Ferdowsi University of Mashhad

Iranian Journal of Numerical Analysis and Optimization

Volume 16, Number 3

September 2026

Serial Number: 38

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 16, Number 3, September 2026

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
- 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 16, Number 3, September 2026

Ferdowsi University of Mashhad - Iran

©2026 All rights reserved. Iranian Journal of Numerical Analysis and Optimization

Iranian Journal of Numerical Analysis and Optimization

Director

M. H. Farahi

Editor-in-Chief

Ali R. Soheili

Managing Editor

M. Gachpazan

EDITORIAL BOARD

Abbasbandi, Saeid*

(Numerical Analysis)

Imam Khomeini International University, Iran.

e-mail: abbasbandy@ikiu.ac.ir

Abdi, Ali*

(Numerical Analysis)

University of Tabriz, Iran.

e-mail: a_abdi@tabrizu.ac.ir

Area, Iván*

(Numerical Analysis)

Universidade de Vigo, Spain.

e-mail: area@uvigo.es

Babaie Kafaki, Saman*

(Optimization)

Semnan University, Iran.

e-mail: sbk@semnan.ac.ir

Babolian, Esmail*

(Numerical Analysis)

Kharazmi University, Iran.

e-mail: babolian@khu.ac.ir

Cardone, Angelamaria*

(Numerical Analysis)

Università degli Studi di Salerno, Italy.

e-mail: ancardone@unisa.it

Dehghan, Mehdi*

(Numerical Analysis)

Amirkabir University of Technology, Iran.

e-mail: mdehghan@aut.ac.ir

Effati, Sohrab*

(Optimal Control & Optimization)

Ferdowsi University of Mashhad, Iran.

e-mail: s-effati@um.ac.ir

Emrouznejad, Ali*

(Operations Research)

Aston University, UK.

e-mail: a.emrouznejad@aston.ac.uk

Farahi, Mohammad Hadi*

(Optimal Control & Optimization)

Ferdowsi University of Mashhad, Iran.

e-mail: farahi@um.ac.ir

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

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

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

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

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

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

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

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

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

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

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

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

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

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

Toutounian, Faezeh*

(Numerical Analysis)

Ferdowsi University of Mashhad, Iran.

e-mail: toutouni@um.ac.ir

Türkyılmazoğlu, Mustafa*

(Applied Mathematics)

Hacettepe University, Turkey.

e-mail: turkyilm@hacettepe.edu.tr

Vahidian Kamyad, Ali*

(Optimal Control & Optimization)

Ferdowsi University of Mashhad, Iran.

e-mail: vahidian@um.ac.ir

Xu, Zeshui*

(Decision Making)

Sichuan University, China.

e-mail: xuzeshui@263.net

Vasagh, Zohreh

(English Text Editor)

Ferdowsi University of Mashhad, Iran.

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

A fitted numerical technique for a class of singularly perturbed delay differential equations 757
A. Mardi and H. Shankar Prasad

Adaptive-batch stochastic gradient descent for constrained optimization based on relaxed barrier functions 790
A. Fillali, S. Addoune and R. Zeghdane

Comparative semi-analytical solutions of Caputo time-fractional Fokker–Planck equations via Laplace transform-based decomposition and iterative methods 815
M. Hossain and M. Aman Ullah

A probabilistic approach to portfolio selection with Risk-Aversion coefficient estimated via logistic regression 852
L. Chin, A. Sukmana, E. Chendra and A. P. Wijaya

Tension spline-quasilinearization scheme for generalized nonlinear time-fractional convection-diffusion equations 876
A.S. Bekela, L.G. Enyadene and M.M. Woldaregay

Robust stochastic gradient descent for linearly constrained problems via adaptive barrier amplification 922
A. Fillali, S. Addoune and R. Zeghdane

A new numerical computational method for solving nonhomogeneous time-fractional nonlinear parabolic problems in unbounded domains 952
S.A. Ghanizadeh and M.H. Akrami

A compact common-weight DEA model for efficiency measurement 991
A. Hadya K, T. Radha Ramanan, M.C. Lineesh and C.M. Sushama

Barycentric rational interpolation method for fractional-order differential equation model of HIV infection of CD4$^+$T-cells 1009
Kh. Divsar and R. Katani

Existence and Ulam–Hyers stability for a singular Caputo-fractional problem with integral boundary conditions 1031
A.E-S. Bouziane and B. Bendoukha

Inverse eigenvalue problems for composite banded tridiagonal matrices: Symmetric and nonsymmetric cases 1064
M. Babaei Zarch

Mathematical model on the effect of primary and secondary toxicants on a biological species: Occurrence of deformity 1083
A.K. Agarwal and A. Agrawal

Analysis of 2D Burger's equation using numerical and semi-analytic techniques 1107
A. Kaur, A. Pasrija and S. Arora

Quasilinearized Genocchi-Polynomial Collocation for Efficient Simulation of a Nonlinear SIRV Anthrax Transmission Model 1133
N. Mohamed, M. El-Gamel and W. Adel

Analysis of the mathematical model's stability describing the prevalence of obesity and its role in diabetes 1161
A. El Mansouri, M. Belam and A. Labzai

Iranian Journal of Numerical Analysis and Optimization
Vol. 16, No. 3, 2026, pp 757–789
https://doi.org/10.22067/ijnao.2026.96102.1757
https://ijnao.um.ac.ir/



Research Article

A fitted numerical technique for a class of singularly perturbed delay differential equations

A. Mardi*, H. Shankar Prasad

Abstract

This paper examines the applicability of a new exponentially fitted scheme on Shishkin mesh for solving a class of singularly perturbed problems in ordinary differential equations containing a small delayed term. The original delayed problem is reduced to an equivalent new problem using Taylor series expansion procedure, and then a three-term recurrence relationship is obtained by approximating the first and second-order derivatives in the reduced problem by their mixed and central difference analogue, respectively. To avoid the nonuniformity in the solution, a fitting factor is introduced into the obtained tri-diagonal scheme, and its value is determined by the use of singular perturbation theory. Thomas algorithm is efficiently used to solve a tridiagonal system of equations. Convergence and stability of the scheme are discussed in detail. Richardson's extrapolation technique is employed to improve the accuracy and rate of convergence of the solution produced by the scheme. Four standard example problems are solved, and the computational results are presented in terms of the maximum absolute point-wise errors and rate of convergence. Computational results are compared with some other

*Corresponding author

Received 21 October 2025; revised 3 January 2026; accepted 4 January 2026

Anupam Mardi
Department of Mathematics, National Institute of Technology Jamshedpur, Jharkhand - 831014, India. e-mail: anupammardiapril@gmail.com

Hari Shankar Prasad
Department of Mathematics, National Institute of Technology Jamshedpur, Jharkhand - 831014, India. e-mail: hsprasad.math@nitjsr.ac.in

How to cite this article
Mardi, A. and Shankar Prasad, H., A fitted numerical technique for a class of singularly perturbed delay differential equations. *Iran. J. Numer. Anal. Optim.*, 2026; 16(3): 757-789. https://doi.org/10.22067/ijnao.2026.96102.1757

published results. Theoretical and computational analysis reveal that the fitted scheme is able to produce accurate and uniformly convergent solutions with first-order accuracy, which is further improved to second-order accuracy by Richardson extrapolation. Furthermore, the method is most suitable for all the values of $(1/N) \gg \varepsilon$, where $\varepsilon \leq 10^{-6}$ and N is the number of subintervals in which the underlying interval is divided.

AMS subject classifications (2020): 34D15 · 34E15 · 34E20 · 34B25

Keywords: Singular perturbation problem · Boundary layer · Differential-difference equation · Fitting factor · Thomas Algorithm · Stability and convergence

1 Introduction

It is well known that the singularly perturbed delay differential equations (SPDDEs) or differential-difference equations with interval boundary conditions (IBCs) are the result of mathematical modeling of various physical and biological phenomena/processes/problems. These SPDDEs with IBCs are most ubiquitous and arise in situations in which the future state of the system depends upon the present as well as the past. There are a number of realistic model singularly perturbed problems in differential-difference equations in which one encounters the SPDDEs with IBCs. The first exit time problems in the modeling of activation of neurons and variational problems in control theory are two examples of such problems. The importance of these singularly perturbed problems in differential-difference equations (SPPDDEs) can be easily realized from the research articles and monographs [34, 37, 41, 55, 46, 23, 6, 22, 69, 74, 53, 38, 5, 11, 8, 19, 72, 73, 54, 71]. These SPDDEs are ordinary/partial differential equations in which the highest order derivative is multiplied by a small singular perturbation parameter ε $(0 < \varepsilon \ll 1)$ and the term(s) contain delay (negative shift) and/or advance (positive shift) argument(s). The solution of such problems exhibits oscillatory and/or boundary/interior layer behavior in the underlying interval of interest. In fact, the solution is of multi-scale character when the ε is small tending to zero. That is there exist thin transition layer(s) where the solution undergoes drastic change, while away from the layer(s) the solution behaves regularly and varies slowly. Here, the word "layer" stands for the small subregion of the independent variable over which the dependent variable experiences the most rapid change and the terms boundary and interior layers are due to the occurrence of the layer in the neighborhood of the end and the interior points of the underlying domain, respectively. These thin layers are usually termed as the boundary layers, shock layers, skin layers and transitional points in the various fields of applied science and engineering.

Development of numerical/analytical methods has been a herculean task due to such abnormal layer behavior of solutions. It has been experienced that the error in solving such problems by applying classical numerical methods cannot be reduced below a certain fixed limit unless specially designed numerical methods or nonuniform meshes are used [47]. Thus, the study of such a pervasive problem is vital if robust and accurate solutions are required. More detailed information can be obtained from the monographs: [18, 10, 9, 56, 57, 55, 8] and the research articles: [20, 52, 75, 63, 67, 29, 26, 28, 36, 33, 61, 66, 25, 65, 27, 1, 62, 30, 40, 39, 59, 60, 15, 17, 7, 32, 21, 76].

This paper considers a class of second-order SPDDEs in ordinary differential equations having small delay (δ) in both the convection and reaction terms:

$$\varepsilon y''(s) + b(s)y'(s-\delta) + c(s)y(s-\delta) + d(s)y(s) = f(s) \tag{1}$$

$$\begin{aligned} \text{with the IBCs: } \quad & y(s) = \alpha(s) \text{ on } -\delta \le s \le 0 \text{ and} \\ & y(s) = \beta(s) \text{ on } 1 \le s \le 1+\delta, \end{aligned} \tag{2}$$

where the functions $b(s)$, $c(s)$, $d(s)$, and $f(s)$ are sufficiently continuously differentiable functions over $[0,1]$ with $b(s) \ge b_* > 0$; α and β are considered to be constants.

The solution of SPDDE (1) with the IBCs (2) exhibits boundary layer behavior at left-end of the interval $[0,1]$ when $b(s) - \delta c(s) \ge M > 0$, where M is considered to be a positive real number. When $b(s) - \delta c(s) \le M < 0$, the solution exhibits boundary layer at the right-end of the underlying domain.

Such SPPDDEs have grabbed a great deal of attention in the recent and past years. In a series of papers, [38, 39, 40, 42, 41], Lange and Miura performed the singular perturbation analysis of boundary-value problem for differential-difference equations with small shifts. In the papers [41] and [42], they employed asymptotic approaches to approximate the solution of BVPs for the two classes of singularly perturbed problems in differential-difference equations with small shifts. In the article [42], they studied the effect of small shifts on the oscillatory behavior of the solution and remarked that the shifts affect oscillatory solutions more than the boundary layer solutions. They pointed out that the delayed term can be approximated by the use of the Taylor series expansion procedure in the case when the delay is of a small order of ε [42]. There are many references in the literature where the shifts are very small and affect the solution considerably. Lange and Miura [42] have demonstrated that the effect of very small shifts on the solution behavior by considering several numerical example problems in which the shifts are of small order of ε. They pointed out that the small shifts affect the solution considerably and cannot be neglected. The biologist Hutchinson [24] also found that "there is a tendency for the time lag to

be reduced as much as possible by natural selection". The arguments for small delay problems are found throughout the literature on epidemics and populations. Thus the small shifts play an important role in practical problems [34]. In the article [35] the authors have devised a B-spline based numerical method to treat singular perturbation problems (SPPs) having delays in the convection and reaction term over a piece-wise uniform mesh. The convergence of the method is thoroughly discussed. The method is achieving almost second-order convergence. In the article [77], the authors have suggested an exponentially fitted mid-point upwind finite difference method for solving a class of convection-diffusion SPDDEs. The proposed method achieves linear order convergence. A nonsymmetric finite difference based scheme has been devised by the authors in the article [78] to treat convection type SPPs having delay shift. The terms containing the delay parameter are expanded up-to second-order using the well-known Taylor series expansion. Theoretical and numerical examinations reveal that the method is achieving second-order convergence. Authors in [58] have carried out a numerical integration technique to treat a class of SPDDEs which yields a tri-diagonal system that has been solved by Thomas algorithm. This scheme has been performed on some example problems to validate the scheme. A detailed discussion has been done on how the change of the shift parameter affects the layer behavior of the solution. The author in [60] has developed a cubic spline-based numerical solving technique for SPDDEs having delays in the convection and reaction terms. The computational results have been compared with some published results. The author has observed that the scheme achieved second-order rate of convergence (ROC) as $\varepsilon \geq h$ but losses an order when $\varepsilon \ll h$ and tending to zero. Similar types of problems are discussed in the paper [3]. Here, the authors have proposed a central finite difference based exponentially fitted scheme for solving it using uniform mesh discretization. The numerical and experimental analysis reveals that the method is of quadratic order convergence for $\varepsilon = O(N^{-1})$. On the other hand, the method achieves linear order convergence for $\varepsilon \ll N^{-1}$ which indicates the fact that the ROC of the method is ε-dependent. The depicted figures for the solution to the example problems, show that the width of the boundary layer decreases as the perturbation parameter goes smaller.

In this article, we have proposed a novel exponentially fitted scheme for numerical solution of SPDDE (1) with the IBCs (2). The negative shift (or delay parameter) induces memory effect into the problem, that is the solution of the problem at any moment depends upon the solution of the problem at some earlier moment. Taylor's series expansion is used to approximate the delayed terms and the first- and second-order derivatives in the reduced equivalent problem are approximated by the first-order mixed finite differences and second-order central differences respectively, which leads to generating a tridiagonal system of equations. Application of Taylor series expansion is justified due to the reason that the delay δ is of a small order of ε. A suitable fitting factor that involves the exponential function has been introduced into the scheme to

avoid the nonuniformity in the solution. The resulting system is solved by "Discrete Invariant Imbedding Algorithm" widely known as "Thomas Algorithm". To improve the accuracy of the solution produced by the proposed scheme and ROC of the scheme, we have applied Richardson extrapolation procedure to it. The stability and convergence of the fitted scheme have been discussed in detail. To confirm the applicability of the proposed scheme, four example problems have been solved for diverse values of δ and ε. The computational results are tabulated in terms of maximum absolute point-wise errors (MAPEs) and the ROCs before and after applying Richardson extrapolation. Numerical experiments show that the scheme is capable of producing better results than the results obtained in [60, 3, 58, 41]. For more clarity, the solutions of the problems have been plotted graphically for different values of ε, N, and δ, which reveals that as δ increases and $\varepsilon \to 0$ the thickness of the boundary layer decreases. Also, the log-log plot of MAPEs vs N for the problems have also been done which shows the relation between the ε, N, and MAPEs.

Here, the novelty of the proposed method is associated with the improvement in the accuracy of the solutions obtained and the ROC achieved, when the Richardson extrapolation procedure is applied. There are some methods/ schemes for which the accuracy of the solutions obtained improves but the ROC does not improve when the Richardson extrapolation procedure is applied on it. In the articles [50, 51], the authors have shown that the Richardson extrapolation does not improve the ROC of the method while there is an improvement in the accuracy of the solutions. Nevertheless, in the case of the proposed scheme, an application of the Richardson extrapolation noticeably improves the accuracy as well as the ROC. This fact strengthens the claim "A numerical solution of required accuracy can be obtained by using Richardson extrapolation" [2, 68]. Furthermore, from the point of view of the wealth of the literature on singular perturbation problems, we always raise a question that whether there are other ways to deal with the SPPs which can be used easily with minimal amount of computational work (problem preparation). Motivated from this question, we have successfully devised a new exponentially fitted finite difference method for solving the considered SPPDDEs. Method is capable of producing highly accurate and uniformly convergent results with quadratic ROC for all the values of $(1/N) \gg \varepsilon$, where $\varepsilon \leq 10^{-6}$ and N is the number of subintervals in which the underlying interval is divided. Furthermore, our method does not depend upon any other parameters as used in [31, 64]

The remaining part of the article has been arranged in the following manner:
In Section 2, an approximated version of the original problem is deduced by the use of Taylor series; In Section 3, Shishkin mesh is described. In Section 4, we have derived the proposed scheme. In Section 5, stability of the proposed method is thoroughly discussed. In Section 6, Richardson's extrapolation procedure is discussed in detail. In Section 7, we have applied the proposed method to some example problems to establish the applicability of the method. The

results produced by the method are presented in tabular form. In Section 8, we have summarized the paper work. Paper ends with the references.

2 Approximate problem and solution properties

Using the Taylor series, we have

$$\left.\begin{aligned} y(s-\delta) &\approx y(s) - \delta\, y'(s) + \mathcal{O}(\delta^2), \\ y'(s-\delta) &\approx y'(s) - \delta\, y''(s) + \mathcal{O}(\delta^2). \end{aligned}\right\} \tag{3}$$

Using the above expansions into (1), we get

$$\mu y''(s) + p(s)y'(s) + q(s)y(s) = f(s), \quad s \in [0,1], \tag{4}$$

with

$$y(0) \approx \alpha(0) \quad \text{and} \quad y(1) \approx \beta(1). \tag{5}$$

where $\mu = \varepsilon - \delta b(s)$, $p(s) = b(s) - \delta c(s)$ and $q(s) = c(s) + d(s)$, respectively, and it is presumed that $q(s) \leq 0$.

The SPDDE (4)–(5) is different from the original problem (1)–(2) by the term $(\mathcal{O}(\delta^2)y'' + \mathcal{O}(\delta^2)y''')$, which is a good approximation of the original problem. The use of Taylor's series expansion in approximating the term containing delay arguments (δ) is valid because of the fact that the values of delay δ are considered to be of $o(\varepsilon)$ [42, 41].

As the new equivalent version of the original problem (4)–(5) differs from the original problem (1)–(2) by the term $(\mathcal{O}(\delta^2)y'' + \mathcal{O}(\delta^2)y''')$, we prefer a new notation say $\lambda(s)$ for the solution of the approximated version of (1)–(2). Thus (4)–(5) can be written as

$$\mathcal{L}\lambda := \mu\lambda''(s) + p(s)\lambda'(s) + q(s)\lambda(s) = f(s), \quad s \in [0,1], \tag{6}$$

with

$$\lambda(0) = \alpha(0) = \alpha_0 \quad \text{and} \quad \lambda(1) = \beta(1) = \beta_1. \tag{7}$$

The solution of the problem (6)–(7) can be decomposed as

$$\lambda = v + w. \tag{8}$$

For any $k_1 > 0$, the regular part v satisfies

$$|v^{(k)}(s)| \leq C \quad \text{for} \quad k = 0, 1, \ldots, k_1, \tag{9}$$

and the layer part w satisfies

$$|w^{(k)}(s)| \leq C\mu^{-k} e^{-p_*(1-s)/\mu} \quad \text{for} \quad k = 0, 1, \ldots, k_1. \tag{10}$$

Furthermore, we have

$$\mathcal{L}v = f \quad \text{and} \quad \mathcal{L}w = 0; \tag{11}$$

see [44].

3 Description of the Shishkin-mesh

In order to obtain the discretized scheme on the Shishkin mesh, we divide the considered interval, $[0, 1]$ into N subintervals such that $X_N = \{0 = s_1 < s_2 < s_3 < \cdots < s_N < s_{N+1} = 1\}$, where N is the multiple of 2. Now, we divide the interval $[0, 1]$ into two subintervals $[0, \tau]$ and $[\tau, 1]$, where the transition parameter τ is given by

$$\tau = \min\left(\frac{1}{2}, \frac{2\mu}{p^*}\ln(N)\right),$$

where p^* is the maximum value of p over $[0, 1]$. The mesh points are given by

$$\begin{aligned} s_{i+1} &= \frac{i}{N/2}\tau, \quad i = 0, 1, 2, \ldots, \frac{N}{2}, \\ s_{\frac{N}{2}+i+1} &= \tau + \frac{i}{N/2}(1-\tau), \quad i = 1, 2, \ldots, \frac{N}{2}, \end{aligned} \tag{12}$$

where the step lengths $h_i = s_{i+1} - s_i \quad i = 1, 2, 3, \ldots, N$ and $h = \max\{h_i\}, \quad i = 1, 2, 3, \ldots, N$. Let the function ϕ denote the mesh generating function over the interval $[0, s_{(N/2)+1}]$ and defined in such a way that $\phi(0) = 0$ and $\phi(\frac{1}{2}) = \frac{\ln(N)}{p^*}$, where ϕ is a continuously differentiable monotonically increasing function. We define another function ψ such that

$$\phi = -\ln\psi. \tag{13}$$

This function is monotonically decreasing with $\psi(0) = 1$ and $\psi(1/2) = e^{-\frac{ln N}{p^*}}$, which gives

$$\psi(s) = e^{-2\frac{(\ln N)s}{p^*}}. \tag{14}$$

4 Description of the proposed scheme

In this section, we will discuss the proposed scheme step by step for obtaining the solution of the problem (6)–(7).

By the theory of singular perturbation, the solution of the problem (6)–(7) can be written as (see . 22–261[56])

$$\lambda(s_i) = \lambda_0(s_i) + (\alpha - \lambda_0(0)) \exp\Big(- \Big(\frac{p(0)}{\mu} - \frac{q(0)}{p(0)}\Big)\Big). \tag{15}$$

Taking $h_i \to 0$ and using (15), we get

$$\lim_{h\to 0} \lambda_i = \lim_{h_i \to 0} \lambda(s_i) = \lambda_0(0) + (\alpha - \lambda_0(0)) \exp\Big(- \Big(\frac{p^2(0) - \mu q(0)}{p(0)}\Big) i\rho\Big), \tag{16}$$

where $\rho = \frac{h}{\mu}$ is called "density factor".

Approximating the first-order derivative and using the mixed finite difference analogue,

$$\lambda_i' = \alpha\left(\frac{\lambda_{i+1}}{2h_i} - \frac{\lambda_{i-1}}{2h_i}\right) + (1-\alpha)\left(\frac{\lambda_{i+1}}{h_i} - \frac{\lambda_i}{h_i}\right) + \mathcal{O}(h_i) \quad \text{with} \quad \alpha = \frac{1}{2};$$

that is,

$$\lambda_i' = \frac{3\lambda_{i+1} - 2\lambda_i - \lambda_{i-1}}{4h_i}, \tag{17}$$

and the second-order derivative by the second-order central difference

$$\lambda_i'' = \frac{\lambda_{i+1} - 2\lambda_i + \lambda_{i-1}}{{h_i}^2}, \tag{18}$$

we obtain

$$\mu\Big[\frac{\lambda_{i+1}}{{h_i}^2} - \frac{2\lambda_i}{{h_i}^2} + \frac{\lambda_{i-1}}{{h_i}^2}\Big] + p_i\Big[\frac{3\lambda_{i+1}}{4h_i} - \frac{2\lambda_i}{4h_i} - \frac{\lambda_{i-1}}{4h_i}\Big] + q_i\lambda_i = f_i. \tag{19}$$

Introducing a fitting factor $\sigma(\rho)$ into (19), we get

$$\sigma(\rho)\mu\Big[\frac{\lambda_{i+1}}{{h_i}^2} - \frac{2\lambda_i}{{h_i}^2} + \frac{\lambda_{i-1}}{{h_i}^2}\Big] + p_i\Big[\frac{3\lambda_{i+1}}{4h_i} - \frac{2\lambda_i}{4h_i} - \frac{\lambda_{i-1}}{4h_i}\Big] + q_i\lambda_i = f_i. \tag{20}$$

Multiplying (19) by h_i and taking $h_i \to 0$, we get

$$\frac{\sigma(\rho)}{\rho}\Big[\lambda_{i+1}-2\lambda_i+\lambda_{i-1}\Big]+p_i\Big[3\lambda_{i+1}-2\lambda_i-\lambda_{i-1}\Big]=0. \tag{21}$$

Using (16), we get

$$\lim_{h_i\to 0}(\lambda_{i+1}-2\lambda_i+\lambda_{i-1})=(\alpha-\lambda_0(0))e^{-i\kappa\rho}\big(e^{\kappa\rho}+e^{-\kappa\rho}-2\big),$$

$$\lim_{h_i\to 0}(3\lambda_{i+1}-2\lambda_i+\lambda_{i-1})=(\alpha-\lambda_0(0))e^{-i\kappa\rho}\big(3e^{-\kappa\rho}-e^{\kappa\rho}-2\big),$$

where $\kappa=\frac{p^2(0)-\mu q(0)}{p(0)}$.

By using the above two equations, the fitting factor is obtained as follows:

$$\begin{aligned}\sigma(\rho)&=\frac{p(0)\rho}{2}\cdot\frac{e^{\frac{k\rho}{2}}-e^{\frac{-k\rho}{2}}}{e^{\frac{k\rho}{2}}+e^{\frac{-k\rho}{2}}}-\frac{p(0)\rho}{4}\\&=\frac{p(0)\rho}{2}\coth\left(\frac{p^2(0)-\mu q(0)}{p(0)}\frac{\rho}{2}\right)-\frac{p(0)\rho}{4}.\end{aligned} \tag{22}$$

Finally, (20) gives the recurrence relation as follows:

$$\mathcal{E}_i\lambda_{i-1}-\mathcal{F}_i\lambda_i+\mathcal{G}_i\lambda_{i+1}=\mathcal{H}_i,\qquad i=2,\dots,N, \tag{23}$$

where

$$\begin{aligned}\mathcal{E}_i&=\frac{\sigma\mu}{{h_i}^2}-\frac{p_i}{4h_i},\\\mathcal{F}_i&=\frac{2\sigma\mu}{{h_i}^2}+\frac{2p_i}{4h_i}-q_i,\\\mathcal{G}_i&=\frac{\sigma\mu}{{h_i}^2}+\frac{3p_i}{4h_i},\\\mathcal{H}_i&=f_i.\end{aligned}$$

Equation (23) with boundary conditions (7) generates $(N-1)\times(N-1)$ system of equations. This system is solved by the Thomas Algorithm. The stability conditions for the algorithm are $\mathcal{E}_i>0$, $\mathcal{F}_i>0$ and $\mathcal{G}_i>0$, $\mathcal{E}_i+\mathcal{G}_i\geq\mathcal{F}_i$. In this method, if $q(s)\leq 0$ holds, then one can easily show the stability of the algorithm.

5 Stability analysis

5.1 Consistency

Let us consider

$$\bar{\tau}_i[\lambda_i] \equiv \mathcal{L}_{h_i}\lambda(s_i) - \mathcal{L}\lambda(s_i), \quad i = 1, 2, \cdots, N+1,$$

where λ is a smooth function over $\mathcal{I} = [0, 1]$ with $\mathcal{L}_{h_i}$ as the finite difference operator. Then the problem finite difference problem (23)–(7) is consistent with the differential problem (6)–(7) if

$$|\bar{\tau}_i[\lambda]| \to 0 \quad \text{when} \quad h \to 0,$$

where $\bar{\tau}_i[\lambda], \quad i = 1, \ldots, N+1$ are the local truncation errors.

5.2 Accuracy

The solution of the problem (6)–(7) is said to be accurate of order p if for a positive constant $\mathcal{C}$ independent of h and μ, there exists a relation of the form:

$$|\bar{\tau}_i[\lambda]| \leq \mathcal{C}h^p, \quad i = 1, 2, \cdots, N+1.$$

Assumption 1. Throughout the paper, we shall assume that

$$\mu \leq \mathcal{C}N^{-1}, \tag{24}$$

which is generally in the case for discretization of the convection-dominated problems.

Assumption 2. $\mathcal{L}^N$ follows the bound

$$|\mathcal{L}^N\lambda(s_i) - \mathcal{L}\lambda(s_i)| \leq 2\mu \int_{x_{i-1}}^{s_{i+1}} |g'''(s)|ds + q_i \int_{s_i}^{s_{i+1}} |g''(s)|ds, \tag{25}$$

where g is any function $g \in C^3[0, 1]$.

Lemma 1 (Kellogg and Tran)**.** If $b(s)$, $c(s)$, $d(s)$ are sufficiently smooth and independent of μ, then the the solution $\lambda(s)$ satisfies the inequality,

$$|\lambda^{(k)}(s)| \leq \mathcal{C}\big(1 + (\varepsilon - \delta b_*)^{-k} e^{\frac{-M's}{(\varepsilon - \delta b_*)}}\big), \quad k = 0, 1, 2, \ldots. \tag{26}$$

Proof. See [48]. □

Theorem 1. Let us consider that the piece-wise differentiable mesh generating function ψ satisfies the condition

$$\frac{\max \phi'}{N} \leq \mathcal{C} \tag{27}$$

and

$$\int_0^{\frac{1}{2}} \phi'(x)^2 dx \leq \mathcal{C}N. \tag{28}$$

Then the error of the scheme (17) satisfies

$$|\lambda(s_i) - \lambda_N(s_i)| \leq, \begin{cases} \mathcal{C}N^{-1} \max |\psi'|, & \text{for } i = 1, \ldots, N/2, \\ \mathcal{C}N^{-1}, & \text{for } i = (N/2)+1, \ldots, N+1. \end{cases} \tag{29}$$

Proof. The mesh size h_i on the fine mesh satisfies

$$h_i = 2\mu\big(\psi(s_i) - \psi(s_{i-1})\big) \leq 2\mu N^{-1} \max \psi' \quad \text{for} \quad i = 2, \ldots, (N/2)+1. \tag{30}$$

Thus

$$h_i \leq \mathcal{C}N^{-1}, \quad \text{for} \quad i = 2, \ldots, N+1, \tag{31}$$

by using (14) and the Assumption 1 and (27).

Now, using the decomposition in [44] of λ, we split the solution of the discrete problem as

$$\mathcal{L}^N v_i^N = f_i \quad \text{for} \quad i = 2, \ldots, N+1, \quad \text{for} \quad v_1^N = v_1, \quad v_{N+1}^N = v_{N+1}$$

and

$$\mathcal{L}^N w_i^N = 0 \quad \text{for} \quad i = 2, \ldots, N+1, \quad \text{for} \quad w_1^N = w_1, \quad w_{N+1}^N = w_{N+1}.$$

□

An immediate consequence of (9), (25), M-matrix property of $\mathcal{L}^N$, and (31),

$$|v_i - v_i^N| \leq \mathcal{C}N^{-1} \qquad \text{for} \quad i = 1, \ldots, N+1. \tag{32}$$

Using the M-matrix property of $\mathcal{L}^N$, one can show that

$$w_i^N \leq \prod_{k=2}^{(N/2)+1} \left(1 + \frac{h_k}{2\mu}\right)^{-1}, \quad \text{for } i = (N/2)+1, \ldots, N+1, \tag{33}$$

from [43, 45, 70]. We have

$$\ln\left(\prod_{k=2}^{(N/2)+1}\left(1+\frac{h_k}{2\mu}\right)\right) \geq \sum_{k=2}^{(N/2)+1}\left(\frac{h_k}{2\mu}-\frac{1}{2}\left(\frac{h_k}{2\mu}\right)^2\right)$$
$$\geq \ln N - \frac{1}{2}\sum_{k=2}^{(N/2)+1}\left(\frac{h_k}{2\mu}\right)^2.$$

Therefore,

$$\prod_{k=2}^{(N/2)+1}\left(1+\frac{h_k}{2\mu}\right)^{-1} \leq N^{-1}\exp\left(\frac{1}{2}\sum_{k=2}^{(N/2)+1}\left(\frac{h_k}{2\mu}\right)^2\right).$$

Now, the representation

$$\frac{h_k}{2\mu} = \int_{s_{k-1}}^{s_k}\phi'(x)\,dx, \quad \text{for } k = 2,\ldots,(N/2)+1,$$

yields

$$\sum_{k=2}^{(N/2)+1}\left(\frac{h_k}{2\mu}\right)^2 \leq \sum_{k=2}^{(N/2)+1}(s_k - s_{k-1})\int_{s_{k-1}}^{s_k}\phi'(x)^2\,dx = N^{-1}\int_0^{1/2}\phi'(x)^2\,dx.$$

Thus, if assumption (28) holds true, then (33) yields

$$w_i^N \leq \mathcal{C}N^{-1} \quad \text{for } i = (N/2)+1,\ldots,N+1.$$

Consequently,

$$|w_i - w_i^N| \leq |w_i| + |w_i^N| \leq \mathcal{C}N^{-1} \quad \text{for } i = (N/2)+1,\ldots,N+1. \tag{34}$$

For the truncation error of the method with respect to the layer part w, inequalities (10) and (25) imply that

$$\bar{\tau}_i := \|\mathcal{L}^N[w_i - w_i^N]\| \leq \mathcal{C}\mu^{-2}\int_{x_{i-1}}^{x_{i+1}} e^{-p_*(1-s)/\mu}\,dx \quad \text{for } i = 2,\ldots,N.$$

Substituting $x = 2\mu\phi(s)$ on the fine part of the mesh, we obtain

$$\bar{\tau}_i \leq \mathcal{C}(e^{p_*/\mu})\mu^{-1}\int_{s_{i-1}}^{s_{i+1}} e^{-2\phi(s)}\phi'(s)\,ds = \mathcal{C}(e^{p_*/\mu})\mu^{-1}\int_{s_{i-1}}^{s_{i+1}} e^{-\phi(s)}|\psi'(s)|\,ds$$

for $i = 2,\ldots,N/2$. Let $s \in [s_{i-1}, s_{i+1}]$ be arbitrary. Then we have

$$e^{-\phi(s)} \leq e^{-s_{i-1}/(2\mu)} = e^{-s_i/(2\mu)} e^{(x_i - s_{i-1})/(2\mu)} = e^{-s_i/(2\mu)} e^{h_i/(2\mu)} \leq \mathcal{C} e^{-s_i/(2\mu)},$$

by (27) and (30). Thus

$$\bar{\tau}_i \leq \mathcal{C}(e^{p_*/\mu})\mu^{-1} N^{-1} e^{-s_i/(2\mu)} \max|\psi'| \quad \text{for } i = 2, \ldots, N/2.$$

Now, the discrete comparison principle based on the barrier function presented in [70] (see also [43, 45]) yields

$$|E_i - E_i^N| \leq \mathcal{C} N^{-1} \max|\psi'| \quad \text{for } i = 2, \ldots, (N/2) + 1. \tag{35}$$

Combining (32), (34), and (35), we are done.

This shows that the proposed new scheme is linearly convergent.

Now, the Richardson extrapolation technique is described for improving the accuracy and ROC of the solution obtained by the proposed scheme (23).

6 Richardson extrapolation

It is well known that the Richardson extrapolation procedure is designed to enhance the ROC of the scheme/method and the accuracy of the computed solutions. This procedure can be described as follows:

Let $X_N \subset X_{2N}$, where X_{2N} is the set of mesh points obtained by bisecting the step size of X_N. Let $\lambda_{2N}(s)$ be the solution obtained for X_{2N}. Let $\lambda_N(s_i)$ be the numerical solution of the fully discretized scheme (23) over X_N. Then using (29) of Theorem 1, one can write

$$\lambda(s_i) - \lambda_N(s_i) = \mathcal{C}(N^{-1}) + R_N, \quad s_i \in X_N. \tag{36}$$

Similarly, by assuming $\lambda_{2N}(s_i)$ as the numerical solution of the fully discretized scheme equation (23) on X_{2N}, we have

$$\lambda(s_i) - \lambda_{2N}(s_i) = \mathcal{C}\left((2N)^{-1}\right) + R_{2N}, \quad s_i \in X_{2N}. \tag{37}$$

The two remainders R_N and R_{2N} are $\mathcal{O}(N^{-2})$. From (36)–(37), using the approaches in [15, 16, 14], we get

$$\lambda(s_i) - \left(2\lambda_{2N}(s_i) - \lambda_N(s_i)\right) \leq \mathcal{C}\left(N^{-2}\right), \quad s_i \in X_{2N}. \tag{38}$$

Now, we define the extrapolation formula as

$$\lambda_N^{ext}(s) = 2\lambda_{2N}(s) - \lambda_N(s), \tag{39}$$

which gives a better approximation of $\lambda(s_i)$ than either $\lambda_N(s_i)$ or $\lambda_{2N}(s_i)$. The truncation error of spatial discretization after applying the extrapolation to the solution can be written as follows:

$$\begin{aligned}\lambda(s_i) - \lambda^{\text{ext}}(s_i) &= \lambda(s_i) - \big(2\lambda_{2N}(s_i) - \lambda_N(s_i)\big),\\ &= 2\big(\lambda(s_i) - \lambda_{2N}(s_i)\big) - \big(\lambda(s_i) - \lambda_N(s_i)\big), \quad (s_i) \in X_{2N}.\end{aligned} \tag{40}$$

Theorem 2 (Error estimation after applying Richardson extrapolation to the scheme)**.** Let $\lambda(s_i)$ be the solution of (6)–(7) and let $\lambda_N^{\text{ext}}(s_i)$ be the extrapolated solution as defined in 39). Then

$$|\lambda(s_i) - \lambda_N^{\text{ext}}(s_i)| \le \mathcal{C}\big(N^{-2}\big), \quad i = 1, \ldots, N+1.$$

Proof. Like the solution $\lambda(s_i)$ of (6)–(7) and the numerical solution $\lambda_N(s_i)$ of (23) in X_N, the numerical solution $\lambda_{2N}(s_i)$ of (23) on X_{2N} can be decomposed into regular $V_{2N}(s_i)$ and singular $W_{2N}(s_i)$ components such that

$$\lambda_{2N}(s_i) = V_{2N}(s_i) + W_{2N}(s_i). \tag{41}$$

Therefore, we have

$$\lambda(s_i) - \lambda_k(s_i) = \begin{cases} v(s_i) - V_{2N}(s_i) + w(s_i) - W_{2N}(s_i), & \text{if } k = 2N,\\ v(s_i) - V_N(s_i) + w(s_i) - W_N(s_i), & \text{if } k = N. \end{cases} \tag{42}$$

The combination of (40) and (42) gives

$$\begin{aligned}\lambda(s_i) - \lambda_N^{\text{ext}}(s_i) &= 2\big(v(s_i) - V_{2N}(s_i)\big) - \big(v(s_i) - V_N(s_i)\big)\\ &\quad + 2\big(w(s_i) - W_{2N}(s_i)\big) - \big(w(s_i) - W_N(s_i)\big)\\ &= v(s_i) - V_N^{\text{ext}}(s_i) + w(s_i) - W_N^{\text{ext}}(s_i),\end{aligned} \tag{43}$$

where $V_N^{\text{ext}}(s_i)$ is the regular component, and $W_N^{\text{ext}}(s_i)$ is the singular component such that $\lambda_N^{\text{ext}}(s_i) = V_N^{\text{ext}}(s_i) + W_N^{\text{ext}}(s_i)$. From (29) and the techniques used in [12, 4] on X_{2N}, we have

$$v(s_i) - V_{2N}(s_i) = \begin{cases} \mathcal{C}\left((2N)^{-1}\right) + (R_1)_{2N}, & 2 \le i \le \dfrac{N}{2} + 1,\\ \mathcal{C}\left((2N)^{-1}\right) + (R_2)_{2N}, & \dfrac{N}{2} + 2 \le i \le N, \end{cases} \tag{44}$$

where $(R_1)_{2N}$ and $(R_2)_{2N}$ are of $\mathcal{O}((h_i/2)^2)$. For $1 \le j \le \frac{N}{2}$, by using (29) and (44), we have

$$2\big(v(s_i) - V_{2N}(s_i)\big) - \big(v(s_i) - V_N(s_i)\big) = \mathcal{C}\left((2N)^{-1}\right) + \mathcal{O}\big((N^{-1})^2\big) - C\left(N^{-1}\right) + \mathcal{O}\big((N^{-1})^2\big) = \mathcal{O}\big(N^{-2}\big).$$

Thus,

$$v(s_i) - V_N^{\text{ext}}(s_i) = \mathcal{O}\big(N^{-2}\big). \tag{45}$$

For $\frac{N}{2} + 2 \le i \le N$, using the same procedure employed above, we can get the same result (45). Now, consider the singular components in the outer layer region and interior layer region.

First, consider the estimate in the outer layer region $[0, 1-\tau]$. From (10), we have

$$\begin{aligned} |w(s_i)| &\le \mathcal{C}\exp\left(-\frac{p_*(1-s_i)}{\mu}\right) \le \mathcal{C}\exp\left(-\frac{p_*(1-(1-\tau))}{\mu}\right) \\ &= \mathcal{C}\exp\left(-\frac{p_*\tau}{\mu}\right) \le \mathcal{C}N^{-2}, \quad \text{since we assumed } \tau = \frac{2\mu}{p^*}\ln N. \end{aligned} \tag{46}$$

For the solution of the given problem, we consider the following barrier function in order to determine the bound of $|W(s_i)|$:

$$\mathcal{S}(s_i) = \mathcal{C}\prod_{r=1}^{i}\left(1 + \frac{p_* h_r}{\mu}\right)\left[\prod_{r=1}^{N}\left(1 + \frac{p_* h_r}{\mu}\right)^{-1}\right]. \tag{47}$$

By the same technique used in [70], we obtain $|W_N(s_i)| \le \mathcal{C}N^{-2}$. Therefore,

$$|w(s_i) - W(s_i)| \le \mathcal{C}N^{-2}. \tag{48}$$

In a similar approach, we can get

$$|w(s_i) - W_{2N}(s_i)| \le \mathcal{C}N^{-2}. \tag{49}$$

From (48)–(49) for the outer layer region, we obtain

$$|w(s_i) - W_N^{\text{ext}}(s_i)| \le \mathcal{C}N^{-2}. \tag{50}$$

For the interior layer region $[1-\tau, 1]$, by following the same procedure in [49, 4, 13], we obtain

$$w(s_i) - W_k(s_i) = \begin{cases} \mathcal{O}\left(N^{-2}\right), & \text{if } k = 2N, \\ \mathcal{O}\left(N^{-2}\right), & \text{if } k = N. \end{cases} \tag{51}$$

Therefore, for the inner layer region, we get

$$|w(s_i) - W_N^{\text{ext}}(s_i)| \leq \mathcal{C}N^{-2}. \tag{52}$$

Hence, the combination of (45), (50), and (52) gives the required result.

□

7 Numerical experiments

To establish the applicability of the proposed numerical scheme, we applied the scheme to some widely discussed example problems and presented the computational results in tabular form in terms of MAPEs and ROCs for diverse values of N and ε. Since the exact solutions of the problems are unknown, "double-mesh principle" [18] is used to compute the MAPEs and ROCs. Thereafter, to increase the accuracy of the solutions and the ROC of the scheme, we applied the Richardson extrapolation technique to it. We calculated the MAPEs (E_ε^N) by the double-mesh principle formula: $E_\varepsilon^N = \max_{1\leq i\leq N+1} |(\lambda_N^{ext})_i - (\lambda_{2N}^{ext})_{2i-1}|$ and the ROCs (R_ε^N) by the formula, $R_\varepsilon^N = \log_2 \frac{E_\varepsilon^N}{E_\varepsilon^{2N}}$.

Problem 1. First, consider the SPDDE with variable coefficient exhibiting left-end boundary layer [60, 3]:
$\varepsilon\lambda''(s) + (1+s)\lambda'(s-\delta) + e^{-2s}\lambda(s-\delta) - 2e^{-s}\lambda(s) = 0$ over [0,1];
with $\lambda(0) = 0$ and $\lambda(1) = 1$.

Problem 2. Now, consider the SPDDE with variable co-efficient exhibiting left-end boundary layer [60]:
$\varepsilon\lambda''(s) + (1+s)\lambda'(s-\delta) + \sin(2s)\lambda(s-\delta) - e^{-s}\lambda(s) = \sin(2s) + 3e^{-s}$ over [0,1];
with $\lambda(0) = 0$ and $\lambda(1) = 1$.

Problem 3. Now, consider the SPDDE with variable co-efficient exhibiting left-end boundary layer [60, 3, 41, 58]:
$\varepsilon\lambda''(x) + e^s\lambda'(s-\delta) + s\lambda(s) = 0$ over [0,1];
with $\lambda(0) = 0$ and $\lambda(1) = 1$.

Problem 4. Finally, consider the nonlinear SPDDE with constant co-efficient exhibiting boundary layer at the left-end of the underlying domain [41, 58]:
$\varepsilon\lambda''(s) + \lambda(s)\lambda'(s-\delta) - \lambda(s) = 0$ over [0,1];
with $\lambda(s) = 1, -\delta \le s \le 0$, and $\lambda(1) = 1$.

The linear form of the problem:
$\varepsilon\lambda''(s) + \lambda'(s-\delta) - \lambda(s) = 0$ over [0,1];
with $\lambda(0) = 1$ and $\lambda(1) = 1$.

Table 1: MAPEs (E_ε^N) and ROCs (R_ε^N) for diverse values of N and ε of Problem 1 for $\delta = 0.5\varepsilon$

ε	$N = 2^3$	$N = 2^4$	$N = 2^5$	$N = 2^6$	$N = 2^7$	$N = 2^8$	$N = 2^9$
before extrapolation							
10^{-6}	8.7523(-003)	4.5022(-003)	2.2956(-003)	1.1627(-003)	5.8468(-004)	2.9337(-004)	1.4705(-004)
R_ε^N	0.9590	0.9718	0.9814	0.9918	0.9949	0.9964	
10^{-7}	8.7520(-003)	4.5019(-003)	2.2953(-003)	1.1625(-003)	5.8439(-004)	2.9307(-004)	1.4715(-004)
R_ε^N	0.9591	0.9719	0.9815	0.9922	0.9957	0.9940	
10^{-8}	8.7520(-003)	4.5019(-003)	2.2952(-003)	1.1624(-003)	5.8436(-004)	2.9305(-004)	1.4718(-004)
R_ε^N	0.9591	0.9719	0.9815	0.9922	0.9957	0.9936	
10^{-9}	8.7520(-003)	4.5019(-003)	2.2952(-003)	1.1624(-003)	5.8436(-004)	2.9304(-004)	1.4719(-004)
R_ε^N	0.9591	0.9719	0.9815	0.9922	0.9958	0.9934	
10^{-10}	8.7520(-003)	4.5019(-003)	2.2952(-003)	1.1624(-003)	5.8436(-004)	2.9305(-004)	1.4718(-004)
R_ε^N	0.9591	0.9719	0.9815	0.9922	0.9957	0.9936	
after extrapolation							
10^{-6}	8.6094(-004)	2.4311(-004)	6.4230(-005)	1.6260(-005)	3.8500(-006)	1.0000(-006)	5.4000(-007)
R_ε^N	1.8243	1.9203	1.9819	2.0784	1.9449	0.8890	
10^{-7}	8.6135(-004)	2.4346(-004)	6.4580(-005)	1.6630(-005)	4.2200(-006)	1.0400(-006)	2.5000(-007)
R_ε^N	1.8229	1.9145	1.9573	1.9785	2.0207	2.0566	
10^{-8}	8.6140(-004)	2.4348(-004)	6.4600(-005)	1.6670(-005)	4.2400(-006)	1.0800(-006)	2.9000(-007)
R_ε^N	1.8229	1.9142	1.9543	1.9751	1.9730	1.8969	
10^{-9}	8.6139(-004)	2.4350(-004)	6.45900(-005)	1.6670(-005)	4.2500(-006)	1.0800(-006)	2.9000(-007)
R_ε^N	1.8227	1.9145	1.9541	1.9717	1.9764	1.8969	
10^{-10}	8.6139(-004)	2.4350(-004)	6.4610(-005)	1.6660(-005)	4.2400(-006)	1.0800(-006)	2.9000(-007)
R_ε^N	1.8227	1.9141	1.9554	1.9743	1.9730	1.8969	

We have implemented the proposed numerical scheme on four standard example Problems 1 to 4 for diverse values of $N = \{2^3, 2^4, \ldots, 2^9\}$ and $\varepsilon = \{10^{-6}, 10^{-7}, 10^{-8}, 10^{-9}\}$ with $\delta = 0.5\varepsilon$ and presented the computational results in Tables 1 to 4 in terms of MAPEs and ROCs before and after applying Richardson extrapolation procedure to it. One can easily observe that the

Table 2: MAPEs (E_ε^N) and ROCs (R_ε^N) for diverse values of N and ε of Problem 2 for $\delta = 0.5\varepsilon$

ε	$N = 2^3$	$N = 2^4$	$N = 2^5$	$N = 2^6$	$N = 2^7$	$N = 2^8$	$N = 2^9$
before extrapolation							
10^{-6}	8.1429(-002)	3.9447(-002)	2.0202(-002)	1.0197(-002)	5.1213(-003)	2.5675(-003)	1.2861(-003)
R_ε^N	1.0456	0.9654	0.9864	0.9936	0.9961	0.9974	
10^{-7}	8.1428(-002)	3.9445(-002)	2.0201(-002)	1.0195(-002)	5.1200(-003)	2.5663(-003)	1.2849(-003)
R_ε^N	1.0457	0.9654	0.9866	0.9936	0.9965	0.9980	
10^{-8}	8.1428(-002)	3.9445(-002)	2.0201(-002)	1.0195(-002)	5.1199(-003)	2.5662(-003)	1.2847(-003)
R_ε^N	1.0457	0.9654	0.9866	0.9937	0.9965	0.9982	
10^{-9}	8.1428(-002)	3.9445(-002)	2.0201(-002)	1.0195(-002)	5.1199(-003)	2.5662(-003)	1.2847(-003)
R_ε^N	1.0457	0.9654	0.9866	0.9937	0.9965	0.9982	
10^{-10}	8.1428(-002)	3.9445(-002)	2.0201(-002)	1.0195(-002)	5.1199(-003)	2.5662(-003)	1.2847(-003)
R_ε^N	1.0457	0.9654	0.9866	0.9937	0.9965	0.9982	
after extrapolation							
10^{-6}	5.5340(-003)	1.7500(-003)	4.8230(-004)	1.2711(-004)	3.3660(-005)	9.7000(-006)	3.6500(-006)
R_ε^N	1.6610	1.8594	1.9239	1.9170	1.7950	1.4101	
10^{-7}	5.5325(-003)	1.7485(-003)	4.8086(-004)	1.2567(-004)	3.2180(-005)	8.2700(-006)	2.1800(-006)
R_ε^N	1.6618	1.8624	1.9360	1.9654	1.9602	1.9236	
10^{-8}	5.5323(-003)	1.7484(-003)	4.8070(-004)	1.2552(-004)	3.2030(-005)	8.1200(-006)	2.0500(-006)
R_ε^N	1.6618	1.8628	1.9372	1.9704	1.9799	1.9859	
10^{-9}	5.5323(-003)	1.7484(-003)	4.8070(-004)	1.2549(-004)	3.2030(-005)	8.1100(-006)	2.0200(-006)
R_ε^N	1.6618	1.8628	1.9376	1.9701	1.9816	2.0053	
10^{-10}	5.5323(-003)	1.7484(-003)	4.8070(-004)	1.2549(-004)	3.2030(-005)	8.0900(-006)	2.0300(-006)
R_ε^N	1.6618	1.8628	1.9376	1.9701	1.9852	1.9947	

application of Richardson extrapolation to the the solution of the original scheme improves the accuracy of the solution in such a way that the ROC of the proposed scheme is improving from 1 to 2. Furthermore, a simple comparison of the computational results presented in the Tables 5, 6, 7, and 8 with the results obtained in [60, 3, 58, 41] clearly reveal that the proposed scheme with Richardson extrapolation is capable of producing highly accurate and uniformly convergent results for all the values of $1/N \gg \varepsilon$, where $\varepsilon \leq 10^{-6}$. To examine the effect of the small delay (δ) on the solution behavior of the example Problems 1 to 4, we plotted the graphs of the obtained solution in Figures 1–8 for $N = 100$ and diverse values of $\delta = \{0.0\varepsilon, 0.3\varepsilon, 0.6\varepsilon, 0.9\varepsilon\}$ and $\varepsilon = \{10^{-1}, 10^{-2}\}$. These figures reveal that the thickness of the boundary layer decreases and form strong boundary layers when δ increases with fixed N and ε. For more clarity, in Figures

Table 3: MAPEs (E_ε^N) and ROCs (R_ε^N) for diverse values of N and ε of Problem 3 for $\delta = 0.5\varepsilon$

ε	$N = 2^3$	$N = 2^4$	$N = 2^5$	$N = 2^6$	$N = 2^7$	$N = 2^8$	$N = 2^9$
before extrapolation							
10^{-6}	2.4316(-002)	1.1480(-002)	5.5608(-003)	2.7345(-003)	1.3557(-003)	6.7510(-004)	3.3700(-004)
R_ε^N	1.0828	1.0458	1.0240	1.0122	1.0059	1.0024	
10^{-7}	2.4315(-002)	1.1479(-002)	5.5605(-003)	2.7342(-003)	1.3554(-003)	6.7480(-004)	3.3670(-004)
R_ε^N	1.0828	1.0457	1.0241	1.0124	1.0062	1.0030	
10^{-8}	2.4315(-002)	1.1479(-002)	5.5604(-003)	2.7342(-003)	1.3554(-003)	6.7480(-004)	3.3670(-004)
R_ε^N	1.0828	1.0457	1.0241	1.0124	1.0062	1.0030	
10^{-9}	2.4315(-002)	1.1479(-002)	5.5604(-003)	2.7342(-003)	1.3554(-003)	6.7470(-004)	3.3670(-004)
R_ε^N	1.0828	1.0457	1.0241	1.0124	1.0064	1.0028	
10^{-10}	2.4315(-002)	1.1479(-002)	5.5604(-003)	2.7341(-003)	1.3554(-003)	6.7470(-004)	3.3670(-004)
R_ε^N	1.0828	1.0457	1.0241	1.0123	1.0064	1.0028	
after extrapolation							
10^{-6}	1.3570(-003)	3.5830(-004)	9.2300(-005)	2.3500(-005)	6.0000(-006)	1.6000(-006)	5.0000(-007)
R_ε^N	1.9212	1.9568	1.9737	1.9696	1.9069	1.6781	
10^{-7}	1.3572(-003)	3.5840(-004)	9.2300(-005)	2.3600(-005)	6.0000(-006)	1.6000(-006)	5.0000(-007)
R_ε^N	1.9210	1.9572	1.9675	1.9758	1.9069	1.6781	
10^{-8}	1.3569(-003)	3.5840(-004)	9.2200(-005)	2.3600(-005)	6.0000(-006)	1.6000(-006)	5.0000(-007)
R_ε^N	1.9207	1.9587	1.9660	1.9758	1.9069	1.6781	
10^{-9}	1.3569(-003)	3.5840(-004)	9.2200(-005)	2.3600(-005)	6.0000(-006)	1.6000(-006)	5.0000(-007)
R_ε^N	1.9207	1.9587	1.9660	1.9758	1.9069	1.6781	
10^{-10}	1.3569(-003)	3.5840(-004)	9.2200(-005)	2.3300(-005)	6.0000(-006)	1.6000(-006)	5.0000(-007)
R_ε^N	1.9207	1.9587	1.9844	1.9573	1.9069	1.6781	

9–12, we have plotted the log-log plots for MAPE vs N (after extrapolation) for Problem 1 to 4 for various values of ε tending to zero. Through these figures, we easily observe that despite of increasing the value of N, for $\varepsilon = 10^{-6}$, the MAPE increases after certain value of N. Thus the method is capable of producing highly accurate and uniformly convergent results for all the values of $(1/N) \gg \varepsilon$ where $\varepsilon \leq 10^{-6}$.

8 Conclusion

A new exponentially fitted numerical scheme is addressed to approximate the solution of a class of SPDDEs having delay terms in the convection and reaction terms. Taylor's series expansion procedure has been used to approximate the terms containing the delay argument (δ). To increase the accuracy of the solution and the ROC of the scheme from 1 to 2, Richardson ex-

Table 4: MAPEs (E_ε^N) and ROCs (R_ε^N) for diverse values of N and ε of Problem 4 for $\delta = 0.5\varepsilon$

ε	$N = 2^3$	$N = 2^4$	$N = 2^5$	$N = 2^6$	$N = 2^7$	$N = 2^8$	$N = 2^9$
before extrapolation							
10^{-6}	1.9855(-002)	1.0658(-002)	5.5309(-003)	2.8184(-003)	1.4226(-003)	7.1450(-004)	3.5787(-004)
R_ε^N	0.8976	0.9464	0.9726	0.9863	0.9935	0.9975	
10^{-7}	1.9856(-002)	1.0659(-002)	5.5314(-003)	2.8189(-003)	1.4231(-003)	7.1497(-004)	3.5833(-004)
R_ε^N	0.8975	0.9464	0.9725	0.9861	0.9931	0.9966	
10^{-8}	1.9856(-002)	1.0659(-002)	5.5315(-003)	2.8189(-003)	1.4231(-003)	7.1502(-004)	3.5838(-004)
R_ε^N	0.8975	0.9463	0.9725	0.9861	0.9930	0.9965	
10^{-9}	1.9856(-002)	1.0659(-002)	5.5315(-003)	2.8189(-003)	1.4231(-003)	7.1502(-004)	3.5838(-004)
R_ε^N	0.8975	0.9463	0.9725	0.9861	0.9930	0.9965	
10^{-10}	1.9856(-002)	1.0659(-002)	5.5315(-003)	2.8189(-003)	1.4231(-003)	7.1502(-004)	3.5838(-004)
R_ε^N	0.8975	0.9463	0.9725	0.9861	0.9930	0.9965	
after extrapolation							
10^{-6}	1.5751(-003)	4.4038(-004)	1.1714(-004)	2.9860(-005)	7.2300(-006)	1.4900(-006)	2.6000(-007)
R_ε^N	1.8386	1.9105	1.9719	2.0461	2.2787	2.5187	
10^{-7}	1.5756(-003)	4.4083(-004)	1.1756(-004)	3.0310(-005)	7.6700(-006)	1.9000(-006)	4.6000(-007)
R_ε^N	1.8376	1.9068	1.9555	1.9825	2.0132	2.0463	
10^{-8}	1.5756(-003)	4.4089(-004)	1.1761(-004)	3.0340(-005)	7.7100(-006)	1.9500(-006)	5.0000(-007)
R_ε^N	1.8374	1.9064	1.9547	1.9764	1.9833	1.9635	
10^{-9}	1.5757(-003)	4.4087(-004)	1.1760(-004)	3.0360(-005)	7.7300(-006)	1.9500(-006)	5.1000(-007)
R_ε^N	1.8376	1.9065	1.9536	1.9736	1.9870	1.9349	
10^{-10}	1.5756(-003)	4.4088(-004)	1.1763(-004)	3.0350(-005)	7.7000(-006)	1.9700(-006)	5.1000(-007)
R_ε^N	1.8374	1.9061	1.9545	1.9788	1.9667	1.9496	

Table 5: Comparison of MAPEs (E_ε^N) for diverse values of N and ε of Problem 1

ε	$N = 2^5$	$N = 2^6$	$N = 2^7$	$N = 2^8$	$N = 2^9$	$N = 2^{10}$
2^{-20}	6.4250(-005)	1.6310(-005)	3.8700(-006)	9.8000(-007)	5.3000(-007)	4.2000(-007)
2^{-24}	6.4580(-005)	1.6650(-005)	4.2200(-006)	1.0600(-006)	2.7000(-007)	8.0000(-008)
2^{-28}	6.4620(-005)	1.6640(-005)	4.2500(-006)	1.0800(-006)	2.9000(-007)	9.0000(-008)
2^{-32}	6.4610(-005)	1.6640(-005)	4.2500(-006)	1.0800(-006)	2.9000(-007)	9.0000(-008)
result in [60]						
2^{-20}	4.7320(-003)	2.4804(-003)	1.2705(-003)	6.4304(-004)	3.2349(-004)	1.6223(-004)
2^{-24}	4.7320(-003)	2.4804(-003)	1.2705(-003)	6.4307(-004)	3.2352(-004)	1.6226(-004)
2^{-28}	4.7320(-003)	2.4804(-003)	1.2705(-003)	6.4307(-004)	3.2352(-004)	1.6226(-004)
2^{-32}	4.7320(-003)	2.4804(-003)	1.2705(-003)	6.4307(-004)	3.2352(-004)	1.6226(-004)
result in [3]						
2^{-20}	0.12360(-001)	0.62211(-002)	0.31206(-002)	0.15628(-002)	0.78203(-003)	0.39117(-003)
2^{-24}	0.12360(-001)	0.62211(-002)	0.31206(-002)	0.15628(-002)	0.78203(-003)	0.39117(-003)
2^{-28}	0.12360(-001)	0.62211(-002)	0.31206(-002)	0.15628(-002)	0.78203(-003)	0.39117(-003)
2^{-32}	0.12360(-001)	0.62211(-002)	0.31206(-002)	0.15628(-002)	0.78203(-003)	0.39117(-003)

trapolation technique is applied. Numerical experiments are performed to show the applicability of the proposed method. Convergence and stability are also examined. Numerical results are noted in tabular form in terms of MAPEs and the ROCs before and after applying the Richard-

Table 6: Comparison of MAPEs (E_ε^N) for diverse values of N and ε of Problem 2

ε	$N=2^5$	$N=2^6$	$N=2^7$	$N=2^8$	$N=2^9$	$N=2^{10}$
10^{-10}	4.8070(-004)	1.2549(-004)	3.2030(-005)	8.0900(-006)	2.0300(-006)	5.2000(-007)
10^{-12}	4.8070(-004)	1.2549(-004)	3.2030(-005)	8.0900(-006)	2.0300(-006)	5.2000(-007)
10^{-15}	4.8070(-004)	1.2549(-004)	3.2030(-005)	8.0900(-006)	2.0300(-006)	5.2000(-007)
10^{-20}	4.8070(-004)	1.2549(-004)	3.2030(-005)	8.0900(-006)	2.0300(-006)	5.2000(-007)
result in [60]						
10^{-10}	7.9338(-003)	4.0004(-003)	2.0070(-003)	1.0060(-003)	5.0411(-004)	2.5064(-004)
10^{-12}	7.9338(-003)	4.0004(-003)	2.0070(-003)	1.0060(-003)	5.0411(-004)	2.5064(-004)
10^{-15}	7.9338(-003)	4.0004(-003)	2.0070(-003)	1.0060(-003)	5.0411(-004)	2.5064(-004)
10^{-20}	7.9338(-003)	4.0004(-003)	2.0070(-003)	1.0060(-003)	5.0411(-004)	2.5064(-004)

Figure 1: Solution graph for Problem 1 for $N=100$, $\varepsilon=\frac{1}{10}$ and various values of δ of $o(\varepsilon)$

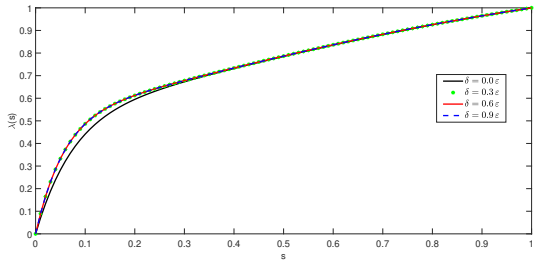
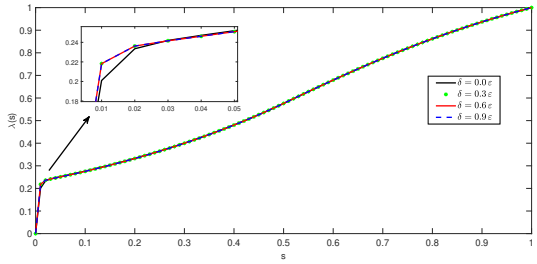


Figure 2: Solution graph for Problem 1 for $N=100$, $\varepsilon=\frac{1}{100}$ and various values of δ of $o(\varepsilon)$



son extrapolation. The results are compared to the results in [60, 3, 58, 41]. The method is successfully achieving quadratic order convergence with higher accuracy. Solution graphs for the considered example problems produced by the proposed numerical technique are plotted for

Table 7: Comparison of MAPEs (E_ε^N) for diverse values of N and ε of Problem 3

ε	$N=2^5$	$N=2^6$	$N=2^7$	$N=2^8$	$N=2^9$	$N=2^{10}$
10^{-9}	9.2200(-005)	2.3600(-005)	6.0000(-006)	1.6000(-006)	5.0000(-007)	3.0000(-007)
10^{-10}	9.2200(-005)	2.3300(-005)	6.0000(-006)	1.6000(-006)	5.0000(-007)	3.0000(-007)
10^{-11}	9.2200(-005)	2.3300(-005)	6.0000(-006)	1.6000(-006)	5.0000(-007)	3.0000(-007)
10^{-12}	9.2200(-005)	2.3300(-005)	6.0000(-006)	1.6000(-006)	5.0000(-007)	3.0000(-007)
10^{-13}	9.2200(-005)	2.3300(-005)	6.0000(-006)	1.6000(-006)	5.0000(-007)	3.0000(-007)
result in [60]						
10^{-9}	4.4394(-003)	2.3357(-003)	1.1972(-003)	6.0599(-003)	3.0464(-004)	1.5287(-004)
10^{-10}	4.4394(-003)	2.3357(-003)	1.1972(-003)	6.0599(-003)	3.0464(-004)	1.5287(-004)
10^{-11}	4.4394(-003)	2.3357(-003)	1.1972(-003)	6.0599(-003)	3.0464(-004)	1.5287(-004)
10^{-12}	4.4394(-003)	2.3357(-003)	1.1972(-003)	6.0599(-003)	3.0464(-004)	1.5287(-004)
10^{-13}	4.4394(-003)	2.3357(-003)	1.1972(-003)	6.0599(-003)	3.0464(-004)	1.5287(-004)
result in [3]						
10^{-9}	4.8030(-003)	2.5439(-003)	1.3085(-003)	6.6353(-004)	3.3410(-004)	1.6763(-004)
10^{-10}	4.8030(-003)	2.5439(-003)	1.3085(-003)	6.6353(-004)	3.3410(-004)	1.6763(-004)
10^{-11}	4.8030(-003)	2.5439(-003)	1.3085(-003)	6.6353(-004)	3.3410(-004)	1.6763(-004)
10^{-12}	4.8030(-003)	2.5439(-003)	1.3085(-003)	6.6353(-004)	3.3410(-004)	1.6763(-004)
10^{-13}	4.8030(-003)	2.5439(-003)	1.3085(-003)	6.6353(-004)	3.3410(-004)	1.6763(-004)
result in [58]						
10^{-9}	0.066(-001)	0.042(-001)	0.023(-001)	0.012(-001)	0.064(-002)	0.032(-002)
10^{-10}	0.066(-001)	0.042(-001)	0.023(-001)	0.012(-001)	0.064(-002)	0.032(-002)
10^{-11}	0.066(-001)	0.042(-001)	0.023(-001)	0.012(-001)	0.064(-002)	0.032(-002)
10^{-12}	0.066(-001)	0.042(-001)	0.023(-001)	0.012(-001)	0.064(-002)	0.032(-002)
10^{-13}	0.066(-001)	0.042(-001)	0.023(-001)	0.012(-001)	0.064(-002)	0.032(-002)
result in [41]						
10^{-9}	0.18(-001)	0.93(-002)	0.47(-002)	0.24(-002)	0.12(-002)	0.59(-003)
10^{-10}	0.18(-001)	0.93(-002)	0.47(-002)	0.24(-002)	0.12(-002)	0.59(-003)
10^{-11}	0.18(-001)	0.93(-002)	0.47(-002)	0.24(-002)	0.12(-002)	0.59(-003)
10^{-12}	0.18(-001)	0.93(-002)	0.47(-002)	0.24(-002)	0.12(-002)	0.59(-003)
10^{-13}	0.18(-001)	0.93(-002)	0.47(-002)	0.24(-002)	0.12(-002)	0.59(-003)

Table 8: Comparison of MAPEs (E_ε^N) for diverse values of N and ε of Problem 4

ε	$N=2^4$	$N=2^5$	$N=2^6$	$N=2^7$	$N=2^8$	$N=2^9$
10^{-9}	4.4087(-004)	1.1760(-004)	3.0360(-005)	7.7300(-006)	1.9500(-006)	5.1000(-007)
10^{-10}	4.4088(-004)	1.1763(-004)	3.0350(-005)	7.7000(-006)	1.9700(-006)	5.1000(-007)
10^{-11}	4.4088(-004)	1.1763(-004)	3.0350(-005)	7.7200(-006)	1.9700(-006)	5.1000(-007)
10^{-12}	4.4088(-004)	1.1763(-004)	3.0350(-005)	7.7200(-006)	1.9700(-006)	5.1000(-007)
10^{-13}	4.4088(-004)	1.1763(-004)	3.0350(-005)	7.7200(-006)	1.9700(-006)	5.1000(-007)
result in [58]						
10^{-9}	0.40(-002)	0.20(-002)	0.10(-002)	0.51(-003)	0.26(-003)	0.13(-003)
10^{-10}	0.40(-002)	0.20(-002)	0.10(-002)	0.51(-003)	0.26(-003)	0.13(-003)
10^{-11}	0.40(-002)	0.20(-002)	0.10(-002)	0.51(-003)	0.26(-003)	0.13(-003)
10^{-12}	0.40(-002)	0.20(-002)	0.10(-002)	0.51(-003)	0.26(-003)	0.13(-003)
10^{-13}	0.40(-002)	0.20(-002)	0.10(-002)	0.51(-003)	0.26(-003)	0.13(-003)
result in [41]						
10^{-9}	0.11(-001)	0.57(-002)	0.29(-002)	0.14(-002)	0.72(-003)	0.36(-003)
10^{-10}	0.11(-001)	0.57(-002)	0.29(-002)	0.14(-002)	0.72(-003)	0.36(-003)
10^{-11}	0.11(-001)	0.57(-002)	0.29(-002)	0.14(-002)	0.72(-003)	0.36(-003)
10^{-12}	0.11(-001)	0.57(-002)	0.29(-002)	0.14(-002)	0.72(-003)	0.36(-003)
10^{-13}	0.11(-001)	0.57(-002)	0.29(-002)	0.14(-002)	0.72(-003)	0.36(-003)

Figure 3: Solution graph for Problem 2 for $N=100$, $\varepsilon=\frac{1}{10}$ and various values of δ of $o(\varepsilon)$

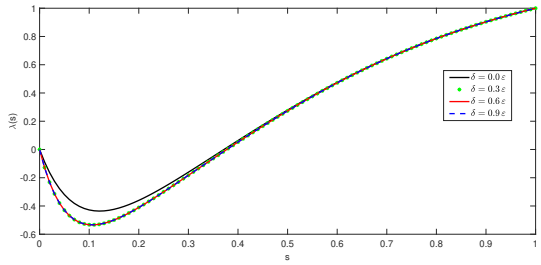


Figure 4: Solution graph for Problem 2 for $N=100$, $\varepsilon=\frac{1}{100}$ and various values of δ of $o(\varepsilon)$

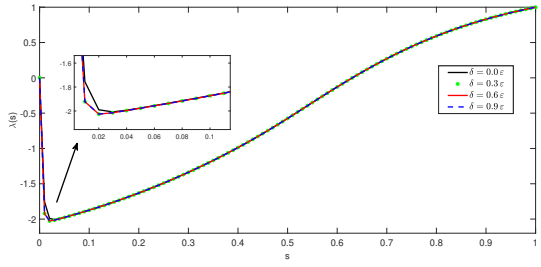


Figure 5: Solution graph for Problem 3 for $N = 100$, $\varepsilon = \frac{1}{10}$ and various values of δ of $o(\varepsilon)$

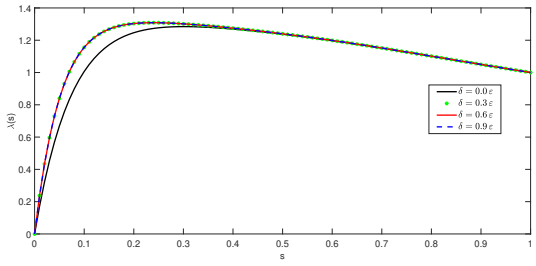


Figure 6: Solution graph for Problem 3 for $N = 100$, $\varepsilon = \frac{1}{100}$ and various values of δ of $o(\varepsilon)$

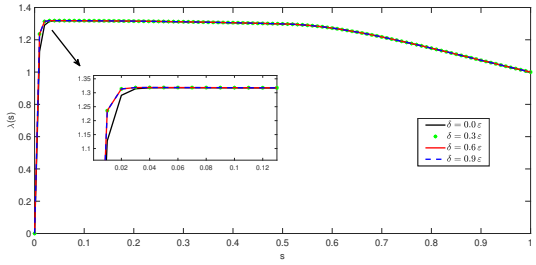


Figure 7: Solution graph for Problem 4 for $N = 100$, $\varepsilon = \frac{1}{10}$ and various values of δ of $o(\varepsilon)$

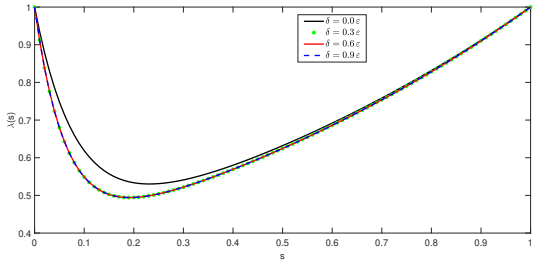


Figure 8: Solution graph for Problem 4 for $N = 100$, $\varepsilon = \frac{1}{100}$ and various values of δ of $o(\varepsilon)$

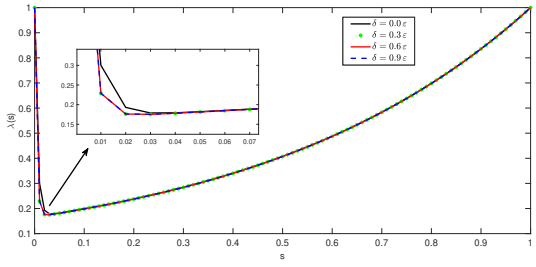


Figure 9: Log-log plot for MAPE vs N for Problem 1 for various values of ε

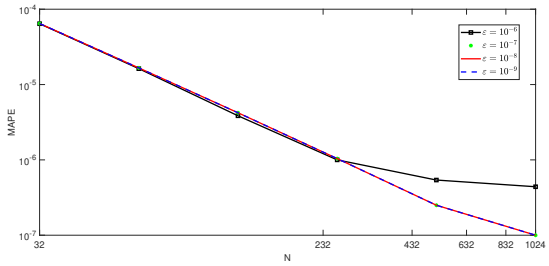


Figure 10: Log-log plot for MAPE vs N for Problem 2 for various values of ε

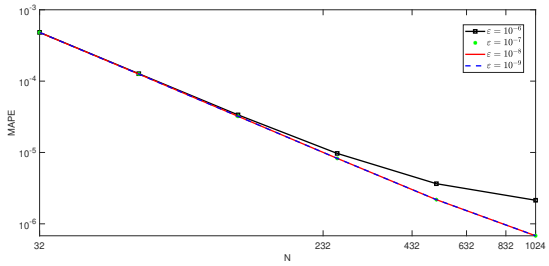


Figure 11: Log-log plot for MAPE vs N for Problem 2 for various values of ε

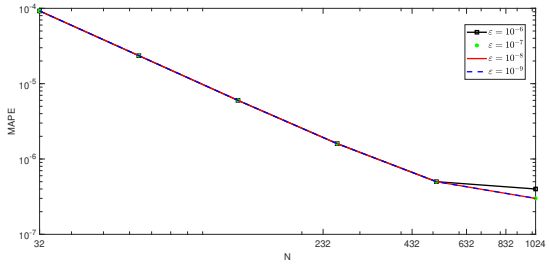
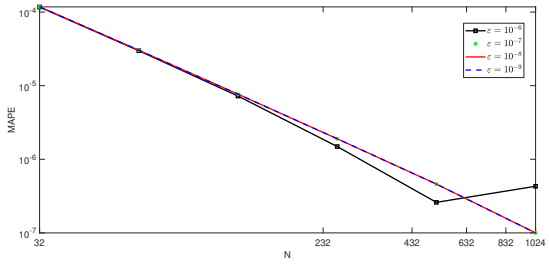


Figure 12: Log-log plot for MAPE vs N for Problem 4 for various values of ε



$N = 100$ and different values of δ and ε. These graphs show that as the delay argument (δ) increases and $\varepsilon \to 0$, the thickness of the boundary layer decreases. For more clarity, in Figures 9, 10, 11, and 12, we have plotted the log-log plots for MAPE vs N (after extrapolation) for the considered Problems 1 to 4 for various values of ε. Through these figures, we easily observe that despite increasing the value of N, for $\varepsilon = 10^{-6}$, the MAPE increases after certain value of N. Thus the method is capable of producing highly accurate and uniformly convergent results for all the values of $(1/N) \gg \varepsilon$ where $\varepsilon \leq 10^{-6}$. Novelty of the proposed method is associated with the improvement in the accuracy of the solutions in such a way that the ROC increases from 1 to 2, when the Richardson extrapolation procedure is applied. For some methods/schemes, the accuracy of the solutions obtained improves, but the ROC does not when the Richardson extrapolation procedure is applied to it [50, 51].

Acknowledgements

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

Funding

This research received no external funding.

Conflicts of Interest

The authors declare no conflicts of interest.

References

[1] Alam, M.J., Prasad, H.S. and Ranjan, R. *A novel fitted method for a class of singularly perturbed differential-difference equations with small delay exhibiting twin layer or oscillatory behavior*, Comput. Math. Math. Phys. 63(12) (2023), 2528–2550.

[2] Andreev, V.F. and Popov, A.M. *Using Richardson's method to construct high-order accurate adaptive grids*, Comput. Math. Model. 10(3) (1999), 227–238.

[3] Angasu, M.A., Duressa, G.F. and Woldaregay, M.M. *Exponentially fitted numerical scheme for singularly perturbed differential equations involving small delays*, J. Appl. Math. Inform. 39(3-4) (2021), 419–435.

[4] Bala, S. and Govindarao, L. and Das, A. and Majumdar, A. *Numerical scheme for partial differential equations involving small diffusion term with non-local boundary conditions*, J. Appl. Math. Comput. 69(6) (2023), 4307–4331.

[5] Balachandran, B. and Kalmár-Nagy, T. and Gilsinn, D.E. *Delay differential equations*, Springer, New York, 2009.

[6] Banasiak, J. and Lachowicz, M. *Methods of small parameter in mathematical biology*, Springer, New York, 2014.

[7] Bawa, R. K. *Spline based computational technique for linear singularly perturbed boundary value problems*, Appl. Math. Comput., 167(1) (2005), 225–236.

[8] Bellen, A. and Zennaro, M. *Numerical methods for delay differential equations*, Springer, Cham, Switzerland, 2013.

[9] Bellman, R. E. and Kalaba, R. E. *Quasilinearization and nonlinear boundary-value problems*, American Elsevier Publishing Company, New York, 1965.

[10] Bender, C. M. and Orszag, S. A. *Advanced mathematical methods for scientists and engineers I: Asymptotic methods and perturbation theory*, Springer Science & Business Media, New York, 2013.

[11] Cooke, K. L. *Differential–difference equations*, in: International Symposium on Nonlinear Differential Equations and Nonlinear Mechanics, Elsevier, 1963, 155–171.

[12] Das, A., Govindarao, L. and Mohapatra, J. *A second-order weighted monotone numerical scheme for time-delayed parabolic initial-boundary-value problem involving a small parameter*, Int. J. Math. Model. Numer. Optim. 12(3) (2022), 233–251.

[13] Das, A. and Natesan, S. *Second-order uniformly convergent numerical method for singularly perturbed delay parabolic partial differential equations*, Int. J. Comput. Math. 95(3) (2018), 490–510.

[14] Das, P. *Comparison of a priori and a posteriori meshes for singularly perturbed nonlinear parameterized problems*, J. Comput. Appl. Math. 290 (2015), 16–25.

[15] Das, P. *A higher order difference method for singularly perturbed parabolic partial differential equations*, J. Difference Equ. Appl. 24(3) (2018), 452–477.

[16] Das, P. *An a posteriori based convergence analysis for a nonlinear singularly perturbed system of delay differential equations on an adaptive mesh*, Numer. Algorithms 81(2) (2019), 465–487.

[17] Das, P. and Natesan, S. *A uniformly convergent hybrid scheme for singularly perturbed system of reaction-diffusion Robin type boundary-value problems*, J. Appl. Math. Comput. 41 (2013), 447–471.

[18] Doolan, E.P. and Miller, J.J.H. and Schilders, W.H.A. *Uniform numerical methods for problems with initial and boundary layers*, Boole Press, Dublin, 1980.

[19] Driver, R.D. *Ordinary and delay differential equations*, Springer, New York, 1977.

[20] Duressa, G.F., Daba, I.T. and Deressa, C.T. *A systematic review on the solution methodology of singularly perturbed differential difference equations*, Mathematics 11(5) (2023), 1108.

[21] Duressa, G.F. and Debela, H.G. *Numerical solution of singularly perturbed differential difference equations with mixed parameters*, J. Math. Model. 9(4) (2021), 691–705.

[22] Fouque, J.-P., Papanicolaou, G. and Sircar, R. and Solna, K. *Singular perturbations in option pricing*, SIAM J. Appl. Math. 63(5) (2003), 1648–1665.

[23] Hek, G. *Geometric singular perturbation theory in biological practice*, J. Math. Biol. 60(3) (2010), 347–386.

[24] Hutchinson, G.E. *Circular causal systems in ecology*, Ann. N. Y. Acad. Sci. 50(4) (1948), 221–246.

[25] Joy, D. and Kumar, S.D. *Liouville–Green transformation technique to solve singularly perturbed delay differential equations of reaction diffusion type*, Math. Methods Appl. Sci. 48(10) (2025), 10314–10331.

[26] Joy, D., Kumar, S.D. and Rihan, F.A. *Advancing numerical solutions for a system of singularly perturbed delay differential equations at linear rate*, Bound. Value Probl. 2025(1) (2025), 15.

[27] Joy, D. and Regal, A.M. *Tension spline approach to singularly perturbed delay differential equations involving large delay*, Gulf J. Math. 15(2) (2023), 166–174.

[28] Kadalbajoo, M.K. and Gupta, V. *A brief survey on numerical methods for solving singularly perturbed problems*, Appl. Math. Comput. 217(8) (2010), 3641–3716.

[29] 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.

[30] Kadalbajoo, M.K. and Patidar, K.C. and Sharma, K.K. *ε-uniformly convergent fitted methods for the numerical solution of the problems arising from singularly perturbed general DDEs*, Appl. Math. Comput. 182(1) (2006), 119–139.

[31] Kadalbajoo, M.K. and Reddy, Y.N. *Numerical solution of singular perturbation problems via deviating arguments*, Appl. Math. Comput. 21(3) (1987), 221–232.

[32] Kadalbajoo, M.K. and Yadaw, A.S. and Kumar, D. *Comparative study of singularly perturbed two-point BVPs via: Fitted-mesh finite difference method, B-spline collocation method and finite element method*, Appl. Math. Comput. 204(2) (2008), 713–725.

[33] Kaur, J. and Kaur, B. *A Review on singularly perturbed problems*, IJRAR (2023).

[34] Kuang, Y. *Delay differential equations*, Academic Press, New York, 1993.

[35] Kumar, D. and Kadalbajoo, M.K. *Numerical treatment of singularly perturbed delay differential equations using B-spline collocation method on Shishkin mesh*, J. Numer. Anal. Ind. Appl. Math. 7(3–4) (2012), 73–90.

[36] Kumar, M. and Singh, P. and Mishra, H.K. *A recent survey on computational techniques for solving singularly perturbed boundary value problems*, Int. J. Comput. Math. 84(10) (2007), 1439–1463.

[37] Lakshmikantham, V. and Wen, L. and Zhang, B. *Theory of differential equations with unbounded delay*, Springer, New York, 2013.

[38] Lange, C.G. and Miura, R.M. *Singular perturbation analysis of boundary value problems for differential-difference equations*, SIAM J. Appl. Math. 42(3) (1982), 502–531.

[39] Lange, C.G. and Miura, R.M. *Singular perturbation analysis of boundary-value problems for differential-difference equations. II. Rapid oscillations and resonances*, SIAM J. Appl. Math. 45(5) (1985), 687–707.

[40] Lange, C.G. and Miura, R.M. *Singular perturbation analysis of boundary-value problems for differential-difference equations. III. Turning point problems*, SIAM J. Appl. Math. 45(5) (1985), 708–734.

[41] Lange, C.G. and Miura, R.M. *Singular perturbation analysis of boundary-value problems for differential-difference equations. V. small shifts with layer behavior*, SIAM J. Appl. Math. 54(1) (1994), 249–272.

[42] Lange, C.G. and Miura, R.M. *Singular perturbation analysis of boundary-value problems for differential-difference equations. VI. small shifts with rapid oscillations*, SIAM J. Appl. Math. 54(1) (1994), 273–283.

[43] Linß, T. *An upwind difference scheme on a novel Shishkin-type mesh for a linear convection–diffusion problem*, J. Comput. Appl. Math. 110(1) (1999), 93–104.

[44] Linß, T. *Solution decompositions for linear convection-diffusion problems*, Z. Anal. Anwend. 21(1) (2002), 209–214.

[45] Linß, T. and Stynes, M. *A hybrid difference scheme on a Shishkin mesh for linear convection–diffusion problems*, Appl. Numer. Math. 31(3) (1999), 255–270.

[46] Longtin, A. and Milton, J.G. *Complex oscillations in the human pupil light reflex with mixed and delayed feedback*, Math. Biosci. 90(1–2) (1988), 183–199.

[47] Miller, J.J.H. *Singular perturbation problems in chemical physics: Analytic and computational methods*, Wiley, New York, 1997.

[48] Miller, J.J.H. and O'Riordan, E. and Shishkin, G.I. *Fitted numerical methods for singular perturbation problems*, World Scientific, Singapore, 1996.

[49] Mukherjee, K. and Natesan, S. *Richardson extrapolation technique for singularly perturbed parabolic convection–diffusion problems*, Computing 92(1) (2011), 1–32.

[50] Munyakazi, J.B. and Patidar, K.C. *Limitations of Richardson's extrapolation for a high-order fitted mesh method for self-adjoint singularly perturbed problems*, J. Appl. Math. Comput. 32(1) (2010), 219–236.

[51] Munyakazi, J.B. and Patidar, K.C. *Performance of Richardson extrapolation on some numerical methods for a singularly perturbed turning point problem whose solution has boundary layers*, J. Korean Math. Soc. 51(4) (2014), 679–702.

[52] Mushahary, P. and Sahu, S.R. and Mohapatra, J. *A parameter uniform numerical scheme for singularly perturbed differential-difference equations with mixed shifts*, J. Appl. Comput. Mech. 6(2) (2020), 344–356.

[53] Norkin, S.B. *Differential equations of the second-order with retarded argument: some problems of the theory of vibrations of systems with retardation*, American Mathematical Society 1972.

[54] Norkin, S.B. *Introduction to the theory and application of differential equations with deviating arguments*, (Vol. 105), Academic Press, 1973.

[55] O'Malley Jr., R.E. *Introduction to singular perturbations*, J. Fluid Mech. 1974.

[56] O'Malley, Jr. R.E. *Singular perturbation methods for ordinary differential equations*, Springer-Verlag, Berlin, 1991.

[57] O'Malley Jr., R.E., Kevorkian J. and Cole, J.D. *Perturbation methods in applied mathematics, and Ali Hasan Nayfeh, Introduction to perturbation techniques*, J. Fluid Mech. 71(4) (1982), 827–828.

[58] Phaneendra, K. and Lalu, M. *Numerical solution of singularly perturbed delay differential equations using Gaussian quadrature method*, J. Phys. Conf. Ser. 1344(1) (2019), 012013.

[59] Prasad, H.S. and Reddy, Y.N. *Numerical solution of singularly perturbed differential-difference equations with small shifts of mixed type by differential quadrature method*, Am. J. Comput. Appl. Math. 2(1) (2012), 46–52.

[60] Ranjan, R. *A novel fitted spline method for the numerical treatment of singularly perturbed differential equations having small delays*, J. Appl. Math. Comput. 69(6) (2023), 4645–4664.

[61] Ranjan, R. and Prasad, H.S. *A novel exponentially fitted finite difference method for a class of 2nd order singularly perturbed boundary value problems with a simple turning point exhibiting twin boundary layers*, J. Ambient Intell. Humaniz. Comput. 13(9) (2022), 4207–4221.

[62] Ranjan, R., Prasad, H.S. and Kiltu, G.G. *A novel fitted three-term method for the numerical treatment of singularly perturbed differential–difference equations*, Mediterr. J. Math. 21(2) (2024), 53.

[63] Raza, A., Khan, A., and Sharma, P. and Ahmad, K. *Solution of singularly perturbed differential difference equations and convection delayed dominated diffusion equations using Haar wavelet*, Math. Sci. 15 (2021), 123–136.

[64] Reddy, Y.N. and Reddy, K.A. *Numerical integration method for general singularly perturbed two-point boundary value problems*, Appl. Math. Comput. 133(2–3) (2002), 351–373.

[65] Regal, A.M. and Kumar, S.D. *Non-polynomial spline method for singularly perturbed differential difference equations with delay and advance terms*, Bull. Karaganda Univ. Math. Ser. 118(2) (2025), 208–221.

[66] Regal, A.M. and Kumar, S.D. *Numerical approximation for singularly perturbed differential equations exhibiting significant positive shift arising in neuronal activity*, Appl. Math. Sci. Eng. 33(1) (2025), 2498414.

[67] Sharma, N. and Kaushik, A. *A uniformly convergent difference method for singularly perturbed parabolic partial differential equations with large delay and integral boundary condition*, J. Appl. Math. Comput. 69(1) (2023), 1071–1093.

[68] Shishkin, G.I. and Shishkina, L.P. *The Richardson extrapolation technique for quasilinear parabolic singularly perturbed convection–diffusion equations*, J. Phys. Conf. Ser. 55(1) (2006), 203–209.

[69] Stein, R.B. *Some models of neuronal variability*, Biophys. J. 7(1) (1967), 37–68.

[70] Stynes, M. and Roos, H.-G. *The midpoint upwind scheme*, Appl. Numer. Math. 23(3) (1997), 361–374.

[71] Tian, H. *The exponential asymptotic stability of singularly perturbed delay differential equations with a bounded lag*, J. Math. Anal. Appl. 270(1) (2002), 143–149.

[72] Tuckwell, H.C. *On the first-exit time problem for temporally homogeneous Markov processes*, J. Appl. Probab. (1976), 39–48.

[73] Villasana, M. and Radunskaya, A. *A delay differential equation model for tumor growth*, J. Math. Biol. 47(3) (2003), 270–294.

[74] Wazewska-Czyzewska, M. and Lasota, A. *Mathematical models of the red cell system*, Mat. Stos. 6(1) (1976), 25–40.

[75] Woldaregay, M.M. *Solving singularly perturbed delay differential equations via fitted mesh and exact difference method*, Res. Math. 9(1) (2022), 2109301.

[76] Woldaregay, M.M. and Duressa, G.F. *Higher-order uniformly convergent numerical scheme for singularly perturbed differential difference equations with mixed small shifts*, Int. J. Differ. Equ. 2020(1) (2020), 6661592.

[77] Woldaregay, M.M. and Duressa, G.F. *Robust mid-point upwind scheme for singularly perturbed delay differential equations*, Comput. Appl. Math. 40(5) (2021), 178.

[78] Woldaregay, M.M. and Duressa, G.F. *Robust numerical scheme for solving singularly perturbed differential equations involving small delays*, Appl. Math. E-Notes, 21 (2021), 622–633.

Iranian Journal of Numerical Analysis and Optimization
Vol. 16, No. 3, 2026, pp 790–814
https://doi.org/10.22067/ijnao.2026.96402.1775
https://ijnao.um.ac.ir/

CC BY NC

Research Article

Adaptive-batch stochastic gradient descent for constrained optimization based on relaxed barrier functions

A. Fillali*, S. Addoune and R. Zeghdane

Abstract

Stochastic gradient descent (SGD) is the cornerstone of large-scale optimization; however, its application to problems with a vast number of constraints remains a significant challenge. Methods based on relaxed logarithmic barrier functions have enabled the use of SGD by sampling constraints, yet the reliance on single samples in these methods often leads to high variance and oscillations that hinder progress near the optimal solution.

*Corresponding author

Received 9 November 2025; revised 17 February 2026; accepted 17 February 2026

Abdelouakil Fillali
Laboratoire d'Analyse Mathématiques et ses Applications (LAMA), Department of Mathematics, University Mohamed El Bachir El Ibrahimi, El Anasser, Bordj Bou Arreridj, 34030, Algeria. e-mail: abdelouakil.fillali@univ-bba.dz

Smail Addoune
Laboratoire d'Analyse Mathématiques et ses Applications (LAMA), Department of Mathematics, University Mohamed El Bachir El Ibrahimi, El Anasser, Bordj Bou Arreridj, 34030, Algeria. e-mail: smail.addoune@univ-bba.dz

Rebiha Zeghdane
Laboratoire d'Analyse Mathématiques et ses Applications (LAMA), Department of Mathematics, University Mohamed El Bachir El Ibrahimi, El Anasser, Bordj Bou Arreridj, 34030, Algeria. e-mail: rebiha.zeghdane@univ-bba.dz

How to cite this article
Fillali, A., Addoune, S. and Zeghdane, R., Adaptive-batch stochastic gradient descent for constrained optimization based on relaxed barrier functions. *Iran. J. Numer. Anal. Optim.*, 2026; 16(3): 790-814. https://doi.org/10.22067/ijnao.2026.96402.1775

In this paper, we propose a novel hybrid algorithm, AdaBS-RBSGD, which integrates the relaxed barrier framework with an adaptive batch size strategy. Our algorithm dynamically adjusts the batch size B_k to be inversely proportional to the estimated optimization loss, systematically reducing gradient variance as the iterates approach the solution.

We provide a rigorous mathematical proof establishing the almost-sure convergence of this hybrid algorithm. Furthermore, our theoretical analysis of error bounds demonstrates that active variance management allows the error term associated with stochastic noise to vanish more rapidly compared to the baseline method. Empirically, results on a benchmark problem show that the proposed algorithm significantly outperforms its single-sample counterpart, successfully breaking the "noise floor," achieving higher stability, and reaching a more accurate final solution. This work presents a practical and theoretically sound methodology for accelerating and improving the precision of stochastic constrained optimization.

AMS subject classifications (2020): Primary 90C15, 90C25; Secondary 49M30, 65K05, 68T05.

Keywords: Constrained optimization, Stochastic gradient descent (SGD), Relaxed logarithmic barrier functions, Adaptive batch sizes, Variance reduction, Large-scale optimization.

1 Introduction

Stochastic gradient descent (SGD) and its variants have become cornerstone algorithms in modern large-scale optimization, driving progress in fields ranging from machine learning to networked systems [4, 14, 25]. These methods are fundamentally rooted in the pioneering work on stochastic approximation by Robbins and Monro [16] and later acceleration techniques by Polyak and Juditsky [15]. While SGD is exceptionally efficient for unconstrained problems [10], its effective application to constrained optimization remains an active area of research [19].

Many classical approaches require full knowledge of the constraint set at each iteration, often necessitating computationally expensive projection steps that are prohibitive when the number of constraints is very large [22]. To overcome this challenge, recent methods have been developed that rely on sampling only a subset of constraints [12, 23, 24]. A particularly promising direction is the adaptive relaxed barrier SGD algorithm proposed by Dimitrieski, Cao, and Ebenbauer [7]. This approach avoids expensive projections by using a carefully adapted relaxed barrier function [8, 9], allowing for sampling both the objective function and the constraints.

While the baseline algorithm guarantees almost-sure convergence [7], it suffers from a critical limitation: the inherent high variance of single-sample gradients causes significant oscillations and stagnation near the optimal solution. To overcome this fundamental bottleneck, we introduce AdaBS-RBSGD (Adaptive-batch relaxed barrier SGD), a novel hybrid algorithm that explicitly

targets this variance. Our key contribution lies in the seamless integration of relaxed barrier functions with a dynamic batch-size adaptation strategy, building upon the foundations laid by [1, 2, 3, 20, 21]. Specifically, unlike traditional methods, our approach adopts the logic from [20] to actively suppress gradient noise by inversely scaling the batch size B_k with the optimization loss —a technique shown to be a powerful alternative to decaying the learning rate [21]. This allows the algorithm to break the "noise floor" and enable high-precision convergence without the computational burden of full constraints. We provide a rigorous theoretical analysis establishing the almost-sure convergence of this hybrid method and demonstrate its clear theoretical and empirical advantages in variance reduction.

2 Problem statement and hybrid algorithm

2.1 Problem statement

In this paper, we consider the constrained optimization problem of finding a minimizer for a finite sum objective function, subject to a large number of inequality constraints. The problem is formally stated as:

$$\min_{x\in\mathbb{R}^d} f(x) = \frac{1}{n}\sum_{i=1}^{n} f_i(x) \quad \text{s.t.} \quad g_j(x) \leq 0, \quad j \in \{1,\dots,m\}, \tag{1}$$

where we assume that the objective functions f_i and constraint functions g_j are continuously differentiable ($\mathcal{C}^1$) functions. We adopt the following standard assumptions [5, 6, 18]:

Assumption 1 (Objective function properties)**.** The objective function f and its components f_i are assumed to possess the following properties for constants $L \geq \mu > 0$ and for any $x, y \in \mathbb{R}^d$:

1. **L-Smoothness:** Each component $f_i(x)$ is L-smooth, meaning its gradient is Lipschitz continuous with constant L:

$$f_i(y) \leq f_i(x) + \nabla f_i(x)^T(y-x) + \frac{L}{2}||y-x||^2$$

2. **convexity:** Each component $f_i(x)$ is convex:

$$f_i(y) \geq f_i(x) + \nabla f_i(x)^T(y-x)$$

3. **Strong Convexity:** The total function $f(x)$ is μ-strongly convex:

$$f(y) \geq f(x) + \nabla f(x)^T (y - x) + \frac{\mu}{2} ||y - x||^2.$$

These assumptions, although restrictive, are not uncommon in convergence proofs for gradient-based optimization algorithms [5, 7, 14].

Assumption 2 (Constraint functions)**.** The constraint functions $g_j(x)$ are affine, defined as:

$$g_j(x) := a_j^T x + b_j,$$

where $a_j \in \mathbb{R}^d$ and $b_j \in \mathbb{R}$. The feasible set $\mathbb{X} := \{x \in \mathbb{R}^d : g_j(x) \leq 0, \forall j\}$ is assumed to have a nonempty interior, which is a fundamental condition for applying barrier-based methods [5, 9]

Assumption 3 (Stochastic batch sampling)**.** At each iteration k, given a batch size $B_k \geq 1$, we sample:

- A batch of objective indices: $\mathcal{I}_k = \{i_s\}_{s=1}^{B_k}$, where each i_s is drawn independently and uniformly at random from $\mathcal{U}\{1, \ldots, n\}$.
- A batch of constraint indices: $\mathcal{J}_k = \{j_s\}_{s=1}^{B_k}$, where each j_s is drawn independently and uniformly at random from $\mathcal{U}\{1, \ldots, m\}$.

This uniform distribution is chosen to ensure that the stochastic batch average gradients are unbiased estimators of the true gradients, thereby satisfying the standard assumptions for SGD algorithms [16]. This is demonstrated by the following function:

1. For the objective function: The expected value of the sampled stochastic batch gradient is equal to the true gradient of the full function [7, 16]:

$$\mathbb{E}\left[\frac{1}{B_k}\sum_{s=1}^{B_k} \nabla f_{i_s}(x)\right] = \frac{1}{B_k}\sum_{s=1}^{B_k} \mathbb{E}[\nabla f_{i_s}(x)] = \frac{1}{n}\sum_{i=1}^{n} \nabla f_i(x) = \nabla f(x).$$

2. For the barrier function: The expected value of the sampled stochastic batch gradient of the barrier function equals the average effect of the gradients over all constraints [7]:

$$\begin{aligned}\mathbb{E}\left[\frac{1}{B_k}\sum_{s=1}^{B_k} \nabla B(g_{j_s}(x), \delta)\right] &= \frac{1}{B_k}\sum_{s=1}^{B_k} \mathbb{E}[\nabla B(g_{j_s}(x), \delta)] \\ &= \frac{1}{m}\sum_{j=1}^{m} \nabla B(g_j(x), \delta).\end{aligned}$$

To handle the constraints stochastically, we use the same relaxed logarithmic barrier function $B(z, \delta)$ as defined in [7]. The core idea is to solve a sequence of unconstrained problems:

$$\min_{x \in \mathbb{R}^d} \Phi_k(x) := f(x) + \frac{1}{m} \sum_{j=1}^{m} B(g_j(x), \delta_k),$$

where the barrier parameter $\delta_k = \delta_\infty + \epsilon_k$ is adapted at each iteration k.

For completeness, we state the formulas for $B(z, \delta)$ and its gradient here, as adopted from [7]. The function is defined piecewise as:

$$B(z, \delta) = \begin{cases} -\delta \ \log(-z) & \text{if } z < -\delta, \\ \frac{1}{2}\left(\frac{(z+2\delta)^2}{\delta} - \delta\right) - \delta \ \log(\delta) & \text{if } z \geq -\delta. \end{cases}$$

Its gradient with respect to z, $\nabla_z B(z, \delta)$, used in the SGD updates, is:

$$\nabla_z B(z, \delta) = \begin{cases} -\delta/z & \text{if } z < -\delta, \\ (z + 2\delta)/\delta & \text{if } z \geq -\delta. \end{cases}$$

2.2 The baseline algorithm

The baseline algorithm proposed by Dimitrieski, Cao, and Ebenbauer [7] is a single-sample SGD method. This corresponds to a special case of Assumption 3 where the batch size is fixed to $B_k = 1$ for all k. At each iteration k, it samples one objective index $i_k \sim \mathcal{U}\{1, \ldots, n\}$ and one constraint index $j_k \sim \mathcal{U}\{1, \ldots, m\}$, and updates x_k using the gradient G_k^{base}:

$$G_k^{base} = \nabla f_{i_k}(x_k) + \nabla B(g_{j_k}(x_k), \delta_k), \tag{2}$$

$$x_{k+1} = x_k - \gamma_k G_k^{base}. \tag{3}$$

As shown in our results, the high variance of G_k^{base} hinders the optimization process, limiting the algorithm to the standard SGD convergence rate of $\mathcal{O}(1/\sqrt{k})$ and causing significant oscillations near the optimal solution.

2.3 The proposed hybrid algorithm

To address the high variance of the baseline, we propose a hybrid algorithm, AdaBS-RBSGD, which integrates the relaxed barrier framework with an adaptive batching mechanism [20]. Instead of a single sample, our method draws a batch of B_k indices: $\mathcal{I}_k = \{i_1, \ldots, i_{B_k}\}$ and $\mathcal{J}_k = \{j_1, \ldots, j_{B_k}\}$. The batch gradient G_k is the average:

$$G_k = \frac{1}{B_k} \sum_{i \in \mathcal{I}_k} \nabla f_i(x_k) + \frac{1}{B_k} \sum_{j \in \mathcal{J}_k} \nabla B(g_j(x_k).\delta_k) \tag{4}$$

The batch size B_k is adapted based on the practical batch loss, L_k^{batch}, which is defined as follows:

$$L_k^{batch} = \frac{1}{B_k} \sum_{i \in \mathcal{I}_k} f_i(x_k) + \frac{1}{B_k} \sum_{j \in \mathcal{J}_k} B(g_j(x_k), \delta_k). \tag{5}$$

We follow the RADADAMP logic from [20] to update the batch size using this loss. The process involves tracking a "noise" term t_k and its exponential moving average $\hat{d}_k$.

First, t_k is calculated as a proxy for the 'optimization noise', using the batch loss L_k^{batch} and batch gradient G_k:

$$t_k = L_k^{batch} + \lambda ||G_k||^2. \tag{6}$$

Next, this term is smoothed using an exponential moving average $\hat{d}_k$:

$$\hat{d}_k = \rho \hat{d}_{k-1} + (1 - \rho) t_k. \tag{7}$$

Finally, the batch size for the next iteration, B_{k+1}, is computed to be inversely proportional to this average $\hat{d}_k$.

The intuition is that as the algorithm approaches the optimum, both the batch loss L_k^{batch} and the gradient norm $||G_k||^2$ will decrease, leading to a smaller t_k and consequently a smaller $\hat{d}_k$. This smaller $\hat{d}_k$ will, in turn, lead to a larger batch size B_{k+1}, actively reducing variance as the solution is approached. The conceptual feedback loop of this process is illustrated in Figure 1. The full algorithm is detailed in Algorithm 1.

⋆ Parameters ρ and λ are adopted from the RADADAMP logic [20]. ρ is the 'memory' or smoothing factor for the exponential moving average $\hat{d}_k$ (7), while λ is a balancing coefficient for the 'noise' term t_k (6) which combines the loss L_k^{batch} and the gradient norm $||G_k||^2$.

