obtained by applying the triangle inequality. □

6 Numerical results

To demonstrate the validity of the proposed scheme for the problem, one model example is presented. As the exact solution of this example is not known, the maximum point-wise error for the given example were computed by using the double mesh principle as

$$E_\varepsilon^{N,M} = \max_{1\leq i\leq N-1}\left|U_i^{N,M}-U_i^{2N,2M}\right|,$$

where $U_i^{N,M}$ is the numerical solution obtained on the mesh $D^N=\Omega_x^N\times\Omega_t^M$ with N mesh intervals in the spatial direction and M mesh intervals in the temporal direction. For any value of N and M, the $\varepsilon-$uniform errors are calculated using

$$E^{N,M} = \max_\varepsilon E_\varepsilon^{N,M}.$$

The rate of convergence of the scheme is calculated by the formula

$$r_\varepsilon^{N,M} = \frac{\log(E_\varepsilon^{N,M})-\log(E_\varepsilon^{2N,2M})}{\log(2)},$$

and the $\varepsilon-$uniform convergence is calculated by

$$r^{N,M} = \frac{\log\left(E^{N,M}\right) - \log\left(E^{2N,2M}\right)}{\log(2)}.$$

Example 1. From [1]

$$\begin{cases} \varepsilon\frac{\partial u}{\partial t} - \varepsilon\frac{\partial^3 u}{\partial t \partial x^2} - 2\frac{\partial^2 u}{\partial x^2} + \frac{1}{2}u\frac{\partial u}{\partial x} = \exp(-t)\sin(\pi x), \\ u(x,0) = \sin(\pi x), & x \in \overline{\Omega}_x, \\ u(0,t) = u(1,t) = 0, & t \in (0,1]. \end{cases}$$

Table 2: $E_\varepsilon^{N,M}$ for $N = M$

ε	32	64	128	256	512
2^0	5.5142e-03	2.8124e-03	1.4209e-03	7.1420e-04	3.5806e-04
2^{-2}	1.9178e-02	1.0254e-02	5.2831e-03	2.6643e-03	1.3257e-03
2^{-4}	5.1476e-02	3.3936e-02	2.1160e-02	1.1892e-02	5.1156e-03
2^{-6}	5.1485e-02	3.3947e-02	2.1168e-02	1.1900e-02	5.1221e-03
2^{-8}	5.1487e-02	3.3950e-02	2.1170e-02	1.1902e-02	5.1239e-03
2^{-10}	5.1488e-02	3.3951e-02	2.1170e-02	1.1903e-02	5.1243e-03
2^{-12}	5.1488e-02	3.3951e-02	2.1170e-02	1.1903e-02	5.1244e-03
2^{-14}	5.1488e-02	3.3951e-02	2.1170e-02	1.1903e-02	5.1245e-03
2^{-16}	5.1488e-02	3.3951e-02	2.1170e-02	1.1903e-02	5.1245e-03
$E^{N,M}$	5.1488e-02	3.3951e-02	2.1170e-02	1.1903e-02	5.1245e-03
$r^{N,M}$	0.60078	0.68143	0.83070	1.2158	

Table 3: $E_{\varepsilon}^{N,M}$ for the proposed method with $N = 60$

ε	N=40	N=80	N=160	N=320	N=640
2^{0}	5.3980e-03	2.8124e-03	1.5094e-03	9.1150e-04	7.7675e-04
2^{-2}	1.9182e-02	1.0254e-02	5.3330e-03	2.7775e-03	1.5158e-03
2^{-4}	5.1414e-02	3.3936e-02	2.1200e-02	1.1953e-02	5.2045e-03
2^{-6}	5.1426e-02	3.3947e-02	2.1207e-02	1.1869e-02	5.2096e-03
2^{-8}	5.1430e-02	3.3950e-02	2.1209e-02	1.1871e-02	5.2110e-03
2^{-10}	5.1431e-02	3.3951e-02	2.1210e-02	1.1871e-02	5.2113e-03
2^{-12}	5.1431e-02	3.3951e-02	2.1210e-02	1.1871e-02	5.2114e-03
2^{-14}	5.1431e-02	3.3951e-02	2.1210e-02	1.1871e-02	5.2114e-03
$E^{N,M}$	5.1431e-02	3.3951e-02	2.1210e-02	1.1871e-02	5.2114e-03
$r^{N,M}$	0.59918	0.67871	8.373	1.1877	

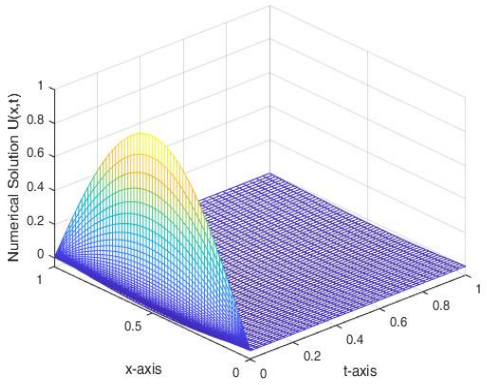


Figure 1: Numerical solution for $N = M = 64$ and $\varepsilon = 2^{-8}$.

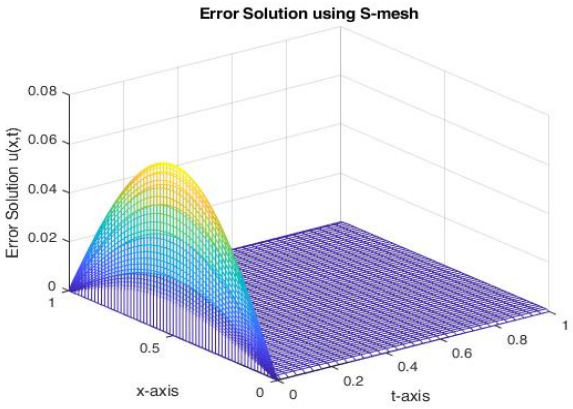


Figure 2: Error solution for $M = N = 64$ and $\varepsilon = 2^{-8}$.

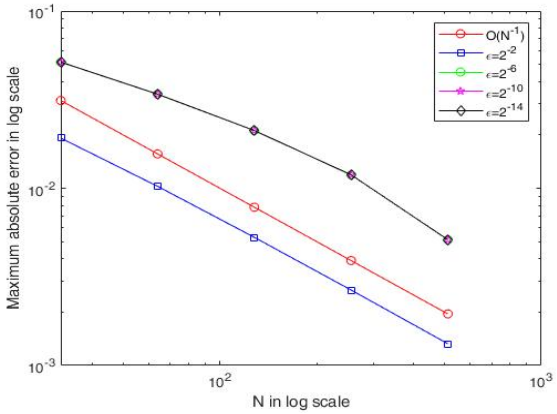


Figure 3: Log-log scale plot for Example 1.

The computed maximum point-wise errors are also presented in Tables 2 and 3. From Table 2, one can observe that as $\varepsilon \to 0$ and time step size decreases with uniform spatial step size, then maximum point-wise also monotonically decreases, and the rate convergence of the method is almost one. Table 3 also yields as temporal step size decreases for fixed spatial step size; then the results of maximum absolute point-wise error also decrease. From Figures 1 and 2, one can also observe that the mesh is dense near the initial, and hence, it indicates that the solution of the model example has an

initial layer at $t = 0$. The log-log plot of the scheme is also displayed in Figure 3, which confirms an agreement of the theoretical and numerical results. Finally, the result from the model example confirms that the proposed numerical method is convergent.

7 Conclusions

A nonpolynomial B-spline collocation method was implemented for singularly perturbed quasilinear Sobolev problems with initial boundary value problems. Newton's linearization method was applied to linearize the nonlinear parts. An implicit Euler method in time variable and cubic trigonometric B-spline collocation was used to approximate the space variable and obtain a three-term recurrence relation. Convergence analysis of the scheme was considered, and the scheme was accurate of order $O(h^2 + \tau)$. The results from the model example indicated the method is accurate for different values of ε, M, and N. In general, the effect of the perturbation parameter indicated that the scheme has a layer at initial points $t = 0$.

Acknowledgements

Authors are grateful to there anonymous referees and editor for their constructive comments.

References

[1] Amiraliyev, G.M. and Amiraliyeva, I.G. *Difference schemes for the singularly perturbed Sobolev equations,* In Difference Equations, Special Functions And Orthogonal Polynomials, 2007, 23–40.

[2] Amiraliyev, G.M., Duru, H. and Amiraliyeva, I.G. *A parameter-uniform numerical method for a Sobolev problem with initial layer,* Numer. Algorithms 44 (2007), 185–203.

[3] Amiraliyev, G.M. and Mamedov, Y.D. *Difference schemes on the uniform mesh for singularly perturbed pseudo-parabolic equations,* Turk. J. Math. 19 (3) (1995), 207–222.

[4] Barenblatt, G.I., Zheltov, I.P. and Kochina, I.N. *Basic concepts in the theory of seepage of homogeneous liquids in fissured rocks [strata]*, J. Appl. Math. Mech. 24 (5) (1960), 1286–1303.

[5] Chen, P.J. and Gurtin, M.E. *On a theory of heat conduction involving two temperatures*, J. Appl. Math. Phys. 19 (1968), 614–627.

[6] Ciftci, I. and Halilov, H. *Dependency of the solution of quasilinear pseudoparabolic equation with periodic boundary condition on* ε, Int. J. Math. Anal. 2 (2008), 881–888.

[7] Duressa, G.F. and Reddy, Y.N. *Domain decomposition method for singularly perturbed differential difference equations with layer behavior*, Int. J. Eng. Sci. 7 (1) (2015), 86–102.

[8] Duru, H. *Difference schemes for the singularly perturbed Sobolev periodic boundary problem*, Appl. Math. Comput. 149 (1) (2004), 187–201.

[9] Geng, F., Tang, Z. and Zhou, Y. *Reproducing kernel method for singularly perturbed one-dimensional initial-boundary value problems with exponential initial layers*, Qual.Theory Dyn. Syst. 17 (1) (2018), 177–187.

[10] Gunes, B. and Duru, H. *A second-order difference scheme for the singularly perturbed Sobolev problems with third type boundary conditions on Bakhvalov mesh*, J. Differ. Equ. 28 (3) (2004), 385–405.

[11] Huilgol, R.R. *A second order fluid of the differential type*, Int. J. Non-Linear Mech. 3 (4) (1968), 471–482.

[12] Jiwari, R. *Local radial basis function-finite difference based algorithms for singularly perturbed Burgers' model*, Math. Comput. Simul. 198 (2022), 106–126.

[13] Jiwrai, R. and Mittal, R.C. *A higher order numerical scheme for singularly perturbed Burger-Huxley equation*, J. Appl. Math. Inform. 29 (3-4) (2011), 813–829.

[14] Jiwari, R., Singh, S. and Singh, P. *Local RBF-FD-based mesh-free scheme for singularly perturbed convection-diffusion-reaction models with variable coefficients*, J. Math. 2022 (2022), 1–11.

[15] Kadalbajoo, M.K. and Patidar, K.C. *Singularly perturbed problems in partial differential equations: A survey*, Appl. Math. Comput 134 (2-3) (2003), 371–429.

[16] Kumar, N., Toprakseven, Ş. and Jiwari, R. *A numerical method for singularly perturbed convection–diffusion–reaction equations on polygonal meshes*, Comput. Appl. Math. 43 (1) (2024), 44.

[17] Mohapatra, J. and Shakti, D. *Numerical treatment for the solution of singularly perturbed pseudo-parabolic problem on an equidistributed grid*, Nonlinear Eng. 9 (1) (2020), 169–174.

[18] Nikolis, A. and Seimenis, I. *Solving dynamical systems with cubic trigonometric splines*, Appl. Math. [E-Notes] 5 (2005), 116–123.

[19] Schoenberg, I.J. *On trigonometric spline interpolation*, J. math. mech. (1964), 795–825.

[20] Van Duijn, C.J., Fan, Y., Peletier, L.A. and Pop, I.S. *Traveling wave solutions for degenerate pseudo-parabolic equations modeling two-phase flow in porous media*, Nonlinear Anal.: Real World Appl. 14 (3) (2013), 1361–1383.

[21] Vijayakumar, V., Udhayakumar, R. and Kavitha, K. *On the approximate controllability of neutral integro-differential inclusions of Sobolev-type with infinite delay*, Evol. Equ. Control Theory. 10 (2) (2021), 271–296.

[22] Zahra, W.K. *Trigonometric B-spline collocation method for solving PHI-four and Allen–Cahn equations*, Mediterr. J. Math. 14 (2017), 1–19.

Iranian Journal of Numerical Analysis and Optimization
Vol. 14, No. 3, 2024, pp 662–680
https://doi.org/10.22067/ijnao.2024.86104.1367
https://ijnao.um.ac.ir/

Research Article

Differential-integral Euler–Lagrange equations

M. Shehata*

Abstract

We study the calculus of variations problem in the presence of a system of differential-integral (D-I) equations. In order to identify the necessary optimality conditions for this problem, we derive the so-called D-I Euler–Lagrange equations. We also generalize this problem to other cases, such as the case of higher orders, the problem of optimal control, and we derive the so-called D-I Pontryagin equations. In special cases, these formulations lead to classical Euler–Lagrange equations. To illustrate our results, we provide simple examples and applications such as obtaining the minimum power for an RLC circuit.

AMS subject classifications (2020): 49J15, 49K15, 34H05.

Keywords: Calculus of variations; Euler–Lagrange equation; Optimal control problems; Differential-integral equation; RLC electrical circuit.

*Corresponding author

Received 28 December 2023; revised 22 February 2024; accepted 16 March 2024

Mohammed Shehata
Department of Basic Science, Bilbeis Higher Institute for Engineering, Sharqia, Egypt.
e-mail: mashehata_ math@yahoo.com, mashehata_ math@bhie.edu.eg

How to cite this article
Shehata, M., Differential-integral Euler–Lagrange equations. *Iran. J. Numer. Anal. Optim.*, 2024; 14(3): 662-680. https://doi.org/10.22067/ijnao.2024.86104.1367

1 Introduction

The calculus of variations began with Johann Bernoulli's Brachistochrone problem at the end of the 17th century. As a result of their work, Euler and Lagrange were able to develop a systematic way of dealing with this kind of problem by introducing what is now known as the Euler–Lagrange equation in the 18th century. This work was then extended in many ways by Bliss, Bolza, Caratheodory, Clebsch, Hahn, Hamilton, Hilbert, Kneser, Jacobi, Legendre, Mayer, Weierstrass, just to quote a few; see [4, 5, 11]. For an interesting historical book on one-dimensional problems of the calculus of variations, see [8].

The classical variational calculus has one major shortcoming despite its great success, it only deals with functionals containing derivatives. Many phenomena in nature can be modeled more accurately using differential integral equations. The application of these equations is found in science, biology, engineering, and economics; see [1, 2, 3, 6, 7, 9, 10, 12, 15, 16]. It is not worthwhile in applications to convert integrals into differentials, especially if there are many integrals of higher orders. In [13], an algorithm has been constructed to compute the exact solutions for the quadratic optimal control problem with integral constraints, and this algorithm has been used to find the optimal solution for single and coupled RC electrical circuits. In this paper, we identify differential-integral (D-I) Euler–Lagrange equations necessary conditions for a new class of variational problems in which a cost functional involves differential and integral operators.

2 Definitions and notations

Definition 1 (Lower and upper integrals)**.** For a given time horizon $[t_0, t_f]$, we define lower and upper integration of a continuous function $x : [t_0, t_f] \to \Re$ by

$$\underline{I}_K x = \int_{t_0}^{t} K(t,\tau)\, x(\tau)\, d\tau, \quad \overline{I}_K x = \int_{t}^{t_f} K(t,\tau)\, x(\tau)\, d\tau$$

with continuous kernel $K(t,\tau)$. We can define lower and upper higher order integrals as follows:

$$\underline{I}^2_{K_1K_2}x = \underline{I}_{K_1}\left(\underline{I}_{K_2}x\right), \qquad \overline{I}^2_{K_1K_2}x = \overline{I}_{K_1}\left(\overline{I}_{K_2}x\right),$$
$$\underline{I}^2_{K}x = \underline{I}_{K}\left(\underline{I}_{K}x\right), \qquad \overline{I}^2_{K}x = \overline{I}_{K}\left(\overline{I}_{K}x\right),$$

and so on.

Definition 2 (Complementary integral)**.** For a given time horizon $[t_0, t_f]$ and continuous function $x : [t_0, t_f] \to \Re$, we define the complement of the integral

$$\underline{I}_K x = \int_{t_0}^{t} K(t,\tau)\, x(\tau)\, d\tau, \text{ by } \overline{I}_{\overline{K}} x = \int_{t}^{t_f} \overline{K}(t,\tau)\, x(\tau)\, d\tau,$$

where $\overline{K}(t,s) := K(s,t)$.

For $K = 1$ we denote it by $\underline{I}_1 x = \underline{I}x, \qquad \overline{I}_1 x = \overline{I}x.$

Applying the Leibniz integral rule $n+1$ times to $\int_{t_0}^{t}(t-\tau)^n$ and $\int_{t_0}^{t_f}(\tau - t)^n$, respectively, we obtain the Cauchy formulas for repeated integration.

Theorem 1 (Cauchy formulas)**.** If $x(t)$ is a continuous function over $[t_0, t_f]$, then

1. $$\int_{t_0}^{t}(t-\tau)^n\, x(\tau)d\tau = n! \underbrace{\int_{t_0}^{t}\int_{t_0}^{\tau}\int_{t_0}^{\tau_1}\cdots\int_{t_0}^{\tau_{n-1}}}_{n+1\, times} x(\tau_n)d\tau_n d\tau_{n-1}\ldots d\tau_1 d\tau,$$
2. $$\int_{t}^{t_f}(\tau-t)^n\, x(\tau)d\tau = n! \underbrace{\int_{t}^{t_f}\int_{\tau}^{t_f}\int_{\tau_1}^{t_f}\cdots\int_{\tau_{n-1}}^{t_f}}_{n+1\, times} x(\tau_n)d\tau_n\, d\tau_{n-1}\ldots d\tau_1 d\tau.$$

From this theorem, we can define lower and upper higher integrals $\underline{I}^n\, x$, $\overline{I}^n x$ by

$$\underline{I}^n\, x = \frac{1}{(n-1)!}\int_{t_0}^{t}(t-\tau)^{n-1}x(\tau)\, d\tau,$$
$$\overline{I}^n x = \frac{1}{(n-1)!}\int_{t}^{t_f}(\tau-t)^{n-1}x(\tau)\, d\tau.$$

Through this paper, $D = \dfrac{d}{dt}$, and in general $D^n = \dfrac{d^n}{dt^n}$.

Note that

(i) $D^n\ (\underline{I}^n\, x(t)) = x(t)$ and $D^n\left(\overline{I}^n\, x(t)\right) = (-1)^n x(t)$.

(ii) $\underline{I}\ (D\, x(t)) = x(t) - x(t_0)$ and $\overline{I}\ (D\, x(t)) = x(t_f) - x(t)$.

3 D-I Euler–Lagrange equations

The first simplest D-I variations problem with fixed ends can be defined as follows: Among all functions $x(t)$ that satisfy the fixed end conditions

$$x(t_0) = x_0, \quad x(t_f) = x_f, \tag{1}$$

find the function for which the functional

$$J(x) = \int_{t_0}^{t_f} f\left(t, x(t), D\,x(t), \underline{I}_{K_1} x(t), \overline{I}_{K_2} x(t)\right) dt, \tag{2}$$

is an extremum. We assume that $f : [t_0, t_f] \times \Re^4 \to \Re$ has continuous first and second partial derivatives with respect to all of its arguments.

To derive the necessary conditions for the extremum, assume that $x = x(t)$ is the desired curve, and take some admissible curve $x = \bar{x}(t)$ close to $x = x(t)$ and include the curves $x = x(t)$ in one parameter family of curves

$$x(t, \epsilon) = x(t) + \epsilon\eta, \quad \eta = \bar{x}(t) - x(t), wheretbelongsto[t_0, t_f].$$

If one considers the values of the functional (2) only on curves of the family $x(t, \epsilon)$, then the functional becomes a function of ϵ:

$$J(y(\epsilon)) = \varphi(\epsilon).$$

This function $\varphi(\epsilon)$ is extremized for $\epsilon = 0$ since for $\epsilon = 0$ we have $x = x(t)$. The necessary conditions for the extremum of the function $\varphi(\epsilon)$ for $\epsilon = 0$ is as we know that $\varphi'(0) = 0$. Therefore we have proved the following lemma.

Lemma 1 (First variation condition)**.** If $x = x(t)$ is a solution to problem (1)–(2), then $\dfrac{\partial}{\partial\epsilon}\left(J(x + \epsilon\eta)\,|_{\epsilon=0} = 0\right.$, for some functions $\eta(t)$ satisfies $\eta(t_0) = \eta(t_f) = 0$.

We also know from the calculus of variations, the following fundamental lemma.

Lemma 2 (The fundamental lemma)**.** If for every continuous function $\eta(t)$

$$\int_{t_0}^{t_f} \Psi(t)\eta(t)dt = 0,$$

where the function $\Psi(t)$ is continuous on the interval $[t_0, t_f]$, then $\Psi(t) \equiv 0$ on that interval.

From the above two lemmas, we will prove the following theorem.

Theorem 2 (D-I Euler Lagrange conditions)**.** If $x = x(t)$ is a solution of problem (1)–(2), then

$$\frac{\partial f}{\partial x} - D\left(\frac{\partial f}{\partial Dx}\right) + \overline{I}_{\overline{K_1}}\left(\frac{\partial f}{\partial \underline{I}_{K_1} x}\right) + \underline{I}_{\overline{K_2}}\left(\frac{\partial f}{\partial \overline{I}_{K_2} x}\right) = 0. \tag{3}$$

Proof. By Lemma 1, if $x = x(t)$ is a solution of Problem 1, then $\frac{\partial}{\partial \epsilon}\left(J(x+\epsilon\eta)\right|_{\epsilon=0} = 0$, for some functions $\eta(t)$ satisfying $\eta(a) = \eta(b) = 0$, and it follows that

$$\int_{t_0}^{t_f} \left[\frac{\partial f}{\partial x}\eta + \frac{\partial f}{\partial Dx} D\eta + \frac{\partial f}{\partial \underline{I}_{K_1} x}\underline{I}_{K_1}\eta + \frac{\partial f}{\partial \overline{I}_{K_2} x}\overline{I}_{K_2}\eta\right] dt = 0. \tag{4}$$

We integrate the second term by parts, and we get

$$\begin{aligned}\int_{t_0}^{t_f} \frac{\partial f}{\partial Dx} D\eta dt &= \left[\frac{\partial f}{\partial x}(t_f)\eta(t_f) - \frac{\partial f}{\partial Dx}(t_0)\eta(t_0)\right] - \int_{t_0}^{t_f} D\left(\frac{\partial f}{\partial Dx}\right)\eta dt \\ &= -\int_{t_0}^{t_f} D\left(\frac{\partial f}{\partial Dx}\right)\eta dt. \end{aligned} \tag{5}$$

By changing the order of the integrations in the third and fourth term in (4), we get

$$\int_{t_0}^{t_f} \frac{\partial f}{\partial \underline{I}_{K_1} x}\underline{I}_{K_1}\eta dt = \int_{t_0}^{t_f} \overline{I}_{\overline{K_1}}\left(\frac{\partial f}{\partial \underline{I}_{K_1} x}\right)\eta dt, \tag{6}$$

$$\int_{t_0}^{t_f} \frac{\partial f}{\partial \overline{I}_{K_2} x}\overline{I}_{K_2}\eta dt = \int_{t_0}^{t_f} \underline{I}_{\overline{K_2}}\left(\frac{\partial f}{\partial \overline{I}_{K_2} x}\right)\eta dt. \tag{7}$$

Thus, substituting (5), (6), and (7) back into (4), gives us

$$\begin{aligned}&\left[\frac{\partial f}{\partial Dx}(t_f)\eta(t_f) - \frac{\partial f}{\partial Dx}(t_0)\eta(t_0)\right] \\ &+ \int_{t_0}^{t_f}\left[\frac{\partial f}{\partial x} - D\left(\frac{\partial f}{\partial Dx}\right) + \overline{I}_{\overline{K_1}}\left(\frac{\partial f}{\partial \underline{I}_{K_1} x}\right) + \underline{I}_{\overline{K_2}}\left(\frac{\partial f}{\partial \overline{I}_{K_2} x}\right)\right]\eta dt = 0.\end{aligned} \tag{8}$$

Finally, from Lemma 2 and $\eta(t_0) = \eta(t_f) = 0$, we obtain the desired D-I Euler–Lagrange equation (3). □

Remark 1. By substituting $t = t_0$ in (3), we obtain the natural condition

$$\left[\frac{\partial f}{\partial x} - D\left(\frac{\partial f}{\partial Dx}\right) + \int_{t_0}^{t_f} \overline{K_1}\left(\frac{\partial f}{\partial \underline{I}_{K_1} x}\right) d\tau\right]_{\tau=t_0} = 0, \tag{9}$$

and by substituting $t = t_f$ in (3), we obtain the natural condition

$$\left[\frac{\partial f}{\partial x} - D\left(\frac{\partial f}{\partial Dx}\right) + \int_{t_0}^{t_f} \overline{K_2}\left(\frac{\partial f}{\partial \overline{I}_{K_2} x}\right) d\tau\right]_{\tau=t_f} = 0. \tag{10}$$

Special cases. There are some special cases of D-I Euler–Lagrange, which are important in many applications:

case 1. If f is independent of $\underline{I}_K x$, then D-I Euler–Lagrange conditions are reduced to so called $(D - \overline{I})$ Euler–Lagrange equation:

$$\frac{\partial f}{\partial x} - D\left(\frac{\partial f}{\partial Dx}\right) + \underline{I}_{\overline{K}}\left(\frac{\partial f}{\partial \overline{I}_K x}\right) = 0,$$

and if $K = (\tau - t)^n$, then $(D - \overline{I})$ Euler–Lagrange conditions become

$$\frac{\partial f}{\partial x} - D\left(\frac{\partial f}{\partial Dx}\right) + \underline{I}^n\left(\frac{\partial f}{\partial \overline{I}^n x}\right) = 0.$$

case 2. If f is independent of $\underline{I}_K x$, then D-I Euler–Lagrange conditions are reduced to so called $(D - \underline{I})$ Euler–Lagrange equation:

$$\frac{\partial f}{\partial x} - D\left(\frac{\partial f}{\partial Dx}\right) + \overline{I}_{\overline{K}}\left(\frac{\partial f}{\partial \underline{I}_K x}\right) = 0,$$

and if $K = (t - \tau)^n$, then $(D - \underline{I})$ Euler–Lagrange conditions become

$$\frac{\partial f}{\partial x} - D\left(\frac{\partial f}{\partial Dx}\right) + \overline{I}^n\left(\frac{\partial f}{\partial \underline{I}^n x}\right) = 0.$$

case 3. If f is independent of both $\underline{I}_{K_1} x$, $\overline{I}_{K_2} x$, then D-I Euler–Lagrange conditions are reduced to the usual Euler–Lagrange equation:

$$\frac{\partial f}{\partial x} - D\left(\frac{\partial f}{\partial Dx}\right) = 0.$$

4 Generalizations

In this section, we generalized the fixed boundaries problem to the cases of integral with deferent kernels, moving boundaries, higher order, and several independent variables.

Integral with different kernel

Consider the functional

$$J(x) = \int_{t_0}^{t_f} f\left(t, x, Dx, \underline{I}_{K_{11}} x, \underline{I}_{K_{12}} x, \ldots, \underline{I}_{K_{1\ell}} x, \overline{I}_{K_{21}} x, \overline{I}_{K_{22}} x, \ldots, \overline{I}_{K_{2k}} x\right) dt, \tag{11}$$

where $f : [t_0, t_f] \times \Re^{2+m+\ell} \to \Re$ has continuous partial derivatives up to the order two with respect to all its arguments. Moreover, t_0 and t_f are specified, and the boundary conditions are

$$x(t_0) = x_0, \quad x(t_f) = x_f.$$

For this case, following the above approach, we obtain the following necessary conditions

$$\frac{\partial f}{\partial Dx} - D\left(\frac{\partial f}{\partial Dx}\right) + \sum_{j=1}^{\ell} \overline{I}_{\overline{K_{1j}}} \left(\frac{\partial f}{\partial \underline{I}_{K_{1j}} x}\right) + \sum_{j=1}^{k} \underline{I}_{\overline{K_{2j}}} \left(\frac{\partial f}{\partial \overline{I}_{K_{2j}} x}\right) = 0. \tag{12}$$

Moving boundaries

Let the terminal conditions at $t = t_0$ and/or at $t = t_f$ not be specified. For this case, following the above approach, we obtain the D-I Euler–Lagrange equation given by (3), and the following transversally conditions:

$$\left.\frac{\partial f}{\partial Dx}\right]_{t=t_0} = 0, \quad \text{if } x(t_0) \text{ is not satisfied and}$$

$$\left.\frac{\partial f}{\partial Dx}\right]_{t=t_f} = 0, \quad \text{if } x(t_f) \text{ is not satisfied.}$$

Higher order

Consider the functional

$$J(x)=\int_{t_0}^{t_f} f\left(t,x,Dx,\ldots,D^m x,\underline{I}_{K_1}x,\underline{I}^2_{K_1}x,\ldots,\underline{I}^\ell_{K_1}x,\overline{I}_{K_2}x,\overline{I}^2_{K_2}x,\ldots,\overline{I}^k_{K_2}x\right)dt, \tag{13}$$

where $f:[t_0,t_f]\times\Re^{1+m+\ell+k}\to\Re$ has continuous partial derivatives up to the order $m+1$ with respect to all its arguments. Moreover, t_0 and t_f are specified, and the boundary conditions are

$$\begin{aligned} x(t_0)&=x_0, & x(t_f)&=x_f,\\ &\vdots & &\vdots\\ D^m x(t_0)&=x_{m0}, & D^m x(t_f)&=x_{mf}. \end{aligned}$$

For this case, following the above approach, we obtain the following necessary conditions:

$$\sum_{i=0}^{m}(-1)^i D^i\left(\frac{\partial f}{\partial D^i x}\right)+\sum_{j=1}^{\ell}\overline{I}^j_{K_1}\left(\frac{\partial f}{\partial\,\underline{I}^j_{K_1}x}\right)+\sum_{j=1}^{k}\underline{I}^j_{K_2}\left(\frac{\partial f}{\partial\,\overline{I}^j_{K_2}x}\right)=0. \tag{14}$$

Several independent variables

Consider the functional

$$\begin{aligned} &J(x_1,\ldots,x_n) \\ &=\int_{t_0}^{t_f} f\left(t,x_1,\ldots,x_n,x_1',\ldots,x_n',\ldots,\underline{I}_{K_1}x_1,\ldots,\underline{I}_{K_1}x_n,\overline{I}_{K_2}x_1,\ldots,\overline{I}_{K_2}x_n\right)dt, \end{aligned} \tag{15}$$

where $x_1,x_2,\ldots,x_n$ are independent functions with continuous first derivatives and $f:[t_0,t_f]\times\Re^{4n}\to\Re$ has continuous first and second partial derivatives with respect to all of its arguments. Moreover, t_0 and t_f are specified, and the boundary conditions are

$$\begin{aligned} x_1(t_0)&=x_{10}, & x_1(t_f)&=x_{1f},\\ &\vdots & &\vdots\\ x_n(t_0)&=x_{n0}, & x_n(t_f)&=x_{nf}. \end{aligned}$$

For this case, following the above approach, we obtain the following necessary conditions:

$$\frac{\partial f}{\partial x_i} - D\left(\frac{\partial f}{\partial D x_i}\right) + \overline{I}_{\overline{K_1}}\left(\frac{\partial f}{\partial \underline{I}_{K_1} x_i}\right) + \underline{I}_{\overline{K_2}}\left(\frac{\partial f}{\partial \overline{I}_{K_2} x_i}\right) = 0, \quad i = 1, 2, \ldots, n. \tag{16}$$

5 D-I optimal control problem

We shall consider the class of control problems where the dynamical system is described by the following ordinary $D - \underline{I}$ equations:

$$Dx = f(t, x, \underline{I}_{K_1} x, \overline{I}_{K_2} x, u), \tag{17}$$

$$x(t_0) = x_0, \quad t_0 \text{ and } t_f \text{ are specified}, \tag{18}$$

where $x(t)$ an n-vector function is determined by $u(t)$ an m-vector function, with $x \in \Re^n$, $u \in \Re^m$.

The performance of the system is measured by the cost functional:

$$J(x) = S(t_f, x(t_f)) + \int_{t_0}^{t_f} L(t, x, \underline{I}_{K_1} x, \overline{I}_{K_2} x, u) dt. \tag{19}$$

The problem is to find the functions $u(t)$ that minimize (or maximize) $)J$. It is assumed that $f(t, x, \underline{I}_{K_1} x, \overline{I}_{K_2} x, u)$ and $L(t, x, \underline{I}_{K_1} x, \overline{I}_{K_2} x, u)$ are continuous for all $t \in [t_0, t_f]$, $x \in \Re^n$, $u \in \Re^m$, and have continuous derivative up to the second order.

Theorem 3 (D-I (Pontryagin))**.** If $u(t)$ is a solution to the problem (17)–(19), then the following equations are satisfied:

state equations

$$Dx = \frac{\partial H}{\partial \lambda} = f(t, x, \underline{I}_{K_1} x, \overline{I}_{K_2} x, u); \tag{20}$$

$$x(t_0) = x_0; \tag{21}$$

adjoint equations

$$-D\lambda = \left(\frac{\partial H}{\partial x}\right)^T + \overline{I}_{\overline{K_1}}\left(\frac{\partial F}{\partial \underline{I}_{K_1} x}\right)^T + \underline{I}_{\overline{K_2}}\left(\frac{\partial F}{\partial \overline{I}_{K_2} x}\right)^T; \tag{22}$$

optimality conditions

$$0 = \left(\frac{\partial H}{\partial u}\right)^T; \tag{23}$$

transversality condition

$$\lambda(t_f) = \left(\frac{\partial S}{\partial x}\right)^T\Bigg]_{t=t_f}; \tag{24}$$

where

$$H = L(t, x, \underline{I}_{K_1}x, \overline{I}_{K_2}x, u) + \lambda^T\left(t, x, \underline{I}_{K_1}x, \overline{I}_{K_2}x, u\right) \tag{25}$$

is the usual Hamiltonian.

Proof. First $S(t_f, x(t_f))$ can be written as

$$S(t_f, x(t_f)) = S(t_0, x(t_0)) + \int_{t_0}^{t_f} \frac{d}{dt}S(t, x(t))dt \tag{26}$$

$$= S(t_0, x(t_0)) + \int_{t_0}^{t_f} \left[\frac{\partial S}{\partial t} + \frac{\partial S}{\partial x}x'\right] dt. \tag{27}$$

Equation (19) becomes

$$J(x) = S(t_0, x(t_0)) + \int_{t_0}^{t_f} L(t, x, \underline{I}_{K_1}x, \overline{I}_{K_2}x, u) + \left[\frac{\partial S}{\partial t} + \frac{\partial S}{\partial x}x'\right] dt. \tag{28}$$

Adjoin the system differential equations (19) to J with multiplier functions $\lambda(t)$ and we have

$$\hat{J}(x) = S(t_0, x(t_0)) + \int_{t_0}^{t_f} H - \lambda^T Dx + \left[\frac{\partial S}{\partial t} + \frac{\partial S}{\partial x}x'\right] dt$$

$$= S(t_0, x(t_0)) + \int_{t_0}^{t_f} F(t, x, Dx, \underline{I}_{K_1}x, \overline{I}_{K_2}x, u, \lambda), \tag{29}$$

where $F(t, x, Dx, \underline{I}_{K_1}x, \overline{I}_{K_2}x, u, \lambda) = H - \lambda^T Dx + \left[\frac{\partial S}{\partial t} + \frac{\partial S}{\partial x}x'\right]$.

Following the same approach in the calculus of variations ((8)) gives

$$\left[\frac{\partial F}{\partial Dx}(t_f)\eta(t_f) - \frac{\partial F}{\partial Dx}(t_0)\eta(t_0)\right]$$
$$+ \int_{t_0}^{t_f} \left[\frac{\partial F}{\partial x} - D\left(\frac{\partial F}{\partial Dx}\right) + \overline{I}_{\overline{K_1}}\left(\frac{\partial F}{\partial \underline{I}_{K_1}x}\right) + \underline{I}_{\overline{K_2}}\left(\frac{\partial F}{\partial \overline{I}_{K_2}x}\right)\right] \eta dt = 0$$

for some $\eta(t_0) = 0$.

From the definition of F and the fact that the D-I Euler equation must be satisfied, we have

$$\begin{aligned}
&\frac{\partial F}{\partial x} - D\left(\frac{\partial F}{\partial Dx}\right) + \overline{I}_{\overline{K_1}}\left(\frac{\partial F}{\partial \underline{I}_{K_1} x}\right) + \underline{I}_{\overline{K_2}}\left(\frac{\partial F}{\partial \overline{I}_{K_2} x}\right) \\
&\quad = \frac{\partial H}{\partial x} + \overline{I}_{\overline{K_1}}\left(\frac{\partial H}{\partial \underline{I}_{K_1} x}\right) + \underline{I}_{\overline{K_2}}\left(\frac{\partial H}{\partial \overline{I}_{K_2} x}\right) + \frac{\partial}{\partial x}\left[S_t + S_x x'\right] \\
&\qquad + D\left(\lambda^T - S_x\right) \\
&\quad = \frac{\partial H}{\partial x} + \overline{I}_{\overline{K_1}}\left(\frac{\partial F}{\partial \underline{I}_{K_1} x}\right) + \underline{I}_{\overline{K_2}}\left(\frac{\partial F}{\partial \overline{I}_{K_2} x}\right) + D\left(\lambda^T\right) = 0. \qquad (30)
\end{aligned}$$

This gives (22). Similarly, λ and u being independent variables, then

$$\begin{aligned}
\frac{\partial F}{\partial \lambda} = \frac{\partial F}{\partial \lambda} = \frac{\partial H}{\partial \lambda} - Dx = 0, \\
\frac{\partial F}{\partial u} = \frac{\partial F}{\partial u} = \frac{\partial H}{\partial u} = 0.
\end{aligned}$$

This gives (20) and (23), respectively. Finally, the transversally or boundary conditions given by the remaining terms of (30) are

$$\frac{\partial F}{\partial x'}(t_f)\eta(t_f) = \left[\frac{\partial S}{\partial x} - \lambda^T\right]\eta(t_f) = 0. \qquad (31)$$

The fact that $\eta(t_f)$ does not vanish, yields (24). □

6 Examples

To illustrate our result, we give some examples.

Example 1. In this example, we want to find the unknown supplied voltage $u(t)$ for the RLC circuit in Figure 1, which minimizes the cost functional given by

$$J = \frac{1}{2}i^2(5) + \frac{1}{2}\int_0^5 u^2(t)dt. \qquad (32)$$

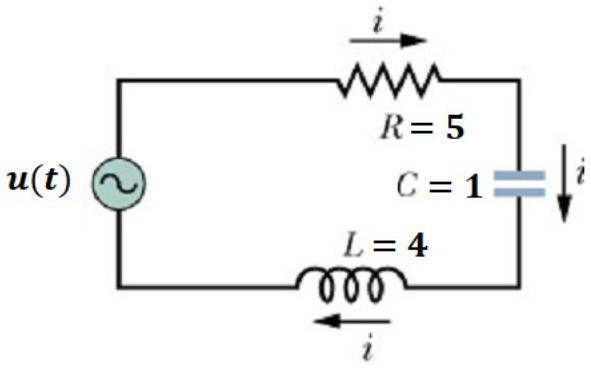


Figure 1: Series RLC circuit.

By applying the Kirchhoff's voltage law, we get

$$4\,\frac{d}{dt}\,i(t) + 5\,i(t) + \int_0^t i(\tau)d\tau = u(t). \tag{33}$$

By applying D-I Pontryagin necessary conditions to this problem with $x \equiv i$, $t_0 = 0$, $t_f = 5$ and

$$H = \frac{1}{2}u^2(t) + \lambda(t)\left[-\frac{5}{4}\,i(t) - \frac{1}{4}\int_0^t i(\tau)d\tau + \frac{1}{4}\,u(t)\right], \tag{34}$$

the optimal control for the problem (32)–(33) is characterized by

$$u(t) = -\frac{1}{4}\,\lambda(t),$$

where $i(t)$ and $\lambda(t)$ satisfy the following equations:

State equations

$$\frac{d\,i(t)}{dt} = -\frac{5}{4}\,i(t) - \frac{1}{4}\int_0^t i(\tau)d\tau - \frac{1}{16}\,\lambda(t), \tag{35}$$

$$i(0) = 1, \tag{36}$$

Adjoint equations

$$\frac{d\,\lambda(t)}{dt} = \frac{5}{4}\lambda(t) + \frac{1}{4}\int_t^5 \lambda(\tau)\,d\tau, \tag{37}$$

$$\lambda(5) = i(5). \tag{38}$$

Remark 2. Equations (35) and (37) provide the necessary conditions for the problem. They constitute two second order D-I equations whose solution contains four constants of integration. To evaluate these, we have 1-equation $i(0) = 1$, 1-equation $\lambda(5) = i(5)$, 1-equation $\underline{I}i(t) = 0$ at $t = 0$ and 1-equation $\overline{I}\lambda(t) = 0$ at $t = 5$.

To solve the adjoint equation (37), let

$$\lambda_1(t) = \int_t^5 \lambda(\tau)\, d\tau, \quad \lambda_2(t) = \frac{d\lambda_1(t)}{dt} = -\lambda(t).$$

Then (37)–(38) can be written in the following matrix form:

$$\begin{bmatrix} \lambda_1(t) \\ \lambda_2(t) \end{bmatrix}' = \begin{bmatrix} 0 & 1 \\ \frac{-1}{4} & \frac{5}{4} \end{bmatrix} \begin{bmatrix} \lambda_1(t) \\ \lambda_2(t) \end{bmatrix}$$

with final conditions:

$$\begin{bmatrix} \lambda_1(5) \\ \lambda_2(5) \end{bmatrix} = \begin{bmatrix} 0 \\ -i(5) \end{bmatrix},$$

which have the solution

$$\begin{bmatrix} \lambda_1(t) \\ \lambda_2(t) \end{bmatrix} = e^{\begin{bmatrix} 0 & 1 \\ \frac{-1}{4} & \frac{5}{4} \end{bmatrix}(t-5)} \begin{bmatrix} 0 \\ -i(5) \end{bmatrix}.$$

Now (see, for example, [14]),

$$\begin{bmatrix} 0 & 1 \\ \frac{-1}{4} & \frac{5}{4} \end{bmatrix} = \begin{bmatrix} \frac{-1}{3} & \frac{4}{3} \\ \frac{-1}{3} & \frac{1}{3} \end{bmatrix} \begin{bmatrix} 1 & 0 \\ 0 & \frac{1}{4} \end{bmatrix} \begin{bmatrix} \frac{-1}{3} & \frac{4}{3} \\ \frac{-1}{3} & \frac{1}{3} \end{bmatrix}^{-1}.$$

Then

$$\begin{aligned} e^{\begin{bmatrix} 0 & 1 \\ \frac{-1}{4} & \frac{5}{4} \end{bmatrix}(t-5)} &= \begin{bmatrix} \frac{-1}{3} & \frac{4}{3} \\ \frac{-1}{3} & \frac{1}{3} \end{bmatrix} \begin{bmatrix} e^{(t-5)} & 0 \\ 0 & e^{\frac{1}{4}(t-5)} \end{bmatrix} \begin{bmatrix} 1 & -4 \\ 1 & -1 \end{bmatrix} \\ &= \frac{1}{3} \begin{bmatrix} -e^{5-t} + 4e^{\frac{1}{4}(5-t)} & 4e^{5-t} - 4e^{\frac{1}{4}(5-t)} \\ -e^{5-t} + e^{\frac{1}{4}(5-t)} & 4e^{5-t} - e^{\frac{1}{4}(5-t)} \end{bmatrix}. \end{aligned}$$

So,

$$\lambda(t) = -\lambda_2(t) = \frac{i(5)}{3}\left[4\, e^{5-t} - \, e^{\frac{1}{4}(5-t)}\right].$$

To solve the state equation (35), let

$$i_1(t) = \int_0^t i(\tau)\, d\tau, \quad i_2(t) = \frac{di_1(t)}{dt} = i(t).$$

Then (35)–(36) can be written in the following nonhomogeneous matrix form

$$\begin{bmatrix} i_1(t) \\ i_2(t) \end{bmatrix}' = \begin{bmatrix} 0 & 1 \\ \frac{-1}{4} & \frac{-5}{4} \end{bmatrix} \begin{bmatrix} i_1(t) \\ i_2(t) \end{bmatrix} + \psi(t, i(5))$$

with initial conditions:

$$\begin{bmatrix} i_1(0) \\ i_2(0) \end{bmatrix} = \begin{bmatrix} 0 \\ 1 \end{bmatrix},$$

where $\psi(t, i(5)) = \begin{bmatrix} 0 \\ \frac{-i(5)}{16}\left[\frac{4}{3}\, e^{5-t} - \frac{1}{3}\, e^{\frac{1}{4}(5-t)}\right] \end{bmatrix}$, which have the solution

$$\begin{aligned} \begin{bmatrix} i_1(t) \\ i_2(t) \end{bmatrix} &= e^{\begin{bmatrix} 0 & 1 \\ \frac{-1}{4} & \frac{-5}{4} \end{bmatrix} t} \begin{bmatrix} 0 \\ 1 \end{bmatrix} + \int_0^t e^{\begin{bmatrix} 0 & 1 \\ \frac{-1}{4} & \frac{-5}{4} \end{bmatrix} (t-\tau)} \psi(\tau, i(5)) d\tau \\ &= \begin{bmatrix} \frac{-4}{3} e^{-t} + \frac{4}{3} e^{\frac{-t}{4}} \\ \frac{4}{3} e^{-t} - \frac{1}{3} e^{\frac{-t}{4}} \end{bmatrix} \\ &\quad - \frac{i(5)}{16} \int_0^t \begin{bmatrix} \left(\frac{-4}{3} e^{\tau - t} + \frac{4}{3} e^{\frac{1}{4}(\tau - t)}\right)\left(\frac{4}{3}\, e^{5-t} - \frac{1}{3}\, e^{\frac{1}{4}(5-t)}\right) \\ \left(\frac{4}{3} e^{\tau - t} - \frac{1}{3} e^{\frac{1}{4}(\tau - t)}\right)\left(\frac{4}{3}\, e^{5-t} - \frac{1}{3}\, e^{\frac{1}{4}(5-t)}\right) \end{bmatrix} d\tau. \end{aligned}$$

So,

$$\begin{aligned} i(t) = i_2(t) &= \frac{4}{3} e^{-t} - \frac{1}{3} e^{\frac{-t}{4}} + \frac{i(5)}{9}\left[e^{5-2t} - \frac{1}{4} e^{\frac{5}{4} - \frac{5}{4}t} - e^{5 - \frac{5}{4}t} + \frac{1}{4} e^{\frac{5}{4} - \frac{1}{2}t}\right] \\ \Rightarrow i(5) &= \frac{4\left(4e^{-5} - e^{-\frac{5}{4}}\right)}{12 - e^{-5} + e^{\frac{-5}{4}}}. \end{aligned}$$

Hence, we obtain the control

$$u(t) = -\frac{\left(4e^{-5} - e^{-\frac{5}{4}}\right)\left(e^{5-t} - e^{\frac{1}{4}(5-t)}\right)}{3\left(12 - e^{-5} + e^{\frac{-5}{4}}\right)} \tag{39}$$

and the current (see Figure 2)

$$i(t) = \frac{4}{3}e^{-t} - \frac{1}{3}e^{\frac{-t}{4}} + \frac{4\left(4e^{-5} - e^{-\frac{5}{4}}\right)\left(e^{5-2t} - \frac{1}{4}e^{\frac{5}{4}-\frac{5}{4}t} - e^{5-\frac{5}{4}t} + \frac{1}{4}e^{\frac{5}{4}-\frac{1}{2}t}\right)}{9\left(12 - e^{-5} + e^{\frac{-5}{4}}\right)}. \tag{40}$$

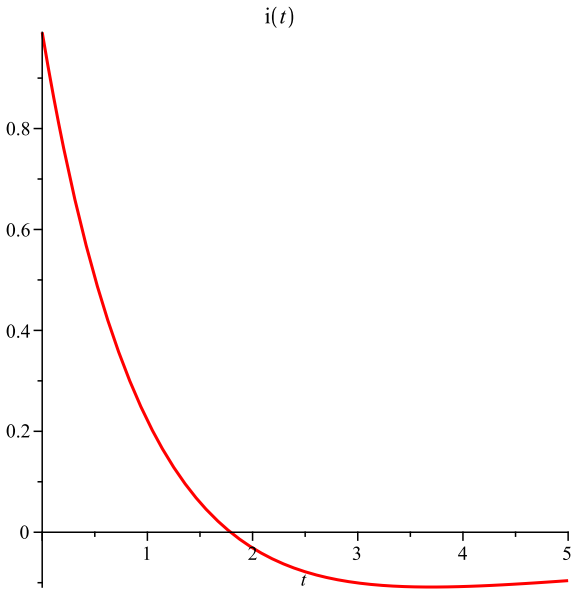


Figure 2: Optimal electrical current.

Example 2. In this example, we want to find $u(t)$ that minimizes the cost functional given by

$$J = \frac{1}{2}\int_0^2 u^2(t)dt + \frac{1}{2}\int_0^2 \left[\int_0^t (t-\tau)x(\tau)\,d\tau\right]^2 dt$$

with constraints

$$Dx(t) = u(t), \quad 0 < t \leq 2,$$
$$x(0) = 1.$$

By applying D-I Pontryagin necessary conditions to this problem, the optimal control is characterized by

$$u = -\lambda,$$

$$\begin{aligned} Dx &= -\lambda, \\ -D\lambda &= \int_t^2 \left\{ (\tau - t) \int_0^\tau (\tau - s) x(s)\, ds \right\} d\tau, \\ x(0) &= 1, \\ \lambda(2) &= 0. \end{aligned}$$

The above system is simplified to the following equations:

$$Dx(t) = -\int_t^2 \left\{ \int_\tau^2 (r - \tau_1) \int_0^r (r - s) x(s)\, ds dr \right\} d\tau_1\, d\tau, \tag{41}$$

$$x(0) = 1. \tag{42}$$

To solve (41)–(42). Let $x_1 = x(t)$, $x_2 = \int_0^t x_1(\tau)\, d\tau$, $x_3(t) = \int_0^t x_2(\tau)\, d\tau$, $x_4(t) = \int_t^2 x_3(\tau)\, d\tau$, $x_5(t) = \int_t^2 x_4(\tau)\, d\tau$ and $x_6(t) = \int_t^2 x_5(\tau)\, d\tau$. Then (41)–(42) is equivalent to the following system:

$$\begin{bmatrix} x_1 \\ x_2 \\ x_3 \\ x_4 \\ x_5 \\ x_6 \end{bmatrix}' = \begin{bmatrix} 0 & 0 & 0 & 0 & 0 & -1 \\ 1 & 0 & 0 & 0 & 0 & 0 \\ 0 & 1 & 0 & 0 & 0 & 0 \\ 0 & 0 & -1 & 0 & 0 & 0 \\ 0 & 0 & 0 & -1 & 0 & 0 \\ 0 & 0 & 0 & 0 & -1 & 0 \end{bmatrix} \begin{bmatrix} x_1 \\ x_2 \\ x_3 \\ x_4 \\ x_5 \\ x_6 \end{bmatrix}$$

with

$$x_1(0) = 1, \quad x_2(0) = 0, \quad x_3(0) = 0, \quad x_4(2) = 0, \quad x_5(2) = 0, \quad x_6(2) = 0,$$

which leads to the graph of $x(t)$ as shown in Figure 3.

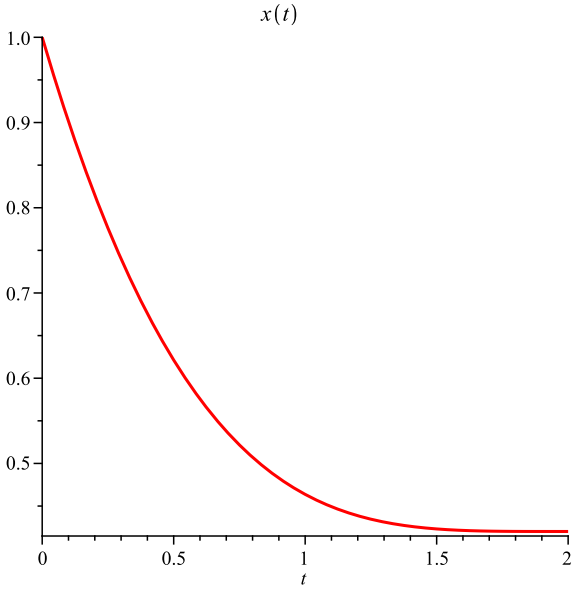


Figure 3: Optimal state solution $x(t)$.

7 Conclusion

In this paper, we have identified D-I Euler–Lagrange equations necessary conditions for a new class of variational problems in which a cost functional involving differential and integral operators. We concluded that if Euler–Lagrange equations contain an integral, then they must contain the complementary integral. We also generalized results to other problems.

Declarations

Conflict of interest: The author declare that he has no conflict of interest.

References

[1] Andrade, B. *On the well-posedness of a Volterra equation with applications in the Navier-Stokes problem*, Math. Methods Appl. Sci. 41 (2018), 750–768.

[2] Angell, T.S. *On the optimal control of systems governed by nonlinear Volterra equations*, J. Optim. Theory Appl. 19 (1976), 29–45.

[3] Belbas, S.A. *A reduction method for optimal control of Volterra integral equations*, Appl. Math. Comput. 197 (2008), 880–890.

[4] Brunt, B. *The calculus of variations*, Universitext, Springer-Verlag, New York, 2004.

[5] Dacorogna, B. *Introduction to the calculus of variations*, Imperial College Press, London 3rd ed., 2014.

[6] Dmitruk, A.V. and Osmolovskii, N.P. *Necessary conditions for a weak minimum in optimal control problems with integral equations subject to state and mixed constraints*, SIAM J. Control Optim. 52 (2014), 3437–3462.

[7] Ebrahimzadeh, A. *A robust method for optimal control problems governed by system of Fredholm integral equations in mechanics*, Iranian Journal of Numerical Analysis and Optimization 13 (2023), 243–261.

[8] Goldstine, H.H. *A history of the calculus of variations from the 17th to the 19th century*, Springer, Berlin, 1980.

[9] Han, S., Lin, P. and Yong, J. *Causal state feedback representation for linear quadratic optimal control problems of singular Volterra integral equations*, arXiv preprint arXiv:2109.07720 (2021).

[10] Kamien, M.I. and Muller, E. *Optimal control with integral state equations*, The Review of Economic Studies 43 (1976), 469–473.

[11] Liberzon, D. *Calculus of variations and optimal control theory: A concise introduction*, Princeton University Press, 2012.

[12] Shehata, M. *From calculus to α calculus*, Progr. Fract. Differ. Appl, Accepted for publication in Volume 10, 3 july (2024).

[13] Shehata, M. *Computing exact solution for linear integral quadratic control problem*, Egyptian Journal of Pure and Applied Science 62 (2024), 33–42.

[14] Shehata, M. and Khalil, A.A. *Algorithm for computing exact solution of the first order linear differential system*, Sohag J. Sci. 7 (2022), 71–77.

[15] Vega, C. *Necessary conditions for optimal terminal time control problems governed by a Volterra integral equation*, J. Optim. Theory Appl. 130 (2006), 79–93.

[16] Vijayakumar, V. *Approximate controllability results for analytic resolvent integro-differential inclusions in Hilbert spaces*, Int. J. Control 91 (2018,) 204–214.

Iranian Journal of Numerical Analysis and Optimization
Vol. 14, No. 3, 2024, pp 681–707
https://doi.org/10.22067/ijnao.2024.86692.1384
https://ijnao.um.ac.ir/

Research Article

An improved imperialist competitive algorithm for solving an inverse form of the Huxley equation

H. Dana Mazraeh, K. Parand*, H. Farahani and S.R. Kheradpisheh

*Corresponding author

Received 04 February 2024; revised 28 April 2024; accepted 01 May 2024

Hassan Dana Mazraeh
Department of Computer and Data Sciences, Faculty of Mathematical Sciences, Shahid Beheshti University, G.C. Tehran, Iran. e-mail: h_danamazraeh@sbu.ac.ir

Kourosh Parand
Department of Computer and Data Sciences, Faculty of Mathematical Sciences, Shahid Beheshti University, G.C. Tehran, Iran.
Department of Cognitive Modeling, Institute for Cognitive and Brain Sciences, Shahid Beheshti University, G.C. Tehran, Iran. e-mail: k_parand@sbu.ac.ir

Hadi Farahani
Department of Computer and Data Sciences, Faculty of Mathematical Sciences, Shahid Beheshti University, G.C. Tehran, Iran. e-mail: h_farahani@sbu.ac.ir

Saeed Reza Kheradpisheh
Department of Computer and Data Sciences, Faculty of Mathematical Sciences, Shahid Beheshti University, G.C. Tehran, Iran. e-mail: s_kheradpisheh@sbu.ac.ir

How to cite this article
Dana Mazraeh, H., Parand, K., Farahani, H. and Kheradpisheh, S.R., An improved imperialist competitive algorithm for solving an inverse form of the Huxley equation. *Iran. J. Numer. Anal. Optim.*, 2024; 14(3): 681-707. https://doi.org/10.22067/ijnao.2024.86692.1384

Abstract

In this paper, we present an improved imperialist competitive algorithm for solving an inverse form of the Huxley equation, which is a nonlinear partial differential equation. To show the effectiveness of our proposed algorithm, we conduct a comparative analysis with the original imperialist competitive algorithm and a genetic algorithm. The improvement suggested in this study makes the original imperialist competitive algorithm a more powerful method for function approximation. The numerical results show that the improved imperialist competitive algorithm is an efficient algorithm for determining the unknown boundary conditions of the Huxley equation and solving the inverse form of nonlinear partial differential equations.

AMS subject classifications (2020): Primary 68W50; Secondary 35A25, 35R30.

Keywords: Huxley equation; Imperialist competitive algorithm; Partial differential equations; Meta-heuristic algorithms; Genetic algorithm.

1 Introduction

The Huxley equation, classified as a nonlinear partial differential equation (NPDE), has the capacity to model a diverse range of phenomena, including biological population dynamics [9] and the propagation of nerves [30]. Its significance lies in its ability to capture the intricate dynamics and interrelationships within these systems, providing valuable insights into their behavior and characteristics. The choice of the Huxley equation as our focus has a dual rationale. First, as previously mentioned, this equation finds numerous practical applications. Second, the selection of this equation, being an NPDE, serves to demonstrate the capability of our proposed algorithm in handling a wide range of inverse forms of NPDEs. Within the realm of partial differential equations (PDEs), an equation is considered "inverse" when one or more of the initial or boundary conditions are missed. In solving inverse forms of PDEs, we utilize data collected from sensors, often referred to as "over-specified conditions," to compensate for the missing condition(s). The primary challenge in solving inverse forms of PDEs lies in the identification

of the missing condition(s). In this paper, we specifically address a scenario in which one of the boundary conditions, denoted as $q(t)$, is missing.

The main purpose of this paper is to present an improved imperialistic competitive algorithm (IICA) for determining the missing boundary condition, $q(t)$. The reason why the imperialistic competitive algorithm (ICA) was chosen is that this algorithm has demonstrated a remarkable capability for solving equations [2, 19, 12]. In this paper, improvements are made to the original ICA to enhance its suitability for estimating a function. Since the ICA is a meta-heuristic algorithm, the results of the IICA and the original ICA are compared to a genetic algorithm (GA), which is one of the well-known and leading algorithms in the realm of meta-heuristic algorithms. In recent years, meta-heuristic algorithms have shown a significant capability in solving inverse forms of linear and nonlinear PDEs and other challenging problems. Also, the convergence of these algorithms has been studied well [6, 3, 21, 20, 29, 17, 22, 28, 10, 24]. Therefore, investigating the capabilities of the new methods and improved algorithms might yield valuable advancements in this field.

The rest of this paper is organized as follows. To calculate the fitness function (cost function) of the algorithms, we need to solve the direct form of the Huxley equation. In section 2, we present the main form of the Huxley equation and the discretization of the Huxley equation using the Crank–Nicolson method [26], which is a finite difference method. This discretization is employed to solve the direct form of the Huxley equation and evaluate the fitness value of a candidate solution accordingly. In section 3, we present the improved ICA in detail, explaining how our improvement makes the original ICA a more powerful method for estimating a function. Since the GA has been widely used and is famous, section 4 provides a brief description of a real-valued GA. In section 5, we present the numerical experiments of the IICA, ICA, and GA and discuss the results. Finally, in section 6, we conclude the paper and state its main outcomes.

2 The Huxley equation and its discretization

In this section, we first present the formulation of the Huxley equation in subsection 2.1, followed by the discretization of this equation in subsection 2.2, which is utilized to construct the fitness function.

2.1 The Huxley equation

The general form of the Huxley equation is as follows:

$$\frac{\partial U}{\partial t} = \frac{\partial^2 U}{\partial x^2} + U(1 - U^\delta)(U^\delta - \gamma), \tag{1a}$$

with initial and boundary conditions:

$$U(x, 0) = f(x), \tag{1b}$$

$$U(0, t) = p(t), \tag{1c}$$

$$U(1, t) = q(t), \tag{1d}$$

where δ is a positive integer, and $\gamma \in (0, 1)$. In this paper, we consider $0 \leq t \leq 1$ and $0 \leq x \leq 1$.

Additionally, the over-specified condition (data coming from a sensor) is as follows:

$$U(a, t) = s(t_j), \quad t_j = k \times j, \quad j = 1, 2, 3, \ldots, M. \tag{2}$$

Here, a represents the location of the sensor, k is the discretization step size of time, $s(t_j)$ is the value measured by the sensor at time t_j, and $x = a$.

2.2 Discretization of the Huxley equation

In this study, we use an implicit finite difference approximation (Crank–Nicolson) method, to discretize (1). As a result, we obtain the following discretized representation for the Huxley equation:

$$
\begin{aligned}
&-r_1U_{i-1,j+1}+(2+2r_1)U_{i,j+1}-r_1U_{i+1,j+1}\\
&\quad=r_1U_{i-1,j}+(2-r_2-2r_1)U_{i,j}+r_1U_{i+1,j}+2r_2{U_{i,j}}^2-r_2{U_{i,j}}^3,\\
&\qquad\qquad i=1,\ldots,N-1,\ j=0,\ldots,N-1, \qquad (3a)\\
&U_{i,0}=f(ih),\ j=0,\ i=1,\ldots,N-1, \qquad (3b)\\
&U_{0,j}=p(jk),\ i=0,\ j=0,1,\ldots,N-1, \qquad (3c)\\
&U_{N,j}=q(jk),\ Nh=1,\ j=0,1,\ldots,N-1, \qquad (3d)
\end{aligned}
$$

where $x=ih,\quad i=0,1,\ldots,N-1$ and h is the step size of the discretization of x, $t=jk,\quad j=0,1,\ldots,N-1$, and k is the step size of the discretization of t, $r_1=k/h^2$ and $r_2=2k$.

Using (3), we obtain the following linear algebraic system of equations:

$$
\begin{pmatrix}
2+2r_1 & -r_1 & 0 & 0 & 0 & 0 & 0\\
-r_1 & 2+2r_1 & -r_1 & 0 & 0 & 0 & 0\\
\vdots & \vdots & \vdots & \vdots & \vdots & \vdots & \vdots\\
0 & 0 & 0 & 0 & -r_1 & 2+2r_1 & -r_1\\
0 & 0 & 0 & 0 & 0 & -r_1 & 2+2r_1
\end{pmatrix}
\begin{pmatrix}
U_{1,j+1}\\ U_{2,j+1}\\ \vdots\\ U_{N-2,j+1}\\ U_{N-1,j+1}
\end{pmatrix}
$$

$$
=\begin{pmatrix}
2-r_2-2r_1 & r_1 & 0 & 0 & 0 & 0 & 0\\
r_1 & 2-r_2-2r_1 & r_1 & 0 & 0 & 0 & 0\\
\vdots & \vdots & \vdots & \vdots & \vdots & \vdots & \vdots\\
0 & 0 & 0 & 0 & r_1 & 2-r_2-2r_1 & r_1\\
0 & 0 & 0 & 0 & 0 & r_1 & 2-r_2-2r_1
\end{pmatrix}
\begin{pmatrix}
U_{1,j}\\ U_{2,j}\\ \vdots\\ U_{N-2,j}\\ U_{N-1,j}
\end{pmatrix}
$$

$$
+r_1\begin{pmatrix}
U_{0,j}+U_{0,j+1}\\ 0\\ \vdots\\ 0\\ U_{N,j}+U_{N,j+1}
\end{pmatrix}
+\begin{pmatrix}
2r_2{U_{1,j}}^2-r_2{U_{1,j}}^3\\
2r_2{U_{2,j}}^2-r_2{U_{2,j}}^3\\
\vdots\\
2r_2{U_{N-2,j}}^2-r_2{U_{N-2,j}}^3\\
2r_2{U_{N-1,j}}^2-r_2{U_{N-1,j}}^3
\end{pmatrix}
$$

where $x=ih,\quad i=0,1,\ldots,N-1$ and h is the step size of the discretization of x, $t=jk,\quad j=0,1,\ldots,N-1$, and k is the step size of the discretization of t, $r_1=k/h^2$ and $r_2=2k$.

In this study, the IICA, ICA, and GA are used to approximate the unknown function $q(t)$ in (1). Specifically, $q(t)$ is treated as a candidate solu-

tion represented as a real-valued vector (coefficients of a polynomial), which is then input into the fitness function for assessment. To evaluate the fitness of a candidate solution, system (3) is solved, and the numerical values of $U(x_i, t_j)$ are computed. Subsequently, the vector $\hat{s}(t_j) = U(x = a, t_j)$ is compared to the vector $s(t_j)$ as described in (2). To perform this comparison, the mean squared error is calculated. Smaller values of the mean squared error between $\hat{s}(t_j)$ and $s(t_j)$ indicate a better approximation of the unknown function $q(t)$. The pseudo-code of the fitness function in this study is as follows:

Algorithm 1: Pseudo-code of the fitness function

Data: Input values: Coefficients of a polynomial approximating $q(t)$

Result: Fitness value of the input values

Function `Fitness(`*Coefficients of a polynomial approximating* $q(t)$`):`

Calculate $U(x_i, t_j)$ using System (3)

return $\frac{1}{\sum_{j=1}^{m}(U(a,t_j)-s_{t_j})^2}$;

In Algorithm 1, as the approximation of $q(t)$ converges towards the exact $q(t)$, the denominator decreases. Consequently, the value of the fitness function increases.

3 Improved imperialistic competitive algorithm (IICA)

The ICA is a powerful meta-heuristic algorithm that has been successfully applied to a wide range of problems in science and engineering. Additionally, in recent years, several authors have attempted to present improved versions of the algorithm to enhance its effectiveness for optimization problems [7, 8, 35, 1, 34, 31, 33, 25, 32, 18, 13, 23]. This research paper represents the first attempt to enhance the original ICA, transforming it into a powerful method for function estimation in differential equations. In this section, we first present the original ICA briefly in Subsection 3.1. Then, in Subsection 3.2, we present the improved version of the ICA, which is a powerful method for function approximation in solving differential equations.

3.1 Original ICA

The ICA is a robust and versatile meta-heuristic optimization technique that has gained great attention in the fields of science and engineering. This algorithm was developed to tackle a wide range of complex problems. The ICA was inspired by the dynamics of imperialistic competition in societies. This algorithm emulates the concept of countries competing for dominance and resources, where each candidate solution to an optimization problem is treated as an independent "country." These countries try to spread their dominance through various interactions, such as assimilation and colonization [4]. The main steps of the original ICA are as follows:

1. initialization: First, initialize a population of candidate solutions (countries) randomly. Each country is considered as follows:

$$\text{country}_i = \{a_1, a_2, \ldots, a_m\}.$$

Here, country_i is the ith candidate solution with size m. In fact, $a_j, 1 \leq j \leq m$ indicate the coefficients of a polynomial as follows:

$$y(x) = a_m x^{m-1} + a_{m-1} x^{m-2} + \cdots + a_2 x^1 + a1.$$

Next, evaluate the fitness of each candidate solution. Then, select the top $N_{impires}$ countries as the imperialists. Finally, form the empires by dividing the remaining countries (colonies) among the imperialists in proportion to the fitness of the imperialists.

2. Assimilation: In every empire, the colonies move towards their imperialist using a randomly adjusted vector, which is scaled by a proximity factor. This stage emulates the impact of imperialism on the colonies and attempts to improve the fitness of each colony. The assimilation operator works to bring the colonies of an empire closer to the characteristics of the imperialist state within the search space. It like guides the colonies to adopt the traits of the imperialist, somewhat similar to how cultural assimilation happens where colonies start to resemble the imperialist in certain ways.

3. Exchanging positions of the imperialist and a colony: A colony may find a better position than the imperialist as it moves closer to it. In this situation, the imperialist and the colony swap their positions and the algorithm continues. The exchange process involves the transfer of colonies between imperialists based on the fitness value. This exchange aims to improve the overall quality of both imperialists and colonies.

4. Imperialistic competition:

   At first, the total power of an empire is evaluated using the following equation:

$$\text{T.C. }_n = \text{Fitness}\left(\text{imperialist}_n\right) + \xi \text{mean}\left\{\text{Fitness}\left(\text{colonies of empire}_n\right)\right\} \tag{4}$$

   Here $T.C_n$ is the total fitness of the nth empire, and ξ is a positive number that is considered to be less than 1. The strength of an empire mostly depends on how strong its imperialist country is. However, the colonies within that empire also have some influence, although it is not very significant. The author suggests that by adjusting a factor called ξ, we can change how much the colonies contribute to the empire's overall power. They recommend setting ξ at 0.1 for a balanced approach.

   Then, the imperialistic competition begins. All empires attempt to acquire colonies belonging to other empires and take control of them. This competitive, imperialistic process results in the gradual weakening of less powerful empires and the strengthening of more dominant ones. This competition is modeled by assigning one of the weakest colonies from the weakest empires to a dominant empire. The dominant empire is the one that wins the competition, which is based on the $T.C$ values of the empires.

5. Eliminating the powerless empires: An empire that has no power will be destroyed in the imperialistic competition and its colonies will be distributed among other empires. Different criteria can be used to model the destruction mechanism and determine when an empire has no power. In the original ICA, an empire collapses when it loses all of its colonies.

Steps 3.1 through 3.1 are repeated until the stop criterion is met. These steps are illustrated in Figure 1. Furthermore, Figure 2 illustrates the flowchart of the original ICA.

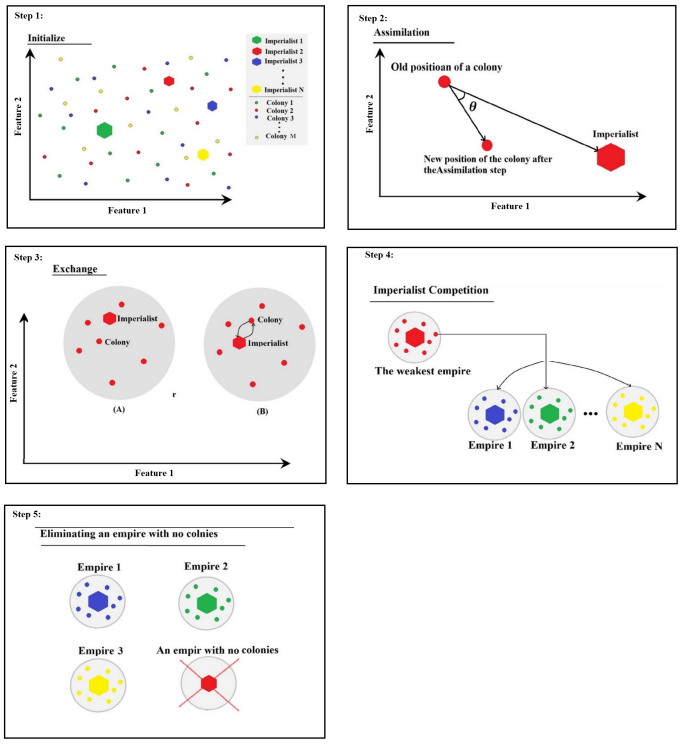


Figure 1: The main steps of the original ICA.

3.2 Improved imperialistic competitive algorithm (IICA)

Our improvements to the original ICA are as follows:

1. Smoothness: This improvement arises from the fact that in many methods, the unknown function that needs to be found is assumed to be smooth [14]. In our paper, we assume that the unknown function $q(t)$ is a polynomial of degree n. The IICA, ICA, and GA attempt to approx-

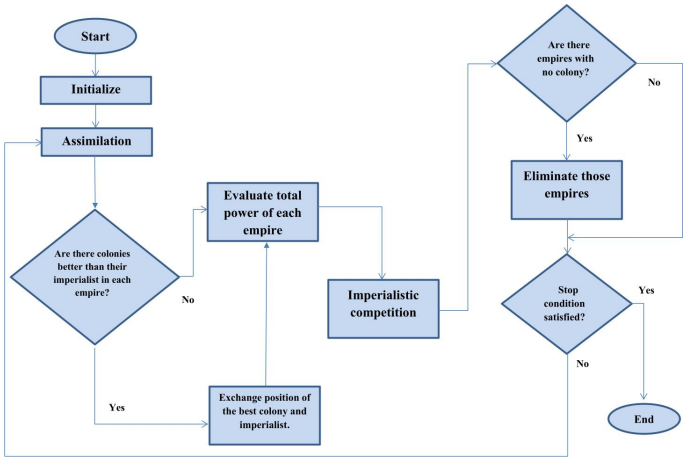


Figure 2: The flowchart of the original ICA.

imate the coefficients of the unknown function $q(t)$ such that the fitness value of the approximated $q(t)$ is maximized. During the middle iterations of the algorithm execution, the values of the coefficients of $q(t)$ may vary too much between two successive values, which might cause a big jump between $q(t_k)$ and $q(t_{k+1})$. Consequently, we introduce a step in which a procedure is executed to make the successive values of the unknown vector $q(t_j)$ much smoother. To apply this smoothness procedure to any elements of the coefficients vector $q(t)$, the following steps are performed:

- Calculate the mean of c_k and c_{k+2} as follows:

$$m = \frac{c_k + c_{k+2}}{2}.$$

- Then, move the value of c_{k+1} toward m as follows:

$$c_{k+1} = \frac{c_{k+1} + \alpha \times m}{1 + \alpha}, \tag{5}$$

where c_j is a coefficient of the candidate solution approximating the unknown $q(t)$ and $\alpha \in \mathbb{R}$ is a hyper-parameter, which should be tuned efficiently. Note that, at the beginning of the algorithm, we have the value $q(t_0) = q(0) = f(0)$ because the initial condition is known. Additionally, for the last element of the candidate solution, we use the

preceding value of m. The effect of the smoothness step is demonstrated through an example in Appendix A.

2. Correction: In many applications of meta-heuristic algorithms, there exists a search space for the values of the unknown vector, typically within the range $[LB, UB]$, where LB and UB stand for lower bound and upper bound respectively. In our algorithm, following the smoothness and assimilation steps, some values may surpass the interval $[LB, UB]$. Consequently, it becomes essential to reposition these values within the valid interval. This procedure is executed as follows:

$$c_k = c_k - \beta,$$

where β represents the amount by which c_j has exceeded $[LB, UB]$. For example, let us consider $UB = 20$, and suppose that c_j has reached a value of 25 after the smoothness and assimilation steps. Following the correction step, c_j will be adjusted to 15.

The whole procedure of the IICA is as follows:

1. Initialization: Generate randomly an initial population and create empires.

2. Assimilation: In every empire, the colonies move towards their imperialism using a randomly adjusted vector.

3. Smoothness: The smoothness procedure is applied to the candidate solutions.

4. Correction: The correction step is applied to the candidate solutions to keep them inside the valid interval.

5. Evaluation: Evaluate the fitness of the candidate solutions.

6. Exchanging position: The imperialist and the colony swap their positions if the colony is better than the imperialist.

7. Evaluation of empires' total power: The total power of the empires is evaluated.

8. Imperialistic competition: The imperialistic competition is done.

9. Collapse: Empires without colonies collapse.

10. Repeat Step 3.2 to Step 3.2, until the predefined number of iterations is not satisfied.

In fact, these two additional steps apply regularization to the coefficients of a polynomial that approximates the unknown $q(t)$, similar to the regularization that is done in the machine learning realm [14]. Figure 3 illustrates the steps of the improved ICA. We repeat steps 2 through 7 until the stop criterion is met. Furthermore, Figure 4 illustrates the flowchart of the improved ICA.

Table 1 presents the parameters of a real-valued GA used in this paper.

Table 1: Parameters of the IICA

Representation	Real valued vectors
Length of countries	Degree of a polynomial
Range of entries	$[-1, 1]$
Initialization	Random
Number of population	50 and 200
Number of empires	5 and 20
α	0.1
Collapse criteria	Having no colony
Termination condition	Number of generation

4 GA for the solution to inverse forms of Huxley equation

The GA, which was mainly developed by Holland [15], is a search method based on the Darwinian principles of biological evolution. This algorithm has been successfully used for various optimization problems. The GA is a stochastic optimization method that uses a population of chromosomes, each representing a possible solution. By applying a genetic operator, each chromosome improves gradually and becomes the basis for the next generation.

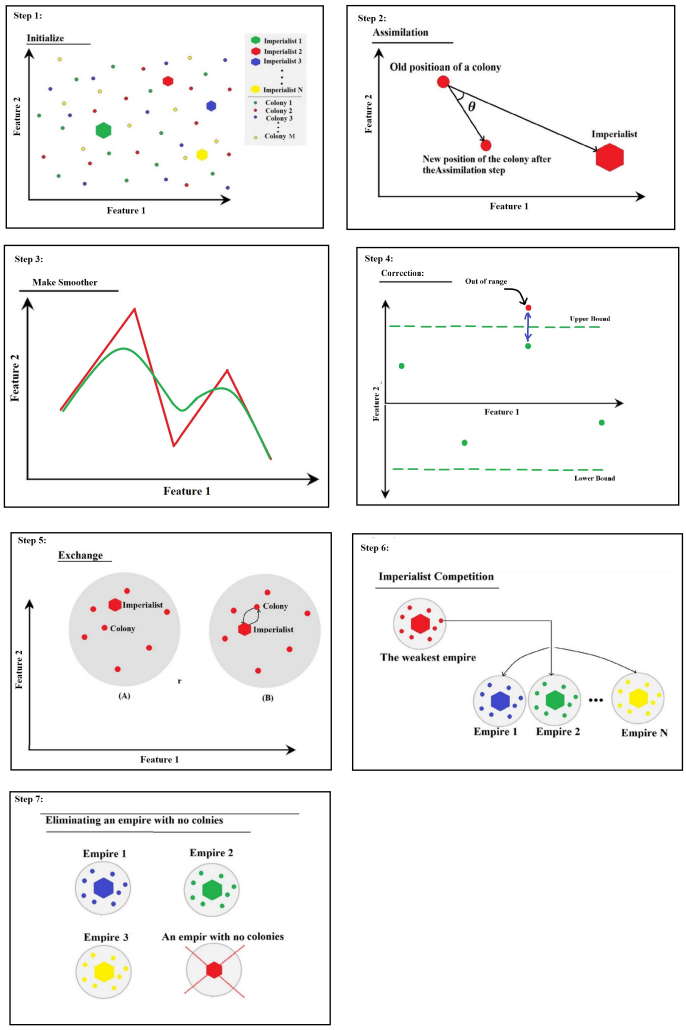


Figure 3: The steps of the improved ICA.

The process continues until the desired number of generations is reached or the predefined fitness value is achieved.

The procedure of a GA is as follows:

1. Generate at random an initial population of chromosomes.
2. Evaluate the fitness of each chromosome in the population.
3. Select some chromosomes as parents.

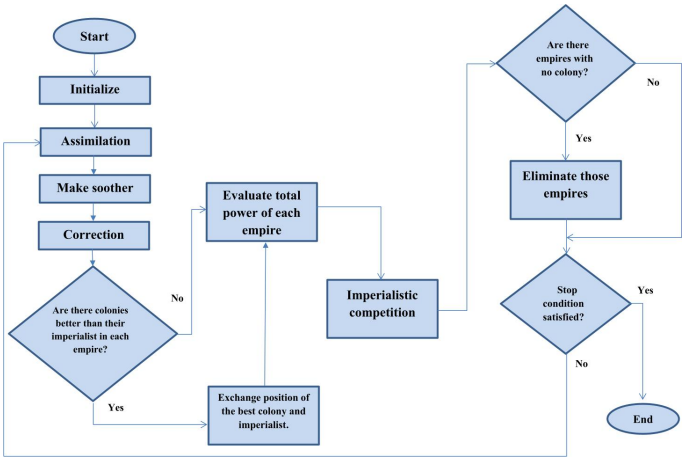


Figure 4: The flowchart of the improved ICA.

4. Apply recombination operation on parents.
5. Apply mutation operation on offspring.
6. Evaluate the fitness of offspring.
7. Update the population.
8. Repeat Step 4 to Step 4, until the predefined number of iterations is not satisfied.

Figure 5 presents the flowchart of the GA used in this paper.

Table 2 presents the parameters of a real-valued GA used in this paper.

To solve an inverse form of the Huxley equation using the GA presented in this section, we consider each candidate solution (chromosome) a real-valued vector as follows:

$$\text{Chromosome}_i = \{a_1, a_2, \ldots, a_m\},$$

where Chromosome_i is the ith candidate solution with size m in the population. Entries $a_j, 1 \leq j \leq m$ indicate the coefficients of a polynomial as follows:

$$y(x) = a_m x^{m-1} + a_{m-1} x^{m-2} + \cdots + a_2 x^1 + a1. \tag{6}$$

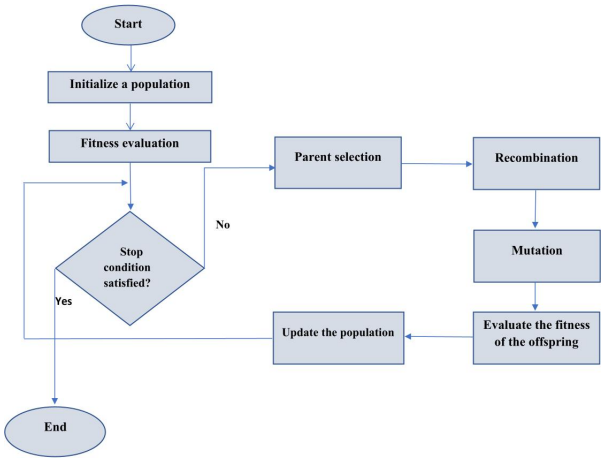


Figure 5: The flowchart of the GA used in this paper.

Table 2: Parameters of the GA

Representation	Real valued vectors
Length of chromosomes	Degree of a polynomial
Recombination	One point crossover
Recombination probability	100%
Mutation	Adding a random value
Mutation probability	$1/n$
Parent selection	Roulette wheel
Survivor selection	Replace the worst
Number of offspring	1
Initialization	Random
Termination condition	Number of generation

Each candidate solution such as (6) is considered as the missing condition $q(t)$ of the Huxley equation.

In our GA, we initially generate a random population of a specific size and evaluate their fitness. Then, as indicated in Figure 5, the main loop iterates the number of iterations times. In each iteration, two candidate solutions are selected as parents based on the roulette wheel selection method [11]. The recombination step, using one-point crossover, is applied to the selected

parents, and a new offspring is created. Subsequently, the mutation step is applied to the offspring by adding a small random value to a randomly chosen entry. The population is then updated by replacing the worst individual with the offspring.

5 Numerical examples

An inverse form of the Huxley equation, when $0 < x < 1, \quad 0 < t < t_M$, is as follows [5]:

$$U_t(x,t) = U_{xx}(x,t) + U(x,t)(1-U(x,t))(U(x,t)-1), \quad 0 < x < 1, \quad 0 < t < T, \tag{7a}$$

$$U(x,0) = \frac{1}{2} + \frac{1}{2}\tanh(\frac{1}{2\sqrt{2}}x), \tag{7b}$$

$$U(0,t) = \frac{1}{2} + \frac{1}{2}\tanh(\frac{-t}{4}), \tag{7c}$$

$$U(1,t) = q(t), \tag{7d}$$

and the over-specified condition

$$s(t_j) = U(0.5, t_j), \quad t_j = k \times j, \quad j = 1, 2, \ldots, N, \tag{7e}$$

where k is the discretization step size time (t), and the function $q(t)$ is missing. In this equation, the exact $U(x,t)$ and $q(t)$ are $\frac{1}{2} + \frac{1}{2}tanh(\frac{1}{2\sqrt{2}}(x - \frac{t}{\sqrt{2}}))$ and $\frac{1}{2} + \frac{1}{2}tanh(\frac{1}{2\sqrt{2}}(1 - \frac{t}{\sqrt{2}}))$, respectively.

The primary goal of this paper is to solve an inverse form of the Huxley equation by estimating its missing condition, denoted as $q(t)$. To assess the accuracy of the estimated function $\hat{q}(t)$, we employ the mean absolute error (MAE) criterion. We calculated the MAE value over the interval $[0, 1]$ with a step size of $h = 0.01$ for each algorithm to demonstrate its precision. Table 3 presents the MAE values comparing the estimated $\hat{q}(t)$ and the exact $q(t)$ for the implementations of IICA, ICA, and GA. The population size is set at 50 for all algorithms, α (the parameter for the smoothness step in Eq. (5)) is fixed at 0.1, and the initial number of empires for both IICA and ICA is 5. It is important to note that, as meta-heuristic algorithms are part of the

stochastic algorithm class, we ran the algorithms three times and reported the best result out of all the outcomes in this paper.

Table 3: The MAE in $[0,1]$ with the step size 0.01 between the exact $q(t)$ and the estimated function $\hat{q}(t)$ found by the IICA, ICA, and GA algorithms. These calculations were based on a population size of 50, an α value of 0.1, and 5 empires

$Num.of.Iter.$	MAE of the GA	MAE of the ICA	MAE of the IICA
100	0.05801	0.01227	0.00636
150	0.02166	0.01171	0.00444
200	0.01402	0.00608	0.00367
250	0.03842	0.00656	0.00417
300	0.04420	0.00392	0.00301
350	0.02635	0.00600	0.00221
400	0.01275	0.00351	0.00175
450	0.00825	0.00301	0.00186
500	0.00337	0.00255	0.00172

Figure 6 displays both the exact function $q(t)$ and the numerically approximated $\hat{q}(t)$ as determined by the IICA, which used 500 iterations, a population size of 50, an α value of 0.1, and 5 empires. The figure shows that the MAE across the interval $[0,1]$, with a step size of $h = 0.01$, is 0.00172. Furthermore, in this section, we will expand our examination to see how increasing the population size and the initial number of empires affects the performance of the IICA.

Figure 7 shows the discrepancy between the exact function $q(t)$ and its numerical approximation $\hat{q}(t)$, as derived by the IICA. This was achieved after 500 iterations, with a population size of 50, an α value of 0.1, and 5 initial empires.

We present the convergence patterns of the IICA, ICA, and GA as depicted in Figure 8, which is derived from Table 3. The figure clearly shows that the IICA converges more rapidly than the ICA, and the ICA, in turn, converges quicker than the GA. The figure also indicates that while the GAs performance varies within the interval, the ICA and IICA demonstrate a consistent improvement as the number of iterations increases.

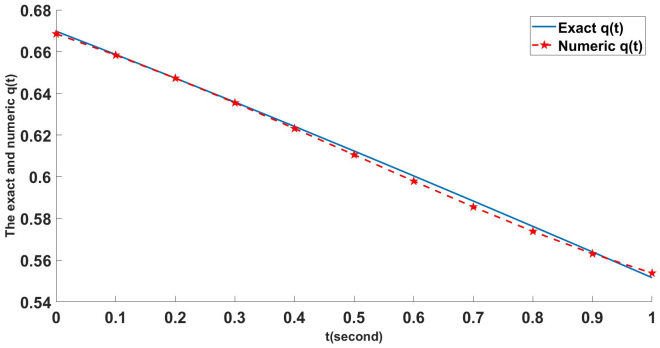


Figure 6: The exact $q(t)$ and the approximated (numeric) $\hat{q}(t)$ found by the IICA with iterations 500, population size 50, $\alpha = 0.1$, and the number of empires 5.

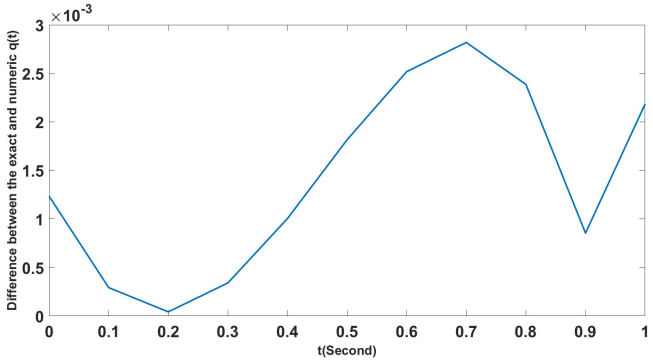


Figure 7: The difference between the exact $q(t)$ and approximated (numeric) $\hat{q}(t)$ found by the IICA with iterations 500, population size 50, $\alpha = 0.1$, and the number of empires 5.

Table 4 displays the MAE in $[0, 1]$ with the step size 0.01 comparisons for the exact function $q(t)$ against the approximated $\hat{q}(t)$ found by the IICA, ICA, and GA algorithms. These results were obtained with a population size of 200, an α value of 0.1, and an initial empire count of 20 for both the IICA and ICA. The table clearly indicates that the precision of the IICA and ICA improves with larger populations and more empires, whereas these changes do not significantly impact the performance of the GA.

Figure 9 showcases the exact $q(t)$ alongside the numerically approximated $\hat{q}(t)$ found by the IICA through 500 iterations, a population size of 200, an α of 0.1, and 20 empires. The figure indicates an MAE within the interval

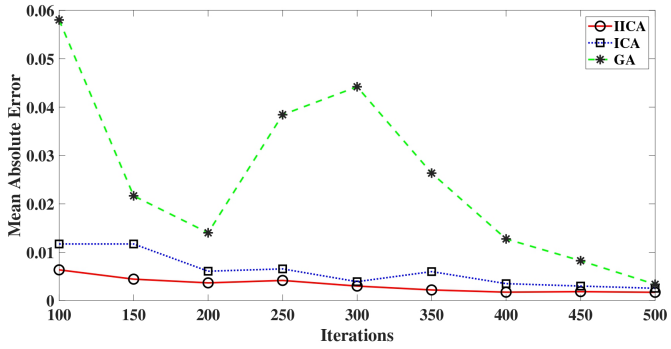


Figure 8: The convergence of the IICA, ICA, and GA extracted from Table 3.

Table 4: The MAE between the approximated $\hat{q}(t)$ and the exact $q(t)$ by the implementation of the IICA, ICA and GA for population size 200, $\alpha = 0.1$, and the number of empires 20

Num.of.Iter.	MAE of the GA	MAE of the ICA	MAE of the IICA
100	0.05280	0.00511	0.00302
150	0.06424	0.00320	0.00282
200	0.02392	0.00309	0.00139
250	0.03457	0.00320	0.00107
300	0.02326	0.00231	0.00104
350	0.03192	0.00203	0.00149
400	0.03222	0.00209	0.00094
450	0.03447	0.00147	0.00110
500	0.02410	0.00156	0.00083

$[0, 1]$, at a step size of 0.01, of 0.00083, which signifies a precise solution in the domain of inverse form of NPDEs.

Figure 10 presents the comparison between the exact $q(t)$ and its numerical approximation $\hat{q}(t)$ as produced by the IICA, following 500 iterations, with a population size of 200, an α of 0.1, and 20 empires. The figure demonstrates that the absolute error is generally on the order of $O(10^{-4})$ across most of the interval. Between approximately 0.2 and 0.5, the error increases to the order of $O(10^{-3})$. While there are fluctuations throughout the interval, the MAE consistently remains at the order of $O(10^{-4})$.

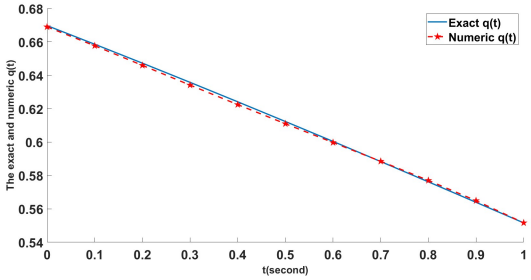


Figure 9: The exact $q(t)$ and approximated (numeric) $\hat{q}(t)$ found by the IICA with iterations 500, population size 200, $\alpha = 0.1$, and the number of empires 20.

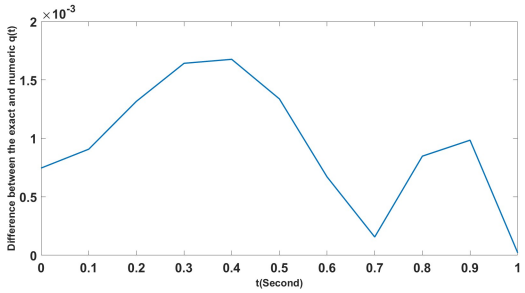


Figure 10: The difference between the exact $q(t)$ and approximated (numeric) $q(t)$ found by the IICA with iterations 500, population size 200, $\alpha = 0.1$, and the number of empires 20.

For illustrative purposes, we present the convergence of the IICA, ICA, and GA extracted from Table 4 in Figure 11. As evident from the figure, the convergence rate of the IICA surpasses that of the ICA and the GA. The error value curve of the IICA consistently lies below those of the original ICA and the GA. The performance of the GA fluctuates between 200 and 500 iterations, while both the ICA and the IICA steadily improve with an increasing number of iterations.

5.1 Discussion

According to our experiments, the best result is achieved when α is set to 0.1 (the parameter for the smoothness step in (5)). Additionally, based on

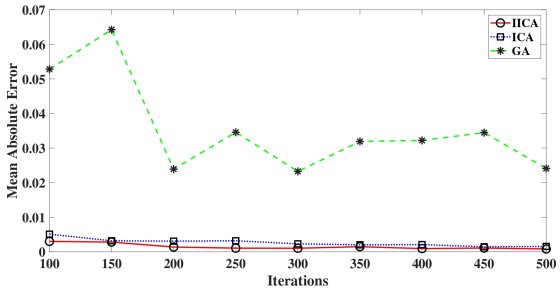


Figure 11: The convergence of the IICA, ICA, and GA extracted from Table 4.

the experiments, when the population size is less than 50, the accuracy of the results is low, and increasing the population size beyond 200 does not considerably improve accuracy. Therefore, we have reported these two values for the population size. Furthermore, according to the results, in general, the performance of the original ICA is better than that of the GA, and the performance of the IICA is better than both of them. Therefore, we focused on the IICA to plot figures and analyze its performance. A good solution using the IICA is obtained when the population size is 200, the size of the initial empires is 20, and the algorithm is iterated 500 times. In this case, the accuracy of the result is on the order of $O(10^{-4})$, which is good accuracy in the realm of the inverse form of NPDEs. Additionally, $[LB, UB]$ is $[-1, 1]$ for all experiments. Due to the nature of the IICA algorithm, the population size has a significant impact on its accuracy. This is the reason why we focused on this parameter. Figures 6 and 9 show the exact $q(t)$ and the numeric $\hat{q}(t)$ found by the IICA for population sizes 50 and 200, respectively. In these figures, it is evident that generally, at the beginning and end of the interval, the accuracy is better than in the middle of the interval. Moreover, for the population size 200 in Figure 9, the figures exactly match, indicating that the proposed improvement has reached a high accuracy. Figures 7 and 10 present the error study found by our proposed algorithm (IICA) for the population sizes 50 and 200, respectively. As can be seen from these figures, the overall accuracy is better for the population size 200 than for the population size 50. Figures 8 and 11 present the convergence of the IICA, the ICA, and the GA with iterations from 100 to 500 for the population sizes 50 and 200,

respectively. It is clear from the figures that our proposed algorithm (IICA) converges better than other algorithms with different population sizes. In fact, the error curve of the IICA is almost always below that of the ICA and the GA. The slope of the convergence curve for the IICA and ICA does not change significantly around the number of iterations of 500. Furthermore, these figures show that the original algorithm converges faster than the GA for a function approximation problem when the unknown function is considered a polynomial. As can be seen from Tables 3 and 4, increasing the size of the population does not help the GA reach a better solution, and this algorithm does not converge to a high accuracy when the number of iterations is increased to the population size of 200. On the other hand, the convergence of the IICA and ICA improves when the size of the population is increased from 50 to 200.

6 Conclusion

The improvements presented in this paper make the original ICA a much more powerful method for solving differential equations and function approximation. Since, in general, in real-world applications, the unknown functions are smooth, the smoothness procedure introduced in this paper helps the algorithm reach a high accuracy in function approximation tasks faster. Furthermore, this paper presents the application of meta-heuristic algorithms in solving inverse forms of NPDEs, which are categorized as ill-posed and challenging problems. The numerical results demonstrate that the IICA can effectively and proficiently solve inverse forms of nonlinear PDEs. Given the prevalence of inverse problems in applied fields, this method holds the potential for solving real-world challenges, which could lead to reduced execution times and enhanced accuracy. In this paper, we considered polynomials as basis functions to approximate the unknown function. In future research, another set of functions, such as orthogonal functions (e.g., Jacobi polynomials, Legendre polynomials, Chebyshev polynomials, and Gegenbauer polynomials), could be considered as the basis functions. Additionally, future research may involve the parallel implementation of the IICA. Moreover, other meta-heuristic algorithms could be employed to tackle this class of problems, al-

lowing for comparative analyses of their outcomes in relation to the results obtained in this study.

Appendix A

Figure 12 illustrates the impact of the smoothness procedure on the vector $x = [3, -7, 15, -6, -17, 18, -19, 9, 13, -4]$ when $\alpha = 0.5$. After applying the smoothness procedure, the resulting vector is denoted as $x' = [0.94, -2, 8.7, -5.4, -9.2, 7.3, -9.95, 6.5, 9.1, -2.2]$.

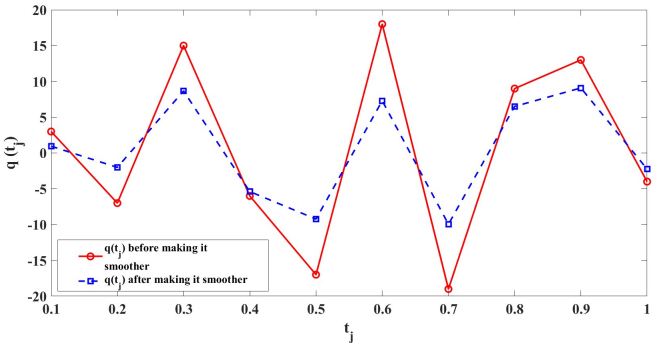


Figure 12: The impact of the smoothness procedure on a vector.

References

[1] Abbasi Molai, A. and Dana Mazraeh, H., *A modified imperialist competitive algorithm for solving nonlinear programming problems subject to mixed fuzzy relation equations*, Int. J. Nonlinear Anal. Appl. 14(3) (2023), 19–32.

[2] Abdollahi, M., Isazadeh, A. and Abdollahi, D. *Imperialist competitive algorithm for solving systems of nonlinear equations*, Comput. Math. Appl. 65(12) (2013), 1894–1908.

[3] Aliyari Boroujeni, A., Pourgholi, R. and Tabasi, S.H. *A new improved teaching–learning-based optimization (ITLBO) algorithm for solv-*

ing nonlinear inverse partial differential equation problems, Comput. Appl. Math. 42(2) (2023), 99.

[4] Atashpaz-Gargari, E. and Lucas, C. *Imperialist competitive algorithm: an algorithm for optimization inspired by imperialistic competition*, In 2007 IEEE, congress on evolutionary computation (pp. 4661-4667). IEEE, 2007.

[5] Babolian, E. and Saeidian, J. *Analytic approximate solutions to Burgers, Fisher, Huxley equations and two combined forms of these equations*, Commun. Nonlinear Sci. Numer. Simul. 14(5) (2009), 1984–1992.

[6] Barrios, D., Malumbres, L. and Rios, J. *Convergence conditions of genetic algorithms*, Int. J. Comput. Math. 68(3-4) (1998), 231–241.

[7] Bilel, N., Mohamed, N., Zouhaier, A. and Lotfi, R. *An improved imperialist competitive algorithm for multi-objective optimization*, Eng. Optim. 48(11) (2016), 1823–1844.

[8] Cai, J., Yang, H., Lai, T. and Xu, K. *A new approach for optimal chiller loading using an improved imperialist competitive algorithm*, Energy and Build. 284 (2023), 112835.

[9] Dai, Y.D., Zhang, H.L., Yuan, Y.P., et al. *The traveling solution of Huxley equation (in Chinese)*, Heilongjiang Sci. Technol. Inf. 11, 57 (2016).

[10] Dana Mazraeh, H., Kalantari, M., Tabasi, S.H., Afzal Aghaei, A., Kalantari, Z. and Fahimi, F. *Solving Fredholm integral equations of the second kind using an improved cuckoo optimization algorithm*, Glob. Anal. Discret. Math. 7(1) (2022), 33–52.

[11] Eiben, A.E. and Smith, J.E. *Introduction to evolutionary computing*, Springer, 2015.

[12] Fathy, A. and Rezk, H. *Parameter estimation of photovoltaic system using imperialist competitive algorithm*, Renew. Energy, 111 (2017), 307–320.

[13] Ghasemi, M., Ghavidel, S., Ghanbarian, M.M., Massrur, H.R. and Gharibzadeh, M. *Application of imperialist competitive algorithm with its modified techniques for multi-objective optimal power flow problem: a comparative study*, Inf. Sci. 281 (2014), 225–247.

[14] Girosi, F., Jones, M. and Poggio, T. *Regularization theory and neural networks architectures*, Neural Comput. 7(2) (1995), 219–269.

[15] Holland, J.H. *Adaptation in natural and artificial systems*, year 1975, publisher: University of Michigan Press.

[16] Loyinmi, A.C. and Akinfe, T.K. *An algorithm for solving the Burgers–Huxley equation using the Elzaki transform*, SN Appl. Sci. 2(1) (2020), 7.

[17] Mazraeh, H.D. and Pourgholi, R. *An efficient hybrid algorithm based on genetic algorithm (GA) and Nelder–Mead (NM) for solving nonlinear inverse parabolic problems*, Iranian Journal of Numerical Analysis and Optimization 8(2) (2018), 119–140.

[18] Molla-Alizadeh-Zavardehi, S., Tavakkoli-Moghaddam, R. and Lotfi, F.H. *A modified imperialist competitive algorithm for scheduling single batch-processing machine with fuzzy due date*, The International Journal of Advanced Manufacturing Technology 85 (2016), 2439–2458.

[19] Nemati, K., Shamsuddin, S.M. and Darus, M. *An optimization technique based on imperialist competition algorithm to measurement of error for solving initial and boundary value problems*, Measurement 48 (2014), 96–108.

[20] Pourgholi, R., Dana, H. and Tabasi, S.H. *Solving an inverse heat conduction problem using genetic algorithm: sequential and multi-core parallelization approach*, Appl. Math. Model. 38(7-8) (2014), 1948–1958.

[21] Rarità, L. *A genetic algorithm to optimize dynamics of supply chains*, In Optimization in Artificial Intelligence and Data Sciences: ODS, First Hybrid Conference, Rome, Italy, September 14-17, 2021 (pp. 107–115). Cham: Springer International Publishing, 2022.

[22] Rarità, L., Stamova, I. and Tomasiello, S. *Numerical schemes and genetic algorithms for the optimal control of a continuous model of supply chains*, Appl. Math. Comput. 388 (2021), 125464.

[23] Razzaghpour, M. and Rusu, A. *Analog circuit optimization via a modified Imperialist Competitive Algorithm*, In 2011 IEEE International Symposium of Circuits and Systems (ISCAS) (pp. 2273–2276). IEEE, 2011.