Algorithm 1 AdaBS-RBSGD (Hybrid adaptive-batch relaxed barrier SGD)

Require: Initial point $x_0 \in \mathbb{R}^d$, total iterations $K \in \mathbb{N}$, initial batch size $B_0 \in \mathbb{N}_{\geq 1}$, max batch size $B_{\max} \in \mathbb{N}_{\geq B_0}$, step-size schedule $\{\gamma_k\}_{k=0}^{K-1} \subset \mathbb{R}^+$, barrier schedule $\{\epsilon_k\}_{k=0}^{K-1} \subset \mathbb{R}^+$, adaptive parameters $\rho \in (0,1)$, $\lambda \in \mathbb{R}^+$, $\delta_\infty \in \mathbb{R}^+$.*

Initialize: $B \leftarrow B_0$, $\hat{d} \leftarrow 0.0$, $d_0 \leftarrow -1.0$

for $k = 0$ to $K-1$ **do**

// Phase 1: Setup Batch Size and Learning Rate

$B_k \leftarrow \lceil B \rceil$

$\gamma'_k \leftarrow \gamma_k$

if $B_k \geq B_{\max}$ **then**

$\gamma'_k \leftarrow \gamma_k \cdot (B_{\max}/B_k)$

$B_k \leftarrow B_{\max}$

end if

Set barrier parameter $\delta_k \leftarrow \delta_\infty + \epsilon_k$

// Phase 2: Compute Stochastic Batch Gradient and Loss

Sample $\mathcal{I}_k = \{i_1, \ldots, i_{B_k}\} \sim \mathcal{U}\{1:n\}^{B_k}$

Sample $\mathcal{J}_k = \{j_1, \ldots, j_{B_k}\} \sim \mathcal{U}\{1:m\}^{B_k}$

$G_k \leftarrow \frac{1}{B_k} \sum_{i \in \mathcal{I}_k} \nabla f_i(x_k) + \frac{1}{B_k} \sum_{j \in \mathcal{J}_k} \nabla B(g_j(x_k), \delta_k)$

$L_k^{batch} \leftarrow \frac{1}{B_k} \sum_{i \in \mathcal{I}_k} f_i(x_k) + \frac{1}{B_k} \sum_{j \in \mathcal{J}_k} B(g_j(x_k), \delta_k)$

// Phase 3: Update Primary Iterate

$x_{k+1} \leftarrow x_k - \gamma'_k G_k$

// Phase 4: Update Batch Size Logic (from [20])

$t_k \leftarrow L_k^{batch} + \lambda \|G_k\|^2$

if $k = 0$ **then**

$\hat{d} \leftarrow t_k$

$d_0 \leftarrow t_k$

else

$\hat{d} \leftarrow \rho \hat{d} + (1-\rho) t_k$

if $\hat{d} > 10^{-8}$ **and** $d_0 > 0$ **then**

$B \leftarrow B_0 \cdot (d_0/\hat{d})$

else

$B \leftarrow B_{\max}$

end if

end if

end for

return x_K

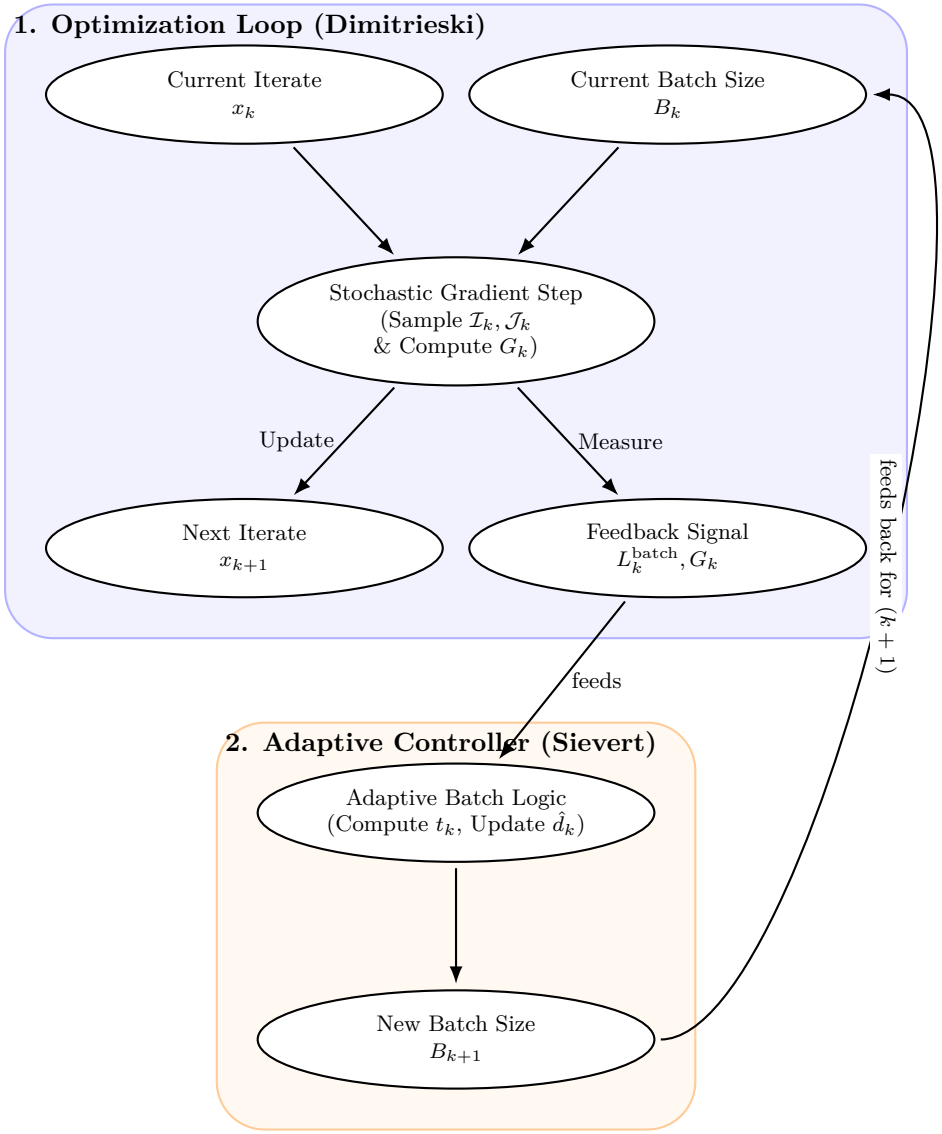


Figure 1: Conceptual diagram of the proposed `AdaBS-RBSGD` algorithm. The system is a feedback loop: (1) The **Optimization Loop** (blue, from [7]) performs the stochastic update and measures the current loss. (2) The **Adaptive Controller** (orange, from [20]) receives this feedback to compute the batch size B_{k+1} for the next iteration, actively managing the gradient variance.

2.4 Design rationale

The proposed AdaBS-RBSGD algorithm is a nontrivial hybrid designed to address the core weaknesses of its parent methods. The relaxed barrier component [7] provides a theoretically-sound mechanism to handle constraints stochastically. However, its primary weakness is the high variance from single-sample gradients. The adaptive batching component [20] directly targets this weakness. By dynamically increasing B_k as the loss decreases, the algorithm actively suppresses the gradient variance, allowing for more stable and precise steps near the optimum.

Noise Handling and Adaptation Speed: The algorithm manages stochastic noise by continuously monitoring the noise proxy t_k (6), which acts as a real-time signal of the variance in the optimization landscape. The responsiveness of the algorithm to changing noise levels is explicitly controlled by the memory factor ρ in the exponential moving average (7). A lower ρ enables rapid adaptation to sudden noise shifts but may lead to instability, whereas a higher ρ (such as our chosen $\rho = 0.999$) ensures a smoother, more gradual adaptation. This conservative approach effectively filters out transient noise spikes, allowing the batch size to grow steadily only when a persistent trend of variance reduction is detected, thus maintaining algorithmic stability.

Implicit Determination of Batch Growth Rate: Unlike heuristic schedules that enforce a fixed growth rate, the rate of increase for B_k in AdaBS-RBSGD is not predetermined but is fully data-driven. It is governed implicitly by the inverse relationship with the smoothed noise metric $\hat{d}_k$ (7). As the optimization loss decreases, $\hat{d}_k$ diminishes, naturally driving B_k upward. Consequently, the "optimal" rate emerges automatically from the landscape geometry: regions of significant progress induce rapid batch scaling, while stagnation leads to slower growth. This mechanism aligns with the practical guideline established by Smith et al. [21], which posits that increasing the batch size is theoretically equivalent to decaying the learning rate γ_k. Thus, practically, our algorithm decouples the manual tuning of batch schedules from the learning rate, allowing the batch size to adjust automatically based on the solver's progress.

3 Convergence analysis

In this section, we provide the formal statement of the convergence theorem for the proposed AdaBS-RBSGD algorithm (Algorithm 1). The rigorous mathematical proof, structured into logical steps, follows in Section 4.

Theorem 1 (Almost sure convergence of AdaBS-RBSGD)**.** Consider the constrained optimization problem (1) and assume Assumptions 1, 2, and 3 hold. Let $\{X_k\}_{k\geq 0}$ be the sequence of

iterates generated by Algorithm 1. Let $\{\gamma_k\}_{k\geq 0}$ and $\{\epsilon_k\}_{k\geq 0}$ be positive sequences such that $\delta_k = \delta_\infty + \epsilon_k$ (with $\delta_\infty > 0$) satisfying

(a) $\sum_{k=0}^{\infty} \gamma_k = \infty$,

(b) $\sum_{k=0}^{\infty} \gamma_k^2 < \infty$,

(c) $\sum_{k=0}^{\infty} \gamma_k \epsilon_k < \infty$.

Then, the sequence of iterates $\{X_k\}_{k\geq 0}$ converges almost surely (a.s.) to the unique minimizer of the relaxed problem, $x^*(\delta_\infty)$. That is:

$$\lim_{k\to\infty} X_k = x^*(\delta_\infty) \quad \text{a.s.}$$

Remark 1 (Convergence conditions and failure modes)**.** The almost-sure convergence established in Theorem 1 relies strictly on Assumptions 1-3. Consequently, the method guarantees convergence primarily for strongly convex objectives subject to affine constraints, provided the step-size conditions in Theorem 1 are met. The algorithm may fail or exhibit instability under the following specific conditions:

1. **Violation of convexity:** Applying the method to nonconvex objectives violates Assumption 1. Without strong convexity, the drift analysis in the proof (specifically the bound on $\langle \nabla\Phi(X_k), X_k\rangle$) no longer holds, risking entrapment in local minima or divergence.

2. **Nonaffine constraints:** Assumption 2 requires affine constraints to ensure the barrier gradient variance remains quadratically bounded (as shown in (11) of the analysis). With general nonlinear constraints, this bound may be violated, potentially leading to explosive gradient variance.

3. **Incompatible parameter decay:** Convergence requires the specific interplay between step-size and barrier parameters ($\sum \gamma_k \epsilon_k < \infty$). If ϵ_k decays too rapidly relative to γ_k, the bias error term will not vanish, preventing convergence to the exact solution.

4 Proof of almost-sure convergence

In this section, we provide a rigorous, step-by-step convergence analysis of the proposed AdaBS-RBSGD algorithm. We rely on the Robbins-Siegmund supermartingale convergence lemma [17] and the techniques for biased stochastic approximation [11] to establish that the iterate sequence $\{X_k\}$ converges almost surely to the optimal solution x^*.

Proof of Theorem 1. We define the Lyapunov function R_k as the squared Euclidean distance between the current iterate X_k and the unique minimizer $x^* := x^*(\delta_\infty)$ of the limit problem:

$$R_k := \|X_k - x^*\|^2.$$

Let $\mathcal{F}_k$ denote the σ-algebra generated by the random variables $\{X_0, \dots, X_k\}$. Our goal is to analyze the conditional expectation $\mathbb{E}[R_{k+1}|\mathcal{F}_k]$.

Step 1: Expansion of the iteration

Recall the update rule: $X_{k+1} = X_k - \gamma_k G_k$. We expand the squared norm:

$$\begin{aligned} R_{k+1} &= \|X_k - \gamma_k G_k - x^*\|^2 \\ &= \|(X_k - x^*) - \gamma_k G_k\|^2 \\ &= \|X_k - x^*\|^2 - 2\gamma_k \langle G_k, X_k - x^* \rangle + \gamma_k^2 \|G_k\|^2. \end{aligned}$$

Taking the conditional expectation with respect to $\mathcal{F}_k$:

$$\mathbb{E}[R_{k+1}|\mathcal{F}_k] = R_k - 2\gamma_k \langle \mathbb{E}[G_k|\mathcal{F}_k], X_k - x^* \rangle + \gamma_k^2 \mathbb{E}[\|G_k\|^2|\mathcal{F}_k]. \tag{8}$$

Measurability note: Before proceeding, it is important to note that since the current iterate X_k is $\mathcal{F}_k$-measurable (fully determined by the history up to step k), the terms $\nabla\Phi(X_k)$ and $\nabla C_k(X_k)$ are deterministic constants with respect to the conditional expectation $\mathbb{E}[\cdot|\mathcal{F}_k]$. Consequently, terms such as $\|\nabla\Phi(X_k)\|$ satisfy $\mathbb{E}[\|\nabla\Phi(X_k)\| \mid \mathcal{F}_k] = \|\nabla\Phi(X_k)\|$, as the expectation operator acts solely on the random batch indices $\mathcal{I}_k$ and $\mathcal{J}_k$ drawn at the current iteration.

Step 2: Analysis of the expected gradient (Drift Term)

We analyze the expected gradient term $\mathbb{E}[G_k|\mathcal{F}_k]$. The batch gradient G_k is the average of B_k independent samples. Let $\nabla\psi_{i,j}(x) = \nabla f_i(x) + \nabla B(g_j(x), \delta_k)$ be a single sample gradient.

$$\mathbb{E}[G_k|\mathcal{F}_k] = \mathbb{E}\left[\frac{1}{B_k}\sum_{s=1}^{B_k} \nabla\psi_{i_s,j_s}(X_k)\middle|\mathcal{F}_k\right] = \mathbb{E}[\nabla\psi_{i,j}(X_k)|\mathcal{F}_k].$$

We decompose this expectation into the target gradient $\nabla\Phi(X_k)$ and the bias term $\nabla C_k(X_k)$. Following the notation in [7], we define the optimality gap of the potential function as:

$$\Phi_0(X_k) := \Phi(X_k) - \Phi(x^*) \geq 0. \tag{9}$$

Substituting the decomposition back into the inner product term in (8):

$$-\langle \mathbb{E}[G_k|\mathcal{F}_k], X_k - x^* \rangle = -\langle \nabla\Phi(X_k), X_k - x^* \rangle - \langle \nabla C_k(X_k), X_k - x^* \rangle.$$

By the strong convexity of $\Phi(x)$**, we have:

$$-\langle \nabla\Phi(X_k), X_k - x^* \rangle \leq -(\Phi_0(X_k) + \frac{\mu}{2}\|X_k - x^*\|^2).$$

Thus, applying the Cauchy-Schwarz inequality, the drift term is bounded by:

$$-2\gamma_k \langle \mathbb{E}[G_k], \tilde{X}_k \rangle \leq -\gamma_k(2\Phi_0(X_k) + \mu R_k) + 2\gamma_k \|\nabla C_k(X_k)\| \sqrt{R_k}. \tag{10}$$

Step 3: Analysis of the variance (Noise Term)

We analyze $\mathbb{E}[\|G_k\|^2|\mathcal{F}_k]$. Since samples are i.i.d., the variance of the average decreases with B_k:

$$\begin{aligned}\mathbb{E}[\|G_k\|^2|\mathcal{F}_k] &= \|\mathbb{E}[G_k]\|^2 + \mathrm{Var}(G_k|\mathcal{F}_k) \\ &= \|\nabla\Phi(X_k) + \nabla C_k(X_k)\|^2 + \frac{1}{B_k}\mathrm{Var}(\nabla\psi_{i,j}(X_k)).\end{aligned}$$

Using the inequality $\|a+b\|^2 \leq 2\|a\|^2 + 2\|b\|^2$ and the bound on single-sample variance:

$$\gamma_k^2 \mathbb{E}[\|G_k\|^2|\mathcal{F}_k] \leq \gamma_k^2 \left(2\|\nabla C_k(X_k)\|^2 + 2\|\nabla\Phi(X_k)\|^2 + \frac{\sigma^2}{B_k}\right). \tag{11}$$

Crucially, the stochastic noise term $\frac{\sigma^2}{B_k}$ is explicitly reduced by the batch size.

Step 4: Synthesis and convergence

Substituting the bounds from (10) and (11) into (8), and utilizing the upper bounds for $\|\nabla C_k(X_k)\|$ and $\|\nabla\Phi(X_k)\|^2$ established in [7] (specifically in [7, Eqs. (8)–(10)]), we derive the following supermartingale inequality:

$$\mathbb{E}[R_{k+1}|\mathcal{F}_k] \leq (1 - \gamma_k\mu + \gamma_k\xi_k)R_k$$

** The function $\Phi(x)$ is strongly convex as it is the sum of a strongly convex function $f(x)$ (Assumption 1) and a sum of convex functions. See [7], Appendix V-D, for the formal proof.

$$- 2\gamma_k(1 - 4\gamma_k\hat{L})\Phi_0(X_k) + \underbrace{\left(\beta_k(\epsilon_k) + \frac{\gamma_k^2\sigma^2}{B_k}\right)}_{Q_k}, \tag{12}$$

where $\hat{L}$ is the modified smoothness constant, and Q_k collects the error terms[***]. We verify the conditions for Robbins-Siegmund:

1. The sequence $\gamma_k\xi_k$ is summable since $\sum\gamma_k\epsilon_k < \infty$.
2. The noise term Q_k is summable since $\sum\gamma_k^2 < \infty$ and $B_k \geq 1$.
3. For sufficiently large k, $(1 - 4\gamma_k\hat{L}) > 0$.

Therefore, R_k converges almost surely to a random variable R_∞. Furthermore, we have $\sum_{k=0}^{\infty}\gamma_k\Phi_0(X_k) < \infty$. Given that $\sum\gamma_k = \infty$, this implies $\liminf_{k\to\infty}\Phi_0(X_k) = 0$. Since Φ_0 is continuous and nonnegative, and R_k converges, we conclude $X_k \to x^*(\delta_\infty)$ almost surely. □

Remark 2 (Theoretical advantage)**.** The derivation in Step 3 highlights the theoretical advantage of the hybrid approach over the single-sample baseline proposed in [7]. Specifically, the variance term for the hybrid algorithm behaves as:

$$Q_{var}^{hybrid} \sim \mathcal{O}\left(\frac{\gamma_k^2(\sigma_\Phi^2 + \sigma_C^2(\epsilon_k))}{B_k}\right).$$

In contrast, the baseline method (where $B_k = 1$) has a variance term of order $\mathcal{O}(\gamma_k^2(\sigma_\Phi^2 + \sigma_C^2(\epsilon_k)))$ (see [7, Eq. 20]). As B_k increases adaptively, the noise term in AdaBS-RBSGD vanishes at a faster rate, allowing the algorithm to achieve lower error floors compared to the baseline.

5 Numerical experiments

In this section, we present a comprehensive empirical investigation to validate the theoretical claims of our proposed AdaBS-RBSGD algorithm. The experiments are designed to demonstrate its practical advantages over the baseline single-sample method from Dimitrieski, Cao, and Ebenbauer [7].

[***] In the baseline proof [7], the total error term is $\gamma_k\epsilon_k r_k + 4\gamma_k^2\sigma_\Phi^2$. Our Q_k term modifies the variance component to account for $B_k > 1$, while preserving the structural terms ξ_k and $\hat{L}$ derived from the barrier relaxation.

5.1 Experimental setup

To ensure a fair and reproducible evaluation, the experiments were conducted on the Google Colab platform with the following specifications:

- **CPU:** Intel(R) Xeon(R) CPU @ 2.20GHz
- **RAM:** 13.61 GB
- **GPU:** NVIDIA Tesla T4 (15 GB VRAM)
- **OS:** Linux

Test problem generation: We utilize the large-scale quadratic programming benchmark framework introduced in [7]. The problem is constructed to ensure that the unconstrained minimizer lies outside the feasible region, thereby enforcing active constraints at the optimal solution. The objective function $f(x)$ is defined as:

$$f(x) = \frac{1}{n}\sum_{i=1}^{n}\sum_{j=1}^{d}\left(\alpha_{i,j}x_j + \log\left(1+e^{-\alpha_{i,j}x_j}\right) + (x_j-\beta)^2\right),$$

where the dimension is set to $d = 50$ and the number of components is $n = 10$. The coefficients $\alpha_{i,j}$ are sampled uniformly from $\mathcal{U}(0.1, 1.1)$. The shift parameter is set to $\beta \approx 2.365$, such that the Euclidean norm of the unconstrained minimizer is approximately $\|x^*_{unc}\|_2 \approx 15$, ensuring it violates the constraints.

The feasible set is defined by $m = 10,000$ affine inequality constraints, generated to form an outer approximation of an ellipsoidal set $\mathcal{E} := \{x \in \mathbb{R}^d : x^\top Qx \leq 100\}$. Here, Q is a positive definite diagonal matrix with entries $q_i \sim \mathcal{U}(1, 1.5)$, and the constraint normal vectors are randomly sampled from the boundary of $\mathcal{E}$. This construction guarantees a compact feasible set satisfying the nonempty interior assumption.

The optimal solution x^* is pre-computed using the L-BFGS-B algorithm on the deterministic barrier objective. Crucially, to ensure statistical robustness, the results presented in this section are averaged over $R = 1000$ **independent simulation trajectories**, providing a significantly more rigorous validation compared to standard evaluations.

5.2 Parameter robustness analysis

To assess the stability of our proposed hybrid algorithm, we conducted a comprehensive grid search over the two key adaptive parameters from the RADADAMP logic [20]: the memory factor ρ (RHO) and the regularization coefficient λ (REG LAMBDA). We evaluated the Mean Final Error $||x_k - x^*||_2$ at $K = 5000$ across a grid of $\rho \in \{0.9, 0.99, 0.999\}$ and $\lambda \in \{10^{-4}, 10^{-3}, 10^{-2}\}$.

Contrarily to findings in previous studies which noted performance degradation at lower memory factors, our implementation demonstrated remarkable robustness. As illustrated in Figure 2, the algorithm maintained a consistently low final error norm (approximately 6.6×10^{-3} to 6.8×10^{-3}) across the entire parameter grid. The uniform distribution of low error values indicates that the adaptive batching mechanism effectively suppresses gradient noise regardless of minor variations in hyperparameter settings.

Based on these results, we proceeded with the conservative settings of $B_0 = 2$, $B_{\max} = 8$, $\rho = 0.999$, and $\lambda = 10^{-3}$ for the main experiments, although our analysis suggests that the algorithm remains stable and precise over a much wider range of configurations.

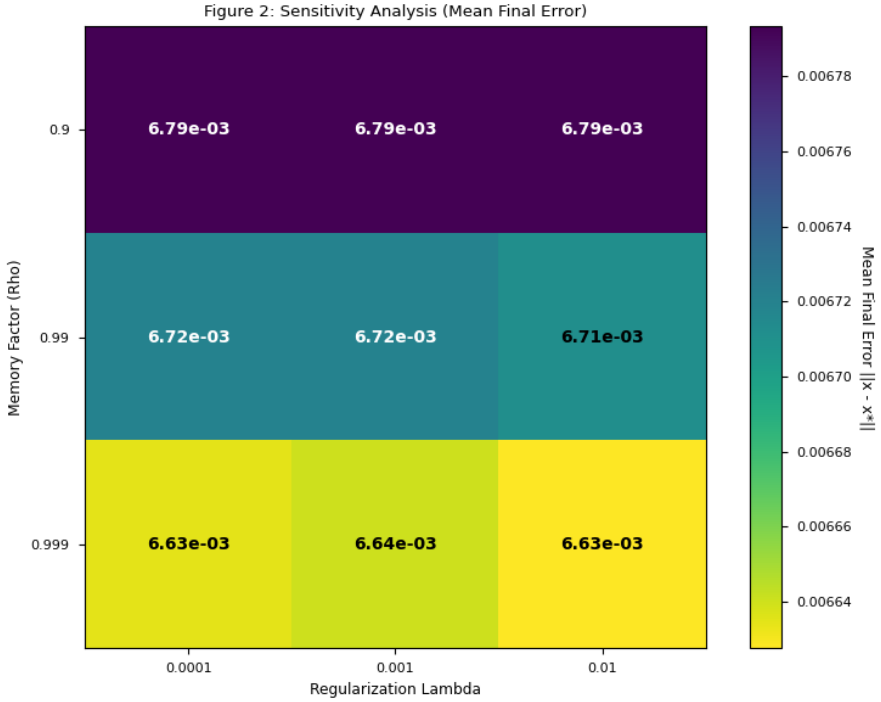


Figure 2: **Robustness Analysis:** Mean Final Error $||x_k - x^*||_2$ at $K = 5000$ for varying memory factors (ρ) and regularization coefficients (λ). The uniform color distribution highlights the stability of the AdaBS-RBSGD algorithm against parameter variations.

5.3 Algorithms for comparison

We compare the performance of two algorithms:

- **Base SGD (Baseline):** The adaptive relaxed barrier SGD algorithm from [7], which corresponds to our Algorithm 1 with a fixed batch size of $B_k = 1$.
- **AdaBS-RBSGD (Ours):** Our proposed hybrid algorithm using the adaptive parameters defined in Section 5.2.

5.4 Performance evaluation

We now evaluate the algorithms based on convergence speed, final solution accuracy, and variance. The results are averaged over $R = 1000$ independent simulation trajectories.

5.4.1 Error vs. iterations

Figure 3 plots the primary performance metric, the Euclidean distance to the optimum $||x_k - x^*||_2$, against the number of iterations (k). The results clearly validate our central hypothesis.

- **Baseline (Red):** The baseline algorithm converges quickly but stagnates around a high-variance "noise floor" (approx. 1.4×10^{-2}). The shaded red area remains large, preventing further progress.
- **AdaBS-RBSGD (Green):** Our hybrid algorithm initially tracks the baseline. However, as it approaches the optimum, the adaptive batching mechanism activates. The algorithm breaks away from the baseline's noise floor, actively suppressing variance. By iteration 10,000, it achieves a final solution that is approximately 64% more accurate ($\approx 4.9 \times 10^{-3}$) than the baseline.

Robustness to Ill-Conditioning: The superior stability observed in Figure 3 also suggests resilience against ill-conditioning. In ill-conditioned landscapes, where high curvature in specific directions causes single-sample SGD to oscillate violently (zig-zagging), the adaptive increase in batch size B_k plays a stabilizing role. By averaging out the stochastic noise components, the algorithm computes a more accurate approximation of the true descent direction. This effectively dampens the erratic oscillations typical of ill-conditioned problems, allowing AdaBS-RBSGD to

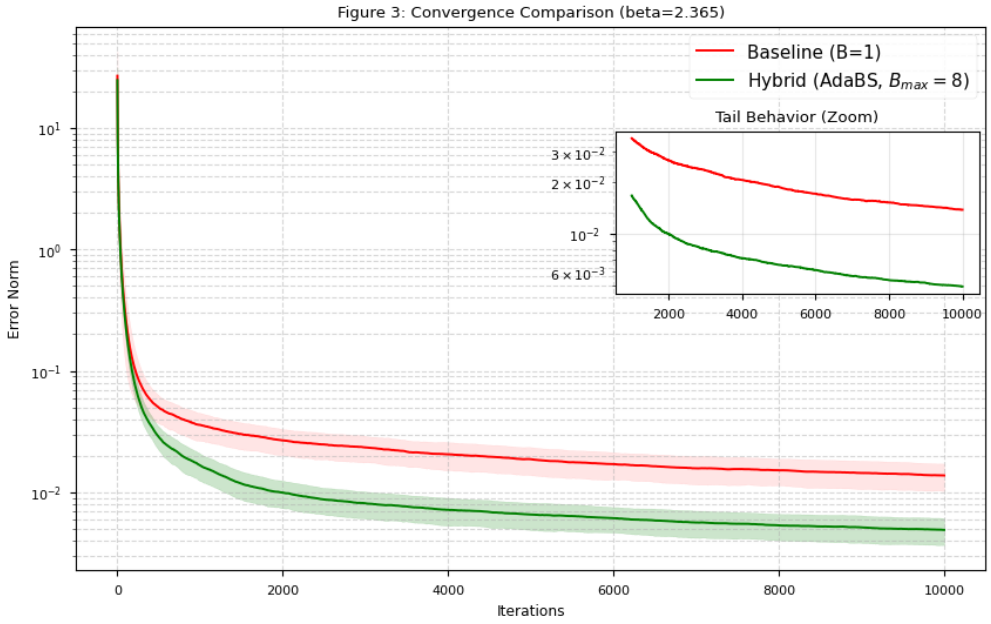


Figure 3: Algorithm Performance Comparison ($n = 10, m = 10000$). The plot shows the mean $||x_k - x^*||_2$ (solid line) and ± 1 standard deviation (shaded area) over $R = 1000$ runs.

maintain steady progress where the baseline stagnates. Regarding nonconvexity, as noted in Remark 1, while the current theory assumes convexity, this variance reduction mechanism is expected to be empirically beneficial in navigating nonconvex landscapes, a subject reserved for future investigation.

5.4.2 Adaptive batching mechanism

Figure 4 explains *how* the hybrid algorithm achieves this variance suppression. It plots the behavior of the mean batch size B_k over the $R = 1000$ trajectories. B_k starts at $B_0 = 2$ allowing fast initial exploration. As the optimization loss decreases, B_k rapidly increases until it hits the cap $B_{\max} = 8$. This transition from a low-batch to a controlled high-batch regime is precisely what actively suppresses the gradient variance and enables continued convergence.

5.4.3 Computational efficiency

Figure 5 provides a critical analysis of computational efficiency by plotting the error against the total number of gradient computations.

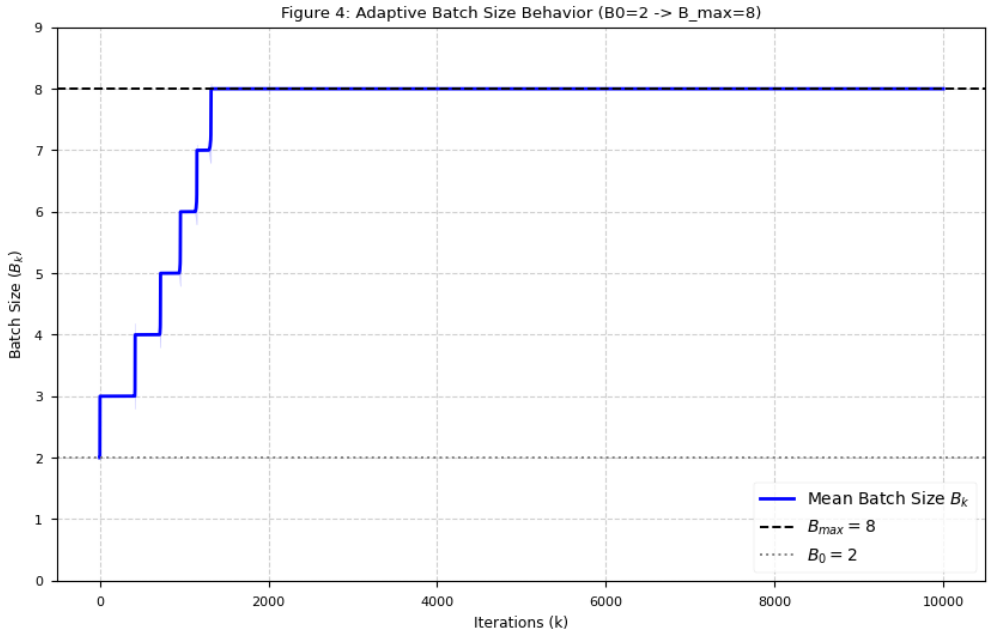


Figure 4: Adaptive batch size behavior (Hybrid algorithm). The plot shows the mean B_k increasing from $B_0 = 2$ to $B_{\max} = 8$ as the algorithm converges.

- **Baseline (Red):** Efficient for low-accuracy solutions but hits a precision ceiling at $\approx 1.4 \times 10^{-2}$, regardless of the computational budget.

- **AdaBS-RBSGD (Green):** While requiring slightly more gradient computations initially, the hybrid method successfully continues to converge past the baseline's ceiling. It trades a marginal increase in computation for a higher precision solution that is impossible for the baseline to achieve.

Scalability with large number of constraints: A key advantage of the adaptive sampling strategy is its scalability. The computational cost per iteration is dominated by the gradient evaluations of the sampled batch, scaling as $\mathcal{O}(B_k \cdot d)$, where d is the dimension of the decision variable. Crucially, this cost is independent of the total number of constraints m. Even as m grows extremely large (high-dimensional constraint space), the batch size remains capped at $B_{\max} \ll m$. Consequently, AdaBS-RBSGD maintains high computational efficiency and is well-suited for large-scale problems where evaluating all constraints is prohibitive.

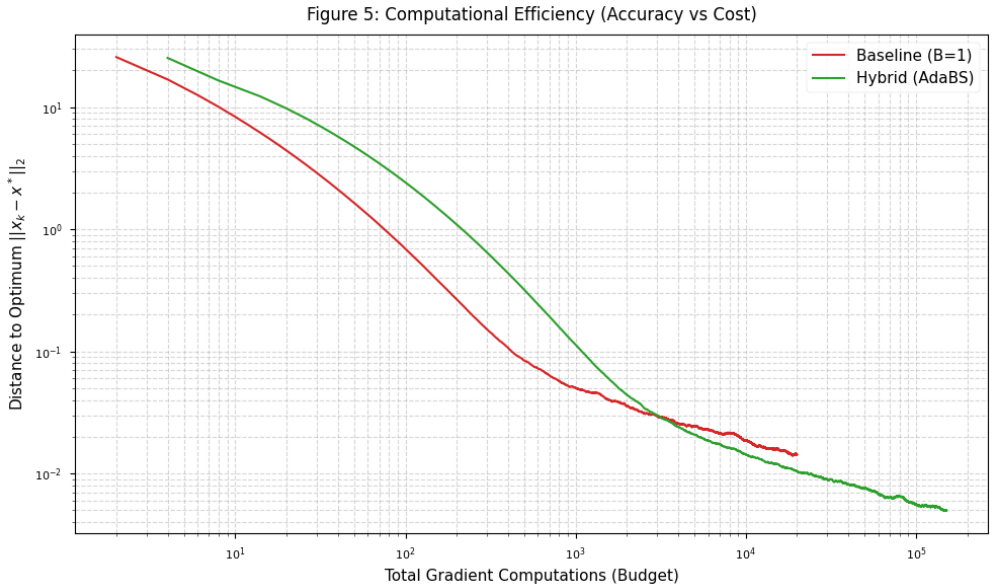


Figure 5: Computational Efficiency. The plot shows the mean $||x_k - x^*||_2$ against the total cumulative gradient computations (log-log scale).

5.4.4 Final solution quality and computational cost

Finally, we analyze the final iterate x_K after $K = 10,000$ iterations. Table 1 summarizes the mean objective value $f(x_K)$, maximum constraint violation, and total wall-clock time.

Table 1: Comparison Results (Mean ± Std. Dev. over $R = 1000$ runs)

Algorithm	**Final Error** $||x - x^*||_2$	**Max Violation**	**Mean Objective**	**Total Time (s)**
Dimitrieski (Baseline)	$1.38 \times 10^{-2} (\pm 3.48 \times 10^{-3})$	0.00(±0.00)	81.7414(±0.0001)	1.79
AdaBS-RBSGD (Ours)	$\mathbf{4.91 \times 10^{-3}} (\pm \mathbf{1.22 \times 10^{-3}})$	0.00(±0.00)	81.7412(±0.0000)	2.53

The table provides two key insights:

1. **Accuracy stability:** Both algorithms find feasible solutions. However, the hybrid algorithm achieves a significantly lower final error norm. More importantly, the standard deviation of our method ($\pm 1.22 \times 10^{-3}$) is approximately a factor of 3 lower than the baseline ($\pm 3.48 \times 10^{-3}$), indicating superior algorithmic stability.

2. **Computational Trade-offs:** The transition from a fixed single-sample batch to an adaptive batch size introduces a fundamental trade-off. By dynamically increasing B_k, the per-iteration computational cost rises (since B_k gradient evaluations are required instead of one). This explains why the hybrid algorithm's total wall-clock time (2.53s)

is higher than the baseline (1.79s). However, this additional computational investment is strictly necessary to suppress the stochastic noise that limits the baseline's precision. Effectively, we trade a controlled increase in runtime for a significant gain in solution accuracy and stability. Practically, this trade-off is kept efficient by the cap $B_{\max} = 8$, ensuring the algorithm remains computationally feasible while achieving a precision level that the fixed-batch baseline cannot reach.

5.5 Sensitivity to batch capacity limit ($B_{\max}$)

To rigorously justify our choice of the batch capacity limit $B_{\max} = 8$, we conducted an ablation study analyzing the trade-off between solution accuracy and computational cost. We evaluated the hybrid algorithm across a range of capacity limits $B_{\max} \in \{2, 4, 8, 16, 32, 64\}$, averaging the results over independent trials to ensure statistical significance.

The results, illustrated in Figure 6, reveal a distinct trade-off point at $B_{\max} = 8$.

- **Significant accuracy gain:** Increasing the limit from $B_{\max} = 2$ to 8 results in a sharp reduction in the final error norm $||x_K - x^*||$ (from $\approx 9.5 \times 10^{-3}$ to $\approx 4.9 \times 10^{-3}$), effectively halving the error with only a minimal increase in execution time.

- **Computational stability regime:** For $B_{\max} \leq 8$, the wall-clock time (dashed red line) remains remarkably stable, showing only a slight linear increase. This indicates that the adaptive batching is efficiently utilizing the available computational resources without overhead.

- **Cost-accuracy Trade-off:** Beyond $B_{\max} = 8$, while the error continues to decrease, the computational cost exhibits a steep, super-linear increase. For instance, achieving the marginal accuracy gain at $B_{\max} = 64$ requires nearly doubling the execution time compared to $B_{\max} = 8$.

Consequently, $B_{\max} = 8$ represents the optimal balance on the Pareto frontier, maximizing solution precision while maintaining the computational efficiency that characterizes our proposed method.

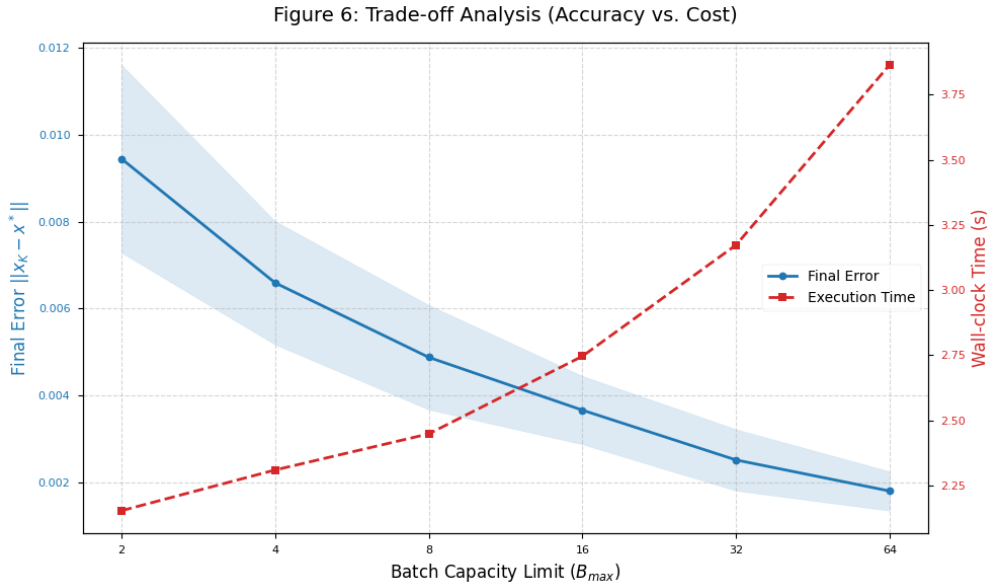


Figure 6: Trade-off Analysis: Final error (solid blue) vs. Wall-clock time (dashed red) for varying batch capacity limits ($B_{\max}$). The plot demonstrates that $B_{\max} = 8$ offers the optimal balance, avoiding the steep computational penalty observed at higher capacities.

6 Conclusion

In this paper, we introduced AdaBS-RBSGD, a novel and efficient hybrid algorithm for solving stochastic constrained optimization problems. Our work builds upon the modern framework established by Dimitrieski, Cao, and Ebenbauer [7], addressing its critical limitation of high gradient variance which leads to significant oscillation near the optimum. To overcome this, we integrated the adaptive batching strategy (RADADAMP) from Sievert-Shah [20]. Our hybrid algorithm dynamically adjusts the batch size B_k to be inversely proportional to the optimization loss, actively suppressing gradient variance as the iterates approach the solution.

Our primary contribution is twofold. First, we provided a rigorous theoretical analysis establishing the almost-sure convergence of this nontrivial hybrid algorithm (Theorem 1). Second, we presented comprehensive numerical experiments that empirically validate our approach. Our results clearly demonstrate that AdaBS-RBSGD successfully breaks away from the high-variance 'noise floor' that stalls the baseline method, allowing our algorithm to achieve a significantly more accurate and stable final solution.

This superior accuracy involves a manageable computational trade-off. As our experiments with the optimized batch cap ($B_{\max} = 8$) confirm, achieving this high-precision solution requires only a moderate increase in the computational budget (approx. 1.4× time increase) compared

to the baseline method. This work thus presents a practical and theoretically-sound method for accelerating stochastic constrained optimization, offering a valuable trade-off for applications where solution precision and stability are paramount.

Future work

The promising results of our AdaBS-RBSGD algorithm open several interesting avenues for future research:

- **Analysis of convergence rate:** Our current analysis proves almost-sure convergence. A valuable theoretical extension would be to derive the explicit convergence rate of the hybrid AdaBS-RBSGD algorithm, similar to analyses found in [13], formally proving that it achieves a faster rate than the baseline, as suggested by our variance analysis.

- **Interaction with step-size:** This work used a pre-defined decaying step-size γ_k. Future work should investigate the synergy between the adaptive batch size B_k and adaptive learning rates (e.g., AdaGrad-style scaling), as both mechanisms respond to the optimization landscape.

- **Nonaffine constraints:** The current analysis, similar to the baseline work [7], is focused on problems with affine (linear) constraints. Investigating the performance, stability, and theoretical guarantees of the AdaBS-RBSGD approach when applied to problems with general nonlinear constraints would be an important next step.

- **Nonuniform sampling:** Our algorithm currently employs standard uniform sampling for constraints. Future work could explore the synergy between adaptive batching and nonuniform or importance sampling techniques. For instance, adaptively sampling constraints that are more frequently violated might accelerate convergence.

- **Extension to non convex objectives:** While our current theoretical framework relies on the assumption of strong convexity to guarantee almost-sure convergence (Assumption 1), stochastic gradient methods are widely used in nonconvex settings (e.g., neural network training). Investigating the empirical performance and establishing local convergence guarantees for AdaBS-RBSGD in nonconvex optimization landscapes represents a significant and promising direction for future research.

Funding

This research received no external funding.

Conflicts of Interest

The authors declare no conflicts of interest.

References

[1] Balles, L., Romero, J. and Hennig, P. *Coupling adaptive batch sizes with learning rates*, In 33rd Conf. Uncertainty in Artif. Intell. (UAI), 675–684, 2017.

[2] Beiser, F., Keith, B., Urbainczyk, S. and Wohlmuth, B. *Adaptive sampling strategies for risk-averse stochastic optimization with constraints,* IMA J. Numer. Anal. 43(6), (2023) 3729–3765.

[3] Bollapragada, R., Byrd, R. and Nocedal, J. *Adaptive sampling strategies for stochastic optimization,* SIAM J. Optim. 28(4), (2018) 3312–3343.

[4] Bottou, L., Curtis, F.E. and Nocedal, J. *Optimization methods for large-scale machine learning,* SIAM Rev. 60(2) (2018), 223–311.

[5] Boyd, S. and Vandenberghe, L. Convex optimization, Cambridge University Press, 2004.

[6] Bubeck, S. *Convex optimization: Algorithms and complexity,* Found. Trends Mach. Learn. 8(3-4) (2015), 231–358.

[7] Dimitrieski, N., Cao, J. and Ebenbauer, C. *Stochastic gradient descent for constrained optimization based on adaptive relaxed barrier functions*, IEEE Control Syst. Lett. 9 (2025) 829–834.

[8] Feller, C. and Ebenbauer, C. *A stabilizing iteration scheme for model predictive control based on relaxed barrier functions,* Automatica 80 (2017), 328–339.

[9] Fiacco, A.V. and McCormick, G.P. *Nonlinear programming: Sequential unconstrained minimization techniques*, John Wiley & Sons, 1968.

[10] Garrigos, G. and Gower, R.M. *Handbook of convergence theorems for (stochastic) gradient methods,* arXiv preprint arXiv:2301.11235 (2023).

[11] Karandikar, R.L. and Vidyasagar, M. *Convergence rates for stochastic approximation: biased noise with unbounded variance, and applications*, J. Optim. Theory Appl. 203(3) (2024) 2412–2450.

[12] Li, M., Grigas, P. and Atamtürk, A. *New penalized stochastic gradient methods for linearly constrained strongly convex optimization,* J. Optim. Theory Appl. 205(2) (2025) 29.

[13] Liu, J. and Yuan, Y. *On almost sure convergence rates of stochastic gradient methods,* In Proc. 35th Annu. Conf. Learn. Theory (COLT), PMLR, 178, 1–21, 2022.

[14] Nocedal, J. and Wright, S.J. *Numerical optimization*, 2nd ed., Springer, 2006.

[15] Polyak, B.T. and Juditsky, A.B. *Acceleration of stochastic approximation by averaging,* SIAM J. Control Optim. 30(4) (1992), 838–855.

[16] Robbins, H. and Monro, S. *A stochastic approximation method,* Ann. Math. Stat. 22(3) (1951), 400-407.

[17] Robbins, H. and Siegmund, D. *A convergence theorem for nonnegative almost supermartingales and some applications*, In Optimizing Methods in Statistics, Academic Press, 233–257, 1971.

[18] Rockafellar, R.T. *Convex analysis*, Princeton University Press, 1970.

[19] Roy, S.K. and Harandi, M. *Constrained stochastic gradient descent: The good practice,* In Int. Conf. Digit. Image Comput. Tech. Appl. (DICTA), IEEE, 1–8, 2017.

[20] Sievert, S. and Shah, S. *Improving the convergence of SGD through adaptive batch sizes,* arXiv preprint arXiv:1910.08222 (2019).

[21] Smith, S.L., Kindermans, P.J. and Le, Q.V. *Don't decay the learning rate, increase the batch size*, In Int. Conf. Learn. Represent. (ICLR), 2018.

[22] Wang, M. and Bertsekas, D.P. *Incremental constraint projection methods for variational inequalities,* Math. Program. 150 (2015), 321–363.

[23] Wang, X., Ma, S. and Yuan, Y. *Penalty methods with stochastic approximation for stochastic nonlinear programming,* Math. Comput. 86(306) (2017) 1793–1820.

[24] Yan, Y. and Xu, Y. *Adaptive primal-dual stochastic gradient method for expectation-constrained convex stochastic programs,* Math. Program. Comput. 14(2) (2022), 319-363.

[25] Zhang, T. *Solving large scale linear prediction problems using stochastic gradient descent algorithms,* In Proc. 21st Int. Conf. Mach. Learn. (ICML), 116, 2004.

Iranian Journal of Numerical Analysis and Optimization
Vol. 16, No. 3, 2026, pp 815–851
https://doi.org/10.22067/ijnao.2026.95409.1728
https://ijnao.um.ac.ir/

CC BY NC

Research Article

Comparative semi-analytical solutions of Caputo time-fractional Fokker–Planck equations via Laplace transform-based decomposition and iterative methods

M. Hossain*, M. Aman Ullah

Abstract

In this study, the Laplace Adomian decomposition method and the Laplace variational iteration method are employed to obtain approximate analytical solutions for both linear and nonlinear time-fractional Fokker–Planck equations (TFFPEs). These methods combine the Laplace transform with the Adomian decomposition method and the variational iteration method, respectively. Fractional derivatives are considered in the Caputo sense. The proposed approaches are straightforward to implement and yield solutions without requiring linearization, discretization, or perturbation. Several examples are solved and presented both graphically and numerically to illustrate the accuracy and stability of the methods. A comparison with the exact solution for the integer-order case confirms that the approximate solutions are in excellent agreement with the exact results. Further comparisons with other methods indicate that the proposed approaches provide improved accuracy while reducing computational effort.

*Corresponding author

Received 13 September 2025; revised 8 February 2026; accepted 23 February 2026

Monowar Hossain
Department of Mathematics, University of Chittagong, Chattogram 4331, Bangladesh. e-mail: monowarcu28@gmail.com

Mohammed Aman Ullah
Department of Mathematics, University of Chittagong, Chattogram 4331, Bangladesh. e-mail: aman@cu.ac.bd

How to cite this article
Hossain, M. and Aman Ullah, M., Comparative semi-analytical solutions of Caputo time-fractional Fokker–Planck equations via Laplace transform-based decomposition and iterative methods . *Iran. J. Numer. Anal. Optim.*, 2026; 16(3): 815-851. https://doi.org/10.22067/ijnao.2026.95409.1728

All tabular and graphical results are generated using Python. Additionally, the stability and convergence of the proposed methods for TFFPEs are analyzed.

AMS subject classifications (2020): 35R11, 26A33, 35Q84, 65M12

Keywords: Caputo fractional derivative, time fractional Fokker–Planck equation, Laplace transform, Mittag-Leffler function, Convergence analysis, Error analysis

1 Introduction

Fractional calculus is a branch of mathematics that deals with derivatives and integrals of arbitrary orders and plays an important role in the mathematical modeling of real-world phenomena such as anomalous diffusion, viscoelastic materials, signal and image processing, electromagnetic field modeling, electrochemical processes, fluid flow problems, reaction-diffusion systems, water wave dynamics, population growth studies, traffic flow analysis, fractional stochastic models, and diverse topics in medical sciences [37, 45, 3].

The Fokker–Planck equation is a statistical physics equation developed by Adriaan Fokker and Max Planck to describe Brownian motion and is commonly used to explain solute transport [39]. The Fokker–Planck equation (FPE) appears in various fields in natural science, including solid-state physics, astrophysics, chemical and biological systems, polymer dynamics, stochastic processes, and circuit theory [39]. Due to the wide range of applications of the FPE, the aim of this study is to obtain analytical and numerical solutions of the linear and nonlinear fractional FPEs.

The linear time-fractional Fokker–Planck equation (TFFPE), also known as the fractional forward Kolmogorov equation [39] has the form for the variables $\Im$ and $\wp$:

$$\frac{\partial^{\zeta}}{\partial \wp^{\zeta}} u(\Im, \wp) = -\frac{\partial}{\partial \Im}\left[P(\Im, \wp) u(\Im, \wp)\right] + \frac{\partial^{2}}{\partial \Im^{2}}\left[R(\Im, \wp) u(\Im, \wp)\right] \tag{1}$$

with the initial condition given by

$$u(\Im, 0) = f(\Im), \quad \Im \in \mathbb{R}. \tag{2}$$

Here, $\frac{\partial^{\zeta}}{\partial \wp^{\zeta}} u(\Im, \wp)$ represents the Caputo fractional derivative of order $\zeta \in (0, 1)$ with respect to time. The function $u(\Im, \wp)$ is considered to be causal in both space and time. The terms $P(\Im, \wp)$ and $R(\Im, \wp) > 0$ denote the drift and diffusion coefficients, respectively, where both coefficients

are assumed to depend only on the spatial variable $\Im$. A related fractional partial differential equation is the fractional backward Kolmogorov equation [39], which can be expressed as follows:

$$\frac{\partial^{\zeta} u}{\partial \wp^{\zeta}} = -\left[P(\Im,\wp)\frac{\partial}{\partial \Im} + R(\Im,\wp)\frac{\partial^2}{\partial \Im^2}\right] u(\Im,\wp). \tag{3}$$

A generalization of (1) to n variables $(\Im_1, \Im_2, \ldots, \Im_n)$ yields the following:

$$\frac{\partial^{\zeta} u}{\partial \wp^{\zeta}} = -\sum_{i=1}^{n} \frac{\partial}{\partial \Im_i}\left[P_i(\Im,\wp)u(\Im,\wp)\right] + \sum_{i=1}^{n}\sum_{j=1}^{n} \frac{\partial^2}{\partial \Im_i \partial \Im_j}\left[R_{ij}(\Im,\wp)u(\Im,\wp)\right] \tag{4}$$

with the initial condition given by

$$u(\Im,0) = f(\Im), \quad \Im = (\Im_1, \Im_2, \ldots, \Im_n) \in \mathbb{R}^n. \tag{5}$$

The nonlinear TFFPE, which extends the traditional FPE, appears in numerous areas of science and engineering, such as nonlinear hydrodynamics, plasma physics, surface physics, polymer dynamics, laser physics, engineering applications, biophysics, neuroscience, population modeling, pattern formation, and selected problems in psychology and marketing [9]. The nonlinear TFFPE can be formulated as follows:

$$\frac{\partial^{\zeta} u}{\partial \wp^{\zeta}} = -\sum_{i=1}^{n} \frac{\partial}{\partial \Im_i}\left[P_i(\Im,\wp,u)u(\Im,\wp)\right] + \sum_{i=1}^{n}\sum_{j=1}^{n} \frac{\partial^2}{\partial \Im_i \partial \Im_j}\left[R_{ij}(\Im,\wp,u)u(\Im,\wp)\right] \tag{6}$$

with the initial condition given by

$$u(\Im,0) = f(\Im), \quad \Im = (\Im_1, \Im_2, \ldots, \Im_n) \in \mathbb{R}^n, \tag{7}$$

where ζ is the order of the fractional derivative $(0 < \zeta \leq 1)$ and n is integer $(n \neq 0)$.

Recently, researchers have developed many methods for constructing approximate analytical solutions of TFFPEs, such as the new iterative method (NIM) [28], He's homotopy perturbation method (HPM) [34], new transform iterative method (NITM) and homotopy perturbation Elzaki transform method (HPETM) [30], Yang transform decomposition method (YTDM) [2], homotopy perturbation Sumudu transform method (HPSTM) [6], fractional power series method (FPSM) [11], fractional iteration method (FIM) [31], shifted Chebyshev collocation method (CPFK) and finite difference method (FDM) [10], homotopy analysis transform method (HATM) [43], asymptotic homotopy perturbation method (AHPM) [19], and finite volume methods (FVM) [48].

Although many analytical and numerical methods have been successfully applied to solve TFFPEs, the Laplace-based hybrid methods, namely the Laplace Adomian decomposition method (LADM) and the Laplace variational iteration method (LVIM), offer distinct advantages. These semi-analytical approaches provide highly accurate approximations that closely match exact solutions even with only a few terms, converge rapidly, and require significantly less computational effort compared to traditional iterative or numerical methods. Additionally, they can handle both linear and nonlinear problems in a unified framework, making them versatile and efficient for a wide range of fractional-order problems. Motivated by the need for fast, reliable, and accurate solutions of TFFPEs with complex drift and diffusion terms, this paper applies and compares LADM and LVIM, provides convergence and stability analysis, and validates the methods with numerical examples, and demonstrates their effectiveness for complex fractional systems.

In 1980, George Adomian introduced the Adomian decomposition method (ADM), which is a well-known methodology to solve linear and nonlinear differential equations of classical as well as fractional order that gives accurate solutions in a concurrent series form. However, ADM does not properly decompose the fractional order parts. So, it was modified with many integral transformations, such as Elzaki, Mohand, Adoodh, Sumudu, Natural, Kashuri-Fundo, Yang, and Shehu [44, 7, 22]. ADM also combined with the Laplace transform leads to a powerful method known as the LADM, which was introduced by Khuri [26]. In this method, we convert a differential equation to an algebraic equation, and the nonlinear terms are decomposed in terms of Adomian polynomials. LADM is more powerful than the standard ADM method [32].

Another well-known powerful method for solving differential equations is the variational iteration method (VIM), introduced by Ji-Huan He [15]. The combined form of the Laplace transform and the VIM is called the LVIM, introduced by Hammouch and Mekkaoui [12]. LVIM is more powerful than VIM. Furthermore, the results obtained using LVIM have been compared with LADM, showing that in some studies LVIM may be more accurate [33], but in our examples both methods produce identical results. Unlike LADM, there is no necessity for the distinct treatment of nonlinear terms.

Recent studies have shown the effectiveness of Laplace-based decomposition and iteration techniques in solving nonlinear and fractional PDEs. For example, LADM has been applied to the fractional Klein–Gordon [40], sine–Gordon [24], multi-dimensional Navier–Stokes [29], Chaffee-Infante equations [5], PHI-four equation [41], voltage of telegraph equation [23], Rotavirus model [1], SEIR epidemic model [8], smoking models [14], COVID-19 SDIQR epidemic models [42]. Likewise, LVIM has been applied to fractional diffusion and reaction-diffusion equations [13], heat-like equations [4], wave-like equations [21], Klein–Gordon and sine–Gordon [35], chemical kinetics and population dynamics [36], physical systems [25], microelectromechanical

systems [47], transport phenomena [18], nonlinear PDEs [17] and many more. These applications highlight the wide utility of transform-based decomposition methods and motivate the present comparative study of LADM and LVIM for TFFPEs.

The rest of this paper is organized as follows. Section 2 presents the necessary mathematical preliminaries. Sections 3 and 4 describe the LADM and LVIM algorithms to solve TFFPE. Section 5 presents the convergence and error analysis of the proposed methods. Section 6 provides numerical examples and comparisons. Section 7 discusses the numerical and graphical results, and Section 8 concludes the paper.

2 Mathematical preliminaries

This section presents fundamental concepts and operators essential for our subsequent analysis of fractional differential equations.

Definition 1 (Caputo Fractional Derivative)**.** For a function $f \in C^n[0,\infty)$ and real number $\zeta > 0$, the Caputo fractional derivative of order ζ is defined as follows [37]:

$$ {}^{C}D^{\zeta}_{\Im} f(\Im) = \begin{cases} \dfrac{1}{\Gamma(n-\zeta)} \displaystyle\int_0^{\Im} (\Im - t)^{n-\zeta-1} f^{(n)}(t)\, dt, & n-1 < \zeta < n, \\ \dfrac{d^n}{d\Im^n} f(\Im), & \zeta = n, \end{cases} \tag{8} $$

where $n \in \mathbb{N}$ and $\Gamma(\cdot)$ represents the gamma function.

For power functions, the Caputo derivative exhibits the following behavior:

$$ {}^{C}D^{\zeta}_{\wp}\, \wp^{\beta} = \begin{cases} \dfrac{\Gamma(\beta+1)}{\Gamma(\beta-\zeta+1)} \wp^{\beta-\zeta}, & \beta > n-1, \\ 0, & \beta \le n-1, \end{cases} \tag{9} $$

where $n-1 < \zeta \le n$ and $\beta \in \mathbb{R}$.

Definition 2 (Laplace transform)**.** The Laplace transform of a function $f(\wp)$ is given by [37, 45]

$$ F(s) = \mathcal{L}\{f(\wp)\} = \int_0^{\infty} f(\wp)\, e^{-s\wp}\, d\wp, \tag{10} $$

where $F(s)$ denotes the Laplace transform of $f(\wp)$.

Definition 3 (Laplace transform of fractional derivatives)**.** For $m-1 < \zeta \le m$, where $m \in \mathbb{N}$, the Laplace transform of the Caputo fractional derivative satisfies [37]

$$\mathcal{L}\left\{{}^{C}D^{\zeta}f(\wp)\right\} = s^{\zeta}F(s) - \sum_{k=0}^{m-1} s^{\zeta-k-1}f^{(k)}(0), \tag{11}$$

where $F(s)$ denotes the Laplace transform of $f(\wp)$.

Definition 4 (Mittag-Leffler function)**.** The one-parameter Mittag-Leffler function $E_{\zeta}(\wp)$ for $\zeta > 0$ is defined through the series expansion [37]:

$$E_{\zeta}(\wp) = \sum_{k=0}^{\infty} \frac{\wp^{k}}{\Gamma(\zeta k + 1)}, \tag{12}$$

where $\Gamma(\cdot)$ denotes the Gamma function.

3 Laplace Adomian decomposition method

The LADM (Laplace Adomian decomposition method) has been successfully applied to a variety of fractional derivatives, including Riemann–Liouville, Caputo, Caputo–Fabrizio–Riemann, Caputo–Fabrizio–Caputo, Atangana–Baleanu–Riemann, and Atangana–Baleanu–Caputo types [20]. In this section, we focus on the implementation of the LADM algorithm for solving the nonlinear TFFPE.

Applying the Laplace transform to both sides of (6) for one variable case, we get

$$\mathcal{L}\left(\frac{\partial^{\zeta}u}{\partial\wp^{\zeta}}\right) = \mathcal{L}\left(-\frac{\partial}{\partial\Im}[P(\Im,\wp,u)u(\Im,\wp)] + \frac{\partial^{2}}{\partial\Im^{2}}[R(\Im,\wp,u)u(\Im,\wp)]\right). \tag{13}$$

Applying the Laplace transform property for fractional derivatives yields

$$s^{\zeta}\bar{u}(\Im,s) - s^{\zeta-1}u(\Im,0) = \mathcal{L}\left(-\frac{\partial}{\partial\Im}[P(\Im,\wp,u)u] + \frac{\partial^{2}}{\partial\Im^{2}}[R(\Im,\wp,u)u]\right). \tag{14}$$

Using the initial condition (7), we obtain

$$\bar{u}(\Im,s) = \frac{f(\Im)}{s} + \frac{1}{s^{\zeta}}\mathcal{L}\left(-\frac{\partial}{\partial\Im}[P(\Im,\wp,u)u] + \frac{\partial^{2}}{\partial\Im^{2}}[R(\Im,\wp,u)u]\right). \tag{15}$$

The standard Laplace decomposition method defines the solution $u(\Im,\wp)$ as an infinite series:

$$u(\Im,\wp) = \sum_{n=0}^{\infty} u_{n}(\Im,\wp). \tag{16}$$

The nonlinear terms $M(u) = P(\Im, \wp, u)u$ and $N(u) = R(\Im, \wp, u)u$ are decomposed using Adomian polynomials:

$$M(u) = \sum_{n=0}^{\infty} M_n, \quad N(u) = \sum_{n=0}^{\infty} N_n, \tag{17}$$

where the Adomian polynomials are computed as

$$M_n = \frac{1}{n!} \frac{d^n}{d\lambda^n} M\left(\sum_{i=0}^{\infty} \lambda^i u_i\right)\Big|_{\lambda=0}, \tag{18}$$

$$N_n = \frac{1}{n!} \frac{d^n}{d\lambda^n} N\left(\sum_{i=0}^{\infty} \lambda^i u_i\right)\Big|_{\lambda=0}. \tag{19}$$

Substituting these into (15) gives

$$\sum_{n=0}^{\infty} u_n(\Im, s) = \frac{f(\Im)}{s} + \frac{1}{s^\zeta}\mathcal{L}\left(-\frac{\partial}{\partial \Im}\sum_{n=0}^{\infty} M_n + \frac{\partial^2}{\partial \Im^2}\sum_{n=0}^{\infty} N_n\right). \tag{20}$$

The recursive solution scheme becomes

$$\begin{aligned} \bar{u}_0(\Im, s) &= \frac{f(\Im)}{s}, \\ \bar{u}_{n+1}(\Im, s) &= \frac{1}{s^\zeta}\mathcal{L}\left(-\frac{\partial}{\partial \Im} M_n + \frac{\partial^2}{\partial \Im^2} N_n\right), \quad n \geq 0. \end{aligned} \tag{21}$$

Applying the inverse Laplace transform, we obtain

$$u_0(\Im, \wp) = f(\Im), \tag{22}$$

$$u_{n+1}(\Im, \wp) = \mathcal{L}^{-1}\left[\frac{1}{s^\zeta}\mathcal{L}\left(-\frac{\partial}{\partial \Im} M_n + \frac{\partial^2}{\partial \Im^2} N_n\right)\right], \quad n \geq 0. \tag{23}$$

The approximate solution is obtained by truncating the series,

$$u(\Im, \wp) \approx \sum_{n=0}^{k} u_n(\Im, \wp). \tag{24}$$

4 Laplace variational iteration method

Applying the Laplace transform to both sides of (6) and using the initial condition (7), we get

$$\bar{u}(\Im, s) = \frac{f(\Im)}{s} + \frac{1}{s^\zeta}\mathcal{L}\left(-\frac{\partial}{\partial \Im}[P(\Im, \wp, u)u] + \frac{\partial^2}{\partial \Im^2}[R(\Im, \wp, u)u]\right). \tag{25}$$

Applying the inverse Laplace transform yields

$$u(\Im,\wp)=f(\Im)+\mathcal{L}^{-1}\left[\frac{1}{s^{\zeta}}\mathcal{L}\left(-\frac{\partial}{\partial\Im}[P(\Im,\wp,u)u]+\frac{\partial^2}{\partial\Im^2}[R(\Im,\wp,u)u]\right)\right]. \tag{26}$$

Differentiating with respect to $\wp$ gives

$$\frac{\partial u(\Im,\wp)}{\partial\wp}=\frac{\partial}{\partial\wp}\mathcal{L}^{-1}\left[\frac{1}{s^{\zeta}}\mathcal{L}\left(-\frac{\partial}{\partial\Im}[P(\Im,\wp,u)u]+\frac{\partial^2}{\partial\Im^2}[R(\Im,\wp,u)u]\right)\right]. \tag{27}$$

The correction functional is constructed as follows:

$$\begin{aligned}u_{n+1}(\Im,\wp)=u_n(\Im,\wp)-\int_0^{\wp}\lambda\Bigg[&\frac{\partial u_n(\Im,\varepsilon)}{\partial\varepsilon}\\&-\frac{\partial}{\partial\varepsilon}\mathcal{L}^{-1}\left(\frac{1}{s^{\zeta}}\mathcal{L}\Big(-\frac{\partial}{\partial\Im}[P(\Im,\varepsilon,u_n)u_n]+\frac{\partial^2}{\partial\Im^2}[R(\Im,\varepsilon,u_n)u_n]\Big)\right)\Bigg]d\varepsilon.\end{aligned} \tag{28}$$

By the variational theory, the Lagrange multiplier λ for (28) can be obtained as follows:

$$1+\lambda|_{\varepsilon=\wp}=0\quad\Rightarrow\quad\lambda=-1.$$

Substituting $\lambda=-1$ into (28) yields the iterative scheme:

$$\begin{aligned}u_{n+1}(\Im,\wp)=u_n(\Im,\wp)+\int_0^{\wp}\Bigg[&\frac{\partial u_n(\Im,\varepsilon)}{\partial\varepsilon}-\frac{\partial}{\partial\varepsilon}\mathcal{L}^{-1}\left(\frac{1}{s^{\zeta}}\mathcal{L}\Big(-\frac{\partial}{\partial\Im}[P(\Im,\varepsilon,u_n)u_n]\right.\\&\left.+\frac{\partial^2}{\partial\Im^2}[R(\Im,\varepsilon,u_n)u_n]\Big)\right)\Bigg]d\varepsilon.\end{aligned} \tag{29}$$

The exact solution is given by

$$u(\Im,\wp)=\lim_{n\to\infty}u_n(\Im,\wp). \tag{30}$$

5 Convergence and error analysis

Definition 5. For a self map $\mathcal{T}:X\to X$ on a set X, a point $\Im$ is called a fixed point of $\mathcal{T}$ if

$$\mathcal{T}(\Im)=\Im. \tag{31}$$

Definition 6. Consider a metric space (Ξ, d). A self-map $\mathcal{T} : X \to X$ is said to be a contraction if there exists a constant γ with $0 \leq \gamma < 1$ such that, for every pair $\Im, \wp \in X$ [27],

$$d\big(\mathcal{T}(\Im), \mathcal{T}(\wp)\big) \leq \gamma\, d(\Im, \wp). \tag{32}$$

Theorem 1. Let (X, d) be a complete non-empty metric space. If $\mathcal{T} : X \to X$ is a contraction mapping, then $\mathcal{T}$ has specifically one fixed point.

Proof. See proof in [27]. □

Theorem 2. Consider a Banach space (X, d), and let $\mathcal{T} : X \to X$ be a self-mapping. Assume that there exist constants $\delta \geq 0$ and $0 \leq \mu < 1$ such that, for all $x, y \in X$,

$$d\big(\mathcal{T}(x), \mathcal{T}(y)\big) \leq \delta\, d(x, \mathcal{T}(x)) + \mu\, d(x, y). \tag{33}$$

If $\mathcal{T}$ has a fixed point q, then $\mathcal{T}$ is said to be Picard-$\mathcal{T}$ stable.

Proof. See proof in [38]. □

5.1 Laplace Adomian decomposition method

Theorem 3 (Picard-$\mathcal{T}$ stability of LADM for nonlinear TFFPEs)**.** Consider a Banach space $(X, \|\cdot\|)$ and a self-map $\mathcal{T} : X \to X$. Then, the iterative procedure (23) of the LADM is defined by

$$\mathcal{T}(u_n(\Im, \wp)) = u_{n+1}(\Im, \wp) = u(\Im, 0) + \mathcal{L}^{-1}\left[\frac{1}{s^{\zeta}}\mathcal{L}\left(A_n(u)\right)\right], \tag{34}$$

where $A_n = -\frac{\partial}{\partial \Im} M_n + \frac{\partial^2}{\partial \Im^2} N_n$ is the Picard $\mathcal{T}$-stable if

1. $\|A_m(u) - A_n(u)\| \leq \mu_1 \|u_m(\Im, \wp) - u_n(\Im, \wp)\|$ for some $\mu_1 \in \mathbb{R}^+$,
2. $\mu = \mu_1 \left\| \frac{\wp^{\zeta}}{\Gamma(\zeta+1)} \right\| < 1$.

Proof. First, we verify that the mapping $\mathcal{T}$ possesses a fixed point. Considering $n, m \in \mathbb{N}$, we obtain

$$\mathcal{T}(u_n(\Im, \wp)) = u(\Im, 0) + \mathcal{L}^{-1}\left[\frac{1}{s^{\zeta}}\mathcal{L}\left(A_n(u)\right)\right], \tag{35}$$

$$\mathcal{T}(u_m(\Im,\wp)) = u(\Im,0) + \mathcal{L}^{-1}\left[\frac{1}{s^\zeta}\mathcal{L}\left(A_m(u)\right)\right]. \tag{36}$$

Subtracting (36) from (35), we obtain

$$\mathcal{T}(u_n(\Im,\wp)) - \mathcal{T}(u_m(\Im,\wp)) = \mathcal{L}^{-1}\left[\frac{1}{s^\zeta}\mathcal{L}\left(A_n(u) - A_m(u)\right)\right]. \tag{37}$$

Taking the norm on the both sides of (37) gives

$$\|\mathcal{T}(u_n(\Im,\wp)) - \mathcal{T}(u_m(\Im,\wp))\| = \left\|\mathcal{L}^{-1}\left[\frac{1}{s^\zeta}\mathcal{L}\left(A_n(u) - A_m(u)\right)\right]\right\|.$$

Using given assumption and the properties of the Laplace transform, we have

$$\begin{aligned}\|\mathcal{T}(u_n(\Im,\wp)) - \mathcal{T}(u_m(\Im,\wp))\| &\le \mu_1\|u_m(\Im,\wp) - u_n(\Im,\wp)\|\left\|\mathcal{L}^{-1}\left[\frac{1}{s^\zeta}\mathcal{L}\left[1\right]\right]\right\| \\ &= \mu_1\|u_n(\Im,\wp) - u_m(\Im,\wp)\|\left\|\frac{\wp^\zeta}{\Gamma(\zeta+1)}\right\| \\ &= \mu\|u_n(\Im,\wp) - u_m(\Im,\wp)\|,\end{aligned}$$

where $\mu = \mu_1\left\|\frac{\wp^\zeta}{\Gamma(\zeta+1)}\right\|$.

Thus, the self-map $\mathcal{T}$ has a fixed point and

$$\|\mathcal{T}(u_n(\Im,\wp)) - \mathcal{T}(u_m(\Im,\wp))\| \le \delta\|u_n(\Im,\wp) - u_m(\Im,\wp)\| + \mu\|u_n(\Im,\wp) - u_m(\Im,\wp)\|. \tag{38}$$

For $\delta = 0$, $\mu = \mu_1\left\|\frac{\wp^\zeta}{\Gamma(\zeta+1)}\right\| < 1$, the self-map $\mathcal{T}$ is a contraction mapping. Therefore, by the Banach fixed-point Theorem 2, the iterative scheme is Picard-$\mathcal{T}$ stable.

□

Theorem 4 (Convergence of LADM)**.** Let $(X, \|\cdot\|)$ be a Banach space and let $\mathcal{T} : X \to X$ be the operator associated with the LADM iteration scheme (34). Then there exists a unique fixed point, and the sequence $\{\psi_n(\Im,\wp)\}_{n=0}^{\infty}$ generated by LADM with an initial value $u_0 \in X$ converges to this fixed point whenever $0 < \eta < 1$, $\|u_0\| < \infty$, and

$$\|u_{n+1}(\Im,\wp)\| \le \eta\|u_n(\Im,\wp)\|, \quad n \ge 0.$$

Proof. Take $u_0 \in X$, and define the sequence of partial sums

$$\psi_n = \sum_{k=0}^{n} u_k, \qquad n \ge 0.$$

From the assumption

$$\|\psi_{n+1} - \psi_n\| = \|u_{n+1}\| \leq \eta \|u_n\| \leq \eta^{n+1} \|u_0\|.$$

For $m > n$, applying the triangle inequality yields

$$\|\psi_m - \psi_n\| \leq \sum_{k=n+1}^{m} \|u_k\| \leq \sum_{k=n+1}^{m} \eta^k \|u_0\| \leq \frac{\eta^{n+1}}{1-\eta} \|u_0\|.$$

Since $0 < \eta < 1$, the right-hand side tends to zero as $n \to \infty$, so $\{\psi_n\}$ is Cauchy in X and hence convergent. Denote the limit by

$$\phi^* = \lim_{n \to \infty} \psi_n.$$

To establish uniqueness, suppose ψ is another fixed point. Then

$$\|\psi^* - \phi\| = \|\mathcal{T}\phi^* - \mathcal{T}\psi\| \leq \eta \|\phi^* - \psi\|.$$

As $0 < \eta < 1$, this forces $\phi^* = \psi$. Therefore, ψ^* is the unique fixed point of $\mathcal{T}$. □

Theorem 5 (Error bound LADM)**.** Assume that $\sum_{k=0}^{n} u_k(\Im, \wp)$ is finite and that $u(\Im, \wp)$ is the corresponding series solution. Considering $\|u_{k+1}(\Im, \wp)\| \leq \eta \|u_k(\Im, \wp)\|$ for some $0 < \eta < 1$, then the maximum absolute error is given by

$$\|u(\Im, \wp) - \sum_{k=0}^{n} u_k(\Im, \wp)\| < \frac{\eta^{n+1}}{1-\eta} \|u_0(\Im, \wp)\|. \tag{39}$$

Proof. Let $\sum_{k=0}^{n} u_k$ be finite; then $\sum_{k=0}^{\infty} u_k < \infty$. We have

$$\begin{aligned} \|u - \sum_{k=0}^{n} u_k\| = \|\sum_{k=n+1}^{\infty} u_k\| &\leq \sum_{k=n+1}^{\infty} \|u_k\| \leq \sum_{k=n+1}^{\infty} \eta^k \|u_0\| \\ &= \eta^{n+1} \|u_0\| \sum_{k=0}^{\infty} \eta^k = \frac{\eta^{n+1}}{1-\eta} \|u_0\|. \end{aligned}$$

□

5.2 Laplace variational iteration method

From the iterative formula (29), we can rewrite it as

$$u_{n+1}(\Im, \wp) = u_n(\Im, \wp) - \mathcal{T}[u_n], \tag{40}$$

where the LVIM operator $\mathcal{T}[u_n]$ is defined as follows:

$$\mathcal{T}[u] = \int_0^{\wp} \left[\frac{\partial u(\Im, \varepsilon)}{\partial \varepsilon} - \frac{\partial}{\partial \varepsilon} \mathcal{L}^{-1} \left(\frac{1}{s^{\zeta}} \mathcal{L} \Big(- \frac{\partial}{\partial \Im} [P(\Im, \varepsilon, u)u] + \frac{\partial^2}{\partial \Im^2} [R(\Im, \varepsilon, u)u] \Big) \right) \right] d\varepsilon. \tag{41}$$

Then the LVIM iteration can be written as

$$u_{n+1}(\Im, \wp) = u_n(\Im, \wp) - \mathcal{T}[u_n]. \tag{42}$$

Define the components ϑ_k, $k = 0, 1, 2, \ldots$, through the recursive relation,

$$\begin{cases} \vartheta_0 = u_0, \\ \vartheta_{k+1} = -\mathcal{T}[\vartheta_0 + \vartheta_1 + \cdots + \vartheta_k], \quad k = 0, 1, 2, \ldots. \end{cases} \tag{43}$$

Then the series solution is given by

$$u(\Im, \wp) = \lim_{n \to \infty} u_n(\Im, \wp) = \sum_{k=0}^{\infty} \vartheta_k(\Im, \wp). \tag{44}$$

Theorem 6 (Convergence of LVIM)**.** Let $\mathcal{T}$, defined in (41), be an operator from a Banach space X to X. The series solution $u(\Im, \wp) = \sum_{k=0}^{\infty} \vartheta_k(\Im, \wp)$, defined in (44), converges if there exists $0 < \eta < 1$ such that

$$\|\vartheta_{k+1}\| \leq \eta \|\vartheta_k\|$$

for all $k \in \mathbb{N} \cup \{0\}$.

Proof. From (43), we have $\vartheta_{k+1} = -\mathcal{T}[\sum_{i=0}^{k} \vartheta_i]$. The condition implies

$$\|\vartheta_{k+1}\| \leq \eta \|\vartheta_k\| \leq \eta^2 \|\vartheta_{k-1}\| \leq \cdots \leq \eta^{k+1} \|\vartheta_0\|.$$

Define the sequence of partial sums $\{S_n\}_{n=0}^{\infty}$ as

$$S_n = \sum_{k=0}^{n} \vartheta_k, \quad n = 0, 1, 2, \ldots. \tag{45}$$

Consider the difference between consecutive partial sums:

$$\|S_{n+1} - S_n\| = \|\vartheta_{n+1}\| \leq \eta^{n+1}\|\vartheta_0\|. \tag{46}$$

For any $n, m \in \mathbb{N}$ with $n > m$, we have

$$\begin{aligned} \|S_n - S_m\| &= \left\| \sum_{k=m+1}^{n} (S_k - S_{k-1}) \right\| \leq \sum_{k=m+1}^{n} \|S_k - S_{k-1}\| \\ &\leq \sum_{k=m+1}^{n} \eta^k \|\vartheta_0\| = \eta^{m+1} \left(\frac{1-\eta^{n-m}}{1-\eta} \right) \|\vartheta_0\|. \end{aligned} \tag{47}$$

Since $0 < \eta < 1$, we obtain

$$\lim_{n,m\to\infty} \|S_n - S_m\| = 0. \tag{48}$$

Hence, $\{S_n\}_{n=0}^{\infty}$ is a Cauchy sequence in the Banach space X, which implies that the series solution $u(\Im, \wp) = \sum_{k=0}^{\infty} \vartheta_k(\Im, \wp)$ converges. □

Theorem 7 (Error estimate for LVIM)**.** If the series solution $\sum_{k=0}^{\infty} \vartheta_k(\Im, \wp)$ converges to $u(\Im, \wp)$, and the truncated series $\sum_{k=0}^{j} \vartheta_k(\Im, \wp)$ is used as an approximation, then the maximum absolute error $\mathcal{E}_j(\Im, \wp)$ is estimated by

$$\mathcal{E}_j(\Im, \wp) = \left\| u(\Im, \wp) - \sum_{k=0}^{j} \vartheta_k(\Im, \wp) \right\| \leq \frac{\eta^{j+1}}{1-\eta} \|\vartheta_0\|, \tag{49}$$

where $0 < \eta < 1$ and $\|\vartheta_{k+1}\| \leq \eta \|\vartheta_k\|$ for all $k \geq 0$.

Proof. From the convergence proof, for $n > j$, we have

$$\|S_n - S_j\| \leq \eta^{j+1} \left(\frac{1-\eta^{n-j}}{1-\eta} \right) \|\vartheta_0\|. \tag{50}$$

Taking the limit as $n \to \infty$, since $S_n \to u(\Im, \wp)$, we have

$$\|u(\Im, \wp) - S_j\| \leq \lim_{n\to\infty} \|S_n - S_j\| \leq \frac{\eta^{j+1}}{1-\eta} \|\vartheta_0\|. \tag{51}$$

Thus,

$$\mathcal{E}_j(\Im, \wp) = \left\| u(\Im, \wp) - \sum_{k=0}^{j} \vartheta_k(\Im, \wp) \right\| \leq \frac{\eta^{j+1}}{1-\eta} \|\vartheta_0\|. \tag{52}$$

□

6 Mathematical examples

In this section, we present the analytical solutions of several linear and nonlinear space-time fractional FPEs and display the solutions both numerically and graphically. Furthermore, we check the non-fractional case ($\zeta = 1$) to ensure consistency with the exact solution.

Example 1. Consider a linear TFFPE (1) with $P(\Im, \wp) = \Im, \quad R(\Im, \wp) = \frac{\Im^2}{2}$:

$$\frac{\partial^\zeta u}{\partial \wp^\zeta} = \left[-\frac{\partial}{\partial \Im}\Im + \frac{\partial^2}{\partial \Im^2}\frac{\Im^2}{2} \right] u(\Im, \wp), \tag{53}$$

subject to the initial condition

$$u(\Im, 0) = f(\Im) = \Im, \quad \Im \in \mathbb{R}. \tag{54}$$

(i) LADM solution

Applying the Laplace transform to both sides of (53), we obtain

$$s^\zeta \bar{u}(\Im, s) - s^{\zeta-1} u(\Im, 0) = \mathcal{L}\left[-\frac{\partial}{\partial \Im}(\Im u) + \frac{\partial^2}{\partial \Im^2}\left(\frac{\Im^2}{2} u\right) \right].$$

Using the initial condition (54), this becomes

$$\bar{u}(\Im, s) = \frac{\Im}{s} + \frac{1}{s^\zeta}\mathcal{L}\left[-\frac{\partial}{\partial \Im}(\Im u) + \frac{\partial^2}{\partial \Im^2}\left(\frac{\Im^2}{2} u\right) \right].$$

The recurrence relation is, therefore

$$\begin{aligned} \bar{u}_0(\Im, s) &= \frac{\Im}{s}, \\ \bar{u}_{n+1}(\Im, s) &= \frac{1}{s^\zeta}\mathcal{L}\left[-\frac{\partial}{\partial \Im}(\Im u_n) + \frac{\partial^2}{\partial \Im^2}\left(\tfrac{\Im^2}{2} u_n\right) \right], \quad n \geq 0. \end{aligned} \tag{55}$$

Applying the inverse Laplace transform, we have

$$u_0(\Im, \wp) = \mathcal{L}^{-1}\left(\frac{\Im}{s}\right) = \Im.$$

For $n = 0$ in (55):

$$u_1(\Im, \wp) = \mathcal{L}^{-1}\left[\frac{1}{s^\zeta}\mathcal{L}\left(-\frac{\partial}{\partial \Im}\Im^2 + \frac{\partial^2}{\partial \Im^2}\frac{\Im^3}{2} \right) \right]$$

$$= \mathcal{L}^{-1}\left(\frac{\Im}{s^{\zeta+1}}\right) = \frac{\Im\,\wp^{\zeta}}{\Gamma(\zeta+1)}.$$

For $n = 1$:

$$\begin{aligned} u_2(\Im,\wp) &= \mathcal{L}^{-1}\left[\frac{1}{s^{\zeta}}\,\mathcal{L}\left(-\frac{\partial}{\partial\Im}\left(\frac{\Im^2\wp^{\zeta}}{\Gamma(\zeta+1)}\right)+\frac{\partial^2}{\partial\Im^2}\left(\frac{\Im^3\wp^{\zeta}}{2\Gamma(\zeta+1)}\right)\right)\right] \\ &= \mathcal{L}^{-1}\left[\frac{1}{s^{\zeta}}\,\mathcal{L}\left(\frac{\Im\wp^{\zeta}}{\Gamma(\zeta+1)}\right)\right] = \mathcal{L}^{-1}\left(\frac{\Im}{s^{2\zeta+1}}\right) = \frac{\Im\,\wp^{2\zeta}}{\Gamma(2\zeta+1)}. \end{aligned}$$

Similarly,

$$u_3(\Im,\wp) = \frac{\Im\,\wp^{3\zeta}}{\Gamma(3\zeta+1)}.$$

Thus, the LADM approximate solution of (53) is

$$u(\Im,\wp) = \Im\left(1+\frac{\wp^{\zeta}}{\Gamma(\zeta+1)}+\frac{\wp^{2\zeta}}{\Gamma(2\zeta+1)}+\frac{\wp^{3\zeta}}{\Gamma(3\zeta+1)}+\ldots\right) = \Im E_{\zeta}(\wp^{\zeta}). \tag{56}$$

(ii) LVIM solution

Applying the Laplace transform to both sides of (53) with initial condition (54), we obtain

$$\bar{u}(\Im,s) = \frac{\Im}{s} + \frac{\mathcal{L}}{s^{\zeta}}\left[-\frac{\partial}{\partial\Im}(\Im u)+\frac{\partial^2}{\partial\Im^2}\left(\frac{\Im^2}{2}u\right)\right]. \tag{57}$$

Applying the inverse Laplace transform yields

$$u(\Im,\wp) = \Im + \mathcal{L}^{-1}\left[\frac{1}{s^{\zeta}}\mathcal{L}\left[-\frac{\partial}{\partial\Im}(\Im u)+\frac{\partial^2}{\partial\Im^2}\left(\frac{\Im^2}{2}u\right)\right]\right]. \tag{58}$$

Differentiating (58) with respect to $\wp$ gives

$$\frac{\partial u(\Im,\wp)}{\partial\wp} = \frac{\partial}{\partial\wp}\mathcal{L}^{-1}\left[\frac{\mathcal{L}}{s^{\zeta}}\left[\left(-\frac{\partial}{\partial\Im}\Im+\frac{\partial^2}{\partial\Im^2}\frac{\Im^2}{2}\right)u\right]\right]. \tag{59}$$

Now, from (29), the iterative formula is

$$\begin{aligned} u_{n+1}(\Im,\wp) &= u_n(\Im,\wp) \\ &\quad - \int_0^{\wp}\left[\frac{\partial u_n(\Im,\varepsilon)}{\partial\varepsilon} - \frac{\partial}{\partial\varepsilon}\mathcal{L}^{-1}\left(\frac{\mathcal{L}}{s^{\zeta}}\left[-\frac{\partial}{\partial\Im}(\Im u_n)+\frac{\partial^2}{\partial\Im^2}\left(\frac{\Im^2}{2}u_n\right)\right]\right)\right]d\varepsilon. \end{aligned} \tag{60}$$

The initial iteration:

$$u_0(\Im, \wp) = \Im.$$

For $n = 0$:

$$\begin{aligned} u_1(\Im, \wp) &= \Im - \int_0^{\wp} \left[\frac{\partial \Im}{\partial \varepsilon} - \frac{\partial}{\partial \varepsilon} \mathcal{L}^{-1} \left[\frac{\mathcal{L}}{s^{\zeta}} \left[-\frac{\partial}{\partial \Im} \Im^2 + \frac{\partial^2}{\partial \Im^2} \frac{\Im^3}{2} \right] \right] \right] d\varepsilon \\ &= \Im + \int_0^{\wp} \left[\frac{\Im\, \zeta\, \varepsilon^{\zeta - 1}}{\Gamma(\zeta + 1)} \right] d\varepsilon = \Im + \frac{\Im \wp^{\zeta}}{\Gamma(\zeta + 1)}. \end{aligned}$$

For $n = 1$:

$$\begin{aligned} u_2(\Im, \wp) &= \Im + \frac{\Im \wp^{\zeta}}{\Gamma(\zeta + 1)} - \int_0^{\wp} \left[\frac{\partial}{\partial \varepsilon} \left(\Im + \frac{\Im \varepsilon^{\zeta}}{\Gamma(\zeta + 1)} \right) \right. \\ &\quad \left. - \frac{\partial}{\partial \varepsilon} \mathcal{L}^{-1} \left[\frac{\mathcal{L}}{s^{\zeta}} \left[-\frac{\partial}{\partial \Im} \left(\Im^2 + \frac{\Im^2 \varepsilon^{\zeta}}{\Gamma(\zeta + 1)} \right) + \frac{\partial^2}{\partial \Im^2} \left(\frac{\Im^3}{2} \left(1 + \frac{\varepsilon^{\zeta}}{\Gamma(\zeta + 1)} \right) \right) \right] \right] \right] d\varepsilon \\ &= \Im + \frac{\Im \wp^{\zeta}}{\Gamma(\zeta + 1)} + \int_0^{\wp} \left[\frac{2\Im\, \zeta\, \varepsilon^{2\zeta - 1}}{\Gamma(2\zeta + 1)} \right] d\varepsilon = \Im + \frac{\Im \wp^{\zeta}}{\Gamma(\zeta + 1)} + \frac{\Im \wp^{2\zeta}}{\Gamma(2\zeta + 1)}. \end{aligned}$$

For $n = 2$:

$$u_3(\Im, \wp) = \Im + \frac{\Im \wp^{\zeta}}{\Gamma(\zeta + 1)} + \frac{\Im \wp^{2\zeta}}{\Gamma(2\zeta + 1)} + \frac{\Im \wp^{3\zeta}}{\Gamma(3\zeta + 1)}.$$

The LVIM solution is

$$u(\Im, \wp) = \Im \left(1 + \frac{\wp^{\zeta}}{\Gamma(\zeta + 1)} + \frac{\wp^{2\zeta}}{\Gamma(2\zeta + 1)} + \frac{\wp^{3\zeta}}{\Gamma(3\zeta + 1)} + \cdots \right) = \Im E_{\zeta}(\wp^{\zeta}). \tag{61}$$

Here, $E_{\zeta}(\wp^{\zeta})$ is the Mittag-Leffler function. The LADM and LVIM solutions at $\zeta = 1$ converge to $u(\Im, \wp) = \Im e^{\wp}$, which is the exact solution for the nonfractional FPE [16]. Our LADM and LVIM solutions (56) and (61) match with the solutions obtained by NITM [30], HPETM [30], YTDM [2], and HPSTM [6].

Table 1: Comparative analysis of HPSTM [6], NITM [30], HPETM [30], YTDM [2], LADM, LVIM, and exact solution for Example 1.

$\wp$	$\Im$	HPSTM	NITM	HPETM	YTDM	LADM	LVIM	Exact
0.2	0.25	0.305333	0.305333	0.305333	0.305333	0.305333	0.305333	0.305351
0.2	0.50	0.610667	0.610667	0.610667	0.610667	0.610667	0.610667	0.610701
0.2	0.75	0.916000	0.916000	0.916000	0.916000	0.916000	0.916000	0.916052
0.2	1.00	1.221333	1.221333	1.221333	1.221333	1.221333	1.221333	1.221403
0.4	0.25	0.372667	0.372667	0.372667	0.372667	0.372667	0.372667	0.372956
0.4	0.50	0.745333	0.745333	0.745333	0.745333	0.745333	0.745333	0.745912
0.4	0.75	1.118000	1.118000	1.118000	1.118000	1.118000	1.118000	1.118869
0.4	1.00	1.490667	1.490667	1.490667	1.490667	1.490667	1.490667	1.491825
0.6	0.25	0.454000	0.454000	0.454000	0.454000	0.454000	0.454000	0.455530
0.6	0.50	0.908000	0.908000	0.908000	0.908000	0.908000	0.908000	0.911059
0.6	0.75	1.362000	1.362000	1.362000	1.362000	1.362000	1.362000	1.366589
0.6	1.00	1.816000	1.816000	1.816000	1.816000	1.816000	1.816000	1.822119

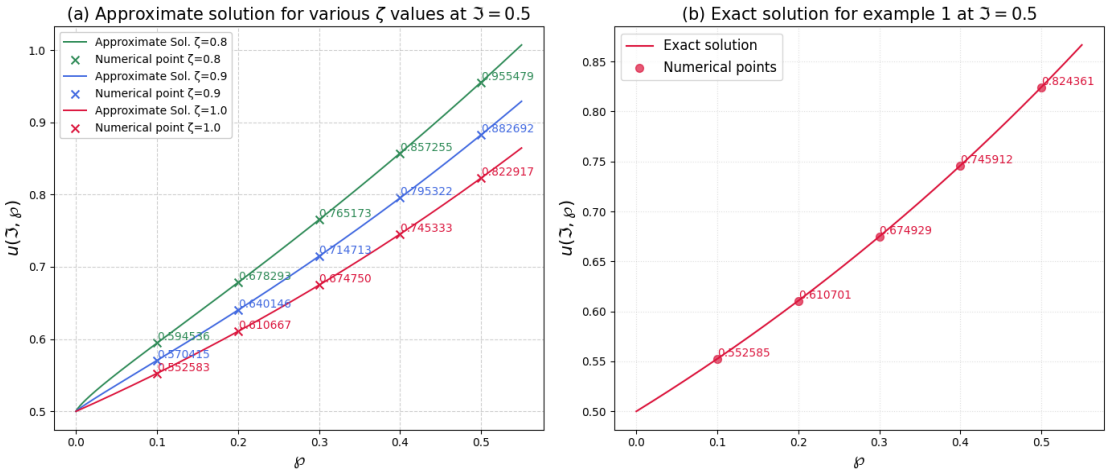


Figure 1: Two-dimensional graphical solutions of Example 1 at $\Im = 0.1$: (a) approximate solutions obtained by the LADM for $\zeta = 0.8, 0.9$ and 1.0, and (b) the corresponding exact solution.

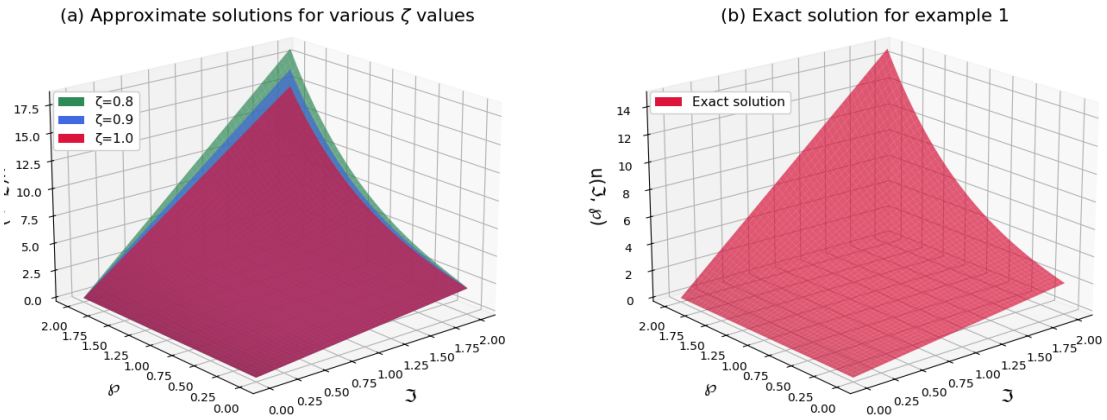


Figure 2: Three-dimensional graphical solutions of Example 1: (a) Approximate solutions for various values of ζ, and (b) Exact solution.

Example 2. Consider a backward Kolmogorov equation (3) with $P(\Im, \wp) = -(1 + \Im)$ and $R(\Im, \wp) = e^{\wp}\Im^2$:

$$\frac{\partial^{\zeta}}{\partial \wp^{\zeta}} u(\Im, \wp) = (1 + \Im)\frac{\partial}{\partial \Im} u(\Im, \wp) + e^{\wp}\Im^2 \frac{\partial^2}{\partial \Im^2} u(\Im, \wp), \tag{62}$$

subject to the initial condition:

$$u(\Im, 0) = 1 + \Im, \quad \Im \in \mathbb{R}. \tag{63}$$

(i) LADM solution

Applying the Laplace transform to both sides of (62) with initial condition (63), we obtain

$$\bar{u}(\Im, s) = \frac{1 + \Im}{s} + \frac{\mathcal{L}}{s^{\zeta}} \left[(1 + \Im)\frac{\partial u}{\partial \Im} + e^{\wp}\Im^2 \frac{\partial^2 u}{\partial \Im^2} \right]. \tag{64}$$

The recursive scheme is

$$\begin{aligned} \bar{u}_0(\Im, s) &= \frac{1 + \Im}{s} \\ \bar{u}_{n+1}(\Im, s) &= \frac{\mathcal{L}}{s^{\zeta}} \left[(1 + \Im)\frac{\partial}{\partial \Im} u_n(\Im, \wp) + e^{\wp}\Im^2 \frac{\partial^2}{\partial \Im^2} u_n(\Im, \wp) \right], \quad n \geq 0. \end{aligned} \tag{65}$$

Applying the inverse Laplace transform, we have

$$u_0(\Im, \wp) = \mathcal{L}^{-1}\left(\frac{1 + \Im}{s} \right) = 1 + \Im.$$

For $n = 0$:

$$u_1(\Im, \wp) = \mathcal{L}^{-1}\left[\frac{\mathcal{L}}{s^\zeta}\left[(1+\Im)\frac{\partial}{\partial \Im}(1+\Im) + e^{\wp}\Im^2\frac{\partial^2}{\partial \Im^2}(1+\Im)\right]\right]$$
$$= \mathcal{L}^{-1}\left[\frac{1+\Im}{s^{\zeta+1}}\right] = (1+\Im)\frac{\wp^\zeta}{\Gamma(\zeta+1)}.$$

For $n = 1$:

$$u_2(\Im, \wp) = \mathcal{L}^{-1}\left[\frac{\mathcal{L}}{s^\zeta}\left[(1+\Im)\frac{\partial}{\partial \Im}\left((1+\Im)\frac{\wp^\zeta}{\Gamma(\zeta+1)}\right) + e^{\wp}\Im^2\frac{\partial^2}{\partial \Im^2}\left((1+\Im)\frac{\wp^\zeta}{\Gamma(\zeta+1)}\right)\right]\right]$$
$$= \mathcal{L}^{-1}\left[\frac{\mathcal{L}}{s^\zeta}\left[\frac{(1+\Im)\,\wp^\zeta}{\Gamma(\zeta+1)}\right]\right] = \mathcal{L}^{-1}\left[\frac{1+\Im}{s^{2\zeta+1}}\right] = (1+\Im)\frac{\wp^{2\zeta}}{\Gamma(2\zeta+1)}.$$

For $n = 2$:

$$u_3(\Im, \wp) = (1+\Im)\frac{\wp^{3\zeta}}{\Gamma(3\zeta+1)}.$$

Thus, the LADM solution of (62) is given by

$$u(\Im, \wp) = (1+\Im)\left(1 + \frac{\wp^\zeta}{\Gamma(\zeta+1)} + \frac{\wp^{2\zeta}}{\Gamma(2\zeta+1)} + \frac{\wp^{3\zeta}}{\Gamma(3\zeta+1)} + \cdots\right) = (1+\Im)E_\zeta(\wp^\zeta). \tag{66}$$

(ii) LVIM solution

Applying the Laplace transform and the inverse Laplace transform to both sides of (62) along with the initial condition, we get

$$u(\Im, \wp) = (1+\Im) + \mathcal{L}^{-1}\left[\frac{\mathcal{L}}{s^\zeta}\left[(1+\Im)\frac{\partial u}{\partial \Im} + e^{\wp}\Im^2\frac{\partial^2 u}{\partial \Im^2}\right]\right]. \tag{67}$$

From (29) and (67), the iterative scheme becomes

$$u_{n+1}(\Im, \wp) = u_n(\Im, \wp) - \int_0^\wp \left[\frac{\partial u_n(\Im, \varepsilon)}{\partial \varepsilon} - \frac{\partial}{\partial \varepsilon}\mathcal{L}^{-1}\left(\frac{\mathcal{L}}{s^\zeta}\left[(1+\Im)\frac{\partial u_n(\Im, \varepsilon)}{\partial \Im} + e^{\wp}\Im^2\frac{\partial^2 u_n(\Im, \varepsilon)}{\partial \Im^2}\right]\right)\right] d\varepsilon. \tag{68}$$

The initial iteration is

$$u_0(\Im, \wp) = 1 + \Im.$$

From (68), for $n = 0$, we have

$$
\begin{aligned}
u_1(\Im,\wp) &= (1+\Im) \\
&\quad - \int_0^{\wp} \left[\frac{\partial(1+\Im)}{\partial\varepsilon} - \frac{\partial}{\partial\varepsilon}\mathcal{L}^{-1}\left[\frac{\mathcal{L}}{s^{\zeta}} \left((1+\Im)\frac{\partial(1+\Im)}{\partial\Im} + \left(e^{\wp}\Im^2\right) \frac{\partial^2(1+\Im)}{\partial\Im^2} \right) \right] \right] d\varepsilon \\
&= (1+\Im) + \int_0^{\wp} \left[\frac{(1+\Im)\,\zeta\,\varepsilon^{\zeta-1}}{\Gamma(\zeta+1)} \right] d\varepsilon = (1+\Im)\left(1 + \frac{\wp^{\zeta}}{\Gamma(\zeta+1)}\right).
\end{aligned}
$$

Similarly, from (68), for $n = 1$, we have

$$
\begin{aligned}
u_2(\Im,\wp) &= u_1(\Im,\wp) \\
&\quad - \int_0^{\wp} \left[\frac{\partial u_1(\Im,\varepsilon)}{\partial\varepsilon} - \frac{\partial}{\partial\varepsilon}\mathcal{L}^{-1}\left[\frac{1}{s^{\zeta}}\mathcal{L}\left((1+\Im)\frac{\partial u_1(\Im,\varepsilon)}{\partial\Im} + \left(e^{\wp}\Im^2\right) \frac{\partial^2 u_1(\Im,\varepsilon)}{\partial\Im^2} \right) \right] \right] d\varepsilon \\
&= (1+\Im)\left(1 + \frac{\wp^{\zeta}}{\Gamma(\zeta+1)}\right) + \int_0^{\wp} \left[(1+\Im)\cdot\frac{2\zeta\varepsilon^{2\zeta-1}}{\Gamma(2\zeta+1)} \right] d\varepsilon \\
&= (1+\Im)\left(1 + \frac{\wp^{\zeta}}{\Gamma(\zeta+1)} + \frac{\wp^{2\zeta}}{\Gamma(2\zeta+1)}\right).
\end{aligned}
$$

Similarly,

$$
u_3(\Im,\wp) = (1+\Im)\left(1 + \frac{\wp^{\zeta}}{\Gamma(\zeta+1)} + \frac{\wp^{2\zeta}}{\Gamma(2\zeta+1)} + \frac{\wp^{3\zeta}}{\Gamma(3\zeta+1)}\right).
$$

The LVIM solution is

$$
u(\Im,\wp) = (1+\Im)\left(1 + \frac{\wp^{\zeta}}{\Gamma(\zeta+1)} + \frac{\wp^{2\zeta}}{\Gamma(2\zeta+1)} + \frac{\wp^{3\zeta}}{\Gamma(3\zeta+1)} + \cdots\right) = (1+\Im)E_{\zeta}(\wp^{\zeta}). \tag{69}
$$

The LADM and LVIM solution at $\zeta = 1$ converge to $u(\Im,\wp) = (1+\Im)e^{\wp}$, which is the exact solution for the nonfractional FPE [16]. Our LADM and LVIM solutions (66) and (69) match with the solutions obtained by NIM [28], NITM [30], HPETM [30], YTDM [2], and FPSM [11].

Table 2: Comparative analysis of FIM, LADM, LVIM, and exact solution for Example 2, including absolute errors for $\zeta = 1$.

$\wp$	$\Im$	FIM [31]	LADM	LVIM	Exact	FIM Er.	LADM Er.	LVIM Er.
0.2	0.1	1.3434667	1.3434667	1.3434667	1.3435430	0.0000764	0.0000764	0.0000764
0.3	0.1	1.4844500	1.4844500	1.4844500	1.4848447	0.0003947	0.0003947	0.0003947
0.4	0.1	1.6397333	1.6397333	1.6397333	1.6410072	0.0012738	0.0012738	0.0012738
0.2	0.6	1.9541333	1.9541333	1.9541333	1.9542444	0.0001111	0.0001111	0.0001111
0.3	0.6	2.1592000	2.1592000	2.1592000	2.1597741	0.0005741	0.0005741	0.0005741
0.4	0.6	2.3850667	2.3850667	2.3850667	2.3869195	0.0018529	0.0018529	0.0018529
0.2	0.9	2.3205333	2.3205333	2.3205333	2.3206652	0.0001319	0.0001319	0.0001319
0.3	0.9	2.5640500	2.5640500	2.5640500	2.5647317	0.0006817	0.0006817	0.0006817
0.4	0.9	2.8322667	2.8322667	2.8322667	2.8344669	0.0022003	0.0022003	0.0022003

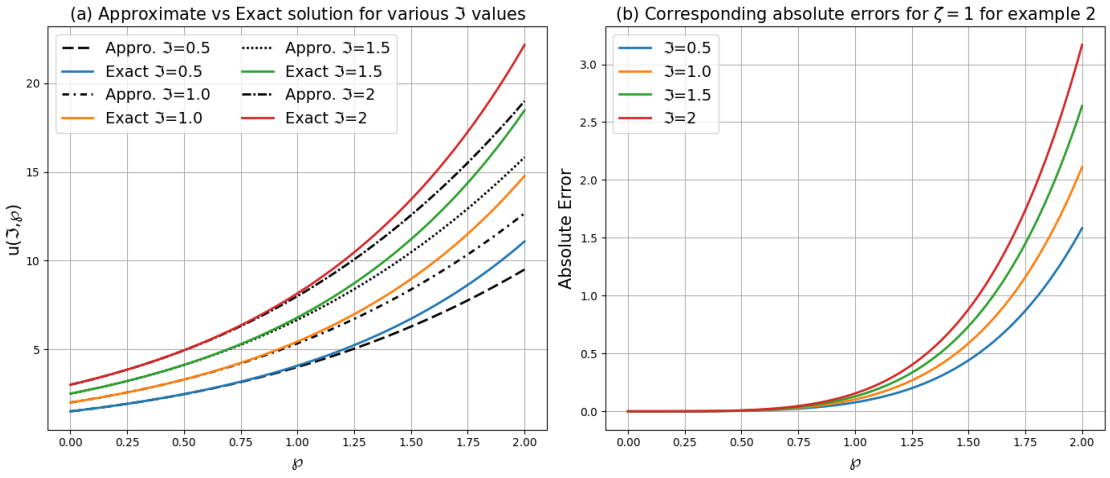


Figure 3: Two-dimensional graphical solutions of Example 2 for different values of $\Im$ when $\zeta = 1$ at : (a) Comparison of the approximate solutions with the exact solution, and (b) The corresponding absolute error .

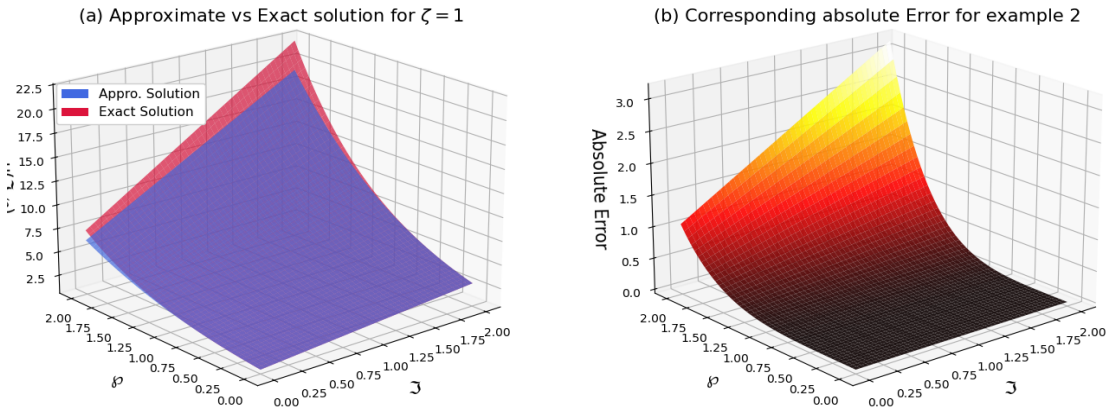


Figure 4: Three-dimensional graphical solutions of Example 2 for $\zeta = 1$: (a) Comparison of the approximate solutions with the exact solution, and (b) The corresponding absolute errors

Example 3. Consider a nonlinear TFFPE (6) with $P_1(\Im, \wp, u) = 4\frac{u}{\Im} - \frac{\Im}{3}$ and $R_{11}(\Im, \wp, u) = u$:

$$\frac{\partial^\zeta u}{\partial \wp^\zeta} = \left[-\frac{\partial}{\partial \Im}\left(4\frac{u^2}{\Im} - \frac{\Im u}{3}\right) + \frac{\partial^2}{\partial \Im^2}\left(u^2\right)\right], \tag{70}$$

subject to the initial condition:

$$u(\Im, 0) = \Im^2, \quad \Im \in \mathbb{R}. \tag{71}$$

(i) LADM solution

Applying the Laplace transform with initial condition, we have

$$\bar{u}(\Im, s) = \frac{\Im^2}{s} + \frac{1}{s^\zeta}\mathcal{L}\left[-\frac{\partial}{\partial \Im}\left(4\frac{u}{\Im} - \frac{\Im}{3}\right)u + \frac{\partial^2}{\partial \Im^2}(u^2)\right]. \tag{72}$$

Applying the inverse Laplace transform, we get

$$u(\Im, \wp) = \Im^2 + \mathcal{L}^{-1}\left[\frac{1}{s^\zeta}\mathcal{L}\left[-\frac{\partial}{\partial \Im}\left(4\frac{u}{\Im} - \frac{\Im}{3}\right)u + \frac{\partial^2}{\partial \Im^2}(u^2)\right]\right]. \tag{73}$$

Now, applying the standard Laplace decomposition method, we get the iterative formula:

$$\sum_{n=0}^{\infty} u_n(\Im, \wp) = \Im^2 + \mathcal{L}^{-1}\left[\frac{1}{s^\zeta}\mathcal{L}\left[-\frac{\partial}{\partial \Im}\left(\frac{4}{\Im}\sum_{n=0}^{\infty} N_n - \frac{\Im}{3}\sum_{n=0}^{\infty} u_n\right) + \frac{\partial^2}{\partial \Im^2}\sum_{n=0}^{\infty} N_n\right]\right], \tag{74}$$

where N_n are Adomian polynomials for u^2 and N_n are Adomian polynomials for $\frac{u^2}{\Im}$:

$$
\begin{aligned}
N_0(u) &= u_0^2,\\
N_1(u) &= 2u_0u_1,\\
N_2(u) &= u_1^2 + 2u_0u_2,\\
&\vdots
\end{aligned}
$$

The solution components are

$$
\begin{aligned}
u_0(\Im,\wp) &= \Im^2,\\
u_1(\Im,\wp) &= \mathcal{L}^{-1}\left[\frac{1}{s^\zeta}\mathcal{L}\left[-\frac{\partial}{\partial\Im}\left(4\frac{u_0^2}{\Im}-\frac{\Im u_0}{3}\right)+\frac{\partial^2}{\partial\Im^2}u_0^2\right]\right]\\
&= \mathcal{L}^{-1}\left[\frac{1}{s^\zeta}\mathcal{L}\left[-\frac{\partial}{\partial\Im}\left(4\Im^3-\frac{\Im^3}{3}\right)+\frac{\partial^2}{\partial\Im^2}\left(\Im^4\right)\right]\right]\\
&= \mathcal{L}^{-1}\left[\frac{1}{s^\zeta}\mathcal{L}\left[-\frac{\partial}{\partial\Im}\left(\frac{11\Im^3}{3}\right)+12\Im^2\right]\right]\\
&= \mathcal{L}^{-1}\left[\frac{1}{s^\zeta}\mathcal{L}\left[-11\Im^2+12\Im^2\right]\right]\\
&= \mathcal{L}^{-1}\left[\frac{1}{s^\zeta}\cdot\frac{\Im^2}{s}\right] = \frac{\Im^2\wp^\zeta}{\Gamma(\zeta+1)},\\
u_2(\Im,\wp) &= \mathcal{L}^{-1}\left[\frac{1}{s^\zeta}\mathcal{L}\left[-\frac{\partial}{\partial\Im}\left(4\frac{2u_0u_1}{\Im}-\frac{\Im u_1}{3}\right)+\frac{\partial^2}{\partial\Im^2}(2u_0u_1)\right]\right]\\
&= \mathcal{L}^{-1}\left[\frac{1}{s^\zeta}\mathcal{L}\left[-\frac{\partial}{\partial\Im}\left(8\Im u_1-\frac{\Im u_1}{3}\right)+\frac{\partial^2}{\partial\Im^2}(2\Im^2u_1)\right]\right]\\
&= \mathcal{L}^{-1}\left[\frac{1}{s^\zeta}\mathcal{L}\left[-\frac{\partial}{\partial\Im}\left(\frac{23}{3}\Im u_1\right)+4u_1+4\Im\frac{\partial u_1}{\partial\Im}+2\Im^2\frac{\partial^2u_1}{\partial\Im^2}\right]\right]\\
&= \mathcal{L}^{-1}\left[\frac{1}{s^\zeta}\mathcal{L}\left[\Im^2\left(\frac{\wp^\zeta}{\Gamma(\zeta+1)}\right)\right]\right] = \mathcal{L}^{-1}\left[\frac{\Im^2}{s^{2\zeta+1}}\right] = \frac{\Im^2\wp^{2\zeta}}{\Gamma(2\zeta+1)}\\
u_3(\Im,\wp) &= \frac{\Im^2\wp^{3\zeta}}{\Gamma(3\zeta+1)}.
\end{aligned}
$$

Thus, the LADM solution of (70) is given by

$$
u(\Im,\wp) = \Im^2\left(1+\frac{\wp^\zeta}{\Gamma(\zeta+1)}+\frac{\wp^{2\zeta}}{\Gamma(2\zeta+1)}+\frac{\wp^{3\zeta}}{\Gamma(3\zeta+1)}+\cdots\right) = \Im^2E_\zeta(\wp^\zeta). \tag{75}
$$

(ii) LVIM solution

From (29) and (73), the iterative scheme becomes

$$u_{n+1}(\Im,\wp) = u_n(\Im,\wp) - \int_0^{\wp}\left[\frac{\partial u_n(\Im,\varepsilon)}{\partial\varepsilon} - \frac{\partial}{\partial\varepsilon}\mathcal{L}^{-1}\left[\frac{\mathcal{L}}{s^{\zeta}}\left(-\frac{\partial}{\partial\Im}\left(4\frac{u_n^2}{\Im} - \frac{\Im u_n}{3}\right) + \frac{\partial^2}{\partial\Im^2}u_n^2\right)\right]\right]d\varepsilon. \tag{76}$$

The initial iteration is

$$u_0(\Im,\wp) = \Im^2.$$

Then for $n = 0$:

$$\begin{aligned}
u_1(\Im,\wp) &= \Im^2 - \int_0^{\wp}\left[\frac{\partial\Im^2}{\partial\varepsilon} - \frac{\partial}{\partial\varepsilon}\mathcal{L}^{-1}\left[\frac{\mathcal{L}}{s^{\zeta}}\left(-\frac{\partial}{\partial\Im}\left(4\frac{\Im^4}{\Im} - \frac{\Im^3}{3}\right) + \frac{\partial^2\Im^4}{\partial\Im^2}\right)\right]\right]d\varepsilon \\
&= \Im^2 - \int_0^{\wp}\left[0 - \frac{\partial}{\partial\varepsilon}\mathcal{L}^{-1}\left[\frac{1}{s^{\zeta}}\mathcal{L}\left(-\frac{\partial}{\partial\Im}\left(4\Im^3 - \frac{\Im^3}{3}\right) + 12\Im^2\right)\right]\right]d\varepsilon \\
&= \Im^2 - \int_0^{\wp}\left[-\frac{\partial}{\partial\varepsilon}\mathcal{L}^{-1}\left[\frac{1}{s^{\zeta}}\mathcal{L}\left(-\frac{\partial}{\partial\Im}\left(\frac{11\Im^3}{3}\right) + 12\Im^2\right)\right]\right]d\varepsilon \\
&= \Im^2 - \int_0^{\wp}\left[-\frac{\partial}{\partial\varepsilon}\mathcal{L}^{-1}\left[\frac{1}{s^{\zeta}}\mathcal{L}\left(-11\Im^2 + 12\Im^2\right)\right]\right]d\varepsilon \\
&= \Im^2 - \int_0^{\wp}\left[-\frac{\partial}{\partial\varepsilon}\mathcal{L}^{-1}\left[\frac{1}{s^{\zeta}}\cdot\frac{\Im^2}{s}\right]\right]d\varepsilon \\
&= \Im^2 - \int_0^{\wp}\left[-\frac{\partial}{\partial\varepsilon}\left(\frac{\Im^2\varepsilon^{\zeta}}{\Gamma(\zeta+1)}\right)\right]d\varepsilon \\
&= \Im^2 + \int_0^{\wp}\left[\frac{\Im^2\zeta\varepsilon^{\zeta-1}}{\Gamma(\zeta+1)}\right]d\varepsilon \\
&= \Im^2 + \frac{\Im^2\wp^{\zeta}}{\Gamma(\zeta+1)}.
\end{aligned}$$

For $n = 1$:

$$\begin{aligned}
u_2(\Im,\wp) &= u_1(\Im,\wp) - \int_0^{\wp}\left[\frac{\partial u_1(\Im,\varepsilon)}{\partial\varepsilon} - \frac{\partial}{\partial\varepsilon}\mathcal{L}^{-1}\left[\frac{\mathcal{L}}{s^{\zeta}}\left(-\frac{\partial}{\partial\Im}\left(4\frac{u_1^2}{\Im} - \frac{\Im u_1}{3}\right) + \frac{\partial^2}{\partial\Im^2}u_1^2\right)\right]\right]d\varepsilon \\
&= \Im^2 + \frac{\Im^2\wp^{\zeta}}{\Gamma(\zeta+1)} - \int_0^{\wp}\left[\frac{\partial}{\partial\varepsilon}\left(\Im^2 + \frac{\Im^2\varepsilon^{\zeta}}{\Gamma(\zeta+1)}\right)\right. \\
&\quad -\frac{\partial}{\partial\varepsilon}\mathcal{L}^{-1}\left[\frac{1}{s^{\zeta}}\mathcal{L}\left(-\frac{\partial}{\partial\Im}\left(\frac{4}{\Im}\left(\Im^2 + \frac{\Im^2\varepsilon^{\zeta}}{\Gamma(\zeta+1)}\right)^2 - \frac{\Im}{3}\left(\Im^2 + \frac{\Im^2\varepsilon^{\zeta}}{\Gamma(\zeta+1)}\right)\right)\right.\right. \\
&\quad \left.\left.\left. + \frac{\partial^2}{\partial\Im^2}\left(\Im^2 + \frac{\Im^2\varepsilon^{\zeta}}{\Gamma(\zeta+1)}\right)^2\right)\right]\right]d\varepsilon \\
&= \Im^2 + \frac{\Im^2\wp^{\zeta}}{\Gamma(\zeta+1)} - \int_0^{\wp}\left[\frac{\Im^2\zeta\varepsilon^{\zeta-1}}{\Gamma(\zeta+1)} - \frac{\partial}{\partial\varepsilon}\mathcal{L}^{-1}\left[\frac{1}{s^{\zeta}}\mathcal{L}\left(\Im^2\left(1 + \frac{\varepsilon^{\zeta}}{\Gamma(\zeta+1)}\right)\right)\right]\right]d\varepsilon \\
&= \Im^2 + \frac{\Im^2\wp^{\zeta}}{\Gamma(\zeta+1)} - \int_0^{\wp}\left[\frac{\Im^2\zeta\varepsilon^{\zeta-1}}{\Gamma(\zeta+1)} - \frac{\partial}{\partial\varepsilon}\left(\frac{\Im^2\varepsilon^{\zeta}}{\Gamma(\zeta+1)} + \frac{\Im^2\varepsilon^{2\zeta}}{\Gamma(2\zeta+1)}\right)\right]d\varepsilon
\end{aligned}$$

$$
\begin{aligned}
&= \Im^2 + \frac{\Im^2 \wp^\zeta}{\Gamma(\zeta+1)} - \int_0^\wp \left[\frac{\Im^2 \zeta \varepsilon^{\zeta-1}}{\Gamma(\zeta+1)} - \left(\frac{\Im^2 \zeta \varepsilon^{\zeta-1}}{\Gamma(\zeta+1)} + \frac{2\Im^2 \zeta \varepsilon^{2\zeta-1}}{\Gamma(2\zeta+1)} \right) \right] d\varepsilon \\
&= \Im^2 + \frac{\Im^2 \wp^\zeta}{\Gamma(\zeta+1)} + \int_0^\wp \left[\frac{2\Im^2 \zeta \varepsilon^{2\zeta-1}}{\Gamma(2\zeta+1)} \right] d\varepsilon \\
&= \Im^2 + \frac{\Im^2 \wp^\zeta}{\Gamma(\zeta+1)} + \frac{\Im^2 \wp^{2\zeta}}{\Gamma(2\zeta+1)}.
\end{aligned}
$$

Similarly,

$$
u_3(\Im, \wp) = \Im^2 \left(1 + \frac{\wp^\zeta}{\Gamma(\zeta+1)} + \frac{\wp^{2\zeta}}{\Gamma(2\zeta+1)} + \frac{\wp^{3\zeta}}{\Gamma(3\zeta+1)} \right).
$$

Hence, the LVIM solution is

$$
u(\Im, \wp) = \Im^2 \left(1 + \frac{\wp^\zeta}{\Gamma(\zeta+1)} + \frac{\wp^{2\zeta}}{\Gamma(2\zeta+1)} + \frac{\wp^{3\zeta}}{\Gamma(3\zeta+1)} + \dots \right) = \Im^2 E_\zeta(\wp^\zeta). \tag{77}
$$

The LADM and LVIM solution at $\zeta = 1$ converges to $u(\Im, \wp) = \Im^2 e^\wp$, which is the exact solution for the nonfractional FPE [16]. Our LADM and LVIM solutions (75) and (77) match with the solutions obtained by HPM [34], NITM [30], HPETM [30], YTDM [2], and HPSTM [6].

Table 3: Comparative analysis of FIM, LADM, LVIM, and exact solution for Example 3, including absolute errors for $\zeta = 1$.

$\Im$	$\wp$	FIM [31]	LADM	LVIM	Exact	FIM Er.	LADM Er.	LVIM Er.
0.1	0.2	0.0124081	0.0122133	0.0122133	0.0122140	0.0001941	0.0000007	0.0000007
0.1	0.3	0.0137434	0.0134950	0.0134950	0.0134986	0.0002452	0.0000036	0.0000036
0.1	0.4	0.0151706	0.0149182	0.0149182	0.0149182	0.0002524	0.0000001	0.0000001
0.6	0.2	0.4466918	0.4396800	0.4396800	0.4397050	0.0069868	0.0000250	0.0000250
0.6	0.3	0.4947610	0.4858200	0.4858200	0.4859492	0.0088118	0.0001292	0.0001292
0.6	0.4	0.5461427	0.5366400	0.5366400	0.5370569	0.0090858	0.0004169	0.0004169
0.9	0.2	1.0050566	0.9892800	0.9892800	0.9893362	0.0157204	0.0000562	0.0000562
0.9	0.3	1.1132122	1.0930950	1.0930950	1.0933856	0.0198265	0.0002906	0.0002906
0.9	0.4	1.2288211	1.2074400	1.2074400	1.2083780	0.0213781	0.0009380	0.0009380

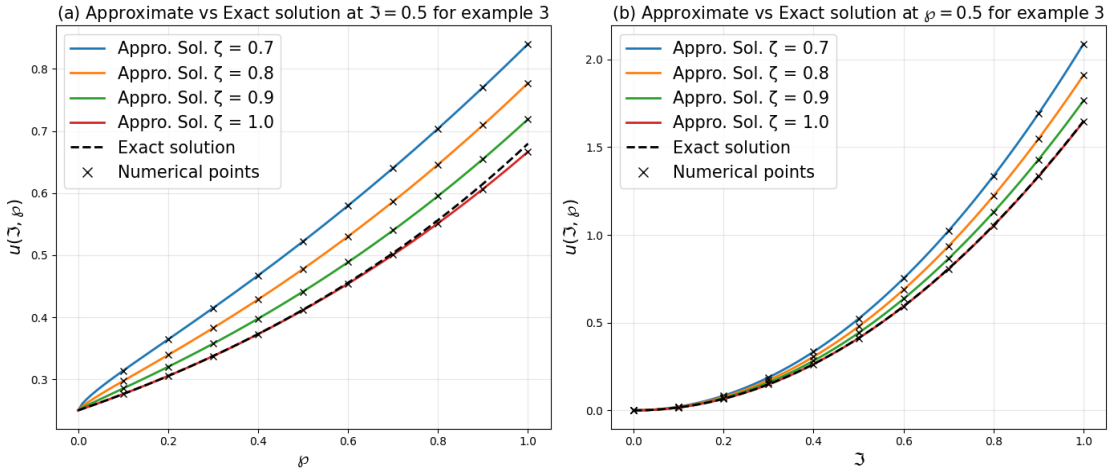


Figure 5: Two-dimensional graphical solutions of Example 3 for different values of ζ: (a) $u(\Im, \wp)$ versus $\wp$, and (b) $u(\Im, \wp)$ versus $\Im$.

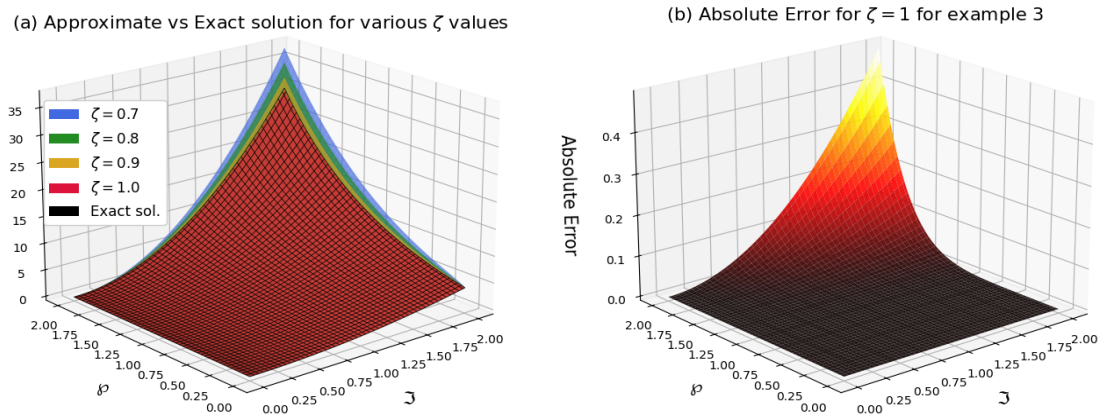


Figure 6: Three-dimensional graphical solutions of Example 3: (a) Comparison of the approximate solutions obtained by the LADM for different values of ζ with the exact solution, and (b) the absolute error surface for $\zeta = 1$.

Example 4. Consider another nonlinear TFFPE:

$$\frac{\partial^{\zeta} u}{\partial \wp^{\zeta}} = -\frac{\partial}{\partial \Im}\left(3u^2 - \frac{\Im u}{2}\right) + \frac{\partial^2}{\partial \Im^2}\left(\Im u^2\right), \tag{78}$$

subject to the initial condition:

$$u(\Im, 0) = \Im, \quad \Im \in \mathbb{R}. \tag{79}$$

(i) LADM solution

Applying the Laplace transform and the inverse Laplace transform to both sides of (78) along with the initial condition, we get

$$u(\Im,\wp)=\Im+\mathcal{L}^{-1}\left[\frac{\mathcal{L}}{s^{\zeta}}\left[-\frac{\partial}{\partial\Im}\left(3u^2-\frac{\Im u}{2}\right)+\frac{\partial^2}{\partial\Im^2}\left(\Im u^2\right)\right]\right]. \tag{80}$$

Now, applying the standard Laplace decomposition method, we get the iterative formula:

$$\sum_{n=0}^{\infty}u_n(\Im,\wp)=\Im+\mathcal{L}^{-1}\left[\frac{\mathcal{L}}{s^{\zeta}}\left[-\frac{\partial}{\partial\Im}\left(3\sum_{n=0}^{\infty}N_n-\frac{\Im}{2}\sum_{n=0}^{\infty}u_n\right)+\frac{\partial^2}{\partial\Im^2}\left(\Im\sum_{n=0}^{\infty}N_n\right)\right]\right], \tag{81}$$

where N_n are Adomian polynomials for u^2 [46]:

$$N_0(u)=u_0^2,\quad N_1(u)=2u_0u_1,\quad N_2(u)=u_1^2+2u_0u_2,\ldots$$

The solution components are

$$\begin{aligned}
u_0(\Im,\wp)&=\Im,\\
u_1(\Im,\wp)&=\mathcal{L}^{-1}\left[\frac{\mathcal{L}}{s^{\zeta}}\left[-\frac{\partial}{\partial\Im}\left(3u_0^2-\frac{\Im u_0}{2}\right)+\frac{\partial^2}{\partial\Im^2}\left(\Im u_0^2\right)\right]\right]\\
&=\mathcal{L}^{-1}\left[\frac{\mathcal{L}}{s^{\zeta}}\left[-\frac{\partial}{\partial\Im}\left(\frac{5\Im^2}{2}\right)+\frac{\partial^2}{\partial\Im^2}\left(\Im^3\right)\right]\right]=\mathcal{L}^{-1}\left[\frac{\Im}{s^{\zeta+1}}\right]=\frac{\Im\wp^{\zeta}}{\Gamma(\zeta+1)},\\
u_2(\Im,\wp)&=\mathcal{L}^{-1}\left[\frac{\mathcal{L}}{s^{\zeta}}\left[-\frac{\partial}{\partial\Im}\left(6u_0u_1-\frac{\Im u_1}{2}\right)+\frac{\partial^2}{\partial\Im^2}\left(2\Im u_0u_1\right)\right]\right]\\
&=\mathcal{L}^{-1}\left[\frac{\mathcal{L}}{s^{\zeta}}\left(-\frac{\partial}{\partial\Im}\left(\frac{11\Im^2}{2}\cdot\frac{\wp^{\zeta}}{\Gamma(\zeta+1)}\right)+\frac{\partial^2}{\partial\Im^2}\left(2\Im^3\frac{\wp^{\zeta}}{\Gamma(\zeta+1)}\right)\right)\right]\\
&=\mathcal{L}^{-1}\left[\frac{\mathcal{L}}{s^{\zeta}}\left[\frac{\Im\wp^{\zeta}}{\Gamma(\zeta+1)}\right]\right]=\mathcal{L}^{-1}\left[\frac{\Im}{s^{2\zeta+1}}\right]=\frac{\Im\wp^{2\zeta}}{\Gamma(2\zeta+1)}\\
u_3(\Im,\wp)&=\frac{\Im\wp^{3\zeta}}{\Gamma(3\zeta+1)}.
\end{aligned}$$

Thus, the LADM solution of (78) is given by

$$u(\Im,\wp)=\Im\left(1+\frac{\wp^{\zeta}}{\Gamma(\zeta+1)}+\frac{\wp^{2\zeta}}{\Gamma(2\zeta+1)}+\frac{\wp^{3\zeta}}{\Gamma(3\zeta+1)}+\cdots\right)=\Im E_{\zeta}(\wp^{\zeta}). \tag{82}$$

(ii) LADM solution

From (29) and (80), the iterative scheme becomes

$$u_{n+1}(\Im,\wp) = u_n(\Im,\wp) - \int_0^{\wp} \left[\frac{\partial u_n(\Im,\varepsilon)}{\partial \varepsilon} \right. \tag{83}$$

$$\left. - \frac{\partial}{\partial \varepsilon}\mathcal{L}^{-1}\left[\frac{\mathcal{L}}{s^{\zeta}} \Big[- \frac{\partial}{\partial \Im}\left(3u_n^2(\Im,\varepsilon) - \frac{\Im\, u_n(\Im,\varepsilon)}{2} \right) + \frac{\partial^2}{\partial \Im^2}\left(\Im\, u_n^2(\Im,\varepsilon) \right) \Big] \right] \right] d\varepsilon. \tag{84}$$

The initial iterationis

$$u_0(\Im,\wp) = \Im.$$

For $n = 0$:

$$u_1(\Im,\wp) = \Im - \int_0^{\wp} \left[\frac{\partial \Im}{\partial \varepsilon} - \frac{\partial}{\partial \varepsilon}\mathcal{L}^{-1}\left[\frac{\mathcal{L}}{s^{\zeta}} \Big[- \frac{\partial}{\partial \Im}\left(3\Im^2 - \frac{\Im^2}{2} \right) + \frac{\partial^2}{\partial \Im^2}\left(\Im^3 \right) \Big] \right] \right] d\varepsilon$$

$$= \Im + \int_0^{\wp} \left[\frac{\Im \zeta \varepsilon^{\zeta-1}}{\Gamma(\zeta+1)} \right] d\varepsilon = \Im\left(1 + \frac{\wp^{\zeta}}{\Gamma(\zeta+1)} \right).$$

For $n = 1$:

$$u_2(\Im,\wp) = u_1(\Im,\wp)$$

$$- \int_0^{\wp} \left[\frac{\partial u_n(\Im,\varepsilon)}{\partial \varepsilon} - \frac{\partial}{\partial \varepsilon}\mathcal{L}^{-1}\left(\frac{\mathcal{L}}{s^{\zeta}} \left[- \frac{\partial}{\partial \Im}\left(3u_1^2 - \frac{\Im u_1}{2} \right) + \frac{\partial^2}{\partial \Im^2}\left(\Im u_1^2 \right) \right] \right) \right] d\varepsilon$$

$$= \Im\left(1 + \frac{\wp^{\zeta}}{\Gamma(\zeta+1)} \right) + \int_0^{\wp} \left[\frac{\Im 2\zeta \varepsilon^{2\zeta-1}}{\Gamma(2\zeta+1)} \right] d\varepsilon$$

$$= \Im\left(1 + \frac{\wp^{\zeta}}{\Gamma(\zeta+1)} + \frac{\wp^{2\zeta}}{\Gamma(2\zeta+1)} \right).$$

For $n = 2$:

$$u_3(\Im,\wp) = \Im\left(1 + \frac{\wp^{\zeta}}{\Gamma(\zeta+1)} + \frac{\wp^{2\zeta}}{\Gamma(2\zeta+1)} + \frac{\wp^{3\zeta}}{\Gamma(3\zeta+1)} \right).$$

The LVIM solution is

$$u(\Im,\wp) = \Im\left(1 + \frac{\wp^{\zeta}}{\Gamma(\zeta+1)} + \frac{\wp^{2\zeta}}{\Gamma(2\zeta+1)} + \frac{\wp^{3\zeta}}{\Gamma(3\zeta+1)} + \cdots \right) = \Im E_{\zeta}(\wp^{\zeta}). \tag{85}$$

The LADM and LVIM solution at $\zeta = 1$ converges to $u(\Im,\wp) = \Im e^{\wp}$, which is the exact solution for the nonfractional and our LADM and LVIM solutions (82) and (85) match with the solutions obtained by HPSTM [6].

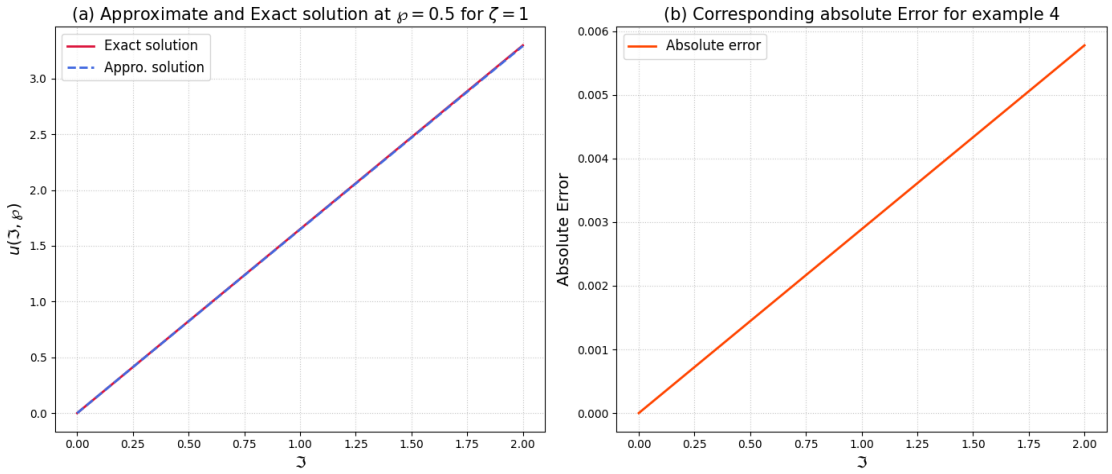


Figure 7: Two-dimensional graphical solutions of Example 4 for $\zeta = 1$: (a) Comparison of approximate and exact solutions, and (b) the corresponding absolute error.

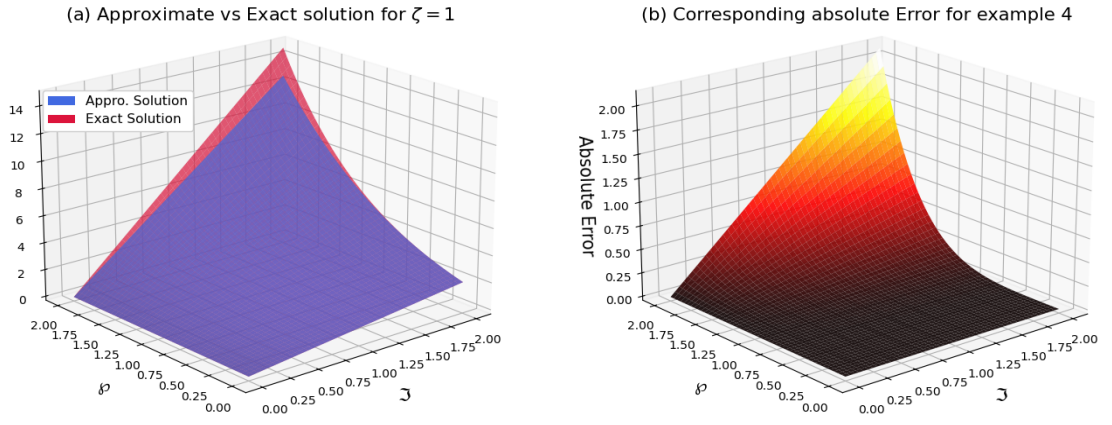


Figure 8: Three-dimensional graphical solutions of Example 4 at $\zeta = 1$: (a) Comparison of approximate and exact solutions, (b) the corresponding absolute error.

Example 5. Consider another TFFPE (4) with $A_1(\Im, \xi) = \Im$, $A_2(\Im, \xi) = 5\xi$, $B_{1,1}(\Im, \xi) = \Im^2$, $B_{1,2}(\Im, \xi) = 1$, $B_{2,1}(\Im, \xi) = 1$, $B_{2,2}(\Im, \xi) = \xi^2$, that is,

$$\frac{\partial^{\zeta} u}{\partial \wp^{\zeta}} = -\frac{\partial}{\partial \Im}\Im u - \frac{\partial}{\partial \xi}5\xi u + \frac{\partial^2}{\partial \Im^2}\Im^2 u + \frac{\partial^2}{\partial \Im \partial \xi}u + \frac{\partial^2}{\partial \xi \partial \Im}u + \frac{\partial^2}{\partial \xi^2}\xi^2 u, \tag{86}$$

Subject to the initial condition:

$$u(\Im,\xi,0)=\Im,\quad \Im,\xi\in\mathbb{R}. \tag{87}$$

(i) LADM solution

Applying the Laplace transform to both sides and using the initial condition, we obtain

$$u(\Im,\xi,s)=\frac{\Im}{s}+\frac{1}{s^{\zeta}}\mathcal{L}\left[-\frac{\partial}{\partial\Im}(\Im u)-\frac{\partial}{\partial\xi}(5\xi u)+\frac{\partial^2}{\partial\Im^2}(\Im^2 u)+\frac{\partial^2}{\partial\Im\partial\xi}u+\frac{\partial^2}{\partial\xi\partial\Im}u+\frac{\partial^2}{\partial\xi^2}(\xi^2 u)\right]. \tag{88}$$

The recurrence relation is

$$\begin{aligned}u_0(\Im,\xi,s)&=\frac{\Im}{s},\\ u_{n+1}(\Im,\xi,s)&=\frac{1}{s^{\zeta}}\mathcal{L}\left[-\frac{\partial}{\partial\Im}(\Im u_n)-\frac{\partial}{\partial\xi}(5\xi u_n)+\frac{\partial^2}{\partial\Im^2}(\Im^2 u_n)+\frac{\partial^2}{\partial\Im\partial\xi}u_n+\frac{\partial^2}{\partial\xi\partial\Im}u_n+\frac{\partial^2}{\partial\xi^2}(\xi^2 u_n)\right].\end{aligned} \tag{89}$$

Applying the inverse Laplace transform, we have

$$u_0(\Im,\xi,\wp)=\mathcal{L}^{-1}\left(\frac{\Im}{s}\right)=\Im.$$

For $n=0$:

$$\begin{aligned}u_1(\Im,\xi,\wp)&=\mathcal{L}^{-1}\left[\frac{1}{s^{\zeta}}\mathcal{L}\left[-\frac{\partial}{\partial\Im}(\Im^2)-\frac{\partial}{\partial\xi}(5\xi\Im)+\frac{\partial^2}{\partial\Im^2}(\Im^3)+\frac{\partial^2}{\partial\Im\partial\xi}\Im+\frac{\partial^2}{\partial\xi\partial\Im}\Im+\frac{\partial^2}{\partial\xi^2}(\xi^2\Im)\right]\right]\\ &=\mathcal{L}^{-1}\left[\frac{1}{s^{\zeta}}\mathcal{L}\left[\Im\right]\right]=\mathcal{L}^{-1}\left[\frac{\Im}{s^{\zeta+1}}\right]=\Im\frac{\wp^{\zeta}}{\Gamma(\zeta+1)}.\end{aligned}$$

For $n=1$:

$$u_2(\Im,\xi,\wp)=\mathcal{L}^{-1}\left[\frac{1}{s^{\zeta}}\mathcal{L}\left[\Im\frac{\wp^{\zeta}}{\Gamma(\zeta+1)}\right]\right]=\mathcal{L}^{-1}\left[\frac{\Im}{s^{2\zeta+1}}\right]=\Im\frac{\wp^{2\zeta}}{\Gamma(2\zeta+1)}.$$

Similarly,

$$u_3(\Im,\xi,\wp)=\Im\frac{\wp^{3\zeta}}{\Gamma(3\zeta+1)}.$$

Thus, the LADM solution is

$$u(\Im,\xi,\wp)=\Im\left(1+\frac{\wp^{\zeta}}{\Gamma(\zeta+1)}+\frac{\wp^{2\zeta}}{\Gamma(2\zeta+1)}+\frac{\wp^{3\zeta}}{\Gamma(3\zeta+1)}+\cdots\right)=\Im E_{\zeta}(\wp^{\zeta}). \tag{90}$$

(ii) LVIM solution

The iterative formula is:

$$u_{n+1}(\Im,\xi,\wp) = u_n(\Im,\xi,\wp) - \int_0^{\wp}\left[\frac{\partial u_n}{\partial\varepsilon} - \frac{\partial}{\partial\varepsilon}\mathcal{L}^{-1}\left[\frac{1}{s^{\zeta}}\mathcal{L}\left[-\frac{\partial}{\partial\Im}(\Im u_n) - \frac{\partial}{\partial\xi}(5\xi u_n) + \frac{\partial^2}{\partial\Im^2}(\Im^2 u_n) + \frac{\partial^2}{\partial\Im\partial\xi}u_n + \frac{\partial^2}{\partial\xi\partial\Im}u_n + \frac{\partial^2}{\partial\xi^2}(\xi^2 u_n)\right]\right]\right]d\varepsilon. \tag{91}$$

The initial iteration is

$$u_0(\Im,\xi,\wp) = \Im.$$

For $n = 0$:

$$u_1(\Im,\xi,\wp) = \Im + \int_0^{\wp}\left[\frac{\Im\zeta\varepsilon^{\zeta-1}}{\Gamma(\zeta+1)}\right]d\varepsilon = \Im + \Im\frac{\wp^{\zeta}}{\Gamma(\zeta+1)}.$$

For $n = 1$:

$$u_2(\Im,\xi,\wp) = \Im + \Im\frac{\wp^{\zeta}}{\Gamma(\zeta+1)} + \int_0^{\wp}\left[\frac{2\Im\zeta\varepsilon^{2\zeta-1}}{\Gamma(2\zeta+1)}\right]d\varepsilon = \Im + \Im\frac{\wp^{\zeta}}{\Gamma(\zeta+1)} + \Im\frac{\wp^{2\zeta}}{\Gamma(2\zeta+1)}.$$

Similarly,

$$u_3(\Im,\xi,\wp) = \Im + \Im\frac{\wp^{\zeta}}{\Gamma(\zeta+1)} + \Im\frac{\wp^{2\zeta}}{\Gamma(2\zeta+1)} + \Im\frac{\wp^{3\zeta}}{\Gamma(3\zeta+1)}.$$

Hence, the LVIM solution is

$$u(\Im,\xi,\wp) = \Im\left(1 + \frac{\wp^{\zeta}}{\Gamma(\zeta+1)} + \frac{\wp^{2\zeta}}{\Gamma(2\zeta+1)} + \frac{\wp^{3\zeta}}{\Gamma(3\zeta+1)} + \cdots\right) = \Im E_{\zeta}(\wp^{\zeta}). \tag{92}$$

The LADM and LVIM solution at $\zeta = 1$ converge to $u(\Im,\xi,\wp) = \Im e^{\wp}$, which is the exact solution for the nonfractional and our LADM and LVIM solutions (90) and (92) match with the solutions obtained by HPM [34].

7 Results and discussion

In Table 1, we present the numerical results of Example 1 for different values of $\wp$ and $\Im$, obtained using HPSTM [6], NITM [30], HPETM [30], YTDM [2], LADM, and LVIM. The results are compared with the exact solution when $\zeta = 1$. It can be observed that LADM and LVIM produce identical results, which are in excellent agreement with the exact solution. Moreover, the consistency of LADM and LVIM with other recently developed methods (HPSTM, NITM, HPETM, and YTDM) further demonstrates the validity, accuracy, and robustness of the proposed approaches.

Tables 2 and 3 summarize the numerical results for Examples 2 (linear) and 3 (nonlinear) for different values of $\Im$ and $\wp$, comparing the FIM [31], LADM, LVIM, and exact solutions at $\zeta = 1$. In both linear and nonlinear cases, LADM and LVIM produce nearly identical results

that closely match the exact solutions, exhibiting minimal absolute errors. While FIM performs adequately for the linear problem, it shows comparatively larger deviations in the nonlinear case. These observations confirm that the proposed Laplace-based hybrid methods are highly accurate, reliable, and computationally efficient, achieving rapid convergence with fewer iterations and lower computational cost compared to traditional approaches. Furthermore, the accuracy of the approximate solutions can be systematically improved by including additional terms in the series expansions.

Since LADM and LVIM solutions yield the same results, we consider only the LADM solution in the graphs. Figures 1–6 illustrate the LADM solutions compared with the exact solutions for Examples 1–3.

In Figure 1, the two-dimensional LADM solutions for $\zeta = 0.8, 0.9, 1$ at $\Im = 0.5$ over $\wp \in [0, 0.5]$ are shown, with cross markers indicating discrete solution points, while graph (b) displays the exact solution at $\zeta = 1$ with circular markers. Excellent agreement between the approximate and exact solutions is observed. Figure 2 presents the corresponding three-dimensional surfaces over $\Im \in [0, 1]$ and $\wp \in [0, 1]$, further confirming this agreement.

For Example 2, Figures 3 and 4 show two-dimensional and three-dimensional comparisons at $\zeta = 1$ across different $\Im$ and $\wp$ values. The two-dimensional absolute error in Figure 3(b) and the three-dimensional error in Figure 4(b) remain very small, demonstrating that LADM converges quickly and accurately even with only the first four terms of the series.

Figures 5 and 6 illustrate the effect of different fractional orders $\zeta = 0.7, 0.8, 0.9, 1$ for Example 3. Two-dimensional plots show LADM solutions for fixed $\Im$ or $\wp$ compared with the exact solution at $\zeta = 1$, and three-dimensional plots display the solution surfaces and absolute errors. In all cases, LADM solutions closely match the exact solutions, confirming the accuracy and reliability of the proposed method.

Figures 7 and 8 present the graphical results for Example 4 at $\zeta = 1$. The two-dimensional plots compare the approximate and exact solutions along with the corresponding absolute error. The three-dimensional plots illustrate the solution surfaces and the absolute error distribution. The results show an excellent agreement between the approximate and exact solutions, demonstrating the accuracy and effectiveness of the proposed method.

8 Conclusion

In this study, the LADM and the LVIM were successfully applied to solve both linear and nonlinear TFFPEs (time-fractional Fokker–Planck equations), as well as related fractional differential equations. The proposed methods proved to be efficient and robust in obtaining analytical and

numerical solutions, achieving excellent agreement with exact solutions after just a few iterations. Compared to existing approaches such as YTDM, NITM, HPETM, HPSTM, and FIM, LADM and LVIM require less computational effort while providing higher accuracy, particularly for nonlinear TFFPEs. Unlike traditional numerical techniques, these methods do not require discretization, linearization, or perturbation, making them highly practical for complex fractional systems. The present work can be extended to more sophisticated fractional models arising in physics, engineering, and finance, demonstrating the versatility of LADM and LVIM. Future research may explore higher-dimensional fractional models or the application of other fractional operators, including Atangana–Baleanu, Caputo–Fabrizio, and conformable derivatives.

Declarations

- **Data Availability**: All data presented in this study were generated using Python, and no external datasets were used. The numerical data and Python codes used for computations and plotting are available from the corresponding author upon reasonable request.
- **Conflict of Interest**: The authors declare no conflict of interest.
- **Funding**: No funding.
- **Monowar Hossain**: Conceptualization, Investigation, Visualization, Validation, Methodology, Software, Formal analysis, Data curation, Writing - original draft ;
  Mohammed Aman Ullah : Supervision, Writing - review and editing

References

[1] Adeniji, A.A., Mogbojuri, O.A., Kekana, M.C. and Fadugba, S.E. *Numerical solution of rotavirus model using Runge-Kutta-Fehlberg method, differential transform method and Laplace Adomian decomposition method*, Alexandria Eng. J., 82 (2023), 323–329.

[2] Alshammari, S., Iqbal, N. and Yar, M. *Fractional-view analysis of space-time fractional Fokker-Planck equations within Caputo operator*, J. Funct. Spaces, 2022(1) (2022), 4471757.

[3] Bekela, A.S. and Deresse, A.T. *A hybrid yang transform adomian decomposition method for solving time-fractional nonlinear partial differential equation*, BMC Res. Notes, 17(1) (2024), 226.

[4] Bhargava, A., Jain, D. and Suthar, D.L. *Applications of the Laplace variational iteration method to fractional heat-like equations*, Partial Differ. Equ. Appl. Math., 8 (2023), 100540.

[5] Bhuyan, B., Sahoo, M., Jena, S.R., and Senapati, A. *Quadratic and cubic order convergence for Kink-Antikink numerical soliton solution profile of (2+1)-dimensional Chaffee-Infante equations in mathematical physics*, Phys. Scr., 100(12) (2025), 125204.

[6] Dubey, R.S., Alkahtani, B.S.T. and Atangana, A. *Analytical solution of space-time fractional Fokker-Planck equation by homotopy perturbation Sumudu transform method*, Math. Probl. Eng., 2015(1) (2015), 780929.

[7] Elzaki, T.M. and Alkhateeb, S.A. *Modification of Sumudu transform "Elzaki transform" and Adomian decomposition method*, Appl. Math. Sci., 9(13) (2015), 603–611.

[8] Farman, M., Saleem, M.U., Ahmad, A. and Ahmad, M.O. *Analysis and numerical solution of SEIR epidemic model of measles with non-integer time fractional derivatives by using Laplace Adomian decomposition method*, Ain Shams Eng. J., 9(4) (2018), 3391–3397.

[9] Frank, T.D. *Stochastic feedback, nonlinear families of Markov processes, and nonlinear Fokker–Planck equations*, Physica A, 331(3–4) (2004), 391–408.

[10] Habenom, H. and Suthar, D.L. *Numerical solution for the time-fractional Fokker–Planck equation via shifted Chebyshev polynomials of the fourth kind*, Adv. Difference Equ., 2020 (2020), 1–16.

[11] Habenom, H.A.I.L.E., Suthar, D.L. and Aychluh, M.U.L.U.A.LE.M. *Solution of fractional Fokker–Planck equation using fractional power series method*, J. Sci. Arts, 48(3) (2019), 593–600.

[12] Hammouch, Z. and Mekkaoui, T. *A Laplace-variational iteration method for solving the homogeneous Smoluchowski coagulation equation*, Appl. Math. Sci., 6(18) (2012), 879–886.

[13] Hammouch, Z. and Mekkaoui, T. *Application of the Laplace variational iteration method to fractional diffusion and reaction–diffusion equations*, Nonlinear Dyn., 87(1) (2017), 487–495.

[14] Haq, F., Shah, K., ur Rahman, G. and Shahzad, M. *Numerical solution of fractional order smoking model via Laplace Adomian decomposition method*, Alexandria Eng. J., 57(2) (2018), 1061–1069.

[15] He, J.H. *Variational iteration method–a kind of non-linear analytical technique: Some examples*, Int. J. Non-Linear Mech., 34(4) (1999), 699–708.

[16] Hesam, S., Nazemi, A.R. and Haghbin, A. *Analytical solution for the Fokker–Planck equation by differential transform method*, Sci. Iran., 19(4) (2012), 1140–1145.

[17] Hilal, E.M.A. and Elzaki, T.M. *Solution of nonlinear partial differential equations by new Laplace variational iteration method*, J. Funct. Spaces, 2014 (2014), 1–5.

[18] Hussain, A. and others *Laplace variational iteration method for fractional transport phenomena*, J. King Saud Univ.-Sci., 31 (2019).

[19] Hussain, W.S., Ali, S., Shah, K., Abdalla, B. and Abdeljawad, T. *Semi analytical scheme for the presentation of solution to fractional Fokker–Planck equation*, Partial Differ. Equ. Appl. Math., 10 (2024), 100740.

[20] Hussein, M.A. *A review on algorithms of Laplace Adomian decomposition method for FPDEs*, Sci. Res. J. Multidiscip., 2 (2022), 1–10.

[21] Jain, D. and Bhargava, A. *Laplace variational iteration method for solving fractional wave-like equations*, J. Comput. Anal. Appl., 31(3) (2023), 377–399.

[22] Jena, S.R. and Sahu, I. *A novel approach for numerical treatment of traveling wave solution of ion acoustic waves as a fractional nonlinear evolution equation on Shehu transform environment*, Phys. Scr., 98(8) (2023), 085231.

[23] Jena, S.R., and Sahu, I. *A reliable method for voltage of telegraph equation in one and two space variables in electrical transmission: approximate and analytical approach*, Phys. Scr., 98(10) (2023), 105216.

[24] Jena, S.R. and Sahu, I. *Approximate and analytical solution of sine–Gordon equation based on Laplace Adomian decomposition method*, Int. J. Syst. Assur. Eng. Manag., 16 (3) (2025), 914–926.

[25] Khan, M.N., Haider, J.A., Wang, Z., Lone, S.A., Almutlak, S.A. and Elseesy, I.E. *Application of Laplace-based variational iteration method to analyze generalized nonlinear oscillations in physical systems*, Mod. Phys. Lett. B, 37(34) (2023), 2350169.

[26] Khuri, S.A. *A Laplace decomposition algorithm applied to a class of nonlinear differential equations*, J. Appl. Math., 1(4) (2001), 141–155.

[27] Kreyszig, E. *Introductory functional analysis with applications*, Wiley, New York, 1978.

[28] Mahdy, A.M. *Numerical solutions for solving model time-fractional Fokker–Planck equation*, Numer. Method Partial Diff. Equ. 37(2) (2021), 1120–1135.

[29] Mahmood, S., Shah, R., Khan, H. and Arif, M. *Laplace Adomian decomposition method for multi-dimensional time fractional model of Navier-Stokes equation*, Symmetry, 11(2) (2019), 149.

[30] Mofarreh, F., Khan, A., Shah, R. and Abdeljabbar, A. *A comparative analysis of fractional-order Fokker–Planck equation*, Symmetry, 15(2) (2023), 430.

[31] Mohamed, A.S., Mahdy, A.M.S. and Mtawa, A.H. *Approximate analytical solution to a time-fractional Fokker–Planck equation*, Bothalia, 45(4) (2015), 57–69.

[32] Mohamed, M.Z. and Elzaki, T.M. *Comparison between the Laplace decomposition method and Adomian decomposition in time-space fractional nonlinear fractional differential equations*, Appl. Math., 9(4) (2018), 448.

[33] Mohamed, M.Z., Elzaki, T.M., Algolam, M.S., Abd Elmohmoud, E.M. and Hamza, A.E. *New modified variational iteration Laplace transform method compares Laplace Adomian decomposition method for solution time-partial fractional differential equations*, J. Appl. Math., 2021(1) (2021) 6662645.

[34] Mousa, M.M. and Kaltayev, A. *Application of He's homotopy perturbation method for solving fractional Fokker–Planck equations*, Z. Naturforsch. A, 64(12) (2009), 788–794.

[35] Nadeem, M. and Li, F. *Modified Laplace variational iteration method for analytical approach of Klein–Gordon and Sine–Gordon equations*, Iran. J. Sci. Technol. Trans. A Sci., 43(4) (2019), 1933–1940.

[36] Nadeem, M. and He, J.H. *He–Laplace variational iteration method for solving nonlinear equations arising in chemical kinetics and population dynamics*, J. Math. Chem., 59(5) (2021), 1234–1245.

[37] Podlubny, I. *Fractional differential equations*, Elsevier, San Diego, 1998.

[38] Qing, Y. and Rhoades, B.E. *T-stability of Picard iteration in metric spaces*, Fixed Point Theory Appl., 2008(1) (2008), 418971.

[39] Risken, H.A. Abbott and Risken, H. *Fokker–Planck equation*, Springer, Berlin, Heidelberg, 1996, pp. 63–95.

[40] Sahu, I. and Jena, S.R. *An efficient technique for time fractional Klein–Gordon equation based on modified Laplace Adomian decomposition technique via hybridized Newton–Raphson scheme arises in relativistic fractional quantum mechanics*, Partial Differ. Equ. Appl. Math., 10 (2024), 100744.

[41] Sahu, I. and Jena, S.R. *The kink–antikink single waves in dispersion systems by generalized PHI-four equation in mathematical physics*, Phys. Scr., 99(5) (2024), 055258.

[42] Sahu, I. and Jena, S.R. *SDIQR mathematical modelling for COVID-19 of Odisha of India with nonlinear ordinary differential equation based on Laplace Adomian decomposition technique*, Model. Earth Syst. Environ., 9(4) (2023), 4031–4040.

[43] Singh, J. and Kumar, R. *An analytical study on fractional Fokker–Planck equation by homotopy analysis transform method*, In: Recent Advances in Mathematics, Statistics and Computer Science, 2016, pp. 313–321.

[44] Subartini, B., Sumiati, I., Sukono, R. and Sulaiman, I.M. *Combined adomian decomposition method with integral transform*, Math. Stat., 9(6) (2021), 976–983.

[45] Tarasov, V.E. *Fractional dynamics: Applications of fractional calculus to dynamics of particles, fields and media*, Springer, Berlin, Heidelberg, 2011.

[46] Wazwaz, A.M. *Partial differential equations and solitary waves theory*, Springer, Berlin, Heidelberg, 2010.

[47] Zhang, Y. and Pang, J. *Laplace-based variational iteration method for nonlinear oscillators in microelectromechanical systems*, Math. Methods Appl. Sci., 49(6) (2026), 5755–5761.

[48] Zhou, S. and Jiang, Y. *Finite volume methods for N-dimensional time fractional Fokker–Planck equations*, Bull. Malays. Math. Sci. Soc., 42 (2019), 3167–3186.

Iranian Journal of Numerical Analysis and Optimization
Vol. 16, No. 3, 2026, pp 852–875
https://doi.org/10.22067/ijnao.2026.95106.1716
https://ijnao.um.ac.ir/

CC BY NC

Research Article

A probabilistic approach to portfolio selection with Risk-Aversion coefficient estimated via logistic regression

L. Chin*, A. Sukmana, E. Chendra and A.P. Wijaya

Abstract

This work presents a probabilistic risk-aversion approach to portfolio selection that combines a logistic regression-based risk assessment framework with a mixed-integer optimization algorithm with mean semi-absolute deviation as the risk criterion. It allows us to incorporate uncertainty in the portfolio selection analysis as well as performance analysis by Monte Carlo

*Corresponding author

Received 27 August 2025; revised 18 December 2025; accepted 18 April 2026

Liem Chin
Center for Mathematics and Society, Department of Mathematics, Faculty of Science, Parahyangan Catholic University, Bandung, Indonesia. e-mail: chin@unpar.ac.id

Agus Sukmana
Center for Mathematics and Society, Department of Mathematics, Faculty of Science, Parahyangan Catholic University, Bandung, Indonesia. e-mail: asukmana@unpar.ac.id

Erwinna Chendra
Center for Mathematics and Society, Department of Mathematics, Faculty of Science, Parahyangan Catholic University, Bandung, Indonesia. e-mail: erwinna@unpar.ac.id

Andreas Parama Wijaya
Center for Mathematics and Society, Department of Mathematics, Faculty of Science, Parahyangan Catholic University, Bandung, Indonesia. e-mail: a.p.wijaya@unpar.ac.id

How to cite this article
Chin, L., Sukmana, A., Chendra, E. and Wijaya, A.P., A probabilistic approach to portfolio selection with Risk-Aversion coefficient estimated via logistic regression. *Iran. J. Numer. Anal. Optim.*, 2026; 16(3): 852-875. https://doi.org/10.22067/ijnao.2026.95106.1716

simulation and compares the portfolios with different levels of risk over time instead of applying a deterministic system of selection based on historical data. The technical indicators of the LQ45 index predict the risk-aversion coefficient and can steer the model to the direction of the market. Under cardinality, quantity, and liquidity constraints, seven portfolios constructed from major Indonesian stock indices can be optimized with a Branch-and-Bound method that operates in GEKKO Python. Simulation results on 10,000 runs are taken to give the portfolios probabilistic rankings of return and risk. The probability values computed using the Monte Carlo method are plotted against the sample size to illustrate the convergence behavior of the algorithm. Eventually, to evaluate the robustness of the algorithm, variations in the probability values are examined by adding normally distributed noise into the model input parameters.

AMS subject classifications (2020): Primary 91G10; Secondary 91G60.

Keywords: Logistic regression; Mean semi-absolute deviation; Monte-Carlo simulation; Portfolio selection; Risk-aversion coefficients.

1 Introduction

After Markowitz introduced the mean-variance portfolio theory [13], many research topics have been based on it. For example, studies by Anagnostopoulos and Mamanis [1] and Cesarone, Scozzari, and Tardella [3] built upon the theory, while Simaan [17] compared it to other theories, such as mean absolute deviation (MAD). There is also research in which portfolio risk is measured using MAD, mean semi-absolute deviation (MSAD), or Gini's mean difference (GMD) instead of the variance used by Markowitz.

Konno and Yamazaki [11] and Konno and Wijayanayake [10] chose MAD as a measure of portfolio risk for two reasons: It is a more straightforward approach, and the Markowitz model, which uses a covariance matrix, is considered impractical due to computational difficulties and implementation complications. MAD accounts for all deviations of a portfolio's return rate from its expected value, both smaller and larger than the expected value. However, a rational investor might only consider deviations below the expected value as actual risk. In other words, variability in the return rate of a portfolio above the average should not be penalized, as investors are more concerned with poor performance (underperformance) than with excessive performance (overperformance). For this reason, MSAD is preferred over MAD, as seen in the work of Sortino and Meer [18]. Although they did not explicitly use the term MSAD, their article introduced the concept of downside deviation, which is the basis of MSAD. They argued that investors are more concerned with the risk of loss, so risk measurement should focus on deviations below the

average. Other researchers who have used MSAD as a measure of portfolio risk include Fang, Lai, and Wang [4] and Yixuan and Zhongfeng [21].

Using GMD for risk measurement, as done by Ji, Lejeune, and Prasad [8] and Jiang, Ji, and Chang [9], is actually better than MSAD. This is because GMD is calculated by averaging all absolute differences between every pair of data points in a data set, whereas MSAD is calculated by averaging the absolute deviations of returns that are below the average return. However, this increases the computational complexity of GMD at $O(n^2)$, compared to MSAD, which is only $O(n)$. For this reason, we chose MSAD as our portfolio risk measure over GMD.

Apart from risk measurement, there are other things that need to be considered in forming a portfolio, that is, risk-aversion adjustment. This adjustment is exhibited by Jiang, Ji, and Chang [9] and Grant, Kwon, and Satchell [6]. The risk-aversion coefficient is needed to reflect the market trend [9]. Finally, there are several practical constraints that need to be considered in portfolio selection: Cardinality (the number of assets in a portfolio) [1, 3, 14, 20], quantity (the proportion of funds for each asset) [1, 3, 20], and transaction cost [4].

In general, research on portfolio optimization uses deterministic methods in formulating problems [22]; therefore, an alternative approach is needed to consider uncertainty and address the problem probabilistically, as proposed by Shadabfar and Cheng [15] and Xiaoxia [19]. The main objective of this research is to provide a consideration for investors by adopting Shadabfar-Cheng's algorithm. Our research introduces several key methodological enhancements and extensions compared to the work of Shadabfar and Cheng [15]. The main differences are as follows.

1. The portfolio risk measure is calculated using the MSAD, rather than the portfolio variance.

2. We incorporate a risk-aversion coefficient that reflects the market trends.

3. The model is extended by incorporating cardinality, quantity, transaction cost, and liquidity constraints.

4. We display the proportion of funds for each asset, which is essential for investors but is often not discussed in many researches.

We use Monte Carlo simulation to present a probabilistic solution for the individual return and risk of portfolios [16].

The remainder of this paper is organized as follows. Section 2 begins with a brief introduction to the fundamental methods. Section 3 then describes the portfolio optimization model we constructed and the probabilistic approach methodology employed in this research. Section 4 is dedicated to the discussion and analysis of the research results, including the convergence

of Monte Carlo simulation and robustness of the algorithm to random perturbations. Finally, Section 5 provides the conclusion based on the obtained findings.

2 Overview

2.1 Logistic regression

Logistic regression is a type of statistical analysis used to model the relationship between one or more independent variables (predictors) and a single categorical or binary dependent variable. Unlike linear regression, which predicts a continuous value, logistic regression predicts the probability that an event occurs, with the output being a value between 0 and 1.

In the context of this study, logistic regression is used to predict the direction of a stock index movement and to determine the risk-aversion coefficient α, as demonstrated by Jiang, Ji, and Chang [9]. Although Jiang et al. actually used both logistic regression and XGBoost for this prediction, the accuracy of the two methods was not significantly different (XGBoost was only 1.01% more accurate than logistic regression). For this reason and because logistic regression is easier to implement and interpret, this study uses logistic regression to predict the direction of the movement of the stock index.

According to James et al. [7], logistic regression can be expressed as

$$\log\left(\frac{p(X)}{1-p(X)}\right) = \beta_0 + \sum_{i=1}^{k} \beta_i X_i, \tag{1}$$

where $X = (X_1, X_2, \ldots, X_k)$ are the k predictors. Equation (1) can be rewritten as

$$p(X) = \frac{e^{\beta_0 + \sum_{i=1}^{k} \beta_i X_i}}{1 + e^{\beta_0 + \sum_{i=1}^{k} \beta_i X_i}}.$$

The probability value $p(X)$ is then used as an input for the risk-aversion coefficient α in the portfolio optimization model [9].

2.2 Branch & Bound method

The Branch and Bound algorithm is a divide-and-conquer methodology for solving discrete optimization problems. We adopted the robust method from [12], which is often partially responsible for successfully solving complex problems.

Consider the integer linear program (ILP):

$$\begin{aligned} z_P := \max \quad & \sum_{j=1}^{n} c_j y_j \\ \text{subject to :} \quad & \\ & \sum_{j=1}^{n} a_{ij} y_j \leq b_i, \text{for } i = 1, 2, \ldots, m, \\ & y_j \in \mathbb{Z}, \text{for } j = 1, 2, \ldots, n. \end{aligned}$$

We assume that the feasible region of the linear programming (LP) relaxation of P is a bounded set. Consequently, the set of feasible solutions for a subproblem P' is a finite set.

1. **Upper Bounds:** Solve the linear programming relaxation of the subproblem P'. Let y^* be its optimal solution. The optimal objective function value of the LP relaxation of P' serves as an upper bound for $z_{P'}$.

2. **Branching Rule:** Select a variable y_k that is not an integer, that is, $y_k^* \notin \mathbb{Z}$. The branching is performed by creating two new subproblems: (a) P' with the additional constraint $y_k \leq \lfloor y_k^* \rfloor$, and (b) P' with the additional constraint $y_k \geq \lceil y_k^* \rceil$.

3. **Lower Bounds:** If the solution y^* of the LP relaxation of P' happens to be in S' (the set of feasible solutions to the integer linear program P'), then the objective value $z^* := c(y^*)$ constitutes a lower bound on $z_{P'}$. This bound can be further refined. Even if the optimal solution is not feasible for P', feasible solutions to P' can be visited during the process of solving its LP relaxation (e.g., using the primal simplex method). Each of these solutions also provides a lower bound on $z_{P'}$. Furthermore, it may be possible to perturb the solution of the LP relaxation to obtain a feasible solution to P'; for instance, if the coefficients a_{ij} are all nonnegative, then rounding down the components of y^* provides a feasible solution.

3 Problem formulation and methodology

3.1 Portfolio model

In this study, portfolio risk is measured using the MSAD, as proposed by Fang, Lai, and Wang [4].

Assume that a portfolio consists of n assets and that r_{ij} represents the return of the ith asset in the jth period, with $i = 1, 2, \ldots, n$ and $j = 1, 2, \ldots, m$. It is also assumed that $m > n$. Let y_i, for $i = 1, 2, \ldots, n$, denote the proportion of investment for the ith asset, with $\sum_{i=1}^{n} y_i = 1$. With these proportions, we assume that investors can buy assets in real numbers.

The semi-absolute deviation of the portfolio return with respect to the expected return in period $j = 1, \ldots, m$ is defined as follows [4]:

$$v_j(y) = \min\left(0, \sum_{i=1}^{n}(r_{ij} - \bar{r}_i)y_i\right) = \frac{1}{2}\left(\sum_{i=1}^{n}(r_{ij} - \bar{r}_i)y_i - \left|\sum_{i=1}^{n}(\bar{r}_i - r_{ij})y_i\right|\right)$$

with $\bar{r}_i$ being the average return of the ith asset.

Thus, the MSAD (mean semi-absolute deviation) of the portfolio return can be expressed as:

$$V(y) = \frac{1}{m}\sum_{j=1}^{m} v_j(y) = \frac{1}{2m}\sum_{j=1}^{m}\left(\sum_{i=1}^{n}(r_{ij} - \bar{r}_i)y_i - \left|\sum_{i=1}^{n}(\bar{r}_i - r_{ij})y_i\right|\right),$$

where $y = (y_1, y_2, \ldots, y_n)^\top$ and $y_i \in [0, 1]$.

The portfolio return without transaction cost is given by $\sum_{i=1}^{n} \bar{r}_i y_i$. If we assume that the transaction cost for each asset is γ_i. Then the total transaction cost is $\sum_{i=1}^{n} \gamma_i y_i$, so we have to subtract the previous return by this cost to get the net of return portfolio, that is,

$$\sum_{i=1}^{n} \bar{r}_i y_i - \sum_{i=1}^{n} \gamma_i y_i.$$

The portfolio optimization model in which risk is measured using MSAD is given by

$$\max \quad z_1 = \sum_{i=1}^{n} \bar{r}_i y_i - \sum_{i=1}^{n} \gamma_i y_i, \tag{2a}$$

$$\min \quad z_2 = \frac{1}{2m}\sum_{j=1}^{m}\left(\sum_{i=1}^{n}(r_{ij} - \bar{r}_i)y_i - \left|\sum_{i=1}^{n}(\bar{r}_i - r_{ij})y_i\right|\right), \tag{2b}$$

subject to

$$\sum_{i=1}^{n} s_i = K, \tag{2c}$$

$$0 \leq y_i \leq \delta_i s_i, \quad i = 1, \ldots, n, \tag{2d}$$

$$l_i \leq \delta_i \leq u_i, \quad i = 1, \ldots, n, \tag{2e}$$

$$\sum_{i=1}^{n} y_i + \sum_{i=1}^{n} \gamma_i y_i = 1, \tag{2f}$$

$$\sum_{i=1}^{n} \lambda_i y_i \geq L, \tag{2g}$$

$$s_i \in \{0, 1\}, \quad i = 1, \ldots, n, \tag{2h}$$

$$y_i \geq 0, \quad i = 1, \ldots, n. \tag{2i}$$

The portfolio model (2) can be explained as follows:

1. The functions z_1 (2a) and z_2 (2b) represent the portfolio return and risk, respectively. Since investors naturally seek to maximize returns and minimize risk, the problem is formulated as one of maximizing returns and minimizing risk.

2. The constraint $\sum_{i=1}^{n} s_i = K$ (2c) is a **cardinality constraint**, where K denotes the number of assets in the portfolio. The binary variable s_i (2h) indicates whether an asset i is included in the portfolio, thereby allowing investors to limit the number of assets.

3. The constraints $0 \leq y_i \leq \delta_i s_i$ (2d) and $l_i \leq \delta_i \leq u_i$ (2e), for $i = 1, \ldots, n$, are **quantity constraints**. They ensure that if an asset i is in the portfolio, its investment proportion is limited to a range between 0 and δ_i. The investor's profile determines the values of l_i and u_i.

4. The constraint $\sum_{i=1}^{n} y_i + \sum_{i=1}^{n} \gamma_i y_i = 1$ (2f) means that the entire investment fund is used for portfolio construction after deducting transaction fees.

5. The constraint on equation (2g) is called a **liquidity constraint**. The liquidity of the ith stock, denoted as λ_i and its calculated as the ratio of the average number of shares traded to the number of listed shares. Investors set the number L for this liquidity constraint

6. The constraint $y_i \geq 0$, for $i = 1, \ldots, n$ (2i), prohibits short-selling in portfolio construction.

The model (2) is a multi-objective optimization problem. To simplify it, the model is transformed into a single-objective optimization problem, as follows:

$$\min \left[-\alpha \left(\sum_{i=1}^{n} \bar{r}_i y_i - \sum_{i=1}^{n} \gamma_i y_i \right) + (1-\alpha) \left(\frac{1}{2m} \sum_{j=1}^{m} \left(\sum_{i=1}^{n} (r_{ij} - \bar{r}_i) y_i - \left| \sum_{i=1}^{n} (\bar{r}_i - r_{ij}) y_i \right| \right) \right) \right].$$

The coefficient $\alpha \in [0,1]$ is referred to as the risk-aversion coefficient, and the investor typically determines its value. However, in this study, α is derived from the predicted value $p(x)$ of a logistic regression model.

If the value of α is greater than or equal to 0.5, the market is predicted to be in a bullish trend. In contrast, if it is less than 0.5, a bearish trend is expected. For $\alpha = 0$, the probability of a market downturn is 1. Consequently, the model is equivalent to a risk-minimization strategy, indicating that the investor is highly risk-averse during a downward market trend. Conversely, when $\alpha = 1$, the probability of a market upturn is 1. The model, therefore, becomes a return-maximization strategy, which means the investor becomes more return-oriented during an upward trend.

By eliminating the absolute value function in the risk term, we can transform the initial optimization model into

$$\min \quad -\alpha R + (1-\alpha) \frac{1}{m} \sum_{j=1}^{m} u_j$$

subject to

$$\begin{aligned}
R = r^\top y - \Gamma^\top y &= \sum_{i=1}^{n} \bar{r}_i y_i - \sum_{i=1}^{n} \gamma_i y_i, \\
u_j + \sum_{i=1}^{n} (\bar{r}_i - r_{ij}) y_i &\geq 0, \quad j = 1, 2, \ldots, m, \\
u_j &\geq 0, \quad j = 1, 2, \ldots, m, \\
\sum_{i=1}^{n} s_i &= K, \\
0 \leq y_i &\leq \delta_i s_i, \quad i = 1, \ldots, n, \\
l_i \leq \delta_i &\leq u_i, \quad i = 1, \ldots, n, \\
\sum_{i=1}^{n} y_i + \sum_{i=1}^{n} \gamma_i y_i &= 1, \\
\sum_{i=1}^{n} \lambda_i y_i &\geq L,
\end{aligned}$$

$$s_i \in \{0,1\}, \quad i = 1,\ldots,n,$$
$$y_i \geq 0, \quad i = 1,\ldots,n.$$

where $r = (\bar{r}_1, \bar{r}_2, \ldots, \bar{r}_n)^\top$ represents the vector of average returns for assets over a certain period, $\Gamma = (\gamma_1, \gamma_2, \ldots, \gamma_n)^\top$ represents the vector of transaction cost for assets, R represents the net portfolio return, and $u_j = \sum_{i=1}^{n}(r_{ij} - \bar{r}_i)y_i$ is an auxiliary non-negative variable.

3.2 Study design

The assets used to analyze the portfolio model in this study are stocks, and all adjusted closed stock and LQ45 index price data were obtained from Yahoo Finance using Python software.

In general, there are two stages to solve this portfolio optimization problem:

1. **Deterministic portfolio formation.** This stage begins with selecting a target portfolio, predicting the risk-aversion coefficient, and constructing an optimum portfolio based on historical data. The final output of this stage is the proportion of funds for each stock, as well as the portfolio's return and risk.

2. **Probabilistic modeling of portfolio optimization.** This stage is conducted to provide investors with an alternative consideration in selecting their portfolio by looking at the best possible portfolios from various existing target portfolio selections.

The discussion continues with a detailed explanation of each stage and is summarized in Figure 1.

We will analyze seven portfolios in this study, which consist of stocks included in seven different indices in Indonesia: IDX30, IDXQ30 (IDX Quality 30), IDXV30 (IDX Value 30), IDXG30 (IDX Growth 30), ECONOMIC30 (IDX Cyclical Economic 30), JII (Jakarta Islamic Index), and IDXSHAGROW (IDX Sharia Growth). Each of these indices is composed of 30 stocks, with the list accessible on the Bursa Efek Indonesia website [23].

To form these seven portfolios, the risk-aversion coefficient first needs to be predicted using a logistic regression model. According to Jiang, Ji, and Chang [9], there are five critical technical indicators used as predictors to forecast the risk-aversion coefficient. However, for simplification, we only used four out of these five indicators: EMV(10) (Ease of Movement), MACD(2,3,3) (Moving Average Convergence/Divergence), RSI(10) (Relative Strength Index), and ADX(3) (Average Directional Movement Index). We can find all the formulas for these four indicators in [9].

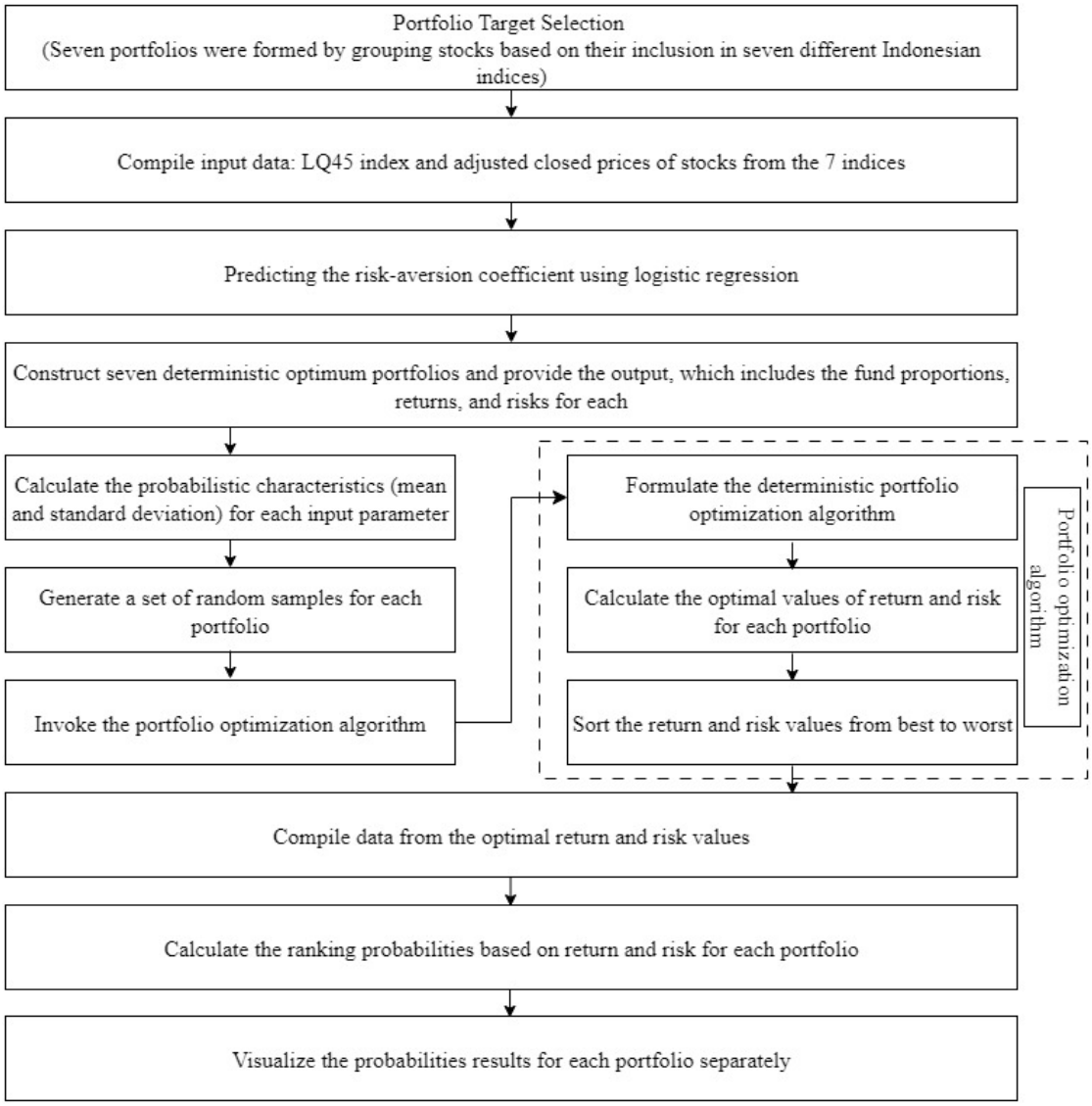


Figure 1: Scheme of the proposed algorithm.

The procedure for calculating the risk-aversion coefficient α is as follows:

1. Collect monthly data for the LQ45 index and use τ periods to build a logistic regression model to determine the risk-aversion coefficient, and a portfolio can finally be formed. In this study, we chose $\tau = 300$ months (25 years). This data is divided into training and testing data with a ratio of 4:1.

2. Based on the obtained LQ45 index values, four technical indicators are calculated and used as predictors to forecast the probability of the LQ45 index appreciating. In addition to the adjusted-closing price, we also need monthly data on the highest and lowest prices of the LQ45 index to calculate EMV(10) and ADX(3), as well as additional trading volume data to calculate EMV(10). The prediction result is then used as the risk-aversion coefficient α in the portfolio model.

The next step is to prepare the daily stock price and trading volume data from the seven indices targeted for portfolio selection. A total of 1,201 price data points are taken to obtain 1,200 daily return data points from these stocks. Meanwhile, we also need 1,200 trading volume and the number of listed stocks data for each stock to calculate their liquidity. Data on the number of listed stocks can be obtained from website IDX [23]. Then, the following steps are performed:

1. Calculate the daily return for each stock.

2. Assuming there are 20 trading days in a month then divide the 1,200 daily return data obtained in the previous step into 60 groups, with each group representing one month. Do this for each stock. Calculate the monthly return r_{ij} for every stock by summing the daily returns in each month. Therefore, each stock has 60 monthly return values.

3. Calculate the expected monthly return of the ith stock (i.e., $\bar{r}_i$) by averaging the monthly returns obtained in the previous step. Do this for all stocks.

4. With the same assumptions in the return calculation (there are 20 trading days in 1 month), we divide the 1,200 trading volume data into 60 groups with each group representing one month. Calculate the average monthly trading volume for every stock so that we can obtain the liquidity of each stock (λ_i) by calculating the ratio of average monthly trading volume to the volume of listed stock.

5. Solve the optimization model to form the seven portfolios discussed, using the GEKKO package in Python. This package already includes the Branch & Bound subroutine to solve the Mixed-Integer Linear Programming (MILP) problem, which is our portfolio optimization model. Given that seven indices are being discussed, the constructed portfolios are named A-G, where Portfolio A consists of stocks from the IDX30 index, Portfolio B consists of stocks from IDXQ30 index, and so on.

The output of this process is the proportion of funds, return, and risk for all portfolios. This output is used as a reference for investors to choose a portfolio that fits their profile. To achieve

this goal, investors need to see the possibilities of the best portfolios from the seven existing portfolios. For this purpose, a probabilistic approach is taken using Monte Carlo simulation, and it is assumed that the daily return of each stock is a normally distributed random variable, as stated by Shadabfar and Cheng [15]. The stages are explained as follows:

1. From the historical data, calculate the mean (α_{ij}^k) and standard deviation (β_{ij}^k) of the daily return for every 20 days for each stock. Here, i denotes the ith stock index ($i \in \{A, B, C, D, E, F, G\}$ since seven indices are used in this study), j denotes the jth stock within an index ($j \in \{1, 2, \ldots, 30\}$ since each index consists of 30 stocks), and k is the kth month ($k \in \{1, 2, \ldots, 60\}$ since the data was taken over 60 months or 5 years). For example: α_{Aj}^1 means the average daily return in the 1st month (remember that each month is assumed to have 20 days) for stock j in stock index A.

2. Perform the following steps to form seven portfolios:

   (a) For $k \in \{1, 2, \ldots, 60\}$, generate 20 random numbers each (these random variables represent the daily return of a stock) from a normal distribution $N(\alpha_{ij}^k, \beta_{ij}^k)$. Thus, each stock has 1200 random numbers representing daily returns over five years.

   (b) Calculate the monthly return for each stock by summing the daily returns per month (each month has 20 days), so for each stock, there are 60 monthly return values.

   (c) Calculate the expected monthly return for each stock by averaging the monthly returns from step b).

   (d) Solve the portfolio optimization model to obtain the return and risk simultaneously. (This is the same as in the deterministic portfolio formation, meaning the same risk-aversion coefficient is used.) The output required at this stage is the return and risk of the seven portfolios formed.

3. Rank the seven portfolios based on the return obtained in end of step 2, so that we have a rank of 1 to 7 for the existing portfolio based on the return.

4. Rank the seven portfolios based on the risk obtained in end of step 2, so that we have a rank of 1 to 7 for the existing portfolio based on the risk.

5. Repeat steps 2 through 4 a total of N times (in this study, we choose $N = 10,000$). As a result, for each portfolio, the number of rankings from 1 to 7 is obtained, for both return and risk. Let these be denoted as n_{il} and m_{il}, where $l = 1, 2, \ldots, 7$. For instance, n_{F1} refers to the number of rank 1 outcomes for the return of Portfolio F over N simulation runs.

6. Calculate the respective probabilities of the rankings obtained in step 5, for both return and risk, using the formula

$$p_{il} = \frac{n_{il}}{N} \quad \text{and} \quad q_{il} = \frac{m_{il}}{N}.$$

We have visualize these results in the form of bar and pie charts. After that, the analysis is continued to see the convergence of the Monte Carlo simulation by creating a graph of the probability of return and risk for a specific portfolio and a specific ranking against the number of simulations N. In this study, Portfolio C for ranking five was chosen, and we ran N from 1 to 10,000.

Finally, the discussion is closed by evaluating the robustness of the proposed algorithm. Robustness evaluation is essential in Monte Carlo simulation because the accuracy of the results critically depends on the assumed probabilistic characteristics of the input parameters, which are often estimated with uncertainty. By evaluating the stability of the simulation outcomes under small perturbations in the input parameters, robustness evaluation ensures that the conclusions drawn from Monte Carlo simulations are not overly sensitive to estimation errors and remain reliable for decision-making. For this purpose, a normal distribution with an absolute coefficient of variation (ACV) is employed. The ACV is adopted because the mean returns of assets may take negative values. The ACV is defined as follows [2]:

$$ACV = \frac{\sigma}{|\mu|},$$

where, μ and σ are mean and standard deviation of the data, respectively. The mean of the normal distribution is assumed to be a scaled version of the mean of the initial random variable, that is

$$\mu_{f_k} = \phi \mu_k,$$

where μ_k denotes the mean of the initial random variable; ϕ represents the noise contribution factor; and μ_{f_k} is the mean of the normal distribution associated with the kth perturbation ($k = 1, 2, \ldots, N_{\text{pert}}$) and N_{pert} is the number of simulations used for robustness evaluation. The proposed algorithm is then applied to compute the optimal probabilistic solutions for the seven portfolios under noise inputs. Specifically, the probability ranking procedure for the seven portfolios is rerun using perturbed inputs, where each random number originally generated from the normal distribution $N(\alpha_{ij}^k, \beta_{ij}^k)$ is augmented by an additive noise term drawn from the normal distribution $N(\mu_{f_k}, \sigma_{f_k})$ with $\sigma_{f_k} = ACV \times |\mu_{f_k}|$.

4 Results and discussion

4.1 Portfolio analysis

To predict the direction of market movement, we used monthly price data for the LQ45 index from July 31, 2000, to July 31, 2025, covering 300 months. The logistic regression model predicted a probability of 0.9219 with an accuracy of 82.7%. Subsequently, this predicted probability was used as the risk-aversion coefficient α. Since the value is closer to 1, the market is expected to be bullish, leading investors to prioritize returns over risk in portfolio construction.

Seven portfolios were constructed using stocks from various indexes: IDX30, IDXQ30, IDXV30, IDXG30, ECONOMIC30, JII, and IDXSHAGROW. Each index consists of 30 stocks. The data consisted of adjusted closing prices collected over five years from July 29, 2020, to July 31, 2025, totalling 1201 data points, assuming 20 trading days per month. Stocks of companies that had their Initial Public Offering (IPO) after July 29, 2020, were excluded to ensure a complete dataset of 1201 data points. The specific stocks excluded were GOTO, MBMA, ADMR, AVIA, CMRY, TPAG, AMMN, and PGEO.

Five key parameters were established for portfolio formation and simulation:

1. The cardinality constraint was set at $K = 8$,
2. The transaction cost of every stock is the same, that is, 0.0015,
3. The number L of liquidity constraint is 0.05,
4. The lower and upper bounds for the quantity constraint for each stock i were $l_i = 0.025$ and $u_i = 0.5$, respectively, and
5. The number of simulations was 10,000.

The time consumption to run this simulation for seven portfolios is around 4.7 hours. The convergence rate of the standard Monte Carlo estimator was evaluated by analyzing the decay of the estimation error with respect to the sample size. This error decreases proportionally to number of simulations N, that is, $O(N^{-1/2})$ [5].

Table 1 compares the returns and risks of the equal-weighted $\frac{1}{n}$ portfolios (which are constructed by allocating the same weight to all stocks that fall within the index) with an alternative solution modeled in this paper. The findings suggest that the equal-weighted $\frac{1}{n}$ portfolios display strikingly low risk. This conclusion correlates to a great diversification, since the exposure of each asset is fixed and so the risk is minimized. But the significantly low risk of the equal-weighted

$\frac{1}{n}$ approach comes with small returns. This kind of return profile may not appeal to investors looking for higher returns, particularly in cases where moderate risk tolerance is considered appropriate. In contrast, the portfolios constructed using the proposed model demonstrate a significant improvement in returns across all portfolio configurations. Although this improvement in return is accompanied by an increase in risk, the trade-off remains reasonable and consistent with portfolio theory, which posits that higher expected returns generally require higher risk exposure. Furthermore, the proposed portfolios also have one additional benefit: There are just eight stocks in each portfolio.

The smaller number of assets implies lower implementation complexity and transaction costs than the equal-weighted $\frac{1}{n}$ strategy that requires investors to allocate capital over the index of stocks they own.

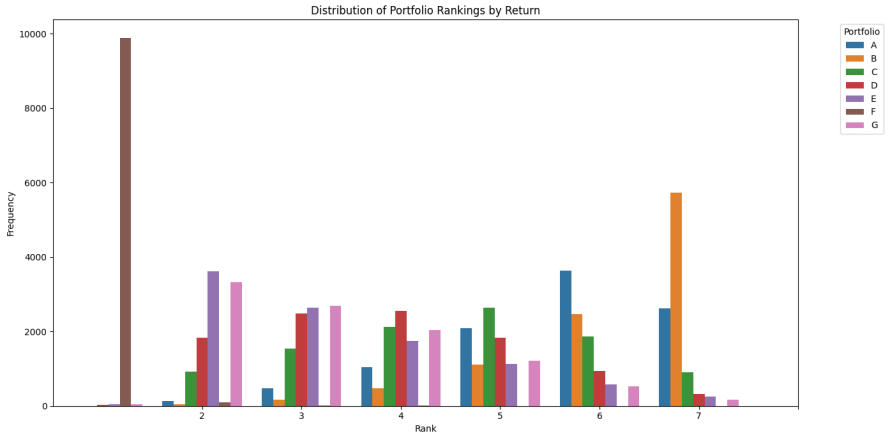


Figure 2: Distribution of return rankings (1-7) for all portfolios. Rank 1 indicates the highest return, while Rank 7 indicates the lowest return.

The overall results indicate that the proposed model strikes a better risk, return, practical feasibility alignment, thus an attractive substitute for those investors seeking improved performance without an unnecessarily difficult diversification process. From this table we can also infer that Portfolio F offers the highest return at 10.67% compared to other portfolios and the highest risk of 7.00%. However, such an investor might have additional considerations to make investment decisions following the approach which was discussed in Subsection 3.2.

Figures 2 and 3 illustrate the distribution of return and risk ranks (1 to 7) for all portfolios. From these two figures, we can see the frequency of a ranking of a portfolio for 10,000 simulations. For example, Portfolio F is the best scenario for returns and the worst for risk because the frequency of rank 1 for returns of this portfolio is the highest compared to other portfolios. Similarly, the frequency of rank 7 (which means the highest risk) is also the highest. Tables 2

Table 1: Summary of Stock Portfolios Performance and Composition

Portfolio A		**Portfolio B**		**Portfolio C**		**Portfolio D**		**Portfolio E**		**Portfolio F**		**Portfolio G**	
Stock Code	Prop.	Stock Code	Prop.	Stock Code	Prop.	Stock Code	Prop.	Stock Code	Prop.	Stock Code	Prop.	Stock Code	Prop.
ADRO	46.9%	ADRO	50%	ADRO	2.5%	BRIS	2.5%	ANTM	2.5%	ADRO	2.5%	ADRO	2.5%
AKRA	2.5%	AKRA	2.5%	AUTO	2.5%	ENRG	50%	BRIS	2.5%	AKRA	2.5%	BRIS	2.5%
AMRT	2.5%	AMRT	2.5%	ELSA	2.5%	ESSA	2.5%	BRMS	34.9%	ANTM	22.2%	DSNG	2.5%
ANTM	18.6%	BFIN	2.5%	ENRG	50%	ISAT	2.5%	BRPT	2.5%	BRMS	2.5%	ISAT	2.5%
BRPT	2.5%	BMRI	4.7%	HRUM	2.5%	ITMG	2.5%	ESSA	2.5%	MEDC	2.5%	ITMG	2.5%
ISAT	21.8%	BNGA	2.5%	ITMG	34.9%	MAPA	2.5%	FILM	50%	PANI	50%	MEDC	2.5%
MEDC	2.5%	BRIS	32.6%	MEDC	2.5%	MIDI	2.5%	MAPA	2.5%	PGAS	2.5%	RAJA	50%
PTBA	2.5%	PTBA	2.5%	SRTG	2.5%	SSIA	34.9%	PTBA	2.5%	PTRO	15.2%	SSIA	34.9%
Return: 3.42%		Return: 3.40%		Return: 4.56%		Return: 4.86%		Return: 5.55%		Return: 10.67%		Return: 5.60%	
Risk: 3.60%		Risk: 3.71%		Risk: 4.83%		Risk: 4.80%		Risk: 6.13%		Risk: 7.00%		Risk: 5.08%	
Equal-weighted $\frac{1}{n}$													
Portfolio A		**Portfolio B**		**Portfolio C**		**Portfolio D**		**Portfolio E**		**Portfolio F**		**Portfolio G**	
Return: 1.27%		Return: 1.01%		Return: 1.35%		Return: 1.88%		Return: 1.60%		Return: 2.37%		Return: 2.06%	
Risk: 0.41%		Risk: 0.37%		Risk: 0.46%		Risk: 0.38%		Risk: 0.45%		Risk: 0.44%		Risk: 0.44%	

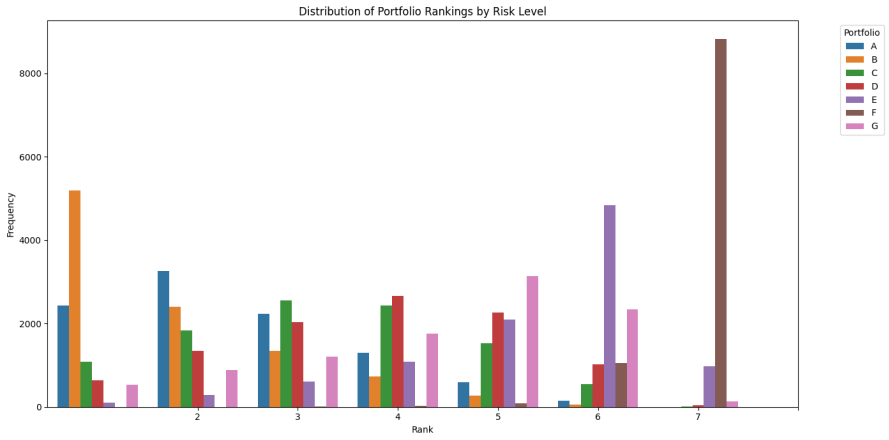


Figure 3: Distribution of risk rankings (1-7) for all portfolios. Rank 1 indicates the lowest risk, while Rank 7 indicates the highest risk.

and 3 show the estimated probabilities of return and risk, respectively. The results indicate that Portfolio F offers the highest probability (0.9884) of the best return but also presents the worst risk to investors with probability 0.8822. Figure 4 is a pie chart that visualizes the estimated probability of return and risk for each portfolio. From the two tables and the figure, we have more information about the seven portfolios. For instance, although the probability of achieving relatively low risk for Portfolio B is the highest (51.98%), but the probability of achieving the return is the lowest (0). Such information is also provide for all other portfolio ranks.

Table 2: Estimated probabilities of return for all portfolios, ranked from 1 (highest return) to 7 (lowest return).

Portfolio	**Return Rank**						
	1	**2**	**3**	**4**	**5**	**6**	**7**
A	0	0.0131	0.0474	0.1051	0.2094	0.3634	0.2616
B	0	0.0052	0.0173	0.0472	0.1104	0.247	0.5729
C	0	0.0919	0.1537	0.2129	0.2639	0.1863	0.0913
D	0.0022	0.1839	0.2489	0.2562	0.1827	0.0934	0.0327
E	0.0051	0.3622	0.2633	0.1739	0.1126	0.058	0.0249
F	0.9884	0.0103	0.0009	0.0004	0	0	0
G	0.0043	0.3334	0.2685	0.2043	0.121	0.0519	0.0166

Table 3: Estimated probabilities of risk for all portfolios, ranked from 1 (lowest risk) to 7 (highest risk).

Portfolio	**Risk Rank**						
	1	**2**	**3**	**4**	**5**	**6**	**7**
A	0.2439	0.3264	0.2241	0.1305	0.06	0.0148	0.0003
B	0.5198	0.2398	0.1345	0.0732	0.0277	0.005	0
C	0.1089	0.1828	0.2558	0.2431	0.153	0.0549	0.0015
D	0.0635	0.1343	0.2032	0.2665	0.2257	0.1026	0.0042
E	0.0105	0.0285	0.0612	0.1085	0.2102	0.4834	0.0977
F	0.0001	0.0002	0.0004	0.0021	0.0093	0.1057	0.8822
G	0.0533	0.088	0.1208	0.1761	0.3141	0.2336	0.0141

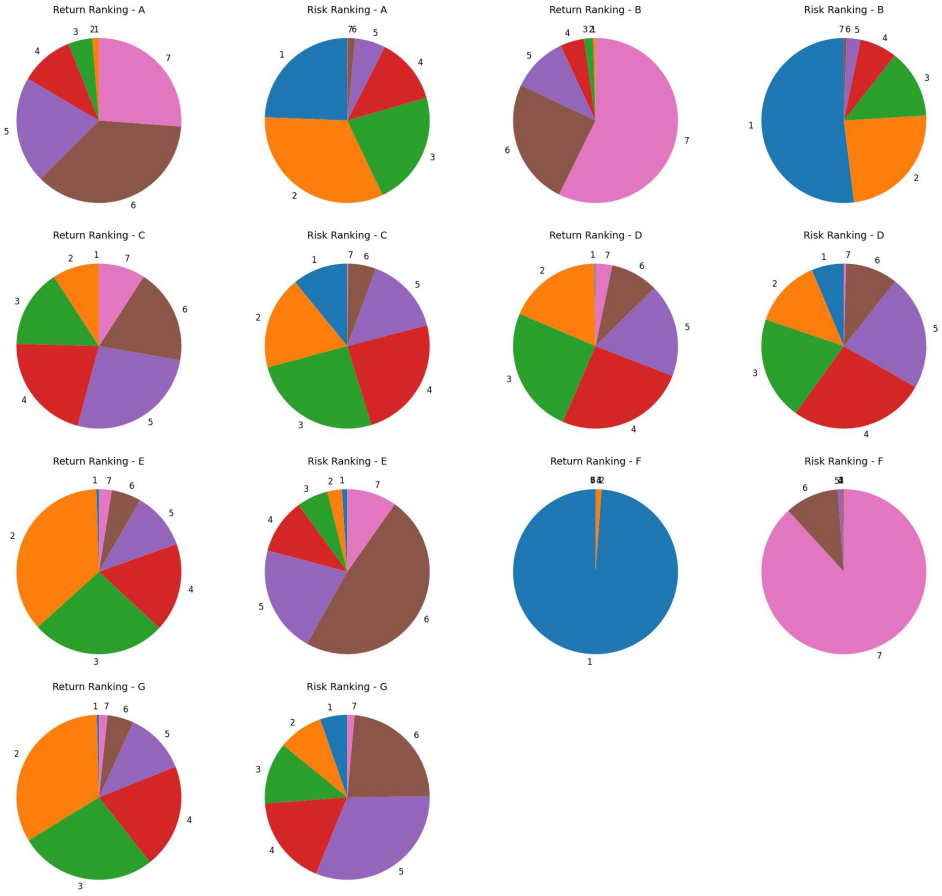


Figure 4: Estimated probability of return and risk for each portfolio, illustrating the distribution of return and risk rankings from 1 to 7.

4.2 Convergence and robustness evaluation

To examine the convergence of the Monte Carlo method in this portfolio optimization problem's probabilistic approach, Figures 5 and 6 show the graphs of return and risk probability values, respectively, against the number of random samples (simulations) for Portfolio C with a rank of 5. The blue curve represents the probability values for Portfolio C's return and risk for 1 to 10,000 simulations, while the dashed red line indicates the probability values for 10,000 simulations. At small random sample sizes, there are significant fluctuations in probability values. However, as the random sample size increases, these fluctuations decrease. Finally, after using approximately 6,000 random samples, the probabilities converge, and the fluctuations become increasingly small and can be ignored. Based on this result, all 10,000 simulations in the previous experiment are considered sufficient.

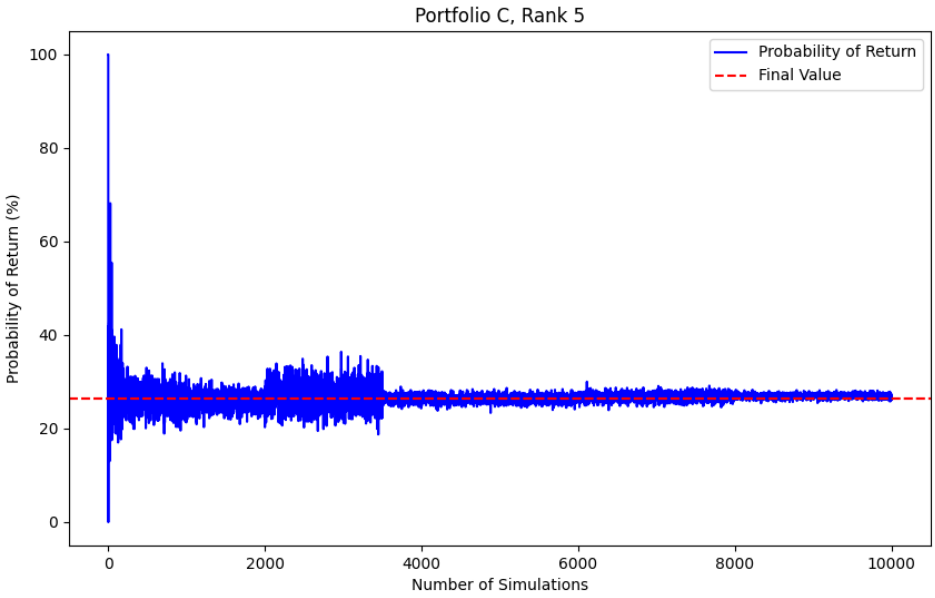


Figure 5: Convergence of the estimated probability of return against the number of simulations for portfolio C (Rank 5).

Besides convergence, the next important thing needed in Monte Carlo simulation is robustness evaluation. This evaluation is essential in Monte Carlo simulation because the accuracy of the results critically depends on the assumed probabilistic characteristics of the input parameters, which are often estimated with uncertainty. By evaluating the stability of the simulation outcomes under small perturbations in the input parameters, robustness analysis ensures that

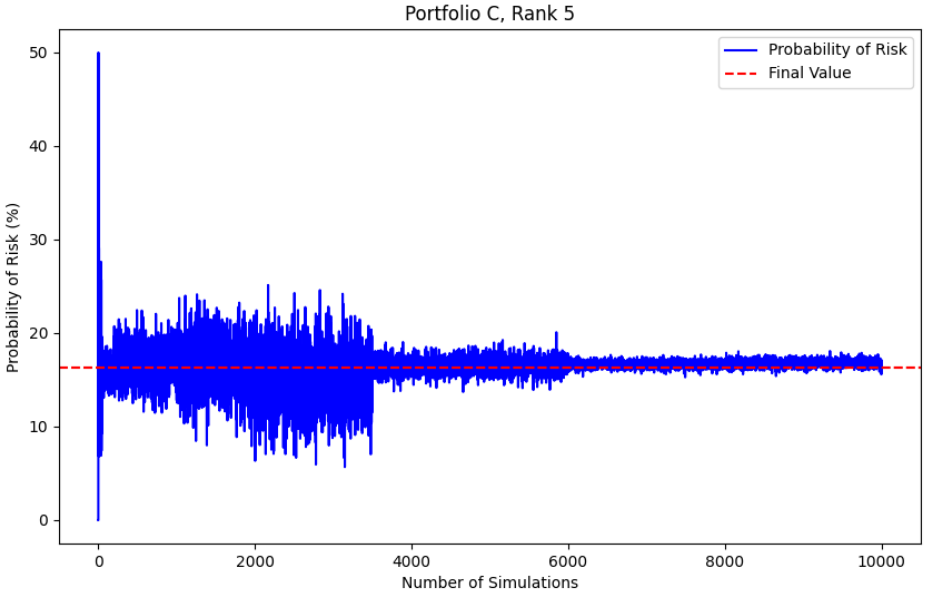


Figure 6: Convergence of the estimated probability of return against the number of simulations for portfolio C (Rank 5).

the conclusions drawn from Monte Carlo simulations are reliable for decision-making. For robustness evaluation, we use $N_{\text{pert}} = 100$. Figure 7 illustrates the box-plots of the return and risk probabilities for Portfolio C with rank 5, respectively. From Figure 7(a), the median value remains stable at approximately 0.261. The interquartile range is relatively narrow, indicating that more than 50% of the simulated probabilities are concentrated within a small interval. This limited dispersion suggests that the estimated return probability exhibits low sensitivity to input perturbations. The presence of only two outliers further indicates that extreme return realizations are rare and do not materially influence the overall distribution. Similarly in Figure 7(b), the median is approximately 0.158, and the interquartile range remains small compared to the overall range of the data, implying that the variability of risk probability is well controlled under input uncertainty. Although several upper outliers are observed, their occurrence is infrequent and does not reflect systematic instability in the risk estimation. Thus, our proposed algorithm is quite robust in estimating the probability of return and risk.

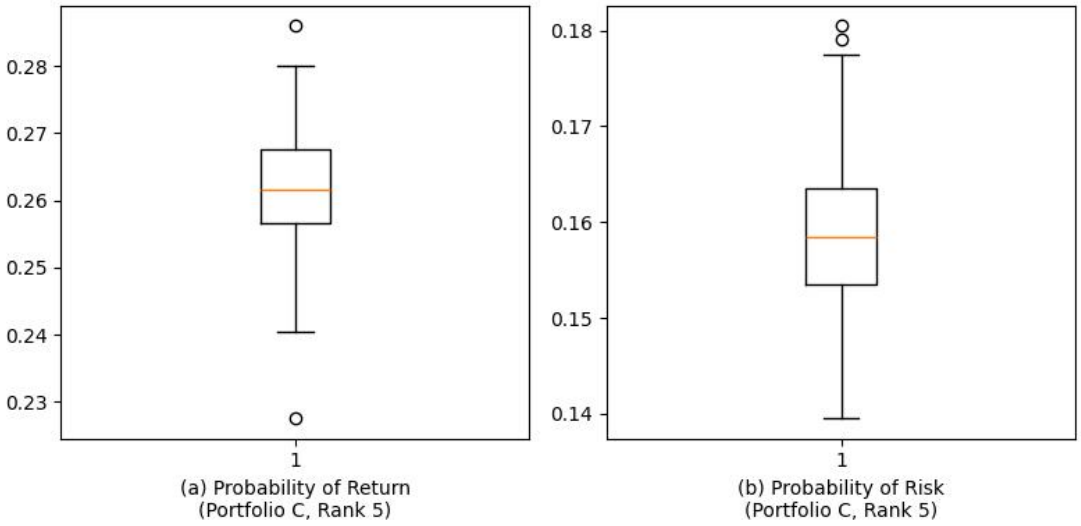


Figure 7: Box-plots of the simulated probabilities for (a) return and (b) risk of Portfolio C with rank 5 under input perturbations.

5 Conclusions

The algorithm proposed in this paper provides an alternative insight for investors with a probabilistic approach to selecting portfolios, including constraints on the number of assets, the quantity and liquidity of each asset, and also transaction cost within the portfolio. Based on five years of data collected between July 29, 2020, and July 31, 2025, with a total of 1,201 data points, investors can not only estimate their expected returns and risks, but also view the probability of an optimal portfolio that is appropriate by considering market trend conditions through the risk-aversion coefficient. A Monte Carlo simulation was used 10,000 times to determine these probabilities, as the convergence graph showed that the simulation had already converged after 6,000 iterations. Eventually, the absence of prominent distributional shifts and the small number of outliers confirm that the proposed portfolio model exhibits satisfactory robustness against small perturbations in the input parameters.

Acknowledgements

The authors gratefully acknowledge the financial support provided by the ”Lembaga Penelitian dan Pengabdian kepada Masyarakat” (LPPM), Parahyangan Catholic University, for this study

(grant number III/LPPM/2025-02/74-P). Authors are also grateful to their anonymous referees and editor for their constructive comments.

Funding

This research received no external funding.

Conflicts of Interest

The authors declare no conflicts of interest.

References

[1] Anagnostopoulos, K.P. and Mamanis, G. *The mean-variance cardinality constrained portfolio optimization problem: An experimental evaluation of five multiobjective evolutionary algorithms*, Expert Syst. Appl. 38 (2011), 14208–14217.

[2] Bindu, K.H, Mukkamala, R., Dey, N. and Rao, C. R. *Coefficient of variation and machine learning applications*, CRC Press, 2019.

[3] Cesarone, F., Scozzari, A. and Tardella, F. *A new method for mean-variance portfolio optimization with cardinality constraints*, Ann. Oper. Res. 205 (2013), 213–234.