[24] Rooholamini, F., Afzal Aghaei, A., Hasheminejad, S.M.H., Azmi, R. and Soltani, S. *Developing Chimp Optimization Algorithm for Function Estimation Tasks*, Computational Mathematics and Computer Modeling with Applications (CMCMA) (2023), 34–44.

[25] Sharifi, M. and Mojallali, H. *Multi-objective modified imperialist competitive algorithm for brushless DC motor optimization*, IETE J. Res. 65(1) (2019), 96–103.

[26] Smith, G.D. *Numerical Solution of Partial Differential Equations: Finite Difference Methods*, Oxford Applied Mathematics and Computing Science Series, Third Edition, year 1986, publisher: Oxford University Press.

[27] Tang, Y. and Zhou, F. *An improved imperialist competition algorithm with adaptive differential mutation assimilation strategy for function optimization*, Expert Syst. Appl. 211 (2023), 118686.

[28] Tomasiello, S. *Numerical solutions of the Burgers–Huxley equation by the IDQ method*, Int. J. Comput. Math. 87(1) (2010), 129–140.

[29] Tomasiello, S. *DQ based methods: theory and application to engineering and physical sciences*, In Handbook of Research on Computational Science and Engineering: Theory and Practice (pp. 316–346). IGI Global. 2012.

[30] Wang, X.Y. *Nerve propagation and wall in liquid crystals*, Phys. Lett. A, 112(8) (1985), 402–406.

[31] Xu, S., Wang, Y. and Lu, P. *Improved imperialist competitive algorithm with mutation operator for continuous optimization problems*, Neural Comput. Appl. 28 (2017), 1667–1682.

[32] Yousefi, M., Yousefi, M. and Darus, A.N. *A modified imperialist competitive algorithm for constrained optimization of plate-fin heat exchangers*, Proc. Inst. Mech. Eng. A: J. Power Energy 226(8) (2012), 1050–1059.

[33] Zandieh, M., Khatami, A.R. and Rahmati, S.H.A. *Flexible job shop scheduling under condition-based maintenance: improved version of imperialist competitive algorithm*, Appl. Soft Comput. 58, (2017), 449–464.

[34] Zhang, Y., Hu, X. and Wu, C. *Improved imperialist competitive algorithms for rebalancing multi-objective two-sided assembly lines with space and resource constraints*, Int. J. Prod. Res. 58(12) (2020), 3589–3617.

[35] Zhang, Y., Wang, Y. and Peng, C. *Improved imperialist competitive algorithm for constrained optimization*, In 2009 International Forum on Computer Science-Technology and Applications (Vol. 1, pp. 204-207). IEEE, 2009.

Iranian Journal of Numerical Analysis and Optimization
Vol. 14, No. 3, 2024, pp 708–735
https://doi.org/10.22067/ijnao.2024.86892.1394
https://ijnao.um.ac.ir/

Research Article

Stability analysis and optimal strategies for controlling a boycotting behavior of a commercial product

O. Aarabate*, S. Belhdid and O. Balatif

Abstract

In this work, we propose a mathematical model that describes citizens' behavior toward a product, where individuals are generally divided into three main categories: potential consumers, boycotters who abstain from it for

*Corresponding author

Received 16 February 2024; revised 15 April 2024; accepted 25 April 2024

Oumaima Aarabate
Laboratory of Fundamental Mathematics and their Applications, Department of Mathematics, Faculty of Sciences, University of Chouaib Doukkali, El jadida, Morocco. e-mail: oumaimaaarabate@gmail.com

Salaheddine Belhdid
Laboratory of Fundamental Mathematics and their Applications, Department of Mathematics, Faculty of Sciences, University of Chouaib Doukkali, El jadida, Morocco. e-mail: salaheddine.belhdid@gmail.com

Omar Balatif
Laboratory of Fundamental Mathematics and their Applications, Department of Mathematics, Faculty of Sciences, University of Chouaib Doukkali, El jadida, Morocco. e-mail: balatif.maths@gmail.com

How to cite this article
Aarabate, O., Belhdid, S. and Balatif, O., Stability analysis and optimal strategies for controlling a boycotting behavior of a commercial product. *Iran. J. Numer. Anal. Optim.*, 2024; 14(3): 708-735. https://doi.org/10.22067/ijnao.2024.86892.1394

various reasons, and actual consumers. Therefore, our work contributes to understanding product boycott behavior and the factors influencing this phenomenon. Additionally, it proposes optimal strategies to control boycott behavior and limit its spread, thus protecting product marketing and encouraging consumer reuse.

We use mathematical theoretical analysis to study the local and global stability, as well as sensitivity analysis to identify parameters with a high impact on the reproduction number R_0. Subsequently, we formulate an optimal control problem aimed at minimizing the number of boycotters and maximizing consumer participation. Pontryagin's maximum principle is employed to characterize the optimal controls. Finally, numerical simulations conducted using MATLAB confirm our theoretical results, with a specific application to the case of the boycott of Centrale Danone by several Moroccan citizens in April 2018.

AMS subject classifications (2020): Primary 03C45; Secondary 90C31, 35F21.

Keywords: Modeling a boycott behavior; Local and global stability; Sensitivity analysis; Optimal control problem.

1 Introduction

Boycotting a product is a conscious decision to refrain from buying or using a particular product as a way to express disapproval or disagreement with the company responsible for producing or selling it. This is due to various reasons that may be related to the company's practices, policies, or ethical standards. Boycott behavior may include actively encouraging others to join in abstaining from the product.

Ireland is where the word "boycott" first appeared in the late 1800s. It comes from the name of Captain Charles Boycott, an English land agent who worked in Ireland and rose to prominence as a representative of harsh landlordism. Charles Boycott was singled out by the Irish Land League in 1880 during the Irish Land War for his inequitable treatment of tenants. Boycott was effectively isolated and found it difficult to manage his land as a result of the League's encouragement of other farmers and laborers to refuse to work for or conduct commerce with him. Over time, the act of isolating

and refusing to comply has been referred to as a "boycott" and has been employed globally as a means of political protest against unfair behaviors or individuals [15].

Similar to what has been witnessed in many countries of the world, especially North African countries; Morocco has also witnessed the emergence of protest movements and boycotts of many goods and products due to high prices and the decline in purchasing power of citizens. On April 20, 2018, this boycott first appeared on social media. The campaign targeted three main suppliers to Morocco-Centrale Danone (dairy products), Sidi Ali Water brand (bottled water), and African gas stations owned by the Aqua Group (gasoline). As for other Arab countries, such as Egypt, Tunisia, and Jordan, they chose to organize general strikes to demand improved living standards, lower prices, and an end to austerity measures. Morocco has taken a different tack by using the boycott to quietly express its anger at high costs and destitution [7]. Moroccans gathered through this boycott to express their dissatisfaction with high prices and the social and economic conditions in which they live.

This and other similar topics have been the subject of many studies and research projects in the social, economic, and political sciences [7, 11, 19, 9, 4, 1, 21]. However, there are still few mathematical studies and research available on this topic [23, 24, 5, 13, 25].

In this paper, we adopt a compartment modeling approach commonly used in epidemiology to model the spread of product boycott behavior in a population. The compartment model is a widely used approach for explaining the transmission of infectious diseases. In epidemiological models, populations are divided into several categories based on their disease status (i.e., "susceptible," "infected," or "removed"), and the process of infection depends on interaction with infected individuals. Likewise, we consider citizens toward a product to be either potential adopters or boycotters and consumers of the product. It closely resembles the phenomenon of contagious, since boycotters have an important impact on prospective consumers not using the product. It is, therefore, reasonable to model product boycotts using the epidemiological approach. Hence, our work contributes to understanding the behavior of product boycotting and the factors influencing this

phenomenon. Additionally, in this work, we propose optimal strategies for controlling boycotting behavior and limiting its spread, thereby safeguarding product marketing and encouraging consumer reuse.

We propose a mathematical model that describes citizens' behavior towards consuming a specific product, where individuals are divided into three basic categories: Potential consumers of the product, boycotters who abstain from purchasing, using, and consuming the product for various reasons, and attempt to influence other individuals to adopt boycott behavior for the product, and the class of consumers already using the product. By using Routh–Hurwitz criteria and constructing Lyapunov functions, we study the local and global stability of the equilibriums. We examine the sensitivity analysis of the model parameters in order to determine which parameters significantly affect the reproduction number R_0. Using the theoretical results of optimal control theory, we also propose optimal strategies to encourage potential customers to purchase and use a company's product and to persuade and satisfy boycotters of the product.

The structure of this article is as follows. Section 2 is split into two parts: the first part contains the proposed mathematical model, while the second part contains some of the model's fundamental characteristics. Section 3 is also divided into parts. We start by analyzing the local and global stability after some numerical simulations, and finally, we discuss the problem of the parameter's sensitivity. The optimal control problem for the suggested model is presented in Section 4, where we also provide some results regarding the existence of the optimal controls and use Pontryagin's maximum principle to characterize them. Numerical simulations are also provided in this section. In Section 5, the paper is brought to its conclusion.

2 Mathematical model and fundamental characteristics

2.1 Mathematical model

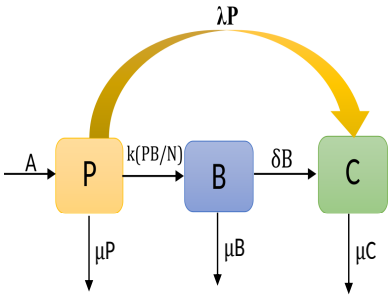


Figure 1: Description of the model.

We consider a mathematical model PBC that captures the behavior of citizens who might use a product, those that boycott the product, and those who consume it. The graphical representation of the proposed model is shown in Figure 1. The total population represented by N is split up into three compartments:

The potential consumers (P) are a category of people that could consume the product. The compartment P is increasing at the rate of A and represents the number of people who can access the product and purchase or consume it. It is decreased when potential consumers become actual consumers at rate λ. It is presumed that potential consumers can also become boycotters of the product at rate k through meaningful interactions with existing boycotters. Finally, the number of potential consumers due to natural death decreases at a rate of μ.

The boycotters (B) who abstain from purchasing, using, and consuming the product for various reasons, and attempt to influence other individuals to adopt boycott behavior for the product. This compartment is increased through effective contact with potential consumers who stop using the product as a result, at a rate k. It is lowered, either by natural death at a rate of μ when boycotters change their opinions about the product and become new actual consumers, at a rate δ.

The actual consumers (C) are those who buy and consume the product. When boycotters change their position and start using the product, the consumer's compartment is increased at a rate of δ. Likewise, it increases at a rate λ when potential consumers are convinced to use the product. Natural death reduces it at the rate μ.

The numbers of people in each of the three classes at time t are represented by the variables $P(t)$, $B(t)$, and $C(t)$, respectively. Time can be measured in years, months, days, or other intervals depending on how frequently survey studies are conducted as needed.

The equation $N(t) = P(t) + B(t) + C(t)$ represents the overall population size at time t. We suppose in this work that N is constant. This model's dynamics are controlled by the nonlinear system of differential equations below.

$$\begin{cases} \dot{P} = A - k\dfrac{PB}{N} - (\mu + \lambda)P, \\ \dot{B} = k\dfrac{PB}{N} - (\delta + \mu)B, \\ \dot{C} = \lambda P + \delta B - \mu C, \end{cases} \tag{1}$$

where $P(0) \geq 0$, $B(0) \geq 0$, and $C(0) \geq 0$ are the given initial states.

2.2 Fundamental characteristics

Since system (1) reflects the population of humans, it is necessary to demonstrate that all of the system's solutions with positive initial data are bounded and will remain positive for all times $t > 0$. The following lemma and theorem will establish this.

2.2.1 The positivity of the model's solutions

Theorem 1. If the initial conditions are positive, that is, $P(0) \geq 0$, $B(0) \geq 0$, and $C(0) \geq 0$, then the solutions $P(t)$, $B(t)$, and $C(t)$ of system (1) are positive for all $t \geq 0$.

Proof. The first equation of system (1) indicates that

$$\frac{dP(t)}{dt} + \left(k\frac{B(t)}{N} + (\lambda + \mu)\right) P(t) \geq 0. \tag{2}$$

Multiplying the inequality (2) by

$$\exp\left[\int_0^t \left(k\frac{B(v)}{N} + (\lambda + \mu)\right) dv\right],$$

we have

$$\frac{dP(t)}{dt}\exp\left[\int_0^t (kB(v)/N+(\lambda+\mu))\,dv\right] + [kB(t)/N+(\lambda+\mu)].P(t)\exp\left[\int_0^t (kB(v)/N+(\lambda+\mu))\,dv\right]\geq 0.$$

Then,

$$\frac{d}{dt}\left[P(t)\exp\left[\int_0^t \left(k\frac{B(v)}{N}+(\lambda+\mu)\right)dv\right]\right]\geq 0. \tag{3}$$

Integrating (3) gives

$$P(t)\geq P(0)\exp\left[\int_0^t \left(-k\frac{B(v)}{N}-(\lambda+\mu)\right)dv\right].$$

So, the solution $P(t)$ is positive.

Likewise, utilizing system (1)'s second and third equations, we have

$$B(t)\geq B(0)\exp\left[-(\delta+\mu)t\right]\geq 0$$

and

$$C(t)\geq C(0)\exp(-\mu t)\geq 0.$$

Thus, we can observe that system (1)'s solutions $P(t)$, $B(t)$, and $C(t)$ are positive for all $t\geq 0$. □

2.2.2 Invariant region

Lemma 1. If the initial conditions are positive, that is, $P(0)\geq 0$, $B(0)\geq 0$, and $C(0)\geq 0$, then the region Ω defined by

$$\Omega=\left\{(P(t),B(t),C(t))\in\mathbb{R}^3_+, P(t)+B(t)+C(t)\leq\frac{A}{\mu}\right\}$$

is positive invariant for system (1).

Proof. When we sum up the system equations (1), we get

$$\frac{dN}{dt}\leq A-(\lambda+\mu)N.$$

Then,

$$N(t)\leq N(0)+At+\int_0^t -\mu N(v)dv.$$

Using a Gronwall lemma, we get

$$N(t) \leq N(0)\exp(-\mu t) + \frac{A}{\mu}(1-\exp(-\mu t)),$$

for the population's initial values as a whole. So, $\lim\limits_{t\to\infty} \sup N(t) = \dfrac{A}{\mu}$. For system (1), it suggests that the region Ω is a positively invariant set.

Thus, the dynamics of the system must be considered in the set Ω. □

3 Analysis of stability and model parameter sensitivity

This section investigates the system (1)'s stability behavior at both a boycotted equilibrium point and a boycott-free equilibrium point. System (1) possesses the subsequent two equilibrium points:

(1) boycott-free equilibrium given by $E_0 = \left(\dfrac{A}{\lambda+\mu}, 0, \dfrac{\lambda A}{\mu(\lambda+\mu)}\right)$. The situation in which there are no boycotters in the population.

(2) boycotted equilibrium point, if $R_0 > 1$, given by $E^* = (P^*, B^*, C^*)$, where $P^* = \dfrac{A(\delta+\mu)}{\mu k}$, $B^* = \dfrac{A(\lambda+\mu)(R_0-1)}{\mu k}$, and $C^* = \dfrac{\lambda A(\delta+\mu)^2 + \delta A(\lambda+\mu)(\delta+\mu)(R_0-1)}{\mu^2(\delta+\mu)k}$. This equilibrium reflects the situation in which a product boycott becomes widespread among the populace.

Here, R_0 is the basic reproduction number given by

$$R_0 = \frac{\mu k}{(\lambda+\mu)(\delta+\mu)}.$$

In the field of epidemiology, the basic reproduction number R_0 denotes the mean quantity of secondary infections among a fully susceptible population.

This threshold, as it relates to our work, denotes the mean number of prospective consumers that a boycotter will convince not to use the product during his interaction time.

In fact, if we assume that $x = (B, C, P)$, then the system (1) may be expressed as

$$\frac{dx}{dt} = \mathcal{F}(x) - \mathcal{W}(x),$$

where

$$\mathcal{F}(x) = \begin{pmatrix} k\dfrac{BP}{N} \\ 0 \\ 0 \end{pmatrix}$$

and

$$\mathcal{W}(x) = \begin{pmatrix} (\delta+\mu)B \\ -\lambda P - \delta B + \mu C \\ -A + k\dfrac{BP}{N} + (\lambda+\mu)P \end{pmatrix}.$$

At the free equilibrium E_0, the Jacobian matrices of $\mathcal{F}(x)$ and $\mathcal{W}(x)$ are

$$D\mathcal{F}(E_0) = \begin{pmatrix} F_{2\times 2} & 0 \\ & 0 \\ 0 \quad 0 & 0 \end{pmatrix}$$

and

$$D\mathcal{W}(E_0) = \begin{pmatrix} W_{2\times 2} & 0 \\ & -\lambda \\ \dfrac{\mu k}{(\lambda+\mu)} \; 0 & 0 \end{pmatrix},$$

where

$$F = \begin{pmatrix} \dfrac{\mu k}{(\lambda+\mu)} & 0 \\ 0 & 0 \end{pmatrix}$$

and

$$W = \begin{pmatrix} \delta+\mu & 0 \\ -\delta & \mu \end{pmatrix}.$$

At last, we have

$$R_0 = \rho\left(FW^{-1}\right) = \frac{\mu k}{(\lambda+\mu)(\delta+\mu)}.$$

3.1 Analysis of local stability

This section examines the boycotted equilibrium's and the boycott-free equilibrium's local stability.

Theorem 2. If $R_0 < 1$, then the boycott-free equilibrium E_0 is locally asymptotically stable; if $R_0 > 1$, then E_0 is unstable.

Proof. At E_0, the Jacobian matrix is provided by

$$J(E_0) = \begin{pmatrix} -(\lambda+\mu) & -\dfrac{\mu k}{(\lambda+\mu)} & 0 \\ 0 & \dfrac{\mu k}{(\lambda+\mu)} - (\delta+\mu) & 0 \\ \lambda & \delta & -\mu \end{pmatrix}.$$

Consequently, eigenvalues of $J(E_0)$'s characteristic equation are

$$\begin{aligned} \zeta_1 &= -\mu, \\ \zeta_2 &= -(\lambda+\mu), \\ \zeta_3 &= (\delta+\mu)(R_0-1). \end{aligned}$$

Clearly, the first and second eigenvalues ζ_1 and ζ_2, respectively, are negative. The third eigenvalue is also negative supplied that $R_0 < 1$.

We deduce that the boycott-free equilibrium E_0 is locally asymptotically stable if $R_0 < 1$, while it is unstable if $R_0 > 1$. □

After that, we assert the following theorem to ascertain the stability of the boycotted equilibrium E^*.

Theorem 3. The boycotted equilibrium E^* is locally asymptotically stable if $R_0 \geq 1$.

Proof. At E^*, the Jacobian matrix is provided by

$$J(E^*) = \begin{pmatrix} -k\dfrac{B^*}{N^*} - (\lambda+\mu) & -k\dfrac{P^*}{N^*} & 0 \\ k\dfrac{B^*}{N^*} & k\dfrac{P^*}{N^*} - (\delta+\mu) & 0 \\ \lambda & \delta & -\mu \end{pmatrix},$$

which is provided by its characteristic equation

$$\zeta^3 + a_1\zeta^2 + a_2\zeta + a_3 = 0,$$

where

$$a_1 = \frac{k\mu}{(\mu+\lambda)} + \mu,$$

$$a_2 = \frac{k\mu^2}{(\mu+\lambda)} + (\mu+\lambda)(\mu+\delta)(R_0-1),$$

$$a_3 = \frac{k\mu^3}{(\mu+\lambda)N} + \mu(\mu+\lambda)(\mu+\delta)(R_0-1).$$

Applying the Routh–Hurwitz criterion [6], if $a_1 > 0$, $a_3 > 0$, and $a_1 a_2 > a_3$, then system (1) is locally asymptotically stable.

Therefore, if $R_0 \geq 1$, then E^* is locally asymptotically stable. □

3.2 Analysis of global stability

The boycotted equilibrium E^* and boycott-free equilibrium E_0 of the model (1), respectively, constitute the global asymptotic stability that we are now concerned with.

Theorem 4. The free equilibrium E_0 of system (1) is globally asymptotically stable on Ω if $R_0 \leq 1$.

Proof. Consider the Lyapunov function $V_1 : \Omega \longrightarrow \mathbb{R}$ in the manner mentioned below:

$$V_1(P,B) = \frac{1}{2}\left[(P-P^0)+B\right]^2 + \frac{(\lambda+\delta+2\mu)N}{k}B.$$

Computing the time derivation of V_1, we get

$$\dot{V}_1(P,B) = (P-P^0+B)\left[A-(\lambda+\mu)P-(\delta+\mu)B\right] + \frac{(\lambda+\delta+2\mu)N}{k}\dot{B}. \quad (4)$$

Due to $A = P^0(\lambda+\mu)$, (4) becomes

$$\begin{aligned}\dot{V}_1(P,B) &= (P-P^0+B)[-(\lambda+\mu)(P-P^0)-(\delta+\mu)B] \\ &\quad + \frac{(\lambda+\delta+2\mu)N}{k}\dot{B} \\ &= -(\lambda+\mu)\left(P-P^0\right)^2 - (\delta+\mu)B^2 \\ &\quad -(\lambda+\delta+2\mu).\frac{(\delta+\mu)N}{k}(1-R_0)B.\end{aligned}$$

Consequently, $\dot{V}_1(P,B) \leq 0$ for $R_0 \leq 1$.

Moreover, if $R_0 \leq 1$, then $\dot{V}_1(P,B) = 0$ is equivalent to $B = 0$ and $P = P^0$.

Thus, Theorem 4 has been proved, and we can now say that by LaSalle's invariance principle [12], the boycott-free equilibrium E_0 is globally asymptotically stable on Ω. □

The global asymptotic stable theorem for the boycott-free equilibrium E^* is then presented as follows.

Theorem 5. The boycotted equilibrium E^* of system (1) is globally asymptotically stable on Ω if $R_0 > 1$.

Proof. Consider the Lyapunov function $V_2 : \Omega \longrightarrow \mathbb{R}$ in the manner mentioned below:

$$V_2(P,B) = Y_1\left[P - P^*\left(1 + ln\left(\frac{P}{P^*}\right)\right)\right] + Y_2\left[B - B^*\left(1 + ln\left(\frac{B}{B^*}\right)\right)\right],$$

where Y_1 and Y_2 are positive constants to be chosen later.

Computing the time derivation of V_2, we get

$$\dot{V}_2(P,B) = \frac{k}{N}(Y_1 - Y_2)(B - B^*)(P - P^*) - AY_1\frac{(P - P^*)^2}{PP^*}.$$

For $Y_1 = Y_2 = 1$, we get

$$\dot{V}_2(P,B) = -A\frac{(P - P^*)^2}{PP^*} \leq 0,$$

and

$\dot{V}_2(P,B) = 0$ is equivalent to $P = P^*$.

Thus, Theorem 5 has been proved, and we can now say that by LaSalle's invariance principle [12], the boycotted equilibrium E^* is globally asymptotically stable on Ω. □

3.3 Sensitivity analysis of the model's parameters

Sensitivity analysis is widely used to determine which parameters significantly affect the reproduction number R_0 or to evaluate a model's resilience to parameter values. In the context of [2, 14, 18], a sensitivity analysis of the model (1) is conducted.

Definition 1. If ξ is a variable that depends differently on t, then its normalized forward sensitivity index (S.I.) is defined as follows:

$$\Upsilon_{\xi}^{t} = \frac{\partial \xi}{\partial t}.\frac{t}{\xi}.$$

Specifically, the following are the computerized S.I.s of the fundamental reproduction number R_0 concerning the model parameters:

$$\begin{cases} \Upsilon_{\mu}^{R_0} = \dfrac{\partial R_0}{\partial \mu}.\dfrac{\mu}{R_0} = \dfrac{\lambda\delta - \mu^2}{(\lambda+\mu)(\delta+\mu)}, \\ \Upsilon_{k}^{R_0} = \dfrac{\partial R_0}{\partial k}.\dfrac{k}{R_0} = 1, \\ \Upsilon_{\delta}^{R_0} = \dfrac{\partial R_0}{\partial \delta}.\dfrac{\delta}{R_0} = -\dfrac{\delta}{\delta+\mu}, \\ \Upsilon_{\lambda}^{R_0} = \dfrac{\partial R_0}{\partial \lambda}.\dfrac{\lambda}{R_0} = -\dfrac{\lambda}{\lambda+\mu}. \end{cases}$$

A positive value of the S.I., that is $\Upsilon_{k}^{R_0}$ indicates that an increase (decrease) in the value of each parameter in this instance results in a proportional increase (decrease) in the basic reproduction number of the disease. Conversely, the negative sign of S.I. suggests that an increase (decrease) in the value of each of the parameters leads to a corresponding decrease (increase) in the basic reproduction number R_0. As an illustration, $\Upsilon_{k}^{R_0} = 1$ implies that a 15% increase or decrease in the effective contact rate k will result in a 15% increase or reduction in the basic reproduction number R_0. In Table 1, we present the sensitivity indices of all model parameters.

Therefore, sensitivity analysis provides information on the appropriate intervention tactics to stop and manage the emergence of a product boycott across the communities outlined in the model (1).

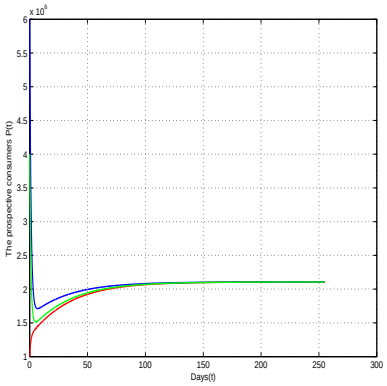
Table 1: Description and S.I. of parameters

Parameter	Description	Value	S.I.
μ	The natural death rate	0.053	+0.077
k	The effective impact rate of boycotters	0.4	+1
δ	The rate at which boycotters convert to actual consumers	0.01	−0.15
λ	The rate of transition of potential consumers to actual consumers	0.6	−0.91

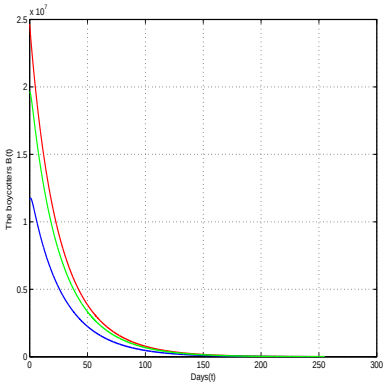
3.4 The numerical simulation

To support our theoretical findings on the stability analysis of the system (1), we provide some numerical simulations in this part. Certain simulation parameters are taken from [16, 17, 10, 20], which discusses the Moroccans' boycott of Centrale Danone [7]. The boycott lasted approximately one year, starting on the 20th of April 2018 [17], so we take $t_f = 255$ days. Each of the two sections of our testing is intended to demonstrate a different feature of the design. First, we aim to test Theorem 4, which states that the boycott free equilibrium E_0 of system (1) is globally stable on Ω. We choose parameters $A = 1375244$, $\lambda = 0.6$, $\delta = 0.01$, $\mu = 0.053$, $k = 0.4$, and $N = 25950000$, and we note that the parameter's model units in this work are in days.

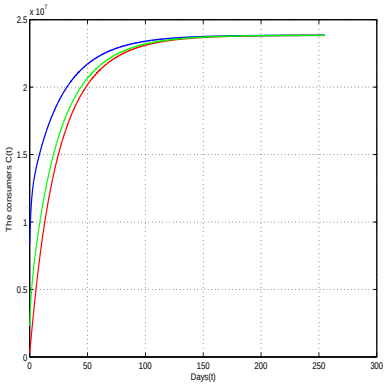
From Figure 2, where $R_0 < 1$, we can easily observe the global stability of the equilibrium $E_0 = (2.106 \times 10^6, 0, 2.384 \times 10^7)$ such that the variables (P), (B), and (C) converge to the equilibrium point E_0.



(a)



(b)



(c)

Figure 2: The convergence of the solutions to the equilibrium point E_0.

The second series of tests (see Figure 3) simulates the spread of boycott behavior of Central Danone's company in the population as a result of high prices and low product quality [7]. By chosen $A = 1375244$, $\lambda = 0.2$, $\delta = 0.01$, $\mu = 0.053$, $k = 0.6$, we have $R_0 > 1$. Then, according to Theorem 5, the equilibrium $E^* = (2.724 \times 10^6, 1.089 \times 10^7, 1.233 \times 10^7)$ is globally stable.

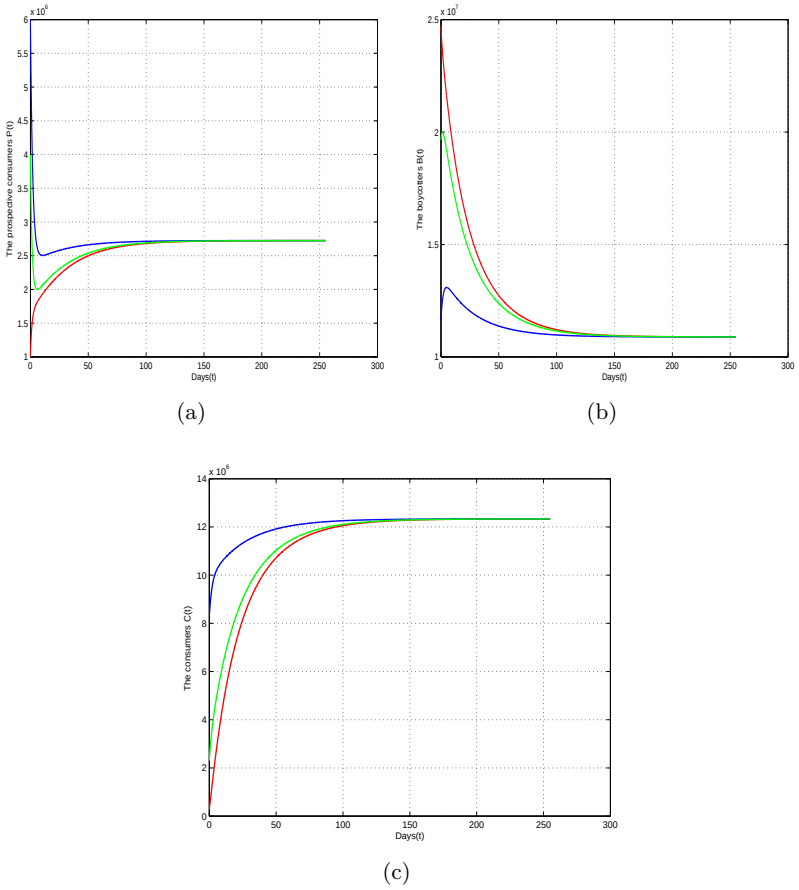


Figure 3: The convergence of the solutions to the equilibrium point E^*.

4 The Problem of optimal control

4.1 Problem synopsis

By targeting its products with boycott campaigns, each producing company aims to protect the level of sales of its product and maintain its loyal customers. To achieve this, appropriate strategies must be adopted to ensure that the number of consumers $C(t)$ is maximized and the number of interrupters $B(t)$ is minimized over the period $[t_0, t_f]$. For that, we propose in this work two controls. The control u_1 represents the effort made to introduce the product and encourage possible consumers to use it by promoting it through social media, including promotions and offers. The control u_2 indicates the efforts to deal with boycotters by understanding the reasons for their position and striving to meet their demands, including developing the product, gradually lowering its price over time, and providing incentives to the product's users. Thus, we consider our controlled mathematical model:

$$\begin{cases} \dot{P}(t) = A - k\dfrac{P(t)B(t)}{N} - (\mu+\lambda)P(t) - u_1(t)P(t), \\ \dot{B}(t) = k\dfrac{P(t)B(t)}{N} - (\delta+\mu)B(t) - u_2(t)B(t), \\ \dot{C}(t) = \lambda P(t) + \delta B(t) - \mu C(t) + u_1(t)P(t) + u_2(t)B(t), \end{cases} \tag{5}$$

with the initial conditions $P_0 \geq 0$, $C_0 \geq 0$, and $B_0 \geq 0$.

The problem is to minimize the objective functional,

$$J(u_1,u_2) = B(t_f) - C(t_f) + \int_{t_0}^{t_f} [B(v) - C(v) + \frac{M_1}{2}u_1^2(v) + \frac{M_2}{2}u_2^2(v)]dv, \tag{6}$$

where t_f is the final time, and the parameters M_1 and M_2 are the strictly positive cost coefficients; they are selected to weigh the relative importance of u_1 and u_2 at time t.

In other words, we look for the optimal controls u_1 and u_2 such that

$$J(u_1^*, u_2^*) = \min_{(u_1,u_2)\in U_{ad}^2} J(u_1,u_2),$$

where U_{ad} is the set of admissible controls defined by

$$U_{ad} = \{u_i(t) : 0 \leq u_i \leq 1, \text{ for i} = 1, 2, \text{ and t} \in [t_0, t_f]\}.$$

4.2 Optimal controls' existence

Fleming and Rishel's result (see [8, Corollary 4.1]) can be used to determine whether the optimal controls exist.

Theorem 6. Take into consideration the system (5) with control problem. An optimal control $(u_1^*, u_2^*) \in U_{ad}^2$ exists such that

$$J(u_1^*, u_2^*) = \min_{(u_1,u_2)\in U_{ad}^2} J(u_1, u_2)$$

if all of the following conditions hold:

1. The set of corresponding state variables and the controls is nonempty.
2. The U_{ad} control set is closed and convex.
3. A linear function in the state and control variables bounds the state system's right side.
4. The integrand $L(P, B, C, u_1, u_2)$ of the objective functional is convex on U_{ad} and there exist constants c_1, $c_2 > 0$, and $\epsilon > 1$ such that:

$$L(P, B, C, u_1, u_2) \geq -c_1 + c_2 \left(|u_1|^2 + |u_2|^2\right)^{\epsilon/2}.$$

Proof. **Condition 1.** To prove that the set of corresponding state variables and the controls is nonempty, a simplified version of an existing result (see [3, Theorem 7.1.1]) is used.

Let $\dot{P} = O_P(t; P, B, C)$, $\dot{B} = O_B(t; P, B, C)$, and $\dot{C} = O_C(t; P, B, C)$, where O_P, O_B, and O_C from the equations system (5)' right-hand side.

Let $u_i(t) = c_i$ for $i = 1, 2$ for some constants, and because all parameters are constants, and P, B, and C are continuous, then O_P, O_B, and O_C are also continuous.

Moreover, the partial derivatives $\frac{\partial O_P}{\partial P}$, $\frac{\partial O_P}{\partial B}$, $\frac{\partial O_P}{\partial C}$, $\frac{\partial O_B}{\partial P}$, $\frac{\partial O_B}{\partial B}$, $\frac{\partial O_B}{\partial C}$, and $\frac{\partial O_C}{\partial P}$, $\frac{\partial O_C}{\partial B}$, $\frac{\partial O_C}{\partial C}$ are all continuous. Consequently, there exists a unique solution (P, B, C) that fulfills the initial conditions.

Therefore, the set of corresponding state variables and the controls is nonempty, and Condition 1 is satisfied.

Condition 2. By definition, U_{ad} is closed. Take any controls $v_1, v_2 \in U_{ad}$ and $\varepsilon \in [0,1]$, then $0 \leq \varepsilon v_1 + (1-\varepsilon)v_2$.

Moreover, we note that $\varepsilon v_1 \leq \varepsilon$ and $(1-\varepsilon)v_2 \leq (1-\varepsilon)$. Then $\varepsilon v_1 + (1-\varepsilon)v_2 \leq \varepsilon + (1-\varepsilon) = 1$.

Therefore, $0 \leq \varepsilon v_1 + (1-\varepsilon)v_2 \leq 1$, for all $v_1, v_2 \in U_{ad}$ and $\varepsilon \in [0,1]$. Hence, U_{ad} is convex and Condition 2 is fulfilled.

Condition 3. Using the differential equations system (5), we get

$$\frac{dN}{dt} \leq A - \mu N.$$

So,

$$\lim_{t\to\infty} \sup N(t) \leq \frac{A}{\mu}.$$

As a result, every solution for model (5) is bounded.

Thus, there exist positive constants R_1, R_2, and R_3 such that for all $t \in [t_0, t_f]$,

$$\begin{aligned} P(t) &\leq R_1, \\ B(t) &\leq R_2, \\ C(t) &\leq R_3. \end{aligned}$$

We take into consideration

$$\begin{cases} O_P = \dot{P}(t) \leq A, \\ O_B = \dot{B}(t) \leq kP(t) - u_2(t)B(t), \\ O_C = \dot{C}(t) \leq \lambda P(t) + \delta B(t) + u_1(t)R_1 + u_2(t)R_2 \end{cases}$$

Then, system (5) can be rewritten in a matrix from as

$$O(t; P, B, C) \leq \bar{A} + BX(t) - RU(t),$$

where $O(t; P, B, C) = \begin{bmatrix} O_P & O_B & O_C \end{bmatrix}^T$, $\bar{A} = \begin{bmatrix} A & 0 & 0 \end{bmatrix}^T$, $X(t) = \begin{bmatrix} P & B & C \end{bmatrix}^T$, $U(t) = \begin{bmatrix} u_1 & u_2 \end{bmatrix}^T$, and

$$B = \begin{bmatrix} 0\ 0\ 0 \\ k\ 0\ 0 \\ \lambda\ \delta\ 0 \end{bmatrix}, \qquad R = \begin{bmatrix} 0 & 0 \\ 0 & B \\ -P & -B \end{bmatrix}.$$

The control vector and state variable vector are given by a linear function. Consequently, we are able to write

$$\begin{aligned} \|O(t;P,B,C)\| &\leq \|\bar{A}\| + \|B\|\|X(t)\| + \|R\|\|U(t)\| \\ &\leq \psi + \Phi\left(\|X(t)\| + \|U(t)\|\right), \end{aligned}$$

where $\psi = \|\bar{A}\|$ and $\Phi = \max\left(\|B\|, \|R\|\right)$.

As a result, we can observe that the sum of the state and control vectors bounds the right side. Consequently, condition 3 is met.

Condition 4. The integrand in the objective functional (6) is convex on U_{ad}. The goal is to demonstrate that there exist constants $c_1, c_2 > 0$ and $\epsilon > 1$ such that the integrand $L(P,B,C,u_1,u_2)$ of the objective functional satisfies

$$\begin{aligned} L(P,B,C,u_1,u_2) &= B(t) - C(t) + \frac{M_1}{2}u_1^2 + \frac{M_2}{2}u_2^2 \\ &\geq -c_1 + c_2\left(|u_1|^2 + |u_2|^2\right)^{\epsilon/2}. \end{aligned}$$

Since the state variables are bounded, let $\epsilon = 2$, $c_1 = 2 \sup_{t\in[t_0,t_f]} (B,C)$, and $c_2 = \inf(\frac{M_1}{2}, \frac{M_2}{2})$. Subsequently, it implies that

$$L(P,B,C,u_1,u_2) \geq -c_1 + c_2\left(|u_1|^2 + |u_2|^2\right)^{\epsilon/2}.$$

□

4.3 The optimal controls' characterization

In this part, we make use of Pontryagin's maximal principle [22]. The key idea is to use the adjoint function to produce the Hamiltonian function by connecting the differential equations system to the objective function. This idea transfers the problem of finding the control to optimize the objective

functional subject to the state of differential equations with initial condition and then finds the control to optimize the Hamiltonian pointwise (concerning the control).

The Hamiltonian H in time t is defined as

$$H(t) = B(t) - C(t) + \frac{M_1}{2}u_1^2(t) + \frac{M_2}{2}u_2^2(t) + \sum_{i=1}^{3}\zeta_i f_i,$$

where f_i represents the right side of the ith state variable's differential equations system (5).

Theorem 7. Given an optimal control $u^* = (u_1^*, u_2^*) \in U_{ad}^2$ and corresponding solutions P^*, B^*, and C^* of corresponding state system (5), there exist adjoint functions ζ_1, ζ_2, and ζ_3 fulfilling

$$\begin{cases} \dot{\zeta}_1 = \zeta_1\{k\frac{B(t)}{N} + \lambda + \mu + u_1(t)\} - \zeta_2 k\frac{B(t)}{N} - \zeta_3\{\lambda + u_1(t)\}, \\ \dot{\zeta}_2 = -1 + \zeta_1 k\frac{P(t)}{N} - \zeta_2\{k\frac{P(t)}{N} - \delta - \mu - u_2(t)\} - \zeta_3\{\delta + u_2(t)\}, \\ \dot{\zeta}_3 = 1 + \zeta_3\mu. \end{cases} \tag{7}$$

At the time t_f, given the transversality conditions, we have

$$\begin{aligned} \zeta_1(t_f) &= 0, \\ \zeta_2(t_f) &= 1, \\ \zeta_3(t_f) &= -1. \end{aligned}$$

Moreover, the optimal controls $u_1^*(t)$ and $u_2^*(t)$ for $t \in [t_0, t_f]$ are provided by

$$u_1^*(t) = \min\left\{1, \max\left\{0, \frac{1}{M_1}P(t)(\zeta_1(t) - \zeta_3(t))\right\}\right\}, \tag{8}$$

$$u_2^*(t) = \min\left\{1, \max\left\{0, \frac{1}{M_2}B(t)(\zeta_3(t) - \zeta_2(t))\right\}\right\}. \tag{9}$$

Proof. The Hamiltonian H in time t, is defined by

$$\begin{aligned} H(t) = B(t) - C(t) + \frac{M_1}{2}u_1^2(t) + \frac{M_2}{2}u_2^2(t) \\ +\zeta_1\{A - k\frac{B(t)P(t)}{N} - (\lambda + \mu)P(t) - u_1(t)P(t)\} \\ +\zeta_2\{k\frac{B(t)P(t)}{N} - (\delta + \mu)B(t) - u_2(t)B(t)\} \end{aligned}$$

$$+\zeta_3\{\lambda P(t) + \delta B(t) - \mu C(t) + u_1(t)P(t) + u_2(t)B(t)\}.$$

Using Pontryagin maximum principle, one may obtain the transversality conditions and adjoint equations, for $t \in [t_0, t_f]$, such that

$$\begin{aligned} \zeta_1(t_f) &= 0, & \dot{\zeta}_1(t) &= -\frac{\partial H}{\partial P}, \\ \zeta_2(t_f) &= 1, & \dot{\zeta}_2(t) &= -\frac{\partial H}{\partial B}, \\ \zeta_3(t_f) &= -1, & \dot{\zeta}_3(t) &= -\frac{\partial H}{\partial C}. \end{aligned}$$

The optimality condition can be used to solve the optimal controls $u_1^*(t)$ and $u_2^*(t)$ for $t \in [t_0, t_f]$. We have

$$\begin{aligned} \frac{\partial H}{\partial u_1} &= M_1 u_1(t) + \zeta_1(t)\{-P(t)\} + \zeta_3(t)\{P(t)\} = 0, \\ \frac{\partial H}{\partial u_2} &= M_2 u_2(t) + \zeta_2(t)\{-B(t)\} + \zeta_3(t)\{B(t)\} = 0. \end{aligned}$$

That is,

$$\begin{aligned} u_1(t) &= \frac{1}{M_1} P(t)(\zeta_1(t) - \zeta_3(t)), \\ u_2(t) &= \frac{1}{M_2} B(t)(\zeta_3(t) - \zeta_2(t)). \end{aligned}$$

It is easy to obtain $u_1^*(t)$ and $u_2^*(t)$ in the form of (8) and (9) by the bounds in U_{ad} of the controls. □

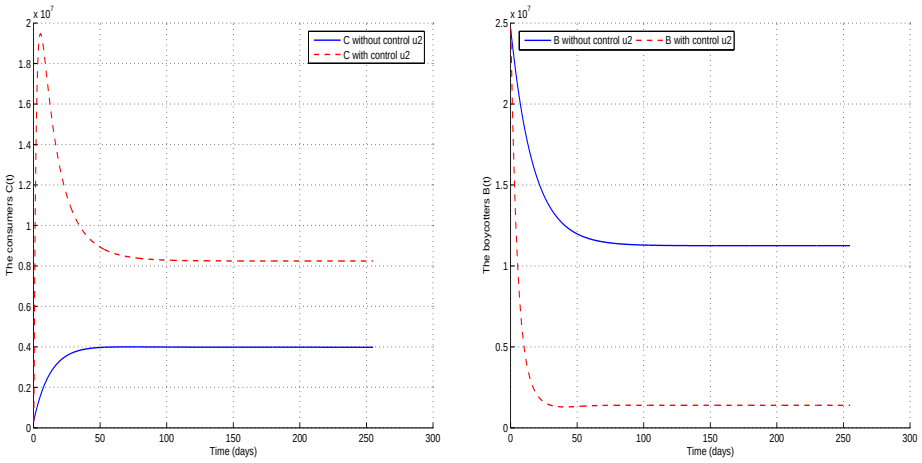
4.4 The numerical simulation

We have starting conditions for the state variables and terminal conditions for the adjoints in our control problem. The optimality system is a two-point boundary value problem with discrete boundary conditions at periods step $i = t_0$ and $i = t_f$. We solve the optimality system iteratively by solving the adjoint system backward after solving the state system forward. We first estimate the controls in the first iteration. Next, we adjust the controls before the subsequent iteration based on the characterization. We continue until the next iteration converges. The following data are used to create and compile

a code in MATLAB: $A = 900000$, $\lambda = 0.05$, $\delta = 0.011$, $\mu = 0.053$, $k = 0.95$, and $N = 25950000$ (the parameter's model units are in days).

Strategy 1: Correcting fallacies and restoring confidence in the product.

In this strategy, we concentrate the efforts, using optimal control u_2, to deal with boycotters by understanding the reasons for their position and striving to meet their demands, including developing the product, addressing negative rumors, restoring their confidence, and gradually lowering its price.



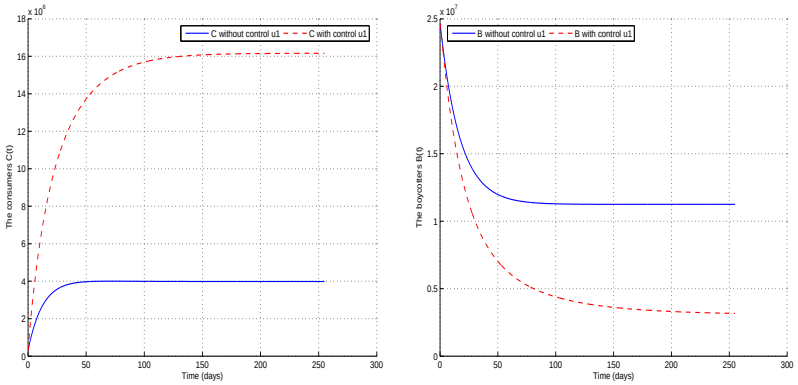
(a) (b)

Figure 4: Optimal consumers and boycotters with and without control u_2^*.

From Figures 4a and 4b, we can see that the number of consumers of the product has grown from 4×10^6 to 8.243×10^6. Also, the number of boycotters has decreased from 1.125×10^7 to 1.392×10^6.

Strategy 2: Publicity and Marketing.

Using the optimal control u_1, this strategy focuses on stimulating and compelling advertising to encourage and motivate potential consumers to use the product and protect them from being affected by boycott campaigns.



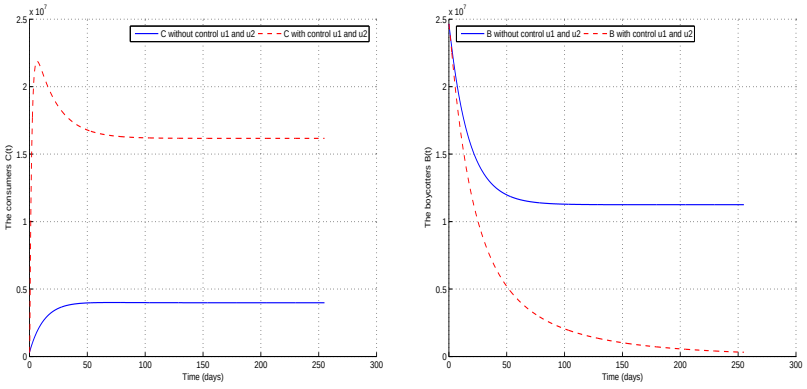
(a) (b)

Figure 5: Optimal consumers and boycotters with and without control u_1^*.

From Figures 5a and 5b, we observe that the number of consumers increased from 4×10^6 to 1.616×10^7. Also, the number of boycotters decreased from 1.125×10^7 to 3.174×10^6.

Strategy 3: Encouraging and motivating potential consumers and targeting boycotters.

This strategy aims to improve the numerical outcomes of cases 1 and 2 by activating simultaneously the two optimal controls u_1 and u_2.



(a) (b)

Figure 6: Optimal consumers and boycotters with and without controls u_1^* and u_2^*.

From Figure 6a, we show clearly that the number of consumers grows significantly from 4×10^6 to 1.617×10^7. Also, Figure 6b shows that the number of boycotters dropped from 1.125×10^7 to 3.255×10^5, which means the proposed strategy is more effective when we combine two optimal controls u_1^* and u_2^*.

Ultimately, we deduce that the suggested approach becomes more successful when we combine the two optimal controls, u_1^* and u_2^*.

Therefore, we observe that as the number of boycotters decreases, their influence on potential consumers diminishes as well, and thus trust is renewed between the product and the consumer. Consequently, a large number of consumers tend to give the product another try, often in a shorter time. This is particularly evident with the introduction of marketing campaigns that address misconceptions, highlight new features of the product, and interact positively with citizens' demands to respect reasonable prices.

5 Conclusion

In this research, we proposed a mathematical model that describes the boycott behavior of citizens regarding a product. We studied the stability analysis of the equilibriums of the proposed model, as well as the sensitivity analysis, in order to know more about the parameters that have a high impact on the reproduction number R_0. Using the results of optimal control theory, we have presented optimal strategies to persuade boycotters of a product to retract their position and thus reduce their influence on potential consumers. We ultimately concluded that both the number of boycotters and their influence on potential consumers decreased, leading to a renewal of trust between the product and the consumer. Consequently, a large number of consumers tend to give the product another try, often in a shorter period, especially with the launch of marketing campaigns that rectify misconceptions, highlight new features of the product, and interact positively with citizens' demands to respect reasonable prices. This study could have other interesting extensions, such as studying stochastic stability and optimal control in a stochastic version of our model through stochastic outcomes. This approach provides an additional degree of realism compared to its de-

terministic counterpart via a stochastic differential equation and includes the effect of a fluctuating environment.

References

[1] Albayati, M.S., Mat, N.K.N., Musaibah, A.S., Aldhaafri, H.S. and Almatari, E.M. *Participate in boycott activities toward danish products from the perspective of Muslim consumer*, Am. J. Econ. Special Issue (2012), 120–124.

[2] Balatif, O., Khajji, B. and Rachik, M. *Mathematical modeling, analysis, and optimal control of abstinence behavior of registration on the electoral lists*, Discrete Dyn. Nature Soc. 2020 (2020).

[3] Boyce, W.E. and DiPrima, R.C. *Elementary differential equations and boundary value problems*, John Wiley and Sons, New York, NY, USA, 2009.

[4] Braunsberger, K. and Buckler, B. *What motivates consumers to participate in boycotts: Lessons from the ongoing Canadian seafood boycott*, J. Bus. Res. 64(3) (2011), 96–102.

[5] Diermeier, D. and Van Mieghem, J.A. *Voting with your Pocketbook - A Stochastic Model of Consumer Boycotts*, Math. Comput. Model. 48 (2008), 1497–1509.

[6] Edelstein-Keshet, L. *Mathematical Models in Biology*, SIAM, 1988.

[7] EL Arraf, S. and Biddou, N. *Corporate Response and Responsibility in the case of consumer boycotts: an Analysis of Centrale Danone crisis in Morocco*, Journal d'Economie, de Management, d'Environnement et de Droit, 2(2) (2019), 29–30.

[8] Fleming, W.H. and Rishel, R.W. *Deterministic and Stochastic Optimal Control*, Springer, New York, NY, USA, 1975.

[9] Hoffmann, S. *Are boycott motives rationalizations?*, J. Consum. Behav. 12(3) (2013), 214–222.

[10] *Indicateurs-Sociaux-ar* [online]. The Ministry of Economy and Finance, (2018). Available from: https://www.finances.gov.ma, (access date : 9 January 2024).

[11] Klein, J.G., Smith, N.C. and John, A. *Why We Boycott: Consumer Motivations for Boycott Participation*, J. Mark. 68(3) (2004), 92–109.

[12] LaSalle, J.P. *The stability of dynamical systems*, Regional Conference Series in Applied Mathematics Vol. 25, SIAM. Philadelphia, PA, USA, 1976.

[13] Li, C. and Ma, Z. *Dynamics analysis of a mathematical model for new product innovation diffusion*, Discrete Dyn. Nature Soc. 2020 (2020), 13 pages.

[14] Makinde, O.D. and Okosun, K.O. *Impact of chemo-therapy on optimal control of malaria disease with infected immigrants*, Biosystems 104(1) (2011), 32–41.

[15] Mayo, C.C. *Captain Boycott 1832-1897* [online]. Comhairle Contae Mhaigh Eo Mayo County Council, (2006). Available from: https://www.mayo.ie/discover/history-heritage/great-battles-conflicts/captain-boycott, (access date : 9 January 2024).

[16] Mesbah, M. *"Khalih Yreeb": Boycott campaign and empowering the role of ordinary citizen* [online]. Moroccan Institute For Policy Analysis. Available from: https://mipa.institute/6734, (access date : 11 January 2024).

[17] Mosameh, M. *The boycott campaign for Central products loses "15 billion centimeters"* [online]. Berrechid news, (6 April 2018). Available from: https://www.berrechidnews.com/2018/06/5074.html, (access date : 01 January 2024).

[18] Nakul, C., Cushing J.M. and Hyman, J.M. *Bifurcation analysis of a mathematical model for malaria transmission*, SIAM J. Appl. Math. 67(1) (2006), 24–45.

[19] Neilson, L.A. *Boycott or buycott? Understanding political consumerism*, J. Consum. Behav. 9(3) (2010), 214–227.

[20] *News note for the high commission for planning about the basic characteristics of an active, working population During the year 2018*, High Commission for Planning, (2018). Available from: https://www.hcp.ma/region-drda.

[21] Palacios-Florencio, B., Revilla-Camacho, M.A., Garzón, D. and Prado-Román, C. *Explaining the boycott behavior: A conceptual model proposal and validation*, J. Consum. Behav. 20(5) (2021), 1313–1325.

[22] Pontryagin, L.S., Boltyanskii, V.G. Gamkrelidze, R.V. and Mishchenko, E.F. *The mathematical theory of optimal processes*, Wiley, New York, NY, USA, 1962.

[23] Zhuo, C., Chen S. and Yan, H. *Mathematical modelling of B2C consumer product supply strategy based on nonessential demand pattern*, J. Math. 2024 (2024), 14 pages.

[24] Zhuo, Z., Chau K.Y., Huang, S. and Kit Ip, Y. *Mathematical modeling of optimal product supply strategies for manufacturer-to-group customers based on semi-real demand patterns*, Int. J. Eng. Bus. Manag. 12 (2020), 1–8.

[25] Zhuo, Z., Chen, S., Yan, H. and He, Y. *A new demand function graph: Analysis of retailer-to-individual customer product supply strategies under a non-essential demand pattern*, Plos one 19(2) (2024), e0298381.

Iranian Journal of Numerical Analysis and Optimization
Vol. 14, No. 3, 2024, pp 736–761
https://doi.org/10.22067/ijnao.2024.86994.1398
https://ijnao.um.ac.ir/

Research Article

Highly accurate collocation methodology for solving the generalized Burgers–Fisher's equation

S. Shallu*, and V.K. Kukreja

Abstract

An improvised collocation scheme is applied for the numerical treatment of the nonlinear generalized Burgers–Fisher's (gBF) equation using splines of degree three. In the proposed methodology, some subsequent rectifications are done in the spline interpolant, which resulted in the magnification of the order of convergence along the space direction. A finite difference approach is followed to integrate the time direction. Von Neumann methodology is opted to discuss the stability of the method. The error bounds and convergence study show that the technique has $(s^4 + \Delta t^2)$ order of convergence. The correspondence between the approximate and analytical solutions is shown by graphs, plotted using MATLAB and by evaluating absolute error.

*Corresponding author

Received 22 February 2024; revised 12 April 2024; accepted 13 April 2024

Shallu Shallu
Department of Mathematics, Punjab Engineering College (Deemed to be University), Chandigarh, 160012, India. e-mail: shallugupta024@gmail.com

Vijay Kumar Kukreja
Department of Mathematics, SLIET Longowal 148106 (Punjab) India. e-mail: vkkukreja@gmail.com

How to cite this article
Shallu, S. and Kukrej, V.K., Highly accurate collocation methodology for solving the generalized Burgers–Fisher's equation. *Iran. J. Numer. Anal. Optim.*, 2024; 14(3): 736-761. https://doi.org/10.22067/ijnao.2024.86994.1398

AMS subject classifications (2020): Primary 35G31; Secondary 65M70.

Keywords: Generalized Burgers–Fisher's equation; Cubic B-splines; Collocation method; Finite difference scheme; Green's function; Von Neumann analysis.

1 Introduction

Lu et al. [17] found that the generalized Burgers–Fisher's (gBF) problem is an extension of generalized Fisher's equation, which is as mentioned hereunder:

$$v_t = \beta v_{xx} + f(v, v_x), \quad x \in (a, b), \quad t \in (t_0, T). \tag{1}$$

Equation (1) can be expressed in the operator form as mentioned below:

$$L \equiv \beta v_{xx} - v_t + f(v, v_x), \tag{2}$$

with the initial condition as:

$$v = v^0, \quad \text{in} \quad [a, b] \times \{t_0\}, \tag{3}$$

and the boundary conditions as:

$$\mathcal{B}v = \Omega, \quad \text{on} \quad \partial\Phi_x \times [t_0, T], \tag{4}$$

where $f(v, v_x) = -\alpha v^\sigma v_x + \gamma v(1 - v^\sigma)$, $\Phi_x = (a, b)$, $\mathcal{B}$ is the boundary operator defined as $\mathcal{B}v = a_1(x, t)v(x, t) + a_2(x, t)v_x(x, t)$. Here v represents the traveling wave phenomena with $\sigma > 0$ and $T > t_0$. Also α, β, and γ correspond to the convection, diffusion, and reaction coefficients, respectively. With $\sigma = 1$, (1) becomes the Burgers–Fisher's equation given below:

$$v_t + \alpha v v_x = \beta v_{xx} + \gamma v(1 - v), \quad x \in (a, b), \quad t \in (t_0, T). \tag{5}$$

This is known as the Burgers–Fisher's equation because it has convective phenomena from the Burgers' problem, diffusion transport along with reaction characteristics from the Fisher's equation. Thus, it is a blending of convection, diffusion, and reaction mechanisms. The proposed problem was

used by Sachdev [26] in self-similarity. When $\gamma = 0$, (5) becomes the Burgers' problem, which was used by Lighthill [16] in the investigation of sound waves in a viscous medium. When $\alpha = 0$, (5) reduces to the modified Fisher's problem, which was used by Murray [23] in mathematical biology.

Since two nonlinear terms occur in (5), therefore analytical methods such as Laplace, Fourier, and other classical approaches to integrate the system become invalid. Due to this, the traveling wave solution of the gBF equation was found by Fan [7] using the extended tanh-function and the Riccati equation. Mickens and Gumel [19] studied the properties of the Burgers–Fisher's problem and worked on its numerical solution using the nonstandard finite difference technique. Kaya and Sayed [15] obtained an explicit series solution of the gBF equation without any transformation and compared it with the numerical solution obtained using the Adomian decomposition technique. This technique was extended by Ismail, Raslan, and Abd Rabboh [12] to analyze the Burgers–Fisher's and Burgers–Huxley's equation. Javidi [13] solved this equation using a combination of pseudospectral Chebyshev and Runge-Kutta fourth-order methods. A variational iteration scheme based on Lagrange multipliers to construct correction functions for the gBF problem was adapted by Moghimi and Hejazi [21]. Wazwaz [38] derived the sets of traveling wave solutions as well as kinks and periodic solutions of the gBF equation using the tanh-coth method. The spectral domain decomposition technique with Chebyshev polynomials for spatial derivatives and RK4 for time integration was used by Golbabai and Javidi [10]. Zhu and Kang [40] applied the B-spline quasi-interpolation technique and opted for forward difference for temporal discretization to solve the Burgers–Fisher's equation. A finite difference technique of sixth-order for space, and the third-order Runge–Kutta method for temporal domain was applied by Sari, Güraslan, and Dağ [30] for the gBF equation.

Bratsos [2] implemented the finite difference technique of order four for space discretization and a predictor-corrector technique for solving the resulting nonlinear system. Sari [29] adapted the polynomial-based differential quadrature technique for space and the SSP-RK scheme of third-order for time to solve the gBF equation. Tatari, Sepehrian, and Alibakhshi [35] used the collocation method with the radial basis function to solve the system of

nonlinear equations by the predictor-corrector method. Zhao et al. [39] used the Legendre–Galerkin formulation for space discretization with Chebyshev–Gauss–Lobatto node points and leapfrog scheme for temporal discretization. Mohammadi [22] used an explicit exponential spline difference scheme for the gBF equation and analyzed convergence, error, and stability properties with the energy method. The limitation of the work was the large computational time. A modified spline collocation technique with the SSPRK-54 scheme was applied by Mittal and Tripathi [20] to analyze the gBF problem. Malik et al. [18] adapted a heuristic genetic algorithm scheme for the gBF equation based on an exp-function hybridization technique.

Chandraker, Awasthi, and Jayaraj [4] applied two implicit finite difference schemes to solve the Burgers–Fisher problem; one was semi-implicit and the other was based on the modified Crank–Nicolson method. Al-Rozbayani and Al-Hayalie [1] applied three different finite difference schemes to solve the Burgers–Fisher's equation. One is an explicit method, the other is an exponential method and the third one is the Du Fort–Frankel method. Hepson [11] implemented an extended B-spline collocation technique to solve the gBF equation. Saeed and Gilani [27] proposed a combination of the CAS wavelet method with a quasi-linearization scheme to solve the gBF equation. Sangwan and Kaur [28] applied a piecewise uniform Shishkin mesh with exponentially fitted splines and for temporal discretization, the implicit Euler method was adopted. The quasilinearization was used to deal with the nonlinear terms. Bratsos and Khaliq [3] adapted an exponential time differencing technique in which a nonlinear system was solved by a second-order modified predictor-corrector scheme.

In this study, we employ an extrapolated collocation algorithm to investigate the gBF equation. This method, previously utilized by Shallu, Kumari, and Kukreja [34, 31, 32], has been successfully applied to solve second-order self-adjoint equations, modified Burgers' equations, as well as RLW and MRLW equations. We enhance the methodology by utilizing improved cubic B-splines for spatial discretization and employing a weighted finite difference method for temporal discretization. These adjustments lead to a notable enhancement in the convergence order in the spatial domain.

The structure of the paper is as follows: In Section 2, we detail the construction and implementation of the improved B-spline collocation methodology for addressing the given problem. Subsequently, we conduct a convergence analysis in the spatial domain. In Section 3, we proceed with the discretization of the temporal domain. Section 4 involves the utilization of the von Neumann method to assess the stability of our proposed approach. In Section 5, we present solved examples to demonstrate the effectiveness of our technique and its superiority over existing data. Finally, in Section 6, we provide a summary of our findings.

2 New cubic B-Spline collocation technique

Consider the uniform subdivision of the Π_x space domain with $s = (b-a)/M$ as the step length of the space domain and $M+1$ is the number of nodal points. The structure of cubic splines $C_{p,3}(x)$ is given in [24]. The numerical solution can be written as follows:

$$W(x,t) = \sum_{p=-1}^{M+1} d_p(t) C_p(x). \tag{6}$$

2.1 Corrections in the second-order derivative

Assume that the spline interpolant $W(x,t)$ fulfills the given constraints:

(I) the interpolatory constraints, for $p = 0, 1, \ldots, M$:

$$W(x_p,t) = v(x_p,t), \tag{7}$$

(II) at the end nodal points, for $p = 0$ and M:

$$W_{xx}(x_p,t) = v_{xx}(x_p,t) - \frac{s^2}{12} v_{xxxx}(x_p,t). \tag{8}$$

Theorem 1. The following relations hold among the cubic spline interpolant (CSI) $W(x,t)$ and the exact solution $v(x,t)$, where $v(x,t)$ satisfy (7) and (8) for $p = 0, 1, \ldots, M$:

$$W_{xx}(x_p,t) = v_{xx}(x_p,t) - \frac{s^2}{12}v_{xxxx}(x_p,t) + O(s^4),$$
$$W_x(x_p,t) = v_x(x_p,t) + O(s^4).$$

In addition,

$$\| W^{(j)} - v^{(j)} \|_\infty = O(s^{4-j}), \quad j = 0,1,2,$$

where $W^{(j)}$ and $v^{(j)}$ represent the jth derivative with respect to "x".

Proof. See [5]. □

Lemma 1. For $v(x,t) \in \mathbb{C}^6[a,b]$, the below mentioned relations hold:

For $p = 0$:

$$v_{xxxx}(x_0,t) = \frac{W_{xx}(x_0,t) - 5W_{xx}(x_1,t) + 4W_{xx}(x_2,t) - W_{xx}(x_3,t)}{x^2} + O(s^2).$$

For $p = 1,2,\ldots,M-1$:

$$v_{xxxx}(x_p,t) = \frac{W_{xx}(x_{p-1},t) - 2W_{xx}(x_p,t) + W_{xx}(x_{p+1},t)}{x^2} + O(s^2).$$

For $p = M$:

$$v_{xxxx}(x_M,t)$$
$$= \frac{W_{xx}(x_M,t) - 5W_{xx}(x_{M-1},t) + 4W_{xx}(x_{M-2},t) - W_{xx}(x_{M-3},t)}{x^2} + O(s^2).$$

Proof. See [5]. □

Corollary 1. For $v(x,t) \in \mathbb{C}^6[a,b]$, the below given relations hold:

For $p = 0,1,\ldots,M$:

$$v_x(x_p,t) = W_x(x_p,t) + O(s^4),$$

For $p = 0$:

$$v_{xx}(x_0,t) = \frac{14W_{xx}(x_0,t) - 5W_{xx}(x_1,t) + 4W_{xx}(x_2,t) - W_{xx}(x_3,t)}{12} + O(s^4),$$

For $p = 1,2,\ldots,M-1$:

$$v_{xx}(x_p,t) = \frac{W_{xx}(x_{p-1},t) + 10W_{xx}(x_p,t) + W_{xx}(x_{p+1},t)}{12} + O(x^4),$$

For $p = M$:

$$v_{xx}(x_M,t)$$

$$= \frac{14W_{xx}(x_M,t) - 5W_{xx}(x_{M-1},t) + 4W_{xx}(x_{M-2},t) - W_{xx}(x_{M-3},t)}{12} + O(s^4).$$

Proof. See [5]. □

2.2 System of equations

At the nodal points, (1) can be expressed as follows:

$$\begin{aligned} v_t(x_p,t) &= \beta v_{xx}(x_p,t) + f(v(x_p,t), v_x(x_p,t)), \quad x_p \in [a,b], \\ \mathcal{B}(v(x_p,t)) &= \Omega(v(x_p,t)), \quad x_p \in \partial\Phi_x. \end{aligned}$$

Substituting the values of $v(x_p,t)$, $v_x(x_p,t)$, and $v_{xx}(x_p,t)$ in the above equations and using Corollary 1, we have

$$\begin{aligned} \frac{\partial}{\partial t}W(x_0,t) &= \frac{\beta}{12}\left[14W_{xx}(x_0,t) - 5W_{xx}(x_1,t) + 4W_{xx}(x_2,t) - W_{xx}(x_3,t)\right] \\ &\quad + \mathfrak{f}(W(x_0,t), W_x(x_0,t)) + O(s^4), \end{aligned} \tag{9}$$

$$\begin{aligned} \frac{\partial}{\partial t}W(x_p,t) &= \frac{\beta}{12}\left[W_{xx}(x_{p-1},t) + 10W_{xx}(x_p,t) + W_{xx}(x_{p+1},t)\right] \\ &\quad + \mathfrak{f}(W(x_p,t), W_x(x_p,t)) + O(s^4), \quad p = 1,2,\ldots,M-1, \end{aligned} \tag{10}$$

$$\begin{aligned} \frac{\partial}{\partial t}W(x_M,t) &= \frac{\beta}{12}[14W_{xx}(x_M,t) - 5W_{xx}(x_{M-1},t) + 4W_{xx}(x_{M-2},t) \\ &\quad - W_{xx}(x_{M-3},t)] + \mathfrak{f}(W(x_M,t), W_x(x_M,t)) + O(s^4). \end{aligned} \tag{11}$$

and the boundary constraints:

$$a_1(x_p,t)W(x_p,t) + a_2(x_p,t)W_x(x_p,t) = \Omega(x_p,t) + O(s^4), \quad p = 0, M. \tag{12}$$

The above relations form a nonlinear vector initial value problem of first-order with (3) as initial constraint.

2.3 Spatial convergence analysis

Let $\hat{L}$ and $\hat{\mathcal{B}}$ be the perturbation operators of L and $\mathcal{B}$. Then, the below given connections hold:

For $p = 1, 2, \ldots, M-1$:

$$\begin{aligned}
\hat{L}W(x_p,t) &\equiv L[W(x_p,t), W_x(x_p,t), W_{xx}(x_p,t) \\
&\quad + \frac{1}{12}[W_{xx}(x_p,t) - 2W_{xx}(x_p,t) + W_{xx}(x_p,t)], \\
\hat{L}W(x_0,t) &\equiv L[W(x_0,t), W_x(x_0,t), W_{xx}(x_0,t) \\
&\quad + \frac{1}{12}[2W_{xx}(x_0,t) - 5W_{xx}(x_1,t) + 4W_{xx}(x_2,t) - W_{xx}(x_3,t)], \\
\hat{L}W(x_M,t) &\equiv L[W(x_M,t), W_x(x_M,t), W_{xx}(x_M,t) \\
&\quad + \frac{1}{12}[2W_{xx}(x_M,t) - 5W_{xx}(x_{M-1},t) + 4W_{xx}(x_{M-2},t) \\
&\quad - W_{xx}(x_{M-3},t)]. \qquad (13) \\
\hat{\mathcal{B}}W(x_p,t) &= \mathcal{B}W(x_p,t), \quad p = 0, M.
\end{aligned}$$

Thus it is deduced that, for the unique CSI that satisfies (7)–(8), the following mentioned connections hold at the nodal points:

$$\hat{L}W(x_p,t) = O(s^4), \quad p = 0, 1, \ldots, M; \quad \hat{\mathcal{B}}W(x_p,t) = O(h^4), \quad p = 0, M.$$

The purpose is to find a cubic spline solution $\hat{v}(x,t)$, such that

$$\hat{L}\hat{v}(x_p,t) = 0, \quad p = 0, 1, \ldots, M; \quad \hat{\mathcal{B}}\hat{v}(x_p,t) = 0, \quad p = 0, M. \qquad (14)$$

Next, Green's function is applied for the establishment of error bounds.

Lemma 2. The coefficient matrix of $v_{xx} = g(x,t)$ having homogeneous boundary constraints has inverse with finite norm.

Proof. Using the steps of formation, the coefficient matrix $\beth$ of the equation $v_{xx} = g(x,t)$ is as mentioned below by using (13):

$$\beth = \frac{1}{12}\begin{bmatrix} 14 & -5 & 4 & -1 & 0 & \dots & 0 \\ 1 & 10 & 1 & 0 & 0 & \dots & 0 \\ 0 & 1 & 10 & 1 & 0 & \dots & 0 \\ \vdots & \vdots & \vdots & \vdots & \vdots & \dots & \vdots \\ 0 & \dots & 0 & 1 & 10 & 1 & 0 \\ 0 & \dots & 0 & 0 & 1 & 10 & 1 \\ 0 & \dots & 0 & -1 & 4 & -5 & 14 \end{bmatrix}$$

Due to the diagonal dominance behavior of matrix, it is invertible. Moreover,

$$\|\beth^{-1}\|_\infty \leq \max_{0\leq p\leq M} \frac{1}{\Delta_p \beth},$$

where

$$\Delta_p \beth = \mid \beth_{pp} \mid - \sum_{j\neq p} \mid \beth_{pj} \mid > 0 \quad \text{for } p = 0, 1, \dots, M.$$

So,

$$\|\beth^{-1}\|_\infty \leq \frac{1}{\min_{0\leq p\leq M} \Delta_p(\beth)} = \frac{12}{14 - (5+4+1)} = 3.$$

Now, onwards $W^{(j)}$, $v^{(j)}$, and $\hat{v}^{(j)}$ are the jth differentiation with respect to space variable. Let $\hat{\beth}$ denote the coefficient matrix of $W^{(1)}(x,t)$ in (9)–(12), that is, $\hat{\beth} = diag(-\frac{5}{h}, 0, \frac{5}{h})$, which is invertible with finite norm. Since the boundary value problem of the form (1) with the boundary constraints (4) can be transformed into the Fredholm integral equation of order two. Let $v^{(2)} = z$ and $\hat{v}^{(2)} = w$ such that z and w fulfill the boundary constraints (4). Then v and $\hat{v}$ can be rebuilt by Green's function as

$$v^{(j)}(x,t) = \int_a^b \frac{\partial^j \mathcal{G}(x,t,r)}{\partial x^j} z(r,t)dr, \quad j = 0, 1,$$
$$\hat{v}^{(j)}(x,t) = \int_a^b \frac{\partial^j \mathcal{G}(x,t,r)}{\partial x^j} w(r,t)dr, \quad j = 0, 1.$$

Let $\delta(x,t)$ be any continuously differentiable function. The operators that are necessary for the establishment of the convergence analysis are given below:

$$\mathcal{A} : \mathbb{C}[a,b] \longrightarrow \mathbb{C}[a,b] \quad \text{such that} \quad \mathcal{A}\delta = \frac{1}{\beta}(G_0\delta_t - f(x,t,G_0\delta, G_1\delta)),$$

where $G_j\delta = \int_a^b \frac{\partial^j \mathcal{G}(x,t,r)}{\partial x^j}\delta(r,t)dr$, $j = 0, 1$ are the operators from $[a, b]$ to $[a, b]$. Let $\mathcal{D}$ represent the piecewise linear interpolation operator at the points $\{(x_p, t)\}_{p=0}^{M}$. Let $\mathcal{S}$ be the following projection operator:

$$\mathcal{S} : \mathbb{C}[a, b] \longrightarrow \mathcal{R}^{M+1} \quad \text{such that} \quad \mathcal{S}\delta = [\delta(x_0, t), \delta(x_1, t), \ldots, \delta(x_M, t)]^T.$$
$$\mathcal{E} : \mathbb{C}[a, b] \longrightarrow \mathbb{C}[a, b], \quad \text{such that} \quad \mathcal{E}\delta = [\mathcal{E}_0\delta, \mathcal{E}_1\delta, \ldots, \mathcal{E}_M\delta]^T,$$

where $\mathcal{E}_p\delta = \frac{1}{\beta}(G_0\delta_t - f(x, t, G_0\delta, \mathcal{E}_p\mathcal{S}G_1\delta))$, in which $\mathcal{E}_p$ denotes the pth row of the coefficient matrix of $v_x(x, t)$. Using above definitions, (1) and (14) can be written as follows:

$$\begin{aligned}(I - \mathcal{A})z &= 0,\\ (\beth\mathcal{S} - \mathcal{E})w &= 0.\end{aligned} \tag{15}$$

Since $\beth$ is an invertible, so

$$(\mathcal{S} - \beth^{-1}\mathcal{E})w = 0.$$

Since w is a linear polynomial, therefore $\mathcal{D}\mathcal{S}w = w$ and

$$(I - \mathcal{D}\beth^{-1}\mathcal{E})w = 0. \tag{16}$$

□

Lemma 3. For the uniform partition of $[a, b]$, $\| \mathcal{D}\beth^{-1}\mathcal{E}\delta - \mathcal{A}\delta \|_\infty \to 0$ as $s \to 0$.

Proof. Th proof holds as follows:

$$\begin{aligned}\| \mathcal{D}\beth^{-1}\mathcal{E}\delta - \mathcal{A}\delta \|_\infty \leq & \| \mathcal{D}\beth^{-1}\mathcal{E}\delta - \mathcal{D}\mathcal{S}\mathcal{A}\delta \|_\infty + \| \mathcal{D}\mathcal{S}\mathcal{A}\delta - \mathcal{A}\delta \|_\infty\\ \leq & \| \mathcal{D} \|_\infty \| \beth^{-1} \|_\infty \| \mathcal{E}\delta - \beth\mathcal{S}\mathcal{A}\delta \|_\infty + \| \mathcal{D}\mathcal{S}\mathcal{A}\delta - \mathcal{A}\delta \|_\infty\\ \leq & \| \mathcal{E}\delta - \beth\mathcal{S}\mathcal{A}\delta \|_\infty + O(s^2).\end{aligned}$$

□

Theorem 2 (see [6])**.** Contemplate the curve $\mathrm{C} = (x, t, v, v_x) \in \mathbb{R}^4$, where $(x, t) \in [a, b] \times [t_0, T]$, and let $v(x, t) \in \mathbb{C}^6[a, b]$ represent the solution of the given equation (1) with the boundary constraint (4), let $f(u, y)$ be adequately smooth near v, and let the hereunder linear problem,

$$v_{xx} - \frac{\partial}{\partial y}\frac{1}{\beta}(u_t - f(u,y))v_x - \frac{\partial}{\partial u}(u_t - f(u,y))v = 0$$

with the boundary constraints (4) be distinctively solvable and acquire Green's function $\mathcal{G}(x,t,r)$. Then, there exist $\epsilon,\ \eta > 0$ (constants) such that

(I) there is no other solution $\hat{w}$ of equation (1) with boundary constraint (4) satisfying $\| v_{xx} - \hat{w}_{xx} \| < \eta$,

(II) for $s < \epsilon$, (16) has a unique spline approximate solution $W(x,.)$ in the same neighborhood of v.

(III) the Newton's method converges in the neighborhood of v for $s < \epsilon$ quadratically, which is used to solve (16).

Theorem 3. Let the presumption of Theorem 2 agree. Then the below given error bound exists:

$$\begin{aligned}
&\| v^{(j)}(x,\cdot) - \hat{v}^{(j}(x,.) \|_\infty = O(s^{4-j}), \quad j = 0,1,2.\\
&| v^{(j)}(x,\cdot) - \hat{v}^{(j)}(x,\cdot) |_{x_p} = O(s^4), \quad j = 0,1.\\
&| v^{(2)}(x,\cdot) - \hat{v}^{(2)}(x,\cdot) |_{x_p} = O(s^2).
\end{aligned}$$

Proof. Consider the equation $W^{(2)} = \hat{\mu},\ \not{\mathcal{B}}W = O(s^4)$. Then there exists a linear polynomial $\bar{w}$ by using Theorem 2, such that

$$\not{\mathcal{B}}\bar{w} = \not{\mathcal{B}}W = O(s^4), \quad \| \bar{w}^{(j)} \|_\infty = O(s^4), \quad j = 0,1.$$

Since $(W^{(2)} - \bar{w}^{(2)}) = \hat{\mu},\ \not{\mathcal{B}}(W - \bar{w}) = 0$ has a unique solution. Therefore using Theorem 2, we have

$$(I - \mathcal{D}\beth^{-1}\mathcal{E})(W^{(2)} - \bar{w}^{(2)}) = O(s^4). \tag{17}$$

Deducting (16) from (17), we have

$$(I - \mathcal{D}\beth^{-1}\mathcal{E})(W^{(2)} - \bar{w}^{(2)} - \hat{v}^{(2)}) = O(s^4).$$

Since $(I - \mathcal{D}\beth^{-1}\mathcal{E})$ is bounded,

$$\| W^{(2)} - \bar{w}^{(2)} - \hat{v}^{(2)} \|_\infty = O(s^4).$$

The equation $(W-\bar{w}-\hat{v})^{(2)}=\bar{\eta}$, $\mathbb{B}(\mathcal{W}-\overline{\exists}-\overline{\triangle})=0$ has unique solution, hence it assures the existence of Green's function such that,

$$\mid (W-\bar{w}-\hat{v})^{(j)} \mid = \int_a^b \frac{\partial^j \mathcal{G}(x,t,r)}{\partial x^j}(W^{(2)}-\bar{w}^{(2)}-\hat{v}^{(2)})dr, \quad j=0,1.$$

Thus,

$$\parallel (W-\bar{w}-\hat{v})^{(j)} \parallel_\infty = O(s^4), \quad j=0,1.$$

So,

$$\parallel (W-\hat{v})^{(j)} \parallel_\infty \leq \parallel (W-\bar{w}-\hat{v})^{(j)} \parallel_\infty + \parallel \bar{w}^{(j)} \parallel_\infty = O(s^4), \quad j=0,1,2. \quad (18)$$

Using Theorem 1, (18), and the triangular inequality implies

$$\parallel (v-\hat{v})^{(j)} \parallel_\infty \leq \parallel (v-W)^{(j)} \parallel_\infty + \parallel (W-\hat{v})^{(j)} \parallel_\infty = O(s^{4-j}), \quad j=0,1,2.$$

□

3 Time discretization

Substitute the approximate values of v, v_x, and v_{xx} in (1), which leads to an initial value problem system as follows:

$$\mathcal{Q}_1 \frac{d}{dt}\mathcal{C}(t) = \frac{1}{h^2}\mathcal{Q}_2 B\mathcal{C}(t) + \mathfrak{F}(t,\mathcal{C}(t)) \quad t\in(t_0,T), \quad (19)$$

with the initial constraint

$$\mathcal{Q}_1\mathcal{C}(t_0) = v^0, \quad (20)$$

where $\mathcal{Q}_1 = tri[1,4,1]$ is a three diagonal matrix, $\mathcal{C}(t)=[d_0(t),d_1(t),\ldots,d_M(t)]^T$, $\mathfrak{F}(t,\mathcal{C}(t))$ is a column vector with the elements $f(t,\mathcal{Q}_{1j}\mathcal{C}(t),\mathcal{Q}_{3j}\mathcal{C}(t))$, $v^0=[v(x_0,t_0),v(x_1,t_0),\ldots,v(x_M,t_0)]^T$, where $B=diag[\beta]$, $\mathcal{Q}_3=\frac{1}{3h}tri[-1,0,1]$ and the matrix $\mathcal{Q}_2$ is as follows:

$$\mathcal{Q}_2 = \frac{1}{2}\begin{bmatrix} 14 & -33 & 28 & -14 & 6 & -1 & 0 & \dots & 0 \\ 1 & 8 & -18 & 8 & 1 & 0 & 0 & \dots & 0 \\ 0 & 1 & 8 & -18 & 8 & 1 & 0 & \dots & 0 \\ \vdots & \vdots & \vdots & \vdots & \vdots & \vdots & \vdots & \dots & \vdots \\ 0 & \dots & 0 & 1 & 8 & -18 & 8 & 1 & 0 \\ 0 & \dots & 0 & 0 & 1 & 8 & -18 & 8 & 1 \\ 0 & \dots & 0 & -1 & 6 & -14 & 28 & -33 & 14 \end{bmatrix}$$

Consider the uniform division of the time domain as $\Gamma_t \equiv \{t_i\}_{i=0}^{m}$ of $[t_0, T]$ with the temporal step size $\Delta t = t^{n+1} - t^n$. Use the weighted finite difference method to discretize (19) as used in [9], with Θ as a parameter and identity matrix I, we have

$$\begin{aligned} &\mathcal{Q}_1 I \frac{\sigma_t}{\Delta t(1-\Theta\sigma_t)}\mathcal{C}^n = \frac{1}{h^2}\mathcal{Q}_2 B\mathcal{C}^n + \mathfrak{F}(t,\mathcal{C}^n), \quad n = 1, 2, \dots, \\ &\left[\mathcal{Q}_1 I - \frac{\Delta t}{h^2}(1-\Theta)\mathcal{Q}_2 BI\right]\mathcal{C}^n - \Delta t(1-\Theta)\mathfrak{F}^n \\ &\quad = \left[\mathcal{Q}_1 I + \frac{\Delta t}{h^2}\Theta\mathcal{Q}_2 BI\right]\mathcal{C}^{n-1} + \Delta t\Theta\mathfrak{F}^{n-1}, \quad n = 1, 2, \dots, \end{aligned} \tag{21}$$

with the following initial constraint:

$$\mathcal{Q}_1\mathcal{C}^0 = v^0.$$

Using the initial constraint, obtain the value of $\mathcal{C}^0$, and using (21), the value of $\mathcal{C}$ can be computed at every successive time level.

Lemma 4. Let $v(\cdot, t) \in \mathbb{C}^3[t_0, T]$ be the exact solution of (1). For $\Theta = \frac{1}{2}$, the time integration methodology has order two, and for $\Theta \in \left(\frac{1}{2}, 1\right]$, it has order one of convergence.

Proof. The proof is given in [14, Theorem 2]. □

Hence, the method is fourth-order convergent in the space and second-order convergent in the time direction for $\Theta = \frac{1}{2}$.

4 Stability analysis

Von Neumann technique is used to analyze the stability technique. Take v as a local constant $P = \max(v)$ and integrate the time domain using a weighted finite difference methodology with $\Theta = \frac{1}{2}$. We get

$$\frac{v_p^{n+1} - v_p^n}{\Delta t} + \alpha P^\sigma \left[\frac{(v_x)_p^{n+1} + (v_x)_p^n}{2}\right] = \beta \frac{(v_{xx})_p^{n+1} + (v_{xx})_p^n}{2} + \gamma(1 - P^\sigma)\left[\frac{v_p^{n+1} + v_p^n}{2}\right].$$

Expressing the $(n+1)$th level in terms of nth time level terms, we have

$$\left(\frac{1}{\Delta t} + \frac{\gamma(P^\sigma - 1)}{2}\right) v_p^{n+1} + \frac{\alpha P^\sigma}{2}(v_x)_p^{n+1} - \frac{\beta}{2}(v_{xx})_p^{n+1} = \left(\frac{1}{\Delta t} - \frac{\gamma(P^\sigma - 1)}{2}\right) v_p^n - \frac{\alpha p}{2}(v_x)_p^n + \frac{\beta}{2}(v_{xx})_p^n. \tag{22}$$

Let

$$e_1 = \frac{1}{\Delta t} + \frac{\gamma(P^\sigma - 1)}{2}; \quad q_1 = \frac{\alpha P^\sigma}{2}; \quad e_2 = \frac{1}{\Delta t} - \frac{\gamma(P^\sigma - 1)}{2}; \quad q_2 = -\frac{\alpha P^\sigma}{2}.$$

With the above substitution, (22) becomes

$$e_1 v_p^{n+1} + q_1 (v_x)_p^{n+1} - \frac{\beta}{2}(v_{xx})_p^{n+1} = e_2 v_p^n + q_2 (v_x)_p^n + \frac{\beta}{2}(v_{xx})_p^n.$$

Substituting the values v, v_x, and v_{xx} and using the improvised cubic B-splines imply

$$e_1(d_{p-1}^{n+1} + 4d_p^{n+1} + d_{p+1}^{n+1}) - \frac{3q_1}{h}(d_{p-1}^{n+1} - d_{p+1}^{n+1}) - \frac{\beta}{4h^2}(d_{p-2}^{n+1} + 8d_{p-1}^{n+1} - 18d_p^{n+1} + 8d_{p+1}^{n+1} + d_{p+2}^{n+1}) = e_2(d_{p-1}^n + 4d_p^n + d_{p+1}^n) - \frac{3q_2}{h}(d_{p-1}^n - d_{p+1}^n) \tag{23}$$

$$+ \frac{\beta}{4h^2}(d_{p-2}^n + 8d_{p-1}^n - 18d_p^n + 8d_{p+1}^n + d_{p+2}^n). \tag{24}$$

After simplifying, (24) becomes

$$-\frac{\beta}{4s^2} d_{p-2}^{n+1} + \left(e_1 - \frac{3q_1}{s} - \frac{2\beta}{s^2}\right) d_{p-1}^{n+1} + \left(4e_1 + \frac{9\beta}{2s^2}\right) d_p^{n+1}$$

$$+\left(e_1+\frac{3q_1}{s}-\frac{2\beta}{s^2}\right)d_{p+1}^{n+1}-\frac{\beta}{4s^2}d_{p+2}^{n+1}$$
$$=\frac{\beta}{4s^2}d_{p-2}^{n}+\left(e_2-\frac{3q_2}{s}+\frac{2\beta}{s^2}\right)d_{p-1}^{n}+\left(4e_2-\frac{9\beta}{2s^2}\right)d_{p}^{n}$$
$$+\left(e_2+\frac{3q_2}{s}+\frac{2\beta}{s^2}\right)d_{p+1}^{n}+\frac{\beta}{4s^2}d_{p+2}^{n}.$$

Moreover,

$$u_1d_{p-2}^{n+1}+u_2d_{p-1}^{n+1}+u_3d_{p}^{n+1}+u_4d_{p+1}^{n+1}+u_1d_{p+2}^{n+1}$$
$$=-u_1d_{p-2}^{n}+u_5d_{p-1}^{n}+u_6d_{p}^{n}+u_7d_{p+1}^{n}-u_1d_{p+2}^{n},$$

where

$$u_1=-\frac{\beta}{4s^2};\quad u_2=e_1-\frac{3q_1}{s}-\frac{2\beta}{s^2};\quad u_3=4e_1+\frac{9\beta}{2s^2};\quad u_4=e_1+\frac{3q_1}{s}-\frac{2\beta}{s^2};$$
$$u_5=e_2-\frac{3q_2}{s}+\frac{2\beta}{s^2};\quad u_6=4e_2-\frac{9\beta}{2s^2};\quad u_7=e_2+\frac{3q_2}{s}+\frac{2\beta}{s^2}.$$

Put $d_p^n=E\eta^n\exp(ip\varphi s)$, where i is the iota, E is the amplitude, and φ is the mode number. We have

$$\eta=\frac{-u_1\exp(-2i\varphi s)+u_5\exp(-i\varphi s)+u_6+u_7\exp(i\varphi s)-u_1\exp(2i\varphi s)}{u_1\exp(-2i\varphi s)+u_2\exp(-i\varphi s)+u_3+u_4\exp(i\varphi s)+u_1\exp(2i\varphi s)}$$
$$=\frac{-2u_1\cos(2\varphi s)+u_6+(u_5+u_7)\cos(\varphi s)+i(u_7-u_5)\sin(\varphi s)}{2u_1\cos(2\varphi s)+u_3+(u_2+u_4)\cos(\varphi s)+i(u_4-u_2)\sin(\varphi s)}$$
$$=\frac{A_1+iB_1}{A_2+iB_2},$$

where

$$A_1=\frac{\beta}{2s^2}\cos(2\varphi s)+\left(2e_2+\frac{4\beta}{s^2}\right)\cos(\varphi s)+4e_2-\frac{9\beta}{2s^2};$$
$$B_1=\frac{6q_2}{s}\sin(\varphi s);$$
$$A_2=-\frac{\beta}{2s^2}\cos(2\varphi s)+\left(2e_1-\frac{4\beta}{s^2}\right)\cos(\varphi s)+4e_1+\frac{9\beta}{2s^2};$$
$$B_2=\frac{6q_1}{s}\sin(\varphi s).$$

We prove $|\eta|\le 1$, that is, $A_1^2+B_1^2\le A_2^2+B_2^2$ for the stability of the technique. As $q_1=-q_2$ therefore $B_1^2=B_2^2$. Next, to prove that $A_2\ge A_1$, that is, $A_2-A_1\ge 0$, we have

$$A_2 - A_1 = -\beta\frac{\cos(2\varphi s)}{s^2} + \left(2(e_1 - e_2) - \frac{8\beta}{s^2}\right)\cos(\varphi s) + 4(e_1 - e_2) + \frac{9\beta}{s^2};$$
$$= -\frac{2\beta}{s^2}\cos^2(\varphi s) + \left(2(e_1 - e_2) - \frac{8\beta}{s^2}\right)\cos(\varphi s) + 4(e_1 - e_2) + \frac{10\beta}{s^2}. \tag{25}$$

For minimum possible value of $A_2 - A_1$, take $\cos(\varphi s) = 1$. So, $A_2 - A_1 = 6(e_1 - e_2) \geq 0$. Hence the technique is unconditionally stable.

5 Numerical examples

In this portion, the gBF problem is analyzed numerically for distinct values of α, β, γ, and σ. The numerical results are represented using tabular form as well as figures and are contrasted with the outcomes in literature as well as with its solitary wave solution. The difference in the results is shown by calculating absolute error defined as

$$\epsilon = |(v_{exact})_p^m - (v_{num})_p^m|; \quad p = 0, 1, 2, \ldots, M,$$

where $(v_{exact})_p^m$ and $(v_{num})_p^m$ are the exact and improved B-spline solutions of degree three, respectively, at the node point x_p.

The solitary wave solution of (1) is given by Wazwaz [37] as follows:

$$v(x,t) = \left[\frac{1}{2} + \frac{1}{2}\tanh\left[\frac{-\alpha\sigma}{2\beta(\sigma+1)}\left(x - \left(\frac{\alpha}{\sigma+1} + \frac{\beta\gamma(\sigma+1)}{\alpha}\right)t\right)\right]\right]^{\frac{1}{\sigma}}, \tag{26}$$

with the below given initial constraints:

$$v(x,0) = \left[\frac{1}{2} + \frac{1}{2}\tanh\left[\frac{-\alpha\sigma x}{2\beta(\sigma+1)}\right]\right]^{\frac{1}{\sigma}},$$

and the boundary constraints,

$$v(0,t) = \left[\frac{1}{2} + \frac{1}{2}\tanh\left[\frac{\alpha\sigma}{2\beta(\sigma+1)}\left(\frac{\alpha}{\sigma+1} + \frac{\beta\gamma(\sigma+1)}{\alpha}\right)t\right]\right]^{\frac{1}{\sigma}},$$
$$v(1,t) = \left[\frac{1}{2} + \frac{1}{2}\tanh\left[\frac{-\alpha\sigma}{2\beta(\sigma+1)}\left(1 - \left(\frac{\alpha}{\sigma+1} + \frac{\beta\gamma(\sigma+1)}{\alpha}\right)t\right)\right]\right]^{\frac{1}{\sigma}}.$$

Example 1. Consider the gBF equation (1) in the domain $[0,1]$ with $\alpha = 0.001$, $\beta = 1$, and $\gamma = 0.001$ as follows:

$$v_t + 0.001v^\sigma v_x = v_{xx} + 0.001v(1 - v^\sigma).$$

The solitary wave solution is given in (26). Table 1 represents the contrast of absolute error with $s = 0.1$ and $\Delta t = 0.0001$ for $t = 0.001, 0.01, 100$, and $\sigma = 1, 4$. The contrast shows that results are superior to the Adomian decomposition scheme [12], compact finite difference method [30], and exponential time differencing method [3]. The CPU time required to compute the absolute error at $t = 0.001$ is 0.043872 sec, at $t = 0.01$, it is 0.053391 sec, and at $t = 100$, it is 6.489983 sec. Figure 1 shows the resemblance between a solitary wave and the approximate solution at distinct times, and Figure 2 represents the three-dimensional surface plot of the approximate solution.

Example 2. Consider the gBF equation (1) in the domain $[0,1]$ with $\alpha = 1$, $\beta = 1$, and $\gamma = 1$ as follows:

$$v_t + v^\sigma v_x = v_{xx} + v(1 - v^\sigma).$$

Table 2 gives the absolute error at distinct times with $s = 0.1$ and $\Delta t = 0.0001$ for $\sigma = 2, 8$. This table demonstrates that results are highly accurate as compared to many existing techniques [12, 30, 3]. The CPU time required to compute the absolute error at $t = 0.0005$ is 0.037888 sec, and at $t = 0.001$, it is 0.043125 sec. The solitary wave behavior and the numerical solution are also represented by graphs. Figure 3 gives the comparison between solitary wave and approximate solution at distinct times and depicts the similarity between them. Figure 4 represents the three-dimensional surface plot of the approximate solution.

Example 3. Consider the gBF equation (1) in the domain $[0,1]$ with $\beta = 1$, $\alpha = 1$, and $\gamma = 0$ as follows:

$$v_t + v^\sigma v_x = v_{xx}.$$

The absolute error at distinct times and different spatial domain points with $h = 0.1$ and $\Delta t = 0.0001$ for $\sigma = 1, 2, 3$ is given in Table 3. Results are found to be more superior as compared to [12, 3]. The CPU time re-

quired to compute the absolute error at $t = 0.001$ is 0.042149 sec, and at $t = 2.0$, it is 6.187059 sec. Figure 5 gives the comparison between solitary wave and approximate solution at distinct times and depicts the similarity between them. Figure 6 represents the three-dimensional surface plot of the approximate solution.

Example 4. Consider the gBF equation (1) in the domain $[0, 1]$ with $\alpha = 0.1$, $\beta = 1$, and $\gamma = -0.0025$ as follows:

$$v_t + 0.1v^\sigma v_x = v_{xx} - 0.0025v(1 - v^\sigma).$$

Table 4 shows the absolute error with space step size $s = 0.1$ and time step size $\Delta t = 0.0001$ for $\sigma = 2, 4$, and 8. From the comparison it is clear that the results with the proposed methodology are superior to many other existing techniques used in [30], [3]. The CPU time required to compute the absolute error at $t = 0.1$ is 0.627901 sec, at $t = 0.5$, it is 0.237965 sec, and at $t = 2.0$, it is 6.489983 sec. Figure 7 demonstrates the resemblance between solitary wave and numerical solution at distinct times, and Figure 8 represents the three-dimensional surface plot of the approximate solution.

6 Conclusion

The gBF problem has been investigated using an innovative approach employing a three-degree spline collocation methodology. Through enhancements made to standard splines, we have achieved significantly improved accuracy in the approximate solution, accompanied by a notable reduction in absolute error. Our improvised methodology demonstrates a convergence order of four in the spatial domain and two in the temporal domain. These results surpass the performance of various established techniques, including the compact finite difference technique, exponential time differencing method, and Adomian decomposition scheme, among others. Furthermore, our method showcases computational efficiency across a range of relevant examples.

Table 1: Absolute error comparison of Example 1 with $\alpha = 0.001$, $s = 0.1$ and $\Delta t = 0.0001$, $\gamma = 0.001$

		$\sigma = 1$					$\sigma = 4$			
t	x	ICSCM	[12]	[30]	[3]	[36]	ICSCM	[30]	[3]	[36]
0.001	0.1	5.21E-15	1.940e-6	1.010e-7	1.150e-8	2.50e-8	1.77e-15	1.75e-8	7.71e-9	4.20e-8
	0.5	1.66E-16	1.940e-6	1.040e-7	3.070e-13	2.50e-8	3.33e-16	1.75e-8	2.07e-13	4.20e-8
	0.9	5.55E-17	1.940e-6	1.010e-7	1.150e-8	2.50e-8	1.11e-16	1.75e-8	7.71e-9	4.20e-8
0.010	0.1	3.25E-14	1.940e-5	7.530e-7	6.020e-8	2.50e-8	5.66e-15	1.27e-6	4.05e-8	4.20e-8
	0.5	1.11E-16	1.940e-5	1.040e-6	8.960e-13	2.50e-8	6.66e-16	1.75e-6	5.56e-13	4.20e-8
	0.9	5.55E-17	1.940e-5	7.530e-7	6.020e-8	2.50e-8	1.11e-16	1.27e-6	4.05e-8	4.20e-8
100	0.1	1.52E-14	-	7.530e-7	1.010e-7	2.50e-8	4.42e-14	-	5.73e-8	4.20e-8
	0.5	2.59E-14	-	1.040e-6	1.500e-11	2.50e-8	1.03e-14	-	3.51e-12	4.20e-8
	0.9	5.55E-15	-	7.530e-7	1.010e-7	2.50e-8	1.66e-15	-	5.73e-8	4.20e-8

Table 2: Absolute error comparison of Example 2 with $\alpha = 1$, $s = 0.1$, $\gamma = 1$, and $\Delta t = 0.0001$

		$\sigma = 2$					$\sigma = 8$			
t	x	ICSCM	[12]	[30]	[3]	[36]	ICSCM	[30]	[3]	[36]
0.0005	0.1	5.15e-11	1.40e-3	7.62e-5	5.67e-6	3.98e-5	2.4139e-9	1.02e-4	2.44e-6	5.16e-5
	0.5	1.35e-12	1.35e-3	9.14e-5	5.75e-9	4.15e-5	5.5742e-12	1.37e-4	1.82e-10	6.08e-5
	0.9	2.08e-11	1.28e-3	1.02e-4	5.95e-6	4.22e-5	1.6251e-9	1.69e-4	3.15e-6	6.91e-5
0.0010	0.1	1.03e-11	2.80e-3	1.50e-4	1.08e-5	3.97e-5	3.4743e-10	2.00e-4	4.65e-6	5.15e-5
	0.5	2.05e-12	2.69e-3	1.83e-4	1.15e-8	4.11e-5	2.8770e-11	2.74e-4	4.02e-10	6.09e-5
	0.9	2.40e-12	2.55e-3	2.00e-4	1.14e-5	4.16e-5	2.3789e-10	3.31e-4	6.00e-6	6.94e-5

Table 3: Absolute error comparison of Example 3 with $\alpha = 1$, $s = 0.1$, $\gamma = 0$, and $\Delta t = 0.0001$

σ	t	x	ICSCM	Ismail [12]	Bratsos [3]	t	ICSCM	Bratsos [3]
1	2	0.1	2.5046e-10	6.43e-5	9.82e-5	20	8.6084e-12	2.65e-6
		0.5	3.7073e-10	6.07e-5	1.45e-5		1.4161e-11	1.45e-7
		0.9	3.2483e-10	4.75e-5	9.29e-5		1.3391e-11	4.05e-6
2	2	0.1	1.1286e-10	1.19e-5	8.34e-5	20	1.1906e-12	2.96e-6
		0.5	6.8167e-10	1.50e-5	4.19e-6		5.1861e-11	7.14e-7
		0.9	1.2681e-10	1.44e-5	9.48e-5		1.2913e-11	5.66e-6
3	0.001	0.1	1.1582e-9	4.44e-4	9.10e-6	10	5.0684e-10	2.46e-5
		0.5	8.0534e-12	1.85e-3	6.75e-9		6.6355e-10	5.11e-6
		0.9	6.8473e-10	9.05e-4	1.09e-5		5.5408e-10	4.35e-5

Table 4: Absolute error comparison of Example 4 with $\alpha = 0.1$, $s = 0.1$, $\gamma = -0.0025$, and $\Delta t = 0.0001$

		$\sigma = 2$			$\sigma = 4$			$\sigma = 8$		
t	x	ICSCM	[30]	[3]	ICSCM	[30]	[3]	ICSCM	[30]	[3]
0.1	0.1	5.040e-14	1.210e-5	9.470e-6	4.787e-13	1.340e-5	6.760e-6	3.140e-12	1.470e-5	4.090e-6
	0.5	1.221e-15	2.900e-5	2.740e-8	6.994e-15	3.490e-5	1.030e-8	6.362e-14	3.830e-5	1.840e-8
	0.9	3.197e-14	1.540e-5	9.570e-6	4.969e-13	1.390e-5	6.920e-8	3.041e-12	1.530e-5	4.240e-6
0.5	0.1	4.252e-14	1.670e-5	9.580e-6	5.746e-13	2.000e-5	6.830e-6	3.148e-12	2.200e-5	4.140e-6
	0.5	2.109e-15	4.690e-5	5.180e-8	2.420e-14	5.640e-5	1.930e-8	2.331e-14	6.220e-5	3.470e-8
	0.9	3.442e-14	1.710e-5	9.660e-6	5.044e-13	2.070e-5	7.010e-6	3.068e-12	2.280e-5	4.300e-6
2.0	0.1	6.051e-14	-	9.590e-6	5.369e-13	-	6.860e-6	3.116e-12	-	4.200e-6
	0.5	3.553e-15	-	5.260e-8	5.329e-15	-	1.890e-8	3.919e-14	-	3.450e-8
	0.9	3.186e-14	-	9.670e-6	4.998e-13	-	7.040e-6	3.069e-12	-	4.350e-6

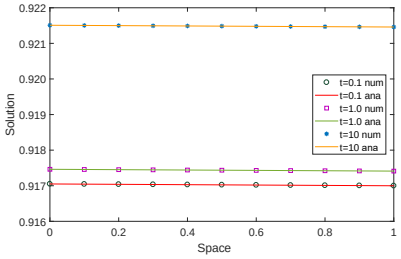


Figure 1: Solution of Example 1 at distinct times with $s = 0.1$, $\Delta t = 0.0001$, and $\sigma = 8$.