[4] Fang, Y., Lai, K. K. and Wang, S.Y. *Portfolio rebalancing model with transaction costs based on fuzzy decision theory*, Eur. J. Oper. Res. 175(2) (2006), 879–893. https://doi.org/10.1016/j.ejor.2005.05.020

[5] Glasserman, P. *Monte Carlo methods in Financial Engineering*, Springer. (2003), 1–3.

[6] Grant, A., Kwon, O. K. and Satchell, S. *Properties of risk aversion estimated from portfolio weights*, J. Asset Manag. 25 (2024), 427–444.

[7] James, G., Witten, D., Hastie, T., Tibshirani, R. and Taylor, J. *An introduction to statistical learning with applications in Python*, Springer. (2023), 138–144.

[8] Ji, R., Lejeune, M.A. and Prasad, S.Y. *Properties, formulations, and algorithms for portfolio optimization using Mean-Gini criteria*, Ann. Oper. Res. 248 (2017), 305–343.

[9] Jiang, Z., Ji, R. and Chang, K. C. *A machine learning integrated portfolio rebalance framework with risk-aversion adjustment*, J. Risk Financial Manag. 13(7) (2020), 155.

[10] Konno, H. and Wijayanayake, A. *Mean-absolute deviation portfolio optimization model under transaction costs*, J. Oper. Res. Soc. Jpn. 42(4) (1999), 422–435.

[11] Konno, H. and Yamazaki, H. *Mean-absolute deviation portfolio optimization model and its applications to Tokyo Stock Market*, Manag. Sci. 37(5) (1991), 519–531.

[12] Lee, J. *A first course in combinatorial optimization*, Cambridge University Press: New York, United States of America. (2004), 177–180.

[13] Markowitz, H. *Portfolio selection*, J. Finance. 7(1) (1952), 77-91.

[14] Mousavi, A. and Michailidis, G. *Cardinality constrained mean-variance portfolios: a penalty decomposition algorithm*, Comput. Optim. Appl. 90 (2025), 631–648.

[15] Shadabfar, M. and Cheng, L. *Probabilistic approach for optimal portfolio selection using a hybrid Monte Carlo simulation and Markowitz model*, Alex. Eng. J. 59 (2020), 3381–3393.

[16] Shadabfar, M. and Wang, Y. *Approximation of the Monte Carlo sampling method for reliability analysis of structures*, Math. Probl. Eng. 2016(1) (2016), 5726565.

[17] Simaan, Y. *Estimation risk in portfolio selection: The mean variance model versus the mean absolute deviation model*, Manag. Sci. 43(10) (1997), 1437–1446.

[18] Sortino, F. A. and Meer, R.v.d. *Downside risk: Capturing what's at stake in investment situations*, J. Portf. Manag. 17(4) (1991), 27–31.

[19] Xiaoxia, H. *Portfolio analysis: from probabilistic to credibilistic and uncertain approaches*, Springer. (2010), 11–60.

[20] Yaman, I. and Dalkılıç, T. E. *A hybrid approach to cardinality constraint portfolio selection problem based on nonlinear neural network and genetic algorithm*, Expert Syst. Appl. 169 (2021), 114517, 294–305.

[21] Yixuan, L. and Zhongfeng, Q. *Mean semi-absolute deviation for uncertain portfolio optimization problem*, J. Uncertain Syst. 6(4) (2021), 299–307.

[22] Yufei, S., Grace, A., Lay, T. K., Guanglu, Z. *Portfolio optimization using a new probabilistic risk measure*, J. Ind. Manag. Optim. 11(4) (2015), 1275–1283.

[23] https://www.idx.co.id/id/data-pasar/data-saham/indeks-saham, last accessed July 24, 2025.

Iranian Journal of Numerical Analysis and Optimization
Vol. 16, No. 3, 2026, pp 876–921
https://doi.org/10.22067/ijnao.2026.96568.1780
https://ijnao.um.ac.ir/

CC BY NC

Research Article

Tension spline-quasilinearization scheme for generalized nonlinear time-fractional convection-diffusion equations

A.S. Bekela*, L.G. Enyadene and M.M. Woldaregay

Abstract

This article presents an efficient numerical technique based on a nonpolynomial tension spline for solving the generalized nonlinear time-fractional convection-diffusion equation with polynomial nonlinear convection terms. The proposed method uses nonpolynomial tension spline spatial discretization with the Crank–Nicolson time-stepping scheme, and employs the quasilinearization technique to handle the nonlinear term $p^v \frac{\partial p}{\partial x}$ for $v = 1, 2, \ldots, 10$. The tension spline introduces a controllable smoothness parameter, which can improve accuracy and reduce spurious oscillations. Additionally, the order of convergence for the suggested scheme is determined by analyzing the truncation error, and we theoretically present the stability of

*Corresponding author

Received 18 November 2025; revised 6 May 2026; accepted 9 May 2026

Alemu Senbeta Bekela
Department of Applied Mathematics, Adama Science and Technology University, Adama, Oromia, Ethiopia.
e-mail: alemusenbeta1@su.edu.et

Lemi Guta Enyadene
Department of Applied Mathematics, Adama Science and Technology University, Adama, Oromia, Ethiopia.
e-mail: senalemi2007@gmail.com

Mesfin Mekuria Woldaregay
Department of Applied Mathematics, Adama Science and Technology University, Adama, Oromia, Ethiopia.
msfnmkr02@gmail.com

How to cite this article
Bekela, A.S. Enyadene, L.G. and Woldaregay, M.M., Tension spline-quasilinearization scheme for generalized nonlinear time-fractional convection-diffusion equations. *Iran. J. Numer. Anal. Optim.*, 2026; 16(3): 876-921.
https://doi.org/10.22067/ijnao.2026.96568.1780

the developed scheme using the von Neumann method. The method is shown to be conditionally stable. The method is validated on two representative test problems, showing good agreement with expected solutions and improved accuracy compared with existing numerical methods. These results illustrate the potential of the approach for solving nonlinear time-fractional convection-diffusion problems.

AMS subject classifications (2020): 35R11, 65M06, 65M12, 65D07, 65H10.

Keywords: Caputo's fractional derivative, Crank-Nicolson scheme, Nonlinear time-fractional convection-diffusion equation, Quasi-linearization, Nonpolynomial tension spline.

1 Introduction

Fractional calculus represents a generalization of ordinary and partial differentiation and integration to arbitrary noninteger orders, thereby unifying and extending the concepts of integer-order differentiation and integration (for the history and development of this field, see, for example, [37]). Compared with classical calculus, fractional derivatives and integrals of arbitrary order provide more accurate representations of real-world phenomena [19]. Only at integer orders $(1, 2, 3, \ldots)$ can classical derivatives give information; they are unable to characterize behaviors that correspond to fractional orders $(0, 1], (1, 2], (2, 3]$ [27, 33]. On the other hand, the fractional-order derivative operator exhibits nonlocal behavior because the derivative at a given point depends on the entire history of the function, not merely its local values [5]. Unlike integer-order derivatives [13], this property improves the modeling performance of systems with memory and hereditary effects [33]. Such characteristics show that fractional derivatives are more versatile in capturing the history and hereditary properties of different dynamical systems and in modeling them [42, 1, 5].

There are numerous definitions of the fractional derivative presented in the literature (e.g., Riemann–liouville, Caputo, Weyl, Grünwald–Letnikov, and Riesz forms) [35]. Multiple definitions help guide researchers in selecting the most convenient one for obtaining the best solutions. The two most frequently utilized are the Riemann–Liouville and Caputo derivatives, which differ mainly in how integration is applied before differentiation. In order to overcome the challenge of determining initial values for the Riemann–Liouville derivative in Laplace transform applications, the Italian mathematician Caputo first proposed the Caputo fractional derivative in the 1960s [37]. The primary benefit of the Caputo approach is that it uses the same initial conditions as traditional (integer-order) differential equations. In recent studies, the Caputo derivatives have

been favored over the Riemann–Liouville derivative for obtaining both exact and approximate solutions of fractional differential equations (FDEs) [35].

Solving FDEs has become essential because of their many applications and their ability to help in understanding complex real-world systems. The time-fractional convection-diffusion equation is a significant type of partial FDE, which is the focus of this investigation. It is challenging to obtain explicit analytic solutions of fractional partial differential equations (FPDEs) because of their peculiar characteristics, which include violations of the semigroup property, Leibniz rule, and chain rule [39]. As a result, FPDEs have attracted considerable interest in approximate analytical and numerical approaches. The perturbation analysis method and differential transform method [4], integral-type variational iterative method [6], and Yang transform decomposition method [7] are examples of approximate analytical methods that have been applied to solve the fractional convection-diffusion equation. One disadvantage of these techniques is that they only take into account the initial condition and overlook the boundary conditions of the fractional convection-diffusion model. Boundary conditions, however, are crucial for describing and simulating real-world processes [41]. Therefore, developing a numerical approach for solving fractional convection-diffusion models with suitable initial and boundary conditions is important.

Nonlinear FPDEs (NFPDEs) play a crucial role in understanding and modeling nonlinear behaviors that frequently occur in real-world physical and engineering systems [13, 33]. NFPDEs have recently been successfully used to apply a large number of mathematical models in mathematical biology, electrodynamics, control theory, fluid mechanics, analytical chemistry, and other fields [34, 7, 13, 25]. For studying real-time events like earthquake propagation, volcanic eruption, diffusion process, cancer models, gene models, soil dynamics in water, cosmology, and seismology, NFPDE are a major option [28, 15]. There are many nonlinear time FPDEs (NTFPDEs). For instance, the nonlinear time-fractional convection-diffusion equation (NTFCDE) [1], nonlinear time-fractional Korteweg-de Vries equation [31], nonlinear telegraph equation [16, 17], time-fractional Burgers equation, nonlinear time-fractional Klein-Gordon equation [29, 18], nonlinear time-fractional sine-Gordon equation [23], nonlinear time-fractional Schrödinger equation [43], nonlinear time-fractional reaction-diffusion equation [9, 2], and nonlinear time-fractional convection-reaction-diffusion equation [38] are the fractional versions of NTFPDEs. In this paper, we focus on the NTFCDE.

The NTFCDE, referred to as the time-fractional Burgers' equation, is one of the most important NFPDEs arising in applied sciences and mathematical physics, and it has interesting applications in physics and astrophysics [6, 13]. NTFCDEs are used in the analyze fluid mechanics modeling, diffusive wave modeling in fluid dynamics, shock wave theory, turbulent fluid mathematical modeling, sound waves in a viscous medium, and many other areas [41]. For the

same reasons, recent studies have focused on examining and developing models for time-fractional Burgers (convection-diffusion) equations. Several techniques have been developed to obtain approximate or numerical solutions because many NFPDEs, including NTFCDEs, cannot be solved analytically. In this section of our manuscript, we briefly review some of them. Esen and Tasbozan [21] used the quadratic B-spline Galerkin method, which is a practical and effective finite element approach, to investigate numerical solutions of the equation. In [22], the authors obtained more accurate numerical solutions by using collocation points and increasing the degree of basis. A Shehu transform variational iteration method for solving NTFCDEs with proportional delay has been proposed by Bekela, Belachew, and Wole [6]. To obtain approximate solutions of the NTFCDE, a numerical method based on the Crank–Nicolson finite difference scheme was proposed in [35]. The formable transform Adomian decomposition method was used by Bekela and Woldaregay [10] to solve nonlinear time-fractional diffusion equations.

The NTFCDE has been discussed in the literature using a number of numerical and semi-analytic approaches. To mention a few, Li, Zhang, and Ran [32] suggested a linearized implicit finite difference scheme for the generalized NTFCDEs. The Haar wavelet scheme and cubic B-spline finite element method were introduced by Esen, Bulut, and Oruç [20]. In recent years, several numerical methods have been studied for solving the NTFCDEs, including the extended cubic B-spline method [3], the Chebyshev wavelet method [36], and combination of cubic B-splines and the differential quadrature method [26] include some of these methods. The anomalous diffusion and wave propagation models were studied in [8]. The generalized NTFCDEs or generalized time-fractional Burgers' equations, have received considerable attention in recent years and have been studied by a number of authors. Readers are referred to [13, 33, 12] and the references therein.

In this article, we consider the following generalized NTFCDE:

$$ {}_0^C D_t^\alpha p(x,t) + (p(x,t))^v \frac{\partial p(x,t)}{\partial x} = d\frac{\partial^2 p(x,t)}{\partial x^2} + f(x,t), (x,t) \in (0,L) \times (0,T), \tag{1} $$

with the initial condition

$$ p(x,0) = g(x), \quad 0 \le x \le L, \tag{2} $$

and the boundary conditions

$$ p(0,t) = g_0(t), \quad p(L,t) = g_1(t), \quad 0 \le t \le T, \tag{3} $$

where d is the diffusion coefficient parameter and v is a positive integer. Also, ${}_0^C D_t^\alpha$ is the Caputo fractional derivative of order α, where $(0 < \alpha \leq 1)$ with respect to t. The functions $g(x), g_0(t), g_1(t)$ and $f(x,t)$ are assumed to be sufficiently smooth in their corresponding domains. There are many interpretations of the concept of fractional-order derivatives, but in this paper we use the sense of Caputo; specifically,

$$ {}_0^C D_t^\alpha p(x,t) = \frac{1}{\Gamma(1-\alpha)} \int_0^t \frac{\partial p(x,s)}{\partial s} \frac{ds}{(t-s)^\alpha}, \qquad 0 < \alpha \leq 1. \tag{4} $$

In order to solve the generalized NTFCDE, this paper proposes a conditionally stable and convergent numerical method based on a nonpolynomial tension spline. First, a sequence of linear time-fractional convection-diffusion is obtained using the quasi-linearization technique [11]. Readers are referred to [40] and [24] for additional work addressing the nonlinear term that appears in the equation using the quasi-linearization technique. Then, the time-fractional derivative is discretized using the Crank–Nicolson difference scheme. The spatial derivatives are further discretized using the nonpolynomial tension spline method, which has not been explored previously for NTFCDEs. Furthermore, some existing numerical approaches for solving NTFCDEs have a low accuracy [13, 21, 20, 35, 25, 44], and we compare the present method with results reported in those references through numerical implementation. Therefore, we have developed a hybrid numerical scheme that combines the quasi-linearization technique, the Crank–Nicolson method, and the nonpolynomial tension spline approach to achieve higher accuracy. The von Neumann analysis shows that the linearized TFPDE is conditionally stable.

The article's remaining content is arranged as follows: The linearization technique called the quasi-linearization process and its approximation are presented in Section 2. Section 3 describes the application of the nonpolynomial tension spline approach in the spatial direction and the Crank–Nicolson scheme in the temporal direction. Section 4 provides a thorough demonstration of the stability and convergence of the proposed scheme. This section also computes the local truncation error of the approximation. Numerical results are presented in Section 5 to support the theoretical findings. A brief discussion of the concluding remarks in Section 6 completes the paper.

2 Linearization technique

In this section, we linearize the generalized NTFCDE in (1) to obtain a sequence of linear time-fractional convection-diffusion equations. Then, these equations are further discretized using a nonpolynomial tension spline based on the Crank–Nicolson scheme.

2.1 Quasi-linearization technique

The method of quasi-linearization linearizes the nonlinear term in the PDE by expanding it about the solution from of the previous step in a sequence of linear TFCDEs that quadratically converge to the solution of the original nonlinear problem. Readers are referred to [11, 40] for additional information regarding the quasi-linearization process and its convergence. For the initial guess p^0, the linearization process yields a sequence $\{p^n\}$ that satisfies the problem's initial condition (2) and boundary conditions (3). We begin by defining the nonlinear operator in (1):

$$\mathcal{N}\left(p, \frac{\partial p}{\partial x}\right) = p^v \frac{\partial p}{\partial x}. \tag{5}$$

The quasi-linearization procedure uses a first-order Taylor expansion about the nth iterate $\left(p^n, (\frac{\partial p}{\partial x})^n\right)$:

$$\begin{aligned}\mathcal{N}\left(p^{n+1}, \left(\frac{\partial p}{\partial x}\right)^{n+1}\right) \approx & \mathcal{N}\left(p^n, \left(\frac{\partial p}{\partial x}\right)^n\right) + \frac{\partial \mathcal{N}}{\partial p}\Big|_{(p^n, (\frac{\partial p}{\partial x})^n)} \left(p^{n+1} - p^n\right) \\ & + \frac{\partial \mathcal{N}}{\partial p_x}\Big|_{(p^n, (\frac{\partial p}{\partial x})^n)} \left(\left(\frac{\partial p}{\partial x}\right)^{n+1} - \left(\frac{\partial p}{\partial x}\right)^n\right).\end{aligned} \tag{6}$$

After differentiation and algebraic simplification, we obtain

$$\begin{aligned}\left(p^{n+1}\right)^v \left(\frac{\partial p}{\partial x}\right)^{n+1} \approx & v\left(p^n\right)^{v-1} \left(\frac{\partial p}{\partial x}\right)^n p^{n+1} + \left(p^n\right)^v \left(\frac{\partial p}{\partial x}\right)^{n+1} \\ & - v\left(p^n\right)^v \left(\frac{\partial p}{\partial x}\right)^n.\end{aligned} \tag{7}$$

Replacing the nonlinear condition term in (1) with this linear approximation gives the iterative sequence of linear equations:

$$\begin{aligned}{}_0^C D_t^\alpha p^{n+1} = & d\left(\frac{\partial^2 p}{\partial x^2}\right)^{n+1} - v\left(p^n\right)^{v-1} \left(\frac{\partial p}{\partial x}\right)^n p^{n+1} - \left(p^n\right)^v \left(\frac{\partial p}{\partial x}\right)^{n+1} \\ & + v\left(p^n\right)^v \left(\frac{\partial p}{\partial x}\right)^n + f^{n+1},\end{aligned} \tag{8}$$

which satisfies the original initial and boundary conditions:

$$p^{n+1}(x, 0) = g(x), \quad 0 \le x \le L, \tag{9}$$

$$p^{n+1}(0,t) = g_0(t), \quad p^{n+1}(L,t) = g_1(t), \quad 0 \le t \le T. \tag{10}$$

From (8), for simplicity of presentation, all nonlinear terms evaluated at the previous time level t_n are denoted collectively by $\bar{p}$, where $\bar{p} = p^n$. However, in the numerical analysis, these quantities are treated explicitly at their original time indices. Then, (8) becomes

$${}_0^C D_t^\alpha p^{n+1} = d\left(\frac{\partial^2 p}{\partial x^2}\right)^{n+1} - v\bar{p}^{v-1}\frac{\partial \bar{p}}{\partial x}p^{n+1} - \bar{p}^v\left(\frac{\partial p}{\partial x}\right)^{n+1} + v\bar{p}^v\frac{\partial \bar{p}}{\partial x} + f^{n+1}. \tag{11}$$

3 Numerical scheme formulation

3.1 The time semi-discretization

In this method, we first divide the spatial domain $[0, L]$ into N_x equal parts, where N_x is the number of nodal points, to obtain the partition Ω_x of the domain defined as $\Omega_x = \{x_m | x_m = mh, m = 0, 1, 2, \ldots, N_x\}$, where $h = \frac{L}{N_x}$. In the same way, the partition of the time domain $[0, T]$ is considered as $\Omega_t = \{t_n | t_n = nk, n = 0, 1, 2, \ldots, M_t\}$, where $k = \frac{T}{M_t}$. The formulas $p_m^n = p(x_m, t_n)$ and $f_m^n = f(x_m, t_n)$ represent the values of the functions p and f at the grid points (x_m, t_n), respectively. Similar to the classical Crank-Nicolson difference scheme, the Caputo fractional order α at $(x_m, t_{n+\frac{1}{2}})$ is defined as follows:

$${}_0^C D_t^\alpha p(x_m, t_{n+\frac{1}{2}}) = \frac{1}{\Gamma(1-\alpha)}\int_0^{t_{n+\frac{1}{2}}} \frac{\partial p(x_m, s)}{\partial s}\left(t_{n+\frac{1}{2}} - s\right)^{-\alpha} ds. \tag{12}$$

Let $t_n = nk$ and $t_{n+\frac{1}{2}} = \left(n + \frac{1}{2}\right)k$. Splitting the integral over the time sub-intervals yields

$$\begin{aligned} {}_0^C D_t^\alpha p(x_m, t_{n+\frac{1}{2}}) = {} & \frac{1}{\Gamma(1-\alpha)}\sum_{i=1}^{n}\int_{(i-1)k}^{ik} \frac{\partial p(x_m, s)}{\partial s}\left((n+\frac{1}{2})k - s\right)^{-\alpha} ds \\ & + \frac{1}{\Gamma(1-\alpha)}\int_{nk}^{(n+\frac{1}{2})k} \frac{\partial p(x_m, s)}{\partial s}\left((n+\frac{1}{2})k - s\right)^{-\alpha} ds. \end{aligned}$$

For the full interval $[t_{i-1}, t_i]$, $i = 1, 2, \ldots, n$, the midpoint is $t_{i-\frac{1}{2}}$. A Taylor expansion of p_m^i and p_m^{i-1} about this midpoint is used to approximate the time derivative. Similarly, for the half-interval $[t_n, t_{n+\frac{1}{2}}]$, the midpoint is $t_{n+\frac{1}{4}}$, and a Taylor expansion of p_m^n and $p_m^{n+\frac{1}{2}}$ around this midpoint is used to approximate the derivative. By substituting these expressions, we obtain

$${}_0^C D_t^\alpha p(x_m, t_{n+\frac{1}{2}})$$

$$= \frac{1}{\Gamma(1-\alpha)} \sum_{i=1}^{n} \int_{(i-1)k}^{ik} \left(\frac{p_m^i - p_m^{i-1}}{k} + O\left(k^2\right) \right) \left(\left(n + \frac{1}{2} \right) k - s \right)^{-\alpha} ds$$
$$+ \frac{1}{\Gamma(1-\alpha)} \int_{nk}^{(n+\frac{1}{2})k} \left(\frac{p_m^{n+1} - p_m^n}{k} + O\left(k\right) \right) \left(\left(n + \frac{1}{2} \right) k - s \right)^{-\alpha} ds.$$

By applying the power rule of integration, we obtain

$$\begin{aligned}
&{}_0^C D_t^\alpha p(x_m, t_{n+\frac{1}{2}}) \\
&= \frac{1}{\Gamma(1-\alpha)} \frac{1}{(1-\alpha)k^\alpha} \sum_{i=1}^{n} \left(p_m^i - p_m^{i-1} \right) \left[\left(n - i + \tfrac{3}{2} \right)^{1-\alpha} - \left(n - i + \tfrac{1}{2} \right)^{1-\alpha} \right] \\
&\quad + \frac{1}{\Gamma(1-\alpha)} \frac{1}{(1-\alpha)} \sum_{i=1}^{n} \left[\left(n - i + \tfrac{3}{2} \right)^{1-\alpha} - \left(n - i + \tfrac{1}{2} \right)^{1-\alpha} \right] O\left(k^{3-\alpha}\right) \\
&\quad + \frac{1}{\Gamma(1-\alpha)} \frac{1}{(1-\alpha)k^\alpha} \frac{p_m^{n+1} - p_m^n}{2^{1-\alpha}} + \frac{1}{\Gamma(1-\alpha)} \frac{1}{(1-\alpha)} \frac{1}{2^{1-\alpha}} O\left(k^{2-\alpha}\right),
\end{aligned}$$

where $O\left(k^{3-\alpha}\right) = O\left(k^2\right) O\left(k^{1-\alpha}\right)$ and $O\left(k^{2-\alpha}\right) = O\left(k\right) O\left(k^{1-\alpha}\right)$. By rearranging, we get

$$\begin{aligned}
&{}_0^C D_t^\alpha p(x_m, t_{n+\frac{1}{2}}) \\
&= \frac{1}{\Gamma(2-\alpha)} \frac{1}{k^\alpha} \frac{p_m^{n+1} - p_m^n}{2^{1-\alpha}} \\
&\quad + \frac{1}{\Gamma(2-\alpha)} \frac{1}{k^\alpha} \sum_{i=1}^{n} \left(p_m^i - p_m^{i-1} \right) \left[\left(n - i + \tfrac{3}{2} \right)^{1-\alpha} - \left(n - i + \tfrac{1}{2} \right)^{1-\alpha} \right] \\
&\quad + \frac{1}{\Gamma(2-\alpha)} \sum_{i=1}^{n} \left[\left(n - i + \tfrac{3}{2} \right)^{1-\alpha} - \left(n - i + \tfrac{1}{2} \right)^{1-\alpha} \right] O\left(k^{3-\alpha}\right) \\
&\quad + \frac{1}{\Gamma(2-\alpha)} \frac{1}{2^{1-\alpha}} O\left(k^{2-\alpha}\right),
\end{aligned} \tag{13}$$

where $\frac{1}{\Gamma(2-\alpha)} = \frac{1}{\Gamma(1-\alpha)} \frac{1}{(1-\alpha)}$. For computational efficiency, we re-index the sum by letting $j = n - i + 1$, so $i = n - j + 1$. With these substitutions, (13) becomes

$$\begin{aligned}
&{}_0^C D_t^\alpha p(x_m, t_{n+\frac{1}{2}}) \qquad (14) \\
&= \frac{1}{\Gamma(2-\alpha)} \frac{1}{k^\alpha} \frac{p_m^{n+1} - p_m^n}{2^{1-\alpha}} \\
&\quad + \frac{1}{\Gamma(2-\alpha)} \frac{1}{k^\alpha} \sum_{j=1}^{n} \left(p_m^{n-j+1} - p_m^{n-j} \right) \left[\left(j + \tfrac{1}{2} \right)^{1-\alpha} - \left(j - \tfrac{1}{2} \right)^{1-\alpha} \right] \\
&\quad + \frac{1}{\Gamma(2-\alpha)} \sum_{j=1}^{n} \left[\left(j + \tfrac{1}{2} \right)^{1-\alpha} - \left(j - \tfrac{1}{2} \right)^{1-\alpha} \right] O\left(k^{3-\alpha}\right) \\
&\quad + \frac{1}{\Gamma(2-\alpha)} \frac{1}{2^{1-\alpha}} O\left(k^{3-\alpha}\right). \qquad (15)
\end{aligned}$$

Let $\sigma_0 = \frac{1}{\Gamma(2-\alpha)k^\alpha}$ and $r_j = \left(j+\frac{1}{2}\right)^{1-\alpha} - (j-\frac{1}{2})^{1-\alpha}$, so that $\sum_{j=1}^{n} r_j = \left(n+\frac{1}{2}\right)^{1-\alpha} - \left(\frac{1}{2}\right)^{1-\alpha}$. By substituting σ_0 and r_j in (14), we get

$$\begin{aligned} &{}_0^C D_t^\alpha p(x_m, t_{n+\frac{1}{2}}) \\ &= \sigma_0 \left[\frac{p_m^{n+1} - p_m^n}{2^{1-\alpha}} + \sum_{j=1}^{n} \left(p_m^{n-j+1} - p_m^{n-j}\right) r_j \right] + \mathcal{R}_1 + \mathcal{R}_2, \end{aligned} \tag{16}$$

where

$$\begin{aligned} &\mathcal{R}_1 + \mathcal{R}_2 \\ &= \frac{1}{\Gamma(2-\alpha)} \left[\sum_{i=1}^{n} r_j O\left(k^{3-\alpha}\right) + \frac{1}{2^{1-\alpha}} O\left(k^{2-\alpha}\right) \right] \\ &= \frac{1}{\Gamma(2-\alpha)} \left[\left(\left(n+\tfrac{1}{2}\right)^{1-\alpha} - (\tfrac{1}{2})^{1-\alpha} \right) O\left(k^{3-\alpha}\right) + \left(\tfrac{1}{2}\right)^{1-\alpha} O\left(k^{2-\alpha}\right) \right]. \end{aligned}$$

Now, substituting $n = t_n/k$, we have

$$\begin{aligned} &\mathcal{R}_1 + \mathcal{R}_2 \\ &= \frac{1}{\Gamma(2-\alpha)} \left[\left(\left(\frac{t_n}{k} + \frac{1}{2}\right)^{1-\alpha} - \left(\frac{1}{2}\right)^{1-\alpha} \right) O\left(k^{3-\alpha}\right) + \left(\tfrac{1}{2}\right)^{1-\alpha} O\left(k^{2-\alpha}\right) \right]. \end{aligned}$$

For small k, we can approximate $\left(\frac{t_n}{k} + \frac{1}{2}\right)^{1-\alpha} \approx \left(\frac{t_n}{k}\right)^{1-\alpha}$, to leading order,

$$\begin{aligned} \mathcal{R}_1 + \mathcal{R}_2 &\approx \frac{1}{\Gamma(2-\alpha)} \left[\left(\frac{t_n}{k}\right)^{1-\alpha} O\left(k^{3-\alpha}\right) + \left(\frac{1}{2}\right)^{1-\alpha} O\left(k^{2-\alpha}\right) \right] \\ &= \frac{1}{\Gamma(2-\alpha)} \left[t_n^{1-\alpha} O\left(k^2\right) + \left(\frac{1}{2}\right)^{1-\alpha} O\left(k^{2-\alpha}\right) \right]. \end{aligned}$$

Since $0 < \alpha < 1$, we have $2 - \alpha < 2$, so $O\left(k^{2-\alpha}\right)$ dominates $O\left(k^2\right)$ for small k. Therefore,

$$\mathcal{R}_1 + \mathcal{R}_2 = O\left(k^{2-\alpha}\right). \tag{17}$$

Expressing the scheme directly in terms of the solution values p_m^i by dropping the truncation term, we rewrite the sum in (16) as

$$\begin{aligned} &{}_0^C D_t^\alpha p(x_m, t_{n+\frac{1}{2}}) \\ &= \sigma_0 \left[\frac{p_m^{n+1} - p_m^n}{2^{1-\alpha}} + r_1 p_m^n + \sum_{i=1}^{n-1} \left(r_{n-i+1} - r_{n-i}\right) p_m^i - r_n p_m^0 \right]. \end{aligned} \tag{18}$$

Thus, (18) is the Crank–Nicolson time-stepping scheme for the Caputo time-fractional derivative at $(x_m, t_{n+\frac{1}{2}})$.

3.2 Nonpolynomial tension spline for space discretization

From space domain discretization using a uniform mesh with step size $\Delta x = h$, we have $\Omega_x = \{x_m | x_m = hm, m = 0, 1, 2, \ldots, N_x, x_{N_x} = L, h = \frac{L}{N_x}\}$, where $\Omega_x = [0, L]$. Let $\mathcal{B}(x, t, \mu)$ be the spline function associated with the tension factor μ, which belongs to the class $C^2[0, L]$. Note that when $\mu \to 0$, this parametric spline function reduced to cubic spline. To interpolate $p(x, t)$, we consider the function at the nodal point (x_m, t_n). Let $\mathcal{B}_m^n$ denote an approximation to $p_m^n = p(x_m, t_n)$, obtained from the segment $\mathcal{B}_{\Delta m}(x, t_n, \mu)$ of the interpolating function passing through the points $(x_m, \mathcal{B}_m^n)$ and $(x_{m+1}, \mathcal{B}_{m+1}^n)$. On the sub-interval $[x_m, x_{m+1}]$, the tension spline function $\mathcal{B}_\Delta(x, t_n, \mu)$ satisfies the following differential equation when $\mu > 0$:

$$\begin{aligned}
&\mathcal{B}_\Delta''(x, t_n, \mu) - \mu \mathcal{B}_\Delta(x, t_n, \mu) \\
&= \left[\mathcal{B}_\Delta''(x_m, t_n, \mu) - \mu \mathcal{B}_\Delta(x_m, t_n, \mu)\right] \frac{x_{m+1} - x}{h} \\
&\quad + \left[\mathcal{B}_\Delta''(x_{m+1}, t_n, \mu) - \mu \mathcal{B}_\Delta(x_{m+1}, t_n, \mu)\right] \frac{x - x_m}{h},
\end{aligned} \tag{19}$$

which satisfies the interpolation conditions

$$\mathcal{B}_\Delta(x_m, t_n, \mu) = p(x_m, t_n), \quad \mathcal{B}_\Delta(x_{m+1}, t_n, \mu) = p(x_{m+1}, t_n). \tag{20}$$

The spline approximation to the second-order derivative $p_{xx}(x, t)$ is given by

$$\mathcal{B}_\Delta''(x_m, t_n, \mu) = M_m^n = (p_{xx})_m^n \text{ and } \mathcal{B}_\Delta''(x_{m+1}, t_n, \mu) = M_{m+1}^n = (p_{xx})_{m+1}^n. \tag{21}$$

To determine the arbitrary constants, we solve (19) and apply the interpolation conditions (20), together with the spline derivative relations in (21), yielding

$$\begin{aligned}
&\mathcal{B}_\Delta(x, t_n, \mu) \\
&= \frac{h^2}{\lambda^2 \sinh(\lambda)} \left[M_{m+1}^n \sinh(\lambda) \frac{x - x_m}{h} + M_m^n \sinh(\lambda) \frac{x_{m+1} - x}{h}\right] \\
&\quad - \left[\frac{h}{\lambda^2} M_m^n - \frac{1}{h^2} p_m^n\right] \frac{x_{m+1} - x}{h} - \left[\frac{h}{\lambda^2} M_{m+1}^n - \frac{1}{h^2} p_{m+1}^n\right] \frac{x - x_{m+1}}{h},
\end{aligned} \tag{22}$$

where $\lambda = \frac{h}{\sqrt{\mu}}$. Now, differentiating (22) and letting $x \to x_m$ on the interval $[x_m, x_{m+1}]$, we obtain

$$\mathcal{B}_\Delta{}'(x_m^+, t_n, \mu) = (p_x)_m^n = \frac{p_{m+1}^n - p_m^n}{h} - h\left[\gamma_1 M_{m+1}^n + \gamma_2 M_m^n\right], \tag{23}$$

where $\gamma_1 = \frac{1}{\lambda^2}\left[1 - \frac{\lambda}{\sinh(\lambda)}\right], \gamma_2 = \frac{2}{\lambda^2}\left[\lambda \coth(\lambda) - 1\right]$. By similar manner, we proceed for the interval $[x_{m-1}, x_m]$ and obtain

$$\mathcal{B}_\Delta^{'}(x_m^-, t_n, \mu) = (p_x)_m^n = \frac{p_m^n - p_{m-1}^n}{h} + h\left[\gamma_2 M_m^n + \gamma_1 M_{m-1}^n\right]. \tag{24}$$

Using the continuity of the first-order derivative of the tension spline function $\mathcal{B}_\Delta(x, t_n, \mu)$ at $x = x_m$, we obtain $\mathcal{B}_\Delta^{'}(x_m^+, t_n, \mu) = \mathcal{B}_\Delta^{'}(x_m^-, t_n, \mu)$ for $m = 1, 2, \ldots, N_x - 1, n = 0, 1, 2, \ldots, M_t$. This equality gives

$$\frac{p_{m+1}^n - p_m^n}{h} - h[\gamma_1 M_{m+1}^n + \gamma_2 M_m^n] = \frac{p_m^n - p_{m-1}^n}{h} + h[\gamma_2 M_m^n + \gamma_1 M_{m-1}^n].$$

After rearranged, this leads to the tridiagonal system

$$\gamma_1 M_{m-1}^n + \gamma_2 M_m^n + \gamma_1 M_{m+1}^n = \frac{1}{h^2}\left[p_{m-1}^n - 2p_m^n + p_{m+1}^n\right] + O\left(h^2\right). \tag{25}$$

Using the nonpolynomial spline function, we approximate the second-order spatial derivative at (x_m, t_n) [30, 14] as follows:

$$\left(\frac{\partial^2 u}{\partial x^2}\right)_m^n = M_m^n + O(h^2). \tag{26}$$

By applying the Crank–Nicholson scheme to (11) at the point $(x_m, t_{n+\frac{1}{2}})$, we obtain

$${}_0^C D_t^\alpha p_m^{n+\frac{1}{2}} = d\left(\frac{\partial^2 p}{\partial x^2}\right)_m^{n+\frac{1}{2}} - v\bar{p}^{v-1}\frac{\partial \bar{p}}{\partial x} p_m^{n+\frac{1}{2}} - \bar{p}^v\left(\frac{\partial p}{\partial x}\right)_m^{n+\frac{1}{2}} + v\bar{p}^v\frac{\partial \bar{p}}{\partial x} + f_m^{n+\frac{1}{2}}. \tag{27}$$

Now, using (18) and (27) within the Crank–Nicolson framework, and for simplicity denoting $\left(\frac{\partial p}{\partial x}\right)_m^n = (p_x)_m^n$, $\left(\frac{\partial^2 p}{\partial x^2}\right)_m^n = (p_{xx})_m^n$, we obtain

$$\begin{aligned}
&\sigma_0\left[\frac{(p_m^{n+1} - p_m^n)}{2^{1-\alpha}} + r_1 p_m^n + \sum_{i=1}^{n}\left(r_{n-i+1} - r_{n-i}\right)p_m^i - p_m^0 r_n\right] \\
&= d\frac{(p_{xx})_m^{n+1} + (p_{xx})_m^n}{2} - v\bar{p}^{v-1}\bar{p}_x\frac{p_m^{n+1} + p_m^n}{2} - \bar{p}^v\frac{(p_x)_m^{n+1} + (p_x)_m^n}{2} \\
&\quad + v\bar{p}^v\bar{p}_x + f^{n+\frac{1}{2}}.
\end{aligned} \tag{28}$$

Using (26) in (28), we get

$$\begin{aligned}\frac{M_m^{n+1}+M_m^n}{2} = & \frac{1}{d}v\bar{p}^{v-1}\bar{p}_x\frac{p_m^{n+1}+p_m^n}{2}+\frac{1}{d}\bar{p}^v\left[\frac{(p_x)_m^{n+1}+(p_x)_m^n}{2}\right]\\ & -\frac{1}{d}v\bar{p}^v\bar{p}_x-\frac{1}{d}f_m^{n+\frac{1}{2}}\\ & +\frac{\sigma_0}{d}\left[\frac{p_m^{n+1}-p_m^n}{2^{1-\alpha}}+r_1p_m^n+\sum_{i=1}^{n}\left(r_{n-i+1}-r_{n-i}\right)p_m^i-p_m^0r_n\right]\\ & +\mathcal{R}_m^{n+\frac{1}{2}}.\end{aligned} \tag{29}$$

Rewrite (25) at $(n+1)^{th}$ time-level as

$$\gamma_1M_{m-1}^{n+1}+\gamma_2M_m^{n+1}+\gamma_1M_{m+1}^{n+1}=\frac{1}{h^2}\left[p_{m-1}^{n+1}-2p_m^{n+1}+p_{m+1}^{n+1}\right], \tag{30}$$

for $m=1,2,\ldots,N_x-1, n=0,1,2,\ldots,M_t-1$. Adding (25) and (30), and then taking their average, we obtain

$$\begin{aligned}&\gamma_1h^2\left(\frac{M_{m-1}^{n+1}+M_{m-1}^n}{2}\right)+\gamma_2h^2\left(\frac{M_m^{n+1}+M_m^n}{2}\right)+\gamma_1h^2\left(\frac{M_{m+1}^{n+1}+M_{m+1}^n}{2}\right)\\ &=\frac{p_{m-1}^{n+1}+p_{m-1}^n}{2}-(p_m^{n+1}+p_m^n)+\frac{p_{m+1}^{n+1}+p_{m+1}^n}{2}.\end{aligned} \tag{31}$$

A use of (29) in (31), yields

$$\begin{aligned}&\gamma_1h^2\left[\frac{1}{d}v\bar{p}^{v-1}\bar{p}_x\frac{p_{m-1}^{n+1}+p_{m-1}^n}{2}+\frac{1}{d}\bar{p}^v\left(\frac{(p_x)_{m-1}^{n+1}+(p_x)_{m-1}^n}{2}\right)\right.\\ &-\frac{1}{d}v\bar{p}^v\bar{p}_x-\frac{1}{d}f_{m-1}^{n+\frac{1}{2}}\\ &\left.+\frac{\sigma_0}{d}\left(\frac{p_{m-1}^{n+1}-p_{m-1}^n}{2^{1-\alpha}}+r_1p_{m-1}^n+\sum_{i=1}^{n}(r_{n-i+1}-r_{n-i})p_{m-1}^i-p_{m-1}^0r_n\right)\right]\\ &+\gamma_2h^2\left[\frac{1}{d}v\bar{p}^{v-1}\bar{p}_x\frac{p_m^{n+1}+p_m^n}{2}+\frac{1}{d}\bar{p}^v\left(\frac{(p_x)_m^{n+1}+(p_x)_m^n}{2}\right)-\frac{1}{d}v\bar{p}^v\bar{p}_x-\frac{1}{d}f_m^{n+\frac{1}{2}}\right.\\ &\left.+\frac{\sigma_0}{d}\left(\frac{p_m^{n+1}-p_m^n}{2^{1-\alpha}}+r_1p_m^n+\sum_{i=1}^{n}(r_{n-i+1}-r_{n-i})p_m^i-p_m^0r_n\right)\right]\\ &+\gamma_1h^2\left[\frac{1}{d}v\bar{p}^{v-1}\bar{p}_x\frac{p_{m+1}^{n+1}+p_{m+1}^n}{2}+\frac{1}{d}\bar{p}^v\left(\frac{(p_x)_{m+1}^{n+1}+(p_x)_{m+1}^n}{2}\right)\right.\\ &-\frac{1}{d}v\bar{p}^v\bar{p}_x-\frac{1}{d}f_{m+1}^{n+\frac{1}{2}}\\ &\left.+\frac{\sigma_0}{d}\left(\frac{(p_{m+1}^{n+1}-p_{m+1}^n)}{2^{1-\alpha}}+r_1p_{m+1}^n+\sum_{i=1}^{n}(r_{n-i+1}-r_{n-i})p_{m+1}^i-p_{m+1}^0r_n\right)\right]\end{aligned}$$

$$= \frac{p_{m-1}^{n+1} + p_{m-1}^{n}}{2} - (p_m^{n+1} + p_m^n) + \frac{p_{m+1}^{n+1} + p_{m+1}^{n}}{2}. \tag{32}$$

By replacing the first-order spatial derivative with the average over four neighboring points, the above equation becomes

$$\begin{aligned}
&\gamma_1 h^2 \left[\frac{1}{d} v\bar{p}^{v-1} \bar{p}_x \frac{p_{m-1}^{n+1} + p_{m-1}^{n}}{2} + \frac{1}{d} \bar{p}^v \left(\frac{p_m^{n+1} + p_m^n - p_{m-1}^{n+1} - p_{m-1}^n}{2h} \right) \right. \\
&- \frac{1}{d} v\bar{p}^v \bar{p}_x - \frac{1}{d} f_{m-1}^{n+\frac{1}{2}} \\
&\left. + \frac{\sigma_0}{d} \left(\frac{p_{m-1}^{n+1} - p_{m-1}^n}{2^{1-\alpha}} + r_1 p_{m-1}^n + \sum_{i=1}^{n} (r_{n-i+1} - r_{n-i}) p_{m-1}^i - p_{m-1}^0 r_n \right) \right] \\
&+ \gamma_2 h^2 \left[\frac{1}{d} v\bar{p}^{v-1} \bar{p}_x \frac{p_m^{n+1} + p_m^n}{2} + \frac{1}{d} \bar{p}^v \left(\frac{p_{m+1}^{n+1} + p_{m+1}^n - p_{m-1}^{n+1} - p_{m-1}^n}{4h} \right) \right. \\
&- \frac{1}{d} v\bar{p}^v \bar{p}_x - \frac{1}{d} f_m^{n+\frac{1}{2}} \\
&\left. + \frac{\sigma_0}{d} \left(\frac{(p_m^{n+1} - p_m^n)}{2^{1-\alpha}} + r_1 u_m^n + \sum_{i=1}^{n} (r_{n-i+1} - r_{n-i}) p_m^i - p_m^0 w_n \right) \right] \\
&+ \gamma_1 h^2 \left[\frac{1}{d} v\bar{p}^{v-1} \bar{p}_x \frac{p_{m+1}^{n+1} + p_{m+1}^n}{2} + \frac{1}{d} \bar{p}^v \left(\frac{p_{m+1}^{n+1} + p_{m+1}^n - p_m^{n+1} - p_m^n}{2h} \right) \right. \\
&- \frac{1}{d} v\bar{p}^v \bar{v}_x - \frac{1}{d} f_{m+1}^{n+\frac{1}{2}} \\
&\left. + \frac{\sigma_0}{d} \left(\frac{p_{m+1}^{n+1} - p_{m+1}^n}{2^{1-\alpha}} + r_1 p_{m+1}^n + \sum_{i=1}^{n} (r_{n-i+1} - r_{n-i}) p_{m+1}^i - p_{m+1}^0 r_n \right) \right] \\
&\quad = \frac{p_{m-1}^{n+1} + p_{m-1}^{n}}{2} - (p_m^{n+1} + p_m^n) + \frac{p_{m+1}^{n+1} + p_{m+1}^{n}}{2}.
\end{aligned} \tag{33}$$

Equation (33) can be rearranged as

$$\begin{aligned}
&\left[-d + \frac{2\gamma_1 \sigma_0 h^2}{2^{1-\alpha}} + \gamma_1 h^2 v\bar{p}^{v-1} \bar{p}_x - \gamma_1 h \bar{p}^v - \frac{\gamma_2 h}{2} \bar{p}^v \right] p_{m-1}^{n+1} \\
&+ \left[2d + \frac{2\gamma_2 \sigma_0 h^2}{2^{1-\alpha}} \gamma_2 h^2 v\bar{p}^{v-1} \bar{p}_x \right] p_m^{n+1} \\
&+ \left[-d + \frac{2\gamma_1 \sigma_0 h^2}{2^{1-\alpha}} + \gamma_1 h^2 v\bar{p}^{v-1} \bar{p}_x + \gamma_1 h \bar{p}^v + \frac{\gamma_2 h}{2} \bar{p}^v \right] p_{m+1}^{n+1} \\
&\quad = \left[d + \frac{2\gamma_2 \sigma_0 h^2}{2^{1-\alpha}} - \gamma_1 h^2 v\bar{p}^{v-1} \bar{p}_x + \gamma_1 h \bar{p}^v + \frac{\gamma_2 h}{2} \bar{p}^v \right] p_{m-1}^n \\
&\qquad + \left[-2d + \frac{2\gamma_2 \sigma_0 h^2}{2^{1-\alpha}} - \gamma_2 h^2 v\bar{p}^{v-1} \bar{p}_x \right] p_m^n \\
&\qquad + \left[d + \frac{2\gamma_1 \sigma_0 h^2}{2^{1-\alpha}} - \gamma_1 h^2 v\bar{p}^{v-1} \bar{p}_x - \gamma_1 h \bar{p}^v - \frac{\gamma_2 h}{2} \bar{p}^v \right] p_{m+1}^n
\end{aligned}$$

$$
\begin{aligned}
&-2\sigma_0 h^2 r_1(\gamma_1 p_{m-1}^n+\gamma_2 p_m^n+\gamma_1 p_{m+1}^n)\\
&+2\sigma_0 h^2 r_n(\gamma_1 p_{m-1}^0+\gamma_2 p_m^0+\gamma_1 p_{m+1}^0)\\
&-2\sigma_0 h^2\sum_i^{n-1}(r_{n-i+1}-r_{n-i})(\gamma_1 p_{m-1}^i+\gamma_2 p_m^i+\gamma_1 p_{m+1}^i)\\
&+4\gamma_1 h^2 v\bar{p}^v\bar{p}_x+2\gamma_2 h^2 v\bar{p}^v\bar{p}_x\\
&+h^2\left[\gamma_1 f_{m-1}^{n+\frac{1}{2}}+\gamma_2 f_m^{n+\frac{1}{2}}+\gamma_1 f_{m+1}^{n+\frac{1}{2}}\right]+2h^2\mathcal{T}_m^{n+\frac{1}{2}},
\end{aligned}
\tag{34}
$$

where $\mathcal{T}_m^{n+\frac{1}{2}}=\gamma_1\mathcal{R}_{m-1}^{n+\frac{1}{2}}+\gamma_2\mathcal{R}_m^{n+\frac{1}{2}}+\gamma_1\mathcal{R}_{m+1}^{n+\frac{1}{2}}$ denotes the truncation error $m=1,2,\ldots,N_x-1$ and $n\geq 0$, which we will be proved later in section 4. Assuming that P_m^n is the approximate solution of p_m^n and removing the truncation error from (34), one obtains the following form of proposed scheme for the generalized NTFCDE in (1):

$$
\begin{aligned}
&\left[-d+\frac{2\gamma_1\sigma_0 h^2}{2^{1-\alpha}}+\gamma_1 h^2 v\bar{P}^{v-1}\bar{P}_x-\gamma_1 h\bar{P}^v-\frac{\gamma_2 h}{2}\bar{P}^v\right]P_{m-1}^{n+1}\\
&+\left[2d+\frac{2\gamma_2\sigma_0 h^2}{2^{1-\alpha}}+\gamma_2 h^2 v\bar{P}^{v-1}\bar{P}_x\right]P_m^{n+1}\\
&+\left[-d+\frac{2\gamma_1\sigma_0 h^2}{2^{1-\alpha}}+\gamma_1 h^2 v\bar{P}^{v-1}\bar{P}_x+\gamma_1 h\bar{P}^v+\frac{\gamma_2 h}{2}\bar{P}^v\right]P_{m+1}^{n+1}\\
&=\left[d+\frac{2\gamma_1\sigma_0 h^2}{2^{1-\alpha}}-\gamma_1 h^2 v\bar{P}^{v-1}\bar{P}_x+\gamma_1 h\bar{P}^v+\frac{\gamma_2 h}{2}\bar{P}^v\right]P_{m-1}^n\\
&\quad+\left[-2d+\frac{2\gamma_2\sigma_0 h^2}{2^{1-\alpha}}-\gamma_2 h^2 v\bar{P}^{v-1}\bar{P}_x\right]P_m^n\\
&\quad+\left[d+\frac{2\gamma_1\sigma_0 h^2}{2^{1-\alpha}}-\gamma_1 h^2 v\bar{P}^{v-1}\bar{P}_x-\gamma_1 h\bar{P}^v-\frac{\gamma_2 h}{2}\bar{P}^v\right]P_{m+1}^n\\
&\quad-2\sigma_0 h^2 r_1(\gamma_1 P_{m-1}^n+\gamma_2 P_m^n+\gamma_1 P_{m+1}^n)\\
&\quad+2\sigma_0 h^2 r_n(\gamma_1 P_{m-1}^0+\gamma_2 P_m^0+\gamma_1 P_{m+1}^0)\\
&\quad-2\sigma_0 h^2\sum_i^{n-1}(r_{n-i+1}-r_{n-i})\left(\gamma_1 P_{m-1}^i+\gamma_2 P_m^i+\gamma_1 P_{m+1}^i\right)\\
&\quad+2h^2 v\bar{P}^v\bar{P}_x(2\gamma_1+\gamma_2)+h^2\left[\gamma_1 f_{m-1}^{n+\frac{1}{2}}+\gamma_2 f_m^{n+\frac{1}{2}}+\gamma_1 f_{m+1}^{n+\frac{1}{2}}\right].
\end{aligned}
\tag{35}
$$

There are N_x-1 equations with N_x+1 unknowns in the system (35). Two more equations are required in order to solve this system. From the boundary conditions in (2) and the initial condition in (3), we have

$$
\begin{aligned}
P_0^n &= g_1(t_n), P_N^n=g_2(t_n), \qquad 0\leq n\leq M_t-1,\\
P_m^0 &= g(x_n), \qquad 0\leq n\leq N_x-1.
\end{aligned}
$$

With these initial and boundary conditions, (35) is written as two systems of equations for ease implementation. The system of equations for $n = 0$ is as follows:

$$AP^1 = BP^0 + C^0, \tag{36}$$

where A and B are tridiagonal matrices of order $N_x - 1$, defined as

$$A = \begin{bmatrix} b_1^1 & c_1^1 & & & & \\ a_2^1 & b_2^1 & c_2^1 & & & \\ & a_3^1 & b_3^1 & c_3^1 & & \\ & & \ddots & \ddots & \ddots & \\ & & & a_{N_x-2}^1 & b_{N_x-2}^1 & c_{N_x-2}^1 \\ & & & & a_{N_x-1}^1 & b_{N_x-1}^1 \end{bmatrix},$$

$$B = \begin{bmatrix} b_1^0 & c_1^0 & & & & \\ a_2^0 & b_2^0 & c_2^0 & & & \\ & a_3^0 & b_3^0 & c_3^0 & & \\ & & \ddots & \ddots & \ddots & \\ & & & a_{N_x-2}^0 & b_{N_x-2}^0 & c_{Nx-2}^0 \\ & & & & a_{N_x-1}^0 & b_{N_x-1}^0 \end{bmatrix},$$

$$P^1 = \left[P_1^1, P_2^1, \cdots, P_{N_x-1}^1\right]^{\mathrm{T}}, P^0 = \left[P_1^0, P_2^0, \cdots, P_{N_x-1}^0\right]^{\mathrm{T}},$$

$$C^0 = 2h^2 \begin{bmatrix} v\left(\bar{P}^v \bar{P}_x\right)_1^0 (2\gamma_1 + \gamma_2) + \gamma_1 f_0^{\frac{1}{2}} + \gamma_2 f_1^{\frac{1}{2}} + \gamma_1 f_2^{\frac{1}{2}} \\ v\left(\bar{P}^v \bar{P}_x\right)_2^0 (2\gamma_1 + \gamma_2) + \gamma_1 f_1^{\frac{1}{2}} + \gamma_2 f_2^{\frac{1}{2}} + \gamma_1 f_3^{\frac{1}{2}} \\ \vdots \\ v\left(\bar{P}^v \bar{P}_x\right)_{N_x-2}^0 (2\gamma_1 + \gamma_2) + \gamma_1 f_{N_x-3}^{\frac{1}{2}} + \gamma_2 f_{N_x-2}^{\frac{1}{2}} + \gamma_1 f_{N_x-1}^{\frac{1}{2}} \\ v\left(\bar{P}^v \bar{P}_x\right)_{N_x-1}^0 (2\gamma_1 + \gamma_2) + \gamma_1 f_{N_x-2}^{\frac{1}{2}} + \gamma_2 f_{N_x-1}^{\frac{1}{2}} + \gamma_1 f_{N_x}^{\frac{1}{2}} \end{bmatrix} - \begin{bmatrix} a^1 P_0^1 - a^0 P_0^0 \\ 0 \\ \vdots \\ 0 \\ c^1 P_{N_x}^1 - c^0 P_{N_x}^0 \end{bmatrix},$$

and the coefficients are

$$
\begin{aligned}
a_1 &= -d + \frac{2\gamma_1\sigma_0 h^2}{2^{1-\alpha}} + \gamma_1 h^2 v\bar{P}^{v-1}\bar{P}_x - \gamma_1 h\bar{P}^v - \frac{\gamma_2 h}{2}\bar{P}^v,\\
b_1 &= 2d + \frac{2\gamma_2\sigma_0 h^2}{2^{1-\alpha}} + \gamma_2 h^2 v\bar{P}^{v-1}\bar{P}_x,\\
c_1 &= -d + \frac{2\gamma_1\sigma_0 h^2}{2^{1-\alpha}} + \gamma_1 h^2 v\bar{P}^{v-1}\bar{P}_x + \gamma_1 h\bar{P}^v + \frac{\gamma_2 h}{2}\bar{P}^v,\\
a_0 &= d + \frac{2\gamma_1\sigma_0 h^2}{2^{1-\alpha}} - \gamma_1 h^2 v\bar{P}^{v-1}\bar{P}_x + \gamma_1 h\bar{P}^v + \frac{\gamma_2 h}{2}\bar{P}^v,\\
b_0 &= -2d + \frac{2\gamma_2\sigma_0 h^2}{2^{1-\alpha}} - \gamma_2 h^2 v\bar{P}^{v-1}\bar{P}_x,\\
c_0 &= d + \frac{2\gamma_1\sigma_0 h^2}{2^{1-\alpha}} - \gamma_1 h^2 v\bar{P}^{v-1}\bar{P}_x - \gamma_1 h\bar{P}^v - \frac{\gamma_2 h}{2}\bar{P}^v.
\end{aligned}
$$

Additionally, the following system of equations is obtained for $n \geq 1$:

$$
\begin{aligned}
AP^{n+1} = \; & BP^n + 2\sigma_0 h^2 r_n\left(\gamma_1 P^0_{m-1} + \gamma_2 P^0_m + \gamma_1 P^0_{m+1}\right) \\
& -2\sigma_0 h^2 \sum_{i}^{n-1}\left(r_{n-i+1} - r_{n-i}\right)\left(\gamma_1 P^i_{m-1} + \gamma_2 P^i_m + \gamma_1 P^i_{m+1}\right) + C^n,
\end{aligned} \tag{37}
$$

where A and B are tri-diagonal matrices of order $N_x - 1$, defined as

$$
A = \begin{bmatrix}
b_1^{n+1} & c_1^{n+1} & & & & \\
a_2^{n+1} & b_2^{n+1} & c_2^{n+1} & & & \\
 & a_3^{n+1} & b_3^{n+1} & c_3^{n+1} & & \\
 & & \ddots & \ddots & \ddots & \\
 & & & a_{N_x-2}^{n+1} & b_{N_x-2}^{n+1} & c_{N_x-2}^{n+1} \\
 & & & & a_{N_x-1}^{n+1} & b_{N_x-1}^{n+1}
\end{bmatrix},
$$

$$
B = \begin{bmatrix}
b^n & c^n & & & & \\
a^n & b^n & c^n & & & \\
 & a^n & b^n & c^n & & \\
 & & \ddots & \ddots & \ddots & \\
 & & & a^n & b^n & c^n \\
 & & & & a^n & b^n
\end{bmatrix},
$$

$$
\begin{aligned}
C^n =& 2h^2 \begin{bmatrix}
v\left(\bar{P}^v\bar{P}_x\right)_1^n(2\gamma_1+\gamma_2)+\gamma_1 f_0^{n+\frac{1}{2}}+\gamma_2 f_1^{n+\frac{1}{2}}+\gamma_1 f_2^{n+\frac{1}{2}} \\
v\left(\bar{P}^v\bar{P}_x\right)_2^n(2\gamma_1+\gamma_2)+\gamma_1 f_1^{n+\frac{1}{2}}+\gamma_2 f_2^{n+\frac{1}{2}}+\gamma_1 f_3^{n+\frac{1}{2}} \\
\vdots \\
v\left(\bar{P}^v\bar{P}_x\right)_{N_x-2}^n(2\gamma_1+\gamma_2)+\gamma_1 f_{N_x-3}^{n+\frac{1}{2}}+\gamma_2 f_{N_x-2}^{n+\frac{1}{2}}+\gamma_1 f_{N_x-1}^{n+\frac{1}{2}} \\
v\left(\bar{P}^v\bar{P}_x\right)_{N_x-1}^n(2\gamma_1+\gamma_2)+\gamma_1 f_{N_x-2}^{n+\frac{1}{2}}+\gamma_2 f_{N_x-1}^{n+\frac{1}{2}}+\gamma_1 f_{N_x}^{n+\frac{1}{2}}
\end{bmatrix} \\
& - \begin{bmatrix}
a^{n+1}U_0^{n+1}-a^nU_0^n \\
0 \\
\vdots \\
0 \\
c^{n+1}U_{N_x}^{n+1}-c^nU_{N_x}^n
\end{bmatrix},
\end{aligned}
$$

and the coefficients are

$$
\begin{aligned}
a^{n+1} &= -d+\frac{2\gamma_1\sigma_0h^2}{2^{1-\alpha}}+\gamma_1h^2v\bar{P}^{v-1}\bar{P}_x-\gamma_1h\bar{P}^v-\frac{\gamma_2h}{2}\bar{P}^v, \\
b^{n+1} &= 2d+\frac{2\gamma_2\sigma_0h^2}{2^{1-\alpha}}+\gamma_2h^2v\bar{P}^{v-1}\bar{P}_x, \\
c^{n+1} &= -d+\frac{2\gamma_1\sigma_0h^2}{2^{1-\alpha}}+\gamma_1h^2v\bar{P}^{v-1}\bar{P}_x+\gamma_1h\bar{P}^v+\frac{\gamma_2h}{2}\bar{P}^v, \\
a^n &= d+\frac{2\gamma_1\sigma_0h^2}{2^{1-\alpha}}-\gamma_1h^2v\bar{P}^{v-1}\bar{P}_x+\gamma_1h\bar{P}^v+\frac{\gamma_2h}{2}\bar{P}^v-2\sigma_0h^2r_1\gamma_1, \\
b^n &= -2d+\frac{2\gamma_2\sigma_0h^2}{2^{1-\alpha}}-\gamma_2h^2v\bar{P}^{v-1}\bar{P}_x-2\sigma_0h^2r_1\gamma_2, \\
c^n &= d+\frac{2\gamma_1\sigma_0h^2}{2^{1-\alpha}}-\gamma_1h^2v\bar{P}^{v-1}\bar{P}_x-\gamma_1h\bar{P}^v-\frac{\gamma_2h}{2}\bar{P}^v-2\sigma_0h^2r_1\gamma_1.
\end{aligned}
$$

The fully discrete scheme results in a tridiagonal linear system with N_x-1 unknowns at each time step. The Thomas algorithm is used to solve this system; during the forward elimination and backward substitution stages, a fixed number of arithmetic operations is required for each grid point. As a result, each time step has a computational cost of $O(N_x)$. Since the solution is advanced over M_t time steps, the algorithm's total computational substitution cost is $O(N_xM_t)$ with memory usage of order $O(N_x)$.

4 Theoretical analysis of the numerical scheme

4.1 Error analysis

Suppose that $\mathcal{T}_m$ is the local truncation error for the developed numerical scheme (25) in m^{th} term. Then,

$$\begin{aligned}\mathcal{T}_m = \frac{1}{h^2}\left[p_{m-1}^n - 2p_m^n + p_{m+1}^n\right] - \gamma_1 M_{m-1}^n - \gamma_2 M_m^n \\ -\gamma_1 M_{m+1}^n + O\left(h^z\right), \qquad z > 0.\end{aligned} \tag{38}$$

Replace M_{m-1}^n, M_m^n and M_{m+1}^n with $\frac{\partial^2 p_{m-1}^n}{\partial x^2}, \frac{\partial^2 p_m^n}{\partial x^2}$ and $\frac{\partial^2 p_{m+1}^n}{\partial x^2}$, respectively, in (38) to get

$$\begin{aligned}\mathcal{T}_m = \frac{1}{h^2}\left[p_{m-1}^n - 2p_m^n + p_{m+1}^n\right] - \gamma_1 \frac{\partial^2 p_{m-1}^n}{\partial x^2} - \gamma_2 \frac{\partial^2 p_m^n}{\partial x^2} \\ -\gamma_1 \frac{\partial^2 p_{m+1}^n}{\partial x^2} + O\left(h^z\right).\end{aligned} \tag{39}$$

Applying the Taylor series expansion for p_{m-1}^n, p_{m+1}^n in terms of p_m^n and its derivatives, we get

$$\begin{aligned}\mathcal{T}_m = \left(2\gamma_1 + \gamma_2 - 1\right)\frac{\partial^2 p_m^n}{\partial x^2} + h^2\left(\gamma_1 - \frac{1}{12}\right)\frac{\partial^4 p_m^n}{\partial x^4} \\ +h^4\left(\frac{\gamma_1}{12} - \frac{1}{360}\right)\frac{\partial^6 p_m^n}{\partial x^6} + O\left(h^z\right).\end{aligned} \tag{40}$$

It is clear from (40) that for suitable values of the parameters $\gamma_1(\gamma_1 = \frac{1}{12})$ and $\gamma_2 = \frac{10}{12}$ such that $2\gamma_1 + \gamma_2 = 1$, the value of z is 2. From the local substitution truncation error $\mathcal{R}_m^{n+\frac{1}{2}}$ in (29), we have

$$\begin{aligned}\mathcal{R}_m^{n+1/2} =& M_m^{n+1/2} - \frac{1}{d}v\bar{p}^{v-1}\bar{p}_x\frac{p_m^{n+1} + p_m^n}{2} - \frac{1}{d}\bar{p}^v\left[\frac{(p_x)_m^{n+1} + (p_x)_m^n}{2}\right] \\ &+ \frac{1}{d}v\bar{p}^v\bar{p}_x + \frac{1}{d}f_m^{n+1/2} \\ &- \frac{\sigma_0}{d}\left[\frac{p_m^{n+1} - p_m^n}{2^{1-\alpha}} + r_1 p_m^n + \sum_{i=1}^{n}(r_{n-i+1} - r_{n-i})p_m^i - p_m^0 r_n\right] \\ =& O\left(h^2\right) + O\left(k^{2-\alpha}\right) = O\left(h^2 + k^{2-\alpha}\right).\end{aligned} \tag{41}$$

Based on the above concepts, truncation error $\mathcal{T}_m^{n+\frac{1}{2}}$ of (34) is given by

$$\mathcal{T}_m^{n+\frac{1}{2}}$$

$$
\begin{aligned}
&= \gamma_1 \mathcal{R}_{m-1}^{n+\frac{1}{2}} + \gamma_2 \mathcal{R}_m^{n+\frac{1}{2}} + \gamma_1 \mathcal{R}_{m+1}^{n+\frac{1}{2}} \\
&= -\frac{\sigma_0\gamma_1}{d}\left[\frac{p_{m-1}^{n+1} - p_{m-1}^{n}}{2^{1-\alpha}} + r_1 p_{m-1}^{n} + \sum_{i=1}^{n}(r_{n-i+1} - r_{n-i})p_{m-1}^{i} - p_{m-1}^{0} r_n\right] \\
&\quad - \frac{\sigma_0\gamma_2}{d}\left[\frac{p_{m}^{n+1} - p_{m}^{n}}{2^{1-\alpha}} + r_1 p_{m}^{n} + \sum_{i=1}^{n}(r_{n-i+1} - r_{n-i})p_{m}^{i} - p_{m}^{0} r_n\right] \\
&\quad - \frac{\sigma_0\gamma_1}{d}\left[\frac{p_{m+1}^{n+1} - p_{m+1}^{n}}{2^{1-\alpha}} + r_1 p_{m+1}^{n} + \sum_{i=1}^{n}(r_{n-i+1} - r_{n-i})p_{m+1}^{i} - p_{m+1}^{0} r_n\right] \\
&\quad - \gamma_1\left[\frac{1}{d} v\bar{p}^{v-1}\bar{p}_x \frac{p_{m-1}^{n+1} + p_{m-1}^{n}}{2} + \frac{1}{d}\bar{p}^{v}\left[\frac{(p_x)_{m-1}^{n+1} + (p_x)_{m-1}^{n}}{2}\right] - \frac{1}{d} v\bar{p}^{v}\bar{p}_x\right] \\
&\quad - \gamma_2\left[\frac{1}{d} v\bar{p}^{v-1}\bar{p}_x \frac{p_{m}^{n+1} + p_{m}^{n}}{2} + \frac{1}{d}\bar{p}^{v}\left[\frac{(p_x)_{m}^{n+1} + (p_x)_{m}^{n}}{2}\right] - \frac{1}{d} v\bar{p}^{v}\bar{p}_x\right] \\
&\quad - \gamma_1\left[\frac{1}{d} v\bar{p}^{v-1}\bar{p}_x \frac{p_{m+1}^{n+1} + p_{m+1}^{n}}{2} + \frac{1}{d}\bar{p}^{v}\left[\frac{(p_x)_{m+1}^{n+1} + (p_x)_{m+1}^{n}}{2}\right] - \frac{1}{d} v\bar{p}^{v}\bar{p}_x\right] \\
&\quad - \left(\gamma_1 M_{m-1}^{n+\frac{1}{2}} + \gamma_2 M_{m}^{n+\frac{1}{2}} + \gamma_1 M_{m+1}^{n+\frac{1}{2}}\right) \\
&\quad + \left(\gamma_1 \frac{1}{d} f_{m-1}^{n+\frac{1}{2}} + \gamma_2 \frac{1}{d} f_{m}^{n+\frac{1}{2}} + \gamma_1 \frac{1}{d} f_{m+1}^{n+\frac{1}{2}}\right) \\
&= (2\gamma_1 + \gamma_2)\left(O(h^2) + O(k^{2-\alpha})\right) = O\left(h^2 + k^{2-\alpha}\right). \qquad (42)
\end{aligned}
$$

Remark 1. A numerical method is said to be consistent if $\mathcal{T}_m^{n+\frac{1}{2}} \to 0$ as $h, k \to 0$

4.2 Stability analysis

In this section, we apply Von Neumann stability analysis to investigate the stability of the proposed numerical scheme. The stability analysis begins by introducing the error term $\mathcal{E}_m^n = p_m^n - P_m^n$, for $m = 0, 1, 2, \ldots, N_x, n = 0, 1, 2, \ldots, M_t$, where P_m^n denotes the numerical approximation to the exact solution of the system defined in (35). Substituting the error expression into the numerical scheme (35), we obtain the following error equation:

$$
\begin{aligned}
&\left[-d + \frac{2\gamma_1\sigma_0 h^2}{2^{1-\alpha}} + \gamma_1 h^2 v\bar{P}^{v-1}\bar{P}_x - \gamma_1 h\bar{P}^{v} - \frac{\gamma_2 h}{2}\bar{P}^{v}\right]\mathcal{E}_{m-1}^{n+1} \\
&+ \left[2d + \frac{2\gamma_2\sigma_0 h^2}{2^{1-\alpha}} + \gamma_2 h^2 v\bar{P}^{v-1}\bar{P}_x\right]\mathcal{E}_{m}^{n+1} \\
&+ \left[-d + \frac{2\gamma_1\sigma_0 h^2}{2^{1-\alpha}} + \gamma_1 h^2 v\bar{P}^{v-1}\bar{P}_x + \gamma_1 h\bar{P}^{v} + \frac{\gamma_2 h}{2}\bar{P}^{v}\right]\mathcal{E}_{m+1}^{n+1}
\end{aligned}
$$

$$
\begin{aligned}
&= \left[d + \frac{2\gamma_1\sigma_0 h^2}{2^{1-\alpha}} - \gamma_1 h^2 v\bar{P}^{v-1}\bar{P}_x + \gamma_1 h\bar{P}^v + \frac{\gamma_2 h}{2}\bar{P}^v\right]\mathcal{E}_{m-1}^n \\
&+ \left[-2d + \frac{2\gamma_2\sigma_0 h^2}{2^{1-\alpha}} - \gamma_2 h^2 v\bar{P}^{v-1}\bar{P}_x\right]\mathcal{E}_m^n \\
&+ \left[d + \frac{2\gamma_1\sigma_0 h^2}{2^{1-\alpha}} - \gamma_1 h^2 v\bar{P}^{v-1}\bar{P}_x - \gamma_1 h\bar{P}^v - \frac{\gamma_2 h}{2}\bar{P}^v\right]\mathcal{E}_{m+1}^n \\
&- 2\sigma_0 h^2 r_1\left(\gamma_1\mathcal{E}_{m-1}^n + \gamma_2\mathcal{E}_m^n + \gamma_1\mathcal{E}_{m+1}^n\right) \\
&+ 2\sigma_0 h^2 r_n\left(\gamma_1\mathcal{E}_{m-1}^0 + \gamma_2\mathcal{E}_m^0 + \gamma_1\mathcal{E}_{m+1}^0\right) \\
&- 2\sigma_0 h^2\sum_i^{n-1}(r_{n-i+1} - r_{n-i})\left(\gamma_1\mathcal{E}_{m-1}^i + \gamma_2\mathcal{E}_m^i + \gamma_1\mathcal{E}_{m+1}^i\right), \qquad (43)
\end{aligned}
$$

with $\mathcal{E}_0^n = \mathcal{E}_{N_x}^n = 0$, for $n > 0, m = 1, 2, \ldots, N_x - 1$. Consider the grid function that is defined as

$$
\mathcal{E}^n(x) = \begin{cases} \mathcal{E}_m^n, & x_{m-\frac{h}{2}} \le x \le x_{m+\frac{h}{2}}, \quad 1 \le m \le N_x - 1, \\ 0, & 0 \le x \le \frac{h}{2} \text{ or } x_{L-\frac{h}{2}} \le x \le L. \end{cases}
$$

Then, the Fourier series expansion of $\mathcal{E}^n(x)$ is given by

$$
\mathcal{E}^n(x) = \sum_{k=-\infty}^{\infty} \eta^n(k) e^{i\frac{2\pi k x}{L}},
$$

where L is right boundary, $i^2 = -1$ and

$$
\eta^n(k) = \frac{1}{L}\int_0^L \mathcal{E}^n(\beta) e^{-i\frac{2\pi k\beta}{L}} d\beta, \qquad n = 1, 2, M_t,
$$

are known as discrete Fourier coefficients. Defining the norm

$$
\|\mathcal{E}^n\|_2 = \left[\sum_{m=1}^{N-1} h|\mathcal{E}_m^n|^2\right]^{\frac{1}{2}} = \left[\int_0^L |\mathcal{E}^n(x)|^2 dx\right]^{\frac{1}{2}}.
$$

Introducing Parseval's identity (for the discrete Fourier transform)

$$
\int_0^L |\mathcal{E}^n(x)|^2 dx = \sum_{k=-\infty}^{\infty} |\eta^n(k)|^2,
$$

and the norm

$$
\|\mathcal{E}^n\|_2 = \left(\sum_{m}^{N_x-1} h|\mathcal{E}_m^n|^2\right)^{\frac{1}{2}} = \left(\int_0^L |\mathcal{E}^n(x)^2 dx|\right)^{\frac{1}{2}}.
$$

Thus, we obtain the following expression:

$$\|\mathcal{E}^n\|_2^2 = \sum_{k=-\infty}^{\infty} |\eta^n(k)|^2.$$

Based on the above analysis discussed in this section, we can write the solution of (43) as $\mathcal{E}_m^n = \eta^n e^{imh\theta}$, where $L = 1$ and $\theta = \frac{2\pi k}{L}$. Now, we use the expression $\mathcal{E}_m^n = \eta^n e^{imh\theta}$ in (43) to obtain

$$\begin{aligned}
&\left[-d + \frac{2\gamma_1\sigma_0 h^2}{2^{1-\alpha}} + \gamma_1 h^2 v\bar{P}^{v-1}\bar{P}_x - \gamma_1 h\bar{P}^v - \frac{\gamma_2 h}{2}\bar{P}^v \right]\eta^{n+1}e^{i(m-1)h\theta} \\
&+ \left[2d + \frac{2\gamma_2\sigma_0 h^2}{2^{1-\alpha}} + \gamma_2 h^2 v\bar{P}^{v-1}\bar{P}_x \right]\eta^{n+1}e^{imh\theta} \\
&+ \left[-d + \frac{2\gamma_1\sigma_0 h^2}{2^{1-\alpha}} + \gamma_1 h^2 v\bar{P}^{v-1}\bar{P}_x + \gamma_1 h\bar{P}^v + \frac{\gamma_2 h}{2}\bar{P}^v \right]\eta^{n+1}e^{i(m+1)h\theta} \\
&\quad = \left[d + \frac{2\gamma_1\sigma_0 h^2}{2^{1-\alpha}} - \gamma_1 h^2 v\bar{P}^{v-1}\bar{P}_x + \gamma_1 h\bar{P}^v + \frac{\gamma_2 h}{2}\bar{P}^v \right]\eta^n e^{i(m-1)h\theta} \\
&\qquad + \left[-2d + \frac{2\gamma_2\sigma_0 h^2}{2^{1-\alpha}} - \gamma_2 h^2 v\bar{P}^{v-1}\bar{P}_x \right]\eta^n e^{imh\theta} \\
&\qquad + \left[d + \frac{2\gamma_1\sigma_0 h^2}{2^{1-\alpha}} - \gamma_1 h^2 v\bar{P}^{v-1}\bar{P}_x - \gamma_1 h\bar{P}^v - \frac{\gamma_2 h}{2}\bar{P}^v \right]\eta^n e^{i(m+1)h\theta} \\
&\qquad - 2\sigma h^2 r_1 \left(\gamma_1 e^{i(m-1)h\theta} + \gamma_2 e^{imh\theta} + \gamma_1 e^{i(m+1)h\theta} \right)\eta^n \\
&\qquad + 2\sigma h^2 r_n \left(\gamma_1 e^{i(m-1)h\theta} + \gamma_2 e^{imh\theta} + \gamma_1 e^{i(m+1)h\theta} \right)\eta^0 \\
&\qquad - 2\sigma_0 h^2 \sum_{j=1}^{n-1} (r_{n-j+1} - r_{n-j}) \left(\gamma_1 e^{i(m-1)h\theta} + \gamma_2 e^{imh\theta} + \gamma_1 e^{i(m+1)h\theta} \right)\eta^j. \qquad (44)
\end{aligned}$$

Divide the entire equation by $e^{imh\theta}$ and apply the trigonometric identities $e^{ih\theta} + e^{-ih\theta} = 2\cos(h\theta), e^{ih\theta} - e^{-ih\theta} = 2i\sin(h\theta)$, we get

$$\begin{aligned}
&\left[2\left(-d + \frac{2\gamma_1\sigma_0 h^2}{2^{1-\alpha}} + \gamma_1 h^2 v\bar{P}^{v-1}\bar{P}_x \right)\cos(h\theta) \right. \\
&\left. + 2i\left(\gamma_1 h\bar{P}^v + \frac{\gamma_2 h}{2}\bar{P}^v \right)\sin(h\theta) + 2d + \frac{2\gamma_2\sigma_0 h^2}{2^{1-\alpha}} + \gamma_2 h^2 v\bar{P}^{v-1}\bar{P}_x \right]\eta^{n+1} \\
&\quad = \left[2\left(d + \frac{2\gamma_1\sigma_0 h^2}{2^{1-\alpha}} - \gamma_1 h^2 v\bar{P}^{v-1}\bar{P}_x \right)\cos(h\theta) \right. \\
&\qquad - 2i\left(\gamma_1 h\bar{P}^v + \frac{\gamma_2 h}{2}\bar{P}^v \right)\sin(h\theta) \\
&\qquad \left. - 2d + \frac{2\gamma_2\sigma_0 h^2}{2^{1-\alpha}} - \gamma_2 h^2 v\bar{P}^{v-1}\bar{P}_x \right]\eta^n
\end{aligned}$$

$$- 2\sigma_0 h^2 r_1 \Big(\gamma_2 + 2\gamma_1 \cos(h\theta)\Big)\eta^n + 2\sigma_0 h^2 r_n \Big(\gamma_2 + 2\gamma_1 \cos(h\theta)\Big)\eta^0 + 2\sigma_0 h^2 \sum_{j=1}^{n-1} \Big(r_{n-j} - r_{n-j+1}\Big)\Big(\gamma_2 + 2\gamma_1 \cos(h\theta)\Big)\eta^j. \tag{45}$$

For $n = 0$, (45) gives

$$\begin{aligned} |\eta^1| &= \frac{\left|\left(\frac{2\sigma_0 h^2}{2^{1-\alpha}} - h^2 v \bar{P}^{v-1} \bar{P}_x\right)\beta_0 - \beta_1 - 2i\beta_2 \sin(h\theta)\right|}{\left|\left(\frac{2\sigma_0 h^2}{2^{1-\alpha}} + h^2 v \bar{P}^{v-1} \bar{P}_x\right)\beta_0 + \beta_1 + 2i\beta_2 \sin(h\theta)\right|}|\eta^0| \\ &= \frac{\sqrt{\left(\left(\frac{2\sigma_0 h^2}{2^{1-\alpha}} - h^2 v \bar{P}^{v-1} \bar{P}_x\right)\beta_0 - \beta_1\right)^2 + 4\big(\beta_2 \sin(h\theta)\big)^2}}{\sqrt{\left(\left(\frac{2\sigma_0 h^2}{2^{1-\alpha}} + h^2 v \bar{P}^{v-1} \bar{P}_x\right)\beta_0 + \beta_1\right)^2 + 4\big(\beta_2 \sin(h\theta)\big)^2}}|\eta^0| \\ &\leq |\eta^0|, \end{aligned} \tag{46}$$

where

$$\beta_0 = \gamma_2 + 2\gamma_1 \cos(h\theta), \quad \beta_1 = 2d\,(1 - \cos(h\theta))\,, \quad \beta_2 = \gamma_1 h \bar{P}^v + \frac{\gamma_2 h}{2}\bar{P}^v. \tag{47}$$

Now, assume that the inequality $|\eta^l| \leq |\eta^0|$ holds for $l = 2, 3, \ldots, n-1$. Then, from (45), we get

$$\begin{aligned} |\eta^{n+1}| =& \frac{\left|\left(\frac{2\sigma_0 h^2}{2^{1-\alpha}} - h^2 v \bar{P}^{v-1} \bar{P}_x\right)\beta_0 - \beta_1 - 2i\beta_2 \sin(h\theta) - 2\sigma_0 h^2 \beta_1 r_1\right|}{\left|\left(\frac{2\sigma_0 h^2}{2^{1-\alpha}} + h^2 v \bar{P}^{v-1} \bar{P}_x\right)\beta_0 + \beta_1 + 2i\beta_2 \sin(h\theta)\right|}|\eta^n| \\ &+ \frac{2\sigma_0 h^2 \beta_1}{\left|\left(\frac{2\sigma_0 h^2}{2^{1-\alpha}} + h^2 v \bar{P}^{v-1} \bar{P}_x\right)\beta_0 + \beta_1 + 2i\beta_2 \sin(h\theta)\right|} \\ &\times \left(r_n |\eta^0| + \sum_{j=1}^{n-1} (r_{n-j} - r_{n-j+1})|\eta^j|\right). \end{aligned} \tag{48}$$

From the assumption $|\eta^l| \leq |\eta^0|$ holds for $l = 2, 3, \ldots, n-1$ and decreasing function of r_j, we have

$$\begin{aligned} r_n |\eta^0| + \sum_{j=1}^{n-1} (r_{n-j} - r_{n-j+1})|\eta^j| &\leq \left(\sum_{j=1}^{n-1} (r_{n-j} - r_{n-j+1}) + r_n\right)|\eta^0| \\ &= r_1 |\eta^0|. \end{aligned} \tag{49}$$

Using (49) in (48), one obtains

$$\begin{aligned}
&|\eta^{n+1}| \\
&\leq \frac{\left|\left(\frac{2\sigma_0 h^2}{2^{1-\alpha}} - h^2 v\bar{P}^{v-1}\bar{P}_x\right)\beta_0 - \beta_1 - 2i\beta_2\sin(h\theta) - 2\sigma_0 h^2\beta_1 r_1\right| + 2\sigma_0 h^2\beta_1 r_1}{\left|\left(\frac{2\sigma_0 h^2}{2^{1-\alpha}} + h^2 v\bar{P}^{v-1}\bar{P}_x\right)\beta_0 + \beta_1 + 2i\beta_2\sin(h\theta)\right|}|\eta^0| \\
&\leq \frac{\left|\left(\frac{2\sigma_0 h^2}{2^{1-\alpha}} - h^2 v\bar{P}^{v-1}\bar{P}_x\right)\beta_0 - \beta_1 - 2\sigma_0 h^2\beta_1 r_1\right| + 2\beta_2\sin(h\theta) + 2\sigma_0 h^2\beta_1 r_1}{\left|\left(\frac{2\sigma_0 h^2}{2^{1-\alpha}} + h^2 v\bar{P}^{v-1}\bar{P}_x\right)\beta_0 + \beta_1 + 2i\beta_2\sin(h\theta)\right|}|\eta^0|,
\end{aligned} \tag{50}$$

where β_0, β_1 and β_2 as defined in (47). As for small $\sin(\theta) \leq \theta$, we have

$$\begin{aligned}
&|\eta^{n+1}| \\
&\leq \frac{\left|\left(\frac{2\sigma_0 h^2}{2^{1-\alpha}} - h^2 v\bar{P}^{v-1}\bar{P}_x\right)\beta_3 - d\theta^2 h^2 - 2d\sigma_0 h^4\theta^2 r_1\right| + 2\beta_2 h\theta + 2d\sigma_0 h^4\theta^2 r_1}{\left|\left(\frac{2\sigma_0 h^2}{2^{1-\alpha}} + h^2 v\bar{P}^{v-1}\bar{P}_x\right)\beta_0 + \beta_1 + 2i\beta_2\sin(h\theta)\right|}|\eta^0|.
\end{aligned} \tag{51}$$

where $\beta_3 = \gamma_2 + 2\gamma_1(1 - \frac{h^2\theta^2}{2})$. Now, we consider two cases: In the first case, we consider $\left(\frac{2\sigma_0 h^2}{2^{1-\alpha}} - h^2 v\bar{P}^{v-1}\bar{P}_x\right)\beta_3 - d\theta^2 h^2 > 2d\sigma_0 h^4\theta^2 r_1$, which yields

$$\begin{aligned}
&|\eta^{n+1}| \\
&\leq \frac{\left(\frac{2\sigma_0 h^2}{2^{1-\alpha}} - h^2 v\bar{P}^{v-1}\bar{P}_x\right)\beta_3 - d\theta^2 h^2 - 2d\sigma_0 h^4\theta^2 r_1 + 2\beta_2 h\theta + 2d\sigma_0 h^4\theta^2 r_1}{\left|\left(\frac{2\sigma_0 h^2}{2^{1-\alpha}} + h^2 v\bar{P}^{v-1}\bar{P}_x\right)\beta_0 + \beta_1 + 2i\beta_2\sin(h\theta)\right|}|\eta^0| \\
&\leq \frac{\left(\frac{2\sigma_0 h^2}{2^{1-\alpha}} - h^2 v\bar{P}^{v-1}\bar{P}_x\right)\beta_3 - d\theta^2 h^2 + 2\beta_2 h\theta}{\left|\left(\frac{2\sigma_0 h^2}{2^{1-\alpha}} + h^2 v\bar{P}^{v-1}\bar{P}_x\right)\beta_0 + \beta_1 + 2i\beta_2\sin(h\theta)\right|}|\eta^0|.
\end{aligned} \tag{52}$$

From the boundedness of $\bar{P}$ and $\bar{P}_x$, we obtain

$$\frac{2h\theta(\gamma_1 h\bar{P}^v + \frac{\gamma_2 h}{2}\bar{P}^v)}{\left|\left(\frac{2\sigma_0 h^2}{2^{1-\alpha}} + h^2 v\bar{P}^{v-1}\bar{P}_x\right)\beta_0 + \beta_1 + 2i\beta_2\sin(h\theta)\right|} \to 0 \text{ as } k \to 0. \tag{53}$$

Using (53) in (52), we obtain

$$|\eta^{n+1}| \leq \frac{\left(\frac{2\sigma_0 h^2}{2^{1-\alpha}} - h^2 v\bar{P}^{v-1}\bar{P}_x\right)\beta_3 - d\theta^2 h^2}{\left|\left(\frac{2\sigma_0 h^2}{2^{1-\alpha}} + h^2 v\bar{P}^{v-1}\bar{P}_x\right)\beta_0 + \beta_1 + 2i\beta_2\sin(h\theta)\right|}|\eta^0| \leq |\eta^0|.$$

Second, we consider $\left(\frac{2\sigma_0 h^2}{2^{1-\alpha}} - h^2 v\bar{P}^{v-1}\bar{P}_x\right)\beta_3 - d\theta^2 h^2 < 2d\sigma_0 h^4\theta^2 r_1$, which gives

$$|\eta^{n+1}| \leq \frac{4d\sigma_0 h^4\theta^2 r_1 - \left(\frac{2\sigma_0 h^2}{2^{1-\alpha}} - h^2 v\bar{P}^{v-1}\bar{P}_x\right)\beta_3 + d\theta^2 h^2 - 2\beta_2 h\theta}{\left|\left(\frac{2\sigma_0 h^2}{2^{1-\alpha}} + h^2 v\bar{P}^{v-1}\bar{P}_x\right)\beta_0 + \beta_1 + 2i\beta_2\sin(h\theta)\right|}|\eta^0|. \tag{54}$$

In order to obtain $|\eta^{n+1}| \leq |\eta^0|$ the next condition must satisfied

$$\frac{4d\sigma_0 h^4\theta^2 r_1 - \left(\frac{2\sigma_0 h^2}{2^{1-\alpha}} - h^2 v\bar{P}^{v-1}\bar{P}_x\right)\beta_3 + d\theta^2 h^2 - 2\beta_2 h\theta}{\left|\left(\frac{2\sigma_0 h^2}{2^{1-\alpha}} + h^2 v\bar{P}^{v-1}\bar{P}_x\right)\beta_0 + \beta_1 + 2i\beta_2\sin(h\theta)\right|} \leq 1,$$

which is hold true when $(3/2) \leq 3^{\alpha}$. Thus, the numerical scheme in (35) is conditionally stable.

4.3 Convergence analysis

In this section, we investigate the convergence of the proposed numerical scheme (35), for which the error is defined as $\xi_m^n = p_m^n - P_m^n, m = 0, 1, 2, \ldots, N_x, n = 0, 1, 2, \ldots, M_t$, at the grid point (x_m, t_n). Then, the error equation is written in the following form:

$$\begin{aligned}
&\left[-d + \frac{2\gamma_1\sigma_0 h^2}{2^{1-\alpha}} + \gamma_1 h^2 v\bar{P}^{v-1}\bar{P}_x - \gamma_1 h\bar{P}^v - \frac{\gamma_2 h}{2}\bar{P}^v\right]\xi_{m-1}^{n+1} \\
&+ \left[2d + \frac{2\gamma_2\sigma_0 h^2}{2^{1-\alpha}} + \gamma_2 h^2 v\bar{P}^{v-1}\bar{P}_x\right]\xi_m^{n+1} \\
&+ \left[-d + \frac{2\gamma_1\sigma_0 h^2}{2^{1-\alpha}} + \gamma_1 h^2 v\bar{P}^{v-1}\bar{P}_x + \gamma_1 h\bar{P}^v + \frac{\gamma_2 h}{2}\bar{P}^v\right]\xi_{m+1}^{n+1} \\
&= \left[d + \frac{2\gamma_1\sigma_0 h^2}{2^{1-\alpha}} - \gamma_1 h^2 v\bar{P}^{v-1}\bar{P}_x + \gamma_1 h\bar{P}^v + \frac{\gamma_2 h}{2}\bar{P}^v\right]\xi_{m-1}^{n} \\
&\quad + \left[-2d + \frac{2\gamma_2\sigma_0 h^2}{2^{1-\alpha}} - \gamma_2 h^2 v\bar{P}^{v-1}\bar{P}_x\right]\xi_m^n \\
&\quad + \left[d + \frac{2\gamma_1\sigma_0 h^2}{2^{1-\alpha}} - \gamma_1 h^2 v\bar{P}^{v-1}\bar{P}_x - \gamma_1 h\bar{P}^v - \frac{\gamma_2 h}{2}\bar{P}^v\right]\xi_{m+1}^n \\
&\quad - 2\sigma_0 h^2 r_1\left(\gamma_1\xi_{m-1}^n + \gamma_2\xi_m^n + \gamma_1\xi_{m+1}^n\right) \\
&\quad + 2\sigma_0 h^2 r_n\left(\gamma_1\xi_{m-1}^0 + \gamma_2\xi_m^0 + \gamma_1\xi_{m+1}^0\right) \\
&\quad - 2\sigma_0 h^2\sum_i^{n-1}\left(r_{n-i+1} - r_{n-i}\right)\left(\gamma_1\xi_{m-1}^i + \gamma_2\xi_m^i + \gamma_1\xi_{m+1}^i\right) \\
&\quad + 2h^2\left(\gamma_1\mathcal{R}_{m-1}^{n+\frac{1}{2}} + \gamma_2\mathcal{R}_m^{n+\frac{1}{2}} + \gamma_1\mathcal{R}_{m+1}^{n+\frac{1}{2}}\right),
\end{aligned} \tag{55}$$

associated with the required conditions in terms of error term $\xi_0^n = \xi_{N_x}^n = \xi_m^0 = 0$ and $\mathcal{R}_m^{n+\frac{1}{2}} \leq C\mathcal{T}_m^{n+\frac{1}{2}} \leq C^1\left(h^2 + k^{2-\alpha}\right)$. Likewise, the grid functions defined in the previous section, we use the following grid functions in the further process of proving the convergence:

$$\xi^n(x) = \begin{cases} \xi_m^n, & x_{m-\frac{h}{2}} \leq x \leq x_{m+\frac{h}{2}}, \quad 1 \leq m \leq N_x - 1, \\ 0, & 0 \leq x \leq \frac{h}{2} \text{ or } x_{L-\frac{h}{2}} \leq x \leq L, \end{cases}$$

and

$$\mathcal{R}^{n+\frac{1}{2}}(x)=\begin{cases}\mathcal{R}_m^{n+\frac{1}{2}}, & x_{m-\frac{h}{2}}\le x\le x_{m+\frac{h}{2}},\quad 1\le m\le N_x-1,\\ 0, & 0\le x\le \frac{h}{2} \text{ or } x_{L-\frac{h}{2}}\le x\le L.\end{cases}$$

Expanding the grid functions $\mathcal{R}^{n+\frac{1}{2}}(x)$ and $\xi^n(x)$ using the Fourier series, we have

$$\xi^n(x)=\sum_{l=-\infty}^{\infty}\delta^n(l)e^{i\frac{2\pi lx}{L}},$$
$$\mathcal{R}^{n+\frac{1}{2}}(x)=\sum_{l=-\infty}^{\infty}r^{n+\frac{1}{2}}(l)e^{i\frac{2\pi lx}{L}},\quad n=0,1,2,\ldots,M_t,$$

where $\delta^n(l)$ and $r^{n+\frac{1}{2}}(l)$ are defined as

$$\delta^n(l)=\frac{1}{L}\int_0^L\xi^n(s)e^{-i\frac{2\pi ls}{L}}ds,\quad r^{n+\frac{1}{2}}(l)=\frac{1}{L}\int_0^L\mathcal{R}^{n+\frac{1}{2}}(s)e^{-i\frac{2\pi ls}{L}}ds.$$

Using Parseval's identity and the L_2-norm, we get

$$\|\xi^n\|_2^2=\sum_{n=1}^{M_t-1}h|\xi_m^n|^2=\int_0^L|\xi^n(s)|^2ds=L\sum_{l=-\infty}^{\infty}|\delta^n(l)|^2,\tag{56}$$

and

$$\|\mathcal{R}^{n+\frac{1}{2}}\|_2^2=\sum_{n=1}^{M_t-1}h|\mathcal{R}_m^{n+\frac{1}{2}}|^2=\int_0^L|\mathcal{R}^{n+\frac{1}{2}}(s)|^2ds=L\sum_{l=-\infty}^{\infty}|r^{n+\frac{1}{2}}(l)|^2.\tag{57}$$

Assuming that $\xi_m^n=\delta^n e^{ihm\theta}$ and $\mathcal{R}_m^{n+\frac{1}{2}}=r^{n+\frac{1}{2}}e^{ihm\theta}$, (55) gives

$$\begin{aligned}\delta^{n+1}=&\frac{\left(\frac{2\sigma_0h^2}{2^{1-\alpha}}-h^2v\bar{P}^{v-1}\bar{P}_x\right)\beta_0-\beta_1-2i\beta_2\sin(h\theta)-2\sigma_0h^2r_1(\gamma_2+2\gamma_1\cos(h\theta))}{\left(\frac{2\sigma_0h^2}{2^{1-\alpha}}+h^2v\bar{P}^{v-1}\bar{P}_x\right)\beta_0+\beta_1+2i\beta_2\sin(h\theta)}\delta^n\\&+\frac{2\sigma_0h^2\beta_0r_n}{\left(\frac{2\sigma_0h^2}{2^{1-\alpha}}+h^2v\bar{P}^{v-1}\bar{P}_x\right)\beta_0+\beta_1+2i\beta_2\sin(h\theta)}\delta^0\\&+\frac{2\sigma_0\beta_1h^2\sum_{j=1}^{n-1}(r_{n-j}-r_{n-j+1})}{\left(\frac{2\sigma_0h^2}{2^{1-\alpha}}+h^2v\bar{P}^{v-1}\bar{P}_x\right)\beta_0+\beta_1+2i\beta_2\sin(h\theta)}\delta^j\\&+\frac{2h^2\beta_0}{\left(\frac{2\sigma_0h^2}{2^{1-\alpha}}+h^2v\bar{P}^{v-1}\bar{P}_x\right)\beta_0+\beta_1+2i\beta_2\sin(h\theta)}r^{n+\frac{1}{2}}.\end{aligned}\tag{58}$$

Lemma 1. If δ^n satisfy (58), then there exists a positive constant C such that

$$|\delta^n|\le C|r^{\frac{1}{2}}|,\qquad n=0,1,2,\ldots,M_t.$$

Proof. We will prove the result using mathematical induction. Since the error equation satisfies the initial condition $\xi_m^0 = 0$, it follows that $r^0 = 0$. Thus, for $n = 0$, (58) gives

$$|\delta^1| = \frac{2h^2\beta_0}{|\left(\frac{2\sigma_0 h^2}{2^{1-\alpha}} + h^2 v\bar{P}^{v-1}\bar{P}_x\right)\beta_0 + \beta_1 + 2i\beta_2\sin(h\theta)|}|r^{\frac{1}{2}}|.$$

It implies that $|\delta^1| \le C|r^{\frac{1}{2}}|$, for some $C \le 1$. Now, we assume that the result is true for n such that $|\delta^n| \le C|r^{\frac{1}{2}}||$. From (58), we get

$$\begin{aligned}
|\delta^{n+1}| \le & \frac{\left|\left(\frac{2\sigma_0 h^2}{2^{1-\alpha}} - h^2 v\bar{P}^{v-1}\bar{P}_x\right)\beta_0 - \beta_1 - 2i\beta_2\sin(h\theta) - 2\sigma_0 h^2\beta_0 r_1\right|}{\left|\left(\frac{2\sigma_0 h^2}{2^{1-\alpha}} + h^2 v\bar{P}^{v-1}\bar{P}_x\right)\beta_0 + \beta_1 + 2i\beta_2\sin(h\theta)\right|}|\delta^n| \\
& + \frac{|2h^2\beta_0|\left(\sigma_0\sum_{j=1}^{n-1}(r_{n-j} - r_{n-j+1})|\delta^j| + \sigma_0 r^n|\delta^0| + |r^{n+\frac{1}{2}}|\right)}{\left|\left(\frac{2\sigma_0 h^2}{2^{1-\alpha}} + h^2 v\bar{P}^{v-1}\bar{P}_x\right)\beta_0 + \beta_1 + 2i\beta_2\sin(h\theta)\right|} \\
\le & \frac{\left|\left(\frac{2\sigma_0 h^2}{2^{1-\alpha}} - h^2 v\bar{P}^{v-1}\bar{P}_x\right)\beta_0 - \beta_1 - 2i\beta_2\sin(h\theta) - 2\sigma_0 h^2\beta_0 r_1\right|}{\left|\left(\frac{2\sigma_0 h^2}{2^{1-\alpha}} + h^2 v\bar{P}^{v-1}\bar{P}_x\right)\beta_0 + \beta_1 + 2i\beta_2\sin(h\theta)\right|}C^1|r^{\frac{1}{2}}| \\
& + \frac{|2\sigma_0\beta_0 h^2 r_1|}{\left|\left(\frac{2\sigma_0 h^2}{2^{1-\alpha}} + h^2 v\bar{P}^{v-1}\bar{P}_x\right)\beta_0 + \beta_1 + 2i\beta_2\sin(h\theta)\right|}C^2|r^{\frac{1}{2}}| \\
& + \frac{|2h^2\beta_0|}{\left|\left(\frac{2\sigma_0 h^2}{2^{1-\alpha}} + h^2 v\bar{P}^{v-1}\bar{P}_x\right)\beta_0 + \beta_1 + 2i\beta_2\sin(h\theta)\right|}|r^{n+\frac{1}{2}}|. \qquad (59)
\end{aligned}$$

Using (41) and (57), we get that there exists a positive constant $C^{n+\frac{1}{2}}$, such that

$$|r^{n+\frac{1}{2}}| \le C^{n+\frac{1}{2}}|r^{\frac{1}{2}}|, \qquad n = 0, 1, 2, \ldots, M_t.$$

Let $C = \max\{C^1, C^2, C^{n+\frac{1}{2}}, n = 0, 1, 2, \ldots, M_t\}$, and let the time step size approach zero. Then we obtain $|\delta^n| \le C|r^{\frac{1}{2}}|$. □

Theorem 1. The proposed numerical scheme (35) is convergent, and the solution satisfies the condition $\|p^n - P^n\| \le C(h^2 + k^{2-\alpha}),\ n = 0, 1, 2, \ldots, M_t$.

Proof. Using the result of Lemma 1 and the Parseval's identity presented in (56) and (57), we get the following result:

$$\|p^n - P^n\| = \|\xi^n\|_2 \le C\|\mathcal{R}^{\frac{1}{2}}\|_2 = C(h^2 + k^{2-\alpha}).$$

Hence, the convergence order relies on the appropriate choice of parameters γ_1 and γ_2. When $\gamma_1 = \frac{1}{12}$ and $2\gamma_1 + \gamma_2 = 1$, the proposed numerical scheme is $O\left(h^2 + k^{2-\alpha}\right)$ accurate. □

Remark 2. Although the nonlinear parameter v can be significantly affect the qualitative behavior of continuous solution, the numerical scheme's convergence orders and stability condition are independent of v. This is because the quasi-linearization process places v only within bounded coefficients, and the error analysis relies solely on the boundedness of these coefficients rather than their specific values. The numerical results recorded in tables and figures confirm that for $v = 1, 2, \ldots, 10$, the errors and convergence rates remain comparable.

5 Numerical simulation

In this section, some numerical experiments are carried out to validate the theoretical results. Let $p(x_m, t_n)$ and $P(x_m, t_n)$ denote the analytical and numerical solutions to the problem (1) at the grid point (x_m, t_n). The errors are measured using the following error norms:

$$E_2^{N_x,M_t} = \max_{1\le n\le M_t} \sqrt{h \sum_{m=1}^{N_x-1} |p(x_m,t_n) - P(x_m,t_n)|^2},$$
$$E_\infty^{N_x,M_t} = \max_{1\le n\le M_t} \left(\max_{1\le m\le N_x-1} |p(x_m,t_n) - P(x_m,t_n)| \right).$$

For problems where the exact solution is not available, the efficiency of the proposed method is evaluated by computing the errors in the L_2 and L_∞-norms using the double mesh principle, defined as follows:

$$E_2^{N_x,M_t} = \max_{1\le n\le M_t} \sqrt{h \sum_{m=1}^{N_x-1} \left|P(x_m,t_n) - \bar{P}(x_{2m},t_{2n})\right|^2}, \tag{60}$$
$$E_\infty^{N_x,M_t} = \max_{1\le n\le M_t} \left(\max_{1\le m\le N_x-1} \left|P(x_m,t_n) - \bar{P}(x_{2m},t_{2n})\right| \right), \tag{61}$$

where $\bar{P}(x_{2m}, t_{2n})$, for $m = 0, 1, 2, \ldots, N_x, n = 0, 1, 2, \ldots, M_t$, denotes the numerical solution obtained on a refined grid by doubling the number of mesh points in both spatial and temporal directions. The spatial and temporal convergence orders of the method are computed using the following formulas:

$$L_q^x = \log_2 \left(\frac{E_q^{N_x,M_t}}{E_q^{2N_x,M_t}} \right), \text{ where } q = 2, \infty, \tag{62}$$

$$L_q^t = \log_2 \left(\frac{E_q^{N_x, M_t}}{E_q^{N_x, 2M_t}} \right), \text{ where } q = 2, \infty. \tag{63}$$

Furthermore, we investigate the combined spatial and temporal convergence orders of the method using the formula:

$$\rho^{N_x, M_t} = \log_2 \left(\frac{E_q^{N_x, M_t}}{E_q^{2N_x, 2M_t}} \right), \text{ where } q = 2, \infty. \tag{64}$$

Based on the structure of the proposed method, a quasi-linearized system of equations is solved at each time level. In this study, the proposed numerical scheme is combined with the Thomas algorithm as an iterative solver to obtain the numerical solution. In practice, the stopping criteria is defined as $\|p_m^n - P_m^n\| \leq 10^{-8}$, where p_m^n denotes the exact solution and P_m^n denotes the numerical solution. For all numerical implementations in this paper, we use the parameter values $\gamma_1 = \frac{1}{12}$ and $\gamma_2 = \frac{10}{12}$.

Example 1. Consider the following generalized nonlinear time-fractional convection-diffusion equation [13]:

$${}_0^C D_t^\alpha p(x,t) + (p(x,t))^v \frac{\partial p(x,t)}{\partial x} = d \frac{\partial^2 p(x,t)}{\partial x^2} + f(x,t), (x,t) \in (0,1) \times (0,1),$$

with initial and boundary conditions

$$\begin{aligned} p(x,0) &= 0, \quad 0 \leq x \leq 1, \\ p(0,t) &= t^2, \quad p(1,t) = et^2, \quad t \geq 0. \end{aligned}$$

The source term is defined as $f(x,t) = \frac{2t^{2-\alpha}e^x}{\Gamma(3-\alpha)} + (t^2 e^x)^{v+1} - dt^2 e^x$. For special value of $\alpha = 1$, the exact solution is $p(x,t) = e^x t^2$.

For this example, the numerical experiments are conducted for various values of α, v, and the mesh sizes, and the corresponding results are presented in tables and figures. Since the exact solution is available only for $\alpha = 1$, the double mesh principle is employed for other values of α. Tables 1–4 present a comprehensive numerical study of the proposed scheme with respect to temporal and spatial discretizations, as well as the influence of the nonlinear convection parameter v and the fractional order α. Table 1 reports the temporal errors, convergence orders, and CPU times of the computed solution for different values of α with a sufficiently fine spatial grid $N_x = 1000$ using formula in (63), ensuring that spatial errors do not influence temporal accuracy. It is observed that the errors in both L_2^t and L_∞^t-norms decrease monotonically as the

time step size is refined, confirming the stability and convergence of the temporal discretization. The computed convergence rates are in good agreement with the theoretical order $2-\alpha$. In particular, for a smaller fractional order $\alpha = 0.2$, the theoretical rate is $2-\alpha = 1.8$, and observe rates are initially close to this value and then slightly decrease toward 1.8 as the mesh is further refined, indicating convergence toward the theoretical estimate. A similar trend is observed for other values of α: for $\alpha = 0.5$, the rates approach the expected order 1.5, while for $\alpha = 0.5$ and $\alpha = 0.9$, the computed rates consistently align with the theoretical orders 1.3 and 1.1, respectively.