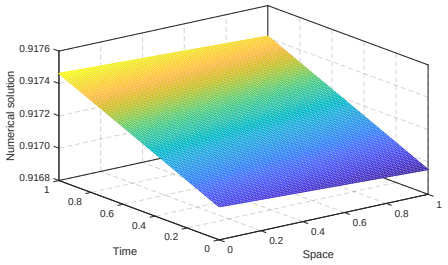


Figure 2: Three-dimensional representation of numerical solution of Example 1 with $s = 0.01$, $\Delta t = 0.001$, and $\sigma = 8$.

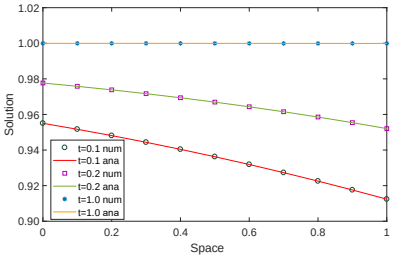


Figure 3: Solution of Example 2 at distinct times with $s = 0.1$, $\Delta t = 0.0001$, and $\sigma = 8$.

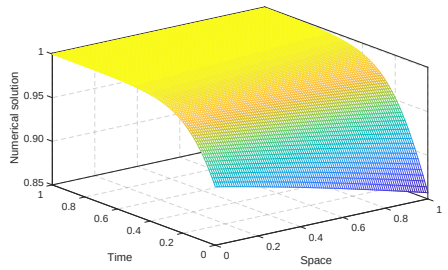


Figure 4: Three-dimensional representation of numerical solution of Example 2 with $s = 0.01$, $\Delta t = 0.001$, and $\sigma = 8$.

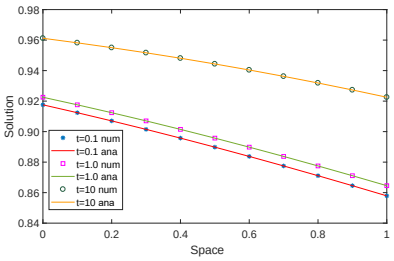


Figure 5: Solution of Example 3 at distinct times with $s = 0.1$, $\Delta t = 0.0001$, and $\sigma = 8$.

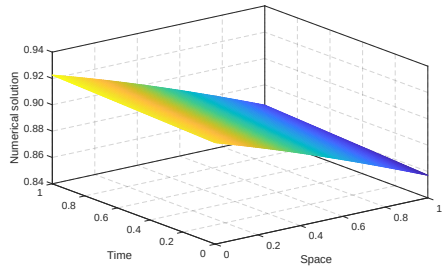


Figure 6: Three-dimensional representation of numerical solution of Example 3 with $s = 0.01$, $\Delta t = 0.001$, and $\sigma = 8$.

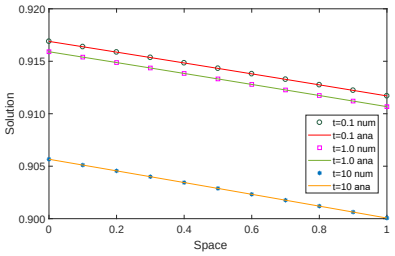


Figure 7: Solution of Example 4 at distinct times with $s = 0.1$, $\Delta t = 0.0001$, and $\sigma = 8$.

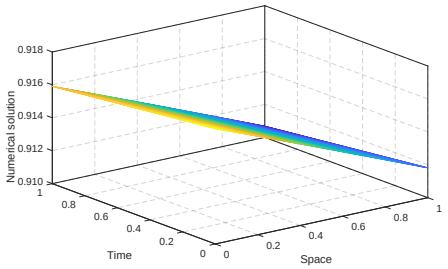


Figure 8: Three-dimensional representation of numerical solution of Example 4 with $s = 0.01$, $\Delta t = 0.001$, and $\sigma = 8$.

References

[1] Al-Rozbayani, A.M. and Al-Hayalie, K.A. *Numerical solution of Burger's-Fisher equation in one-dimensional using finite differences methods*, integration 9 (2018), 10.

[2] Bratsos, A.G. *A fourth order improved numerical scheme for the generalized Burgers-Huxley equation*, Am. J. Comput. Math. 1(03) (2011), 152–158.

[3] Bratsos, A.G. and Khaliq, A.Q.M. *An exponential time differencing method of lines for Burgers-Fisher and coupled Burgers equations*, J. Comput. Appl. Math. 356 (2019), 182–197.

[4] Chandraker, V., Awasthi, A. and Jayaraj, S. *Numerical treatment of Burger-Fisher equation*, Procedia Technology 25 (2016), 1217–1225.

[5] Daniel, J.W. and Swartz, B.K. *Extrapolated collocation for two-point boundary-value problems using cubic splines*, IMA J. Appl. Math. 16(2) (1975), 161–174.

[6] De Boor, C. and Swartz, B. *Collocation at Gaussian points*, SIAM J. Numer. Anal. 10(4) (1973), 582–606.

[7] Fan, E. *Extended tanh-function method and its applications to nonlinear equations*, Phys. Lett. A 277(4-5) (2000), 212–218.

[8] Ghasemi, M. *A new superconvergent method for systems of nonlinear singular boundary value problems*, Int. J. Comput. Math. 90(5) (2013), 955–977.

[9] Ghasemi, M. *An efficient algorithm based on extrapolation for the solution of nonlinear parabolic equations*, Int. J. Nonlinear Sci. Numer. Simul. 19(1) (2018), 37–51.

[10] Golbabai, A. and Javidi, M. *A spectral domain decomposition approach for the generalized Burgers–Fisher equation*, Chaos Solitons Fractals 39(1) (2009), 385–392.

[11] Hepson, O.E. *An extended cubic B–spline finite element method for solving generalized Burgers–Fisher equation*, arXiv preprint arXiv:1612.03333 (2016).

[12] Ismail, H.N., Raslan, K. and Abd Rabboh, A.A. *Adomian decomposition method for Burger's–Huxley and Burger's–Fisher equations*, Appl. Math. Comput. 159(1) (2004), 291–301.

[13] Javidi, M. *Spectral collocation method for the solution of the generalized Burger–Fisher equation*, Appl. Math. Comput. 174(1) (2006), 345–352.

[14] Kadalbajoo, M.K., Tripathi, L.P. and Kumar, A. *A cubic B-spline collocation method for a numerical solution of the generalized Black–Scholes equation*, Math. Comput. Model. 55(3-4) (2012), 1483–1505.

[15] Kaya, D. and El-Sayed, S.M. *A numerical simulation and explicit solutions of the generalized Burgers-Fisher equation*, Appl. Math. Comput. 152(2) (2004), 403–413.

[16] Lighthill, M.J. *In surveys in mechanics*, Cambridge Cambridge University Press, Viscosity effects in sound waves of finite amplitude, 1956.

[17] Lu, B.Q., Xiu, B.Z., Pang, Z.L. and Jiang, X.F. *Exact traveling wave solution of one class of nonlinear diffusion equations*, Phys. Lett. A, 175(2) (1993), 113–115.

[18] Malik, S.A., Qureshi, I.M., Amir, M., Malik, A.N. and Haq, I. *Numerical solution to generalized Burgers-Fisher equation using exp-function method hybridized with heuristic computation*, PloS one 10(3) (2015), e0121728.

[19] Mickens, R.E. and Gumel, A. *Construction and analysis of a non-standard finite difference scheme for the Burgers-Fisher equation*, J. Sound. Vib. 257(4) (2002), 791–797.

[20] Mittal, R.C. and Tripathi, A. *Numerical solutions of generalized Burgers–Fisher and generalized Burgers–Huxley equations using collocation of cubic B-splines*, Int. J. Comput. Math. 92(5) (2015), 1053–1077.

[21] Moghimi, M. and Hejazi, F.S. *Variational iteration method for solving generalized Burger–Fisher and Burger equations*, Chaos Solitons Fractals 33(5) (2007), 1756–1761.

[22] Mohammadi, R. *Spline solution of the generalized Burgers-Fisher equation*, Appl. Anal. 91(12) (2012), 2189–2215.

[23] Murray, J.D. *Mathematical biology*, Berlin Springer-Verlag, 1989.

[24] Prenter, P.M. *Splines and variational methods*, New York Wiley-interscience publication, 1975.

[25] Russell, R.D. and Shampine, L.F. *A collocation method for boundary value problems*, Numer. Math. 19 (1971), 1–28.

[26] Sachdev, P.L. *Self-similarity and beyond exact solutions of nonlinear problems*, New York, Chapman & Hall/CRC, 2000.

[27] Saeed, U. and Gilani, K. *CAS wavelet quasi-linearization technique for the generalized Burger–Fisher equation*, Math. Sci. 12(1) (2018), 61–69.

[28] Sangwan, V. and Kaur, B. *An exponentially fitted numerical technique for singularly perturbed Burgers-Fisher equation on a layer adapted mesh*, Int. J. Comput. Math. 96(7) (2019), 1502–1513.

[29] Sari, M. *Differential quadrature solutions of the generalized Burgers-Fisher equation with a strong stability preserving high-order time integration*, Math. Comput. Appl. 16(2) (2011), 477–486.

[30] Sari, M., Gürarslan, G. and Dağ, İ. *A compact finite difference method for the solution of the generalized Burgers–Fisher equation*, Numer. Methods Partial Differ. Equ. 26(1) (2010), 125–134.

[31] Shallu and Kukreja, V.K. *An improvised collocation algorithm with specific end conditions for solving modified Burgers equation*, Numer. Methods Partial Differ. Equ. 37(1) (2021), 874–896.

[32] Shallu and Kukreja, V.K. *Analysis of RLW and MRLW equation using an improvised collocation technique with SSP-RK43 scheme*, Wave Motion 105 (2021), 102761.

[33] Shallu and Kukreja, V.K. *An improvised collocation algorithm to solve generalized Burgers'-Huxley equation*, Arab. J. Math. 11(2) (2022), 379–396.

[34] Shallu, Kumari, A. and Kukreja, V.K. *An improved extrapolated collocation technique for singularly perturbed problems using cubic B-spline functions*, Mediterr. J. Math. 18(4) (2021), 1–29.

[35] Tatari, M., Sepehrian, B. and Alibakhshi, M. *New implementation of radial basis functions for solving Burgers-Fisher equation*, Numer. Methods Partial Differ. Equ. 28(1) (2012), 248–262.

[36] Verma, A.K. and Kayenat, S. *On the stability of Micken's type NSFD schemes for generalized Burgers Fisher equation*, J. Differ. Equ. Appl. 25(12) (2019), 1706–1737.

[37] Wazwaz, A.M. *The tanh method for generalized forms of nonlinear heat conduction and Burgers–Fisher equations*, Appl. Math. Comput. 169(1) (2005), 321–338.

[38] Wazwaz, A.M. *Analytic study on Burgers, Fisher, Huxley equations and combined forms of these equations*, Appl. Math. Comput. 195(2) (2008), 754–761.

[39] Zhao, T., Li, C., Zang, Z. and Wu, Y. *Chebyshev–Legendre pseudo-spectral method for the generalised Burgers–Fisher equation*, Appl. Math. Model. 36(3) (2012), 1046–1056.

[40] Zhu, C.G. and Kang, W.S. *Numerical solution of Burgers–Fisher equation by cubic B-spline quasi-interpolation*, Appl. Math. Comput. 216(9) (2010), 2679–2686.

Iranian Journal of Numerical Analysis and Optimization
Vol. 14, No. 3, 2024, pp 762–795
https://doi.org/10.22067/ijnao.2024.86894.1393
https://ijnao.um.ac.ir/

Research Article

Uniformly convergent numerical solution for caputo fractional order singularly perturbed delay differential equation using extended cubic B-spline collocation scheme

N.A. Endrie*, and G.F. Duressa

Abstract

This article presents a parameter uniform convergence numerical scheme for solving time fractional order singularly perturbed parabolic convection-diffusion differential equations with a delay. We give a priori bounds on the exact solution and its derivatives obtained through the problem's asymptotic analysis. The Euler's method on a uniform mesh in the time direction

*Corresponding author

Received 16 February 2024; revised 20 May 2024; accepted 27 May 2024

Nuru Ahmed Endrie
Department of Mathematics,College of Natural and Computational Science, Arba Minch University, Arba Minch, Ethiopia. e-mail: nuruahmed222@gmail.com

Gemechis File Duressa
Department of Mathematics, College of Natural and Computational Science, Jimma University, Jimma, Ethiopia. e-mail: gammeef@gmail.com

How to cite this article
Endrie, N.A. and Duressa, G.F., Uniformly convergent numerical solution for caputo fractional order singularly perturbed delay differential equation using extended cubic B-spline collocation scheme. *Iran. J. Numer. Anal. Optim.*, 2024; 14(3): 762-795. https://doi.org/10.22067/ijnao.2024.86894.1393

and the extended cubic B-spline method with a fitted operator on a uniform mesh in the spatial direction is used to discretize the problem. The fitting factor is introduced for the term containing the singular perturbation parameter, and it is obtained from the zeroth-order asymptotic expansion of the exact solution. The ordinary B-splines are extended into the extended B-splines. Utilizing the optimization technique, the value of μ (free parameter, when the free parameter μ tends to zero the extended cubic B-spline reduced to convectional cubic B-spline functions) is determined. It is also demonstrated that this method is better than some existing methods in the literature.

AMS subject classifications (2020): Primary 65L11; Secondary 65N12.

Keywords: Singularly perturbed problem; Fractional derivative; Artificial viscosity; Delay differential equation.

1 Introduction

In this work, we consider the singularly perturbed parabolic delay differential equation of fractional order in time,

$$\begin{aligned}\mathfrak{L}y(x,t) \equiv D_t^{\gamma} y(x,t) - \varepsilon \frac{\partial^2 y(x,t)}{\partial x^2} + q(x)\frac{\partial y(x,t)}{\partial x} + r(x,t)y(x,t) \\ = -s(x,t)y(x,t-\delta) + f(x,t), \qquad (x,t) \in \Omega = (0,1)\times(0,\mathfrak{T}], \quad (1)\end{aligned}$$

with

$$\begin{cases} y(x,t) = \varphi_b(x,t), & \text{for } (x,t) \in [0,1]\times[-\delta,0], \\ y(0,t) = \varphi_l(t), y(1,t) = \varphi_r(t), & \text{for } \ t \in (0,\mathfrak{T}), \end{cases} \quad (2)$$

where D_t^{γ} is the Caputo fractional derivative of order $0 < \gamma < 1$, δ is delay parameter, and $0 < \varepsilon \ll 1$ is the singular perturbation parameter. For the domain $\bar{\Omega} = [0,1]\times[0,\mathfrak{T}]$, if

$$q(x) \geq \beta > 0, \quad r(x,t) \geq 0, \quad s(x,t) \geq \alpha > 0,$$

are bounded and smooth functions, then initial data and boundary conditions are also smooth and bounded in their respective domains. The solution of

model problem (1) has a boundary layer of regular type at $x = 1$ with a width of $\mathcal{O}(\varepsilon)$.

Parameter-dependent differential equations, whose solution behavior depends on the magnitude of the parameters, are used to model many physical and biological phenomena. If the highest-order derivative of a differential equation is multiplied by a small positive parameter, $\varepsilon(0 < \varepsilon < 1)$, then the differential equation is said to be singularly perturbed. Such issues arise in modeling of reaction-diffusion processes, chemical reactor theory, aerodynamics, elasticity, quantum mechanics, plasma dynamics, and many other related domains [3].

Fractional calculus has an origin as old as classical calculus, although it was not used for a very long period to solve scientific and engineering problems. Indeed, fractional calculus started attracting the attention of scientists and researchers in recent decades due to its numerous applications [6, 35]. Noninteger derivatives were first introduced by Leibnitz in 1695, as far as the authors can tell.

Derivatives of arbitrary order were mentioned by Euler and Fourier, but no examples or applications were provided. The honor of being the first to apply in real-world scenarios belongs to Niels Henrik Abel [1] in 1823. However, as stated in [6], fractional calculus began to be essential by Riemann and Liouville. Fractional-order differential equations are used to model a wide range of real-world phenomena, including protein dynamics, dielectric relaxation phenomena in polymeric materials, visco-elastic behavior, transport of passive tracers carried by fluid flow in a porous medium in groundwater hydrology, transport dynamics in systems subject to anomalous diffusion, and long-term memory in financial time series [21, 15].

Singularly perturbed delay differential equations (SPDDEs) are employed to model physical problems that evolve based on both their current condition and history. To make a model more realistic, it may be important to represent former system states in addition to the current state. Delay differential equations (DDEs) are useful for describing time-dependent phenomena that rely on a past state [17]. Because delay differential equations have so many applications in the fields, including bio-sciences, control theory, economics, material science, medicine, robotics, and more, there has been a major rise

in interest in studying problems during the past several decades. The field of delay differential equations theory is extensive, with notable works including in [7, 11, 12, 33, 22, 19, 22, 23, 24], and there are various real-world examples of delay differential equations in the works by Nelson and Perelson [32], Villasana and Radunskaya [41], and Zhao [44]. Singularly perturbed problem (SPP) solutions are not smooth and contain boundary layer-related singularities. When the perturbation parameter (ε) and mesh length are lowered, even advanced numerical algorithms do not perform consistently well. The results of classical numerical methods on uniform meshes fail to provide a reasonably accurate approximate solution of the exact solution, and the truncation error becomes unbounded as the singular perturbation parameter tends to zero unless a large number of mesh points are used in the approximation process [13]. However, this highlights the numerical method's computational inefficiency. When the number of mesh points grows, the resulting algebraic system of equations may become ill-conditioned. The shortcoming encourages the creation of a suitable numerical approach whose accuracy is independent of the perturbation parameter, highlighting the key advantage of the proposed method [16].

Xu [43] has proposed the extended cubic B-spline, a generalization of the B-spline. In [42] investigation, the three extended B-splines with degrees 4, 5, and 6 were provided. To modify the shape of the cubic B-spline curve for extended B-splines, a free parameter is added to the cubic B-spline base functions. The degree of the piecewise polynomials is raised, and a one-free parameter is included, but the continuity of the extended cubic B-splines stays in the order of 3. This encourages us to develop an extended cubic B-spline trial function as part of a numerical technique [10]. The spline-based approach has gained a lot of popularity these days among the various algorithms for solving SPDDEs. Daba and Duressa [8] gave a uniform convergent numerical method for the singularly perturbed parabolic convection-diffusion equation with a small delay and advance parameter in the spatial variable of the reaction term using an extended cubic B-spline approach. Additionally, they [9] suggested a uniformly convergent numerical solution based on a cubic B-spline and uniform mesh for this problem. Kumar and Kadalbajoo [26] suggested a parameter-uniform numerical method for the problem using a

cubic B-spline on a Shishkin mesh. Kumar and Kadalbajoo [25] and Negero and Duressa [31] developed a parameter uniform convergent method to solve time-dependent singularly perturbed delay parabolic convection-diffusion initial boundary value problems, respectively, using the cubic B-spline collocation method on a piecewise uniform Shishkine mesh and a uniform mesh. In [18], they devised a fitted extended cubic B-spline collocation method to solve singularly perturbed parabolic equations with nonsmooth convection coefficient and discontinuous source terms.

The numerical solution of time-fractional singularly perturbed ordinary differential equations (ODEs) and partial differential equations (PDEs) has not received much attention in the literature. Bijura [5] presented fractionally ordered nonlinear SPPs using higher-order asymptotic solutions. Using the finite element method, Roop [36] developed the numerical solution of fractional ODEs. Qasem and Muhammed [2] used the Pade approximation to estimate the solution of fractional-order nonlinear singularly perturbed two-point boundary-value problems. The matched asymptotic scheme for fractional-order boundary layer problems has been expanded in [4]. Sayevand and Pichaghchi[39] tackled the fractional order boundary value problem by presenting a method to solve singularly perturbed ODEs. Based on the characteristics of a local fractional derivative, they defined the local fractional derivative and expanded the matching asymptotic expansion approach. A linear B-spline operational matrix of fractional derivatives for singularly perturbed ODEs and PDEs has been proposed in [38]. Sahoo and Vikas [37] devised a finite difference method to address a class of time-fractional singularly perturbed convection-diffusion problems. Kumar and Vigo-Aguiar [27] constructed by discretizing time domains using uniform step size and piece-wise-uniform Shishkin meshes for space domains in the study of delay parabolic and time-fractional SPDEs.

To most of our understanding, there is only one paper in the literature that discusses the construction and analysis of a numerical scheme for the class of SPFODDEs under review [27]. This article aims to present and analyze implicit Euler's scheme for time discretization and spatial discretization based on the extended cubic B-spline method by introducing fitting factors. These

methods yield robust numerical results while preserving important features of the corresponding continuous problems.

This article has been organized into the following sections as follows: The preliminary notions are defined in section 2. In Section 3, the formulation of the continuous problem is discussed along with an analytical solution and an analysis of the derivative behavior using defined bounds. We analyze implicit Euler's scheme for time discretization and spatial discretization based on the extended cubic B-spline method by introducing the fitting factor presented in section 4. Section 5 discusses the uniform convergence analysis of the approach. The numerical experiments carried out to confirm theoretical findings and show the method's accuracy are described in detail in Section 6. An overview of the paper's main conclusions is given in the concluding section.

2 Preliminaries

The definitions and tools needed for this study are provided in this section (see [28, 29, 27]).

Definition 1 (Singularly-perturbed problem)**.** If the highest-order derivative of a differential equation is multiplied by a small parameter ε, where ε is the perturbation parameter and $0 < \varepsilon << 1$, the differential equation is considered singularly perturbed.

Definition 2 (Gamma function)**.** If z is a complex number with a nonnegative real part, then the gamma function ($\mathbb{R}(z) > 0$) is given by the following definition:

$$\Gamma(z) = \int_0^\infty x^{z-1} e^{-x} dx. \tag{3}$$

Definition 3 (Caputo fractional derivative)**.** For $m \in \mathbb{N}$ and $\gamma \in (m-1, m)$, the Caputo fractional derivative of a function $g(t)$ with lower limit zero is defined as

$$D_0^\gamma g(t) = \frac{1}{\Gamma(m-\gamma)} \int_0^t \frac{g^{(m)}(s)}{(t-s)^{\gamma-m+1}} ds. \tag{4}$$

Definition 4. The function $v(x,t)$ can be defined as the γ-order differentiation, with lower bound zero, of a function $m \in \mathbb{N}$ with regard to $t$in the

Caputo sense, as follows:

$$\frac{\partial^{\gamma} v(x,t)}{\partial t^{\gamma}} = \begin{cases} \frac{1}{\Gamma(m-\gamma)} \int_0^t \frac{\partial^m v(x,s)}{\partial s^m} \frac{1}{(t-s)^{\gamma-m+1}} ds & \text{if } \gamma \in (m-1, m), \\ \frac{\partial^m v(x,t)}{\partial t^m} & \text{if } \gamma = m. \end{cases} \tag{5}$$

3 Properties of continuous problem

Assuming sufficiently smoothness of $\varphi_l(t), \varphi_r(t)$, and $\varphi_b(x,t)$ and satisfying the following compatibility conditions at the corner points $(0,0)$, $(1,0)$, and $(0,-\delta)$ as well as the delay term, the existence and uniqueness of the solution of (1)–(2) can be established. Let

$$\begin{cases} \varphi_b(0,0) = \varphi_l(0), \\ \varphi_b(1,0) = \varphi_r(0), \end{cases} \tag{6}$$

and

$$\begin{cases} \left.\frac{d^{\gamma}\varphi_l}{d^{\gamma}t}\right|_{t=0} - \varepsilon \left.\frac{\partial^2 \varphi_b}{\partial x^2}\right|_{(0,0)} + q(0) \left.\frac{\partial \varphi_b}{\partial x}\right|_{(0,0)} + r(0,0)\varphi_b(0,0) \\ \qquad = -s(0,0)\varphi_b(0,-\delta) + f(0,0), \\ \left.\frac{d^{\gamma}\varphi_l}{d^{\gamma}t}\right|_{t=0} - \varepsilon \left.\frac{\partial^2 \varphi_b}{\partial x^2}\right|_{(1,0)} + q(1) \left.\frac{\partial \varphi_b}{\partial x}\right|_{(1,0)} + r(1,0)\varphi_b(1,0) \\ \qquad = -s(1,0)\varphi_b(0,-\delta) + f(1,0). \end{cases} \tag{7}$$

The reduced problem obtained by putting $\varepsilon = 0$ in (1) is

$$D_t^{\gamma} y(x,t) + q(x)\frac{\partial y(x,t)}{\partial x} + r(x,t)y(x,t) = -s(x,t)y(x,t-\delta) + f(x,t), \quad (x,t) \in \Omega. \tag{8}$$

This is a hyperbolic partial differential equation of first order. Because (8) contains first-order derivatives, the reduced problem is not required to meet the boundary conditions. Thus, the solution to the problem in (1) displays a boundary layer.

Now, we will show that the operator $\mathfrak{L}$ satisfies the maximum principle.

Lemma 1 (Continuous maximum principle)**.** Consider the function $\phi(x,t) \in C^2(\Omega) \cap C^0(\bar{\Omega})$, with $\mathfrak{L}\phi(x,t) \geq 0$ in Ω and $\phi(x,t) \geq 0$, for all $(x,t) \in \Lambda = \{0\} \times (0,\mathfrak{T}] \cup \{1\} \times (0,\mathfrak{T}] \cup [0,1] \times [-\delta, 0]$. Then $\phi(x,t) \geq 0$, for all $(x,t) \in \bar{\Omega}$.

Proof. Let us assume that there exists $(\varsigma, \iota) \in \bar{\Omega}$ with

$$\phi(\varsigma,\iota) = \min_{(x,t)\in\bar{\Omega}} \phi(x,t), \quad \text{and} \quad \phi(\varsigma,\iota) < 0.$$

Based on this assumptions, one may confirm that $(\varsigma,\iota) \notin \Lambda$, which implies that $(\varsigma,\iota) \in \Omega$. Using the operator $\mathfrak{L}$ on $\phi(x,t)$, we get

$$\mathfrak{L}\phi(x,t) = D_t^{\gamma}\phi(x,t) - \varepsilon\phi_{xx}(x,t) + q(x)\phi_x(x,t) + r(x,t)\phi(x,t).$$

At the point of minimum (ς,ι), we obtain

$$\mathfrak{L}\psi((\varsigma,\iota)) = D_t^{\gamma}\phi(\varsigma,\iota) - \varepsilon\phi_{xx}(\varsigma,\iota) + q(\varsigma)\phi_x(\varsigma,\iota) + r(\varsigma,\iota)\psi(\varsigma,\iota).$$

The function ϕ has minimum at the point (ς,ι), so $D_t^{\gamma}\phi \geq 0$, $\phi_x = 0$, $\phi_{xx} \geq 0$ at point (ς,ι), and $r(\varsigma,\iota) \geq 0$ for $(\varsigma,\iota) \in \Omega$. Therefore, we have

$$\mathfrak{L}\psi(\varsigma,\iota) < 0.$$

This contradicts our assumption $\mathfrak{L}\phi(x,t)$ in Ω.

Thus, we conclude that $\phi(x,t) \geq 0$, for all $(x,t) \in \bar{\Omega}$. □

Lemma 2. The differential equation (1)–(2) has a solution $y(x,t)$ that satisfies this estimate:

$$|y(x,t) - \varphi_b(x,0)| \leq Ct, \qquad (x,t) \in \bar{\Omega},$$

in which C is a constant that does not depend on ε.

Proof. See reference [30]. □

Lemma 3. With its initial and boundary conditions in (2), the solution to problem (1) is bounded as follows:

$$|y(x,t)| \leq C, \quad \text{for all } (x,t) \in \bar{\Omega}. \tag{9}$$

Proof. From Lemma 2

$$\begin{aligned}|y(x,t)| &= |y(x,t) - \varphi_b(x,0) + \varphi_b(x,0)| \\ &\leq |y(x,t) - \varphi_b(x,0)| + |\varphi_b(x,0)| \\ &\leq Ct + |\varphi_b(x,0)| \\ &\leq Ct + C \\ &\leq C \qquad \text{since } t \in (0,\mathfrak{T}],\ t \text{ is bounded.}\end{aligned}$$

□

4 Numerical schemes

We are going to develop the numerical scheme in this section as well. After discretizing the temporal derivative using implicit Euler's scheme, we discretize the spatial derivative based on the extended cubic B-spline approach by applying a fitting factor on a uniform mesh to solve the resulting system of ordinary differential equations.

4.1 Temporal discretization

We first partition the time domain $[0,\mathfrak{T}]$ into M_τ subintervals having uniform step size $\tau = \mathfrak{T}/M_\tau$. We chose M_τ so that for some positive integer $k \in (0, M_\tau)$, $\delta = k\tau$ needs to be a mesh point. A collection of all mesh points in the time direction is represented by the set Ω^{M_τ}; we then have $\Omega^{M_\tau} = \{t_0 = 0 < t_1 < t_2 < \cdots < t_k = \delta < t_{M_\tau - 1} < t_{M_\tau} = \mathfrak{T}\}$. We employ $\Omega_\delta^{M_\tau}$ as the collection of all mesh points between zero and $-\delta$; $\Omega_\delta^{M_\tau} = \{t_{-k} = -\delta < t_{-k+1} < \cdots < t_{-1} < t_0 = 0\}$.

According to Definition 4,

$$\begin{aligned}z(x,t_{j+1}) &= \frac{\partial^\gamma y(x,t_{j+1})}{\partial t^\gamma} \\ &= \frac{1}{\Gamma(1-\gamma)} \int_0^{t_{j+1}} \frac{\partial y(x,t_{j+1})}{\partial t} (t_{j+1} - \eta)^{-\gamma} d\eta \\ &= \frac{\tau)^{-\gamma}}{\Gamma(2-\gamma)} \sum_{i=0}^{j} B_i \left(y(x,t_{j-i+1}) - y(x,t_{j-i}) \right) + \mathcal{R}_\tau\end{aligned}$$

$$= \sigma \sum_{i=0}^{j} B_i \left(y(x, t_{j-i+1}) - y(x, t_{j-i}) \right) + \mathcal{R}_\tau$$

$$= \sigma y(x, t_{j+1}) - \sigma y(x, t_{j+1}) \sigma \sum_{i=1}^{j} B_i \left(y(x, t_{j-i+1}) - y(x, t_{j-i}) \right) + \mathcal{R}_\tau,$$

where

$$\mathcal{R}_\tau = O(\tau) \int_0^{t_{j+1}} (t_{j+1} - \eta)^{-\gamma} d\eta \quad \text{is} \quad \text{the} \quad \text{truncation} \quad \text{error},$$

and

$$\sigma = \frac{\tau)^{-\gamma}}{\Gamma(2-\gamma)}, \quad B_i = (i+1)^{1-\gamma} - (i)^{1-\gamma}.$$

Hence we obtain

$$z(x, t_{j+1}) = \sigma y(x, t_{j+1}) - \sigma y(x, t_j) + \sigma \sum_{i=1}^{j} B_i \left(y(x, t_{j-i+1}) - y(x, t_{j-i}) \right) + \mathcal{R}_\tau. \tag{10}$$

Substituting (10) into (1) On Ω^{M_τ}, we get

$$z(x, t_{j+1}) - \varepsilon \frac{\partial^2 y^{j+1}(x)}{\partial x^2} + q(x) \frac{\partial y^{j+1}(x)}{\partial x} + r^{j+1}(x) y^{j+1}(x) = -s^{j+1}(x) y^{j-k+1}(x) + f^{j+1}(x).$$

Once the expressions are rearranged and the operator form has been put in, we get

$$\tilde{\mathfrak{L}} y^{j+1}(x) = -\varepsilon \frac{\partial^2 y^{j+1}(x)}{\partial x^2} + q(x) \frac{\partial y^{j+1}(x)}{\partial x} + \nu^{j+1}(x) y^{j+1}(x) = F^j(x) \tag{11}$$

for $j = 1, 2, \ldots, M_\tau$ with

$$\begin{cases} y(x,t) = \varphi_b(x, t_j), \quad \text{for}(x,t) \in [0,1] \times [-\delta, 0], \\ y(0, t_j) = \varphi_l(t_j), y(1, t_j) = \varphi_r(t_j), \quad \text{for} \quad t \in (0, \mathfrak{T}), \end{cases} \tag{12}$$

where

$$\nu(x, t_{j+1}) = r^{j+1}(x) + \sigma,$$

$$F^{j+1}(x) = \begin{cases} -s^{j+1}(x)\varphi_b(x,t_{j-k+1}) + f(x,t_{j+1}) + \sigma B_j \varphi_b(x,t_j) \\ +\sigma \sum_{i=1}^{j} B_i \left(y(x,t_{j-i+1}) - y(x,t_{j-i})\right), \quad \text{for } j = 1,2,\ldots,k, \\ \\ -s^{j+1}(x)y^{j-k+1}(x) + f^{j+1}(x) + \sigma\psi_b(x,t_j) \\ +\sigma \sum_{i=1}^{j} B_i \left(y(x,t_{j-i+1}) - y(x,t_{j-i})\right), \quad \text{for } j = k+1,\ldots,M_\tau. \end{cases}$$

After some rearrangement of (11) we obtain

$$\left(1 + \alpha_0 \mathfrak{L}^*_{\varepsilon,\delta}\right) y^{j+1}(x) = F^{j+1}(x), \tag{13}$$

where

$$\alpha_0 = \Gamma(2-\gamma)\Delta t^\gamma,$$

$$\mathfrak{L}^*_{\varepsilon,\delta} = -\varepsilon\frac{\partial^2}{\partial x^2} + q(x)\frac{\partial}{\partial x} + r^{j+1}(x),$$

$$F^{j+1}(x) = \begin{cases} -\alpha_0 r^{j+1}(x)\varphi_b(x,t_{j-k+1}) + \alpha_0 fj + 1(x) + \varphi_b(x,t_j) \\ +\sum_{i=1}^{j} B_i \left(y(x,t_{j-i+1}) - y(x,t_{j-i})\right), \quad \text{for } j = 1,2,\ldots,k, \\ \\ -\alpha_0 r^{j+1}(x)y^{j-k+1}(x) + \alpha_0 fj + 1(x) + \varphi_b(x,t_j) \\ +\sum_{i=1}^{j} B_i \left(y(x,t_{j-i+1}) - y(x,t_{j-i})\right), \quad \text{for } j = k+1,\ldots,M_\tau. \end{cases}$$

Lemma 4 (Semi-discrete Maximum Principle)**.** Let $\psi(x,t_{j+1})$ be a smooth function such that $\psi(x,t_{j+1}) \geq 0$ and, $\psi(x,t_{j+1}) \geq 0$, for all $(x,t_{j+1}) \in \Lambda = \{0\}\times(0,\mathfrak{T}] \cup \{1\}\times(0,\mathfrak{T}] \cup [0,1]\times[-\delta,0]$. Then $(1+\mathfrak{L}^*_{\varepsilon,\delta})\psi(x,t_{j+1}) \geq 0$ in $\tilde{\Omega}$ implies that $\psi(x,t_{j+1}) \geq 0$, for all $(x,t_{j+1}) \in \bar{\Omega}$.

Proof. Suppose that there exists $(\iota,t_{j+1}) \in \bar{\Omega}$ with

$$\psi(\iota,t_{j+1}) = \min_{(x,t_{j+1})\in\bar{\Omega}} \psi(x,t_{j+1}), \quad \text{and} \quad \psi(\iota,t_{j+1}) < 0,$$

and that $\psi(\iota,t_{j+1}) < 0$. Then

$$(\iota,t_{j+1}) \notin \{(0,t_{j+1}),(1,t_{j+1})\} \quad \text{and} \quad \psi_x(\iota,t_{j+1}) = 0, \psi_{xx}(\iota,t_{j+1}) > 0.$$

Applying the operator $\mathfrak{L}^*_{\varepsilon,\delta}$ on $\psi(x,t_{j+1})$, we get

$$(1+\alpha_0\mathfrak{L}^*_{\varepsilon,\delta})\psi(x,t_{j+1}) = \psi(\iota,t_{j+1}) + \alpha_0\left(-\varepsilon\psi_{xx}(x,t_{j+1}) + q(x)\psi_x(x,t_{j+1})\right.$$

$$+r(x,t_{j+1})\psi(x,t_{j+1}))\,.$$

At the point of minimum (ι,t_{j+1}), we obtain

$$(1+\alpha_0\mathfrak{L}^*_{\varepsilon,\delta})\psi(\iota,t_{j+1}) = \psi(\iota,t_{j+1}) + \alpha_0(-\varepsilon\psi_{xx}(\iota,t_{j+1}) + q(\iota)\varphi_x(\iota,t_{j+1}) + r(\iota,t_{j+1})\psi(\iota,t_{j+1})).$$

At the point (ι,t_{j+1}), the function ψ has minimum, so $\psi_x=0$, $\psi_{xx}\geq 0$ at point (ι,t_{j+1}) and $r(\iota,t_{j+1})\geq 0$ for $(\iota,t_{j+1})\in\Omega$. Therefore, we have

$$(1+\alpha_0\mathfrak{L}^*_{\varepsilon,\delta})\psi(\iota,t_{j+1})<0,$$

which contradicts our assumption $(1+\mathfrak{L}^*_{\varepsilon,\delta})\psi(x,t_{j+1})$ in Ω.

Therefore, we conclude that $\psi(x,t_{j+1})\geq 0,$ for all $(x,t_{j+1})\in\bar{\Omega}$.

Hence from the above prove the operator $(1+\alpha_0\mathfrak{L}^*_{\varepsilon,\delta})$ satisfies the maximum principle, and consequently

$$\left\|(1+\alpha_0\mathfrak{L}^*_{\varepsilon,\delta})^{-1}\right\| \leq \frac{1}{1+\theta\tau}. \tag{14}$$

□

Lemma 5. [Truncation error] The local truncation error corresponding to the semi-discretized problem (12) satisfies

$$\left|\mathcal{R}_\tau^{j+1}\right| \leq C\tau^{2-\gamma}. \tag{15}$$

Proof. From semi-discretized problem, we have

$$\begin{aligned}
\mathcal{R}_\tau^{j+1} &= \frac{O(\tau)}{\Gamma(1-\gamma)}\int_0^{t_{j+1}}(t_{j+1}-\eta)^{-\gamma}d\eta \\
&= \frac{O(\tau)}{\Gamma(1-\gamma)}\int_0^{(j+1)\tau}((j+1)\tau-\eta)^{-\gamma}d\eta \\
&= \frac{O(\tau)}{\Gamma(1-\gamma)}\frac{((j+1)\tau)^{1-\tau}}{1-\gamma} \\
&= \frac{((j+1)\tau)^{1-\tau}}{\Gamma(2-\gamma)}O(\tau)(\tau^{1-\tau}) \\
&\leq \frac{((j+1)\tau)^{1-\tau}}{\Gamma(2-\gamma)}\tau^{2-\tau} \\
&\leq C\tau^{2-\tau}.
\end{aligned}$$

Therefore, we obtain

$$\left|\mathcal{R}_{\tau}^{j+1}\right| \leq C\tau^{2-\tau}.$$

□

Lemma 6. [Global error bound:] The global error estimation at t_{j+1} satisfies

$$\|E_{j+1}\| \leq C\tau^{2-\gamma}.$$

Proof. Since the function $y(x, t_{j+1})$ satisfies

$$(1 + \alpha_0 \mathfrak{L}_{\varepsilon,\delta}^{*}) y(x, t_{j+1}) = F^{j+1}(x), \tag{16}$$

and also the solution of the continuous problem (1)–(2) is smooth enough, then we have

$$\begin{aligned} F^{j+1}(x) &= (1 + \alpha_0 \mathfrak{L}_{\varepsilon,\delta}^{*}) y(x, t_{j+1}) + \mathcal{R}_{\tau}^{j+1} \\ &= (1 + \alpha_0 \mathfrak{L}_{\varepsilon,\delta}^{*}) y(x, t_{j+1}) + C\tau^{2-\gamma}, \end{aligned} \tag{17}$$

From (16)–(17), the error corresponding to (13) satisfies the following boundary value problem:

$$\begin{aligned} (1 + \alpha_0 \mathfrak{L}_{\varepsilon,\delta}^{*}) E_{j+1} &= C\tau^{2-\gamma}, \\ \Longrightarrow E_{j+1} &= (1 + \alpha_0 \mathfrak{L}_{\varepsilon,\delta}^{*})^{-1} \tau^{2-\gamma}, \end{aligned}$$

$$\|E_{j+1}\| \leq \frac{1}{1 + \theta\tau} C\tau^{2-\gamma}.$$

hence, we obtain the result

$$\|E_{j+1}\| \leq C\tau^{2-\gamma}.$$

□

Theorem 1. The semi-discretize solution $y(x, t_{j+1})$ and its derivatives satisfy the following bounds:

$$\left| \frac{d^i y(x, t_{j+1})}{dx^i} \right| \leq C(1 + \varepsilon^{-i} \exp(-\beta(1 - x)/\varepsilon)), \qquad \text{for} \quad i = 0, 1, 2, 3, 4.$$

Proof. For the proof, refer [14]. □

We can write (11) as operator form,

$$\tilde{\mathfrak{L}}_{\varepsilon}^{\tau} y^{j+1}(x) = F^{j}(x), \tag{18}$$

where $\tilde{\mathfrak{L}}_{\varepsilon}^{\tau} y(x) = -\varepsilon \frac{\partial^2 y(x)}{\partial x^2} + q(x)\frac{\partial y(x)}{\partial x} + \nu(x)y(x)$ and

$$F^{j}(x) = \begin{cases} -s^{j+1}(x)\varphi_b(x, t_{j-k+1}) + f(x, t_{j+1}) + \sigma B_j \varphi_b(x, t_j) \\ +\sigma \sum_{i=1}^{j} B_i \left(y(x, t_{j-i+1}) - y(x, t_{j-i}) \right), \qquad \text{for } j = 1, 2, \ldots, k, \\ \\ -s^{j+1}(x) yj - k + 1(x) + f^{j+1}(x) + \sigma \psi_b(x, t_j) \\ +\sigma \sum_{i=1}^{j} B_i \left(y(x, t_{j-i+1}) - y(x, t_{j-i}) \right), \qquad \text{for } j = k+1, \ldots, M_\tau. \end{cases}$$

4.2 Spatial discretization

To solve the semi-discretized problem (11), we use the extended cubic B-spline collocation scheme. To take into consideration the exponential properties of exact solution on the uniform mesh, artificial viscosity will be introduced. Thus, an artificial viscosity $\sigma(x, \varepsilon)$ replaces the perturbation parameter ε, which disrupts the highest derivative.

4.3 Extended cubic B-spline collocation method

We divided the spatial domain using uniform mesh such that the set $\Omega_x^{N_h}$ is the collection of all mesh points in the spacial direction; with $x_i = ih,\ i = 0, 1, 2, \ldots, N_h$.

The extended cubic B-spline G_i of degree 4 for $\mu \in (-8, 1)$, has the following form [10]:

$$G_i(x) = \frac{1}{24h^4}\begin{cases} 4h(1-\mu)(x-x_i)^3 + 3\mu(x-x_{i-2})^4, & x \in [x_{i-1}, x_{i-1}], \\ (4-\mu)h^4 + 12h^3(x-x_{i-1}) + 6h^2(2+\mu)(x-x_{i-1})^2 & \\ -12h(x-x_{i-1})^3 - 3\mu(x-x_{i-1})^4, & x \in [x_{i-1}, x_i], \\ (4-\mu)h^4 + 12h^3(x_{i+1}-x) + 6h^2(2+\mu)(x_{i+1}-x)^2 & \\ -12h(x_{i+1}-x)^3 - 3\mu(x_{i+1}-x)^4, & x \in [x_i, x_{i+1}], \\ 4h(1-\mu)(x_{i+2}-x)^3 + 3\mu(x_{i+2}-x)^4, & x \in [x_{i+1}, x_{i+2}], \\ 0, & \text{otherwise.} \end{cases} \tag{19}$$

Consider that the approximation y_i to the exact solution $Y(x,\mu)$ at the point (x, t_{j+1}). It can be defined as follows using combinations of the cubic B-splines and unknown time-dependent parameters:

$$Y(x,\mu) = \sum_{k=-1}^{N_h+1} a_k G_k, \tag{20}$$

where a_k are time dependent parameters to be determined from the collocation method with the boundary and initial conditions.

Outside of the region $[x_{i-1}, x_{i+2}]$, the extended cubic B-splines and their four principle derivatives vanish. Principle four spline functions cover the interval $[x_{i-1}, x_i]$. Thus, the $y(x,t)$ variation over the element can be written as

$$Y(x,\mu) = \sum_{k=i-1}^{i+2} a_k G_k, \tag{21}$$

where a_{i-1}, a_i, a_{i+1}, and a_{i+2} are the element parameters. Equation (21) can be used to compute the values of the cubic B-spline $G_k(x,\mu)$ and its successive derivatives $G'_k(x,\mu), G''_k(x,\mu)$ at the knots. These values are provided in Table 1 below.

Table 1: Values of $G_k(x,\mu)$ and its principle two derivatives at the node points

G_i	x_{i-2}	x_{i-1}	x_i	x_{i+1}	x_{i+1}
$G_i(x_i,\mu)$	0	$\frac{4-\mu}{12}$	$\frac{8+\mu}{12}$	$\frac{4-\mu}{24}$	0
$G_i'(x_i,\mu)$	0	$-\frac{1}{2h}$	0	$\frac{1}{2h}$	0
$G_i''(x_i,\mu)$	0	$\frac{2+\mu}{2h^2}$	$-\frac{2+\mu}{h^2}$	$\frac{2+\mu}{2h^2}$	0

Substituting the values of Table 1 in (11) and its first and second derivatives at node x_i gives

$$\begin{aligned} y(x_i,\mu) &= \frac{4-\mu}{24}a_{i-1} + \frac{8+\mu}{12}a_i + \frac{4-\mu}{24}a_{i+1}, \\ y'(x_i,\mu) &= -\frac{1}{2h}a_{i-1} + \frac{1}{2h}a_{i+1}, \\ y''(x_i,\mu) &= \frac{2+\mu}{2h^2}a_{i-1} - \frac{2+\mu}{h^2}a_i + \frac{2+\mu}{2h^2}a_{i+1}. \end{aligned} \tag{22}$$

Substituting (22) into (11), then we obtain

$$\begin{aligned} &-\frac{\xi(i)(2+\mu)}{2h^2}\left(a_{i-1}-2a_i+a_{i+1}\right) - \frac{q_i}{2h}\left(a_{i-1}-a_{i+1}\right) \\ &+\nu_i^{j+1}\left(\frac{4-\mu}{24}a_{i-1} + \frac{8+\mu}{12}a_i + \frac{4-\mu}{24}a_{i+1}\right) = F_i^j. \end{aligned} \tag{23}$$

Let us introduce the artificial viscosity $\xi(x_i,\varepsilon)$ into (11). Artificial diffusion (or artificial viscosity) is added to the term in the given differential equation that contains the singular perturbation parameter to generate the discretization scheme. This artificial diffusion is introduced by means of fitting factor $\xi_i(\varepsilon) = \xi(x_i,\varepsilon)$. The zero order asymptotic solution of (18) exists and unique (see [29, 34]) given as

$$y(x) = y_0(x) + [\varphi_r - y_0(1)]\exp\left(-\int_0^1\left(\frac{q(x)}{\varepsilon} - \frac{r(x)}{q(x)}\right)dx\right) + O(\varepsilon). \tag{24}$$

Approximation for $q(x)$ and $r(x)$ confined to their first terms about $x=1$ from Taylor's series can be obtained as

$$y(x) = y_0(x) + [\varphi_r - y_0(1)]\exp\left(-\frac{q(x)(1-x)}{\varepsilon}\right), \tag{25}$$

where $y_0(x)$ is the solution of the reduced problem. The convection-diffusion problem in (2) has a right layer, and we have the uniform discretization

point $x_i = ih$ and $\rho = \frac{h}{\varepsilon}$. By taking the limit $h \longrightarrow 0$, for (25) at x_{i-1}, x_i and x_{i+1}, then we obtain

$$\begin{aligned}
\lim_{h\to 0} y_i &= y_0(0) + [\varphi_r - y_0(1)]\, e^{-\frac{q(x)}{\varepsilon}(1-x)},\\
\lim_{h\to 0} y_{i-1} &= y_0(0) + [\varphi_r - y_0(1)]\, e^{-\frac{q(x)}{\varepsilon}(1-x)} e^{-q(0)\rho},\\
\lim_{h\to 0} y_{i+1} &= y_0(0) + [\varphi_r - y_0(1)]\, e^{-\frac{q(x)}{\varepsilon}(1-x)} e^{q(0)\rho}.
\end{aligned} \tag{26}$$

Now, we determine the fitting factor ξ by considering the fitted operator (18); that is,

$$\lim_{h\to 0} \xi_i = \lim_{h\to 0} \frac{q(i)}{2+\mu} \left(\frac{a_{i-1} - a_{i+1}}{a_{i-1} - 2a_i + a_{i+1}} \right). \tag{27}$$

By substituting (26) into (27) and simplifying it, we have

$$\xi(i) = \frac{\rho q(i)}{2+\mu} \coth\left(\frac{q(i)\rho}{2}\right). \tag{28}$$

Hence by using the artificial viscosity into (23) and simplifying it, then we get

$$\begin{aligned}
&\left[-\frac{\xi(i)(2+\mu)}{2h^2} - \frac{q_i}{2h} + \nu_i^{j+1}\frac{4-\mu}{24}\right] a_{i-1} + \left[\frac{\xi(2+\mu)}{2h^2} + \frac{8+\mu}{12}\right] a_i\\
&+ \left[-\frac{\xi(2+\mu)}{2h^2} - \frac{q_i}{2h} + \frac{4-\mu}{24}\nu_i^{j+1}\right] a_{i+1} = F_i^j.
\end{aligned} \tag{29}$$

Let

$$\begin{aligned}
\mathfrak{H}_i^- &= -\frac{\xi(2+\mu)}{2h^2} - \frac{q_i}{2h} + \nu_i^{j+1}\frac{4-\mu}{24},\\
\mathfrak{H}_i^0 &= \frac{\xi(2+\mu)}{2h^2} + \frac{8+\mu}{12},\\
\mathfrak{H}_i^+ &= -\frac{\xi(2+\mu)}{2h^2} - \frac{q_i}{2h} + \frac{4-\mu}{24}\nu_i^{j+1}.
\end{aligned}$$

Then

$$\mathfrak{H}_i^- a_{i-1} + \mathfrak{H}_i^0 a_i + \mathfrak{H}_i^+ a_{i+1} = F_i^j. \tag{30}$$

For the given boundary conditions, we have

$$\begin{aligned}
&\frac{4-\mu}{24}a_{-1} + \frac{8+\mu}{12}a_0 + \frac{4-\mu}{24}a_1 = \varphi_l(t_{j+1}),\\
&\frac{4-\mu}{24}a_{N_h-1} + \frac{8+\mu}{12}a_{N_h} + \frac{4-\mu}{24}a_{N_h+1} = \varphi_r(t_{j+1}).
\end{aligned} \tag{31}$$

For $(N_h+3)\times(N_h+3)$ systems, (30)–(31) provide the (N_h+3) unknowns $a_{-1}, a_0, a_1, \ldots, a_{N_h+1}$. The (N_h+1) system of equations in (N_h+1) unknowns $a_0, a_1, \ldots, a_{N_h}$, can be expressed in the matrix form by eliminating a_{-1} and a_{N_h+1} from (30)–(31),

$$HA = F, \tag{32}$$

where

$$H = \begin{bmatrix} -2\left(\frac{8+\mu}{4-\mu}\right)\mathfrak{H}_0^- + \mathfrak{H}_0^0 & -\mathfrak{H}_0^- + \mathfrak{H}_0^+ & & & & & \\ \mathfrak{H}_1^- & \mathfrak{H}_1^0 & \mathfrak{H}_1^+ & & & & \\ \ddots & \ddots & \ddots & & & & \\ & \mathfrak{H}_i^- & \mathfrak{H}_i^0 & \mathfrak{H}_i^+ & & & \\ & & \ddots & \ddots & \ddots & & \\ & & & \mathfrak{H}_{Nh-1}^- & \mathfrak{H}_{Nh-1}^0 & \mathfrak{H}_{Nh-1}^+ & \\ & & & & \mathfrak{H}_{N_h}^- - \mathfrak{H}_{N_h}^+ & -2\left(\frac{8+\mu}{4-\mu}\right)\mathfrak{H}_{N_h}^- + \mathfrak{H}_{N_h}^0 \end{bmatrix},$$

$$A = [a_0 \quad a_1 \quad \ldots \quad a_{N_h-1} \quad a_{N_h}]^T$$

and

$$F = \left[F_0^j - \frac{24}{4-\mu}H_0^-\varphi_l(t_{j+1}) \quad F_1^j \quad F_2^j \quad \ldots \quad F_{N_h-1}^j \quad F_{N_h}^j - \frac{24}{4-\mu}H_{N_h}^-\varphi_r(t_{j+1})\right]^T.$$

5 Convergence analysis

Lemma 7. Consider the extended cubic B-spline

$$G = \{G_{-1}(x,\mu), G_0(x,\mu), G_1(x,\mu), \ldots, G_{N_h}(x,\mu), G_{N_h+1}(x,\mu)\}$$

given in (19). It satisfies the inequality

$$\sum_{i=-1}^{N_h+1} |(G_i(x,\mu))| \leq \frac{7}{4}. \tag{33}$$

Proof. We start from the known properties

$$\left|\sum_{i=-1}^{N_h+1} G_i(x_i,\mu)\right| = \sum_{i=-1}^{N_h+1} |G_i(x_i,\mu)|,$$

where $G_i(x,\mu)$ is nonzero at three nodal points only. Thus, using Table 1, at every nodal value x_i, we obtain

$$\sum_{i=-1}^{N_h+1} |G_i(x_i,\mu)| = |G_{i-1}(x_i,\mu)| + |G_i(x_i,\mu)| + |G_{i+1}(x_i,\mu)|$$
$$= \frac{4-\mu}{24} + \frac{8+\mu}{12} + \frac{4-\mu}{24} = 1 < \frac{7}{4}.$$

From Table 1, for $x_{i-1} \le x \le x_i$, we have

$$|G_i(x_i,\mu)| \le \frac{8+\mu}{12}, \quad |G_{i-1}(x_i,\mu)| \le \frac{8+\mu}{12}.$$

Similarly, for $x_{i-1} \le x \le x_i$, we get

$$|G_{i+1}(x_i,\mu)| \le \frac{8+\mu}{12}, \quad |G_{i-2}(x_i,\mu)| \le \frac{8+\mu}{12}.$$

Now, for any point $x \in [x_{i-1}, x_i]$, we obtain

$$\sum_{i=-1}^{N_h+1} |G_i(x_i,\mu)| = |G_{i-1}(x_i,\mu)| + |G_i(x_i,\mu)| + |G_{i+1}(x_i,\mu)|$$
$$= \frac{4-\mu}{24} + \frac{8+\mu}{12} + \frac{4-\mu}{24} = \frac{20-\mu}{12}.$$

Since $-8 < \mu < 1$, thus

$$\sum_{i=-1}^{N_h+1} |G_i(x_i,\mu)| = \frac{20-\mu}{12} < \frac{7}{4}.$$

□

Let $\bar{\Psi}$ be a unique cubic spline interpolate obtained from an approximately solution $Y(x,\mu)$ of the problems (11) to the given solution $y(x)$. Then

$$\bar{\psi}(x) = \sum_{i=-1}^{N_h+1} \bar{A} G_i(x,\mu). \tag{34}$$

For $z > 0$, let $\alpha(z) = z\coth(z)$ satisfy $\alpha(0) = 1, \alpha(z) = \alpha(-z)$. Then $|\alpha(z) - 1| \le Cz^2$ for $0 < z \le 1$. Since $\coth z \longrightarrow 1$ as $z \longrightarrow \infty$, so $|\alpha(z) - 1| \le Cz$. Hence for $z > 0$, we have

$$|\xi(z) - 1| \le \frac{Cz^2}{1+z} \quad \text{and} \quad \frac{\varepsilon(h/\varepsilon)^2}{h/\varepsilon + 1} = \frac{h^2}{h+\varepsilon}. \tag{35}$$

Lemma 8. Set a cubic spline interpolant $\bar{\Psi} \in C^2(0,1)$ to a solution $Y(x)$. For $x \in (x_i, x_{i+1})$, the standard cubic spline interpolation approximate holds

if $Y(x) \in C^4(0,1)$. According to Hall's estimate [20], we have

$$\left|Y^(k)(x) - \bar{\Psi}(x)^{(k)}\right| \leq c_i \left\|Y^{(4)}\right\| N_h^{-(4-k)}, \tag{36}$$

where c_i's are constants independent of h and N_h.

Theorem 2. [Parameter uniform convergence] Let $S(x,\mu)$ be the collocation approximation from the space of splines to the solution, $Y^{j+1}(x)$ be the approximate solution of the semi-discretized problem (11), and let $y(x_i, t_{j+1})$ be the continuous solution of (1) and (2). Therefore, the following error bound is valid for suitably large N:

$$\left\|Y^{j+1}(x_i) - y^{j+1}(x_i)\right\| \leq \frac{N_h^{-2}}{N_h^{-1} + \varepsilon}. \tag{37}$$

Proof. To prove the theorem, we start by using Lemma 5. We get the bounds

$$\begin{aligned}
&\left|Y^{j+1}(x_i) - \bar{\Psi}(x_i)\right| \leq c_0 \left\|\frac{d^4 Y^{j+1}(x_i)}{dx^4}\right\| N_h^{-4},\\
&\left|\frac{dY^{j+1}(x_i)}{dx} - \frac{d\bar{\Psi}(x_i)}{dx}\right| \leq c_1 \left\|\frac{d^4 Y^{j+1}(x_i)}{dx^4}\right\| N_h^{-3},\\
&\left|\frac{d^2 Y^{j+1}(x_i)}{dx^2} - \frac{d^2\bar{\Psi}(x_i)}{dx^2}\right| \leq c_2 \left\|\frac{d^4 Y^{j+1}(x_i)}{dx^4}\right\| N_h^{-2}.
\end{aligned} \tag{38}$$

Using the triangle inequality, we have

$$\left|Y^{j+1}(x_i) - y^{j+1}(x_i)\right| \leq \left|Y^{j+1}(x_i) - \bar{\Psi}(x_i)\right| + \left|\bar{\Psi}(x_i) - y^{j+1}(x_i)\right|.$$

The collocating condition gives

$$\tilde{\mathfrak{L}}_\varepsilon^{h,\tau} y^{j+1}(x_i) = \tilde{\mathfrak{L}}_\varepsilon^{h,\tau} Y^{j+1}(x_i).$$

Assume that $\tilde{\mathfrak{L}}_\varepsilon^{h,\tau}\bar{\Psi}(x_i) = \bar{F}(x_i, t_j)$, which satisfies the boundary conditions, $\bar{\Psi}(x_0) = \bar{\Psi}(x_{N_h+1})$. Then

$$\begin{aligned}
\left|\tilde{\mathfrak{L}}_\varepsilon^{h,\tau} y^{j+1}(x_i) - \tilde{\mathfrak{L}}_\varepsilon^{h,\tau}\bar{\Psi}(x_i)\right| &= \left|\tilde{\mathfrak{L}}_\varepsilon^{h,\tau} Y^{j+1}(x_i) - \tilde{\mathfrak{L}}_\varepsilon^{h,\tau}\bar{\Psi}(x_i)\right|\\
&= \left|-\varepsilon\left(\frac{d^2 y^{j+1}(x_i)}{dx^2} - \xi_i(\varepsilon)\frac{d^2\bar{\Psi}(x_i)}{dx^2}\right)\right|\\
&\quad + \left|q(x_i)\left(\frac{dy^{j+1}(x_i)}{dx} - \frac{d\bar{\Psi}(x_i)}{dx}\right)\right|
\end{aligned}$$

$$
\begin{aligned}
&+\left|\nu^{j+1}(x-i)(y^{j+1}(x_i)-\bar{\Psi}(x_i))\right| \\
&\leq |\varepsilon|\,|\xi|\left\|\frac{d^2y^{j+1}(x_i)}{dx^2}\right\| \\
&+|\varepsilon|\,|\xi|\left|\frac{d^2y^{j+1}(x_i)}{dx^2}-\xi_i(\varepsilon)\frac{d^2\bar{\Psi}(x_i)}{dx^2}\right| \\
&+|q(x_i)|\left|\frac{dy^{j+1}(x_i)}{dx}-\frac{d\bar{\Psi}(x_i)}{dx}\right| \\
&+\left|\nu^{j+1}\right|\left|y^{j+1}(x_i)-\bar{\Psi}(x_i)\right|. \qquad (39)
\end{aligned}
$$

Now, using (35) and Lemma 1, then we obtain

$$
\begin{aligned}
&\max_{0\leq i,j\leq N_h,M_\tau}\left|y^{j+1}(x_i)-\bar{\Psi}(x_i)\right| \leq \frac{N_h^{-2}}{N_h^{-1}+\varepsilon} \\
&\Rightarrow \left\|y^{j+1}(x_i)-\bar{\Psi}(x_i)\right\| \leq \frac{N_h^{-2}}{N_h^{-1}+\varepsilon}. \qquad (40)
\end{aligned}
$$

The coefficient matrix associated with (20) is of size $(N_h+1)\times(N_h+1)$ with its elements. For $i=1,2,\ldots,N_h-1$, we have

$$
\begin{aligned}
&\mathfrak{H}_i^- < 0, \quad \text{since all terms are positive,} \\
&\mathfrak{H}_i^0 > 0, \quad \text{since all terms are positive,} \\
&\mathfrak{H}_i^+ < 0, \quad \text{since all terms are positive} \quad \coth\left(\frac{q(i)\varepsilon}{2h}\right) \geq 1.
\end{aligned}
$$

Thus, the coefficient matrix of the proposed method, satisfies the properties of M-matrix. This implies that the inverse matrix exists and it is nonnegative. This implies [40]

$$\left|H^{-1}\right| < CN_h^{-2}. \qquad (41)$$

From (32) and $\mathfrak{L}_\varepsilon^{h,\tau}y^{j+1}(x_i)-\tilde{\mathfrak{L}}_\varepsilon^{h,\tau}\bar{\Psi}(x_i)$, we get the result

$$H(A-\bar{A}) = F-\bar{F}, \qquad (42)$$

where $A-\bar{A}=(a_0-\bar{a}_0,\quad a_1-\bar{a}_1,\quad a_2-\bar{a}_2,\ldots,\quad a_{N_h}-\bar{a}_{N_h})$ and $F-\bar{F}=\big(F(x_0,t_j)-\bar{F}(x_0,t_j),\quad F(x_1,t_j)-\bar{F}(x_1,t_j),\ldots,F(x_{N_h},t_j)-\bar{F}(x_{N_h},t_j)\big)$. Using (41), so the boundary conditions are bounded. Therefore, (39) and (42) give

$$\left|A-\bar{A}\right| \leq \frac{N_h^{-2}}{N_h^{-1}+\varepsilon}.$$

Hence, by using (21) for $Y(x,\mu)$ and Lemma 34 for $\bar{\Psi}(x)$, we get

$$\left|Y(x_i,\mu)-\bar{\Psi}(x_i)\right| = \left|A-\bar{A}\right| \sum_{0}^{N_h+1} |G_i(x_i,\mu)| \leq \frac{N_h^{-2}}{N_h^{-1}+\varepsilon}. \tag{43}$$

Thus, using (40) and (43), also the triangle inequality, we obtain our result

$$\left\|Y^{j+1}(x_i)-y^{j+1}(x_i)\right\| \leq \frac{N_h^{-2}}{N_h^{-1}+\varepsilon}. \tag{44}$$

□

Theorem 3. If y and Y be exact and cubic B-spline approximation solution of the problem (1), respectively, then the following error bound holds:

$$\max_{0\leq i,j\leq M_h,M_t} |y(x_i,t_j)-Y(x_i,\mu)| \leq C\left(\frac{h^2}{h+\varepsilon}+(\tau)^{2-\gamma}\right). \tag{45}$$

Proof. By combining Theorems 6 and 2, we get our result. □

6 Numerical result

In this section, we show two numerical examples that demonstrate the accuracy of the method and the result of the error analysis. Separate tables display the error and corresponding convergence rates for each of these two test examples. Since the exact solution to the example is unknown, double mesh will be used in this article to determine the accuracy of the numerical solution. The maximum point-wise absolute error is determined as

$$E_{\varepsilon}^{N_h,M_\tau} = \max_{0\leq i,j\leq N_h,M_\tau} \left|Y_{i,j}^{N_h,M_\tau}(x_i,t_j)-Y_{i,j}^{2N_h,2M_\tau}(x_{2i},t_{2j})\right|,$$

where N_h and M_τ are the number of mesh points in the spatial and temporal directions, respectively. The parameter uniform error estimation is defined as

$$e^{N_h,M_\tau} = \max_{\varepsilon}\{E_{\varepsilon}^{N_h,M_\tau}\}.$$

Next, we also determine the rate of convergence of the method by using the formula

$$RoC_{\varepsilon}^{N_h,M_\tau} = \log_2\left(\frac{E_{\varepsilon}^{N_h,M_\tau}}{E_{\varepsilon}^{2N_h,2M_\tau}}\right).$$

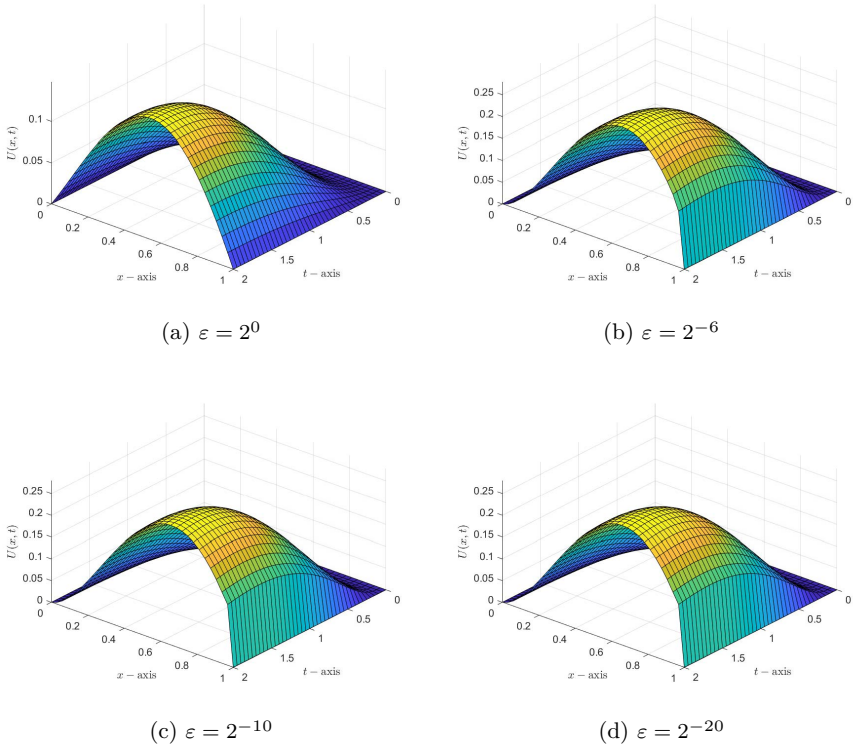
The parameter uniform rate of convergence is defined as

$$R^{N_h,M_\tau} = \max_{\varepsilon}\{RoC_{\varepsilon}^{N_h,M_\tau}\}.$$

Example 1. Consider the time-fractional SPPPDE

$$\begin{aligned} D_t^{\gamma} y(x,t) - \varepsilon \frac{\partial^2 y(x,t)}{\partial x^2} + \left(2 - x^2\right) \frac{\partial y(x,t)}{\partial x} + ((x+1)(t+1))y(x,t) \\ = y(x,t-1) + 10t^2 \exp(-t)x(1-x), \end{aligned}$$

on $(x,t) \in \Omega = (0,1) \times (0,\mathfrak{T}]$, with initial and boundary conditions $\varphi_b(x,t) = 0$, $\varphi_l(t) = 0$ and $\varphi_r(t) = 0$.



(a) $\varepsilon = 2^0$ (b) $\varepsilon = 2^{-6}$

(c) $\varepsilon = 2^{-10}$ (d) $\varepsilon = 2^{-20}$

Figure 1: Three-dimensional plot of the numerical solution for Example 1 for different values of ε with $\gamma = 0.5, N_h = 32$, and $M_\tau = 40$.

Table 2: Absolute maximum error and rate of convergence for Example 1 for different values of ε, with fix $\gamma = 0.5$

$(N_h, M_\tau) \Rightarrow$	(16,20)	(32,40)	(64,80)	(128, 160)	(256, 320)	(512, 400)
$\varepsilon = 2^{0}$	4.1605e-03	1.9136e-03	8.5272e-04	3.6840e-04	1.5434e-04	6.2758e-05
	1.1205	1.1662	1.2108	1.2551	1.2983	-
$\varepsilon = 2^{-2}$	7.8335e-03	3.3678e-03	1.3897e-03	5.5457e-04	2.1515e-04	8.1537e-05
	1.2179	1.2771	1.3253	1.3661	1.3998	-
$\varepsilon = 2^{-4}$	9.5332e-03	3.9931e-03	1.6048e-03	6.2416e-04	2.3637e-04	8.7649e-05
	1.2554	1.3151	1.3624	1.4009	1.4312	-
$\varepsilon = 2^{-6}$	9.8688e-03	4.1357e-03	1.6591e-03	6.4279e-04	2.4217e-04	8.9302e-05
	1.2548	1.3177	1.3680	1.4083	1.4393	-
$\varepsilon = 2^{-8}$	9.8893e-03	4.1566e-03	1.6705e-03	6.4735e-04	2.4366e-04	8.9726e-05
	1.2505	1.3151	1.3676	1.4097	1.4413	-
$\varepsilon = 2^{-10}$	9.8893e-03	4.1568e-03	1.6711e-03	6.4798e-04	2.4397e-04	8.9826e-05
	1.2504	1.3147	1.3668	1.4093	1.4415	-
$\varepsilon = 2^{-12}$	9.8893e-03	4.1568e-03	1.6711e-03	6.4798e-04	2.4398e-04	8.9840e-05
	1.2504	1.3147	1.3668	1.4092	1.4414	-
$\varepsilon = 2^{-14}$	9.8893e-03	4.1568e-03	1.6711e-03	6.4798e-04	2.4398e-04	8.9840e-05
	1.2504	1.3147	1.3668	1.4092	1.4414	-
$\varepsilon = 2^{-20}$	9.8893e-03	4.1568e-03	1.6711e-03	6.4798e-04	2.4398e-04	8.9840e-05
	1.2504	1.3147	1.3668	1.4092	1.4414	-
$\varepsilon = 2^{-25}$	9.8893e-03	4.1568e-03	1.6711e-03	6.4798e-04	2.4398e-04	8.9840e-05
	1.2504	1.3147	1.3668	1.4092	1.4414	-
$\varepsilon = 2^{-30}$	9.8893e-03	4.1568e-03	1.6711e-03	6.4798e-04	2.4398e-04	8.9840e-05
	1.2504	1.3147	1.3668	1.4092	1.4414	-
e^{N_h, M_τ}	9.8893e-03	4.1568e-03	1.6711e-03	6.4798e-04	2.4398e-04	8.9840e-05
R^{N_h, M_τ}	1.2504	1.3147	1.3668	1.4092	1.4414	-

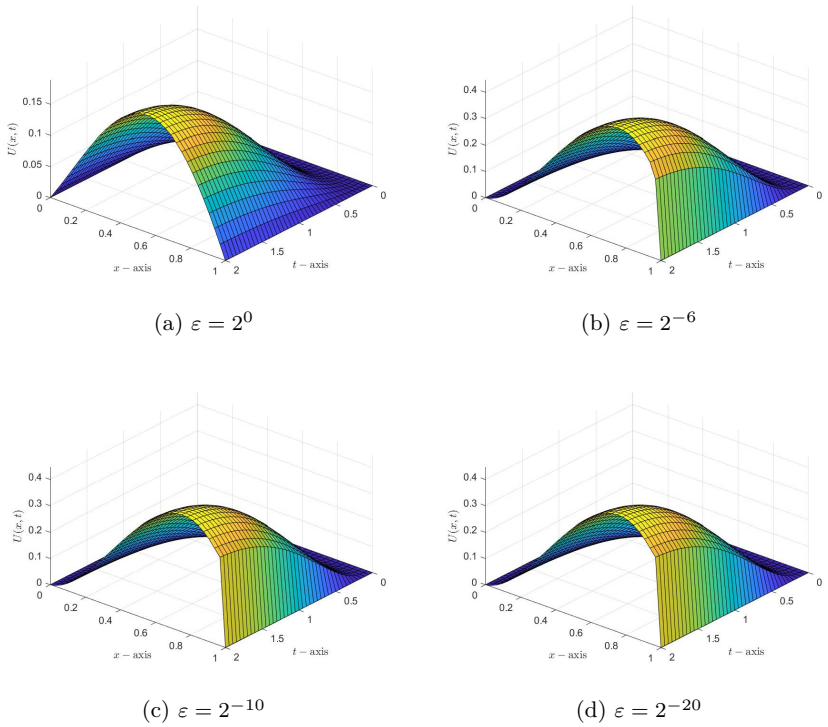
Table 3: Comparison of absolute maximum error and rate of convergence for Example 1 for different values of ε, with fix $\gamma = 0.5$

$(N_h, M_\tau) \Rightarrow$	(16,20)	(32,40)	(64,80)	(128, 160)
Proposed Method				
$\varepsilon = 2^{-6}$	9.8688e-03	4.1357e-03	1.6591e-03	6.4279e-04
	1.2548	1.3177	1.3680	-
$\varepsilon = 2^{-8}$	9.8893e-03	4.1566e-03	1.6705e-03	6.4735e-04
	1.2505	1.3151	1.3676	-
$\varepsilon = 2^{-10}$	9.8893e-03	4.1568e-03	1.6711e-03	6.4798e-04
	1.2504	1.3147	1.3668	-
$\varepsilon = 2^{-12}$	1.4000e-02	5.5760e-03	2.1233e-03	7.8272e-04
	1.3281	1.3929	1.4398	-
$\varepsilon = 2^{-14}$	1.4000e-02	5.5760e-03	2.1233e-03	7.8272e-04
	1.3281	1.3929	1.4398	-
$\varepsilon = 2^{-20}$	9.8893e-03	4.1568e-03	1.6711e-03	6.4798e-04
	1.2504	1.3147	1.3668	-
$\varepsilon = 2^{-25}$	9.8893e-03	4.1568e-03	1.6711e-03	6.4798e-04
	1.2504	1.3147	1.3668	-
$\varepsilon = 2^{-30}$	9.8893e-03	4.1568e-03	1.6711e-03	6.4798e-04
	1.2504	1.3147	1.3668	-
e^{N_h, M_τ}	9.8893e-03	4.1568e-03	1.6711e-03	6.4798e-04
R^{N_h, M_τ}	1.2504	1.3147	1.3668	-
Method in reference [27]				
$\varepsilon = 2^{-6}$	1.0088E-02	4.9401e-03	2.0143e-03	7.1385E-04
	1.0300	1.2943	1.4966	-
$\varepsilon = 2^{-8}$	1.1863e-02	6.3546e-03	3.3404e-03	1.8221e-03
	0.9006	0.9278	0.8744	-
$\varepsilon = 2^{-10}$	1.2246e-02	6.6457e-03	3.4625e-03	1.7661e-03
	0.8818	0.9406	0.9712	-
$\varepsilon = 2^{-12}$	1.2336e-02	6.7141e-03	3.5082e-03	1.7930e-03
	0.8776	0.9365	0.9683	-
$\varepsilon = 2^{-20}$	1.2365e-02	6.7364e-03	3.5230e-03	1.8022e-03
	0.8762	0.9352	0.9670	-
$\varepsilon = 2^{-25}$	1.2365e-02	6.7364e-03	3.5230e-03	1.8022e-03
	0.8762	0.9352	0.9670	-
$\varepsilon = 2^{-30}$	1.2365e-02	6.7364e-03	3.5230e-03	1.8022e-03
	0.8762	0.9352	0.9670	-
e^{N_h, M_τ}	1.2365e-02	6.7364e-03	3.5230e-03	1.8022e-03
R^{N_h, M_τ}	0.8762	0.9352	0.9670	-

Example 2. Consider the time-fractional SPPPDE

$$D_t^{\gamma} y(x,t) - \varepsilon \frac{\partial^2 y(x,t)}{\partial x^2} + \left(2 - x^2\right) \frac{\partial y(x,t)}{\partial x} + xy(x,t) = y(x,t-1) + 10t^2 \exp(-t)x(1-x),$$

on $(x,t) \in \Omega = (0,1) \times (0,\mathfrak{T}]$, with initial and boundary conditions $\varphi_b(x,t) = 0$, $\varphi_l(t) = 0$ and $\varphi_r(t) = 0$.



(a) $\varepsilon = 2^0$ (b) $\varepsilon = 2^{-6}$

(c) $\varepsilon = 2^{-10}$ (d) $\varepsilon = 2^{-20}$

Figure 2: Three-dimensional plot of the numerical solution for Example 2 for different values of ε with $\gamma = 0.5, M_x = 32$, and $M_t = 40$.

The numerical results are described in terms of maximum absolute errors and numerical rate of convergence in Tables 2 and 4. These results are compared with those of a previously developed numerical approach found in the literature in [27], using Tables 3 and 5. Additionally, the log-log plot (Figure 3) and the numerical solution for Examples 1 and 2 (refer to Figures

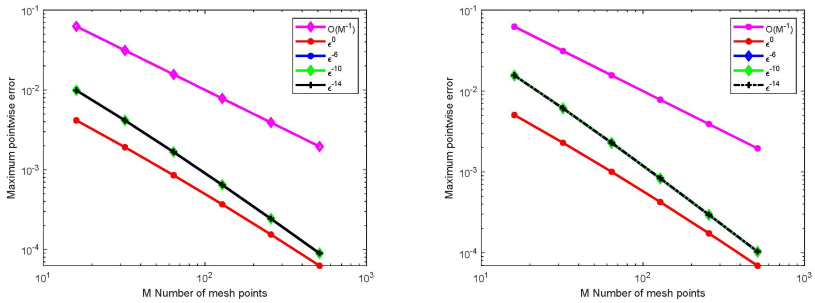
Table 4: Maximum error and rate of convergence for Example 2 for different values of ε, with fix $\gamma = 0.5$

$(N_h, M_\tau) \Rightarrow$	(16,20)	(32,40)	(64,80)	(128, 160)	(256, 320)	(512, 400)
$\varepsilon = 2^{0}$	5.0614e-03	2.2923e-03	1.0036e-03	4.2492e-04	1.7424e-04	6.9357e-05
	1.1427	1.1916	1.2400	1.2861	1.3290	-
$\varepsilon = 2^{-2}$	1.1193e-02	4.6126e-03	1.8150e-03	6.8997e-04	2.5570e-04	9.3106e-05
	1.2789	1.3456	1.3954	1.4321	1.4575	-
$\varepsilon = 2^{-4}$	1.4655e-02	5.8083e-03	2.1878e-03	7.9856e-04	2.8584e-04	1.0112e-04
	1.3352	1.4087	1.4540	1.4822	1.4992	-
$\varepsilon = 2^{-6}$	1.5546e-02	6.0863e-03	2.2785e-03	8.2711e-04	2.9408e-04	1.0330e-04
	1.3529	1.4175	1.4620	1.4919	1.5094	-
$\varepsilon = 2^{-8}$	1.5610e-02	6.1249e-03	2.2975e-03	8.3410e-04	2.9620e-04	1.0386e-04
	1.3497	1.4146	1.4618	1.4937	1.5119	-
$\varepsilon = 2^{-10}$	1.5610e-02	6.1253e-03	2.2986e-03	8.3507e-04	2.9663e-04	1.0399e-04
	1.3496	1.4140	1.4608	1.4932	1.5122	-
$\varepsilon = 2^{-12}$	1.5610e-02	6.1253e-03	2.2986e-03	8.3507e-04	2.9666e-04	1.0401e-04
	1.3496	1.4140	1.4608	1.4931	1.5121	-
$\varepsilon = 2^{-14}$	1.5610e-02	6.1253e-03	2.2986e-03	8.3507e-04	2.9666e-04	1.0401e-04
	1.3496	1.4140	1.4608	1.4931	1.5121	-
$\varepsilon = 2^{-20}$	1.5610e-02	6.1253e-03	2.2986e-03	8.3507e-04	2.9666e-04	1.0401e-04
	1.3496	1.4140	1.4608	1.4931	1.5121	-
$\varepsilon = 2^{-30}$	1.5610e-02	6.1253e-03	2.2986e-03	8.3507e-04	2.9666e-04	1.0401e-04
	1.3496	1.4140	1.4608	1.4931	1.5121	-
e^{N_h, M_τ}	1.5610e-02	6.1253e-03	2.2986e-03	8.3507e-04	2.9666e-04	1.0401e-04
R^{N_h, M_τ}	1.3496	1.4140	1.4608	1.4931	1.5121	-

1 and 2) demonstrate the ε-uniform convergence of the scheme. A boundary layer, as shown in Figures 1 and 2, is located at the right side of the space domain in the numerical solution of Examples 1 and 2 above. Figures 1 and 2 also display the computed solutions $y_{i,j}$ for various perturbation parameter values, along with the influence of fractional order. Figure 3 displays the log-log plots of the maximum absolute errors against the number of meshes for both cases, demonstrating the developed numerical scheme's convergent nature regardless of the perturbation value. The suggested scheme is ε-uniformly convergent, as illustrated by the numerical results shown in Tables 2 and 4, by combining extended cubic B-spline collocation with artificial viscosity numerical method in the spatial direction with the implicit Euler's method in the temporal direction. We can see that, for each value of ε, the

Table 5: Comparison of maximum error and rate of convergence for Example 2 for different values of ε, with fix $\gamma = 0.5$

$(N_h, M_\tau) \Rightarrow$	(16,20)	(32,40)	(64,80)	(128, 160)
Proposed Method				
$\varepsilon = 2^{-6}$	1.5546e-02	6.0863e-03	2.2785e-03	8.2711e-04
	1.3529	1.4175	1.4620	-
$\varepsilon = 2^{-8}$	1.5610e-02	6.1249e-03	2.2975e-03	8.3410e-04
	1.3497	1.4146	1.4618	-
$\varepsilon = 2^{-10}$	1.5610e-02	6.1253e-03	2.2986e-03	8.3507e-04
	1.3496	1.4140	1.4608	-
$\varepsilon = 2^{-12}$	1.5610e-02	6.1253e-03	2.2986e-03	8.3507e-04
	1.3496	1.4140	1.4608	-
$\varepsilon = 2^{-14}$	1.5610e-02	6.1253e-03	2.2986e-03	8.3507e-04
	1.3496	1.4140	1.4608	-
$\varepsilon = 2^{-20}$	1.5610e-02	6.1253e-03	2.2986e-03	8.3507e-04
	1.3496	1.4140	1.4608	-
$\varepsilon = 2^{-30}$	1.5610e-02	6.1253e-03	2.2986e-03	8.3507e-04
	1.3496	1.4140	1.4608	-
e^{N_h, M_τ}	1.5610e-02	6.1253e-03	2.2986e-03	8.3507e-04
R^{N_h, M_τ}	1.3496	1.4140	1.4608	-
Method in reference ([27])				
$\varepsilon = 2^{-6}$	1.5818e-02	7.8811e-03	2.9140e-03	8.1121E-04
	1.0051	1.4354	1.8449	-
$\varepsilon = 2^{-8}$	2.1516e-02	9.5195e-03	4.7373e-03	2.2942e-03
	1.1765	1.0068	1.0461	
$\varepsilon = 2^{-10}$	2.4877e-02	1.1527e-02	5.4771e-03	2.6471e-03
	1.1098	1.0735	1.0490	-
$\varepsilon = 2^{-12}$	2.5768e-02	1.2070e-02	5.7971e-03	2.8303e-03
	1.0942	1.0580	1.0344	-
$\varepsilon = 2^{-15}$	2.6068e-02	1.2257e-02	5.9063e-03	2.8933e-03
	1.0887	1.0533	1.0295	-
$\varepsilon = 2^{-25}$	2.6069e-02	1.2258e-02	5.9068e-03	2.8935e-03
	1.0886	1.0533	1.0296	-
$\varepsilon = 2^{-30}$	2.6069e-02	1.2258e-02	5.9068e-03	2.8935e-03
	1.0886	1.0533	1.0296	-
e^{N_h, M_τ}	2.6069e-02	1.2258e-02	5.9068e-03	2.8935e-03
R^{N_h, M_τ}	1.0886	1.0533	1.0296	-



(a) Log-log plot of the max error for Example 1

(b) Log-log plot of the max error for Example 2

Figure 3: Log-log plot of maximum absolute errors for Examples 1 and 2 for different values of ε.

maximum point-wise error decreases as N_h, M_τ grows from the results in Tables 2 and 4. It is evident that, for every N_h, M_τ, the maximum point-wise error is $\varepsilon \longrightarrow 0$ stable. By utilizing these two examples, we verify that the suggested numerical technique is more accurate, stable, and ε-uniformly convergent, with a convergence rate that is almost one. The execution of the proposed method is done by using the MATLAB R2022b software package.

7 Conclusion

We solved the time delay singularly perturbed parabolic convection-diffusion problem with the time-fractional order of derivative using the extended cubic B-spline collocation method. The solution to the problem showed a boundary layer on the right side of the spatial domain. The layer region of the solution has a steep gradient due to the existence of ε. Because of the rapidly changing solution behavior in the layer region, it is computationally challenging to determine the solution analytically or using standard numerical approaches. To control this effect, we came up with a plan that makes use of an extended cubic B-spline collocation scheme in the spatial direction and an implicit Euler's scheme in the temporal direction. It has been demonstrated that the developed numerical approach is stable and converges uniformly.

Two model problems have been taken into consideration for the numerical experimentation for various values of the perturbation parameter and fractional order derivatives in order to confirm the method's compatibility. The scheme was shown to have an order of convergence of $O(\frac{N_h^{-2}}{N_h^{-1}+1} + \tau^{2-\gamma})$ and to be ε-uniformly convergent.

Acknowledgements

The authors would like to thank the anonymous reviewers for their helpful feedback and recommendations that helped to make the article more successful.

References

[1] Abel, N.H. *œuvres complètes de Niels Henrik Abel*, Vol. 1. Cambridge University Press, 2012.

[2] Al-Mdallal, Q.M. and Syam, M.I. *An efficient method for solving nonlinear singularly perturbed two points boundary-value problems of fractional order*, Commun. Nonlinear Sci. Numer. Simul. 17(6) (2012), 2299–2308.

[3] Anilay, W.T., Duressa, G.F. and Woldaregay, M.M. *Higher order uniformly convergent numerical scheme for singularly perturbed reaction-diffusion problems*, Kyungpook Math. J. 61(3) (2021), 591–612.

[4] Atangana, A. and Doungmo Goufo, E.F. *Extension of matched asymptotic method to fractional boundary layers problems*, Math. Probl. Eng. 2014(1) (2014), 107535.

[5] Bijura, A.M. *Nonlinear singular perturbation problems of arbitrary real orders*, No. IC–2003/130, Trieste; Italy, The Abdus Salam International Center for Theoretical Physics, 2003.

[6] Choudhary, R., Singh, S. and Kumar, D. *A second-order numerical scheme for the time-fractional partial differential equations with a time delay*, Comput. Appl. Math. 41(3) (2022), 1–28.

[7] Cooke, K.L. *Differential—difference equations,* In International symposium on nonlinear differential equations and nonlinear mechanics, Elsevier, 1963, 155–171.

[8] Daba, I.T. and Duressa, G.F. *Extended cubic b-spline collocation method for singularly perturbed parabolic differential-difference equation arising in computational neuroscience*, Int. J. Numer. Methods Biomed. Eng. 37(2) (2021), e3418.

[9] Daba, I.T. and Duressa, G.F. *Collocation method using artificial viscosity for time dependent singularly perturbed differential–difference equations*, Math. Comput. Simul. 192 (2022), 201–220.

[10] Dağ, İ.D.R.İ.S., Irk, D.U.R.S.U.N. and Sarı, M. *The extended cubic b-spline algorithm for a modified regularized long wave equation*, Chinese Physics B 22(4) (2013), 040207.

[11] Diekmann, O., Van Gils, S.A., Lunel, S.M. and Walther, H.O. *Delay equations: functional-, complex-, and nonlinear analysis*, volume 110. Springer Science & Business Media, 2012.

[12] Driver, R.D. *Ordinary and delay differential equations*, volume 20. Springer Science & Business Media, 2012.

[13] Farrell, P., Hegarty, A., Miller, J.M., O'Riordan, E. and Shishkin, G.I. *Robust computational techniques for boundary layers*, Chapman and hall/CRC, 2000.

[14] Gelu, F.W. and Duressa, G.F. *Hybrid method for singularly perturbed robin type parabolic convection–diffusion problems on Shishkin mesh*, Partial Differ. Equ. Appl. Math. 8 (2023), 100586.

[15] Gerasimov, A.N. *A generalization of linear laws of deformation and its application to problems of internal friction*, Akad. Nauk SSSR. Prikl. Mat. Meh 12(3) (1948), 251–260.

[16] Hailu, W.S. and Duressa, G.F. *Accelerated parameter-uniform numerical method for singularly perturbed parabolic convection-diffusion problems with a large negative shift and integral boundary condition*, Results Appl. Math. 18 (2023), 100364.

[17] Hailu, W.S. and Duressa, G.F. *Uniformly convergent numerical scheme for solving singularly perturbed parabolic convection-diffusion equations with integral boundary condition*, Differ. Equ. Dyn. Syst. (2023), 1–27.

[18] Hailu, W.S. and Duressa, G.F. *A robust collocation method for singularly perturbed discontinuous coefficients parabolic differential difference equations*, Res. Math. 11(1) (2024), 2301827.

[19] Hale, J.K. and Lunel, S.M.V. *Introduction to functional differential equations*, volume 99. Springer Science and Business Media, 2013.

[20] Hall, C.A. *On error bounds for spline interpolation*, J. Approx. Theory 1(2) (1968), 209–218.

[21] Kaslik, E. and Sivasundaram, S. *Analytical and numerical methods for the stability analysis of linear fractional delay differential equations*, J. Comput. Appl. Math. 236(16) (2012), 4027–4041.

[22] Kolmanovskii, V. and Myshkis, A. *Applied theory of functional differential equations*, volume 85. Springer Science and Business Media, 2012.

[23] Kolmanovskii, V.B. and Nosov, V.R. *Stability of functional differential equations*, volume 180, Elsevier, 1986.

[24] Kuang, Y. *Delay differential equations: with applications in population dynamics*, Academic press, 1993.

[25] 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.

[26] Kumar, D. and Kadalbajoo, M.K. *A parameter-uniform numerical method for time-dependent singularly perturbed differential-difference equations*, Appl. Math. Model. 35(6) (2011), 2805–2819.

[27] Kumar, K., P, P.C. and Vigo-Aguiar, J. *Numerical solution of time-fractional singularly perturbed convection–diffusion problems with a delay in time*, Mathematical Methods in the Applied Sciences, 44(4) (2021), 3080–3097.

[28] Miller, J.J., O'riordan, E. 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.

[29] Miller, K.S. and Ross, B. *An introduction to the fractional calculus and fractional differential equations*, Willey New York, 1993.

[30] 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.

[31] 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.

[32] Nelson, P.W. and Perelson, A.S. *Mathematical analysis of delay differential equation models of HIV-1 infection*, Math. Biosci. 179(1) (2002), 73–94.

[33] Norkin, S.B. *Introduction to the theory and application of differential equations with deviating arguments*, Academic Press, 1973.

[34] O'malley, R.E. *Singular perturbation methods for ordinary differential equations*, volume 89. Springer, 1991.

[35] Prabhakar, T.R. *A singular integral equation with a generalized Mittag Leffler function in the kernel*, Yokohama Math. J. 19(1) (1971), 7–15.

[36] Roop, J.P. *Numerical approximation of a one-dimensional space fractional advection–dispersion equation with boundary layer*, Comput. Math. Appl. 56(7) (2008), 1808–1819.

[37] Sahoo, S.K. and Gupta, V. *A robust uniformly convergent finite difference scheme for the time-fractional singularly perturbed convection-diffusion problem*, Comput. Math. Appl. 137 (2023), 126–146.

[38] Sayevand, K. and Pichaghchi, K. *Efficient algorithms for analyzing the singularly perturbed boundary value problems of fractional order*, Commun. Nonlinear Sci. Numer. Simul. 57 (2018), 136–168.

[39] Sayevand, K. and Pichaghchi, K. *A novel operational matrix method for solving singularly perturbed boundary value problems of fractional multi-order*, Int. J. Comput. Math. 95(4) (2018), 767–796.

[40] Varah, J.M. *A lower bound for the smallest singular value of a matrix*, Linear Algebra Appl. 11(1) (1975), 3–5.

[41] Villasana, M. and Radunskaya, A. *A delay differential equation model for tumor growth*, J. Math. Biol. 47 (2003), 270–294.

[42] Xu, G., Wang, G.Z. and Chen, X.D. *Free-form deformation with rational dms-spline volumes*, J. Comput. Sci. Technol. 23(5) (2008), 862–873,

[43] Xuli, H. *An extension of the cubic uniform b-spline curve*, Journal of Computer Aided Design and Computer Graphics 15(5) (2003), 576–578.

[44] Zhao, T. *Global periodic-solutions for a differential delay system modeling a microbial population in the chemostat*, J. Math. Anal. Appl. 193(1) (1995), 329–352.

Iranian Journal of Numerical Analysis and Optimization
Vol. 14, No. 3, 2024, pp 796–832
https://doi.org/10.22067/ijnao.2024.87084.1403
https://ijnao.um.ac.ir/

Research Article

Finite element analysis for microscale heat equation with Neumann boundary conditions

M.H. Hashim and A.J. Harfash*

Abstract

In this paper, we explore the numerical analysis of the microscale heat equation. We present the characteristics of numerical solutions obtained through both semi- and fully-discrete linear finite element methods. We establish a priori estimates and error bounds for both semi-discrete and fully-discrete finite element approximations. Additionally, the existence and uniqueness of the semi-discrete and fully-discrete finite element approximations have been confirmed. The study explores error bounds in various spaces, comparing the semi-discrete to the exact solutions, the semi-discrete against the fully-discrete solutions, and the fully-discrete solutions

*Corresponding author

Received 29 February 2024; revised 22 May 2024; accepted 13 June 2024

Mohammed Homod Hashim
Department of Mathematics, College of Sciences, University of Basrah, Basrah, Iraq.
e-mail: mohammed.hmmod@uobasrah.edu.iq

Akil Jassim Harfash
Department of Mathematics, College of Sciences, University of Basrah, Basrah, Iraq.
e-mail: akil.harfash@uobasrah.edu.iq

How to cite this article
Hashim, M.H. and Harfash, A.J., Finite element analysis for microscale heat equation with Neumann boundary conditions. *Iran. J. Numer. Anal. Optim.*, 2024; 14(3): 796-832. https://doi.org/10.22067/ijnao.2024.87084.1403

with the exact ones. A practical algorithm is introduced to address the system emerging from the fully-discrete finite element approximation at every time step. Additionally, the paper presents numerical error calculations to further demonstrate and validate the results.

AMS subject classifications (2020): Primary 65M60; Secondary 65M12, 35D30.

Keywords: Finite element; Microscale heat equation; Convergence; Weak solution.

1 Introduction

The microscale heat transport equation holds significance as a crucial model in microtechnology. Diverging from the classical heat diffusion model, microscale heat transport incorporates temperature derivatives of second and third order concerning time and space, introducing a more intricate representation of heat transfer dynamics at the microscale. The equation governing the microscale heat transport, capturing the thermal characteristics of thin films and other microstructures, is expressed as follows: [32]:

$$\frac{1}{\alpha}(\gamma_t + \gamma_q\gamma_{tt}) = \gamma_q\Delta\gamma_t + \Delta\gamma + s, \qquad \text{in} \quad \Im \times [0, \Re], \tag{1}$$

$$\frac{\partial\gamma}{\partial\nu} = 0, \qquad \text{on} \quad \partial\Im \times [0, \Re], \tag{2}$$

$$\gamma(\cdot, 0) = \gamma_1^0, \qquad \gamma_t(\cdot, 0) = \gamma_2^0 \qquad \text{in} \quad \Im, \tag{3}$$

where $\Im$ is an open bounded domain in $\mathbb{R}^n$ $(n = 1, 2, 3)$, γ is the temperature, α and γ_q are positive constants. Here α is the thermal diffusivity. Also, γ_q's represent the time lag of the heat flux and the temperature gradient [31].

The microscale heat transport equation serves as a mathematical representation to elucidate heat transfer phenomena occurring at extremely small scales, predominantly in micro- and nanoscale systems. This equation holds significance as a foundational tool in the realm of microscale heat transfer, finding applications in diverse areas such as the phonon-electron interaction model [25], the single energy equation [27, 28], the phonon scattering model [20], the phonon radiative transfer model [21], and the lagging behavior model

[27, 26]. These models can be effectively described and analyzed using the microscale heat transport equation.

A limited number of researchers have addressed the numerical solution of the one-dimensional microscale heat transport equation. Qui and Tien [24] employed the Crank–Nicolson technique to solve the phonon-electron interaction model. Joshi and Majumdar [21] utilized an explicit upstream difference method to address the phonon radiative transfer model in a one-dimensional medium. Zhang and Zhao [31] tackled the one-dimensional microscale heat transport equation using a fourth-order compact scheme, demonstrating its unconditional stability. In their works [33, 32], they extended their approach to solving the two- and three-dimensional microscale heat transport equations with second-order accuracy in both time and space. Additionally, a compact finite difference scheme with fourth-order spatial accuracy and second-order temporal accuracy for the three-dimensional microscale heat transport equation was developed by Harfash [11].

Recently, a compact difference scheme for the microscale heat transport equation was formulated in [6]. This scheme demonstrated superior precision compared to a previously suggested method, offering higher-order accuracy. The study established the unconditional stability and convergence of the developed compact difference scheme. In the work presented in [22], the application of the localized radial basis function partition of the unity method has been investigated for solving the microscale heat transport equation. The proposed algorithm involves a two-phase discretization of the unknown solution. In the study documented in [1], the focus was on developing a novel meshless numerical approach for solving the heat transport equation. To achieve this objective, the Crank–Nicolson finite difference method was employed to discretize the time derivative. The time-discrete scheme's unconditional stability and convergence were subsequently verified through an energy method.

It is worth noting that all previous numerical studies to solve the microscale heat transport equation were not concerned with performing theoretical analysis of the numerical solutions in terms of studying the solution spaces and the convergence of the numerical solution to the analytical solution. Also, the study of error in previous studies was incomplete due to the

failure to specify the numerical solution spaces. In this study, the system will be approximated by the finite element method for space and the finite difference method for time. A comprehensive study of the numerical solution will be conducted. The study includes finding the spaces of the numerical solution and studying the error. In addition, we perform the convergence analysis of the approximate weak form to the continuous weak form. Many recent studies have involved the use of the finite element method to solve various problems [5, 29, 23, 4, 10, 30].

To address our approaches, we see that under the transformation $\psi = \gamma + \gamma_q \gamma_t$, the system (1)–(3) becomes as follows:
(Λ) Find $\{\gamma, \psi\}$ such that

$$\frac{1}{\alpha}\partial_t \psi = \Delta \psi + s, \qquad \text{in } \Im \times [0, \Re], \tag{4}$$

$$\gamma_q \partial_t \gamma = \psi - \gamma, \qquad \text{in } \Im \times [0, \Re], \tag{5}$$

$$\frac{\partial \gamma}{\partial \nu} = 0, \qquad \text{on } \partial\Im \times [0, \Re], \tag{6}$$

$$\gamma(\cdot, 0) = \gamma_1^0, \qquad \psi(\cdot, 0) = \gamma_1^0 + \gamma_q \gamma_2^0 := \psi^0, \qquad \text{in } \Im. \tag{7}$$

Next, we introduce a weak formulation of the system (4) and (5) in the following form:
(Λ) Find $\psi(\mathbf{x}, t), \gamma(\mathbf{x}, t) \in H^1(\Im)$ such that $\psi(\mathbf{x}, 0) = \psi^0(\mathbf{x}), \gamma(\mathbf{x}, 0) = \gamma_1^0(\mathbf{x})$ and for a.e. $t \in (0, \Re)$ and for all $\Upsilon \in H^1(\Im)$,

$$\frac{1}{\alpha}(\partial_t \psi, \Upsilon) + (\nabla \psi, \nabla \Upsilon) = s(1, \Upsilon), \qquad \text{in } \Im \times [0, \Re], \tag{8}$$

$$\gamma_q(\partial_t \gamma, \Upsilon) + (\gamma, \Upsilon) = (\psi, \Upsilon), \qquad \text{in } \Im \times [0, \Re]. \tag{9}$$

This passage outlines the structure of the paper. Section 2 is dedicated to defining the notation used throughout this study. In section 3, the semi-discrete approximation of Problem (Λ) is discussed, with subsection 3.1 focusing on global existence and subsection 3.2 on uniqueness. Subsection 3.3 presents the error bounds associated with the semi-discrete approximation. The fully-discrete finite element approximation of the Problem (Λ) is covered in section 4, where stability bounds are also derived. The following subsections, 4.1, 4.2, and 4.3, analyze the existence, uniqueness, and convergence of the solution, respectively, and subsection 4.4 examines the error estimates

for the fully-discrete approximation. Lastly, section 5 describes a numerical algorithm for implementing the fully-discrete approximation of the Problem (Λ), including calculations of numerical errors.

2 Finite element spaces and associated results

Consider the finite element space $S^\hbar$, which is a subset of $H^1(\Im)$, defined by

$$S^\hbar := \{\varepsilon \in C(\overline{\Im}) : \varepsilon|_\tau \text{ is linear } \text{ for all } \tau \in \mathcal{T}^\hbar\}.$$

The set $\{\kappa_j\}_{j=1}^J$ represents the standard basis functions for $S^\hbar$, which adhere to the property that $\kappa_j(\varsigma_i) = \delta_{ij}$ for all $i, j = 1, \ldots, J$. Here, $\mathcal{N}^\hbar := \{\varsigma_j\}_{j=0}^J$ denotes the collection of nodes corresponding to the partition $\mathcal{T}^\hbar$. Additionally, we introduce

$$S^\hbar_{\geq 0} := \{\varepsilon \in S^\hbar : \varepsilon(\varsigma_j) \geq 0, j = 1, \ldots, J\}$$

$$\subset H^1_{\geq 0} := \{\varepsilon \in H^1(\Im) : \varepsilon \geq 0 \ a.e. \in \Im\}.$$

The operator $\Pi^\hbar : C(\overline{\Im}) \to S^\hbar$ represents the Lagrange interpolation operator, which can also be referred to as the piecewise linear interpolant. This operator ensures that

$$\Pi^\hbar \varepsilon(\varsigma_j) := \varepsilon(\varsigma_j), \quad \text{for } \ j = 1, \ldots, J.$$

Furthermore, we introduce a discrete L^2 inner (or semi-inner) product on $S^\hbar(C(\overline{\Im}))$ as

$$(u, v)^\hbar := \int_{\Im} \Pi^\hbar(u(\mathbf{x})v(\mathbf{x}))d\mathbf{x} = \sum_{j=1}^{J} \widehat{M}_{jj} \ u(\varsigma_j)\, v(\varsigma_j), \tag{10}$$

where $\widehat{M}_{jj} = (\kappa_j, \kappa_j)^\hbar = (1, \kappa_j) > 0$. By observing (10), it is straightforward to confirm that

$$(\varepsilon_1, \varepsilon_2)^\hbar = (\Pi^\hbar \varepsilon_1, \varepsilon_2)^\hbar = (\Pi^\hbar \varepsilon_1, \Pi^\hbar \varepsilon_2)^\hbar \qquad \text{for all } \varepsilon_1, \varepsilon_2 \in C(\overline{\Im}).$$

In the context of the finite element space $S^\hbar$, several established results are notable. The induced discrete semi-norm on $C(\overline{\Im})$ and the norm on $S^\hbar$ are

both represented by $|\cdot|_h$, which is defined by the expression $[(\cdot,\cdot)^\hbar]^{1/2}$. It has been proven that the semi-norm $|\cdot|_h$ is equivalent to the norm $\|\cdot\|_0$, which is defined as $[(\cdot,\cdot)]^{1/2}$. This relationship can be described as

$$\|\vartheta\|_0^2 \le |\vartheta|_\hbar^2 \le (\ell+2)\|\vartheta\|_0^2 \qquad \text{for all } \vartheta \in S^\hbar. \tag{11}$$

The Poincaré inequality, given that $\hbar$ is sufficiently small, can be expressed in the following form:

$$((\zeta,\zeta)^\hbar)^{\frac{1}{2}} = |\zeta|_\hbar \le C_p(|\zeta|_1 + |(\zeta,1)^\hbar|). \tag{12}$$

We define, for any $\lambda(x) \in S^\hbar$, that

$$|\lambda^\hbar|_{\hbar,\zeta} := (\int_\Im \Pi^\hbar\{|\lambda(x)^\hbar|^\zeta dx)^{\frac{1}{\zeta}} \equiv (\sum_{i=0}^{k} \widehat{M}_{jj}\lambda(x_i)^\hbar|^\zeta)^{\frac{1}{\zeta}} \quad if \quad 0 \le \zeta < \infty,$$

and

$$|\lambda^\hbar|_{\hbar,\zeta} := \max_{0\le j\le k} |\lambda(x_j)^\hbar| \qquad\qquad if \quad \zeta = \infty.$$

We now revisit some well-established results concerning the space $S^\hbar$ under the assumption that $\mathcal{T}^\hbar$ forms a quasi-uniform partitioning: For any $\tau \in \mathcal{T}^\hbar, \xi \in S^\hbar, 1 \le p,q \le \infty$ and $m,l \in \{0,1\}$ with $l \le m$, we have

$$\|\xi\|_{m,p,"\tau"} \le C\hbar_{"\tau"}^{l-m+\ell\min(0,\frac{1}{p}-\frac{1}{q})}\|\xi\|_{l,q,"\tau"},$$

where the abbreviation "τ" means "with" or "without" τ. The inequality stated above is commonly referred to as "the inverse inequality, " as documented in . 75–771[9]. Additionally, it remains valid when replacing $\|\cdot\|$ with $|\cdot|$, as indicated in . 140–1421[7].

For future reference, we introduce the subsequent inverse inequalities, derived from the quasi-uniform condition as outlined in of [7, Theorem 3.2.6],

$$|\xi|_{1,p,"\tau"} \le C\hbar_{"\tau"}^{-1}|\xi|_{0,p,"\tau"}, \quad 1 \le p \le \infty,$$

$$|\xi|_{m,p,"\tau"} \le C\hbar_{"\tau"}^{-\ell(\frac{1}{q}-\frac{1}{p})}|\xi|_{m,q,"\tau"}, \quad 1 \le q \le p \le \infty, \quad m \in \{0,1\}.$$

We also need the following interpolation results for every $\xi \in W^{1,s}(\Im)$, where $s \in [2,\infty]$ if $\ell = 1$, and $s \in (\ell,\infty]$ if ℓ is either 2 or 3:

$$|(I-\Pi^{\hbar})\xi|_{m,s} \leq C\hbar^{1-m}|\xi|_{1,s}, \quad m \in \{0,1\}, \tag{13}$$

$$\lim_{\hbar\to 0} |(I-\Pi^{\hbar})\xi|_{1,s} = 0, \tag{14}$$

(see [9, Theorem 1.103 and Corollary 1.110], respectively). We also bring to mind the following useful result (e.g., [8]): For any $\xi_1, \xi_2 \in S^{\hbar}$, we have

$$|(\xi_1,\xi_2) - (\xi_1,\xi_2)^{\hbar}| \leq C\hbar^{1+m}|\xi_1|_{m,n_1}|\xi_2|_{1,n_2}, \tag{15}$$

for $m \in \{0,1\}$ and $1 \leq n_1, n_2 \leq \infty$ with $\frac{1}{n_1} + \frac{1}{n_2} = 1$.

We can conveniently introduce the "inverse Laplacian Green's operator" denoted as $\mathcal{G} : (H^1(\Im))' \to H^1(\Im)$ such that

$$(\nabla\mathcal{G}\breve{v}_1, \nabla\breve{v}_2) = \langle \breve{v}_1, \breve{v}_2\rangle \qquad \text{for all} \quad \breve{v}_2 \in H^1(\Im),$$

where $\langle\cdot,\cdot\rangle$ denotes the duality pairing between $(H^1(\Im))'$ and $H^1(\Im)$ such that

$$\langle \hbar, \breve{v}\rangle = (\hbar, \breve{v}) \quad \text{for all} \quad \hbar \in L^2(\Im) \text{ and } \breve{v} \in H^1(\Im). \tag{16}$$

3 A semi-discrete approximation

We define the following semi-discrete approximation of the system (4)–(7):
$(\Lambda^{\hbar})$ Find $\{\gamma^{\hbar}, \psi^{\hbar}\} \in S^{\hbar} \times S^{\hbar}$ such that for a.e. $t \in (0, \Re)$

$$\frac{1}{\alpha}(\frac{\partial\psi^{\hbar}}{\partial t}, \Upsilon^{\hbar})^{\hbar} + (\nabla\psi^{\hbar}, \nabla\Upsilon^{\hbar}) = s(1,\Upsilon^{\hbar})^{\hbar} \qquad \text{for all } \Upsilon^{\hbar} \in S^{\hbar}, \tag{17}$$

$$\gamma_q(\frac{\partial\gamma^{\hbar}}{\partial t}, \Upsilon^{\hbar}) + (\gamma^{\hbar}, \Upsilon^{\hbar}) = (\psi^{\hbar}, \Upsilon^{\hbar}) \qquad \text{for all } \Upsilon^{\hbar} \in S^{\hbar}, \tag{18}$$

$$\gamma^{\hbar}(\mathbf{x}, 0) = \mathbb{P}^{\hbar}\gamma(\mathbf{x}). \tag{19}$$

3.1 Global existence

Theorem 1. Suppose that $\Im \subset \mathbb{R}^d$ (with $d \leq 3$) is an open, bounded, convex domain. Let $\gamma^0 \in H^1(\Im)$. Then, the system (17)–(19) admits a solution $\psi^{\hbar}, \gamma^{\hbar}$ that satisfies

$$\psi^{\hbar} \in L^{\infty}(0,\Re;H^{1}(\Im)) \cap L^{\infty}(0,\Re;L^{2}(\Im)) \cap L^{2}(0,\Re;H^{1}(\Im)) \cap L^{2}(\Im_{\Re}),$$

$$\gamma^{\hbar} \in L^{2}(0,\Re;H^{1}(\Im)) \cap L^{\infty}(0,\Re;L^{2}(\Im)) \cap L^{\infty}(\Im_{\Re}),$$

$$\frac{\partial \psi^{\hbar}}{\partial t} \cap \frac{\partial \gamma^{\hbar}}{\partial t} \in L^{2}(\Im_{\Re}).$$

Proof. Selecting $\Upsilon^{\hbar} = \psi^{\hbar}$ in (17) and $\Upsilon^{\hbar} = \gamma^{\hbar}$ in (18), we obtain

$$\frac{1}{\alpha}(\frac{\partial \psi^{\hbar}}{\partial t}, \psi^{\hbar})^{\hbar} + (\nabla \psi^{\hbar}, \nabla \psi^{\hbar}) = s(1, \psi^{\hbar})^{\hbar}, \tag{20}$$

$$\gamma_q(\frac{\partial \gamma^{\hbar}}{\partial t}, \gamma^{\hbar}) + (\gamma^{\hbar}, \gamma^{\hbar}) = (\psi^{\hbar}, \gamma^{\hbar}). \tag{21}$$

Now, by utilizing (20) and (21), we derive the following result:

$$\frac{1}{2\alpha}\frac{d}{dt}|\psi^{\hbar}|_{\hbar}^{2} + \frac{\gamma_q}{2}\frac{d}{dt}\|\gamma^{\hbar}\|_0^2 + \|\gamma^{\hbar}\|_0^2 + |\nabla \psi^{\hbar}|_0^2 = \int_{\Im} \Pi^{\hbar}(\psi^{\hbar}, \gamma^{\hbar})\, d\mathbf{x} + s\int_{\Im} \Pi^{\hbar}\psi^{\hbar}\, d\mathbf{x}.$$

Through the application of Young's inequality, Hölder's inequality, and (11), it can be established that

$$\begin{aligned}\frac{d}{2dt}\left[\frac{1}{\alpha}|\psi^{\hbar}|_{\hbar}^{2} + \gamma_q\|\gamma^{\hbar}\|_0^2\right] + \|\gamma^{\hbar}\|_0^2 + \|\nabla \psi^{\hbar}\|_0^2 &\le C\left[\|\psi^{\hbar}\|_0^2 + \|\gamma^{\hbar}\|_0^2\right] + C(|\Im|, s) \\ &\le C\left[|\psi^{\hbar}|_{\hbar}^{2} + \|\gamma^{\hbar}\|_0^2\right] + C(|\Im|, s).\end{aligned}$$

By applying the Grönwall lemma and integrating equation (21) over the time interval $(0, \Re)$, we obtain

$$\begin{aligned}&\frac{1}{2\alpha}|\psi^{\hbar}(\Re)|_{\hbar}^{2} + \frac{\gamma_q}{2}\|\gamma^{\hbar}(\Re)\|_0^2 + \int_0^{\Re} |\gamma^{\hbar}|_{\hbar}^{2}\, dt + \int_0^{\Re} |\nabla \psi^{\hbar}|_{\hbar}^{2}\, dt \\ &\le \frac{1}{2\alpha}|\psi^{\hbar}(0)|_{\hbar}^{2} + \frac{\gamma_q}{2}\|\gamma^{\hbar}(0)\|_0^2 + C(|\Im|, s).\end{aligned} \tag{22}$$

From (22), it follows that

$$\begin{aligned}\|\psi^{\hbar}\|_{L^{\infty}(0,\Re;L^{2}(\Im))} \le C, \qquad \|\gamma^{\hbar}\|_{L^{\infty}(0,\Re;L^{2}(\Im))} \le C, \\ \|\gamma^{\hbar}\|_{L^{2}(\Im_{\Re})} \le C, \qquad \|\psi^{\hbar}\|_{L^{2}(0,\Re;H^{1}(\Im))} \le C.\end{aligned} \tag{23}$$

Now, by choosing $\Upsilon^{\hbar} = \frac{\partial \psi^{\hbar}}{\partial t}$ in (17), we have

$$\frac{1}{\alpha}\|\frac{\partial \psi^{\hbar}}{\partial t}\|_0^2 + \frac{d}{2dt}\|\nabla \psi^{\hbar}\|_0^2 = s(1, \frac{\partial \psi^{\hbar}}{\partial t})^{\hbar}.$$

Through the application of Hölder and Young's inequalities, we can state that

$$\frac{1}{2\alpha}\|\frac{\partial\psi^{\hbar}}{\partial t}\|_0^2+\frac{d}{2dt}\|\nabla\psi^{\hbar}\|_0^2\leq C(s^2,|\Im|,\alpha). \tag{24}$$

Integrating both sides of (24) from 0 to t results in

$$\frac{1}{2\alpha}\int_0^{\Re}\|\frac{\partial\psi^{\hbar}}{\partial t}\|_0^2\,dt+\frac{1}{2}\|\nabla\psi^{\hbar}(T)\|_0^2\leq\frac{1}{2}\|\nabla\psi^{\hbar}(0)\|_0^2+C(s^2,|\Im|,\alpha). \tag{25}$$

From (25) and considering the assumptions that $\psi^0\in H^1(\Im)$, it can be concluded that

$$\|\frac{\partial\psi^{\hbar}}{\partial t}\|_{L^2(\Im_{\Re})}\leq C,\qquad \|\psi^{\hbar}\|_{L^2(0,\Re;H^1(\Im))}\leq C. \tag{26}$$

Now, by choosing $\Upsilon^{\hbar}=\frac{\partial\gamma^{\hbar}}{\partial t}$ in (18), we obtain that

$$\gamma_q\|\frac{\partial\gamma^{\hbar}}{\partial t}\|_0^2+\frac{d}{2dt}\|\gamma^{\hbar}\|_0^2\leq(\psi^{\hbar},\frac{\partial\gamma^{\hbar}}{\partial t})^{\hbar}. \tag{27}$$

Through the application of Young's inequality, we can express that

$$\frac{\gamma_q}{2}\|\frac{\partial\gamma^{\hbar}}{\partial t}\|_0^2+\frac{d}{2dt}\|\gamma^{\hbar}\|_0^2\leq\frac{1}{2\gamma_q}|\psi^{\hbar}|_{\hbar}^2. \tag{28}$$

Integrating (28) over $(0,1)$ and utilizing (28), along with the observation that $L^{\infty}(0,\Re,L^2(\Im))\hookrightarrow L^2(\Im_{\Re})$, we deduce that

$$\frac{\gamma_q}{2}\int_0^{\Re}\|\frac{\partial\gamma^{\hbar}}{\partial t}\|_0^2\,dt+\frac{1}{2}\|\gamma^{\hbar}(\Re)\|_0^2\leq\frac{1}{2}\|\gamma^{\hbar}(0)\|_0^2+\frac{1}{2\gamma_q}\int_0^{\Re}|\psi^{\hbar}|_{\hbar}^2\,dt. \tag{29}$$

From (29) and considering the assumptions that $\gamma_1^0\in H^1(\Im)$, it can be concluded that

$$\|\frac{\partial\gamma^{\hbar}}{\partial t}\|_{L^2(\Im_{\Re})}\leq C,\qquad \|\gamma^{\hbar}\|_{L^{\infty}(\Im_{\Re})}\leq C.$$

Now, by setting $\Upsilon^{\hbar}=\Delta\gamma^{\hbar}$ in (18), it follows that

$$\frac{\gamma_q}{2}\frac{d}{dt}\|\nabla\gamma^{\hbar}\|_0^2+\|\nabla\gamma^{\hbar}\|_0^2=(\nabla\psi^{\hbar},\nabla\gamma^{\hbar}).$$

By applying Young's inequality, we arrive at the conclusion that

$$\frac{\gamma_q}{2}\frac{d}{dt}\|\nabla\gamma^{\hbar}\|_0^2+\frac{1}{2}\|\nabla\gamma^{\hbar}\|_0^2\leq\frac{1}{2}\|\nabla\psi^{\hbar}\|_0^2. \tag{30}$$

When integrating both sides of (30) from 0 to t, the result is

$$\gamma_q\|\nabla\gamma^\hbar(\Re)\|_0^2+\int_0^\Re\|\nabla\gamma^\hbar\|_0^2dt\leq\int_0^\Re\|\nabla\psi^\hbar\|_0^2dt+\gamma_q\|\nabla\gamma^\hbar(0)\|_0^2. \tag{31}$$

From (23) and (31), considering the assumptions that $\gamma_1^0\in H^1(\Im)$, it can be concluded that

$$\|\psi^\hbar\|_{L^\infty(0,\Re;H^1(\Im))}\leq C,\ \ \|\gamma^\hbar\|_{L^2(0,\Re;H^1(\Im))}\leq C.$$

□

3.2 Uniqueness

Let $\psi_1^\hbar,\psi_2^\hbar$ and $\gamma_1^\hbar,\gamma_2^\hbar$ represent two sets of solutions from the semi-discrete approximations given by (17) and (18), respectively. By defining $\psi^\hbar=\psi_1^\hbar-\psi_2^\hbar$ and $\gamma^\hbar=\gamma_1^\hbar-\gamma_2^\hbar$ and subsequently subtracting the semi-discrete approximations ($\Im^\hbar$), we obtain the following result:

$$\frac{1}{\alpha}(\frac{\partial\psi^\hbar}{\partial t},\Upsilon^\hbar)^\hbar+(\nabla\psi^\hbar,\nabla\Upsilon^\hbar)=0, \tag{32}$$

and

$$\gamma_q(\frac{\partial\gamma^\hbar}{\partial t},\Upsilon^\hbar)+(\gamma^\hbar,\Upsilon^\hbar)=(\psi^\hbar,\Upsilon^\hbar). \tag{33}$$

By selecting $\Upsilon^\hbar=\alpha\psi^\hbar$ for (32) and $\Upsilon^\hbar=\frac{1}{\gamma_q}\gamma^\hbar$ for (33), respectively, we deduce the that

$$\frac{d}{2dt}|\psi^\hbar|_\hbar^2+\alpha|\psi^\hbar|_1^2=0, \tag{34}$$

and

$$\frac{d}{2dt}\|\gamma^\hbar\|_0^2+\frac{1}{\gamma_q}\|\gamma^\hbar\|_0^2=\frac{1}{\gamma_q}(\psi^\hbar,\gamma^\hbar). \tag{35}$$

From (34) and (35), it follows that

$$\frac{d}{2dt}|\psi^\hbar|_\hbar^2+\frac{d}{2dt}\|\gamma^\hbar\|_0^2+\alpha|\psi^\hbar|_1^2+\frac{1}{\gamma_q}\|\gamma^\hbar\|_0^2=\frac{1}{\gamma_q}(\psi^\hbar,\gamma^\hbar).$$

Applying Young's inequality along with (11), we deduce that

$$\frac{d}{2dt}[|\psi^\hbar|_\hbar^2+\|\gamma^\hbar\|_0^2]+\alpha|\psi^\hbar|_1^2+\frac{1}{\gamma_q}\|\gamma^\hbar\|_0^2\leq\frac{1}{2\gamma_q}\|\psi^\hbar\|_0^2+\frac{1}{2\gamma_q}\|\gamma^\hbar\|_0^2$$

$$\leq \frac{1}{2\gamma_q}|\psi^\hbar|_\hbar^2 + \frac{1}{2\gamma_q}\|\gamma^\hbar\|_0^2.$$

The application of Grönwall's lemma leads to the conclusion as

$$|\psi^\hbar|_\hbar^2 + \|\gamma^\hbar\|_0^2 + \alpha\int_0^\Re |\psi^\hbar|_1^2 dt + \frac{1}{\gamma_q}\int_0^\Re \|\gamma^\hbar\|_0^2 dt \leq e^{C\Re}[|\psi^\hbar(0)|_\hbar^2 + \|\gamma^\hbar(0)\|_0^2].$$

Therefore, if $\psi_1^\hbar(0) = \psi_2^\hbar(0)$ and $\gamma_1^\hbar(0) = \gamma_2^\hbar(0)$, then we can deduce the uniqueness of the solutions such that $\psi_1^\hbar(t) = \psi_2^\hbar(t)$ and $\gamma_1^\hbar(t) = \gamma_2^\hbar(t)$ for all t.

3.3 Error estimate

Theorem 2. Assuming that the conditions of Theorem 1 are met and given

$$\|\psi\|_{L^2(0,\Re;H^2(\Im))} \leq C, \tag{36}$$

then the solution $\{\psi, \gamma\}$ adheres to the specified error constraints

$$\|e_\psi\|_{L^\infty(0,\Re;L^2(\Im))} + \|e_\gamma\|_{L^\infty(0,\Re;L^2(\Im))} + \|e_\psi\|_{L^2(0,\Re;H^1(\Im))} + \|e_\gamma\|_{L^2(\Im_\Re)} \leq C\hbar^2, \tag{37}$$

where

$$e_\psi = \psi - \psi^\hbar, \quad e_\gamma = \gamma - \gamma^\hbar. \tag{38}$$

Proof. To begin with, we establish the following definitions:

$$e_\psi^A = \psi - \Pi^\hbar\psi, \quad e_\gamma^A = \gamma - \Pi^\hbar\gamma, \tag{39}$$

$$e_\psi^\hbar = \Pi^\hbar\psi - \psi^\hbar, \quad e_\gamma^\hbar = \Pi^\hbar\gamma - \gamma^\hbar. \tag{40}$$

From (38)–(40), we have that

$$e_\psi^\hbar = e_\psi - e_\psi^A, \quad e_\gamma^\hbar = e_\gamma - e_\gamma^A. \tag{41}$$

By applying (40) and (38)–(40), it is straightforward to determine that

$$\|e_\psi^A\|_0 \leq C\hbar^2\|\psi\|_2, \quad \|e_\gamma^A\|_0 \leq C\hbar^2\|\gamma\|_2,$$

$$|e_\psi^A|_1 \leq C\hbar\|\psi\|_2, \quad |e_\gamma^A|_1 \leq C\hbar\|\gamma\|_2, \tag{42}$$

$$\|e_\psi^A\|_1 \le C\hbar\|\psi\|_2, \quad \|e_\gamma^A\|_1 \le C\hbar\|\gamma\|_2,$$
$$\|e_\psi^\hbar\|_0 \le \|e_\psi\|_0 + C\hbar^2\|\psi\|_2, \quad \|e_\gamma^\hbar\|_0 \le \|e_\gamma\|_0 + C\hbar^2\|\gamma\|_2,$$
$$|e_\psi^\hbar|_1 \le |e_\psi|_1 + C\hbar\|\psi\|_2, \quad |e_\gamma^\hbar|_1 \le |e_\gamma|_1 + C\hbar\|\gamma\|_2,$$
$$\|e_\psi^\hbar\|_1 \le \|e_\psi\|_0 + |e_\psi|_1 + C\hbar\|\psi\|_2, \quad \|e_\gamma^A\|_1 \le \|e_\gamma\|_0 + |e_\gamma|_1 + C\hbar\|\gamma\|_2. \quad (43)$$

Subtracting (8) and (9) from (17) and (18), respectively, and choosing $\Upsilon = \Upsilon^\hbar$ yields that

$$(\partial_t\psi, \Upsilon) - (\partial_t\psi^\hbar, \Upsilon)^\hbar + \alpha(\nabla e_\psi, \nabla\Upsilon) = 0, \quad (44)$$

and

$$(\partial_t\gamma, \Upsilon) - (\partial_t\gamma^\hbar, \Upsilon)^\hbar + \frac{1}{\gamma_q}[(\gamma, \Upsilon) - (\gamma^\hbar, \Upsilon)^\hbar] = \frac{1}{\gamma_q}[(\psi, \Upsilon) - (\psi^\hbar, \Upsilon)^\hbar]. \quad (45)$$

Selecting $\Upsilon = e_\psi^\hbar$ in (44) and $\Upsilon = e_\gamma^\hbar$ in (45) results in

$$(\partial_t\psi, e_\psi^\hbar) - (\partial_t\psi^\hbar, e_\psi^\hbar)^\hbar + \alpha(\nabla e_\psi, \nabla e_\psi^\hbar) = 0, \quad (46)$$

and

$$(\partial_t\gamma, e_\gamma^\hbar) - (\partial_t\gamma^\hbar, e_\gamma^\hbar) + \frac{1}{\gamma_q}[(\gamma, e_\gamma^\hbar) - (\gamma^\hbar, e_\gamma^\hbar)] = \frac{1}{\gamma_q}[(\psi, e_\gamma^\hbar) - (\psi^\hbar, e_\gamma^\hbar)]. \quad (47)$$

By employing (41) and adding and subtracting the terms $(\partial_t\psi^\hbar, e_\psi^\hbar)$ and $(\partial_t\gamma^\hbar, e_\gamma^\hbar)$ to (46), as well as the terms $(\gamma^\hbar, e_\gamma^\hbar)$ and $(\psi^\hbar, e_\gamma^\hbar)$ to (47), it can be deduced that

$$(\partial_t e_\psi, e_\psi^\hbar) + \{(\partial_t\psi^\hbar, e_\psi^\hbar) - (\partial_t\psi^\hbar, e_\psi^\hbar)^\hbar\} + \alpha(\nabla e_\psi, \nabla e_\psi) = \alpha(\nabla e_\psi, \nabla e_\psi^A), \quad (48)$$

and

$$(\partial_t\gamma, e_\gamma^\hbar) - (\partial_t\gamma^\hbar, e_\gamma^\hbar) + \frac{1}{\gamma_q}[(\gamma, e_\gamma^\hbar) - (\gamma^\hbar, e_\gamma^\hbar)] = \frac{1}{\gamma_q}[(\psi, e_\gamma^\hbar) - (\psi^\hbar, e_\gamma^\hbar)]. \quad (49)$$

From (48) and (49), it can be observed that

$$\frac{d}{2dt}\|e_\psi\|_0^2 + \frac{d}{2dt}\|e_\gamma\|_0^2 + \alpha|e_\psi|_1^2 + \frac{1}{\gamma_q}\|e_\gamma\|_0^2$$
$$\le \alpha(\nabla e_\psi, \nabla e_\psi^A) + [(\partial_t\psi^\hbar, e_\psi^\hbar)^\hbar - (\partial_t\psi^\hbar, e_\psi^\hbar)] + \frac{1}{\gamma_q}[(\psi, e_\gamma^\hbar) - (\psi^\hbar, e_\gamma^\hbar)]. \quad (50)$$

We bound each term on the right-hand side of (50) separately. Initially, by employing Cauchy–Schwarz and Young inequalities, along with (42), it can be inferred that

$$\alpha(\nabla e_\psi, \nabla e_\psi^A) \leq \frac{\alpha}{4}|e_\psi|_1^2 + C\hbar^2\|\psi\|_2^2. \tag{51}$$

By applying (15), Young's inequality, and (43), it can be established that

$$[(\partial_t\psi^\hbar, e_\psi^\hbar)^\hbar - (\partial_t\psi^\hbar, e_\psi^\hbar)] \leq C\hbar^2\|\partial_t\psi^\hbar\|_0^2 + \frac{1}{2}\|e_\psi\|_0^2 + \frac{\alpha}{4}|e_\psi|_1^2 + C\hbar^2\|\psi\|_2^2. \tag{52}$$

By employing (40), Young's inequality, and (43), we determine that

$$\frac{1}{\gamma_q}[(\psi, e_\gamma^\hbar) - (\psi^\hbar, e_\gamma^\hbar)^\hbar] = \frac{1}{\gamma_q}(e_\psi^\hbar, e_\gamma^\hbar) \leq \frac{1}{2}\|e_\psi^\hbar\|_0^2 + C\|e_\gamma^\hbar\|_0^2. \tag{53}$$

By inserting equations (51)–(53) into (50), it can be deduced that

$$\begin{aligned}&\frac{d}{2dt}[\|e_\psi\|_0^2 + \|e_\gamma\|_0^2] + \frac{\alpha}{2}|e_\psi|_1^2 + \frac{1}{\gamma_q}\|e_\gamma\|_0^2 \\ &\leq C\|e_\psi\|_0^2 + C\|e_\gamma\|_0^2 + C\hbar^2\|\partial_t\psi^\hbar\|_0^2 + C\hbar^2\|\psi\|_2^2.\end{aligned} \tag{54}$$

Multiplying (54) by 2 and applying Grönwall's lemma, we obtain that

$$\begin{aligned}&\|e_\psi(\Re)\|_0^2 + \|e_\gamma(\Re)\|_0^2 + \alpha\int_0^\Re |e_\psi|_1^2 dt + \frac{2}{\gamma_q}\int_0^\Re \|e_\gamma\|_0^2 dt \\ &\leq e^{2C\Re}[\|e_\psi(0)\|_0^2 + \|e_\gamma(0)\|_0^2] + e^{2C\Re}[C\hbar^2\|\partial_t\psi^\hbar\|_{L^2(\Im_\Re)} \\ &\quad + C\hbar^2\|\psi\|_{L^2(0,\Re;H^2(\Im))}].\end{aligned} \tag{55}$$

In order to bound the right-hand side of (55), initially, we observe from (54) and taking into account that $\psi^0, \gamma_1^0 \in H^1$, yielding that

$$\|e_\psi(0)\|_0^2 \equiv \|\psi^0 - p^\hbar\psi^0\|_0^2 \leq C\hbar^2\psi^0 \leq C\hbar^2,$$

and

$$\|e_\gamma(0)\|_0^2 \equiv \|\gamma_1^0 - p^\hbar\gamma_1^0\|_0^2 \leq C\hbar^2\gamma_1^0 \leq C\hbar^2.$$

Additionally, based on Theorem 1, the remaining terms on the right-hand side of (55) are bounded. Consequently, we ultimately obtain that

$$\|e_\psi\|_{L^\infty(0,\Re;L^2(\Im))} + \|e_\gamma\|_{L^\infty(0,\Re;L^2(\Im))} + \|e_\psi\|_{L^2(0,\Re;H^1(\Im))} + \|e_\gamma\|_{L^2(\Im_\Re)} \leq C\hbar^2.$$

□

4 A fully-discrete approximation

Suppose that N is a positive integer, and let $\Delta t := \frac{\Re}{N}$ represent the time step. We investigate the following fully-discrete finite element approximation of the system (1)–(7):

$(\Lambda^{h,\Delta t})$ For $n \geq 1$ find $\{\Psi^n,\ \Gamma^n\} \in [S^\hbar]^2$ such that for all $\Upsilon \in S^\hbar$

$$\frac{1}{\alpha}\left(\frac{\Psi^n - \Psi^{n-1}}{\Delta t},\ \Upsilon\right)^\hbar + (\nabla\Psi^n,\ \nabla\Upsilon) = s(1,\Upsilon)^\hbar \qquad \text{for all } \Upsilon \in S^\hbar, \quad (56)$$

$$\gamma_q\left(\frac{\Gamma^n - \Gamma^{n-1}}{\Delta t},\ \Upsilon\right) + (\Gamma^n,\ \Upsilon) = (\Psi^n,\ \Upsilon) \qquad \text{for all } \Upsilon \in S^\hbar, \quad (57)$$

$$\Psi^0 = \mathbb{P}^\hbar\psi, \qquad \Gamma_1^0 = \mathbb{P}^\hbar\gamma.$$

Theorem 3. Suppose $\psi^0, \gamma_1^0 \in L^2(\Im)$ with $|\psi^0(\cdot)| \leq 1$ almost everywhere in $\Im$ and $\Gamma^0 \in H^1(\Im) \cap L^2(\Im)$. Then, for all $\Delta t \leq \frac{\gamma_q(1-\theta)}{1+\alpha s\gamma_q}$, where $\theta \in (0,1)$, the problem $(\Lambda^{h,\Delta t})$ has a solution Ψ^n, Γ^n, $n = 1, \ldots, N$, satisfying that

$$\max_{m=1,\ldots,N}[\|\Psi^n\|_\hbar^2 + \|\nabla\Psi^n\|_0^2 + \|\Gamma^n\|_\hbar^2] + \sum_{n=1}^{m}\Delta t\|\nabla\Psi^n\|_0^2 + \sum_{n=1}^{N}\Delta t\|\Gamma^n\|_\hbar^2$$
$$+ \sum_{n=1}^{m}[\|\Psi^n - \Psi^{n-1}\|_\hbar^2 + \|\nabla(\Psi^n - \Psi^{n-1})\|_\hbar^2 + \|\Gamma^n - \Gamma^{n-1}\|_\hbar^2] \leq C.$$

Proof. Selecting $\Upsilon = \alpha\Delta t\Psi^n$ and $\Upsilon = \frac{\Delta t}{\gamma_q}\Gamma^n$ in (56) and (57), respectively, we derive that

$$(\Psi^n - \Psi^{n-1},\ \Psi^n)^\hbar + \alpha\Delta t(\nabla\Psi^n,\ \nabla\Psi^n) = \alpha\Delta t s(1,\Psi^n)^\hbar, \quad (58)$$

$$(\Gamma^n - \Gamma^{n-1},\ \Gamma^n) + \frac{\Delta t}{\gamma_q}(\Gamma^n,\ \Gamma^n) = \frac{\Delta t}{\gamma_q}(\Psi^n,\ \Gamma^n). \quad (59)$$

By inserting (58) into (59), we find that

$$(\Psi^n - \Psi^{n-1},\ \Psi^n)^\hbar + (\Gamma^n - \Gamma^{n-1},\ \Gamma^n) + \frac{\Delta t}{\gamma_q}(\Gamma^n,\ \Gamma^n) + \alpha\Delta t(\nabla\Psi^n,\ \nabla\Psi^n)$$
$$= \alpha\Delta t s(1,\Psi^n)^\hbar + \frac{\Delta t}{\gamma_q}\Gamma^n(\Psi^n,\ \Gamma^n), \quad (60)$$

Utilizing the subsequent straightforward identity,

$$2x(x-y) = x^2 - y^2 + (x-y)^2 \qquad \text{for all } x, y \in R, \tag{61}$$

Hölder and Young's inequalities, and (59)–(60), we find that

$$\begin{aligned}
&[1 - \Delta t(\frac{1}{\gamma_q} + \alpha s)]\|\Psi^n\|_{\hbar}^2 + \|\Gamma^n\|_{\hbar}^2 + \|\Psi^n - \Psi^{n-1}\|_{\hbar}^2 \\
&+ \|\Gamma^n - \Gamma^{n-1}\|_{\hbar}^2 + 2\alpha\Delta t\|\nabla\Psi^n\|_0^2 + \frac{\Delta t}{\gamma_q}|\Gamma^n|_{\hbar}^2 \\
&\leq \|\Psi^{n-1}\|_{\hbar}^2 + \|\Gamma^{n-1}\|_{\hbar}^2 + C(s, \Delta t, \alpha, |\Im|).
\end{aligned}$$

Subsequently, it can be concluded that

$$\begin{aligned}
&[1 - \Delta t(\frac{1 + \alpha\gamma_q s}{\gamma_q})][\|\Psi^n\|_{\hbar}^2 + \|\Gamma^n\|_{\hbar}^2] + \|\Psi^n - \Psi^{n-1}\|_{\hbar}^2 \\
&+ \|\Gamma^n - \Gamma^{n-1}\|_{\hbar}^2 + 2\alpha\Delta t\|\nabla\Psi^n\|_0^2 + \frac{\Delta t}{\gamma_q}|\Gamma^n|_{\hbar}^2 \\
&\leq \|\Psi^{n-1}\|_{\hbar}^2 + \|\Gamma^{n-1}\|_{\hbar}^2 + C(s, \Delta t, \alpha, |\Im|).
\end{aligned} \tag{62}$$

Since $\Psi^n, \Gamma^n \geq 0, \quad \Delta t \leq \frac{\gamma_q(1-\theta)}{1+\alpha s\gamma_q}$, thus we have $\theta \leq 1 - \Delta t(\frac{1+\alpha\gamma_q s}{\gamma_q})$. Then we can find that

$$\frac{1}{1 - \Delta t(\frac{1+\alpha\gamma_q s}{\gamma_q})} = 1 + \frac{\Delta t(\frac{1+\alpha\gamma_q s}{\gamma_q})}{1 - \Delta t(\frac{1+\alpha\gamma_q s}{\gamma_q})} \leq 1 + \frac{\Delta t(1 + \alpha\gamma_q s)}{\gamma_q\theta}.$$

It is deduced from (62) that

$$\begin{aligned}
&\|\Psi^n\|_{\hbar}^2 + \|\Gamma^n\|_{\hbar}^2 + 2\alpha\Delta t\|\nabla\Psi^n\|_0^2 + \frac{\Delta t}{\gamma_q}\|\Gamma^n\|_{\hbar}^2 + \|\Psi^n - \Psi^{n-1}\|_{\hbar}^2 + \|\Gamma^n - \Gamma^{n-1}\|_{\hbar}^2 \\
&\leq \frac{1}{1 - \Delta t(\frac{1+\alpha\gamma_q s}{\gamma_q})}[\|\Psi^{n-1}\|_{\hbar}^2 + \|\Gamma^{n-1}\|_{\hbar}^2 + C] \\
&\leq 1 + \frac{\Delta t(1 + \alpha\gamma_q s)}{\gamma_q\theta}[\|\Psi^{n-1}\|_{\hbar}^2 + \|\Gamma^{n-1}\|_{\hbar}^2 + C].
\end{aligned}$$

Summing the given equation over $n = 1, \ldots, m$ for $m \leq N$ and incorporating assumptions about initial conditions, lead to

$$\begin{aligned}
&\max_{m=1,\ldots,N}[\|\Psi^m\|_{\hbar}^2 + \|\Gamma^m\|_{\hbar}^2] + 2\alpha\sum_{n=1}^{m}\Delta t\|\nabla\Psi^n\|_0^2 + \frac{1}{\gamma_q}\sum_{n=1}^{N}\Delta t|\Gamma^n|_{\hbar}^2 \\
&+ \sum_{n=1}^{m}[\|\Psi^n - \Psi^{n-1}\|_{\hbar}^2 + \|\Gamma^n - \Gamma^{n-1}\|_{\hbar}^2] \leq e^{\frac{\Re(1+\alpha\gamma_q s)}{\gamma_q\theta}}[|\Psi^0|_{\hbar}^2 + |\Gamma^0|_{\hbar}^2 + C] \leq C.
\end{aligned}$$

Opting for $\Upsilon = \alpha\frac{\Psi^n-\Psi^{n-1}}{\Delta t}$ in (56), we deduce that

$$|\frac{\Psi^n-\Psi^{n-1}}{\Delta t}|_{\hbar}^2+\frac{1}{\Delta t}(\nabla\Psi^n,\ \nabla(\Psi^n-\Psi^{n-1}))=\frac{s}{\Delta t}(1,\Psi^n-\Psi^{n-1})^{\hbar}.$$

Utilizing Hölder and Young's inequalities, along with (61), we obtain that

$$|\frac{\Psi^n-\Psi^{n-1}}{\Delta t}|_{\hbar}^2+\frac{\alpha}{\Delta t}|\nabla\Psi^n|_{\hbar}^2+\frac{\alpha}{\Delta t}|\nabla(\Psi^n-\Psi^{n-1})|_{\hbar}^2\leq\frac{\alpha}{\Delta t}|\nabla\Psi^{n-1}|_{\hbar}^2+C(\Delta t,|\Im|,s\alpha).$$

Summing the aforementioned equation over $n=1,\ldots,m$, where $m\leq N$ and employing assumptions regarding initial conditions, result in

$$\begin{aligned}&\max_{m=1,\ldots,N}[\|\nabla\Psi^m\|_{\hbar}^2]+\sum_{n=1}^{m}[|\frac{\Psi^n-\Psi^{n-1}}{\Delta t}|_{\hbar}^2+\frac{\alpha}{\Delta t}|\nabla(\Psi^n-\Psi^{n-1})|_{\hbar}^2]\\&\leq\frac{\alpha}{\Delta t}|\nabla\Psi^0|_{\hbar}^2+C(\Delta t,|\Im|,s,\alpha)\leq C.\end{aligned}$$

□

4.1 Existence of the approximation

Following the methodology similar to what is described in [12, 13, 14, 15, 18, 16, 19, 2, 17, 3], we introduce the functions as follows: $A_\psi : S^{\hbar}\times S^{\hbar}\to S^{\hbar}$ and $A_\gamma : S^{\hbar}\times S^{\hbar}\to S^{\hbar}$ such that for all $\Upsilon\in S^{\hbar}$, we have

$$(A_\psi(\Psi,\Gamma),\ \Upsilon)^{\hbar}=(\Psi-\Psi^{n-1},\ \Upsilon)^{\hbar}+\alpha\Delta t(\nabla\Psi,\ \nabla\Upsilon)-s\alpha\Delta t(1,\ \Upsilon),\quad(63)$$

$$(A_\gamma(\Psi,\Gamma),\ \Upsilon)=(\Gamma-\Gamma^{n-1},\ \Upsilon)+\frac{\Delta t}{\gamma_q}(\Gamma,\ \Upsilon)-\frac{\Delta t}{\gamma_q}(\Psi,\ \Upsilon),\quad(64)$$

respectively. We initially observe that the continuous piecewise linear functions A_ψ and A_γ are distinctly determinable based on their values at the nodal points $\mathcal{N}^{\hbar}$. This uniqueness becomes evident when we set $\Upsilon\equiv\kappa_j$, where $j=0,\ldots,J$, in (63) and (64). Subsequently, we derive solvable square matrix systems as follows:

$$\widehat{M}A_\psi(\Psi,\Gamma)=S_1,\qquad\widehat{M}A_\gamma(\Psi,\Gamma)=S_2,$$

where $\widehat{M}$ is the lumped mass matrix introduced and S_1 and S_2 are given vectors in terms of the nodal values of Ψ, Γ, Ψ^{n-1}, and Γ^{n-1}. Thus, the functions A_ψ and A_γ are well defined.

By considering (63) and (64), the problem $(\Lambda^{h,\Delta t})$ can be reformulated as

For given $\{\Psi^0, \Gamma^0\} \in S^\hbar \times S^\hbar$, find $\{\Psi^n, \Gamma^n\} \in S^\hbar \times S^\hbar$, $n \geq 1$ such that

$$A_\psi(\Psi, \Gamma) = 0, \qquad A_\gamma(\Psi, \Gamma) = 0.$$

Lemma 1. For any given $R > 0$, the functions $A_\psi : [S^\hbar]^2 R \to S^\hbar$ and $A_\gamma : [S^\hbar]^2_R \to S^\hbar$ exhibit continuity in the following manner

$$[S^\hbar]^2_R = \left\{ \{\Upsilon_1, \Upsilon_2\} \in S^\hbar \times S^\hbar : |\Upsilon_1|^2_\hbar + |\Upsilon_2|^2_\hbar \leq R^2 \right\}.$$

Proof. Consider $\Psi_1, \Gamma_1, \Psi_2, \Gamma_2 \in [S^\hbar]^2$. From (63), for all $\Upsilon \in S^\hbar$, we have

$$(A_\psi(\Psi_1, \Gamma_1) - A_\psi(\Psi_2, \Gamma_2),\ \Upsilon)^\hbar = (\Psi_1 - \Psi_2,\ \Upsilon)^\hbar + \alpha \Delta t (\nabla \Psi_1 - \nabla \Psi_2,\ \nabla \Upsilon). \tag{65}$$

Choosing $\Upsilon = A_\Psi(\Psi_1, \Gamma_1) - A_\Psi(\Psi_2, \Gamma_2)$ in (65) yields on noting the Cauchy–Schwarz inequality, (**??**), and (11), that

$$\begin{aligned}
&|A_\psi(\Psi_1, \Gamma_1) - A_\psi(\Psi_2, \Gamma_2)|^2_\hbar = (\Psi_1 - \Psi_2,\ A_\gamma(\Psi_1, \Gamma_1) - A_\psi(\Psi_2, \Gamma_2))^\hbar \\
&+ \Delta t \alpha (\nabla \Psi_1 - \nabla \Psi_2,\ \nabla (A_\psi(\Psi_1, \Gamma_1) - A_\psi(\Psi_2, \Gamma_2))) \\
&\quad \leq |\Psi_1 - \Psi_2|_\hbar |A_\psi(\Psi_1, \Gamma_1) - A_\psi(\Psi_2, \Gamma_2|_\hbar \\
&\qquad + \Delta t \alpha |\Psi_1 - \Psi_2|_1 |A_\psi(\Psi_1, \Gamma_1) - A_\psi(\Psi_2, \Gamma_2|_1 \\
&\quad \leq |\Psi_1 - \Psi_2|_\hbar |A_\psi(\Psi_1, \Gamma_1) - A_\psi(\Psi_2, \Gamma_2|_\hbar \\
&\qquad + C\hbar^{-1} \Delta t \alpha [|\Psi_1 - \Psi_2|_\hbar |A_\psi(\Psi_1, \Gamma_1) - A_\psi(\Psi_2, \Gamma_2|_\hbar] \\
&\qquad + (1 + C\hbar^{-1} \Delta t \alpha)[|\Psi_1 - \Psi_2|_\hbar |A_\psi(\Psi_1, \Gamma_1) - A_\psi(\Psi_2, \Gamma_2|_\hbar].
\end{aligned} \tag{66}$$

Based on (66), it follows that

$$|A_\psi(\Psi_1, \Gamma_1) - A_\psi(\Psi_2, \Gamma_2)|_\hbar \leq C(\hbar^{-1}, \Delta t, \alpha) |\Psi_1 - \Psi_2|_\hbar.$$

From (64), we deduce for all $\Upsilon \in S^\hbar$ that

$$(A_u(\Psi_1,\Gamma_1)-A_\gamma(\Psi_2,\Gamma_2),\Upsilon)=(\Gamma_1-\Gamma_2,\Upsilon)+\frac{\Delta t}{\gamma_q}(\Gamma_1-\Gamma_2,\Upsilon)-\frac{\Delta t}{\gamma_q}(\Psi_1-\Psi_2,\Upsilon). \tag{67}$$

By defining $\Upsilon = A_\gamma(\Psi_1,\Gamma_1) - A_\gamma(\Psi_2,\Gamma_2)$ and utilizing the Cauchy–Schwarz inequality along with (11), we obtain

$$\begin{aligned}&\|A_\gamma(\Psi_1,\Gamma_1) - A_\gamma(\Psi_2,\Gamma_2)\|_0^2\\ &\le (1+\frac{\Delta t}{\gamma_q})\|\Gamma_1-\Gamma_2\|_0\|A_\gamma(\Psi_1,\Gamma_1)-A_\gamma(\Psi_2,\Gamma_2)\|_0\\ &\quad+\frac{\Delta t}{\gamma_q}\|\Psi_1-\Psi_2\|_0\|A_\gamma(\Psi_1,\Gamma_1)-A_\gamma(\Psi_2,\Gamma_2)\|_0.\end{aligned}$$

By employing (11), it follows that

$$\begin{aligned}|A_\gamma(\Psi_1,\Gamma_1) - A_\gamma(\Psi_2,\Gamma_2)|_\hbar^2 \le& C|\Gamma_1-\Gamma_2|_\hbar|A_\gamma(\Psi_1,\Gamma_1)-A_\gamma(\Psi_2,\Gamma_2)|_\hbar\\ &+C|\Psi_1-\Psi_2|_\hbar|A_\gamma(\Psi_1,\Gamma_1)-A_\gamma(\Psi_2,\Gamma_2)|_\hbar.\end{aligned}$$

Hence, we can conclude that

$$|A_\gamma(\Psi_1,\Gamma_1) - A_\gamma(\Psi_2,\Gamma_2)|_\hbar \le C[|\Gamma_1-\Gamma_2|_\hbar + |\Psi_1-\Psi_2|_\hbar]. \tag{68}$$

The outcomes (67) and (68) demonstrate that A_ψ and A_γ, respectively, are Lipschitz continuous. □

Theorem 4. Let $\{\Psi^{n-1},\Gamma^{n-1}\} \in S^\hbar \times S^\hbar$ be given solution to the $(n-1)$th step of $(\Lambda^{h,\Delta t})$ for some $n = 1,2,\dots,N$. Then, for all $h > 0$, and for all $\Delta t \le \frac{1+2\alpha s\gamma_q}{4\gamma_q}$, there exists a solution $\{\Psi^n,\Gamma^n\} \in [S^\hbar]_R^2$ to the nth step of $(\Lambda^{h,\Delta t})$.

Proof. By way of contradiction, assume that for $R > 0$, there is no $\Psi^n,\Gamma^n \in S^\hbar \times S^\hbar$ such that $A_\psi(\Psi,\Gamma) = A_\gamma(\Psi,\Gamma) = 0$. Noting the continuity of $A_\psi(\Psi,\Gamma)$ and $A_\gamma(\Psi,\Gamma)$ on $[S^\hbar]^2R$, we define the continuous function $\mathtt{B} : [S^\hbar]^2R \to [S^\hbar]_R^2$ as

$$\mathtt{B}(\Psi,\Gamma) = (\mathtt{B}_\psi(\Psi,\Gamma),\mathtt{B}_\gamma(\Psi,\Gamma)),$$

where $\mathtt{B}_\psi(\Psi,\Gamma)$ and $\mathtt{B}_\gamma(\Psi,\Gamma)$ are given by

$$\begin{aligned}\mathtt{B}_{\psi}(\Psi,\Gamma) &:= \frac{-RA_{\psi}(\Psi,\Gamma)}{|(A_{\psi}(\Psi,\Gamma),A_{\gamma}(\Psi,\Gamma))|_{S^{\hbar}\times S^{\hbar}}},\\ \mathtt{B}_{\gamma}(\Psi,\Gamma) &:= \frac{-RA_{\psi}(\Psi,\Gamma)}{|(A_{\psi}(\Psi,\Gamma),A_{\gamma}(\Psi,\Gamma))|_{S^{\hbar}\times S^{\hbar}}}.\end{aligned} \tag{69}$$

where $|(\cdot,\cdot)|_{[S^{\hbar}]^2_R}$ is the standard norm on $[S^{\hbar}]^2_R$ defined by

$$|(\mathtt{X}_1,\mathtt{X}_2)|_{S^{\hbar}\times S^{\hbar}} = (\sum_{i=1}^{2}|\mathtt{X}_i|^2_{\hbar})^{\frac{1}{2}}.$$

Observing the continuity of A_{ψ} and A_{γ}, as indicated in Lemma 1, we can deduce that the function $\mathtt{B}$ is continuous. Therefore, considering that $[S^{\hbar}]^2R$ is a convex and compact subset of $S^{\hbar}\times S^{\hbar}$, Schauder's theorem implies the existence of $\Psi,\Gamma\in[S^{\hbar}]^2R$ as a fixed point of $\mathtt{B}$, i.e.,

$$\mathtt{B}(\Psi,\Gamma) = (\mathtt{B}_{\psi}(\Psi,\Gamma),\mathtt{B}_{\gamma}(\Psi,\Gamma)) = (\Psi,\Gamma).$$

Additionally, we observe from (69) that the fixed point Ψ,Γ satisfies

$$|\Psi|^2_{\hbar}+\|\Gamma\|^2_0 = \|\mathtt{B}_{\psi}(\Psi,\Gamma)\|^2_0+|\mathtt{B}_{\gamma}(\Psi,\Gamma)|^2_{\hbar} = (\Psi,\Gamma) = R^2. \tag{70}$$

To establish a contradiction for R sufficiently large, we select $\Upsilon\equiv\Psi$ in (63) and $\Upsilon\equiv\Gamma$ in (64), leading to the determination that

$$(A_{\psi}(\Psi,\Gamma),\ \Psi)^{\hbar} = (\Psi-\Psi^{n-1},\ \Psi)^{\hbar}+\alpha\Delta t(\nabla\Psi,\ \nabla\Psi)-s\alpha\Delta t(1,\ \Psi)^{\hbar}, \tag{71}$$

$$(A_{\gamma}(\Psi,\Gamma),\ \Gamma) = (\Gamma-\Gamma^{n-1},\ \Gamma)+\frac{\Delta t}{\gamma_q}(\Gamma,\ \Gamma)-\frac{\Delta t}{\gamma_q}(\Psi,\ \Gamma). \tag{72}$$

By merging (71) and (72) and considering (61), (11), as well as Hölder and Young's inequalities, we obtain, for R sufficiently large, that

$$\begin{aligned}&(A_{\psi}(\Psi,\Gamma),\ \Psi)^{\hbar}+(A_{\psi}(\Psi,\Gamma),\ \Gamma)^{\hbar}\\ &= (\Psi-\Psi^{n-1},\ \Psi)^{\hbar}+(\Gamma-\Gamma^{n-1},\ \Gamma)^{\hbar}+\alpha\Delta t(\nabla\Psi,\ \nabla\Psi)\\ &\quad+\frac{\Delta t}{\gamma_q}(\Gamma,\ \Gamma)-s\alpha\Delta t(1,\ \Psi)-\frac{\Delta t}{\gamma_q}(\Psi,\ \Gamma)\\ &\geq \frac{1}{2}|\Psi|^2_{\hbar}+\frac{1}{2}|\Psi^{n-1}|^2_{\hbar}+\frac{1}{2}\|\Psi\|^2_0+\frac{1}{2}\|\Gamma^{n-1}\|^2_0+\alpha\Delta t|\Psi|^2_1\\ &\quad+\frac{\Delta t}{\gamma_q}\|\Gamma\|^2_0-\frac{\Delta t}{2\gamma_q}\|\Psi\|^2_0-s\alpha\Delta t|\Psi|^2_{\hbar}\end{aligned}$$

$$\begin{aligned}
&\geq \frac{1}{2}|\Psi|_{\hbar}^2+\frac{1}{2}|\Psi^{n-1}|_{\hbar}^2+\frac{1}{2}|\Gamma|_{\hbar}^2+\frac{1}{2}|\Gamma^{n-1}|_{\hbar}^2+\alpha\Delta t|\Psi|_1^2\\
&\quad+\frac{\Delta t}{\gamma_q}|\Gamma|_{\hbar}^2-\frac{\Delta t}{2\gamma_q}|\Psi|_{\hbar}^2-s\alpha\Delta t|\Psi|_{\hbar}^2\\
&\geq (\frac{1}{2}-\frac{\Delta t}{2\gamma_q}-s\alpha\Delta t)|\Psi|_{\hbar}^2+(\frac{1}{2}+\frac{\Delta t}{\gamma_q})|\Gamma|_{\hbar}^2+\alpha\Delta t|\Psi|_1^2+\frac{1}{2}|\Psi^{n-1}|_{\hbar}^2+\frac{1}{2}|\Gamma^{n-1}|_{\hbar}^2\\
&\geq \min\{(\frac{1}{2}-\frac{\Delta t}{2\gamma_q}-s\alpha\Delta t),(\frac{1}{2}+\frac{\Delta t}{\gamma_q})\}R^2+\alpha\Delta t|\Psi|_1^2+C(\Psi^{n-1},\Gamma^{n-1})\\
&\geq 0. \qquad (73)
\end{aligned}$$

Observing that Ψ,Γ is a fixed point of the function $\mathtt{B}$, and taking into account (69) and (73), it follows, for R sufficiently large, that

$$(\Psi,\ \Psi)^{\hbar}+(\Gamma,\ \Gamma)=\frac{-R[(A_{\psi}(\Psi,\Gamma),\Psi)^{\hbar}+(A_{\gamma}(\Psi,\Gamma),\Gamma)]}{|(A_{\psi}(\Psi,\Gamma),A_{\gamma}(\Psi,\Gamma)|_{S^{\hbar}\times S^{\hbar}}}<0. \qquad (74)$$

It comes from (70) that

$$(\Psi,\ \Psi)^{\hbar}+(\Gamma,\ \Gamma)=|\Psi|_{\hbar}^2+\|\Gamma\|_0^2\geq C[|\Psi|_{\hbar}^2+|\Gamma|_{\hbar}^2]\geq R^2\geq 0,$$

which contradicts (74). The contradiction establishes the existence of $\Psi^n,\Gamma^n\in S^{\hbar}\times S^h$ such that $A_{\psi}(\Psi^n,\Gamma^n)=A_{\gamma}(\Psi^n,\Gamma^n)=0$. In other words, it confirms the existence of a solution Ψ^n,Γ^n for the nth step of $(\Lambda^{h,\Delta t})$. □

4.2 Uniqueness of approximation

By selecting $\Upsilon=\Psi=\Psi_1^n-\Psi_2^n$ in (56) and $\Upsilon=\Gamma=\Gamma_1^n-\Gamma_2^n$ in (57), and subsequently subtracting the corresponding fully-discrete approximations for both Ψ and Γ, we arrive at the following equations:

$$\frac{1}{\Delta t\alpha}(\Psi,\Psi)^{\hbar}+(\nabla\Psi,\ \nabla\Psi)=0, \qquad (75)$$

$$\frac{\gamma_q}{\Delta t}(\Gamma,\Gamma)+(\Gamma,\Gamma)=(\Psi,\Gamma). \qquad (76)$$

Multiplying equation (75) by $\Delta t\alpha$ and combining this result with equation (76) results in

$$|\Psi|_{\hbar}^2+\Delta t\alpha|\Psi|_1^2+\frac{\gamma_q}{\Delta t}\|\Gamma\|_0^2+\|\Gamma\|_0^2=(\Psi,\Gamma). \qquad (77)$$

By using Young's inequality and (11), we have

$$\frac{1}{2}|\Psi|_{\hbar}^2 + \frac{\gamma_q}{\Delta t}\|\Gamma\|_0^2 \leq 0.$$

This leads to the conclusion that $\Psi_1^n = \Psi_2^n$ and $\Gamma_1^n - \Gamma_2^n$ for all $n \geq 1$, as needed.

4.3 Existence of weak solution

Firstly, let us introduce the following definitions:

$$\Psi(t) := (\frac{t - t_{n-1}}{\Delta t})\Psi^n + (\frac{t_n - t}{\Delta t})\Psi^{n-1}, \quad t \in [t_{n-1}, t_n] \quad n \geq 1, \tag{78}$$

$$\Gamma(t) := (\frac{t - t_{n-1}}{\Delta t})\Gamma^n + (\frac{t_n - t}{\Delta t})\Gamma^{n-1}, \quad t \in [t_{n-1}, t_n] \quad n \geq 1, \tag{79}$$

and

$$\Psi^+(t) := \Psi^n, \quad \Psi^-(t) := \Psi^{n-1}, \quad t \in (t_{n-1}, t_n], \quad n \geq 1, \tag{80}$$

$$\Gamma^+(t) := \Gamma^n, \quad \Gamma^-(t) := \Gamma^{n-1}, \quad t \in (t_{n-1}, t_n], \quad n \geq 1. \tag{81}$$

Considering (78), (79), along with (80) and (81), we derive that

$$\frac{\partial \Psi}{\partial t} = \frac{\Psi^+ - \Psi^-}{\Delta t} = \frac{\Psi^+ - \Psi}{t_n - t} = \frac{\Psi - \Psi^-}{t - t_{n-1}}, \quad t \in (t_{n-1}, t_n) \quad n \geq 1, \tag{82}$$

$$\frac{\partial \Gamma}{\partial t} = \frac{\Gamma^+ - \Gamma^-}{\Delta t} = \frac{\Gamma^+ - \Gamma}{t_n - t} = \frac{\Gamma - \Gamma^-}{t - t_{n-1}}, \quad t \in (t_{n-1}, t_n) \quad n \geq 1. \tag{83}$$

Leveraging the aforementioned information, we can reformulate the problem $(\Lambda^{h,\Delta t})$ in the following manner:

$$\int_0^{\Re} (\frac{\partial \Psi}{\partial t},\ X)^{\hbar}\, dt + \alpha \int_0^{\Re} (\nabla \Psi^+,\ \nabla X)\, dt = \alpha s \int_0^{\Re} (1,\ X)^{\hbar}\, dt, \tag{84}$$

$$\int_0^{\Re} (\frac{\partial \Gamma}{\partial t},\ X)\, dt + \frac{1}{\gamma_q}\int_0^{\Re} (\Gamma^+,\ X)\, dt = \frac{1}{\gamma_q}\int_0^{\Re} (\Psi^+,\ X)\, dt. \tag{85}$$

Theorem 5. Assuming that the conditions of Theorem 3 are met, there exists a subsequence of $\Psi^{\pm}, \Gamma^{\pm}$ that solves (84) and (85) as $\hbar \to 0$ such that

$$\Psi, \Psi^{\pm} \rightharpoonup^* \psi \quad \textit{and} \quad \Gamma, \Gamma^{\pm} \rightharpoonup^* \gamma \quad \textit{in} \quad L^{\infty}(0, \Re; L^2(\Im)), \tag{86}$$

$$\Psi, \Psi^{\pm} \rightharpoonup^{*} \psi \qquad in \quad L^{\infty}(0, \Re; H^{1}(\Im)),$$
$$\Psi, \Psi^{\pm} \rightharpoonup \psi \qquad in \quad L^{2}(0, \Re; H^{1}(\Im)), \tag{87}$$
$$\Gamma, \Gamma^{\pm} \to \gamma \qquad in \quad L^{2}(\Im_{\Re})$$
$$\frac{\partial \Psi}{\partial t} \to \frac{\partial \psi}{\partial t} \qquad and \qquad \frac{\partial \Gamma}{\partial t} \to \frac{\partial \gamma}{\partial t} \qquad in \quad L^{2}(\Im_{\Re}). \tag{88}$$

Proof. The aforementioned convergence results stem from the bounds provided in (58), taking into account that the spaces $L^{2}(0, \Re; H^{1}(\Im))$ and $L^{2}(\Im_{\Re})$ are reflexive Banach spaces. Additionally, the spaces $L^{\infty}(0, \Re; H^{1}(\Im))$ and $L^{\infty}(0, \Re; L^{2}(\Im))$ serve as the dual spaces of $L^{1}(0, \Re; (H^{1}(\Im))')$ and $L^{1}(0, \Re; L^{2}(\Im))$, respectively. While these dual spaces are separable Banach spaces, they are not reflexive. □

Theorem 6. Assuming that the conditions of Theorem 5 are satisfied, the functions Ψ, Γ constitute a global weak solution in the following sense, for all $\Upsilon \in L^{2}(0, \Re; H^{1}(\Im))$:

$$\int_{0}^{\Re} (\frac{\partial \psi}{\partial t},\ \Upsilon)^{\hbar}\, dt + \alpha \int_{0}^{\Re} (\nabla \psi,\ \nabla \Upsilon)\, dt = \alpha s \int_{0}^{\Re} (1,\ \Upsilon)^{\hbar}\, dt, \tag{89}$$

$$\int_{0}^{\Re} (\frac{\partial \gamma}{\partial t},\ \Upsilon)\, dt + \frac{1}{\gamma_q} \int_{0}^{\Re} (\gamma,\ \Upsilon)\, dt = \frac{1}{\gamma_q} \int_{0}^{\Re} (\psi,\ \Upsilon)\, dt. \tag{90}$$

Proof. If we substitute $X \equiv \Pi^{\hbar} \Upsilon$ into (84), then we obtain that

$$\int_{0}^{\Re} (\frac{\partial \Psi}{\partial t},\ \Pi^{\hbar} \Upsilon)^{\hbar}\, dt + \alpha \int_{0}^{\Re} (\nabla \Psi^{+},\ \nabla \Pi^{\hbar} \Upsilon)\, dt = \alpha s \int_{0}^{\Re} (1,\ \Pi^{\hbar} \Upsilon)^{\hbar}\, dt.$$

For any $\Upsilon \in L^{2}(0, \Re; H^{1}(\Im))$ and $\widetilde{\Upsilon} \in H^{1}(0, \Re; H^{1}(\Im))$, we have

$$\begin{aligned}
\int_{0}^{\Re} (\frac{\partial \Psi}{\partial t},\ \Pi^{\hbar} \Upsilon)^{\hbar} dt = & \int_{0}^{\Re} \left[\left(\frac{\partial \Psi}{\partial t},\ \Pi^{\hbar} (\Upsilon - \widetilde{\Upsilon}) \right)^{\hbar} - \left(\frac{\partial \Psi}{\partial t},\ \Pi^{\hbar} (\Upsilon - \widetilde{\Upsilon}) \right) \right] dt \\
& + \int_{0}^{\Re} \left[\left(\frac{\partial \Psi}{\partial t},\ \Pi^{\hbar} \widetilde{\Upsilon} \right)^{\hbar} - \left(\frac{\partial \Psi}{\partial t},\ \Pi^{\hbar} \widetilde{\Upsilon} \right) \right] dt \\
& + \int_{0}^{\Re} \left(\frac{\partial \Psi}{\partial t},\ (\Pi^{\hbar} - I) \Upsilon \right) dt \\
& + \int_{0}^{\Re} \left(\frac{\partial \Psi}{\partial t},\ \Upsilon \right) dt
\end{aligned}$$

$$:= K_{1,1} + K_{1,2} + K_{1,3} + K_{1,4}. \tag{91}$$

Utilizing (15), (13), (14), Hölder's inequality, and (58), provides that

$$\begin{aligned}
|K_{1,1}| &\equiv \left| \int_0^{\Re} \left[\left(\frac{\partial \Psi}{\partial t},\ \Pi^{\hbar}(\Upsilon - \widetilde{\Upsilon}) \right)^{\hbar} - \left(\frac{\partial \Psi}{\partial t},\ \Pi^{\hbar}(\Upsilon - \widetilde{\Upsilon}) \right) \right] dt \right| \\
&\leq \int_0^{\Re} \left| \left[\left(\frac{\partial \Psi}{\partial t},\ \Pi^{\hbar}(\Upsilon - \widetilde{\Upsilon}) \right)^{\hbar} - \left(\frac{\partial \Psi}{\partial t},\ \Pi^{\hbar}(\Upsilon - \widetilde{\Upsilon}) \right) \right] \right| dt \\
&\leq Ch \int_0^{\Re} \| \frac{\partial \Psi}{\partial t} \|_0 \quad |\Pi^{\hbar}(\Upsilon - \widetilde{\Upsilon})|_1 \, dt \\
&\leq C\hbar \| \frac{\partial \Psi}{\partial t} \|_{L^2(\Im_{\Re})} \| \Upsilon - \widetilde{\Upsilon} \|_{L^2(0,\Re;H^1(\Im))} \\
&\leq C\hbar \| \Upsilon - \widetilde{\Upsilon} \|_{L^2(0,\Re;H^1(\Im))} \quad \to 0 \quad as \quad \hbar \to 0.
\end{aligned} \tag{92}$$

It can be deduced from (15), (13), (14), Hölder's inequality, and (86) that

$$\begin{aligned}
|K_{1,2}| &\equiv \left| \int_0^{\Re} \left[\left(\frac{\partial \Psi}{\partial t},\ \Pi^{\hbar}\widetilde{\Upsilon} \right)^{\hbar} - \left(\frac{\partial \Psi}{\partial t},\ \Pi^{\hbar}\widetilde{\Upsilon} \right) \right] dt \right| \\
&\leq \int_0^{\Re} \left| \left(\Psi,\ \frac{\partial(\Pi^{\hbar}\widetilde{\Upsilon})}{\partial t} \right)^{\hbar} - \left(\Psi,\ \frac{\partial(\Pi^{\hbar}\widetilde{\Upsilon})}{\partial t} \right) \right| dt \\
&\quad + \left| \left(\Psi(\cdot,\Re),\ \Pi^{\hbar}\widetilde{\Upsilon}(\cdot,\Re) \right)^{\hbar} - \left(\Psi(\cdot,\ \Re), \Pi^{\hbar}\widetilde{\Upsilon}(\cdot,\ \Re) \right) \right| \\
&\quad + \left| \left(\Psi(\cdot,0),\ \Pi^{\hbar}\widetilde{\Upsilon}(\cdot,0) \right)^{\hbar} - \left(\Psi(\cdot,\ 0), \Pi^{\hbar}\widetilde{\Upsilon}(\cdot,\ 0) \right) \right| \\
&\leq C\hbar \int_0^{\Re} \|\Psi\|_0 | \frac{\partial(\Pi^{\hbar}\widetilde{\Upsilon})}{\partial t} |_1 \, dt + Ch \|\Psi(\cdot,\Re)\|_0 |\Pi^{\hbar}\widetilde{\Upsilon}(\cdot,\Re)|_1 \\
&\quad + Ch \|\Psi(\cdot,0)\|_0 |\Pi^{\hbar}\widetilde{\Upsilon}(\cdot,0)|_1 \\
&\leq Ch \|\Psi\|_{L^{\infty}(0,\Re;L^2(\Im))} \|\widetilde{\Upsilon}\|_{H^1(0,\Re;H^1(\Im))} \\
&\leq Ch \|\widetilde{\Upsilon}\|_{H^1(0,\Re;H^1(\Im))} \quad \to \quad 0 \quad as \quad \hbar \to 0.
\end{aligned} \tag{93}$$

To analyze the term $K_{1,3}$, we proceed by utilizing (16), applying Hölder's inequality, and considering (88), resulting in

$$\begin{aligned}
|K_{1,3}| \equiv | \int_0^{\Re} (\frac{\partial \Psi}{\partial t},\ (\Pi^{\hbar} - I)\Upsilon) \, dt| &= | \int_0^{\Re} \langle \frac{\partial \Psi}{\partial t},\ (\Pi^{\hbar} - I)\Upsilon \rangle \, dt| \\
&\leq \| \frac{\partial \Psi}{\partial t} \|_{L^2(\Im_{\Re})} \| (\Pi^{\hbar} - I)\Upsilon \|_{L^2(0,\Re;H^1(\Im))}
\end{aligned}$$

$$\leq C\|(\Pi^{\hbar}-I)\Upsilon\|_{L^2(0,\Re;H^1(\Im))}. \tag{94}$$

From (16) and the outcome of weak convergence as stated in (88), it can be concluded for any $\Upsilon \in L^2(0,\Re;H^1(\Im))$ that

$$K_{1,4} \equiv \int_0^{\Re} (\frac{\partial \Psi}{\partial t},\ \Upsilon)\, dt = \int_0^{\Re} \langle \frac{\partial \Psi}{\partial t},\ \Upsilon\rangle dt \to \int_0^{\Re} \langle \frac{\partial \psi}{\partial t},\ \Upsilon\rangle\, dt \quad as \quad \hbar \to 0. \tag{95}$$

Combining (91)–(95) and the denseness of $H^1(0,\Re;H^1(\Im))$ in $L^2(0,\Re;H^1(\Im))$ yields for all $\Upsilon \in L^2(0,\Re;H^1(\Im))$, that

$$\int_0^{\Re} (\frac{\partial \Psi}{\partial t},\ \Pi^{\hbar}\Upsilon)^{\hbar} dt \to \int_0^{\Re} \langle \frac{\partial \psi}{\partial t},\ \Upsilon\rangle\, dt \quad as \quad \hbar \to 0.$$

By applying Hölder's inequality, considering (14), and taking into account (87), we obtain the following outcome for any $\Upsilon \in L^2(0,\Re;H^1(\Im))$

$$\begin{aligned}
\int_0^{\Re} (\nabla\Psi^{+},\ \nabla\Pi^{\hbar}\Upsilon)\, dt &= \int_0^{\Re} (\nabla\Psi^{+},\ \nabla(\Pi^{\hbar}-I)\Upsilon) dt + \int_0^{\Re} (\nabla\Psi^{+},\ \nabla\Upsilon) dt \\
&\leq \left| \int_0^{\Re} (\nabla\Psi^{+},\ \nabla(\Pi^{\hbar}-I)\Upsilon) dt + \int_0^{\Re} (\nabla\Psi^{+},\ \nabla\Upsilon) dt \right| \\
&\leq \int_0^{\Re} |(\nabla\Psi^{+},\ \nabla(\Pi^{\hbar}-I)\Upsilon)| dt + \int_0^{\Re} |(\nabla\Psi^{+},\ \nabla\Upsilon)| dt \\
&\leq \int_0^{\Re} |\Psi^{+}|_1 |(\Pi^{\hbar}-I)\Upsilon|_1\, dt + \int_0^{\Re} |(\nabla\Psi^{+},\ \nabla\Upsilon)| dt \\
&\leq \|\Psi^{+}\|_{L^2(0,\Re;H^1(\Im))} \|(\Pi^{\hbar}-I)\Upsilon\|_{L^2(0,\Re;H^1(\Im))} \\
&\quad + \int_0^{\Re} |(\nabla\Psi^{+},\ \nabla\Upsilon)| dt \\
&\to \int_0^{\Re} (\nabla\psi,\ \nabla\Upsilon)\, dt \quad as \quad \hbar \to 0.
\end{aligned}$$

Now, if we set $X \equiv \Pi^{\hbar}\Upsilon$ in (85), then we can determine that

$$\int_0^{\Re} (\frac{\partial \Gamma}{\partial t},\ \Pi^{\hbar}\Upsilon)\, dt + \frac{1}{\gamma_q} \int_0^{\Re} (\Gamma^{+},\ \Pi^{\hbar}\Upsilon)\, dt = \frac{1}{\gamma_q} \int_0^{\Re} (\Psi^{+},\ \Pi^{\hbar}\Upsilon)\, dt.$$

For any $\Upsilon \in L^2(0,\Re;H^1(\Im))$, it holds that

$$\begin{aligned}
\int_0^{\Re} (\frac{\partial \Gamma}{\partial t},\ \Pi^{\hbar}\Upsilon)^{\hbar} dt &= \int_0^{\Re} \left(\frac{\partial \Gamma}{\partial t},\ (\Pi^{\hbar}-I)\Upsilon \right) dt + \int_0^{\Re} \left(\frac{\partial \Gamma}{\partial t},\ \Upsilon \right) dt \\
&:= K_{2,1} + K_{2,2}.
\end{aligned}$$

By employing (16), Hölder's inequality, and taking into account (88), it can be deduced that

$$\begin{aligned}|K_{2,1}| \equiv |\int_0^{\Re}(\frac{\partial\Gamma}{\partial t},\ (\Pi^{\hbar}-I)\Upsilon)\,dt| &= |\int_0^{\Re}\langle\frac{\partial\Gamma}{\partial t},\ (\Pi^{\hbar}-I)\Upsilon\rangle\,dt| \\ &\leq \|\frac{\partial\Gamma}{\partial t}\|_{L^2(\Im_{\Re})}\|(\Pi^{\hbar}-I)\Upsilon\|_{L^2(0,\Re;H^1(\Im))} \\ &\leq C\|(\Pi^{\hbar}-I)\Upsilon\|_{L^2(0,\Re;H^1(\Im))}.\end{aligned}$$

Derived from (16) and the outcome of weak convergence as stated in (88), it can be concluded for any $\Upsilon \in L^2(0,\Re;H^1(\Im))$ that

$$K_{2,2} \equiv \int_0^{\Re}(\frac{\partial\Gamma}{\partial t},\ \Pi^{\hbar}\Upsilon)dt \to \int_0^{\Re}\langle\frac{\partial\gamma}{\partial t},\ \Upsilon\rangle\,dt \quad as \quad \hbar\to 0.$$

Now, we have for all $\Upsilon \in L^2(0,\Re;H^1(\Im))$, that

$$\int_0^{\Re}(\Gamma^{+},\ \Pi^{\hbar}\Upsilon)^{\hbar}dt = \int_0^{\Re}(\Gamma^{+},\ (\Pi^{\hbar}-I)\Upsilon)dt + \int_0^{\Re}(\Gamma^{+},\ \Upsilon)dt := K_{3,1}+K_{3,2}. \tag{96}$$

Through the application of the Hölder inequality, considering (13), (58), and taking into consideration (11), we can establish that

$$\begin{aligned}|K_{3,1}| = \left|\int_0^{\Re}(\Gamma^{+},\ (\Pi^{\hbar}-I)\Upsilon)dt\right| &\leq C\hbar\|\Gamma^{+}\|_{L^2(\Im_{\Re})}\|\Upsilon\|_{L^2(0,\Re;H^1(\Im))} \\ \leq C\hbar\|\Upsilon\|_{L^2(0,\Re;H^1(\Im))} \quad &\to \quad 0 \quad as \quad \hbar\to 0.\end{aligned} \tag{97}$$

Consolidating (96)–(97) results in the following expression for all $\Upsilon \in L^2(0,\Re;H^1(\Im))$:

$$\int_0^{\Re}(\Gamma^{+},\ \Upsilon)dt \to \int_0^{\Re}(\gamma,\ \Upsilon)dt \quad as \quad \hbar\to 0.$$

Similarly to (96), we can show that

$$\int_0^{\Re}(\Psi^{+},\ \Upsilon)dt \to \int_0^{\Re}(\psi,\ \Upsilon)dt \quad as \quad \hbar\to 0.$$

Thus, the proof of (89) and (90) has been completed. □

4.4 Error estimate of the approximation

Initially, we can rephrase (56) and (57) in the following form:
Find $\Psi^n(\cdot,\Re),\Gamma^n(\cdot,\Re)\in H^1(0,\Re;S^\hbar)$ such that $\Psi^0:=\mathbb{P}^\hbar\psi^0$ and $\Gamma^0:=\mathbb{P}^\hbar\gamma_1^0$,

$$\left(\frac{\partial\Psi}{\partial t},\ \Upsilon\right)^\hbar+\alpha(\nabla\Psi^+,\ \nabla\Upsilon)=\alpha s(1,\Upsilon)^\hbar \qquad \text{for all } \Upsilon\in S^\hbar, \tag{98}$$

$$\left(\frac{\partial\Gamma}{\partial t},\ \Upsilon\right)+\frac{1}{\gamma_q}(\Gamma^+,\ \Upsilon)=\frac{1}{\gamma_q}(\Psi^+,\ \Upsilon) \qquad \text{for all } \Upsilon\in S^\hbar. \tag{99}$$

Theorem 7. Assuming the validity of the results from Theorem 2, it follows that

$$\|\psi-\Psi^+\|_{L^\infty(0,\Re;L^2(\Im))}+\|\gamma-\Gamma^+\|_{L^\infty(0,\Re;L^2(\Im))}+\|\psi-\Psi^+\|_{L^2(0,\Re;H^1(\Im))}+\|\gamma-\Gamma^+\|_{L^2(\Im_\Re)}\le C\Delta t.$$

Proof. Let

$$\begin{aligned} E_1&=\psi^\hbar-\Psi, \qquad E_1^+=\psi^\hbar-\Psi^+,\\ E_2&=\gamma^\hbar-\Gamma, \qquad \Gamma_1^+=\gamma^\hbar-\Gamma^+. \end{aligned} \tag{100}$$

Then, based on (100), it follows that

$$\begin{aligned} E_1^+-E_1&=\Psi-\Psi^+=(t-t_n)\frac{\partial\Psi}{\partial t},\\ E_1^--E_1&=\Psi-\Psi^-=(t-t_{n-1})\frac{\partial\Psi}{\partial t}, \end{aligned} \tag{101}$$

$$\begin{aligned} E_2^+-E_2&=\Gamma-\Gamma^+=(t-t_n)\frac{\partial\Gamma}{\partial t},\\ E_2^--E_2&=\Gamma-\Gamma^-=(t-t_{n-1})\frac{\partial\Gamma}{\partial t}. \end{aligned} \tag{102}$$

By employing (100), (101), and (102), we obtain the ensuing set of inequalities:

$$|E_1^+|_\hbar\le|E_1|_\hbar+|\Psi^+-\Psi^-|_\hbar, \tag{103}$$

$$|E_1|_\hbar\le|E_1^+|_\hbar+|\Psi^+-\Psi^-|_\hbar,$$

$$|E_2^+|_\hbar\le|E_2|_\hbar+|\Gamma^+-\Gamma^-|_\hbar, \tag{104}$$

$$|E_2|_\hbar \leq |E_2^+|_\hbar + |\Gamma^+ - \Gamma^-|_\hbar.$$

Upon substituting $\Upsilon^\hbar = X$ into both (17) and (18) and subsequently subtracting (98) from (17) and (99) from (18), we obtain the following result:

$$\left(\frac{\partial E_1}{\partial t},\ X\right)^\hbar + \alpha(\nabla {E_1}^+,\ \nabla X) = 0, \tag{105}$$

$$\left(\frac{\partial E_2}{\partial t},\ X\right) + \frac{1}{\gamma_q}(E_2^+,\ X) = \frac{1}{\gamma_q}(E_1^+,\ X). \tag{106}$$

By choosing $X = E_1^+$ in (105) and $X = E_2^+$ in (106), we reach the following expressions:

$$\left(\frac{\partial E_1}{\partial t},\ E_1^+\right)^\hbar + \alpha(\nabla {E_1}^+,\ \nabla E_1^+) = 0,$$

$$\left(\frac{\partial E_2}{\partial t},\ E_2^+\right) + \frac{1}{\gamma_q}(E_2^+,\ E_2^+) = \frac{1}{\gamma_q}(E_1^+,\ E_2^+).$$

Upon utilizing (101), (102), applying Young's inequality, considering (11), (103), and taking into account (104), we can conclude that

$$\begin{aligned}
&\frac{d}{2dt}[|E_1|_\hbar^2 + \|E_2\|_0^2] + \alpha|E_1|_1^2 + \frac{1}{\gamma_q}\|E_2\|_0^2 \\
&\leq \frac{1}{2\gamma_q}\|E_1^+\|_0^2 + \frac{1}{2\gamma_q}\|E_2^+\|_0^2 + (\frac{\partial E_1}{\partial t}, \Psi^+ - \Psi)^\hbar + (\frac{\partial E_2}{\partial t}, \Gamma^+ - \Gamma) \\
&\leq C|E_1^+|_\hbar^2 + \frac{1}{2\gamma_q}\|E_2^+\|_0^2 + (\frac{\partial E_1}{\partial t}, \Psi^+ - \Psi)^\hbar + (\frac{\partial E_2}{\partial t}, \Gamma^+ - \Gamma) \\
&\leq C|E_1|_\hbar^2 + C|\Psi^+ - \Psi^-|_\hbar^2 + \frac{1}{2\gamma_q}\|E_2\|_0^2 + C|\Gamma^+ - \Gamma^-|_\hbar^2 \\
&\quad + (\frac{\partial E_1}{\partial t}, \Psi^+ - \Psi)^\hbar + (\frac{\partial E_2}{\partial t}, \Gamma^+ - \Gamma).
\end{aligned} \tag{107}$$

Subsequently, through the utilization of (12), (100), and (11), we derive the following expression:

$$(\frac{\partial E_1}{\partial t}, \Psi^+ - \Psi)^\hbar \leq |\frac{\partial \psi^\hbar}{\partial t}|_\hbar |\Psi^+ - \Psi^-|_\hbar + \frac{1}{\Delta t}|\Psi^+ - \Psi^-|_\hbar^2, \tag{108}$$

$$\begin{aligned}
&(\frac{\partial E_2}{\partial t}, \Gamma^+ - \Gamma) \leq \|\frac{\partial \gamma^\hbar}{\partial t}\|_0 \|\Gamma^+ - \Gamma^-\|_0 + \frac{1}{\Delta t}\|\Gamma^+ - \Gamma^-\|_0^2 \\
&\leq |\frac{\partial \gamma^\hbar}{\partial t}|_\hbar |\Gamma^+ - \Gamma^-|_\hbar + \frac{1}{\Delta t}|\Gamma^+ - \Gamma^-|_\hbar^2.
\end{aligned} \tag{109}$$

By substituting (108) and (109) into (107), we arrive at the ensuing inequality:

$$\begin{aligned}
&\frac{d}{2dt}[|E_1|_\hbar^2+\|E_2\|_0^2]+\alpha|E_1|_1^2+\frac{1}{\gamma_q}\|E_2\|_0^2\\
&\le C|E_1|_\hbar^2+\frac{1}{2\gamma_q}\|E_2\|_0^2+C|\Psi^+-\Psi^-|_\hbar^2+C|\Gamma^+-\Gamma^-|_\hbar^2+|\frac{\partial\psi^\hbar}{\partial t}|_\hbar|\Psi^+-\Psi^-|_\hbar\\
&\quad+\frac{1}{\Delta t}|\Psi^+-\Psi^-|_\hbar^2+|\frac{\partial\gamma^\hbar}{\partial t}|_\hbar|\Gamma^+-\Gamma^-|_\hbar+\frac{1}{\Delta t}|\Gamma^+-\Gamma^-|_\hbar^2.
\end{aligned}$$

Multiplying the above outcomes by 2 and employing the Grönwall lemma, while taking into account that $E_1(0)=E_2(0)=0$, yield the following result:

$$\begin{aligned}
&|E_1|_\hbar^2+\|E_2\|_0^2+2\alpha\int_0^\Re|E_1|_1^2dt+\frac{2}{\gamma_q}\int_0^\Re\|E_2\|_0^2dt\\
&\le e^{C\Re}\int_0^\Re[|\Psi^+-\Psi^-|_\hbar^2+|\Gamma^+-\Gamma^-|_\hbar^2+\frac{1}{\Delta t}|\Psi^+-\Psi^-|_\hbar^2+\frac{1}{\Delta t}|\Gamma^+-\Gamma^-|_\hbar^2\\
&\quad+|\frac{\partial\psi^\hbar}{\partial t}|_\hbar|\Psi^+-\Psi^-|_\hbar+|\frac{\partial\gamma^\hbar}{\partial t}|_\hbar|\Gamma^+-\Gamma^-|_\hbar]dt. \qquad (110)
\end{aligned}$$

The terms on the right-hand side of (110) can be bounded by applying Theorem 3, leading to the determination that

$$\begin{aligned}
&\int_0^\Re[|\Psi^+-\Psi^-|_\hbar^2+|\Gamma^+-\Gamma^-|_\hbar^2]dt\\
&=\sum_{n=1}^N\int_{t_{n-1}}^{t_n}[|\Psi^n-\Psi^{n-1}|_\hbar^2+|\Gamma^n-\Gamma^{n-1}|_\hbar^2]dt\le C\Delta t,
\end{aligned}$$

$$\begin{aligned}
&\int_0^\Re\frac{1}{\Delta t}[|\Psi^+-\Psi^-|_\hbar^2+|\Gamma^+-\Gamma^-|_\hbar^2]dt\\
&=\sum_{n=1}^N\int_{t_{n-1}}^{t_n}[|\Psi^n-\Psi^{n-1}|_\hbar^2+|\Gamma^n-\Gamma^{n-1}|_\hbar^2]dt\le C.
\end{aligned}$$

To bound the fifth and sixth terms on the right-hand side of (110), we employ the Cauchy–Schwarz inequality and utilize the results of Theorems 3 and 2, to find that

$$\int_0^\Re[|\frac{\partial\psi^\hbar}{\partial t}|_\hbar|\Psi^+-\Psi^-|_\hbar\le C(\sum_{n=1}^N\int_{t_{n-1}}^{t_n}|\Psi^+-\Psi^-|_\hbar^2)^{\frac{1}{2}}\le C\Delta t,$$

$$\int_0^{\Re}[|\frac{\partial\gamma^{\hbar}}{\partial t}|_{\hbar}|\Gamma^{+}-\Gamma^{-}|_{\hbar}\leq C(\sum_{n=1}^{N}\int_{t_{n-1}}^{t_n}|\Gamma^{+}-\Gamma^{-}|_{\hbar}^{2})^{\frac{1}{2}}\leq C\Delta t.$$

This pertains to the subsequent outcome

$$|E_1|_{\hbar}^2+\|E_2\|_0^2+2\alpha\int_0^{\Re}|E_1|_1^2dt+\frac{2}{\gamma_q}\int_0^{\Re}\|E_2\|_0^2dt\leq C\Delta t. \tag{111}$$

Therefore, based on (111), it follows that

$$\|E_1\|_{L^{\infty}(0,\Re;L^2(\Im))}+\|E_2\|_{L^{\infty}(0,\Re;L^2(\Im))}+\|E_1\|_{L^2(0,\Re;H^1(\Im))}+\|E_2\|_{L^2(\Im_{\Re})}\leq C\Delta t.$$

5 Numerical results

This section delves into the numerical analysis of the microscale heat equation. An iterative method has been developed to solve the nonlinear equation system presented by the problem $(\Lambda^{h,\Delta t})$. Additionally, it presents and examines the numerical error outcomes for the Dirichlet non-homogeneous boundary conditions applied to the microscale heat equation in one, two, and three dimensions.

5.1 Numerical algorithm

We begin by presenting a practical algorithm designed to solve the nonlinear algebraic system generated by the approximate problem $(\Lambda^{h,\Delta t})$ at each time level: $(\Lambda_k^{h,\Delta t})$: Given $\{\Psi^{n,0},\Gamma^{n,0}\}\in[S^{\hbar}]^2$, then for $k\geq 1$ find $\{\Psi^{n,k},\Gamma^{n,k}\}\in[S^{\hbar}]^2$ such that for all $\Upsilon\in S^{\hbar}$

$$\left(\frac{\Psi^{n,k}-\Psi^{n-1}}{\Delta t},\ \Upsilon\right)^{\hbar}+\alpha(\nabla\Psi^{n,k},\ \nabla\Upsilon)=\alpha s(1,\Upsilon)^{\hbar}, \tag{112}$$

$$\left(\frac{\Gamma^{n,k}-\Gamma^{n-1}}{\Delta t},\ \Upsilon\right)+\frac{1}{\gamma_q}(\Gamma^{n,k},\ \Upsilon)=\frac{1}{\gamma_q}(\Psi^{n,k},\ \Upsilon). \tag{113}$$

Initiating with $\Psi^0\equiv\Pi^{\hbar}\psi^0$ and $\Gamma^0\equiv\Pi^{\hbar}\gamma^0$, and for $n\geq 1$, we initialize $\Psi^{n,0}\equiv\Psi^{n-1}$ and $\Gamma^{n,0}\equiv\Gamma^{n-1}$. Equations (112) and (113) can be reformulated as a system of $2\times(J+1)^d$ linear equations by testing (112) and

(113) with basis functions $\varphi_j, j = 0, \ldots, J$. For the numerical experiments, we choose a tolerance $TOL = 10^{-7}$ and define the stopping criteria based on this tolerance. Now,

$$\max\left\{|\Psi^{n,k} - \Psi^{n,k-1}|_{0,\infty}, |\Gamma^{n,k} - \Gamma^{n,k-1}|_{0,\infty}\right\} < TOL, \tag{114}$$

that is, for k satisfying (114), we set $\Psi^n \equiv \Psi^{n,k}, \Gamma^n \equiv \Gamma^{n,k}$.

The programs were developed in MATLAB, and the resulting linear systems were addressed using the Gauss–Seidel iteration method. Although a formal proof of the convergence of $\Psi^{n,k}, \Gamma^{n,k}$ towards Ψ^n, Γ^n for a fixed n has not been established, practical observations have shown promising convergence characteristics. It was observed that the iterative method consistently achieved good convergence (requiring only a few iterations to meet the stopping criteria at each time step).

5.2 Error computations

To evaluate the error, we introduce a slight alteration to Problem (Λ) by incorporating source terms $f(\mathbf{x}, t)$ and $h(\mathbf{x}, t)$. This modification transforms the system denoted by (4)–(5) into the following configuration: (Λ) Find $\{\psi, \gamma\}$ such that

$$\partial_t \psi = \alpha \Delta \psi + \alpha s + f(\mathbf{x}, t),$$

$$\partial_t \gamma = \frac{1}{\gamma_q}\psi - \frac{1}{\gamma_q}\gamma + h(\mathbf{x}, t).$$

Hence, we propose the following fully-discrete finite element approximation for $(\Lambda_k^{h,\Delta t})$:
$(\Lambda_k^{h\,\Delta t})$: Given $\{\Psi^{n,0}, \Gamma^{n,0}\} \in [S^\hbar]^2$, then for $k \geq 1$ find $\{\Psi^{n,k}, \Gamma^{n,k}\} \in [S^\hbar]^2$, such that for all $\Upsilon \in S^\hbar$

$$\left(\frac{\Psi^{n,k} - \Psi^{n-1}}{\Delta t},\ \Upsilon\right)^\hbar + \alpha(\nabla \Psi^{n,k},\ \nabla \Upsilon) - \alpha s(1, \Upsilon)^\hbar = (f(\mathbf{x}, t_n),\ \Upsilon), \tag{115}$$

$$\left(\frac{\Gamma^{n,k} - \Gamma^{n-1}}{\Delta t},\ \Upsilon\right) + \frac{1}{\gamma_q}(\Gamma^{n,k},\ \Upsilon) - \frac{1}{\gamma_q}(\Psi^{n,k},\ \Upsilon) = (h(\mathbf{x}, t_n),\ \Upsilon). \tag{116}$$

5.2.1 One-dimensional error

In this section, we present two numerical examples that solve the system denoted by (115) and (116) under homogeneous Dirichlet boundary conditions and given initial conditions. For all examples, for the sake of simplicity, we set $\alpha = 1, \gamma_q = 1, s = 0$, with the spatial domain $\Im = [0, 1]$ and the time duration $\Re = 1$. The initial and boundary conditions, along with the source terms $f(x, t)$ and $h(x, t)$, are determined based on the specific analytical solution for each example. Initially, we divided the domain $\Im = [0, 1]$ uniformly into J intervals to create a square mesh. Let $h = 1/J$ represent the mesh size for the element, and let $\Delta t = 10^{-6}$ be the time step. The analytical solutions are defined as follows:

(i) $\psi = \exp(t + x^3)$,

(ii) $\psi = 1 + \exp(0.5 * t + x^2)$.

For this example, the errors in the L^1, L^2, and L^∞-norms are detailed in Tables 1–2.

Table 1: Discrete L^1, L^2, L^∞-norms error for Example (i)

	$\|\psi - \Psi\|$			$\|\gamma - \Gamma\|$		
J	L^1	L^2	L^∞	L^1	L^2	L^∞
10	2.3E-02	8.1E-03	3.4E-02	5.4E-02	1.9E-02	8.0E-02
20	5.5E-03	1.4E-03	8.6E-03	1.3E-02	3.3E-03	2.0E-02
25	3.5E-03	7.9E-04	5.5E-03	8.4E-03	1.9E-03	1.3E-02
40	1.3E-03	2.4E-04	2.2E-03	3.2E-03	5.7E-04	5.1E-03
50	8.6E-04	1.4E-04	1.4E-03	2.0E-03	3.2E-04	3.3E-03
100	2.1E-04	2.4E-05	3.4E-04	5.1E-04	5.7E-05	8.2E-04

5.2.2 Two-dimensional error

For a two-dimensional example involving the system described by (115) and (116) with Dirichlet boundary conditions and an initial condition, the spatial

Table 2: Discrete L^1, L^2, L^∞-norms error for Example (ii)

	$\|\psi - \Psi\|$			$\|\gamma - \Gamma\|$		
J	L^1	L^2	L^∞	L^1	L^2	L^∞
10	3.9E-03	1.4E-03	5.4E-03	8.0E-03	2.8E-03	1.1E-02
20	9.4E-04	2.3E-04	1.4E-03	1.9E-03	4.7E-04	2.8E-03
25	6.0E-04	1.3E-04	8.8E-04	1.2E-03	2.7E-04	1.8E-03
40	2.3E-04	4.0E-05	3.4E-04	4.7E-04	8.1E-05	7.0E-04
50	1.5E-04	2.3E-05	2.2E-04	2.9E-04	4.6E-05	4.4E-04
100	1.0E-05	1.3E-06	2.0E-05	2.1E-05	2.5E-06	4.0E-05

domain is simplified to $\Im = [0,1] \times [0,1]$. The time interval is set to $[0, \Re] = [0,1]$. For ease of analysis, we select $\alpha = 1$, $\gamma_q = 1$, and $s = 0$. The initial and boundary conditions, along with the source terms $f(x,y,t)$ and $h(x,y,t)$, should be chosen in line with the specific analytical solution pertinent to each case. The time step selected for the computations is $\Delta t = 10^{-6}$. The analytical solution be

(i) $\psi = \exp(t + x + y)$,

(ii) $\psi = \exp(t + x^2 + y^2)$.

The errors in the L^1, L^2, and L^∞-norms for the corresponding simulations are provided in Tables 3–4.

5.2.3 Three-dimensional error

In this section, we introduce a three-dimensional example, considering the system denoted by (115) and (116) with Dirichlet boundary and initial conditions. For the purposes of this numerical example, we simplify the spatial domain to $\Im = [0,1] \times [0,1] \times [0,1]$, with the time interval set to $[0, \Re] = [0,1]$. We set the parameters $\alpha = 1$, $\gamma_q = 1$, and $s = 0$. The initial and boundary conditions, along with the source terms $f(x,y,z,t)$ and $h(x,y,z,t)$, are to be established based on the specific analytical solution for each example. The time step is chosen as $\Delta t = 10^{-5}$. The analytical solution will be

Table 3: Discrete L^1, L^2, L^∞-norms error for Example (i)

	$\|\psi - \Psi\|$			$\|\gamma - \Gamma\|$		
J	L^1	L^2	L^∞	L^1	L^2	L^∞
10	4.4E-04	5.4E-05	7.8E-04	1.0E-03	1.3E-04	1.8E-03
20	1.0E-04	6.1E-06	2.0E-04	2.4E-04	1.4E-05	4.6E-04
25	6.4E-05	3.0E-06	1.3E-04	1.5E-04	7.0E-06	2.9E-04
40	2.5E-05	7.4E-07	5.0E-05	5.7E-05	1.7E-06	1.2E-04
50	1.7E-05	3.9E-07	3.3E-05	3.6E-05	8.5E-07	7.4E-05
100	7.4E-06	9.1E-08	1.7E-05	2.1E-05	2.5E-07	4.4E-05

Table 4: Discrete L^1, L^2, L^∞-norms error for Example (ii)

	$\|\psi - \Psi\|$			$\|\gamma - \Gamma\|$		
J	L^1	L^2	L^∞	L^1	L^2	L^∞
10	9.6E-03	1.2E-03	1.8E-02	2.3E-02	2.8E-03	4.2E-02
20	2.2E-03	1.3E-04	4.6E-03	5.3E-03	3.2E-04	1.1E-02
25	1.4E-03	6.7E-05	2.9E-03	3.3E-03	1.6E-04	6.9E-03
40	5.4E-04	1.6E-05	1.1E-03	1.3E-03	3.8E-05	2.7E-03
50	3.4E-04	8.1E-06	7.4E-04	8.0E-04	1.9E-05	1.7E-03
100	7.6E-05	9.0E-07	1.7E-04	1.8E-04	2.1E-06	3.9E-04

$$\psi = \exp(t + x^2 + y^2 + z^2).$$

The errors in the L^1, L^2, and L^∞-norms for the corresponding simulations are detailed in Table 5.

Acknowledgments

We are grateful to the anonymous referee for their insightful comments, which have helped enhance the manuscript.

Table 5: Discrete L^1, L^2, L^∞-norms error

	$\|\psi - \Psi\|$			$\|\gamma - \Gamma\|$		
J	L^1	L^2	L^∞	L^1	L^2	L^∞
10	1.3E-02	5.4E-04	3.2E-02	3.0E-02	1.3E-03	7.5E-02
20	2.9E-03	4.1E-05	8.2E-03	6.7E-03	9.7E-05	1.9E-02
25	1.8E-03	1.8E-05	5.2E-03	4.1E-03	4.3E-05	1.2E-02
40	6.2E-04	3.2E-06	2.0E-03	1.4E-03	7.5E-06	4.6E-03
50	3.7E-04	1.4E-06	1.2E-03	8.5E-04	3.2E-06	2.9E-03

References

[1] Abbaszadeh, M. and Dehghan, M. *Investigation of heat transport equation at the microscale via interpolating element-free Galerkin method,* Eng. Comput. 38(Suppl 4) (2022), 3317–3333.

[2] Al-Juaifri, G.A. and Harfash, A.J. F*inite element analysis of nonlinear reaction–diffusion system of Fitzhugh–Nagumo type with Robin boundary conditions,* Math. Comput. Simul. 203 (2023), 486–517.

[3] Al-Musawi, G.A. and Harfash, A.J. *Finite element analysis of extended Fisher-Kolmogorov equation with Neumann boundary conditions,* Appl. Numer. Math. 201 (2024), 41–71.

[4] Baharlouei, S., Mokhtari, R. and Chegini, N. *Solving two-dimensional coupled burgers equations via a stable hybridized discontinuous Galerkin method,* Iran. J. Num. Anal. Optim. 13(3) (2023), 397–425.

[5] Biazar, J. and Salehi, F. *Chebyshev Galerkin method for integro-differential equations of the second kind,* Iran. J. Num. Anal. Optim. 6(1) (2016), 31–43.

[6] Castro, MA. , Rodríguez F., Cabrera, J., and Martín, J.A. *A compact difference scheme for numerical solutions of second order dual-phase-lagging models of microscale heat transfer,* J. Comput. Appl. Math. 291 (2016), 432–440.

[7] Ciarlet, P.G. *The finite element method for elliptic problems*, SIAM, (2002).

[8] Ciavaldini, J.F. *Analyse numerique d'un problème de Stefan à deux phases par une methode d'éléments finis,* SIAM J. Num. Anal. 12(3) (1975), 464–487.

[9] Ern, A. and Guermond, J.L. *Theory and practice of finite elements*, Springer Science & Business Media, 159 (2004)

[10] Haghighi, D., Abbasbandy, S. and Shivanian, E. *Applying the meshless fragile points method to solve the two-dimensional linear schrödinger equation on arbitrary domains,* Iran. J. Num. Anal. Optim. 13(1) (2023), 1–18.

[11] Harfash, A.J. *High accuracy finite difference scheme for three-dimensional microscale heat equation,* J. Comput. Appl. Math. 220(1-2) (2008),335–346.

[12] Hashim, M.H. and Harfash, A.J. *Finite element analysis of a Keller–Segel model with additional cross-diffusion and logistic source. part I: Space convergence,* Comput. Math. with Appl. 89(1) (2021), 44–56.

[13] Hashim, M.H. and Harfash, A.J. *Finite element analysis of a Keller-Segel model with additional cross-diffusion and logistic source. part II: Time convergence and numerical simulation,* Comput. Math. with Appl. 109(1) (2022), 216–234.

[14] Hashim, M.H. and Harfash, A.J. *Finite element analysis of attraction-repulsion chemotaxis system. Part I: Space convergence,* Commun. Appl. Math. Comput. 4(3) (2022), 1011–1056.

[15] Hashim, M.H. and Harfash, A.J. *Finite element analysis of attraction-repulsion chemotaxis system. Part II: Time convergence, error analysis and numerical results,* Commun. Appl. Math. Comput. 4(3) (2022), 1057–1104.

[16] Hassan, S.M. and Harfash, A.J. *Finite element analysis of a two-species chemotaxis system with two chemicals,* Appl. Num. Math. 182 (2022), 148–175.

[17] Hassan, S.M. and Harfash, A.J. *Finite element analysis of the two-competing-species Keller–Segel chemotaxis model,* Comput. Math. Model. 33(4) (2022), 443–471.

[18] Hassan, S.M. and Harfash, A.J. *Finite element approximation of a Keller–Segel model with additional self-and cross-diffusion terms and a logistic source,* Commun. Nonlinear Sci. Num. Simul. 104 (2022), 106063.

[19] Hassan, S.M. and Harfash, A.J. *Finite element analysis of chemotaxis-growth model with indirect attractant production and logistic source,* Int. J. Comput. Math. 100(4) (2023), 745–774.

[20] Joseph, D.D. and Preziosi, L. *Heat waves,* Rev. Modern Phys. 61(1) (1989), 41.

[21] Joshi, A.A. and Majumdar, A. *Transient ballistic and diffusive phonon heat transport in thin films,* J. Appl. Phys. 74(1) (1993), 31–39.

[22] Nikan, O., Avazzadeh, Z., and Tenreiro Machado, J.A. *Numerical treatment of microscale heat transfer processes arising in thin films of metals,* Int. Commun. Heat Mass Transf. 132 (2022), 105892.

[23] Pajand, M.R., Moghaddam, N.G. and Ramezani, M.R. *Review of the strain-based formulation for analysis of plane structures part ii: Evaluation of the numerical performance,* Iran. J. Num. Anal. Optim. 11(2) (2021), 485–511.

[24] Qiu, T.Q. and Tien, C.L. *Short-pulse laser heating on metals,* Int. J. Heat Mass Transf. 35(3) (1992), 719–726.

[25] Qiu, T.Q. and Tien, C.L. *Heat transf. mechanisms during short-pulse laser heating of metals*, J. Heat Transf. 115(4) (1993), 835–841.

[26] Tzou, D.Y. *Experimental support for the lagging behavior in heat propagation,* J. Thermophys. Heat Transf. 9(4) (1995), 686–693.

[27] Tzou, D.Y. *The generalized lagging response in small-scale and high-rate heating,* Int. J. Heat Mass Transf. 38(17) (1995), 3231–3240.

[28] Tzou, D.Y. *A unified field approach for heat conduction from macro-to micro-scales,* J. Heat Transf. 117(1) (1995), 8–16.

[29] Yeganeh, S., Mokhtari, R. and Fouladi, S. *Using a LDG method for solving an inverse source problem of the time-fractional diffusion equation,* Iran. J. Num. Anal. Optim. 7(2) (2017), 115–135.

[30] Youssri, Y.H. and Atta, A.G. *Modal spectral Tchebyshev Petrov–Galerkin stratagem for the time-fractional nonlinear burgers' equation,* Iran. J. Num. Anal. Optim. 14(1) (2024), 172–199.

[31] Zhang, J. and Zhao, J.J. *High accuracy stable numerical solution of 1D microscale heat transport equation,* Comm. Num. Methods Eng. 17(11) (2001), 821–832.

[32] Zhang, J. and Zhao, J.J. *Iterative solution and finite difference approximations to 3d microscale heat transport equation,* Math. Comput. Simul. 57(6) (2001), 387–404.