Regarding the stability condition stated in Remark 3, namely $3^{\alpha} \geq \frac{3}{2}$ (implying $\alpha \geq 0.37$), it is important to emphasize that this is a sufficient not necessary condition. The numerical results clearly show that the scheme remains stable and accurate even for $\alpha = 0.2$, where the condition is violated. This confirms that the theoretical stability bound is conservative, as commonly observed in von Neumann-type analyses and energy-based estimates. The inclusion of $\alpha = 0.2$ is therefore deliberate, demonstrating the robustness of the method beyond the theoretically guarantee range.

Table 1: Temporal errors, convergence orders, and CPU times of the computed solution for Example 1 with different fractional orders α and time step sizes M_t, using $N_x = 1000$ and $d = v = 1$.

α	Norms	$M_t = 32$	$M_t = 64$	$M_t = 128$	$M_t = 256$	$M_t = 512$	$M_t = 1024$
	L_2^t	1.14×10^{-4}	3.01×10^{-5}	7.81×10^{-6}	2.02×10^{-6}	5.79×10^{-7}	1.66×10^{-7}
		1.928	1.945	1.949	1.805	1.798	—
0.2	L_∞^t	1.60×10^{-4}	4.21×10^{-5}	1.09×10^{-5}	2.82×10^{-6}	7.87×10^{-7}	2.26×10^{-7}
		1.924	1.947	1.953	1.843	1.800	—
	CUP(s)	3.3953	6.9406	17.5882	41.9596	128.6501	420.4133
	L_2^t	1.98×10^{-4}	6.21×10^{-5}	1.85×10^{-5}	5.55×10^{-6}	1.82×10^{-6}	6.03×10^{-7}
		1.676	1.744	1.740	1.609	1.593	—
0.5	L_∞^t	2.67×10^{-4}	8.33×10^{-5}	2.47×10^{-5}	7.65×10^{-6}	2.50×10^{-6}	8.28×10^{-7}
		1.684	1.752	1.692	1.6114	1.598	—
	CUP(s)	0.5482	2.8717	6.5998	5.9386	32.0136	66.3795
	L_2^t	2.53×10^{-4}	9.11×10^{-5}	3.40×10^{-5}	1.31×10^{-5}	5.26×10^{-6}	2.12×10^{-6}
		1.475	1.419	1.373	1.321	1.306	—
0.7	L_∞^t	3.50×10^{-4}	1.25×10^{-4}	4.69×10^{-5}	1.80×10^{-5}	7.21×10^{-6}	2.91×10^{-6}
		1.478	1.421	1.378	1.324	1.307	—
	CUP(s)	0.1335	0.2927	0.6866	1.7674	4.3070	14.4766
	L_2^t	2.43×10^{-4}	9.86×10^{-5}	4.23×10^{-5}	1.92×10^{-5}	8.98×10^{-6}	4.19×10^{-6}
		1.300	1.220	1.136	1.102	1.099	—
0.9	L_∞^t	3.35×10^{-4}	1.36×10^{-4}	5.83×10^{-5}	2.64×10^{-5}	1.23×10^{-6}	5.74×10^{-6}
		1.304	1.222	1.142	1.102	1.099	—
	CUP(s)	0.1335	0.2927	0.6866	1.7674	4.3070	14.4766

Table 2: Spatial errors, convergence orders, and CPU times of the computed solution for Example 1 with different values of α and the space step sizes N_x and $M_t = 1000$ with $d = v = 1$.

α	Norms	$N_x = 4$	$N_x = 8$	$N_x = 16$	$N_x = 32$	$N_x = 64$
	L_2^x	1.13×10^{-3}	2.73×10^{-4}	6.76×10^{-5}	1.67×10^{-5}	4.07×10^{-6}
		2.047	2.017	2.013	2.039	—
0.2	L_∞^x	1.53×10^{-3}	3.82×10^{-4}	9.45×10^{-5}	2.35×10^{-5}	5.72×10^{-6}
		2.005	2.016	2.008	2.039	—
	CUP(s)	17.2782	14.5134	14.3909	14.6191	12.0917
	L_2^x	1.09×10^{-3}	2.65×10^{-4}	6.53×10^{-5}	1.60×10^{-5}	3.74×10^{-6}
		2.047	2.020	2.027	2.099	—
0.4	L_∞^x	1.49×10^{-3}	3.70×10^{-4}	9.14×10^{-5}	2.25×10^{-5}	5.25×10^{-6}
		2.010	2.017	2.022	2.098	—
	CUP(s)	17.0886	14.4452	14.3717	14.4228	11.8596
	L_2^x	1.02×10^{-3}	2.47×10^{-4}	5.96×10^{-5}	1.33×10^{-5}	2.12×10^{-6}
		2.053	2.052	2.165	2.648	—
0.7	L_∞^x	1.40×10^{-3}	3.44×10^{-4}	8.36×10^{-5}	1.87×10^{-5}	2.90×10^{-6}
		2.024	2.044	2.158	2.689	—
	CUP(s)	17.0548	14.6582	14.3671	14.3693	11.8885

Table 2 presents the spatial discretization errors and CPU times of the computed solution using a sufficiently fine time step $M_t = 1000$, ensuring that the temporal errors do not influence spatial accuracy. The results clearly demonstrate that the method achieves second-order accuracy in space for all considered values of α. The observed convergence rates remain uniformly close to 2 across all refinement levels, confirming that the spatial discretization is independent of the fractional-order α and is not affected by temporal memory effects.

Tables 3 and 4 examine the effect of the nonlinear parameter v in the term $p^v p_x$. It is observed Table 3 that the errors in both L_2 and L_∞-norms decrease consistently with mesh refinement, and the corresponding convergence rates remain stable for all values $v = 1, 2, \ldots, 10$. This indicates that the proposed scheme maintains its accuracy even for strongly nonlinear cases. Moreover, under further mesh refinement, and as also observed in Table 1, the numerical convergence orders approach the expected theoretical values more closely. Table 4 reports pointwise errors for fixed discretization parameters $N_x = 160$ and $M_t = 256$, for various values of α and v. The results indicates that the numerical errors remain uniformly small across all tested values of α and v, demonstrating the stability and robustness of the scheme. As α increases, a slight improvement in accuracy is observed, which is consistent with the smoother behavior of solutions for larger fractional orders. In line with Remark 2, the convergence behavior and stability of the scheme are not significantly influenced by v. This is reflected in the numerical results, where the errors and convergence rates remain comparable across all tested values of v, demonstrating

Table 3: Pointwise errors and corresponding convergence orders for Example 1 in L_2 and L_∞ norms for different values of v with $\alpha = 0.5$ and $d = 1$.

v	Norms	$N_x = 10$ $M_t = 16$	$N_x = 20$ $M_t = 32$	$N_x = 40$ $M_t = 64$	$N_x = 80$ $M_t = 128$	$N_x = 160$ $M_t = 256$	$N_x = 320$ $M_t = 512$
1	L_2	1.533×10^{-3}	5.262×10^{-4}	1.684×10^{-4}	5.110×10^{-5}	1.487×10^{-5}	4.186×10^{-6}
		1.54	1.64	1.72	1.78	1.83	
	L_∞	2.069×10^{-3}	7.071×10^{-4}	2.254×10^{-4}	6.800×10^{-5}	1.968×10^{-5}	5.515×10^{-6}
		1.55	1.65	1.73	1.79	1.84	—
2	L_2	1.522×10^{-3}	5.256×10^{-4}	1.684×10^{-4}	5.110×10^{-5}	1.487×10^{-5}	4.186×10^{-6}
		1.53	1.64	1.72	1.78	1.83	—
	L_∞	2.053×10^{-3}	7.063×10^{-4}	2.253×10^{-4}	6.800×10^{-5}	1.968×10^{-5}	5.515×10^{-6}
		1.54	1.65	1.73	1.79	1.84	—
3	L_2	5.567×10^{-3}	1.219×10^{-3}	2.827×10^{-4}	6.651×10^{-5}	1.542×10^{-5}	4.186×10^{-6}
		2.19	2.11	2.09	2.11	1.88	—
	L_∞	9.333×10^{-3}	1.958×10^{-3}	4.482×10^{-4}	1.063×10^{-4}	2.519×10^{-5}	5.842×10^{-6}
		2.25	2.13	2.08	2.08	2.11	—
4	L_2	1.289×10^{-2}	2.635×10^{-3}	6.058×10^{-4}	1.457×10^{-4}	3.551×10^{-5}	8.629×10^{-6}
		2.29	2.12	2.06	2.04	2.04	—
	L_∞	2.445×10^{-2}	5.568×10^{-3}	1.142×10^{-3}	2.415×10^{-4}	5.901×10^{-5}	1.450×10^{-5}
		2.13	2.29	2.24	2.03	2.02	—
⋮	⋮	⋮	⋮	⋮	⋮	⋮	⋮
10	L_2	4.903×10^{-2}	1.245×10^{-2}	2.924×10^{-3}	7.070×10^{-4}	1.696×10^{-4}	4.004×10^{-5}
		1.98	2.09	2.05	2.06	2.08	—
	L_∞	9.751×10^{-2}	2.833×10^{-2}	8.120×10^{-3}	2.208×10^{-3}	5.768×10^{-4}	1.460×10^{-4}
		1.78	1.80	1.88	1.94	1.98	—

Table 4: Pointwise errors for Example 1 in L_2 and L_∞ norms for different values of α and v using $N_x = 160, M_t = 256$ and $d = 1$.

v	Norms	$\alpha = 0.4$	$\alpha = 0.6$	$\alpha = 0.8$	$\alpha = 0.9$	$\alpha = 1$
1	L_2	1.367×10^{-5}	1.421×10^{-5}	2.402×10^{-5}	2.219×10^{-5}	4.760×10^{-7}
	L_∞	1.830×10^{-5}	1.941×10^{-5}	3.294×10^{-5}	3.045×10^{-5}	6.528×10^{-7}
2	L_2	1.367×10^{-5}	1.421×10^{-5}	2.325×10^{-5}	2.137×10^{-5}	3.834×10^{-6}
	L_∞	1.830×10^{-5}	1.866×10^{-5}	3.178×10^{-5}	2.921×10^{-5}	5.495×10^{-6}
3	L_2	1.714×10^{-5}	1.421×10^{-5}	2.306×10^{-5}	2.116×10^{-5}	1.792×10^{-5}
	L_∞	2.714×10^{-5}	2.248×10^{-5}	3.155×10^{-5}	2.895×10^{-5}	2.776×10^{-5}
4	L_2	3.654×10^{-5}	3.409×10^{-5}	3.086×10^{-5}	3.104×10^{-5}	3.680×10^{-5}
	L_∞	5.992×10^{-5}	5.778×10^{-5}	5.496×10^{-5}	5.513×10^{-5}	6.004×10^{-5}
⋮	⋮	⋮	⋮	⋮	⋮	⋮
10	L_2	1.697×10^{-4}	1.695×10^{-4}	1.692×10^{-4}	1.692×10^{-4}	1.695×10^{-4}
	L_∞	5.768×10^{-4}	5.768×10^{-4}	5.767×10^{-4}	5.767×10^{-4}	5.767×10^{-4}

the robustness of the method with respect to nonlinear effects. A numerical comparison of the suggested method with the current numerical approaches documented in [13], [21], [22], and [35] is presented in Tables 5–7. Comparing the suggested approach to the previously published schemes, the numerical results show that it achieves better convergence characteristics and higher accuracy. Results specifically show that the current approach is robust and efficient in all tested problem settings, confirming its efficacy and potential benefits for real-world applications across a broad range of v values.

Table 5: Comparison results for Example 1 in the L_2 and L_∞ norms for $\alpha = 0.4, d = 1$, and different values of v and mesh sizes.

v	Norms	$N_x = 16, M_t = 10$		$N_x = 32, M_t = 20$		$N_x = 64, M_t = 40$	
		Method [13]	Present	Method [13]	Present	Method [13]	Present
2	L_2	1.08×10^{-2}	6.314×10^{-3}	2.53×10^{-3}	1.338×10^{-3}	6.15×10^{-4}	3.614×10^{-4}
	L_∞	1.73×10^{-2}	8.962×10^{-3}	4.09×10^{-3}	1.917×10^{-3}	9.92×10^{-4}	5.141×10^{-4}
3	L_2	1.61×10^{-2}	8.561×10^{-3}	3.75×10^{-3}	1.611×10^{-3}	9.15×10^{-4}	4.148×10^{-4}
	L_∞	3.56×10^{-2}	1.263×10^{-2}	8.64×10^{-3}	2.482×10^{-3}	2.10×10^{-3}	6.167×10^{-4}

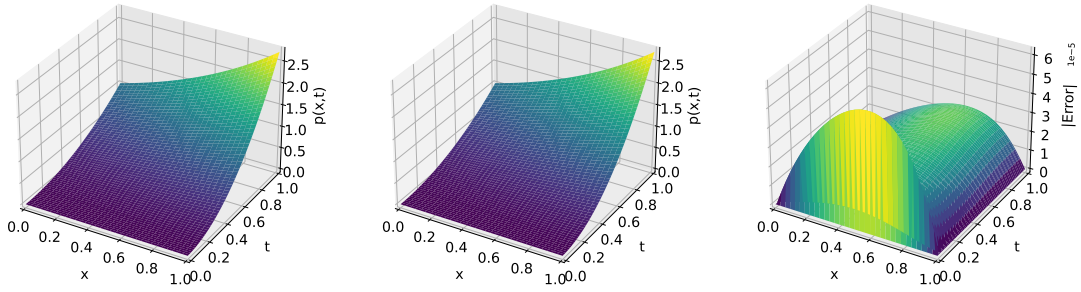
Table 6: L_∞ and L_2 errors comparisons for Example 1 with $M_t = 4000$, $\alpha = 0.5$ and $v = 1$.

N_x	Method [21]		Method [22]		Present method	
	L_2	L_∞	L_2	L_∞	L_2	L_∞
10	1.633×10^{-3}	2.296×10^{-3}	1.765×10^{-3}	3.101×10^{-3}	5.136×10^{-4}	7.113×10^{-4}
20	4.477×10^{-4}	6.250×10^{-4}	4.056×10^{-4}	8.128×10^{-4}	1.246×10^{-4}	1.737×10^{-4}
40	1.618×10^{-4}	2.273×10^{-4}	6.774×10^{-5}	2.095×10^{-4}	3.081×10^{-5}	4.320×10^{-5}
80	9.262×10^{-5}	1.331×10^{-4}	4.575×10^{-5}	6.920×10^{-5}	7.580×10^{-6}	1.063×10^{-5}

Table 7: L_∞ and L_2 errors comparisons for Example 1 with $\alpha = 0.5, v = 1$ and $N_x = 40$.

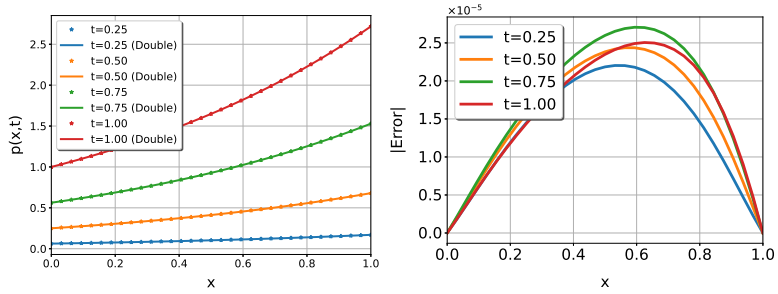
M_t	Method [35]		Present method	
	L_2	L_∞	L_2	L_∞
1000	9.204×10^{-4}	2.401×10^{-3}	7.474×10^{-6}	1.089×10^{-5}
2000	4.549×10^{-4}	1.198×10^{-3}	6.639×10^{-6}	9.759×10^{-6}
4000	2.225×10^{-4}	5.972×10^{-4}	6.363×10^{-6}	9.385×10^{-6}
8000	8.344×10^{-5}	2.364×10^{-4}	6.271×10^{-6}	9.259×10^{-6}

Figure 1 displays the numerical solution and absolute error surface plots, while Figure 2 displays the two-dimensional (2D) numerical solution and error plots for Example 1 at $\alpha = 0.4, d = v = 1$. Graphs of absolute errors for various values of α and v at various time levels $t = [0.25, 0.5, 0.75, 1.0]$ are shown in Figures 3–8. All errors in Figures 2–8 are computed using the double mesh principle, based on the difference between the numerical solutions obtained on the meshes $(N_x, M_t) = (64, 128)$ and $(N_x, M_t) = (128, 256)$ for Example 1.



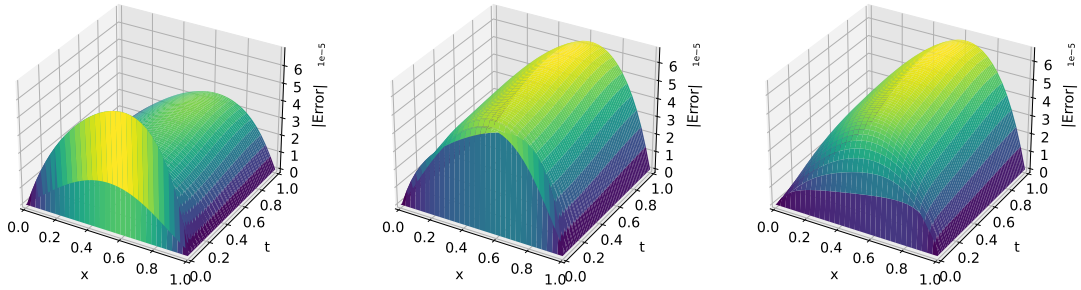
(a) $N_x = 64, M_t = 128$ (b) $N_x = 128, M_t = 256$ (c) Error

Figure 1: 3D representations of the solution behavior and corresponding absolute error for Example 1 with $\alpha = 0.4, v = 1$ and $d = 1$.



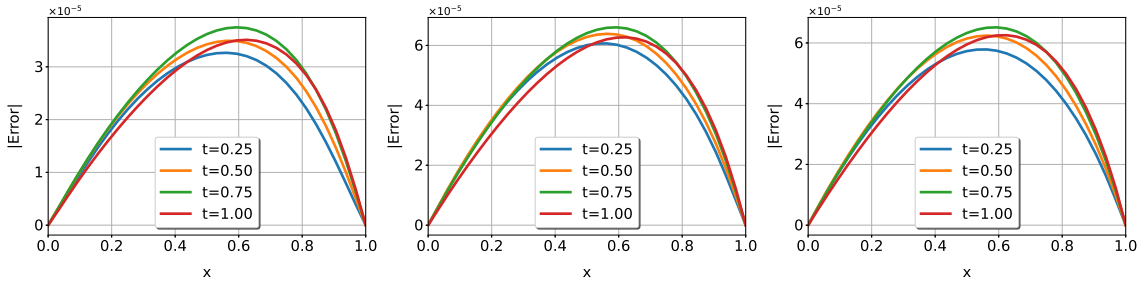
(a) Solution comparison (b) Error

Figure 2: 2D representations of the solution comparison and the corresponding absolute error for Example 1 with $\alpha = 0.4, v = 1$ and $d = 1$.



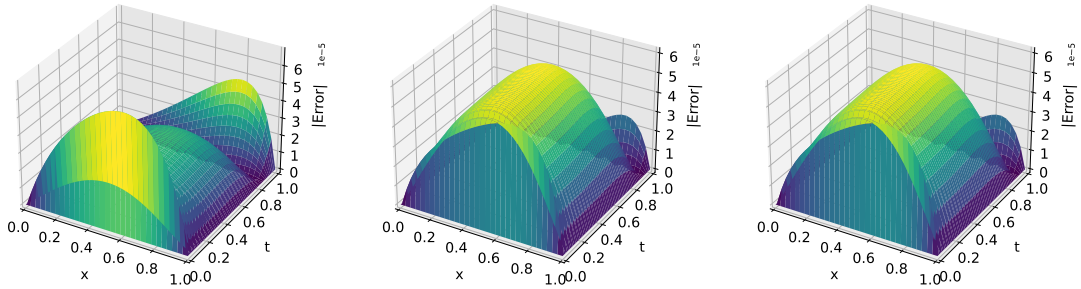
(a) Error for $\alpha = 0.5$ (b) Error for $\alpha = 0.7$ (c) Error for $\alpha = 0.9$

Figure 3: 3D representations of the absolute error behaviors for Example 1 with $v = 1$ and $d = 1$.



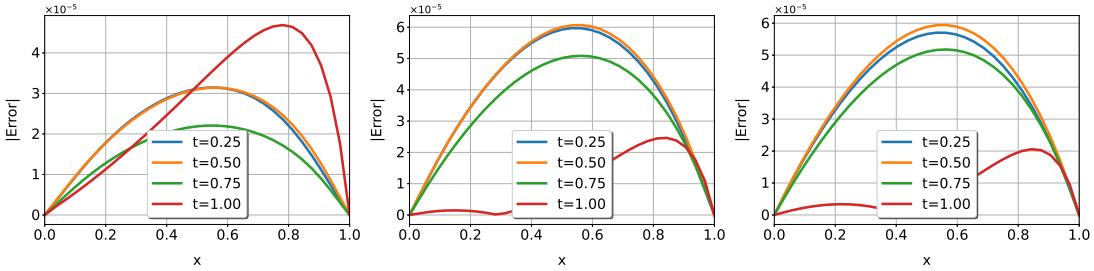
(a) Error for $\alpha = 0.5$ (b) Error for $\alpha = 0.7$ (c) Error for $\alpha = 0.9$

Figure 4: 2D representations of the absolute error behaviors for Example 1 at $v = d = 1$.



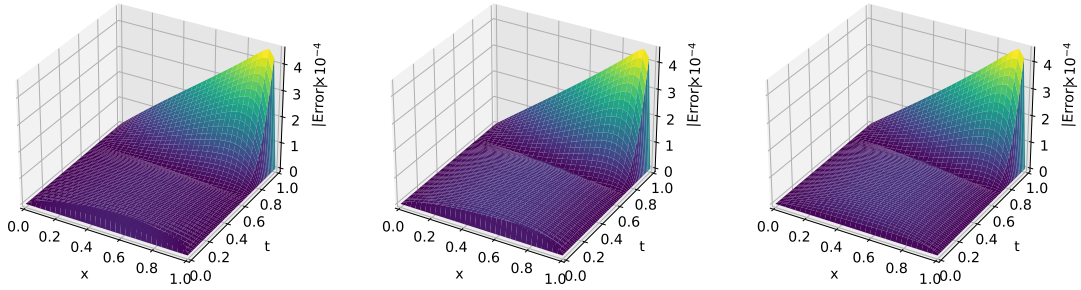
(a) Error for $\alpha = 0.5$ (b) Error for $\alpha = 0.7$ (c) Error for $\alpha = 0.9$

Figure 5: 3D representations of the absolute error behaviors for Example 1 at $v = 2, d = 1$.



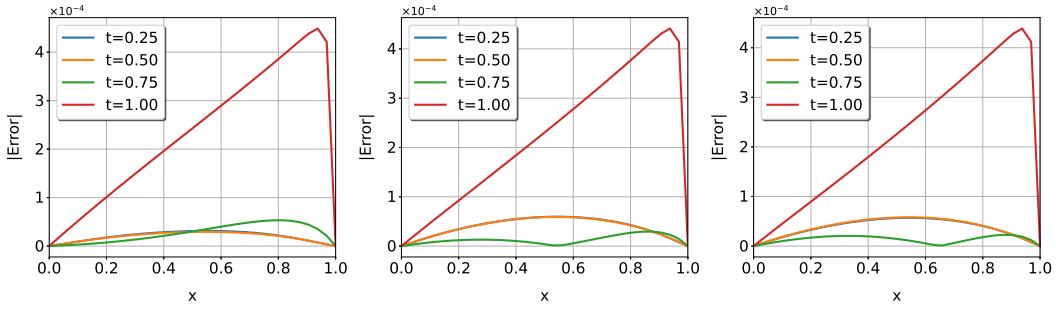
(a) Error for $\alpha = 0.5$ (b) Error for $\alpha = 0.7$ (c) Error for $\alpha = 0.9$

Figure 6: 2D representations of the absolute error behaviors for Example 1 at $v = 2, d = 1$.



(a) Error for $\alpha = 0.5$ (b) Error for $\alpha = 0.7$ (c) Error for $\alpha = 0.9$

Figure 7: 3D representations of the absolute error behaviors for Example 1 at $v = 4, d = 1$.



(a) Error for $\alpha = 0.5$ (b) Error for $\alpha = 0.7$ (c) Error for $\alpha = 0.9$

Figure 8: 2D representations of the absolute error behaviors for Example 1 at $v = 4, d = 1$.

Example 2. Consider the following generalized nonlinear time-fractional convection-diffusion equation [13]:

$$ {}_0^C D_t^\alpha p(x,t) + (p(x,t))^v \frac{\partial p(x,t)}{\partial x} = d\frac{\partial^2 p(x,t)}{\partial x^2} + f(x,t), (x,t) \in (0,1)\times(0,1), $$

with initial and boundary conditions

$$ p(x,0) = 0, \quad 0 \le x \le 1, $$
$$ p(0,t) = 0, \quad p(1,t) = 0, \quad t \ge 0, $$

where the source term is $f(x,t) = \frac{2t^{2-\alpha}\sin(2\pi x)}{\Gamma(3-\alpha)} + 2\pi t^4 \sin(2\pi x)\cos(2\pi x) + 4dt^2\pi^2 \sin(2\pi x)$, and $p(x,t) = t^2 \sin(2\pi x)$ is the exact solution of this problem for special case $\alpha = 1$ and $v = 1$ [35, 25].

Table 8: Pointwise errors and corresponding convergence orders for Example 2 in the L_2 and L_∞ norms for $\alpha = 0.5, d = 1$, and different values of v and mesh sizes.

v	Norms	$N_x = 10$ $M_t = 16$	$N_x = 20$ $M_t = 32$	$N_x = 40$ $M_t = 64$	$N_x = 80$ $M_t = 128$	$N_x = 160$ $M_t = 256$	$N_x = 320$ $M_t = 512$
1	L_2	6.03×10^{-3}	1.24×10^{-3}	3.02×10^{-4}	7.24×10^{-5}	1.72×10^{-5}	4.08×10^{-6}
		2.27	2.05	2.06	2.07	2.08	—
	L_∞	8.65×10^{-3}	1.67×10^{-3}	4.27×10^{-4}	1.02×10^{-4}	2.44×10^{-5}	5.77×10^{-6}
		2.37	1.97	2.06	2.07	2.08	—
2	L_2	1.66×10^{-2}	3.87×10^{-3}	1.01×10^{-3}	2.59×10^{-4}	6.61×10^{-5}	1.68×10^{-5}
		2.10	1.93	1.97	1.98	1.97	—
	L_∞	2.53×10^{-2}	6.56×10^{-3}	1.71×10^{-3}	4.36×10^{-4}	1.11×10^{-4}	2.82×10^{-5}
		1.95	1.94	1.97	1.98	1.97	—
3	L_2	5.07×10^{-3}	1.24×10^{-3}	3.02×10^{-4}	7.24×10^{-5}	1.72×10^{-5}	4.08×10^{-6}
		2.02	2.05	2.06	2.07	2.08	—
	L_∞	6.83×10^{-3}	1.67×10^{-3}	4.27×10^{-4}	1.02×10^{-4}	2.44×10^{-5}	5.77×10^{-6}
		2.02	1.97	2.06	2.07	2.08	—
4	L_2	1.95×10^{-2}	2.80×10^{-3}	8.02×10^{-4}	2.10×10^{-4}	5.41×10^{-5}	1.39×10^{-5}
		2.80	1.80	1.93	1.96	1.96	—
	L_∞	2.61×10^{-2}	4.59×10^{-3}	1.33×10^{-3}	3.51×10^{-4}	9.03×10^{-5}	2.32×10^{-5}
		2.51	1.78	1.93	1.96	1.96	—
⋮	⋮	⋮	⋮	⋮	⋮	⋮	⋮
10	L_2	7.05×10^{-3}	1.61×10^{-3}	5.90×10^{-4}	1.63×10^{-4}	4.31×10^{-5}	1.13×10^{-5}
		2.12	1.46	1.85	1.92	1.93	—
	L_∞	1.41×10^{-2}	2.57×10^{-3}	1.00×10^{-3}	2.79×10^{-4}	7.34×10^{-5}	1.92×10^{-5}
		2.46	1.36	1.84	1.93	1.93	—

Table 8 represents the pointwise errors and corresponding convergence orders for Example 2 in the L_2 and L_∞ norms for different values of the nonlinear parameter v and progressively refined mesh sizes. It is observed that the errors decrease consistently as the mesh size is refined, and the computed convergence orders approach second-order accuracy in both norms. This confirms that the proposed scheme maintains its expected spatial accuracy while simultaneously

capturing the temporal behavior under mesh refinement. Moreover, the results demonstrate that the convergence rates remain stable across a wide range of nonlinearities $v = 1, 2, \ldots, 10$. Minor variations in the convergence order for coarse meshes are observed, particularly for large values of v, but these effects diminish as the mesh is refined. This behavior is consistent with the pre-asymptotic regime and does not affect the overall convergence trend. Table 9 reports pointwise errors for fixed discretization parameters $N_x = 160$ and $M_t = 256$, for various values of α and v. The results indicate that the numerical errors remain uniformly small across all tested values of α and v, demonstrating the stability and robustness of the scheme. As α increases, a slight improvement in accuracy is observed, which is consistent with the smoother behavior of solutions for larger fractional orders. In line with Remark 2, the numerical results in both tables confirm that the convergence behavior and stability of the scheme are essentially independent of the nonlinear parameter v. The comparable error magnitudes and consistent convergence rates across all tested values of v demonstrate that the proposed method effectively handles nonlinear effects without loss of accuracy. Overall, Tables 8 and 9 further validate the theoretical findings, showing that the scheme is second-order accurate, stable, and robust for a wide range of fractional orders and nonlinearities.

Table 9: Pointwise errors for Example 2 in L_2 and L_∞ norms for $N_x = 160, M_t = 256, d = 1$, and different values of α and v.

v	Norms	$\alpha = 0.4$	$\alpha = 0.6$	$\alpha = 0.8$	$\alpha = 0.9$	$\alpha = 1$
1	L_2	1.92×10^{-5}	1.45×10^{-5}	1.24×10^{-5}	1.21×10^{-5}	3.31×10^{-6}
	L_∞	2.72×10^{-5}	2.05×10^{-5}	2.01×10^{-5}	1.98×10^{-5}	5.59×10^{-6}
2	L_2	6.72×10^{-5}	6.49×10^{-5}	6.16×10^{-5}	5.72×10^{-5}	4.78×10^{-5}
	L_∞	1.12×10^{-5}	1.09×10^{-5}	1.04×10^{-5}	9.79×10^{-5}	8.38×10^{-5}
3	L_2	1.92×10^{-5}	1.45×10^{-5}	1.33×10^{-5}	1.31×10^{-5}	1.06×10^{-5}
	L_∞	2.72×10^{-5}	2.05×10^{-5}	1.88×10^{-5}	1.85×10^{-5}	1.60×10^{-5}
4	L_2	5.56×10^{-5}	5.25×10^{-5}	4.74×10^{-5}	4.20×10^{-5}	3.23×10^{-5}
	L_∞	9.23×10^{-5}	8.81×10^{-5}	3.25×10^{-5}	7.27×10^{-5}	5.70×10^{-5}
⋮	⋮	⋮	⋮	⋮	⋮	⋮
10	L_2	4.54×10^{-5}	4.03×10^{-5}	3.25×10^{-5}	2.68×10^{-5}	1.91×10^{-5}
	L_∞	7.65×10^{-5}	6.97×10^{-5}	5.89×10^{-5}	4.98×10^{-5}	3.61×10^{-5}

A numerical comparison of the suggested method with the current numerical approaches documented in [22, 35, 25, 44, 21] is presented in Tables 10–14. The numerical results demonstrate that the proposed method achieves higher accuracy compared with the previously reported schemes. Specifically, the findings support its efficacy and possible benefits for real-world uses across a broad range of v values. The tension parameters of the nonpolynomial spline offer a vital

benefit over a conventional cubic B-spline because they can prevent unwarranted spurious oscillations. The key feature of tension spline is the tension parameter μ, which allows local control of the spline shape and can effectively suppress oscillations that often arise in convection-dominated regimes. Tables 6, 10 and 14 demonstrate that our algorithm can be better used than common B-spline algorithms [21, 22].

Table 10: Example 2: L_∞ and L_2 errors for $M_t = 4000$, $\alpha = 0.5$, and $v = d = 1$.

N_x	Method [22]		Method [35]		Present method	
	L_2	L_∞	L_2	L_∞	L_2	L_∞
10			2.26×10^{-2}	3.04×10^{-2}	7.14×10^{-3}	1.01×10^{-2}
20			5.44×10^{-3}	7.70×10^{-3}	4.77×10^{-4}	7.48×10^{-4}
40	1.22×10^{-3}	1.73×10^{-3}	1.22×10^{-3}	1.72×10^{-3}	8.20×10^{-5}	1.31×10^{-4}
80	1.77×10^{-4}	2.53×10^{-4}	1.68×10^{-4}	2.38×10^{-4}	2.09×10^{-5}	3.02×10^{-5}
100	5.22×10^{-5}	7.65×10^{-5}	4.23×10^{-5}	5.99×10^{-5}	1.35×10^{-5}	1.90×10^{-5}

Table 11: Comparison for Example 2 in L_2 and L_∞ norms for different values of α with $M_t = 1000$ and $v = d = 1$.

α	Norms	$N_x = 8$		$N_x = 16$		$N_x = 32$	
		Method [25]	Present	Method [25]	Present	Method [25]	Present
0.5	L_2	3.32×10^{-2}	1.80×10^{-2}	8.14×10^{-3}	1.06×10^{-3}	2.02×10^{-3}	1.31×10^{-4}
	L_∞	4.70×10^{-2}	2.54×10^{-2}	1.15×10^{-2}	1.44×10^{-3}	2.86×10^{-3}	2.26×10^{-4}
0.9	L_2	3.29×10^{-2}	1.78×10^{-2}	8.06×10^{-3}	1.05×10^{-3}	2.00×10^{-3}	1.29×10^{-4}
	L_∞	4.65×10^{-2}	2.52×10^{-2}	1.14×10^{-2}	1.43×10^{-3}	2.83×10^{-3}	2.24×10^{-4}

Table 12: Comparison for Example 2 in L_2 and L_∞ norms for different values of α with $M_t = N_x$ and $v = d = 1$.

α	Norms	$N_x = 10$		$N_x = 20$		$N_x = 40$	
		Method [25]	Present	Method [25]	Present	Method [25]	Present
0.5	L_2	2.11×10^{-2}	1.31×10^{-2}	5.21×10^{-3}	3.25×10^{-3}	1.30×10^{-3}	7.91×10^{-4}
	L_∞	2.88×10^{-2}	1.77×10^{-2}	7.37×10^{-3}	4.37×10^{-3}	1.88×10^{-3}	1.11×10^{-3}
0.9	L_2	2.08×10^{-2}	1.05×10^{-2}	5.15×10^{-3}	2.26×10^{-3}	1.28×10^{-3}	4.55×10^{-4}
	L_∞	2.84×10^{-2}	1.42×10^{-2}	7.28×10^{-3}	3.05×10^{-3}	1.81×10^{-3}	6.43×10^{-4}

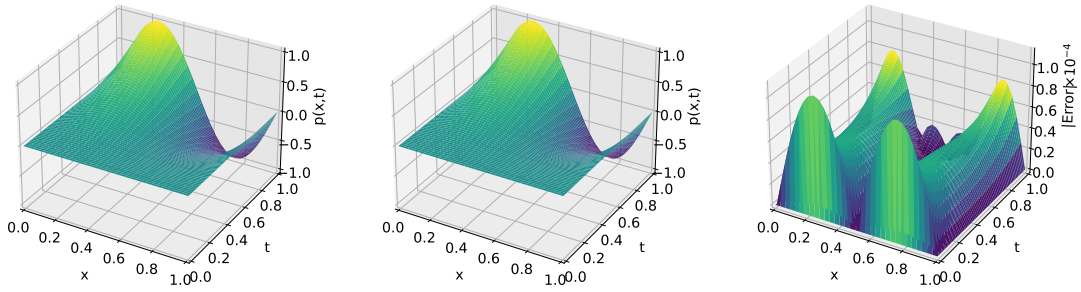
In Figure 9, we display the surface plots of the numerical solution and the absolute error for Example 1 at $\alpha = 0.4, d = v = 1$. All errors in Figures 10–13 are computed using the double mesh principle, based on the difference between the numerical solutions obtained on the meshes $(N_x, M_t) = (64, 128)$ and $(N_x, M_t) = (128, 256)$ for Example 2. In Figures 10–13 illustrate the distributions of absolute errors for different values of α and v.

Table 13: Comparison for Example 2 in L_2 and L_∞ norms for different values of α with $h = 0.01$ and $v = 1, d = 1$.

t	Norms	$\alpha = 0.5$		$\alpha = 1$	
		Method [44]	Present	Method [44]	Present
0.02	L_2	3.764×10^{-3}	1.206×10^{-4}	4.333×10^{-3}	5.332×10^{-5}
	L_∞	8.596×10^{-3}	1.705×10^{-4}	9.738×10^{-3}	7.597×10^{-5}
0.04	L_2	2.577×10^{-3}	5.011×10^{-4}	2.968×10^{-3}	2.123×10^{-4}
	L_∞	5.303×10^{-3}	7.087×10^{-4}	6.010×10^{-3}	3.003×10^{-4}
0.06	L_2	2.166×10^{-3}	1.249×10^{-3}	2.494×10^{-3}	5.998×10^{-4}
	L_∞	5.303×10^{-3}	1.766×10^{-3}	6.010×10^{-3}	8.483×10^{-4}
0.08	L_2	3.921×10^{-3}	2.246×10^{-3}	4.513×10^{-3}	1.223×10^{-3}
	L_∞	8.596×10^{-3}	3.176×10^{-3}	9.738×10^{-3}	1.729×10^{-3}

Table 14: Comparison for Example 2 in L_2 and L_∞ norms for $\alpha = 0.5$ with $N_x = 120, M_t = 2000$, $v = d = 1$ and $T = 0.1$.

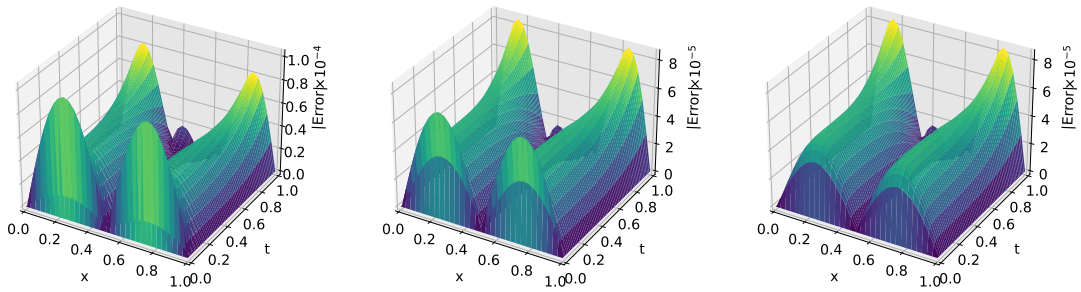
d	Method [21]		Method [25]		Present method	
	L_2	L_∞	L_2	L_∞	L_2	L_∞
0.1	1.501×10^{-5}	2.126×10^{-5}	6.666×10^{-7}	9.427×10^{-7}	1.185×10^{-8}	2.019×10^{-8}
0.01	1.045×10^{-6}	2.001×10^{-6}	1.338×10^{-7}	1.919×10^{-7}	2.744×10^{-8}	4.775×10^{-8}
0.005	7.583×10^{-7}	2.341×10^{-6}	8.573×10^{-8}	1.261×10^{-7}	2.947×10^{-8}	5.143×10^{-8}



(a) $N_x = 64, M_t = 128$ (b) $N_x = 128, M_t = 256$ (c) Error

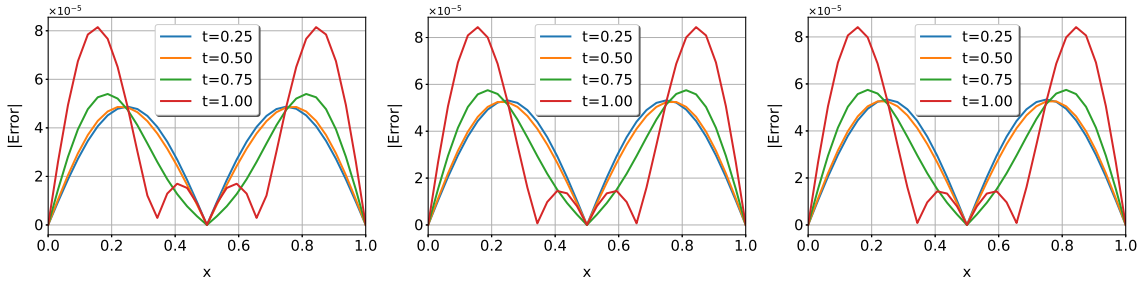
Figure 9: 3D representations of solution behaviors and absolute error corresponding to Example 2 for $\alpha = 0.4$ and $v = d = 1$.

Remark 3. The stability condition, which is $3^\alpha \geq \frac{3}{2}$ and leads to $\alpha \geq 0.37$, is not a necessary condition; rather, it is sufficient. It can be observed from Tables 1 and 2 that for $\alpha = 0.2$, the condition $3^\alpha \geq \frac{3}{2}$ is violated, but the results are of the same precision.



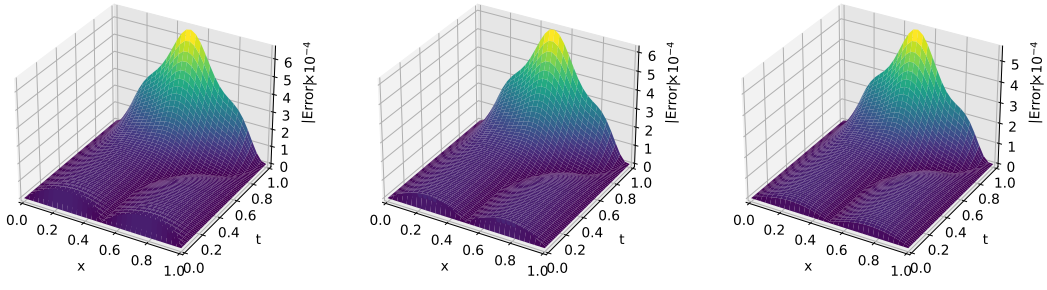
(a) Error for $\alpha = 0.5$ (b) Error for $\alpha = 0.7$ (c) Error for $\alpha = 0.9$

Figure 10: 3D representations of the absolute error behaviors for Example 2 at $v = d = 1$.



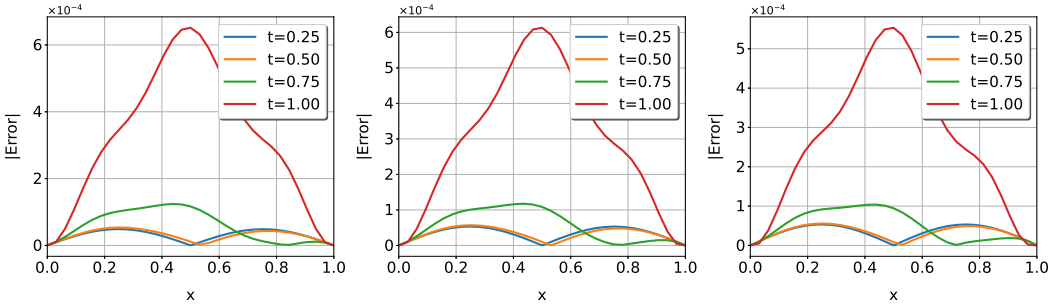
(a) Error for $\alpha = 0.5$ (b) Error for $\alpha = 0.7$ (c) Error for $\alpha = 0.9$

Figure 11: 2D representations of the absolute error behaviors for Example 2 at $v = d = 1$.



(a) Error for $\alpha = 0.5$ (b) Error for $\alpha = 0.7$ (c) Error for $\alpha = 0.9$

Figure 12: 3D representations of the absolute error behaviors for Example 2 at $v = 2, d = 1$.



(a) Error for $\alpha = 0.5$ (b) Error for $\alpha = 0.7$ (c) Error for $\alpha = 0.9$

Figure 13: 2D representations of the absolute error behaviors for Example 2 at $v = 2, d = 1$.

6 Conclusion

This paper develops a numerical method for the generalized nonlinear time-fractional convection-diffusion equation. First, the quasi-linearization process is used to obtain a linear-time fractional convection-diffusion sequence. Next, we apply the scheme on uniform partitions, which consists of the nonpolynomial tension spline scheme in the spatial direction and the Crank–Nicolson discretization in the temporal direction. Analytical and some numerical results from the literature are compared with the results. This comparison demonstrates the effectiveness and efficiency of the proposed approach, which can also be applied to a variety of scientific and physical fields. Additionally, the error norms L_2 and L_∞ are calculated and provided in tables to demonstrate the accuracy of the present approach. The accuracy of the proposed scheme is demonstrated using two test problems for a range of parameter values. The results of these different tests, along with comparisons with other results and the error norms L_2 and L_∞, are displayed in tables and figures. Finally, it has been demonstrated that the present approach is suitable and widely applicable to FDEs. In the future, we will extend this method to Caputo sense time-FDEs near $t = 0$, which exhibit weakly singular behavior.

Acknowledgements

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

Conflict of interest

The author declares that he has no competing interests.

References

[1] Abolhasani, M., Abbasbandy, S. and Allahviranloo, T. *A new variational iteration method for a class of fractional convection-diffusion equations in large domains*, Math. 5(26), (2017), 1–15.

[2] Ahmad, H., Farooq, M., Khan, I., Nawaz, R., Fewster-Young, N. and Askar, S. *Analysis of nonlinear fractional-order Fisher equation using two reliable techniques*, Open Phys. 22(1), (2024), 1–8.

[3] Akram, T., Abbas, M., Riaz, M.B., Ismail, A.I. and Ali, N.M. *An efficient numerical technique for solving time fractional Burgers equation*, Alex. Eng. J. 59(4), (2020), 2201–2220.

[4] Alam, K.N., Ara, A. and Mahmood, A. *Numerical solutions of time-fractional Burgers equations: A comparison between generalized differential transformation technique and homotopy perturbation method*, Int. J. Numer. Methods Heat Fluid Flow, 22(2), (2012), 175–193.

[5] Atangana, A. *Fractal-fractional differentiation and integration: connecting fractal calculus and fractional calculus to predict complex system*, Chaos Solitons Fractals. 102 (2017), 396–406.

[6] Bekela, A.S., Belachew, M.T. and Wole, G.A. *A numerical method using Laplace-like transform and variational theory for solving time-fractional nonlinear partial differential equations with proportional delay*, Adv. differ. equ. 2020(1), (2020), 1–19.

[7] Bekela, A.S. and Deresse, A.T. *A hybrid yang transform Adomian decomposition method for solving time-fractional nonlinear partial differential equation*, BMC Res. Notes, 17(226), (2024), 1–20.

[8] Bekela, A.S. and Deresse, A.T. *An efficient numerical method for nonlinear time-fractional hyperbolic partial differential equations based on fractional Shehu transform iterative method*, J. Appl. Math. 2025 (2025), 1–22.

[9] Bekela, A.S., Enyadene, L.G. and Woldaregay, M. M. *Analysis and application of a hybrid tension-spline method for nonlinear time-fractional reaction–diffusion equations*, Discov. Appl. Sci. 8(457), (2026), 1–32.

[10] Bekela, A.S. and Woldaregay, M.M. *Formable transform Adomian decomposition method for solving nonlinear time-fractional diffusion equation*, Partial Differ. Equ. Appl. Math. 15 (2025), 1–10.

[11] Bellman, R.E. and Kalaba, R.E. *Quasilinearization and nonlinear boundary-value problems*, Amer. Elsevier Publ. Co., New York, 3, 1965.

[12] Chawla, R., Deswal, K., Kumar, D. and Baleanu, D. *Numerical simulation for generalized time-fractional Burgers' equation with three distinct linearization schemes*, J. Comput. Nonlinear Dyn. 18(4), (2023), 1–16.

[13] Chawla, R., Kumar, D. and Singh, S. *A second-order scheme for the generalized time-fractional Burgers' equation*, J. Comput. Nonlinear Dyn. 19(1), (2024), 1–13.

[14] 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.

[15] Delkhosh, M. *Introduction of derivatives and integrals of fractional order and its applications*, Appl. Math. Phys. 1(4), (2013), 103–119.

[16] Deresse, A.T. and Bekela, A.S. *A deep learning approach: physics-informed neural networks for solving a nonlinear telegraph equation with different boundary conditions*, BMC Res. Notes, 18(77), (2025), 1–21.

[17] Deresse, A.T., Bekela, A.S. and Dufera, T.T. *MT-PINNs: multi-term physics-informed neural networks for solving initial boundary value problems of 2D and 3D nonlinear telegraph equations*, Bound. Value Probl. 2025(146), (2025), 1–29.

[18] Deresse, A.T. and Bekela, A.S. and Dufera, T.T. *Leveraging advanced deep neural networks algorithm for solving multi-dimensional forward and inverse problems of Klein-Gordon equation with quadratic, cubic, and fifth-degree polynomials nonlinearity*, Comput. Math. Model. 36(3), (2025), 489–516.

[19] Du, M., Wang, Z. and Hu, H. *Measuring memory with the order of fractional derivative*, Sci. Rep. 3(1), (2013), 1–3.

[20] Esen, A., Bulut, F. and Oruç, O. *A unified approach for the numerical solution of time-fractional Burgers' type equations*, Eur. Phys. J. Plus, 131(4), (2016), 1–13.

[21] Esen, A. and Tasbozan, O. *Numerical solution of time-fractional Burgers equation*, Acta Univ. Sapientiae Math. 7(2), (2015), 167–185.

[22] Esen, A. and Tasbozan, O. *Numerical solution of time-fractional Burgers equation by cubic B-spline finite elements*, Mediterr. J. Math. 13(3), (2016), 1325–1337.

[23] Fang, J., Nadeem, M., Habib, M., Karim, S. and Wahash, H.A. *A new iterative method for the approximate solution of Klein-Gordon and Sine-Gordon equations*, J. Funct. Spaces, 2022(2), (2022), 1–9.

[24] Gowrisankar, S. and Natesan, S. *An efficient robust numerical method for singularly perturbed Burgers' equation*, Appl. Math. Comput. 346(1), (2019), 385–394.

[25] Guan, J., Rahman, K. and Guo, Z. *Radial basis function finite difference method for solving the generalized time-fractional Burgers equation with three types of boundary conditions*, Chaos Solitons Fractals. 199(3), (2025), 116754.

[26] Hashmi, M.S., Wajiha, M., Yao, S., Ghaffar, A. and Inc, M. *Cubic spline based differential quadrature method: A numerical approach for fractional Burger equation*, Results Phys. 26 (2021), 1–11.

[27] Inc, M. *The approximate and exact solutions of the space-and time-fractional Burgers equations with initial conditions by variational iteration method*, J. Math. Anal. Appl. 345(1), (2008), 476–484.

[28] Jacobs, B.A. and Harley, C. *Application of nonlinear time-fractional partial differential equations to image processing via hybrid Laplace transform method*, J. Math. 2018 (2018), 1–9.

[29] Kanth, A.S.V.R., Aruna, K., Raghavendar, K., Rezazadeh, H. and Inc, M. *Numerical solutions of nonlinear time fractional Klein-Gordon equation via natural transform decomposition method and iterative Shehu transform method*, J. Ocean Eng. Sci. (2021).

[30] Kanth, A.S.V.R. and Sirswal, D. *Analysis and numerical simulation for a class of time-fractional diffusion equation via tension spline*, Numer. Algor. 79(2), (2018), 479–497.

[31] Kumar, R. and Jain, S. *Time fractional generalized Korteweg-de Vries equation: Explicit series solutions and exact solutions*, J. Frac. Calc. Nonlinear Sys. 1(1), (2021), 62–77.

[32] Li, D., Zhang, C. and Ran, M. *A linear finite difference scheme for generalized time fractional Burgers equation*, Appl. Math. Model. 40 (11-12), (2016), 6069–6081.

[33] Majeed, A., Kamran, M. and Rafique, M. *An approximation to the solution of time-fractional modified Burgers' equation using extended cubic B-spline method*, Comput. Appl. Math. 39(4), (2020), 1–21.

[34] Naeem, M., Aljahdaly, N.H., Shah, R. and Weera, W. *The study of fractional-order convection–reaction-diffusion equation via an Elzake Atangana–Baleanu operator*, AIMS Math. 7(10), (2022), 18080–18098.

[35] Onal, M. and Esen, A. *A Crank–Nicolson approximation for the time-fractional Burgers equation*, Appl. math. nonlinear sci. 5(2), (2020), 177–184.

[36] Oru, O., Esen, A. and Bulut, F. *A unified finite difference Chebyshev wavelet method for numerically solving time-fractional Burgers' equation*, Discrete Contin. Dyn. Syst., Ser. S. 12(3), (2019), 533–542.

[37] Podlubny, I. *Fractional differential equations: an introduction to fractional derivatives, fractional differential equations, to methods of their solution and some of their applications*, Elsevier, 1998.

[38] Poojitha, S. and Awasthi, A. *Eloquent numerical approach for solving generalized time-fractional convection-diffusion-reaction problems*, Phys. Scr. 99(12), (2024), 1–19.

[39] Priyendhu, K.S., Prakash, P. and Lakshmanan, M. *Invariant subspace method to the initial and boundary value problem of the higher dimensional nonlinear time-fractional PDEs*, Commun. Nonlinear Sci. Numer. Simul. 122 (2023), 107245.

[40] Rubin, S.G. and Graves, J.R.A. *Viscous flow solutions with a cubic spline approximation*, Comput. Fluids, 3(1), (1975), 1–36.

[41] Salama, F.M. *An efficient explicit group method for time-fractional Burgers equation*, Front. Phys. 13, (2025), 1–18.

[42] Sugimoto, N. *Burgers equation with a fractional derivative; hereditary effects on nonlinear acoustic waves*, J. Fluid Mech. 225 (1991), 631–653.

[43] Vijayaram, S. and Balasubramaniam, P. *Optimal control for nonlinear time-fractional Schrudinger equation: an application to quantum optics*, Phys. Scr. 99(9), (2024), 095115.

[44] Yousif, M.A. and Hamasalh, F.K. *Novel simulation of the time-fractional Burgers-Fisher equations using a nonpolynomial spline fractional continuity method*, AIP Adv. 12(11), (2022), 1–13.

Iranian Journal of Numerical Analysis and Optimization
Vol. 16, No. 3, 2026, pp 922–951
https://doi.org/10.22067/ijnao.2026.95961.1753
https://ijnao.um.ac.ir/

CC BY NC

Research Article

Robust stochastic gradient descent for linearly constrained problems via adaptive barrier amplification

A. Fillali*, S. Addoune and R. Zeghdane

Abstract

Methods based on stochastic gradient descent using relaxed barrier functions provide a powerful framework for linearly constrained optimization but often suffer from sensitivity to hyperparameter settings and slow recovery from infeasible regions. Our investigation reveals that the optimal performance of standard relaxed barrier methods is confined to a narrow and unstable, "special-case" optimization path, limiting their practical robustness. This paper introduces a novel algorithm designed to overcome these limitations through a dynamic barrier amplification mechanism. This approach adaptively intensifies the corrective force of the barrier gradient

*Corresponding author

Received 14 October 2025; revised 23 February 2026; accepted 22 May 2026

Abdelouakil Fillali
Laboratoire d'Analyse Mathématiques et ses Applications (LAMA), Department of Mathematics, University Mohamed El Bachir El Ibrahimi, Bordj Bou Arreridj, 34030, Algeria. e-mail: abdelouakil.fillali@univ-bba.dz

Smail Addoune
Laboratoire d'Analyse Mathématiques et ses Applications (LAMA), Department of Mathematics, University Mohamed El Bachir El Ibrahimi, Bordj Bou Arreridj, 34030, Algeria. e-mail: smail.addoune@univ-bba.dz

Rebiha Zeghdane
e-mail: rebiha.zeghdane@univ-bba.dz
Laboratoire d'Analyse Mathématiques et ses Applications (LAMA), Department of Mathematics, University Mohamed El Bachir El Ibrahimi, Bordj Bou Arreridj, 34030, Algeria.

How to cite this article
Fillali, A., Addoune, S. and Zeghdane, R., Robust stochastic gradient descent for linearly constrained problems via adaptive barrier amplification. *Iran. J. Numer. Anal. Optim.*, 2026; 16(3): 922-951. https://doi.org/10.22067/ijnao.2026.95961.1753

in proportion to the magnitude of constraint violations, ensuring a strong push towards the feasible set specifically when iterates become highly infeasible. We provide a rigorous theoretical analysis, preserving the almost sure convergence guarantees of the baseline algorithm and deriving an explicit linear convergence rate to a neighborhood of the solution. Numerical results on challenging linearly constrained quadratic programming problems demonstrate that our algorithm exhibits superior robustness: It significantly reduces constraint violations in ill-conditioned scenarios compared to the standard method, acting as a reliability safety net, while maintaining competitive efficiency in nominal conditions.

AMS subject classifications (2020): Primary 90C15, 90C25; Secondary 49M30, 65K05, 68T05.

Keywords: Linearly constrained optimization, Stochastic gradient descent, Relaxed barrier functions, Dynamic barrier amplification, Robust optimization, Convergence analysis

1 Introduction

Stochastic gradient descent (SGD) and its variants have established themselves as the cornerstone of modern large-scale optimization, driving advancements in fields ranging from machine learning to complex networked systems [2, 3, 15, 24]. While SGD exhibits exceptional efficiency in unconstrained settings, its extension to constrained optimization remains a challenging and active area of research [19].

Classical approaches often demand full knowledge of the constraint set, necessitating computationally expensive projection operations that become prohibitive in high-dimensional spaces [20]. To mitigate this, recent methodologies have emerged that rely on sampling subsets of constraints, utilizing techniques, such as penalty functions [12, 21] or primal-dual frameworks [22, 23]. Among these, the *adaptive relaxed barrier SGD* proposed by Dimitrieski, Cao, and Ebenbauer [6] stands out as a promising direction, leveraging a relaxed logarithmic barrier to avoid projections entirely [7, 8].

However, despite its theoretical guarantees, our investigation uncovers a critical limitation in this baseline approach: Its performance is heavily dependent on a "highly sensitive" optimization path. Specifically, when initialized far from the feasible region or when faced with ill-conditioned linearly constrained problems, the uniform learning rate fails to generate a sufficient corrective signal. This architectural vulnerability leads not merely to slow convergence, but to significant constraint violations, rendering the method unreliable for applications where feasibility is paramount.

In this paper, we address this reliability gap by introducing a novel algorithm: Robust SGD with dynamic barrier amplification. Unlike standard methods that treat all gradients equally, our approach introduces a mechanism that dynamically intensifies the corrective barrier force in direct proportion to constraint violations. This effectively creates an adaptive "safety net," ensuring robust recovery from infeasible regions without compromising efficiency in standard conditions.

Our contributions are summarized as follows:

- **Algorithmic Innovation:** We introduce the *dynamic barrier amplification* mechanism, which adaptively modulates the barrier gradient to prioritize feasibility precisely when needed.

- **Theoretical Rigor:** We provide a comprehensive convergence analysis that preserves the almost sure convergence guarantees of the baseline algorithm. Furthermore, addressing a gap in prior work [6], we derive an explicit linear convergence rate to a neighborhood of the solution using the expected smoothness framework.

- **Empirical Robustness:** Through rigorous "stress testing" on linearly constrained quadratic problems, we demonstrate that our algorithm acts as a robust safeguard. While performing identically to the baseline in nominal settings, it significantly suppresses constraint violations in ill-conditioned scenarios where the standard method fails.

The remainder of this paper is organized as follows: Section 2 formalizes the problem statement and details the proposed dynamic barrier amplification methodology. Section 3 provides the theoretical framework, establishing almost sure convergence. Section 4 presents the derivation of the linear convergence rate. Section 5 reports the numerical experiments and robustness analysis. Finally, Section 6 concludes the paper and outlines directions for future research.

2 The proposed methodology

In this section, we formally introduce the proposed algorithm. We begin by defining the linearly constrained problem formulation and the necessary assumptions, incorporating specific relaxations suggested by the analysis of stochastic approximation. We then detail our novel modification, the dynamic barrier amplification strategy.

2.1 Problem formulation and assumptions

We address the following linearly constrained optimization problem:

$$\min_{x\in\mathbb{R}^d} f(x) := \frac{1}{n}\sum_{i=1}^{n} f_i(x) \quad \text{s.t.} \quad g_j(x) \le 0, \qquad \text{for all } j \in \{1,\dots,m\}, \tag{1}$$

where the objective components $f_i : \mathbb{R}^d \to \mathbb{R}$ are continuously differentiable, and the constraint functions are affine, defined as $g_j(x) := a_j^T x + b_j$ with $a_j \in \mathbb{R}^d$ and $b_j \in \mathbb{R}$.

We adopt the following standard assumptions, which are common in the analysis of stochastic gradient methods [4, 5, 18].

Assumption 1 (Objective function properties)**.** The objective function components f_i are L-smooth and convex. Furthermore, the aggregate function $f(x)$ is μ-strongly convex for some $\mu > 0$. These assumptions are standard in convergence proofs for gradient-based optimization [6, 15].

Remark 1. This condition is satisfied if at least one component function f_i is strongly convex, even if the others are merely convex.

Assumption 2 (Feasibility)**.** The feasible set $\mathbb{X} := \{x \in \mathbb{R}^d : g_j(x) \le 0, \text{ for all } j\}$ satisfies Slater's condition, meaning it has a nonempty interior. This is a fundamental requirement for the validity of barrier-based methods [4, 8]

Assumption 3 (Stochastic sampling)**.** At each iteration k, an objective index $i_k \sim \mathcal{U}\{1,\dots,n\}$ and a constraint index $j_k \sim \mathcal{U}\{1,\dots,m\}$ are sampled independently and uniformly at random. This ensures that the stochastic gradients are unbiased estimators satisfying standard SGD assumptions [6]:

$$\mathbb{E}[\nabla f_{i_k}(x)] = \nabla f(x),$$
$$\mathbb{E}[\nabla B(g_{j_k}(x), \delta_k)] = \frac{1}{m}\sum_{j=1}^{m} \nabla B(g_j(x), \delta_k).$$

2.2 Baseline algorithm: Adaptive relaxed barrier

The foundation of our work is the adaptive relaxed barrier SGD algorithm proposed by Dimitrieski, Cao, and Ebenbauer [6]. This method addresses the constrained problem (1) using

a stochastic update rule that combines the objective gradient with the gradient of a relaxed logarithmic barrier function, $B(g, \delta)$. The standard update rule is given by

$$x_{k+1} = x_k - \gamma_k \left(\nabla f_{i_k}(x_k) + \nabla B(g_{j_k}(x_k), \delta_k) \right). \tag{2}$$

While effective, this approach applies a uniform learning rate γ_k to both the objective and barrier gradients. This can lead to slow progress, particularly when an iterate is far outside the feasible region and requires a stronger corrective signal.

In this context, the specific relaxed logarithmic barrier function $B(z, \delta)$ employed in (2) is defined as follows:

$$B(z, \delta) := \begin{cases} -\delta \log(-z), & z < -\delta, \\ \frac{1}{2}\left(\frac{(z+2\delta)^2}{\delta} - \delta \right) - \delta \log(\delta), & z \geq -\delta. \end{cases}$$

2.3 Dynamic barrier amplification strategy

To address the limitation of the baseline method, we introduce a novel mechanism called dynamic barrier amplification. The core idea is to adaptively increase the influence of the barrier gradient precisely when a constraint is violated. This is achieved by introducing a dynamic amplification factor, denoted by A_k, which modulates the barrier term. The amplification factor is a function of the constraint violation $g_{j_k}(x_k)$ and is defined as

$$A_k = 1 + \lambda \cdot \tanh\left(\max(0, g_{j_k}(x_k)) \right),$$

where $\lambda > 0$ is a hyperparameter that controls the intensity of the amplification.

The hyperbolic tangent function, `tanh`, is specifically chosen for its intrinsic saturation properties. As the constraint violation increases, `tanh` provides a smooth yet significant corrective boost that naturally saturates, preventing overly aggressive updates that could lead to numerical instability.

The intuition behind the design of A_k is effective. When a constraint is satisfied ($g_{j_k}(x_k) \leq 0$), the `max` term becomes zero, and thus $A_k = 1$. In this case, our algorithm's update step defaults to the original relaxed barrier method (see (2)). However, when a constraint is violated ($g_{j_k}(x_k) > 0$), A_k becomes greater than 1, applying a stronger corrective force to push the iterate back towards the feasible region more aggressively. The behavior of this amplification factor is illustrated in Figure 1.

This leads to our proposed amplified rate SGD update rule:

$$x_{k+1} = x_k - \gamma_k \left(\nabla f_{i_k}(x_k) + A_k \cdot \nabla B(g_{j_k}(x_k), \delta_k) \right). \tag{3}$$

By dynamically amplifying the barrier gradient, this update rule prioritizes constraint satisfaction when needed, thereby aiming for faster and more robust overall convergence.

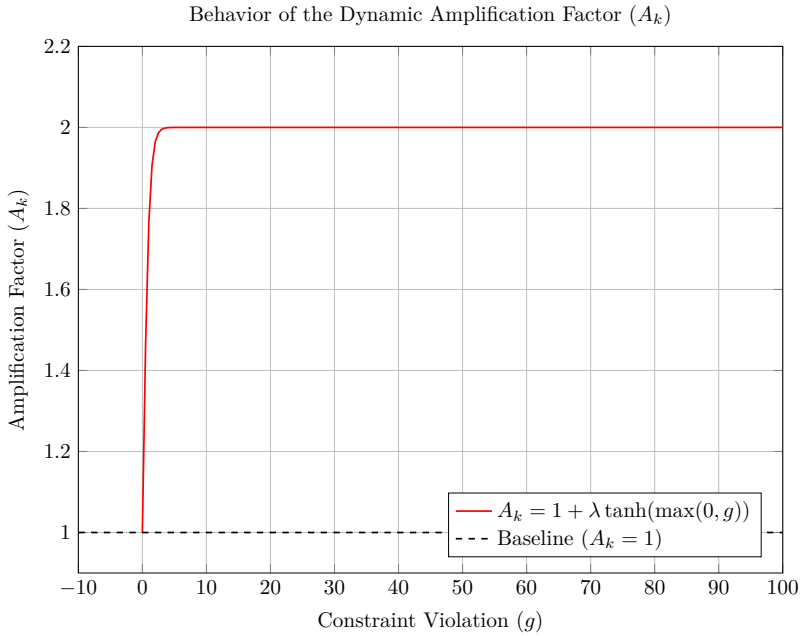


Figure 1: Behavior of the Dynamic Amplification Factor (A_k) as a function of the constraint violation (g), using the hyperbolic tangent (tanh) function with $\lambda = 1.0$. The factor remains at its baseline value of 1 in the feasible region ($g \leq 0$) and increases smoothly upon violation ($g > 0$), naturally saturating at an upper bound of $1 + \lambda$. This inherent stability removes the need for an explicit upper cap.

3 Theoretical framework and convergence analysis

Our convergence analysis relies on the well-established theory of stochastic gradient methods, rooted in the pioneering work on stochastic approximation [16, 11, 9]. Recent advancements have focused on providing a deeper, quantitative understanding of this convergence by establishing explicit *almost sure convergence rates* for various stochastic algorithms [13]. The analysis proves that the introduction of the stabilized dynamic barrier amplification factor, A_k, preserves the almost sure convergence guarantees of the original algorithm [6]. The baseline algorithm is proven to converge with probability one for strongly convex objectives and affine inequality constraints,

under specific conditions on the step-size (γ_k) and barrier adaptation (δ_k) rules [6]. Our proof builds upon the detailed derivations for the stochastic gradient components presented in the appendices of Dimitrieski, Cao, and Ebenbauer [6].

Our analysis extends this framework to formally prove that the introduction of the dynamic barrier amplification factor, A_k, preserves this convergence guarantee. The proof leverages the fact that the modifications introduced by A_k are inherently bounded due to the natural saturation properties of the hyperbolic tangent function, allowing for the application of the same convergence theorems.

3.1 Preliminaries and Lyapunov function

We begin by restating our proposed update rule from (3):

$$x_{k+1} = x_k - \gamma_k(\nabla f_{i_k}(x_k) + A_k \nabla B(g_{j_k}(x_k), \delta_k)).$$

Our goal is to prove that the sequence of iterates $\{X_k\}$ generated by this rule converges almost surely to $x^*(\delta_\infty)$[⋆].

Let $\tilde{X}_k = X_k - x^*(\delta_\infty)$ be the error at iteration k. We define the Lyapunov function $V_k = ||\tilde{X}_k||^2$. The core of our proof is to analyze the conditional expectation of V_{k+1} given the filtration $\mathcal{F}_k$[⋆⋆], denoted by $\mathbb{E}_k[\cdot] = \mathbb{E}[\cdot|\mathcal{F}_k]$.

Expanding the Lyapunov function, we have

$$\mathbb{E}_k[||\tilde{X}_{k+1}||^2] = ||\tilde{X}_k||^2 - 2\gamma_k \mathbb{E}_k[G_k^T \tilde{X}_k] + \gamma_k^2 \mathbb{E}_k[||G_k||^2], \tag{4}$$

where $G_k = \nabla f_{i_k}(x_k) + A_k \nabla B(g_{j_k}(x_k), \delta_k)$ is the amplified stochastic gradient.

⋆ Here, $x^*(\delta_\infty)$ denotes the unique global minimizer of the regularized objective function $\Phi(x)$ (as defined in (5)), with the limiting barrier parameter $\delta_\infty > 0$.

⋆⋆ $\mathcal{F}_k$ represents the σ-algebra generated by the history of the algorithm up to iteration k, i.e., $\{x_0, i_0, j_0, \ldots, i_{k-1}, j_{k-1}\}$.

3.2 Decomposition of the amplified stochastic gradient

To analyze (4), we first decompose the gradient term G_k. Following Dimitrieski, Cao, and Ebenbauer [6], the relaxed barrier gradient can be split as $\nabla B(g_{j_k}, \delta_k) = \nabla B(g_{j_k}, \delta_\infty) - \nabla C(g_{j_k}, \delta_k)$, where ∇C captures the difference due to the decaying barrier parameter $\epsilon_k = \delta_k - \delta_\infty$.

To formalize the analysis, we first define the regularized objective function, $\Phi(X_k)$, as follows:

$$\Phi(X_k) := f(X_k) + \frac{1}{m}\sum_{j=1}^{m} B(g_j(X_k), \delta_\infty). \tag{5}$$

Using this definition, the amplified stochastic gradient G_k can be decomposed. This decomposition is key, as it isolates the original stochastic gradient term, $\nabla\Phi_{i_k,j_k}(X_k)$[***], from the new terms introduced by our amplification mechanism:

$$G_k = \underbrace{\nabla f_{i_k}(X_k) + \nabla B(g_{j_k}(X_k), \delta_\infty)}_{\nabla\Phi_{i_k,j_k}(X_k)} + (A_k - 1)\nabla B(g_{j_k}(X_k), \delta_\infty) - A_k \nabla C(g_{j_k}(X_k), \delta_k).$$

3.3 Bounding the Lyapunov recurrence

We now bound the terms in (4).

3.3.1 Bounding the cross-term

The expectation of the cross-term is

$$\mathbb{E}_k[G_k^T \tilde{X}_k] = \mathbb{E}_k[\nabla\Phi_{i_k,j_k}^T \tilde{X}_k] + \mathbb{E}_k[((A_k - 1)\nabla B_{j_k,\infty})^T \tilde{X}_k] - \mathbb{E}_k[(A_k \nabla C_{j_k})^T \tilde{X}_k]. \tag{6}$$

1. The first term provides the driver for convergence. Due to the μ-strong convexity of $\Phi(\cdot)$ (from Assumption 1), we have $\mathbb{E}_k[\nabla\Phi_{i_k,j_k}^T \tilde{X}_k] = \nabla\Phi(X_k)^T \tilde{X}_k \geq \Phi_0(X_k) + \frac{\mu}{2}||\tilde{X}_k||^2$, where $\Phi_0(X_k) := \Phi(X_k) - \Phi(x^*(\delta_\infty))$.

2. The second term is nonnegative. When a constraint is satisfied ($g_{j_k}(x_k) \leq 0$), the amplification factor $A_k = 1$, making $(A_k - 1) = 0$ and the term zero. When a constraint is violated ($g_{j_k}(x_k) > 0$), the factor $(A_k - 1)$ is positive. Furthermore, the inner product is non-

[***] The stochastic gradient of the regularized objective function, $\nabla\Phi_{i_k,j_k}$, is defined based on (5). For a single random sample (i_k, j_k), it is the sum of the objective gradient and the final barrier gradient: $\nabla\Phi_{i_k,j_k}(X_k) := \nabla f_{i_k}(X_k) + \nabla B(g_{j_k}(X_k), \delta_\infty)$.

negative. This follows from the first-order condition for the convexity of the barrier function $B(g_{j_k}(\cdot), \delta_\infty)$, which implies that

$$(\nabla B_{j_k,\infty})^T \tilde{X}_k \geq B(g_{j_k}(x_k), \delta_\infty) - B(g_{j_k}(x^*(\delta_\infty)), \delta_\infty) > 0.$$

Therefore, the conditional expectation $\mathbb{E}_k[((A_k - 1)\nabla B_{j_k,\infty})^T \tilde{X}_k]$ is nonnegative.

3. The third term, $-\mathbb{E}_k[(A_k \nabla C_{j_k})^T \tilde{X}_k]$.

First, by applying the Cauchy–Schwarz inequality to the inner product, we have

$$(A_k \nabla C_{j_k})^T \tilde{X}_k \leq ||A_k \nabla C_{j_k}|| \cdot ||\tilde{X}_k||.$$

The amplification factor is inherently bounded due to the `tanh` function, such that $A_k = 1 + \lambda \tanh(\max(0, g_{j_k}(x_k))) \leq 1 + \lambda$. Since A_k is a positive scalar, we can write

$$||A_k \nabla C_{j_k}|| = A_k ||\nabla C_{j_k}|| \leq (1 + \lambda)||\nabla C_{j_k}||.$$

Taking the conditional expectation $\mathbb{E}_k[\cdot]$ yields

$$\mathbb{E}_k[(A_k \nabla C_{j_k})^T \tilde{X}_k] \leq (1 + \lambda)\mathbb{E}_k[||\nabla C_{j_k}|| \cdot ||\tilde{X}_k||].$$

The remaining term $\mathbb{E}_k[||\nabla C_{j_k}||||\tilde{X}_k||]$ is precisely the baseline (nonamplified) error term. Following the citation style of the main paper, the bound for this term is explicitly derived in the baseline algorithm's analysis [6, Appendix V-B], resulting in the inequality stated in [6, p. 4, Eq. (8)]. By substituting this known upper bound, we get the final result:

$$\mathbb{E}_k[(A_k \nabla C_{j_k})^T \tilde{X}_k] \leq (1 + \lambda) \left[\|\tilde{X}_k\|^2 \left(\overline{a} \frac{\epsilon_k}{\delta_k \delta_\infty} + \hat{c}\epsilon_k \overline{a} + \overline{b} \frac{\epsilon_k}{\delta_k \delta_\infty} \right) + \frac{\hat{c}\epsilon_k}{4} \overline{a} + \frac{1}{4} \overline{b} \frac{\epsilon_k}{\delta_k \delta_\infty} \right].$$

To handle this amplified error and proceed with the proof, we drop the nonnegative second term from (6) to find an upper bound for the full cross-term. By substituting the bounds for the first and third terms, we get

$$\begin{aligned} -2\gamma_k \mathbb{E}_k[G_k^T \tilde{X}_k] \leq & -2\gamma_k \left(\Phi_0(X_k) + \frac{\mu}{2} ||\tilde{X}_k||^2 \right) \\ & + 2\gamma_k (1 + \lambda) \left[||\tilde{X}_k||^2 \left(\overline{a} \frac{\epsilon_k}{\delta_k \delta_\infty} + \hat{c}\epsilon_k \overline{a} + \overline{b} \frac{\epsilon_k}{\delta_k \delta_\infty} \right) \right. \\ & \left. + \frac{\hat{c}\epsilon_k}{4} \overline{a} + \frac{1}{4} \overline{b} \frac{\epsilon_k}{\delta_k \delta_\infty} \right]. \end{aligned}$$

3.3.2 Bounding the squared norm

To analyze the Lyapunov recurrence, we must find an upper bound for the term $\mathbb{E}_k[\|G_k\|^2]$. Using the inequality $\|u+v+w\|^2 \leq 3\|u\|^2 + 3\|v\|^2 + 3\|w\|^2$, which is an application of the Cauchy–Schwarz inequality, we can bound the squared norm of the amplified stochastic gradient G_k. The decomposition from section 3.2 leads to

$$\mathbb{E}_k[\|G_k\|^2] \leq 3\mathbb{E}_k[\|\nabla\Phi_{i_k,j_k}\|^2] + 3\mathbb{E}_k[\|(A_k-1)\nabla B_{j_k,\infty}\|^2] + 3\mathbb{E}_k[\|A_k\nabla C_{j_k}\|^2]. \tag{7}$$

We bound each of the three terms on the right-hand side as follows:

- Term 1: Bounding $\mathbb{E}_k^{\mathcal{F}}[||\nabla\Phi_{i_k,j_k}(X_k)||^2]$. This term represents the variance of the stochastic gradient of the regularized objective function Φ. As proven in detail in the [6, Appendix V-D] of the baseline algorithm, this term is bounded by the following inequality:

  $$\mathbb{E}_k[||\nabla\Phi_{i_k,j_k}(X_k)||^2] \leq 4\hat{L}\Phi_0(X_k) + 2\sigma_\Phi,$$

  where $\hat{L}$ is a positive constant related to the smoothness of the function, and σ_Φ is a bounded constant representing the variance at the optimal solution.

- Term 2: Bounding $\mathbb{E}_k[||(A_k-1)\nabla B_{j_k,\infty}||^2]$p. The second term is a novel artifact of the amplification mechanism, To establish a controllable upper bound, we leverage the fact that $A_k - 1 = \lambda\tanh(\max(0, g_{j_k}))$. Since $|\tanh(\cdot)| \leq 1$, we have $|A_k - 1| \leq \lambda$. This leads to the bound:

  $$\mathbb{E}_k\left[||(A_k-1)\nabla B_{j_k,\infty}||^2\right] \leq \lambda^2\mathbb{E}_k\left[||\nabla B_{j_k,\infty}||^2\right].$$

  The problem is now reduced to bounding the remaining component, $\mathbb{E}_k[||\nabla B_{j_k,\infty}||^2]$, which is part of the original stochastic variance analyzed in the baseline algorithm by Dimitrieski, Cao, and Ebenbauer [6].

  The boundedness of this term is formally justified by the proof that the variance at the optimum, σ_Φ, is a finite value ($\sigma_\Phi < \infty$), as established in the baseline algorithm's analysis [6].[†] This analysis is presented in detail in [6, Appendix V-D].

  This analysis ensures that the term is bounded by a finite constant, which leads to the following overall bound:

  $$\mathbb{E}_k[||(A_k-1)\nabla B_{j_k,\infty}||^2] \leq C_B,$$

[†] In the source paper [6], σ_Φ is defined in Eq. (20) using the term $||\nabla\Phi_{i,j}||^2$. The stochastic gradient $\nabla\Phi_{i,j}$ is itself defined on page 3 as the sum of the objective and final barrier gradients: $\nabla f_i + \nabla B(g_j, \delta_\infty)$. Therefore, the finiteness of σ_Φ necessarily implies the boundedness of the final barrier gradient term.

where $C_B > 0$ is a constant that depends on $1 + \lambda$. and the final barrier variance.

- Term 3: Bounding $\mathbb{E}_k[||A_k \nabla C_{j_k}||^2]$. This term arises from the amplification of the transient error caused by the decaying barrier parameter. Using the fact that the amplification factor is inherently bounded ($A_k \leq 1 + \lambda$), we can establish an upper bound:

$$\mathbb{E}_k \left[||A_k \nabla C_{j_k}||^2\right] \leq (1 + \lambda)^2 \cdot \mathbb{E}_k \left[||\nabla C_{j_k}||^2\right].$$

The remaining component, $\mathbb{E}_k[||\nabla C_{j_k}||^2]$, is precisely the expected squared error from the baseline algorithm, which is analyzed by Dimitrieski, Cao, and Ebenbauer [6]. The bound for this term is explicitly derived in [6, p. 4, Eq. (9)], with the full justification provided in [6, Appendix V-C, p. 8]. By substituting this result, we obtain the complete upper bound for our amplified error term:

$$\mathbb{E}_k[||A_k \nabla C_{j_k}||^2] \leq (1 + \lambda)^2 \cdot \left(3\hat{a} \frac{\epsilon_k^2 ||\tilde{X}_k||^2}{\delta_k^2 \delta_\infty^2} + 3\hat{c}^2 \epsilon_k^2 \overline{a} + \frac{3\overline{b}\epsilon_k^2}{\delta_k^2 \delta_\infty^2} \right).$$

To formalize this step of the proof, we rely on the following key lemma, which is a direct application of the Robbins–Siegmund Theorem [17].

Lemma 1 (Application of the Robbins–Siegmund Theorem [17])**.** Let $\{V_k\}$, $\{a_k\}$, and $\{b_k\}$ be sequences of nonnegative random variables, adapted to a filtration $\{\mathcal{F}_k\}$, such that for all $k \geq 0$,

$$\mathbb{E}[V_{k+1}|\mathcal{F}_k] \leq (1 - a_k)V_k + b_k \quad \text{a.s..}$$

If $\sum_{k=0}^{\infty} a_k = \infty$ a.s. and $\sum_{k=0}^{\infty} b_k < \infty$ a.s., then V_k converges almost surely to a finite random variable, and $\sum_{k=0}^{\infty} a_k V_k < \infty$ almost surely.

3.4 Main convergence result

By substituting the bounds derived in section 3.3.2 back into the Lyapunov recurrence (4) and collecting terms, we arrive at a relation of the form shown in (8). For a complete step-by-step derivation of this result, see Appendix 6.

$$\mathbb{E}_k[\|\tilde{X}_{k+1}\|^2] \leq (1 - \gamma_k(\mu - \xi_k'))\|\tilde{X}_k\|^2 - 2\gamma_k(1 - C_1\gamma_k)\Phi_0(X_k) + Q_k'. \tag{8}$$

To enhance the clarity of the proof, we now provide a more explicit description of the summable error term Q'_k from (8). This term, which is an augmented version of the error sequence from the baseline algorithm [6], collects all residual, bounded error components from the Lyapunov analysis that do not depend on the error norm $||\tilde{X}_k||^2$ or the objective distance $\Phi_0(X_k)$. It incorporates new bounded terms that arise from our dynamic barrier amplification mechanism. The constant C_1 is related to the smoothness parameter $\hat{L}$.

Our modified term, ξ'_k, is an augmented version of the original error sequence ξ_k from [6, p. 4]. It includes additional components related to our amplification mechanism, but importantly, it remains a vanishing sequence (i.e., $\lim_{k\to\infty} \xi'_k = 0$). While the original term collected errors from the decaying barrier, our augmented term ξ'_k now incorporates the inherent bound of the `tanh` function. This ensures that even the amplified error terms are properly accounted for in the convergence proof. The explicit formula for ξ'_k is constructed by scaling the components of the original ξ_k based on how the amplification factor, bounded by $1 + \lambda$, impacts the cross-term and squared-norm bounds in the Lyapunov analysis:

$$\xi'_k := 2(1+\lambda)\left(\frac{\epsilon_k}{\delta_k\delta_\infty}(\bar{a}+\bar{b}) + \hat{c}\epsilon_k\bar{a}\right) + 9(1+\lambda)^2\gamma_k\frac{\epsilon_k^2\hat{a}}{\delta_k^2\delta_\infty^2}. \tag{9}$$

Crucially, the step-size rules for γ_k and ϵ_k, adopted from the baseline algorithm, are sufficient to satisfy the conditions of Lemma 1, thereby enabling the proof of our main convergence result.

Theorem 1 (Almost sure convergence)**.** Consider the amplified rate SGD algorithm defined by the update rule in (3). Let Assumptions 1, 2, and 3 hold. Then the sequence of iterates $\{X_k\}$ converges almost surely to the unique minimizer $x^*(\delta_\infty)$ of the regularized objective function:

$$\lim_{k\to\infty} \|X_k - x^*(\delta_\infty)\|^2 = 0 \quad \text{a.s.}$$

Proof. The proof of almost sure convergence is established by analyzing the behavior of the Lyapunov function $V_k = \|X_k - x^*(\delta_\infty)\|^2$, which represents the squared Euclidean distance to the minimizer. We aim to show that $V_k \to 0$ a.s. by employing the Robbins–Siegmund lemma [17], a fundamental result in martingale convergence theory. The proof proceeds in three steps.

Step 1: The Lyapunov recurrence relation. The starting point is the master recurrence inequality, derived in detail in Appendix 6. This inequality governs the conditional expectation of the Lyapunov function with respect to the filtration $\mathcal{F}_k$ (the history of the process up to iteration k):

$$\mathbb{E}[V_{k+1}|\mathcal{F}_k] \leq (1 - \gamma_k(\mu - \xi'_k))V_k - 2\gamma_k(1 - C_1\gamma_k)\Phi_0(X_k) + Q'_k,$$

where $\mu > 0$ is the strong convexity modulus, $\{\xi'_k\}$ is a vanishing sequence capturing the perturbation from the amplification mechanism, and $\{Q'_k\}$ is a sequence of summable residual errors.

The term $-2\gamma_k(1-C_1\gamma_k)\Phi_0(X_k)$ is nonpositive. This holds because $\gamma_k > 0$ by definition, the step-size is chosen sufficiently small such that $(1 - C_1\gamma_k) > 0$ for large k, and most importantly, $\Phi_0(X_k) = \Phi(X_k) - \Phi(x^*(\delta_\infty)) \geq 0$ by the definition of $x^*(\delta_\infty)$ as the minimizer of $\Phi(x)$.

By dropping this nonpositive term, we obtain a looser but still valid upper bound, which simplifies the application of the subsequent lemma:

$$\mathbb{E}[V_{k+1}|\mathcal{F}_k] \leq (1 - \gamma_k(\mu - \xi'_k))V_k + Q'_k. \tag{10}$$

Step 2: Verification of the conditions of Lemma 1.

To apply Lemma 1 (the Robbins–Siegmund Theorem) [17] to our recurrence relation in (10), we map the terms as follows:

- $a_k := \gamma_k(\mu - \xi'_k)$,
- $b_k := Q'_k$.

We now verify that the conditions of the lemma are met:

1. Condition 1 ($\sum a_k = \infty$): Since $\lim_{k\to\infty} \xi'_k = 0$ and $\mu > 0$, there exists a k_0 such that for all $k \geq k_0$, we have $(\mu - \xi'_k) \geq \mu/2 > 0$. Given that the step-size rule satisfies $\sum \gamma_k = \infty$, it follows that the series $\sum a_k$ must also diverge. Thus, the condition is satisfied.
2. Condition 2 ($\sum b_k < \infty$): The sequence $\{Q'_k\}$ is constructed to contain all summable residual error terms. Therefore, $\sum_{k=0}^{\infty} b_k = \sum_{k=0}^{\infty} Q'_k < \infty$ by definition. Thus, the condition is satisfied.

Step 3: Final conclusion via proof by contradiction.

With both conditions of the Robbins–Siegmund lemma satisfied, we can conclude two key facts: (i) the series $\sum a_k$ diverges to infinity almost surely, and (ii) the series $\sum a_k V_k < \infty$ almost surely.

We now show that these two facts can only hold simultaneously if V_k converges to zero. We proceed by contradiction. Assume that V_k does not converge to zero, but rather to a small positive constant $c > 0$. In this case, for a sufficiently large k, the series $\sum a_k V_k$ would behave like $c\sum a_k$. Since $\sum a_k$ diverges to infinity, it follows that $c\sum a_k$ must also diverge to infinity.

This leads to a direct contradiction with the second fact established by the lemma, which guarantees that $\sum a_k V_k$ is finite. Since our assumption leads to a contradiction, it must be false. Therefore, the only possible conclusion is that V_k must converge to zero.

$$\lim_{k\to\infty} V_k = \lim_{k\to\infty} \|X_k - x^*(\delta_\infty)\|^2 = 0 \quad \text{a.s.}$$

This completes the proof. □

While this result provides a solid theoretical foundation by confirming that our mechanism preserves the almost sure convergence guarantees of the baseline method by Dimitrieski, Cao, and Ebenbauer [6], it does not quantify the algorithm's efficiency. To provide this deeper, quantitative insight—a contribution that extends beyond the original theoretical framework, as Dimitrieski et al. established almost sure convergence but left the derivation of an explicit convergence rate as an avenue for future research—we now derive the algorithm's explicit nonasymptotic convergence rate.

4 Convergence rate analysis: An expected smoothness approach

In this section, we provide a modern and streamlined convergence analysis for the proposed amplified rate SGD algorithm. Instead of relying on restrictive assumptions such as the uniform boundedness of stochastic gradient variance, we adopt the more general framework of *expected smoothness* [10]. This approach not only simplifies the proof but also provides a clearer intuition by directly linking the convergence rate to two fundamental quantities: An effective smoothness constant and the gradient noise at the optimum.

4.1 Preliminaries and key assumptions

Our analysis focuses on minimizing the regularized objective function, previously defined in (5):

$$\Phi(x) := f(x) + \frac{1}{m}\sum_{j=1}^{m} B(g_j(x), \delta_\infty).$$

Let $x^* := x^*(\delta_\infty)$ denote the unique minimizer of $\Phi(x)$. Our proof relies on the following standard assumptions.

Assumption 4 (Effective strong convexity)**.** The regularized objective function $\Phi(x)$ is μ-strongly convex. Consequently, for any $x \in \mathbb{R}^d$, it satisfies

$$\langle \nabla\Phi(x), x - x^* \rangle \geq \Phi(x) - \Phi(x^*) + \frac{\mu}{2}\|x - x^*\|^2.$$

Assumption 5 (Expected smoothness and finite noise)**.** Let $G_k(x) = \nabla f_{i_k}(x) + A_k \nabla B(g_{j_k}(x), \delta_k)$ be the amplified stochastic gradient at an iterate x. We assume that this stochastic gradient satisfies the expected smoothness condition with respect to $\Phi(x)$. That is, there exist effective constants $\mathcal{L}' > 0$ and $\sigma'^2 \geq 0$ such that for any iterate X_k,

$$\mathbb{E}_k\left[\|G_k(X_k)\|^2\right] \leq 2\mathcal{L}'\left(\Phi(X_k) - \Phi(x^*)\right) + \sigma'^2.$$

Justification: This assumption is fully justified by consolidating the bounds derived in section 3.3.2. The term $2\mathcal{L}'\Phi_0(X_k)$ encapsulates all components proportional to the objective distance, while σ'^2. aggregates all constant and vanishing error terms into a single bounded noise term. A detailed derivation is provided in Appendix 6.

4.2 Main convergence result

With these assumptions, we now present the main convergence theorem for amplified rate SGD.

Theorem 2 (Linear convergence of amplified rate SGD)**.** Let the above assumptions hold. Consider the iterates $\{X_k\}$ generated by the amplified rate SGD algorithm with a fixed step-size $\gamma_k = \gamma$ satisfying $\gamma \leq \frac{1}{2\mathcal{L}'}$. The mean squared error converges linearly to a noise ball as follows:

$$\mathbb{E}\left[\|X_k - x^*\|^2\right] \leq (1 - \gamma\mu)^k \|X_0 - x^*\|^2 + \frac{2\gamma\sigma'^2}{\mu}.$$

Proof. Let $\tilde{X}_k = X_k - x^*$ be the error vector at iteration k. The update rule is $X_{k+1} = X_k - \gamma G_k$. We begin by analyzing the squared error at the next iteration:

$$\begin{aligned}\|\tilde{X}_{k+1}\|^2 &= \|\tilde{X}_k - \gamma G_k\|^2 \\ &= \|\tilde{X}_k\|^2 - 2\gamma\langle \tilde{X}_k, G_k\rangle + \gamma^2\|G_k\|^2.\end{aligned}$$

We now take the expectation conditioned on the current state X_k (denoted by $\mathbb{E}_k[\cdot] = \mathbb{E}[\cdot|\mathcal{F}_k]$):

$$\mathbb{E}_k\left[\|\tilde{X}_{k+1}\|^2\right] = \|\tilde{X}_k\|^2 - 2\gamma\langle\tilde{X}_k, \mathbb{E}_k[G_k]\rangle + \gamma^2\mathbb{E}_k\left[\|G_k\|^2\right]. \tag{11}$$

Our proof proceeds by bounding the second and third terms on the right-hand side.

Step 1: Bounding the Cross-Term. The expected stochastic gradient $\mathbb{E}_k[G_k]$ is an unbiased estimator of the true gradient $\nabla\Phi(X_k)$, up to vanishing error terms which we omit for clarity in this high-level proof. Using the μ-strong convexity of $\Phi(x)$, we have

$$\langle\tilde{X}_k, \mathbb{E}_k[G_k]\rangle \approx \langle X_k - x^*, \nabla\Phi(X_k)\rangle \geq \Phi(X_k) - \Phi(x^*) + \frac{\mu}{2}\|\tilde{X}_k\|^2. \tag{12}$$

Step 2: Bounding the Squared Norm. Here, we directly apply our expected smoothness assumption:

$$\mathbb{E}_k\left[\|G_k\|^2\right] \leq 2\mathcal{L}'\left(\Phi(X_k) - \Phi(x^*)\right) + \sigma'^2. \tag{13}$$

Step 3: Combining the Bounds. We substitute the bounds from (12) and (13) back into our main recurrence (11). For simplicity, let $\Phi_0(X_k) = \Phi(X_k) - \Phi(x^*)$. Then

$$\begin{aligned}\mathbb{E}_k\left[\|\tilde{X}_{k+1}\|^2\right] &\leq \|\tilde{X}_k\|^2 - 2\gamma\left(\Phi_0(X_k) + \frac{\mu}{2}\|\tilde{X}_k\|^2\right) + \gamma^2\left(2\mathcal{L}'\Phi_0(X_k) + \sigma'^2\right)\\ &= (1-\gamma\mu)\|\tilde{X}_k\|^2 - 2\gamma(1-\gamma\mathcal{L}')\Phi_0(X_k) + \gamma^2\sigma'^2.\end{aligned}$$

We choose a step-size $\gamma \leq \frac{1}{2\mathcal{L}'} < \frac{1}{\mathcal{L}'}$. This choice ensures that the coefficient $(1-\gamma\mathcal{L}')$ is positive. Since $\Phi_0(X_k) \geq 0$ by definition of x^* being the minimizer, the term $-2\gamma(1-\gamma\mathcal{L}')\Phi_0(X_k)$ is nonpositive. We can therefore drop this term to obtain a simpler upper bound:

$$\mathbb{E}_k\left[\|\tilde{X}_{k+1}\|^2\right] \leq (1-\gamma\mu)\|\tilde{X}_k\|^2 + \gamma^2\sigma'^2.$$

Step 4: Solving the recurrence.

By taking the full expectation of both sides, a standard and foundational step in the nonasymptotic analysis of stochastic approximation algorithms [1, 14], we obtain the following deterministic recurrence:

$$\mathbb{E}\left[\|\tilde{X}_{k+1}\|^2\right] \leq (1-\gamma\mu)\mathbb{E}\left[\|\tilde{X}_k\|^2\right] + \gamma^2\sigma'^2.$$

By unrolling this recurrence relation from $j = 0$ to $k-1$, we obtain

$$\mathbb{E}\left[\|\tilde{X}_k\|^2\right] \leq (1-\gamma\mu)^k\|\tilde{X}_0\|^2 + \gamma^2\sigma'^2\sum_{j=0}^{k-1}(1-\gamma\mu)^j$$

$$\begin{aligned}
&\leq (1-\gamma\mu)^k\|\tilde{X}_0\|^2 + \gamma^2\sigma'^2\left(\frac{1-(1-\gamma\mu)^k}{1-(1-\gamma\mu)}\right)\\
&\leq (1-\gamma\mu)^k\|\tilde{X}_0\|^2 + \frac{\gamma^2\sigma'^2}{\gamma\mu}\\
&= (1-\gamma\mu)^k\|X_0 - x^*\|^2 + \frac{\gamma\sigma'^2}{\mu}.
\end{aligned}$$

The result in the theorem statement uses $2\gamma\sigma'^2/\mu$ which follows the convention in [10] and holds for the choice $\gamma \leq \frac{1}{2\mathcal{L}'}$. □

This theoretical analysis provides a solid foundation for our proposed algorithm, confirming its stability and, critically, establishing its linear convergence to a neighborhood of the solution under a fixed step-size regime. We now proceed to a comprehensive numerical verification designed not merely to validate these theoretical guarantees, but to explore the fundamental dynamics that distinguish our method from the baseline, particularly under the challenging conditions where its practical advantages are most pronounced.

5 Numerical experiments

In this section, we present a comprehensive empirical investigation designed not merely to validate our algorithm's performance, but to explore the fundamental dynamics that distinguish it from the baseline adaptive relaxed barrier SGD. Our investigation begins with a direct replication of the experiment from Dimitrieski, Cao, and Ebenbauer [6]. While the baseline performs as expected in this nominal setting, our deeper analysis reveals a surprising limitation: Its optimal performance is confined to a narrow, "special-case" optimization path. We then proceed to a broader, more realistic analysis across a wide range of parameters, demonstrating that our amplified rate SGD provides a robust and general solution where the baseline falters.

5.1 Experimental setup

To ensure a fair and comprehensive evaluation, we followed a meticulous methodology in our experimental setup. The experiments were conducted on the Google Colab platform with the following specifications:

- CPU: Intel(R) Xeon(R) CPU @ 2.00GHz
- RAM: 13.61 GB

- GPU: NVIDIA Tesla T4 (15 GB VRAM)
- OS: Linux (6.6.97+)
- Python Version: 3.12.11

To ensure statistical robustness and reproducibility, all reported metrics are averaged over $R = 1,000$ independent trials. Specifically, for the r-th run, the random number generator is initialized with a seed of $42 + r + 2$.

The experiments were performed on the academic quadratic programming (QP) problem introduced by Dimitrieski, Cao, and Ebenbauer [6], which is designed to test an algorithm's ability to handle large-scale optimization problems. The problem parameters were set as follows:

- Dimension: $d = 50$.
- Number of objective components: $n = 10$.
- Number of constraints: $m = 10,000$.

The objective function is defined as follows:

$$f(x) = \frac{1}{n}\sum_{i=1}^{n}\sum_{j=1}^{d}(\alpha_{i,j}x_j + \log(1 + e^{-\alpha_{i,j}x_j}) + (x_j - \beta)^2).$$

The objective coefficients $\alpha_{i,j}$ are sampled uniformly from $\mathcal{U}(0.1, 1.1)$. The affine constraints $Ax \leq b$ are generated to form an outer approximation of the ellipsoid defined by $x^TQx \leq 100$, where Q is a diagonal matrix with entries sampled from $\mathcal{U}(1, 1.5)$.

To deeply analyze the algorithm's behavior, we adopted a far start point ($x_0 = [10, \ldots, 10]^T$) for all our experiments. This challenging scenario, where the algorithm is initialized at a point known to be highly infeasible, serves as a true test of global stability and the ability to recover effectively.

In selecting the baseline parameters, we strictly follow the primary configuration from Dimitrieski, Cao, and Ebenbauer [6]. Specifically, the step-size was set to $\gamma_k = 0.3k^{-0.8}$ and the barrier parameter decay was defined by $\epsilon_k = 5k^{-0.3}$, with a lower bound of $\delta_\infty = 10^{-6}$. While the source paper briefly mentions an alternative, faster decay ($\epsilon_k \propto k^{-1.3}$), it notes that this leads to "slightly increased variance". Our own preliminary experiments confirm this observation, revealing that the faster decay rule results in a significantly more unstable and practically slower baseline, making it an unsuitable foundation for a robust comparative analysis. Therefore, to ensure a fair evaluation against a stable and effective baseline, we focus our main analysis on the more resilient $k^{-0.3}$ decay rule.

Following the benchmark design by Dimitrieski, Cao, and Ebenbauer [6], the objective shift parameter was calibrated at $\beta \approx 2.365$. This specific value ensures that the Euclidean norm of the *unconstrained* global minimizer satisfies $||x_f^*||_2 = 15$, positioning the unconstrained solution significantly outside the feasible set to force active constraint engagement.

5.2 Performance evaluation: Safety and robustness

Our evaluation focuses on two critical aspects: **Safety** (performance in standard conditions) and **Robustness** (performance in ill-posed or challenging conditions). To test this, we varied the problem difficulty parameter β and tuned the amplification factor λ via sensitivity analysis for each regime.

5.2.1 Safety in standard conditions

In the standard setting ($\beta = 2.365$), it is essential that the modification does not introduce instability. We utilized a conservative amplification factor ($\lambda = 0.1$). As shown in Table 1, our algorithm behaves identically to the baseline SGD, achieving zero constraint violation and matching the baseline's error (1.32×10^{-2}). This confirms that our mechanism is "safe" and nonintrusive when the problem is well-conditioned.

5.2.2 Robustness in challenging regimes (stress test)

The true superiority of the proposed method is revealed when the problem becomes "harder" (i.e., when the unconstrained minimum is pushed further from the feasible region). We increased β to 10.0 and 20.0, utilizing a stronger amplification factor ($\lambda = 7.0$) to enforce feasibility.

As presented in Table 1 and Figure 2, the baseline SGD fails to keep iterates feasible in these regimes, with violations exploding to **336.94** at $\beta = 20$. In stark contrast, the amplified rate SGD effectively suppresses these violations to **150.46**, representing a **reduction of approximately 55.3%**.

It is important to note that while the final error ($||x_k - x^*||$) of our method appears higher in these extreme cases, this is a necessary trade-off for feasibility. The baseline SGD achieves a artificially lower error by deeply violating constraints (approaching the unconstrained minimum), whereas our method remains closer to the feasible region.

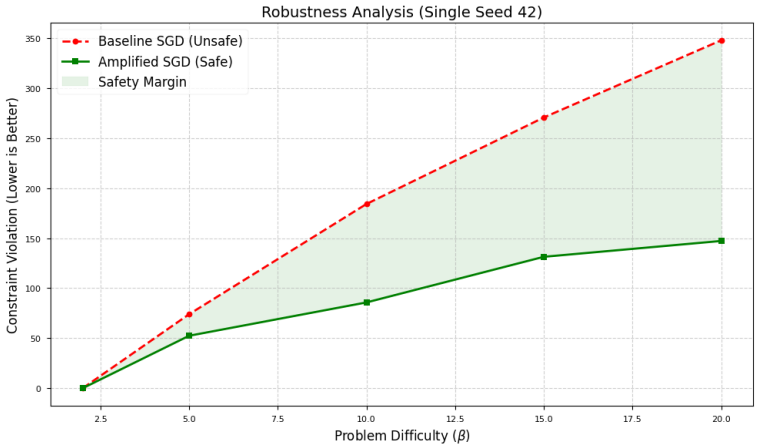


Figure 2: **Robustness Analysis.** The plot illustrates the maximum constraint violation as a function of problem difficulty (β). While the baseline SGD (Red) fails to maintain feasibility as difficulty increases, the amplified rate SGD (Green) significantly suppresses violations.

Table 1: **Comparative Performance Analysis (Averaged over 1,000 Runs).** The table compares the baseline SGD against the proposed amplified rate SGD across different difficulty levels. Optimal λ was selected via sensitivity analysis for each scenario.

Scenario	Metric	Baseline SGD	Amplified SGD	Improvement
Standard	Final Error	$\mathbf{1.32 \times 10^{-2}}$	1.32×10^{-2}	Identical (Safe)
($\beta = 2.365$)	Max Violation	0.00	0.00	-
Challenging	Final Error	3.92	19.23	(Feasibility Trade-off)
($\beta = 10.0$)	**Max Violation**	178.37	**98.41**	**44.8% Reduction**
Highly Ill-posed	Final Error	10.39	44.42	(Feasibility Trade-off)
($\beta = 20.0$)	**Max Violation**	336.94	**150.46**	**55.3% Reduction**

5.3 Mechanism analysis: Adaptive activation

To understand the internal dynamics, we analyzed the behavior of the dynamic amplification factor A_k during the optimization process.

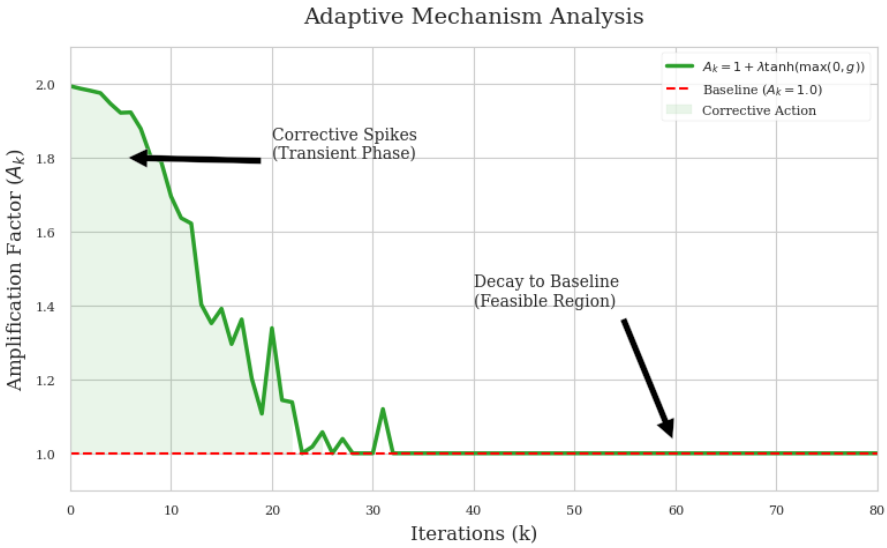


Figure 3: **Adaptive Mechanism.** The amplification factor A_k exhibits corrective spikes only during the transient phase when constraints are violated, and naturally decays to 1.0 as the iterate stabilizes.