[33] Zhang, J. and Zhao, J.J. *Unconditionally stable finite difference scheme and iterative solution of 2D microscale heat transport equation,* J. Comput. Phys. 170(1) (2001), 261–275.

Iranian Journal of Numerical Analysis and Optimization
Vol. 14, No. 3, 2024, pp 833–874
https://doi.org/10.22067/ijnao.2024.87599.1425
https://ijnao.um.ac.ir/

Research Article

Numerical method for the solution of high order Fredholm integro-differential difference equations using Legendre polynomials

P.T. Pantuvo, G. Ajileye*, R. Taparki and O.O. Aduroja

Abstract

*Corresponding author

Received 13 April 2024; revised 10 July 2024; accepted 12 July 2024

Peter Tsoke Pantuvo
Department of Mathematics and Statistics, Federal University Wukari, Taraba State, Nigeria. e-mail: pantuvo.tsoke@fuwukari.edu.ng

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

Richard Taparki
Department of Mathematical Sciences, Taraba State University, Jalingo, Taraba State, Nigeria. e-mail: richardtaparki01@gmail.com

Ojo Olamiposi Aduroja
Department of Mathematics, University of Ilesa, Ilesa, Osun State, Nigeria. e-mail: olamiposi.aduroja@unilesa.edu.ng

How to cite this article
Pantuvo, P.T., Ajileye, G., Taparki, R., and Aduroja, O.O., Numerical method for the solution of high order Fredholm integro-differential difference equations using Legendre polynomials. *Iran. J. Numer. Anal. Optim.*, 2024; 14(3): 833-874. https://doi.org/10.22067/ijnao.2024.87599.1425

This research paper deals with the numerical method for the solution of high-order Fredholm integro-differential difference equations using Legendre polynomials. We obtain the integral form of the problem, which is transformed into a system of algebraic equations using the collocation method. We then solve the algebraic equation using Newton's method. We establish the uniqueness and convergence of the solution. Numerical problems are considered to test the efficiency of the method, which shows that the method competes favorably with the existing methods and, in some cases, approximates the exact solution.

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

Keywords: Collocation; Fredholm; Integro-differential; Linear and nonlinear; Approximate solution.

1 Introduction

The theory of integral equations is one of the most important branches of mathematics. Currently, considerable interest in mixed integro-differential difference equations has been stimulated due to their numerous applications in the areas of engineering, science, and medicine. In integro-differential difference equations, the unknown function appears to be under the integration sign, and it may also include the derivatives and functional arguments of the unknown function [28]. integro-differential difference equations can be grouped into Fredholm integro-differential difference equations and Volterra integro-differential difference equations. The upper bound of the integral part of the Volterra type is variable, while it is a fixed number for that of the Fredholm type [16].

Many numerical methods have been presented in open literature for solving integro-differential difference equations and integro-differential equations, include the Adomian decompositions method by [19], the collocation method [4, 2, 13], Hybrid linear multistep method [17, 6, 21], Homotopy analysis method [18], Bernoulli matrix method [10], Differential transform method [15], Shifted Legendre polynomials [23], Bernstein Polynomials Method [22], Differential transformation [12], Chebyshev polynomials[24], Lucas series and

polynomials [14], Optimal Auxiliary Function Method (OAFM) [30], Block pulse functions operational matrices [26], and Spectral Homotopy Analysis Method [8]. Ajileye and Aminu [5] presented the standard collocation method to solve first-order Volterra integro-differential equations. Assuming an approximation solution, the class of integro-differential equations was restated in terms of the derived polynomial. After solving for the unknown, we collocated the resultant equation at many points within the range $[0, 1]$, yielding a system of linear algebraic equations. Ajileye et al. [7] introduced a collocation method for the computational solution of the integro-differential equations with Fredholm- Volterra fractional order. They first obtained the problem in linear integral form, which they then converted into a set of linear algebraic equations using standard collocation points.

This research paper considers the integro-differential difference equation of the form

$$\sum_{n=0}^{\alpha} P_n(x)\, y^{(n)} = \sum_{m=0}^{M} Q_m(x)\, y^{(m)}(x-\tau) + g(x) + \lambda \int_a^b K(x,t)\, y^L(t-\tau)\, d\tau, \quad x \in [a,b] = [-1,1], \tag{1}$$

with the initial condition

$$y^{(m-1)}(a) = y_{m-1}, \tag{2}$$

where $g, P, Q \in C([a,b], \mathbb{R})$, $K \in \left([a,b]^2, \mathbb{R}\right)$, λ and y_{m-1} are known constants. $P_\alpha(x) = 1$, $\alpha > M$.

2 Basic definitions

In this section, we define some basic terms that would be encountered in this research.

Definition 1 (Integral equation [9])**.** Given an integral equation

$$y(x) = u(x) + \int_{x_0}^{x_f} k((t,s), y(s))\, ds, \tag{3}$$

then if

(i) $k(t,s) = k(s,t)$, then it is symmetry.

(ii) $k(t,s) = k(a+b-t,\ a+b-s)$ is linear then the kernel is centrosymmetric.

(iii) If $\mathrm{k}(t,s,y(s)) = k(t-s)\ g(s,y(s))$ then the equation is called convolution integral equation and if $g(s,y(s)) = y(s)$, it is called linear.

Definition 2 (Normed space [9])**.** Let X be a nonvector space over k. A norm on x is a function $\|\cdot\| : X \to X$ such that for all $x, y \in X$ and $\alpha \in X$

(i) $\|x\| \geq 0$,

(ii) $\|x\| = 0$ if and only if $x = 0$,

(iii) $\|\alpha x\| = |\alpha|\,\|x\|$,

(iv) $\|x+y\| \leq \|x\| + \|y\|$.

A vector space X on which there is a norm is called a normed space.

Definition 3 (Banach space [9])**.** Banach space is a complete normed space.

Definition 4 (Lipschitzian continuity [9])**.** Let $(X, \|\cdot\|)$ be a normed space. A mapping $T : X \to X$ is L-Lipschitz if there exists $L > 0$ such that $\|Tx - Ty\|_\infty \leq L\,\|x-y\|_\infty$ for all $x, y \in X$.

Definition 5 (q- contraction [9])**.** Let $(X, \|\cdot\|)$ be a normed space. The mapping $T : X \to X$ is a $q-$contraction if $\|Tx_1 - Tx_2\|_\infty \leq q\,\|x_1 - x_2\|_\infty, q \in [0,1)$ fixed for all $x_1, x_2 \in X$.

Definition 6 (Strict q-contraction [9])**.** Let $(X, \|\cdot\|)$ be a norm space. The mapping $T : X \to X$ is strict q-contraction when

$$\|T^n x_1 - T^n x_2\|_\infty \leq q^n\,\|x_1 - x_2\|_\infty \quad \text{for all } x_1, x_2 \in X. \tag{4}$$

Definition 7 (nth integration [20])**.** Let $u(x)$ be an integrable function; then

$${}_aI_x^k\,(u(x)) = \frac{1}{\Gamma(k)} \int_a^x (x-t)^{k-1}\,u(t)dt, \tag{5}$$

$${}_aI_x^k \left(u^{(k)}(x)\right) = u(x) - \sum_{i=0}^{k-1} \frac{x^i}{i!} u^{(i)}(a). \tag{6}$$

Definition 8 (Legendre polynomial [1]). Legendre polynomial on the interval $[-1,1]$ can be determined with the aid of the recurrence formulas

$$L_{n+1}(x)=\frac{2n+1}{n+1}xL_n(x)-\frac{n}{n+1}L_{n-1}(x),\qquad n=1,2,\ldots, \tag{7}$$

where $L_0(x)=1$, $L_1(x)=x$. In order to use these polynomials on the interval $x\in[0,1]$, shifted Legendre polynomial is then defined by the recurrence formula

$$p_{n+1}(x)=\frac{(2n+1)(2x-1)}{(n+1)}p_n(x)-\frac{n}{n+1}p_{n-1}(x), \tag{8}$$

where $p_0=1$, $p_1(x)=2x-1$. The analytical form of degree n is defined as

$$p_n(x)=\sum_{k=0}^{n}\frac{(-1)^{n+k}-\Gamma(n+k+1)}{\Gamma(n-k+1)(\Gamma(k+1))^2}x^k. \tag{9}$$

Theorem 1 (Banach's fixed point theorem [25]). Let $(X,\|\cdot\|)$ be a complete norm space; then each contraction mapping $T:X\to X$ has a unique fixed point x of T in X, such that $T(x)=x$.

3 Methodology

This section considers the development of our method, which was achieved by developing the integral form of the modeled (1) and obtaining the algebraic equations using some lemmas.

Lemma 1. Let $y\in C([a,b],\mathbb{R})$ be the solution to (1) and (2), let $K\in C\left([a,b]^2,\mathbb{R}\right)$, and let g and $Q\in C([a,b],\mathbb{R})$. Then (1) and (2) are equivalent to

$$\begin{aligned}y(x)=H(x)&+\frac{1}{\Gamma(\alpha)}\sum_{m=0}^{M}\int_{x_0}^{x}(x-t)^{\alpha-1}Q_m(t)\,y^{(m)}(t)\,dt\mathbf{M}_{-1}\\&-\frac{1}{\Gamma(\alpha)}\sum_{n=0}^{\alpha-1}\int_{x_0}^{x}(x-t)^{\alpha-1}P_n(t)\,y^{(n)}(t)\,dt\\&+\frac{\lambda}{\Gamma(\alpha)}\int_{x_0}^{x}(x-t)^{\alpha-1}\left[\int_a^b K(t,\tau)\,y^L(\tau)\,d\tau\right]dt\mathbf{M}_{-1}^{L},\end{aligned} \tag{10}$$

where

$$H(x)=\sum_{n=0}^{\alpha-1}\frac{x^i}{i!}y_i+\frac{1}{\Gamma(\alpha)}\int_{x_0}^{x}(x-t)^{\alpha-1}g(t)\,dt.$$

Proof. Equation (10) can be written as

$$y^{(\alpha)}(x)=\sum_{m=0}^{M}Q_m(x)\,y^{(m)}(x-\tau)-\sum_{n=0}^{\alpha-1}P_n(x)\,y^{(n)}(x)$$
$$+g(x)+\lambda\int_a^b K(x,t)\,y^L(t-\tau)\,dt. \tag{11}$$

Using [29] gives

$$y(x-\tau)=y(x)\,\mathbf{M}_{-1}, \tag{12}$$

where

$$\mathbf{M}_{-1}=\begin{bmatrix}\binom{0}{0}(-\tau)^0 & \binom{1}{0}(-\tau)^0 & \cdots & \binom{N}{0}(-\tau)^N\\ 0 & \binom{1}{1}(-\tau)^0 & \cdots & \binom{N}{1}(-\tau)^{N-1}\\ \vdots & \vdots & \vdots & \vdots\\ 0 & 0 & \cdots & \binom{N}{N}(-\tau)^0\end{bmatrix}. \tag{13}$$

Substituting (12) into (11) gives

$$y^{(\alpha)}(x)=\sum_{m=0}^{M}Q_m(x)\,y^{(m)}(x)\,\mathbf{M}_{-1}-\sum_{n=0}^{\alpha-1}P_n(x)\,y^{(n)}(x)$$
$$+g(x)+\lambda\int_{x_0}^{b}K(x,t)\,y^L(t)\,dt\mathbf{M}_{-1}^L. \tag{14}$$

Using (6) in (14) ,we have

$$y(x)=\sum_{i=0}^{n-1}\frac{x^i}{i!}y^i(a)+\frac{1}{\Gamma(\alpha)}\sum_{m=0}^{M}\int_{x_0}^{x}(x-t)^{\alpha-1}Q_m(t)\,y^{(m)}(t)\,dt\,|\mathbf{M}_{-1}|$$
$$-\frac{1}{\Gamma(\alpha)}\sum_{n=0}^{\alpha-1}\int_{x_0}^{x}(x-t)^{\alpha-1}P_n(t)\,y^{(n)}(t)\,dt$$
$$+\frac{1}{\Gamma(\alpha)}\int_{x_0}^{x}(x-t)^{\alpha-1}g(x)\,dt$$

$$+\frac{\lambda}{\Gamma(\alpha)}\int_{x_0}^{x}(x-t)^{\alpha-1}\left[\int_a^b K(x,\tau)\,y^L(\tau)\,d\tau\right]dt\left|\mathbf{M}_{-1}^L\right|,$$

$$\begin{aligned}y(x)=H(x)&+\frac{1}{\Gamma(\alpha)}\sum_{m=0}^{M}\int_{x_0}^{x}(x-t)^{\alpha-1}Q_m(t)\,y^{(m)}(t)\,dt\,|\mathbf{M}_{-1}|\\&-\frac{1}{\Gamma(\alpha)}\sum_{n=0}^{\alpha-1}\int_{x_0}^{x}(x-t)^{\alpha-1}P_n(t)\,y^{(n)}(t)\,dt\\&+\frac{\lambda}{\Gamma(\alpha)}\int_{x_0}^{x}(x-t)^{\alpha-1}\left[\int_a^b K(t,\tau)\,y^L(\tau)\,d\tau\right]dt\left|\mathbf{M}_{-1}^L\right|,\end{aligned}\tag{15}$$

where

$$H(x)=\sum_{n=0}^{\alpha-1}\frac{x^i}{i!}y_i+\frac{1}{\Gamma(\alpha)}\int_{x_0}^{x}(x-t)^{\alpha-1}g(t)\,dt.$$

□

3.1 Method of solution

Let the solution to (10) be approximated by

$$y(x)=\mathbf{P}(x)\,\mathbf{A},\tag{16}$$

where

$$\begin{aligned}P_0(x)=1,\quad P_1(x)=x,\\\mathbf{P}(x)=[P_0(x)\quad P_1(x)\dots P_N(x)]\end{aligned}$$

is the polynomial defined by

$$P_n(x)=\sum_{m=0}^{M}\frac{(-1)^m\,\Gamma(2n-2m+1)}{2^n\Gamma(m+1)\,\Gamma(n-m+1)\,\Gamma(n-2m+1)}x^{n-2m},\tag{17}$$

$$P_n(x)=\sum_{m=0}^{M}Q(n;m)\,x^{n-2m},$$

where

$$Q(n;m)=\sum_{m=0}^{M}\frac{(-1)^m\,\Gamma(2n-2m+1)}{2^n\Gamma(m+1)\,\Gamma(n-m+1)\,\Gamma(n-2m+1)},$$

$$M=floor\left(\frac{n}{2}\right)\text{ and }\mathbf{A}=[a_0\ \ a_1\ \ \ldots\ \ a_N]^T$$

are constant to be determined.

Equation (16) can be written in the form

$$y(x)=\mathbf{X}(x)\,\mathbf{DA}, \tag{18}$$

where (i) when N is even, we have

$$\mathbf{X}(x)=\left[1\ x^2\ \ldots,\ x^4\ \ldots x^{2n}\right],\quad n=0,1,\ldots,$$

$$\mathbf{D}_{even}=\begin{bmatrix} D(0;0) & 0 & 0 & \cdots & 0\\ D(2;1) & D(2;0) & 0 & \cdots & 0\\ \vdots & \vdots & \vdots & \cdots & \vdots\\ D\left(N;\frac{N}{2}\right) & D\left(N;\frac{N}{2}-1\right) & D\left(N;\frac{N}{2}-2\right) & \cdots & D(N;0)\end{bmatrix}^T. \tag{19}$$

(ii) when N is odd, we have

$$\mathbf{X}(x)=[x\ \ x^3\ \ x^5\ \ \ldots\ \ x^{2n+1}],\quad n=0,(1),\frac{N-1}{2}, \tag{20}$$

$$\mathbf{D}_{odd}=\begin{bmatrix} D(1;0) & 0 & 0 & \cdots & 0\\ D(3;1) & D(3;0) & 0 & \cdots & 0\\ \vdots & \vdots & \vdots & \cdots & \vdots\\ D\left(N;\frac{N-1}{2}\right) & D\left(N;\frac{N-3}{2}\right) & D\left(N;\frac{N-5}{2}\right) & \cdots & D(N;0)\end{bmatrix}^T. \tag{21}$$

Hence,

$$y^{(n)}(x)=\mathbf{X}^{(n)}(x)\,\mathbf{DA} \tag{22}$$

writing

$$\mathbf{P}(x)=\mathbf{X}(x)\,\mathbf{D}. \tag{23}$$

Lemma 2. Let $y\in C([a,b],\mathbb{R})$ be defined by (18); then

$$y^{(m)}(x)=\frac{\Gamma(n+1)}{\Gamma(n-m+1)}x^{n-m}\mathbf{DA}, \tag{24}$$

Proof. Given

$$y(x) = x(x)\,\mathbf{DA},$$

then

$$y^{(m)}(x) = \frac{d^m}{dx^m} x(x)\,\mathbf{DA},$$

$$y^{(m)}(x) = \frac{d^m}{dx^m} x^n \mathbf{DA},\ n = 0(1)N.$$

We prove by induction, when

$$m = 1, \qquad y^{(1)} = nx^{n-1}\mathbf{DA},$$

$$m = 2, \qquad y^{(2)} = n(n-1)x^{n-2}\mathbf{DA},$$

$$m = 3, \qquad y^{(3)} = n(n-1)(n-2)x^{n-3}.\mathbf{DA}$$

Therefore, at

$$\begin{aligned} m = n, \qquad y^{(n)} &= (n-1)(n-2)\cdots(n-m+1)x^{n-m}\mathbf{DA} \\ &= \frac{n(n-1)(n-2)\cdots(n-m+1)(n-m)!}{(n-m)!}x^{n-m}\mathbf{DA} \\ &= \frac{\Gamma(n+1)}{\Gamma(n-m+1)}x^{n-m}\mathbf{DA} \end{aligned} \tag{25}$$

which is the expected result. □

Lemma 3. Let $y \in C([a,b],\mathbb{R})$ be defined by (18), let $K \in C\left([a,b]^2,\mathbb{R}\right)$ be defined by $K(x,t) = x^i t^j$, and let

$$V_1 = \int_a^b K(x,t)\,y^L(t-\tau)\,dt; \tag{26}$$

then (26) is equivalent to

$$V(x;n) = \int_a^b x^i t^j \underbrace{VUU^TVV^TU\ldots U^TV}_{L \text{ times}} dt,$$

where

$$V(x;n) = x^n\left|\mathbf{M}_{-1}\right|, U = \mathbf{DA}.$$

Proof. Let

$$V_1(x) = \int_a^b K(x,t)\,y^L(t-\tau)\,dt.$$

Using (18)

$$y(x) = \mathbf{X}(x)\,\mathbf{DA}$$

then

$$y(x-\tau) = \mathbf{X}(x-\tau)\,\mathbf{DA} = x^n|\mathbf{M_{-1}}|\,\mathbf{DA},\ n = 0(1)N.$$

Hence

$$\begin{aligned} y^L(x-\tau) &= (x\,|\mathbf{M}_{-1}|\,\mathbf{DA})^L \\ &= (x^n\,|\mathbf{M}_{-1}|\,\mathbf{DA})^L, \quad n = 0(1)N \\ &= \underbrace{(x\mathbf{M}_{-1}\mathbf{DA})\,(x^n\,|\mathbf{M}_{-1}|\,\mathbf{DA})^T\,(x^n\,|\mathbf{M}_{-1}|\,\mathbf{DA})\cdots(x^n\,|\mathbf{M}_{-1}|\,\mathbf{DA})^T}_{L \text{ times}}, \\ V_1(x;n) &= \int_a^b x^i t^j \underbrace{VUU^TVV^TU\ldots U^TV}_{L \text{ times}} dt, \end{aligned}$$

where

$$V(x;n) = x^n\,|\mathbf{M}_{-1}|,\ U = \mathbf{DA}.$$

□

Lemma 4. Let y(x) be approximated by (18); then (10) is equivalent to

$$\sum_{n=0}^{\alpha} P_n(x)\frac{\Gamma(k+1)}{\Gamma(k-n+1)}x^{k-n}\mathbf{DA} - \sum_{m=0}^{M} Q_m(x)\frac{\Gamma(k+1)}{\Gamma(k-m+1)}x^{k-m}\,|\mathbf{M}_{-1}|\,\mathbf{DA} - \lambda\int_a^b x^i t^j \underbrace{VUU^TVV^TU\ldots U^TV}_{L \text{ times}} dt = g(x).$$

Proof. It holds that

$$y(x-\tau) = \mathbf{X}(x)\,\mathbf{M}_{-1}\mathbf{DA},$$

$$y(x-\tau) = x^k\mathbf{M}_{-1}\mathbf{DA},\ k = 0(1)N,$$

and

$$y(x) = x^k\mathbf{DA},\ k = 0(1)N.$$

Substituting $y(x-\tau)$ and $y(x)$ in (1), we have

$$\sum_{n=0}^{\alpha} P_n(x)\frac{\Gamma(k+1)}{\Gamma(k-n+1)}x^{k-n}\mathbf{DA} - \sum_{m=0}^{M} Q_m(x)\frac{\Gamma(k+1)}{\Gamma(k-m+1)}x^{k-m}\,|\mathbf{M}_{-1}|\,\mathbf{DA}$$

$$-\lambda\int_a^b x^i t^j \underbrace{VUU^TVV^TU\ldots U^TV}_{L\text{ times}}dt = g(x). \tag{27}$$

Collocating (27) gives

$$\sum_{n=0}^{\alpha} P_n(x_i)\frac{\Gamma(k+1)}{\Gamma(k-n+1)}x_i^{k-n}\mathbf{DA} - \sum_{m=0}^{M} Q_m(x_i)\frac{\Gamma(k+1)}{\Gamma(k-m+1)}x_i^{k-m}|\mathbf{M}_{-1}|\mathbf{DA} - \lambda\int_a^b x_i^i t^j \underbrace{VUU^TVV^TU\ldots U^TV}_{L\text{ times}}dt = g(x_i),$$

$$\mathbf{F}(x_i) = \mathbf{W}(x_i) - \mathbf{g}(x_i) = 0, \tag{28}$$

where

$$\mathbf{W}(x_i) = \sum_{n=0}^{\alpha} P_n(x_i)\frac{\Gamma(k+1)}{\Gamma(k-n+1)}x_i^{k-n}\mathbf{DA} - \sum_{m=0}^{M} Q_m(x_i)\frac{\Gamma(k+1)}{\Gamma(k-m+1)}x_i^{k-m}|\mathbf{M}_{-1}|\mathbf{DA} - \lambda\int_a^b x^i t^j \underbrace{VUU^TVV^TU\ldots U^TV}_{L\text{ times}}dt = g(x).$$

□

3.2 Uniqueness of the method

In this section, we assume that the solution to (1) and (2) exists. We then establish the uniqueness of solution and present solutions from the method of solution.

Theorem 2 (Uniqueness theorem)**.** Let $T : C([a,b],\mathbb{R}) \to C([a,b],\mathbb{R})$ be a mapping, let $y \in C([a,b],\mathbb{R})$ be the solution to (10), and let $C([a,b],\mathbb{R})$ be a Banach space.

In order to apply the uniqueness of solution, we have to establish the following:

i. Continuity of T,

ii. T is a $q-$*contraction*,

iii. T is strict contraction.

In order to prove the uniqueness theorem, we use the following hypothesis [3]:

$$
\begin{aligned}
H_1: &\quad P^* = \sum_{n=0}^{N-1} \sup_{x\in J} |P_n(x)|, \\
H_2: &\quad K^* = \sup_{x\in J} \int_a^b |K(t,s)|\, ds, \\
H_3: &\quad Q^* = \sum_{n=0}^{N-1} \sup_{x\in J} |Q_m(x)|, \\
H_4: &\quad \left|y_1^{(m)} - y_2^{(m)}\right| \le L_m |y_1 - y_2| \text{ for all } m \ge 0, \\
H_5: &\quad \left|y_1^L - y_2^L\right| \le H^L |y_1 - y_2|, \\
H_6: &\quad \sup_{x\in J} \left|y_N^{(m)}\right| = \zeta_m, \\
H_7: &\quad \sup_{x\in J} \left|y_N^{(n)}\right| = \zeta_n,
\end{aligned}
$$

where $J = [-1,1]$.

Theorem 3 (Continuity)**.** Let $T : C([a,b],\mathbb{R}) \to C([a,b],\mathbb{R})$ be a mapping, let $y \in C([a,b],\mathbb{R})$ be a solution to (10) and let $C([a,b],\mathbb{R})$ be a Banach space. If $\lim_{h\to\infty} y_h(x) = y(x)$, then T is continuous on $C([a,b],\mathbb{R})$ if $\|Ty_h - Ty\|_\infty \to 0$ as $h \to \infty$.

Proof. It holds that

$$
\begin{aligned}
&|(Ty_h)(x) - (Ty)(x)| \\
&\le \frac{1}{\Gamma(\alpha)} \sum_{n=0}^{\alpha-1} \int_a^x (x-t)^{\alpha-1} |P_n(t)| \left|y_h^{(n)}(t) - y^{(n)}(t)\right| dt \\
&\quad + \frac{1}{\Gamma(\alpha)} \left[\sum_{m=0}^{\alpha} \int_a^x (x-t)^{\alpha-1} \left|Q_m^{(m)}(t)\right| \left|y_h(t) - y^{(m)}(t)\right| dt\right] |\mathbf{M}_{-1}| \\
&\quad + \frac{|\lambda|}{\Gamma(\alpha)} \int_a^x (x-t)^{\alpha-1} \left[\int_a^b |K(t,s)| \left|y_h^L(s) - y^L(s)\right| ds\right] dt \left|\mathbf{M}_{-1}^L\right|.
\end{aligned}
$$

Using hypothesis H_4, we have

$$
\begin{aligned}
&|(Ty_h)(x) - (Ty)(x)| \\
&\leq \frac{L_n}{\Gamma(\alpha)} \sum_{n=0}^{\alpha-1} \int_a^x (x-t)^{\alpha-1} |P_n(t)| \, |y_h - y| \, dt \\
&\quad + \frac{L_m}{\Gamma(\alpha)} \left[\sum_{m=0}^{\alpha} \int_a^x (x-t)^{\alpha-1} |Q_m(t)| \, |y_h - y| \, dt \right] |\mathbf{M}_{-1}| \\
&\quad + \frac{|\lambda| H^L}{\Gamma(\alpha)} \int_a^x (x-t)^{\alpha-1} \left[\int_a^b |K(t,s)| \, |y_h(s) - y(s)| \, ds \right] dt \left| \mathbf{M}_{-1}^L \right|.
\end{aligned}
$$

Taking the supremum of both sides gives

$$
\begin{aligned}
&\sup_{x \in J} |(Ty_h)(x) - (Ty)(x)| \\
&\leq \frac{L_n}{\Gamma(\alpha)} \sum_{n=0}^{\alpha-1} \int_a^x (x-t)^{\alpha-1} \sup_{x \in J} |P_n(t)| \sup_{x \in J} |y_h(t) - y(t)| \, dt \\
&\quad + \frac{L_m}{\Gamma(\alpha)} \left[\sum_{m=0}^{\alpha} \int_a^x (x-t)^{\alpha-1} \sup_{x \in J} |Q_m(t)| \sup_{x \in J} |y_h(t) - y(t)| \, dt \right] |\mathbf{M}_{-1}| \\
&\quad + \frac{|\lambda_1| H^L}{\Gamma(\alpha)} \int_a^x (x-t)^{\alpha-1} \left[\sup_{x \in J} \int_a^b |K(t,s)| \sup_{x \in J} |y_h(s) - y(s)| \, ds \right] dt \left| \mathbf{M}_{-1}^L \right|.
\end{aligned}
$$

Using the hypothesis, we have

$$
\begin{aligned}
&\|Ty_h - Ty\|_\infty \\
&\leq \frac{L_m P^*}{\Gamma(\alpha)} \|y_h - y\|_\infty \int_a^x (x-t)^{\alpha-1} \, dt \\
&\quad + \frac{L_n Q^*}{\Gamma(\alpha)} \|y_h - y\|_\infty \int_a^x (x-t)^{\alpha-1} \, dt \, |\mathbf{M}_{-1}| \\
&\quad + \frac{H^L |\lambda|}{\Gamma(\alpha)} K^* |y_h - y| \left| \mathbf{M}_{-1}^L \right| \int_a^x (x-t)^{\alpha-1} \, dt \\
&\leq \frac{(x-a)^\alpha P^* L_m}{\Gamma(\alpha+1)} \|y_h - y\|_\infty + \frac{(x-a) Q P^* L_n}{\Gamma(\alpha+1)} |\mathbf{M}_{-1}| \, \|y_h - y\|_\infty \\
&\quad + \frac{H^L |\lambda| K^* \left| \mathbf{M}_{-1}^L \right|}{\Gamma(\alpha+1)} \|y_h - y\|_\infty \\
&\leq \frac{1}{\Gamma(\alpha+1)} \left(P^* L_m + Q^* |\mathbf{M}_{-1}| L_n + H^L |\lambda| K^* \left| \mathbf{M}_{-1}^L \right| \right) \|y_h - y\|_\infty .
\end{aligned}
$$

Since $\lim_{h \to \infty} y_h \to y$, hence

$$
\|Ty_h - Ty\|_\infty \to 0 \text{ as } h \to \infty.
$$

Therefore, T is continuous. □

Theorem 4 (q-contraction)**.** Let $T : C([a,b],\mathbb{R}) \to C([a,b],\mathbb{R})$ be a mapping and let $C([a,b],\mathbb{R})$ be a Banach space. Then T is q-contraction if

$$q = \frac{1}{\Gamma(\alpha+1)}\left(P^* L_n + Q^* L_m \left|\mathbf{M}_{-1}\right| + H^L \left|\lambda\right| K^* \left|\mathbf{M}_{-1}^L\right|\right) < 1. \tag{29}$$

Proof. Using Theorem 1, we have

$$\begin{aligned} Ty(x) = H(x) &+ \frac{1}{\Gamma(\alpha)} \sum_{m=0}^{M} \int_{x_0}^{x} (x-t)^{\alpha-1} Q_m(t)\, y^{(m)}(t) \left|\mathbf{M}_{-1}\right| \\ &- \frac{1}{\Gamma(\alpha)} \sum_{n=0}^{\alpha-1} \int_{x_0}^{x} (x-t)^{\alpha-1} P_n(t)\, y^{(n)}(t)\, dt \\ &+ \frac{\lambda}{\Gamma(\alpha)} \int_{x_0}^{x} (x-t)^{\alpha-1} \left[\int_a^b K(t,\tau)\, y^L(\tau)\, d\tau\right] dt \left|\mathbf{M}_{-1}^L\right|. \end{aligned}$$

Then

$$\begin{aligned} &\left|(Ty_1)(x) - (Ty_2)(x)\right| \\ &\le \frac{1}{\Gamma(\alpha)} \sum_{m=0}^{M} \int_{x_0}^{x} (x-t)^{\alpha-1} \left|Q_m(t)\right| \left|y_1^{(m)}(t) - y_2^{(m)}(t)\right| dt \left|\mathbf{M}_{-1}\right| \\ &\quad + \frac{1}{\Gamma(\alpha)} \left[\sum_{n=0}^{\alpha-1} \int_{x_0}^{x} (x-t)^{\alpha-1} \left|P_n(t)\right| \left|y_1^{(n)}(t) - y_2^{(n)}(t)\right| dt\right] \\ &\quad + \frac{\left|\lambda\right| H^L}{\Gamma(\alpha)} \int_{x_0}^{x} (x-t)^{\alpha-1} \left[\int_a^b \left|K(t,\tau)\right| \left|y_1(\tau) - y_2(\tau)\right| d\tau\right] dt \left|\mathbf{M}_{-1}^L\right|, \end{aligned}$$

$$\begin{aligned} &\left|(Ty_1)(x) - (Ty_2)(x)\right| \\ &\le \frac{L_m}{\Gamma(\alpha)} \sum_{m=0}^{M} \int_{x_0}^{x} (x-t)^{\alpha-1} \left|Q_m(t)\right| \left|y_1 - y_2\right| dt \left|\mathbf{M}_{-1}\right| \\ &\quad + \frac{L_n}{\Gamma(\alpha)} \left[\sum_{n=0}^{\alpha-1} \int_{x_0}^{x} (x-t)^{\alpha-1} \left|P_n(t)\right| \left|y_1 - y_2\right| dt\right] \\ &\quad + \frac{\left|\lambda\right| H^L}{\Gamma(\alpha)} \int_{x_0}^{x} (x-t)^{\alpha-1} \left[\int_a^b \left|K(t,\tau)\right| \left|y_1(\tau) - y_2(\tau)\right| d\tau\right] dt \left|\mathbf{M}_{-1}^L\right| \end{aligned}$$

Taking the supremum of both sides gives

$$
\begin{aligned}
&\sup_{x\in J}\left|\left(Ty_1\right)(x)-\left(Ty_2\right)(x)\right| \\
&\leq \frac{L_m}{\Gamma(\alpha)}\sum_{m=0}^{M}\int_{x_0}^{x}(x-t)^{\alpha-1}\sup_{x\in J}\left|Q_m(t)\right|\sup_{x\in J}\left|y_1-y_2\right|dt\left|\mathbf{M}_{-1}\right| \\
&\quad+\frac{L_n}{\Gamma(\alpha)}\left[\sum_{n=0}^{\alpha-1}\int_{x_0}^{x}(x-t)^{\alpha-1}\sup_{x\in J}\left|P_n(t)\right|\sup_{x\in J}\left|y_1-y_2\right|dt\right] \\
&\quad+\frac{|\lambda|H^L}{\Gamma(\alpha)}\int_{x_0}^{x}(x-t)^{\alpha-1}\left[\sup_{x\in J}\int_a^b\left|K(t,\tau)\right|\sup_{x\in J}\left|y_1(\tau)-y_2(\tau)\right|d\tau\right]dt\left|\mathbf{M}_{-1}^{L}\right| \\
&\leq \frac{1}{\Gamma(\alpha+1)}\left(\frac{P^*L_n}{\Gamma(\alpha)}+Q^*\frac{L_m}{\Gamma(\alpha)}\left|\mathbf{M}_{-1}\right|+H^L|\lambda|K^*\left|\mathbf{M}_{-1}^{L}\right|\right)\left\|y_1-y_2\right\|_\infty
\end{aligned}
$$

Since q is T contraction, then

$$
q=\frac{1}{\Gamma(\alpha+1)}\left(P^*L_n+Q^*L_m\left|\mathbf{M}_{-1}\right|+H^L|\lambda|K^*\left|\mathbf{M}_{-1}^{L}\right|\right)<1. \tag{30}
$$

□

Theorem 5 (Convergence of solution)**.** Let $\left(C\left([a,b],\mathbb{R}\right),\|\cdot\|\right)$ be a norm space, let $y(x)$ and $y_N(t)$ be the exact and approximate solution of (10), respectively. Then

$$
\left\|y_N-y\right\|_\infty\leq\frac{\left\|H_N-H\right\|_\infty+\zeta_mP_n^*+\zeta_nQ_m^*\left|\mathbf{M}_{-1}\right|}{1-q}, \tag{31}
$$

where

$$
q=\frac{1}{\Gamma(\alpha+1)}\left(P^*L_n+Q^*L_m\left|\mathbf{M}_{-1}\right|+H^L|\lambda|K^*\left|\mathbf{M}_{-1}^{L}\right|\right).
$$

Proof. Let $H(t),Q(t)$, and $\mathrm{P}(t)$ be expanded in Legendre polynomial. Then

$$
\begin{aligned}
&\left|y_N(x)-y(x)\right| \\
&\leq\left|H_N(x)-H(x)\right| \\
&\quad+\frac{1}{\Gamma(\alpha)}\sum_{n=0}^{\alpha-1}\int_a^x(x-t)^{\alpha-1}\left|P_n^N(t)\,y_N^{(n)}(t)-P_n(t)\,y^{(n)}(t)\right|dt \\
&\quad+\frac{1}{\Gamma(\alpha)}\sum_{m=0}^{M}\int_a^x(x-t)^{\alpha-1}\left|Q_m^N(t)\,y_N^{(m)}(t)-Q_m(t)\,y^{(m)}(t)\right|dt\left|\mathbf{M}_{-1}\right| \\
&\quad+\frac{|\lambda|}{\Gamma(\alpha)}\int_a^x(x-t)^{\alpha-1}\left[\int_a^b\left|K(t,s)\right|\left|y_N^L(s)-y^L(s)\right|ds\right]dt\left|\mathbf{M}_{-1}^{L}\right|
\end{aligned}
$$

$$
\begin{aligned}
&\leq |H_N(x) - H(x)| \\
&\quad + \frac{1}{\Gamma(\alpha)} \sum_{n=0}^{\alpha-1} \int_a^x (x-t)^{\alpha-1} \left| y_N^{(m)}(t) \right| \left| P_n^N(t) - P_n(t) \right| dt \\
&\quad + \frac{1}{\Gamma(\alpha)} \sum_{n=0}^{\alpha-1} \int_a^x (x-t)^{\alpha-1} |P_n(t)| \left| y_N^{(m)}(t) - y^{(m)}(t) \right| dt \\
&\quad + \frac{1}{\Gamma(\alpha)} \sum_{m=0}^{M} \int_a^x (x-t)^{\alpha-1} \left| y_N^{(n)}(t) \right| \left| Q_m^N(t) - Q_m(t) \right| dt \left| \mathbf{M}_{-1} \right| \\
&\quad + \frac{1}{\Gamma(\alpha)} \sum_{m=0}^{M} \int_a^x (x-t)^{\alpha-1} |Q_m(t)| \left| y_N^{(n)}(t) - y^{(n)}(t) \right| dt \left| \mathbf{M}_{-1} \right| \\
&\quad + \frac{|\lambda|}{\Gamma(\alpha)} \int_a^x (x-t)^{\alpha-1} \left[\int_a^b |K(t,s)| \left| y_N^L(s) - y^L(s) \right| ds \right] dt \left| \mathbf{M}_{-1}^L \right| \\
&\leq \|H_N(x) - H(x)\| + \frac{1}{\Gamma(\alpha)} \sum_{n=0}^{\alpha-1} \int_a^x (x-t)^{\alpha-1} \left| y_N^{(m)}(t) \right| \left| P_n^N(t) - P_n(t) \right| dt \\
&\quad + \frac{L_1}{\Gamma(\alpha)} \sum_{n=0}^{\alpha-1} \int_a^x (x-t)^{N-1} |P_n(t)| \left| y_N(t) - y(t) \right| dt \\
&\quad + \frac{1}{\Gamma(\alpha)} \sum_{m=0}^{M} \int_a^x (x-t)^{\alpha-1} \left| y_N^{(n)}(t) \right| \left| Q_m^N(t) - Q_m(t) \right| dt \left| \mathbf{M}_{-1} \right| \\
&\quad + \frac{L_2}{\Gamma(\alpha)} \sum_{m=0}^{M} \int_a^x (x-t)^{\alpha-1} |Q_m(t)| \left| y_N(t) - y(t) \right| dt \left| \mathbf{M}_{-1} \right| \\
&\quad + \frac{|\lambda| H^L}{\Gamma(\alpha)} \int_a^x (x-t)^{\alpha-1} \left[\int_a^b |K(t,s)| \left| y_N(s) - y(s) \right| ds \right] dt \left| \mathbf{M}_{-1}^L \right|.
\end{aligned}
$$

Taking supremum of both sides, we have

$$
\begin{aligned}
&\sup_{x \in J} |y_N(x) - y(x)| \\
&\leq \sup_{x \in J} |H_N(x) - H(x)| \\
&\quad + \frac{1}{\Gamma(\alpha)} \sum_{n=0}^{\alpha-1} \int_a^x (x-t)^{\alpha-1} \sup_{x \in J} \left| y_N^{(m)}(t) \right| \sup_{x \in J} |P_N^n(t) - P_n(t)| dt \\
&\quad + \frac{L_m}{\Gamma(\alpha)} \sum_{n=0}^{\alpha-1} \int_a^x (x-t)^{\alpha-1} \sup_{x \in J} |P_N(t)| \sup_{x \in J} |y_N(t) - y(t)| dt \\
&\quad + \frac{1}{\Gamma(\alpha)} \sum_{m=0}^{M} \int_a^x (x-t)^{\alpha-1} \sup_{x \in J} \left| y_N^{(n)}(t) \right| \sup_{x \in J} |Q_n^m(t) - Q_m(t)| dt \left| \mathbf{M}_{-1} \right|
\end{aligned}
$$

$$+\frac{L_n}{\Gamma(\alpha)}\sum_{n=0}^{\alpha-1}\int_a^x (x-t)^{\alpha-1}\sup_{x\in J}|Q_m(t)|\sup_{x\in J}|y_N(t)-y(t)|\,dt\,|\mathbf{M}_{-1}|$$
$$+\frac{|\lambda|H^L}{\Gamma(\alpha)}\int_a^x (x-t)^{\alpha-1}\left[\int_a^b |K(t,s)|\sup_{x\in J}|y_N(s)-y(s)|\,ds\right]dt\left|\mathbf{M}_{-1}^L\right|.$$

Therefore,

$$\begin{aligned}\|y_N-y\|_\infty \le \|H_N-H\|_\infty &+\frac{\zeta_m P_n^*}{\Gamma(\alpha)}\int_a^x (x-t)^{N-1}\,dt\\ &+\frac{L_m P^*\|y_N-y\|_\infty}{\Gamma(\alpha)}\int_a^x (x-t)^{\alpha-1}\,dt\\ &+\frac{\zeta_n Q_m^*}{\Gamma(\alpha)}\int_a^x (x-t)^{N-1}\,dt\,|\mathbf{M}_{-1}|\\ &+\frac{L_n Q^*\|y_N-y\|_\infty}{\Gamma(\alpha)}\int_a^x (x-t)^{\alpha-1}\,|\mathbf{M}_{-1}|\\ &+\frac{|\lambda|H^L\|y_N-y\|_\infty K^*}{\Gamma(\alpha)}\int_a^x (x-t)^{\alpha-1}\,dt\left|\mathbf{M}_{-1}^L\right|,\end{aligned}$$

$$\|y_N-y\|_\infty \le \frac{\|H_N-H\|_\infty+\frac{\zeta_m P_n^*}{\Gamma(\alpha+1)}+\frac{\zeta_n Q_m^*}{\Gamma(\alpha+1)}|\mathbf{M}_{-1}|}{1-\frac{L_m P^*}{\Gamma(\alpha+1)}-\frac{L_n Q^*}{\Gamma(\alpha+1)}-\frac{|\lambda|H^L K^*}{\Gamma(\alpha+1)}|\mathbf{M}_{-1}|}. \tag{32}$$

Hence, simplification of (32) gives the required result as follows:

$$\|y_N-y\|_\infty \le \frac{\|H_N-H\|_\infty+\zeta_m P_n^*+\zeta_n Q_m^*|\mathbf{M}_{-1}|}{1-q},$$

where

$$q=\frac{1}{\Gamma(\alpha+1)}\left(P^*L_n+Q^*L_m|\mathbf{M}_{-1}|+H^L|\lambda|K^*\left|\mathbf{M}_{-1}^L\right|\right).$$

□

3.3 Numerical examples

In this section, we present numerical examples to test the efficiency of the method. The results are presented in tables as we consider Chebyshev's points.

Problem 1. [27] Consider a third order Fredholm integro-differential difference equations

$$\begin{aligned} & y^{(3)}(x) - xy'(x) + y''(x-1) - xy(x-1) \\ & = -(x+1)(\sin(x-1) + \cos x) - \cos 2 + 1 + \int_{-1}^{1} y(t-1)\, dt \end{aligned}$$

subject to

$$y(0) = 0, \quad y'(0) = 1, \quad y''(0) = 0$$

with the exact solution

$$y(x) = \sin x.$$

To show q-contraction for Problem 1, we have

$$\begin{aligned} Ty(x) = & \frac{1}{\Gamma(\alpha)} \int_0^x (x-t)^{\alpha-1} ty'(t)\, dt + \frac{1}{\Gamma(\alpha)} \int_0^x (x-t)^{\alpha-1} y''(t)\, dt\, |\mathbf{M}_{-1}| \\ & + \frac{1}{\Gamma(\alpha)} \int_0^x (x-t)^{\alpha-1} ty(t)\, dt\, |\mathbf{M}_{-1}| + \frac{1}{\Gamma(\alpha)} \int_0^x (x-t) f(t)\, dt \\ & + \frac{1}{\Gamma(\alpha)} \int_0^x (x-t)^{\alpha-1} \left[\int_{-1}^{1} y(s)\, ds \right] dt\, |\mathbf{M}_{-1}|, \end{aligned} \tag{33}$$

$$\begin{aligned} Ty_1(x) = & \frac{1}{\Gamma(3)} \int_0^x (x-t)^2 ty_1'(t)\, dt + \frac{1}{\Gamma(3)} \int_0^x (x-t)^2 y_1''(t)\, dt\, |\mathbf{M}_{-1}| \\ & + \frac{1}{\Gamma(3)} \int_0^x (x-t)^2 ty_1(t)\, dt\, |\mathbf{M}_{-1}| + \frac{1}{\Gamma(3)} \int_0^x (x-t) f(t)\, dt \\ & + \frac{1}{\Gamma(3)} \int_0^x (x-t)^{N-1} \left[\int_{-1}^{1} y_1(s)\, ds \right] dt\, |\mathbf{M}_{-1}|, \end{aligned} \tag{34}$$

$$\begin{aligned} Ty_2(x) = & \frac{1}{\Gamma(3)} \int_0^x (x-t)^2 ty_2'(t)\, dt + \frac{1}{\Gamma(3)} \int_0^x (x-t)^2 y_2''(t)\, dt\, |\mathbf{M}_{-1}| \\ & + \frac{1}{\Gamma(3)} \int_0^x (x-t)^2 ty_2(t)\, dt\, |\mathbf{M}_{-1}| + \frac{1}{\Gamma(3)} \int_0^x (x-t) f(t)\, dt \\ & + \frac{1}{\Gamma(3)} \int_0^x (x-t)^2 \left[\int_{-1}^{1} y_2(s)\, ds \right] dt\, |\mathbf{M}_{-1}|, \end{aligned} \tag{35}$$

$$|Ty_1 - Ty_2| \leq \frac{1}{\Gamma(3)} \int_0^x (x-t)^2 |t|\, |y_1'(t) - y_2'(t)|\, dt$$

$$+\frac{1}{\Gamma(3)}\int_0^x (x-t)^2 \left|y_1''(t)-y_2''(t)\right| dt \left|\mathbf{M}_{-1}\right|$$
$$+\frac{1}{\Gamma(3)}\int_0^x (x-t)^2 |t| \left|y_1(t)-y_2(t)\right| dt \left|\mathbf{M}_{-1}\right|$$
$$+\frac{1}{\Gamma(3)}\int_0^x (x-t)^2 \left[\int_{-1}^1 \left|y_1(s)-y_2(s)\right| ds\right] dt \left|\mathbf{M}_{-1}\right|. \quad (36)$$

Using H_4 $\left|y_1^{(m)}(t)-y_2^{(m)}(t)\right| \leq L_m \left|y_1-y_2\right|$ gives

$$\begin{aligned}
&|Ty_1-Ty_2| \\
&\leq \frac{1}{\Gamma(3)}\int_0^x (x-t)^2 |t| L_m \left|y_1-y_2\right| dt \\
&\quad +\frac{1}{\Gamma(3)}\int_0^x (x-t)^2 L_n \left|y_1-y_2\right| dt \left|\mathbf{M}_{-1}\right| \\
&\quad +\frac{1}{\Gamma(3)}\int_0^x (x-t)^2 |t| \left|y_1(t)-y_2(t)\right| dt \left|\mathbf{M}_{-1}\right| \\
&\quad +\frac{1}{\Gamma(3)}\int_0^x (x-t)^2 \left[\int_{-1}^1 \left|y_1(s)-y_2(s)\right| ds\right] dt \left|\mathbf{M}_{-1}\right|. \qquad (37)
\end{aligned}$$

Taking supremum of (37) gives

$$\begin{aligned}
&\sup_{x\in J}|Ty_1-Ty_2| \\
&\leq \frac{1}{\Gamma(3)}\int_0^x (x-t)^2 \sup_{x\in J}|t| \sup_{x\in J}\left|y_1'(t)-y_2'(t)\right| dt \\
&\quad +\frac{1}{\Gamma(3)}\int_0^x (x-t)^2 \sup_{x\in J}\left|y_1''(t)-y_2''(t)\right| dt \left|\mathbf{M}_{-1}\right| \\
&\quad +\frac{1}{\Gamma(3)}\int_0^x (x-t)^2 \sup_{x\in J}|t| \sup_{x\in J}\left|y_1(t)-y_2(t)\right| dt \left|\mathbf{M}_{-1}\right| \\
&\quad +\frac{1}{\Gamma(3)}\int_0^x (x-t)^2 \left[\int_{-1}^1 \sup_{x\in J}\left|y_1(s)-y_2(s)\right| ds\right] dt \left|\mathbf{M}_{-1}\right|, \qquad (38)
\end{aligned}$$

$$\begin{aligned}
\left\|Ty_1-Ty_2\right\|_\infty \leq{}& \frac{L_n}{\Gamma(4)}\left\|y_1-y_2\right\|_\infty + \frac{L_m}{\Gamma(4)}\left\|y_1-y_2\right\|_\infty \left|\mathbf{M}_{-1}\right| \\
&+\frac{1}{\Gamma(4)}\left\|y_1-y_2\right\|_\infty \left|\mathbf{M}_{-1}\right| + \frac{K^*}{\Gamma(4)}\left\|y_1-y_2\right\|_\infty \left|\mathbf{M}_{-1}\right| \quad (39)
\end{aligned}$$

where

$$K^* = \int_{-1}^1 \left|K(s,t)\, ds\right| = 2.$$

Since $\mathrm{K}^* = 2$,

$$\|Ty_1 - Ty_2\|_\infty \leq \left[\frac{1}{\Gamma 4} L_1 + L_2 \left|\mathbf{M}_{-1}\right| + 3\left|\mathbf{M}_{-1}\right|\right] \|y_1 - y_2\|_\infty,$$

$$\text{for } q - \textit{contraction} \quad \frac{1}{\Gamma 4} L_1 + L_2 \left|\mathbf{M}_{-1}\right| + 3\left|\mathbf{M}_{-1}\right| < 1. \tag{40}$$

To show the convergence of solution for Problem 1, we have

$$\begin{aligned} Ty_N(x) = H_N(x) &+ \frac{1}{\Gamma(\alpha)} \int_0^x (x-t)^{\alpha-1} t y_N'(t)\, dt \\ &+ \frac{1}{\Gamma(\alpha)} \int_0^x (x-t)^{\alpha-1} y_N''(t)\, dt \left|\mathbf{M}_{-1}\right| \\ &+ \frac{1}{\Gamma(\alpha)} \int_0^x (x-t)^{\alpha-1} t y_N(t)\, dt \left|\mathbf{M}_{-1}\right| \\ &+ \frac{1}{\Gamma(\alpha)} \int_0^x (x-t)^{\alpha-1} \left[\int_{-1}^1 y_N(s)\, ds\right] dt \left|\mathbf{M}_{-1}\right|, \end{aligned} \tag{41}$$

$$\begin{aligned} Ty(x) = H(x) &+ \frac{1}{\Gamma(\alpha)} \int_0^x (x-t)^{\alpha-1} t y'(t)\, dt \\ &+ \frac{1}{\Gamma(\alpha)} \int_0^x (x-t)^{\alpha-1} y''(t)\, dt \left|\mathbf{M}_{-1}\right| \\ &+ \frac{1}{\Gamma(\alpha)} \int_0^x (x-t)^{\alpha-1} t y(t)\, dt \left|\mathbf{M}_{-1}\right| \\ &+ \frac{1}{\Gamma(\alpha)} \int_0^x (x-t)^{\alpha-1} \left[\int_{-1}^1 y(s)\, ds\right] dt \left|\mathbf{M}_{-1}\right|, \end{aligned}$$

$$\begin{aligned} |Ty_N(x) - Ty(x)| \leq |H_N(x) - H(x)| &+ \frac{1}{\Gamma(3)} \int_0^x (x-t)^2 |t| \left|y_N'(t) - y'(t)\right| dt \\ &+ \frac{1}{\Gamma(3)} \int_0^x (x-t)^2 \left|y_N''(t) - y''(t)\right| dt \left|\mathbf{M}_{-1}\right| \\ &+ \frac{1}{\Gamma(3)} \int_0^x (x-t)^2 |t| \left|y_N(t) - y(t)\right| dt \left|\mathbf{M}_{-1}\right| \\ &+ \frac{1}{\Gamma(3)} \int_0^x (x-t)^2 \left[\int_{-1}^1 \left|y_N(s) - y(s)\right| ds\right] dt \left|\mathbf{M}_{-1}\right| \end{aligned} \tag{42}$$

Using (H_4) $\left|y_1^{(m)} - y_2^{(m)}\right| \leq L_m |y_1 - y_2|$,

$$|Ty_N(x) - Ty(x)| \leq |H_N(x) - H(x)| + \frac{L_1}{\Gamma(\alpha)} \int_0^x (x-t)^\alpha |t| \left|y_N - y\right| dt$$

$$+\frac{L_2}{\Gamma(3)}\int_0^x (x-t)^2 |y_N - y|\, dt\, |\mathbf{M}_{-1}|$$
$$+\frac{1}{\Gamma(3)}\int_0^x (x-t)^2 |t|\, |y_N - y|\, dt\, |\mathbf{M}_{-1}|$$
$$+\frac{1}{\Gamma(3)}\int_0^x (x-t)^2 \left[\int_{-1}^1 |y_N - y|\, ds\right] dt\, |\mathbf{M}_{-1}|, \quad (43)$$

$$\sup_{x\in J} |Ty_N(x) - Ty(x)|$$
$$\leq \sup_{x\in J} |H_N(x) - H(x)| + \frac{L_1}{\Gamma(\alpha)}\int_0^x (x-t)^{\alpha} \sup_{x\in J}|t| \sup_{x\in J}|y_N - y|\, dt$$
$$+\frac{L_2}{\Gamma(3)}\int_0^x (x-t)^2 \sup_{x\in J}|y_N - y|\, dt\, |\mathbf{M}_{-1}|$$
$$+\frac{1}{\Gamma(3)}\int_0^x (x-t)^2 \sup_{x\in J}|t| \sup_{x\in J}|y_N - y|\, dt\, |\mathbf{M}_{-1}|$$
$$+\frac{1}{\Gamma(3)}\int_0^x (x-t)^2 \left[\int_{-1}^1 \sup_{x\in J}|y_N - y|\, ds\right] dt\, |\mathbf{M}_{-1}|, \quad (44)$$

$$\|y_N - y\|_\infty \leq \|H_N - H\|_\infty + \frac{L_1}{\Gamma(4)}\|y_N - y\|_\infty$$
$$+\frac{L_2}{\Gamma(4)}\|y_N - y\|_\infty |\mathbf{M}_{-1}| + \frac{1}{\Gamma(4)}\|y_N - y\|_\infty |\mathbf{M}_{-1}|$$
$$+\frac{K^*}{\Gamma(4)}\|y_N - y\|_\infty |\mathbf{M}_{-1}|, \quad (45)$$

$$K^* = \int_{-1}^1 |K(s,t)\, ds| = 2,$$
$$\left[1 - \frac{L_1}{\Gamma 4} - \frac{L_2}{\Gamma 4}|\mathbf{M}_{-1}| - \frac{3|\mathbf{M}_{-1}|}{\Gamma 4}\right]\|y_N - y\|_\infty \leq \|H_N(x) - H(x)\|,$$
$$\|y_N - y\|_\infty \leq \frac{\Gamma 4\, \|H_N(x) - H(x)\|_\infty}{\Gamma 4 - L_1 - L_2|\mathbf{M}_{-1}| - 3|\mathbf{M}_{-1}|},$$

$$\|y_N - y\|_\infty \leq \frac{\Gamma(4)\|H_N(x) - H(x)\|_\infty}{\Gamma(4) - \Gamma(4)q} \leq \frac{\Gamma(4)\|H_N(x) - H(x)\|_\infty}{\Gamma(4)(1-q)}$$
$$\leq \frac{\|H_N(x) - H(x)\|_\infty}{(1-q)}, \quad (46)$$

since $q < 1$, $\|y_N - y\|_\infty$ exists. Furthermore since H is not affected by the approximate solution, this implies that $H_N - H = 0$. Hence,

$\|y_N - y\|_\infty \leq 0$, which shows that it converges.

Solving Problem 1 numerically gives the following solution.

Solution 1. Comparing with (1) , $l = r = 0$, $b = 1$, $a = -1$, we have

$$\begin{aligned} P_3(x) &= 1, \qquad P_2(x) = 0, \qquad P_1(x) = -x, \\ Q_2(x) &= 1, \qquad Q_1(x) = 0, \qquad Q_0(x) = x, \\ g(x) &= -(x+1)(\sin(x-1) + \cos(x)) - \cos 2 + 1, \\ \lambda &= 1, \qquad k(x,t) = 1. \end{aligned}$$

Using $n = 3$,

$$y^{(3)}(x) = \frac{\Gamma(n+1)}{\Gamma(n-m+1)} x^{n-m} \mathbf{DA}, \quad n = 0(1)N, \quad m = 3,$$

$$-xy'(x) = -\frac{\Gamma(n+1)}{\Gamma(n-m+1)} x^{n-m+1} \mathbf{DA}, \quad n = 0(1)N, \quad m = 1.$$

Using Lemma 1,

$$y^{(2)}(x-1) = \frac{\Gamma(n+1)}{\Gamma(n-m+1)} x^{n-m} |\mathbf{M}_{-1}| \mathbf{DA}, \quad n = 0(1)N, \quad m = 2,$$

$$-xy(x-1) = \frac{\Gamma(n+1)}{\Gamma(n-m+1)} x^{n-m+1} |\mathbf{M}_{-1}| \mathbf{DA}, \quad n = 0(1)N, \quad m = 1.$$

Using Lemma 4, then

$$\begin{aligned} \int_{-1}^{1} y(\tau - 1)\, dt &= \int_{-1}^{1} y(t)\, dt\, |\mathbf{M}_{-1}| \mathbf{DA} \\ &= \int_{-1}^{1} (x-t)^0 t^n dt\, |\mathbf{M}_{-1}| \mathbf{DA}, \end{aligned}$$

$$\begin{aligned} (x-t)^0 &= \left| \frac{\Gamma(n+1)}{\Gamma(n+2)} x^{n+1} \right|_{-1}^{1} |\mathbf{M}_{-1}| \mathbf{DA} \\ &= \frac{\Gamma(n+1)}{\Gamma(n+2)} - \frac{\Gamma(n+1)}{\Gamma(n+2)} (-1)^{n+1} |\mathbf{M}_{-1}| \mathbf{DA}. \end{aligned}$$

Taking $N = 3$ for illustration,

$$\mathbf{A}=\left[a_0 \;\; a_1 \;\; a_2 \;\; a_3\right]^T,\ \mathbf{X}(x)=\left[1 \;\; x \;\; x^2 \;\; x^3\right],\ \mathbf{X}(x_i)=\begin{bmatrix}1&0&0&0\\1&\frac{1}{3}&\frac{1}{9}&\frac{1}{27}\\1&\frac{2}{3}&\frac{4}{9}&\frac{8}{27}\\1&1&1&1\end{bmatrix}.$$

Using Lemma 2 with $m=3$

$$y'''(x)=\begin{bmatrix}0&0&0&6\end{bmatrix}\mathbf{DA},\quad y'''(x_i)=\begin{bmatrix}0&0&0&6\\0&0&0&6\\0&0&0&6\\0&0&0&6\end{bmatrix}\mathbf{DA},$$

$$-xy'(x)=\begin{bmatrix}0&-x&-2x^2&-3x^3\end{bmatrix}\mathbf{DA},\quad xy'(x_i)=\begin{bmatrix}0&1&-2&3\\0&\frac{1}{3}&-\frac{2}{9}&\frac{1}{9}\\0&-\frac{1}{3}&-\frac{2}{9}&\frac{-1}{9}\\0&-1&-2&-3\end{bmatrix}\mathbf{DA},$$

$$y''(x-1)=\begin{bmatrix}0&0&2&6x\end{bmatrix}\mathbf{DA},\quad y''(x_i-1)=\begin{bmatrix}0&0&2&-12\\0&0&2&-8\\0&0&2&-4\\0&0&2&0\end{bmatrix}\mathbf{DA},$$

$$-xy(x-1)=\begin{bmatrix}-x&-x^2&-x^3&-x^4\end{bmatrix}\mathbf{DA},\quad x_iy(x_i-1)=\begin{bmatrix}1&-2&4&-8\\\frac{1}{3}&\frac{-4}{9}&\frac{16}{27}&\frac{-64}{81}\\\frac{-1}{3}&\frac{2}{9}&\frac{-4}{27}&\frac{8}{81}\\-1&0&0&0\end{bmatrix}\mathbf{DA},$$

$$\int_{-1}^{1}y(t-1)\,dt=\begin{bmatrix}2&-2&\frac{8}{3}&4\\2&-2&\frac{8}{3}&-4\\2&-2&\frac{8}{3}&-4\\2&-2&\frac{8}{3}&-4\end{bmatrix}\mathbf{DA}.$$

Then

$$\mathbf{W}=y_1^3(x_i)-x_iy^1(x_i)+y''(x_i-1)-x_iy(x_i-1)-\int_{-1}^{1}y(t-1)\,dt,$$

$$\mathbf{W} = \begin{bmatrix} -1 & 1 & \frac{4}{3} & -7 \\ \frac{-5}{3} & \frac{17}{9} & \frac{-8}{27} & \frac{107}{81} \\ \frac{-7}{3} & \frac{17}{9} & \frac{-28}{27} & \frac{485}{81} \\ -3 & 1 & \frac{-8}{3} & 7 \end{bmatrix}.$$

Applying the initial conditions,

$$\mathbf{WW} = \begin{bmatrix} 1 & 0 & 0 & 0 \\ 0 & 1 & 0 & 0 \\ 0 & 0 & 2 & 0 \\ -3 & 1 & \frac{-8}{3} & 7 \end{bmatrix},$$

$$\mathbf{GG} = \begin{bmatrix} 0 & 1 & 0 & 0.33554 \end{bmatrix}^T.$$

For solving

$$\mathbf{F}(A) = \mathbf{WWA} - \mathbf{GG} = \mathbf{0},$$

using Newton Raphson's method gives

$$\mathbf{A} = \begin{bmatrix} 0 & 1 & 0 & \frac{-1060}{111167} \end{bmatrix}.$$

Substituting into the approximate solution gives

$$y_3(x) = -0.09423x^3 + x.$$

Solving at $N = 5, 7, 10$, and 12, we have

$$y_5 = -0.01145575875x^5 + 0.1569475319x^4 - 0.2923832453x^3 + x,$$

$$\begin{aligned} y_7 = {} & 0.00007972949456x^7 - 0.001012399319x^6 + 0.007758248986x^5 \\ & +0.01453036551x^4 - 0.18476635x^3 + x, \end{aligned}$$

$$\begin{aligned} y_{10} = {} & -0.000001697045863x^{10} + 0.00001105299936x^9 \\ & -0.000004369833648x^8 - 0.0004760258412x^7 + 0.0009938127921x^6 \\ & +0.008908454521x^5 - 0.01420598521x^4 - 0.1491471867x^3 + x, \end{aligned}$$

$$y_{12} = 0.000000005662168848x^{12} - 0.0000002473722641x^{11}$$

Table 1: Results of Problem 1

x_i	$exact$	$N=7$	$N=10$	$N=12$	$N=15$
-0.2	-0.19866933	-0.19874343	-0.19858323	-0.19867251	-0.1986693
-0.4	-0.38941834	-0.39009317	-0.38863325	-0.38944734	-0.38941803
-0.6	-0.56464247	-0.56719006	-0.56167599	-0.56475205	-0.56464128
-0.8	-0.71735609	-0.72400265	-0.70961228	-0.71764214	-0.71735297
-1.0	-0.84147098	-0.8555408	-0.82507367	-0.84207669	-0.84146437

Table 2: Absolute error for Problem 1

x_i	ERR_7 [27]	ERR_7	ERR_{10}	ERR_{12}	ERR_{15}
-0.2	2.37×10^{-5}	7.41×10^{-5}	8.61×10^{-5}	3.18×10^{-6}	3.0×10^{-8}
-0.4	1.15×10^{-4}	6.75×10^{-4}	7.85×10^{-4}	2.90×10^{-5}	3.10×10^{-7}
-0.6	8.13×10^{-4}	2.55×10^{-3}	2.97×10^{-3}	1.10×10^{-4}	1.19×10^{-6}
-0.8	2.12×10^{-3}	6.67×10^{-3}	7.74×10^{-3}	2.86×10^{-4}	3.12×10^{-6}
-1.0	4.82×10^{-3}	1.41×10^{-2}	1.64×10^{-2}	6.06×10^{-4}	6.61×10^{-6}

$$
\begin{aligned}
&-0.00000003837633285x^{10}+0.000002937637513x^{9}\\
&-0.00000009352857036x^{8}-0.000204436741x^{7}+0.00002155603046x^{6}\\
&+0.008345804444x^{5}-0.0003081052941x^{4}-0.1662866423x^{3}+x,
\end{aligned}
$$

$$
\begin{aligned}
y_{15}=&-7.406875094e-13x^{15}+1.298432652e-13x^{14}\\
&+0.000000001564193771x^{13}+1.326069211e-11x^{12}\\
&-0.00000002504317927x^{11}-0.000000009095730742x^{10}\\
&+0.000002760036704x^{9}-0.000000022116615 03x^{8}\\
&-0.000198555331x^{7}+0.0000005103898555x^{6}\\
&+0.008333628614x^{5}-0.000007295143911x^{4}\\
&-0.1666576686x^{3}+x.
\end{aligned}
$$

Problem 2. [27] Let us consider the integro-differential difference equation with variable coefficient

$$y''(x)+xy(x)-xy(x-1)+y'(x-1)+y(x-1)=e^{-x}+e+\int_{-1}^{0}ty(t-1)$$

subject to the initial condition

$$y(0) = 1, \qquad y'(0) = -1$$

with the exact solution is given by $y(x) = e^{-x}$.

To show q-contraction for Problem 2 gives

$$\begin{aligned} Ty(x) = & \frac{1}{\Gamma(\alpha)} \int_0^x (x-t)^{\alpha-1} t y'(t)\, dt + \frac{1}{\Gamma(\alpha)} \int_0^x (x-t)^{N-1} y(t)\, dt \\ & - \frac{1}{\Gamma(\alpha)} \int_0^x (x-t)^{\alpha-1} y'(t)\, dt\, |\mathbf{M}_{-1}| - \frac{1}{\Gamma(\alpha)} \int_0^x (x-t)\, y(t)\, dt \\ & + \frac{1}{\Gamma(\alpha)} \int_0^x (x-t)\, f(t)\, dt \\ & + \frac{1}{\Gamma(\alpha)} \int_0^x (x-t)^{N-1} \left[\int_{-1}^0 t y(s)\, ds \right] dt\, |\mathbf{M}_{-1}|, \end{aligned} \tag{47}$$

$$\begin{aligned} Ty_1(x) = & \frac{1}{\Gamma(2)} \int_0^x (x-t)\, t y_1'(t)\, dt + \frac{1}{\Gamma(2)} \int_0^x (x-t)\, y_1(t)\, dt \\ & - \frac{1}{\Gamma(2)} \int_0^x (x-t)\, y_1'(t)\, dt\, |\mathbf{M}_{-1}| - \frac{1}{\Gamma(2)} \int_0^x (x-t)\, y_1(t)\, dt \\ & + \frac{1}{\Gamma(2)} \int_0^x (x-t)\, f(t)\, dt \\ & + \frac{1}{\Gamma(2)} \int_0^x (x-t)^{N-1} \left[\int_{-1}^0 t y_1(s)\, ds \right] dt\, |\mathbf{M}_{-1}|, \end{aligned} \tag{48}$$

$$\begin{aligned} Ty_2(x) = & \frac{1}{\Gamma(2)} \int_0^x (x-t)\, t y_2'(t)\, dt + \frac{1}{\Gamma(2)} \int_0^x (x-t)\, y_2(t)\, dt \\ & - \frac{1}{\Gamma(2)} \int_0^x (x-t)\, y_2'(t)\, dt\, |\mathbf{M}_{-1}| - \frac{1}{\Gamma(2)} \int_0^x (x-t)\, y_2(t)\, dt \\ & + \frac{1}{\Gamma(2)} \int_0^x (x-t)\, f(t)\, dt \\ & + \frac{1}{\Gamma(2)} \int_0^x (x-t)^{N-1} \left[\int_{-1}^0 t y_2(s)\, ds \right] dt\, |\mathbf{M}_{-1}|, \end{aligned} \tag{49}$$

$$\begin{aligned} |Ty_1 - Ty_2| \leq & \frac{1}{\Gamma(2)} \int_0^x (x-t)\, |t|\, |y_1'(t) - y_2'(t)|\, dt \\ & + \frac{1}{\Gamma(2)} \int_0^x (x-t)\, |y_1(t) - y_2(t)|\, dt \end{aligned}$$

$$+\frac{1}{\Gamma(2)}\int_0^x (x-t)\left|y_1'(t)-y_2'(t)\right|dt\left|\mathbf{M}_{-1}\right|$$
$$+\frac{1}{\Gamma(2)}\int_0^x (x-t)\left|y_1(t)-y_2(t)\right|dt$$
$$+\frac{1}{\Gamma(2)}\int_0^x (x-t)\left[\int_{-1}^0 |t|\left|y_1(s)-y_2(s)\right|ds\right]dt\left|\mathbf{M}_{-1}\right| \quad (50)$$

Using (H_4) $\left|y_1^{(m)}(t)-y_2^{(m)}(t)\right| \le L_m\left|y_1-y_2\right|$ gives

$$|Ty_1-Ty_2| \le \frac{L_1}{\Gamma(2)}\int_0^x (x-t)|t|\left|y_1-y_2\right|dt+\frac{1}{\Gamma(2)}\int_0^x (x-t)\left|y_1-y_2\right|dt$$
$$+\frac{L_2}{\Gamma(2)}\int_0^x (x-t)\left|y_1-y_2\right|dt\left|\mathbf{M}_{-1}\right|$$
$$+\frac{1}{\Gamma(2)}\int_0^x (x-t)\left|y_1-y_2\right|dt$$
$$+\frac{1}{\Gamma(2)}\int_0^x (x-t)^{N-1}\left[\int_{-1}^0 |t|\left|y_1-y_2\right|ds\right]dt\left|\mathbf{M}_{-1}\right|. \quad (51)$$

Taking supremum of (51) gives

$$\sup_{x\in J}|Ty_1-Ty_2|$$
$$\le \frac{L_1}{\Gamma(2)}\int_0^x (x-t)\sup_{x\in J}|t|\sup_{x\in J}\left|y_1-y_2\right|dt$$
$$-\frac{L_2}{\Gamma(2)}\int_0^x (x-t)\sup_{x\in J}\left|y_1-y_2\right|dt\left|\mathbf{M}_{-1}\right|$$
$$+\frac{1}{\Gamma(2)}\int_0^x (x-t)\left[\int_{-1}^0 \sup_{x\in J}|t|\sup_{x\in J}\left|y_1-y_2\right|ds\right]dt\left|\mathbf{M}_{-1}\right|, \quad (52)$$

$$\left\|Ty_1-Ty_2\right\|_\infty \le \frac{L_1}{\Gamma(3)}\left\|y_1-y_2\right\|_\infty-\frac{L_2}{\Gamma(3)}\left\|y_1-y_2\right\|_\infty\left|\mathbf{M}_{-1}\right|$$
$$+\frac{K^*}{\Gamma(3)}\left\|y_1-y_2\right\|_\infty\left|\mathbf{M}_{-1}\right|,\ K^*=\int_{-1}^0 |K(s,t)|\,ds=1 \quad (53)$$

$$\left\|Ty_1-Ty_2\right\|_\infty \le \frac{1}{\Gamma 3}\left[L_1-L_2\left|\mathbf{M}_{-1}\right|-K^*\left|\mathbf{M}_{-1}\right|\right]\left\|y_1-y_2\right\|_\infty, \text{since } \mathrm{K}^*=1$$
$$\le \frac{1}{\Gamma 3}\left[L_1-L_2\left|\mathbf{M}_{-1}\right|-\left|\mathbf{M}_{-1}\right|\right]\left\|y_1-y_2\right\|_\infty$$
$$\text{for } q-\text{contraction } \frac{1}{\Gamma 3}\left[L_1-L_2\left|\mathbf{M}_{-1}\right|-\left|\mathbf{M}_{-1}\right|\right]<1. \quad (54)$$

To show the convergence of solution for Problem 2, we have

$$\begin{aligned} Ty_N(x) = H_N(x) &+ \frac{1}{\Gamma(\alpha)}\int_0^x (x-t)^{\alpha-1} t y_N'(t)\,dt \\ &+ \frac{1}{\Gamma(\alpha)}\int_0^x (x-t)^{\alpha-1} y_N(t)\,dt \\ &- \frac{1}{\Gamma(\alpha)}\int_0^x (x-t)^{\alpha-1} y_N'(t)\,dt\,|\mathbf{M}_{-1}| \\ &- \frac{1}{\Gamma(\alpha)}\int_0^x (x-t)^{\alpha-1} y_N(t)\,dt \\ &+ \frac{1}{\Gamma(\alpha)}\int_0^x (x-t)^{\alpha-1}\left[\int_{-1}^0 t y_N(s)\,ds\right]dt\,|\mathbf{M}_{-1}|, \end{aligned} \tag{55}$$

$$\begin{aligned} Ty(x) = H(x) &+ \frac{1}{\Gamma(\alpha)}\int_0^x (x-t)^{\alpha-1} t y'(t)\,dt \\ &+ \frac{1}{\Gamma(\alpha)}\int_0^x (x-t)^{\alpha-1} y(t)\,dt \\ &- \frac{1}{\Gamma(\alpha)}\int_0^x (x-t)^{\alpha-1} y'(t)\,dt\,|\mathbf{M}_{-1}| \\ &- \frac{1}{\Gamma(\alpha)}\int_0^x (x-t)^{\alpha-1} y(t)\,dt \\ &+ \frac{1}{\Gamma(\alpha)}\int_0^x (x-t)^{\alpha-1}\left[\int_{-1}^0 t y(s)\,ds\right]dt\,|\mathbf{M}_{-1}|, \end{aligned} \tag{56}$$

$$\begin{aligned} |Ty_N(x) - Ty(x)| = |H_N(x) - H(x)| &+ \frac{1}{\Gamma(2)}\int_0^x (x-t)\,|t|\,|y_N'(t) - y'(t)|\,dt \\ &+ \frac{1}{\Gamma(2)}\int_0^x (x-t)\,|y_N(t) - y(t)|\,dt \\ &+ \frac{1}{\Gamma(2)}\int_0^x (x-t)\,|y_N'(t) - y'(t)|\,dt\,|\mathbf{M}_{-1}| \\ &+ \frac{1}{\Gamma(2)}\int_0^x (x-t)\,|y_N(t) - y(t)|\,dt \\ &+ \frac{1}{\Gamma(2)}\int_0^x (x-t)\left[\int_{-1}^0 |t|\,|y_N(s) - y(s)|\,ds\right]dt\,|\mathbf{M}_{-1}| \end{aligned} \tag{57}$$

Using $\mathrm{H}_4 = \left|y_1^{(m)} - y_2^{(m)}\right| \leq L_m\,|y_1 - y_2|$, we have

$$\begin{aligned} |Ty_N(x) - Ty(x)| = |H_N(x) - H(x)| &+ \frac{L_n}{\Gamma(2)}\int_0^x (x-t)\,|t|\,|y_N - y|\,dt \\ &+ \frac{1}{\Gamma(2)}\int_0^x (x-t)\,|y_N - y|\,dt \end{aligned}$$

$$+\frac{L_m}{\Gamma(2)}\int_0^x (x-t)\,|y_N-y|\,dt\,|\mathbf{M}_{-1}|$$
$$+\frac{1}{\Gamma(2)}\int_0^x (x-t)\,|y_N-y|\,dt$$
$$+\frac{1}{\Gamma(2)}\int_0^x (x-t)\left[\int_{-1}^0 |t|\,|y_N-y|\,ds\right]dt\,|\mathbf{M}_{-1}|\,. \quad (58)$$

Taking supremum of both sides gives

$$\sup_{x\in J}|Ty_N(x)-Ty(x)| \le \sup_{x\in J}|H_N(x)-H(x)|$$
$$+\frac{L_n}{\Gamma(2)}\int_0^x (x-t)\sup_{x\in J}|t|\sup_{x\in J}|y_N-y|\,dt$$
$$+\frac{1}{\Gamma(2)}\int_0^x (x-t)\sup_{x\in J}|y_N-y|\,dt$$
$$+\frac{L_m}{\Gamma(2)}\int_0^x (x-t)\sup_{x\in J}|y_N-y|\,dt\,|\mathbf{M}_{-1}|$$
$$+\frac{1}{\Gamma(2)}\int_0^x (x-t)\sup_{x\in J}|y_N-y|\,dt$$
$$+\frac{1}{\Gamma(2)}\int_0^x (x-t)\left[\int_{-1}^0 \sup_{x\in J}|t|\sup_{x\in J}|y_N-y|\,ds\right]dt\,|\mathbf{M}_{-1}|\,. \quad (59)$$

$$\|y_N-y\|_\infty \le \|H_N-H\|_\infty + \frac{L}{\Gamma(3)}\|y_N-y\|_\infty + \frac{1}{\Gamma(3)}\|y_N-y\|_\infty$$
$$+\frac{L}{\Gamma(3)}\|y_N-y\|_\infty|\mathbf{M}_{-1}| + \frac{1}{\Gamma(3)}\|y_N-y\|_\infty$$
$$+\frac{K^*}{\Gamma(2)}\|y_N-y\|_\infty|\mathbf{M}_{-1}|\,, \quad (60)$$

$$K^* = \int_{-1}^0 |K(s,t)|\,ds = 1,$$

$$(1-\Gamma(3)-L_1-L_2|\mathbf{M}_{-1}|-|\mathbf{M}_{-1}|)\|y_N-y\|_\infty \le \|H_N-H\|_\infty\,, \quad (61)$$

$$\frac{\Gamma(3)\|H_N-H\|_\infty}{\Gamma(3)-L_1-L_2|\mathbf{M}_{-1}|-|\mathbf{M}_{-1}|} \le \frac{\Gamma(3)\|H_N-H\|_\infty}{\Gamma(3)-\Gamma(3)q} \le \frac{\|H_N-H\|_\infty}{1-q}. \quad (62)$$

Since $q<1$, $\|y_N-y\|_\infty$ exists. Furthermore since H is not affected by the approximate solution, this implies that $H_N-H=0$. Hence,

$$\|y_N-y\|_\infty \le 0, \text{ which shows that it converges.}$$

Solving Problem 2 numerically gives as follows.

Solution 2. Comparing with (1),

$$P_2(x)=1,\quad P_1(x)=1,\quad P_0(x)=x,\quad Q_1(x)=1,\quad Q_0(x)=x-1,$$
$$g(x)=e^{-x}+e,\quad L=0,\lambda=1,\quad k(x,t)=t.$$

Hence,

$$y''(x)=\frac{\Gamma(n+1)}{\Gamma(n-m+1)}x^{n-m}\mathbf{DA},\quad m=2,\ n=0(1)N,$$

$$xy(x)=\frac{\Gamma(n+1)}{\Gamma(n-m+1)}x^{n-m+1}\mathbf{DA},\quad n=0(1)N,$$

$$xy(x-1)=\frac{\Gamma(n+1)}{\Gamma(n-m+1)}x^{n-m}\mathbf{M}_{-1}\mathbf{DA},\quad m=0,\ n=0(1)N,$$

$$y'(x-1)=\frac{\Gamma(n+1)}{\Gamma(n-m+1)}x^{n-m}\mathbf{M}_{-1}\mathbf{DA},\quad m=1,\ n=0(1)N,$$

$$y(x-1)=\frac{\Gamma(n+1)}{\Gamma(n-m+1)}x^{n-m}\mathbf{M}_{-1}\mathbf{DA},\quad m=0,\ n=0(1)N$$

$$\int_{-1}^{0}ty(t-1)\,dt=\int_{-1}^{0}t(X(t))\,dt\mathbf{M}_{-1}\mathbf{DA},$$

$$\int_{-1}^{0}t^{n+1}dt\mathbf{M}_{-1}\mathbf{DA}=\left[\frac{t^{n+2}}{n+2}\right]_{-1}^{0}\mathbf{M}_{-1}\mathbf{DA}=\frac{-(-1)^{n+2}}{n+2}\mathbf{M}_{-1}\mathbf{DA}.$$

Substituting into the approximate solution gives

$$y_3(x)=-0.136348347x^3+0.589442798x^2-x+1,$$

$$\begin{aligned}y_5(x)=&-0.0134885103x^5-0.0135768498x^4-0.0370560278x^3\\&+0.556269874x^2-x+1,\end{aligned}$$

$$\begin{aligned}y_7(x)=&-0.000626560072x^7+0.00170800372x^6-0.0066859407x^5\\&+0.0477358013x^4-0.18169637x^3+0.492416409x^2-x+1,\end{aligned}$$

$$\begin{aligned}y_{10}(x)=&-0.000000191511561x^{10}-0.00000417255283x^9+0.0000275111315x^8\\&-0.000173222035x^7+0.00137098295x^6-0.00841936174x^5\\&+0.0413648129x^4-0.165914444x^3+0.500378553x^2-1x+1,\end{aligned}$$

Table 3: Results of Problem 2

x_i	*exact*	$N=5$	$N=7$	$N=10$	$N=12$	$N=15$
-0.2	1.2214028	1.2225298	1.2212289	1.2214114	1.2214032	1.2214027
-0.4	1.4918247	1.4911653	1.4917137	1.4918302	1.4918249	1.4918247
-0.6	1.8221188	1.8075506	1.82332	1.8220587	1.822116	1.8221189
-0.8	2.2255409	2.1738442	2.2304976	2.2252932	2.2255296	2.2255412
-1.0	2.7182818	2.5932376	2.7308691	2.7176529	2.718253	2.7182825

Table 4: Absolute error for Problem 2

x_i	ERR_{13}[27]	$N=3$	$N=7$	$N=10$	$N=12$
-0.2	2.27×10^{-5}	3.266×10^{-3}	1.739×10^{-4}	8.600×10^{-6}	4.000×10^{-7}
-0.4	1.43×10^{-5}	1.121×10^{-2}	1.110×10^{-4}	5.500×10^{-6}	2.000×10^{-7}
-0.6	1.57×10^{-4}	1.953×10^{-2}	1.201×10^{-3}	6.010×10^{-5}	2.800×10^{-6}
-0.8	6.49×10^{-4}	2.151×10^{-2}	4.957×10^{-3}	2.477×10^{-4}	1.130×10^{-5}
-1.0	1.65×10^{-3}	7.509×10^{-3}	1.259×10^{-2}	6.289×10^{-4}	2.880×10^{-5}

$$\begin{aligned} y_{12}(x) = {} & 0.00000000418649794x^{12} - 0.0000000216397711x^{11} + 0.000000244349171x^{10} \\ & -0.00000282966002x^9 + 0.0000249459268x^8 - 0.000197246671x^7 \\ & +0.00138804522x^6 - 0.00833728371x^5 + 0.0416528216x^4 - 0.166632138x^3 \\ & +0.500017372x^2 - x + 1, \end{aligned}$$

$$\begin{aligned} y_{15}(x) = {} & -5.30985676e^{-13}x^{15} + 1.49454482e^{-11}x^{14} - 1.60904091e^{-10}x^{13} \\ & +0.00000000202630179x^{12} - 0.0000000251416995x^{11} + 0.000000276291033x^{10} \\ & -0.00000275410796x^9 + 0.0000247983934x^8 - 0.000198437964x^7 \\ & +0.00138890723x^6 - 0.00833324776x^5 + 0.0416669666x^4 - 0.166667415x^3 \\ & +0.499999624x^2 - x + 1. \end{aligned}$$

Problem 3. [11] Consider the third-order nonlinear Fredholm integro-differential difference equation

$$u'''(x)+\frac{1}{2}u''+xu'(x)+2u'(x-1)+\frac{1}{2}xu(x)+u(x-1) = e+\int_{-1}^{0} tu^2(t-1)\,dt$$

with the following initial condition

$$u(0)=1, u'(0)=-\frac{1}{2}, u''(0)=\frac{1}{4},$$

the exact solution

$$u(x)=e^{-\frac{x}{2}}.$$

To show q-contraction for Problem 3, we have

$$\begin{aligned} Tu(x) &= \frac{1}{\Gamma(\alpha)}\int_0^x (x-t)^{\alpha-1}\frac{1}{2}u''(t)\,dt + \frac{1}{\Gamma(\alpha)}\int_0^x (x-t)^{\alpha-1} tu'(t)\,dt \\ &+\frac{2}{\Gamma(\alpha)}\int_0^x (x-t)^{\alpha-1} u'(t)\,dt\mathbf{M}_{-1} + \frac{1}{2\Gamma(\alpha)}\int_0^x (x-t)^{\alpha-1} tu(t)\,dt \\ &+\frac{1}{\Gamma(\alpha)}\int_0^x (x-t)^{\alpha-1} u(t)\,dt\mathbf{M}_{-1} + \frac{1}{\Gamma(\alpha)}\int_0^x (x-t) f(t)\,dt \\ &+\frac{1}{\Gamma(\alpha)}\int_0^x (x-t)^{\alpha-1}\left[\int_{-1}^0 tu^2(s)\,ds\right]dt\mathbf{M}_{-1}, \end{aligned} \tag{63}$$

$$\begin{aligned} Tu_1(x) &= \frac{1}{\Gamma(3)}\int_0^x (x-t)^2\frac{1}{2}u_1''(t)\,dt + \frac{1}{\Gamma(3)}\int_0^x (x-t)^2 tu_1'(t)\,dt \\ &+\frac{2}{\Gamma(3)}\int_0^x (x-t)^2 u_1'(t)\,dt\mathbf{M}_{-1} + \frac{1}{2\Gamma(3)}\int_0^x (x-t)^2 tu_1(t)\,dt \\ &+\frac{1}{\Gamma(3)}\int_0^x (x-t)^2 u_1(t)\,dt\mathbf{M}_{-1} + \frac{1}{\Gamma(3)}\int_0^x (x-t) f(t)\,dt \\ &+\frac{1}{\Gamma(3)}\int_0^x (x-t)^2\left[\int_{-1}^0 tu_1^2(s)\,ds\right]dt\mathbf{M}_{-1}, \end{aligned} \tag{64}$$

$$\begin{aligned} Tu_2(x) &= \frac{1}{\Gamma(3)}\int_0^x (x-t)^2\frac{1}{2}u_2''(t)\,dt + \frac{1}{\Gamma(3)}\int_0^x (x-t)^2 tu_2'(t)\,dt \\ &+\frac{2}{\Gamma(3)}\int_0^x (x-t)^2 u_2'(t)\,dt\mathbf{M}_{-1} + \frac{1}{2\Gamma(3)}\int_0^x (x-t)^2 tu_2(t)\,dt \\ &+\frac{1}{\Gamma(3)}\int_0^x (x-t)^2 u_2(t)\,dt\mathbf{M}_{-1} + \frac{1}{\Gamma(3)}\int_0^x (x-t) f(t)\,dt \\ &+\frac{1}{\Gamma(3)}\int_0^x (x-t)^2\left[\int_{-1}^0 tu_2^2(s)\,ds\right]dt\mathbf{M}_{-1}, \end{aligned} \tag{65}$$

$$\begin{aligned} |Ty_1 - Ty_2| &\le \frac{1}{\Gamma(3)}\int_0^x (x-t)^2 |u_1''(t)-u_2''(t)|\,dt \\ &+\frac{1}{\Gamma(3)}\int_0^x (x-t)^2 |t|\,|u_1'(t)-u_2'(t)|\,dt \end{aligned}$$

$$
\begin{aligned}
&+\frac{1}{\Gamma(3)}\int_0^x (x-t)^2 \left|u_1'(t)-u_2'(t)\right| dt\mathbf{M}_{-1}\\
&+\frac{1}{\Gamma(3)}\int_0^x (x-t)^2 |t| \left|u_1(t)-u_2(t)\right| dt\\
&+\frac{1}{\Gamma(3)}\int_0^x (x-t)^2 \left|u_1(t)-u_2(t)\right| dt\mathbf{M}_{-1}\\
&+\frac{1}{\Gamma(3)}\int_0^x (x-t)^2 \left[\int_{-1}^0 |t| \left|u_1^2(s)-u_2^2(s)\right| ds\right] dt\mathbf{M}_{-1}. \qquad (66)
\end{aligned}
$$

Using $H_4 = \left|y_1^{(m)}(t)-y_2^{(m)}(t)\right| \leq L_m |y_1-y_2|$, gives

$$
\begin{aligned}
|Tu_1-Tu_2| \leq{}& \frac{L_n}{\Gamma(3)}\int_0^x (x-t)^2 |u_1-u_2| dt + \frac{L_m}{\Gamma(3)}\int_0^x (x-t)^2 |t| |u_1-u_2| dt\\
&+\frac{L_m}{\Gamma(3)}\int_0^x (x-t)^2 |u_1-u_2| dt\mathbf{M}_{-1}\\
&+\frac{1}{\Gamma(3)}\int_0^x (x-t)^2 |t| |u_1-u_2| dt\\
&+\frac{1}{\Gamma(3)}\int_0^x (x-t)^2 |u_1-u_2| dt\mathbf{M}_{-1}\\
&+\frac{L^2}{\Gamma(3)}\int_0^x (x-t)^2 \left[\int_{-1}^0 |t| |u_1-u_2| ds\right] dt\mathbf{M}_{-1}. \qquad (67)
\end{aligned}
$$

Taking supremum of (67) gives

$$
\begin{aligned}
&\sup_{x\in J}|Tu_1-Tu_2|\\
\leq{}& \frac{L_n}{\Gamma(3)}\int_0^x (x-t)^2 \sup_{x\in J}|u_1-u_2| dt\\
&+\frac{L_m}{\Gamma(3)}\int_0^x (x-t)^2 \sup_{x\in J}|t| \sup_{x\in J}|u_1-u_2| dt\\
&+\frac{L_m}{\Gamma(3)}\int_0^x (x-t)^2 \sup_{x\in J}|u_1-u_2| dt\,|\mathbf{M}_{-1}|\\
&+\frac{1}{\Gamma(3)}\int_0^x (x-t)^2 \sup_{x\in J}|t| \sup_{x\in J}|u_1-u_2| dt\\
&+\frac{1}{\Gamma(3)}\int_0^x (x-t)^2 \sup_{x\in J}|u_1-u_2| dt\,|\mathbf{M}_{-1}|\\
&+\frac{L^2}{\Gamma(3)}\int_0^x (x-t)^2 \left[\int_{-1}^0 \sup_{x\in J}|t| \sup_{x\in J}|u_1-u_2| ds\right] dt\,|\mathbf{M}_{-1}|, \qquad (68)
\end{aligned}
$$

$$\|Ty_1-Ty_2\|_\infty$$

$$
\begin{aligned}
&\leq \frac{L_n}{\Gamma(3)} \|y_1 - y_2\|_\infty \int_0^x (x-t)^2 \, dt + \frac{L_m}{\Gamma(3)} \|y_1 - y_2\|_\infty \int_0^x (x-t)^2 \, dt \\
&+ \frac{L_m}{\Gamma(3)} \|y_1 - y_2\|_\infty \int_0^x (x-t)^2 \, dt \, |\mathbf{M}_{-1}| \\
&+ \frac{1}{\Gamma(3)} \|y_1 - y_2\|_\infty \int_0^x (x-t)^2 \, dt \\
&+ \frac{1}{\Gamma(3)} \|y_1 - y_2\|_\infty \int_0^x (x-t)^2 \, dt \, |\mathbf{M}_{-1}| \\
&+ \frac{L^2 K^*}{\Gamma(3)} \|y_1 - y_2\|_\infty \int_0^x (x-t)^2 \, dt \, |\mathbf{M}_{-1}| \, , \qquad (69)
\end{aligned}
$$

$$
\begin{aligned}
\|Ty_1 - Ty_2\|_\infty \leq \frac{L_1}{\Gamma(4)} \|y_1 - y_2\|_\infty + \frac{L_2}{\Gamma(4)} \|y_1 - y_2\|_\infty \\
+ \frac{L_2}{\Gamma(4)} \|y_1 - y_2\|_\infty |\mathbf{M}_{-1}| + \frac{1}{\Gamma(4)} \|y_1 - y_2\|_\infty \\
+ \frac{1}{\Gamma(4)} \|y_1 - y_2\|_\infty |\mathbf{M}_{-1}| \\
+ \frac{L^2 K^*}{\Gamma(4)} \|y_1 - y_2\|_\infty |\mathbf{M}_{-1}| \, ,
\end{aligned}
$$

$$
K^* = \int_{-1}^0 |K(s,t)| \, ds = 1,
$$

$$
\|Ty_1 - Ty_2\|_\infty \leq \frac{1}{\Gamma(4)} \left[L_1 + L_2 + L_2 |\mathbf{M}_{-1}| + 1 + |\mathbf{M}_{-1}| + L^2 |\mathbf{M}_{-1}| \right] \|y_1 - y_2\|_\infty
$$

$$
\text{for q-contraction, } \frac{1}{\Gamma(4)} \left[L_1 + L_2 + L_2 |\mathbf{M}_{-1}| + 1 + |\mathbf{M}_{-1}| + L^2 |\mathbf{M}_{-1}| \right] < 1. \qquad (70)
$$

To show the convergence of solution for Problem 3, we have

$$
\begin{aligned}
y_N(x) = H_N(x) + \frac{1}{\Gamma(\alpha)} \int_0^x (x-t)^{\alpha-1} \frac{1}{2} u_N''(t) \, dt + \frac{1}{\Gamma(\alpha)} \int_0^x (x-t)^{\alpha-1} t u_N'(t) \, dt \\
+ \frac{2}{\Gamma(\alpha)} \int_0^x (x-t)^{\alpha-1} u_N'(t) \, dt \, |\mathbf{M}_{-1}| + \frac{1}{2\Gamma(\alpha)} \int_0^x (x-t)^{\alpha-1} t u_N(t) \, dt \\
+ \frac{1}{\Gamma(\alpha)} \int_0^x (x-t)^{\alpha-1} u_N(t) \, dt \, |\mathbf{M}_{-1}| \\
+ \frac{1}{\Gamma(\alpha)} \int_0^x (x-t)^{\alpha-1} \left[\int_{-1}^0 t u_N^2(s) \, ds \right] dt \, |\mathbf{M}_{-1}| \, , \qquad (71)
\end{aligned}
$$

$$y(x) = H(x) + \frac{1}{\Gamma(\alpha)} \int_0^x (x-t)^{\alpha-1} \frac{1}{2} u''(t)\, dt + \frac{1}{\Gamma(\alpha)} \int_0^x (x-t)^{\alpha-1} t u'(t)\, dt$$
$$+ \frac{2}{\Gamma(\alpha)} \int_0^x (x-t)^{\alpha-1} u'(t)\, dt\, |\mathbf{M}_{-1}| + \frac{1}{2\Gamma(\alpha)} \int_0^x (x-t)^{\alpha-1} t u(t)\, dt$$
$$+ \frac{1}{\Gamma(\alpha)} \int_0^x (x-t)^{\alpha-1} u(t)\, dt\, |\mathbf{M}_{-1}|$$
$$+ \frac{1}{\Gamma(\alpha)} \int_0^x (x-t)^{\alpha-1} \left[\int_{-1}^0 t u^2(s)\, ds \right] dt\, |\mathbf{M}_{-1}|, \tag{72}$$

$$|y_N(x) - y(x)| \le |H_N(x) - H(x)| + \frac{1}{\Gamma(3)} \int_0^x (x-t)^2 |u_N''(t) - u''(t)|\, dt$$
$$+ \frac{1}{\Gamma(3)} \int_0^x (x-t)^2 |t|\, |u_N'(t) - u'(t)|\, dt$$
$$+ \frac{1}{\Gamma(3)} \int_0^x (x-t)^2 |u_N'(t) - u'(t)|\, dt\, |\mathbf{M}_{-1}|$$
$$+ \frac{1}{\Gamma(3)} \int_0^x (x-t)^2 |t|\, |u_N(t) - u(t)|\, dt$$
$$+ \frac{1}{\Gamma(3)} \int_0^x (x-t)^2 |u_N(t) - u(t)|\, dt\, |\mathbf{M}_{-1}|$$
$$+ \frac{1}{\Gamma(3)} \int_0^x (x-t)^2 \left[\int_{-1}^0 |t|\, |u_N^2(s) - u^2(s)|\, ds \right] dt\, |\mathbf{M}_{-1}| \tag{73}$$

Applying hypothesis (H_4) $\left|y_1^{(m)} - y_2^{(m)}\right| \le L_m |y_1 - y_2|$ for all $m \ge 0$, we have

$$|y_N(x) - y(x)|$$
$$\le |H_N(x) - H(x)| + \frac{L_n}{\Gamma(3)} \int_0^x (x-t)^2 |u_N - u|\, dt$$
$$+ \frac{L_n}{\Gamma(3)} \int_0^x (x-t)^2 |t|\, |u_N - u|\, dt$$
$$+ \frac{L_m}{\Gamma(3)} \int_0^x (x-t)^2 |u_N - u|\, dt\, |\mathbf{M}_{-1}|$$
$$+ \frac{1}{\Gamma(3)} \int_0^x (x-t)^2 |t|\, |u_N - u|\, dt$$
$$+ \frac{1}{\Gamma(3)} \int_0^x (x-t)^2 |u_N - u|\, dt\, |\mathbf{M}_{-1}|$$
$$+ \frac{L^2}{\Gamma(3)} \int_0^x (x-t)^2 \left[\int_{-1}^0 |t|\, |u_N - u|\, ds \right] dt\, |\mathbf{M}_{-1}|. \tag{74}$$

Taking supremum of both sides gives

$$\begin{aligned}
&\sup_{x\in J}|y_N(x)-y(x)|\\
&\leq \sup_{x\in J}|H_N(x)-H(x)|+\frac{L_n}{\Gamma(3)}\int_0^x (x-t)^2\sup_{x\in J}|u_N-u|\,dt\\
&\quad+\frac{L_m}{\Gamma(3)}\int_0^x (x-t)^2\sup_{x\in J}|t|\sup_{x\in J}|u_N-u|\,dt\\
&\quad+\frac{L_m}{\Gamma(3)}\int_0^x (x-t)^2\sup_{x\in J}|u_N-u|\,dt\,|\mathbf{M}_{-1}|\\
&\quad+\frac{1}{\Gamma(3)}\int_0^x (x-t)^2\sup_{x\in J}|t|\sup_{x\in J}|u_N-u|\,dt\\
&\quad+\frac{1}{\Gamma(3)}\int_0^x (x-t)^2\sup_{x\in J}|u_N-u|\,dt\,|\mathbf{M}_{-1}|\\
&\quad+\frac{L^2}{\Gamma(3)}\int_0^x (x-t)^2\left[\int_{-1}^0\sup_{x\in J}|t|\sup_{x\in J}|u_N-u|\,ds\right]dt\,|\mathbf{M}_{-1}|, \qquad (75)
\end{aligned}$$

$$\begin{aligned}
\|u_N-u\|_\infty &\leq \|H_N-H\|_\infty+\frac{L_1}{\Gamma(4)}\|u_N-u\|_\infty\\
&\quad+\frac{L_2}{\Gamma(4)}\|u_N-u\|_\infty+\frac{L_2}{\Gamma(4)}\|u_N-u\|_\infty\mathbf{M}_{-1}\\
&\quad+\frac{1}{\Gamma(4)}\|u_N-u\|_\infty+\frac{1}{\Gamma(4)}\|u_N-u\|_\infty\mathbf{M}_{-1}\\
&\quad+\frac{L^2K^*}{\Gamma(4)}\|u_N-u\|_\infty\mathbf{M}_{-1}, \qquad (76)
\end{aligned}$$

$$K^*=\int_{-1}^0|K(s,t)|\,ds=1,$$

$$\begin{bmatrix}1-\frac{L_1}{\Gamma(3)}+\frac{L_2}{\Gamma(3)}+\frac{L_2}{\Gamma(3)}|\mathbf{M}_{-1}|\\ +\frac{1}{\Gamma(3)}+\frac{1}{\Gamma(3)}|\mathbf{M}_{-1}|+\frac{L^2}{\Gamma(3)}|\mathbf{M}_{-1}|\end{bmatrix}\|u_N-u\|_\infty\leq\|H_N-H\|_\infty,$$

$$\begin{aligned}
\|u_N-u\|_\infty &\leq \frac{\Gamma3\,\|H_N-H\|_\infty}{\Gamma3-L_1-L_2-L_2|\mathbf{M}_{-1}|-1-|\mathbf{M}_{-1}|+L^2|\mathbf{M}_{-1}|}\\
&\leq \frac{\Gamma3\,\|H_N-H\|_\infty}{\Gamma3-\Gamma3q}\leq\frac{\Gamma3\,\|H_N-H\|_\infty}{\Gamma3\,(1-q)}\leq\frac{\|H_N-H\|_\infty}{1-q}. \qquad (77)
\end{aligned}$$

Since $q<1$, $\|u_N-u\|_\infty$ exists. Furthermore since H is not affected by the approximate solution, this implies that $H_N-H=0$. Hence,

$\|y_N - y\|_\infty \leq 0$, which shows that it converges.

Solving Problem 3 numerically gives as follows.

Solution 3. Using $N = 2$ for illustration

$$\mathbf{A} = [a_0 \quad a_1 \quad a_2]^T, \ \mathbf{X} = [1 \quad x \quad x^2],$$

$$\mathbf{M}_{-1}(\tau) = \begin{bmatrix} 1 & -\tau & \tau^2 \\ 0 & 1 & -2\tau \\ 0 & 0 & 1 \\ 0 & 0 & 0 \end{bmatrix}, \text{when } \tau = 1, \mathbf{M}_{-1} = \begin{bmatrix} 1 & -1 & 1 \\ 0 & 1 & -2 \\ 0 & 0 & 1 \\ 0 & 0 & 0 \end{bmatrix},$$

$$u'''(x_i) = \begin{bmatrix} 0 & 0 & 0 \\ 0 & 0 & 0 \\ 0 & 0 & 0 \end{bmatrix}, \ \frac{1}{2}u''(x_i) = \begin{bmatrix} 0 & 0 & 1 \\ 0 & 0 & 1 \\ 0 & 0 & 1 \end{bmatrix},$$

$$x_i u'(x_i) = \begin{bmatrix} 0 & 0 & 1 \\ 0 & \frac{1}{2} & \frac{1}{2} \\ 0 & 1 & 2 \end{bmatrix}, x_i u(x_i) = \begin{bmatrix} 0 & 2 & -4 \\ 0 & 2 & -2 \\ 0 & 2 & 0 \end{bmatrix},$$

$$u(x_i - 1) = \begin{bmatrix} 0 & 0 & 0 \\ \frac{1}{4} & \frac{1}{8} & \frac{1}{16} \\ \frac{1}{2} & \frac{1}{2} & \frac{1}{2} \end{bmatrix}, \ \int_1^0 tu^2(t-1)\,dt = \begin{bmatrix} 1 & -1 & 1 \\ 1 & -\frac{1}{2} & \frac{1}{4} \\ 1 & 0 & 0 \end{bmatrix}.$$

Solving A using Newton's Raphson's method gives

$$A = \begin{bmatrix} 1 & \frac{-1}{2} & \frac{1}{8} \end{bmatrix}.$$

Substituting into the approximate solution gives

$$y_2(x) = 1 - \frac{1}{2}x + \frac{1}{8}x^2,$$

which converges to the exact solution.

4 Discussion of results

In this section, we discussed the results obtained from the solved problems using our developed method and the advantages of the new method over the

existing methods in the literature. We also established the uniqueness and convergence of the solution.

Theorem 2 was used to establish the uniqueness of the method by first establishing that T is continuous, T is q-contraction, and T is strict contraction using some hypothesis. Theorem 3 shows the proof for continuity. Theorem 4 proves the q-contraction. Using the theorem, we proved that the result is a q-contraction, which shows the uniqueness of the method.

Theorem 5 was used to show the convergence of the solution, and it was established that the method converges.

Problem 1: The approximation gives $y_3(x) = -0.09423x^3 + x$, and solving at $N = 5, 7, 10, 12$ and 15, we obtained Table 1, which shows the result obtained from solving Problem 1 at $x_i = -0.2$ to -1.0 at various values of N and the exact solution. Table 2 shows the error of Problem 1, and it indicates that as our N increases, the error result becomes more consistent, particularly when $N = 12$ and $N = 15$. It can be seen that the error is small and more consistent across all values of x_i and the values of N considered. For instance, the least error in [27] at $N = 15$ is 2.37×10^{-5} while the least error in our method is 0.10×10^{-8} at $N = 15$, this clearly shows that our method performed better.

Problem 2: The approximation gives

$$y_3(x) = -0.136348347x^3 + 0.589442798x^2 - x + 1,$$

and solving at $N = 5, 7, 10, 12$ and 15, we obtained Table 3, which shows the result obtained from solving Problem 2 at $x_i = -0.2$ to -1.0 at various values of N and the exact solution. Table 4 shows the error of Problem 2, and it indicates that as our N increases, the error result becomes more consistent, particularly when $N = 12$ and $N = 15$. It can be seen that the error is small and more consistent across all values of x_i and the values of N considered. For instance, comparing the error of [27] at $N = 15$ and that of our method at $N = 12$, this clearly shows that our method performs better.

Problem 3: The approximation gives $y_2(x) = 1 - \frac{1}{2}x + \frac{1}{8}x^2$, which shows that the results converge to the exact solution.

The numerical method is observed to be consistent and converges faster to the exact solution, as shown in Problems 1, 2, and 3. It is also observed that as N increases, the solution gets better. Hence the stability of the method.

Hence, from the results obtained, one may simply conclude that the numerical method derived is more efficient and computationally reliable than the existing methods in the literature.

5 Conclusion

In conclusion, a new numerical method for solving high-order integro-differential difference equations using Legendre polynomials with some conditions solved Fredholm differential difference equations. Our method has proven to be effective and efficient when compared to other methods of solution. In some of the examples, the result gave the exact solution, and for others, as we increase our value of N, the result approaches the exact solution so fast after a few iterations. The comparison of results also shows that our method performed favorably. All of the computations in this paper were performed using MATLAB 15.

References

[1] Abbas, S. and Mehdi, D. *A new operational matrix for solving fractional order differential equations,* Comput. Math. Appl. 59 (2010), 1326–1336.

[2] Adesanya, A.O., Yahaya, Y.A. Ahmed, B. and Onsachi, R.O. *Numerical solution of linear integral and integro-differential equations using Boubakar collocation method*, Inter. J. Math. Anal. Optim. Theory Appl. 2 (2019), 592–598.

[3] Ahmed, A.H., Kirtiwant, P.G. and Shakir, M.A. *The approximate solutions of fractional integro-differential equations by using modified adomian decomposition method*, Khayyam J. Math. 5 (1) (2019), 21-39.

[4] Ajileye, G. and Aminu, F.A. *A numerical method using collocation approach for the solution of Volterra-Fredholm integro-differential equations*, African Scientific Reports 1 (2022), 205–211.

[5] 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.

[6] Ajileye, G. Amoo, S.A. and Ogwumu, O.D. *Hybrid block method algorithms for solution of first order initial value problems in ordinary differential equations*, J. Appl. Comput. Math. 7(2018) 390.

[7] Ajileye, G., James, A.A., Ayinde, A.M. and Oyedepo, T. *Collocation approach for the computational solution of Fredholm-Volterra fractional order of integro-differential equations*, J. Nig. Soc. Phys. Sci. 4 (2022), 834.

[8] Atabakan, Z.P., Nasab, A.K., Kiliçman, A. and Eshkuvatov, Z.K. *Numerical solution of nonlinear Fredholm integro-differential equations using spectral homotopy analysis method*, Math. Probl. Eng. 9 (7) (2013) 674364.

[9] Berinde, V. *Iterative approximation of fixed points*, Romania. Editura Efemeride, Baia Mare, 2002.

[10] Bhraway, A.H. Tohidi, E. and Soleymani, F. *A new Bernoulli matrix method for solving high order linear and nonlinear Fredholm integro-differential equations with piecewise interval*, Appl. Math. Comput. 219 (2012), 482–497.

[11] Biazar, J. and Gholami, P.M. *Application of variational iteration method for linear and nonlinear integro-differential-difference equations*, Int. Math. Forum5 (2010), 3335–3341.

[12] Darania, P. and Ebadian, A. *A method for the numerical solution of the integro-differential equations*, Appl. Math. Comput. 188 (2007), 657–668.

[13] Elmaci, D. and Baykus Savasaneril, N. *The Lucas Polynomial solution of linear Volterra-Fredholm integral equations*, Matrix Sci. Math. 6(1) (2022), 21–25.

[14] Elmaci, D. and Baykus Savasaneril, N. *Solutions of high-order linear Volterra integro-differential equations via Lucas polynomials*, Montes Taurus J. Pure Appl. Math. 5 (1) (2023), 22–33.

[15] Ercan, C. and Kharerah, T. *Solving a class of Volterra integral system by the differential transform method*, Int. J. Nonlinear Sci. 16 (2013), 87–91.

[16] Gulsu, M. and Ozturk, Y. *On the numerical solution of linear Fredholm-Volterra integro-differential Difference Equations with Piecewise Intervals*, Appl. Appl. Math. Comput. 7(3) (2012), 556–557.

[17] James A.A. and Ajileye, G., Ayinde A.M. and Dunama, W. *Hybrid-block method for the solution of second order non-linear differential equations*, J. Adv. Math. Comput. Sci. 37(12) (2022), 156–169.2456-9968.

[18] Karakoc, S.B.G., Eryilmaz, A. and Basbuk, M. *The approximate solutions of Fredholm integro-differential difference equations with variable coefficients via homotopy analysis method*, Math. Probl. Eng. (2013) Article ID: 261645.

[19] Khan, R.H. and Bakodah, H.O. *Adomian decomposition method and its modification for nonlinear Abel's integral equations*, Comput. Math. Appl. 7 (2013), 2349–2358.

[20] Matar, M.M. *Nonlocal integro-differential equations with arbitrary fractional order*, Konuralp J. Math. 4(1) (2016), 114–121.

[21] Mehdiyeva, G. Ibrahimov, V. and Imanova, M. *On the construction of the multistep methods to solving the initial-value problem for ODE and the Volterra integro-differential equations*, IAPE, Oxford, United Kingdom, 2019.

[22] Oyedepo, T., Ayinde, M.A., Adenipekun, A.E. and Ajileye, G. *Least-squares collocation Bernstein method for solving system of linear fractional integro-differential equations*, Int. J. Comput. Appl. 183(22) (2021), 0975–8887.

[23] Oyedepo, T., Ayoade, A.A. Ajileye, G. and Ikechukwu, N. J. *Legendre computational algorithm for linear integro-differential equations*, Cumhuriyet Science Journal 44(3) (2023), 561-566.

[24] Oyedepo, T., Ishola, C.Y., Ayoade, A.A. and Ajileye, G. *Collocation computational algorithm for Volterra-Fredholm integro-differential equations*, Electron. J. Math. Anal. Appl. 11(2) (2023), 1–9.

[25] Palais, R.S. *A simple proof of the Banach contraction principle*, J. Fixed Point Theory Appl. 2 (2007) 221–223.

[26] Rahmani, L., Rahimi, B. and Mordad, M. *Numerical solution of Volterra-Fredholm integro-differential equation by block pulse functions and operational matrices*, Gen. Math. Notes 4 (2) (2011), 7–48.

[27] Taiwo, O.A., Alimi, A.T. and Akanmu, M.A. *Numerical solutions for linear Fredholm integro-differential difference equations with variable coefficients by collocation methods*, JEPER 1 (2) (2014), 175–185.

[28] Volterra, V. *Theory of functionals and of integral and integro-differential equations,* Dover Publications, 2005.

[29] Yalcinbas, S. and Akkaya, T. *A numerical approach for solving linear integro-differential-difference equations with Boubaker polynomial bases*, Ain Shams Eng. J. 3(2) (2012), 153–161.

[30] Zada, L., Al-Hamami, M., Nawaz, R., Jehanzeb, S., Morsy, A., Abdel-Aty, A. and Nisar, K.S. *A new approach for solving Fredholm integro-differential equations*. Inform. Sci. Lett. 10(3) (2021), 407–415.

Iranian Journal of Numerical Analysis and Optimization
Vol. 14, No. 3, 2024, pp 875–899
https://doi.org/10.22067/ijnao.2024.86362.1375
https://ijnao.um.ac.ir/

Research Article

A pseudo−operational collocation method for optimal control problems of fractal−fractional nonlinear Ginzburg−Landau equation

T. Shojaeizadeh*, E. Golpar-Raboky* and Parisa Rahimkhani

Abstract

The presented work introduces a new class of nonlinear optimal control problems in two dimensions whose constraints are nonlinear Ginzburg−Landau equations with fractal−fractional (FF) derivatives. To acquire their approximate solutions, a computational strategy is expressed using the FF derivative in the Atangana−Riemann−Liouville (A-R-L) concept with the

*Corresponding author

Received 12 January 2024; revised 12 April 2024; accepted 13 May 2024

T. Shojaeizadeh
Department of Mathematics, Qom Branch, Islamic Azad University, Qom, Iran. e-mail: Ta.Shojaeizadeh@iau.ac.ir

E. Golpar-Raboky
Department of Mathematics, University of Qom, Qom, Iran. e-mail: g.raboky@qom.ac.ir

Parisa Rahimkhani
Faculty of Science, Mahallat Institute of Higher Education, Mahallat, Iran. e-mail: rahimkhani.parisa@mahallat.ac.ir

How to cite this article
Shojaeizadeh, T., Golpar-Rabok, E. and Rahimkhani, P., A pseudo−operational collocation method for optimal control problems of fractal−fractional nonlinear Ginzburg−Landau equation. *Iran. J. Numer. Anal. Optim.*, 2024; 14(3): 875-899. https://doi.org/10.22067/ijnao.2024.86362.1375

Mittage-Leffler kernel. The mentioned scheme utilizes the shifted Jacobi polynomials (SJPs) and their operational matrices of fractional and FF derivatives. A method based on the derivative operational matrices of SJR and collocation scheme is suggested and employed to reduce the problem into solving a system of algebraic equations. We approximate state and control functions of the variables derived from SJPs with unknown coefficients into the objective function, the dynamic system, and the initial and Dirichlet boundary conditions. The effectiveness and efficiency of the suggested approach are investigated through the different types of test problems.

AMS subject classifications (2020): Primary 35R30; Secondary 65M32, 35K20.

Keywords: Fractal−fractional (FF) derivative; Shifted Jacobi polynomials (SJPs); Operational matrices; Nonlinear Ginzburg−Landau equation; Optimal control problem.

1 Introduction

The Ginzburg−Landau equation is one of the most studied nonlinear partial differential equations in physics and engineering. This equation describes diverse types of phenomena, including superconductivity, Bose-Einstein, superfluidity, nonlinear waves, second-order phase transitions, condensation, liquid crystals, and strings in field theory [1]. There are many numerical and analytical schemes for solving this equation, for instance, see [10, 12, 18, 20, 25, 31].

Atangana [2] introduced the idea of FF derivation. The FF derivatives have been found very useful in many science and engineering applications. Since the fractals can be realized in nature as a fractal process or fractal media, it is interesting to derive the fractal or FF equations. The fractional partial differential equations appear in chaotic dynamics [32], long-range dissipation [22], and material science [26]. Fractional integrals and derivatives are a robust framework that can be applied to describe processes with various levels of complexity [11].

The fractional generalization of the Ginzburg–Landau equation was introduced in [30]. This equation can be used to describe the dynamical processes

in a medium with fractal dispersion and capture some long-range interactions of a system that can not be captured by traditional integer order differential equations. It is well has been evaluated from different aspects of this equation [29, 19, 28, 33]. Recently, Ding et al. studied higher-order numerical algorithm for the two-dimensional nonlinear spatial fractional complex Ginzburg−Landau equation [6].

Orthogonal polynomials have been extensively employed in solving optimal control problems involving fractional partial differential equations [4, 5, 9, 23, 27].

In [15] a numerical method for solving the model of the nonlinear Ginzburg−Landau equation in a FF sense is presented.

Regarding numerical methods for the FF equations, the critical step is the approximation of the fractional or FF derivatives.

Although, some approximate schemes for solving the FF model of nonlinear Ginzburg−Landau equation have been presented, for the first time we propose a scheme for solving the optimal control problem of FF nonlinear Ginzburg−Landau equation. The method uses SJPs for its numerical solution.

Using the FF derivative in the A-R-L concept and fractional derivatives in Caputo and Atangana-Baleanu-Caputo sense, optimal control of FF advection-diffusion-reaction equations is provided. These classes of problems are solved an operational matrix with high accuracy. Here , we consider the following optimal control problem:

$$(P)\ \min_{\mathbf{u}\in\mathbf{U}_{ad}}\ \mathcal{J}(\mathbf{y},\mathbf{u}) := \|\mathbf{y}(s,t)-\hat{\mathbf{y}}(s,t)\ \|^2_{L^2_{\omega(\epsilon,f)}(\Omega)} +\epsilon^2\|\mathbf{u}(s,t)-\hat{\mathbf{u}}(s,t)\ \|^2_{L^2_{\omega(\epsilon,f)}(\Omega)}, \tag{1}$$

with a nonlinear FF dynamic equation

$$\begin{aligned} &{}^{FFM}_{0}\mathcal{D}^{\alpha,\beta}\mathbf{y}(s,t) - (r_1+i\mu_1)\mathbf{y}_{ss}(s,t) + (r_2+i\mu_2)|\mathbf{y}(s,t)|^2\mathbf{y}(s,t) \\ &\quad - (r(s)+i\mu(s))\mathbf{y}(s,t) = f(s,t)+\mathbf{u}(s,t), \end{aligned} \tag{2}$$

on the domain $(s,t)\in\Omega$ with the initial condition

$$\mathbf{y}(s,0) = v(s), \tag{3}$$

and the boundary conditions

$$\mathbf{y}(0,t) = k(s), \quad \mathbf{y}(1,t) = g(s), \tag{4}$$

where, $\Omega := [0,1] \times [0,1]$. In the above relations, the state variables $\mathbf{y}(s,t)$ and the control variables $\mathbf{u}(s,t)$ are undetermined complex functions $\hat{\mathbf{y}}(s,t)$, $\hat{\mathbf{u}}(s,t)$, v, k and g are complex determined functions, r_1, r_2, μ_1 and, μ_2 are known constants and $r(s)$ and, $\mu(s)$ are real functions; in addition, ϵ in the transition process is the weight of the control action, and $\mathbf{U}_{ad} = \{\mathbf{u} \in L_2(\Omega) : u_1 \leq \mathbf{u} \leq u_2, u_1, u_2 \in \mathbb{R} \cup \pm\infty\}$ has determined the collection of admissible controls. Here, ${}^{FFM}_{0}\mathcal{D}^{\alpha,\beta}$ denotes the FF derivative operator of order $(\alpha,\beta) \in (0,1)$ in the A-R-L sense with Mittag–Leffler non-singular kernel [2, 3].

In the presented plan, we solve it by converting the main problem into a system of algebraic equations. For this aim, the functions y and u are approximated by SJPs with unknown coefficients. By substituting these approximations into the objective function, a nonlinear algebraic equation with unknown coefficients is derived. By substituting the mentioned approximations in the dynamic system and the initial and boundary conditions and utilizing the FF derivative operational matrix of SJPs, we derive a system of nonlinear algebraic equations. Finally, by using Lagrangian multipliers, we connect the algebraic equations obtained from the nonlinear FF dynamic equation and the initial and boundary conditions with the algebraic equation created by the objective function, and the optimal solution is achieved using the constrained extremum method.

2 Fractal-Fractional calculus

Here, we describe the definitions and basic features of FF calculus in the Atangana-Riemann-Liouville- Caputo sense.

Definition 1. [13]. The two-parameter Mittag-Lefler function is defined as follows:

$$E_{\zeta,\eta}(t) = \sum_{k=0}^{\infty} \frac{t^k}{\Gamma(k\zeta + \eta)}, \tag{5}$$

where $\zeta, \eta \in \mathbb{R}^+$, and $t \in \mathbb{R}$. Please remember that for $\eta = 1$ it is considered as $E_\zeta(t) = E_{\zeta,1}(t)$.

Definition 2. [2, 3]. The FF derivative of the continuous function $z(s,t)$ of order (α, β) in the A-R-L sense with Mittag-Leffler kernel is defined by

$$^{FFM}\quad {}_0\mathcal{D}_t^{\alpha,\beta} z(s,t) = \frac{c(\alpha)}{1-\alpha}\frac{\partial}{\partial t^\beta}\int_0^t z(s,\tau)E_\alpha(\frac{-\alpha(t-\tau)^\alpha}{1-\alpha})d\tau, \tag{6}$$

where $(\alpha, \beta) \in (0,1)$, $c(\alpha) = 1 - \alpha + \frac{\alpha}{\Gamma(\alpha)}$ and

$$\frac{\partial z(s,t)}{\partial t^\beta} = \lim_{\Delta t \to 0} \frac{z(s,t+\Delta t) - z(s,t)}{(t+\Delta t)^\beta - t^\beta}. \tag{7}$$

Remark 1. The aforementioned definition can be expressed as follows:

$$^{FFM}\quad {}_0\mathcal{D}_t^{\alpha,\beta} z(s,t) = \frac{c(\alpha)t^{1-\beta}}{\beta(1-\alpha)}\frac{\partial}{\partial t}\int_0^t z(s,\tau)E_\alpha(\frac{-\alpha(t-\tau)^\alpha}{1-\alpha})d\tau. \tag{8}$$

Corollary 1. [14]. Let $\alpha, \beta \in (0,1)$ and $r \in \mathbb{N}\bigcup\{0\}$. Then, we have

$$^{FFM}\quad {}_0\mathcal{D}_t^{\alpha,\beta} t^r = \frac{c(\alpha)r!t^{r-\beta+1}}{\beta(1-\alpha)}E_{\alpha,r+1}(\frac{-\alpha t^\alpha}{1-\alpha}). \tag{9}$$

3 The shifted Jacobi polynomials and their properties

The well-known SJPs on $[0,1]$ can be defined by the following explicit analytic formula: [4, 16]

$$b_i^{(e,g)}(t) = \sum_{k=0}^{i} \pi_k^{(i)} t^k, \tag{10}$$

where

$$\pi_k^{(i)} = (-1)^{i-k}\binom{i+e+f+k}{k}\binom{i+f}{i-k}, \tag{11}$$

$i \in \mathbb{N}\cup\{0\}$, $e, f > -1, e+f \neq -1$. Concerning the weight function $\omega^{e,f}(t) = t^f(1-t)^e$ on $[0,1]$ for SJPs, the orthogonality condition is Demonstrated by

$$\int_0^1 b_i^{(e,f)}(t)b_j^{(e,f)}(t)\omega^{(e,f)}(t)dt = \lambda_i\delta_{ij}, \tag{12}$$

that δ_{ij} is Kronecker's delta function and

$$\lambda_i = \frac{\Gamma(i+e+1)\Gamma(i+f+1)}{(2i+e+f+1)\Gamma(i+e+f+1)\Gamma(i+1)}.$$

Any assumed function $y \in L^2_{\omega^{(e,f)}}[0,1]$ in $(n+1)$ terms of SJPs can be written as follows

$$y(t) \simeq \sum_{i=0}^{n} y_i b_i^{(e,f)}(t) \triangleq Y^T \Phi_n(t), \tag{13}$$

where

$$\begin{aligned} Y &= [y_0,\ y_1,\ \ldots,\ y_n]^T, \\ \Phi_n(t) &\triangleq \left[b_0^{(e,f)}(t),\ b_1^{(e,f)}(t),\ \ldots,\ b_n^{(e,f)}(t)\right]^T, \end{aligned} \tag{14}$$

and

$$y_i = \frac{1}{\lambda_i}\int_0^1 y(t) b_i^{(e,f)}(t)\omega^{(e,f)}(t)dt, \quad i = 0,1,\cdots,n.$$

In the same way, a bivariate function $y(s,t) \in L^2_{\omega^{(e,f)}}(\Omega)$ can be expanded by the SJPs as

$$y(s,t) \simeq \sum_{i=0}^{m}\sum_{j=0}^{n} y_{ij} b_i^{(e,f)}(s) b_j^{(e,f)}(t) \triangleq \Phi_m^T(s) Y \Phi_n(t), \tag{15}$$

where the entries of the unknown matrix $Y = [y_{ij}]$ (coefficients matrix of $(m+1)\times(n+1)$ dimensional) are obtained from the following equation

$$y_{ij} = \frac{1}{\lambda_i \lambda_j}\int_0^1\int_0^1 y(s,t) b_i^{(e,f)}(s) b_j^{(e,f)}(t)\omega^{(e,f)}(s)\omega^{(e,f)}(t)dxdt, \tag{16}$$

for $i = 0,1,\cdots,m, \quad j = 0,1,\cdots,n$. The first-order derivative of the vector $\Phi_n(t)$ can be expressed by [8, 7, 21]

$$\frac{d\Phi_n(t)}{dt} = \mathbf{D}^{(1)}\Phi_n(t) \tag{17}$$

where, $\mathbf{D}^{(1)}$ is the $(n+1)\times(n+1)$ operational matrix of derivative given by

$$\mathbf{D}^{(1)} = (d_{ij}) = \begin{cases} C_1(i,j), & i > j, \\ \\ 0, & \text{otherwise}, \end{cases} , \tag{18}$$

where

$$C_1(i,j) = \frac{(i+e+f+1)(i+e+f+2)_j(j+e+2)_{i-j-1}\Gamma(j+e+f+1)}{(i-j-1)!\Gamma(2j+e+f+1)}$$

$$\times_3 F_2 \left(\begin{matrix} -i+1+j,\ i+j+e+f+2,\ j+e+1 \\ \\ j+e+2, \quad 2j+e+f+2, \end{matrix} ; 1 \right)$$

For example, for even n we have

$$\mathbf{D}^{(1)} = \begin{pmatrix} 0 & 0 & 0 & \dots & 0 & 0 \\ C_1(1,0) & 0 & 0 & \dots & 0 & 0 \\ C_1(2,0) & C_1(2,1) & 0 & \dots & 0 & 0 \\ C_1(3,0) & C_1(3,1) & C_1(3,2) & \dots & 0 & 0 \\ \vdots & \vdots & \vdots & \dots & \vdots & \vdots \\ C_1(n,0) & C_1(n,1) & C_1(n,2) & \dots & C_1(n,n-1) & 0 \end{pmatrix}.$$

Remark 2. [8]. Recall that the shifted factorial $(a)_n$ is defined by

$$(a)_n = \frac{\Gamma(a+n)}{\Gamma(a)},$$

and the hypergeometric function $\times_3 F_2$ is defined by

$$\times_3 F_2 \left(\begin{matrix} -n,\ a, & b \\ \\ c, \quad 1+a+b-c-n, & \end{matrix} ; 1 \right) = \frac{(c-a)_n(c-n)_n}{(c)_n(c-a-b)_n}.$$

Remark 3. Generally, the r-derivative operational matrix of SJPs of $\Phi_n(t)$ can be given by:

$$\frac{d^r \Phi_n(t)}{dt^r} = \mathbf{D}^{(r)} \Phi_n(t), \tag{19}$$

where $r \in \mathbb{N}$ and $\mathbf{D}^{(r)}$ denotes the $r-$th power of $\mathbf{D}^{(1)}$.

Theorem 1. [24]. Assume that $\alpha, \beta \in (0,1)$. The FF derivative of order (α, β) in the A-R-L sense of $\Phi_n(t)$ in (14) is achieved as:

$$^{FFM}\quad {}_0\mathcal{D}_t^{\alpha,\beta}\Phi_n(t) \simeq \mathbf{E}^{(\alpha,\beta)}\Phi_n(t), \tag{20}$$

that $\mathbf{E}^{(\alpha,\beta)} = [\epsilon_{ij}^{(\alpha,\beta)}]$ is called FF derivative operational matrix of the SJPs, and its entries for $1 \leq i, j \leq n+1$ are yielded as follows

$$\epsilon_{ij}^{(\alpha,\beta)} = \frac{c(\alpha)}{\lambda_{j-1}\beta(1-\alpha)} \sum_{m=0}^{i-1}\sum_{r=0}^{j-1}\sum_{l=0}^{\infty} (\frac{\alpha}{1-\alpha})^l \\ \frac{\pi_m^{(i-1)}\pi_r^{(j-1)} m!\Gamma(e+1)\Gamma(\alpha l-\beta+f+m+r+2)}{\Gamma(\alpha l+m+1)\Gamma(\alpha l-\beta+e+f+m+r+3)}, \tag{21}$$

in which $\pi_m^{(i-1)}$ and $\pi_r^{(j-1)}$ are presented in (11).

4 Convergence analysis

Here in two dimensions, the convergence analysis of SJPs expansion is explored. Set

$$\mathcal{W}^{(e,f)}(s,t) = \omega^{(e,f)}(s)\omega^{(e,f)}(t), \tag{22}$$

where $\omega^{e,f}(z) = z^f(1-z)^e,\ z \in [0,1]$.

Theorem 2. Suppose $\tau \in C^{m+n+1}(\Omega)$ and $|\ \frac{\partial^{n+m+1}}{\partial x^{n+m+1-i}\partial t^i}\Im(s,t)\ | \leqslant \Delta$, for $0 \leq i \leq n+m+1$. Let $\mathcal{Y} = span\{b_i^{(e,f)}(s)\ b_j^{(e,f)}(t),\ 0 \leq i \leq m,\ 0 \leq j \leq n\}$, be a vector subspace with finite dimension of $L^2(\Omega)$. If $\Im_{mn}(s,t)$ is a unique best approximation of τ out of $\mathcal{Y}$ obtained from the proposed method, then the error upper bound satisfies the following relation:

$$\|\Im(s,t) - \Im_{mn}(s,t)\|_{L^2_{\mathcal{W}^{(e,f)}}} = \int_0^1\int_0^1 (\Im(s,t) - \Im_{mn}(s,t))^2 \mathcal{W}^{(e,f)}(s,t)dsdt \\ \leqslant \frac{\Delta 2^{2(m+n+1)}}{(n+m+1)!}. \tag{23}$$

Proof. From Maclaurin's expansion for $\Im(s,t)$, we have

$$\Im(s,t) = \wp(s,t) + \frac{1}{(n+m+1)!}(s\frac{\partial}{\partial s} + t\frac{\partial}{\partial t})^{n+m+1}\Im(\xi_0 s, \xi_0 t), \quad \xi_0 \in (0,1),$$

where

$$\wp(s,t)=\sum_{k=0}^{n+m}\frac{1}{k!}(s\frac{\partial}{\partial s}+t\frac{\partial}{\partial t})^{k}\ \Im(0,0).$$

Thusly

$$|\Im(s,t)-\wp(s,t)|=|\frac{1}{(n+m+1)!}(s\frac{\partial}{\partial s}+t\frac{\partial}{\partial t})^{n+m+1}\Im(\xi_0 s,\xi_0 t)|,\quad \xi_0\in(0,1).$$

Since $\Im_{mn}(s,t)$ is the best approximation of $\Im(s,t)$, we acquire

$$\|\ \Im-\Im_{mn}\|^2_{L^2_{\mathcal{W}^{(e,f)}}}\leq\|\ \tau-p\|^2_{L^2_{\mathcal{W}^{(e,f)}}}.$$

By the definition of the L^2-norm and binomial expansion $(s\frac{\partial}{\partial s}+t\frac{\partial}{\partial t})^{n+m+1}$, we will have

$$\begin{aligned}
&\|\ \Im(s,t)-\wp(s,t)\ \|^2_{L^2_{\mathcal{W}^{(e,f)}}}\\
&=\int_0^1\int_0^1[\frac{1}{(n+m+1)!}(s\frac{\partial}{\partial s}+t\frac{\partial}{\partial t})^{n+m+1}\Im(\xi_0 s,\xi_0 t)]^2\mathcal{W}^{(e,f)}(s,t)dsdt\\
&=\int_0^1\int_0^1[\frac{1}{(n+m+1)!}\sum_{i=0}^{n+m+1}\binom{n+m+1}{i}s^{n+m+1-i}t^i\frac{\partial^{n+m+1}}{\partial s^{n+m+1-i}\partial t^i}\\
&\quad\Im(\xi_0 s,\xi_0 t)]^2\mathcal{W}^{(e,f)}(s,t)dsdt\\
&\leq\frac{\Delta^2}{(n+m+1)!^2}\int_0^1\int_0^1\sum_{i=0}^{n+m+1}\sum_{j=0}^{n+m+1}\binom{n+m+1}{i}\\
&\quad\binom{n+m+1}{j}s^{n+m+1-i}t^i s^{n+m+1-j}t^j\mathcal{W}^{(e,f)}(s,t)dsdt.
\end{aligned}$$

Since, $0\leq s,t\leq 1$, we have

$$\begin{aligned}
\|\ \Im-\wp\|^2_{L^2_{\mathcal{W}^{(e,f)}}}&\leq\tfrac{\Delta^2}{(n+m+1)!^2}\int_0^1\int_0^1\sum_{i=0}^{n+m+1}\sum_{j=0}^{n+m+1}\binom{n+m+1}{i}\\
&\quad\binom{n+m+1}{j}\mathcal{W}^{(e,f)}(s,t)dsdt\\
&=\tfrac{\Delta^2 2^{2(m+n+1)}}{(n+m+1)!^2}
\end{aligned}$$

□

which is the desired result.

Corollary 2. If $\Im(s,t)$ is an infinity differential function on Ω, then

$$\|\Im(s,t)-\Im_{mn}(s,t)\|_{L^2_{\mathcal{W}^{(e,f)}}} \to 0 \ as \ n,m \to \infty$$

5 Expression of the proposed approach

In the present section, we will solve the introduced problem in Eqs. (1)-(4) numerically. For this purpose, we first decompose the complex state and control variables and functions of the problem in their real and imaginary parts as follows

$$\begin{aligned} y(s,t) &= y_1(s,t)+iy_2(s,t), & u(s,t) &= u_1(s,t)+iu_2(s,t),\\ \hat{y}(s,t) &= \hat{y}_1(s,t)+i\hat{y}_2(s,t), & v(s) &= v_1(s)+iv_2(s),\\ g(t) &= g_1(t)+ig_2(t), & k(t) &= k_1(t)+ik_2(t), \end{aligned} \tag{24}$$

where, $y_j(s,t), \hat{y}_j(s,t), u_j(s,t), v_j(s), g_j(t)$ and $k_j(t)$ are real functions for $j=1,2$. Thus, the called problem can be illustrated in a coupled system of nonlinear FF differential equations as

$$\begin{aligned} FFM \quad & {}_0\mathcal{D}^{\alpha,\beta}y_1(s,t) - r_1 y_{1ss}(s,t) + \mu_1 y_{2ss}(s,t)\\ & +r_2({y_1}^2(s,t)+{y_2}^2(s,t))y_1(s,t) - \mu_2({y_1}^2(s,t)+{y_2}^2(s,t))y_2(s,t)\\ & -r(s)y_1(s,t)+\mu(s)y_2(s,t) = u_1(s,t)+f_1(s,t),\\ FFM \quad & {}_0\mathcal{D}^{\alpha,\beta}y_2(s,t) - r_1 y_{2ss}(s,t) - \mu_1 y_{1ss}(s,t)\\ & +r_2({y_1}^2(s,t)+{y_2}^2(s,t))y_2(s,t) + \mu_2({y_1}^2(s,t)+{y_2}^2(s,t))y_1(s,t)\\ & -r(s)y_1(s,t)-\mu(s)y_1(s,t) = u_2(s,t)+f_2(s,t), \end{aligned} \tag{25}$$

for $i=1,2$ with the initial conditions

$$y_i(s,0) = v_i(s), \tag{26}$$

and the boundary conditions

$$\begin{aligned} y_i(0,t) &= g_i(t),\\ y_i(1,t) &= k_i(t). \end{aligned} \tag{27}$$

Now, the state and control variables are approximated in terms of the $SJPs$ as follows for $k = 1, 2$

$$y_k(s,t) \simeq \Phi_m^T(s)\mathbf{Y}_k\Phi_n(t), \tag{28}$$

$$u_k(s,t) \simeq \Phi_m^T(s)\mathbf{U}_k\Phi_n(t), \tag{29}$$

where the vectors $\Phi_m(s)$ and $\Phi_n(t)$ are introduced in Eq. (14), and $\mathbf{Y}_k = ({y^k}_{ij})$, and $\mathbf{U}_k = (u_{ij}^k)$ are the unknown coefficients matrices of $(m+1) \times (n+1)$ dimensional. Set

$$\begin{aligned}\mathcal{B}(s,t) \triangleq \Big[&b_0^{(e,f)}(s)b_0^{(e,f)}(t), \ldots, b_m^{(e,f)}(s)b_0^{(e,f)}(t) \mid \ldots \\ &\mid b_0^{(e,f)}(s)b_n^{(e,f)}(t), \ldots, b_m^{(e,f)}(s)b_n^{(e,f)}(t)\Big]^T.\end{aligned} \tag{30}$$

Considering Eq. (15), we can write Eqs. (28) and (29) as follows:

$$y_k(s,t) \simeq \Phi_m^T(s)\mathbf{Y_k}\Phi_n(t) = \mathcal{B}^T(s,t)vec(\mathbf{Y_k}), \tag{31}$$

and

$$u_k(s,t) \simeq \Phi_m^T(s)\mathbf{U_k}\Phi_n(t) = \mathcal{B}^T(s,t)vec(\mathbf{U_k}), \tag{32}$$

where $k = 1, 2$, and

$$\begin{aligned}vec(\mathbf{Y_k}) &= \left[y_{00}^k, \ldots, y_{m0}^k \mid \ldots \mid y_{0n}^k, \ldots, y_{mn}^k\right]^T, \\ vec(\mathbf{U_k}) &= \left[u_{00}^k, \ldots, u_{m0}^k \mid \ldots \mid u_{0n}^k, \ldots, u_{mn}^k\right]^T.\end{aligned}$$

The following results are obtained from Eqs. (28) and (20):

$$^{FFM}\quad {}_0\mathcal{D}_t^{\alpha,\beta}y_k(s,t) \simeq \Phi_m^T(s)\mathbf{Y}_k\mathbf{E}^{(\alpha,\beta)}\Phi_n(t). \tag{33}$$

Also, from Remark 3 and two times derivative with respect to s on both sides of Eq. (28) yields

$$y_{k\,ss}(s,t) \simeq \Phi_m^T(s)(\mathbf{D}^{(2)})^T\mathbf{Y}_k\Phi_n(t). \tag{34}$$

Regarding relations Eqs. (26)–(27) and Eq. (31), for $k = 1, 2$

$$\begin{aligned}
&\mathcal{B}(s,0)vec(\mathbf{Y_k}) = v_k(s),\\
&\mathcal{B}(0,t)vec(\mathbf{Y_k}) = k_k(t),\\
&\mathcal{B}(1,t)vec(\mathbf{Y_k}) = g_k(t).
\end{aligned} \tag{35}$$

Inserting Eqs. (31)–(34) into Eq. (25) gives

$$\begin{aligned}
\mathcal{Z}_1(s,t,) \triangleq \mathcal{B}(s,t)\Big[\Big((\mathbf{E}^{(\alpha,\beta)T}\otimes I_{m+1}) - r_1(I_{n+1}\otimes(\mathbf{D}^{(2)})^T) - r(s) + r_2((\mathcal{B}(s,t)vec(\mathbf{Y_1}))^2\\
+ (\mathcal{B}(s,t)vec(\mathbf{Y_2}))^2)\Big)vec(\mathbf{Y_1}) + \Big(\mu_1(I_{n+1}\otimes(\mathbf{D}^{(2)})^T) + \mu(s)\\
- \mu_2((\mathcal{B}(s,t)vec(\mathbf{Y_1}))^2 + (\mathcal{B}(s,t)vec(\mathbf{Y_2}))^2)\Big)vec(\mathbf{Y_2}) - vec(\mathbf{U_1})\Big] - f_1(s,t) \simeq 0,\\
\mathcal{Z}_2(s,t,) \triangleq \mathcal{B}(s,t)\Big[\Big((\mathbf{E}^{(\alpha,\beta)T}\otimes I_{m+1}) - r_1(I_{n+1}\otimes(\mathbf{D}^{(2)})^T) - r(s) + r_2((\mathcal{B}(s,t)vec(\mathbf{Y_1}))^2\\
+ (\mathcal{B}(s,t)vec(\mathbf{Y_2}))^2)\Big)vec(\mathbf{Y_2}) - \Big(\mu_1(I_{n+1}\otimes(\mathbf{D}^{(2)})^T) - \mu(s)\\
- \mu_2((\mathcal{B}(s,t)vec(\mathbf{Y_1}))^2 + (\mathcal{B}(s,t)vec(\mathbf{Y_2}))^2)\Big)vec(\mathbf{Y_1}) - vec(\mathbf{U_2})\Big] - f_2(s,t) \simeq 0,
\end{aligned} \tag{36}$$

where, I_{m+1} and I_{n+1} are identity matrices of $m+1$ and $n+1$ orders, respectively and $\otimes$ represents Kronecker's product [17]. Finally, from Eqs. (35) and (36), a system of $2(m+1)(n+1)$ algebraic equations can be written as for $k = 1, 2$:

$$\begin{cases}
\mathcal{C}^k_{i,j} \triangleq \mathcal{Z}_k(\xi_i,\eta_j) = 0, & i = 2,\ldots,m, \quad j = 2,\ldots,n+1,\\
\tilde{\mathcal{H}}_k \triangleq \mathcal{B}(\xi_i,0)vec(\mathbf{Y_k}) - v_k(\xi_i) = 0 & i = 1,\ldots,m+1,\\
\tilde{\mathcal{M}}_k \triangleq \mathcal{B}(0,\eta_j)vec(\mathbf{Y_k}) - k_k(\eta_j) = 0, & j = 2,\ldots,n+1,\\
\tilde{\mathcal{N}}_k \triangleq \mathcal{B}(1,\eta_j)vec(\mathbf{Y_k}) - g_k(\eta_j) = 0, & j = 2,\ldots,n+1,
\end{cases} \tag{37}$$

where

$$\begin{aligned}
\xi_i &= \frac{1}{2}(1 - cos(\frac{2(i-1)\pi}{2m})), && i = 1,\ldots,m+1,\\
\eta_j &= \frac{1}{2}(1 - cos(\frac{2(j-1)\pi}{2n})), && j = 1,\ldots,n+1.
\end{aligned}$$

Then, the performance index of the examined problem is approximated using SJPs. First, the desired function is approximated as:

$$\hat{y}_k(s,t) \simeq \Phi_m^T(s)\widehat{\mathbf{Y}}_{\mathbf{k}}\Phi_n(t), \quad k=1,2. \tag{38}$$

Inserting Eqs. (28), (29), and (38) into Eq. (1), we get

$$\begin{aligned}
\mathcal{J}(y,u) &\simeq \mathbb{J}_{m,n}(\mathbf{Y_1},\mathbf{Y_2},\mathbf{U_1},\mathbf{U_2})\\
&= \int_0^1\int_0^1 \Big[(\phi_m(s)^T\mathbf{Y_1}\phi_n(t)-\phi_m(s)^T\widehat{\mathbf{Y}}_1\phi_n(t))(\phi_m(s)^T\mathbf{Y}_1\phi_n(t)-\phi_m(s)^T\widehat{\mathbf{Y}}_1\phi_n(t))^T\Big]\\
&\quad \mathcal{W}^{(e,f)}(s,t)dsdt\\
&\quad + \int_0^1\int_0^1 \Big[(\phi_m(s)^T\mathbf{Y_2}\phi_n(t)-\phi_m(s)^T\widehat{\mathbf{Y}}_2\phi_n(t))(\phi_m(s)^T\mathbf{Y}_2\phi_n(t)-\phi_m(s)^T\widehat{\mathbf{Y}}_2\phi_n(t))^T\Big]\\
&\quad \mathcal{W}^{(e,f)}(s,t)dsdt\\
&\quad + \epsilon^2\Big(\int_0^1\int_0^1 \Big[(\phi_m(s)^T\mathbf{U_1}\phi_n(t)-\phi_m(s)^T\widehat{\mathbf{U}}_1\phi_n(t))(\phi_m(s)^T\mathbf{U}_1\phi_n(t)-\phi_m(s)^T\widehat{\mathbf{U}}_1\phi_n(t))^T\Big]\\
&\quad \mathcal{W}^{(e,f)}(s,t)dsdt\\
&\quad + \int_0^1\int_0^1 \Big[(\phi_m(s)^T\mathbf{U_2}\phi_n(t)-\phi_m(s)^T\widehat{\mathbf{U}}_2\phi_n(t))(\phi_m(s)^T\mathbf{U}_2\phi_n(t)-\phi_m(s)^T\widehat{\mathbf{U}}_2\phi_n(t))^T\Big]\\
&\quad \mathcal{W}^{(e,f)}(s,t)dsdt\Big).
\end{aligned}$$

Because expressions $\int_0^1\int_0^1(\phi_m(s)^T\widehat{\mathbf{Y}}_1\phi_n(t))^2\mathcal{W}^{(e,f)}(s,t)dsdt$, $\int_0^1\int_0^1(\phi_m(s)^T\widehat{\mathbf{Y}}_2\phi_n(t))^2\mathcal{W}^{(e,f)}(s,t)dsdt$, $\int_0^1\int_0^1(\phi_m(s)^T\widehat{\mathbf{U}}_1\phi_n(t))^2\mathcal{W}^{(e,f)}(s,t)dsdt$, and $\int_0^1\int_0^1(\phi_m(s)^T\widehat{\mathbf{U}}_2\phi_n(t))^2\mathcal{W}^{(e,f)}(s,t)dsdt$ do not have any effective role in minimization due to being positive, so according to Eq. (12), the above equation is expressed as follows

$$\mathbb{J}_{m,n}((\mathbf{Y_1},\mathbf{Y_2},\mathbf{U_1},\mathbf{U_2})) = vec(\mathbf{Y_1})^T(\Upsilon_n\otimes\Upsilon_m)vec(\mathbf{Y_1}) + vec(\mathbf{Y_2})^T(\Upsilon_n\otimes\Upsilon_m)vec(\mathbf{Y_2}) \tag{39}$$

$$- 2vec(\widehat{\mathbf{Y}}_1)^T(\Upsilon_n\otimes\Upsilon_m)vec(\mathbf{Y_1}) - 2vec(\widehat{\mathbf{Y}}_2)^T(\Upsilon_n\otimes\Upsilon_m)vec(\mathbf{Y_2}) \tag{40}$$

$$+ \epsilon^2[vec(\mathbf{U_1})^T(\Upsilon_n\otimes\Upsilon_m)vec(\mathbf{U_1}) + vec(\mathbf{U_2})^T(\Upsilon_n\otimes\Upsilon_m)vec(\mathbf{U_2}) \tag{41}$$

$$- 2vec(\widehat{\mathbf{U}}_1)^T(\Upsilon_n\otimes\Upsilon_m)vec(\mathbf{U_1}) - 2vec(\widehat{\mathbf{U}}_2)^T(\Upsilon_n\otimes\Upsilon_m)vec(\mathbf{U_2})], \tag{42}$$

where

$$\Upsilon_k = diag(\lambda_0,\ldots,\lambda_k).$$

The discussed optimal control problem has now transformed into a finite-dimensional optimization. To solve the obtained optimization problem, we employ the Lagrangian multipliers scheme. First define

$$\mathcal{J}^*(y,u) \simeq \mathbb{J}^*(\mathbf{Y_1},\mathbf{Y_2},\mathbf{U_1},\mathbf{U_2},\Sigma) = \mathbb{J}_{m,n}(\mathbf{Y_1},\mathbf{Y_2},\mathbf{U_1},\mathbf{U_2}) + \Sigma^T\mathcal{C}, \tag{43}$$

where

$$\mathcal{C} = \big[\mathcal{C}^1_{2,2}, \ldots, \mathcal{C}^1_{2,n+1} | \ldots | \mathcal{C}^1_{m,2}, \ldots, \mathcal{C}^1_{m,n+1} | \mathcal{C}^2_{2,2}, \ldots, \mathcal{C}^2_{2,n+1} | \ldots | \mathcal{C}^2_{m,2}, \ldots, \mathcal{C}^2_{m,n+1} | \tilde{\mathcal{H}}_1 | \tilde{\mathcal{H}}_2 | \tilde{\mathcal{M}}_1 | \tilde{\mathcal{M}}_2 | \tilde{\mathcal{N}}_1 | \tilde{\mathcal{N}}_2 \big],$$

and

$$\Sigma = \Big[\varsigma_1 \ \varsigma_2 \ \cdots \ \varsigma_{2(m+1)(n+1)} \Big]^T,$$

where the Lagrange multipliers vector is denoted with Σ. Now the optimality conditions for $k = 1, 2$ are derived from the following equations:

$$\begin{cases} \dfrac{\partial \mathcal{J}^*(y,u)}{\partial vec(\mathbf{Y_k})} = 0, \\ \dfrac{\partial \mathcal{J}^*(y,u)}{\partial vec(\mathbf{U_k})} = 0, \\ \dfrac{\partial \mathcal{J}^*(y,u)}{\partial \Sigma} = 0. \end{cases}$$

The above system of algebraic equations can be solved by Newton's iterative method or by using Matlab software packages. The numerical solutions of $y(s,t)$ and $u(s,t)$ are specified by determining $\mathbf{Y}_k$ and $\mathbf{U}_k$ and placing them in Eqs. (31) and (32), respectively.

6 Illustrative examples

In this section, using some test problems, the accuracy and efficiency of the described method in Section 5 have been investigated. To achieve this goal, the maximum absolute error (MAE) and root mean square (RMS) are calculated. All calculations and results have been done using the Fmincon package in MATLAB software. The accuracy of the obtained results from the proposed method is calculated using MAE and RMS, which is defined as follows:

$$MAE = \max_{1 \le i \le m+1} \max_{1 \le j \le n+1} |f(\xi_i, \eta_j) - \widetilde{f}(\xi_i, \eta_j)|,$$

$$RMS = \sqrt{\frac{1}{(m+1)(n+1)}(\sum_{i=1}^{m+1}\sum_{j=1}^{n+1}|f(\xi_i,\eta_j) - \widetilde{f}(\xi_i,\eta_j)|^2)},$$

where $\widetilde{f}$ represents the numerical solution of f in the collocation points of (ξ_i, η_j). Note that in all test problems, the first 29 terms of the infinite series in the Mittag-Leffler functions are used in numerical calculations.

Example 1. Consider the following objective function:

$$\mathcal{J}(y,u) = \int_0^1\int_0^1 \Big((y(s,t) - t^4e^{is})^2 - \epsilon^2(u(s,t) - ((1+2i)t^4 + (1+i)t^{12} - s^2(1+is)t^4)e^{is})^2\mathcal{W}^{(e,f)}(s,t)\Big)dsdt, \tag{44}$$

where $\epsilon = 1$. Subject to the following nonlinear time FF Ginzburg–Landau equation [14]:

$${}_0^{FFM}D^{\alpha,\beta}y(s,t) - (1+2i)y_{ss}(s,t) + (1+i)|y(s,t)|^2y(s,t) - s^2(1+is)y(s,t) = f(s,t) + u(s,t),$$

where

$$f(s,t) = (\tfrac{c(\alpha)4!t^{5-\beta}}{\beta(1-\alpha)}E_{\alpha,5}(\tfrac{-\alpha t^\alpha}{1-\alpha}))e^{is}\,.$$

The exact solution of state and control functions is mentioned below,

$$y(s,t) = t^4e^{is},$$
$$u(s,t) = ((1+2i)t^4 + (1+i)t^{12} - s^2(1+is)t^4)e^{is}.$$

From the analytical solution of $y(s,t)$, we acquire the initial and boundary conditions. For the computational solution of this example, we have used the introduced method in Section 5 with values of $m = n$. For some choices (α,β) and $(e,f) = (0,0)$, MAE and RMS values of state and control variables are shown in Figures $1-4$. AE graphs for state and control variables with $m = n = 7,\ \alpha = 0.25, \beta = 0.85$ and $(e,f) = (0,0)$ are depicted in Figure 5.

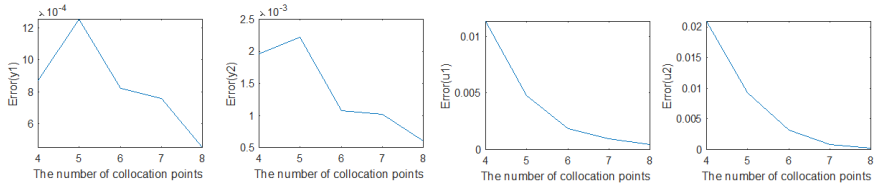


Figure 1: The RMS of the presented method for the state and control functions with $\alpha = 0.35$ and $\beta = 0.35$ in Example 1.

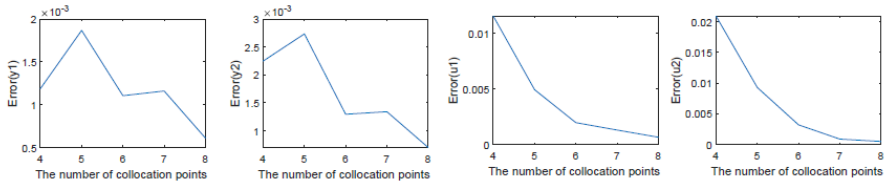


Figure 2: The RMS of the presented method for the state and control functions with $\alpha = 0.55$ and $\beta = 0.35$ in Example 1.

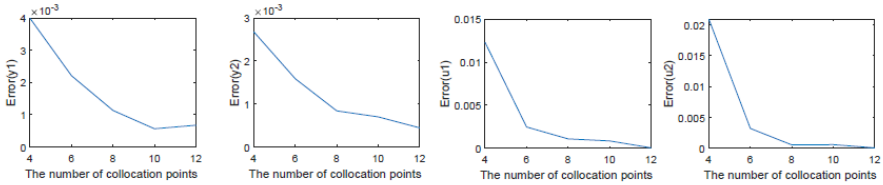


Figure 3: The RMS of the presented method for the state and control functions with $\alpha = 0.75$ and $\beta = 0.35$ in Example 1.

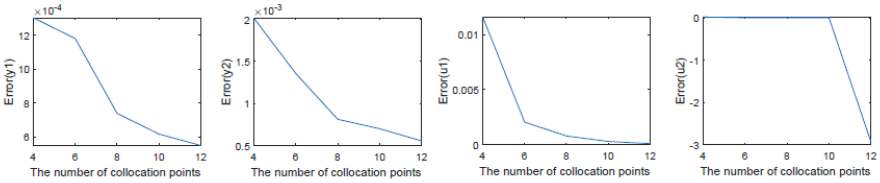


Figure 4: The RMS of the presented method for the state and control functions with $\alpha = 0.75$ and $\beta = 0.65$ in Example 1.

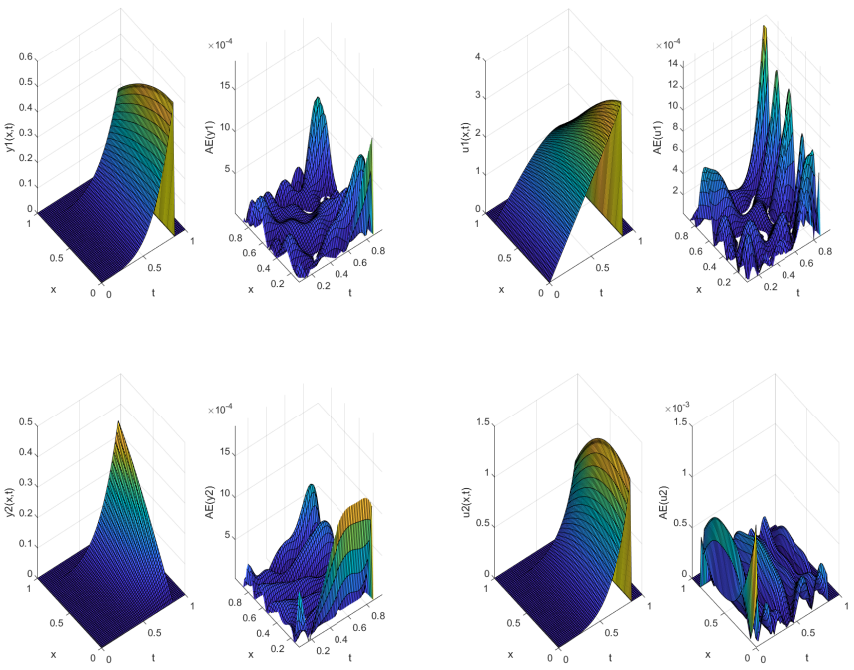


Figure 5: Numerical solution (SJPs) and error function (AE) surfaces of $y(s,t)$ (left) and $u(s,t)$ (right) with (m=n=7) in Example 1.

Example 2. Consider the problem of optimal control with $\epsilon = 10^{-1}$ as follows:

$$\mathcal{J}(y,u) = \int_0^1 \int_0^1 \Big((y(s,t) - t^2 sin(t) e^{is})^2 - \epsilon^2 (u(s,t) - (5it^2 sin(t) + 2t^6 sin^3(t) - it^2 sin(t) e^{-is}) e^{is})^2 \mathcal{W}^{(e,f)}(s,t) \Big) ds dt, \tag{45}$$

with the nonlinear time FF dynamical system: [14]

$${}_0^{FFM} D^{\alpha,\beta} y(s,t) - 5i y_{ss}(s,t) + 2|y(s,t)|^2 y(s,t) - ie^{-is} y(s,t) = f(s,t) + u(s,t),$$

where,

$$f(s,t) = \left(\frac{c(\alpha) t^{4-\beta}}{\beta(1-\alpha)} \sum_{k=0}^{\infty} (-1)^k (2k+3)(2k+2) t^{2k} E_{\alpha,2k+4}\left(\frac{-\alpha t^{\alpha}}{1-\alpha}\right)\right) e^{is}.$$

The exact solution of state and control functions is mentioned following,

$$y(s,t) = t^2 sin(t) e^{is},$$
$$u(s,t) = (5it^2 sin(t) + 2t^6 sin^3(t) - it^2 sin(t) e^{-is}) e^{is}.$$

From the analytical solution of $y(s,t)$, we acquire the initial and boundary conditions. For the computational solution of this example, we have used the introduced method in section 5 with values of $m = n$. For some choices (α, β) and $(e, f) = (0.5, 0.5)$, MAE and RMS values of state and control functions are shown in Figures $6 - 9$. AE graphs for state and control functions with $m = n = 7$, $\alpha = 0.75, \beta = 0.25$ and $(e, f) = (0.5, 0.5)$ are depicted in Figure 10.

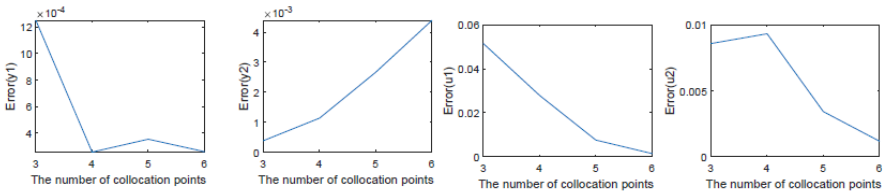


Figure 6: The RMS of the presented method for the state and control functions with $\alpha = 0.35$ and $\beta = 0.35$ in Example 2.

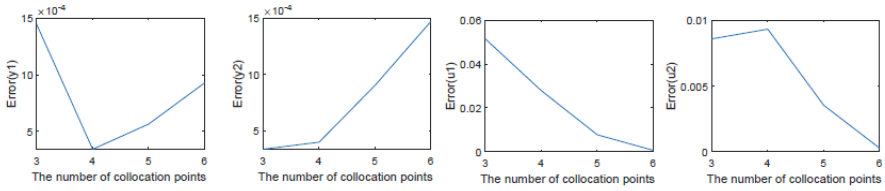


Figure 7: The RMS of the presented method for the state and control functions with $\alpha = 0.65$ and $\beta = 0.35$ in Example 2.

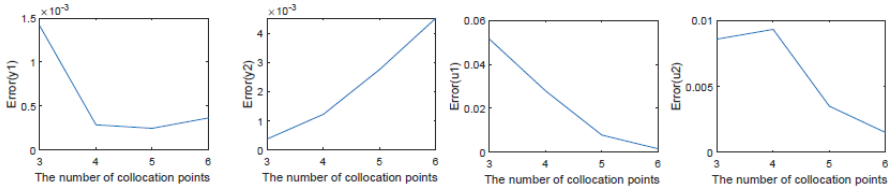


Figure 8: The RMS of the presented method for the state and control functions with $\alpha = 0.75$ and $\beta = 0.15$ in Example 2.

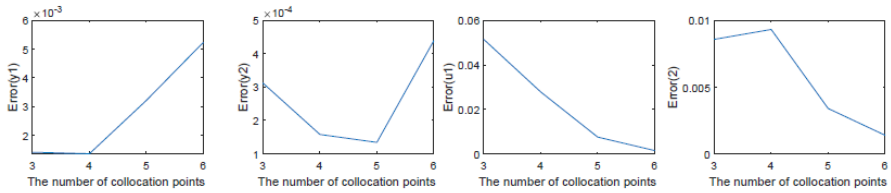


Figure 9: The RMS of the presented method for the state and control functions with $\alpha = 0.75$ and $\beta = 0.45$ in Example 2.

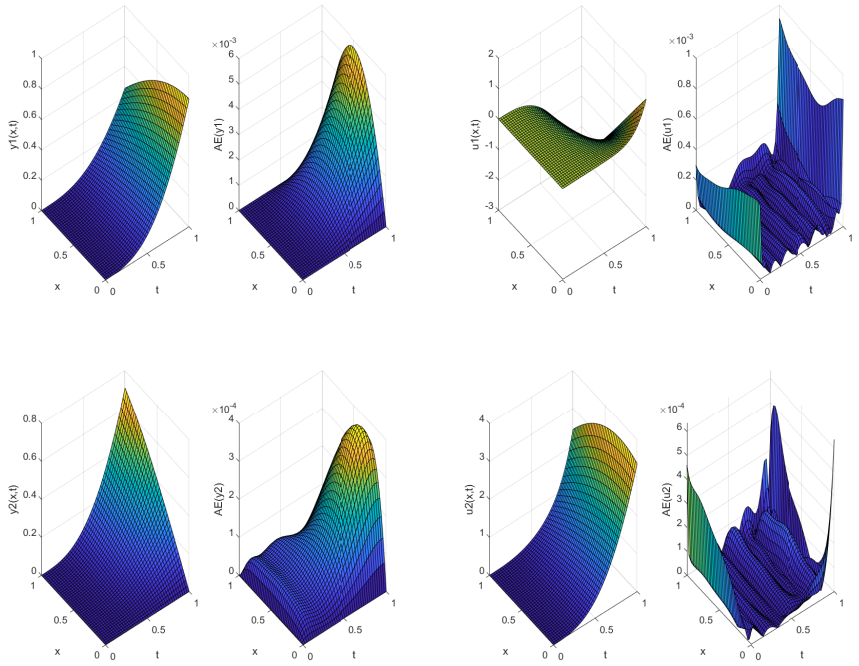


Figure 10: Numerical solution (SJPs) and error function (AE) surfaces of $y(s,t)$ (left) and $u(s,t)$ (right) with (m=n=7) in Example 2.

Example 3. Consider the problem of optimal control with $\epsilon = 1.1$ as follows:

$$\mathcal{J}(y,u) = \int_0^1 \int_0^1 \Big((y(s,t) - it^3 e^{-(t+is)})^2 - \epsilon^2 (u(s,t) - ie^{-is}(2t^3 e^{-t} + 3it^9 e^{-3t} - (2s+1+3is^2)t^3 e^{-t}))^2 \mathcal{W}^{(e,f)}(s,t) \Big) ds dt, \tag{46}$$

with the FF dynamical system: [14]

$${}_0^{FFM}D^{\alpha,\beta}y(s,t) - 2y_{ss}(s,t) + 3i|y(s,t)|^2y(s,t) - (2s+1+3is^2)y(s,t) = f(s,t) + u(s,t),$$

where,

$$f(s,t) = i(\tfrac{c(\alpha)t^{4-\beta}}{\beta(1-\alpha)}\sum_{k=0}^{\infty}(-1)^k(k+3)(k+2)(k+1)t^kE_{\alpha,k+4}(\tfrac{-\alpha t^\alpha}{1-\alpha}))e^{-is}.$$

The exact solution of state and control functions is mentioned following,

$$y(s,t) = it^3e^{-(t+is)},$$

$$u(s,t) = ie^{-ix}(2t^3e^{-t} + 3it^9e^{-3t} - (2s+1+3is^2)t^3e^{-t}).$$

The homogeneous initial and boundary conditions are obtained from the analytic solution of $y(s,t)$. For the numerical solution of this example, we have used the introduced method in section 5 with values of $m = n$. For some choices (α,β) and (e,f), MAE and RMS values of state and control variables are shown in Figures $11-14$. AE graphs for state and control functions with $m = n = 7,\ \alpha = 0.25, \beta = 0.25$ and $(e,f) = (0,1)$ are depicted in Figure 15.

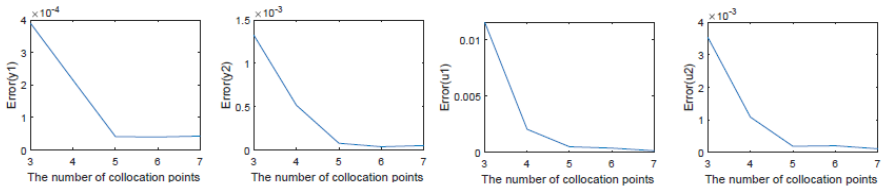


Figure 11: The RMS of the presented method for the state and control functions with $\alpha = 0.25$ and $\beta = 0.25$ in Example 3.

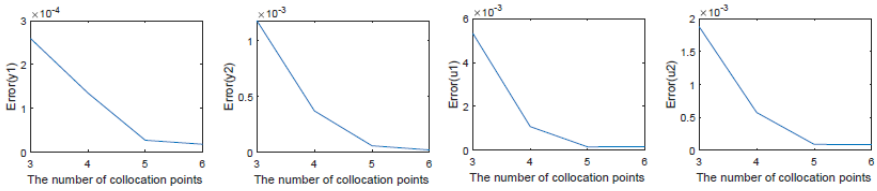


Figure 12: The RMS of the presented method for the state and control functions with $\alpha = 0.65$ and $\beta = 0.25$ in Example 3.

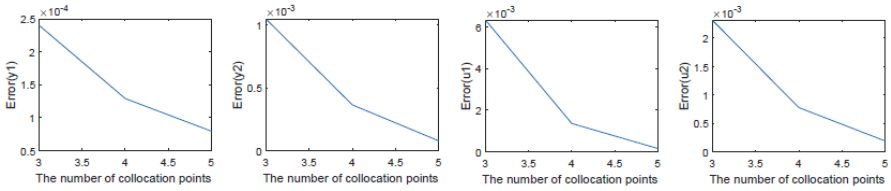


Figure 13: The RMS of the presented method for the state and control functions with $\alpha = 0.80$ and $\beta = 0.25$ in Example 3.

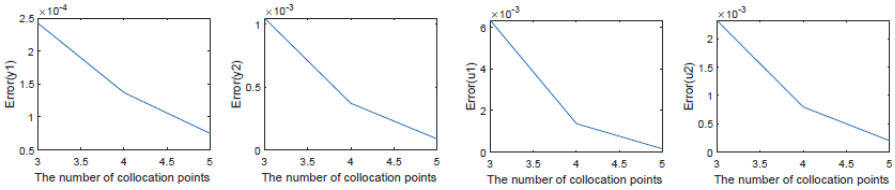


Figure 14: The RMS of the presented method for the state and control functions with $\alpha = 0.80$ and $\beta = 0.65$ in Example 1.

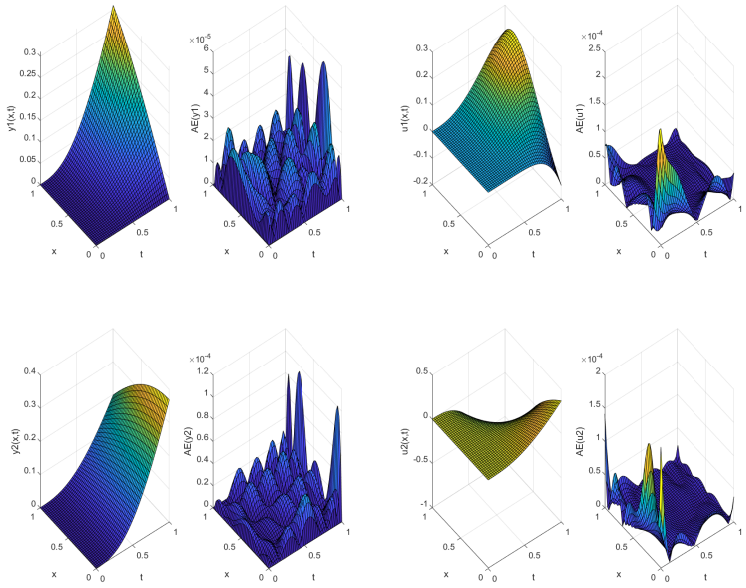


Figure 15: Numerical solution (SJPs) and error function (AE) surfaces of $y(s,t)$ (left) and $u(s,t)$ (right) with (m=n=7) in Example 3.

7 Conclusion

In this paper we introduced a novel class of optimal control with the nonlinear Ginzburg−Landau equation. To express this new class, we have used the FF derivative in the A-R-L sense with Mittag-Leffler non-singular kernel. For the numerical solution of this class of optimal control problems, an efficient method based on the shifted Jacobi polynomials has been proposed. To transform the main problem into a system of nonlinear algebraic equations, we have used the FF derivative operational matrix of SJPs and the collocation method. By presenting three numerical examples, we have investigated and evaluated the accuracy of the mentioned scheme.

References

[1] Aranson, I.S. and Kramer, L. *The world of the complex Ginzburg-Landau equation*, Reviews of Modern Physics 74(1) (2002), 99.

[2] Atangana, A. *Fractal-fractional differentiation and integration: connecting fractal calculus and fractional calculus to predict complex system*, Chaos, Solitons and Fractals, 102 (2017), 396–406.

[3] Atangana, A. and Qureshi, S. *Modeling attractors of chaotic dynamical systems with fractal–fractional operators*, Chaos, Solitons and Fractals 123 (2019), 320–337.

[4] Bhrawy, A.H., Doha, E.H., Baleanu, D. and Ezz-Eldien, S.S. *A spectral tau algorithm based on Jacobi operational matrix for numerical solution of time fractional diffusion-wave equations*, Journal of Computational Physics 293 (2015), 142–156.

[5] Darehmiraki, M., Farahi, M.H. and Effati, S. *A novel method to solve a class of distributed optimal control problems using Bezier curves*, Journal of Computational and Nonlinear Dynamics 11(6) (2016).

[6] Ding, H. and Li, C. *High-order numerical algorithm and error analysis for the two dimensional nonlinear spatial fractional complex Ginzburg-*

Landau equation, Communications in Nonlinear Science and Numerical Simulation 120 (2023) , 107160.

[7] Doha, E.H. *On the construction of recurrence relations for the expansion and connection coefficients in series of Jacobi polynomials*, Journal of Physics A: Mathematical and General 37(3) (2004), 657.

[8] Doha, E.H., Bhrawy, A.H. and Ezz-Eldien, S.S. *A new Jacobi operational matrix: an application for solving fractional differential equations*, Applied Mathematical Modelling 36(10) (2012), 4931–4943.

[9] Du, N., Wang, H. and Liu, W. *A fast gradient projection method for a constrained fractional optimal control*, Journal of Scientific Computing 68(1) (2016), 1–20.

[10] Goyal, A., Raju, T.S. and Kumar, C.N. *Lorentzian-type soliton solutions of ac-driven complex Ginzburg –- Landau equation*, Applied Mathematics and Computation 218(24) (2012), 11931–11937.

[11] Gu, X.M., Shi, L. and Liu, T. *Well-posedness of the fractional Ginzburg–Landau equation*, Applicable Analysis 98(14) (2019), 2545–2558.

[12] Gunzburger, M.D., Hou, L.S. and Ravindran, S.S. *Analysis and approximation of optimal control problems for a simplified Ginzburg-Landau model of superconductivity*, Numerische Mathematik 77 (1997), 243–268.

[13] Hasegawa, A. *Optical Solitons in Fibers*, In Optical solitons in fibers, Springer, Berlin, Heidelberg, 1989, 1–74

[14] Heydari, M.H., Atangana, A. and Avazzadeh, Z. *Chebyshev polynomials for the numerical solution of fractal-fractional model of nonlinear Ginzburg –- Landau equation*, Eng. Comput. (2019), 1–12.

[15] Heydari, M.H., Atangana, A. and Avazzadeh, Z. *Chebyshev polynomials for the numerical solution of fractal–fractional model of nonlinear Ginzburg –- Landau equation*, Engineering with Computers 2021(37) (2021), 1377–1388.

[16] Heydari, M.H., Avazzadeh, Z. and Atangana, A. *Shifted Jacobi polynomials for nonlinear singular variable-order time fractional Emden–Fowler equation generated by derivative with non-singular kernel*, Advances in Difference Equations 2021(1) (2021), 1–15.

[17] Kilicman, A. and Al Zhour, Z.A.A. *Kronecker operational matrices for fractional calculus and some applications*, Applied Mathematics and Computation 187(1) (2007), 250–265.

[18] Li, B. and Zhang, Z. *A new approach for numerical simulation of the time-dependent Ginzburg -– Landau equations*, Journal of Computational Physics 303 (2015), 238–250.

[19] Li, M., Huang, C. and Wang, N. *Galerkin finite element method for the nonlinear fractional Ginzburg -– Landau equation*, Applied Numerical Mathematics 118 (2017), 131–149.

[20] Lopez, V. *Numerical continuation of invariant solutions of the complex Ginzburg –- Landau equation*, Communications in Nonlinear Science and Numerical Simulation 61 (2018), 248–270.

[21] Luke, Y.L. *The special functions and their approximations*, Academic press, 53.

[22] Mainardi, F. and Gorenflo, R. *On Mittag-Leffler-type functions in fractional evolution processes*, J. Comput. Appl. Math. 118 (2000), 283–299.

[23] Mophou, G.M. *Optimal control of fractional diffusion equation*, Computers and Mathematics with Applications 61(1) (2011), 68–78.

[24] Shojaeizadeh, T., Mahmoudi, M. and Darehmiraki, M. *Optimal control problem of advection-diffusion-reaction equation of kind fractal-fractional applying shifted Jacobi polynomials*, Chaos, Solitons and Fractals 143 (2021), 110568.

[25] Shokri, A. and Bahmani, E. *Direct meshless local Petrov-–Galerkin (DMLPG) method for 2D complex Ginzburg–Landau equation*, Engineering Analysis with Boundary Elements 100 (2019), 195–203.

[26] Tarasov, V. *Fractional dynamics: applications of fractional calculus to dynamics of particles, fields and media*, Berlin: Springer-Verlag, jointly with Higher Education Press, Beijing, 2011.

[27] Toledo-Hernandez, R., Rico-Ramirez, V., Rico-Martinez, R., Hernandez-Castro, S. and Diwekar, U.M. *A fractional calculus approach to the dynamic optimization of biological reactive systems. Part II: Numerical solution of fractional optimal control problems*, Chemical Engineering Science 117 (2014), 239–247.

[28] Wang, N. and Huang, C. *An efficient split-step quasi-compact finite difference method for the nonlinear fractional Ginzburg–Landau equations*, Computers and Mathematics with Applications 75(7) (2018), 2223–2242.

[29] Wang, P. and Huang, C. *An implicit midpoint difference scheme for the fractional Ginzburg - Landau equation*, Journal of Computational Physics 312 (2016), 31–49.

[30] Weitzner, H. and Zaslavsky, G.M. *Commun Nonlinear*, Sci. Numer. Simulation 8 (2003), 273–281.

[31] Yan, Y., Moxley III, F.I. and Dai, W. *A new compact finite difference scheme for solving the complex Ginzburg–-Landau equation*, Applied Mathematics and Computation 260 (2015), 269–287.

[32] Zaslavsky, G. *Hamiltonian chaos and fractional dynamics*, Oxford: Oxford University Press, 2005.

[33] Zeng, W., Xiao, A. and Li, X. *Error estimate of Fourier pseudo–spectral method for multidimensional nonlinear complex fractional Ginzburg–-Landau equations*, Applied Mathematics Letters 93 (2019), 40–45.

Iranian Journal of Numerical Analysis and Optimization
Vol. 14, No. 3, 2024, pp 900–937
https://doi.org/ 10.22067/ijnao.2024.87373.1412
https://ijnao.um.ac.ir/

Research Article

A numerical computation for solving delay and neutral differential equations based on a new modification to the Legendre wavelet method

N.M. El-Shazly* and M.A. Ramadan

Abstract

The goal of this study is to use our suggested generalized Legendre wavelet method to solve delay and equations of neutral differential form with proportionate delays of different orders. Delay differential equations have some application in the mathematical and physical modelling of real-world problems such as human body control and multibody control systems, electric circuits, dynamical behavior of a system in fluid mechanics, chemical en-

*Corresponding author

Received 23 March 2024; revised 24 May 2024; accepted 31 May 2024

Naglaa M. El-Shazly
Mathematics and Computer Science Department, Faculty of Science, Menoufia University, Menoufia, Egypt. e-mail: Naglaamoh1@yahoo.com

Mohamed A. Ramadan
Mathematics and Computer Science Department, Faculty of Science, Menoufia University, Menoufia, Egypt. e-mail: ramadanmohamed13@yahoo.com

How to cite this article
El-Shazly, N.M. and Ramadan, M.A., A numerical computation for solving delay and neutral differential equations based on a new modification to the Legendre wavelet method. *Iran. J. Numer. Anal. Optim.*, 2024; 14(3): 900-937. https://doi.org/ 10.22067/ijnao.2024.87373.1412

gineering, infectious diseases, bacteriophage infection's spread, population dynamics, epidemiology, physiology, immunology, and neural networks.

The use of orthonormal polynomials is the key advantage of this method because it reduces computational cost and runtime. Some examples are provided to demonstrate the effectiveness and accuracy of the suggested strategy. The method's accuracy is reported in terms of absolute errors. The numerical findings are compared to other numerical approaches in the literature, particularly the regular Legendre wavelets method, and show that the current method is quite effective in order to solve such sorts of differential equations.

AMS subject classifications (2020): Primary 34K40; Secondary 65L05, 40C05.

Keywords: Generalized Legendre wavelets; orthonormal polynomials delay differential equations; neutral differential equations; accuracy.

1 Introduction

Delay differential equations are important in the mathematical and physical modelling of real-world problems such as human body control and multi-body control systems, electric circuits, the dynamical behavior of a system in fluid mechanics, chemical engineering [18, 33], infectious diseases, bacteriophage infection's spread [9], population dynamics, epidemiology, physiology, immunology, neural networks, and the application of Legendre wavelet for solving differential pharmacology.

For solving nonlinear differential equations with proportional delays, there are several numerical techniques, like the Runge–Kutta–Fehlberg methods [25], Adomian decomposition method [8, 14, 26], Hermite wavelet-based method [32], Aboodh transformation method [1], power series method [5], decomposition method [34], differential transform method [22], Iterative variational approach [19], Pade's series-based approach and power method [35], spectral method [2], variable multistep methods [21], quasilinearization technique [28], polynomial least squares method [10], homotopy perturbation method [31], first kind Bessel's functions [39, 40], Legendre polynomials of shifted form [41, 44], and the first Boubaker polynomial approach [12]. More-

over, solving equations of nonlinear ordinary differential type using collocation methods with the use of Bessel polynomials are studied in [38, 42, 43]. In addition, numerous numerical approaches, such as the One-leg-method [36] and Chebyshev polynomials [30], have been employed to approximate the solutions of neutral differential equations. Gümgüm, Özdek,. and Özaltun [16, 17] have presented Legendre wavelet solutions of both high order nonlinear ordinary delay differential equations and neutral differential equations with proportional delays. Nisar et al. [24] presented the efficient and significant solutions to a nonlinear fractional model. Zhang et al. [45] investigated the multiple solitons, lump solitons, and interaction with two stripe soliton solutions, for the fractional gCBS-BK equation. Amer and Olorode [3] presented a numerical evaluation of a novel slot-drill enhanced oil recovery technology for tight rocks.

In this paper, we use our suggested generalized Legendre wavelet approach ($GLWM$) [13] to solve delay and neutral differential equations with proportionate delays of different orders in the following form:

$$\sum_{p=0}^{3}\sum_{q=0}^{Q} R_{pq}(t)y^{(p)}(t-\eta_{pq}(t)) + \sum_{r=0}^{2}\sum_{s=0}^{r} S_{rs}(t)y^{(r)}(\delta_{rs}t)y^{(s)}(\gamma_{rs}t) = h(t), \quad Q \leq 3, \tag{1}$$

with the initial conditions

$$y^{(p)}(0) = \alpha_p \qquad p = 0\ ,\ 1\ ,\ 2, \tag{2}$$

where the given continuous functions $R_{pq}(t), S_{rs}(t), h\ (t)$ and the variable delays $\eta_{pq}(t)$ on $0 \leq t < 1$, δ_{rs} and γ_{rs} constants are assigned to indicate proportional delays.

The mechanism of our suggested method is reducing computational cost and runtime with the use of orthonormal polynomials. We provide various examples to demonstrate the effectiveness and accuracy of the suggested strategy. We reported the method's accuracy in terms of absolute errors. The numerical findings are compared to other numerical approaches in the literature, particularly the regular Legendre wavelets method ($RLWM$), and

manifest that the current method is quite effective in order to solve such sorts of differential equations.

2 Definitions and preliminaries

2.1 Legendre wavelet and its properties

Considering a single function "mother wavelet" $\psi(t)$, from which wavelets represent a family of functions by dilating and transforming this single function. This family of continuous wavelets [15] has the following form:

$$\psi_{a,b}(t) = |a|^{-\frac{1}{2}}\,\psi\left(\frac{t-b}{a}\right), \qquad a,\ b\ \in R\,,\ \ a \neq 0. \tag{3}$$

The Legendre wavelets on the interval $[0\ ,\ 1)$ is defined by

$$\psi_{n,m}(t) = \begin{cases} \sqrt{m+\frac{1}{2}}\ 2^{\frac{k}{2}} L_m(2^k t - \hat{n}), & \frac{\hat{n}-1}{2^k}\ \leq\ t\ <\ \frac{\hat{n}+1}{2^k}\ ; \\ 0 & otherwise, \end{cases} \tag{4}$$

for which k is positive integer, $n = 1\ ,\ 2\ , \ldots ,\ 2^{k-1}$ and $\hat{n} = 2n-1$, the order of the Legendre polynomial is denoted by $m = 0\ ,\ 1\ ,\ 2\ ,\ \ldots\ ,\ M$ and the normalized time is denoted by t. The Legendre polynomials acquired in the above definition are defined as follows:

$$\begin{aligned} &L_0(t) = 1\ \ , \\ &L_1(t) = t\ \ , \\ &L_{m+1}(t) = \frac{2m+1}{m+1}\ t\ L_m(t) - \frac{m}{m+1}\ L_{m-1}(t), \qquad m = 1,\ 2,\ 3,\ \ldots , \end{aligned} \tag{5}$$

which are orthogonal over $[-1, 1]$ with weighting function $w(t) = 1$, for more details (see [4]). After shifting the Legendre polynomials by $t = 2x - 1$, the shifted Legendre polynomials $L_m(x) = L_m^*(2x-1)$ that are orthogonal on $[0, 1)$ can be denoted as follows:

$$L_m(x) = \sum_{s=0}^{m} (-1)^{m+s} \frac{(m+s)!x^s}{(m-s)!(s!)^2}.$$

2.2 Generalized Legendre wavelet expansion

In this subsection, we offer a generalization for the Legendre wavelets method, denoted by $GLWM$ [13], for solving delay and neutral differential equations with proportionate delays of different orders defined as in (1). The proposed $GLWM$ on the interval $[0,1)$ are defined by

$$\psi^{(\mu)}{}_{n,m}(t) = \begin{cases} \sqrt{m+\frac{1}{2}}\ \mu^{\frac{k}{2}} L_m(\mu^k t - \hat{n}), & \frac{\hat{n}-1}{\mu^k} \leq t < \frac{\hat{n}+1}{\mu^k}\ ; \\ 0 & otherwise \end{cases} \tag{6}$$

for which k is positive integer, $n = 1\ ,\ 2\ ,\ldots,\ \mu^{k-1}\ ,\ \mu \geq 3$ and $\hat{n} = 2n-1$ and the order of the Legendre Polynomial is denoted by $m = 0\ ,\ 1\ ,\ 2\ ,\ \ldots\ ,\ M$ and the normalized time is denoted by t.

3 Discussion and results

3.1 Function approximation

A function $f(t)$ defined on $[0,1)$ may be expanded as infinite series of the type seen below:

$$f(t) = \sum_{n=1}^{\infty}\sum_{m=0}^{\infty} c_{n,m}\ \psi^{\mu}{}_{n,m}, \tag{7}$$

where $c_{n,m} =< f,\ \psi^{\mu}{}_{n,m} > = \int_0^1 f(t)\ \psi^{\mu}{}_{n,m}(t)\ dt$.

After trimming, (7) can be written as follows:

$$f(t) \approx \sum_{n=1}^{\mu^{k-1}}\sum_{m=0}^{M} c_{n,m}\ \psi^{\mu}{}_{n,m}(t) = C^T \Psi\ (t), \tag{8}$$

where

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

and

$$\Psi(t) = \left[\psi^{\mu}_{1,0}\psi^{\mu}{}_{1,1}\dots\psi^{\mu}_{1,M}\dots\psi^{\mu}_{2,0}\psi^{\mu}_{2,1}\dots\psi^{\mu}_{2,M}\dots\psi^{\mu}_{\mu^{k-1},0}\psi^{\mu}_{\mu^{k-1},1}\dots\psi^{\mu}_{\mu^{k-1},M}\right]^{T}.$$

3.2 Generalized Legendre wavelet operational matrix of differentiation

The qth derivative of the vector $\Psi(t)$, defined in (6) can be obtained by

$$\frac{d^q}{dt^q}\Psi(t) = D^q\Psi(t)\ , \tag{9}$$

where D^q is the qth power of the $\mu^{k-1}(M+1)\times\mu^{k-1}(M+1)$ operational matrix of differentiation D, defined in [23] as follows:

$$D = \begin{pmatrix} F & 0 & \cdots & 0 \\ 0 & F & \cdots & \vdots \\ \vdots & & \ddots & 0 \\ 0 & \cdots & 0 & F \end{pmatrix},$$

where F is a $(M+1)\times(M+1)$ submatrix of the type

$$F = \mu^k \times \begin{pmatrix} 0 & 0 & 0 & \dots & 0 & 0 & \\ \sqrt{3} & 0 & 0 & \dots & 0 & 0 & \\ 0 & \sqrt{3}\sqrt{5} & 0 & \dots & 0 & 0 & \\ \sqrt{7} & 0 & \sqrt{5}\sqrt{7} & \dots & 0 & 0 & \\ \vdots & \vdots & \vdots & \dots & \vdots & \vdots & \\ \sqrt{2M+1} & 0 & \sqrt{5}\sqrt{2M+1} & \dots & \sqrt{2M-1}\sqrt{2M+1} & 0 & \}M \text{ is odd} \\ 0 & \sqrt{3}\sqrt{2M+1} & 0 & \dots & \sqrt{2M-1}\sqrt{2M+1} & 0 & \}M \text{ is even} \end{pmatrix}.$$

3.3 The use of the operational differentiation matrix

To address the problem presented in (1) and (2), we first find the approximated solution considering the truncated series in (8), utilizing generalized Legendre wavelets as

$$y(t) \approx \sum_{n=1}^{\mu^{k-1}} \sum_{m=0}^{M} c_{n,m}\ {\psi^{\mu}}_{n,m}(t) = C^T \Psi(t)\ , \tag{10}$$

where the coefficients $c_{n,m}$ are to be determined. Using (9) to approximate the pth derivative as

$$y^p(t) = C^T \frac{d^p}{dt^p} \Psi(t)\ = C^T D^p \Psi(t). \tag{11}$$

Substituting (10) and (11) into (1) implies that

$$\begin{aligned} &\sum_{p=0}^{3} \sum_{q=0}^{Q} R_{pq}(t) C^T D^p \Psi(t - \eta_{pq}(t)) \\ &+ \sum_{r=0}^{2} \sum_{s=0}^{r} S_{rs}(t) C^T D^r \Psi\ (\delta_{rs} t)\ C^T D^s \Psi(\gamma_{rs} t)\ = h(t)\ \ ,\ \ Q \leq 3. \end{aligned} \tag{12}$$

We need $\mu^{k-1}(M+1)$ equations to determine the unbeknown coefficients $c_{n,m}$ of the vector C. The first three equations are derived using the initial conditions (2), (3), and (4) as

$$\begin{aligned} &y(0) = C^T D\ \Psi\ (0)\ , \\ &y^{(p)}\ (0) = C^T D^p\ \Psi\ (0)\ , \qquad p = 1,\ 2\ . \end{aligned}$$

and $\mu^{k-1}(M\ +\ 1)\ -\ 3$ equations are obtained by substituting the first $(\mu^{k-1}(M\ +\ 1))\ -\ 3$ roots of shifted Legendre polynomial $P_{\mu^{k-1}(M+1)}(t)$ in (12).

Then, using MATLAB, we can solve the obtained system of nonlinear equations and the approximated solution in (10) is obtained.

3.4 Convergence criteria of the proposed GLWM

In this subsection, we discuss the theoretical analysis of the convergence of our approach to solve (1).

We want to prove that $y(t) = \sum_{n=1}^{\infty} \sum_{m=0}^{\infty} c_{n,m} \psi^{\mu}_{n,m}(t)$ defined in (10) using the GLWM converges to $y(t)$.

Let $L^2(R)$ be the Hilbert space. We have shown that $\psi^{(\mu)}_{n,m}(t) = \sqrt{m+\frac{1}{2}}\mu^{\frac{k}{2}} L_m(\mu^k t - \hat{n})$ forms an orthonormal basis [13].

Let $y(t) = \sum_{i=0}^{M} h_{ni}\psi^{\mu}_{ni}(t)$ be a solution of (1) such that $h_{1i} = \langle y(t), \psi^{\mu}_{1i}(t)\rangle$ for $n = 1$ in which $\langle\ .\ ,\ .\ \rangle$ denotes the inner product.

Let we denote $\psi^{\mu}_{ni}(t) = \psi^{\mu}(t)$ and $\alpha_j = \langle y(t), \psi^{\mu}(t)\rangle$

$$y(t) = \sum_{i=1}^{M} \langle y(t), \psi^{\mu}_{1i}(t)\rangle\ \ \psi^{\mu}_{1i}(t).$$

Consider the sequences of partial sums

$$W_{n-1} = \sum_{j=1}^{n-1} \alpha_j\ \psi^{\mu}(t_j) \qquad \text{and} \qquad W_{m-1} = \sum_{j=1}^{m-1} \alpha_j\ .\psi^{\mu}(t_j)$$

Then,

$$\begin{aligned}\langle y(t), W_{n-1}\rangle &= \left\langle y(t), \sum_{j=1}^{n-1} \alpha_j\ \psi^{\mu}(t_j) \right\rangle = \sum_{j=1}^{n-1} \bar{\alpha}_j\ \ \langle y(t), \psi^{\mu}(t_j)\rangle \\ &= \sum_{j=1}^{n-1} \bar{\alpha}_j\ \alpha_j = \sum_{j=1}^{n-1} |\alpha_j|^2 .\end{aligned}$$

Moreover,

$$\begin{aligned}\parallel W_{n-1} - W_{m-1} \parallel^2 &= \left\| \sum_{j=m}^{n-1} \alpha_j\ \psi^{\mu}(t_j) \right\|^2 \\ &= \left\langle \sum_{i=m}^{n-1} \alpha_i\ \psi^{\mu}(t_i)\ ,\ \sum_{j=m}^{n-1} \alpha_j\ \psi^{\mu}(t_j) \right\rangle \\ &= \sum_{i=m}^{n-1}\sum_{j=m}^{n-1} \alpha_i\ \bar{\alpha}_j\ \ \langle \psi^{\mu}(t_i)\ ,\ \psi^{\mu}(t_j)\rangle = \sum_{i=m}^{n-1} |\ \alpha_i\ |^2 .\end{aligned}$$

As $n \to \infty$, by Bessel's inequality, we get that $\sum_{i=m}^{n-1} |\ \alpha_i\ |^2$ is convergent, it yields that $\{W_{n-1}\}$ is a Cauchy sequence and it converges to W (say).

Now, we have

$$\begin{aligned} < W - y(t), \psi^{\mu}(t_j) > &= < W, \psi^{\mu}(t_j) > - < y(t), \psi^{\mu}(t_j) > \\ &= < \lim_{n\to\infty} W_{n-1} \ , \ \psi^{\mu}(t_j) > \ - \ \alpha_j \\ &= \lim_{n\to\infty} < W_{n-1} \ , \ \psi^{\mu}(t_j) > \ - \ \alpha_j \\ &= \lim_{n\to\infty} < \sum_{j=1}^{n-1} \alpha_j \ \psi^{\mu}(t_j) \ , \ \psi^{\mu}(t_j) > \ - \alpha_j \\ &= \alpha_j - \alpha_j = 0, \end{aligned}$$

which is satisfied only in the case if $y(t) = W$. Thus, $y(t) = \sum_{j=1}^{\infty} \alpha_j \ \psi^{\mu}(t_j)$.

3.5 Error bound

Suppose that the function $y(t)$ defined in $[0, \ 1]$ is m times continuously differentiable function. Then there exists a mean error bound for the approximation of $\sum_{n=1}^{\mu^{k-1}} \sum_{m=0}^{M} c_{n,m} \ \psi^{\mu}(t) = C^T \Psi^{\mu}(t)$ to $y(t)$ as follows [37]:

$$\| \ y - C^T \psi^{\mu}(t) \ \| \leq \frac{1}{m! \mu^{mk}} \sup_{t\in[0, \ 1]} \left| y^{(m)}(t) \ \right| \ .$$

We divide the interval $[0, \ 1]$ into subintervals $\left[\frac{\hat{n}-1}{\mu^k} \ , \ \frac{\hat{n}+1}{\mu^k} \right]$. So we can approximate $y(t)$ to the polynomial $C^T \Psi^{\mu}(t)$ of mth degree, taking into consideration a minimum error for these subintervals. Therefore, we can utilize the maximum error estimation for this polynomial that insets $y(t)$, that is,

$$\begin{aligned}
\left\| y - C^T\psi^{\mu}(t) \right\|^2 &= \int_0^1 \left[y(t) - C^T\psi^{\mu}(t)\right]^2 \quad dt \\
&= \sum_{n=1}^{\mu^{k-1}} \int_{\frac{2n-2}{\mu^k}}^{\frac{2n}{\mu^k}} \left[\, y(t) - C^T\psi^{\mu}(t)\right]^2 \, dt \\
&\leq \sum_{n=1}^{\mu^{k-1}} \int_{\frac{2n-2}{\mu^k}}^{\frac{2n}{\mu^k}} \left[y(t) - y^*(t)\right]^2 \, dt \\
&\leq \sum_{n=1}^{\mu^{k-1}} \int_{\frac{2n-2}{\mu^k}}^{\frac{2n}{\mu^k}} \left[\frac{1}{m!\mu^{mk}} \sup_{t\in[0,\,1]} \left|y^{(m)}(t)\right|\right]^2 \quad dt \\
&\leq \int_0^1 \left[\frac{1}{m!\mu^{mk}} \sup_{t\in[0,\,1]} \left|y^{(m)}(t)\right|\right]^2 dt \\
&= \frac{1}{m!\mu^{mk}} \sup_{t\in[0,\,1]} \left|y^{(m)}(t)\right|^2 ,
\end{aligned}$$

where $y^*(t)$ denotes the mth order interpolation of $y(t)$. Taking the square roots of both sides yields the desired outcome.

4 Numerical examples

In this section, we demonstrate the advantage and high accuracy of our proposed GLWM by applying it to various conventional delay differential equations. All the numerical test examples of our program were carried out by MATLAB R2015a.

Example 1. Assume the equation of the second-order neutral differential form through proportional delays shown below [17]:

$$\begin{aligned}
&y''(t) = \frac{3}{4}y(t) + y(\frac{t}{2}) + y'(\frac{t}{2}) + \frac{1}{2}y''(\frac{t}{2}) - t^2 - t + 1, \qquad 0 \leq t \leq 1, \\
&y(0) = 0\,, \qquad y'(0) = 0.
\end{aligned} \tag{13}$$

The exact solution of this initial value problem is given by $y(t) = t^2$.

We first apply the GLWM for $M = 2, k = 1, \mu = 3$.

For this choice of M, k, μ, the function approximation for $y(t)$ will take the summation form,

$$y(t) \approx \sum_{n=1}^{\mu^{k-1}} \sum_{m=0}^{M} c_{n\,,\,m}\, \psi^{\mu}_{n,m}(t) = \sum_{n=1}^{1} \sum_{m=0}^{2} c_{n\,,\,m}\, \psi^{\mu}_{n,m}(t) = C^T \Psi\,, \quad (14)$$

where $C_{3\times 1} = [\, c_{1,0} \;\; c_{1,1} \;\; c_{1,2} \,]^T$ and $\Psi_{3\times 1}(t) = \left[\psi^{\mu}_{1,0}(t)\; \psi^{\mu}_{1,1}(t)\; \psi^{\mu}_{1,2}(t) \right]^T$, where the generalized Legendre wavelets $\psi^{\mu}_{1,m}(t) \quad , \quad m = 0,\, 1\,,\, 2$ in this case, are given by

$$\psi^{\mu}_{1,0}(t) = \begin{cases} \frac{\sqrt{2}\sqrt{3}}{2}, & 0 \le t < \frac{2}{3}, \\ 0, & otherwise, \end{cases}$$

$$\psi^{\mu}_{1,1}(t) = \begin{cases} \frac{3\sqrt{2}}{2}(3t-1), & 0 \le t < \frac{2}{3}, \\ 0, & otherwise, \end{cases}$$

$$\psi^{\mu}_{1,2}(t) = \begin{cases} \frac{\sqrt{15}}{\sqrt{2}}\,(\frac{3}{2}(3t-1)^2 - \frac{1}{2}), & 0 \le t < \frac{2}{3}, \\ 0, & otherwise. \end{cases}$$

Thus, $y(t)$ and $y(\frac{t}{2})$ can be approximated as

$$y(t) = c_{1,0}\,\frac{\sqrt{2}\sqrt{3}}{2} + c_{1,1}\frac{3\sqrt{2}}{2}(3t-1) + c_{1,2}\,\frac{\sqrt{15}}{\sqrt{2}}\,(\frac{3}{2}(3t-1)^2 - \frac{1}{2})\,,$$

$$y(\frac{t}{2}) = c_{1,0}\,\frac{\sqrt{2}\sqrt{3}}{2} + c_{1,1}\frac{3\sqrt{2}}{2}(3\frac{t}{2}-1) + c_{1,2}\,\frac{\sqrt{15}}{\sqrt{2}}\,(\frac{3}{2}(3\frac{t}{2}-1)^2 - \frac{1}{2})\,.$$

To calculate the first and second derivatives of $y(t)$, we use the 3×3 operational matrix of differentiation$P$ and P^2 in the form

$$P = \begin{pmatrix} 0 & 0 & 0 \\ 3\sqrt{3} & 0 & 0 \\ 0 & 3\sqrt{15} & 0 \end{pmatrix}, \qquad P^2 = \begin{pmatrix} 0 & 0\,0 \\ 0 & 0\,0 \\ 9\sqrt{45} & 0\,0 \end{pmatrix}$$

as follows:

$$\begin{aligned} y'(t) &= c_{1,1} 3\sqrt{3}\psi^{\mu}_{1,0} + c_{1,2} 3\sqrt{15}\psi^{\mu}_{1,1} \\ &= c_{1,1} 3\sqrt{3}\frac{\sqrt{2}\sqrt{3}}{2} + c_{1,2} 3\sqrt{15}\frac{3\sqrt{2}}{2}(3t-1), \end{aligned}$$

$$y''(t) = c_{1,2}\, 9\sqrt{45}\, \psi^{\mu}_{1,0} = c_{1,2}\, 9\sqrt{45}\, \frac{\sqrt{2}\sqrt{3}}{2}\,,$$

and hence $y''(\frac{t}{2}) = c_{1,2}\, 9\sqrt{45}\, \frac{\sqrt{2}\sqrt{3}}{2}$.

Using these approximations, (13) takes the form

$$
\begin{aligned}
c_{1,2}9\sqrt{45}\frac{\sqrt{2}\sqrt{3}}{2} &= \frac{3}{4}(c_{1,0}\frac{\sqrt{2}\sqrt{3}}{2} + c_{1,1}\frac{3\sqrt{2}}{2}(3t-1) + c_{1,2}\,\frac{\sqrt{15}}{\sqrt{2}}(\frac{3}{2}(3t-1)^2 - \frac{1}{2}) \\
&+ (c_{1,0}\frac{\sqrt{2}\sqrt{3}}{2} + c_{1,1}\frac{3\sqrt{2}}{2}(3\frac{t}{2}-1) + c_{1,2}\frac{\sqrt{15}}{\sqrt{2}}(\frac{3}{2}(3\frac{t}{2}-1)^2 - \frac{1}{2}) \\
&+ c_{1,1}3\sqrt{3}\frac{\sqrt{2}\sqrt{3}}{2} + c_{1,2}3\sqrt{15}\frac{3\sqrt{2}}{2}(3\frac{t}{2}-1) \\
&+ \frac{1}{2}c_{1,2}\,9\sqrt{45}\frac{\sqrt{2}\sqrt{3}}{2} - t^2 - t\, + 1.
\end{aligned} \tag{15}
$$

It should be noted that in order to find the unknown coefficients, $c_{1,0}\ c_{1,1}\ c_{1,2}$, we need three equations. Two equations are obtained from the initial conditions in (13) as follows:

$$
\begin{aligned}
c_{1,0}\,\frac{\sqrt{2}\sqrt{3}}{2} \;-\; c_{1,1}\frac{3\sqrt{2}}{2} \;+\; c_{1,2}\,\frac{\sqrt{15}}{\sqrt{2}} &= 0\,, \\
c_{1,1}\,3\sqrt{3}\,\frac{\sqrt{2}\sqrt{3}}{2} \;-\; c_{1,2}\,3\sqrt{15}\,\frac{3\sqrt{2}}{2} &= 0\,.
\end{aligned}
$$

We can gain the third equation by inserting the first root of third-order shifted generalized Legendre polynomial, given by $t = 0.07513$, in (15). Solving this 3×3 nonlinear system gives

$$
\begin{aligned}
C_{3\times1} &= [c_{1,0} \quad c_{1,1} \quad c_{1,2}]^T \\
&= [0.12096245643373 \quad 0.104756560175784 \quad 0.027048027531119]^T.
\end{aligned}
$$

Hence, the approximate solution of [17, Example 1] using our proposed $GLWM$ is obtained as

$$
\begin{aligned}
y(t) &= C^T\Psi \\
&= [0.12096245643373 \quad 0.104756560175784 \quad 0.027048027531119]^T \times \\
&\begin{bmatrix} \frac{\sqrt{2}\sqrt{3}}{2} \\ \frac{3\sqrt{2}}{2}(3t-1) \\ \frac{\sqrt{15}}{\sqrt{2}}\,(\frac{3}{2}(3t-1)^2 - \frac{1}{2}) \end{bmatrix}.
\end{aligned}
$$

Along with the absolute errors compared to the exact solution, the estimates of the approximation can be evaluated at the locations in the prescribed interval, $0 \le t < \frac{2}{3}$ and summarized in the table (Table 1) below.

Table 1: Approximate solution and the absolute error of in [17, Example 1] using our $GLWM$ for $M = 2; k = 1; \mu = 3$

t	**Exact solution**	**Approximate solution**	**Absolute error**
0.1	0.01	0.009999999999990999	9.0015e-15
0.2	0.04	0.03999999999999143	8.5733e-15
0.3	0.09	0.08999999999999162	8.3786e-15
0.4	0.16	0.1599999999999916	8.4176e-15
0.5	0.25	0.2499999999999913	8.6902e-15
0.6	0.36	0.3599999999999908	9.1964e-15

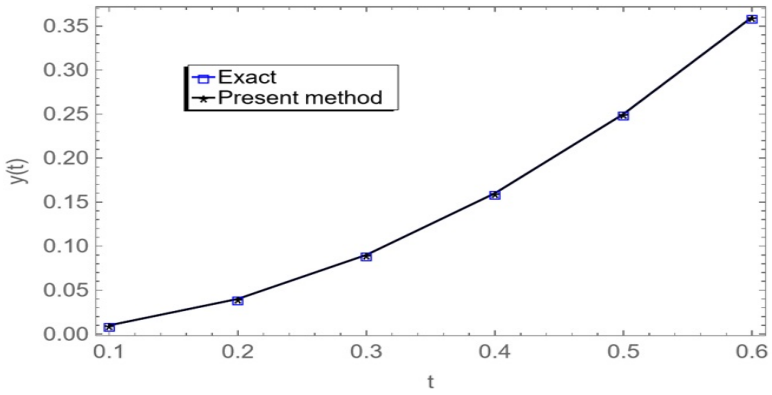


Figure 1: Approximate solution against the exact solution for Example 1

In Table 2 below, absolute error comparisons between the proposed approach $GLWM$ and the $RLWM$ for the same M ($M = 2$) and other numerical methods, namely, the Runge–Kutta method of two-stage order-one case (RKM) [25], One-leg θ method with $\theta = 0.8$ [36], variational iteration method (VIM) with $n = 6$ [11], homotopy perturbation method (HPM) with $n = 6$ [7], reproducing kernel Hilbert space method (RKHSM) with $n = 100$ [20], Legendre–Gauss collocation method (LCM) with $n = 10$ [6], homotopy analysis method (HAM) with $n = 6$ [29] are provided. Also, we present solutions on this interval for comparison because the numerical approaches mentioned previously produced solutions in the same range. From Table 1, Figure 1, and Table 2, we can presume that the current method is more ac-

curate, effective, and promising when compared to other numerical methods, particularly with the normal Legendre wavelet method.

Table 2: Comparison of the absolute error of the suggested method with other numerical methods

t	**Present method** $GLWM(M=2)$	$RLWM$[17] $(M=2)$	**One-leg θ Method**[36]	**RKM**[25]
0.1	9.0015e-15	3.43e-11	6.10e-03	1.00e−03
0.2	8.5733e-15	7.79e-11	2.58e-02	2.02e−03
0.3	8.3786e-15	1.98e-10	6.47e-02	3.07e−03
0.4	8.4176e-15	3.26e-10	1.37e-01	4.17e−03
0.5	8.6902e-15	4.62e-10	2.81e-01	5.34e−03

t	**VIM** $n=6$ [11]	**RKHSM** $n=100$ [2]	**HPM** $n=6$ [7]	**HAM** $n=6$ [20]	**LCM** $n=10$ [6]
0.1	1.67e−04	9.57e−06	1.67e−04	2.25e−08	6.59e−17
0.2	7.15e−04	1.95e−04	7.15e−04	9.81e−08	1.37e−17
0.3	1.73e−03	2.94e−04	1.72e−03	2.44e−07	5.67e−18
0.4	3.30e−03	3.93e−04	3.30e−03	4.90e−07	6.98e−17
0.5	5.55e−03	4.92e−04	5.55e−03	8.69e−07	2.13e−17

Example 2. Consider the following equation of the first order neutral differential form through proportional delay [17]:

$$y'(t) = -y(t) + 0.1y(0.8t) + 0.5y'(0.8t) + (0.32t - 0.5)e^{-0.8t} + e^{-t}, 0 \le t \le 1$$
$$y(0) = 0. \tag{16}$$

The exact solution of this initial value problem is given by $y(t) = te^{-t}$.