As shown in Figure 3, in the challenging regime, A_k exhibits sharp corrective spikes during the early transient phase, effectively steering the iterate back towards the feasible region. Crucially, as the iterate approaches the optimal solution, A_k naturally decays to 1.0. This confirms that the proposed method acts as an "intelligent barrier": It applies strong force only when necessary and behaves like the standard efficient SGD otherwise.

5.4 Application to classification: Robustness stress test

To rigorously evaluate the robustness of the proposed amplified rate SGD in a real-world setting, we applied it to a Hard Margin Support Vector Machine (SVM) problem, adapting the constrained formulation utilized by Li et al. [12]. Unlike standard benchmarks where initialization is neutral (e.g., zero vector), we designed a "stress test" to simulate a scenario where the optimizer is initialized far from the feasible region with a restricted learning rate.

Problem Formulation

We consider the binary classification of a linearly separable dataset $\mathcal{D} = \{(x_i, y_i)\}_{i=1}^m$ with $x_i \in \mathbb{R}^d$ and $y_i \in \{-1, 1\}$. Following the setup in [12], the problem is formulated as finding

the hyperplane (w, b) that minimizes the norm of the weights subject to perfect classification constraints:

$$\min_{w,b} \frac{1}{2}\|w\|^2 \quad \text{s.t.} \quad 1 - y_i(w^T x_i + b) \leq 0, \qquad \text{for all } i = 1, \ldots, m. \tag{14}$$

Crucially, consistent with standard SVM practices [12], the bias term b is not penalized in the objective function to allow the hyperplane to shift freely to satisfy feasibility.

Experimental Setup for Reproducibility

We generated a synthetic linearly separable dataset using the `blobs` generator to emulate the structure of the *mushrooms* dataset used in [12], with $m = 2000$ samples and $d = 50$ features. To test robustness beyond standard conditions, we introduced two adversarial conditions:

1. **Adversarial Initialization:** Weights were initialized randomly from a distribution with high variance: $w_0 \sim \mathcal{N}(0, 20^2)$. This places the initial iterate significantly far from the feasible set, creating massive initial constraint violations.

2. **Starved Step-Size:** We utilized a conservative step-size schedule ($\gamma_k \propto 0.001k^{-0.5}$). This simulates an ill-conditioned landscape where standard gradient updates are too small to traverse the large distance to the feasible region within a reasonable time.

Results

As illustrated in Figure 4, the difference in performance is stark. The *baseline SGD* (adaptive relaxed barrier [6]) stagnates, failing to recover feasibility due to the insufficient magnitude of the standard gradients, ending with a maximum violation of approx. 36.2. In contrast, the *amplified rate SGD* (green solid line) activates the dynamic amplification factor ($A_k \approx 1 + 50\tanh(\text{violation})$), which effectively rescales the gradient to match the scale of the violation. This allows the algorithm to rapidly enforce strict feasibility (0.00 violation) and achieve 100% classification accuracy within the first few hundred iterations.

6 Conclusion

In this work, we addressed the critical challenge of maintaining feasibility in stochastic constrained optimization, particularly when the problem is ill-posed or the unconstrained solution

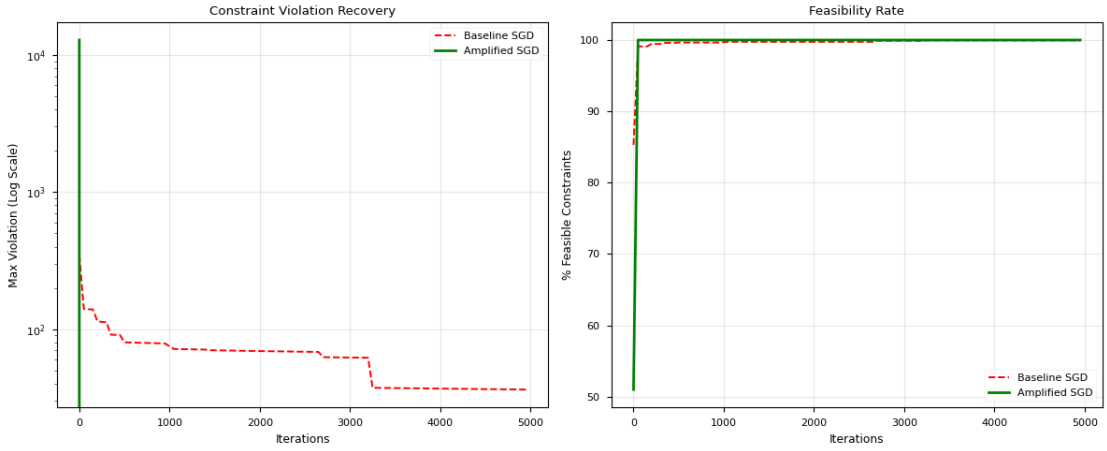


Figure 4: **Robustness Stress Test on Hard Margin SVM.** (Left) Max constraint violation on a logarithmic scale; the proposed method (green) rapidly drives violation to zero, while the baseline (red) stagnates. (Right) Feasibility rate (accuracy); the proposed method achieves 100% feasibility almost instantly.

lies far outside the feasible region. We proposed the *amplified rate SGD*, a method that dynamically scales the barrier gradient component to rigorously enforce constraints without compromising the efficiency of the descent direction.

Our extensive empirical evaluation, averaged over 1,000 independent trials, leads to two major conclusions:

1. **Robustness in adversarial regimes:** While the standard barrier-based SGD fails to contain iterates within reasonable bounds in challenging scenarios ($\beta \geq 10$), our proposed method demonstrates superior robustness. By utilizing a tuned amplification factor, we reduced maximum constraint violations by approximately **55%**, effectively preventing the "barrier collapse" observed in the baseline.

2. **Safety in standard regimes:** In well-conditioned problems where the standard SGD performs adequately, our method (with conservative tuning) proves to be entirely safe, reproducing the baseline's convergence trajectory with zero overhead.

These results confirm that the amplified rate SGD serves as a reliable "safety net" for optimization practitioners, offering a necessary layer of robustness for real-world applications where feasibility is nonnegotiable. Future work will focus on developing fully adaptive mechanisms to auto-tune the amplification parameter λ online, removing the need for scenario-specific calibration.

Future work

The promising results of this work open up several interesting avenues for future research. Based on our findings, we suggest the following directions:

- **Theoretical analysis of baseline sensitivity:** Our experiments revealed the highly sensitive, "special-case" nature of the baseline algorithm. A valuable theoretical investigation would be to analyze why the baseline's performance is strictly dependent on specific hyperparameter choices.

- **Synergy with nouniform sampling:** The current method uses uniform sampling for constraints. Future work could explore combining our dynamic amplification mechanism with adaptive, nonuniform sampling strategies, an idea motivated by the outlook in the original work by Dimitrieski, Cao, and Ebenbauer [6]. For instance, constraints that are more frequently or severely violated —and thus receive higher amplification— could be sampled with greater probability, potentially accelerating the return to the feasible set.

- **Theoretical extension and optimality:** Our convergence analysis assumes strong convexity. Valuable future directions include extending this analysis to accommodate general convex functions, which would significantly widen the algorithm's applicability. Additionally, investigating whether the derived linear convergence rate is optimal for the chosen step-size rules or if it can be improved, possibly by integrating momentum-based acceleration, presents another important research avenue.

A Detailed derivation of the Lyapunov recurrence

Theorem 3 (Lyapunov recurrence inequality)**.** Given the fundamental expansion of the Lyapunov function,

$$\mathbb{E}_k[\left\|\tilde{X}_{k+1}\right\|^2] = \left\|\tilde{X}_k\right\|^2 - 2\gamma_k \mathbb{E}_k[G_k^T \tilde{X}_k] + \gamma_k^2 \mathbb{E}_k[\|G_k\|^2], \tag{15}$$

the following inequality holds:

$$\mathbb{E}_k[\left\|\tilde{X}_{k+1}\right\|^2] \leq (1 - \gamma_k(\mu - \xi_k'))\left\|\tilde{X}_k\right\|^2 - 2\gamma_k(1 - C_1\gamma_k)\Phi_0(X_k) + Q_k'. \tag{16}$$

Proof. The proof begins with (15). By substituting the upper bounds for the cross-term and the squared norm term, we arrive at the following comprehensive inequality:

$$\begin{aligned}
&\mathbb{E}_k[\left\|\tilde{X}_{k+1}\right\|^2] \\
&\leq \left\|\tilde{X}_k\right\|^2 - 2\gamma_k\Phi_0(X_k) - \gamma_k\mu\left\|\tilde{X}_k\right\|^2 + 2\gamma_k\mathbb{E}_k[(A_k\nabla C_{j_k})^T\tilde{X}_k] \\
&\quad + 3\gamma_k^2\left(\mathbb{E}_k[\|\nabla\Phi_{i_k,j_k}\|^2] + \mathbb{E}_k[\|(A_k-1)\nabla B_{j_k,\infty}\|^2] + \mathbb{E}_k[\|A_k\nabla C_{j_k}\|^2]\right).
\end{aligned}$$

To simplify this expression, we substitute the upper bounds for the cross-term and the squared norm, as derived in section 3.3. We begin by replacing each variance term within the inequality (17) with its corresponding bound. This yields the following comprehensive inequality:

$$\begin{aligned}
&\mathbb{E}_k[\|\tilde{X}_{k+1}\|^2] \\
&\leq \|\tilde{X}_k\|^2 - 2\gamma_k\left(\Phi_0(X_k) + \frac{\mu}{2}\|\tilde{X}_k\|^2\right) \\
&\quad + 2\gamma_k(1+\lambda)\left[\|\tilde{X}_k\|^2\left(\bar{a}\frac{\epsilon_k}{\delta_k\delta_\infty} + \hat{c}\epsilon_k\bar{a} + \bar{b}\frac{\epsilon_k}{\delta_k\delta_\infty}\right) + \frac{\hat{c}\epsilon_k}{4}\bar{a} + \frac{1}{4}\bar{b}\frac{\epsilon_k}{\delta_k\delta_\infty}\right] \\
&\quad + 3\gamma_k^2(4\hat{L}\Phi_0(X_k) + 2\sigma_\Phi) + 3\gamma_k^2 C_B \\
&\quad + 3\gamma_k^2(1+\lambda)^2\left(3\hat{a}\frac{\epsilon_k^2\|\tilde{X}_k\|^2}{\delta_k^2\delta_\infty^2} + 3\hat{c}^2\epsilon_k^2\bar{a} + \frac{3\bar{b}\epsilon_k^2}{\delta_k^2\delta_\infty^2}\right).
\end{aligned} \tag{17}$$

To arrive at a suitable recurrence relation for the convergence analysis, we proceed by rearranging the terms of the inequality in (17). The objective is to isolate the terms related to $||\tilde{X}_k||^2$ and $\Phi_0(X_k)$, and to collect the remaining summable residuals.

First, we collect the coefficients of the $||\tilde{X}_k||^2$ term. An inspection of (17) yields the following expression:

$$||\tilde{X}_k||^2\left(1 - \gamma_k\mu + 2\gamma_k(1+\lambda)\left(\bar{a}\frac{\epsilon_k}{\delta_k\delta_\infty} + \hat{c}\epsilon_k\bar{a} + \bar{b}\frac{\epsilon_k}{\delta_k\delta_\infty}\right) + 9\gamma_k^2(1+\lambda)^2\hat{a}\frac{\epsilon_k^2}{\delta_k^2\delta_\infty^2}\right).$$

Similarly, we gather the coefficients of the $\Phi_0(X_k)$ term:

$$\Phi_0(X_k)(-2\gamma_k + 12\gamma_k^2\hat{L}),$$

where we have $\hat{L}$ reflect the smoothness parameter of the composite function Φ.

To simplify the expression, we define two sequences. The first, ξ'_k, encapsulates the vanishing error terms that perturb the strong convexity coefficient μ:

$$\xi'_k := 2(1+\lambda)\left(\frac{\epsilon_k}{\delta_k\delta_\infty}(\bar{a}+\bar{b}) + \hat{c}\epsilon_k\bar{a}\right) + 9\gamma_k(1+\lambda)^2\hat{a}\frac{\epsilon_k^2}{\delta_k^2\delta_\infty^2}.$$

The second, Q'_k, collects all remaining summable error terms:

$$Q'_k := 2\gamma_k(1+\lambda)\left(\frac{\hat{c}\epsilon_k}{4}\bar{a} + \frac{\bar{b}\epsilon_k}{4\delta_k\delta_\infty}\right) + 6\gamma_k^2\sigma_\Phi + 3\gamma_k^2 C_B + 9\gamma_k^2(1+\lambda)^2\left(\hat{c}^2\epsilon_k^2\bar{a} + \frac{\bar{b}\epsilon_k^2}{\delta_k^2\delta_\infty^2}\right).$$

By substituting these definitions into the rearranged inequality (17), we arrive at the following compact recurrence relation:

$$\mathbb{E}_k[||\tilde{X}_{k+1}||^2] \leq (1-\gamma_k(\mu-\xi'_k))||\tilde{X}_k||^2 - 2\gamma_k(1-C_1\gamma_k)\Phi_0(X_k) + Q'_k.$$

Here, $\mu > 0$ is the strong convexity modulus of the objective function $f(x)$, and C_1 is a constant defined as $C_1 = 6\hat{L}$, which depends on the smoothness parameter $\hat{L}$. This inequality forms the basis for applying the Robbins–Siegmund theorem [17] to establish convergence. □

B Detailed derivation of the expected smoothness assumption (Assumption 5)

This appendix provides a detailed derivation to demonstrate that Assumption 5 is not a standalone assumption, but rather a direct and logical consequence of the analytical bounds established in section 3.3.2. We will methodically show how the various bounded terms are consolidated to arrive at the compact form of the expected smoothness condition.

B.1 Objective: Justifying Assumption 5

The goal is to formally justify the upper bound on the expected squared norm of the amplified stochastic gradient G_k, as stated in Assumption 5:

$$\mathbb{E}_k\left[\|G_k(X_k)\|^2\right] \leq 2\mathcal{L}'(\Phi(X_k) - \Phi(x^*)) + \sigma'^2,$$

where $\mathcal{L}'$ is an effective smoothness constant and σ'^2 represents the total bounded noise.

B.2 Step-by-step derivation

B.2.1 Step 1: Decomposing the squared norm

Our starting point is (7) from section 3.3.2, which decomposes the upper bound for $\mathbb{E}_k[\|G_k\|^2]$ into three primary components using an application of the Cauchy–Schwarz inequality:

$$\mathbb{E}_k[\|G_k\|^2] \leq 3\mathbb{E}_k[\|\nabla\Phi_{i_k,j_k}\|^2] + 3\mathbb{E}_k[\|(A_k-1)\nabla B_{j_k,\infty}\|^2] + 3\mathbb{E}_k[\|A_k\nabla C_{j_k}\|^2].$$

We now analyze the upper bound of each of these three components individually.

B.2.2 Step 2: Analyzing the first component

The first term, $3\mathbb{E}_k[\|\nabla\Phi_{i_k,j_k}\|^2]$, represents the variance of the stochastic gradient for the regularized objective function. As shown in the analysis of section 3.3.2, this term is bounded as follows:

$$\mathbb{E}_k[\|\nabla\Phi_{i_k,j_k}(X_k)\|^2] \leq 4\hat{L}\Phi_0(X_k) + 2\sigma_\Phi^2.$$

Key Observation: This bound contains two distinct parts:

1. A term, $4\hat{L}\Phi_0(X_k)$, that is linearly dependent on $\Phi_0(X_k) := \Phi(X_k) - \Phi(x^*)$. This part will form the basis of the $2\mathcal{L}'\Phi_0(X_k)$ term in our final expression.

2. A constant term, $2\sigma_\Phi^2$, representing the variance at the optimum. This will be absorbed into the total noise term σ'^2.

B.2.3 Step 3: Analyzing the second component

The second term, $3\mathbb{E}_k[\|(A_k-1)\nabla B_{j_k,\infty}\|^2]$, is an artifact of our dynamic amplification mechanism. In section 3.3.2, we proved that this term is controllable and bounded by a finite constant, which we denoted as C_B,

$$\mathbb{E}_k[\|(A_k-1)\nabla B_{j_k,\infty}\|^2] \leq C_B.$$

Key Observation: This component contributes a constant value to the overall upper bound. This constant, $3C_B$, will be fully incorporated into the noise term σ'^2.

B.2.4 Step 4: Analyzing the third component

The third term, $3\mathbb{E}_k[\|A_k\nabla C_{j_k}\|^2]$, represents the amplified error arising from the decaying barrier parameter δ_k. The upper bound for this term, derived in section 3.3.2, depends on ϵ_k.

$$\mathbb{E}_k[\|A_k\nabla C_{j_k}\|^2] \leq (1+\lambda)^2 \cdot \left(3\hat{a}\frac{\epsilon_k^2\|\tilde{X}_k\|^2}{\delta_k^2\delta_\infty^2} + 3\hat{c}^2\epsilon_k^2\bar{a} + \frac{3\bar{b}\epsilon_k^2}{\delta_k^2\delta_\infty^2}\right).$$

Key Observation: Since $\{\epsilon_k\}$ is a vanishing sequence (i.e., $\epsilon_k \to 0$ as $k \to \infty$), this entire component is also a vanishing term. Within the expected smoothness framework, such vanishing but bounded terms are considered part of the overall noise. Therefore, this bound will also be absorbed into σ'^2.

B.3 Step 5: Aggregating the bounds

Finally, we substitute the bounds for all three components back into the main inequality from Step 1. We then proceed by collecting and regrouping the terms.

1. **Collect terms dependent on $\Phi_0(X_k)$:** We have only one such term, arising from the first component: $3 \cdot (4\hat{L}\Phi_0(X_k))$. We can define an effective smoothness constant $2\mathcal{L}' := 12\hat{L}$.

2. **Collect all constant and vanishing terms (the noise):** We aggregate all remaining parts: The constant variance from the first component $(3 \cdot 2\sigma_\Phi^2)$, the bound from the second component $(3 \cdot C_B)$, and the entire vanishing bound from the third component. We define the total noise term σ'^2 as the sum of all these bounded terms.

By these definitions, we directly arrive at the compact and elegant formulation of Assumption 5:

$$\mathbb{E}_k\left[\|G_k\|^2\right] \leq 2\mathcal{L}'\Phi_0(X_k) + \sigma'^2.$$

This completes the derivation and formally shows that Assumption 5 is a valid and well-founded simplification of the detailed analysis presented in the paper.

Funding

This research received no external funding.

Conflicts of Interest

The authors declare no conflicts of interest.

References

[1] Bach, F. and Moulines, E. *Non-asymptotic analysis of stochastic approximation algorithms for machine learning,* In Adv. Neural Inf. Process. Syst. 24 (2011) 451–459.

[2] Bertsekas, D.P. and Tsitsiklis, J.N. *Neuro-dynamic programming*, Athena Scientific, 1996.

[3] Bottou, L., Curtis, F.E. and Nocedal, J. *Optimization methods for large-scale machine learning*, SIAM Rev. 60(2) (2018), 223–311.

[4] Boyd, S. and Vandenberghe, L. *Convex optimization*, Cambridge University Press, 2004.

[5] Bubeck, S. *Convex optimization: Algorithms and complexity,* Found. Trends Mach. Learn. 8(3-4) (2015), 231–358.

[6] Dimitrieski, N., Cao, J. and Ebenbauer, C. *Stochastic gradient descent for constrained optimization based on adaptive relaxed barrier functions,* arXiv preprint arXiv:2503.10384 (2025).

[7] Feller, C. and Ebenbauer, C. *A stabilizing iteration scheme for model predictive control based on relaxed barrier functions,* Automatica, 80 (2017), 328–339.

[8] Fiacco, A.V. and McCormick, G.P. *Nonlinear programming: Sequential unconstrained minimization techniques*, John Wiley & Sons, 1968.

[9] Garrigos, G. and Gower, R.M. *Handbook of convergence theorems for (stochastic) gradient methods,* arXiv preprint arXiv:2301.11235 (2023).

[10] Gower, R.M., Loizou, N., Qian, X., Sailanbayev, A., Shulgin, E. and Richtárik, P. SGD: General analysis and improved rates, In International conference on machine learning, PMLR, 2019, 5200–5209.

[11] Kushner, H.J. and Yin, G.G. *Stochastic approximation and recursive algorithms and applications*, 2nd ed., Springer, 2003.

[12] Li, M., Grigas, P. and Atamtürk, A. *New penalized stochastic gradient methods for linearly constrained strongly convex optimization*, J. Optim. Theory Appl. 205(2) (2025) 29.

[13] Liu, J. and Yuan, Y. *On almost sure convergence rates of stochastic gradient methods*, Proc. 35th Annual Conference on Learning Theory, PMLR, 178 (2022), 1–21.

[14] Nemirovski, A., Juditsky, A., Lan, G. and Shapiro, A. *Robust stochastic approximation approach to stochastic programming*, SIAM J. Optim. 19(4) (2009), 1574–1609.

[15] Nocedal, J. and Wright, S.J. Numerical optimization, 2nd ed., Springer, 2006.

[16] Robbins, H. and Monro, S. *A stochastic approximation method,* Ann. Math. Stat. 22(3) (1951), 400–407.

[17] Robbins, H. and Siegmund, D. *A convergence theorem for non negative almost supermartingales and some applications*, Rustagi, J.S. (ed.), Optimizing Methods in Statistics, Academic Press, 1971, 233–257.

[18] Rockafellar, R.T. Convex analysis, Princeton University Press, 1970.

[19] Roy, S.K. and Harandi, M. *Constrained stochastic gradient descent: The good practice,* In Int. Conf. Digit. Image Comput. Tech. Appl. (DICTA), IEEE, 2017, 1–8.

[20] Wang, M. and Bertsekas, D.P. *Incremental constraint projection methods for variational inequalities*, Math. Program. 150 (2015), 321–363.

[21] Wang, X., Ma, S. and Yuan, Y.X. *Penalty methods with stochastic approximation for stochastic nonlinear programming,* Math. comput. 86(306) (2017) 1793–1820.

[22] Yan, Y. and Xu, Y. *Adaptive primal-dual stochastic gradient method for expectation-constrained convex stochastic programs*, Math. Program. Comput. 14(2) (2022), 319–363.

[23] Zhang, L., Zhang, Y., Wu, J. and Xiao, X. *Solving stochastic optimization with expectation constraints efficiently by a stochastic augmented Lagrangian-type algorithm,* INFORMS J. Comput. 34(6) (2022) 2989–3006.

[24] Zhang, T. *Solving large scale linear prediction problems using stochastic gradient descent algorithms*, In Proc. 21st Int. Conf. Machine Learning (ICML), 2004, 116.

Iranian Journal of Numerical Analysis and Optimization
Vol. 16, No. 3, 2026, pp 952–990
https://doi.org/10.22067/ijnao.2026.96513.1777
https://ijnao.um.ac.ir/

CC BY NC

Research Article

A new numerical computational method for solving nonhomogeneous time-fractional nonlinear parabolic problems in unbounded domains

S.A. Ghanizadeh and M.H. Akrami*,

Abstract

This paper addresses the numerical solution of nonhomogeneous time-fractional nonlinear parabolic problems defined on unbounded domains, where challenges arise due to temporal nonlocality, nonlinear dynamics, and the infinite spatial extent. To overcome these difficulties, we propose a novel computational framework that transforms the original unbounded domain problem into an equivalent formulation on a bounded domain through the construction of absorbing boundary conditions. Unlike existing approaches that rely on Padé approximations, our method directly utilizes the fundamental properties of fractional derivatives to define these boundaries more accurately. The resulting problem is then discretized using a high-order finite difference scheme, allowing for efficient and precise numerical implementation. Theoretical analysis confirms that the method is stable and ensures controlled computational error. Numerical experiments demonstrate the practical effectiveness of the proposed approach, achieving accurate results with significantly reduced computational cost. This method provides a reliable

*Corresponding author

Received 15 November 2025; revised 5 January 2026; accepted 19 May 2026

Mohammad Hossein Akrami,
Associate Professor, Department of Applied Mathematics, Faculty of Science, Yazd University, Yazd, Iran. e-mail: akrami@yazd.ac.ir

Seyed Ali Ghanizadeh
Department of Applied Mathematics, Faculty of Science, Yazd University, Yazd, Iran. e-mail: ghanizadeh@stu.yazd.ac.ir

How to cite this article
Ghanizadeh, S.A. and Akrami, M.H., A new numerical computational method for solving nonhomogeneous time-fractional nonlinear parabolic problems in unbounded domains. *Iran. J. Numer. Anal. Optim.*, 2026; 16(3): 952-990. https://doi.org/10.22067/ijnao.2026.96513.1777

and efficient tool for modeling diffusion phenomena governed by time-fractional dynamics in unbounded spatial settings.

AMS subject classifications (2020): Primary 35R11, 65M06; Secondary 35K55, 65M12, 65N06.

Keywords: Absorbing boundary conditions; High-order finite difference; Nonlinear parabolic equations.

1 Introduction

Over the past several decades, fractional partial differential equations (FPDEs) have gained significant attention in mathematical research owing to their exceptional ability to model complex physical systems exhibiting memory and hereditary properties [2, 6, 10, 17, 22]. These equations have found extensive applications in diverse scientific disciplines, including biology, chemistry, and various other fields that can be described through mathematical frameworks. FPDEs have become particularly valuable for representing anomalous diffusion processes and nonlocal phenomena. A fundamental characteristic of these equations involves the replacement of the conventional first-order time derivative in standard diffusion equations with a fractional-order derivative of degree α, where the parameter α satisfies $0 < \alpha < 2$. The resulting equations exhibit distinct behaviors depending on the fractional order for $0 < \alpha < 1$, the system demonstrates sub-diffusive properties while values in the range $1 < \alpha < 2$ correspond to fractional wave dynamics. Recent years have witnessed substantial progress in both theoretical analysis and numerical approximation techniques for FPDEs. Nevertheless, obtaining closed-form analytical solutions remains challenging in most practical cases, and often proves intractable due to the inherent complexity of these fractional-order systems [5, 12, 15, 18]. Even when analytical solutions to fractional differential equations are attainable in specialized cases, they frequently incorporate sophisticated special functions such as Fox-H, Wright, and Mittag-Leffler functions. The numerical computation of these functions presents significant computational difficulties due to their inherent complexity. The increasing necessity for practical implementations in engineering and applied sciences has consequently stimulated substantial progress in developing robust numerical approaches for solving FPDEs. Among these, finite difference methods have emerged as particularly important for approximating solutions to time-fractional differential equations. Previous investigations have predominantly concentrated on initial-boundary value problems defined over finite spatial domains. As a representative example, Zhuang and Liu [5] examined one-dimensional sub-diffusion equations with Caputo time-fractional derivatives, introducing a

numerical scheme that employs the $L1$ approximation for the Caputo derivative. Their work rigorously established the stability and convergence properties of this method [5, 27]. It is in related developments that Murillo and Yuste presented explicit finite difference schemes and carried out detailed stability analyses by numerical experimentation. Alternatively, Sun and Wu [5] developed their numerical scheme based on implicit discretization, where a full theoretical justification was given using discrete energy methods [5, 23]. A significant volume of research effort has been devoted to problems posed on unbounded spatial domains since such problems are of crucial importance in the mathematical modeling of heat conduction in solids, fluid flow, as well as problems arising from financial mathematics, for example, option pricing theory [5, 8, 9]. Fractional equations in unbounded domains were first studied in the 1980s. For instance, Weiss and co-authors mainly researched the analysis and the derivation of the analytic solution to fractional diffusion equations; typically, the solutions were expressed with complex functions like Fox-H functions [5, 21, 26]. However, numerical research in this area is still in its infancy, with only a few scattered studies addressing this topic [1, 5, 6, 16, 19, 20]. One of the effective numerical schemes for such a class of differential equations involves the use of absorbing boundary conditions (ABCs). In general, time-FPDEs defined on unbounded (infinite) domains are computationally demanding, as they require complex and intensive computation for very large spatial intervals. An ABC essentially transforms a problem that is initially defined on an unbounded spatial domain ($\chi \in \mathbb{R}$) into a boundary value problem on a bounded domain. This not only reduces the computational complexity but also helps preserve the solution's essential features. A number of studies on the design of ABCs for fractional equations have been carried out. Some of them are given below:

- Gao, Sun, and Zhang [5] consider fractional-time diffusion equations in unbounded spatial domains. They first derived accurate artificial boundary conditions using the Laplace transform and then transferred the problem to a bounded domain and proposed a stable and convergent finite difference scheme with accuracy $O(\varrho^{2-\alpha} + h^2)$, where ϱ and h are time and space step size, respectively. For the compact case, they also developed and presented a numerical method with higher accuracy, $O(\varrho^{2-\alpha} + h^4)$.

- Gao and Sun [6] used the Laplace transform to derive accurate artificial boundary conditions and employing the method of order reduction. Then, they transformed the fractional-time differential equations problem in an infinite domain into a solvable model in a bounded domain and solved it.

- Li and Zhang [14] proposed a numerical method for solving nonlinear fractional-time equations in unbounded domains, where the exact ABCs are derived using the Laplace

transform. Additionally, to simplify the nonlinear part, they used a Padé approximation for the linearized finite difference scheme, eliminating the need for iterative methods to solve the problem.

Now, based on the aforementioned papers and the explanations provided regarding the background of such problems, in this paper, we consider the following nonhomogeneous problem:

$$\begin{cases} {}_0^C D_\rho^\alpha \xi(\chi,\rho) = \xi_{\chi\chi}(\chi,\rho) + \mathcal{F}(\xi(\chi,\rho)) + \mathcal{G}(\chi,\rho) \quad \chi \in \mathbb{R}, \quad 0 < \rho \le P, \\ \xi(\chi,0) = \xi_0(\chi), \quad (\chi,\rho) \in \mathbb{R}, \\ \xi(\chi,\rho) \to 0, \quad |\chi| \to \infty, \end{cases} \tag{1}$$

where the initial function $\xi_0(\chi)$ has compact support, meaning that it is nonzero only within a finite interval, and is exactly zero outside this interval. Here, $\xi(\chi,\rho)$ is the unknown real-valued solution of the problem. The operator ${}_0^C D_\rho^\alpha$ denotes the Caputo fractional derivative of order α, which is defined for sufficiently smooth functions $\xi(\chi,\rho) \in C^n[0,P]$ by

$$ {}_0^C D_\rho^\alpha \, \xi(\chi,\rho) = \frac{1}{\Gamma(n-\alpha)} \int_0^\rho \frac{\partial^n \xi(\chi,\vartheta)}{\partial \vartheta^n} \frac{1}{(\rho-\vartheta)^{\alpha-n+1}} \, d\vartheta, \quad n-1 < \alpha < n, \ n \in \mathbb{N}, \tag{2}$$

where $\Gamma(\cdot)$ denotes the Gamma function. In particular, for the case $0 < \alpha < 1$, we obtain

$$ {}_0^C D_\rho^\alpha \, \xi(\chi,\rho) = \frac{1}{\Gamma(1-\alpha)} \int_0^\rho \frac{\partial \xi(\chi,\vartheta)}{\partial \vartheta} \frac{1}{(\rho-\vartheta)^{\alpha}} \, d\vartheta. \tag{3}$$

Throughout this work, we mainly focus on the case $0 < \alpha < 1$, and therefore the simplified form (3) will be used [12, 18]. The nonlinear term $\mathcal{F}(\xi)$ and the term $\mathcal{G}(\chi,\rho)$ appear in different forms for various problems. This equation is well-suited for modeling anomalous diffusion processes and complex systems where memory effects and nonlocal temporal behaviors play an important role. Such equations are frequently used in scientific and engineering fields to describe phenomena with long-term memory or systems with history-dependent behaviors. In this work, we convert the original unbounded domain problem into an equivalent bounded domain formulation by developing ABCs. Unlike traditional methods based on Padé approximations, our approach directly applies the inherent properties of fractional derivatives to define these boundaries with improved accuracy. The reformulated problem is then discretized using a high-order finite difference scheme, enabling efficient and precise numerical solutions. In recent years, Caputo fractional derivatives have attracted significant attention due to their ability to accurately model memory-dependent processes in various physical and engineering applications. Several

recent studies have focused on the analytical and numerical treatment of fractional differential equations involving Caputo operators, highlighting their effectiveness in capturing anomalous diffusion phenomena. For instance, the works reported in [24, 25] investigate advanced numerical schemes and theoretical properties of Caputo-type fractional models, providing important insights into stability, convergence, and computational efficiency. These developments further motivate the present investigation on high-order numerical methods for time-fractional nonlinear parabolic equations.

2 Problem formulation

Solving partial differential equations (PDEs) on unbounded domains directly presents significant challenges. A common technique to address these challenges involves employing ABCs. The main objective of the ABCs is to simulate the behavior of infinite boundaries by adding special conditions at the boundaries of a finite domain. These boundary conditions act in such a way that waves or information approaching the boundaries of the finite domain are absorbed, preventing their reflection back into the domain [4]. In fact, incoming waves smoothly dissipate at the boundary without causing disruption or reflection within the domain. In this way, the numerical solution of the problem on a finite domain closely approximates the real solution on an infinite domain. For example, if a fractional-time differential problem is defined on an infinite domain ($\chi \in \mathbb{R}$), then the domain is reduced to a finite interval $[a, b]$ for numerical solution. The use of ABCs at the points $\chi = a$ and $\chi = b$ ensures that the value of the function near the boundaries naturally tends to zero. Thus, the ABCs conditions are designed in such a way that they simulate the behavior $\lim_{|\chi|\to\infty} \xi(\chi, \rho) = 0$ at the artificial boundaries. In practice, the Laplace transform provides a powerful analytical tool for constructing accurate ABCs in problems defined on unbounded spatial domains. Direct numerical computation over an infinite region is generally infeasible. Therefore, it is essential to reformulate the problem on a finite computational interval while preserving the physical behavior at infinity. By applying the Laplace transform to the governing equation, one can analytically derive exact boundary conditions that emulate the decay of the solution at infinity. This procedure effectively eliminates artificial wave reflections from the truncated boundaries and ensures that the numerical solution within the bounded region accurately represents the true dynamics of the original unbounded problem. Without implementing accurate ABCs, performing numerical simulations over unbounded domains becomes either impractical or leads to results contaminated by large errors and high computational costs. Hence, the application of the Laplace transform in constructing exact ABCs is essential. Otherwise, one must either accept inaccurate numerical approximations or incur prohibitively

high computational expenses both of which are far from ideal for reliable numerical modeling. It is important to note that the nonlinear term is restricted to the interior computational domain, while the exterior region is assumed to be linear. Under this assumption, the ABCs derived through the Laplace transform remain mathematically valid and physically consistent. In this study, taking into account that the solution $\xi(\chi,\rho)$ vanishes as $|\chi|\to\infty$, a direct relationship between its spatial derivative and the fractional time derivative of order $(\frac{\alpha}{2})$ is derived. This relation serves as the exact ABC, ensuring that the outgoing energy and information approaching the boundaries are naturally absorbed instead of being reflected back into the computational domain. Consequently, artificial reflections are effectively eliminated, leading to a substantial improvement in the accuracy and stability of the numerical solution. In this study, our focus is on the fractional-time diffusion equation in an infinite spatial domain, as shown in (1). First, we consider two artificial boundary points for the left and right sides of the domain as follows:

$$\begin{cases} \Lambda_- := \{\chi|\chi = I_-\}, \\ \Lambda_+ := \{\chi|\chi = I_+\}. \end{cases}$$

Then, we divide the entire domain $\mathbb{R}$ into three sections:

$$\begin{cases} \Delta_- := \{\chi \mid -\infty < \chi < I_-\}, \\ \Delta_i := \{\chi \mid I_- \le \chi \le I_+\}, \\ \Delta_+ := \{\chi \mid I_+ < \chi < +\infty\}. \end{cases} \tag{4}$$

The exterior of the domain is denoted by $\Delta_e = \Delta_- \cup \Delta_+$. The constants I_- and I_+ are chosen such that the initial function $\xi_0(\chi)$ takes nonzero values only within the interior region of the domain Δ_i and is exactly zero outside of it. The form of the operator equation (1) in Δ_e is written as follows:

$$ {}_0^C D_\rho^\alpha \xi(\chi,\rho) = \mathcal{L}\xi(\chi,\rho) + \mathcal{N}\xi(\chi,\rho), \quad \chi \in \Delta_e, \tag{5}$$

where

$$\mathcal{L}\xi = \xi_{\chi\chi}, \qquad \mathcal{N}\xi = \mathcal{F}(\xi) + \mathcal{G}(\chi,\rho).$$

Considering the linear part of (5), we have

$$ {}_0^C D_\rho^\alpha \xi(\chi,\rho) = \xi_{\chi\chi}(\chi,\rho). \tag{6}$$

The Laplace transform of the Caputo fractional derivative for $0<\alpha<1$ is given by [18]:

$$\begin{aligned}\mathcal{L}[{}_0^C D_\rho^\alpha \xi(\chi,\rho)](\vartheta) =& \vartheta^\alpha \mathcal{L}[\xi(\chi,\rho)](\vartheta) - \vartheta^{\alpha-1}\xi(\chi,0)\\ =& \vartheta^\alpha \tilde{\xi}(\chi,\vartheta) - \vartheta^{\alpha-1}\xi(\chi,0).\end{aligned} \tag{7}$$

Thus, we first restrict the problem (1) to the domain Δ_e, where $\mathcal{F}(\xi(\chi,\rho)) \approx 0$ and $\mathcal{G}(\chi,\rho) \approx 0$. Now, considering the initial condition $\xi_0(\chi) = 0$ for $\chi \in \Delta_e$, (6) can be rewritten in the Laplace space as follows:

$$\vartheta^\alpha \tilde{\xi}(\chi,\vartheta) = \tilde{\xi}_{\chi\chi}(\chi,\vartheta), \quad \mathrm{Re}(\vartheta) > 0. \tag{8}$$

Equation (8) has a general solution given as follows:

$$\tilde{\xi}(\chi,\vartheta) = \mathcal{A}_1(\vartheta)e^{-\sqrt{\vartheta^\alpha}\chi} + \mathcal{A}_2(\vartheta)e^{\sqrt{\vartheta^\alpha}\chi}, \tag{9}$$

where $\mathcal{A}_1(\vartheta)$ and $\mathcal{A}_2(\vartheta)$are unknown parameters. Given that $\xi(\chi,\vartheta) \to 0$ as $|\chi| \to \infty$, we have

$$\tilde{\xi}(\chi,\vartheta) = \begin{cases} \mathcal{A}_1(\vartheta)e^{-\sqrt{\vartheta^\alpha}\chi}, & \chi \in \Delta_+,\\ \mathcal{A}_2(\vartheta)e^{\sqrt{\vartheta^\alpha}\chi}, & \chi \in \Delta_-. \end{cases} \tag{10}$$

Differentiating $\tilde{\xi}(\chi,\vartheta)$ with respect to χ, yields

$$\frac{\partial\tilde{\xi}(\chi,\vartheta)}{\partial\chi} = \begin{cases} -\sqrt{\vartheta^\alpha}\,\tilde{\xi}(\chi,\vartheta) = -\vartheta^{\frac{\alpha}{2}}\tilde{\xi}(\chi,\vartheta), & \chi \in \Delta_+,\\ \sqrt{\vartheta^\alpha}\,\tilde{\xi}(\chi,\vartheta) = \vartheta^{\frac{\alpha}{2}}\tilde{\xi}(\chi,\vartheta), & \chi \in \Delta_-. \end{cases} \tag{11}$$

Now, applying the inverse Laplace transform to (11), implies

$$\frac{\partial\xi(\chi,\rho)}{\partial\chi} = \begin{cases} -\frac{1}{\Gamma(1-\frac{\alpha}{2})}\int_0^\rho (\rho-\vartheta)^{-\frac{\alpha}{2}}\frac{\partial\xi(\chi,\vartheta)}{\partial\vartheta}\,d\vartheta = -{}_0^C D_\rho^{\frac{\alpha}{2}}\xi(\chi,\rho), & \chi \in \Delta_+,\\ \frac{1}{\Gamma(1-\frac{\alpha}{2})}\int_0^\rho (\rho-\vartheta)^{-\frac{\alpha}{2}}\frac{\partial\xi(\chi,\vartheta)}{\partial\vartheta}\,d\vartheta = {}_0^C D_\rho^{\frac{\alpha}{2}}\xi(\chi,\rho), & \chi \in \Delta_-. \end{cases} \tag{12}$$

By taking the limit as $\chi \to I_+$ and $\chi \to I_-$, and assuming the smoothness of $\xi(\chi,\rho)$ at $\chi = I_-$ and $\chi = I_+$, we obtain the following relation:

$$\begin{cases} \frac{\partial\xi(I_+,\rho)}{\partial\chi} = -{}_0^C D_\rho^{\frac{\alpha}{2}}\xi(I_+,\rho),\\ \frac{\partial\xi(I_-,\rho)}{\partial\chi} = {}_0^C D_\rho^{\frac{\alpha}{2}}\xi(I_-,\rho). \end{cases} \tag{13}$$

In this work, the interrelation between the Caputo and Riemann–Liouville fractional derivatives is implicitly exploited in the analytical formulation of the ABCs. Specifically, since the function $\xi(\chi,\rho)$ is defined such that all its integer-order time derivatives vanish at the initial point ($\rho = 0$), the Caputo and Riemann–Liouville derivatives become equivalent in this setting. Under this

assumption, the composition property of the Riemann–Liouville derivative can be consistently extended to the Caputo derivative. Consequently, the following relation holds and is analytically justified based on the fundamental properties of fractional calculus:

$$ {}_0^C D_\rho^{\frac{\alpha}{2}} \left({}_0^C D_\rho^{\frac{\alpha}{2}} \, \xi(\chi,\rho) \right) = {}_0^C D_\rho^{\alpha} \, \xi(\chi,\rho). $$

Now, applying the fractional derivative ${}_0^C D_\rho^{\frac{\alpha}{2}}$ to relation (13), gives

$$ \begin{cases} {}_0^C D_\rho^{\alpha} \xi(\chi,\rho) = -{}_0^C D_\rho^{\frac{\alpha}{2}} \xi_\chi(\chi,\rho) & \chi = I_+, \\ {}_0^C D_\rho^{\alpha} \xi(\chi,\rho) = {}_0^C D_\rho^{\frac{\alpha}{2}} \xi_\chi(\chi,\rho) & \chi = I_-. \end{cases} \tag{14} $$

At this stage, the exact artificial boundary conditions (14) for the problem (1) have been obtained. Considering (14), we observe that these boundary conditions involve fractional time derivatives of order $\frac{\alpha}{2}$, which is half of the time derivative order in the governing equation (1). Thus, the original problem (1) is reduced to an IVP problem in the bounded domain Δ_i and can be written as follows:

$$ \begin{aligned} &{}_0^C D_\rho^{\alpha} \xi(\chi,\rho) = \xi_{\chi\chi}(\chi,\rho) + \mathcal{F}(\xi(\chi,\rho)) + \mathcal{G}(\chi,\rho), \quad (\chi,\rho) \in \Delta_i \times (0,P], \\ &{}_0^C D_\rho^{\frac{\alpha}{2}} \frac{\partial \xi(I_-,\rho)}{\partial \chi} = \xi_{\chi\chi}(I_-,\rho) + \mathcal{F}(\xi(I_-,\rho)) + \mathcal{G}(I_-,\rho), \quad \rho \in (0,P], \\ &{}_0^C D_\rho^{\frac{\alpha}{2}} \frac{\partial \xi(I_+,\rho)}{\partial \chi} = -(\xi_{\chi\chi}(I_+,\rho) + \mathcal{F}(\xi(I_+,\rho)) + \mathcal{G}(I_+,\rho)), \quad \rho \in (0,P], \\ &u(\chi,0) = \varphi(\chi), \qquad \chi \in \Delta_i = [I_-, I_+], \\ &u(0,\rho) = \phi(\rho), \qquad \rho \in (0,P]. \end{aligned} \tag{15} $$

Now, we transformed PDE with unbounded domain (1) into an equivalent bounded domain formulation by developing ABCs (15). This reformulation enables the solution of the modified problem, governed by (15), using an appropriate numerical methodology. In (15), the functions $\varphi(\chi)$ and $\phi(\rho)$ are known functions determined by the given source term and initial data of the original problem (1), and they arise naturally from the reformulation of the nonhomogeneous term after imposing the ABCs.

3 Numerical technique

This section presents a novel numerical methodology for solving the boundary-value problem governed by (15). It is organized into three main components:

(1) the derivation of a high-order approximation scheme for fractional derivatives,
(2) the development of an efficient high-order finite difference framework based on this approximation, and
(3) the implementation of the proposed numerical scheme to solve the considered problem.

Compared with classical time discretization methods for Caputo fractional derivatives, such as the $L1$ scheme with accuracy $\mathcal{O}(\varrho^{2-\alpha})$, the proposed high-order finite difference approximation achieves a temporal accuracy of $\mathcal{O}(\varrho^{3-\alpha})$. This improvement significantly reduces the numerical error for long-time simulations and allows the use of relatively larger time step sizes without sacrificing accuracy. Moreover, the higher-order approximation enhances the stability properties of the numerical scheme, particularly in the presence of ABCs involving fractional derivatives of order $\alpha/2$. As a result, the proposed method provides a more accurate and efficient computational framework compared with standard low-order schemes, especially for nonlinear fractional diffusion problems posed on unbounded spatial domains.

3.1 Approximation of fractional derivative

This part establishes high-order numerical approximations for the Caputo fractional derivative. The accuracy of this method depends on the parameter $\alpha \in (0,1)$. First, we introduce a $(3-\alpha)$-order approximation, which was proposed in [13]. For simplification, in this subsection, we consider one variable function $\xi(\rho)$, and suppose that $\xi(\rho) \in C^6[0,P]$, that $\xi^{(5)}(\rho)$ is continuous, and that $\xi^{(6)}(\rho)$ exists in the interval $[0,P]$. Additionally, $0 = \rho_0 < \rho_1 < \ldots < \rho_N = P$ is given for a positive integer $N > 0$, where $\rho_k = k\varrho$, $k = 0,1,\ldots,N$, and $\varrho = \frac{P}{N}$ is the time step size. First, we write the Taylor expansion of the function $\xi'(\vartheta)$ around the point $\rho = \rho_j$, where $0 \le j < k$. We have

$$\begin{aligned}\xi^{'}(\vartheta) =& \xi^{'}(\rho_j) + \xi^{''}(\rho_j)(\vartheta-\rho_j) + \frac{\xi^{'''}(\rho_j)}{2!}(\vartheta-\rho_j)^2 \\ &+ O((\vartheta-\rho_j)^3), \qquad \vartheta \in (\rho_j, \rho_{j+1}).\end{aligned} \tag{16}$$

This expansion shows that the first derivative of the function at the point ϑ is computed as a linear combination of higher-order derivatives at the point ρ_j. Now, consider the following relations for the first and second derivatives:

$$\xi^{'}(\rho_j) = \frac{\xi(\rho_{j+1}) - \xi(\rho_{j-1})}{2\varrho} - \frac{\xi^3(\rho_j)}{3!}\varrho^2 + O(\varrho^4), \tag{17}$$

and

$$\xi^{''}(\rho_j) = \frac{\xi(\rho_{j+1}) - 2\xi(\rho_j) + \xi(\rho_{j-1})}{\varrho^2} - \frac{\xi^4(\rho_j)}{12}\varrho^2 + O(\varrho^4). \tag{18}$$

By substituting (17) and (18) into (16), we arrive at the following relation:

$$\begin{aligned}\xi^{'}(\vartheta) =& \frac{\xi(\rho_{j+1}) - \xi(\rho_{j-1})}{2\varrho} + \frac{\xi(\rho_{j+1}) - 2\xi(\rho_j) + \xi(\rho_{j-1})}{\varrho^2}(\vartheta - \rho_j) - \frac{\xi^3(\rho_j)}{3!}\varrho^2 \\ &+ \frac{\xi^3(\rho_j)}{2!}(\vartheta - \rho_j)^2 + O((\vartheta - \rho_j)^3), \quad 0 < \vartheta - \rho_j < \varrho. \end{aligned} \tag{19}$$

This approximate relation defines the derivative $\xi'(\rho)$ by considering the integrals corresponding to the third-order derivatives over the specified interval. By applying relation (19), the Caputo derivative approximation with the condition $(0 < \alpha < 1)$ is written as follows:

$$\begin{aligned}{}_0^C D_\rho^\alpha \xi(\rho)|_{\rho=\rho_k} =& \frac{1}{\Gamma(1-\alpha)} \int_0^{\rho_k} (\rho_k - \vartheta)^{-\alpha} \xi^{'}(\vartheta) d\vartheta \\ =& \frac{1}{\Gamma(1-\alpha)} \sum_{j=0}^{k-1} \int_{\rho_j}^{\rho_{j+1}} (\rho_k - \vartheta)^{-\alpha} \xi^{'}(\vartheta) d\vartheta \\ =& \frac{1}{\Gamma(1-\alpha)} \sum_{j=0}^{k-1} \int_{\rho_j}^{\rho_{j+1}} (\rho_k - \vartheta)^{-\alpha} \Big(\frac{\xi(\rho_{j+1}) - \xi(\rho_{j-1})}{2\varrho} \\ &+ \frac{\xi(\rho_{j+1}) - 2\xi(\rho_j) + \xi(\rho_{j-1})}{\varrho^2}(\vartheta - \rho_j) - \frac{\xi^3(\rho_j)}{3!}\varrho^2 \\ &+ \frac{\xi^3(\rho_j)}{2!}(\vartheta - \rho_j)^2 \Big) d\vartheta + O((\vartheta - \rho_j)^3) \\ =& \frac{\varrho^{-\alpha}}{\Gamma(3-\alpha)} \sum_{j=0}^{k-1} (\omega_{1,k-j}(\xi^{j+1} - \xi^{j-1}) \\ &+ \omega_{2,k-j}(\xi^{j+1} - 2\xi^j + \xi^{j-1})) + \mathcal{R}^k, \end{aligned} \tag{20}$$

where $\mathcal{R}^k \leq C\varrho^{3-\alpha}$ [13]. Relation (20) represents the Caputo fractional-time derivative applied to solve the problem in (1). Following the approximation framework established in [3, Lemma 2.1], weighting coefficients defined in (20) are derived as follows:

$$\begin{cases} \omega_{1,n} = \frac{2-\alpha}{2} \left((n)^{1-\alpha} - (n-1)^{1-\alpha} \right), \\ \omega_{2,n} = (n)^{2-\alpha} - (n-1)^{2-\alpha} - (2-\alpha)(n-1)^{1-\alpha}. \end{cases} \tag{21}$$

Therefore, fractional-time derivative of relation (20) is written as follows:

$$ {}_0^C D_\rho^\alpha \xi(\rho)|_{\rho=\rho_k} = \frac{\varrho^{-\alpha}}{\Gamma(3-\alpha)} \sum_{j=0}^{k-1} (\omega_{1,k-j}(\xi^{j+1} - \xi^{j-1}) + \omega_{2,k-j}(\xi^{j+1} - 2\xi^j + \xi^{j-1})) + O(\varrho^{3-\alpha}). \quad (22) $$

In formula (22), if $j = 0$, then the value $\xi^{j-1} = \xi^{-1}$ is defined outside the interval $[0, P]$. Here, several choices are presented for approximating ξ^{-1} . In numerical computations, values of the neighboring function are mainly used to approximate ξ^{-1} such that

$$ \xi^{-1} = \xi(0) - \varrho\xi'(0) + \frac{\varrho^2}{2}\xi''(0) + O(\varrho^3). $$

1. When $\xi'(0) = 0$ and $\xi''(0) = 0$, we have $\xi^{-1} = \xi^0 + O(\varrho^3)$, where the convergence order of formula (22) is $O(\varrho^{3-\alpha})$.

2. When $\xi'(0) = 0$ and $\xi''(0) \neq 0$, we have $\xi^{-1} = \xi^0 + \frac{\varrho^2}{2}\xi''(0) + O(\varrho^3)$. In this case, the convergence order of formula (22) is $O(\varrho^2)$.

3. When $\xi'(0) \neq 0$, the convergence order is $O(\varrho)$.

Lemma 1. [7] The coefficients $\omega_{1,k-j}$ and $\omega_{2,k-j}$ in relation (22) have the following properties:

(I) Initial values of the coefficients:

$$ \omega_{1,1} = \frac{2-\alpha}{2}, \quad \omega_{2,1} = 1. $$

Assume that $0 < \alpha < 1$.

(II) Constraints and behavior of the coefficients:

$$ \begin{cases} 0 < \omega_{1,r-j+1} < \omega_{1,r-j} \leq \frac{2-\alpha}{2} < 1, \\ 0 < \omega_{2,r-j+1} < \omega_{2,r-j} \leq 1. \end{cases} $$

(III) Difference between adjacent coefficients:

$$ \begin{cases} \omega_{1,r-j+1} - \omega_{1,r-j} > \omega_{1,r-j} - \omega_{1;r-j-1}, & r-j \geq 2, \\ \omega_{2,r-j+1} - \omega_{2,r-j} > \omega_{2,r-j} - \omega_{2,r-j-1}, & r-j \geq 2. \end{cases} $$

(IV) Sum of coefficients:

$$\omega_{1,2}+\omega_{1,1}+\omega_{2,2}-\omega_{2,1}\begin{cases}>0, & \alpha\in(0,\alpha_0),\\ \leq 0, & \alpha\in[\alpha_0,1),\end{cases}$$

which $\alpha_0\approx 0.68$ is the root of the equation

$$(6-\alpha)2^{-\alpha}+\alpha-4=0,\quad \alpha\in(0,1),$$

and also,

$$\omega_{1,r-j}+\omega_{1,r-j-1}+\omega_{2,r-j}-\omega_{2,r-j-1}>0,\quad r-j\geq 3.$$

(V) Special combinations of the coefficients:

$$-\omega_{1,2}+2\omega_{2,1}-\omega_{2,2}>0,$$

$$\omega_{1,3}-\omega_{1,1}+\omega_{2,3}-2\omega_{2,2}+\omega_{2,1}\begin{cases}>0, & \alpha\in(0,\alpha_1),\\ \leq 0, & \alpha\in[\alpha_1,1),\end{cases}$$

that $\alpha_1\approx 0.37$ is the root of the equation

$$-2^{3-\alpha}+3\cdot 2^{2-\alpha}-2^{1-\alpha}+2^{-\alpha}-\frac{3(2-\alpha)^2}{2^{1-\alpha}}-3\alpha^2+6=0,\quad \alpha\in(0,1),$$

and also,

$$\omega_{1,r-j+1}-\omega_{1,r-j-1}+\omega_{2,r-j+1}-2\omega_{2,r-j}+\omega_{2,r-j-1}<0,\quad r-j\geq 3.$$

In this work, a high-order finite difference approximation with temporal accuracy of $O(3-\alpha)$ is adopted for discretizing the fractional-time derivative, as presented in (22). This scheme achieves higher accuracy than the classical $L1$ method, whose convergence order is $O(2-\alpha)$. The enhanced accuracy enables the method to produce reliable results even for relatively large time step sizes. Moreover, since the ABCs are analytically and exactly formulated, numerical disturbances due to wave reflections at the boundaries are effectively eliminated, which plays a crucial role in maintaining the stability of the proposed numerical scheme (28).

3.2 Construction of the finite difference method

Assume that $\varrho = \frac{P}{N}$ and $h = \frac{I_+ - I_-}{M}$ are the time and spatial step sizes, respectively, where M and N are given positive integers. The spacing between the grid points in the temporal and spatial dimensions is considered as follows:

$$\begin{cases} \chi_i = ih, & 0 \le i \le M, \\ \rho_k = k\varrho, & 0 \le k \le N, \end{cases}$$

and the computational domain is partitioned into two structured grids:

$$\Pi_\varrho = \{\rho_k \mid 0 = \rho_0 < \rho_1 < \cdots < \rho_N = P\},$$

$$\Pi_h = \{\chi_i \mid \chi_i = I_- + ih,\ 0 \le i \le M\}.$$

Assume that ξ_i^k is the numerical approximation of $\xi(\chi_i, \rho_k)$, we define the space of grid functions defined on $\Pi_\tau \times \Pi_h$ as follows:

$$\mathcal{V} = \{\xi_i^k \mid 0 \le i \le M,\ 0 \le k \le N\}.$$

To facilitate the subsequent numerical analysis, the following systematic notation is formally defined:

$$\begin{cases} \sigma_\chi \xi_{i-\frac{1}{2}}^k = \frac{\xi_i^k - \xi_{i-1}^k}{h}, \quad \sigma_\chi^2 \xi_i^k = \frac{\sigma_\chi \xi_{i+\frac{1}{2}}^k - \sigma_\chi \xi_{i-\frac{1}{2}}^k}{h}, \quad \mathcal{F}_v = \frac{\partial \mathcal{F}}{\partial \xi}\Big|_{\xi=v}, \\ \xi_{i+\frac{1}{2}}^k = \frac{1}{2}(\xi_{i+1}^k + \xi_i^k), \quad \sigma_\chi \xi_{i+\frac{1}{2}}^k = \frac{1}{h}(\xi_{i+1}^k - \xi_i^k). \end{cases}$$

For each $\xi, v \in \Pi_h$, the inner product and norm are expressed as

$$(\xi, v) = h \sum_{i=1}^{M} (\xi_{i-\frac{1}{2}} v_{i-\frac{1}{2}}),$$

$$\|\xi\| = \sqrt{(\xi, \xi)}, \quad \|\xi\|_1 = \sqrt{h \sum_{i=1}^{M} (\sigma_\chi \xi_{i-\frac{1}{2}})^2}, \quad \|\xi\|_\infty = \max_{0 \le i \le M} |\xi_i|.$$

The fractional-time derivative of relation (22) is rewritten as follows:

$$\begin{aligned} D_\varrho^\alpha \xi_i^k =& \mu_1 \sum_{j=0}^{k-1} (\omega_{1,k-i}(\xi_i^{j+1} - \xi_i^{j-1}) + \omega_{2,k-i}(\xi_i^{j+1} - 2\xi_i^j + \xi_i^{j-1})) \\ &+ O(\varrho^{3-\alpha}), \end{aligned} \tag{23}$$

where $\mu_1 = \frac{\varrho^{-\alpha}}{\Gamma(3-\alpha)}$. To address the computational complexity associated with the fractional differentiation of the nonlinear term $\mathcal{F}(\xi_i^k)$, the composition property of the Riemann–Liouville derivative together with its equivalence to the Caputo form is employed. This analytical manipulation enables the fractional derivative of order $(\frac{\alpha}{2})$ to act solely on $\mathcal{F}(\xi_i^k)$, while its influence is consistently controlled through the exact ABCs derived earlier. As a result, the nonlinear contribution is incorporated in a mathematically consistent and numerically stable manner within the proposed scheme.

In this paper, only one continuous mathematical model is investigated, namely the nonhomogeneous time-fractional nonlinear parabolic equation posed on an unbounded spatial domain, as introduced in (1). The formulation presented in (15) does not represent a new model; rather, it is an equivalent reformulation obtained by imposing exact ABCs on the original unbounded domain problem. The introduction of the auxiliary variable $v(\chi,\rho) = \xi_\chi(\chi,\rho)$ is carried out solely to facilitate the numerical discretization and the implementation of the ABCs involving fractional derivatives of order $\alpha/2$. This auxiliary reformulation does not alter the governing equation nor introduce additional modeling assumptions. Consequently, the continuous problems defined by (1) and (15) are mathematically equivalent, and the numerical scheme developed in section 3 is derived directly from this equivalent formulation.

3.3 Applying numerical method to the main problem

In this section, we address the numerical solution of problem (15). First, we consider the following change of variable:

$$\begin{cases} v(\chi,\rho) = \xi_\chi(\chi,\rho), \\ v_\chi(\chi,\rho) = \xi_{\chi\chi}(\chi,\rho), \end{cases}$$

where function $v(\chi,\rho)$ is used as an intermediate function to maintain the continuity of the derivatives of $\xi(\chi,\rho)$ during the discretization. In this case, the problem (15) can be rewritten as the following equation:

$$\begin{aligned}
&{}_0^C D_\rho^\alpha \xi(\chi,\rho) = v_\chi(\chi,\rho) + \mathcal{F}(\xi(\chi,\rho)) + \mathcal{G}(\chi,\rho), \quad (\chi,\rho) \in \Delta_i \times (0,P], \\
&v(\chi,\rho) = \xi_\chi(\chi,\rho), \quad \chi \in \Delta_i, \\
&{}_0^C D_\rho^{\frac{\alpha}{2}} v(\chi,\rho) = v_\chi(\chi,\rho) + \mathcal{F}(\xi(\chi,\rho)) + \mathcal{G}(\chi,\rho), \quad \chi \in \Delta_-, \\
&v(\chi,\rho) = \xi_\chi(\chi,\rho), \quad \chi \in \Delta_-, \\
&{}_0^C D_\rho^{\frac{\alpha}{2}} v(\chi,\rho) = -(v_\chi(\chi,\rho) + \mathcal{F}(\xi(\chi,\rho)) + \mathcal{G}(\chi,\rho)), \quad \chi \in \Delta_+,
\end{aligned} \tag{24}$$

$$v(\chi,\rho)=\xi_{\chi}(\chi,\rho),\quad \chi\in\Delta_{+}$$
$$\xi(\chi,0)=\varphi(\chi),\quad \chi\in\Delta_{i},$$
$$\xi(0,\rho)=\phi(\rho),\quad \rho>0.$$

We observe that problem (1) is equivalent to problem (24), as their solutions coincide. Thus, we formulate the following lemma to establish the uniqueness of the solution to problem (24).

Remark 1. (a) In the numerical implementation, the nonlinear term $\mathcal{F}(\xi_i^k)$ is replaced by its previous-time value $\mathcal{F}(\xi_i^{k-1})$. This semi-implicit (time-lagging) treatment linearizes the discrete system at each time-step and significantly reduces the computational cost. (b) To analyze the effect of this replacement, we define $T_i^k=\mathcal{F}(\xi_i^k)-\mathcal{F}(\xi_i^{k-1})$. Under the Lipschitz continuity condition $|\mathcal{F}(\xi_i^k)-\mathcal{F}(\xi_i^{k-1})|\le L|\xi_i^k-\xi_i^{k-1}|$, it follows that $|T_i^k|\le L|\xi_i^k-\xi_i^{k-1}|$. Hence the lagging error is of higher order and does not affect the overall accuracy, stability, or convergence of the proposed scheme.

Lemma 2. Assume $\xi(\chi,\rho)\in C^{4,6}([I_-,I_+]\times[0,P])$, and let the temporal grid be $\rho_k=k\varrho$. Suppose the ghost value is chosen so that $\xi^{-1}=\xi^0+\mathcal{O}(\varrho^3)$ this holds if $\partial_\rho\xi(\chi,0)=\partial_\rho^2\xi(\chi,0)=0$. Then the local truncation error $\mathcal{R}_i^k$ of the scheme (24) satisfies

$$\mathcal{R}_i^k=\mathcal{O}(\varrho^{3-\alpha}+h^2),$$

uniformly for $0\le i\le M$ and $1\le k\le N$.

Proof. The smoothness assumption $\xi\in C_\chi^4C_\rho^6([I_-,I_+]\times[0,P])$ ensures that the remainder terms in the temporal Taylor expansions are uniformly bounded with respect to both χ and ρ. Moreover, since the temporal grid is uniform ($\rho_k=k\varrho$), the mean value theorem yields

$$\begin{aligned}\partial_\rho\xi(\chi_i,\rho_j)&=\partial_\rho\xi(\chi_i,\rho_k)+O(|\rho_j-\rho_k|)\\&=\partial_\rho\xi(\chi_i,\rho_k)+O(\varrho),\end{aligned}$$

and a similar relation holds for $\partial_\rho^2\xi$. Hence, replacing ρ_j by ρ_k inside the discrete fractional sums introduces an additional term of order $O(\varrho^{3-\alpha})$, which is absorbed into the overall truncation error. The choice $\xi_{-1}=\xi_0+O(\varrho^3)$, corresponding to the conditions $\xi_\rho(\chi,0)=\xi_{\rho\rho}(\chi,0)=0$, eliminates any lower–order boundary contributions. We then proceed with the Taylor expansions for fixed i,k and each $j=0,\dots,k-1$.

$$\begin{cases} \xi_i^{j+1} - \xi_i^{j-1} = 2\varrho\,\partial_\rho \xi(\chi_i,\rho_j) + \dfrac{2\varrho^3}{6}\partial_\rho^3 \xi(\chi_i,\rho_j) + \mathcal{O}(\varrho^5), \\ \xi_i^{j+1} - 2\xi_i^{j} + \xi_i^{j-1} = \varrho^2\,\partial_\rho^2 \xi(\chi_i,\rho_j) + \mathcal{O}(\varrho^4), \\ \xi_{i+1}^{k} - 2\xi_i^{k} + \xi_{i-1}^{k} = h^2\,\partial_\chi^2 \xi(\chi_i,\rho_k) + \mathcal{O}(h^4). \end{cases} \tag{25}$$

Recall the discrete Caputo approximation used in the scheme (22):

$$\begin{aligned} {}_0^C D_\rho^\alpha \xi(\rho)|_{\rho=\rho_k} =& \frac{\varrho^{-\alpha}}{\Gamma(3-\alpha)} \sum_{j=0}^{k-1} (\omega_{1,k-j}(\xi^{j+1} - \xi^{j-1}) \\ &+ \omega_{2,k-j}(\xi^{j+1} - 2\xi^j + \xi^{j-1})) + O(\varrho^{3-\alpha}), \end{aligned}$$

where the remainder $\mathcal{O}(\varrho^{3-\alpha})$ originates from integrating the higher-order terms of the Taylor expansions against the kernel $(\rho_k - \vartheta)^{-\alpha}$ on each temporal subinterval; under the stated smoothness the constant in this bound is uniform in k (see [7, Lemma 2.1]). Substitute the expansions (25) into the discrete Caputo sum to obtain

$$\begin{aligned} &\frac{\varrho^{-\alpha}}{\Gamma(3-\alpha)} \sum_{j=0}^{k-1} \Big(\omega_{1,k-j}(\xi_i^{j+1} - \xi_i^{j-1}) + \omega_{2,k-j}(\xi_i^{j+1} - 2\xi_i^j + \xi_i^{j-1})\Big) \\ &= \frac{\varrho^{-\alpha}}{\Gamma(3-\alpha)} \sum_{j=0}^{k-1} \Big(2\varrho\,\omega_{1,k-j}\,\partial_\rho \xi(\chi_i,\rho_j) + \varrho^2 \omega_{2,k-j}\,\partial_\rho^2 \xi(\chi_i,\rho_j)\Big) + \mathcal{O}(\varrho^{3-\alpha}). \end{aligned}$$

By the regularity of $\partial_\rho \xi$ and $\partial_\rho^2 \xi$ on $[0,\rho_k]$ and using the boundedness and positivity properties of the weights $\{\omega_{1,m}\}, \{\omega_{2,m}\}$ 1, we may replace the temporal arguments ρ_j by ρ_k up to a temporal interpolation error of order $\mathcal{O}(\varrho)$ inside the sums. Consequently,

$$\frac{\varrho^{-\alpha}}{\Gamma(3-\alpha)} \Big(2\varrho\partial_\rho \xi(\chi_i,\rho_k) \sum_{j=0}^{k-1} \omega_{1,k-j} + \varrho^2 \partial_\rho^2 \xi(\chi_i,\rho_k) \sum_{j=0}^{k-1} \omega_{2,k-j}\Big) + \mathcal{O}(\varrho^{3-\alpha}).$$

A direct comparison with the continuous Caputo derivative shows that the leading discrete terms reproduce the continuous operator up to the $\mathcal{O}(\varrho^{3-\alpha})$ remainder (this is the standard consistency calculation for the chosen weights; the details follow from the definition of $\omega_{1,n}, \omega_{2,n}$ and standard quadrature estimates). Together with the spatial discretization error $\mathcal{O}(h^2)$ from (25), we obtain a bound for the local truncation error of the fully discrete scheme:

$$|\mathcal{R}_i^k| \le C_1\big(\varrho^{3-\alpha} + h^2\big),$$

for some constant $C_1 > 0$ independent of ϱ, h, uniformly in i, k. Finally, the hypothesis $\xi_{-1} = \xi_0 + \mathcal{O}(\varrho^3)$ guarantees that no lower-order boundary-term arising from the ghost value degrades the stated order. Therefore $\mathcal{R}_i^k = \mathcal{O}(\varrho^{3-\alpha} + h^2)$, which proves the lemma. □

Assume that $\xi(\chi, \rho) \in C^{4,6}_{\chi,\rho}(\mathbb{R})$, and define the discrete values $\xi_i^k = \xi(\chi_i, \rho_k)$, $V_i^k = v(\chi_i, \rho_k)$ and source term $g_i^k = \mathcal{G}(\chi_i, \rho_k)$, function $f(\xi(\chi_i, \rho_{k-1}))$ is defined as $\mathcal{F}(\xi(\chi_i, \rho_{k-1}))$. By employing the fractional derivative approximation formula given in relation (22) and using Lemma 2, together with Taylor expansion, we obtain

$$
\begin{aligned}
&{}_0^C D_\rho^\alpha \xi_{i+\frac{1}{2}}^k = \sigma_\chi V_{i+\frac{1}{2}}^k + \mathcal{F}(\xi_{i+\frac{1}{2}}^{k-1}) + \mathcal{G}_{i+\frac{1}{2}}^k + R_{i+\frac{1}{2}}^k,\\
&V_{i+\frac{1}{2}}^k = \sigma_\chi \xi_{i+\frac{1}{2}}^k + Z_{i+\frac{1}{2}}^k,\\
&{}_0^C D^{\frac{\alpha}{2}} V_0^k = \left(\sigma_\chi V_0^k + \mathcal{F}(\xi_0^{k-1}) + \mathcal{G}_0^k\right) + T^k,\\
&V_0^k = \sigma_\chi \xi_0^k + A^k,\\
&{}_0^C D^{\frac{\alpha}{2}} V_M^k = -\left(\sigma_\chi V_M^k + \mathcal{F}(\xi_M^{k-1}) + \mathcal{G}_M^k\right) + W^k,\\
&V_M^k = \sigma_\chi \xi_M^k + Q^k,\\
&\xi_i^0 = \varphi(\chi_i), \qquad \chi \in \Delta_i,\\
&\xi_0^k = \phi(\rho_k), \qquad \rho > 0,
\end{aligned}
\tag{26}
$$

where the remainder functions are denoted by R, Z, T, A, W, and Q. There exists a positive constant C_0 such that [6]

$$
\begin{aligned}
&|R_{i+\frac{1}{2}}^k| \le C_0\left(\varrho^{3-\alpha} + h^2\right), \quad |Z_{i+\frac{1}{2}}^k| \le C_0 h^2, \quad |T^k| \le C_0(\varrho^{3-\frac{\alpha}{2}} + h^2),\\
&|A^k| \le C_0 h^2, \qquad |W^k| \le C_0(\varrho^{3-\frac{\alpha}{2}} + h^2), \quad |Q^k| \le C_0 h^2,
\end{aligned}
\tag{27}
$$

By assuming that $\xi(\chi, \rho) \in C^{4,6}_{\chi,\rho}(\mathbb{R})$ and applying the fractional derivative approximation in (22) together with Lemma 2 and the Taylor expansion, the discrete formulation (26) is obtained. The remainder terms, denoted by R, Z, T, A, W, and Q, represent higher–order truncation errors, whose magnitudes are bounded as in (27). Since these terms are of order $\mathcal{O}(\varrho^{3-\alpha} + h^2)$ or smaller, their effects can be neglected in the implementation stage $R = Z = T = A = W = Q = 0$. Consequently, the simplified and implementable finite difference scheme with accuracy $\mathcal{O}(\varrho^{3-\alpha} + h^2)$ and guaranteed stability is obtained, as presented in (28).

$$
\begin{aligned}
&{}_0^C D_\rho^\alpha \xi_{i+\frac{1}{2}}^k = \sigma_\chi v_{i+\frac{1}{2}}^k + f(\xi_{i+\frac{1}{2}}^{k-1}) + g_{i+\frac{1}{2}}^k,\\
&v_{i+\frac{1}{2}}^k = \sigma_\chi \xi_{i+\frac{1}{2}}^k,
\end{aligned}
$$

$$
\begin{aligned}
&{}_0^C D_\rho^{\frac{\alpha}{2}} v_0^k = \left(\sigma_\chi v_0^k + f(\xi_0^{k-1}) + g_0^k\right), \\
&v_0^k = \sigma_\chi \xi_0^k, \\
&{}_0^C D_\rho^{\frac{\alpha}{2}} v_M^k = -\left(\sigma_\chi v_M^k + f(\xi_M^{k-1}) + g_M^k\right), \\
&v_M^k = \sigma_\chi \xi_M^k, \\
&\xi_i^0 = \varphi(\chi_i), \qquad \chi \in \Delta_i, \\
&\xi_0^k = \phi(\rho_k), \qquad \rho > 0.
\end{aligned} \tag{28}
$$

For the interior grid points $1 \leqslant i \leqslant M-1$, the classical second-order central difference approximation is applied for the spatial derivative, which provides an accuracy of order $\mathcal{O}(h^2)$. However, at the boundary points χ_0 and χ_M, the external function values ξ_{-1} and ξ_{M+1} are not available, and thus the standard central difference formula cannot be directly employed. To preserve the second-order spatial accuracy and maintain the consistency of the numerical solution, Lemma 3 constructs suitable boundary approximations by expressing the spatial derivatives at the boundaries in terms of the known nodal values and the exact ABCs derived in Section 2. These boundary relations are incorporated into Theorem 1, where the complete discrete system of equations is established by combining the interior scheme (28) with the boundary approximations from Lemma 3. Consequently, Theorem 3.4 ensures that the number of algebraic equations equals the number of unknowns, making the proposed finite difference formulation fully discrete and directly implementable. We approximate the second-order integer derivative as follows:

Lemma 3. Assume that $\xi(\chi) \in C^4[I_-, I_+]$, that $h = \frac{I_+ - I_-}{M}$, and that M is a given positive integer. Then

$$
\begin{aligned}
&\xi''(\chi_0) - \frac{2}{h}\left(\frac{\xi(\chi_1) - \xi(\chi_0)}{h} - \xi'(\chi_0)\right) = -\frac{h}{3}\xi^{(3)}(\chi_0 + \theta_1 h), \quad \theta_1 \in (0,1), \\
&\xi''(\chi_M) - \frac{2}{h}\left(\xi'(\chi_M) - \frac{\xi(\chi_M) - \xi(\chi_{M-1})}{h}\right) = \frac{h}{3}\xi^{(3)}(\chi_M - \theta_2 h), \quad \theta_2 \in (0,1).
\end{aligned} \tag{29}
$$

Proof. Lemma 3 can be easily proven using Taylor's theorem. □

Theorem 1. for $1 \leq i \leq M-1$, $1 \leq k \leq N$ the finite difference scheme (28) is equivalent to

$$
\begin{aligned}
\frac{1}{2}\left({}_0^C D_\rho^\alpha \xi_{i-\frac{1}{2}}^k + {}_0^C D_\rho^\alpha \xi_{i+\frac{1}{2}}^k\right) =& \sigma_\chi^2 \xi_i^k + \frac{1}{2}\left(f(\xi_{i-\frac{1}{2}}^{k-1}) + f(\xi_{i+\frac{1}{2}}^{k-1})\right) \\
&+ \frac{1}{2}\left(g_{i-\frac{1}{2}}^k + g_{i+\frac{1}{2}}^k\right), \\
{}_0^C D_\rho^\alpha \xi_{\frac{1}{2}}^k =& \frac{2}{h}\left(\sigma_\chi \xi_{\frac{1}{2}}^k - {}_0^C D_\rho^{\frac{\alpha}{2}} \xi_0^k + f(\xi_0^{k-1})\right) + g_{\frac{1}{2}}^k, \\
{}_0^C D_\rho^\alpha \xi_{M-\frac{1}{2}}^k =& \frac{2}{h}\left(-{}_0^C D_\rho^{\frac{\alpha}{2}} \xi_M^k - \sigma_\chi \xi_{M+\frac{1}{2}}^k - f(\xi_M^{k-1})\right) + g_{M-\frac{1}{2}}^k,
\end{aligned}
$$

$$\begin{aligned}\xi_i^0 =&\xi(\chi_i), \quad 0 \le i \le M,\\ \xi_0^k =&\varphi(\rho_k),\end{aligned} \tag{30}$$

and

$$\begin{aligned}v_0^k &= \sigma_\chi \xi_{\frac{1}{2}}^k - \frac{h}{2}\left({}_0^C D_\rho^\alpha \xi_{\frac{1}{2}}^k - f(\xi_0^{k-1}) - g_{\frac{1}{2}}^k\right),\\ v_M^k &= \sigma_\chi \xi_{M-\frac{1}{2}}^k + \frac{h}{2}\left({}_0^C D_\rho^\alpha \xi_{M-\frac{1}{2}}^k - f(\xi_M^{k-1}) - g_{M-\frac{1}{2}}^k\right),\\ v_i^k &= \sigma_\chi \xi_{i-\frac{1}{2}}^k + \frac{h}{2}\left({}_0^C D_\rho^\alpha \xi_{i-\frac{1}{2}}^k - f(\xi_i^{k-1}) - g_{i-\frac{1}{2}}^k\right), \quad 1 \le i \le M,\ 1 \le k \le N.\end{aligned}$$

Proof. We start from the discrete relations at the half-grid points, as presented in (28) of the manuscript:

$$\frac{1}{2}\left(v_{i+1}^k - v_i^k\right) = \frac{h}{2}\left({}_0^C D_\rho^\alpha \xi_{i+\frac{1}{2}}^k - \mathcal{F}(\xi_{i+\frac{1}{2}}^{k-1}) - \mathcal{G}_{i+\frac{1}{2}}^k\right), 0 \le i \le M-1,\ 1 \le k \le N, \tag{31}$$

$$v_{i+\frac{1}{2}}^k = \sigma_\chi\, \xi_{i+\frac{1}{2}}^k. \tag{32}$$

Here we employ the semi-implicit treatment $\mathcal{F}(\xi_i^k) \mapsto \mathcal{F}(\xi_i^{k-1})$ and omit the higher-order remainder terms (see the bounds in (27)), the neglected terms are of order $\mathcal{O}(\varrho^{3-\alpha} + h^2)$. To derive an explicit expression for v_{i+1}^k, we add (31) and (32) evaluated at the same half-point. Using the standard identity $v_{i+\frac{1}{2}}^k = \frac{1}{2}(v_{i+1}^k + v_i^k)$, which follows from the definition of the half-grid variables, we obtain after simple algebra

$$v_{i+1}^k = \sigma_\chi\, \xi_{i+\frac{1}{2}}^k + \frac{h}{2}\left({}_0^C D_\rho^\alpha \xi_{i+\frac{1}{2}}^k - \mathcal{F}(\xi_{i+\frac{1}{2}}^{k-1}) - \mathcal{G}_{i+\frac{1}{2}}^k\right).$$

Shifting the index $i \mapsto i-1$ yields the equivalent relation

$$v_i^k = \sigma_\chi\, \xi_{i-\frac{1}{2}}^k + \frac{h}{2}\left({}_0^C D_\rho^\alpha \xi_{i-\frac{1}{2}}^k - \mathcal{F}(\xi_{i-\frac{1}{2}}^{k-1}) - \mathcal{G}_{i-\frac{1}{2}}^k\right), \qquad 1 \le i \le M. \tag{33}$$

Next, to obtain an expression for v_i^k in terms of the right half-point, we use (31) at index i together with (32) at $i+\frac{1}{2}$. Eliminating the difference $v_{i+1}^k - v_i^k$ between these two relations gives

$$v_i^k = \sigma_\chi\, \xi_{i+\frac{1}{2}}^k - \frac{h}{2}\left({}_0^C D_\rho^\alpha \xi_{i+\frac{1}{2}}^k - \mathcal{F}(\xi_{i+\frac{1}{2}}^{k-1}) - \mathcal{G}_{i+\frac{1}{2}}^k\right), \qquad 0 \le i \le M-1. \tag{34}$$

Now, combine (33) and (34) for interior indices $1 \le i \le M-1$. Adding the two expressions for v_i^k and dividing by two yields

$$\frac{1}{2}\left({}_0^C D_\rho^\alpha \xi_{i-\frac{1}{2}}^k + {}_0^C D_\rho^\alpha \xi_{i+\frac{1}{2}}^k\right) = \sigma_\chi^2 \xi_i^k + \frac{1}{2}\left(\mathcal{F}(\xi_{i-\frac{1}{2}}^{k-1}) + \mathcal{F}(\xi_{i+\frac{1}{2}}^{k-1})\right) + \frac{1}{2}\left(\mathcal{G}_{i-\frac{1}{2}}^k + \mathcal{G}_{i+\frac{1}{2}}^k\right),$$

which is exactly the interior equation stated in Theorem 1. In the interior domain, the nonlinear term $\mathcal{F}(\xi_i^k)$ is treated semi-implicitly, $\mathcal{F}(\xi_i^k)$ is replaced by its previous time level $\mathcal{F}(\xi_i^{k-1})$ to ensure numerical stability and to avoid solving a nonlinear system at each step. However, in the boundary relations derived from the discrete ABCs, this term naturally appears with the current time index k, since those relations are obtained directly from the continuous ABC operators and therefore do not involve any time-lagging approximation. This difference in the treatment of $\mathcal{F}$ does not affect the overall consistency, as both approximations preserve the same truncation order $\mathcal{O}(\varrho^{3-\alpha}+h^2)$ (27). Finally, to derive the right boundary formula, substitute (33) with $i=M$ and use the ABC given in the manuscript:

$$ {}_0^C D_\rho^{\alpha/2} v_M^k = -\big(\sigma_\chi v_M^k + \mathcal{F}(\xi_M^{k-1}) + \mathcal{G}_M^k\big), \qquad v_M^k = \sigma_\chi \xi_M^k. $$

Note that ${}_0^C D_\rho^{\alpha/2} v_M^k = {}_0^C D_\rho^{\alpha/2}(\sigma_\chi \xi_M^k)$. Substituting this relation and rearranging terms gives

$$ {}_0^C D_\rho^{\alpha} \xi_{M-\frac{1}{2}}^k = \frac{2}{h}\Big(- {}_0^C D_\rho^{\alpha/2}(\sigma_\chi \xi_M^k) - \sigma_\chi \xi_{M+\frac{1}{2}}^k - \mathcal{F}(\xi_M^{k-1})\Big) + \mathcal{G}_{M-\frac{1}{2}}^k. $$

Here $\xi_{M+\frac{1}{2}}^k$ denotes the boundary-adjacent value treated via the discrete ABC approximation Lemma 3. The left boundary relation is obtained analogously. Combining the interior and boundary equations shows that the total number of discrete equations matches the number of unknown nodal values, so the fully discrete system is well-posed and can be solved uniquely at each time level. This completes the proof. □

4 Stability and convergence of the numerical method

In this section, we present several preliminary lemmas that will be essential for establishing the stability and convergence of the proposed finite difference scheme. These lemmas provide useful discrete inequalities and auxiliary estimates that simplify the forthcoming analysis.

Lemma 4. [5, 11] Suppose the network function $\xi = \{\xi^k \mid 0 \le k \le N\}$ is defined on the temporal grid Π_ϱ. Then, the following inequality holds:

$$ \varrho \sum_{k=1}^{n} \big({}_0^C D_\rho^{\alpha} \xi^k\big)\, \xi^k \ge \frac{\rho_n^{-\alpha}}{2\Gamma(1-\alpha)} \varrho \sum_{k=1}^{n} \big(\xi^k\big)^2 - \frac{\rho_n^{1-\alpha}}{2\Gamma(2-\alpha)} \big(\xi^0\big)^2, \quad 1 \le n \le N, \tag{35} $$

where ${}_0^C D_\rho^\alpha \xi^k$ for $k \geq 1$ is defined according to the relation (22). This lemma plays a crucial role in establishing the positivity and coercivity of the discrete Caputo operator, which are essential properties for the stability and convergence analysis of fractional time-stepping schemes.

Lemma 5. [5] Let ξ be a grid function defined on the spatial grid Π_h with $\xi_0 = 0$. Then the following discrete inverse inequality holds:

$$\|\xi^k\|_\infty^2 \leq C \, \|\sigma_\chi \xi^k\|^2, \tag{36}$$

where C is a positive constant independent of h.

The above lemmas provide the fundamental discrete estimates required for analyzing the stability and convergence properties of the proposed scheme.

Theorem 2 (Initial estimate)**.** Assume that $\{\xi_i^k : 0 \leq i \leq M,\ 0 \leq k \leq N\}$ is the grid solution of the discrete scheme (37)–(44), with the notation $\rho_n = n\varrho$ and the discrete norms $\|\cdot\|$ (discrete L^2) and $\|\cdot\|_\infty$ (discrete supremum). Moreover,

$$ {}_0^C D_\rho^\alpha \xi_{i+\frac{1}{2}}^k = \sigma_\chi v_{i+\frac{1}{2}}^k + f(\xi_{i+\frac{1}{2}}^{k-1}) + g_{i+\frac{1}{2}}^k, \qquad 0 \leqslant i \leqslant M-1, \tag{37}$$

$$ v_{i+\frac{1}{2}}^k = \sigma_\chi \xi_{i+\frac{1}{2}}^k + z_{i+\frac{1}{2}}^k \qquad 0 \leqslant i \leqslant M-1, \tag{38}$$

$$ {}_0^C D_\rho^{\frac{\alpha}{2}} v_0^k = \sigma_\chi v_0^k + f(\xi_0^{k-1}) + g_0^k + t^k, \tag{39}$$

$$ v_0^k = \sigma_\chi \xi_0^k + a^k, \tag{40}$$

$$ {}_0^C D_\rho^{\frac{\alpha}{2}} v_M^k = -\sigma_\chi v_M^k - f(\xi_M^{k-1}) - g_M^k + w^k, \tag{41}$$

$$ v_M^k = \sigma_\chi \xi_M^k + q^k, \tag{42}$$

$$ \xi_i^0 = \varphi(\chi_i), \qquad \chi \in \Delta_i, \tag{43}$$

$$ \xi_0^k = 0, \qquad \rho > 0. \tag{44}$$

Suppose furthermore that the remainder terms $z_{i+\frac{1}{2}}^k, t^k, a^k, w^k, q^k$ satisfy the bounds from Lemma 2 and relation (27):

$$|z_{i+\frac{1}{2}}^k| \leq C_0 h^2, \quad |a^k|, |q^k| \leq C_0 h^2, \quad |t^k|, |w^k| \leq C_0(\varrho^{3-\frac{\alpha}{2}} + h^2),$$

for some $C_0 > 0$ independent of h, ϱ. Let

$$\mu_n := \frac{\rho_n^{-\alpha}}{\Gamma(1-\alpha)}, \qquad 1 \leq n \leq N.$$

Then for every $1 \leq n \leq N$ it holds that

$$\varrho\sum_{k=1}^{n}\|\xi^k\|_\infty^2+\mu_n\,\varrho\sum_{k=1}^{n}\|\xi^k\|^2\le C_1\|\xi^0\|^2+C_2\varrho\sum_{k=1}^{n}(f(\xi^{k-1}),\xi^k)+C_3\varrho\sum_{k=1}^{n}\|g^k\|^2+C_4\varrho\sum_{k=1}^{n}(q^k)^2+C_5\varrho\sum_{k=1}^{n}\big((\xi_M^k)^2+(\sigma_\chi\xi_M^k)^2\big),\tag{45}$$

where $C_1,\dots,C_5>0$ are constants depending only on α, C_0 the remainder bounds and universal constants but independent of h, ϱ and n.

Proof. Using the discrete inner product and applying relations (37)–(38), we obtain

$$\begin{aligned}({}_0^C D_\rho^\alpha\xi^k,\,\xi^k)&=(\sigma_\chi v^k,\,\xi^k)+(f(\xi^{k-1}),\,\xi^k)+(g^k,\,\xi^k),\\(v^k,\,v^k)&=(\sigma_\chi\xi^k,\,v^k)+(z^k_{i+\frac12},\,v^k),\qquad 1\le k\le N.\end{aligned}\tag{46}$$

Adding these two relations gives

$$\begin{aligned}({}_0^C D_\rho^\alpha\xi^k,\,\xi^k)+\|v^k\|^2=&(\sigma_\chi v^k,\,\xi^k)+(\sigma_\chi\xi^k,\,v^k)+(f(\xi^{k-1}),\,\xi^k)+(g^k,\,\xi^k)\\&+(z^k_{i+\frac12},\,v^k),\qquad 1\le k\le N.\end{aligned}\tag{47}$$

Next, using the discrete integration-by-parts identity together with the boundary conditions (42) and (44), we have

$$\begin{aligned}(\sigma_\chi\xi^k,\,v^k)+(\sigma_\chi v^k,\,\xi^k)&=h\sum_{i=1}^{M}\big[(\sigma_\chi\xi^k_{i-\frac12})v^k_{i-\frac12}+(\sigma_\chi v^k_{i-\frac12})\xi^k_{i-\frac12}\big]\\&=\xi_M^k v_M^k-\xi_0^k v_0^k=\xi_M^k(\sigma_\chi\xi_M^k+q^k),\end{aligned}\tag{48}$$

since $\xi_0^k=0$ for all k. Combining (47) and (48) yields

$$\begin{aligned}({}_0^C D_\rho^\alpha\xi^k,\,\xi^k)+\|v^k\|^2=&(f(\xi^{k-1}),\,\xi^k)+(g^k,\,\xi^k)+\xi_M^k\sigma_\chi\xi_M^k+\xi_M^k q^k\\&+(z^k_{i+\frac12},\,v^k),\qquad 1\le k\le N.\end{aligned}\tag{49}$$

Multiplying both sides of (49) by ϱ, summing over $k=1,\dots,n$, and applying Lemma 4, we obtain

$$\frac{\rho_n^{-\alpha}}{2\Gamma(1-\alpha)}\,\varrho\sum_{k=1}^{n}\|\xi^k\|^2+\varrho\sum_{k=1}^{n}\|v^k\|^2$$

$$\leq \frac{\rho_n^{1-\alpha}}{2\Gamma(2-\alpha)}\,\|\xi^0\|^2 + \varrho\sum_{k=1}^{n}(f(\xi^{k-1}),\,\xi^k) + \varrho\sum_{k=1}^{n}(g^k,\,\xi^k)$$
$$+\,\varrho\sum_{k=1}^{n}\xi_M^k\,\sigma_\chi\xi_M^k + \varrho\sum_{k=1}^{n}(z_{i+\frac{1}{2}}^k,\,v^k) + \varrho\sum_{k=1}^{n}\xi_M^k\,q^k,\quad 1\leq n\leq N. \tag{50}$$

Applying the Cauchy–Schwarz and Young inequalities to the right-hand side of (50) yields the following estimate:

$$\frac{\rho_n^{-\alpha}}{2\Gamma(1-\alpha)}\,\varrho\sum_{k=1}^{n}\|\xi^k\|^2 + \varrho\sum_{k=1}^{n}\|v^k\|^2$$
$$\leq \frac{\rho_n^{1-\alpha}}{2\Gamma(2-\alpha)}\,\|\xi^0\|^2 + \varrho\sum_{k=1}^{n}(f(\xi^{k-1}),\,\xi^k)$$
$$+\,\varrho\sum_{k=1}^{n}\left(\frac{\rho_n^{-\alpha}}{4\Gamma(1-\alpha)}\,\|\xi^k\|^2 + \rho_n^{\alpha}\Gamma(1-\alpha)\,\|g^k\|^2\right)$$
$$+\,\varrho\sum_{k=1}^{n}\left(\frac{1}{2}\|v^k\|^2 + \frac{1}{2}\|z_{i+\frac{1}{2}}^k\|^2\right) + \varrho\sum_{k=1}^{n}\left(\frac{1}{2}(\xi_M^k)^2 + \frac{1}{2}(q^k)^2\right)$$
$$+\,\varrho\sum_{k=1}^{n}\left(\frac{1}{2}(\xi_M^k)^2 + \frac{1}{2}(\sigma_\chi\xi_M^k)^2\right),\quad 1\leq n\leq N. \tag{51}$$

Using the estimate (51) and moving the small left-hand terms to the left, we obtain

$$\frac{\rho_n^{-\alpha}}{4\Gamma(1-\alpha)}\,\varrho\sum_{k=1}^{n}\|\xi^k\|^2 + \frac{1}{2}\,\varrho\sum_{k=1}^{n}\|v^k\|^2$$
$$\leq \frac{\rho_n^{1-\alpha}}{2\Gamma(2-\alpha)}\,\|\xi^0\|^2 + \varrho\sum_{k=1}^{n}(f(\xi^{k-1}),\,\xi^k)$$
$$+\,\rho_n^{\alpha}\Gamma(1-\alpha)\,\varrho\sum_{k=1}^{n}\|g^k\|^2 + \frac{1}{2}\,\varrho\sum_{k=1}^{n}\|z_{i+\frac{1}{2}}^k\|^2 + \varrho\sum_{k=1}^{n}\left(\frac{1}{2}(q^k)^2 + (\xi_M^k)^2\right)$$
$$+\,\frac{1}{2}\,\varrho\sum_{k=1}^{n}(\sigma_\chi\xi_M^k)^2, \tag{52}$$

for every $1\leq n\leq N$. Using relation (38) together with Lemma 5, we have the pointwise bound

$$\|\xi^k\|_\infty^2 \leq \|\sigma_\chi\xi^k\|^2 = \|v^k - z_{i+\frac{1}{2}}^k\|^2 \leq 2\|v^k\|^2 + 2\|z_{i+\frac{1}{2}}^k\|^2,\qquad 1\leq k\leq N. \tag{53}$$

Now, multiply (53) by ϱ and sum over $k=1,\ldots,n$, then combine with (52). Using the fact that $\frac{1}{2}\varrho\sum_{k=1}^{n}\|v^k\|^2$ appears on the left of (52), we obtain after elementary algebra the inequality

$$\varrho\sum_{k=1}^{n}\|\xi^k\|_\infty^2+\frac{\rho_n^{-\alpha}}{\Gamma(1-\alpha)}\,\varrho\sum_{k=1}^{n}\|\xi^k\|^2-4\varrho\sum_{k=1}^{n}(\xi_M^k)^2-2\varrho\sum_{k=1}^{n}(\sigma_\chi\xi_M^k)^2$$
$$\le\frac{2\rho_n^{1-\alpha}}{\Gamma(2-\alpha)}\,\|\xi^0\|^2+4\varrho\sum_{k=1}^{n}(f(\xi^{k-1}),\,\xi^k)+4\rho_n^{\alpha}\Gamma(1-\alpha)\,\varrho\sum_{k=1}^{n}\|g^k\|^2$$
$$+\quad C_z\,\varrho\sum_{k=1}^{n}\|z_{i+\frac{1}{2}}^k\|^2+\varrho\sum_{k=1}^{n}(q^k)^2, \tag{54}$$

for every $1\le n\le N$, where $C_z>0$ is a numerical constant (here one may take $C_z=4$). Finally, by introducing the notation

$$\mu_n=\frac{\rho_n^{-\alpha}}{\Gamma(1-\alpha)},$$

and substituting this definition into (54), we simplify the inequality as

$$\varrho\sum_{k=1}^{n}\|\xi^k\|_\infty^2+\mu_n\,\varrho\sum_{k=1}^{n}\|\xi^k\|^2-4\varrho\sum_{k=1}^{n}(\xi_M^k)^2-2\varrho\sum_{k=1}^{n}(\sigma_\chi\xi_M^k)^2$$
$$\le\frac{2\mu_n\rho_n}{1-\alpha}\,\|\xi^0\|^2+4\varrho\sum_{k=1}^{n}(f(\xi^{k-1}),\xi^k)+\frac{4}{\mu_n}\,\varrho\sum_{k=1}^{n}\|g^k\|^2+2\varrho\sum_{k=1}^{n}(q^k)^2. \tag{55}$$

This inequality coincides precisely with relation (45), which completes the proof of Theorem 2. □

The purpose of Theorem 2 (Initial Estimate) is to establish a discrete *a priori* energy bound for the numerical solution. This estimate ensures that the grid solution $\{\xi_i^k\}$ remains uniformly bounded at all temporal levels and that the discrete ABCs do not introduce instability into the system. Therefore, this theorem provides the theoretical foundation for the subsequent stability and convergence analysis of the proposed numerical method.

Theorem 3 (Stability)**.** Based on the discrete scheme (30) and under the assumptions of Theorem 2, the proposed finite difference method is *unconditionally stable*. Specifically, let ξ_i^k $(0\le i\le M,\,1\le k\le N)$ denote the numerical solution with homogeneous initial data $\varphi(\rho_k)=0$ for all $1\le k\le N$. Then, the following stability estimate holds:

$$\varrho\sum_{k=1}^{n}\|\xi^k\|_\infty^2+\frac{\mu_n}{4}\,\varrho\sum_{k=1}^{n}\|\xi^k\|^2-2\varrho\sum_{k=1}^{n}(\xi_M^k)^2-\varrho\sum_{k=1}^{n}(\sigma_\chi\xi_M^k)^2$$
$$\le\frac{\mu_n\rho_n}{2(1-\alpha)}\,\|\xi^0\|^2+\varrho\sum_{k=1}^{n}\big(f(\xi^{k-1}),\xi^k\big)+\frac{1}{\mu_n}\,\varrho\sum_{k=1}^{n}\|g^k\|^2, \tag{56}$$

where $\mu_n=\dfrac{\rho_n^{-\alpha}}{\Gamma(1-\alpha)}$, and C_0 denotes the remainder constant introduced in relation (27).

Proof. The proof follows the same energy argument as in Theorem 2. □

In the following, we examine the convergence of the difference method described in relations (37)–(43). We assume that

$$e_i^k = U_i^k - \xi_i^k, \quad \varepsilon_i^k = V_i^k - v_i^k, \quad 0 \le i \le M, \quad 0 \le k \le N,$$

where U_i^k and V_i^k denote the numerical approximations of $\xi(\chi_i, \rho_k)$ and $v(\chi_i, \rho_k)$, respectively.

Theorem 4 (Convergence)**.** Assume that $\xi(\chi, \rho) \in C^{4,4}_{\chi,\rho}([I_-, I_+] \times [0, P])$, and let $\{\xi_i^k \mid 0 \le i \le M,\ 0 \le k \le N\}$ and $\{U_i^k \mid 0 \le i \le M,\ 0 \le k \le N\}$ be the exact and numerical solutions of the problems (15) and (37)–(43), respectively. Then there exists a constant $C_1 > 0$ such that

$$\sqrt{\varrho \sum_{k=1}^{n} \big(\|e^k\|_\infty^2 - 4(e_M^k)^2 - 2(\sigma_\chi e_M^k)^2 + \mu \|e^k\|^2 \big)} \le C_1(\varrho^{3-\alpha} + h^2), \quad 1 \le n \le N, \tag{57}$$

where $e_i^k = U_i^k - \xi_i^k$ is the discretization error. The constant C_1 is given by

$$C_1 = 2C_0 T^{\alpha+\frac{1}{2}} \sqrt{\tfrac{1}{2} + \Gamma(1-\alpha)}, \tag{58}$$

and μ is defined in Theorem 2, while C_0 is specified in (27).

Proof. Following the approach of Theorem 2, we subtract the exact relation (28) from the discrete scheme (26) to obtain the error system:

$$\begin{aligned}
&{}_0^C D_\rho^\alpha e_{i+\frac{1}{2}}^k = \sigma_\chi \varepsilon_{i+\frac{1}{2}}^k + \big(f(U_i^{k-1}) - f(\xi_i^{k-1}) \big) + R_{i+\frac{1}{2}}^k, \quad 0 \le i \le M-1,\ 1 \le k \le N,\\
&\varepsilon_{i+\frac{1}{2}}^k = \sigma_\chi e_{i+\frac{1}{2}}^k + Z_{i+\frac{1}{2}}^k,\\
&{}_0^C D_\rho^{\frac{\alpha}{2}} \varepsilon_0^k = \sigma_\chi \varepsilon_0^k + \big(f(U_0^{k-1}) - f(\xi_0^{k-1}) \big) + T^k, \quad \varepsilon_0^k = \sigma_\chi e_0^k + A^k,\\
&{}_0^C D_\rho^{\frac{\alpha}{2}} \varepsilon_M^k = -\sigma_\chi \varepsilon_M^k - \big(f(U_M^{k-1}) - f(\xi_M^{k-1}) \big) + W^k, \quad \varepsilon_M^k = \sigma_\chi e_M^k + Q^k,\\
&e_i^0 = 0, \quad \chi \in \Delta_i, \qquad e_0^k = 0.
\end{aligned} \tag{59}$$

Applying Theorem 2 to this error system yields

$$\begin{aligned}
&\varrho \sum_{k=1}^{n} \|e^k\|_\infty^2 - 4\varrho \sum_{k=1}^{n} (e_M^k)^2 - 2\varrho \sum_{k=1}^{n} (\sigma_\chi e_M^k)^2 + \mu\varrho \sum_{k=1}^{n} \|e^k\|^2\\
&\qquad \le \frac{4}{\mu} \varrho \sum_{k=1}^{n} \|R^k\|^2 + 2\varrho \sum_{k=1}^{n} \|Q^k\|^2, \quad 1 \le n \le N.
\end{aligned} \tag{60}$$

Using the remainder bounds from (27), $|R^k|, |Z^k|, |T^k|, |W^k|, |Q^k| \le C_0(\varrho^{3-\alpha} + h^2)$, we have

$$\begin{aligned}
&\varrho \sum_{k=1}^{n} \|e^k\|_\infty^2 + \mu\varrho \sum_{k=1}^{n} \|e^k\|^2 - 4\varrho \sum_{k=1}^{n} (e_M^k)^2 - 2\varrho \sum_{k=1}^{n} (\sigma_\chi e_M^k)^2 \\
&\quad \le 2\rho_n C_0^2 h^4 + \frac{4}{\mu}\big(C_0(\varrho^{3-\alpha} + h^2)\big)^2 \\
&\quad = 2\rho_n C_0^2 h^4 + 4\rho_n^{\alpha+1}\Gamma(1-\alpha)\big(C_0(\varrho^{3-\alpha} + h^2)\big)^2 \\
&\quad \le 4C_0^2 T^{1+\alpha}\Big(\tfrac{1}{2} + \Gamma(1-\alpha)\Big)(\varrho^{3-\alpha} + h^2)^2. \qquad (61)
\end{aligned}$$

Taking square roots of both sides yields

$$\begin{aligned}
&\sqrt{\varrho \sum_{k=1}^{n} \big(\|e^k\|_\infty^2 - 4(e_M^k)^2 - 2(\sigma_\chi e_M^k)^2 + \mu\|e^k\|^2\big)} \\
&\le 2C_0 T^{\alpha+\frac{1}{2}} \sqrt{\tfrac{1}{2} + \Gamma(1-\alpha)}(\varrho^{3-\alpha} + h^2), \qquad (62)
\end{aligned}$$

which proves (57)–(58). □

5 Numerical examples

In this section, two representative numerical examples are provided to comprehensively evaluate the accuracy, stability, and efficiency of the proposed computational scheme. In Example 1, the objective is to examine the spatiotemporal behavior of the numerical solution and to verify the convergence and stability properties of the method under various parameter settings. In Example 3, the performance of the proposed finite difference scheme given by (28) is compared with the numerical approach reported in [14], allowing for both quantitative and qualitative analysis in terms of accuracy and computational cost. All simulations were implemented in `MATLAB R2022a`. The numerical computations were performed on an `Asus K53SC` laptop equipped with an `Intel Core i5-2430M` processor and `8 GB RAM`. The obtained numerical results demonstrate that the proposed method yields highly accurate and stable solutions while maintaining computational efficiency even on a mid-range hardware platform.

Example 1. Consider the following time-fractional equation:

$$ {}_0^C D_\rho^\alpha \xi(\chi, \rho) = \xi_{\chi\chi} + \mathcal{F}(\xi(\chi, \rho)) + \mathcal{G}(\xi(\chi, \rho)), \quad 0 \le \rho \le P, \quad \chi \in \mathbb{R}, $$

where the nonlinear term is considered as follows:

$$\mathcal{F}(\xi) = \xi^2.$$

The source term is obtained in the following form:

$$\mathcal{G}(\chi,\rho) = \frac{e^{-\chi^2} 6\rho^{3-\alpha}}{\Gamma(4-\alpha)} - e^{-\chi^2}(\rho^3+1)(4\chi^2-2) - e^{-2\chi^2}(\rho^3+1)^2.$$

This problem also has the initial condition:

$$\xi(\chi,0) = e^{-\chi^2},$$

and the exact solution is

$$\xi(\chi,\rho) = e^{-\chi^2}(\rho^3+1).$$

In this example, the temporal and spatial discretizations are specified by setting the number of temporal grid points to $N = 320$ and the number of spatial grid points to $M = 2^9$. The fractional order is considered for two representative cases, $\alpha \in \{0.37, 0.68\}$. The spatial step size is given by $h = 1/M$, while the temporal step sizes are chosen as $\varrho \in \left\{\frac{1}{10}, \frac{1}{20}, \frac{1}{40}, \frac{1}{80}, \frac{1}{160}, \frac{1}{320}\right\}$. The spatial computational domain is defined as $I = [a,b] = [-2,2]$, and the final time is taken as $P = 1$. This example is designed to assess the numerical accuracy of the proposed method for solving time-fractional differential equations with prescribed initial conditions and known analytical solutions. Both numerical and graphical results are provided to illustrate the spatiotemporal behavior of the computed solution and to compare it with the exact analytical solution. The numerical error is defined as follows:

$$E(h,\varrho) = \|e_h^k\|_\infty = \|U_h^k - \xi(\chi_h,\rho_k)\|.$$

The convergence order for the temporal and spatial variations is calculated as follows:

$$\text{order}_1 = \log_2 \frac{E(h,\varrho)}{E(h,\varrho/2)}, \quad \text{order}_2 = \log_2 \frac{E(h,\varrho)}{E(h/2,\varrho)}.$$

Using a fixed spatial step size of $h = 2^{-9}$ and based on the numerical results summarized in Table 1, it is evident that the absolute error $\|e_h^k\|_\infty$ decreases monotonically as the temporal step size ϱ is refined from $\frac{1}{10}$ to $\frac{1}{320}$. This consistent reduction in error confirms the temporal convergence of the proposed scheme. Furthermore, the observed convergence rates, denoted by order_1, for both fractional orders $\alpha = 0.37$ and $\alpha = 0.68$ remain approximately second order, with values ranging between 2.00 and 2.42. The slight reduction of the observed rates compared to the theoretical convergence rate $O(\varrho^{3-\alpha})$ is attributed to the boundary approximations, which

slightly limit the global convergence. Nevertheless, the method achieves the expected high-order accuracy at interior points. In particular, the obtained convergence orders slightly exceed those of the classical $L1$ scheme, thereby validating the improved temporal discretization strategy developed in this study. Additionally, the CPU time reported in the last column of Table 1 shows a gradual and proportional increase as the temporal step size decreases. This trend indicates that the computational cost grows moderately with increasing accuracy, confirming the overall efficiency and robustness of the implemented algorithm. Referring to Table 2, the spatial

Table 1: Error results and temporal convergence rates for Example 1 corresponding to various values of α and ϱ.