We first apply the GLWM for $M = 4, k = 1, \mu = 3$.

For this choice of M, k, μ, the function approximation for $y(t)$ will take the summation form:

$$y(t) \approx \sum_{n=1}^{\mu^{k-1}} \sum_{m=0}^{M} c_{n,m}\psi^{\mu}_{n,m}(t) = \sum_{n=1}^{1} \sum_{m=0}^{4} c_{n,m}\psi^{\mu}_{n,m}(t) = C^T\Psi, \tag{17}$$

where $C_{5\times1} = [c_{1,0} \quad c_{1,1} \quad c_{1,2} \quad c_{1,\,3} \quad c_{1,\,4}]^T$ and

$$\Psi_{5\times1}(t) \;=\; \left[\, \psi^{\mu}_{1,0}(t)\; \psi^{\mu}_{1,1}(t)\; \psi^{\mu}_{1,2}(t) \quad \psi^{\mu}_{1,3}(t) \;\; \psi^{\mu}_{1,4}(t) \,\right]^{T},$$

where the generalized Legendre wavelets $\psi^{\mu}_{1,m}(t), m = 0, 1, 2, 3, 4$, which in this case, are given by

$$\psi^{\mu}_{1,0}(t) = \begin{cases} \frac{\sqrt{6}}{2}, & 0 \;\le\; t \;<\; \frac{2}{3}, \\ 0, & otherwise, \end{cases}$$

$$\psi^{\mu}_{1,1}(t) = \begin{cases} \frac{3\sqrt{2}}{2}(3t-1), & 0 \;\le\; t \;<\; \frac{2}{3}, \\ 0, & otherwise, \end{cases}$$

$$\psi^{\mu}_{1,2}(t) = \begin{cases} \frac{\sqrt{30}}{4}\;(3(3t-1)^2-1), & 0 \;\le\; t \;<\; \frac{2}{3}, \\ 0, & otherwise, \end{cases}$$

$$\psi^{\mu}_{1,3}(t) = \begin{cases} \frac{\sqrt{42}}{4}\;(3t-1)\,(5(3t-1)^2-3), & 0 \;\le\; t \;<\; \frac{2}{3}, \\ 0, & otherwise, \end{cases}$$

$$\psi^{\mu}_{1,4}(t) = \begin{cases} \frac{3\sqrt{6}}{16}\;(35(3t-1)^4-30(3t-1)^2+3), & 0 \;\le\; t \;<\; \frac{2}{3}, \\ 0, & otherwise. \end{cases}$$

So, we can approximate $y(t)$ and $y(0.8t)$ as

$$\begin{aligned} y(t) =& c_{1,0}\frac{\sqrt{6}}{2} + c_{1,1}\frac{3\sqrt{2}}{2}(3t-1) + c_{1,2}\frac{\sqrt{30}}{4}(3(3t-1)^2-1) \\ &+ c_{1,3}\frac{\sqrt{42}}{4}(3t-1)(5(3t-1)^2-3) \\ &+ c_{1,4}\frac{3\sqrt{6}}{16}(35(3t-1)^4-30(3t-1)^2+3), \end{aligned}$$

$$\begin{aligned} y(0.8t) =& c_{1,0}\frac{\sqrt{6}}{2} + c_{1,1}\frac{3\sqrt{2}}{2}(3(0.8t)-1) + c_{1,2}\frac{\sqrt{30}}{4}\,(3(3(0.8t)-1)^2-1) \\ &+ c_{1,3}\frac{\sqrt{42}}{4}(3(0.8t)-1)(5(3(0.8t)-1)^2-3) \\ &+ c_{1,4}\frac{3\sqrt{6}}{16}(35(3(0.8t)-1)^4-30(3(0.8t)-1)^2+3). \end{aligned}$$

In order to approximate the first derivative of $y(t)$, we use the 5×5 operational matrix of differentiation P in the form

$$P = \begin{pmatrix} 0 & 0 & 0 & 0 & 0 \\ 3\sqrt{3} & 0 & 0 & 0 & 0 \\ 0 & 3\sqrt{15} & 0 & 0 & 0 \\ 3\sqrt{7} & 0 & 3\sqrt{35} & 0 & 0 \\ 0 & 3\sqrt{27} & 0 & 3\sqrt{63} & 0 \end{pmatrix},$$

as follows:

$$y'(t) = \left(c_{1,1}3\sqrt{3} + c_{1,3}3\sqrt{7}\right)\psi_{1,0}^{\mu} + \left(c_{1,2}3\sqrt{15} + c_{1,4}3\sqrt{27}\right)\psi_{1,1}^{\mu} + c_{1,3}3\sqrt{35}\psi_{1,2}^{\mu} + c_{1,4}3\sqrt{63}\psi_{1,3}^{\mu}.$$

Using these approximations, (16) takes the form

$$\begin{aligned}
&(c_{1,1}3\sqrt{3} + c_{1,3}3\sqrt{7})\frac{\sqrt{6}}{2} + (c_{1,2}3\sqrt{15} + c_{1,4}3\sqrt{27})\,(\frac{3\sqrt{2}}{2}(3t-1)) \\
&+c_{1,3}3\sqrt{35}\left(\frac{\sqrt{30}}{4}(3(3t-1)^2-1)\right) + c_{1,4}3\sqrt{63}\left(\frac{\sqrt{42}}{4}(3t-1)(5(3t-1)^2-3)\right) \\
&= -\begin{bmatrix} c_{1,0}\frac{\sqrt{6}}{2} + c_{1,1}\frac{3\sqrt{2}}{2}(3t-1) + c_{1,2}\frac{\sqrt{30}}{4}(3(3t-1)^2-1) \\ +c_{1,3}\frac{\sqrt{42}}{4}(3t-1)(5(3t-1)^2-3) \\ +c_{1,4}\frac{3\sqrt{6}}{16}\,(35(3t-1)^4 - 30(3t-1)^2+3) \end{bmatrix} \\
&\quad +0.1\begin{bmatrix} c_{1,0}\frac{\sqrt{6}}{2} + c_{1,1}\frac{3\sqrt{2}}{2}(3(0.8t)-1) + c_{1,2}\frac{\sqrt{30}}{4}(3(3(0.8t)-1)^2-1) \\ +c_{1,3}\frac{\sqrt{42}}{4}(3(0.8t)-1)(5(3(0.8t)-1)^2-3) \\ +c_{1,4}\frac{3\sqrt{6}}{16}(35(3(0.8t)-1)^4 - 30(3(0.8t)-1)^2+3) \end{bmatrix} \\
&\quad +0.5\begin{bmatrix} \left(c_{1,1}3\sqrt{3} + c_{1,3}3\sqrt{7}\right)\frac{\sqrt{6}}{2} + \left(c_{1,2}3\sqrt{15} + c_{1,4}3\sqrt{27}\right)\left(\frac{3\sqrt{2}}{2}(3(0.8t)-1)\right) \\ +c_{1,3}3\sqrt{35}\,\left(\frac{\sqrt{30}}{4}(3(3(0.8t)-1)^2-1)\right) \\ +c_{1,4}3\sqrt{63}\left(\frac{\sqrt{42}}{4}(3(0.8t)-1)\,(5(3(0.8t)-1)^2-3)\right) \end{bmatrix} \\
&\quad + (0.32t-0.5)e^{-0.8t} + e^{-t}. \qquad (18)
\end{aligned}$$

Note that in order to determine the unknown coefficients $c_{1,0}$ $c_{1,1}$ $c_{1,2}$ $c_{1,3}$ $c_{1,4}$, we need five equations. One equation is obtained from the initial conditions in (16) as follows:

$$c_{1,0}\frac{\sqrt{6}}{2} - c_{1,1}\frac{3\sqrt{2}}{2} + c_{1,2}\frac{\sqrt{30}}{2} - c_{1,3}\frac{\sqrt{42}}{2} + c_{1,4}\frac{3\sqrt{6}}{2} = 0.$$

The second, third, fourth, and fifth equations are obtained by inserting the smaller four roots of the sixth-order shifted generalized Legendre polynomial,

that are given by $t_1 = 0.03127,\ \ t_2 = 0.1538,\ \ t_3 = 0.3333,\ \ t_4 = 0.5128$, in (18).

Solving this nonlinear 5×5 system gives

$$\begin{aligned} C_{5\times 1} &= [\ c_{1,0} \quad c_{1,1} \quad c_{1,2} \quad c_{1,3} \quad c_{1,4}\ ]^T \\ &= [0.17673294249913 \quad 0.07841472133705 \quad 0.01644085313212 \\ &\quad 0.00146577426114 \quad 0.00009124056578]^T. \end{aligned}$$

Hence, the approximate solution of Example 2 using our proposed $GLWM$ is obtained as

$$\begin{aligned} y(t) &= C^T \Psi \\ &= [0.17673294249913 \quad 0.07841472133705 \quad 0.01644085313212 \\ &\quad 0.00146577426114 \quad 0.00009124056578]^T \\ &\quad * \begin{bmatrix} \frac{\sqrt{6}}{2} \\ \frac{3\sqrt{2}}{2}(3t-1) \\ \frac{\sqrt{30}}{4}\ (3(3t-1)^2-1) \\ \frac{\sqrt{42}}{4}\ (3t-1)\ (5(3t-1)^2-3) \\ \frac{3\sqrt{6}}{16}\ (35(3t-1)^4-30(3t-1)^2+3) \end{bmatrix}. \end{aligned}$$

Along with the absolute errors compared to the exact solution, we can evaluate the approximation at the locations in the prescribed interval, $0 \leq t < \frac{2}{3}$ and summarized in the table (Table 3) below.

Table 3: Comparison of the absolute error for Example 2 of the suggested method with other numerical methods

t	**suggested method** $GLWM$ $(M=4), \mu=3$	$RLWM$ [17] $(M=4)$	**One-leg θ Method**[36]	**Two-stage order-one Runge-Kutta method** [25]	**Variational iteration method** $n=6$ [11]	**RKHSM** $n=100$ [7]	**HPM** $n=6$ [7]
0.1	**6.44e-07**	1.19e−05	4.65e−03	8.68e−04	1.30e−03	1.42e−04	1.06e−03
0.2	**3.78e-06**	2.01e−05	1.45e−02	1.49e−03	2.14e−03	1.17e−04	1.35e−03
0.3	**2.50e-06**	2.40e−05	2.57e−02	1.90e−03	2.63e−03	9.45e−04	1.18e−03
0.4	**3.33e-06**	2.15e−06	3.60e−02	2.16e−03	2.84e−03	7.59e−04	7.61e−04
0.5	**5.88e-06**	2.76e−05	4.43e−02	2.28e−03	2.83e−03	6.03e−04	2.32e−04
0.6	**1.35e-05**	2.13e−05	5.03e−02	2.31e−03	2.67e−03	4.73e−04	2.98e−04

In Table 3 and Figure 2, absolute error comparisons between the proposed approach $GLWM$ and the $RLWM$ for the same M $(M = 2)$ and other

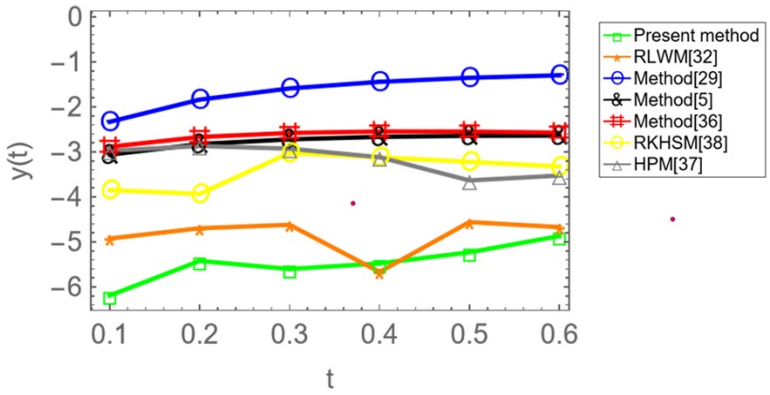


Figure 2: Absolute error for Example 2 using the presented method against the other methods listed in Table 3

numerical methods, Runge–Kutta method of two-stage order-one case (RKM) [**26**], One-leg θ method with $\theta = 0.8$ [36], Variational iteration method (VIM) with $n = 6$ [11], Homotopy perturbation method (HPM) with $n = 6$ [7], Reproducing Kernel Hilbert space method (RKHSM) with $n = 100$ [20], Legendre–Gauss collocation method (LCM) with $n = 10$ [6], Homotopy analysis method(HAM) with $n = 6$ [29] are provided.

We can presume that the current method is more effective and promising when compared to other numerical solutions, particularly with the normal Legendre wavelet method. Moreover, the absolute errors compared to the exact solution, we can evaluate the approximation at the locations in the prescribed interval, $0 \leq t < \frac{1}{2}$, for two values of $\mu = 3$, $\mu = 4$ and summarized in the next table, Table 4, given as.

As, one can see the absolute error is improved as we increase the values of the parameter μ .

Example 3. Consider the following third order nonlinear equation with proportional delay [16]:

$$\begin{aligned} &y'''(t) + 1 - 2y^2(\frac{t}{2}) = 0 \ , \qquad 0 \leq t < 1 \ . \\ &y(0) = 0 \ , \ y'(0) = 1 \ , \ y''(0) = 0, \end{aligned} \tag{19}$$

with the exact solution of the type $y(t) = \sin(t)$.

Table 4: Comparison of the absolute error for Example 2 of our proposed method GLWM in two cases $\mu = 3$, $\mu = 4$.

t	**Present method** $GLWM(M = 4)$, $\mu = 3$	**Present method** $GLWM(M = 4)$, $\mu = 4$
0.1	**6.44e-07**	**7.52 e-07**
0.2	**3.78e-06**	**1.49e-08**
0.3	**2.50e-06**	**8.55e-07**
0.4	**3.33e-06**	**1.64e-06**
0.5	**5.88e-06**	**2.43e-05**

We first apply the GLWM for $M = 5, k = 1, \mu = 3$.

For this choice of M, k , μ, the function approximation for $y(t)$ will take the summation form:

$$y(t) \approx \sum_{n=1}^{\mu^{k-1}} \sum_{m=0}^{M} c_{n\,,\,m}\, \psi^{\mu}_{n,m}(t) = \sum_{n=1}^{1} \sum_{m=0}^{5} c_{n\,,\,m}\, \psi^{\mu}_{n,m}(t) = C^T \Psi\,, \quad (20)$$

where $C_{6\times 1} = [\, c_{1,0} \quad c_{1,1} \quad c_{1,2} \quad c_{1,\,3} \quad c_{1,\,4} \quad c_{1,\,5}]^T$ and

$$\Psi_{6\times 1}(t) = \left[\psi^{\mu}_{1,0}(t)\ \psi^{\mu}_{1,1}(t)\ \psi^{\mu}_{1,2}(t) \quad \psi^{\mu}_{1,3}(t)\ \psi^{\mu}_{1,4}(t) \quad \psi^{\mu}_{1,5}(t)\ \right]^T,$$

where the generalized Legendre wavelets are $\psi^{\mu}_{1,m}(t)$, $m = 0,\ 1\ ,\ 2\ ,\ 3,\ 4,\ 5$.

So, we can approximate $y(t)$ and $y(t/2)$ as

$$\begin{aligned} y(t) =& c_{1,0}\frac{\sqrt{6}}{2} + c_{1,1}\frac{3\sqrt{2}}{2}(3t-1) + c_{1,2}\frac{\sqrt{30}}{4}(3(3t-1)^2-1) \\ &+ c_{1,3}\frac{\sqrt{42}}{4}(3t-1)\ (5(3t-1)^2-3) \\ &+ c_{1,4}\frac{3\sqrt{6}}{16}(35(3t-1)^4 - 30(3t-1)^2 + 3) \\ &+ c_{1,5}\frac{\sqrt{66}}{16}\left(63(3t-1)^5 - 70(3t-1)^3 + 15(3t-1)\right), \end{aligned}$$

$$\begin{aligned} y(t/2) =& c_{1,0}\frac{\sqrt{6}}{2} + c_{1,1}\frac{3\sqrt{2}}{2}(3(t/2)-1) + c_{1,2}\frac{\sqrt{30}}{4}\,(3(3(t/2)-1)^2-1) \\ &+ c_{1,3}\frac{\sqrt{42}}{4}(3(t/2)-1)(5(3(t/2)-1)^2-3) \\ &+ c_{1,4}\frac{3\sqrt{6}}{16}(35(3(t/2)-1)^4 - 30(3(t/2)-1)^2+3) \\ &+ c_{1,5}\frac{\sqrt{66}}{16}\left(63(3(t/2)-1)^5 - 70(3(t/2)-1)^3 + 15(3(t/2)-1)\right). \end{aligned}$$

To approximate the first, second, and third derivatives of $y(t)$, we use the 6×6 operational matrix of differentiation P in the form

$$P = \begin{pmatrix} 0 & 0 & 0 & 0 & 0 & 0 \\ 3\sqrt{3} & 0 & 0 & 0 & 0 & 0 \\ 0 & 3\sqrt{15} & 0 & 0 & 0 & 0 \\ 3\sqrt{7} & 0 & 3\sqrt{35} & 0 & 0 & 0 \\ 0 & 3\sqrt{27} & 0 & 3\sqrt{63} & 0 & 0 \\ 3\sqrt{11} & 0 & 3\sqrt{55} & 0 & 3\sqrt{99} & 0 \end{pmatrix},$$

$$P^2 = \begin{pmatrix} 0 & 0 & 0 & 0 & 0\ 0 \\ 0 & 0 & 0 & 0 & 0\ 0 \\ 9\sqrt{45} & 0 & 0 & 0 & 0\ 0 \\ 0 & 9\sqrt{15}\sqrt{35} & 0 & 0 & 0\ 0 \\ 0 & 0 & 9\sqrt{35}\sqrt{63} & 0 & 0\ 0 \\ 0 & 9\sqrt{15}\sqrt{55}+9\sqrt{27}\sqrt{99} & 0 & 9\sqrt{63}\sqrt{99} & 0\ 0 \end{pmatrix},$$

and

$$P^3 = \begin{pmatrix} 0 & 0 & 0 & 0\ 0\ 0 \\ 0 & 0 & 0 & 0\ 0\ 0 \\ 0 & 0 & 0 & 0\ 0\ 0 \\ 27\sqrt{35}\sqrt{45} & 0 & 0 & 0\ 0\ 0 \\ 0 & 27\sqrt{15}\sqrt{35}\sqrt{63} & 0 & 0\ 0\ 0 \\ 27\sqrt{45}\sqrt{55}+27\sqrt{81}\sqrt{99}+27\sqrt{7}\sqrt{63}\sqrt{99} & 0 & 27\sqrt{35}\sqrt{63}\sqrt{99} & 0\ 0\ 0 \end{pmatrix},$$

as follows:

$$
\begin{aligned}
y'(t) &= \left(c_{1,1}3\sqrt{3} + c_{1,3}3\sqrt{7} + c_{1,5}3\sqrt{11}\right)\psi_{1,0}^{\mu} + \left(c_{1,2}3\sqrt{15} + c_{1,4}3\sqrt{27}\right)\psi_{1,1}^{\mu} \\
&\quad + \left(c_{1,3}3\sqrt{35} + c_{1,5}3\sqrt{55}\right)\ \psi_{1,2}^{\mu} + c_{1,4}3\sqrt{63}\ \psi_{1,3}^{\mu} + c_{1,5}\ 3\sqrt{99}\psi_{1,4}^{\mu} \\
&= \left(c_{1,1}3\sqrt{3} + c_{1,3}3\sqrt{7} + c_{1,5}3\sqrt{11}\right)\frac{\sqrt{6}}{2} \\
&\quad + \left(c_{1,2}3\sqrt{15} + c_{1,4}3\sqrt{27}\right)\left(\frac{3\sqrt{2}}{2}(3t-1)\right) \\
&\quad + \left(c_{1,3}3\sqrt{35} + c_{1,5}3\sqrt{55}\right)\left(\frac{\sqrt{30}}{4}(3(3t-1)^2-1)\right) \\
&\quad + c_{1,4}3\sqrt{63}\left(\frac{\sqrt{42}}{4}(3t-1)(5(3t-1)^2-3)\right) \\
&\quad + c_{1,5}3\sqrt{99}\left(\frac{3\sqrt{6}}{16}(35(3t-1)^4 - 30(3t-1)^2 + 3)\right),
\end{aligned}
$$

$$
\begin{aligned}
y''(t) &= \left(c_{1,2}9\sqrt{45}\right)\psi_{1,0}^{\mu} + \left(c_{1,3}9\sqrt{35}\sqrt{15} + c_{1,5}\left(9\sqrt{55}\sqrt{15} + 9\sqrt{27}\sqrt{99}\right)\right)\psi_{1,1}^{\mu} \\
&\quad + \left(c_{1,4}9\sqrt{63}\sqrt{35}\right)\psi_{1,2}^{\mu} + c_{1,5}9\sqrt{99}\sqrt{63}\psi_{1,3}^{\mu} \\
&= \left(c_{1,2}9\sqrt{45}\right)\frac{\sqrt{6}}{2} + \left(c_{1,3}9\sqrt{35}\sqrt{15} + c_{1,5}\left(9\sqrt{55}\sqrt{15} + 9\sqrt{27}\sqrt{99}\right)\right) \\
&\quad \times \left(\frac{3\sqrt{2}}{2}(3t-1)\right) + \left(c_{1,4}9\sqrt{63}\sqrt{35}\right)\left(\frac{\sqrt{30}}{4}(3(3t-1)^2-1)\right) \\
&\quad + c_{1,5}9\sqrt{99}\sqrt{63}\left(\frac{\sqrt{42}}{4}(3t-1)(5(3t-1)^2-3)\right),
\end{aligned}
$$

$$
\begin{aligned}
y'''(t) &= \left(c_{1,3}27\sqrt{35}\sqrt{45} + c_{1,5}\left(27\sqrt{55}\sqrt{45} + 27\sqrt{81}\sqrt{99} + 27\sqrt{7}\sqrt{63}\sqrt{99}\right)\right)\psi_{1,0}^{\mu} \\
&\quad + \left(c_{1,4}27\sqrt{63}\sqrt{35}\sqrt{15}\right)\psi_{1,1}^{\mu} + \left(c_{1,5}27\sqrt{63}\sqrt{35}\sqrt{99}\right)\psi_{1,2}^{\mu} \\
&= \left(c_{1,3}27\sqrt{35}\sqrt{45} + c_{1,5}\left(27\sqrt{55}\sqrt{45} + 27\sqrt{81}\sqrt{99} + 27\sqrt{7}\sqrt{63}\sqrt{99}\right)\right)\frac{\sqrt{6}}{2} \\
&\quad + \left(c_{1,4}27\sqrt{63}\sqrt{35}\sqrt{15}\right)\left(\frac{3\sqrt{2}}{2}(3t-1)\right) \\
&\quad + \left(c_{1,5}27\sqrt{63}\sqrt{35}\sqrt{99}\right)\ \left(\frac{\sqrt{30}}{4}(3(3t-1)^2-1)\right).
\end{aligned}
$$

Using these approximations, (19) takes the form

$$\left(c_{1,3}27\sqrt{35}\sqrt{45}+c_{1,5}\left(27\sqrt{55}\sqrt{45}+27\sqrt{81}\sqrt{99}+27\sqrt{7}\sqrt{63}\sqrt{99}\right)\right)\frac{\sqrt{6}}{2}$$
$$+\left(c_{1,4}27\sqrt{63}\sqrt{35}\sqrt{15}\right)\left(\frac{3\sqrt{2}}{2}(3t-1)\right)$$
$$+\left(c_{1,5}27\sqrt{63}\sqrt{35}\sqrt{99}\right)\left(\frac{\sqrt{30}}{4}(3(3t-1)^2-1)\right)+1$$
$$-2\begin{pmatrix} c_{1,0}\frac{\sqrt{6}}{2}+c_{1,1}\frac{3\sqrt{2}}{2}(3(t/2)-1)+c_{1,2}\frac{\sqrt{30}}{4}(3(3(t/2)-1)^2-1) \\ +c_{1,3}\frac{\sqrt{42}}{4}(3(t/2)-1)(5(3(t/2)-1)^2-3) \\ +c_{1,4}\frac{3\sqrt{6}}{16}(35(3(t/2)-1)^4-30(3(t/2)-1)^2+3) \\ +c_{1,5}\frac{\sqrt{66}}{16}\left(63(3(t/2)-1)^5-70(3(t/2)-1)^3+15(3(t/2)-1)\right) \end{pmatrix}^2$$
$$=0. \tag{21}$$

Note that in order to determine the unknown coefficients $c_{1,0}$ $c_{1,1}$ $c_{1,2}$ $c_{1,3}$ $c_{1,4}$ $c_{1,5}$, we need six equations. Three equations are obtained from the initial conditions in (19) as follows:

$$c_{1,0}\frac{\sqrt{6}}{2}-c_{1,1}\frac{3\sqrt{2}}{2}+c_{1,2}\frac{2\sqrt{30}}{4}-c_{1,3}\frac{2\sqrt{42}}{4}+c_{1,4}\frac{24\sqrt{6}}{16}-c_{1,5}\frac{8\sqrt{66}}{16}=0,$$
$$\left(c_{1,1}3\sqrt{3}+c_{1,3}3\sqrt{7}+c_{1,5}3\sqrt{11}\right)\frac{\sqrt{6}}{2}-\left(c_{1,2}3\sqrt{15}+c_{1,4}3\sqrt{27}\right)\left(\frac{3\sqrt{2}}{2}\right)$$
$$+\left(c_{1,3}3\sqrt{35}+c_{1,5}3\sqrt{55}\right)\left(\frac{2\sqrt{30}}{4}\right)-c_{1,4}3\sqrt{63}\left(\frac{2\sqrt{42}}{4}\right)$$
$$+c_{1,5}3\sqrt{99}\left(\frac{24\sqrt{6}}{16}\right)=1,$$
$$\left(c_{1,2}9\sqrt{45}\right)\frac{\sqrt{6}}{2}-\left(c_{1,3}9\sqrt{35}\sqrt{15}+c_{1,5}\left(9\sqrt{55}\sqrt{15}+9\sqrt{27}\sqrt{99}\right)\right)\left(\frac{3\sqrt{2}}{2}\right)$$
$$+\left(c_{1,4}9\sqrt{63}\sqrt{35}\right)\left(\frac{2\sqrt{30}}{4}\right)-c_{1,5}9\sqrt{99}\sqrt{63}\left(\frac{2\sqrt{42}}{4}\right)=0.$$

The fourth, fifth, and sixth equations are obtained by inserting the smaller three roots of the seventh order shifted generalized Legendre polynomial, $t_1=0.02251, t_2=0.1129, t_3=0.2538$, in (21).

Solving this nonlinear 6×6 system gives

$$C_{6\times 1} = [\ c_{1,0} \quad c_{1,1} \quad c_{1,2} \quad c_{1,3} \quad c_{1,4} \quad c_{1,5}\ ]^T$$
$$= [0.26223389730166 \quad 0.14684295578193 \quad -0.00438943865859$$
$$-0.00071526649488 \quad 0.00001058352043 \quad 0.00000106184998\ ]^T.$$

Hence, the approximate solution of Example 3 using our proposed $GLWM$ is obtained as

$$y(t) = C^T\Psi$$
$$= [0.26223389730166 \quad 0.14684295578193 \quad -0.00438943865859$$
$$-0.00071526649488 \quad 0.00001058352043 \quad 0.00000106184998]^T$$
$$* \begin{bmatrix} \frac{\sqrt{6}}{2} \\ \frac{3\sqrt{2}}{2}(3t-1) \\ \frac{\sqrt{30}}{4}\ (3(3t-1)^2-1) \\ \frac{\sqrt{42}}{4}\ (3t-1)\ (5(3t-1)^2-3) \\ \frac{3\sqrt{6}}{16}\ (35(3t-1)^4-30(3t-1)^2+3) \\ \frac{\sqrt{66}}{16}\ \big(63(3t-1)^5-70(3t-1)^3+15(3t-1)\ \big) \end{bmatrix}.$$

Along with the absolute errors compared to the exact solution, we can evaluate the approximation at the locations in the prescribed interval, $0 \le t < \frac{2}{3}$, and summarized in the table (Table 5) below.

Table 5: Numerical results and the absolute error for **Example 3** for our proposed method GLWM using fifth- and sixth-order polynomials ($M = 5\ ,\ 6$)

t	Exact solution	Approximate solution $M=5$ **Present method** $GLWM$ $(M=5, \mu=3, k=1)$	Absolute Error	Approximate solution $M=6$ **Present method** $GLWM$ $(M=6, \mu=3, k=1)$	Absolute Error
0.1	0.09983341665	0.09983341651	1.369e-10	0.09983341665	1.969e-12
0.2	0.1986693308	0.198669332	1.2394e-09	0.1986693307	9.529e-11
0.3	0.2955202067	0.2955202026	4.0622e-09	0.2955202072	5.532e-10
0.4	0.3894183423	0.3894183507	8.4262e-09	0.3894183404	1.956e-9
0.5	0.4794255386	0.4794258702	3.3162e-07	0.4794255601	2.152e-8
0.6	0.5646424734	0.5646445308	2.0574e-06	0.5646427814	3.08e-7

In Table 6, absolute error comparisons between the proposed approach $GLWM$ and the $RLWM$ and other numerical methods is shown.

Example 4. Assume the following equation of the second order nonlinear differential form through proportional delay [16]:

$$y''(t) + 2\,y(t) - y^2(t) + y(t^3/8) = \sin\ t - \sin^2 t + \sin(t^3/8), \quad 0 \le t \le 1,$$
$$y(0) = 0, \qquad y'(0) = 1, \tag{22}$$

Table 6: Comparison of the absolute errors with other numerical methods

t	Absolute Error for present method $GLWM$ $(M=5)$ $(\mu=3,k=1)$	Absolute Error for present method $GLWM$ $(M=6)$ $(\mu=3,k=1)$	Absolute Error for Legendre wavelets method $RLWM$ $(M=5)$	Absolute Error for Legendre wavelets method $RLWM$ $(M=6)$	Decomposition Method $E13$ [34]	Adomian decomposition Method $E9$ [14]
0.1	1.369e-10	1.969e-12	2.54e-09	5.37e-10	0.0	1.02e-15
0.2	1.2394e-09	9.529e-11	3.24e-09	1.39e-09	0.0	5.28e-13
0.3	4.0622e-09	5.532e-10	2.11e-08	1.59e-09	0.0	2.02e-11
0.4	8.4262e-09	1.956e-9	1.44e-08	7.06e-09	0.0	2.69e-10
0.5	3.3162e-07	2.152e-8	1.21e-07	3.52e-09	2.61e-09	2.00e-09
0.6	2.0574e-06	3.08e-7	1.42e-07	3.27e-08	1.04e-08	1.03e-08

with the exact solution of the form $y(t) = \sin t$. Comparison between approximate solution and the absolute error of [16, Example 3] using our GLWM for $M = 5, 6; k = 1; \mu = 3$ is listed (in Table 7) below. Also, comparison between the absolute error for Example 4 of the present method with the RLWM of [16, Example 3] is listed in Table 8.

Table 7: Approximate solution and the absolute error of [16, Example 3] using our $GLWM$ for M=5, 6; $k = 1$; $\mu = 3$

t	Exact Solution	Approximate solution M=5; $k = 1$; $\mu = 3$	Approximate solution M=6; $k = 1$; $\mu = 3$
0.1	0.09983341665	0.09983341629	0.09983341679
0.2	0.1986693308	0.1986693294	0.1986693306
0.3	0.2955202067	0.2955202132	0.2955202065
0.4	0.3894183423	0.3894183346	0.3894183434
0.5	0.4794255386	0.4794255315	0.4794255352
0.6	0.5646424734	0.5646429917	0.5646425129

Along with the absolute errors compared to the exact solution, we can evaluate the approximation at the locations in the prescribed interval, $0 \leq t < \frac{1}{2}$, for two values of $\mu = 3, \mu = 4$ and summarized in the table (Table 9) below.

Example 5. Consider the following third order nonlinear differential equation with proportional delay [27]:

Table 8: Comparison of the absolute error for Example 4 of the present method with the RLWM of [16, Example 3]

t	**Present method** $GLWMM = 5$, $\mu = 3, k = 1$	$RLWM$ [13] $M = 6$, $\mu = 3,\ k = 1$	**RLWM [16]** $M = 5, k = 0$	**RLWM [16]** $M = 6, k = 0$
0.1	**3.562e-10**	**1.3936e-10**	8.963065387e-09	3.389353381e-10
0.2	**1.414e-9**	**1.846e-10**	2.720358344e-08	3.618011279e-09
0.3	**6.549e-9**	**1.6084e-10**	2.394278514e-08	3.060617093e-09
0.4	**7.734e-9**	**1.1379e-9**	6.937304025e-08	7.998320783e-09
0.5	**7.082e-9**	**3.3891e-9**	1.053035117e-07	7.465058682e-09
0.6	**5.183e-7**	**3.9459e-8**	1.310158346e-07	1.884611267e-08

Table 9: Comparison of the absolute error for Example 4 of our proposed method GLWM in two cases $\mu = 3$, $\mu = 4$

t	**Present method** $GLWM$ $(M = 5), \mu = 3$	**Present method** $GLWM$ $(M = 5), \mu = 4$	**Present method** $GLWM$ $(M = 6), \mu = 3$	**Present method** $GLWM$ $(M = 5), \mu = 4$
0.1	**3.562e-10**	**2.158e-10**	**1.3936e-10**	**2.9452e-11**
0.2	**1.414e-9**	**5.474e-10**	**1.846e-10**	**6.4737e-11**
0.3	**6.549e-9**	**1.051e-9**	**1.6084e-10**	**1.5656e-10**
0.4	**7.734e-9**	**7.761e-9**	**1.1379e-9**	**2.8361e-10**
0.5	**7.082e-9**	**2.619e-7**	**3.3891e-9**	**3.3032e-8**

$$\begin{aligned} &y'''(t) = -y(t) - y(t-0.3)\ + e^{(-t+0.3)}, \qquad 0 \le t \le 1, \\ &y(0) = 0\ \ ,\ \ y'(0) = -1,\ \ y''(0) = 1, \end{aligned} \tag{23}$$

with the exact solution of the form $y(t)\ = e^{-t}$.

We first apply the GLWM for $M = 9, k = 1, \mu = 3$.

For this choice of M,k,μ, the function approximation for $y(t)$ will take the summation form,

$$y(t)\ \approx\ \sum_{n=1}^{\mu^{k-1}} \sum_{m=0}^{M} c_{n\ ,\ m}\ \psi_{n,m}^{\mu}(t)\ =\ \sum_{n=1}^{1} \sum_{m=0}^{9} c_{n\ ,\ m}\ \psi_{n,m}^{\mu}\ (t)\ =\ C^T \Psi\ , \tag{24}$$

where $C_{10\times 1} = [\ c_{1,0}\ \ c_{1,1}\ \ c_{1,2}\ \ c_{1,\,3}\ \ c_{1,\,4}\ \ c_{1,5}\ \ c_{1,6}\ \ c_{1,7}\ \ c_{1,8}\ \ c_{1,9}\ ]^T$ and

$$\Psi_{10\times 1}(t) = \begin{bmatrix} \psi^{\mu}_{1,0}(t) & \psi^{\mu}_{1,1}(t) & \psi^{\mu}_{1,2}(t) & \psi^{\mu}_{1,3}(t) & \psi^{\mu}_{1,4}(t) \\ \psi^{\mu}_{1,5}(t) & \psi^{\mu}_{1,6}(t) & \psi^{\mu}_{1,7}(t) & \psi^{\mu}_{1,8}(t) & \psi^{\mu}_{1,9}(t) \end{bmatrix}^T,$$

where the generalized Legendre wavelets are $\psi^{\mu}_{1,m}(t), m = 0, 1, 2, \ldots, 9$.

Thus, $y(t)$ and $y(t-0.3)$ can be approximated as

$$\begin{aligned} y(t) =& c_{1,0}\frac{\sqrt{6}}{2} + c_{1,1}\frac{3\sqrt{2}}{2}(3t-1) + c_{1,2}\frac{\sqrt{30}}{4}(3(3t-1)^2-1) \\ &+ c_{1,3}\frac{\sqrt{42}}{4}(3t-1)(5(3t-1)^2-3) \\ &+ c_{1,4}\frac{3\sqrt{6}}{16}\,(35(3t-1)^4 - 30(3t-1)^2 + 3) \\ &+ c_{1,5}\frac{\sqrt{66}}{16}\left(63(3t-1)^5 - 70(3t-1)^3 + 15(3t-1)\right) \\ &+ c_{1,6}\frac{\sqrt{78}}{32}\left(231(3t-1)^6 - 315(3t-1)^4 + 105(3t-1)^2 - 5\right) \\ &+ c_{1,7}\frac{3\sqrt{10}}{32}\left(429(3t-1)^7 - 693(3t-1)^5 + 315(3t-1)^3 - 35(3t-1)\right) \\ &+ c_{1,8}\frac{\sqrt{102}}{256}(6435(3t-1)^8 - 12012(3t-1)^6 + 6930(3t-1)^4 \\ &- 1260(3t-1)^2 + 35) + c_{1,9}\Bigg(\frac{\sqrt{114}}{256}\left(12155(3t-1)^9 - 25740(3t-1)^7\right. \\ &\left.+18018(3t-1)^5 - 4620(3t-1)^3 + 315(3t-1)\right)\Bigg), \end{aligned}$$

$$\begin{aligned} y(t-0.3) =& c_{1,0}\frac{\sqrt{6}}{2} + c_{1,1}\frac{3\sqrt{2}}{2}(3(t-0.3)-1) + c_{1,2}\frac{\sqrt{30}}{4}(3(3(t-0.3)-1)^2-1) \\ &+ c_{1,3}\frac{\sqrt{42}}{4}(3(t-0.3)-1)(5(3(t-0.3)-1)^2-3) \\ &+ c_{1,4}\frac{3\sqrt{6}}{16}(35(3(t-0.3)-1)^4 - 30(3(t-0.3)-1)^2 + 3) \\ &+ c_{1,5}\frac{\sqrt{66}}{16}\left(63(3(t-0.3)-1)^5 - 70(3(t-0.3)-1)^3 + 15(3(t-0.3)-1)\right) \\ &+ c_{1,6}\frac{\sqrt{78}}{32}\left(231(3(t-0.3)-1)^6 - 315(3(t-0.3)-1)^4\right. \\ &\left.+105(3(t-0.3)-1)^2 - 5\right) \\ &+ c_{1,7}\frac{3\sqrt{10}}{32}\left(429(3(t-0.3)-1)^7\right. \end{aligned}$$

$$
\begin{aligned}
&-693(3(t-0.3)-1)^5+315(3(t-0.3)-1)^3-35(3(t-0.3)-1)\,\big)\\
&+c_{1,8}\frac{\sqrt{102}}{256}\big(6435(3(t-0.3)-1)^8-12012(3(t-0.3)-1)^6\\
&+6930(3(t-0.3)-1)^4-1260(3(t-0.3)-1)^2+35\big)\\
&+c_{1,9}\left(\frac{\sqrt{114}}{256}\left(\begin{array}{l}12155(3(t-0.3)-1)^9-25740(3(t-0.3)-1)^7+\\18018(3(t-0.3)-1)^5-4620(3(t-0.3)-1)^3\\+315(3(t-0.3)-1)\end{array}\right)\right),
\end{aligned}
$$

In order to approximate the first, second, and third derivatives of $y(t)$, we use the 10×10 operational matrix of differentiation P in the form

$$
P=\begin{pmatrix}
0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0\\
3\sqrt{3} & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0\\
0 & 3\sqrt{15} & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0\\
3\sqrt{7} & 0 & 3\sqrt{35} & 0 & 0 & 0 & 0 & 0 & 0 & 0\\
0 & 3\sqrt{27} & 0 & 3\sqrt{63} & 0 & 0 & 0 & 0 & 0 & 0\\
3\sqrt{11} & 0 & 3\sqrt{55} & 0 & 3\sqrt{99} & 0 & 0 & 0 & 0 & 0\\
0 & 3\sqrt{39} & 0 & 3\sqrt{91} & 0 & 3\sqrt{143} & 0 & 0 & 0 & 0\\
3\sqrt{15} & 0 & 3\sqrt{75} & 0 & 3\sqrt{135} & 0 & 3\sqrt{195} & 0 & 0 & 0\\
0 & 3\sqrt{51} & 0 & 3\sqrt{119} & 0 & 3\sqrt{187} & 0 & 3\sqrt{255} & 0 & 0\\
3\sqrt{19} & 0 & 3\sqrt{95} & 0 & 3\sqrt{171} & 0 & 3\sqrt{247} & 0 & 3\sqrt{323} & 0
\end{pmatrix},
$$

$$
P^2 = 1.0e+03
$$

$$
*\begin{pmatrix}
0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0\\
0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0\\
0.060 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0\\
0 & 0.2062 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0\\
0.2700 & 0 & 0.4226 & 0 & 0 & 0 & 0 & 0 & 0 & 0\\
0 & 0.7238 & 0 & 0.7108 & 0 & 0 & 0 & 0 & 0 & 0\\
0.6814 & 0 & 1.3061 & 0 & 1.0708 & 0 & 0 & 0 & 0 & 0\\
0 & 1.6301 & 0 & 2.0289 & 0 & 1.5029 & 0 & 0 & 0 & 0\\
1.3359 & 0 & 2.7382 & 0 & 2.8944 & 0 & 2.0069 & 0 & 0 & 0\\
0 & 2.9897 & 0 & 4.0479 & 0 & 3.9033 & 0 & 2.5829 & 0 & 0
\end{pmatrix},
$$

and

$$P^{\,3} = 1.0e+05$$

$$* \begin{pmatrix} 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0\ 0\ 0 \\ 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0\ 0\ 0 \\ 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0\ 0\ 0 \\ 0.0107 & 0 & 0 & 0 & 0 & 0 & 0 & 0\ 0\ 0 \\ 0 & 0.0491 & 0 & 0 & 0 & 0 & 0 & 0\ 0\ 0 \\ 0.0940 & 0 & 0.1261 & 0 & 0 & 0 & 0 & 0\ 0\ 0 \\ 0 & 0.3187 & 0 & 0.2550 & 0 & 0 & 0 & 0\ 0\ 0 \\ 0.3953 & 0 & 0.6945 & 0 & 0.4486 & 0 & 0 & 0\ 0\ 0 \\ 0 & 1.1453 & 0 & 1.2636 & 0 & 0.7200 & 0 & 0\ 0\ 0 \\ 1.1651 & 0 & 2.2579 & 0 & 2.0655 & 0 & 1.0821 & 0\ 0\ 0 \end{pmatrix},$$

as follows:

$$\begin{aligned}
y'(t) =& (c_{1,1}3\sqrt{3} + c_{1,3}3\sqrt{7} + c_{1,5}3\sqrt{11} + c_{1,7}3\sqrt{15} + c_{1,9}3\sqrt{19})\psi^{\mu}_{1,0} \\
&+(c_{1,2}3\sqrt{15} + c_{1,4}3\sqrt{27} + c_{1,6}\ 3\sqrt{39} + c_{1,8}3\sqrt{51})\psi^{\mu}_{1,1} \\
&+(c_{1,3}3\sqrt{35} + c_{1,5}3\sqrt{55} + c_{1,7}3\sqrt{75} + c_{1,9}3\sqrt{95})\psi^{\mu}_{1,2} \\
&+(c_{1,4}3\sqrt{63} + c_{1,6}3\sqrt{91} + c_{1,8}3\sqrt{119})\psi^{\mu}_{1,3} \\
&+(c_{1,5}\ 3\sqrt{99} + c_{1,7}3\sqrt{135} + c_{1,9}3\sqrt{171})\psi^{\mu}_{1,4} \\
&+(c_{1,6}3\sqrt{143} + c_{1,8}3\sqrt{187})\psi^{\mu}_{1,5} + (c_{1,7}3\sqrt{195} + c_{1,9}3\sqrt{247})\psi^{\mu}_{1,6} \\
&+(c_{1,8}3\sqrt{255})\psi^{\mu}_{1,7} + (c_{1,9}3\sqrt{323})\psi^{\mu}_{1,8} \\
=& (c_{1,1}3\sqrt{3} + c_{1,3}3\sqrt{7} + c_{1,5}3\sqrt{11} + c_{1,7}\ 3\sqrt{15} + c_{1,9}3\sqrt{19})\frac{\sqrt{6}}{2} \\
&+(c_{1,2}\ 3\sqrt{15} + c_{1,4}3\sqrt{27} + c_{1,6}\ 3\sqrt{39} + c_{1,8}3\sqrt{51})(\frac{3\sqrt{2}}{2}(3t-1)) \\
&+(c_{1,3}3\sqrt{35} + c_{1,5}3\sqrt{55} + c_{1,7}\ 3\sqrt{75} + c_{1,9}3\sqrt{95})(\frac{\sqrt{30}}{4}(3(3t-1)^2-1)) \\
&+(c_{1,4}3\sqrt{63} + c_{1,6}3\sqrt{91} + c_{1,8}3\sqrt{119})\ (\frac{\sqrt{42}}{4}(3t-1)(5(3t-1)^2-3)) \\
&+\left(c_{1,5}3\sqrt{99} + c_{1,7}3\sqrt{135} + c_{1,9}3\sqrt{171}\right)\left(\frac{3\sqrt{6}}{16}(35(3t-1)^4 - 30(3t-1)^2 + 3)\right) \\
&+\left(c_{1,6}3\sqrt{143} + c_{1,8}3\sqrt{187}\right)\frac{\sqrt{66}}{16}\left(63(3t-1)^5 - 70(3t-1)^3 + 15(3t-1)\right) \\
&+\left(c_{1,7}3\sqrt{195} + c_{1,9}3\sqrt{247}\right)\frac{\sqrt{78}}{32}\ \left(231(3t-1)^6 - 315(3t-1)^4 + 105(3t-1)^2 - 5\right) \\
&+\left(c_{1,8}3\sqrt{255}\right)\frac{3\sqrt{10}}{32}\left(429(3t-1)^7 - 693(3t-1)^5 + 315(3t-1)^3 - 35(3t-1)\right) \\
&+\left(c_{1,9}3\sqrt{323}\right)\frac{\sqrt{102}}{256}\left(6435(3t-1)^8 - 12012(3t-1)^6\right. \\
&\left.+6930(3t-1)^4 - 1260(3t-1)^2 + 35\right)
\end{aligned}$$

$$
\begin{aligned}
y''(t) = &\left(c_{1,2}27\sqrt{5} + c_{1,4}270 + c_{1,6}189\sqrt{13} + c_{1,8}324\sqrt{17}\right)\frac{\sqrt{6}}{2} \\
&+ \left(c_{1,3}45\sqrt{21} + c_{1,5}126\sqrt{33} + c_{1,7}243\sqrt{45} + c_{1,9}396\sqrt{57}\right)\left(\frac{3\sqrt{2}}{2}(3t-1)\right) \\
&+ \left(c_{1,4}189\sqrt{5} + c_{1,6}162\sqrt{65} + c_{1,8}297\sqrt{85}\right)\left(\frac{\sqrt{30}}{4}(3(3t-1)^2-1)\right) \\
&+ \left(c_{1,5}81\sqrt{77} + c_{1,7}198\sqrt{105} + c_{1,9}351\sqrt{133}\right)\left(\frac{\sqrt{42}}{4}(3t-1)(5(3t-1)^2-3)\right) \\
&+ \left(c_{1,6}99\sqrt{117} + c_{1,8}234\sqrt{153}\right)\left(\frac{3\sqrt{6}}{16}(35(3t-1)^4 - 30(3t-1)^2 + 3)\right) \\
&+ \left(c_{1,7}117\sqrt{165} + c_{1,9}270\sqrt{209}\right)\frac{\sqrt{66}}{16}\left(63(3t-1)^5 - 70(3t-1)^3 + 15(3t-1)\right) \\
&+ \left(c_{1,8}135\sqrt{221}\right)\frac{\sqrt{78}}{32}\left(231(3t-1)^6 - 315(3t-1)^4 + 105(3t-1)^2 - 5\right) \\
&+ \left(c_{1,9}153\sqrt{285}\right)\frac{3\sqrt{10}}{32}\left(429(3t-1)^7 - 693(3t-1)^5 + 315(3t-1)^3 - 35(3t-1)\right)
\end{aligned}
$$

$$
\begin{aligned}
y'''(t) = &1.0e+05 \\
&* \begin{pmatrix}
(0.0107c_{1,3} + 0.0940c_{1,5} + 0.3953c_{1,7} + 1.1651c_{1,9})\frac{\sqrt{6}}{2} \\
+ (0.0491c_{1,4} + 0.3187c_{1,6} + 1.1453c_{1,8})\left(\frac{3\sqrt{2}}{2}(3t-1)\right) \\
+ (0.1261c_{1,5} + 0.6945c_{1,7} + 2.2579c_{1,9})\left(\frac{\sqrt{30}}{4}(3(3t-1)^2-1)\right) \\
+ (0.2550c_{1,6} + 1.2636c_{1,8})\left(\frac{\sqrt{42}}{4}(3t-1)(5(3t-1)^2-3)\right) \\
+ (0.4486c_{1,7} + 2.0655c_{1,9})\left(\frac{3\sqrt{6}}{16}(35(3t-1)^4 - 30(3t-1)^2 + 3)\right) \\
+ (0.7200c_{1,8})\frac{\sqrt{66}}{16}\left(63(3t-1)^5 - 70(3t-1)^3 + 15(3t-1)\right) \\
+ (1.0821c_{1,9})\frac{\sqrt{78}}{32}\left(231(3t-1)^6 - 315(3t-1)^4 + 105(3t-1)^2 - 5\right)
\end{pmatrix}
\end{aligned}
$$

Using these approximations, (24) takes the form

$$
\begin{aligned}
&1.0e+05 \\
&* \begin{pmatrix}
(0.0107c_{1,3} + 0.0940c_{1,5} + 0.3953c_{1,7} + 1.1651c_{1,9})\frac{\sqrt{6}}{2} \\
+ (0.0491c_{1,4} + 0.3187c_{1,6} + 1.1453c_{1,8})\left(\frac{3\sqrt{2}}{2}(3t-1)\right) \\
+ (0.1261c_{1,5} + 0.6945c_{1,7} + 2.2579c_{1,9})\left(\frac{\sqrt{30}}{4}(3(3t-1)^2-1)\right) \\
+ (0.2550c_{1,6} + 1.2636c_{1,8})\left(\frac{\sqrt{42}}{4}(3t-1)(5(3t-1)^2-3)\right) \\
+ (0.4486c_{1,7} + 2.0655c_{1,9})\left(\frac{3\sqrt{6}}{16}(35(3t-1)^4 - 30(3t-1)^2 + 3)\right) \\
+ (0.7200c_{1,8})\frac{\sqrt{66}}{16}\left(63(3t-1)^5 - 70(3t-1)^3 + 15(3t-1)\right) \\
+ (1.0821c_{1,9})\frac{\sqrt{78}}{32}\left(231(3t-1)^6 - 315(3t-1)^4 + 105(3t-1)^2 - 5\right)
\end{pmatrix}
\end{aligned}
$$

$$
\begin{aligned}
= -\Bigg(& c_{1,0}\frac{\sqrt{6}}{2} + c_{1,1}\frac{3\sqrt{2}}{2}(3t-1) + c_{1,2}\frac{\sqrt{30}}{4}(3(3t-1)^2-1) \\
& + c_{1,3}\frac{\sqrt{42}}{4}(3t-1)(5(3t-1)^2-3) \\
& + c_{1,4}\frac{3\sqrt{6}}{16}(35(3t-1)^4 - 30(3t-1)^2 + 3) \\
& + c_{1,5}\frac{\sqrt{66}}{16}\left(63(3t-1)^5 - 70(3t-1)^3 + 15(3t-1)\right) \\
& + c_{1,6}\frac{\sqrt{78}}{32}\left(231(3t-1)^6 - 315(3t-1)^4 + 105(3t-1)^2 - 5\right) \\
& + c_{1,7}\frac{3\sqrt{10}}{32}\left(429(3t-1)^7 - 693(3t-1)^5 + 315(3t-1)^3 - 35(3t-1)\right) \\
& + c_{1,8}\frac{\sqrt{102}}{256}\left(6435(3t-1)^8 - 12012(3t-1)^6 + 6930(3t-1)^4\right. \\
& \left. -1260(3t-1)^2 + 35\right) + c_{1,9}\Bigg(\frac{\sqrt{114}}{256}\left(12155(3t-1)^9 - 25740(3t-1)^7\right. \\
& \left. +18018(3t-1)^5 - 4620(3t-1)^3 + 315(3t-1)\right)\Bigg) \Bigg) \\
- \Bigg(& c_{1,0}\frac{\sqrt{6}}{2} + c_{1,1}\frac{3\sqrt{2}}{2}(3(t-0.3)-1) + c_{1,2}\frac{\sqrt{30}}{4}(3(3(t-0.3)-1)^2-1) \\
& + c_{1,3}\frac{\sqrt{42}}{4}(3(t-0.3)-1)(5(3(t-0.3)-1)^2-3) \\
& + c_{1,4}\frac{3\sqrt{6}}{16}(35(3(t-0.3)-1)^4 - 30(3(t-0.3)-1)^2 + 3) \\
& + c_{1,5}\frac{\sqrt{66}}{16}\left(63(3(t-0.3)-1)^5 - 70(3(t-0.3)-1)^3 + 15(3(t-0.3)-1)\right) \\
& + c_{1,6}\frac{\sqrt{78}}{32}\left(231(3(t-0.3)-1)^6 - 315(3(t-0.3)-1)^4\right. \\
& \left. +105(3(t-0.3)-1)^2 - 5\right) + c_{1,7}\frac{3\sqrt{10}}{32}\left(429(3(t-0.3)-1)^7\right. \\
& \left. -693(3(t-0.3)-1)^5 + 315(3(t-0.3)-1)^3 - 35(3(t-0.3)-1)\right) \\
& + c_{1,8}\frac{\sqrt{102}}{256}\left(6435(3(t-0.3)-1)^8 - 12012(3(t-0.3)-1)^6\right. \\
& \left. +6930(3(t-0.3)-1)^4 - 1260(3(t-0.3)-1)^2 + 35\right) \\
& + c_{1,9}\left(\frac{\sqrt{114}}{256}\begin{pmatrix} 12155(3(t-0.3)-1)^9 - 25740(3(t-0.3)-1)^7 + \\ 18018(3(t-0.3)-1)^5 - 4620(3(t-0.3)-1)^3 + \\ 315(3(t-0.3)-1) \end{pmatrix}\right) \Bigg) \\
& + e^{(-t+0.3)}, \qquad (25)
\end{aligned}
$$

Note that to determine the unknown coefficients $c_{1,0},\ c_{1,1},\ c_{1,2},\ c_{1,\,3},\ c_{1,\,4},\ c_{1,5},\ c_{1,6},\ c_{1,7},\ c_{1,8},\ c_{1,9}$, we need ten equations.

Three equations are obtained using the initial conditions in (23) as follows:

$$c_{1,0}\tfrac{\sqrt{6}}{2} - c_{1,1}\tfrac{3\sqrt{2}}{2} + c_{1,2}\tfrac{2\sqrt{30}}{4} - c_{1,3}\tfrac{2\sqrt{42}}{4} + c_{1,4}\tfrac{24\sqrt{6}}{16} - c_{1,5}\tfrac{8\sqrt{66}}{16}$$
$$+c_{1,6}\tfrac{26\sqrt{78}}{32} - c_{1,7}\tfrac{48\sqrt{10}}{32} + c_{1,8}\tfrac{128\sqrt{102}}{256} - c_{1,9}\tfrac{128\sqrt{114}}{256} = 0,$$

$$\left(c_{1,1}3\sqrt{3} + c_{1,3}3\sqrt{7} + c_{1,5}3\sqrt{11} + c_{1,7}3\sqrt{15} + c_{1,9}3\sqrt{19}\right)\tfrac{\sqrt{6}}{2}$$
$$-\left(c_{1,2}3\sqrt{15} + c_{1,4}3\sqrt{27} + c_{1,6}3\sqrt{39} + c_{1,8}3\sqrt{51}\right)\left(\tfrac{3\sqrt{2}}{2}\right)$$
$$+\left(c_{1,3}3\sqrt{35} + c_{1,5}3\sqrt{55} + c_{1,7}3\sqrt{75} + c_{1,9}3\sqrt{95}\right)\left(\tfrac{2\sqrt{30}}{4}\right)$$
$$-\left(c_{1,4}3\sqrt{63} + c_{1,6}3\sqrt{91} + c_{1,8}3\sqrt{119}\right)\left(\tfrac{2\sqrt{42}}{4}\right)$$
$$+\left(c_{1,5}3\sqrt{99} + c_{1,7}3\sqrt{135} + c_{1,9}3\sqrt{171}\right)\left(\tfrac{24\sqrt{6}}{16}\right)$$
$$-\left(c_{1,6}3\sqrt{143} + c_{1,8}3\sqrt{187}\right)\tfrac{8\sqrt{66}}{16} + \left(c_{1,7}3\sqrt{195} + c_{1,9}3\sqrt{247}\right)\tfrac{16\sqrt{78}}{32}$$
$$-\left(c_{1,8}3\sqrt{255}\right)\tfrac{48\sqrt{10}}{32} + \left(c_{1,9}3\sqrt{323}\right)\tfrac{128\sqrt{102}}{256} = -1,$$

$$\left(c_{1,2}27\sqrt{5} + c_{1,4}270 + c_{1,6}189\sqrt{13} + c_{1,8}324\sqrt{17}\right)\tfrac{\sqrt{6}}{2}$$
$$-\left(c_{1,3}45\sqrt{21} + c_{1,5}126\sqrt{33} + c_{1,7}243\sqrt{45} + c_{1,9}396\sqrt{57}\right)\left(\tfrac{3\sqrt{2}}{2}\right)$$
$$+\left(c_{1,4}189\sqrt{5} + c_{1,6}162\sqrt{65} + c_{1,8}297\sqrt{85}\right)\left(\tfrac{2\sqrt{30}}{4}\right)$$
$$-\left(c_{1,5}81\sqrt{77} + c_{1,7}198\sqrt{105} + c_{1,9}351\sqrt{133}\right)\left(\tfrac{2\sqrt{42}}{4}\right)$$
$$+\left(c_{1,6}99\sqrt{117} + c_{1,8}234\sqrt{153}\right)\left(\tfrac{24\sqrt{6}}{16}\right) - \left(c_{1,7}117\sqrt{165} + c_{1,9}270\sqrt{209}\right)\tfrac{8\sqrt{66}}{16}$$
$$+\left(c_{1,8}135\sqrt{221}\right)\tfrac{16\sqrt{78}}{32} - \left(c_{1,9}153\sqrt{285}\right)\tfrac{48\sqrt{10}}{32} = 1.$$

The reminder seven equations are obtained by inserting the smaller three roots of the 11th-order shifted Legendre polynomial, $t_1 = 0.008698$,$t_2 = 0.04498$, $t_3 = 0.1069, t_4 = 0.1889$,$t_5 = 0.2837$,$t_6 = 0.383$,$t_7 = 0.4778$, in (25).

Solving this nonlinear 10×10 system gives

$$C_{10\times 1} = [\ c_{1,0}\quad c_{1,1}\quad c_{1,2}\ \ c_{1,\,3}\quad c_{1,\,4}\quad c_{1,5}\ \ c_{1,6}\ \ c_{1,7}\ \ c_{1,8}\ \ c_{1,9}\ ]^T$$
$$= [\ 0.59593988797141\quad -0.113848030223\quad 0.00976752484854$$
$$-0.00054936985920\quad 0.00002304551981\quad -0.00000077145558$$
$$0.00000002149144\ \ -0.00000000051302\ \ 0.00000000001061$$
$$0.00000000000022\ \ ]^T$$

Hence, the approximate solution of Example 3 using our proposed $GLWM$ is obtained as

$$
\begin{aligned}
y(t) &= C^T\Psi \\
&= [\,0.59593988797141 \quad -0.113848030223 \quad 0.00976752484854 \\
&\quad -0.00054936985920 \quad 0.00002304551981 \quad -0.00000077145558 \\
&\quad 0.00000002149144 \quad -0.00000000051302 \quad 0.00000000001061 \quad 0.00000000000022\,]^T \\
&* \begin{bmatrix}
\frac{\sqrt{6}}{2} \\
\frac{3\sqrt{2}}{2}(3t-1) \\
\frac{\sqrt{30}}{4}\,(3(3t-1)^2-1) \\
\frac{\sqrt{42}}{4}\,(3t-1)\,(5(3t-1)^2-3) \\
\frac{3\sqrt{6}}{16}\,(35(3t-1)^4-30(3t-1)^2+3) \\
\frac{\sqrt{66}}{16}\left(63(3t-1)^5-70(3t-1)^3+15(3t-1)\right) \\
\frac{\sqrt{78}}{32}\left(231(3t-1)^6-315(3t-1)^4+105(3t-1)^2-5\right) \\
\frac{3\sqrt{10}}{32}\left(429(3t-1)^7-693(3t-1)^5+315(3t-1)^3-35(3t-1)\right) \\
\frac{\sqrt{102}}{256}\left(6435(3t-1)^8-12012(3t-1)^6+6930(3t-1)^4-1260(3t-1)^2+35\right) \\
\frac{\sqrt{114}}{256}\left(12155(3t-1)^9-25740(3t-1)^7+18018(3t-1)^5-4620(3t-1)^3+315(3t-1)\right)
\end{bmatrix}.
\end{aligned}
$$

Along with the absolute errors compared to the exact solution, we can evaluate the approximation at the locations in the prescribed interval, $0 \le t < \frac{2}{3}$ and summarized in the table (Table 10) below.

Table 10: Approximate solutions of [27, Example 3] using the $RLWM$ and $GLWM$ for $M = 9$

t	**Exact Solution**	**Approximate solution of** $RLWM$ $M = 9; k = 0$	**Approximate solution of** $GLWM$ $M = 9; k = 1; \mu = 3$
0.1	0.9048374180359596	0.9048374180282546	0.9048374180356493
0.2	0.8187307530779818	0.8187307530802117	0.8187307530782875
0.3	0.7408182206817179	0.740818220690352	0.7408182206816675
0.4	0.6703200460356393	0.6703200460269125	0.6703200460363143
0.5	0.6065306597126334	0.6065306597153067	0.6065306597129507
0.6	0.5488116360940264	0.5488116361078827	0.5488116360935682

In Table 11, absolute error comparisons for [27, Example 3] of the present method with the RLWM, Hermite Polynomial Collocation Method, H-CLSM, H-DLSM, Chebyshev Polynomial Collocation Method, C-CLSM and C-DLSM are as follows:

Table 11: Comparison of the absolute error for [27, Example 3] of the present method with the RLWM, Hermite Polynomial Collocation Method, H-CLSM, H-DLSM, Chebyshev Polynomial Collocation Method, C-CLSM and C-DLSM.

t	Absolute error of $GLWM$ $M=9$, $\mu=3$, $k=1$	Absolute error of $RLWM$ $M=9$, $k=0$	Absolute error of Hermite Polynomial Collocation Method	Absolute error of H-CLSM	Absolute error of H-DLSM	Absolute error of Chebyshev Polynomial Collocation Method	Absolute error of C-CLSM	Absolute error of C-DLSM
0.2	3.06e-13	2.23e-12	6.20e-09	3.38e-10	1.38e-12	3.70e-07	3.05e-09	3.53e-12
0.4	6.75e-13	8.73e-12	5.76e-08	4.85e-09	7.33e-12	2.38e-06	9.42e-09	5.78e-11
0.6	4.58e-13	1.39e-11	1.79e-07	1.07e-08	1.77e-11	5.97e-06	2.68e-08	1.78e-10

5 Conclusion

As demonstrated in this study, the current method produces more accurate findings than the other methods, especially the regular Legendre wavelets method. This method has a substantially lower maximum absolute error than the other numerical and semi-analytical ones for, simply solving, the delay and neutral differential equations with proportion at delays of different orders using our suggested $GLWM$, as demonstrated in this paper. We hope to see the same accuracy in the author's future research of fractional differential equations based on the accurate results derived from these polynomials in this work.

Acknowledgment

We express our sincere thanks to the anonymous referees for valuable suggestions that improved the final manuscript.

References

[1] Aboodh, K.S., Farah, R.A., Almardy, I.A. and Osman, A.K. *Solving delay differential equations by Aboodh transformation method*, International

Journal of Applied Mathematics & Statistical Sciences, 7(2) (2018), 55–64.

[2] Ali, I., Brunner, H. and Tang, T. *A spectral method for pantograoh-type delay differential equations and its convergence analysis*, J. Comput. Math. 27(2-39) (2009), 254–265.

[3] Amer, H. and Olorode, O. *Numerical evaluation of a novel Slot-Drill Enhanced Oil Recovery Technology for Tight Rocks*, SPE J. 27(4) (2022), 2294–2317.

[4] Balaji, S. *Legendre wavelet operational matrix method for solution of fractional order Riccati differential equation*, J. Egypt. Math. Soc., 23 (2) (2015), 263–270.

[5] Benhammouda, B., Leal, H.V. and Martinez, L.H. *Procedure for exact solutions of nonlinear Pantograph delay differential equations*, British Journal of Mathematics and Computer Science, 4(19) (2014), 2738–2751.

[6] Bhrawy, A.H., Assas, L.M., Tohidi, E. and A. Alghamdi, M. *Legendre–Gauss collocation method for neutral functional-differential equations with proportional delays*, Adv. Differ. Eq., 63, (2013) 1–16.

[7] Biazar, J. and Ghanbari, B. *The homotopy perturbation method for solving neutral functional-differential equations with proportional delays*, J. King Saud Univ. Sci., 24 (2012), 33–37.

[8] Blanco-Cocom, L., Estrella, A.G. and Avila-Vales, E. *Solving delaydifferential systems with history functions by the Adomian decomposition method*, Appl. Math. Comput. 218 (2012), 5994–6011.

[9] Bocharova, G.A. and Rihanb, F.A. *Numerical modeling in biosciences using delay differential equations*, J. Comput. Appl. Math. 125 (2000), 183–199.

[10] Căruntu, B. and Bota, C. *Analytical approximate solutions for a general class of nonlinear delay differential equations*, Sci. World J. (2014), 631416.

[11] Chen, X. and Wang, L. *The variational iteration method for solving a neutral functional- differential equation with proportional delays*, Comput. Math. Appl., 59 (2010), 2696–2702.

[12] Davaeifar, S. and Rashidinia, J. *Solution of a system of delay differential equations of multipantograph type*, J. Taibah Univ. Sci. 11 (2017), 1141–1157.

[13] El-Shazly, N.M., Ramadan, M.A. and Radwan, T. *Generalized Legendre wavelets,definition, properties and their applications for solving linear differential equations*, Egyptian Journal of Pure and Applied Science, 62(1) (2024), 20–32.

[14] Evans, D.J. and Raslan, K.R. *The Adomian decomposition method for solving delay differential equations*, Int. J. Comput. Math., 82(1) (2005), 49–54.

[15] Gu, J.S. and Jiang, W.S. *The Haar wavelets operational matrix of integration*, Int. J. Syst. Sci., 27 (7) (1996), 623–628.

[16] Gümgüm, S., Özdek, D.E. and Özaltun, G. *Legendre wavelet solution of high order nonlinear ordinary delay differential equations*, Turk. J. Math. 43 (2019), 1339–1352.

[17] Gümgüm, S., Özdek, D., Özaltun, E.G. and Bildik, N. *Legendre wavelet solution of neutral differential equations with proportional delays*,J. Appl. Math. Comput. 61 (2019), 389–404.

[18] Ha, P. *Analysis and numerical solutions of delay differential-algebraic equations*, Ph. D, Technical University Of Berlin, Berlin, Germany, 2015.

[19] Khader, M.M. *Numerical and theoretical treatment for solving linear and nonlinear delay differential equations using variational iteration method*, Arab Journal of Mathematical Sciences, 19(2) (2013), 243–256.

[20] Lv, X. and Gao, Y. *The RKHSM for solving neutral functional-differential equations with proportional delays*, Math. Methods Appl. Sci., 36 (2013), 642–649.

[21] Martin, J.A. and Garcia, O. *Variable multistep methods for delay differential equations*, Math. Comput. Model. 35(2002), 241–257.

[22] Mirzaee, F. and Latifi, L. *Numerical solution of delay differential equations by differential transform method*, Journal of Sciences (Islamic Azad University), 20(78/2) (2011), 83–88.

[23] Mohammadi, F. and Hosseini, M.M. *A new Legendre wavelet operational matrix of derivative and its applications in solving the singular ordinary differential equations*, J. Frankl. Inst., 348 (2011), 1787–1796.

[24] Nisar, K.S., Ilhan, O. A., Manafian, J., Shahriari, M. and Soybaş, D. *Analytical behavior of the fractional Bogoyavlenskii equations with conformable derivative using two distinct reliable methods*, Results i Phys. 22 (2021), 103975.

[25] Oberle, H.J. and Pesch, H.J. *Numerical treatment of delay differential equations by Hermite interpolation*, Numer. Math. 37 (1981), 235–255.

[26] Ogunfiditimi, F.O. *Numerical solution of delay differential equations using the Adomian decomposition method*, nt. J. Eng. Sci. 4(5)(2015), 18–23.

[27] Pushpam, A.E.K. and Kayelvizhi, C. *Solving delay differential equations using least square method based on successive integration technique*, Mathematical Statistician and Engineering Applications, 72(1) (2023), 1104–1115.

[28] Ravi-Kanth, A.S.V. and Kumar, P.M.M. *A numerical technique for solving nonlinear singularly perturbed delay differential equations*, Math. Model. Anal. 23(1) (2018), 64–78.

[29] Sakar, M.G. *Numerical solution of neutral functional-differential equations with proportional delays*, Int. J. Optim. Control Theor. Appl., 7(2) (2017), 186–194.

[30] Sedaghat, S., Ordokhani, Y. and Dehghan, M. *Numerical solution of delay differential equations of pantograph type via Chebyshev polynomials*, Commun. Nonlinear Sci. Numer. Simul., 17 (2012), 4815–4830.

[31] Shakeri, F. and Dehghan, M. *Solution of delay differential equations via a homotopy perturbation method*, Math. Comput. Model. 48(2008), 486–498.

[32] Shiralashetti, S.C., Hoogarand, B.S. and Kumbinarasaiah, S. *Hermite wavelet based‘ method for the numerical solution of linear and nonlinear delay differential equations*, International Journal of Engineering Science and Mathematics, 6(8) (2017), 71–79.

[33] Stephen, A. G. and Kuang, Y. *A delay reaction-diffusion model of the spread of bacteriophage infection*, Society for Industrial and Applied Mathematics, 65(2) (2005), 550–566,

[34] Taiwo, O.A. and Odetunde, O.S. *On the numerical approximation of delay differential equations by a decomposition method* , Asian Journal of Mathematics & Statistics, 3(4)(2010), 237–243.

[35] Vanani, S.K. and Aminataei, A. *On the numerical solution of nonlinear delay differential equations*, Journal of Concrete and Applicable Mathematics, 8(4)(2010), 568–576.

[36] Wang, W. and Li, S. *On the one-leg-methods for solving nonlinear neutral functional differential equations*, Appl. Math. Comput., 193(1) (2007), 285–301.

[37] Yousefi, S.A. *Legendre scaling function for solving generalized Emden–Fowler equations*, Int. J. Inf. Syst. Sci. 3 (2007), 243–250.

[38] Yüzbaşı, Ş. *A numerical approach for solving a class of the nonlinear Lane-Emden type equations arising in astrophysics*, Math. Method. Appl. Sci. 34 (2011), 2218–2230.

[39] Yüzbaşı, Ş. *An efficient algorithm for solving multi-pantograph equation system*, Comput. Math. Appl. 64 (2012), 589–603.

[40] 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 (2012), 1691–1705.

[41] Yüzbaşı, Ş. *Shifted Legendre method with residual error estimation for delay linear Fredholm integro-differential equations*, J. Taibah Uni. Sci. 11(2)(2017), 344–352.

[42] Yüzbaşı, Ş. *A numerical scheme for solutions of a class of nonlinear differential equations*, J. Taibah Uni. Sci. 11 (2017), 1165–1181.

[43] Yüzbaşı, Ş. and Şahin, N. *On the solutions of a class of nonlinear ordinary differential equations by the Bessel polynomials*, J. Numer. Math. 20(1) (2012), 55–79.

[44] Yüzbaşı, Ş. and Sezer, M. *Shifted Legendre approximation with the residual correction to solve pantograph-delay type differential equations*, Appl. Math. Model. 39 (2015), 6529–6542.

[45] Zhang, M., Xie, X., Manafian, J., Ilhan, O.A. and Singh, G. *Characteristics of the new multiple rogue wave solutions to the fractional generalized CBS-BK equation*, J. Adv. Res. 38 (2022), 131–142.

Iranian Journal of Numerical Analysis and Optimization
Vol. 14, No. 3, 2024, pp 938–969
https://doi.org/10.22067/ijnao.2024.87481.1418
https://ijnao.um.ac.ir/

Research Article

Extending quasi-GMRES method to solve generalized Sylvester tensor equations via the Einstein product

M.M. Izadkhah*,

Abstract

This paper aims to extend a Krylov subspace technique based on an incomplete orthogonalization of Krylov tensors (as a multidimensional extension of the common Krylov vectors) to solve generalized Sylvester tensor equations via the Einstein product. First, we obtain the tensor form of the quasi-GMRES method, and then we lead to the direct variant of the proposed algorithm. This approach has the great advantage that it uses previous data in each iteration and has a low computational cost. Moreover, an upper bound for the residual norm of the approximate solution is found. Finally, several experimental problems are given to show the acceptable accuracy and efficiency of the presented method.

AMS subject classifications (2020): 15A69,6 5F08, 65F10.

*Corresponding author

Received 04 Aprail 2024; revised 01 June 2024; accepted 16 June 2024

Mohammad Mahdi Izadkhah
Department of Computer Science, Faculty of Computer and Industrial Engineering, Birjand University of Technology, Birjand, Iran. e-mail: izadkhah@birjandut.ac.ir

How to cite this article
Izadkhah, M.M., Extending quasi-GMRES method to solve generalized Sylvester tensor equations via the Einstein product. *Iran. J. Numer. Anal. Optim.*, 2024; 14(3): 938-969. https://doi.org/10.22067/ijnao.2024.87481.1418

Keywords: Generalized Sylvester tensor equations; Einstein product; Quasi-GMRES method; Convergence analysis.

1 Introduction

As a common notation in the research literature, tensors are written in calligraphic font, for example, $\mathcal{A}$. For a positive integer N, an Nth order tensor (in some literature N-mode tensor, e.g., [6]) $\mathcal{A} = (a_{i_1 \cdots i_N})(1 \leq i_j \leq I_j, j = 1, \ldots, N)$ is a multidimensional N-way array with I $(I = I_1 I_2 \cdots I_N)$ entries [25]. Let $\mathbb{R}^{I_1 \times \cdots \times I_N}$ be the set of Nth order tensors of size $I_1 \times \cdots \times I_N$ over the real field $\mathbb{R}$. The tensor $\mathcal{O} \in \mathbb{R}^{I_1 \times \cdots \times I_N}$ with all entries zero denotes the zero tensor.

In this paper, we suggest an efficient iterative method to solve the generalized Sylvester tensor equation

$$\mathcal{A} \star_N \mathcal{X} \star_M \mathcal{B} + \mathcal{C} \star_N \mathcal{X} \star_M \mathcal{D} = \mathcal{F}, \tag{1}$$

where $\mathcal{A}, \mathcal{C} \in \mathbb{R}^{I_1 \times \cdots \times I_N \times I_1 \times \cdots \times I_N}$, $\mathcal{B}, \mathcal{D} \in \mathbb{R}^{K_1 \times \cdots \times K_M \times K_1 \times \cdots \times K_M}$, $\mathcal{F} \in \mathbb{R}^{I_1 \times \cdots \times I_N \times K_1 \times \cdots \times K_M}$ are known tensors, and $\mathcal{X} \in \mathbb{R}^{I_1 \times \cdots \times I_N \times K_1 \times \cdots \times K_M}$ is an unknown tenor to be determined. We denote the Einstein product by $\star_N$, which will be described in detail in Section 2.

Tensor equations arise from various fields of science and engineering multidimensional applications, including signal processing, data mining, thermal radiation, information retrieval, and three-dimensional (3D) microscopic heat transfer problems in heat transfer, and so many other modern applications in machine learning [31, 32, 33, 34, 39, 44].

Tensor equations involving the Einstein product have been studied in [7, 15, 38], which have many applications in continuum physics, engineering, isotropic, and anisotropic elastic models [26]. For example, Wang and Xu [41] introduced some iterative methods for solving different types of these tensor equations. Huang, Xie, and Ma [23] proposed the Krylov subspace methods to solve a class of tensor equations via the Einstein product. Huang and Ma [22] presented an iterative algorithm to solve the generalized Sylvester tensor equation. In [21], they also presented the global least squares methods based

on tensor form to solve the tensor equation (1). Liang, Zheng, and Zhao [30] discussed the tensor inversion and its applications for solving the tensor equations via the Einstein product.

The high order Sylvester tensor equation via the Tucker product of tensors is as follows:

$$\mathcal{X} \times_1 A_1 + \mathcal{X} \times_2 A_2 + \cdots + \mathcal{X} \times_N A_N = \mathcal{D}, \tag{2}$$

where $A_j \in \mathbb{R}^{I_j \times I_j}$, $j = 1, 2, \ldots, N$, $\mathcal{D} \in \mathbb{R}^{I_1 \times \cdots \times I_N}$ are known, and $\mathcal{X} \in \mathbb{R}^{I_1 \times \cdots \times I_N}$ is unknown. The product $\times_k$ will be defined in the next Section.

Recently, Li, Wang, and Zhang [29] proposed a modified conjugate residual method to solve the generalized coupled variant of (2), and Dehdezi and Karimi [11] extended the conjugate gradient squared method and the conjugate residual squared method to obtain their iterative solutions. Zhang, Ding, and Li [45] mainly focused on proposing the tensor form of the generalized product-type biconjugate gradient method to solve the generalized Sylvester quaternion tensor equations (2). Heyouni, Movahed, and Tajaddini [20] used the Hessenberg process instead of the Arnoldi process to generate a basis of the Krylov subspace and then proposed an iterative method to solve the real tensor equation. In addition, Zhang and Wang [44] introduced the CGNR and CGNE methods for the third-order Sylvester tensor equation (2).

Let us contemplate the following partial differential equation (see, e.g., [3, 21]):

$$\begin{cases} -\Delta u + c^T \nabla u = f, & \text{in } \Omega = [0,1]^N, \\ u = 0, & \text{on } \partial\Omega. \end{cases}$$

The use of the finite-difference discretization together with a second-order convergent scheme for the convection term leads us to a linear system that is expressed in the form (2). Chen and Lu [9] established the projection method to solve the tensor equation (2). They also applied the Kronecker product preconditioner to accelerate the convergence of the iterative method. Later, Beik, Movahed, and Ahmadi-Asl [6] derived the Krylov subspace methods to solve the Sylvester tensor equation (2) in the case of 3-mode tensors. Shi, Wei, and Ling [37] investigated the backward error and perturbation bounds for the tensor equation (2) for the 3-mode tensors.

The high order Sylvester tensor equation, which uses the Einstein product, is defined in [38] and is given by

$$\mathcal{A} \star_N \mathcal{X} + \mathcal{X} \star_M \mathcal{B} = \mathcal{C}, \tag{3}$$

where $\mathcal{A} \in \mathbb{R}^{I_1 \times \cdots \times I_N \times I_1 \times \cdots \times I_N}$, $\mathcal{B} \in \mathbb{R}^{J_1 \times \cdots \times J_M \times J_1 \times \cdots \times J_M}$, $\mathcal{C} \in \mathbb{R}^{I_1 \times \cdots \times I_N \times J_1 \times \cdots \times J_M}$ and $\mathcal{X} \in \mathbb{R}^{I_1 \times \cdots \times I_N \times J_1 \times \cdots \times J_M}$. It is noteworthy that the Sylvester tensor equation given in (3) comes from the discretization of the linear partial differential equation by the finite difference, finite element, and spectral methods in high dimension [19, 27, 28, 26].

Recently, Sun et al. [38] investigated the generalized inverses of tensors via the Einstein product. Using the generalized inverses of tensors, they also gave the general solutions of the tensor equation (3). Behera and Mishra [4] derived further results on generalized inverses of tensors via the Einstein product. Later, Wang and Xu [41] considered the iterative algorithms for solving the tensor equation (3). Moreover, Dehdezi and Karimi [12] presented an extended version of a gradient-based iterative method for solving large multilinear systems via the Einstein product. They introduced a new preconditioner to accelerate the convergence rate of the new iterative methods. As the gradient-based and the gradient-based least-squares algorithms, Dehdezi [10] derived iterative methods for the Sylvester-transpose tensor equation as (1). Erfanifar and Hajarian [16] also proposed a method for solving the nonlinear tensor equation

$$\mathcal{X} + \mathcal{A}^T \star_M \mathcal{X}^{-1} \star_N \mathcal{A} = \mathcal{I}$$

along with the Einstein product.

Brown and Hindmarsh [8] and then Jia [24] analyzed an incomplete generalized minimal residual method for solving large unsymmetric linear systems with low computational cost, which is a truncated version of the generalized minimal residual method (GMRES) [35]. Later, Saad and Wu [36] extracted a direct form of the incomplete generalized minimal residual method, abbreviated by DQGMRES, using QR decomposition of the Hessenberg matrix that appeared in the incomplete GMRES method. This motivates us to present an effective high order iterative algorithm as the DQGMRES method based

on the tensor format to solve the generalized Sylvester tensor equation (1) via the Einstein product.

The outline of this paper is as follows. In Section 2, we concisely recall some definitions and properties of tensor operators that are useful in the rest of the paper. In Section 3, we derive the tensor form of the DQGMRES method for solving the generalized Sylvester tensor equation (1) via the Einstein product. In Section 4, we analyze the convergence properties of the proposed method and find an upper bound for the residual norm of the approximate solution. Moreover, in Section 5, we report some numerical experiments on solving (1) using the presented method to illustrate its effectiveness and accuracy. Finally, a conclusion is drawn in Section 6.

2 Preliminaries

In this section, some preliminary definitions, and a number of technical lemmas are given, which will be used in what follows.

Definition 1. [38] Let N, M, L be the positive integers, let $\mathcal{A} \in \mathbb{R}^{I_1 \times \cdots \times I_N \times K_1 \times \cdots \times K_M}$, and let $\mathcal{B} \in \mathbb{R}^{K_1 \times \cdots \times K_M \times J_1 \times \cdots \times J_L}$. The Einstein product of two tensors $\mathcal{A}$ and $\mathcal{B}$ is defined by the operation $\star_M$ via

$$(\mathcal{A} \star_M \mathcal{B})_{i_1 \cdots i_N j_1 \cdots j_L} = \sum_{k_M=1}^{K_M} \cdots \sum_{k_1=1}^{K_1} a_{i_1 \cdots i_N k_1 \cdots k_M} b_{k_1 \cdots k_M j_1 \cdots j_L}. \tag{4}$$

Thus $\mathcal{A} \star_M \mathcal{B} \in \mathbb{R}^{I_1 \times \cdots \times I_N \times J_1 \times \cdots \times J_L}$ and the associative law of this tensor product holds.

For $\mathcal{A} = (a_{i_1 \cdots i_N j_1 \cdots j_M}) \in \mathbb{R}^{I_1 \times \cdots \times I_N \times J_1 \times \cdots \times J_M}$, let $\mathcal{B} = (b_{i_1 \cdots i_M j_1 \cdots j_N}) \in \mathbb{R}^{J_1 \times \cdots \times J_M \times I_1 \times \cdots \times I_N}$ be a tensor with $b_{i_1 \cdots i_M j_1 \cdots j_N} = a_{j_1 \cdots j_N i_1 \cdots i_M}$. We call $\mathcal{B}$ the transpose of $\mathcal{A}$ and denote it by $\mathcal{A}^T$.

When $N = M = 1$, the tensor equation (1) reduces to

$$AXB + CXD = F, \tag{5}$$

which is the generalized Sylvester matrix equation and arises frequently from the areas of systems and control theory [13, 14]. According to the repre-

sentation (5), the tensor equation (1) is called generalized Sylvester tensor equation.

Definition 2. [38] Let $\mathcal{A} = (a_{i_1\cdots i_N i_1\cdots i_N}) \in \mathbb{R}^{I_1\times\cdots\times I_N\times I_1\times\cdots\times I_N}$. The trace of $\mathcal{A}$ is defined as

$$\operatorname{tr}(\mathcal{A}) = \sum_{i_N=1}^{I_N}\cdots\sum_{i_1=1}^{I_1} a_{i_1\cdots i_N i_1\cdots i_N}. \tag{6}$$

The inner product of two tensors $\mathcal{X},\mathcal{Y} \in \mathbb{R}^{I_1\times\cdots\times I_N\times J_1\times\cdots\times J_M}$ is defined as

$$\langle\mathcal{X},\mathcal{Y}\rangle = \operatorname{tr}(\mathcal{Y}^T \star_N \mathcal{X}) = \sum_{j_M=1}^{J_M}\cdots\sum_{j_1=1}^{J_1}\sum_{i_N=1}^{I_N}\cdots\sum_{i_1=1}^{I_1} x_{i_1\cdots i_N j_1\cdots j_M} y_{i_1\cdots i_N j_1\cdots j_M}. \tag{7}$$

Therefore, the tensor norm induced by the inner product (7) is acquired as

$$\|\mathcal{A}\| = \sqrt{\langle\mathcal{X},\mathcal{X}\rangle} = \sqrt{\sum_{j_M=1}^{J_M}\cdots\sum_{j_1=1}^{J_1}\sum_{i_N=1}^{I_N}\cdots\sum_{i_1=1}^{I_1} |x_{i_1\cdots i_N j_1\cdots j_M}|^2}, \tag{8}$$

which is called the tensor Frobenius norm.

Let us set $I = I_1 I_2 \ldots I_N$ and, similarly, $J = J_1 J_2 \ldots J_N$, $K = K_1 K_2 \ldots K_M$, and $L = L_1 L_2 \ldots L_M$.

Definition 3. The transformation $\Phi_{IJ} : \mathbb{R}^{I_1\times\cdots\times I_N\times J_1\times\cdots\times J_N} \to \mathbb{R}^{I\times J}$ with $\Phi_{IJ}(\mathcal{A}) = A$ is defined component-wisely as

$$(\mathcal{A})_{i_1\cdots i_N j_1\cdots j_N} \to (A)_{st},$$

where $\mathcal{A} \in \mathbb{R}^{I_1\times\cdots\times I_N\times J_1\times\cdots\times J_N}$, $A \in \mathbb{R}^{I\times J}$, $s = i_N + \sum_{p=1}^{N-1}\left((i_p-1)\prod_{q=p+1}^{N} I_q\right)$, and $t = j_N + \sum_{p=1}^{N-1}\left((j_p-1)\prod_{q=p+1}^{N} J_q\right)$.

Routine computations verify that the tensor equation (1) is equivalent to the following large system of linear equations:

$$\mathcal{M}x = b, \tag{9}$$

with $x = \operatorname{vec}(\Phi_{IK}(\mathcal{X}))$, $b = \operatorname{vec}(\Phi_{IK}(\mathcal{F}))$, and

$$\mathcal{M} = B^T \otimes A + D^T \otimes C,$$

where $A = \Phi_{II}(\mathcal{A})$, $B = \Phi_{KK}(\mathcal{B})$, $C = \Phi_{II}(\mathcal{C})$, and $D = \Phi_{KK}(\mathcal{D})$.

The notation $\otimes$ represents the Kronecker product and the operator "vec" corresponds to a vector; see [17] for more details. The system of linear equations (9) is consistent if and only if (1) is consistent, which means that the coefficient matrix $\mathcal{M}$ needs to be nonsingular. In this study, it is assumed that the tensor equation (1) has a unique solution.

The jth *frontal slice* of an Nth order tensor $\mathcal{X} \in \mathbb{R}^{I_1 \times \cdots \times I_N}$ (also known as the column tensor of $\mathcal{X}$) is denoted by

$$\underbrace{\mathcal{X}_{:\cdots:j}}_{(N-1)-\text{times}}, \quad \text{for } j = 1, 2, \ldots, I_N,$$

which is a tensor in $\mathbb{R}^{I_1 \times \cdots \times I_{N-1}}$ and is obtained by fixing the last index.

Definition 4. The operator $\times_n$ stands for the n-mode matrix product of a tensor $\mathcal{X} \in \mathbb{R}^{I_1 \times \cdots \times I_N}$ with a matrix $A \in \mathbb{R}^{J \times I_n}$ as $\mathcal{X} \times_n A_n$, which is an Nth order tensor of size $I_1 \times I_2 \times \cdots I_{k-1} \times J \times I_{k+1} \cdots \times I_N$. For each element, we have

$$(\mathcal{X} \times_n A)_{i_1 \cdots i_{n-1} j i_{n+1} \cdots i_N} = \sum_{i_n=1}^{I_n} x_{i_1 \cdots i_N} a_{j i_n}. \tag{10}$$

Definition 5. The operator $\bar{\times}_n$ (for $n = 1, 2, \ldots, N$) represents the n-mode (vector) product of a tensor $\mathcal{X} \in \mathbb{R}^{I_1 \times \cdots \times I_N}$ with a vector $v \in \mathbb{R}^{I_n}$ is indicated by $\mathcal{X} \bar{\times}_n v$, which is an $(N-1)$th order tensor of size $I_1 \times I_2 \times \cdots I_{n-1} \times I_{n+1} \cdots \times I_N$. The elements are defined as follows:

$$(\mathcal{X} \bar{\times}_n v)_{i_1 i_2 \cdots i_{n-1} i_{n+1} \cdots i_N} = \sum_{i_n=1}^{I_n} x_{i_1 i_2 \cdots i_N} v_{i_n}.$$

Based on Definitions 4 and 5, one can establish some simple calculation rules for the matrix and the vector k-mode products representations; see [25] for further details.

Lemma 1. If $\mathcal{X} \in \mathbb{R}^{I_1 \times \cdots \times I_N}$, $A \in \mathbb{R}^{J_k \times I_k}$ and $v \in \mathbb{R}^{J_k}$, then

$$\mathcal{X} \times_k A \bar{\times}_k v = \mathcal{X} \bar{\times}_k (A^T v).$$

We can see the validity of the following proposition in [25], which is useful for our development.

Proposition 1. If $\mathcal{X} \in \mathbb{R}^{I_1 \times \cdots \times I_N}$, $A \in \mathbb{R}^{J_k \times I_k}$, and $B \in \mathbb{R}^{P_k \times J_k}$, then

$$(\mathcal{X} \times_k A) \times_k B = \mathcal{X} \times_k (BA).$$

Proposition 2. Let $\mathcal{X} \in \mathbb{R}^{I_1 \times \cdots \times I_N}$ be an Nth order tensor and $v = e_j$ such that e_j is the jth column of the identity matrix $I^{(I_N)}$. Then

$$\mathcal{X} \bar{\times}_N v = \mathcal{X}_{::\cdots:j}, \quad j = 1, 2, \ldots, I_N.$$

Consider two N-mode tensors $\mathcal{X}$ and $\mathcal{Y}$. We define $\boxtimes^{(N)}$ product for $N = 1, 2, \ldots$, by beginning 1-mode tensor as a vector and developing the 2-mode tensor as a matrix. In point of fact, the $\boxtimes^{(1)}$ and $\boxtimes^{(2)}$ products are naturally written in the following forms:

$$\mathcal{X} \boxtimes^{(1)} \mathcal{Y} = \mathcal{X}^T \mathcal{Y}, \quad \mathcal{X}, \mathcal{Y} \in \mathbb{R}^{I_1},$$

and

$$\mathcal{X} \boxtimes^{(2)} \mathcal{Y} = \mathcal{X}^T \mathcal{Y}, \quad \mathcal{X} \in \mathbb{R}^{I_1 \times I_2}, \mathcal{Y} \in \mathbb{R}^{I_1 \times \tilde{I}_2}.$$

In general case, the $\boxtimes^{(N)}$ product between two tensors $\mathcal{X} \in \mathbb{R}^{I_1 \times I_2 \times \cdots I_{N-1} \times I_N}$ and $\mathcal{Y} \in \mathbb{R}^{I_1 \times I_2 \times \cdots I_{N-1} \times \tilde{I}_N}$ is defined as an $I_N \times \tilde{I}_N$ matrix whose (i, j)th element is

$$\left[\mathcal{X} \boxtimes^{(N)} \mathcal{Y}\right]_{ij} = \mathrm{tr}(\mathcal{X}_{::\cdots:i} \boxtimes^{(N-1)} \mathcal{Y}_{::\cdots:j}), \quad N = 2, 3, \ldots.$$

The following proposition from [6] presents some constructive relations for the $\boxtimes^{(N+1)}$ product and the $\bar{\times}_k$ vector product, which are useful for the convergence analysis of the proposed method.

Proposition 3. Suppose that $\mathcal{B} \in \mathbb{R}^{I_1 \times I_2 \times \cdots I_N \times m}$ is an $(N+1)$-mode tensor with the N-mode column tensors $\mathcal{B}_1, \mathcal{B}_2, \ldots, \mathcal{B}_m \in \mathbb{R}^{I_1 \times I_2 \times \cdots I_N}$ and $z = (z_1, z_2, \ldots, z_m)^T \in \mathbb{R}^m$. For an arbitrary $(N+1)$-mode tensor $\mathcal{A}$ with N-mode column tensors $\mathcal{A}_1, \mathcal{A}_2, \ldots, \mathcal{A}_m$, we have the following statements:

$$\mathcal{A} \boxtimes^{(N+1)} (\mathcal{B} \bar{\times}_{N+1} z) = \left(\mathcal{A} \boxtimes^{(N+1)} \mathcal{B}\right) z, \tag{11}$$

and

$$(\mathcal{B} \bar{\times}_{N+1} z) \boxtimes^{(N+1)} \mathcal{A} = z^T \left(\mathcal{B} \boxtimes^{(N+1)} \mathcal{A}\right). \tag{12}$$

In the spirit of the fact that $\|\mathcal{X}\|^2 = \text{tr}(\mathcal{X} \boxtimes^{(N)} \mathcal{X}) = \mathcal{X} \boxtimes^{(N+1)} \mathcal{X}$, and also using Proposition 3, the next proposition is acquired.

Definition 6. The set of N-mode tensors $\mathcal{V}_1, \mathcal{V}_2, \ldots, \mathcal{V}_m \in \mathbb{R}^{I_1 \times I_2 \times \cdots I_N}$ is called orthonormal if

$$\langle \mathcal{V}_i, \mathcal{V}_j \rangle = 0, \quad i, j = 1, 2, \ldots, m (i \neq j),$$

and $\langle \mathcal{V}_i, \mathcal{V}_i \rangle = 1$ for $i = 1, 2, \ldots, m$.

Remark 1. Suppose that $\mathcal{A}$ is a given $(N+1)$-mode tensor with the column tensors $\mathcal{A}_1, \mathcal{A}_2, \ldots, \mathcal{A}_m \in \mathbb{R}^{I_1 \times I_2 \times \cdots I_N}$. If the set of N-mode tensors $\mathcal{A}_1, \mathcal{A}_2, \ldots, \mathcal{A}_m$ is orthonormal, then

$$\mathcal{A} \boxtimes^{(N+1)} \mathcal{A} = I^{(m)}.$$

Proposition 4. Let $\mathcal{X} \in \mathbb{R}^{I_1 \times I_2 \times \cdots \times I_N}$ be an N-mode tensor, and let $v \in \mathbb{R}^{I_N}$. Then,

$$\|\mathcal{X} \bar{\times}_N v\| \leq \|\mathcal{X}\| \|v\|_2.$$

Remark 2. In the case that the frontal slices of a tensor $\mathcal{F} \in \mathbb{R}^{I_1 \times I_2 \times \cdots \times I_N}$ is orthonormal, then for $v \in \mathbb{R}^{I_N}$, Remark 1 concludes

$$\|\mathcal{F} \bar{\times}_N v\| = \|v\|_2.$$

3 Tensor form of the quasi-GMRES method

By using given tensors $\mathcal{A}, \mathcal{C} \in \mathbb{R}^{I_1 \times \cdots \times I_N \times I_1 \times \cdots \times I_N}$, $\mathcal{B}, \mathcal{D} \in \mathbb{R}^{K_1 \times \cdots \times K_M \times K_1 \times \cdots \times K_M}$, we define the following linear operator:

$$\mathcal{L} : \mathbb{R}^{I_1 \times \cdots \times I_N \times K_1 \times \cdots \times K_M} \to \mathbb{R}^{I_1 \times \cdots \times I_N \times K_1 \times \cdots \times K_M},$$

as

$$\mathcal{X} \mapsto \mathcal{L}(\mathcal{X}) := \mathcal{A} \star_N \mathcal{X} \star_M \mathcal{B} + \mathcal{C} \star_N \mathcal{X} \star_M \mathcal{D}.$$

Based on the above definition, the generalized Sylvester tensor equation (1) is stated as $\mathcal{L}(\mathcal{X}) = \mathcal{F}$.

Thanks to the above discussion, the kth tensor Krylov subspace associated with the linear operator $\mathcal{L}$ and a tensor $\mathcal{V} \in \mathbb{R}^{I_1 \times \cdots \times I_N \times K_1 \times \cdots \times K_M}$ is defined

as

$$\mathcal{K}_k(\mathcal{L},\mathcal{V}) = \text{span}\{\mathcal{V}, \mathcal{L}(\mathcal{V}), \dots \mathcal{L}^{k-1}(\mathcal{V})\},$$

where $\mathcal{L}^i(\mathcal{V}) = \mathcal{L}(\mathcal{L}^{i-1}(\mathcal{V}))$ and $\mathcal{L}^0(\mathcal{V}) = \mathcal{V}$.

First, we introduce a useful alternative to the well-known Arnoldi process by truncating the orthogonalization process [24]. In this way, we achieve a strategy with low computational cost and a small truncation parameter m. It is emphasized that the truncation parameter m for the kth tensor Krylov subspace must be satisfied $2 \leq m \leq k$. Here, we start with the tensor form of the incomplete orthogonalization process (IOP_BTF), described by Algorithm 2.

Algorithm 2: IOP_BTF

1. **Input:** Given tensors $\mathcal{A}, \mathcal{C} \in \mathbb{R}^{I_1\times\cdots\times I_N\times I_1\times\cdots\times I_N}$, $\mathcal{B}, \mathcal{D} \in \mathbb{R}^{K_1\times\cdots\times K_M\times K_1\times\cdots\times K_M}$ and $\mathcal{V} \in \mathbb{R}^{I_1\times\cdots\times I_N\times K_1\times\cdots\times K_M}$.
2. Set $\beta = \|\mathcal{V}\|$ and $\mathcal{V}_1 = \mathcal{V}/\beta$
3. For given k, define $(k+1)\times k$ matrix $\bar{H}_k$, and set $\bar{H}_k = 0$;
4. **for** $j = 1, 2, \dots, k$ **do**
5. Compute $\mathcal{W}_j = \mathcal{L}(\mathcal{V}_j)$
6. **for** $i = \max\{1, j-m+1\}, \dots, j$ **do**
7. $h_{ij} = \langle \mathcal{W}_j, \mathcal{V}_i \rangle$
8. $\mathcal{W}_j = \mathcal{W}_j - h_{ij}\mathcal{V}_i$
9. **end**
10. Compute $h_{j+1,j} = \|\mathcal{W}_j\|$ and $\mathcal{V}_{j+1} = \frac{\mathcal{W}_j}{h_{j+1,j}}$
11. **end**
12. **Output:** Tensors $\mathcal{V}_j$, for $j = 1, 2, \dots, k+1$ and matrix $\bar{H}_k$.

It is plain to verify that the IOP_BTF strategy produces the locally orthonormal basis $\mathcal{V}_1, \mathcal{V}_2, \dots, \mathcal{V}_k$ (only the last m tensors $\mathcal{V}_i$'s are orthonormal) for the tensor Krylov subspace $\mathcal{K}_k(\mathcal{L},\mathcal{V})$ [35].

Let $\bar{H}_k = [h_{ij}]_{(k+1)\times k}$ be the matrix whose nonzero entries are those computed in lines 7 and 10 of Algorithm 2. We denote H_k as the matrix obtained from $\bar{H}_k$ by deleting its last row. Note that the Hessenberg matrix H_k has a band structure with a bandwidth $m+1$. Assume that $\tilde{\mathcal{V}}_k$ is the $(M+N+1)$-mode tensor with the frontal slices $\mathcal{V}_1, \mathcal{V}_2, \dots, \mathcal{V}_k$ obtained by

Algorithm 2 with the truncation parameter m. Beik, Movahed, and Ahmadi-Asl [6] have proven the following statement for the Arnoldi_BTF process

$$[\mathcal{L}(\mathcal{V}_1), \cdots, \mathcal{L}(\mathcal{V}_k)] = \tilde{\mathcal{V}}_{k+1} \times_{(N+M+1)} \bar{H}_k^T, \tag{13}$$

which is also satisfied for the IOP_BTF strategy in Algorithm 2.

Here, we briefly recall how the well-known GMRES method can be extended based on the basis of the tensor form. Let $\mathcal{X}_0 \in \mathbb{R}^{I_1 \times \cdots \times I_N \times K_1 \times \cdots \times K_M}$ be a given initial tensor guess for the exact solution of (1) with the corresponding residual tensor $\mathcal{R}_0 = \mathcal{F} - \mathcal{L}(\mathcal{X}_0) \in \mathbb{R}^{I_1 \times \cdots \times I_N \times K_1 \times \cdots \times K_M}$. For the approximate solution $\mathcal{X}_k$ computed at the kth iterative step of the GMRES_BTF method [23], we consider

$$\mathcal{X}_k \in \mathcal{X}_0 + \mathcal{K}_k(\mathcal{L}, \mathcal{R}_0),$$

and

$$\|\mathcal{F} - \mathcal{L}(\mathcal{X}_k)\| = \min_{\mathcal{X} \in \mathcal{X}_0 + \mathcal{K}_k(\mathcal{L}, \mathcal{R}_0)} \|\mathcal{F} - \mathcal{L}(\mathcal{X})\|. \tag{14}$$

So, the quasi-GMRES method (QGMRES) consists of performing the IOP_BTF and constructing $\mathcal{X}_k = \mathcal{X}_0 + \tilde{\mathcal{V}}_k \bar{\times}_{(M+N+1)} y_k$, where y_k is obtained as the condition (14) holds true; see [23] for more details.

As Saad and Wu mentioned in [36], the dimension of the Krylov subspace in the GMRES method increases by one at each step, which makes the procedure impractical for large dimensions. There are two standard remedies to this problem. The first is to restart the algorithm. In a simple way, the dimension is fixed, and the algorithm is restarted as many times as necessary, defining the initial vector defined as the latest approximation from the previous outer iteration. An alternative is to truncate the long-recurrence of the Arnoldi process as described in the IOP_BTF strategy in Algorithm 2.

Following the incomplete GMRES method presented by Brown and Hindmarsh [8], we now describe the QGMRES method based on the tensor format (QGMRES_BTF) in Algorithm 3.