α	τ	$\|e_h^k\|_\infty$	$order_1$	CPU
0.37	1/10	3.622383e-2	–	0.15
	1/20	9.036367e-3	2.0031	0.27
	1/40	2.098646e-3	2.1063	0.40
	1/80	4.456779e-4	2.2354	0.54
	1/160	8.283166e-5	2.4277	0.85
	1/320	1.783819e-5	2.2152	1.45
0.68	1/10	3.634468e-2	–	0.14
	1/20	9.077627e-3	2.0014	0.21
	1/40	2.129381e-3	2.0919	0.29
	1/80	4.610580e-4	2.2074	0.56
	1/160	9.052296e-5	2.3486	0.83
	1/320	2.168393e-5	2.0617	1.46

convergence behavior of the proposed scheme is analyzed by fixing the temporal step size at $\varrho = 2^{-9}$. It is clearly observed that the absolute error $\|e_h^k\|_\infty$ decreases consistently as the spatial mesh is refined, with h reduced from 1/10 to 1/320. For both fractional orders $\alpha = 0.37$ and $\alpha = 0.68$, the convergence rate, denoted as order_2, remains remarkably stable and very close to the theoretical second-order accuracy, with reported values around 1.99 across all grid refinements. This uniform convergence confirms that the spatial discretization employed in the scheme achieves the expected second-order accuracy, in agreement with the central-difference approximation used for the second spatial derivative. The obtained numerical results demonstrate that the proposed method maintains high spatial accuracy while preserving stability and efficiency. Furthermore, the CPU time values shown in the last column of Table 2 indicate a moderate and proportional increase in computational cost as the grid is refined, confirming the scalability and computational efficiency of the implemented algorithm. In Figure 1, the numerical solution obtained by the present method, (28), is compared with the exact analytical solution for Example 1. The visual comparison reveals excellent agreement between the numerical and exact solutions throughout the entire spatiotemporal domain. This strong consistency verifies the ro-

bustness and accuracy of the proposed numerical scheme in capturing the dynamic behavior of the underlying time-fractional problem, both in the temporal and spatial dimensions. The plot

Table 2: Error results and spatial convergence rates for Example 1 corresponding to various values of α and h.

α	h	$\|e_h^k\|_\infty$	$order_2$	CPU
0.37	1/10	6.804864e-2	–	0.19
	1/20	1.765156e-2	1.9468	0.21
	1/40	4.443487e-3	1.9900	0.32
	1/80	1.112211e-3	1.9983	0.46
	1/160	2.781080e-4	1.9997	0.81
	1/320	6.952907e-5	2.0000	1.43
0.68	1/10	7.850191e-2	–	0.14
	1/20	2.075293e-2	1.9194	0.16
	1/40	5.231265e-3	1.9881	0.30
	1/80	1.309223e-3	1.9984	0.45
	1/160	3.273427e-4	1.9998	0.83
	1/320	8.183599e-5	2.0000	1.46

presented in Figure 2 illustrates the comparison between the numerical and analytical solutions when the spatial step size is fixed at $h = 2^{-9}$ and the temporal step size ϱ is successively refined. It is clearly observed that, as ϱ decreases, the numerical solution progressively approaches the exact analytical solution across the entire computational domain. This uniform improvement in accuracy demonstrates the temporal convergence of the proposed scheme, (28), confirming its numerical consistency with the theoretical prediction based on the Caputo fractional derivative formulation. The close visual agreement between the numerical and exact solutions further verifies the reliability and precision of the method in capturing the transient dynamics of the time-fractional model. The graphical results displayed in Figures 3 and 4 provide a clear and

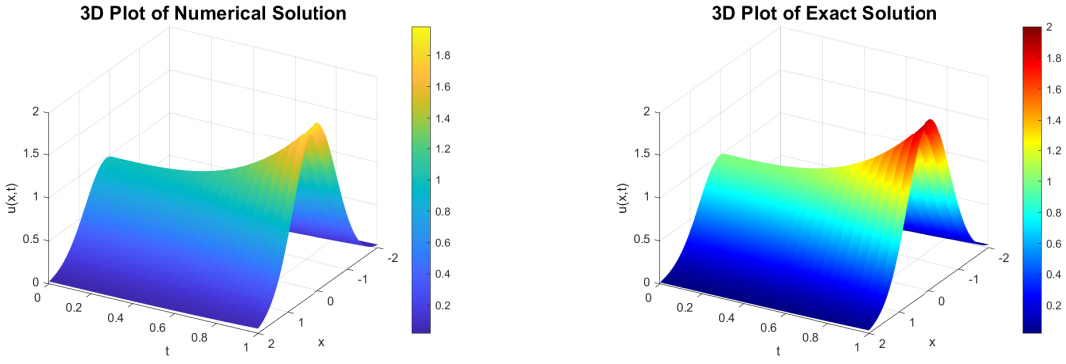


Figure 1: Comparison between the exact solution and the numerical solution of Example 1.

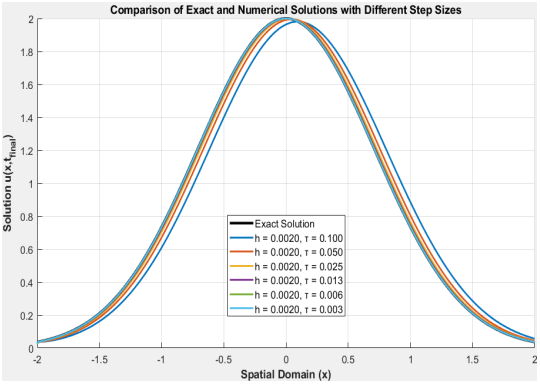


Figure 2: Convergence of the numerical approximation to the exact solution for Example 1.

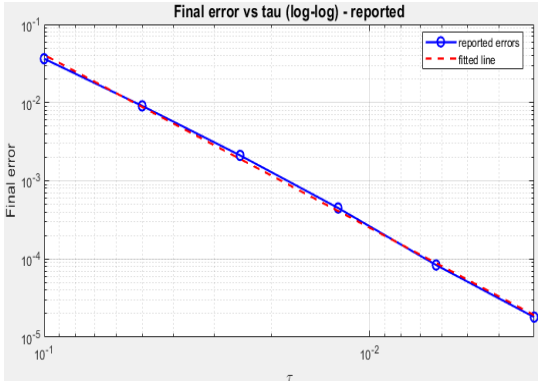


Figure 3: Logarithmic error plot versus ϱ for Example 1.

convincing verification of the theoretical temporal accuracy and stability of the proposed numerical scheme (28). As shown in Figure 3, the log-log representation of the maximum absolute error $\|e_h^k\|_\infty$ versus the time-step size ϱ exhibits an almost perfectly linear decay. The slope of the fitted line corresponds to an observed convergence rate of approximately $p \approx 2.45$, which agrees very well with the theoretical prediction $\mathcal{O}(\varrho^{3-\alpha})$ for $\alpha = 0.37$. This strong agreement between the numerical and theoretical rates confirms that the proposed temporal discretization accurately captures the high-order behavior expected from the fractional derivative formulation. Moreover, the excellent linearity of the data points in the logarithmic scale indicates that the error decreases uniformly with time-step refinement, providing strong evidence of asymptotic temporal consistency and robustness of the scheme. Figure 4 further illustrates the evolution of the absolute error with respect to time for different temporal step sizes. It is clearly observed that as ϱ becomes smaller, the magnitude of the error significantly decreases, and its temporal profile remains smooth and stable throughout the entire simulation period. For sufficiently fine time steps, the error remains nearly constant, indicating that the proposed scheme effectively suppresses error propagation and prevents numerical oscillations. Only for coarser grids does a slight increase in the error magnitude appear near the final time, which is expected due to the accumulation of local truncation errors inherent to fractional-order models. Overall, the results shown in Figures 3 and 4 demonstrate that the proposed method achieves excellent agreement with the theoretical convergence rate, while maintaining remarkable numerical stability even over long-time integrations. These findings confirm the efficiency and reliability of the developed high-order scheme for accurately resolving the transient dynamics of time-fractional nonlinear parabolic equations on unbounded domains. The consistent convergence behavior, combined with the absence of numerical instabilities, strongly validates both the mathematical formulation and the computational implementation of the proposed approach.

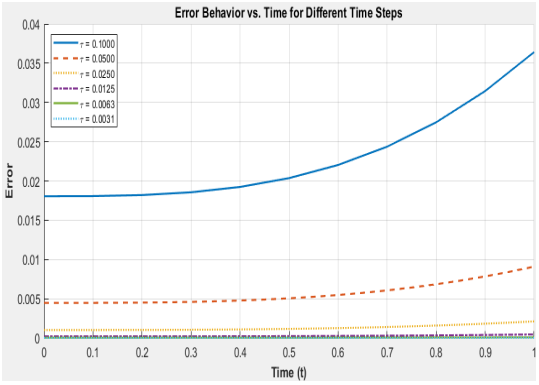


Figure 4: Error behavior for various ϱ values confirming the convergence of the numerical method (28) in Example 1.

Summary: The numerical results presented in Tables 1–2 and Figures 2–4 confirm the high accuracy, stability, and efficiency of the proposed numerical scheme (28). The method exhibits a monotonic decrease in error with both temporal and spatial refinements, achieving convergence rates close to the theoretical orders, $O(\varrho^{3-\alpha})$ in time and $O(h^2)$ in space. Excellent agreement between numerical and analytical solutions validates the correctness of the temporal and spatial discretizations, while the smooth and bounded error evolution confirms unconditional stability. CPU time analysis indicates that computational cost grows moderately with mesh refinement, maintaining a favorable balance between accuracy and efficiency. Overall, the results demonstrate that the proposed approach provides a robust, accurate, and computationally efficient framework for solving time-fractional nonlinear parabolic problems on unbounded domains, effectively fulfilling its theoretical design objectives across all tested configurations. The following Example 2 is introduced to further demonstrate the applicability of the proposed numerical method to physically meaningful problems. Unlike Example 1, which is primarily designed to verify the theoretical convergence behavior of the scheme, the present example considers a nonzero initial condition and a smooth polynomial temporal evolution that commonly arises in diffusion and heat transfer processes with memory effects. The exact solution is chosen as $\xi(\chi,\rho) = e^{-\chi^2}(1+\rho)^3$ which allows the Caputo fractional derivative to be evaluated explicitly without involving special functions such as the Mittag-Leffler function. This choice ensures that the source term can be constructed exactly while preserving analytical simplicity and numerical stability. From a numerical standpoint, this example provides a more comprehensive test of the proposed higher order discretization and absorbing boundary treatment, as it simultaneously incorporates nonlinear effects, fractional temporal memory, and nonzero initial data. Hence, it offers a realistic and robust validation of the effectiveness and accuracy of the proposed method.

Example 2. Consider the following nonlinear time–fractional diffusion equation posed on an unbounded spatial domain:

$$ {}_0^C D_\rho^\alpha \xi(\chi,\rho) = \xi_{\chi\chi} + \mathcal{F}(\xi(\chi,\rho)) + \mathcal{G}(\chi,\rho), \quad 0<\alpha<1, \quad 0 \le \rho \le P, \quad \chi \in \mathbb{R}, $$

where the nonlinear reaction term is defined as

$$ \mathcal{F}(\xi) = \xi^2 + \xi. $$

The initial condition is given by

$$ \xi(\chi,0) = e^{-\chi^2}. $$

We choose the exact solution in the form

$$ \xi(\chi,\rho) = e^{-\chi^2}(1+\rho)^3, $$

the source term is constructed as

$$ \begin{aligned} \mathcal{G}(\chi,\rho) = e^{-\chi^2}\left[\frac{4!}{\Gamma(4-\alpha)}\rho^{3-\alpha} + 3\frac{3!}{\Gamma(3-\alpha)}\rho^{2-\alpha} + 3\frac{2!}{\Gamma(2-\alpha)}\rho^{1-\alpha}\right] \\ - (4\chi^2-2)e^{-\chi^2}(1+\rho)^3 - e^{-2\chi^2}(1+\rho)^6. \end{aligned} $$

Table 3: Temporal convergence results for Example 2 with $h = 1/2^9$.

α	τ	$\|e_h^k\|_\infty$	$order_1$	CPU
0.37	1/10	1.471500e-1	–	0.28
	1/20	3.662159e-2	2.0065	0.35
	1/40	8.584786e-3	2.0928	0.47
	1/80	1.864492e-3	2.2029	0.84
	1/160	3.687139e-4	2.3382	1.13
	1/320	8.914893e-5	2.0482	1.97
0.68	1/10	1.474604e-1	–	0.17
	1/20	3.677890e-2	2.0033	0.28
	1/40	8.663713e-3	2.0858	0.41
	1/80	1.903987e-3	2.1859	0.72
	1/160	3.884649e-4	2.2931	1.17
	1/320	9.102469e-5	2.0934	2.00

Example 3. Consider the following problem:

$$ {}_0^C D_\rho^\alpha \xi(\chi,\rho) = \xi_{\chi\chi} + \mathcal{F}(\xi) \quad \chi \in \mathbb{R}, \quad 0<\rho \le P, $$

with the initial condition given by

Table 4: Spatial convergence results for Example 2 with $\tau = 1/2^9$.

α	h	$\|e_h^k\|_\infty$	$order_2$	CPU
0.37	1/10	3.270424e-1	–	0.15
	1/20	8.452788e-2	1.9519	0.27
	1/40	2.124029e-2	1.9926	0.45
	1/80	5.313965e-3	1.9989	0.67
	1/160	1.328601e-3	1.9998	1.08
	1/320	3.321506e-4	1.9999	2.05
0.68	1/10	3.510207e-1	–	0.13
	1/20	9.244171e-2	1.9249	0.24
	1/40	2.326518e-2	1.9903	0.40
	1/80	5.820151e-3	1.9990	0.65
	1/160	1.455053e-3	1.9999	1.15
	1/320	3.637556e-4	2.0000	2.07

$$\xi(\chi,0) = \left(\frac{1}{1+e^{-\frac{6\chi}{4}}}\right)^{\frac{1}{3}},$$

and the exact solution defined as

$$\xi(\chi,\rho) = \left(\frac{1}{1+e^{-\frac{15\rho+6\chi}{4}}}\right)^{\frac{1}{3}},$$

The nonlinear function $\mathcal{F}(\xi)$ is defined by

$$\mathcal{F}(\xi) = \xi(1-\xi^6),$$

The numerical method employed in [14] to solve the above problem is described as follows:

$$\begin{aligned}
&{}_0^C D_\rho^\alpha \xi_i^k = \sigma_\chi^2 \xi_i^k + \mathcal{F}(\xi_i^{k-1}), \quad 1 \le k \le N, \quad 1 \le i \le M-1,\\
&\sigma_\chi \xi_1^k = \Phi_k + 3s_0^{\alpha/2}\xi_0^k + \frac{h}{2}\left({}_0^C D_\rho^\alpha \xi_0^k - \mathcal{F}(\xi_0^{k-1})\right),\\
&\sigma_\chi \xi_{M-1}^k = \Psi_k - 3s_0^{\alpha/2}\xi_M^k - \frac{h}{2}\left({}_0^C D_\rho^\alpha \xi_M^k - \mathcal{F}(\xi_M^{k-1})\right),\\
&{}_0^C D_\rho^\alpha \Phi_k = \left(\mathcal{F}(\xi_0^{k-1}) - 3s_0^\alpha\right)\Phi_k + \left(8s_0^{3\alpha/2} - 3s_0^{\alpha/2}\mathcal{F}(\xi_0^{k-1})\right)\xi_0^k + 3s_0^{\alpha/2}\mathcal{F}(\xi_0^{k-1}),\\
&{}_0^C D_\rho^\alpha \Psi_k = \left(\mathcal{F}(\xi_M^{k-1}) - 3s_0^\alpha\right)\Psi_k - \left(8s_0^{3\alpha/2} - 3s_0^{\alpha/2}\mathcal{F}(\xi_M^{k-1})\right)\xi_M^k - 3s_0^{\alpha/2}\mathcal{F}(\xi_M^{k-1}),\\
&\xi_i^0 = \xi_0(\chi_i), \quad \Phi_0 = \varphi(0), \quad \Psi_0 = \phi(0).
\end{aligned} \tag{63}$$

Table 5 presents the maximum errors $\|e_h^k\|_\infty$ at the final time P for different time step sizes ϱ and fractional orders α, with a fixed spatial step size $h = 2^{-8}$. The results indicate that the proposed method achieves significantly higher temporal accuracy compared to the method in

[14], with a convergence rate (Order1) between 2.31 and 2.63, consistent with the theoretical order $O(\varrho^{3-\alpha} + h^2)$. In contrast, the method of [14] has a theoretical temporal accuracy of $O(\varrho + h^2)$, which explains why the observed convergence rate in time (Order1) is limited to approximately 0.98–1.00. As the time step ϱ decreases from 1/10 to 1/160, the maximum error of the proposed method drops sharply; for example, with $\alpha = 0.37$, the error decreases from 5.71×10^{-2} to 3.88×10^{-5}, while the corresponding error for the method in [14] reduces only from 2.47×10^{-2} to 1.60×10^{-3}. This demonstrates a significantly faster convergence of the proposed method to the exact solution. Increasing the fractional order from $\alpha = 0.37$ to $\alpha = 0.68$ slightly reduces the temporal convergence rate, as predicted by the term $(3-\alpha)$ in the theoretical order, but the proposed method consistently outperforms the reference method for all α, maintaining lower error values and stable behavior. The smooth, monotonic decrease of errors with decreasing ϱ confirms the robust numerical stability of the proposed scheme, with no oscillations or divergence observed even for very fine temporal grids. Overall, these results indicate that the proposed method achieves both higher temporal accuracy and faster convergence than the method in [14], and its superior performance is especially pronounced on finer temporal grids, highlighting its suitability for applications where temporal precision is critical. The observed

Table 5: Temporal errors and convergence of the proposed method for Example 3.

α	ϱ	Our method (28)	order1	Method in [14]	order1
0.37	1/10	$5.7138196e-2$	–	$2.465416e-2$	–
	1/20	$9.2080193e-3$	2.6334	$1.246783e-2$	0.98
	1/40	$1.4874461e-3$	2.6300	$6.285612e-3$	0.99
	1/80	$2.4005411e-4$	2.6314	$3.170451e-3$	0.99
	1/160	$3.8839013e-5$	2.6277	$1.595221e-3$	0.99
0.68	1/10	$3.1182567e-2$	–	$3.945612e-2$	–
	1/20	$6.2297534e-3$	2.3234	$1.985421e-2$	0.99
	1/40	$1.2475707e-3$	2.3200	$9.963412e-3$	0.99
	1/80	$2.4960448e-4$	2.3214	$4.993210e-3$	0.99
	1/160	$5.0064596e-5$	2.3177	$2.500843e-3$	1.00

differences in temporal convergence rates between Examples 1 and 3 are expected due to the distinct problem setups. In Example 1, a small spatial domain $[-2, 2]$ with a nonzero source term was considered, whereas Example 3 involves a larger domain $[-20, 20]$ under homogeneous conditions ($\mathcal{G}(\chi, \rho) = 0$). The larger computational domain in Example 3 allows for smoother spatial variations, reducing the relative impact of spatial discretization on the overall error and enabling the temporal convergence to closely approach the theoretical order. Conversely, in Example 1, the presence of the source term and the smaller domain induces steeper gradients and

localized dynamics, which slightly diminish the observed temporal convergence rate. Moreover, the homogeneous nature of Example 3 promotes uniform and stable numerical behavior, allowing the method to achieve higher temporal accuracy ($Order1$ between 2.32 and 2.63) compared to Example 1 ($Order1$ between 2.0 and 2.42). Therefore, these differences reflect the natural influence of domain size, source term, and boundary conditions, rather than any deficiency in the numerical scheme. Figures 5 and 6 present the two-dimensional and three-dimensional plots of the numerical solution obtained by applying the method introduced in (28) to the problem described in Example 3. The results demonstrate a clear convergence of the numerical solution to the analytical solution, exhibiting excellent agreement in both spatial and temporal dimensions. These plots confirm that the proposed numerical scheme performs consistently with the theoretical predictions regarding accuracy, stability, and convergence. The visual evidence, combined with the quantitative analysis presented in Table 5, reinforces the reliability and robustness of the method for capturing the solution behavior with high fidelity across the computational domain.

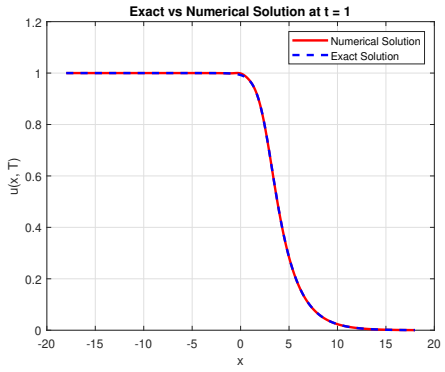


Figure 5: Convergence behavior of the numerical solution toward the exact solution for Example 3.

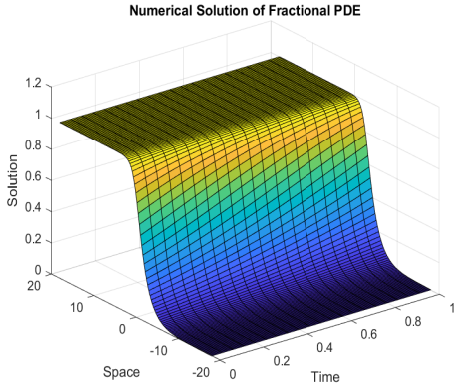


Figure 6: Numerical solution plot for Example 3.

6 Conclusions

This study presents a novel, rigorously analyzed, and computationally efficient framework for the numerical solution of nonhomogeneous time-fractional nonlinear parabolic equations defined on unbounded spatial domains. The core challenges of temporal nonlocality, inherent nonlinearity, and the infinite spatial extent were addressed through a fundamentally new approach. The

methodology's cornerstone is the derivation of highly accurate ABCs, constructed directly from the fundamental properties of fractional derivatives. This principled approach circumvents the limitations of existing techniques based on Padé approximations, ensuring a faithful representation of the underlying fractional dynamics at domain boundaries. By effectively truncating the infinite domain to a finite computational region while minimizing spurious reflections, these ABCs transform the original intractable problem into an equivalent, well-posed formulation on a bounded domain. The resulting bounded domain problem was discretized using a high-order finite difference scheme, facilitating stable and precise numerical implementation. A comprehensive theoretical analysis established the unconditional stability of the proposed method and provided rigorous error bounds, guaranteeing the reliability of the obtained solutions. Extensive numerical experiments unequivocally demonstrate the practical efficacy and robustness of the framework. The method achieves high accuracy in capturing the complex, memory-dependent behaviors characteristic of fractional-order systems, even in the presence of strong nonlinearities. Numerical results further confirm the superior accuracy and stability of the proposed method compared to existing approaches for unbounded domains, while maintaining a significantly reduced computational burden. Consequently, this work provides a powerful, reliable, and efficient computational tool for simulating anomalous diffusion and related complex phenomena governed by time-fractional dynamics in unbounded settings. The framework has direct relevance to diverse applications across physics, biology, and engineering. Moreover, it lays a solid foundation for future extensions to multi-dimensional scenarios, stochastic fractional PDEs, and problems involving variable coefficients, highlighting the potential for further development and application in complex fractional systems.

Future directions

Building upon the results of this study, several promising research directions can be explored. The proposed numerical framework can be extended to multi-dimensional time-fractional nonlinear parabolic equations defined on unbounded domains. Another important direction is the generalization of the ABCs to space-fractional and distributed-order models. Furthermore, the development of adaptive time-stepping and spatial refinement strategies could significantly improve computational efficiency for long-time simulations. Extending the present high-order scheme to coupled systems, stochastic FPDEs, and problems with variable coefficients also represents a meaningful avenue for future research. Finally, applying the proposed method to real-world

models arising in physics, biology, and engineering may further demonstrate its practical relevance and versatility.

Author declaration

The authors have no conflicts to disclose.

Funding

This research received no external funding.

Conflicts of Interest

The authors declare no conflicts of interest.

References

[1] Agrawal, O.P. *A general solution for the fourth-order fractional diffusion-wave equation*, Fract. Calc. Appl. Anal. 3(1) (2000), 1–12.

[2] Akrami, M.H., Poya, A. and Zirak, M.A. *Solving the general form of the fractional Black–Scholes with two assets through reconstruction variational iteration method*, Results. Appl. Math. 22 (2024), 100444.

[3] Cao, J., Li, C. and Chen, Y. *High-order approximation to Caputo derivatives and Caputo-type advection-diffusion equations (II)*, Fract. Calc. Appl. Anal. 18 (2015), 735–761.

[4] Engquist, B. and Majda, A. *Absorbing boundary conditions for the numerical simulation of waves*, Math. Comput. 31(139) (1977), 629–651.

[5] Gao, G., Sun, H. and Sun, Z.Z. *The finite difference approximation for a class of fractional sub-diffusion equations on a space unbounded domain*, J. Comput. Phys. 236 (2013), 443–460.

[6] Gao, G., Sun, H. and Zhang, Y.N. *A finite difference scheme for fractional sub-diffusion equations on an unbounded domain using artificial boundary conditions*, J. Comput. Phys. 231(7) (2012), 2865–2879.

[7] Gao, G.H., Sun, Z.Z. and Zhang, H.W. *A new fractional numerical differentiation formula to approximate the Caputo fractional derivative and its applications*, J. Comput. Phys. 259 (2014), 33–50.

[8] Han, H. and Huang, Z. *A class of artificial boundary conditions for heat equation in unbounded domains*, Comput. Math. Appl. 43(6-7) (2002), 889–900.

[9] Han, H., Sun, Z.Z. and Wu, X.N. *Convergence of a difference scheme for the heat equation in a long strip by artificial boundary conditions*, Numer. Methods Partial Differ. Equ. 24(1) (2008), 272–295.

[10] Hashemi, M.S., Mirzazadeh, M., Bayram, M. and El Din, S.M. *Numerical approximation of the Cauchy non-homogeneous time-fractional diffusion-wave equation with Caputo derivative using shifted Chebyshev polynomials*, Alex. Eng. J. 81 (2023), 118–129.

[11] Jiang, S., Zhang, J., Zhang, Q. and Zhang, Z. *Fast evaluation of the Caputo fractional derivative and its applications to fractional diffusion equations*, Commun. Comput. Phys. 21(3) (2017), 650–678.

[12] Kilbas, A.A. and Srivastava, H. M. and Trujillo, J. J. *Theory and applications of fractional differential equations*, Elsevier 204, 2006.

[13] Li, C., Wang, R., Du, H. and Pagnini, G. *High-order approximation to Caputo derivatives and Caputo-type advection-diffusion equations*, Commun. Appl. Ind. Math. 299 (2015), 159–175.

[14] Li, D. and Zhang, J. *Efficient implementation to numerically solve the nonlinear time fractional parabolic problems on unbounded spatial domain*, J. Comput. Phys. 322 (2016), 415–428.

[15] Magin, R. *Fractional calculus in bioengineering, part 1*, Crit. Rev. Biomed. Engin. 32(1) (2004), 1–34.

[16] Metzler, R. and Klafter, J. *Boundary value problems for fractional diffusion equations*, Phys. A: Stat. Mech. Appl. 278(1-2) (2000), 107–125.

[17] Nikan, O., Rashidinia, J. and Jafari, H. *Numerically pricing American and European options using a time fractional Black–Scholes model in financial decision-making*, Alex. Eng. J. 112 (2025), 235–245.

[18] Podlubny, I. *Fractional differential equations: An introduction to fractional derivatives, fractional differential equations, to methods of their solution and some of their applications*, Elsevier, 198, 1998.

[19] Povstenko, Y. *Signaling problem for time-fractional diffusion-wave equation in a half-plane*, Fract. Calc. Appl. Anal. 11(3) (2008), 331–349.

[20] Povstenko, Y. *Signaling problem for time-fractional diffusion-wave equation in a half-space in the case of angular symmetry*, Nonlinear Dyn. 59 (2010), 593–605.

[21] Schneider, W.R. and Wyss, W. *Fractional diffusion and wave equations*, J. Math. Phys. 30(1) (1989), 134–144.

[22] Senfiazad, P., Heydari, M.H., Bayram, M. and Baleanu, D. *A numerical method based on the shifted Jacobi polynomials for a class of tempered fractional quadratic integro-differential equations*, Results Appl. Math. 27 (2025), 100601.

[23] Sun, Z. Z. and Wu, X. *A fully discrete difference scheme for a diffusion-wave system*, Appl. Numer.Math. 56(2) (2006), 193–209.

[24] Wang, F., Ahmad, I., Ahmad, H., Alsulami, M.D., Alimgeer, K.S., Cesarano, C. and Nofal, T. A. *Meshless method based on RBFs for solving three-dimensional multi-term time fractional PDEs arising in engineering phenomenons*, J. King Saud Univ. Sci. 33(8) (2021), 101604.

[25] Wang, F., Khan, M.N., Ahmad, I., Ahmad, H., Abu-Zinadah, H. and Chu, Y.M. *Numerical solution of traveling waves in chemical kinetics: time-fractional fishers equations*, Fract. 30(02) (2022), 2240051.

[26] Wyss, W. *The fractional diffusion equation*, J. Math. Phys. 27(11) (1986), 2782–2785.

[27] Zhuang, P. and Liu, F. *Implicit difference approximation for the time fractional diffusion equation*, J. Appl. Math. Comput. 22 (2006), 87–99.

Iranian Journal of Numerical Analysis and Optimization
Vol. 16, No. 3, 2026, pp 991–1008
https://doi.org/10.22067/ijnao.2025.94912.1705
https://ijnao.um.ac.ir/



Research Article



A compact common-weight DEA model for efficiency measurement

A. Hadya K*, T. Radha Ramanan, M.C. Lineesh and C.M. Sushama

Abstract

This paper proposes a novel common-weight DEA model that combines multi-objective linear fractional programming and goal programming to determine a common set of weights for evaluating decision-making units (DMUs). The model formulates a single linear program without requiring iterative procedures, thereby enhancing computational efficiency. To validate the model's effectiveness, we compare the proposed method with a satisfaction degree-based model, a distance-based model, and a credibility-based common-weight method. Our model demonstrates competitive efficiency scores and a high degree of discrimination among DMUs, while maintaining a lower computational complexity. This research contributes by offering a practical and computationally efficient alternative to existing models.

*Corresponding author

Received 14 August 2025; revised 10 December 2025; accepted 28 December 2025

Aysha Hadya Kollathodi
Department of Mathematics, Farook College, Kozhikode, Kerala, India. e-mail: ayshahadya@farookcollege-.ac.in

Thiyagarajan Radha Ramanan
Department of Mechanical Engineering, National Institute of Technology Calicut, Kerala, India.

Meppadi Chalil Lineesh
Department of Mathematics, National Institute of Technology Calicut, Kerala, India

Chalil Madathil Sushama
Department of Mathematics, National Institute of Technology Calicut, Kerala, India

How to cite this article
Hadya K, A., Radha Ramanan, T., Lineesh, M.C. and Sushama, C.M., A compact common-weight DEA model for efficiency measurement. *Iran. J. Numer. Anal. Optim.*, 2026; 16(3): 991-1008. https://doi.org/10.22067/ijnao.2025.94912.1705

AMS subject classifications (2020): 90C08, 90C29, 90C90

Keywords: Common-weights, Data envelopment analysis, Efficiency evaluation, Multi-objective linear fractional programming, Goal programming

1 Introduction

Charnes, Cooper, and Rhodes [3] initially suggested the data envelopment analysis (DEA) technique for determining the effectiveness of decision-making units (DMUs). These DMUs can be anything, like companies, schools, hospitals, or even government organizations, that use multiple inputs (resources) to produce multiple outputs (results). Traditional DEA models evaluate the efficiency of DMUs that use multiple inputs (resources) to produce multiple outputs (results). Conventional DEA models, like the CCR and BCC models, let each DMU select its own input and output weights.

This implies that each unit can maximize its efficiency score by modifying the relative importance of its resources and outputs. However, because of this flexibility, many DMUs may be classified as efficient, particularly when there are fewer units than inputs and outputs. This reduces the model's ability to clearly distinguish between the best and worst performers, making it difficult to rank or compare units properly.

To solve this problem, DEA models based on common-weights were developed, assuming a single set of weights for all DMUs, so each unit is measured against the same standard. The idea behind using common-weights is to make the evaluation fairer and to improve the comparability between different units. The concept of using common-weights to improve fairness in DEA was first suggested by Roll, Cook, and Golany [17], who proposed a method for measuring each unit's effectiveness using a standard set of weights. This laid the groundwork for further research into how to better compare units using common-weights.

One of the primary advantages of using DEA models with common-weights is that they allow for better ranking of all DMUs, including both the efficient and inefficient ones. In traditional DEA, many units might be considered efficient, even if they are not truly performing at their best. Common-weight models, however, help provide a clear ranking based on a uniform set of criteria, which is especially useful for decision-making and resource allocation. Charnes et al. [2] and Golany and Roll [5] emphasized how using common-weights can make the evaluation process fairer and more objective, especially in areas where consistent and transparent comparisons are important.

Over the years, many researchers have developed different methods for calculating these common-weights. One approach involves using multi-objective optimization, where several goals (such as minimizing input excess and maximizing output efficiency) are considered at the same time. For example, Kao and Hung [7] proposed using goal programming to find the best common-weights by considering multiple efficiency-related goals. Wang, Greatbanks, and Yang [20] and Wu et al. [21] explored how integrating cross-efficiency evaluations into the common-weight model could help reduce bias in the results and provide more reliable rankings. Cross-efficiency, introduced by Seiford and Zhu [18], helps to avoid giving too much weight to units that perform well only when evaluated using their own customized weights.

In addition to mathematical improvements, many researchers have also focused on creating hybrid methods that combine DEA with other techniques like multi-objective linear programming (MOLP) or mixed-integer programming. These methods help improve the accuracy of efficiency evaluations by adding constraints or preferences according to the objectives of the decision-maker. For example, there are studies in which hybrid methods which can be used in industries like banking and healthcare to better evaluate the performance of various institutions [23, 13]. common-weight and weight restricted DEA models also enhance discrimination and comparability of efficiency scores. Applications include internet banking efficiency [11], eco-efficiency of urban waste services [14], dynamic performance evaluation of airlines [24], and climate competitiveness assessment of EU regions using entropy-based common-weights [9]. These works highlight the wide applicability and effectiveness of common-weight DEA frameworks. In recent years, some researchers have even incorporated behavioral theories, like prospect theory, into DEA models. Yu et al. [25], for instance, used prospect theory to account for how people make decisions under risk and uncertainty, leading to more realistic evaluations.

Recent studies have introduced innovative approaches to common-weight DEA, addressing various challenges in performance evaluation. In 2022, Omrani, Fahimi, and Emrouznejad [15] introduced a common-weight credibility DEA model that uses credibility theory to handle uncertainty in inputs and outputs. This model gives a strong evaluation method for industries like aviation, where data can vary a lot. In 2023, Majidi, Ebrahimnejad, and Azizi [12] suggested a fuzzy DEA model with common-weights that looks at both optimistic and pessimistic views when dealing with undesirable outputs. By using fuzzy arithmetic, this approach cuts down on computation time and offers a fairer assessment of DMUs under uncertain data. Most recently, in 2024, Kim and co-authors [10] presented a cluster-analysis-based common-weight DEA model. They group DMUs by their production patterns and then assign one weight vector to all units in each group, which gives more accurate and comparable efficiency measures in complex industries like iron and steel manufacturing. In addition to fuzzy and credibility-based approaches, robust DEA models have also been proposed to explicitly handle data uncertainty using robust optimization

techniques [16, 1]. Together, these studies show a growing focus on making common-weight DEA models more robust, fair, and useful in situations where data are difficult or varied.

The existing models, especially those using iterative procedures, take a lot of time to compute — particularly when there are a lot of DMUs, inputs, and outputs [22]. These models solve the problem several times in different rounds to get the final result. This increases the overall complexity and makes them less suitable for large datasets. Also, many models give high efficiency scores to a large number of DMUs, which reduces the ability to clearly identify top performers [4, 19]. To overcome these issues, we have developed a new DEA model that works without any iteration. It solves the problem in a single step, which saves time and reduces computational effort. At the same time, it gives similar efficiency results when compared to existing models. But unlike them, our model identifies fewer DMUs as fully efficient. This shows that it is stricter and better at distinguishing between high and average performers. We also found that our model gives a more realistic satisfaction degree, which is useful when studying behavior or preferences of the DMUs. Overall, the proposed model is simpler, faster, and more practical, especially for large-scale problems. It helps in better decision-making and provides a strong alternative to more complex and time-consuming models.

2 Methodology

Assume that n DMUs with m inputs and s outputs need to be assessed. To determine a set of common-weights for inputs and outputs consider the following multi-objective linear fractional programming (MOLFP):

$$Max \quad \left(\frac{\sum_{r=1}^{s} u_r y_{r1}}{\sum_{i=1}^{m} v_i x_{i1}}, \frac{\sum_{r=1}^{s} u_r y_{r2}}{\sum_{i=1}^{m} v_i x_{i2}}, \ \ldots, \ \frac{\sum_{r=1}^{s} u_r y_{rn}}{\sum_{i=1}^{m} v_i x_{in}}\right), \tag{1}$$

$$s.t. \quad \sum_{r=1}^{s} u_r y_{rj} - \sum_{i=1}^{m} v_i x_{ij} \leq 0, \quad j = 1, \ldots, n, \tag{2}$$

$$u_r \geq 0, \quad r = 1, \ldots, s, \tag{3}$$

$$v_i \geq 0, \quad i = 1, \ldots, m. \tag{4}$$

After solving the CCR model for each DMU_j [3], we have

$$Max \quad \sum_{r=1}^{s} u_r y_{rj}, \tag{5}$$

$$s.t \quad \sum_{i=1}^{m} v_i x_{ij} = 1, \tag{6}$$

$$\sum_{r=1}^{s} u_r y_{rj} - \sum_{i=1}^{m} v_i x_{ij} \leq 0, \quad j = 1, \ldots, n, \tag{7}$$

$$u_r \geq 0, \quad r = 1, \ldots, s, \tag{8}$$

$$v_i \geq 0, \quad i = 1, \ldots, m. \tag{9}$$

The efficiency of each DMU_j can be evaluated as

$$\theta_j = \frac{\sum_{r=1}^{s} u_r y_{rj}}{\sum_{i=1}^{m} v_i x_{ij}}. \tag{10}$$

Now, DMUs can be divided into two groups: $E = \{j : \theta_j = 1\}$ and $I = \{\, j : \theta_j < 1 \,\}$.
Let $\theta^+ = \max\limits_{j \in I} \theta_j$ and $\theta^- = \min\limits_{j \in I} \theta_j$.

Let N_j and D_j be the numerator and denominator of the jth objective function. The optimum values of the numerator and denominator can be readily derived as 1 for CCR efficient units ($j \in E$). However, the efficiency ratings of CCR efficient units should not be lower than those of the most efficient CCR inefficient unit, which is θ^+.

Based on these conditions, the membership functions for CCR efficient units are defined as follows [6]:

$$CN_i^E(x) = \begin{cases} \dfrac{N_i(x) - \theta^+}{1 - \theta^+}, & \theta^+ \leq N_i(x) \leq 1, \\ 0, & \text{otherwise}, \end{cases} \tag{11}$$

$$CD_i^E(x) = \begin{cases} \dfrac{\theta^+ - D_i(x)}{\theta^+ - 1}, & 1 \leq D_i(x) \leq \theta^+, \\ 0, & \text{otherwise}. \end{cases} \tag{12}$$

Similarly, for CCR inefficient units ($j \in I$) the ideal values of numerator and denominator are θ^+ and $\frac{1}{\theta^+}$, respectively, and the efficiency scores should not be lower than the lowest-performing CCR inefficient unit, which is θ^-. Therefore, the membership functions for CCR inefficient units will be as follows [6]:

$$C_I^{N_i}(x) = \begin{cases} \dfrac{N_i(x) - \theta^-}{\theta^+ - \theta^-}, & \theta^- \leq N_i(x) \leq \theta^+, \\ 0, & \text{otherwise}, \end{cases} \tag{13}$$

$$C_I^{D_i}(x) = \begin{cases} \frac{\frac{1}{\theta^-} - D_i(x)}{\frac{1}{\theta^-} - \frac{1}{\theta^+}}, & \frac{1}{\theta^-} \geq D_i(x) \geq \frac{1}{\theta^+}, \\ 0, & \text{otherwise.} \end{cases} \tag{14}$$

Development of the common-weight model

Now we formulate a goal programming model by taking 1 as the aspiration level for each membership function (11), (12), (13) and (14). So the goals can be formulated as

$$C_E^{N_i}(x) = 1 \quad \Rightarrow \quad \frac{N_i(x) - \theta^+}{1 - \theta^+} = 1, \tag{15}$$

$$C_E^{D_i}(x) = 1 \quad \Rightarrow \quad \frac{\frac{1}{\theta^+} - D_i(x)}{\frac{1}{\theta^+} - 1} = 1, \tag{16}$$

$$C_I^{N_i}(x) = 1 \quad \Rightarrow \quad \frac{N_i(x) - \theta^-}{\theta^+ - \theta^-} = 1, \tag{17}$$

$$C_I^{D_i}(x) = 1 \quad \Rightarrow \quad \frac{\frac{1}{\theta^-} - D_i(x)}{\frac{1}{\theta^-} - \frac{1}{\theta^+}} = 1. \tag{18}$$

Simplifying and substituting numerator and denominator functions in (15), (16), (17) and (18), we get

$$\sum_{r=1}^{s} u_r y_{rj} = 1, \qquad j \in E, \tag{19}$$

$$-\sum_{i=1}^{m} v_i x_{ij} = -1, \qquad j \in E, \tag{20}$$

$$\sum_{r=1}^{s} u_r y_{rj} = \theta^+, \qquad j \in I, \tag{21}$$

$$-\sum_{i=1}^{m} v_i x_{ij} = -\frac{1}{\theta^+}, \qquad j \in I. \tag{22}$$

By introducing deviational variables d_j^+ and d_j^- for $j = 1, 2, \ldots, n$ and $i = 1, 2$, we can formulate the goal programming model as

$$
\begin{aligned}
\text{Minimize} \quad & \sum_{i=1}^{2}\sum_{j=1}^{n}\left(d_{ij}^{-}+d_{ij}^{+}\right) \\
\text{subject to} \quad & \sum_{r=1}^{s} u_r y_{rj} + d_{1j}^{-} - d_{1j}^{+} = 1, && \text{for all } j \in E, \\
& -\sum_{i=1}^{m} v_i x_{ij} + d_{2j}^{-} - d_{2j}^{+} = -1, && \text{for all } j \in E, \\
& \sum_{r=1}^{s} u_r y_{rj} + d_{1j}^{-} - d_{1j}^{+} = \theta^{+}, && \text{for all } j \in I, \\
& -\sum_{i=1}^{m} v_i x_{ij} + d_{2j}^{-} - d_{2j}^{+} = -\frac{1}{\theta^{+}}, && \text{for all } j \in I, \\
& \sum_{r=1}^{s} u_r y_{rj} - \sum_{i=1}^{m} v_i x_{ij} \le 0, && j = 1,\ldots,n, \\
& d_{ij}^{-} \ge 0, \quad d_{ij}^{+} \ge 0, && i = 1,2;\ j = 1,\ldots,n \\
& u_r \ge 0, && r = 1,\ldots,s, \\
& v_i \ge 0, && i = 1,\ldots,m.
\end{aligned}
\tag{23}
$$

From the goal programming model (23) we can obtain a common set of weights u_r for outputs and v_i $(i = 1, 2, \ldots, m)$ for inputs. These common-weights allow us to compare and rank the DMUs on a common basis.

3 Theoretical properties of the proposed model

Before analyzing the performance of the proposed goal programming model, we establish the feasibility and boundedness of the model to ensure that optimal solutions can be obtained for the proposed model.

Theorem 1. The common-weight model (23) is feasible.

Proof. From the output-oriented CCR model, for a CCR-efficient DMU j, the optimal weights satisfy $\sum u_r y_{rj} = 1$ and $\sum v_i x_{ij} = 1$. These equalities match the goal targets for the efficient unit exactly. Consequently, both deviational variables d_j^+ and d_j^-, as well as d_i^+ and d_i^-, can be set to zero, ensuring feasibility for DMU j without any penalty.

For all other DMUs k, the same weights u_r, v_i may not lead to exact satisfaction of the corresponding equality constraints in the goal programming model. However, the model includes nonnegative deviational variables specifically to allow for flexibility in these cases. By assigning

appropriate positive values to these variables, the equalities can always be satisfied, regardless of how far the weighted input-output ratios are from the aspiration levels of 1 or other bounds. Furthermore, the inequality constraint $\sum v_i x_{ik} \leq 1$ is guaranteed to hold for all k because it is a defining constraint of the CCR model.

Lastly, all nonnegativity constraints for the weights and deviational variables are respected by construction: the CCR model ensures nonnegative weights, and the deviational variables are explicitly constrained to be nonnegative in the goal programming formulation. □

Theorem 2. The optimal value of model (23) is bounded below, and the weights u_r and v_i are bounded above under the assumption that every input and output value is greater than zero.

Proof. Since the model minimizes the sum of nonnegative deviational variables and each variable in the model is restricted to be greater than or equal to zero, the objective function is clearly bounded below by zero. Now assume, for contradiction, that the weights u_r or v_i are unbounded above. Then, due to the positive input and output data, the expressions $\sum u_r y_{rj}$ or $\sum v_i x_{ij}$ could become arbitrarily large, violating the equality and inequality constraints of the model — particularly, those requiring sums to equal fixed finite values like 1. Thus, any increase in weights must be offset by changes in deviation variables, which increases the objective function value, preventing unbounded solutions from being optimal. Therefore, the optimal value of the model is bounded. □

4 Numerical example

To illustrate the applicability and effectiveness of the developed common-weight DEA model (23) based on MOLFP and goal programming, we consider a well-known dataset from Kao and Hung [7], originally from Kao and Yang [8], which assesses the performance of 17 forest districts across Taiwan. Each forest district is defined by four input variables, budget (x_1), initial stocking (x_2), labor (x_3), and land (x_4) and three output variables, main product (y_1), soil conservation (y_2), and recreation (y_3). The data for the 17 forest districts is given in Table 1 below:

Table 2 shows the common-weights given to each input and output used in our model. These weights represent how important each input (like budget, labor, land) and output (like main product, soil conservation, recreation) is when calculating the efficiency of the different forest districts. A special feature of our model is that it gives a nonzero weight to every input and output. This is different from many other models using common-weights, where sometimes some weights become zero. When a weight is zero, it means that particular input or output is ignored completely in the analysis, which can give an incomplete or unfair result.

Table 1: Input and Output Data for 17 Forest Districts

DMU	X_1 **(dollars)**	X_2 **(m^3)**	X_3 **(persons)**	X_4 **(ha)**	Y_1 **(m^3)**	Y_2 **(m^3)**	Y_3 **(visits)**
1	51.62	11.23	49.22	33.52	40.49	14.89	166.71
2	85.78	123.98	55.13	108.46	43.51	173.93	6.45
3	66.65	104.18	257.09	13.65	139.74	115.96	0.00
4	27.87	107.60	14.00	146.43	25.47	131.79	0.00
5	51.28	117.51	32.07	84.50	46.20	144.99	0.00
6	36.05	193.32	59.52	8.23	46.88	190.77	822.92
7	25.83	105.80	9.51	227.20	19.40	120.09	0.00
8	123.02	82.44	87.35	98.80	43.33	125.84	404.69
9	61.95	99.77	33.00	86.37	45.43	79.60	52.62
10	80.33	104.65	53.30	79.06	27.28	132.49	42.67
11	205.92	183.49	144.16	59.66	14.09	196.29	16.15
12	82.09	104.94	46.51	127.28	44.87	108.53	0.00
13	202.21	187.74	149.39	93.65	44.97	184.77	0.00
14	67.55	82.83	44.37	60.85	26.04	85.00	23.95
15	72.60	132.73	44.67	173.48	5.55	135.65	24.13
16	84.83	104.28	159.12	171.11	11.53	110.22	49.09
17	71.77	88.16	69.19	123.14	44.83	74.54	6.14

By making sure that every factor has some weight, our model ensures that all aspects of performance are considered. This makes the evaluation more balanced and realistic. The weights assigned by the proposed model highlight the relative importance of each factor in shaping overall efficiency. For example, V_2 (Initial stocking) receives a higher weight of 0.005975, showing that this input strongly influences the efficiency of the 17 forest districts and plays a key role in distinguishing performance among them. In contrast, V_3 (Labor input) has a much smaller weight of 0.000832, suggesting that variations in labor have little impact on overall efficiency in this dataset. These differences illustrate the model's ability to identify which inputs and outputs most significantly affect efficiency, offering useful insight for decision-makers about where resources or focus can have the greatest effect. By looking at these weight patterns, practitioners can gain clearer insight into the practical implications of the evaluation and see which factors most strongly influence performance.

Table 2: Weights of inputs and outputs using proposed model

Weight	U_1	U_2	U_3	V_1	V_2	V_3	V_4
Value	0.002058	0.006144	0.000032	0.002091	0.005975	0.000832	0.001772

4.1 Comparison with satisfaction degree model

To evaluate the performance of the proposed common-weight model (23), in Table 3, we compare the efficiency scores and satisfaction degrees with those obtained using the iterative satisfaction degree-based approach [22].

Table 3: Efficiency scores comparison

DMU	Proposed Model	Rank	Satisfaction Degree Model	Rank
1	0.9793	5	1.0000	1
2	1.0000	1	0.9999	3
3	0.9997	3	0.6979	12
4	0.8867	8	0.9583	4
5	1.0000	1	1.0000	1
6	0.9898	4	0.9004	5
7	0.7093	16	0.8197	8
8	0.8743	9	0.8609	7
9	0.6711	17	0.7099	11
10	0.9479	7	0.8764	6
11	0.7543	15	0.6524	14
12	0.8614	10	0.7388	9
13	0.8368	13	0.6258	15
14	0.9547	6	0.7233	10
15	0.8598	11	0.6888	13
16	0.7640	14	0.5194	17
17	0.8405	12	0.5736	16

From the table, it is observed that the proposed model identifies two DMUs (DMU 2 and DMU 5) as fully efficient with scores slightly above or equal to 1.000, in close agreement with the satisfaction degree-based model. For the majority of the DMUs, our model delivers efficiency scores that are either comparable or significantly higher. For instance, DMU 3 receives a high score of 0.9997 from our model, while it was evaluated at only 0.6979 in the satisfaction model. Similarly, DMUs 13, 16, and 17 also see considerable improvement.

Examining Table 3 shows the effect of near-efficient DMUs under the proposed model. DMU 3 (0.9997), DMU 1 (0.9793), and DMU 6 (0.9898) are all close to the frontier, yet their ranks shift slightly compared to other methods. These small differences, typically within a few positions, reflect the model's enhanced discrimination among nearly identical units. Importantly, this effect influences ranking rather than the absolute efficiency classification, indicating that the magnitude is modest and does not compromise evaluation reliability.

To further evaluate the alignment between the models, satisfaction degrees derived from each are also compared in Table 4. The satisfaction degree reflects how closely a DMU's efficiency aligns with the ideal value (typically 1), interpreted through a satisfaction function [22].

Table 4: Satisfaction degree comparison

DMU	Proposed Model	Satisfaction Model
1	0.9793	1.0000
2	1.0000	0.9999
3	0.9997	0.6979
4	0.8867	0.9583
5	1.0000	1.0000
6	0.9898	0.8989
7	0.7093	0.8197
8	0.8743	0.8584
9	0.6711	0.6979
10	0.9479	0.9320
11	0.7543	0.6979
12	0.8614	0.8912
13	0.8368	0.7826
14	0.9547	0.9353
15	0.8598	0.9030
16	0.7640	0.6979
17	0.8405	0.8346

To statistically assess the difference in satisfaction degrees between our proposed model and the existing satisfaction degree-based model, the Wilcoxon signed-rank test was used. Pairs of ordinal or nonnormally distributed data can be compared using this nonparametric test.

The test yielded a test statistic $W = 89.000$ with a corresponding z-value of 0.592 and p-value = 0.579. Because the p-value is greater than the standard 0.05 threshold, we do not reject the null hypothesis, suggesting that there is insufficient statistical evidence to support a difference in satisfaction levels between the two models.

We compare the time complexity of our proposed model with that of the satisfaction degree-based model from previous studies. Both models aim to solve a linear programming (LP) problem, but they differ in their structure and the number of iterations required to achieve a satisfactory result.

Our model is formulated with a total of $4n + m + s$ variables and $7n + m + s$ constraints. It is solved as a single LP problem, meaning all the variables and constraints are processed simultaneously without the need for iterative procedures. This characteristic makes our model more computationally efficient, especially as the number of DMUs increases.

In contrast, the satisfaction degree-based model proposed in previous studies involves $3n + m + s + 1$ variables and $6n + m + s + 1$ constraints per iteration. A key difference is that this model requires multiple iterations to achieve a high level of precision. According to the authors of the satisfaction model, 14 iterations are required to reduce the approximation error to less than 0.0001. While this iterative process allows the model to fine-tune its results, it also increases computational time, making it less efficient for large-scale problems.

To compare the computational complexity of the two models, we analyze the overall time complexity involved in solving an LP problem. The computational effort required by algorithms such as the interior-point method or ellipsoid method is approximately

$$O\big((n+m)^3\big),$$

where n is the number of variables and m is the number of constraints. This cubic complexity arises due to the inherent structure of the LP problem and the algorithms used to solve it.

Assuming we denote by $v = m + s$ the total count of input and output variables, the computational load of each model can be expressed in closed form. Our proposed approach requires solving a single LP with $4n + v$ variables and $7n + v$ constraints, yielding a worst-case time complexity of

$$T_{\text{proposed}} = O\big((4n + v + 7n + v)^3\big) = O\big((11n + 2v)^3\big).$$

By contrast, the satisfaction degree model must solve $t = 14$ iterations of an LP that, in each pass, has approximately $3n + v$ variables and $6n + v$ constraints. Its per-iteration cost is

$$O\big((3n + v + 6n + v)^3\big) = O\big((9n + 2v)^3\big),$$

so that cumulatively:

$$T_{\text{iterative}} = 14 \times O\big((9n + 2v)^3\big).$$

Since $(11n + 2v)^3 < 14\,(9n + 2v)^3$ for all $n, v > 0$, our single-LP formulation always requires less total computation than 14 iterations of the satisfaction degree model. Indeed, one finds that if the iterative scheme needed as few as follows:

$$t > \left(\frac{11n + 2v}{9n + 2v}\right)^3 \approx 1.82$$

iterations (i.e., $t \geq 2$), the cumulative cost of the iterative model would exceed that of our method even after just two iterations.

This highlights the clear efficiency advantage of our proposed model, especially for medium- to large-scale DEA applications. In conclusion, our model provides a clear advantage by being faster and simpler to compute, while still delivering similar satisfaction levels as the satisfaction-based model. The satisfaction-based model requires multiple iterations to reach the desired accuracy, whereas our model achieves the same results in just one step. This makes our model more efficient and practical, especially for larger or time-sensitive problems. By reducing the complexity without sacrificing performance, our approach offers a more efficient solution for real-world applications.

4.2 Comparison with distance-based model

In this section, we further compare the proposed model with the results obtained from two alternative models found in the literature [19]. The models were developed to evaluate the efficiency of the same set of DMUs as in our study. Table 5 summarizes the efficiency scores obtained from both models and their respective rankings.

Table 5: Efficiency score comparison with distance-based model

DMU	Model 1	Model 2	Proposed Model
1	1.000	1.000	0.9793
2	1.000	1.000	1.0000
3	0.925	1.000	0.9997
4	1.000	1.000	0.8867
5	1.000	1.000	1.0000
6	1.000	0.966	0.9898
7	0.863	0.874	0.7093
8	0.834	0.847	0.8743
9	0.676	0.678	0.6711
10	0.883	0.878	0.9479
11	0.668	0.653	0.7543
12	0.702	0.717	0.8614
13	0.625	0.623	0.8368
14	0.710	0.712	0.9547
15	0.732	0.722	0.8598
16	0.666	0.670	0.7640
17	0.567	0.592	0.8405

In our model, only the best-performing DMUs are marked as efficient. Specifically, DMUs 2 and 5 are considered the most efficient. On the other hand, in the models used for comparison, a larger number of DMUs are identified as efficient, with five DMUs being marked as fully efficient.

While having more efficient DMUs might seem beneficial, it can create confusion and make it harder to clearly distinguish between the best-performing units and the others. By limiting the efficient DMUs to just those that truly perform well, our model focuses on the top performers, which helps to identify where excellence lies.

This method has the advantage of being clearer, as it avoids labeling too many units as efficient, which could reduce the importance of those that genuinely stand out. When fewer DMUs are marked as efficient, it makes it easier to focus on areas that need improvement and allows others to learn from the best. In simple terms, our model is more selective, making the results more meaningful and useful for decision-making. It highlights the best performers, making the analysis clearer and more actionable for anyone looking to improve efficiency.

4.3 Comparison with credibility-based common-weight model

In recent literature, many common-weight DEA models are developed for decision-making problems with uncertain data. Therefore, along with the comparison with satisfaction degree-based and distance-based models, we also compare the proposed method with a recent credibility-based common-weight DEA model, namely, Omrani, Fahimi, and Emrouznejad [15].

Although the proposed model is formulated for crisp data, this comparison is included to examine how the proposed compact common-weight structure performs when incorporated into a recent uncertainty-based framework. For this purpose, all credibility calculations and uncertainty modeling steps of [15] are retained, and only the final common-weight model is replaced by the proposed common-weight formulation.

For comparison, we use the numerical example of Omrani, Fahimi, and Emrouznejad [15], which evaluates the efficiency of 14 Iranian airlines under fuzzy conditions. The results of the comparison between Omrani's common-weight model (for $\lambda = 0.8$) and the proposed compact CSW model are given in Table 6.

In the results of Omrani et al., the efficiency score equal to 1 is assigned to more than one DMU, leading to repetition of fully efficient units. This limits the ability of the model to clearly distinguish the best-performing DMUs. In contrast, the proposed model avoids repetition of efficiency value 1 and produces a more spread-out set of efficiency scores. Omrani et al.'s model also shows repetition in the rankings. In particular, DMUs 2, 9, and 14 share the same rank (Rank 4), indicating insufficient discrimination. The proposed model clearly differentiates these DMUs by assigning distinct efficiency scores and ranks.

Table 6: Efficiency scores and ranks comparison with Omrani et al. model

DMU	Omrani et al. Efficiency	Omrani et al. Rank	Proposed Model Efficiency	Proposed Model Rank
1	0.866	11	0.8689	10
2	0.946	4	0.9285	8
3	0.938	7	0.9127	9
4	0.849	12	0.9448	6
5	0.967	3	1.0000	1
6	1.000	1	0.9522	4
7	0.769	14	0.9678	2
8	1.000	1	0.9461	5
9	0.946	4	0.7439	13
10	0.874	10	0.8689	11
11	0.906	9	0.9427	7
12	0.838	13	0.6817	14
13	0.946	4	0.8501	12
14	0.941	8	0.9583	3

Overall, the results in Table 6 indicate that replacing the final common-weight stage of Omrani et al.'s method with the proposed compact common-weight model leads to better discrimination without altering the underlying credibility-based efficiency evaluation framework.

5 Conclusion

This study introduced a noniterative common-weight DEA model capable of generating efficiency scores in a single computational step. By avoiding iterative procedures, the model enhances computational efficiency, especially when working with large number of inputs, outputs, and DMUs. Additionally, by detecting fewer perfectly efficient units, it increases the analysis's discriminatory power, thus enabling a more precise ranking of performance.

The results produced by the proposed model are comparable to those of existing methods but require less computational time and effort, making it a practical choice for real-world applications where resources are limited.

Although some DMUs achieve scores very close to one, the uniform evaluation framework of the proposed model highlights small differences among these near-efficient units, resulting in small shifts in their ranking. This behavior reflects the model's ability to provide finer discrimination. At the same time, the study has been conducted on selected datasets, and future research could examine the model's performance across other sectors and contexts. While the present study

shows that the proposed model can be used within uncertainty-based DEA frameworks at the common-weight stage, a full extension to directly handle fuzzy data, as well as the inclusion of weight restrictions and decision-maker preferences, can be taken up in future research.

Overall, the proposed model offers a fast, straightforward, and effective alternative for common-weight DEA, contributing to improved performance evaluation and supporting future methodological developments in the field.

Funding

This research received no external funding.

Conflicts of Interest

The authors declare no conflicts of interest.

References

[1] Arabmaldar, A., Hatami-Marbini, A., Loske, D., Hammerschmidt, M. and Klumpp, M. *Robust data envelopment analysis with variable budgeted uncertainty*, Eur. J. Oper. Res. 315 (2024), 626–641. DOI: 10.1016/j.ejor.2023.11.043

[2] Charnes, A., Cooper, W.W., Huang, Z.M. and Sun, D.B. *Polyhedral cone-ratio DEA models with an illustrative application to large commercial banks*, J. Econometrics 46 (1990), 73–91. DOI: 10.1016/0304-4076(90)90048-X

[3] Charnes, A., Cooper, W.W. and Rhodes, E. *Measuring the efficiency of decision making units*, Eur. J. Oper. Res. 2 (1978), 429–444. DOI: 10.1016/0377-2217(78)90138-8

[4] Contreras, I. *A review of the literature on DEA models under common set of weights*, J. Model. Manag. 15 (2020), 1277–1300. DOI: 10.1108/JM2-02-2019-0043

[5] Golany, B. and Roll, Y. *An application procedure for DEA*, Omega 17 (1989), 237–250. DOI: 10.1016/0305-0483(89)90029-7

[6] Hajiagha, S.H.R., Hashemi, S.S. and H. Amoozad Mahdiraji. *DEA with common set of weights based on a multi objective fractional programming problem*, Int. J. Ind. Eng. Prod. Res., 25 (2014), 207–214.

[7] Kao, C. and Hung, H.T. *Data envelopment analysis with common-weights: the compromise solution approach*, J. Oper. Res. Soc. 56 (2005), 1196–1203. DOI: 10.1057/palgrave.jors.2601924

[8] Kao, C. and Yang, Y. C. *Reorganization of forest districts via efficiency measurement*, Eur. J. Oper. Res. 58 (1992), 356–362. DOI: 10.1016/0377-2217(92)90066-I

[9] Karman, A. and Banaś, J. *The evaluation of climate change competitiveness via DEA models and Shannon entropy: EU regions*, Entropy 26 (2024), 732. DOI: 10.3390/e26090732

[10] Kim, N. H., He, F., Hong, K. R., Kim, H. C. and Han, S. M. *A new common weights DEA model based on cluster analysis*, Oper. Res. 24 (2024), 29. DOI: 10.1007/s12351-024-00838-5

[11] Kraidi, A.A., Daneshvar, S. and Adesina, K.A. *Weight-restricted approach on constant returns to scale DEA models: Efficiency of internet banking in Turkey*, Heliyon 10 (2024), e31008. DOI: 10.1016/j.heliyon.2024.e31008

[12] Majdi, M., Ebrahimnejad, A. and Azizi, A. *Common-weights fuzzy DEA model in the presence of undesirable outputs with ideal and anti-ideal points: development and prospects*, Complex Intell. Syst. 9 (2023), 6223–6240. DOI: 10.1007/s40747-023-01030-6

[13] Mirmozaffari, Mirpouya and Shadkam, Elham and Khalili, Seyed Mohammad and Yazdani, Maziar. *Developing a novel integrated generalised data envelopment analysis (DEA) to evaluate hospitals providing stroke care services*, Bioengineering 8 (2021), 207. DOI: 10.3390/bioengineering8120207

[14] Molinos-Senante, M., Maziotis, A. and Sala-Garrido, R. *Estimating the eco-efficiency of urban waste services towards sustainable waste management*, Sustain. Dev. 32 (2024), 5677–5691. DOI: 10.1002/sd.2983

[15] Omrani, H., Fahimi, P. and Emrouznejad, A. *A common weight credibility data envelopment analysis model for evaluating decision making units with an application in airline performance*, RAIRO Oper. Res. 56 (2022), 911–930. DOI: 10.1051/ro/2022031

[16] Peykani, P., Nouri, M., Shojaie, S.E., Ghiyaasi, F. and Esmaeli, A. *Efficiency analysis of intellectual capital under deep uncertainty: A robust DEA approach*, in International

Conference on Data Envelopment Analysis, Springer Nature Switzerland (2024), pp. 27–41. DOI: 10.1007/978-3-031-98177-7_3

[17] Roll, Y., Cook, W.D. and Golany, B. *Controlling factor weights in DEA*, IIE Trans. 23 (1991), 2–9. DOI: 10.1080/07408179108963835

[18] Seiford, L.M. and Zhu, J. *Profitability and marketability of the top 55 U.S. commercial banks*, Management Sci. 45 (1999), 1270–1288. DOI: 10.1287/mnsc.45.9.1270

[19] Shahghobadi, S., Ghomashi, A. and Moradi, F. *Non-dominated DEA cross efficiency scores; a secondary goal approach*, Iran. J. Optim. 2 (2022), 101.

[20] Wang, Y.M., Greatbanks, R. and Yang, J.B. *Interval efficiency assessment using DEA*, Fuzzy Sets Syst. 153 (2005), 347–370. DOI: 10.1016/j.fss.2004.12.011

[21] Wu, J., Chu, J., Sun, J. and Zhu, Q. *DEA cross-efficiency evaluation based on Pareto improvement*, Eur. J. Oper. Res. 248 (2016), 571–579. DOI: 10.1016/j.ejor.2015.07.042

[22] Wu, J., Chu, J., Zhu, Q., Li, Y. and Liang, L. *Determining common weights in data envelopment analysis based on the satisfaction degree*, J. Oper. Res. Soc. 67 (2016), 1446–1458. DOI: 10.1057/jors.2016.35

[23] Yang, J. B., Wong, B. Y. H., Xu, D. L., Liu, X. B. and Steuer, R. E. *Integrated bank performance assessment and management planning using hybrid minimax reference point–DEA approach*, Eur. J. Oper. Res. 207 (2010), 1506–1518. DOI: 10.1016/j.ejor.2010.07.001

[24] Yekta, A. P., Noveiri, M. J. S., Maghbouli, M. and Edalatpanah, S. A. *Dynamic DEA with common weights: A case study of Iranian airlines*, J. Uncertain Syst. 17 (2024), 2440001. DOI: 10.1142/S1752890924400014

[25] Yu, Y., Zhu, W., Shi, Q. and Zhuang, S. *Common set of weights in data envelopment analysis under prospect theory*, Expert Syst. 38(2021), e12602. DOI: 10.1111/exsy.12602

Iranian Journal of Numerical Analysis and Optimization
Vol. 16, No. 3, 2026, pp 1009–1030
https://doi.org/10.22067/ijnao.2026.93855.1660
https://ijnao.um.ac.ir/

CC BY NC

Research Article

Barycentric rational interpolation method for fractional-order differential equation model of HIV infection of CD4$^+$T-cells

Kh. Divsar and R. Katani*

Abstract

In this work, we study HIV infection dynamics through fractional-order differential equations (FDEs), which have gained significant attention due to their ability to model complex biological processes with greater accuracy. The fractional-order models provide a more flexible framework by incorporating memory effects and hereditary properties, which are crucial in understanding the progression of HIV infection in CD4$^+$T-cells. This paper introduces a novel numerical method for solving the FDE model of HIV infection by transforming the FDE system into a system of weakly singular Volterra integral equations and applying a collocation method based on barycentric rational interpolation. A convergence analysis is given, and the estimated error is determined. This approach aims to enhance the accuracy and applicability of numerical solutions in modeling HIV dynamics. Numerical examples are provided to demonstrate the applicability and effectiveness of the proposed method.

*Corresponding author

Received 1 June 2025; revised 3 May 2026; accepted 9 May 2026

Khatoon Divsar
Faculty of sciences, Yasouj University, Yasouj - Iran. e-mail: divsark@gmail.com

Roghayeh Katani
Faculty of sciences, Yasouj University, Yasouj - Iran. e-mail: katani@yu.ac.ir

This article was suggested by the scientific committee of the 17th seminar on differential equation and dynamical system for publication in IJNAO, which was accepted after independent review

How to cite this article

Divsar, Kh. and Katani, R., Barycentric rational interpolation method for fractional-order differential equation model of HIV infection of CD4$^+$T-cells. *Iran. J. Numer. Anal. Optim.*, 2026; 16(3): 1009-1030. https://doi.org/10.22067/ijnao.2026.93855.1660

AMS subject classifications (2020): Primary 65R20; Secondary , 45D05.

Keywords: Model of HIV infection of CD4$^+$T-cells; Fractional differential equation; Barycentric rational interpolation; Error analysis.

1 Introduction

The mathematical modeling of HIV infection dynamics has long been a critical tool for understanding the complex interplay between viral replication and the immune response. Among the key components of such models, the interaction of HIV with CD4$^+$T-cells (a primary target of the virus) has been extensively studied using systems of differential equations. Perelson, Kirschner, and Boer [27] proposed an ordinary differential equation (ODE) model with four components: uninfected healthy CD4$^+$T-cells, latently infected CD4$^+$T-cells, actively infected CD4$^+$T-cells, and free virus. Other authors have proposed many models that take this model as their inspiration. Culshaw and Ruan [9] simplified the model of Perelson et al. into one consisting of only three components— healthy CD4$^+$T-cells, infected CD4$^+$T-cells, and free virus —and they introduced a discrete time delay to the model.

These integer-order models have provided invaluable insights, but they are inherently local, meaning the rate of change of the system's state depends only on its current state. Recent studies have shown that classical integer-order HIV models, while successful in capturing short-term viral dynamics, are often limited in representing the long-term and history-dependent processes inherent in immunological systems. Traditional ODE models assume that the rate of change of viral and cellular populations depends only on their present state, thereby neglecting the influence of past biological activity on current infection progression [11, 28]. However, clinical observations indicate that HIV dynamics exhibit long-term memory effects such as delayed immune activation, progressive loss of CD4$^+$T-cell regenerative capacity, and persistent viral reservoirs that continue to influence the system over time. Fractional-order models, incorporating derivatives of noninteger order, provide a mathematically rigorous framework for representing such hereditary and nonlocal effects through their intrinsic memory kernels [16, 7]. This makes fractional modeling particularly suitable for chronic viral infections like HIV, where immune responses and viral replication are governed not only by current viral load but also by accumulated infection history.

Recent advances in fractional calculus have motivated the use of fractional-order differential equations (FDEs) to better capture memory effects and nonlocal dynamics inherent in biological systems, offering a more nuanced representation of HIV pathogenesis. These models offer a

more realistic representation of HIV infection dynamics by accounting for the complex interactions and memory effects in biological systems [5]. The ability of fractional differential equations to accurately capture the dynamics of HIV infection in $CD4^+$T-cells is highlighted in previous works. For example, Sokhanvar and Askari-Hemmat [30] proposed a Galerkin method based on Legendre multi-wavelet functions to obtain approximate solutions of a fractional model for HIV infection of $CD4^+$T cells. In [8], a Laplace Adomian decomposition method was introduced for solving this model. Also, an Adams–Bashforth–Moulton method for the fractional model of HIV infection of $CD4^+$T-cells was presented in [19]. Whereas the methods mentioned above provide valuable tools for solving fractional HIV models, the development of efficient and high-order numerical schemes remains an active area of research. Recent literature has continued to explore novel approaches to handle the complexities of fractional systems. For instance, a Haar collocation method has been applied to a first-order model of HIV infection $CD4^+$T-Cells model in [3]. A quasi-linearization and Lagrangian method has been used for solving an integer order model of the HIV infection of $CD4^+$T-cells in [26]. This approach is based on the Lagrangian method by using the collocation points of transformed Hermite polynomials. Furthermore, the authors in [33] have proposed a numerical method based on novel Bessel polynomials for solving the fractional-order HIV-1 infection model of CD T-cells. This method considers the impact of antiviral drug treatment for solving a fractional-order HIV infection model, highlighting its utility as a fundamental computational tool. Alshammari et al. proposed an HIV model incorporating the impact of antiretroviral medication via the Caputo–Fabrizio derivative and presented a numerical study that investigates the complex dynamics of HIV infection in $CD4^+$T cells [2]. These recent contributions highlight the ongoing quest for numerical methods that combine computational efficiency with high precision for fractional-order biological models.

Despite these advances, a significant challenge persists: the nonlocal nature of fractional derivatives often leads to numerical methods that are computationally expensive or suffer from stability issues, especially when dealing with the weak singularities inherent in the kernel of the equivalent Volterra integral equation form. Many existing methods, including some of those cited, can be limited by low-order convergence, complex implementation, or a lack of a robust theoretical foundation for handling such singularities efficiently. This creates a clear gap for a high-order, computationally efficient, and numerically stable scheme specifically designed for systems of fractional differential equations with weak singular kernels.

In this paper, we consider the FDE model proposed by Dinga and Yea [10]. This model consists of three components and is actually a fractional version of the Culshaw and Ruan model [9]. We address the fractional-order HIV infection model of $CD4^+$T-cells by transforming the system of FDEs into a system of weakly singular Volterra integral equations (SWSVIEs). This reformulation not only mitigates the analytical complexities of fractional operators but also pro-

vides a robust framework for numerical approximation. In the following, we propose a novel collocation method based on barycentric rational interpolation to solve the resulting SWSVIEs. Our contribution is a tailored numerical scheme that directly addresses the gap of efficiently solving SWSVIEs arising from fractional biological models. The proposed method leverages the well-documented stability and high accuracy of barycentric rational interpolation, which is particularly adept at handling singular integrands without requiring specialized graded meshes, thus maintaining a simpler implementation profile than many competing high-order methods. This interpolation has emerged as a powerful tool in the numerical solution of integral equations, particularly Volterra integral equations. Huang et al. applied collocation methods based on barycentric rational interpolation to Volterra integro-differential equations with weakly singular kernels [35]. The authors in [22] combined discrete techniques on graded meshes with barycentric rational interpolation to address time-dependent partial integro-differential equations with weakly singular kernels. The method achieved second-order convergence in time. In [23], reducible quadrature rules were explored using barycentric rational interpolation, which do not require additional moment calculations. These rules demonstrated high-order convergence and extensive stability regions. Also, Li and Huang [21] developed a composite quadrature method for auto-convolution Volterra integral equations, based on linear barycentric rational interpolation. By leveraging this approach, we develop an efficient algorithm that avoids the stiffness often encountered in traditional discretization methods for FDEs. Numerical experiments demonstrate the method's precision, convergence, and applicability to real-world scenarios, underscoring its potential as a reliable tool for simulating fractional-order biomedical systems.

This paper is organized as follows: Section 2 introduces the fractional HIV infection model and its transformation to a Volterra integral system. Section 3 provides details of the barycentric rational collocation technique. The convergence analysis and the corresponding order of convergence are established in Section 4. Furthermore, Section 5 introduces an error estimator, which is particularly useful for problems where an exact solution is unknown. Numerical results demonstrating the efficacy of the proposed method are presented in Section 6, followed by a discussion and biological interpretation in Section 7. Finally, concluding remarks and future research directions are discussed in Section 8.

2 Fractional differential model

We first recall from [28] the main definitions and outcomes of fractional calculus that will be used in the following.

Definition 1. The operator J_a^α of order $\alpha \geq 0$ of the function y is given by

$$J_a^\alpha y(t) = \begin{cases} \frac{1}{\Gamma(\alpha)} \int_a^t \frac{y(s)}{(t-s)^{1-\alpha}} ds, & \alpha > 0, \\ y(t), & \alpha = 0, \end{cases}$$

and called the Riemann–Liouville fractional integral operator.

Definition 2. The Caputo fractional derivative of order $n-1 < \alpha \leq n$ $(n \in \mathbb{N})$ for $y \in AC^n([a,T])$ is defined as

$${}_a^C D_t^\alpha y(t) = \begin{cases} \frac{1}{\Gamma(n-\alpha)} \int_a^t \frac{y^{(n)}(s)}{(t-s)^{1+\alpha-n}} ds, & n-1 < \alpha < n, \\ y^{(n)}(t), & \alpha = n. \end{cases}$$

Proposition 1. TPodlubny) Based on the previous definitions, we have

$${}_0^C D_t^\alpha (J_0^\alpha y(t)) = y(t), \quad J_0^\alpha \left({}_0^C D_t^\alpha y(t)\right) = y(t) - \sum_{i=0}^{n-1} \frac{t^i}{i!} y^i(0). \tag{1}$$

2.1 Conversion to an SWSVIE

Consider FDE model proposed by Dinga and Yea [10] with some initial conditions in the form

$$\begin{aligned} {}_0^C D_t^\alpha T(t) &= s - \mu_T T(t) + rT(t)\left(1 - \frac{T(t)+I(t)}{T_{\max}}\right) - \kappa_1 V(t) T(t), \\ {}_0^C D_t^\alpha I(t) &= \kappa_1' V(t) T(t) - \mu_I I(t), \\ {}_0^C D_t^\alpha V(t) &= N\mu_b I(t) - \kappa_1 V(t) T(t) - \mu_V V(t), \quad 0 \leq t \leq R < \infty, \\ T(0) &= T_0, \quad I(0) = I_0, \quad V(0) = V_0, \end{aligned} \tag{2}$$

where $T(t)$ represents the concentration of healthy CD4$^+$T-cells at time t, $I(t)$ represents the concentration of infected CD4$^+$T-cells, and $V(t)$ the concentration of free HIV at time t. Here, $0.5 < \alpha \leq 1$ is restricted. When $0 < \alpha \leq 0.5$, we do not think fractional derivatives can approximately describe the rate of change in number [10].

Following, note that s is the source of CD4$^+$T-cells from precursors, μ_T is the natural death rate of CD4$^+$T-cells ($\mu_T T_{\max} > s$, see [27, p. 85]), r is their growth rate (thus, $r > \mu_T$ in general), and $T_{\max}$ is their carrying capacity. The parameter κ_1, represents the rate of infection of T-cells with free virus. κ_1' is the rate at which infected cells become actively infected. μ_I is a blanket death term for infected cells, to reflect the assumption that we do not initially know whether the cells die naturally or by bursting. In addition, μ_b is the lytic death rate for infected cells. Since

N viral particles are released by each lysing cell, this term is multiplied by the parameter N to represent the source for free virus (assuming a one-time initial infection). Finally, μ_V is the loss rate of virus.

The initial conditions for infection by free virus are $T(0) = T_0, I(0) = 0, V(0) = V_0$, where

$$T_0 = \frac{r - \mu_T + \left[(r-\mu_T)^2 + 4rsT_{\max}^{-1}\right]^{1/2}}{2rT_{\max}^{-1}}.$$

The parameter values given in Table 1.

By applying the operator J_0^α to both sides of (2) and by using Proposition 1, we obtain the equivalent system

$$\begin{aligned}
T(t) &= T_0 + \frac{1}{\Gamma(\alpha)} \int_0^t (t-\tau)^{\alpha-1} \left(s - \mu_T T + rT(1 - \frac{T+I}{T_{\max}}) - \kappa_1 VT \right) d\tau, \\
I(t) &= I_0 + \frac{1}{\Gamma(\alpha)} \int_0^t (t-\tau)^{\alpha-1} \left(\kappa_1' VT - \mu_I I \right) d\tau, \\
V(t) &= V_0 + \frac{1}{\Gamma(\alpha)} \int_0^t (t-\tau)^{\alpha-1} \left(N\mu_b I - \kappa_1 VT - \mu_V V \right) d\tau.
\end{aligned} \tag{3}$$

Consider

$$\begin{aligned}
K_1(t,T,I,V) &:= s - \mu_T T + rT\left(1 - \frac{T+I}{T_{\max}}\right) - \kappa_1 VT, \\
K_2(t,T,I,V) &:= \kappa_1' VT - \mu_I I, \\
K_3(t,T,I,V) &:= N\mu_b I - \kappa_1 VT - \mu_V V.
\end{aligned}$$

Therefore, we can obtain the matrix form

$$\mathbf{f}(t) = \mathbf{g} + \int_0^t \mathbf{p}(t,\tau)\mathbf{K}(\tau,\mathbf{f})d\tau,$$

where

$$\begin{aligned}
\mathbf{f} &= (T,I,V)^\top, & \mathbf{p}(t,\tau) &= \left((t-\tau)^{\alpha-1}, (t-\tau)^{\alpha-1}, (t-\tau)^{\alpha-1}\right)^\top, \\
\mathbf{g} &= (T_0, I_0, V_0)^\top, & \mathbf{K}(\tau,\mathbf{f}) &= \frac{1}{\Gamma(\alpha)} \left(K_1(\tau,\mathbf{f}^\top), K_2(\tau,\mathbf{f}^\top), K_3(\tau,\mathbf{f}^\top) \right).
\end{aligned}$$

Table 1: Parameters and variables [10]

		Values
T	Uninfected CD4$^+$T-cell population size	$T_0 = 1000\ mm^{-3}$
I	Infected CD4$^+$T-cell density	$I_0 = 0$
V	Initial density of HIV RNA	$V_0 = 10^{-3}\ mm^{-3}$
μ_T	Natural death rate of CD4$^+$T-cells	$0.02\ day^{-1}$
μ_I	Blanket death rate of infected CD4$^+$T-cells	$0.26\ day^{-1}$
μ_b	Lytic death rate for infected cells	$0.24\ day^{-1}$
μ_v	Death rate of free virus	$2.4\ day^{-1}$
κ_1	Rate CD4$^+$T-cells become infected with virus	$2.4 \times 10^{-5}\ mm^3\ day^{-1}$
κ_1'	Rate infected cells become active	$2 \times 10^{-5}\ mm^3\ day^{-1}$
r	Growth rate of CD4$^+$T-cells population	$0.03\ day^{-1}$
N	Number of virions produced by infected CD4$^+$T-cells	Varies
$T_{\max}$	Maximal population level of CD4$^+$T-cells	$1500\ mm^{-3}$
s	Source term for uninfected CD4$^+$T-cells	$10\ day^{-1}\ mm^{-3}$
T_0	CD4$^+$ T-cells population for HIV-negative persons	$1000\ mm^{-3}$

3 Numerical method

Generally, the singularities of the kernel functions of the system (3) may cause a nonsmooth solution, and its derivative to be singular near the integral domain boundary. In these cases, by a smoothing transformation, we can recover the smooth property of the solution, (for more details, see [14, 17, 20]). In the following, we assume that the solutions of the system (3) are smooth.

Given $n+1$ strictly ordered interpolation points $t_0 < t_1 < \cdots < t_n$ and the associated values $y_0, y_1, \ldots, y_n$, for each $i = 0, 1, \ldots, n-d$, $0 \le d \le n$, let p_i be the polynomials of degree at most d which interpolate the $d+1$ data pairs $\{(t_i, y_i), (t_{i+1}, y_{i+1}), \ldots, (t_{i+d}, y_{i+d})\}$ of $y(t)$, $t \in [0, R]$, with $y_k = y(t_k)$, $k = 0, 1, \ldots, n$. Then the Floater and Hormann rational interpolation function [12] is

$$r(t) = \frac{\sum_{i=0}^{n-d} \lambda_i(t) p_i(t)}{\sum_{i=0}^{n-d} \lambda_i(t)}, \qquad \lambda_i(t) = \frac{(-1)^i}{(t-t_i)(t-t_{i+1})\cdots(t-t_{i+d})}, \tag{4}$$

where $\lambda_i(t)$ are the original rational interpolation weights and d is the rational interpolation parameter.

Lemma 1. [12] For any $d \in \{0, 1, \ldots, n\}$, the rational interpolation $r(t)$ has no poles.

Lemma 2. [12] For any $d \in \{1, \ldots, n\}$ and $y \in C^{d+2}[0,1]$, then the error bound can be obtained as

$$\| y - r \| \leq \begin{cases} h^{d+1} \frac{\|u^{(d+2)}\|}{d+2}, & \text{if } n-d \text{ is odd,} \\ h^{d+1} \left(\frac{\|u^{(d+2)}\|}{d+2} + \frac{\|u^{(d+1)}\|}{d+1} \right), & \text{if } n-d \text{ is even,} \end{cases}$$

with $h = \max\limits_{0 \leq i \leq n-1} (x_{i+1} - x_i)$.

Furthermore, for equidistant nodes, the modified barycentric Floater–Hormann weight functions are introduced as $\alpha_i = (-1)^{i-d} \sum\limits_{k \in J_i} \begin{pmatrix} d \\ i-k \end{pmatrix}$ where $J_i = \{k \in I : i-d \leq k \leq i, I = \{0, 1, \ldots, n\}\}$ is an index set. As shown in [12], (4) can be rewritten as

$$r_n[y](t) = \sum_{i=0}^{n} \phi_i(t) y_i, \quad \phi_i(t) = (\frac{\alpha_i}{t - t_i}) / (\sum_{k=0}^{n} \frac{\alpha_k}{t - t_k}), \tag{5}$$

where $\phi_i(t), i = 0, 1, \ldots, n$ are called barycentric rational interpolation basis functions, which satisfy the Lagrange property $\phi_i(t_k) = \delta_{ik}, k = 0, 1, \ldots, n$ with δ_{ik} being the Kronecker-delta functions. Consider $T(t_i) \approx T_i$, $I(t_i) \approx I_i$, $V(t_i) \approx V_i$, therefore the solution of system (3) may be approximated by the representation $K_l(t, T(t), I(t), V(t)) \approx r_n[K_l](t)$, for $l = 1, 2, 3$ and we will have

$$\begin{aligned} T_j &= T_0 + \frac{1}{\Gamma(\alpha)} \sum_{i=0}^{n} K_{1,i} \int_0^{t_j} (t_j - \tau)^{\alpha-1} \phi_i(\tau) d\tau, \\ I_j &= I_0 + \frac{1}{\Gamma(\alpha)} \sum_{i=0}^{n} K_{2,i} \int_0^{t_j} (t_j - \tau)^{\alpha-1} \phi_i(\tau) d\tau, \\ V_j &= V_0 + \frac{1}{\Gamma(\alpha)} \sum_{i=0}^{n} K_{3,i} \int_0^{t_j} (t_j - \tau)^{\alpha-1} \phi_i(\tau) d\tau, \qquad j = 1, 2, \ldots, n, \end{aligned} \tag{6}$$

where the collocation nodes t_j are the same as the interpolation nodes and $K_{l,i} := K_l(t_i, T_i, I_i, V_i)$, for $l = 1, 2, 3$.

The calculation of the above integral is crucial to achieve a highly accurate solution of (3); by using variable transformation $\tau = \frac{t_j}{2} s + \frac{t_j}{2}$ we obtain

$$\int_0^{t_j} (t_j - \tau)^{\alpha-1} \phi_i(\tau) d\tau = (\frac{t_j}{2})^{\alpha} \int_{-1}^{1} (1-s)^{\alpha-1} \phi_i(\frac{t_j}{2} s + \frac{t_j}{2}) ds,$$

then we can use the Gauss–Jacobi quadrature rule with the weights ω_q and the corresponding nods σ_q to approximate them. Then, a discrete system can be obtained as

$$
\begin{aligned}
T_j &= T_0 + \frac{1}{\Gamma(\alpha)} \sum_{i=0}^{n} K_{1,i} \sum_{q=1}^{N} \omega_q (1-\sigma_q)^{\alpha-1} \phi_i(\frac{t_j}{2}\sigma_q + \frac{t_j}{2}),\\
I_j &= I_0 + \frac{1}{\Gamma(\alpha)} \sum_{i=0}^{n} K_{2,i} \sum_{q=1}^{N} \omega_q (1-\sigma_q)^{\alpha-1} \phi_i(\frac{t_j}{2}\sigma_q + \frac{t_j}{2}),\\
V_j &= V_0 + \frac{1}{\Gamma(\alpha)} \sum_{i=0}^{n} K_{3,i} \sum_{q=1}^{N} \omega_q (1-\sigma_q)^{\alpha-1} \phi_i(\frac{t_j}{2}\sigma_q + \frac{t_j}{2}), \qquad j = 1, 2, \ldots, n.
\end{aligned} \tag{7}
$$

The above system can be rewritten in matrix form

$$
\mathbf{f}_j = \mathbf{g} + \sum_{i=0}^{n} \mathbf{K}_i \sum_{q=1}^{N} \omega_q \mathbf{p}(t_j, \sigma_q) \phi_i(\frac{t_j}{2}\sigma_q + \frac{t_j}{2}),
$$

where

$$
\begin{aligned}
&\mathbf{f}_j = (T_j, I_j, V_j)^\top, && \mathbf{g} = (T_0, I_0, V_0)^\top,\\
&\mathbf{p}(t_j, \sigma_q) = \left((t_j-\sigma_q)^{\alpha-1}, (t_j-\sigma_q)^{\alpha-1}, (t_j-\sigma_q)^{\alpha-1}\right)^\top,\\
&\mathbf{K}_i = \frac{1}{\Gamma(\alpha)} \left(K_1(t_i, T_i, I_i, V_i), K_2(t_i, T_i, I_i, V_i), K_3(t_i, T_i, I_i, V_i)\right).
\end{aligned}
$$

Finally, the obtained system can be solved by Newton's iterative method, therefore we can obtain the values of the unknown functions at the mesh points by solving system (7) for $j = 1, 2, \ldots, n$.

4 Convergence analysis

Theorem 1. Let the kernels K_l $(l = 1, 2, 3)$ satisfy a Lipschitz condition with respect to all variables except for the first. Then the solution of the discrete system (6) is convergent to the solution of the system (3), and the order of convergence is $O(h^{\alpha+d+1})$.

Proof. Let $\mathbf{e}_j := \mathbf{f} - \mathbf{f}_j$, without loss of generality, suppose that $\| \mathbf{e}_j \|_\infty = \max_j |T(t_j) - T_j|$. Therefore by subtracting (6) from (3), we have

$$
\begin{aligned}
T(t_j) - T_j =& \frac{1}{\Gamma(\alpha)} \int_0^{t_j} (t_j-\tau)^{\alpha-1} K_1(\tau, T, I, V) d\tau\\
&- \frac{1}{\Gamma(\alpha)} \sum_{i=0}^{n} K_{1,i} \int_0^{t_j} (t_j-\tau)^{\alpha-1} \phi_i(\tau) d\tau\\
=& \frac{1}{\Gamma(\alpha)} \int_0^{t_j} (t_j-\tau)^{\alpha-1} \Big[K_1(\tau, T, I, V)
\end{aligned}
$$

$$-\sum_{i=0}^{n} K_1(t_i, T(t_i), I(t_i), V(t_i))\phi_i(\tau)\Big] d\tau$$
$$+\frac{1}{\Gamma(\alpha)}\sum_{i=0}^{n}[K_1(t_i, T(t_i), I(t_i), V(t_i))$$
$$-K_1(t_i, T_i, I_i, V_i)]\int_0^{t_j}(t_j-\tau)^{\alpha-1}\phi_i(\tau)d\tau. \quad (8)$$

To guarantee the existence and uniqueness of the solution, we assume that the kernels $K_l(l = 1, 2, 3)$ satisfy a Lipschitz condition with respect to all variables except for the first [18, 25]. Applying Lemma 2 and using Lipschitz condition leads to

$$\| \mathbf{e}_j \|_\infty \leq \frac{h^{d+1}}{(d+2)\Gamma(\alpha)} \| \frac{\partial^{d+2}K}{\partial t^{d+2}} \| \int_0^{t_j}(t_j-\tau)^{\alpha-1}d\tau$$
$$+\frac{L}{\Gamma(\alpha)}\sum_{i=0}^{n}\| \mathbf{f}(t_j)-\mathbf{f}_j \| \int_0^{t_j}(t_j-\tau)^{\alpha-1}\phi_i(\tau)d\tau,$$

where $n-d$ is odd and L is the Lipschitz constant. Equivalently,

$$\| \mathbf{e}_j \|_\infty \leq \frac{h^{\alpha+d+1}j^\alpha}{\alpha(d+2)\Gamma(\alpha)} \| \frac{\partial^{d+2}K}{\partial t^{d+2}} \|$$
$$+\frac{Lj^\alpha h^\alpha}{2\Gamma(\alpha)}\sum_{i=0}^{n}\| \mathbf{e}_i \| \int_{-1}^{1}(1-s)^{\alpha-1}\phi_i(\frac{t_j}{2}s+\frac{t_j}{2})ds$$
$$\leq C_1 h^{\alpha+d+1} + C_2 h^\alpha \sum_{i=0}^{n}\| \mathbf{e}_i \|,$$

where C_1, C_2 are constants. Similarly, for the case where $n-d$ is even, we can obtain

$$\| \mathbf{e}_j \|_\infty \leq \frac{h^{d+1}j^\alpha}{\alpha\Gamma(\alpha)}\left(\frac{1}{d+2}\| \frac{\partial^{d+2}K}{\partial t^{d+2}} \| + \frac{1}{d+1}\| \frac{\partial^{d+1}K}{\partial t^{d+1}} \|\right)$$
$$+\frac{Lj^\alpha h^\alpha}{2\Gamma(\alpha)}\sum_{i=0}^{n}\| \mathbf{e}_i \| \int_{-1}^{1}(1-s)^{\alpha-1}\phi_i(\frac{t_j}{2}s+\frac{t_j}{2})ds$$
$$\leq C_1 h^{\alpha+d+1} + C_2 h^\alpha \sum_{i=0}^{n}\| \mathbf{e}_i \| .$$

Now, consider $\| \mathbf{e} \|_\infty = \max_{j=1,2,\ldots,n} \| \mathbf{e}_j \|_\infty = \| \mathbf{e}_n \|_\infty$. Then we will have

$$\| \mathbf{e}_n \|_\infty \leq \frac{C_1}{1-C_2h^\alpha}h^{\alpha+d+1} + \frac{C_2}{1-C_2h^\alpha}h^\alpha \sum_{i=0}^{n-1}\| \mathbf{e}_i \|,$$

and using the Grönwall's inequality (see [24]), yields

$$\| \mathbf{e}_n \|_\infty \leq \frac{C_1}{1 - C_2 h^\alpha} h^{\alpha+d+1} \exp\left(\frac{C_2}{1 - C_2 h^\alpha} t_n\right).$$

Hence $\| \mathbf{e}_n \|_\infty \to 0$ as $h \to 0$ and the order of convergence is $O(h^{\alpha+d+1})$ which completes the proof. □

5 Error estimation

In this section, we obtain the error estimation (see [1, 31]) for HIV model (2) utilizing barycentric rational interpolation method. The residual functions $R_{1,n}(t)$, $R_{2,n}(t)$, and $R_{3,n}(t)$ are

$$\begin{aligned}
R_{1,n}(t) &= {}_0^C D_t^\alpha T_n(t) - s + \mu_T T_n(t) - r T_n(t)\left(1 - \frac{T_n(t) + I_n(t)}{T_{\max}}\right) \\
&\quad + \kappa_1 V_n(t) T_n(t), \\
R_{2,n}(t) &= {}_0^C D_t^\alpha I_n(t) - \kappa_1' V_n(t) T_n(t) + \mu_I I_n(t), \\
R_{3,n}(t) &= {}_0^C D_t^\alpha V_n(t) - N\mu_b I_n(t) - \kappa_1 V_n(t) T_n(t) + \mu_V V_n(t), \quad 0 \leq t \leq R < \infty.
\end{aligned} \tag{9}$$

Let the error functions be

$$\begin{aligned}
e_{1,n}(t) &= T(t) - T_n(t), \\
e_{2,n}(t) &= I(t) - I_n(t), \\
e_{3,n}(t) &= V(t) - V_n(t),
\end{aligned} \tag{10}$$

where $T(t)$, $I(t)$, and $V(t)$ are exact solutions. Subtracting (9) from (2) leads to

$$\begin{aligned}
{}_0^C D_t^\alpha (T(t) - T_n(t)) =& -\mu_T (T(t) - T_n(t)) + r(T(t) - T_n(t)) \\
& - \frac{r}{T_{\max}} \left[T^2(t) - T_n^2(t) + T(t)I(t) - T_n(t)I_n(t)\right] \\
& - \kappa_1 (V(t)T(t) - V_n(t)T_n(t)) - R_{1,n}(t), \\
{}_0^C D_t^\alpha (I(t) - I_n(t)) =& \kappa_1' (V(t)T(t) - V_n(t)T_n(t)) - \mu_I (I(t) - I_n(t)) - R_{2,n}(t), \\
{}_0^C D_t^\alpha (V(t) - V_n(t)) =& N\mu_b (I(t) - I_n(t)) - \kappa_1 (V(t)T(t) - V_n(t)T_n(t)) \\
& - \mu_V (V(t) - V_n(t)) - R_{3,n}(t),
\end{aligned} \tag{11}$$

By adding and subtracting some terms and using relations (10), we have

$$
\begin{aligned}
{}_0^C D_t^\alpha e_{1,n}(t) =& -\mu_T e_{1,n}(t) + r e_{1,n}(t) - \frac{r}{T_{\max}}\left[T^2(t) \pm 2T(t)T_n(t) \pm T_n^2(t)\right.\\
&\left. -T_n^2(t) + T(t)I(t) - T_n(t)I_n(t) \pm I(t)T_n(t)\right]\\
&-\kappa_1\left[V(t)T(t) - V_n(t)T_n(t) \pm V(t)T_n(t)\right] - R_{1,n}(t),\\
{}_0^C D_t^\alpha e_{2,n}(t) =& \kappa_1'\left[V(t)T(t) - V_n(t)T_n(t) \pm V(t)T_n(t)\right] - \mu_I e_{2,n}(t) - R_{2,n}(t),\\
{}_0^C D_t^\alpha e_{3,n}(t) =& N\mu_b e_{2,n}(t) - \kappa_1\left[V(t)T(t) - V_n(t)T_n(t) \pm V(t)T_n(t)\right]\\
&-\mu_V e_{3,n}(t) - R_{3,n}(t).
\end{aligned} \tag{12}
$$

Then we obtain

$$
\begin{aligned}
{}_0^C D_t^\alpha e_{1,n}(t) =& -\mu_T e_{1,n}(t) + r e_{1,n}(t) - \frac{r}{T_{\max}}\left[e_{1,n}^2(t) + 2T_n e_{1,n}(t) + T_n e_{2,n}(t)\right.\\
&\left. + I e_{1,n}(t)\right] - \kappa_1 V e_{1,n}(t) - \kappa_1 T_n e_{3,n}(t) - R_{1,n}(t),\\
{}_0^C D_t^\alpha e_{2,n}(t) =& \kappa_1' V e_{1,n}(t) + \kappa_1' T_n e_{3,n}(t) - \mu_I e_{2,n}(t) - R_{2,n}(t),\\
{}_0^C D_t^\alpha e_{3,n}(t) =& N\mu_b e_{2,n}(t) - \kappa_1 V e_{1,n}(t) - \kappa_1 T_n e_{3,n}(t) - \mu_V e_{3,n}(t) - R_{3,n}(t).
\end{aligned} \tag{13}
$$

The initial conditions for the approximate solution of the HIV model are $T_n(0) = T_0$, $I_n(0) = I_0$, and $V_n(0) = V_0$. Therefore the initial values for system (13) will be $e_{1,n}(0) = e_{2,n}(0) = e_{3,n}(0) = 0$.

In the following, we estimate $e_{1,n}(t)$ $e_{2,n}(t)$ and $e_{3,n}(t)$ by using proposed barycentric interpolation method (Section 3). To accomplish this, we apply the operator J_0^α to both sides of (13); then we have

$$
\begin{aligned}
e_{1,n}(t) &= e_{1,n}(0) + \frac{1}{\Gamma(\alpha)}\int_0^t (t-\tau)^{\alpha-1} k_1(\tau, e_{1,n}, e_{2,n}, e_{3,n})d\tau,\\
e_{2,n}(t) &= e_{2,n}(0) + \frac{1}{\Gamma(\alpha)}\int_0^t (t-\tau)^{\alpha-1} k_2(\tau, e_{1,n}, e_{2,n}, e_{3,n})d\tau,\\
e_{3,n}(t) &= e_{3,n}(0) + \frac{1}{\Gamma(\alpha)}\int_0^t (t-\tau)^{\alpha-1} k_3(\tau, e_{1,n}, e_{2,n}, e_{3,n})d\tau,
\end{aligned}
$$

where

$$
\begin{aligned}
k_1(t, e_{1,n}, e_{2,n}, e_{3,n}) =& (r-\mu_T)e_{1,n}(t) - \frac{r}{T_{\max}}\left[e_{1,n}^2(t) + 2T_n e_{1,n}(t) + T_n e_{2,n}(t)\right.\\
&\left. + I e_{1,n}(t)\right] - \kappa_1 V e_{1,n}(t) - \kappa_1 T_n e_{3,n}(t) - R_{1,n}(t),\\
k_2(t, e_{1,n}, e_{2,n}, e_{3,n}) =& \kappa_1' V e_{1,n}(t) + \kappa_1' T_n e_{3,n}(t) - \mu_I e_{2,n}(t) - R_{2,n}(t),
\end{aligned}
$$

$$k_3(t, e_{1,n}, e_{2,n}, e_{3,n}) = N\mu_b e_{2,n}(t) - \kappa_1 V e_{1,n}(t) - \kappa_1 T_n e_{3,n}(t) - \mu_V e_{3,n}(t) - R_{3,n}(t).$$

In the same manner as described in section 3, we consider $e_{1,n}(t_i) \approx e^i_{1,n}$, $e_{2,n}(t_i) \approx e^i_{2,n}$, $e_{3,n}(t_i) \approx e^i_{3,n}$ and $k_l(t, e_{1,n}, e_{2,n}, e_{3,n}) \approx r_n[k_l](t) = \sum_{i=0}^{n} \phi_i(t) k_{l,i}$ where $k_{l,i} := k_l(t_i, e^i_{1,n}, e^i_{2,n}, e^i_{3,n}), l = 1, 2, 3$. Also, interpolate data pairs and barycentric rational interpolation basis function are similar to Section 3.

Finally, approximation the above integrals and using Gauss–Jacobi quadrature rule and solve the obtained system leads to estimate of the error functions $e_{1,n}(t)$, $e_{2,n}(t)$ and $e_{3,n}(t)$.

6 Numerical results

In this section, we conduct experimental tests to evaluate the convergence and measure the error of the proposed method. To do this, we carry out the numerical experiments employing the software "MATLAB R2016a" to find approximate solution. The order of the Gauss–Jacobi quadrature is chosen to balance the errors between interpolation and quadrature, for Gauss–Jacobi quadrature of order N we have

$$\left| \int_{-1}^{1} (1-s)^{\alpha-1} f(s) ds - \sum_{q=1}^{N} \omega_q f(\sigma_q) \right| \leq \frac{C}{N^{2m}} \|f^{(2m)}\|_\infty,$$

where m depends on the smoothness of f. Therefore, we should choose N such that

$$\text{Quadrature Error} \approx \text{Interpolation Error}.$$

This leads to the condition

$$\frac{C_{quad}}{N^{2m}} \approx C_{interp} h^{d+1}.$$

Therefore the optimal N satisfies

$$N \approx \left(\frac{C_{quad}}{C_{interp} h^{d+1}} \right)^{\frac{1}{2m}}.$$

First, we have to apply the method presented for model (2) with parameter values reported in Table 1. We consider the initial conditions $T(0) = 1000$, $I(0) = 0$ and $V(0) = 0.001$ in the time period of 100 days with the step length size $h = 0.1$.

The diagrams of the concentration of healthy CD4$^+$T cells (T), infected CD4$^+$T cells (I), and the concentration of HIV virus (V) in this time period for different values of α are shown

in Figures 1–3, respectively. The figures show that for the initial concentration $V = 10^{-3}mm^{-3}$ of HIV virus, after almost ten days, the concentration of healthy CD4$^+$T cells decreases rapidly, and the concentration of infected CD4$^+$T cells and HIV virus increase rapidly in the same proportion and after almost twenty days, the concentration of healthy cells becomes very low, while the concentration of infected cells and HIV virus remains high. Table 2 presents the

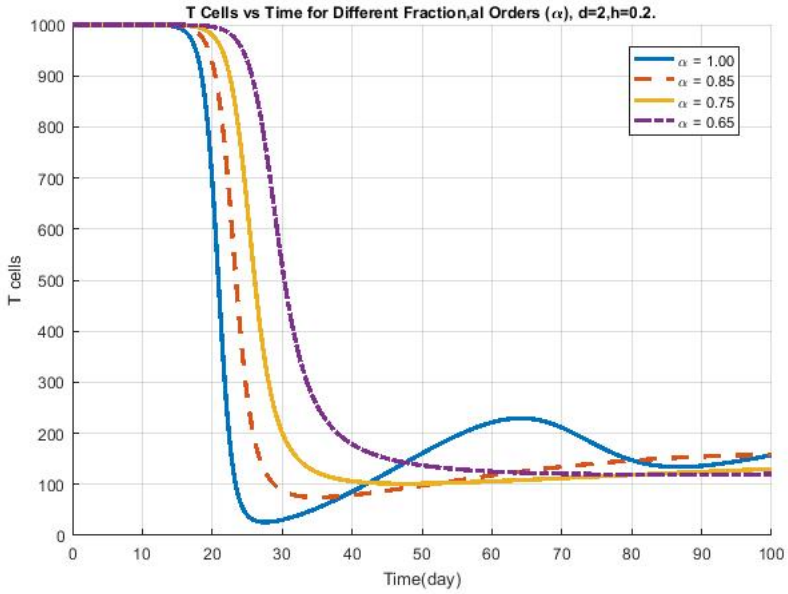


Figure 1: Rate of the concentration of healthy CD4$^+$T-cells ($T(t)$), for $d = 2$, $h = 0.2$.

estimated errors in the time period of 100 days with the step length size $h = 0.2$ at specific mesh points for $d = 2$. The results of this table show that the estimated errors for $\alpha = 0.65$ are significantly smaller, suggesting better stability for fractional orders.