Algorithm 3: QGMRES_BTF

1. **Input:** Given tensors $\mathcal{A}, \mathcal{C} \in \mathbb{R}^{I_1 \times \cdots \times I_N \times I_1 \times \cdots \times I_N}$, $\mathcal{B}, \mathcal{D} \in \mathbb{R}^{K_1 \times \cdots \times K_M \times K_1 \times \cdots \times K_M}$ and $\mathcal{F} \in \mathbb{R}^{I_1 \times \cdots \times I_N \times J_1 \times \cdots \times J_M}$, truncation parameter m, and initial guess $\mathcal{X}_0 \in \mathbb{R}^{I_1 \times \cdots \times I_N \times K_1 \times \cdots \times K_M}$.
2. Compute $\mathcal{R}_0 = \mathcal{F} - \mathcal{L}(\mathcal{X}_0)$, set $\beta = \|\mathcal{R}_0\|$ and $\mathcal{V}_1 = \mathcal{R}_0/\beta$
3. For given k, define $(k+1) \times k$ matrix $\bar{H}_k$, and set $\bar{H}_k = 0$
4. **for** $j = 1, 2, \ldots, k$ **do**
5. Compute $\mathcal{W}_j = \mathcal{L}(\mathcal{V}_j)$
6. **for** $i = \max\{1, j-m+1\}, \ldots, j$ **do**
7. $h_{ij} = \langle \mathcal{W}_j, \mathcal{V}_i \rangle$
8. $\mathcal{W}_j = \mathcal{W}_j - h_{ij}\mathcal{V}_i$
9. **end**
10. Compute $h_{j+1,j} = \|\mathcal{W}_j\|$ and $\mathcal{V}_{j+1} = \frac{\mathcal{W}_j}{h_{j+1,j}}$
11. **end**
12. Solve the problem $y_k = \operatorname{argmin}_{y \in \mathbb{R}^k} \|\bar{H}_k y - \beta e_1\|$
13. Compute $\mathcal{X}_k = \mathcal{X}_0 + \tilde{\mathcal{V}}_k \bar{\times}_{(M+N+1)} y_k$
14. **Output:** Approximate solution $\mathcal{X}_k$.

Constructing of $\hat{\mathcal{V}}_{k+1}$ and its first frontal slice as $\mathcal{V}_1 = \mathcal{R}_0/\beta$, yields $\mathcal{R}_0 = \hat{\mathcal{V}}_{k+1} \bar{\times}_{(M+N+1)} (\beta e_1)$, where e_1 is the first column of the identity matrix $I^{(k+1)}$. By using Lemma 1 and also making use of (13), the residual tensor $\mathcal{R}_k$ for the QGMRES_BTF approximate solution $\mathcal{X}_k$ generated by Algorithm 3 is given by

$$\begin{aligned} \mathcal{R}_k &= \mathcal{R}_0 - (\tilde{\mathcal{V}}_{k+1} \times_{(M+N+1)} \bar{H}_k^T) \bar{\times}_{(M+N+1)} y_k, \\ &= \tilde{\mathcal{V}}_{k+1} \bar{\times}_{(M+N+1)} [\beta e_1 - \bar{H}_k y_k]. \end{aligned}$$

The norm of the residual tensor $\mathcal{R}_k$ is then formulated as

$$\|\mathcal{R}_k\| = \|\tilde{\mathcal{V}}_{k+1} \bar{\times}_{(M+N+1)} [\beta e_1 - \bar{H}_k y_k]\|, \tag{15}$$

where, as before in Algorithm 3, y_k minimizes the norm $\|\beta e_1 - \bar{H}_k y\|_2$ over all vectors y in $\mathbb{R}^k$. This approach does not minimize the actual norm of the residual tensor over $\mathcal{X}_0 + \mathcal{K}_k(\mathcal{L}, \mathcal{R}_0)$. This idea leads us to minimize the norm $\|\beta e_1 - \bar{H}_k y\|_2$ by the QR factorization method. We implement the

direct variant of the QGMRES idea motivated from [36] by using the Givens rotation matrices to transform $\bar{H}_k$ and βe_1, to get

$$\bar{R}_k = \begin{pmatrix} R_k \\ 0 \end{pmatrix}, \text{ and } \bar{g}_k = (\gamma_1, \gamma_2, \ldots, \gamma_{k+1})^T, \tag{16}$$

respectively, in which R_k is an upper triangular matrix. Actually, we construct the following unitary matrix of order $k+1$

$$\mathbf{Q}_k = \Omega_k \cdots \Omega_2 \Omega_1, \tag{17}$$

where the $(k+1) \times (k+1)$ Givens rotation matrices

$$\Omega_i = \begin{pmatrix} I^{(i-1)} & & & \\ & c_i & s_i & \\ & -s_i & c_i & \\ & & & I^{(k-i)} \end{pmatrix} \equiv \begin{pmatrix} c_i & s_i \\ -s_i & c_i \end{pmatrix}, \quad i = 1, 2, \ldots, k, \tag{18}$$

are used with $c_i^2 + s_i^2 = 1$ in which $I^{(n)}$ indicates the identity matrix of order n. We now construct the following pre-multiplication operations on the kth column of $\bar{H}_k$:

$$\Omega_{k-1}\Omega_{k-2}\cdots\Omega_{k-m} \begin{pmatrix} \vdots \\ 0 \\ 0 \\ h_{k-m+1,k} \\ \vdots \\ h_{kk} \\ h_{k+1,k} \\ \vdots \end{pmatrix} = \begin{pmatrix} \vdots \\ 0 \\ t_{k-m,k} \\ t_{k-m+1,k} \\ \vdots \\ t_{kk} \\ h_{k+1,k} \\ \vdots \end{pmatrix}. \tag{19}$$

By adopting Ω_k in the kth column of the result vector in (19), we get

$$\begin{pmatrix} c_k & s_k \\ -s_k & c_k \end{pmatrix} \begin{pmatrix} \vdots \\ 0 \\ t_{k-m,k} \\ \vdots \\ t_{kk} \\ h_{k+1,k} \\ \vdots \end{pmatrix} = \begin{pmatrix} \vdots \\ 0 \\ t_{k-m,k} \\ \vdots \\ t_{kk} \\ 0 \\ \vdots \end{pmatrix}, \tag{20}$$

with $c_k = \frac{t_{kk}}{\sqrt{t_{kk}^2+h_{k+1,k}^2}}$ and $s_k = \frac{h_{k+1,k}}{\sqrt{t_{kk}^2+h_{k+1,k}^2}}$. For elements of $\bar{g}_k$, we have the recurrence relations $\gamma_{k+1} = -s_k\gamma_k$ and $\gamma_k = c_k\gamma_k$, with the initial term $\gamma_1 = \beta$.

Then, for any vector $y \in \mathbb{R}^k$, one has

$$\begin{aligned} \|\beta e_1 - \bar{H}_k y\|_2^2 &= \|\mathbf{Q}_k(\beta e_1 - \bar{H}_k y)\|_2^2 \\ &= \|\bar{g}_k - \bar{R}_k y\|_2^2 \\ &= |\gamma_{k+1}|^2 + \|g_k - R_k y\|_2^2. \end{aligned} \tag{21}$$

The minimum of the left-hand side is reached when the second term on the right-hand side of (21) has disappeared. Since R_k is nonsingular, the minimum of (21) is obtained by $y_k = R_k^{-1}g_k$, in which g_k is the vector obtained by removing the last element γ_{k+1} from $\bar{g}_k$. We therefore have $\|\beta e_1 - \bar{H}_k y_k\|_2 = |\gamma_{k+1}|$.

Following the above discussion and making use of Lemma 1, we obtain

$$\begin{aligned} \mathcal{X}_k &= \mathcal{X}_0 + \tilde{\mathcal{V}}_k \bar{\times}_{N+1} y_k \\ &= \mathcal{X}_0 + \tilde{\mathcal{V}}_k \bar{\times}_{N+1} (R_k^{-1} g_k) \\ &= \mathcal{X}_0 + (\tilde{\mathcal{V}}_k \times_{N+1} R_k^{-T}) \bar{\times}_{N+1} g_k \\ &= \mathcal{X}_0 + \tilde{\mathcal{P}}_k \bar{\times}_{N+1} g_k \\ &= \mathcal{X}_0 + \tilde{\mathcal{P}}_{k-1} \bar{\times}_{N+1} g_{k-1} + \gamma_k \mathcal{P}_k \\ &= \mathcal{X}_{k-1} + \gamma_k \mathcal{P}_k, \end{aligned}$$

where $\tilde{\mathcal{P}}_k = \tilde{\mathcal{V}}_k \times_{N+1} R_k^{-T}$ with the frontal slices $\mathcal{P}_i$'s. By using Proposition 1, we conclude that $\tilde{\mathcal{V}}_k = \tilde{\mathcal{P}}_k \times_{N+1} R_k^T$, and straightforward computations yield

$$\begin{aligned}
&\mathcal{P}_1 = \mathcal{V}_1/t_{11},\\
&\mathcal{P}_2 = (\mathcal{V}_2 - t_{12}\mathcal{P}_1)/t_{22},\\
&\quad\vdots\\
&\mathcal{P}_k = t_{kk}^{-1}\left(\mathcal{V}_k - \sum_{i=k-m}^{k-1} t_{ik}\mathcal{P}_i\right),
\end{aligned}$$

where t_{ik} for $i = k-m,\ldots,k-1,k$ are the elements of the kth column of the upper triangular matrix R_k in (20). We can describe the DQGMRES algorithm based on the tensor format (DQGMRES_BTF) for solving the generalized Sylvester tensor equation (1) via the Einstein product as done in Algorithm 4.

Algorithm 4: DQGMRES_BTF

1. **Input:** Given tensors $\mathcal{A},\mathcal{C} \in \mathbb{R}^{I_1\times\cdots\times I_N\times I_1\times\cdots\times I_N}$, $\mathcal{B},\mathcal{D} \in \mathbb{R}^{K_1\times\cdots\times K_M\times K_1\times\cdots\times K_M}$ and $\mathcal{F} \in \mathbb{R}^{I_1\times\cdots\times I_N\times J_1\times\cdots\times J_M}$, truncation parameter m, and initial guess $\mathcal{X}_0 \in \mathbb{R}^{I_1\times\cdots\times I_N\times K_1\times\cdots\times K_M}$.
2. Compute $\mathcal{R}_0 = \mathcal{F} - (\mathcal{A}\star_N\mathcal{X}_0\star_M\mathcal{B} + \mathcal{C}\star_N\mathcal{X}_0\star_M\mathcal{D})$, $\gamma_1 = \|\mathcal{R}_0\|$, $\mathcal{V}_1 = \mathcal{R}_0/\gamma_1$
3. **for** $k = 0,1,\ldots$ until convergence **do**
4. Compute $h_{ik}, i = \max\{1,k-m+1\},\ldots,k$, and $\mathcal{V}_{k+1}$ as in lines 2 to 10 of Algorithm 2
5. Update the QR factorization of $\bar{H}_k$ according (19) and (20): i.e.
6. Apply $\Omega_i, i = k-m,\ldots,k-1$, to the kth column of $\bar{H}_k$
7. Compute the rotation coefficients c_k and s_k
8. Apply rotation Ω_k to the last column of $\bar{H}_k$ and to $\bar{g}_k$; i.e. compute
9. $\gamma_{k+1} = -s_k\gamma_k, \quad \gamma_k = c_k\gamma_k$
10. $t_{kk} = \sqrt{h_{k+1,k}^2 + t_{kk}^2}$
11. $\mathcal{P}_k = \left(\mathcal{V}_k - \sum_{i=k-m}^{k-1} t_{ik}\mathcal{P}_i\right)/t_{kk}$
12. $\mathcal{X}_k = \mathcal{X}_{k-1} + \gamma_k\mathcal{P}_k$
13. If $|\gamma_{k+1}|$ is small enough the Stop
14. **End**
15. **Output:** Approximate solution $\mathcal{X}_k$ for (1)

4 Convergence analysis of the QGMRES_BTF method

We prove here some convergence results for the DQGMRES_BTF method. The next theorem provides a representation of the residual tensor $\mathcal{R}_k$ of the DQGMRES_BTF method.

Lemma 2. Let $\tilde{\mathcal{V}}_k$ be an $(M+N+1)$-mode tensor with column tensors $\mathcal{V}_i$ for $i = 1, 2, \ldots, k$ which is generated by Algorithm 2 and $\mathbf{Q}_k$ the unitary matrix specified in (17). The residual tensor $\mathcal{R}_k$ of the DQGMRES_BTF method is then given by

$$\mathcal{R}_k = \gamma_{k+1}\tilde{\mathcal{V}}_{k+1} \times_{(M+N+1)} \mathbf{Q}_k \bar{\times}_{(M+N+1)} e_{k+1}, \tag{22}$$

where γ_{k+1} is the last element of $\bar{g}_k$ in (16).

Proof. As discussed earlier in (16), the kth residual iterate of the DQGMRES_BTF method has the following form:

$$\begin{aligned}\mathcal{R}_k &= \mathcal{R}_0 - \mathcal{L}(\tilde{\mathcal{V}}_k)\bar{\times}_{(M+N+1)} y_k \\ &= \tilde{\mathcal{V}}_{k+1}\bar{\times}_{(M+N+1)}(\beta e_1 - \bar{H}_k y_k) \\ &= \tilde{\mathcal{V}}_{k+1}\bar{\times}_{(M+N+1)}(\mathbf{Q}_k^T(\bar{g}_k - \bar{R}_k y_k)).\end{aligned}$$

In view of (21), one can see that y_k minimizes the 2-norm of $\bar{g}_k - \bar{R}_k y$ over y and thus annihilates all components of the right-hand side $\bar{g}_k$ except the last one, which is equal to $\gamma_{k+1}e_{k+1}$. Now, it follows that

$$\mathcal{R}_k = \tilde{\mathcal{V}}_{k+1}\bar{\times}_{(M+N+1)}(\mathbf{Q}_k^T(\gamma_{k+1}e_{k+1})).$$

Finally, making use of Lemma 1 completes the proof of the lemma. □

Next, we present a suitable upper bound for the residual norm of the DQGMRES_BTF method, which depends on the specific parameter computed in the proposed Algorithm 4 in a cost-effective way. For this purpose, we prove the following Lemma.

Lemma 3. The residual $\mathcal{R}_k$ obtained by the DQGMRES_BTF algorithm with the truncation parameter m for the generalized Sylvester tensor equations of the form (1) satisfies the following inequality:

$$\|\mathcal{R}_k\| \leq |\gamma_{k+1}|\sqrt{k-m+1}.$$

Proof. From Lemmas 1 and 2, we have

$$\mathcal{R}_k = \gamma_{k+1}\tilde{\mathcal{V}}_{k+1}\bar{\times}_{(M+N+1)}(\mathbf{Q}_k^T e_{k+1}).$$

Let $q = \mathbf{Q}_k^T e_{k+1}$ be the unit vector with components $\eta_1, \eta_2, \ldots, \eta_{k+1}$. Then by using Proposition 4, we get

$$\begin{aligned}
\|\mathcal{R}_k\| &= |\gamma_{k+1}|\|\tilde{\mathcal{V}}_{k+1}\bar{\times}_{(M+N+1)}q\| \\
&\leq |\gamma_{k+1}|\left(\left\|\sum_{i=1}^{m+1}\eta_i\mathcal{V}_i\right\| + \left\|\sum_{i=m+2}^{k+1}\eta_i\mathcal{V}_i\right\|\right) \\
&\leq |\gamma_{k+1}|\left(\left[\sum_{i=1}^{m+1}\eta_i^2\right]^{1/2} + \sum_{i=m+2}^{k+1}|\eta_i|\|\mathcal{V}_i\|\right) \\
&\leq |\gamma_{k+1}|\left(\left[\sum_{i=1}^{m+1}\eta_i^2\right]^{1/2} + \sqrt{k-m}\left[\sum_{i=m+2}^{k+1}\eta_i^2\right]^{1/2}\right) \\
&\leq |\gamma_{k+1}|\sqrt{k-m+1},
\end{aligned}$$

where the last inequality follows by the Cauchy–Schwarz inequality and $\|q\|_2 = 1$. □

The next corollary is come to the conclusion by Lemma 3 together with the useful relation between the last elements γ_{k+1} and γ_k of $\bar{g}_k$ and $\bar{g}_{k-1}$, respectively; that is, $\gamma_{k+1} = -s_k\gamma_k$.

Corollary 1. Let $\mathcal{R}_k$ be the residual tensor of the DQGMRES_BTF kth iterate. Then

$$\|\mathcal{R}_k\| \leq |s_1 s_2 \cdots s_k|\|\mathcal{R}_0\|\sqrt{k-m+1},$$

where s_i's are defined as (18).

An extension of the Gram–Schmidt orthogonalization process based on the tensor format concludes the following lemma for the linear independent tensors $\mathcal{V}_i$ for $i = 1, 2, \ldots, k$, which are generated by Algorithm 2.

Lemma 4. Suppose that $\tilde{\mathcal{V}}_{k+1}$ is an $(M+N+1)$-order tensor with the $k+1$ frontal slices $\mathcal{V}_i$ for $i = 1, 2, \ldots, k+1$ obtained by using the IOP_BTF

Algorithm 2. Then, there is an $(k+1)\times(k+1)$ nonsingular matrix $\mathbf{U}$ such that $\tilde{\mathcal{V}}_{k+1}=\tilde{\mathcal{F}}_{k+1}\times_{(M+N+1)}\mathbf{U}$, where $\tilde{\mathcal{F}}_{k+1}$ is an $(M+N+1)$-order tensor with the $k+1$ orthonormal frontal slices $\mathcal{F}_i$ for $i=1,2,\ldots,k+1$; that is,

$$\tilde{\mathcal{F}}_{k+1}\boxtimes^{(M+N+1)}\tilde{\mathcal{F}}_{k+1}=I^{(k+1)}. \tag{23}$$

Proof. The proof is a direct result of the tensor form of the Gram–Schmidt orthogonalization process described in [35]. □

In the last theorem of this section, an inequality is found that can be usefully applied in the convergence analysis of the DQGMRES_BTF method. This is a comparison of the residual tensor obtained after k steps of using the DQGMRES_BTF method with that of the GMRES_BTF method [9].

Theorem 1. Assume that $\tilde{\mathcal{V}}_{k+1}$ is an $(M+N+1)$-order tensor with $k+1$ frontal slices $\mathcal{V}_i$ for $i=1,2,\ldots,k+1$ obtained by using the IOP_BTF Algorithm 2, $\tilde{\mathcal{V}}_{k+1}=\tilde{\mathcal{F}}_{k+1}\times_{(M+N+1)}\mathbf{U}^T$, where $\tilde{\mathcal{F}}_{k+1}$ is satisfied (23) and $\mathbf{U}^T$ is nonsingular. Let $\mathcal{R}_k^Q$ and $\mathcal{R}_k^G$ be the residual obtained after k steps of using DQGMRES_BTF and GMRES_BTF methods, respectively. Then

$$\|\mathcal{R}_k^Q\|\leq\kappa_2(\mathbf{U})\|\mathcal{R}_k^G\|, \tag{24}$$

where $\kappa_2(\mathbf{U})$ is the condition number of the matrix $\mathbf{U}$.

Proof. Consider the subset of $\mathcal{K}_{k+1}(\mathcal{L},\mathcal{V}_1)$ given by

$$\mathcal{N}=\{\mathcal{R}:\mathcal{R}=\tilde{\mathcal{V}}_{k+1}\bar{\times}_{(M+N+1)}t;\ t=\beta e_1-\bar{H}_k y;\ y\in\mathbb{R}^k\}.$$

Denote by $\mathbf{y}_k$ the minimizer of $\|\beta e_1-\bar{H}_k y\|_2$ over y and $\mathbf{t}_k=\beta e_1-\bar{H}_k\mathbf{y}_k$. Thus, Lemma 2 concludes that $\mathcal{R}_k^Q=\tilde{\mathcal{V}}_{k+1}\bar{\times}_{(M+N+1)}\mathbf{t}_k$. For any member $\mathcal{R}\in\mathcal{N}$, there exists $\mathbf{t}$ such that $\mathcal{R}=\tilde{\mathcal{F}}_{k+1}\times_{(M+N+1)}\mathbf{U}^T\bar{\times}_{N+1}\mathbf{t}$, which is defined by Lemma 1, it is equivalent to $\mathcal{R}=\tilde{\mathcal{F}}_{k+1}\bar{\times}_{(M+N+1)}(\mathbf{U}\mathbf{t})$. Hence, Proposition 3 yields $\mathbf{U}\mathbf{t}=\tilde{\mathcal{F}}_{k+1}\boxtimes^{(M+N+1)}\mathcal{R}$. Since $\mathbf{U}$ is nonsingular, thus

$$\mathbf{t}=\mathbf{U}^{-1}(\tilde{\mathcal{F}}_{k+1}\boxtimes^{(M+N+1)}\mathcal{R}).$$

From the unitary property of $\tilde{\mathcal{F}}_{k+1}$, we deduce that

$$\|\mathcal{R}_k^Q\|=\|\mathbf{U}\mathbf{t}_k\|_2\leq\|\mathbf{U}\|_2\|\mathbf{t}_k\|_2. \tag{25}$$

Note that $\|\mathbf{t}_k\|_2$ is the minimum of the 2-norm of $\beta e_1 - \bar{H}_k y$ over y. Therefore,

$$\begin{aligned}\|\mathbf{t}_k\|_2 &= \|\mathbf{U}^{-1}(\tilde{\mathcal{F}}_{k+1} \boxtimes^{(M+N+1)} \mathcal{R}_k^Q)\| \\ &\leq \|\mathbf{U}^{-1}(\tilde{\mathcal{F}}_{k+1} \boxtimes^{(M+N+1)} \mathcal{R})\| \\ &\leq \|\mathbf{U}^{-1}\|_2 \|\tilde{\mathcal{F}}_{k+1} \boxtimes^{(M+N+1)} \mathcal{R}\|.\end{aligned}$$

It is convenient to obtain $\|\tilde{\mathcal{F}}_{k+1} \boxtimes^{(M+N+1)} \mathcal{R}\| = \|\mathcal{R}\|$, and then

$$\begin{aligned}\|\mathbf{t}_k\|_2 &\leq \|\mathbf{U}^{-1}\|_2 \|\mathcal{R}\|, \quad \text{for all } \mathcal{R} \in \mathcal{N} \\ &\leq \|\mathbf{U}^{-1}\|_2 \|\mathcal{R}_k^G\|.\end{aligned}$$

Consequently, equation (25) is revealed as

$$\begin{aligned}\|\mathcal{R}_k^Q\| &\leq \|\mathbf{U}\|_2 \|\mathbf{U}^{-1}\|_2 \|\mathcal{R}_k^G\| \\ &= \kappa_2(\mathbf{U}) \|\mathcal{R}_k^G\|.\end{aligned}$$

The result is now concluded. □

5 Numerical results

In this section, we present some numerical results to illustrate the effectiveness and accuracy of the proposed DQGMRES_BTF method for solving several types of the generalized Sylvester tensor equation (1) via the Einstein product. To this end, we compare the DQGMRES_BTF method with the CGNR_BTF method given in [10], the CGNE_BTF method proposed in [12] as the tensor format of the CGNR and CGNE algorithms in [35], respectively. We compare also our results with those of the RNSD_BTF method proposed in [5]. All computations were performed using double-precision floating-point arithmetic in MATLAB codes. The computer we used is a system with the specification Intel(R) Core(TM) i3 CPU 2.13GHz, 4G RAM, and 64-bit operating system. In all examples, we choose zero tensor $\mathcal{X}_0 = \mathcal{O}$ as the initial guess. It must be emphasized that no preconditioning was used for any of the test problems. We consider the stopping criterion

$$ERR \equiv \frac{\|\mathcal{R}_k\|}{\|\mathcal{R}_0\|} \leq 10^{-6},$$

where $\mathcal{R}_k$ is the residual tensor corresponding to the approximate solution $\mathcal{X}_k$; that is,

$$\mathcal{R}_k = \mathcal{F} - \mathcal{A} \star_N \mathcal{X}_k \star_M \mathcal{B} - \mathcal{C} \star_N \mathcal{X}_k \star_M \mathcal{D}.$$

If the stopping criterion mentioned above does not apply, then we consider the maximum number of iterations Max-Iter = 1000 in each example. In all tensor computations, we get help from the MATLAB Tensor Toolbox, developed by Bader and Kolda [1, 2] to implement MATLAB.

Example 1. [7, 23] Consider the 3D Poisson problem

$$\begin{cases} -\nabla^2 v = f, & \text{in } \Omega = \{(x, y, z), 0 < x, y, z < 1\}, \\ v = 0, & \text{on } \partial\Omega \end{cases} \tag{26}$$

where f is a given function and

$$\nabla^2 v = \frac{\partial^2 v}{\partial x^2} + \frac{\partial^2 v}{\partial y^2} + \frac{\partial^2 v}{\partial z^2}.$$

Several problems in physics and mechanics are modeled by (26), where the solution v means, for example, temperature, electromagnetic potential, or displacement of an elastic membrane fixed at the boundary. Now, we consider an approximation of the unknown function $v(x, y, z)$ in (26) corresponding to the uniform mesh step sizes, namely, Δx in the x-direction, Δy in the y-direction, and Δz in the z-direction, satisfy $\Delta x = \Delta y = \Delta z = h = \frac{1}{N+1}$. By the standard central finite difference formulas for the three dimensions, we obtain the following difference relationship:

$$6v_{ijk} - v_{i-1,j,k} - v_{i+1,j,k} - v_{i,j-1,k} - v_{i,j+1,k} - v_{i,j,k-1} - v_{i,j,k+1} = h^3 f_{ijk}. \tag{27}$$

Hence, the higher order tensor representation of the 3D discretized Poisson problem (26) as described in (27) is given by

$$\bar{\mathcal{A}}_N \star_3 \mathcal{V} = \mathcal{F}, \tag{28}$$

where the Laplacian tensor $\bar{\mathcal{A}}_N \in \mathbb{R}^{N\times N\times N\times N\times N\times N}$ and $\mathcal{V}, \mathcal{F} \in \mathbb{R}^{N\times N\times N}$. Both $\mathcal{V}$ and $\mathcal{F}$ are discretized on the unit cube. The entries on the tensor block $(\bar{\mathcal{A}}_N)^{(2,4,6)}_{l,m,n}$ of $\bar{\mathcal{A}}_N$ in (28) follow a seven-point stencil as

$$
\begin{aligned}
((\bar{\mathcal{A}}_N)^{(2,4,6)}_{\alpha,\beta,\gamma})_{\alpha,\beta,\gamma} &= \frac{6}{h^3},\\
((\bar{\mathcal{A}}_N)^{(2,4,6)}_{\alpha,\beta,\gamma})_{\alpha-1,\beta,\gamma} &= ((\bar{\mathcal{A}}_N)^{(2,4,6)}_{\alpha,\beta,\gamma})_{\alpha+1,\beta,\gamma} = -\frac{1}{h^3},\\
((\bar{\mathcal{A}}_N)^{(2,4,6)}_{\alpha,\beta,\gamma})_{\alpha,\beta-1,\gamma} &= ((\bar{\mathcal{A}}_N)^{(2,4,6)}_{\alpha,\beta,\gamma})_{\alpha,\beta+1,\gamma} = -\frac{1}{h^3},\\
((\bar{\mathcal{A}}_N)^{(2,4,6)}_{\alpha,\beta,\gamma})_{\alpha,\beta,\gamma-1} &= ((\bar{\mathcal{A}}_N)^{(2,4,6)}_{\alpha,\beta,\gamma})_{\alpha,\beta,\gamma+1} = -\frac{1}{h^3},
\end{aligned}
$$

for $\alpha, \beta, \gamma = 2, \ldots, N-1$. We use the notation $(\bar{\mathcal{A}}_N)^{(2,4,6)}_{l,m,n} = \bar{\mathcal{A}}_N(:, l, :, m, :, n)$ for the block tensors of $\bar{\mathcal{A}}_N$. For different grids $N = 4, 6, 8$, the iteration number and the CPU time of the CGNR_BTF and the CGNE_BTF methods are reported in Table 1, compared with the proposed DQGMRES_BTF method with the truncation parameter $m = 5$. The corresponding convergence histories of the numerical results are depicted in Figures 1 and 2 with the truncation parameter $m = 10$ of the DQGMRES_BTF method for $N = 8$ and $N = 10$, respectively. These results show that the DQGMRES_BTF algorithm is more effective and less expensive than the other solvers.

Table 1: Results of the iteration number (Iter) and CPU time (Time) for Example 1 with different Grids and the truncation parameter $m = 10$.

Methods	CGNE_BTF		CGNR_BTF		DQGMRES_BTF	
Grid	Time	Iter	Time	Iter	Time	Iter
$4 \times 4 \times 4$	0.1635	6	0.1626	6	0.0126	6
$6 \times 6 \times 6$	5.9706	19	5.9789	19	0.1311	19
$8 \times 8 \times 8$	67.7331	38	67.5701	38	0.5693	26

Example 2. Let us consider Sylvester tensor equation,

$$\mathcal{A} \star_N \mathcal{X} + \mathcal{X} \star_M \mathcal{B} = \mathcal{C}, \tag{29}$$

where various cases for the coefficient tensors $\mathcal{A}$ and $\mathcal{B}$ are given by

(a) $\mathcal{A} = \text{tenrand}([4\ 2\ 4\ 2])$, $\mathcal{B} = \text{tenrand}([5\ 3\ 5\ 3])$,

(b) $\mathcal{A} = \text{tenrand}([6\ 4\ 6\ 4])$, $\mathcal{B} = \text{tenrand}([8\ 5\ 8\ 5])$,

(c) $\mathcal{A} = \text{tenrand}([10\ 5\ 10\ 5])$, $\mathcal{B} = \text{tenrand}([12\ 6\ 12\ 6])$.

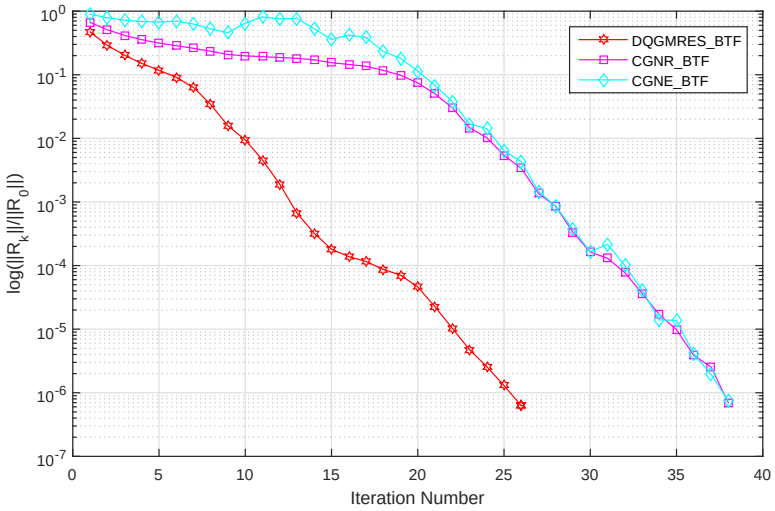


Figure 1: Comparison of convergence histories for Example 1 with Grid $N = 8$ and the truncation parameter $m = 10$.

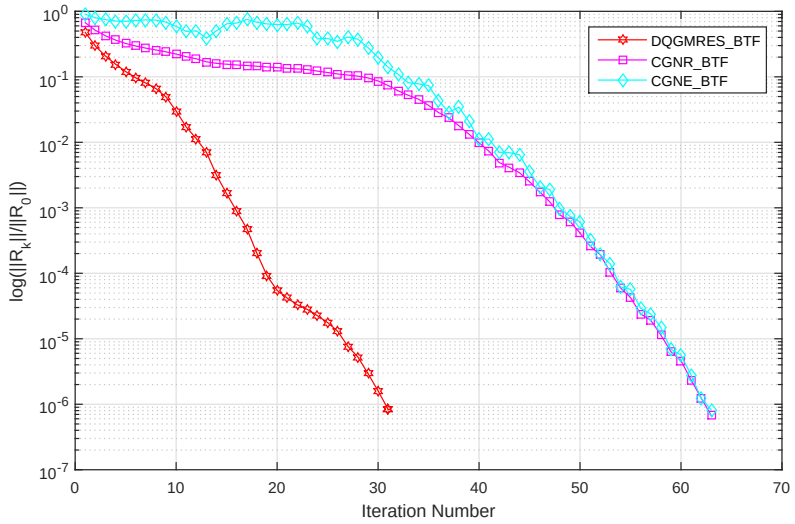


Figure 2: Comparison of convergence histories for Example 1 with Grid $N = 10$ and the truncation parameter $m = 10$.

The iteration number and CPU time of the CGNR_BTF, CGNE_BTF, and RNSD_BTF methods are reported in Table 2, compared with the proposed DQGMRES_BTF method with the truncation parameter $m = 5$ for cases (a)-(b). The corresponding convergence histories of the numerical results are depicted in Figure 3 with the truncation parameter $m = 5$ of the DQGMRES_BTF method. Theses results show that the DQGMRES_BTF algorithm is more effective and less expensive than the other methods. If we apply the DQGMRES_BTF algorithm, we obtain the more efficient approximate solution of Example 2. The result curves for case (c) are depicted in Figure 4. These results confirm the acceptable convergence of the proposed DQGMRES_BTF method. In other words, we can say that the proposed method is efficient for solving this type of tensor equation equipped with the Einstein product for small truncation parameters. The corresponding convergence histories of the numerical results for large-size $\mathcal{A} = tenrand([20\ 10\ 20\ 10])$ and $\mathcal{B} = tenrand([10\ 10\ 10\ 10])$ are depicted in Figure 5 with the truncation parameter $m = 5$ of the DQGMRES_BTF method and superior property of the DQGMRES_BTF method is observed compared to those of the CGNR_BTF method.

Table 2: Results of iteration number (Iter) and CPU time (Time) for Example 2.

Methods	RNSD_BTF		CGNE_BTF		CGNR_BTF		DQGMRES_BTF	
	Time	Iter	Time	Iter	Time	Iter	Time	Iter
case (a)	1.8173	†	0.9166	200	0.6954	151	0.1387	29
case (b)	17.7077	†	14.6543	416	6.8674	195	1.0540	29
case (c)	54.3223	†	41.6364	121	16.6119	303	2.8529	21

According to Definition 3, one can reduce the Sylvester tensor equation (29) to the associated Sylvester matrix equation and then solve it by the block QGMRES method [18]. We present the numerical results of applying the block QGMRES method to the reduced matrix equation and compare them with those of the DQGMRES_BTF method with the small truncation parameter $m = 5$. The advantage of the DQGMRES_BTF method is the short elapsed CPU time. We use the global conjugate gradient method [18] to solve the minimization problem in the block QGMRES method with inner max-iter=1000. As the conclusion, in the case $I_1 = 10, I_2 = 6, J_1 = 10, J_2 =$

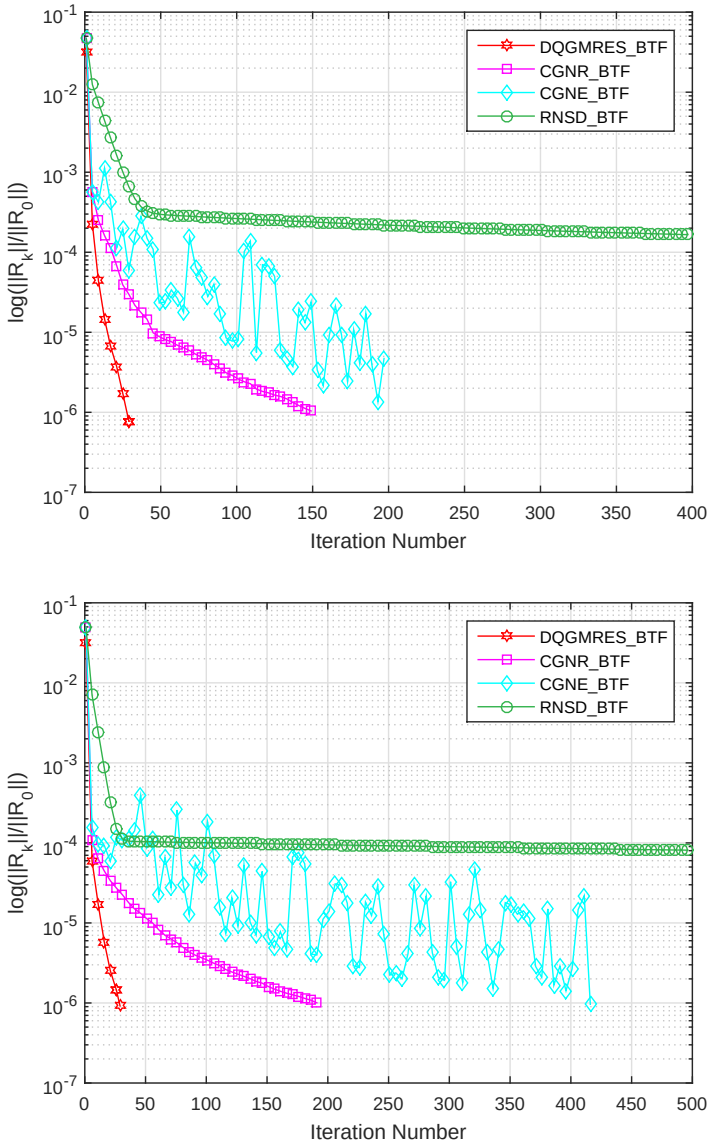


Figure 3: Comparison of convergence histories for case (a) (Up) and case (b) (Down) in Example 2 with the truncation parameter $m = 5$.

4, the results of Figure 6 (Up) have been obtained. It seems we have a better number of iterations for the block QGMRES, but the elapsed time is worse than the DQGMRES-BTF methods as depicted in Figure 6 (Down).

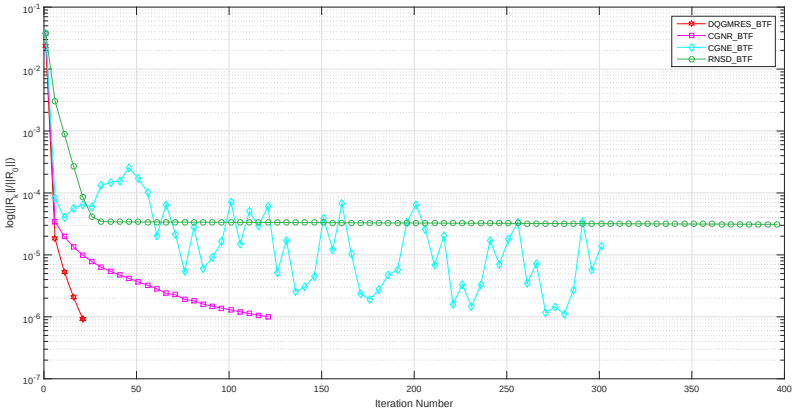


Figure 4: Comparison of convergence histories for case (c) in Example 2 with the truncation parameter $m = 5$.

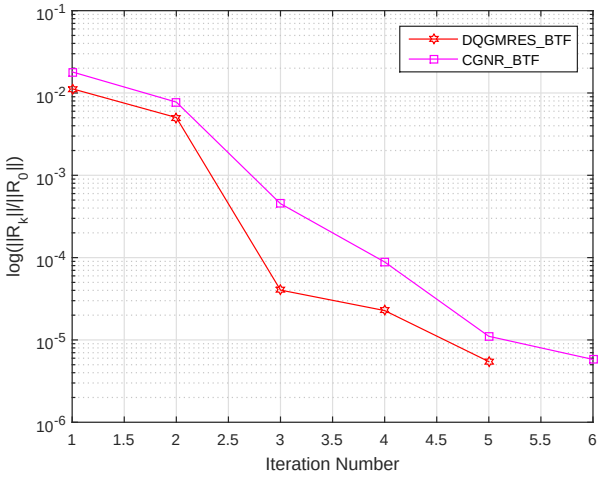


Figure 5: Comparison of convergence histories in Example 2 with the truncation parameter $m = 5$.

Example 3. Consider generalized Sylvester tensor equation

$$\mathcal{A} \star_N \mathcal{X} \star_M \mathcal{B} + \mathcal{C} \star_N \mathcal{X} \star_M \mathcal{D} = \mathcal{F},$$

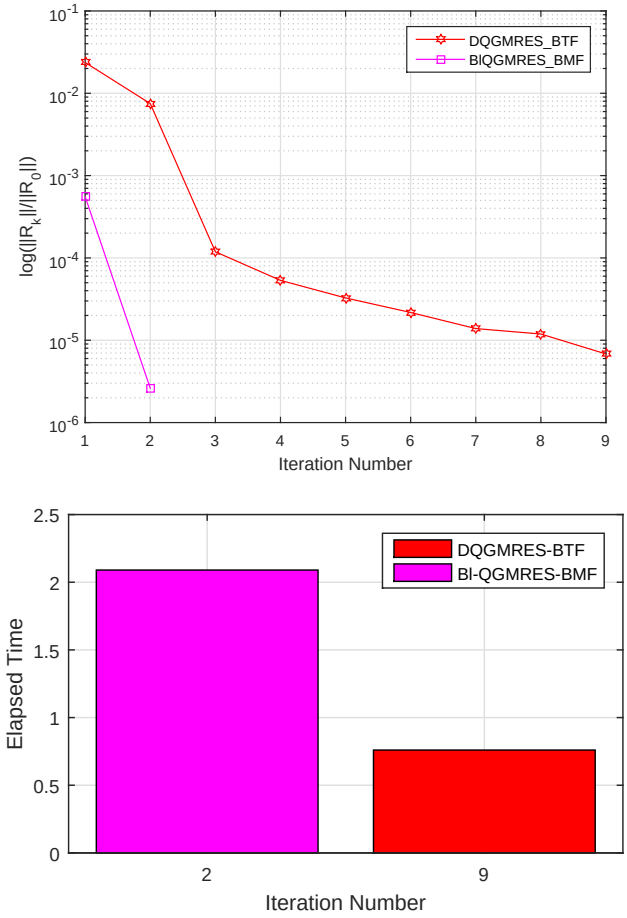


Figure 6: Comparison of the convergence behavior (Up) and the elapsed times (Down) for the block QGMRES and DQGMRES_BTF methods in Example 2 with the truncation parameter $m = 5$.

where $\mathcal{A} = \text{tenrand}([6\ 6\ 6\ 6])$, $\mathcal{B} = \text{tenrand}([8\ 8\ 8\ 8])$, $\mathcal{C} = \text{tenrand}([6\ 6\ 6\ 6])$, $\mathcal{D} = \text{tenrand}([8\ 8\ 8\ 8])$.

In Table 3, we report the numerical results of the iteration number and the CPU time of the CGNR_BTF and CGNE_BTF methods, compared to the proposed DQGMRES_BTF method with the different truncation parameters m. The convergence curves of the numerical results are depicted in Figure 7 with the truncation parameter $m = 5$ of the DQGMRES_BTF method. The

effectiveness of the DQGMRES_BTF method and the less elapsed time are shown in Table 3 and Figure 7.

Table 3: Results of iteration number (Iter) and CPU time (Time) for Example 3 with various truncation parameters m in the DQGMRES_BTF method.

Methods	Time(Iter)
CGNR_BTF	11.8351(70)
CGNE_BTF	32.6890(191)
DQGMRES_BTF	
m=5	2.7033(16)
m=10	2.8392(16)
m=15	2.5292(15)

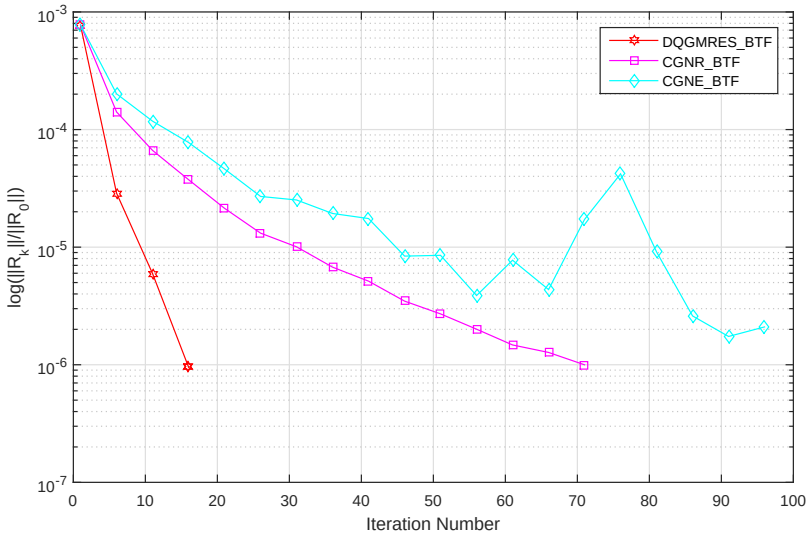


Figure 7: Comparison of convergence histories for Example 3 with the truncation parameter $m = 5$.

6 Conclusion

In this paper, we proposed an iterative method for solving generalized Sylvester tensor equations via the Einstein product using the tensor form of the QGMRES method. We present some useful results on tensor computations and propose a direct variant of the QGMRES method to utilize previous data and practical implementation of the method. Also, some results proved to illustrate a prior convergence behavior of the new method. In the numerical results of the experimental problems, we observed that the presented method has more efficiency and accuracy properties with low computational cost compared to the other tensor equation solvers such as CGNR, CGNE, and RNSD methods.

Acknowledgements

The author is very much indebted to the anonymous referees for their constructive suggestions and helpful comments, which greatly improved the original manuscript of this paper.

Data Availability The data that support the findings of this study are available from the corresponding author upon reasonable request.

References

[1] Bader, B.W. and Kolda, T.G. *Efficient MATLAB computations with sparse and factored tensors*, SIAM J. Sci. Comput. 30(1) (2007), 205–231.

[2] Bader, B.W. and Kolda, G.T. *Tensor Toolbox for MATLAB, Version 3.6*, Available online at https://www.tensortoolbox.org., 2023.

[3] Ballani, J. and Grasedyck, L. *A projection method to solve linear systems in tensor form*, Numer. Linear Algebra Appl. 20 (2013), 27–43.

[4] Behera, R. and Mishra, D. *Further results on generalized inverses of tensors via the Einstein product*, Linear Multilinear Algebra. 65 (2017), 1662–1682.

[5] Beik, F.P.A. and Ahmadi-Asl, S. *Residual norm steepest descent based iterative algorithms for Sylvester tensor equations*, J. Math. Model. 2 (2015), 115–131.

[6] Beik, F.P.A., Movahed, F.S. and Ahmadi-Asl, S. *On the krylov subspace methods based on the tensor format for positive definite sylvester tensor equations*, Numer. Linear Algebra. Appl. 23 (2016), 444–466.

[7] Brazell, M., Li, N., Navasca, C. and Tamon, C. *Solving multilinear systems via tensor inversion*, SIAM J. Matrix Anal. Appl. 34 (2013), 542–570.

[8] Brown, P.N. and Hindmarsh, A.C. *Reduced storage methods in sti ODE systems*, Appl. Math. Comput. 31 (1989), 40.

[9] Chen, Z. and Lu, L.Z. *A projection method and Kronecker product precondetioner for solving Sylvester tensor equations*, Science China 55 (2012), 1281–1292.

[10] Dehdezi, E.K. *Iterative methods for solving Sylvester transpose-tensor equation* $\mathcal{A}\star_N \mathcal{X} \star_M \mathcal{B}+\mathcal{C}\star_M \mathcal{X}^T \star_N \mathcal{D} = \mathcal{E}$, Operations Research Forum. 2 (2021), 64.

[11] Dehdezi, E.K. and Karimi, S. *Extended conjugate gradient squared and conjugate residual squared methods for solving the generalized coupled Sylvester tensor equations*, T. I. Meas. Control. 43 (2021), 519–527.

[12] Dehdezi, E.K. and Karimi, S. *A gradient based iterative method and associated preconditioning technique for solving the large multilinear systems*, Calcolo. 58 (2021), 51.

[13] Ding, F. and Chen. T. *Gradient based iterative algorithms for solving a class of matrix equations*, IEEE Trans Autom Control. 50 (2005), 1216–1221.

[14] Ding, F. and Chen, T. *Iterative least squares solutions of coupled Sylvester matrix equations*, Syst. Control Lett. 54 (2005), 95–107.

[15] Ding, W. and Wei, Y. *Solving multi-linear system with $\mathcal{M}$-tensors*, J Sci Comput. 68 (2016), 689–715.

[16] Erfanifar, R. and Hajarian, M. *Several efficient iterative algorithms for solving nonlinear tensor equation $\mathcal{X}+\mathcal{A}^T\star_M\mathcal{X}^{-1}\star_N\mathcal{A}=\mathcal{I}$ with Einstein product*, Comput. Appl. Math. 43 (2024), 84.

[17] Graham, A. *Kronecker products and matrix calculus: with applications*, Courier Dover Publications, 2018.

[18] Guennouni, A.E., Jbilou, K. and Riquet, A.J. *Block Krylov subspace methods for solving large Sylvester equations*, Numeric. Algorithms. 29 (2002), 75–96.

[19] Grasedyck, L. *Existence and computation of low Kronecker-rank approximations for large linear systems of tensor product structure*, Computing. 72 (2004), 247–265.

[20] Heyouni, M., Movahed, F.S. and Tajaddini, A. *A tensor format for the generalized hessenberg method for solving sylvester tensor equations*, J. Comput. Appl. Math. 377 (2020), 112878.

[21] Huang, B. and Ma, C. *Global least squares methods based on tensor form to solve a class of generalized Sylvester tensor equations*, Appl. Math. and Comput. 369 (2020), 124892.

[22] Huang, B. and Ma, C. *An iterative algorithm to solve the generalized Sylvester tensor equations*, Linear Multilinear Algebra. 68(6) (2020), 1175–1200.

[23] Huang, B., Xie, Y. and Ma, C. *Krylov subspace methods to solve a class of tensor equations via the Einstein product*, Numer. Linear Algebra Appl. 40(4) (2019), e2254.

[24] Jia, Z. *On IGMRES(q), incomplete generalized minimal residual methods for large unsymmetric linear systems*, Technical Report 94-047, Depart-

ment of Mathematics, University of Bielefeld, Sonderforschungsbereich 343, 1994. Last revision March, 1995.

[25] Kolda, T.G. and Bader, B.W. *Tensor decompositions and applications*, SIAM Rev. 51 (2009), 455–500.

[26] Lai, W.M., Rubin, D. and Krempl, E. *Introduction to continuum mechanics*, Oxford: Butterworth Heinemann, 2009.

[27] Li, B.W., Sun, Y.S. and Zhang, D.W. *Chebyshev collocation spectral methods for coupled radiation and conduction in a concentric spherical participating medium*, ASME J Heat Transfer. 131 (2009), 062701–62709.

[28] Li, B.W., Tian, S., Sun, Y.S. and Hu, Z.M. *Schur-decomposition for 3D matrix equations and its application in solving radiative discrete ordinates equations discretized by Chebyshev collocation spectral method*, J. Comput. Phys. 229 (2010), 1198–1212.

[29] Li, T., Wang, Q.-W. and Zhang, X.-F. *A Modified Conjugate Residual Method and Nearest Kronecker Product Preconditioner for the Generalized Coupled Sylvester Tensor Equations*, Mathematics. 10 (2022), 1730.

[30] Liang, L., Zheng, B. and Zhao, R.J. *Tensor inversion and its application to the tensor equations with Einstein product*, Linear Multilinear Algebra. 67 (2018), 843–870.

[31] Malek, A. and Masuleh, S.H.M. *Mixed collocation-finite difference method for 3D microscopic heat transport problems, J. Comput. Appl. Math.* 217 (2008), 137–147.

[32] Malek, A., Bojdi, Z.K. and Golbarg, P.N.N. *Solving fully three-dimensional microscale dual phase lag problem using mixed-collocation finite difference discretization*, J. Heat Transfer. 134 (2012), 0945041–0945046.

[33] Masuleh, S.H.M. and Phillips, T.N. *Viscoelastic flow in an undulating tube using spectral methods*, Computers & Fluids 33 (2004), 10751095.

[34] Qi, L. and Luo, Z. *Tensor analysis: Spectral theory and special tensors*, SIAM, Philadelphia, 2017.

[35] Saad, Y. *Iterative methods for sparse linear systems*, PWS press, New York, 1995.

[36] Saad, Y. and Wu, K. *DQGMRES: a direct quasi-minimal residual algorithm based on incomplete orthogonalization*, Numeric. Linear Algebra Appl. 3(4) (1996), 329–343.

[37] Shi, X., Wei, Y. and Ling, S. *Backward error and perturbation bounds for high order Sylvester tensor equation*, Linear Multilinear Algebra. 61 (2013), 1436–1446.

[38] Sun, L., Zheng, B., Bu, C. and Wei, Y. *Moore-Penrose inverse of tensors via Einstein product*, Linear Multilinear Algebra. 64 (2016), 686–698.

[39] Tzou, D.V. *Macro to Micro Heat Transfer*, Taylor & Francis: Washington, 1996.

[40] Wang, Q.W. and Wang, X. *A system of coupled two-sided sylvester-type tensor equations over the quaternion algebra*, Taiwanese J. Math. 24 (2020), 1399–1416.

[41] Wang, Q.W. and Xu, X. *Iterative algorithms for solving some tensor equations*, Linear Multilinear Algebra. 67(7) (2019), 1325–1349.

[42] Wang, Q.W., Xu, X.J. and Duan, X.F. *Least squares solution of the quaternion sylvester tensor equation*, Linear Multilinear Algebra. 69 (2021), 104–130.

[43] Xie, M.Y. and Wang, Q.W. *Reducible solution to a quaternion tensor equation*, Front. Math. China. 15 (2020), 1047–1070.

[44] Zhang, X.-F. and Wang, Q.-W. *Developing iterative algorithms to solve Sylvester tensor equations*, Appl. Math. Comput. 409 (2021), 126403.

[45] Zhang, X.-F., Ding, W. and Li, T. *Tensor form of GPBiCG algorithm for solving the generalized Sylvester quaternion tensor equations*, J. Franklin Institute. 360 (2023), 5929–5946.

Iranian Journal of Numerical Analysis and Optimization
Vol. 14, No. 3, 2024, pp 970–990
https://doi.org/10.22067/ijnao.2024.86481.1379
https://ijnao.um.ac.ir/

Research Article

A stabilized simulated annealing-based Barzilai–Borwein method for the solution of unconstrained optimization problems

H. Sharma* and R.K. Nayak

Abstract

The Barzilai–Borwein method offers efficient step sizes for large-scale unconstrained optimization problems. However, it may not guarantee global convergence for nonquadratic objective functions. Simulated annealing-based on Barzilai–Borwein (SABB) method addresses this issue by incorporating a simulated annealing rule. This work proposes a novel step-size strategy for the SABB method, referred to as the SABB*m* method. Furthermore, we introduce two stabilized variants: SABB*stab* and SABB*mstab*. SABB*stab* combines a simulated annealing rule with a stabilization step to ensure convergence. SABB*mstab* builds upon SABB*stab*,

*Corresponding author

Received 19 January 2024; revised 08 May 2024; accepted 11 May 2024

Hitarth Sharma
Department of Mathematics, International Institute of Information Technology, Bhubaneswar, Odisha, India, 751029. e-mail: hitarth@iiit-bh.ac.in

Rupaj Kumar Nayak
Department of Mathematics, International Institute of Information Technology, Bhubaneswar, Odisha, India, 751029. e-mail: rupaj@iiit-bh.ac.in

How to cite this article
Sharma, H. and Kumar Nayak, R., A stabilized simulated annealing-based Barzilai–Borwein method for the solution of unconstrained optimization problems. *Iran. J. Numer. Anal. Optim.*, 2024; 14(3): 970-990. https://doi.org/10.22067/ijnao.2024.86481.1379

incorporating the modified step size derived from the SABBm method. The effectiveness and competitiveness of the proposed methods are demonstrated through numerical experiments on CUTEr benchmark problems.

AMS subject classifications (2020): Primary 65K05; Secondary 90C06, 90C30.

Keywords: Unconstrained optimization; Barzilai–Borwein method; Simulated annealing method; Stabilized BB method.

1 Introduction

Researchers have shown considerable interest in unconstrained optimization problems due to their significant theoretical importance and practical applicability in the field of optimization. Its applications span various fields, including engineering, physics, finance, machine learning, and more. Furthermore, it is applicable for addressing a range of problems, including parameter estimation, function fitting, optimization of cost functions, and various others. The general form of an unconstrained optimization problem is

$$\min_{x\in\mathbb{R}^n} f(x), \tag{1}$$

where $f:\mathbb{R}^n\to\mathbb{R}$ is continuously differentiable. The iterative formula in the classical steepest-descent method [6] for the problem (1) is of the form

$$x_{k+1} = x_k + \alpha_k d_k, \tag{2}$$

where the search direction $d_k\in\mathbb{R}^n$ is determined as the negative gradient of f at x_k as

$$d_k = -\nabla f(x_k), \tag{3}$$

and the step size α_k is determined by

$$\alpha_k = \arg\min_{\alpha} f(x_k+\alpha d_k). \tag{4}$$

The above method is simple. However, it performs poorly as it exhibits linear convergence and is influenced by ill-conditioning [1]. Barzilai and Borwein [3] introduced two novel step sizes to be utilized together with the direction

of the negative gradient. Equation (2) requires less computation than (3), and the algorithm was also less sensitive to ill-conditioning. The step size α_k of the Barzilai–Borwein (BB) method [3] is given by

$$\alpha_k = \frac{s_{k-1}^\top s_{k-1}}{s_{k-1}^\top y_{k-1}} \tag{5}$$

or

$$\tilde{\alpha}_k = \frac{s_{k-1}^\top y_{k-1}}{y_{k-1}^\top y_{k-1}}, \tag{6}$$

where $s_{k-1} = x_k - x_{k-1}$, $y_{k-1} = \nabla f_k - \nabla f_{k-1}$. This approach has an R-superlinear convergence rate for a two-dimensional strictly convex quadratic objective function, and it outperforms the linearly convergent classical steepest-descent technique [6]. Subsequently, Raydan [20] established the global convergence of this method for n-dimensional strictly convex quadratic objective functions. The R-linear convergence rate was demonstrated by Dai and Liao [7]. On the other hand, Fletcher [11] contended that R-linear convergence is only anticipated in general cases. It is further confirmed that the BB method does not guarantee global convergence when the objective function is nonquadratic unless it is integrated with certain globalizing schemes.

The first nonmonotone line search framework for Newton's methods was presented by Grippo, Lampariello, and Lucidi [13], which has been applied in a number of later papers on nonmonotone line search techniques (see [4, 14, 16, 26]). Furthermore, nonmonotone line search approaches are also used by certain spectral gradient methods for optimization problems (see [23, 25]). Although these techniques are often effective, they do have certain drawbacks. Specifically, in the nonmonotone line search technique, the maximum value discards any good function value generated in an iteration. Furthermore, as shown in [13, 22], the selection of integer $M(M > 0)$ can have a significant impact on the numerical performance of nonmonotonic line search algorithms. Dai and Zhang [8] created an adaptive nonmonotone line search to overcome these two shortcomings, which is utilized in conjunction with the two-point gradient approach for optimization problems. A novel nonmonotone line search technique was later proposed by Zhang and Hager

[24], which is shown by numerical experiments to perform better than both monotone and traditional nonmonotone strategies.

Kirkpatrick, Gelatt, and Vecchi [15] first used the Simulated Annealing (SA) method initially introduced by Metropolis et al. [17], for addressing combinatorial optimization problems. Numerous authors have thoroughly examined the later SA approach for both discrete and continuous optimization issues. The SA approach is popular among researchers because it is capable of avoiding trapping in local minima. However, for large-scale problems, it is not acceptable because of the enormous computing cost.

Numerous hybrid algorithms that combine the SA method with other optimization techniques have been published, owing to their theoretical guarantee of convergence, enhanced performance in numerous real-world situations, and ease of implementation. Dong, Li, and Peng [10] presented a hybrid method that combines the BB method and the SA method. This nonmonotonic technique is called the Simulated Annealing-Based Barzilai–Borwein (SABB) method. Global convergence is also established under certain moderate assumptions in their research.

Numerous researchers have observed that the BB method might produce steps that deviate iterations too far from the optimal solution. Furthermore, it fails to converge even for strongly convex functions. The stabilized BB method, as proposed by Burdakov, Dai, and Huang [5], is a stabilized version of the BB method. The approach involves constraining the distance between each pair of consecutive iterates, a strategy that frequently reduces the number of BB iterations. In [5], the global convergence of this approach is also demonstrated for strongly convex functions with Lipschitz gradients. Barzilai and Borwein [3], followed by Raydan [20], demonstrated global convergence for the case of a strictly convex quadratic function, irrespective of the number of variables involved.

While SABB outperforms other similar algorithms, such as the adaptive two-point step-size gradient algorithm by Dai and Zhang [8] and the Barzilai–Borwein gradient method (GBB) by Raydan [21], it can generate too long steps, which may lead to the discarding of significant intermediate iterates. Motivated by this challenge, we introduce some novel variants of the SABB method.

In this paper, we first briefly discuss some existing methods to solve unconstrained optimization problems in Section 1. Then, in Section 2, we propose a modified version of the SABB method referred to as the SABBm method and two stabilized versions of the SABB method referred to as SABB*stab* and SABB*mstab*, respectively. In Section 3, we present the numerical results obtained through rigorous experimentation and analysis. This section provides valuable insights into the effectiveness and efficiency of the proposed variants. Finally, Section 4 provides a concise summary of the conclusions derived from our extensive research findings. Within this section, we deliberate on the significance of each variant and its implications concerning unconstrained optimization problems.

1.1 Simulated annealing-based Barzilai–Borwein (SABB) method

When applied to nonquadratic objective functions, the original BB technique does not provide global convergence. To tackle this problem, the BB method was combined with the SA approach and introduced as the SABB method by the authors [10]. In order to approve the BB step, the method incorporates an SA criterion. If the BB step is deemed unacceptable, then an Armijo line search method is employed. Under certain mild conditions, the global convergence of the SABB technique is established. The BB step of the SABB method given by Dong, Li, and Peng [10] is

$$\alpha_k = \max\left\{\alpha_{\min}, \min\left\{\alpha_{\max}, \frac{s_{k-1}^\top s_{k-1}}{s_{k-1}^\top y^{k-1}}\right\}\right\}, \tag{7}$$

where $s_{k-1} = x_k - x_{k-1}, y_{k-1} = \nabla f_k - \nabla f_{k-1}$ and $\alpha_{\min}, \alpha_{\max} > 0$ are two fixed parameters.

1.2 Stabilized BB (BB*stab*) method

Occasionally, the BB approach produces excessively long steps, causing the iterates to deviate too far from the optimal solution. The objective function might not converge even if it is strongly convex. In that situation, a basic modification can make it convergent, and the stabilized version preserves the rapid local convergence of the BB method as the number of stabilization steps is established to be finite. The step size of the stabilized BB method given by Burdakov, Dai, and Huang [5] is

$$\alpha_k = \min\{\alpha_k^{\mathrm{BB}}, \alpha_k^{\mathrm{Stab}}\} \tag{8}$$

where

$$\alpha_k^{BB} = \frac{s_{k-1}^{\top} s_{k-1}}{s_{k-1}^{\top} y_{k-1}} \tag{9}$$

with $s_{k-1} = x_k - x_{k-1}, y_{k-1} = \nabla f_k - \nabla f_{k-1} = g_k - g_{k-1}$, and

$$\alpha_k^{\mathrm{Stab}} = \frac{D}{\|g_{k-1}\|}, \tag{10}$$

where $D > 0$ is a parameter to be chosen in a particular way.

2 Variants of the SABB method

In this section, we introduce our proposed variants of the SABB method. The step-size selection strategy within SABB is critical for achieving efficient convergence towards the minimum. Therefore, the proposed SABB variants are distinguished by their step-size selection criteria.

2.1 Modified SABB (SABB*m*) method

Mu and Liu [18] employed a BB method to solve the unconstrained optimization subproblem arising within the Augmented Lagrangian Method (ALM) for large-scale binary quadratic programming. This choice leverages the BB method's low computational cost and effectiveness in achieving near-optimal

solutions, making the resulting ALM approach a viable option for these large-scale problems. Here, we propose a new step size for the SABB method, which is a modified version of the step size given in [18]. The step size of our method has an additional parameter $\exp(d)$, where $d \in (0,1)$. The inclusion of $\exp(d)$ and $\xi^{l_{\min}}$ in the SABB step size is motivated by their demonstrated effectiveness. They enable the modified step size to achieve the same results in fewer iterations and with reduced computation time. Here we introduce a new step for the SABB technique, which is presented as

$$\alpha_k^{\text{SABBm}} = \begin{cases} \exp(d)\xi^{l_{\min}}\alpha_0 & \text{if } k = 0, \\ \exp(d)\xi^{l_{\min}} \min\left\{\alpha_{\max}, \max\left\{\alpha_{\min}, \frac{\|s_k\|_2^2}{s_k^\top y_k}\right\}\right\} & \text{if } k \geq 1, \end{cases} \tag{11}$$

where $\alpha_{\min}, \alpha_{\max} > 0$ are two fixed parameters, $s_k = x_{k+1} - x_k, y_k = \nabla f_{k+1} - \nabla f_k$, ξ is the given parameter, the chosen parameters $d, \sigma \in (0,1)$, and $l_{\min}$ is the smallest nonnegative integer l satisfying

$$f(x_k + \xi^l \alpha_k \nabla f_k) \leq f(x_k) + \sigma \xi^l \nabla^\top f_k \alpha_k \nabla f_k, \tag{12}$$

where

$$\alpha_k = -\min\left\{\alpha_{\max}, \max\left\{\alpha_{\min}, \frac{\|s_k\|_2^2}{s_k^\top y_k}\right\}\right\}.$$

The approach used in (12) to obtain the value of $l_{\min}$ is inspired by the nonmonotone line search framework developed by Grippo, Lampariello, and Lucidi [13]. We now propose the SABBm as follows:

2.1.1 Convergence of SABBm method

To establish global convergence, we require the following remarks and assumptions.

Remark 1. Armijo line search: let m_k be the smallest integer that satisfies $f(x_k - \delta^m \nabla f_k) \leq f_k - c\delta^m \|\nabla f_k\|^2$ where $\delta \in (0,1)$. Then $\alpha_{k+1}^{\text{SABB}m} = \delta^{m_k}$.

Assumption 1. The set $Z_0 = \{x \in \mathbb{R}^n : f(x) \leq f(x_0) + (1-\gamma)^{-1} T_0 \eta\}$ is bounded and closed.

Algorithm 5: SABBm method

1: **Give an initial point** $x_0 \in \mathbb{R}^n$. Let $\epsilon > 0$, $0 < \alpha_{\min} < \alpha_{\max} < \infty$, $\alpha_0 \in (\alpha_{\min}, \alpha_{\max})$, $T_0 > 0, c, \gamma, \xi, d, \sigma \in (0,1), \eta \in \mathbb{Z}_+$. Set $k := 0$ and $\alpha_0^{\text{SABBm}} = \exp(d)\xi^{l_{\min}}\alpha_0$

2: If $\|\nabla f_k\| < \epsilon$, then stop

3: Compute $z_k = x_k - \alpha_k^{\text{SABB}m}\nabla f_k$
and $\Delta f_k = f(z_k) - (f_k - c\alpha_k^{\text{SABBm}}\|\nabla f_k\|^2)$

4: Let $p_k = e^{-\frac{\Delta f_k}{T_k}}$ and pick a random number $r_k \in \left(e^{-\eta}, e^{-\frac{1}{\eta}}\right)$.

5: If $p_k \geq r_k$, let $x_{k+1} = z_k$ and go to step 7.

6: Otherwise, let $\alpha_k^{\text{SABB}m}$ be a step size determined by the Armijo line search and $x_{k+1} = x_k - \alpha_k^{\text{SABB}m}\nabla f_k$.

7: Compute $\alpha_{k+1}^{\text{SABB}m}$ using (11).

8: Let $T_{k+1} = \gamma T_k; k = k+1$ and go to step 2.

Lemma 1. The step size $\alpha_k^{\text{SABB}m}$ in the SABBm method satisfies the condition $\alpha_k^{\text{SABB}m} \geq \alpha$, for all $k \geq 0$, where $\alpha \geq 0$ is a constant.

Proof. There are two cases for the step size of the SABBm method: one is by the Armijo line search, while the other is by (11). In the case of the Armijo line search method from [19], by the property of termination in finite steps, there exists an integer $N > 0$ such that $\alpha_k^{\text{SABB}m} > \delta^N$, for all $k \geq 0$. In the second case, from (11), it is clear that $\alpha_k^{SABB}m > \alpha_{\min}$. Let $\alpha = \min\{\delta^N, \alpha_{\min}\}$. The result is obtained, and this concludes the proof. □

Lemma 2. If Assumption 1 holds, then the SABBm method is well defined, and the sequence generated by the SABBm method is $\{x_k\} \subset Z_0$.

Proof. In Algorithm 5, we consider $p_k = e^{-\frac{\Delta f_k}{T_k}}$, and if $p_k \geq r_k$, then $\Delta f_k \leq -T_k \ln r_k$, which implies

$$f(z_k) \leq f_k - c\alpha_k^{\text{SABB}m}\|\nabla f_k\|^2 - T_k \ln r_k.$$

Let

$$\mu_k = \begin{cases} 1 & \text{if } p_k \geq r_k, \\ 0 & \text{otherwise.} \end{cases}$$

Then, the functional value sequence $\{f_k\}$ generated by SABBm satisfies

$$f_{k+1} \leq f_k - c\alpha_k^{\mathrm{SABB}m}\|\nabla f_k\|^2 - \mu_k T_k \ln r_k. \tag{13}$$

Since $e^{-\eta} < r_k < e^{-\frac{1}{\eta}}$ and $\left(e^{-\eta}, e^{-\frac{1}{\eta}}\right) \subset (0,1)$, we have $\ln r_k < -\frac{1}{\eta} < 0$. Again since, $0 \leq \mu_k \in \{0,1\}$ and $T_k > 0$, we have $-\mu_k T_k \ln r_k \geq 0$. Hence, if the step size $\alpha_k^{\mathrm{SABB}m}$ is determined by the Armijo line search method, we get

$$f_{k+1} \leq f_k - c\alpha_k^{\mathrm{SABB}m}\|\nabla f_k\|^2. \tag{14}$$

Thus, one can conclude that (14) implies (13). Therefore, based on the termination property within a finite number of steps of the Armijo line search [19], the method is well defined.
By (13), we get

$$\begin{aligned}
f_{k+1} &\leq f_k - c\alpha_k^{\mathrm{SABB}m}\|\nabla f_k\|^2 - \mu_k T_k \ln r_k \\
&\leq f_k - \mu_k T_k \ln r_k \\
&\leq f_k - T_k \ln r_k (\text{as } \mu_k = 1 \text{ if } p_k \geq r_k;\ \text{otherwise } \mu_k = 0) \\
&= f_0 - \sum_{i=1}^{k} T_i \ln r_i \\
&= f_0 - \sum_{i=1}^{k} \gamma^i T_0 \ln r_i \\
&\leq f_0 - T_0\eta \sum_{i=1}^{k} \gamma^i \\
&< f_0 + (1-\gamma)^{-1} T_0 \eta
\end{aligned}$$

for all $k \geq 0$. Therefore, $\{x_k\} \subset Z_0$. □

Here is the proof of the global convergence of the SABBm method.

Theorem 1. Under Assumption 1, $\lim_{k\to\infty} \|\nabla f(x_k)\| = 0$, where the sequence $\{x_k\}$ is generated by the SABBm method.

Proof. By Lemma 2, we have $\{x_k\} \subset Z_0$. Since Z_0 is compact, the sequence $\{x_k\}$ is convergent. By (13), for $k \geq 0$

$$f_{k+1} \leq f_k - c\alpha_k^{\mathrm{SABB}m}\|\nabla f_k\|^2 - \mu_k T_k \ln r_k$$

$$\leq f_k - c\alpha_k^{\mathrm{SABB}m}\|\nabla f_k\|^2 - T_k \ln r_k,$$

where $-T_k \ln r_k > 0$ and $\mu_k \in \{0, 1\}$. Combining this with Lemma 1, we get

$$c\alpha\|\nabla f_k\|^2 \leq c\alpha_k^{\mathrm{SABB}m}\|\nabla f_k\|^2 \leq f_k - f_{k+1} - T_k \ln r_k. \tag{15}$$

Summarizing (15) from $k = 0$ to K, we get

$$c\alpha \sum_{k=0}^{K}\|\nabla f_k\|^2 \leq f_0 - f_{K+1} - \sum_{k=0}^{K} T_k \ln r_k. \tag{16}$$

By Assumption 1 and the continuity of $f(x)$ on Z_0, $f(x) \geq \beta$ holds for all $x \in Z_0$, where $\beta < \infty$ is a real number. Therefore, by taking limits on (16) as $K \to \infty$, we obtain

$$\begin{aligned}
\sum_{k=0}^{\infty}\|\nabla f_k\|^2 &= \sum_{k\geq 0}\|\nabla f_k\|^2 \\
&\leq \frac{f_0 - f_{K+1} - \sum_{k\geq 0} T_k \ln r_k}{c\alpha} \\
&= \frac{f_0 - f_{K+1} - \sum_{k\geq 0} \gamma^k T_0 \ln r_k}{c\alpha} \\
&\leq \frac{f_0 - f_{K+1} + T_0\eta \sum_{k\geq 0} \gamma^k}{c\alpha} \\
&\leq \frac{f_0 - f_{K+1} + (1-\gamma)^{-1}\eta T_0}{c\alpha} \\
&\leq (f_0 - \beta + (1-\gamma)^{-1}\eta T_0)/c\alpha.
\end{aligned}$$

□

2.2 Stabilized SABB (SABB*stab*) method

The method proposed here combines the SABB method from [10] with the BB*stab* method from [5], resulting in a modified version of the BB*stab* method [5]. Two-step size values, namely α_k^{SABB} and α_k^{stab}, are computed using (7) and (10), respectively, and their minimum value is used as the required step size $\alpha_k^{\mathrm{SABB}stab}$. Note that it utilizes two distinct initial values x_0 and x_1

instead of that of SABB and SABBm. We now present the SABB$stab$ algorithm.

Algorithm 6: SABB$stab$ method

1: Give two initial points $x_0, x_1 \in \mathbb{R}^n$ such that $x_0 \neq x_1$ and a scalar $\Delta > 0$. Let $\epsilon > 0, c \in (0,1), 0 < \alpha_{\min} < \alpha_{\max} < \infty$, $\alpha_1 \in [\alpha_{\min}, \alpha_{\max}], T_0 > 0, \gamma \in (0,1), \eta \in \mathbb{Z}_+$. Set $k := 1$

2: If $\|\nabla f_k\| < \epsilon$, then stop

3: Compute $z_k = x_k - \alpha_k \nabla f_k$ and $\Delta f_k = f(z_k) - (f_k - c\alpha_k \|\nabla f_k\|^2)$

4: Let $p_k = e^{-\frac{\Delta f_k}{T_k}}$ and pick a random number $r_k \in \left(e^{-\eta}, e^{-\frac{1}{\eta}}\right)$.

5: If $p_k \geq r_k$ let $x_{k+1} = z_k$ and go to Step 6. Otherwise, let α_k^{BB} be a step size determined by the Armijo line search and $x_{k+1} = x_k - \alpha_k^{BB} \nabla f_k$.

6: Compute $\alpha_{k+1}^{\mathrm{BB}} = \max\left\{\alpha_{\min}, \min\left\{\alpha_{\max}, \frac{s_k^\top s_k}{s_k^\top y^k}\right\}\right\}$, where $s_k = x_{k+1} - x_k, y_k = \nabla f_{k+1} - \nabla f_k$ and Compute $\alpha_{k+1}^{\mathrm{stab}} = \frac{\Delta}{\|\nabla f_k\|}$.

7: Compute $\alpha_{k+1} = \min\{\alpha_{k+1}^{\mathrm{BB}}, \alpha_{k+1}^{\mathrm{stab}}\}$

8: Let $T_{k+1} = \gamma T_k; k = k+1$ and go to first Step 3.

2.2.1 Convergence of SABB*stab* Method

Assumption 2. There exist positive constants $\Lambda_1 \leq \Lambda_2$, and the function $f : \mathbb{R}^n \to \mathbb{R}^1$ is twice continuously differentiable such that

$$\Lambda_1 \|\nabla v\|^2 \leq \ v^\top \nabla^2 f(x) v \leq \Lambda_2 \|v\|^2 \ \text{ for all } \ x, v \in \mathbb{R}^n.$$

Assumption 3. For some $\rho > 0$ and $L \geq 0$, the following property holds:

$$\left\|\nabla^2 f(x) - \nabla^2 f(x^*)\right\| \leq \ L\|x - x^*\| \ \text{ for all } \ x \in \ B_\rho(x^*),$$

where L is a Lipschitz constant, ρ is a radius, and $B_\rho(x^*) = \{x \in \mathbb{R}^n : \|x - x^*\| \leq \rho\}$.

If Assumptions 1, 2, and 3 hold, then our proposed method SABB$stab$ is well defined by the Lemma 2. Note that in Algorithm 6, if $\alpha_{k+1} = \alpha_{k+1}^{stab}$, then it is convergent, according to [5, Theorem 3.2] method, and if $\alpha_{k+1} = \alpha_{k+1}^{\mathrm{BB}}$, then it is convergent, according to [10, Theorem 1].

2.3 Modified stabilized SABB (SABB*mstab*) method

In this section, we present a stabilized version of the SABBm method termed SABB*mstab*, which is our main proposal. Similar to the previous method, this approach also utilizes two different initial values, x_0 and x_1. In this method, we propose to choose

$$\alpha_k = \min\{\alpha_k^{SABB}m, \alpha_k^{stab}\}.$$

We now present the SABB*mstab* algorithm below.

Algorithm 7: SABB*mstab* method

1: **Give initial points** $x_0, x_1 \in \mathbb{R}^n$ such that $x_0 \neq x_1$ and a scalar $\Delta > 0$. Let $\epsilon > 0, c \in (0,1), 0 < \alpha_{\min} < \alpha_{\max} < \infty$, $\alpha_1 \in [\alpha_{\min}, \alpha_{\max}], T_0 > 0, \gamma \in (0,1), \eta \in \mathbb{Z}_+$. Set $k := 1$

2: If $\|\nabla f_k\| < \epsilon$, then stop

3: Compute $z_k = x_k - \alpha_k \nabla f_k$ and $\Delta f_k = f(z_k) - (f_k - c\alpha_k \|\nabla f_k\|^2)$

4: Let $p_k = e^{-\frac{\Delta f_k}{T_k}}$ and pick a random number $r_k \in \left(e^{-\eta}, e^{-\frac{1}{\eta}}\right)$.

5: If $p_k \geq r_k$ let $x_{k+1} = z_k$ and go to step 6. Otherwise, let α_k^{BB} be a step size determined by the Armijo line search and $x_{k+1} = x_k - \alpha_k \nabla f_k$.

6: Compute $\alpha_{k+1} = \min\{\alpha_{k+1}^{SABBm}, \alpha_{k+1}^{stab}\}$, where α_{k+1}^{SABBm} is computed from (11) and $\alpha_{k+1}^{\text{stab}} = \frac{\Delta}{\|\nabla f_k\|}$.

7: Let $T_{k+1} = \gamma T_k; k = k+1$ and go to Step 2.

Note that in the SABB*mstab* method, we only replace the step size of the SABB*stab* with the step given in (11). Since SABB*stab* is convergent, SABB*mstab* method is also convergent.

3 Numerical experiments

In this section, we perform numerical experiments to showcase the effectiveness of all variants of the SABB method. In our entire study, we discuss two types of algorithms. The performances of the proposed methods SABBm, SABB*stab*, and SABB*mstab* are compared with the performance of the SABB method described in the research [10]. For all our experiments, we imple-

mented the algorithms using RStudio. The computations were performed on a laptop equipped with an Intel(R) Core(TM) i3 CPU at 3.20 GHz and 4GB of memory. We focus on demonstrating the performance differences between stabilized and nonstabilized variants of SABB. The algorithms SABB and SABBm are terminated when the number of iterations exceeds 3000, or $\|\nabla f_k\| \leq 10^{-6}$. In the stabilized version, we compute two step sizes in each iteration and choose their minimum. The step-size computation becomes a time-consuming task; therefore, we terminate the algorithms SABB$stab$ and SABB$mstab$ when the iterations exceed 10^5, or $\|\nabla f_k\| \leq 10^{-6}$. In all four methods, if the above termination criteria fail, then we claim that the method fails, and we reflect it in Table 2 as "F".

We observe that the selection of α_0, $\alpha_{\min}$, and $\alpha_{\max}$ plays a crucial role in the computation. Similarly, for the SABB$stab$ and SABB$mstab$ methods, along with these three values, the selection of the second initial value x_1 plays a crucial role in the convergence of these methods. For these two methods, we need the value of Δ also, which we compute using the formula $\Delta = \|x_k - x_{k-1}\|$ for $k = 1, 2, \ldots, n$. To identify suitable parameters for our methods, we solved instances of different dimensions available in the CUTEr library and observed that the accuracy in our computation is much better when we set $2^{-30} \leq \alpha_{\min} \leq \alpha_{\max} \leq 2^{20}$, $\alpha_0 = 2^{-10}(\alpha_{\min} + \alpha_{\max}), \gamma = 0.99, T_0 = 1000, c = 10^{-4}$, and $\eta = 20$.

The test functions reflected in Table 1 are taken from the CUTEr library [2, 12]. Table 2 presents the numerical results, where $IP(n)$ denotes the problem serial number with the dimension of the problem. We have taken the dimension from 2 to 500. In the columns under the methods such as SABB, SABBm, SABB$stab$, and SABB$mstab$, three computations are mentioned, such as *feval*, *iter*, and *cput*, which represent the number of function evaluations, the number of iterations, and CPU time (measured in seconds), respectively.

Table 1: Test functions

IP	Function Name	IP	Function Name
1	Extended Freudenstein and Roth Function:	40	Extended BD1 function (Block Diagonal):
2	Extended Trigonometric Function:	41	Extended Maratos Function:
3	Extended Rosenbrock Function:	42	Perturbed quadratic diagonal Function:
4	Generalized Rosenbrock Function:	43	Extended Wood Function:
5	Extended White and Holst function:	44	Quadratic QF1 Function:
6	TRIDIA function (CUTE):	45	Extended quadratic penalty QP1 Function:
7	Extended Beale function:	46	Extended quadratic penalty QP2 Function:
8	Extended Penalty function:	47	Quadratic QF2 Function:
9	ARGLINB function (CUTE):	48	Extended quadratic exponential EP1 function:
10	FLETCHCR function (CUTE):	49	POWER function (CUTE):
11	ARWHEAD function (CUTE):	50	ENGVAL1 function (CUTE):
12	EG2 function (CUTE):	51	EDENSCH function (CUTE):
13	Partial Perturbed Quadratic function (CUTE):	52	CUBE function (CUTE):
14	Almost Perturbed Quadratic function:	53	Extended quadratic exponential EP1 function:
15	NONDIA function (CUTE):	54	Perturbed Tridiagonal Quadratic function:
16	Staircase 1 function:	55	Staircase 2 function:
17	LIARWHD function (CUTE):	56	DQDRTC function (CUTE):
18	Extended Freudenstein and Roth Function:	57	Perturbed Tridiagonal Quadratic function:
19	BDQRTIC function (CUTE):	58	BIGGSB1 Function (CUTE):
20	NONDQUAR function (CUTE):	59	Extended DENSCHNB Function (CUTE):
21	Broyden Tridiagonal function (CUTE):	60	Generalized Quartic Function:
22	ARGLINC function (CUTE):	61	Diagonal 8 Function:
23	BDEXP function (CUTE):	62	Full Hessian FH3 Function:
24	NONSCOMP function (CUTE):	63	SINCOS Function:
25	QUARTC function (CUTE):	64	HIMMELH Function (CUTE):
26	Extended DENSCHNF Function (CUTE):	65	Raydan 1 Function:
27	DIXON3DQ Function (CUTE):	66	Raydan 2 Function:
28	COSINE Function (CUTE):	67	DIXMAANA Function:
29	Diagonal 7 Function:	68	DIXMAANB Function:
30	Diagonal 9 Function:	69	DIXMAANC Function:
31	HIMMELBG Function (CUTE):	70	DIXMAAND Function:
32	Diagonal 1 Function:	71	DIXMAANF Function:
33	Diagonal 2 Function:	72	DIXMAANH Function:
34	Diagonal 3 Function:	73	Extended TET Function (Three Exponential Terms):
35	Generalized Tridiagonal 1 Function:	74	Diagonal 4 Function:
36	Extended Tridiagonal 1 Function:	75	Diagonal 5 Function:
37	Generalized PSC1 Function:	76	Extended Himmelblau Function:
38	Extended PSC1 Function:	77	Generalized White and Holst Function:
39	Full Hessian FH2 Function:		

Table 2: Test results

IP(n)	SABB			SABB*m*			SABB*stab*			SABB*mstab*		
	feval	iter	cput	feval	iter	cput	feval	iter	cput	feval	iter	cput
1(2)	573	142	0.1159	53	12	0.0131	545	136	0.1844	48	12	0.0153
2(2)	873	217	0.3042	97	23	0.0291	681	170	0.4070	72	18	0.0565
3(2)	12005	3000	4.4121	3001	749	0.9078	38297	9574	25.9366	5172	1293	3.4904
4(2)	12005	3000	5.3538	3001	749	1.3600	38297	9574	25.1429	5172	1293	3.1331
5(2)	12005	3000	5.9533	3653	912	1.7136	39965	9991	28.1861	5396	1349	3.7478
6(2)	641	159	0.1028	521	129	0.0807	665	166	0.1515	540	135	0.1224
7(2)	3489	871	0.9035	3229	806	0.7592	400001	100000	5.4008	400000	100000	2.1504
8(2)	1157	288	0.2649	941	234	0.2008	1189	297	0.9345	972	243	0.3556
9(2)	20	4	0.0168	24	5	0.0161	33	8	0.0491	32	8	0.0130
10(10)	1906	503	3.2912	1886	498	3.0315	5	1	0.0181	4	1	0.0150
11(10)	104	25	0.2370	116	28	0.2753	385	96	2.1929	308	77	0.9040
12(10)	12005	3000	10.8701	12005	3000	10.1288	53605	13401	1.6342	43408	10852	1.3426
13(10)	249	61	0.9032	201	49	0.6771	213	53	1.0810	168	42	0.8606
14(10)	529	131	0.5164	429	106	0.3814	441	110	0.6078	356	89	0.5343
15(10)	6789	1696	4.3888	6209	1551	3.9080	51641	12910	2.5466	50620	12655	1.8658
16(10)	2244	560	2.7465	2503	625	3.2653	4813	1203	7.0123	56936	14234	1.7390
17(10)	1593	397	0.7971	1561	389	0.8241	3641	910	2.1399	3568	892	2.0244
18(40)	383	96	1.6555	445	112	2.2936	F	F	F	F	F	F
19(40)	606	152	3.4532	797	200	4.4500	4001	1000	5.0125	1700	425	1.5056
20(40)	12001	3000	55.9570	11999	3000	1.2298	8001	2000	3.1065	8000	2000	3.4526
21(40)	171	42	1.8844	167	41	1.8661	F	F	F	F	F	F
22(40)	1324	332	56.2828	11999	3000	10.4524	1957	489	35.0812	1840	460	23.7643
23(40)	12005	3000	2.2816	12005	3000	2.3259	32001	8000	15.1335	32000	8000	13.5988
24(40)	11972	3000	20.2500	11982	3000	21.5313	8001	2000	2.2935	8000	2000	2.0169
25(40)	1377	343	3.5877	1373	342	3.2536	8001	2000	2.1647	8000	2000	1.7061
26(40)	104	25	0.3547	108	26	0.3713	4001	1000	2.4536	264	66	10.2070
27(40)	1467	368	3.5020	1398	351	3.0541	5	1	0.0238	4	1	0.0369
28(40)	225	55	0.6246	165	40	0.4936	4001	1000	1.8497	4000	1000	1.9239
29(40)	44	10	0.1541	44	10	0.1378	4001	1000	1.5144	1628	407	41.7341
30(40)	11997	3000	29.1681	11982	3000	28.9967	4001	1000	1.2257	4000	1000	1.1827
31(40)	12005	3000	32.1075	12005	3000	32.4967	32001	8000	8.7434	32000	8000	8.7530
32(40)	12003	3000	6.6654	340	84	11.8590	F	F	F	F	F	F
33(40)	224	55	8.7581	236	58	3.8310	4001	1000	3.0450	4000	1000	3.0237
34(40)	12003	3000	2.8101	228	56	2.8166	621	155	30.4833	820	205	38.2410
35(40)	113	27	3.0366	109	26	2.4956	F	F	F	F	F	F
36(40)	1209	301	20.4789	1425	355	23.4677	40001	10000	11.3231	40000	10000	12.7731
37(40)	12003	3000	8.3764	12003	3000	8.2833	40001	10000	25.0045	40000	10000	25.0573
38(40)	57	13	1.4227	57	13	1.4582	1249	312	1.7140	1248	312	1.7120
39(40)	953	240	12.6938	1377	346	18.4661	F	F	F	F	F	F
40(40)	300	74	5.5401	304	75	5.8676	F	F	F	F	F	F
41(40)	64	15	0.9178	64	15	0.9510	F	F	F	F	F	F
42(40)	196	48	2.1041	288	71	2.6848	7201	1800	31.1102	7188	1797	27.9218
43(40)	3456	885	51.5988	6748	1731	1.6669	4001	1000	3.2047	4000	1000	2.1038
44(40)	247	61	1.9867	291	72	1.9010	817	204	8.7129	812	203	8.4563
45(40)	112	27	1.3176	96	23	1.0800	1469	367	32.8439	1464	366	33.5307
46(40)	247	62	3.4999	477	121	6.5672	5	1	0.1461	4	1	0.1123
47(40)	334	83	3.0507	335	83	2.9854	4001	1000	49.9540	4000	1000	49.8391
48(40)	313	77	5.7917	325	80	5.8702	7085	1771	28.6868	7072	1768	27.9670
49(50)	6503	1630	12.5130	7159	1796	11.1532	5	1	0.0179	4	1	0.0170
50(50)	2461	614	7.5193	1125	280	3.3699	797	199	3.6244	780	195	3.6969
51(50)	433	107	1.8298	421	104	1.7742	749	187	4.4861	736	184	4.4655
52(50)	11996	3000	28.3663	11998	3000	28.9950	5	1	0.0229	4	1	0.0181
53(50)	41	9	1.8297	41	9	1.0646	37	9	1.9401	36	9	1.2722
54(50)	436	108	24.9522	412	102	27.8671	1981	495	4.1695	1976	494	1.6517
55(50)	2346	588	6.9273	2375	595	7.4707	5	1	1.0590	4	1	0.7714
56(50)	158	39	10.1396	169	42	6.9630	9013	2253	2.7833	9012	2253	3.6513
57(50)	327	81	11.9068	327	81	7.0950	1341	335	1.1654	1340	335	1.2965
58(50)	1286	322	40.9906	1223	306	25.9469	5	1	0.0379	4	1	0.0302
59(100)	657	163	5.8164	533	132	4.3743	657	164	8.2231	532	133	6.4424
60(100)	561	139	7.0155	453	112	4.9969	397	99	7.3207	320	80	6.3334

Table 2 Continued...

61(200)	357	88	17.6759	289	71	10.9239	405	101	32.2711	328	82	22.8033
62(200)	25	5	1.5282	80	19	4.6503	21	5	1.7963	48	12	3.9969
63(200)	569	141	33.5531	401	99	22.7348	1413	353	2.1739	1156	289	1.6964
64(200)	621	154	24.6155	505	125	20.2527	645	161	39.3083	524	131	31.9428
65(200)	12005	3000	5.3382	12005	3000	5.5383	2533	633	1.6214	2100	525	1.1722
66(200)	1449	361	39.0737	1057	263	28.5491	893	223	35.2179	732	183	28.7968
67(300)	45	10	16.0129	45	10	16.3452	261	65	19.8307	260	65	21.2720
68(300)	37	8	12.1318	37	8	12.1640	493	123	43.9275	492	123	1.1385
69(300)	37	8	11.2262	37	8	11.6436	929	232	7.1460	928	232	7.2529
70(300)	72	17	23.1260	72	17	23.8315	1881	470	15.0397	1876	469	15.0572
71(300)	968	241	35.4163	981	244	35.2606	4001	1000	41.7712	4000	1000	39.0339
72(300)	589	147	24.1729	683	171	27.5352	F	F	F	F	F	F
73(500)	109	26	1.0931	121	29	49.6553	77	19	46.0448	76	19	35.9698
74(500)	33	7	4.4168	56	13	4.8235	5	1	0.7420	4	1	0.5555
75(500)	40	9	13.7751	40	9	8.7212	5	1	1.0040	4	1	0.6785
76(500)	136	33	26.7670	136	33	17.6005	2893	723	12.9185	2884	721	9.4026
77(500)	61	14	1.0407	61	14	52.2529	4525	1131	1.0296	4516	1129	24.5784

In Table 3, we present the number of problems for which the method achieves the least iter, the least cput, and the least feval, respectively. The observations from Table 3 lead to the conclusion that the SABBm method exhibits superior performance compared to SABB, and SABB$mstab$ outperforms SABB$stab$ in the stabilized version. We also observe that for problems $10, 14, 27, 46, 49, 52, 55, 58, 74$, and 75, the stabilized versions SABB$mstab$ and SABB$stab$ exhibit superior performance in terms of iterations, function evaluations, and time.

Table 3: Least table

Metric	**SABB**	**SABB**m	**SABB**$stab$	**SABB**$mstab$
feval	31	**33**	1	23
iter	34	**36**	14	19
cput	17	**30**	8	22

We employed the performance profile of Dolan and Moré [9] to compare the performance of the proposed methods. We construct the performance profile graphs for three key metrics: function evaluations (feval), number of iterations (iter), and CPU time (cput). Figure 1 illustrates the function evaluations performance profile. Figure 2 depicts the performance profile for number of iterations, and Figure 3 shows the performance profile of CPU time. Across all three metrics (Figures 1–3), SABBm is almost the top curve, indicating its superior efficiency compared to the other methods.

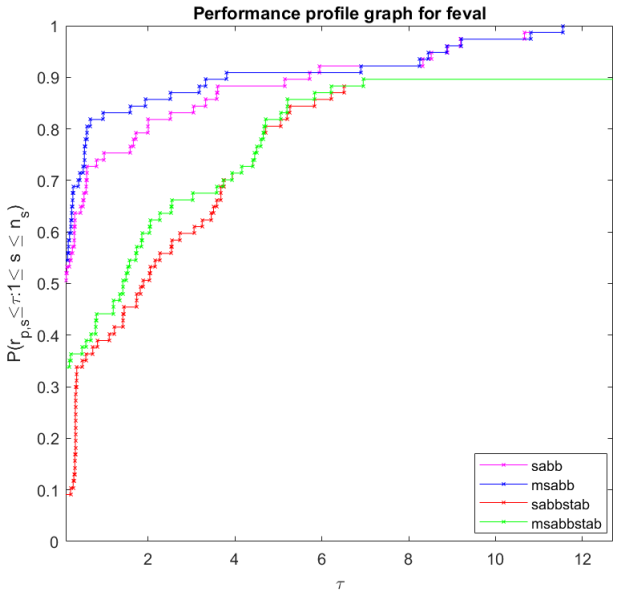


Figure 1: Performance profile for functional evaluations.

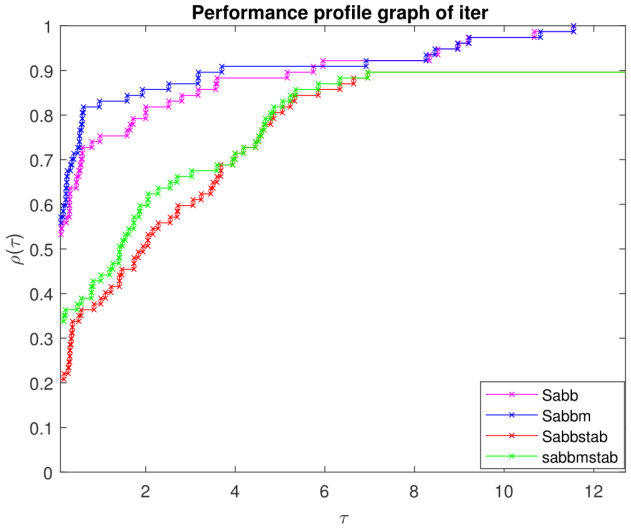


Figure 2: Performance profile for number of iterations.

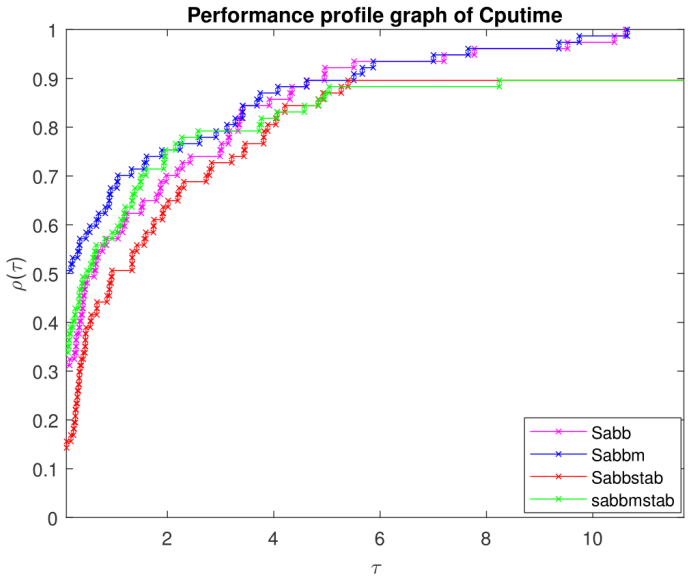


Figure 3: Performance profile for CPU time.

4 Conclusion

This study proposes three novel variants of the SABB algorithm for solving unconstrained optimization problems. These variants hybridize the BBStab and SABB approaches. The performance of the proposed methods is evaluated on a set of 77 benchmark problems from the CUTEr library (details in Table 2). The results reveal that SABBm emerges as the most efficient algorithm in terms of function evaluations, iterations, and CPU time. However, SABB$stab$ and SABB$mstab$ outperform SABBm in a small subset of problems regarding the number of iterations and computational time.

Compliance with Ethical Standards

- This article does not contain any studies involving animals performed by any of the authors.

- This article does not contain any studies involving human participants performed by any of the authors.

Acknowledgments

This research work is fully supported by the National Board for Higher Mathematics (NBHM), Department of Atomic Energy, Govt. of India (Grant no. 02011/22/2021 NBHM (R.P.)/R&D II/900 Date 06/08/2021).

Conflict of interest

The authors have no conflicts of interest to declare that are relevant to the content of this article.

References

[1] Akaike, H. *On a successive transformation of probability distribution and its application to the analysis of the optimum gradient method*, Annals of the Institute of Statistical Mathematics 11(1) (1959), 1–16.

[2] Andrei, N. *An unconstrained optimization test functions collection*, Adv. Model. Optim. 10(1) (2008), 147–161.

[3] Barzilai, J. and Borwein, J.M. *Two-point step size gradient methods*, IMA J. Numer. Anal. 8(1) (1988), 141–148.

[4] Birgin, E.G., Martínez, J.M. and Raydan, M. *Nonmonotone spectral projected gradient methods on convex sets*, SIAM J. Optim. 10(4) (2000), 1196–1211.

[5] Burdakov, Y., Dai, O. and Huang, N. *Stabilized Barzilai–Borwein method*, J. Comput. Math. 37(6) (2019), 916–936.

[6] Cauchy, A. *Méthode générale pour la résolution des systemes d'équations simultanées*, Comp. Rend. Sci. Paris 25 (1847), 536–538.

[7] Dai, Y.-H. and Liao, L.-Z. *R-linear convergence of the barzilai and borwein gradient method*, IMA J. Numer. Anal. 22(1) (2002), 1–10.

[8] Dai, Y.-H. and Zhang, H. *Adaptive two-point stepsize gradient algorithm*, Numer. Algorithms 27 (2001), 377–385.

[9] Dolan, E.D. and Moré, J.J. *Benchmarking optimization software with performance profiles*, Math. Program. 91(2) (2002), 201–213.

[10] Dong, W.-L., Li, X. and Peng, Z. *A simulated annealing-based Barzilai–Borwein gradient method for unconstrained optimization problems*, Asia-Pac. J. Oper. Res. 36(04) (2019), 1950017.

[11] Fletcher, R. *Low storage methods for unconstrained optimization*, Dundee Department of Mathematics and Computer Science, University of Dundee, 1988.

[12] Gould, N.I.M., Orban, D. and Toint, P.L. *Cutest: a constrained and unconstrained testing environment with safe threads for mathematical optimization*, Comput. Optim. Appl. 60(3) (2015), 545–557.

[13] Grippo, L., Lampariello, F. and Lucidi, S. *A nonmonotone line search technique for newton's method*, SIAM J. Numer. Anal. 23(4) (1986) 707–716.

[14] Han, J. and Liu, G. *Global convergence analysis of a new nonmonotone BFGS algorithm on convex objective functions*, Comput. Optim. Appl. 7 (1997), 277–289.

[15] Kirkpatrick, S., Gelatt, C.D. and Vecchi, M.P. *Optimization by simulated annealing*, Science, 220(4598) (1983), 671–680.

[16] Liu, G.H. and Peng, J.M. *The convergence properties of a nonmonotonic algorithm*, J. Comput. Math. 1 (1992), 65–71.

[17] Metropolis, N., Rosenbluth, A.W., Rosenbluth, M.N., Teller, A.H. and Teller, E. *Equation of state calculations by fast computing machines*, J. Comput. Math. 21(6) (1953), 1087–1092.

[18] Mu, X. and Liu, W. *An augmented lagrangian method for binary quadratic programming based on a class of continuous functions*, Optim. Lett. 10(3) (2016), 485–497.

[19] Nocedal, J. and Wright, S.J. *Numerical optimization*, Springer, 1999.

[20] Raydan, M. *On the barzilai and borwein choice of steplength for the gradient method*, IMA J. Numer. Anal. 13(3) (1993), 321–326.

[21] Raydan, M. *The barzilai and borwein gradient method for the large scale unconstrained minimization problem*, SIAM J. Optim. 7(1) (1997), 26–33.

[22] Toint, P.L. *An assessment of nonmonotone linesearch techniques for unconstrained optimization*, SIAM J. Sci. Comput. 17(3) (1996), 725–739.

[23] Wang, C., Liu, Q. and Yang, X. *Convergence properties of nonmonotone spectral projected gradient methods*, J. Comput. Appl. Math. 182(1) (2005), 51–66.

[24] Zhang, H. and Hager, W.W. *A nonmonotone line search technique and its application to unconstrained optimization*, SIAM J. Optim. 14(4) (2004), 1043–1056.

[25] Zhensheng, Yu. *Solving bound constrained optimization via a new nonmonotone spectral projected gradient method*, Appl. Numer. Math. 58(9) (2008), 1340–1348.

[26] Zhou, J.L. and Tits, A.L. *Nonmonotone line search for minimax problems*, J. Optim. Theory Appl. 76(3) (1993), 455–476.

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 opti-

mization. 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. 14, No. 3, pp 638-990

Nonpolynomial B-spline collocation method for solving singularly perturbed quasilinear Sobolev equation 638
F. Edosa Merga and G. File Duressa

Differential-integral Euler–Lagrange equations 662
M. Shehata

An improved imperialist competitive algorithm for solving an inverse form of the Huxley equation 681
H. Dana Mazraeh, K. Parand, H. Farahani and S.R. Kheradpisheh

Stability analysis and optimal strategies for controlling a boycotting behavior of a commercial product 708
O. Aarabate, S. Belhdid and O. Balatif

Highly accurate collocation methodology for solving the generalized Burgers–Fisher's equation 736
S. Shallu and V.K. Kukrej

Uniformly convergent numerical solution for caputo fractional order singularly perturbed delay differential equation using extended cubic B-spline collocation scheme 762
N.A. Endrie and G.F. Duressa

Finite element analysis for microscale heat equation with Neumann boundary conditions 796
M.H. Hashim and A.J. Harfash

Numerical method for the solution of high order Fredholm integro-differential difference equations using Legendre polynomials 833
P.T. Pantuvo, G. Ajileye, R. Taparki and O.O. Aduroja

A pseudo−operational collocation method for optimal control problems of fractal−fractional nonlinear Ginzburg−Landau equation 875
T. Shojaeizadeh, E. Golpar-Rabok and P. Rahimkhani

A numerical computation for solving delay and neutral differential equations based on a new modification to the Legendre wavelet method 900
N.M. El-Shazly and M.A. Ramadan

Extending quasi-GMRES method to solve generalized Sylvester tensor equations via the Einstein product 938
M.M. Izadkhah

A stabilized simulated annealing-based Barzilai–Borwein method for the solution of unconstrained optimization problems 970
H. Sharma and R. Kumar Nayak

web site: https://ijnao.um.ac.ir
Email: ijnao@um.ac.ir
ISSN-Print: 2423-6977
ISSN-Online: 2423-6969