To demonstrate the efficiency of the proposed method and provide comparisons with existing numerical approaches, consider [34]

$$\begin{aligned}
\frac{dT}{dt} &= 0.1 - 0.02 + T\left(1 - \frac{T+I}{1500}\right) - 0.0027VT,\\
\frac{dI}{dt} &= -0.0027VT - 0.3I,\\
\frac{dV}{dt} &= 10(0.3)I - 2.4V, \qquad\qquad 0 \le t \le 1,
\end{aligned} \tag{14}$$

with $T(0) = 0.1$, $I(0) = 0$ and $V(0) = 0.1$.

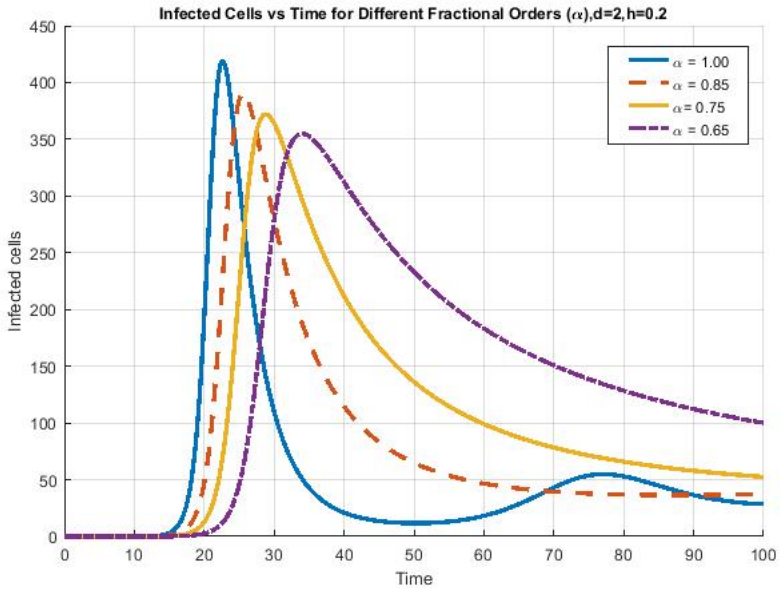


Figure 2: Rate of the concentration of infected CD4$^+$T-cells ($I(t)$)for $d = 2$, $h = 0.2$.

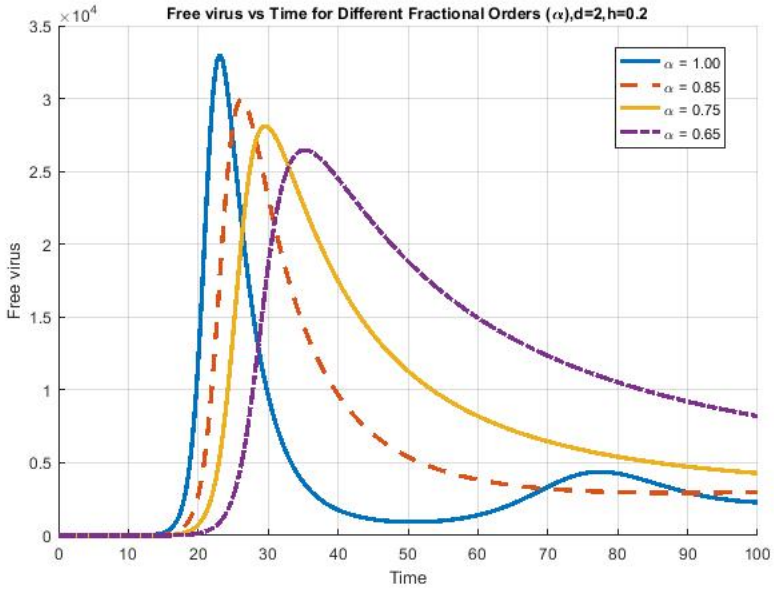


Figure 3: Rate of the concentration of free HIV virus ($V(t)$)for $d = 2$, $h = 0.2$.

Table 2: Estimated errors for parameter values of Table 1, $h = 0.2$, $d = 2$.

	$\alpha = 1$			$\alpha = 0.65$		
t	$e_{1,n}(t)$	$e_{2,n}(t)$	$e_{3,n}(t)$	$e_{1,n}(t)$	$e_{2,n}(t)$	$e_{3,n}(t)$
1	2.456e-08	5.712e-09	4.025e-09	0	0	0
10	3.104e-07	1.321e-08	1.182e-08	5.172e-11	4.065e-12	3.421e-12
20	6.461e-07	7.051e-08	4.315e-08	2.854e-10	8.421e-11	9.603e-11
40	2.603e-06	3.402e-07	9.124e-08	6.106e-10	2.207e-10	1.557e-11
60	1.781e-05	7.051e-05	4.315e-06	5.854e-09	9.217e-10	4.103e-10
80	6.821e-05	2.157e-04	1.052e-05	6.421e-08	3.165e-09	7.551e-10
100	1.004e-04	4.203e-04	5.167e-05	3.301e-07	6.781e-09	2.445e-09

In Tables 3–5, the values of the approximate solutions at some mesh points (with step length $h = 0.1$ and $d = 3$) are compared with the Runge–Kutta method, the Bessel collocation method [32], exponential Galerkin method [34], and Haar collocation algorithm [3].

The method's results in these tables are often the closest to the Runge–Kutta solution, sometimes even surpassing the other collocation methods. The close agreement with the Runge–Kutta method (the benchmark for integer-order ODEs) for the model from [34] confirms that our proposed method is numerically accurate and produces biologically plausible results.

Table 3: Comparison for uninfected cells $T(t)$.

t	Runge–Kutta	Bessel coll. [32]	Galerkin ($N = 5$) [34]	Haar coll. [3]	presented method
0.2	0.2087297222	0.2038616561	0.1982953765	0.2073431784	0.2087654367
0.4	0.4059409955	0.3803309335	0.4183153468	0.3930921357	0.4058215441
0.6	0.7635801781	0.6954623767	0.7603331972	0.6453087042	0.7634567891
0.8	1.4119574363	1.2759624442	1.4077147917	1.3978540981	1.4076542124
1.0	2.5867778755	2.3832277428	2.5915947135	2.5709802316	2.5854678912

7 Discussion and biological interpretation

The numerical simulations presented in this work provide insights into the progression of HIV infection and demonstrate the capability of the proposed numerical scheme to accurately capture the key biological characteristics of viral dynamics. The dynamics depicted in Figures 1–3 for the

Table 4: Comparison for infected $I(t)$.

t	Runge–Kutta	Bessel coll. [32]	Galerkin ($N = 5$) [34]	Haar coll. [3]	presented method
0.2	0.0000060315	0.0000062478	0.0000059641	0.0000070416	0.0000071256
0.4	0.0000131530	0.0000129355	0.0000131340	0.0000140251	0.0000134567
0.6	0.0000212106	0.0000203526	0.0000212682	0.0000230126	0.0000223456
0.8	0.0000301518	0.0000283730	0.0000301754	0.0000309231	0.0000312345
1.0	0.0000399942	0.0000369084	0.0000400377	0.0000403894	0.0000398122

Table 5: Comparison for free virus particles $V(t)$.

t	Runge–Kutta	Bessel coll. [32]	Galerkin ($N = 5$) [34]	Haar coll. [3]	presented method
0.2	0.0618798121	0.0618799185	0.0618799035	0.0617912741	0.0618797124
0.4	0.0382948730	0.0382949349	0.0382947890	0.0380026874	0.0382947891
0.6	0.0237045402	0.0237043186	0.0237046061	0.0236011879	0.0237041237
0.8	0.0146803506	0.0146795698	0.0146803810	0.0139672890	0.0146808541
1.0	0.0091008270	0.0090993030	0.0091008486	0.0089053621	0.00910084567

model (2) with the parameters from Table 1 recapitulate the well-established clinical progression of untreated HIV infection. The simulation can be interpreted as a representation of two key phases:

- The Acute Infection Phase (Days 0-10): The initial conditions, particularly the low viral load ($V(0) = 0.001$), simulate the moment of infection. The subsequent rapid decline in healthy CD4$^+$T-cells ($T(t)$) and the simultaneous rapid increase in both infected cells ($I(t)$) and free virus ($V(t)$) are hallmarks of the acute HIV syndrome. Biologically, this represents the uncontrolled replication and dissemination of the virus following its entry into the host. The virus specifically targets and hijacks the machinery of CD4$^+$T-cells. The immune activation leads to the observed sharp drop in $T(t)$, while the rise in $I(t)$ and $V(t)$ directly reflects the successful establishment and spread of the infection within the host.

- Transition to the Chronic Phase (After $\sim$ 20 days): The simulation shows the system reaching a new, pathological equilibrium where the concentration of healthy CD4$^+$T-cells remains critically low, while the populations of infected cells and free virus persist at high

levels. This state models the transition to the chronic phase of HIV infection. In a clinical setting, this corresponds to a patient with a high viral load and a low CD4 count, leaving them severely immunocompromised and vulnerable to opportunistic infections. The model's ability to capture this set-point is crucial, as the level at which this equilibrium is established is a strong predictor of the rate of disease progression to AIDS (The acute and chronic phases of HIV infection are discussed in [15], while [6] specifically addresses the establishment of a viral set-point during the asymptomatic stage and the critically low CD4 count characteristic of chronic infection).

An important contribution of the present study is the incorporation and analysis of fractional-order dynamics. The results demonstrate that solutions corresponding to fractional orders (e.g., $\alpha = 0.65$) display significantly improved numerical stability and lower estimation errors compared with the classical integer-order case (consistent with [4, 29] showing that fractional models are more biologically plausible for primary HIV infection). From a biological perspective, this observation suggests that fractional operators may better represent the intrinsic memory and hereditary properties of HIV infection, where the current state of the immune system is influenced by the history of viral–immune interactions. Such interpretations are compatible with experimental evidence indicating that viral replication, cell activation, and immune response processes involve delays and cumulative biological effects that cannot always be represented adequately by first-order derivative models. By capturing these nonlocal effects, the fractional formulation may be considered closer to the real underlying biological mechanisms, improving the descriptive and predictive power of mathematical models in immunology.

Furthermore, comparisons with existing numerical schemes, including Runge–Kutta methods, the Bessel collocation method, exponential Galerkin approaches, and Haar wavelet-based algorithms, indicate that the proposed method achieves superior computational performance in terms of accuracy and error control.

In general, the findings of this study highlight that the proposed numerical method advances current computational tools for HIV modeling by improving solution quality and enabling the study of biological memory effects. These advantages can be valuable in future investigations of therapeutic strategies, viral mutation dynamics, immune system degradation, and long-term patient response modeling.

8 Conclusion

This paper proposed a numerical method for solving a model of HIV infection of $CD4^+$T-cells. We considered the model as a fractional differential equations system and converted it to an

equivalent SWSVIEs (system of weakly singular Volterra integral equations). The obtained system was solved by a collocation method based on barycentric rational interpolation. We showed the numerical results for different values of α. The proposed method is convergent with high accuracy, simple, and can be applied to solve other cases of fractional differential equations.

The parameter d which controls the degree of the local polynomial interpolation used in the Floater–Hormann family of barycentric rational interpolations plays a crucial role in both the performance and numerical stability of the method. While higher d can improve theoretical accuracy for smooth functions, it also increases the risk of oscillations (a phenomenon akin to the Runge effect).

Although one cannot prescribe a universal optimal d, but for equidistant nodes, larger d generally improves stability in the first form [13]. Also, the numerical results show that both error and cost depend significantly on d. Future work could include a systematic sensitivity analysis or a heuristic for choosing d based on node distribution and desired accuracy.

Funding

This research received no external funding.

Conflicts of Interest

The authors declare no conflicts of interest.

References

[1] Abbasi, Z., Izadi, M. and Hosseini, M.M. *A highly accurate matrix method for solving a class of strongly nonlinear BVP arising in modeling of human shape corneal,* Math. Meth. Appl. Sci. 46(2) (2023) 1511–1527.

[2] Alshammari, A.O., Ahmad, I., Jan, R. and Idris, S.A. *Fractional-calculus analysis of the dynamics of CD4$^+$ T cells and human immunodeficiency viruses*, Eur. Phys. J. Spec. Top. 234(8) (2025) 1991–2003. https://doi.org/10.1140/epjs/s11734-024-01192-5

[3] Amin, R., YÜZBAŞI, Ş. and Nazir, S. *Efficient numerical scheme for the solution of HIV infection CD4$^+$ T-cells using Haar wavelet technique*, Comput. Model. Eng. Sci. 131(2)

(2022) 639–653. https://doi.org/10.32604/cmes.2022.019154.

[4] Arafa A.A.M., Rida S.Z., Khalil M., *Fractional modeling dynamics of HIV and $CD4^+$ T-cells during primary infection*, Nonlinear Biomed. Phys. 6 (2012) 1–7.

[5] Arshad, S., Baleanu, D., Bu, W. and Tang, Y., *Effects of HIV infection on $CD4^+$ T-cell population based on a fractional-order model*, Adv. Differ. Equ. 2017 (2017) 92. https://doi.org/10.1186/s13662-017-1143-0.

[6] Bajaria, S.H., Webb, G., Cloyd, M. and Kirschner, D. *Dynamics of naive and memory CD4+ T lymphocytes in HIV-1 disease progression*, J. Acquir. Immune Defic. Syndr. 30(1) (2002) 41–58.

[7] Baleanu, D., Diethelm, K., Scalas, E. and Trujillo J.J., *Fractional Calculus Models and Numerical Methods*, Series on Complexity, Nonlinearity and Chaos, World Scientific, 2012. https://doi.org/10.1142/8180.

[8] Hasani-Lichae, B., Biazar, J. and Ayati, Z., *The fractional differential model of HIV-1 infection of $CD4^+$ T-cells with description of the effect of antiviral drug treatment*, Comput. Math. Methods Med. 201 (1) (2019) 4059549. https://doi.org/10.1155/2019/4059549.

[9] Culshaw, R.V. and Ruan, S. *A delay-differential equation model of HIV infection of $CD4^+$ T-cells*, Math. Biosci. 165 (2000) 27–39.

[10] Dinga, Y. and Yea, H., *A fractional-order differential equation model of HIV infection of $CD4^+$ T-cells*, Math. Comput. Model. 50 (2009) 386–392.

[11] Diethelm, K. , *The analysis of fractional differential equations*, Springer, Berlin, 2010.

[12] Floater, M.S. and Hormann, K., *Barycentric rational interpolation with no poles and high rates of approximation*, J. Numer. Math. 107 (2) (2007) 315–331.

[13] Fuda, C., *Numerical stability of barycentric interpolation,Doctoral Dissertation submitted to the Faculty of Informatics of the Università della Svizzera italiana*, Doctoral Dissertation submitted to the Faculty of Informatics of the Università della Svizzera italiana, 2024.

[14] Galperin, E.A., Kansa, E.J., Makroglou, A. and Nelson, S.A., *Variable transformations in the numerical solution of second kind Volterra integral equations with continuous and weakly singular kernels; extensions to Fredholm integral equations*, J. Comput. Appl. Math. 115 (1) (2000) 193–211.

[15] Hernandez-Vargas, E.A. and Middleton, R.H. *Modeling the three stages in HIV infection,* J. Theor. Biol. 7(320) (2013)33–40. doi:10.1016/J.JTBI.2012.11.028.

[16] Hilfer, R. (Ed.), *Applications of fractional calculus in physics*. World Scientific, Singapore, 2000.

[17] Katani, R. and McKee, S., *A product integration method for nonlinear second kind Volterra integral equations with a weakly singular kernel (with application to fractional differential equations)*, Appl. Math. Lett. 163 (2025) 109403.

[18] Katani, R. and Shahmorad, S., *A block by block method for solving system of Volterra integral equations with continuous and Abel kernels*, Math. Model. Anal. 20 (6) (2015), 737–753.

[19] Khader, M.M. *The modeling dynamics of HIV and $CD4^+$ T-cells during primary infection in fractional order: numerical simulation*, Mediterr. J. Math. 15(3) (2018) 139. https://doi.org/10.1007/s00009-018-1178-9.

[20] Laurie, D.P., *Periodizing transformations for numerical integration*, J. Comp. Appl. Math. 66 (1-2)(1983) 337–344.

[21] Li, M. and Huang, C. *The linear barycentric rational quadrature method for auto-convolution Volterra integral equations*. J. Sci. Comput. 78 (2019) 549–564. https://doi.org/10.1007/s10915-018-0779-6.

[22] Liu, H., Ma, Y., Li, H. and Zhang, W., *Combination of discrete technique on graded meshes with barycentric rational interpolation for solving a class of time-dependent partial integro-differential equations with weakly singular kernels*, Comput. Math. Appl. 141 (2023) 159–169, https://doi.org/10.1016/j.camwa.2023.04.018.

[23] Ma, J., *A class of reducible quadrature rules for the second-kind Volterra integral equations using barycentric rational interpolation*, J. Comput. Appl. Math. 445 (2024) 115803. https://doi.org/10.1016/j.cam.2024.115803.

[24] McKee, S., Tang, T. and Diogo, T., *An Euler-type method for two-dimensional Volterra integral equations of the first kind*. IMA J. Numer. Anal. 20(3) (2000) 423–440.

[25] O'Regan D., Agarwal R.P. and Perera K. ,*Nonlinear integral equations singular in the dependent variable*, Appl. Math. Lett. 20 (2007) 1137-1141.

[26] Parand, K., Kalantari, Z., Delkhosh, M., *Quasilinearization-Lagrangian method to solve the HIV infection model of CD4$^+$ T cells*, SeMA 75 (2018) 271–283. https://doi.org/10.1007/s40324-017-0133-1

[27] Perelson, A.S., Kirschner, D.E. and Boer, R. De, *Dynamics of HIV infection of CD4C T-cells*, Math. Biosci. 114 (1993) 81–125.

[28] Podlubny, I., *Fractional differential equations: An introduction to fractional derivatives, fractional differential equations, to methods of their solution and some of their applications*, Elsevier, 1998.

[29] Salah, E.Y., Sontakke, B., Abdo, M.S., Shatanawi, W., Abodayeh, K., Albalwi, M.D. and Kamrujjaman, Md., *Conformable fractional-order modeling and analysis of HIV/AIDS transmission dynamics*, Int. J. Differ. Equ. 2024(1) (2024) 1958622.

[30] Sokhanvar, E. and Askari Hemmat, A., *Numerical solution of a fractional model for HIV infection of CD4$^+$ T cells via Legendre multiwavelet functions*, Int. J. Bioautomation 24 (4) (2020) 359–370. doi:10.7546/ijba.2020.24.4.000634.

[31] Yildirim, G. and Yüzbasi, S.,*Numerical solutions of SIRD model of Covid-19 by utilizing Pell-Lucascollocation method*, Turk. J. Math. 48 (6) (2024) 1156–1182.https://doi.org/10.55730/1300-0098.3567

[32] Yüzbasi, S., *A numerical approach to solve the model for HIV infection of CD4$^+$ T cells*, Appl. Math. Model. 36 (2012) 5876–5890.

[33] Yüzbasi, S. and Izadi, M., *Bessel-quasilinearization technique to solve the fractional-order HIV-1 infection of CD4$^+$ T-cells considering the impact of antiviral drug treatment*, Appl. Math. Comput. 431 (2022) 127319. https://doi.org/10.1016/j.amc.2022.127319.

[34] Yüzbasi, S. and Karacayir, M., *An exponential Galerkin method for solutions of HIV infection model of CD4$^+$ T-cells* Comput. Biol. Chem. 67(2017)205–212. DOI10.1016/j.compbiolchem.2016.12.006

[35] Zhao, Z. and Huang, C. *Collocation methods based on barycentric rational interpolation for Volterra integro-differential equations with weakly singular kernels*, Math. Methods Appl. Sci. 48 (2023) 8007–8023. https://doi.org/10.1002/mma.9468.

Iranian Journal of Numerical Analysis and Optimization
Vol. 16, No. 3, 2026, pp 1031–1063
https://doi.org/10.22067/ijnao.2026.93179.1636
https://ijnao.um.ac.ir/

CC BY NC

Research Article

Existence and Ulam–Hyers stability for a singular Caputo-fractional problem with integral boundary conditions

A.E-S. Bouziane*, and B. Bendoukha

Abstract

In this article, we study a singular fractional-order Caputo differential problem involving boundary conditions with linear operators. The analysis is carried out in an appropriate Banach space framework. First, by applying the Banach contraction principle, we establish a uniqueness result for the considered problem. Then, using Krasnosel'skii's fixed point theorem, we derive an existence theorem under suitable assumptions on the nonlinear term. In the second part of the paper, we investigate different types of stability for the obtained solutions, namely Ulam–Hyers stability and Ulam–Hyers–Rassias stability. Sufficient conditions ensuring these stability properties are presented and proved. The obtained theoretical results highlight the influence of the singular fractional operator and the associated boundary conditions. Finally, two illustrative examples are discussed in order to demonstrate the applicability and effectiveness of the established results.

*Corresponding author

Received 2 May 2025; revised 11 May 2026; accepted 15 May 2026

Aboubaker El-Saddik Bouziane
Department of Mathematics, Laboratory of Pure and Applied Mathematics, Faculty of Exact Sciences and Computer Science, Abdelhamid Ibn Badis University, Mostaganem, Algeria. e-mail: aboubaker27@gmail.com

Berrabah Bendoukha
Department of Mathematics, Laboratory of Pure and Applied Mathematics, Faculty of Exact Sciences and Computer Science, Abdelhamid Ibn Badis University, Mostaganem, Algeria. e-mail: bbendoukha@gmail.com

How to cite this article
Bouziane, A.E-S. and Bendoukha, B., Existence and Ulam–Hyers stability for a singular Caputo-fractional problem with integral boundary conditions. *Iran. J. Numer. Anal. Optim.*, 2026; 16(3): 1031-1063. https://doi.org/10.22067/ijnao.2026.93179.1636

AMS subject classifications (2020): 26A33, 34A08, 34D20, 37C25, 47A05.

Keywords: Bounded linear operator; Caputo derivative; Fixed point; Krasnosel'skii fixed point theorem; Ulam–Hyers–Rassias stability.

1 Introduction

In recent decades, fractional calculus (FC) has attracted the interest of many mathematicians and researchers. It gives the possibility to represent complex physical phenomena more accurately and efficiently than ordinary calculus. Models involving fractional differentiation highlight intermediate behaviors that cannot be done by models involving natural differentiation. Because of that, FC has, over time, become an essential tool in the description of complex dynamic systems with memory, nonlocal and hereditary properties. Compared to differential equations involving ordinary derivatives, fractional differential equations offer the possibility of better modeling in many areas such as viscolelastic materials, where the material's response depends on its entire deformation history (see [6, 7]), processes that exhibit memory effects like in biology (see [2, 24, 28]) and medicine (see [2, 4, 5, 13, 14, 20, 23]), in economy and finance see ([8, 11, 19]), and others.

The use of fractional-order differential equations, particularly the Caputo equation, has significantly improved the understanding of boundary problems. This methodology effectively translates initial conditions, leading to important insights regarding the existence of solutions and their stability. To investigate these problems, various techniques are employed, including fixed-point theorems, Ulam–Hyers and Ulam–Hyers–Rassias stability concepts see [1, 3, 12, 15, 21, 22, 25, 27] and references therein.

Our starting point in the present work is the paper [22] in which, Rezapour and Shabibi considered the singular Caputo fractional integro-differential problem

$$\begin{cases} {}^{C}D^{\alpha}x(t) + f(t, x(t)) = 0, \quad t \in]0,1[, \quad n = [\alpha] + 1, \\ x^{(j)}(0) = 0, \quad 0 \leqslant j \neq 2 \leqslant n-1, \\ ax(1) = I^{p}x(1), \quad a \geqslant 1, \quad \alpha \geqslant 3, \quad p \geqslant 1, \quad 2\Gamma(p) \geqslant \Gamma(\alpha), \end{cases}$$

where I^{p} is the Riemann–Liouville fractional integral of order p and $f : [0,1] \times \mathbb{R}^{+} \longrightarrow \mathbb{R}^{+}$ is singular at $t = 0$. They firstly proved that this problem is equivalent to an integral equation of the form:

$$y(t)=\int_0^1 G(t,s)f(s,y(s))ds,\quad G(t,s)=\frac{-(t-s)^{\alpha-1}}{\Gamma(\alpha)}\delta_{t,s}+t^2\Phi(s), \tag{1}$$

where $\Phi(s)$ is a given function and $\delta_{t,s}=1$ if $0\leqslant s\leqslant t$ and 0 otherwise.

Using the representation (1), they gave in particular a sufficient uniqueness condition by means of the Banach contraction principle and a sufficient existence condition by using the compressive form of the Guo-Krasnosel'skii fixed point theorem. In 2020, Kheiryan and Rezapour [15] investigated the Ulam stability of this problem via its integral formulation. Formula (1) implies that the kernel $G(t,s)$ is a one-dimensional perturbation of the Green's function $g(t,s):=\frac{-(t-s)^{\alpha-1}}{\Gamma(\alpha)}$ of the equation ${}^{C}D^{\alpha}x(t)+f(t,x(t))=0$.

In the present work, we discuss the existence, the uniqueness, the Ulam–Hyers and the Ulam–Hyers–Rassias stability of solutions for the singular with two integral conditions boundary-value problem:

$${}^{C}D^{\alpha}y(\tau)+\psi(\tau,y(\tau))=0,\quad \tau\in[0,1],\ \alpha\geqslant 3, \tag{2}$$

$$y^{(j)}(0)=0,\ j\in\{0,1,\ldots,n-1\}\setminus\{j_1,j_2\},\quad n=[\alpha]+1, \tag{3}$$

$$A_k\left(y^{(q_k)}(1)\right)=\frac{1}{\Gamma(p_k)}\int_0^1(1-\theta)^{p_k-1}y(\theta)d\theta,\quad p_k\ \in\mathbb{N}^*,\ k=\overline{1,2}, \tag{4}$$

where $j_1,\ j_2,\ q_1,\ q_2,\ p_1,\ p_2$ are nonnegative integers such that

$$0\leqslant q_1\neq q_2\leqslant j_1\neq j_2\leqslant[\alpha], \tag{5}$$

the function ψ is defined on $[0,1]\times\mathbb{E}\longrightarrow\mathbb{E}$ with the following singularity:

$$\lim_{\tau\longrightarrow 0^+}\|\psi(\tau,\theta)\|_{\mathbb{E}}=+\infty,\quad \text{for all}\ \ \theta\in\mathbb{E} \tag{6}$$

and satisfying some conditions that will be specified later, $\mathbb{E}$ is a real Banach space, and A_1 and A_2 are two commuting bounded linear operators in $\mathbb{E}$. Compared to [22], the present work contains the following novelties: The function ψ is Banach space valued, the number of integral conditions is greater, instead of the one-dimensional (axial) distortion represented by the scalar a in [22], we have now two multi-dimensional (spatial) distortions represented by the operators A_1 and A_2.

The article is arranged as follows: In Section 2, we recall some preliminary concepts including basic definitions and properties. In Section 3, we first give the integral equivalent formulation of the problem proving in particular that the kernel is a two-dimensional perturbation of the Green's operator $\widetilde{G}(\tau,\theta)=\frac{-(\tau-\theta)^{\alpha-1}}{\Gamma(\alpha)}I_{\mathbb{E}}$ of the equation ${}^{C}D^{\alpha}x(\tau)+\psi(\tau,x(\tau))=0$, $I_{\mathbb{E}}$ is the identity in $\mathbb{E}$. Using the integral formulation, we give a sufficient criterion for the uniqueness

and existence of a solution by means of the Banach fixed point theorem and the Krasnosel'skii fixed point theorem successively. Some Ulam–Hyers and Ulam–Hyers–Rassias stability results are also obtained. In Section 4, two illustrative examples are given.

2 Preliminaries

In what follows, $\mathbb{E}$ is a real Banach space equipped with the norm $\|\cdot\|_{\mathbb{E}}$, $C(J,\mathbb{E})$ is the Banach space of continuous functions from $J=[0,1]$ into $\mathbb{E}$ with the norm

$$\|y\|_{C(J,\mathbb{E})}=\sup\left\{\|y(\tau)\|_{\mathbb{E}}:\ \tau\in J\right\}.$$

Moreover, $L^1(J,\mathbb{R})$ is the space of measurable functions $l:J\longrightarrow\mathbb{R}$ that are Lebesgue integrable with the norm

$$\|l\|_{L^1(J,\mathbb{R})}=\int_J|l(\theta)|d\theta.$$

Likewise, $L^1(J,\mathbb{E})$ is the space of measurable functions $l:J\longrightarrow\mathbb{E}$ that are Bochner integrable with the norm [9, 18]

$$\|l\|_{L^1(J,\mathbb{E})}=\int_J\|l(\theta)\|_{\mathbb{E}}d\theta.$$

Finally, $\mathcal{B}(\mathbb{E})$ is the Banach algebra of all bounded linear operators in $\mathbb{E}$, equipped with the operator norm

$$\|A\|_{\mathcal{B}(\mathbb{E})}=\inf\left\{M>0:\|A(x)\|_{\mathbb{E}}\leqslant M\|x\|_{\mathbb{E}},\quad\text{for all}\ \ x\in\mathbb{E}\right\}.$$

The following definitions and lemmas are well known in the FC of real valued functions. They can be naturally extended to the Banach space valued case as follows.

Definition 1. [16] Let $\alpha>0$ and let $[\alpha]:=n-1$ be its integer part. The Riemann–Liouville fractional integral of order α of a function $l:J\longrightarrow\mathbb{E}$ is defined by

$$I_0^{\alpha}l(\tau)=\frac{1}{\Gamma(\alpha)}\int_0^{\tau}(\tau-\theta)^{\alpha-1}l(\theta)d\theta,$$

provided that this integral exists.

Definition 2. [16] The right-side Caputo fractional derivative of order α of a function $l:J\longrightarrow\mathbb{E}$ is defined by

$$ {}^{C}D_0^{\alpha} l(\tau) := \begin{cases} \frac{1}{\Gamma(n-\alpha)} \int_0^{\tau} \frac{l^{(n)}(\theta)}{(\tau-\theta)^{\alpha+1-n}} d\theta & (n-1 < \alpha < n), \\ \\ l^{(\alpha)}(\tau) & (\alpha = n). \end{cases} $$

In the both previous definitions, $\Gamma(\cdot)$ is the Gamma function

$$ \Gamma(\varsigma) = \int_0^{\infty} e^{-u} u^{\varsigma-1} du, \quad \Gamma(\varsigma+1) = \varsigma\Gamma(\varsigma), \quad Re(\varsigma) > 0. $$

Lemma 1. [16]

(i) If $g \in L^1(J, \mathbb{E}),\ \alpha > 0,\ \beta > 0$, then

$$ I_0^{\alpha} I_0^{\beta} g(\tau) = I_0^{\alpha+\beta} g(\tau). $$

(ii) Let $g \in L^1(J, \mathbb{E}),\ \beta > \alpha > 0$; then

$$ {}^{C}D_0^{\alpha}\ I_0^{\beta} g(\tau) = I_0^{\beta-\alpha} g(\tau). $$

(iii) If $\alpha > 0,\ \xi > -1$, then

$$ I_0^{\alpha} \left(\tau^{\xi} I_{\mathbb{E}}\right) = \frac{\Gamma(\xi+1)}{\Gamma(\alpha+\xi+1)} \tau^{\alpha+\xi} I_{\mathbb{E}}. $$

(iv) If $\alpha > 0,\ \xi > -1$, then

$$ {}^{C}D_0^{\alpha} \left(\tau^{\xi} I_{\mathbb{E}}\right) = \frac{\Gamma(\xi+1)}{\Gamma(\alpha-\xi+1)} \tau^{\alpha-\xi} I_{\mathbb{E}}, $$

where $I_{\mathbb{E}}$ is the identity operator in $\mathbb{E}$.

Lemma 2. [16] Let $y \in C(J, \mathbb{E}),\ 0 < n-1 \leqslant \alpha < n$ and let $\phi \in L^1(J,\ \mathbb{E})$. Then, the general solution of the fractional differential equation

$$ {}^{C}D^{\alpha} y(\tau) + \phi(\tau) = 0,\ (0 \leqslant \tau \leqslant 1) $$

is given by

$$ y(\tau) = \frac{-1}{\Gamma(\alpha)} \int_0^{\tau} (\tau-\theta)^{\alpha-1} \phi(\theta) d\theta + c_0 + c_1\tau + \cdots + c_{n-1}\tau^{n-1}, $$

where $c_0, c_1, \ldots, c_{n-1}$ are constants vectors in $\mathbb{E}$.

Definition 3. Let $(\mathcal{Q}, d)$ be a complet metric space. A map $H : \mathcal{Q} \longrightarrow \mathcal{Q}$ is called a contraction if there exists $0 \leqslant L < 1$ such that

$$d(Hy_1,\ Hy_2) \leqslant Ld(y_1,\ y_2), \quad \text{for all} \quad y_1,\ y_2 \in \mathcal{Q}.$$

Lemma 3. [26] (Banach contraction principle) Let $\mathcal{Q}$ be a complete metric space and let $H : \mathcal{Q} \longrightarrow \mathcal{Q}$ be a contraction. Then, there exists a unique $y \in \mathcal{Q}$ such that $Hy = y$.

Theorem 1 (Krasnosel'skii's fixed point theorem [17])**.** Let $\mathcal{S}$ be a closed, convex, and nonempty subset of a Banach space $\mathcal{L}$. Let $M,\ N$ be operators in $\mathcal{L}$ such that
(a) $My + Nz \in \mathcal{S}$ whenever $y, z \in \mathcal{S}$;
(b) M is compact and continuous;
(c) N is a contraction.
Then, there exists $x \in \mathcal{S}$ such that $Mx + Nx = x$.

The following lemma regroups some well-known facts about the bounded linear operators in Banach spaces [29].

Lemma 4. Let $\mathbb{E}$ be a Banach space.
1) If $A \in \mathcal{B}(\mathbb{E})$ is a bijection, then A^{-1} is also an element of $\mathcal{B}(\mathbb{E})$.
2) If $A,\ B$ are two commuting elements of $\mathcal{B}(\mathbb{E})$ and A is a bijection, then A^{-1} and B also commute.
3) If $A,\ B$ are two elements of $\mathcal{B}(\mathbb{E})$, then

$$\|AB\|_{\mathcal{B}(\mathbb{E})} \leqslant \|A\|_{\mathcal{B}(\mathbb{E})} \|B\|_{\mathcal{B}(\mathbb{E})}.$$

4) If $A,\ B \in \mathcal{B}(\mathbb{E})$ are such that A is bijective and $\|B\|_{\mathcal{B}(\mathbb{E})} < \frac{1}{\|A^{-1}\|_{\mathcal{B}(\mathbb{E})}}$, then also $A + B$ is bijective.

3 Main results

3.1 Auxiliary results

We begin this section by introducing the following notations that will be used through the reminder of the article. Let $\mathcal{B}_0$ be a set of commuting elements of $\mathcal{B}(\mathbb{E})$ and let j, p, q be nonnegative integers such that $q \leqslant j \leqslant \alpha - q - 1$. For each $A \in \mathcal{B}_0$, we set

$$w(A, j, p, q) := \frac{\Gamma(j+1)}{\Gamma(j-q+1)} A - \frac{\Gamma(j+1)}{\Gamma(j+p+1)} I_{\mathbb{E}}, \tag{7}$$

$$z(\theta, A, p, q) = \frac{(1-\theta)^{\alpha-q-1}}{\Gamma(\alpha-q)} A - \frac{(1-\theta)^{\alpha+p-1}}{\Gamma(\alpha+p)} I_{\mathbb{E}}, \quad \theta \in J, \tag{8}$$

where $I_{\mathbb{E}}$ is the identity in $\mathbb{E}$. We also set

$$\mathcal{W} = \left\{ w(A, j, p, q) : \ A \in \mathcal{B}_0, \ j, p, q \in \mathbb{Z}^+ \right\};$$

$$\mathcal{Z} = \left\{ z(\theta, A, p, q) : \ \theta \in J, \ A \in \mathcal{B}_0, \ p, q \in \mathbb{Z}^+ \right\}.$$

Remark 1. The following properties are verified:

1. Each of $\mathcal{W}$ and $\mathcal{Z}$ is a subset of commuting elements of $\mathcal{B}(\mathbb{E})$ with the following norm estimations:

$$\|w(A, j, p, q)\|_{\mathcal{B}(\mathbb{E})} \leqslant \frac{\Gamma(j+1)}{\Gamma(j-q+1)} \|A\|_{\mathcal{B}(\mathbb{E})} + \frac{\Gamma(j+1)}{\Gamma(j+p+1)}, \tag{9}$$

$$\|z(\theta, A, p, q)\|_{\mathcal{B}(\mathbb{E})} \leqslant \frac{\|A\|_{\mathcal{B}(\mathbb{E})}}{\Gamma(\alpha-q)} + \frac{1}{\Gamma(\alpha+p)}, \qquad \text{for all} \ \ \theta \in J. \tag{10}$$

2. Each element of $\mathcal{W}$ commutes with each element of $\mathcal{Z}$.

Suppose now that the nonnegative integers $j_1, \ j_2, \ p_1, \ p_2, \ q_1, \ q_2$ satisfy (5) and $\mathcal{B}_0 = \{A_1, \ A_2\}$, where $A_1, \ A_2$ are the commuting elements of $\mathcal{B}(\mathbb{E})$ appearing in the relation (4) and let

$$U = w(A_1, j_1, p_1, q_1) w(A_2, j_2, p_2, q_2) - w(A_1, j_2, p_1, q_1) w(A_2, j_1, p_2, q_2), \tag{11}$$

$$\begin{aligned} \Phi_1(\theta) = \ & w(A_2, j_2, p_2, q_2) z(\theta, A_1, p_1, q_1) \\ & - w(A_1, j_2, p_1, q_1) z(\theta, A_2, p_2, q_2), \quad \theta \in J, \end{aligned} \tag{12}$$

$$\begin{aligned} \Phi_2(\theta) = \ & w(A_1, j_1, p_1, q_1) z(\theta, A_2, p_2, q_2) \\ & - w(A_2, j_1, p_2, q_2) z(\theta, A_1, p_1, q_1), \quad \theta \in J. \end{aligned} \tag{13}$$

Using the commutativity of the operators, one can check that

$$w(A_1, j_1, p_1, q_1)\Phi_1(\theta) + w(A_1, j_2, p_1, q_1)\Phi_2(\theta) = Uz(\theta, A_1, p_1, q_1) \quad (\theta \in J) \tag{14}$$

and

$$w(A_2, j_1, p_2, q_2)\Phi_1(\theta) + w(A_2, j_2, p_2, q_2)\Phi_2(\theta) = Uz(\theta, A_2, p_2, q_2), \ (\theta \in J). \tag{15}$$

Lemma 5. Let $\phi \in L^1(J, \mathbb{E})$. If the operator U is invertible, then the unique solution of the boundary value problem

$$({}^C D^\alpha y)(\tau) + \phi(\tau) = 0, \qquad 0 < \tau \leqslant 1, \quad \alpha \geqslant 3 \tag{16}$$

with the conditions (3), (4), (5) is given by

$$y(\tau) = \int_0^1 G(\tau, \theta)\,(\phi(\theta))\,d\theta, \tag{17}$$

where

$$G(\tau,\theta) = \begin{cases} \frac{-(\tau-\theta)^{\alpha-1}}{\Gamma(\alpha)} I_{\mathbb{E}} + \tau^{j_1} U^{-1}\Phi_1(\theta) + \tau^{j_2} U^{-1}\Phi_2(\theta), & 0 \leqslant \theta < \tau \leqslant 1, \\ \tau^{j_1} U^{-1}\Phi_1(\theta) + \tau^{j_2} U^{-1}\Phi_2(\theta), & 0 \leqslant \tau \leqslant \theta \leqslant 1. \end{cases} \tag{18}$$

Proof. Note firstly that

$$\Phi_1(\theta) \in \mathcal{B}(\mathbb{E}), \ \Phi_2(\theta) \in \mathcal{B}(\mathbb{E}), \ G(\tau,\theta) \in \mathcal{B}(\mathbb{E}), \quad \text{for all} \ \ (\theta, \ \tau) \in J^2.$$

According to Lemma 2, the solution of the problem (16) has the form

$$y(\tau) = \frac{-1}{\Gamma(\alpha)} \int_0^\tau (\tau - \theta)^{\alpha-1}\phi(\theta)d\theta + c_0 + \tau c_1 + \cdots + \tau^{n-1} c_{n-1},$$

where $c_0, c_1, \ldots, c_{n-1}$ are constant vectors in $\mathbb{E}$. Using the initial conditions at $\tau = 0$, we obtain that $c_k = 0$ for $k \notin \{j_1, j_2\}$ and then,

$$y(\tau) = \frac{-1}{\Gamma(\alpha)} \int_0^\tau (\tau - \theta)^{\alpha-1}\phi(\theta)d\theta + \tau^{j_1} c_{j_1} + \tau^{j_2} c_{j_2} \quad (\tau \in J).$$

The integral conditions

$$A_k \left(y^{(q_k)}(1) \right) = \frac{1}{\Gamma(p_k)} \int_0^1 (1-\theta)^{p_k - 1} y(\theta)d\theta, \quad k = \overline{1,2}$$

give us

$$A_k\left(\frac{-1}{\Gamma(\alpha-q_k)}\int_0^1(1-\theta)^{\alpha-q_k-1}\phi(\theta)d\theta+\frac{\Gamma(j_1+1)}{\Gamma(j_1-q_k+1)}c_{j_1}+\frac{\Gamma(j_2+1)}{\Gamma(j_2-q_k+1)}c_{j_2}\right)$$

$$=\frac{-1}{\Gamma(\alpha+p_k)}\int_0^1(1-\theta)^{\alpha+p_k-1}\phi(\theta)d\theta+\frac{\Gamma(j_1+1)}{\Gamma(j_1+p_k+1)}c_{j_1}+\frac{\Gamma(j_2+1)}{\Gamma(j_2+p_k+1)}c_{j_2},$$

or equivalently,

$$\begin{aligned}&\left(\frac{\Gamma(j_1+1)}{\Gamma(j_1-q_k+1)}A_k-\frac{\Gamma(j_1+1)}{\Gamma(j_1+p_k+1)}I_{\mathbb{E}}\right)(c_{j_1})\\&+\left(\frac{\Gamma(j_2+1)}{\Gamma(j_2-q_k+1)}A_k-\frac{\Gamma(j_2+1)}{\Gamma(j_2+p_k+1)}I_{\mathbb{E}}\right)(c_{j_2})\\&=\int_0^1\left(\frac{(1-\theta)^{\alpha-q_k-1}}{\Gamma(\alpha-q_k)}A_k-\frac{(1-\theta)^{\alpha+p_k-1}}{\Gamma(\alpha+p_k)}I_{\mathbb{E}}\right)(\phi(\theta))d\theta.\end{aligned}$$

Taking in account formulas (7) and (8), we get finally that the constant vectors c_{j_1} and c_{j_2} satisfy the following linear system:

$$\begin{cases}w(A_1,j_1,p_1,q_1)(c_{j_1})+w(A_1,j_2,p_1,q_1)(c_{j_2})=\int_0^1 z(\theta,A_1,p_1,q_1)\,(\phi(\theta))\,d\theta,\\[2ex] w(A_2,j_1,p_2,q_2)(c_{j_1})+w(A_2,j_2,p_2,q_2)(c_{j_2})=\int_0^1 z(\theta,A_2,p_2,q_2)\,(\phi(\theta))\,d\theta.\end{cases}$$

By applying the operator $w(A_2,j_2,p_2,q_2)$ to the first equation and the operator $w(A_1,j_2,p_1,q_1)$ to the second equation and then subtracting, we get

$$\begin{aligned}U\,(c_{j_1})&=\int_0^1\begin{pmatrix}w(A_2,j_2,p_2,q_2)z(\theta,A_1,p_1,q_1)\\[2ex] -\,w(A_1,j_2,p_1,q_1)z(\theta,A_2,p_2,q_2)\end{pmatrix}(\phi(\theta))\,d\theta\\&=\int_0^1\Phi_1(\theta)\,(\phi(\theta))\,d\theta.\end{aligned}$$

Similarly, by applying the operator $w(A_2,j_1,p_2,q_2)$ to the first equation and the operator $w(A_1,j_1,p_1,q_1)$ to the second equation and then subtracting, we get

$$\begin{aligned}U\,(c_{j_2})&=\int_0^1\begin{pmatrix}w(A_1,j_1,p_1,q_1)z(\theta,A_2,p_2,q_2)\\[2ex] -\,w(A_2,j_1,p_2,q_2)z(\theta,A_1,p_1,q_1)\end{pmatrix}(\phi(\theta))\,d\theta\\&=\int_0^1\Phi_2(\theta)\,(\phi(\theta))\,d\theta.\end{aligned}$$

Since the operator U is autonomous with respect to $\tau \in J$ and invertible by hypothesis, we have

$$(c_{j_1}, c_{j_2}) = \left(\int_0^1 U^{-1}\Phi_1(\theta)\left(\phi(\theta)\right) d\theta,\ \int_0^1 U^{-1}\Phi_2(\theta)\left(\phi(\theta)\right) d\theta \right).$$

Replacing in the expression of $y(\tau)$, we get finally

$$\begin{aligned} y(\tau) &= \frac{-1}{\Gamma(\alpha)} \int_0^\tau (\tau-\theta)^{\alpha-1}\phi(\theta)d\theta + \tau^{j_1}c_{j_1} + \tau^{j_2}c_{j_2} \\ &= \frac{-1}{\Gamma(\alpha)} \int_0^\tau (\tau-\theta)^{\alpha-1}\phi(\theta)d\theta + \tau^{j_1} \int_0^1 U^{-1}\Phi_1(\theta)\left(\phi(\theta)\right) d\theta \\ &\quad + \tau^{j_2} \int_0^1 U^{-1}\Phi_2(\theta)\left(\phi(\theta)\right) d\theta \\ &= \int_0^1 G(\tau,\theta)\left(\phi(\theta)\right) d\theta. \end{aligned}$$

Conversely, assume that the function $y(\tau),\ \tau \in J$ admits the representation given by (17), (18). Observing that

$$y(\tau) = -I_{0^+}^{\alpha}\phi(\tau) + \tau^{j_1} \int_0^1 U^{-1}\Phi_1(\theta)\left(\phi(\theta)\right) d\theta + \tau^{j_2} \int_0^1 U^{-1}\Phi_2(\theta)\left(\phi(\theta)\right) d\theta.$$

It is not difficult to check that $({}^{C}D^{\alpha}y)(\tau) + \phi(\tau) = 0,\ 0 < \tau \leqslant 1$ and that the initial conditions (3) are all satisfied. Let us prove the condition (4) for $k = 1$ (the case where $k = 2$ can be obtained by the same way). Direct calculations give

$$\begin{aligned} A_1\left(y^{(q_1)}(1)\right) &= \int_0^1 \frac{-(1-\theta)^{\alpha-q_1-1}}{\Gamma(\alpha-q_1)} A_1\left(\phi(\theta)\right) d\theta \\ &\quad + \int_0^1 \frac{\Gamma(j_1+1)}{\Gamma(j_1-q_1+1)} A_1U^{-1}\Phi_1(\theta)\left(\phi(\theta)\right) d\theta \\ &\quad + \int_0^1 \frac{\Gamma(j_2+1)}{\Gamma(j_2-q_1+1)} A_1U^{-1}\Phi_2(\theta)\left(\phi(\theta)\right) d\theta, \end{aligned}$$

$$\begin{aligned} I_{0^+}^{p_1}y(1) &= \int_0^1 \frac{-(1-\theta)^{\alpha+p_1-1}}{\Gamma(\alpha+p_1)} \phi(\theta)d\theta \\ &\quad + \int_0^1 \frac{\Gamma(j_1+1)}{\Gamma(p_1+j_1+1)} U^{-1}\Phi_1(\theta)\left(\phi(\theta)\right) d\theta \\ &\quad + \int_0^1 \frac{\Gamma(j_2+1)}{\Gamma(p_1+j_2+1)} U^{-1}\Phi_2(\theta)\left(\phi(\theta)\right) d\theta. \end{aligned}$$

Hence, we deduce that

$$I_{0^+}^{p_1} y(1) - A_1 \left(y^{(q_1)}(1) \right) = \int_0^1 z(\theta, A_1, p_1, q_1) \left(\phi(\theta) \right) d\theta - \int_0^1 w(A_1, j_1, p_1, q_1) U^{-1} \Phi_1(\theta) \left(\phi(\theta) \right) d\theta - \int_0^1 w(A_1, j_2, p_1, q_1) U^{-1} \Phi_2(\theta) \left(\phi(\theta) \right) d\theta.$$

By (14), $I_{0^+}^{p_1} y(1) - A_1 \left(y^{(q_1)}(1) \right) = 0$. □

In the best of our knowledge, the kernel $G(\tau, \theta)$ in the previous lemma is obtained for the first time. It is an uniformly bounded linear operator according to the norm-majoration (22) below. Moreover,

$$G(0, \theta) = 0 = G(\tau, 1), \quad \text{for all } \tau,\ \theta \in [0, 1].$$

Function $\widetilde{G}(\tau, \theta) = -\frac{(\tau - \theta)^{\alpha - 1}}{\Gamma(\alpha)} I_{\mathbb{E}}$ is such that ${}^C D^{\alpha} \left(\int_0^1 \widetilde{G}(\tau, \theta) h(\theta) d\theta \right) = h(\tau)$. Thus, it can be regarded as the Green's operator of the fractional differential equation ${}^C D^{\alpha} x(\tau) + h(\tau) = 0$.

If we set $\mathbb{E} = \mathbb{R},\ q_1 = q_2 := q,\ j_1 = j_2 := j,\ p_1 = p_2 := p$ and $A_1 = A_2 := a,\ a \geqslant 1$, then the problem (2)–(4) becomes

$${}^C D^{\alpha} y(\tau) + \psi(\tau, y(\tau)) = 0,\ \tau \in [0, 1],\ \alpha \geqslant 3, \tag{19}$$

$$y^{(j)}(0) = 0,\ j \in \{0, 1, \ldots, n-1\} \setminus \{j\},\ n = [\alpha] + 1, \tag{20}$$

$$a y^{(q)}(1) = \frac{1}{\Gamma(p)} \int_0^1 (1 - \theta)^{p-1} y(\theta) d\theta,\ p\ \in \mathbb{N}^*. \tag{21}$$

Setting in this new problem $q = 0,\ j = 2$, we retrieve the problem investigated in [22].

Remark 2. a) Since all the real numbers $\alpha - 1,\ \alpha - q_1 - 1,\ \alpha - q_2 - 1$ are positive (see the condition (5)), all the functions

$$\theta \longmapsto (1 - \theta)^{\alpha - q_1 - 1}, \quad \theta \longmapsto (1 - \theta)^{\alpha - q_2 - 1}, \quad (\tau, \theta) \longmapsto (\tau - \theta)^{\alpha - 1}$$

are continuous. This leads to the continuity of the operator-valued functions $\Phi_1(\theta),\ \Phi_2(\theta)$ from J into $\mathcal{B}(\mathbb{E})$ and the continuity of the operator-valued function $G(\tau, \theta)$ from $J \times J$ into $\mathcal{B}(\mathbb{E})$.
b) Obviously,

$$\sup_{\theta \in J} \left(\left\| U^{-1} \Phi_k(\theta) \right\|_{\mathcal{B}(\mathbb{E})} \right) \leqslant \left\| U^{-1} \right\|_{\mathcal{B}(\mathbb{E})} \left(\sup_{\theta \in J} \left\| \Phi_k(\theta) \right) \right\|_{\mathcal{B}(\mathbb{E})} \Big) < +\infty, \quad k = \overline{1, 2},$$

and by (18), we have

$$\delta = \sup_{\theta,\tau\in J}\left(\|G(\tau,\theta)\|_{\mathcal{B}(\mathbb{E})}\right) \leqslant \Delta := \frac{1}{\Gamma(\alpha)} + \left\|U^{-1}\right\|_{\mathcal{B}(\mathbb{E})}\sum_{k=1}^{2}\sup_{\theta\in J}\left(\|\Phi_k(\theta))\,\|_{\mathcal{B}(\mathbb{E})}\right). \tag{22}$$

3.2 Existence of solutions

In the present and the following sections, ψ is the function listed in the problem (2). We shall also suppose that there exist three real constants $k > 0,\ m > 0$, and $0 < \sigma < 1$ such that for all $\tau \in [0,\ 1]$,

$$\|\tau^{\sigma}\psi(\tau, y_1(\tau)) - \tau^{\sigma}\psi(\tau, y_2(\tau))\|_{\mathbb{E}} \leqslant k\|y_1 - y_2\|_{C(J,\mathbb{E})}, \quad \text{for all } (y_1, y_2) \in (C(J,\mathbb{E}))^2, \tag{23}$$

and

$$\sup_{\tau\in J}\|\tau^{\sigma}\psi(\tau, 0)\|_{\mathbb{E}} \leqslant m. \tag{24}$$

Lemma 6. For the function ψ, the following holds:

$$\int_{\tau_0}^{\tau}\|\psi(\theta, y(\theta))\|_{\mathbb{E}}d\theta \leqslant (k\|y\|_{C(J,\mathbb{E})} + m)\frac{\tau^{1-\sigma} - \tau_0^{1-\sigma}}{1-\sigma}$$

for all $(\tau, y) \in [0,\ 1] \times C(J,\mathbb{E})$. In particular,

$$\int_{0}^{\tau}\|\psi(\theta, y(\theta))\|_{\mathbb{E}}d\theta = \lim_{\tau_0\longrightarrow 0^+}\int_{\tau_0}^{\tau}\|\psi(\theta, y(\theta))\|_{\mathbb{E}}d\theta \leqslant (k\|y\|_{C(J,\mathbb{E})} + m)\frac{\tau^{1-\sigma}}{1-\sigma}.$$

Proof. Indeed, if $0 \leqslant \theta \leqslant 1$ and $y \in C(J,\mathbb{E})$, then we get by using the conditions (23) and (24),

$$\begin{aligned}\|\psi(\theta, y(\theta))\|_{\mathbb{E}} &\leqslant \|\psi(\theta, y(\theta)) - \psi(\theta, 0)\|_{\mathbb{E}} + \|\psi(\theta, 0)\|_{\mathbb{E}}\\ &= \theta^{-\sigma}\|\theta^{\sigma}\psi(\theta, y(\theta)) - \theta^{\sigma}\psi(\theta, 0)\|_{\mathbb{E}} + \theta^{-\sigma}\|\theta^{\sigma}\psi(\theta, 0)\|_{\mathbb{E}}\\ &\leqslant (k\|y\|_{C(J,\mathbb{E})} + m)\theta^{-\sigma}.\end{aligned}$$

Thus, for $0 < \tau_0 < \tau \leqslant 1$ and $y \in C(J,\mathbb{E})$,

$$\int_{\tau_0}^{\tau}\|\psi(\theta, y(\theta))\|_{\mathbb{E}}d\theta \leqslant (k\|y\|_{C(J,\mathbb{E})} + m)\int_{\tau_0}^{\tau}\theta^{-\sigma}d\theta$$

$$= (k\|y\|_{C(J,\mathbb{E})} + m)\frac{\tau^{1-\sigma} - \tau_0^{1-\sigma}}{1-\sigma}.$$

□

It follows from this lemma that the function $[0,\ 1] \ni \theta \longmapsto \psi(\theta, y(\theta))$ belongs to $L^1([0,\ 1], \mathbb{E})$ for every fixed in $C(J, \mathbb{E})$ function y. So, the fractional integral $I_0^{\alpha}\psi(\theta, y(\theta)),\ 0 \leqslant \tau \leqslant 1$ is well defined.

Theorem 2. Let the operator U be invertible, let the constants σ and k satisfy formulas (23) and (24). If $0 < k < \frac{1-\sigma}{\delta}$, then the problem (2)–(4) admits a unique solution $y \in C(J, \mathbb{E})$. Moreover,

$$\|y\|_{C(J,\mathbb{E})} \leqslant \frac{m\delta}{1-\sigma-k\delta}.$$

Proof. Recall firstly that according to (22),

$$\delta = \sup_{\theta,\tau\in J}\left(\|G(\tau,\theta)\|_{\mathcal{B}(\mathbb{E})}\right) < \infty.$$

Consider in $C(J, \mathbb{E})$ the self mapping Ψ, defined by

$$y \longmapsto \Psi y(\tau) = \int_0^1 G(\tau,\theta)\psi(\theta, y(\theta))d\theta. \tag{25}$$

Therefore, Ψ is well defined and according to Lemma 5, the fixed points of Ψ are the solutions of the problem (2)–(4). So, to prove the theorem, we need to prove that under the given conditions, the operator Ψ has a unique fixed point. For instance, let $y_1,\ y_2$ be two elements of $C(J, \mathbb{E})$ and $0 \leqslant \tau \leqslant 1$. Then

$$\begin{aligned}
\|\Psi y_1(\tau) - \Psi y_2(\tau)\|_{\mathbb{E}} &= \left\|\int_0^1 G(\tau,\theta)\psi(\theta, y_1(\theta))d\theta - \int_0^1 G(\tau,\theta)\psi(\theta, y_2(\theta))d\theta\right\|_{\mathbb{E}} \\
&\leqslant \int_0^1 \|G(\tau,\theta)\|_{\mathcal{B}(\mathbb{E})}\|\psi(\theta, y_1(\theta)) - \psi(\theta, y_2(\theta))\|_{\mathbb{E}}d\theta \\
&\leqslant \delta\int_0^1 \theta^{-\sigma}\|\theta^{\sigma}\psi(\theta, y_1(\theta)) - \theta^{\sigma}\psi(\theta, y_2(\theta))\|_{\mathbb{E}}d\theta \\
&\leqslant k\delta\|y_1 - y_2\|_{C(J,\mathbb{E})}\int_0^1 \theta^{-\sigma}d\theta = \frac{k\delta}{1-\sigma}\|y_1 - y_2\|_{C(J,\mathbb{E})}.
\end{aligned}$$

Consequently,

$$\|\Psi y_1 - \Psi y_2\|_{C(J,\mathbb{E})} \leqslant \frac{k\delta}{1-\sigma}\|y_1 - y_2\|_{C(J,\mathbb{E})}.$$

Since $0 < \frac{k\delta}{1-\sigma} < 1$, then by the Banach contraction principle (Lemma 3), the operator Ψ has a unique fixed point y in $C(J, \mathbb{E})$, which is the unique solution of the problem (2)–(4). On the

other hand, taking in account Lemma 6, we have

$$\|y\|_{C(J,\mathbb{E})} = \|\Psi y\|_{C(J,\mathbb{E})} \leqslant \frac{k\|y\|_{C(J,\mathbb{E})} + m}{1-\sigma}\delta = \frac{k\delta}{1-\sigma}\|y\|_{C(J,\mathbb{E})} + \frac{m\delta}{1-\sigma}.$$

Consequently,

$$\left(1 - \frac{k\delta}{1-\sigma}\right)\|y\|_{C(J,\mathbb{E})} \leqslant \frac{m\delta}{1-\sigma},$$

or equivalently,

$$\|y\|_{C(J,\mathbb{E})} \leqslant \frac{m\delta}{1-\sigma-k\delta}.$$

□

Next, we give an existence result.

Theorem 3. Let the operator U be invertible, and assume that (23) is verified with the constants $k,\ \sigma$ such that

$$0 < k < \frac{1-\sigma}{\|U^{-1}\|_{\mathcal{B}(\mathbb{E})}\sum_{i=1}^{2}\sup_{\xi\in J}\left(\|\Phi_i(\xi)\|_{\mathcal{B}(\mathbb{E})}\right)}.$$

Assume also that there exists a function $\mu \in C(J,\mathbb{R}^+)$ such that

$$\tau^{\sigma}\|\psi(\tau,u)\|_{\mathbb{E}} \leqslant \mu(\tau), \quad \text{for all} \quad \tau \in J, \quad \text{for all} \quad u \in \mathbb{E}. \tag{26}$$

Then, the problem (2)–(4) has at least one solution.

Proof. For the proof, we shall use the Krasnosel'skii's theorem (Theorem 1). For instance, consider the maps $M,\ N$ defined in $C(J,\mathbb{E})$ by

$$My(\tau) = \frac{-1}{\Gamma(\alpha)}\int_0^{\tau}(\tau-\theta)^{\alpha-1}\psi(\theta,y(\theta))d\theta,$$
$$Ny(\tau) = \sum_{i=1}^{2}\tau^{j_i}\int_0^{1}U^{-1}\Phi_i(\theta)\psi(\theta,y(\theta))d\theta.$$

Observe that $\Psi = M + N$, and let

$$R \geqslant \|\mu\|_{C(J,\mathbb{R}^+)}\frac{1+\Gamma(\alpha)\|U^{-1}\|_{\mathcal{B}(\mathbb{E})}\sum_{i=1}^{2}\sup_{\xi\in J}\left(\|\Phi_i(\xi)\|_{\mathcal{B}(\mathbb{E})}\right)}{\Gamma(\alpha)(1-\sigma)},$$

$$\mathcal{Q}_R = \left\{y \in C(J,\mathbb{E}) :\ \|y\|_{C(J,\mathbb{E})} \leqslant R\right\}.$$

Hence $\mathcal{Q}_R$ is a closed, bounded and convex nonempty subset of $C(J,\mathbb{E})$, the proof will be given in several steps.

Step 1: *$My_1 + Ny_2 \in \mathcal{Q}_R$, for $y_1,\ y_2 \in \mathcal{Q}_R$.*

For $\tau \in J$ and $y_1, y_2 \in \mathcal{Q}_R$, we have

$$\begin{aligned}
&\|My_1(\tau) + Ny_2(\tau)\|_{\mathbb{E}} \\
&\quad\leqslant \|My_1(\tau)\|_{\mathbb{E}} + \|Ny_2(\tau)\|_{\mathbb{E}} \\
&\quad\leqslant \frac{1}{\Gamma(\alpha)}\int_0^\tau \|(\tau-\theta)^{\alpha-1}\psi(\theta,y_1(\theta))\|_{\mathbb{E}}d\theta \\
&\qquad+\sum_{i=1}^{2}\tau^{j_i}\int_0^1 \|U^{-1}\Phi_i(\theta)\psi(\theta,y_2(\theta))\|_{\mathbb{E}}d\theta \\
&\quad\leqslant \frac{1}{\Gamma(\alpha)}\int_0^\tau \|\psi(\theta,y_1(\theta))\|_{\mathbb{E}}d\theta \\
&\qquad+\sum_{i=1}^{2}\int_0^1 \|U^{-1}\|_{\mathcal{B}(\mathbb{E})}\sup_{\xi\in J}\left(\|\Phi_i(\xi)\|_{\mathcal{B}(\mathbb{E})}\right)\|\psi(\theta,y_2(\theta))\|_{\mathbb{E}}d\theta \\
&\quad\leqslant \frac{1}{\Gamma(\alpha)}\int_0^\tau \mu(\theta)\theta^{-\sigma}d\theta \\
&\qquad+\sum_{i=1}^{2}\|U^{-1}\|_{\mathcal{B}(\mathbb{E})}\sup_{\xi\in J}\left(\|\Phi_i(\xi)\|_{\mathcal{B}(\mathbb{E})}\right)\int_0^1 \mu(\theta)\theta^{-\sigma}d\theta \\
&\quad\leqslant \|\mu\|_{C(J,\mathbb{R}^+)}\left(\frac{1}{\Gamma(\alpha)}+\sum_{i=1}^{2}\|U^{-1}\|_{\mathcal{B}(\mathbb{E})}\sup_{\xi\in J}\left(\|\Phi_i(\xi)\|_{\mathcal{B}(\mathbb{E})}\right)\right)\int_0^1 \theta^{-\sigma}d\theta \\
&\quad= \|\mu\|_{C(J,\mathbb{R}^+)}\frac{1+\Gamma(\alpha)\|U^{-1}\|_{\mathcal{B}(\mathbb{E})}\sum_{i=1}^{2}\sup_{\xi\in J}\left(\|\Phi_i(\xi)\|_{\mathcal{B}(\mathbb{E})}\right)}{\Gamma(\alpha)(1-\sigma)} \\
&\quad\leqslant R.
\end{aligned}$$

Thus, $y_1, y_2 \in \mathcal{Q}_R \Longrightarrow My_1 + Ny_2 \in \mathcal{Q}_R$.

Step 2: *N is a contraction.*

For $y_1, y_2 \in C(J,\mathbb{E})$ and $\tau \in J = [0,\ 1]$, we have

$$\begin{aligned}
&\|Ny_1(\tau) - Ny_2(\tau)\|_{\mathbb{E}} \\
&\quad\leqslant \sum_{i=1}^{2}\tau^{j_i}\int_0^1 \|U^{-1}\Phi_i(\theta)\left(\psi(\theta,y_1(\theta)) - \psi(\theta,y_2(\theta))\right)\|_{\mathbb{E}}d\theta \\
&\quad\leqslant \|U^{-1}\|_{\mathcal{B}(\mathbb{E})}\left(\sum_{i=1}^{2}\sup_{\xi\in J}\left(\|\Phi_i(\xi)\|_{\mathcal{B}(\mathbb{E})}\right)\right) \\
&\qquad\times\int_0^1 \theta^{-\sigma}\theta^{\sigma}\|\psi(\theta,y_1(\theta)) - \psi(\theta,y_2(\theta))\|_{\mathbb{E}}d\theta
\end{aligned}$$

$$\leqslant k\|y_1 - y_2\|_{C(J,\mathbb{E})}\|U^{-1}\|_{\mathcal{B}(\mathbb{E})}\left(\sum_{i=1}^{2}\sup_{\xi\in J}\left(\|\Phi_i(\xi)\|_{\mathcal{B}(\mathbb{E})}\right)\right)\int_0^1 \theta^{-\sigma}d\theta$$
$$= k\frac{\|U^{-1}\|_{\mathcal{B}(\mathbb{E})}\sum_{i=1}^{2}\|\Phi_i\|_{C(J,\mathbb{E})}}{1-\sigma}\|y_1 - y_2\|_{C(J,\mathbb{E})}.$$

By hypothesis, $0 < k\frac{\|U^{-1}\|_{\mathcal{B}(\mathbb{E})}\sum_{i=1}^{2}\|\Phi_i\|_{C(J,\mathbb{E})}}{1-\sigma} < 1$. Thus, N is a contraction.

Step 3: *M is compact and continuous.*

The proof will be given in three Claims

Claim 1: *M is continuous.* Let y_1 and y_2 be two arbitrary elements of $C(J,\mathbb{E})$ and let $\tau \in J$. Then

$$\begin{aligned}\|My_1(\tau) - My_2(\tau)\|_{\mathbb{E}} &\leqslant \left\|\frac{1}{\Gamma(\alpha)}\int_0^\tau (\tau-\theta)^{\alpha-1}(\psi(\theta,y_1(\theta)) - \psi(\theta,y_2(\theta)))d\theta\right\|_{\mathbb{E}}\\ &\leqslant \frac{1}{\Gamma(\alpha)}\int_0^1 \theta^{-\sigma}\theta^{\sigma}\|\psi(\theta,y_1(\theta)) - \psi(\theta,y_2(\theta))\|_{\mathbb{E}}d\theta\\ &\leqslant \frac{k}{\Gamma(\alpha)}\|y_1 - y_2\|_{C(J,\mathbb{E})}\int_0^1 \theta^{-\sigma}d\theta\\ &\leqslant \frac{k}{(1-\sigma)\Gamma(\alpha)}\|y_1 - y_2\|_{C(J,\mathbb{E})}.\end{aligned}$$

Hence,

$$y_1,\ y_2 \in C(J,\mathbb{E}) \Longrightarrow \|My_1 - My_2\|_{C(J,\mathbb{E})} \leqslant \frac{k}{(1-\sigma)\Gamma(\alpha)}\|y_1 - y_2\|_{C(J,\mathbb{E})}.$$

The last relation expresses that the operator M is $\dfrac{k}{(1-\sigma)\Gamma(\alpha)}$-Lypschitzian and so, is continuous.

Claim 2: *M maps bounded sets into bounded sets in $C(J,\mathbb{E})$.*

It is enough to show that there exists a positive constant r such that for each $y \in \mathcal{Q}_R$, one has $\|M(y)\|_{C(J,\mathbb{E})} \leqslant r$. In fact, we have

$$\begin{aligned}\|(My)(\tau)\|_{\mathbb{E}} &\leqslant \frac{1}{\Gamma(\alpha)}\int_0^1 (\tau-\theta)^{\alpha-1}\|\psi(\theta,y(\theta))\|_{\mathbb{E}}d\theta\\ &\leqslant \frac{1}{\Gamma(\alpha)}\int_0^1 \theta^{-\sigma}\theta^{\sigma}\|\psi(\theta,y(\theta))\|_{\mathbb{E}}d\theta\\ &\leqslant \frac{1}{\Gamma(\alpha)}\int_0^1 \theta^{-\sigma}\mu(\theta)d\theta\\ &\leqslant \frac{\|\mu\|_{C(J,\mathbb{R}^+)}}{\Gamma(\alpha)}\int_0^1 \theta^{-\sigma}d\theta.\end{aligned}$$

The operator M is uniformly bounded in $\mathcal{Q}_R$ as

$$\|My\|_{C(J,\mathbb{E})} = \sup_{\tau\in J}\left(\|My(\tau)\|_{\mathbb{E}}\right) \leqslant r := \frac{\|\mu\|_{C(J,\mathbb{R}^+)}}{(1-\sigma)\Gamma(\alpha)}, \quad \text{for all} \quad y \in \mathcal{Q}_R.$$

Claim 3: *M maps each $\mathcal{Q}_R$ into an equicontinuous subset of $C(J,\mathbb{E})$.*
Let $\tau_1, \tau_2 \in J$ such that $\tau_1 < \tau_2$ and let $y \in \mathcal{Q}_R$. Then,

$$\begin{aligned}
&\|My(\tau_1) - My(\tau_2)\|_{\mathbb{E}} \\
&\quad= \frac{1}{\Gamma(\alpha)}\left\|\int_0^{\tau_2}(\tau_2-\theta)^{\alpha-1}\psi(\theta,y(\theta))d\theta - \int_0^{\tau_1}(\tau_1-\theta)^{\alpha-1}\psi(\theta,y(\theta))d\theta\right\|_{\mathbb{E}} \\
&\quad\leqslant \frac{1}{\Gamma(\alpha)}\left\|\int_0^{\tau_1}\left((\tau_2-\theta)^{\alpha-1}-(\tau_1-\theta)^{\alpha-1}\right)\psi(\theta,y(\theta))d\theta\right\|_{\mathbb{E}} \\
&\qquad+\frac{1}{\Gamma(\alpha)}\left\|\int_{\tau_1}^{\tau_2}(\tau_2-\theta)^{\alpha-1}\psi(\theta,y(\theta))d\theta\right\|_{\mathbb{E}} \\
&\quad\leqslant \frac{\|\mu\|_{C(J,\mathbb{R}^+)}}{\Gamma(\alpha)}\int_0^{\tau_1}\left((\tau_2-\theta)^{\alpha-1}-(\tau_1-\theta)^{\alpha-1}\right)\theta^{-\sigma}d\theta \\
&\qquad+\frac{\|\mu\|_{C(J,\mathbb{R}^+)}}{\Gamma(\alpha)}\int_{\tau_1}^{\tau_2}(\tau_2-\theta)^{\alpha-1}\theta^{-\sigma}d\theta \\
&\quad\leqslant \frac{\|\mu\|_{C(J,\mathbb{R}^+)}}{\Gamma(\alpha)}\int_0^{\tau_1}\left((\tau_2-\theta)^{\alpha-1}-(\tau_1-\theta)^{\alpha-1}\right)\theta^{-\sigma}d\theta \\
&\qquad+\frac{\|\mu\|_{C(J,\mathbb{R}^+)}}{\Gamma(\alpha)}\int_{\tau_1}^{\tau_2}\theta^{-\sigma}d\theta \\
&\quad= \frac{\|\mu\|_{C(J,\mathbb{R}^+)}}{\Gamma(\alpha)}\int_0^{\tau_1}\left((\tau_2-\theta)^{\alpha-1}-(\tau_1-\theta)^{\alpha-1}\right)\theta^{-\sigma}d\theta \\
&\qquad+\|\mu\|_{C(J,\mathbb{R}^+)}\frac{\tau_2^{1-\sigma}-\tau_1^{1-\sigma}}{(1-\sigma)\Gamma(\alpha)} \\
&\quad= \frac{\|\mu\|_{C(J,\mathbb{R}^+)}}{\Gamma(\alpha)}\left(\int_0^{\tau_1}\left((\tau_2-\theta)^{\alpha-1}-(\tau_1-\theta)^{\alpha-1}\right)\theta^{-\sigma}d\theta\right. \\
&\qquad\left.+\frac{\tau_2^{1-\sigma}-\tau_1^{1-\sigma}}{(1-\sigma)}\right).
\end{aligned}$$

Clearly, $\lim_{\tau_1\longrightarrow\tau_2}\|My(\tau_1)-My(\tau_2)\|_{\mathbb{E}}$ tends to zero independently on $y \in C(J,\mathbb{E})$. Thus, M is equicontinuous and so, it is a relatively compact operator on $\mathcal{Q}_R$. As a consequence of Claims 1, 2, 3 and together with the Arzelá -Ascoli theorem, we conclude that M is compact. According to Theorem 1, we deduce that $Mx + Nx$ has a fixed point $x \in \mathcal{Q}_R$, which is a solution of the problem (2)–(4). □

3.3 Ulam Stability

In this section, we investigate different types of Ulam's stability of the fractional boundary value problem (2). For instance, we adopt the following definitions.

Definition 4. [10] The fractional boundary value problem (2) is said to be Ulam–Hyers stable if there exists a real number $C_\psi > 0$ depending only on ψ and such that for every $\varepsilon > 0$ and every solution $y \in C(J, \mathbb{E})$ of the inequality

$$\|{}^C D_0^\alpha y(\tau) + \psi(\tau, y(\tau))\|_{\mathbb{E}} \leqslant \varepsilon, \ \tau \in [0,1]\,, \tag{27}$$

with the conditions

$$\begin{cases} y^{(j)}(0) = 0, \quad j \in \{0,1,\ldots,n-1\} \setminus \{j_1, j_2\}\,, \ n = [\alpha] + 1, \\ A_k \left(y^{(q_k)}(1)\right) = \dfrac{1}{\Gamma(p_k)} \int_0^1 (1-\theta)^{p_k - 1} y(\theta) d\theta, \ p_k \ \in \mathbb{N}^*, \ k = \overline{1,2}, \end{cases} \tag{28}$$

there exists a solution $y_0 \in C(J, \mathbb{E})$ of the fractional value problem (2) satisfying

$$\|y(\tau) - y_0(\tau)\|_{\mathbb{E}} \leqslant C_\psi \varepsilon, \ \tau \in [0,1]\,.$$

Definition 5. [10] The fractional boundary value problem (2) is said to be generalized Ulam–Hyers stable if there exists a function $\omega \in C(J, \mathbb{R}^+)$ with $\omega(0) = 0$ such that for every $\varepsilon > 0$ and every solution $y \in C(J, \mathbb{E})$ of the inequality (27) with the conditions (28), there exists a solution $y_0 \in C(J, \mathbb{E})$ of the fractional value problem (2) satisfying

$$\|y(\tau) - y_0(\tau)\|_{\mathbb{E}} \leqslant \omega(\varepsilon), \ \tau \in [0,1]\,.$$

Definition 6. [10] The fractional boundary value problem (2) is said to be Ulam–Hyers–Rassias stable with respect to the function $\varphi : [0,1] \longrightarrow \mathbb{R}^+$ if there exists a real number $C_{\psi,\varphi} > 0$ depending only on ψ, φ and such that for every $\varepsilon > 0$ and every solution $y \in C(J, \mathbb{E})$ of the inequality

$$\|{}^C D_0^\alpha y(\tau) + \psi(\tau, y(\tau))\|_{\mathbb{E}} \leqslant \varepsilon\varphi(\tau), \ \tau \in [0,1]\,, \tag{29}$$

with the conditions (28), there exists a solution $y_0 \in C(J, \mathbb{E})$ of the fractional value problem (2) satisfying

$$\|y(\tau) - y_0(\tau)\|_{\mathbb{E}} \leqslant C_{\psi,\varphi} \varepsilon\varphi(\tau), \ \tau \in [0,1]\,.$$

Remark 3. 1) A function $y \in C(J, \mathbb{E})$ is a solution of the inequality (27) if and only if there exists a function $h_y : [0,1] \longrightarrow \mathbb{E}$ such that

$$ {}^{C}D_0^{\alpha} y(\tau) + \psi(\tau, y(\tau)) = h_y(\tau) \quad \text{and} \quad \|h_y(\tau)\|_{\mathbb{E}} \leqslant \varepsilon, \quad (0 \leqslant \tau \leqslant 1). \tag{30} $$

2) Ulam–Hyers stability $\Longrightarrow$ Generalized Ulam–Hyers stability.

Theorem 4. Let $\psi : [0,1] \times \mathbb{E} \longrightarrow \mathbb{E}$ be a function satisfying (6), (23), and (24), and suppose that the operator U is invertible. If $0 < k < \frac{1-\sigma}{\delta}$, then the problem (2) is Ulam–Hyers stable.

Proof. Recall once again that according to (22),

$$ \delta = \sup_{\theta,\tau \in J} \left(\|G(\tau,\theta)\|_{\mathcal{B}(\mathbb{E})} \right) < \infty. $$

It comes from Theorem 2 that under the assumptions, the problem (2) has a unique solution y_0. Let now $\varepsilon > 0$ and let $y \in C(J, \mathbb{E})$ be any solution of the inequality (27) with the conditions (28). The function

$$ [0,1]^2 \ni (\tau,\theta) \longmapsto \widetilde{\psi_y}(\tau,\theta) = \psi(\tau,\theta) - h_y(\tau) $$

satisfies all the conditions (6), (23), and (24) with the majorating constant $m + \varepsilon$ instead of m. Moreover, y is a solution of the problem

$$ {}^{C}D^{\alpha} z(\tau) + \widetilde{\psi_y}(\tau, z(\tau)) = 0, \quad (0 \leqslant \tau \leqslant 1) $$

with the conditions (28). By Lemma 5

$$ \begin{aligned} y(\tau) &= \int_0^1 G(\tau,\theta) \left(\psi(\theta, y(\theta)) - h_y(\theta) \right) d\theta \\ &= \int_0^1 G(\tau,\theta)\psi(\theta, y(\theta)) d\theta - \int_0^1 G(\tau,\theta) h_y(\theta) d\theta \quad (0 \leqslant \tau \leqslant 1). \end{aligned} $$

Consequently, for every $\tau \in [0,1]$,

$$ \begin{aligned} &\|y(\tau) - y_0(\tau)\|_{\mathbb{E}} \\ &\quad \leqslant \left\| \int_0^1 G(\tau,\theta)\psi(\theta, y(\theta)) d\theta - \int_0^1 G(\tau,\theta)\psi(\theta, y_0(\theta)) d\theta \right\|_{\mathbb{E}} \\ &\qquad + \left\| \int_0^1 G(\tau,\theta) h_y(\theta) d\theta \right\|_{\mathbb{E}} \\ &\quad \leqslant \int_0^1 \|G(\tau,\theta)\|_{\mathcal{B}(\mathbb{E})} \|\psi(\theta, y(\theta)) - \psi(\theta, y_0(\theta))\|_{\mathbb{E}} d\theta \\ &\qquad + \int_0^1 \|G(\tau,\theta)\|_{\mathcal{B}(\mathbb{E})} \|h_y(\theta)\|_{\mathbb{E}} d\theta \\ &\quad \leqslant \delta \int_0^1 \theta^{-\sigma} \|\theta^{\sigma}\psi(\theta, y(\theta)) - \theta^{\sigma}\psi(\theta, y_0(\theta))\|_{\mathbb{E}} d\theta + \delta\varepsilon \end{aligned} $$

$$\leqslant \frac{\delta}{1-\sigma} k\|y-y_0\|_{C(J,\mathbb{E})} + \delta\varepsilon.$$

Finally,

$$\tau \in J \Longrightarrow \|y(\tau)-y_0(\tau)\|_{\mathbb{E}} \leqslant \|y-y_0\|_{C(J,\mathbb{E})} \leqslant C_\psi \varepsilon, \quad \text{with } C_\psi \geqslant \frac{(1-\sigma)\delta}{1-\sigma-k\delta}.$$

□

Remark 4. Setting $\omega(\varepsilon) = C_\psi\varepsilon$, it follows that $\omega(0) = 0$ and the problem (2) is generalized Ulam–Hyers stable.

Next result concerns the Ulam–Hyers–Rassias stability.

Theorem 5. Let $\psi : [0,1] \times \mathbb{E} \longrightarrow \mathbb{E}$ be a function satisfying (6), (23), (24), let the operator U be invertible and $0 < k < \frac{1-\sigma}{\delta}$. Then, the problem (2) is Ulam–Hyers–Rassias stable with respect to any Riemann integrable function $\varphi : [0,1] \longrightarrow \mathbb{R}^+$ for which there exists a constant $\delta_\varphi > 0$ such that

$$\text{for all } \tau \in [0,1] : \int_0^1 \varphi(\theta)d\theta \leqslant \delta_\varphi \varphi(\tau). \tag{31}$$

Proof. Let $\varepsilon > 0$ and let $\varphi : [0,1] \longrightarrow \mathbb{R}^+$ be a Riemann integrable function satisfying (31). Let also $y \in C(J,\mathbb{E})$ satisfy the inequality (29) with the conditions (28). As in the remark 3, this implies that there exists a function $h_y(\tau)$ satisfying $h_y(\tau) \leqslant \varepsilon\varphi(\tau)$ for all $\tau \in [0,1]$ and y is a solution of the problem

$${}^C D_0^\alpha z(\tau) + \widetilde{\psi_y}(\tau, z(\tau)) = 0, \quad 0 \leqslant \tau \leqslant 1$$

with the conditions (28). The function $\widetilde{\psi_y}(\tau, z(\tau))$ obviously satisfies the conditions (6) and (23). Moreover, for $0 \leqslant \tau \leqslant 1$,

$$\begin{aligned}\|\tau^\sigma \widetilde{\psi_y}(\tau,0)\|_{\mathbb{E}} &= \|\tau^\sigma(\psi(\tau,0) - h_y(\tau))\|_{\mathbb{E}} \leqslant \|\tau^\sigma\psi(\tau,0)\|_{\mathbb{E}} + \|\tau^\sigma h_y(\tau)\|_{\mathbb{E}} \\ &\leqslant m + \tau^\sigma\varepsilon\varphi(\tau) \leqslant m + \varepsilon \sup_{0\leqslant\tau\leqslant 1} \varphi(\tau).\end{aligned}$$

So, the condition (24) is also satisfied by $\widetilde{\psi_y}$. According to Lemma 5, we have the representation

$$y(\tau) = \int_0^1 G(\tau,\theta)(\psi(\theta, y(\theta)) - h_y(\theta))d\theta.$$

Therefore, if y_0 is the unique solution of problem (2)–(4), then for every $\tau \in [0,1]$,

$$\|y(\tau) - y_0(\tau)\|_{\mathbb{E}}$$

$$
\begin{aligned}
&\leqslant \int_0^1 \|G(\tau,\theta)\|_{\mathcal{B}(\mathbb{E})}\|\psi(\theta,y(\theta))-\psi(\theta,y_0(\theta))\|_{\mathbb{E}}d\theta\\
&\quad+\int_0^1 \|G(\tau,\theta)\|_{\mathcal{B}(\mathbb{E})}\|h_y(\theta)\|_{\mathbb{E}}d\theta\\
&\leqslant \delta\int_0^1 \theta^{-\sigma}\|\theta^{\sigma}\psi(\theta,y(\theta))-\theta^{\sigma}\psi(\theta,y_0(\theta))\|_{\mathbb{E}}d\theta\\
&\quad+\delta\varepsilon\int_0^1 \varphi(\theta)d\theta\\
&\leqslant \frac{\delta}{1-\sigma}k\|y-y_0\|_{C(J,\mathbb{E})}+\delta\varepsilon\int_0^1\varphi(\theta)d\theta\\
&\leqslant \frac{\delta}{1-\sigma}k\|y-y_0\|_{C(J,\mathbb{E})}+\delta\varepsilon\delta_{\varphi}\varphi(\tau).
\end{aligned}
$$

Finally,

$$
\text{for all } \ \tau\in[0,\ 1]:\ \|y(\tau)-y_0(\tau)\|_{\mathbb{E}}\leqslant C_{\psi,\varphi}\varepsilon\varphi(\tau),\quad \text{with } \ C_{\psi,\varphi}\geqslant\frac{(1-\sigma)\delta_{\varphi}\delta}{1-\sigma-k\delta}.
$$

□

4 Examples

In this section, we treat two illustrative examples that differ only in the nature of the operators used. In the first example, we suppose that A_1 and A_2 are scalar; in the second example, we consider operators of multiplication by θ and θ^2, respectively. Note firstly that for the operators Φ_1 and Φ_2 defined by (12) and (13), we have the following norm estimations:

$$
\begin{aligned}
\|\Phi_1\|_{\mathcal{B}(\mathbb{E})} &\leqslant \|w(A_2,j_2,p_2,q_2)\|_{\mathcal{B}(\mathbb{E})}\|z(\theta,A_1,p_1,q_1)\|_{\mathcal{B}(\mathbb{E})}\\
&\quad+\|w(A_1,j_2,p_1,q_1)\|_{\mathcal{B}(\mathbb{E})}\|z(\theta,A_2,p_2,q_2)\|_{\mathcal{B}(\mathbb{E})}\\
&\leqslant\left(\frac{\Gamma(j_2+1)}{\Gamma(j_2-q_2+1)}\|A_2\|_{\mathcal{B}(\mathbb{E})}+\frac{\Gamma(j_2+1)}{\Gamma(j_2+p_2+1)}\right)\\
&\quad\times\left(\frac{1}{\Gamma(\alpha-q_1)}\|A_1\|_{\mathcal{B}(\mathbb{E})}+\frac{1}{\Gamma(\alpha+p_1)}\right)\\
&\quad+\left(\frac{\Gamma(j_2+1)}{\Gamma(j_2-q_1+1)}\|A_1\|_{\mathcal{B}(\mathbb{E})}+\frac{\Gamma(j_2+1)}{\Gamma(j_2+p_1+1)}\right)\\
&\quad\times\left(\frac{1}{\Gamma(\alpha-q_2)}\|A_2\|_{\mathcal{B}(\mathbb{E})}+\frac{1}{\Gamma(\alpha+p_2)}\right)
\end{aligned}
$$

and similarly,

$$\begin{aligned}\|\Phi_2\|_{\mathcal{B}(\mathbb{E})} &\leqslant \|w(A_1,j_1,p_1,q_1)\|_{\mathcal{B}(\mathbb{E})}\|z(\theta,A_2,p_2,q_2)\|_{\mathcal{B}(\mathbb{E})}\\ &\quad+\|w(A_2,j_1,p_2,q_2)\|_{\mathcal{B}(\mathbb{E})}\|z(\theta,A_1,p_1,q_1)\|_{\mathcal{B}(\mathbb{E})}\\ &\leqslant \left(\frac{\Gamma(j_1+1)}{\Gamma(j_1-q_1+1)}\|A_1\|_{\mathcal{B}(\mathbb{E})}+\frac{\Gamma(j_1+1)}{\Gamma(j_1+p_1+1)}\right)\\ &\quad\times\left(\frac{1}{\Gamma(\alpha-q_2)}\|A_2\|_{\mathcal{B}(\mathbb{E})}+\frac{1}{\Gamma(\alpha+p_2)}\right)\\ &\quad+\left(\frac{\Gamma(j_1+1)}{\Gamma(j_1-q_2+1)}\|A_2\|_{\mathcal{B}(\mathbb{E})}+\frac{\Gamma(j_1+1)}{\Gamma(j_1+p_2+1)}\right)\\ &\quad\times\left(\frac{1}{\Gamma(\alpha-q_1)}\|A_1\|_{\mathcal{B}(\mathbb{E})}+\frac{1}{\Gamma(\alpha+p_1)}\right).\end{aligned}$$

Let $\mathbb{E}$ be a real Banach space and let $\psi : J\times\mathbb{E}\longrightarrow\mathbb{E}$ be given by

$$\psi(\tau,u)=\begin{cases}\beta\ln(e+\tau)\dfrac{7+\cos(\pi\tau)}{\sqrt{\tau}(1+\|u\|_{\mathbb{E}})}u_0,\ \tau\in]0,1],\ u\in\mathbb{E},\\[2ex] 0, \qquad\qquad\qquad\qquad \tau=0,\ u\in\mathbb{E},\end{cases}$$

where $\beta>0$ and u_0 is a fixed element in $\mathbb{E}$.
The function ψ satisfies the conditions

$$\lim_{\tau\longrightarrow 0^+}\|\psi(\tau,u)\|_{\mathbb{E}}=+\infty,\quad \text{for all}\ \ u\in\mathbb{E},$$

$$\|\sqrt{\tau}\psi(\tau,u)-\sqrt{\tau}\psi(\tau,v)\|_{\mathbb{E}}\leqslant 8\beta e\|u_0\|_{\mathbb{E}}\quad \|u-v\|_{\mathbb{E}},$$
$$\text{for all}\ \ \tau\in J,\quad \text{for all}\ \ (u,v)\in\mathbb{E}^2,$$

and

$$\|\sqrt{\tau}\psi(\tau,u)\|_{\mathbb{E}}\leqslant\mu(\tau):=8\beta e\|u_0\|_{\mathbb{E}},\ \ \text{for all}\ \ \tau\in J,\ \ \text{for all}\ \ u\in\mathbb{E}.$$

For numerical simulation, we consider $\mathbb{E}=\mathbb{R}$, $u_0=1$, and $\beta=1$. The nonlinear function ψ is implemented accordingly, and the singularity at $\tau=0$ is handled by setting $\psi(0,y)=0$.

Figure 1 highlights the amplification effect generated by increasing values of β.

Figure 2 emphasizes the strong growth generated by the singular term $\dfrac{1}{\sqrt{\tau}}$.

Figure 3 shows the damping effect produced by the term $(1+|y|)$ as $|y|$ increases.

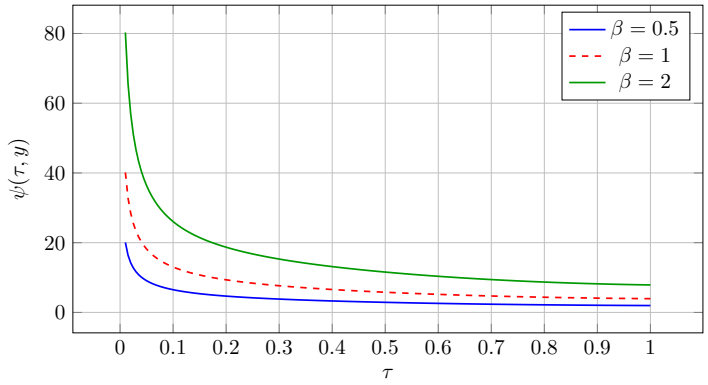


Figure 1: Influence of the parameter β on the nonlinear function $\psi(\tau, y) = \beta \ln(e+\tau)\dfrac{7+\cos(\pi\tau)}{\sqrt{\tau}(1+|y|)}$ on the interval $[0, 1]$.

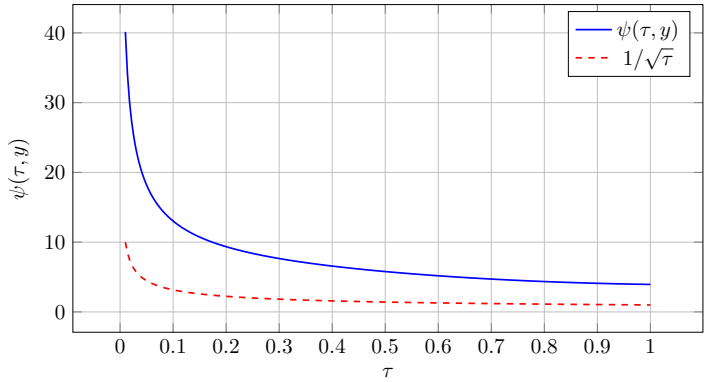


Figure 2: Singular dynamics of the nonlinear function $\psi(\tau, y)$ near $\tau = 0$ on the interval $[0, 1]$.

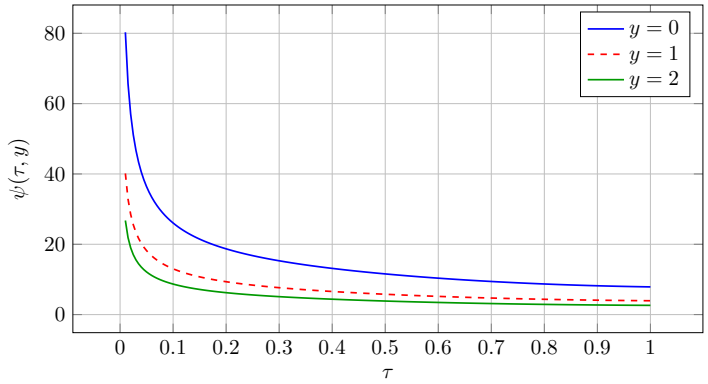


Figure 3: Effect of the state variable y on the nonlinear function $\psi(\tau, y)$ on the interval $[0, 1]$.

Example 1. Consider now in the Banach space $C(J,\mathbb{E})$ the following Caputo-fractional problem:

$$(S1):\begin{cases} {}_1^C\partial_0^{\alpha} y(\tau)+\psi(\tau,y(\tau))=0, & 0<\tau<1, \alpha\geqslant 3,\\ y^{(j)}(0)=0, & j\notin\{3,4\},\\ a_1y(1)=\int_0^1(1-s)y(s)ds, & a_1\in\mathbb{R}^*,\\ a_2y'(1)=\int_0^1 y(s)ds, & a_2\in\mathbb{R}^*. \end{cases}$$

In this example, A_1 and A_2 are the scalar operators $a_1I_{\mathbb{E}}$, $a_2I_{\mathbb{E}}$, respectively, and

$$\sigma=\frac{1}{2},\ m=8\beta e\|u_0\|_{\mathbb{E}},\ (j_1,j_2)=(3,4),\ (q_1,q_2)=(0,1),\ (p_1,p_2)=(2,1).$$

Consequently, the operator U (formula (11)) is defined in $\mathbb{E}$ as follows:

$$Ux=\left(\frac{600a_1a_2+30a_1-60a_2+1}{600}\right)x,\quad x\in\mathbb{E}.$$

Moreover, U is bounded with

$$\|U\|_{\mathcal{B}(\mathbb{E})}=\frac{|600a_1a_2+30a_1-60a_2+1|}{600}.$$

Also, U is invertible if $600a_1a_2+30a_1-60a_2+1\neq 0$; its inverse is then given by the formula

$$U^{-1}x=\left(\frac{600}{600a_1a_2+30a_1-60a_2+1}\right)x,\quad x\in\mathbb{E}$$

and $\left\|U^{-1}\right\|_{\mathcal{B}(\mathbb{E})}=\frac{1}{\|U\|_{\mathcal{B}(\mathbb{E})}}$.
Using the norm estimations given in the beginning of the section and the relation $\alpha\Gamma(\alpha)=\Gamma(\alpha+1),\ \alpha>0$, we get

$$\begin{aligned}\|\Phi_1\|_{\mathcal{B}(\mathbb{E})} &\leqslant \Lambda_1\\ &:=\frac{1}{30\Gamma(\alpha+2)}\begin{pmatrix}(30\alpha^3+120\alpha^2+90\alpha)|a_1a_2|+(6\alpha^2+36\alpha+30)|a_1|\\ +(\alpha^3-\alpha+120)|a_2|+(\alpha+7)\end{pmatrix}\end{aligned}$$

and

$$\begin{aligned}\|\Phi_2\|_{\mathcal{B}(\mathbb{E})} &\leqslant \Lambda_2\\ &:=\frac{1}{20\Gamma(\alpha+2)}\begin{pmatrix}(20\alpha^3+60\alpha^2+40\alpha)|a_1a_2|+(5\alpha^2+25\alpha+20)|a_1|\\ +(\alpha^3-\alpha+60)|a_2|+(\alpha+6)\end{pmatrix}\end{aligned}$$

By (22), we can write

$$\delta \leqslant \Delta \leqslant \widetilde{\Delta} := \frac{1}{\Gamma(\alpha)} + \frac{1}{\|U\|_{\mathcal{B}(\mathbb{E})}}(\Lambda_1 + \Lambda_2). \tag{32}$$

Thus, the problem $(S1)$ has at least one solution if

$$k := 8\beta e\|u_0\|_{\mathbb{E}} < \frac{1-\sigma}{\widetilde{\Delta} - \frac{1}{\Gamma(\alpha)}} = \frac{\Gamma(\alpha)}{2\left(\Gamma(\alpha)\widetilde{\Delta} - 1\right)}.$$

If

$$k := 8\beta e\|u_0\|_{\mathbb{E}} < \frac{1-\sigma}{\widetilde{\Delta}} = \frac{1}{2\widetilde{\Delta}} < \frac{1}{2\delta}, \tag{33}$$

then the problem $(S1)$ admits a unique solution y_0 and it is Ulam–Hyers stable.

The function $\varphi(\tau) = 1 + \tau^2$, for all $\tau \in [0,1]$ is Riemann integrable and

$$\int_0^1 \varphi(\theta)d\theta = \frac{4}{3} = \delta_\varphi \leqslant \frac{4}{3}\varphi(\tau) \quad (0 \leqslant \tau \leqslant 1).$$

Here is the graphical representation of the function $\varphi(\tau) = 1 + \tau^2$ on $[0,1]$.

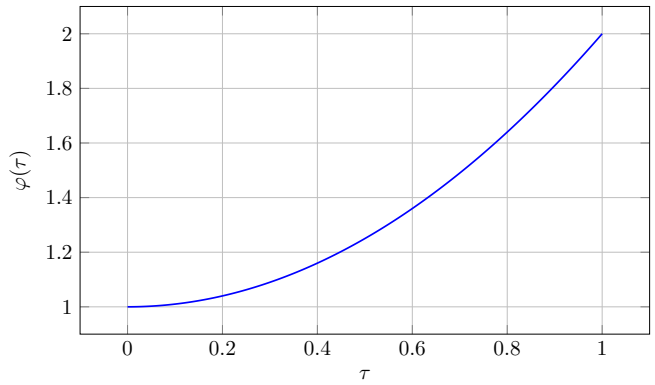


Figure 4: Plot of $\varphi(\tau) = 1 + \tau^2$.

Figure 4 confirms that the function is increasing.

Thus, for k satisfying (33), the problem $(S1)$ is Ulam–Hyers–Rassias stable with respect to φ.

The following numerical simulations illustrate the behavior of the solution corresponding to Example 1 for several integer and noninteger fractional orders $\alpha \geqslant 3$. Figure 5 presents the two-dimensional evolution of the solution $y(\tau)$ with respect to the variable τ, while Figure 6

provides a three-dimensional visualization showing the combined influence of both the fractional order α and the variable τ on the dynamics of the solution. These graphical representations clearly highlight the effect induced by the fractional operator as well as the singular behavior near $\tau = 0$.

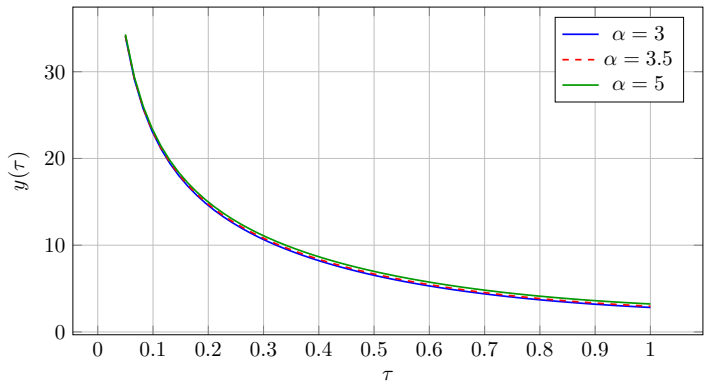


Figure 5: Numerical solution of Example 1 for integer and noninteger fractional orders $\alpha \geqslant 3$.

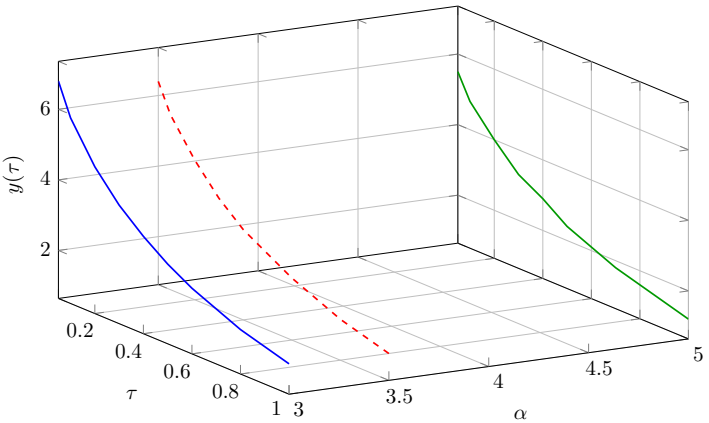


Figure 6: Three-dimensional numerical solution of Example 1 for integer and noninteger fractional orders $\alpha \geqslant 3$.

Example 2. Let now $\mathbb{E} = C(J, \mathbb{R})$ and consider in the Banach space $C(J, \mathbb{E})$ equipped with the norm

$$\|y\|_{C(J,\mathbb{E})} = \max_{\tau \in J} \|y(\tau, .)\|_{\mathbb{E}} = \max_{(\theta,\tau) \in J^2} |y(\tau, \theta)|$$

the following Caputo-fractional problem

$$(S2): \begin{cases} {}_1^C\partial_0^\alpha y(\tau,\theta)+\psi(\tau,y(\tau,\theta))=0, & 0<\tau<1,\ 0\leqslant\theta\leqslant 1, \alpha\geqslant 3,\\ y_\tau^{(j)}(r,\theta)=0, & j\notin\{3,4\},\ 0\leqslant\theta\leqslant 1,\\ \theta y(1,\theta)=\int_0^1(1-s)y(s,\theta)ds, & 0\leqslant\theta\leqslant 1,\\ \theta^2 y_\tau^{(1)}(1,\theta)=\int_0^1 y(s,\theta)ds, & 0\leqslant\theta\leqslant 1, \end{cases}$$

where ${}_1^C\partial_0^\alpha$ is the Caputo partial derivative of order α with respect to the first argument τ, $y_\tau^{(j)}(r,\theta)=\frac{d^j}{d\tau^j}y(\tau,\theta)\mid_{\tau=r}$ and the function ψ is the same used in the first example.

We firstly prove that the problem $(S2)$ can be expressed under the form (2), (3), (4). Indeed, let A_1 and A_2 be the linear operators defined in $\mathbb{E}=C(J,\mathbb{R})$ by

$$(A_1z)(\theta)=\theta z(\theta),\quad (A_2z)(\theta)=\theta^2 z(\theta),\quad \text{for all}\ \ \theta\in J.$$

Now A_1 and A_2 commute and are bounded with $\|A_1\|_{\mathcal{B}(\mathbb{E})}=\|A_2\|_{\mathcal{B}(\mathbb{E})}=1$. Using them, $(S2)$ can be rewritten in the desired form:

$$(S2): \begin{cases} {}_1^C\partial_0^\alpha y(\tau,\theta)+\psi(\tau,y(\tau,\theta))=0, & 0<\tau<1,\ 0\leqslant\theta\leqslant 1, \alpha\geqslant 3,\\ y_\tau^{(j)}(r,\theta)=0, & j\notin\{3,4\},\ 0\leqslant\theta\leqslant 1,\\ A_1y(1,.)(\theta)=\int_0^1(1-s)y(s,\theta)ds, & 0\leqslant\theta\leqslant 1\\ A_2y_\tau^{(1)}(1,.)(\theta)=\int_0^1 y(s,\theta)ds, & 0\leqslant\theta\leqslant 1. \end{cases}$$

In the present case, the operator U (formula (11)) is defined in $\mathbb{E}$ as follows:

$$Ux(\theta)=\left(\frac{600\theta^3-60\theta^2+30\theta+1}{600}\right)x(\theta).$$

The function $h(\theta)=\dfrac{600\theta^3-60\theta^2+30\theta+1}{600},\ \theta\in J$ is continuous, increases and

$$h(0)=\frac{1}{600}<h(\theta)<h(1)=\frac{571}{600},\quad 0<\theta<1.$$

Figure 7 illustrates the behavior of the function $h(\theta)$ on the interval $[0,1]$.

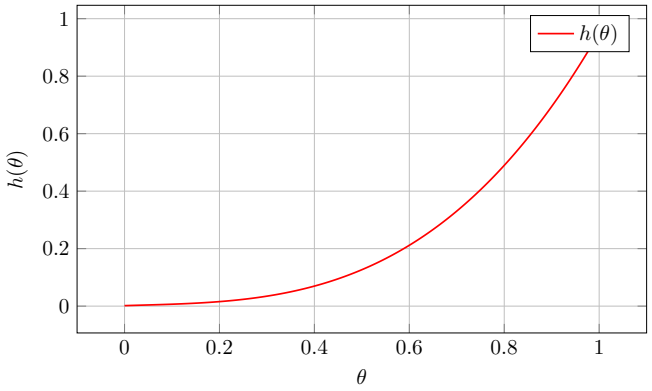


Figure 7: Plot of the function $h(\theta) = \dfrac{600\theta^3 - 60\theta^2 + 30\theta + 1}{600}$ on $[0, 1]$. The graph confirms the monotone increasing behavior of the operator coefficient appearing in Example 2.

We can thus conclude that U is bounded and $\|U\|_{\mathcal{B}(\mathbb{E})} = h(1) = \frac{571}{600}$. Moreover, U is also invertible with

$$U^{-1}x(\theta) = \frac{1}{h(\theta)}x(\theta) = \frac{600}{600\theta^3 - 60\theta^2 + 30\theta + 1}x(\theta),$$

$$\left\|U^{-1}\right\|_{\mathcal{B}(\mathbb{E})} = \frac{1}{h(0)} = 600.$$

Since $\|A_1\|_{\mathcal{B}(\mathbb{E})} = \|A_2\|_{\mathcal{B}(\mathbb{E})} = 1$, then to estimate the corresponding $\|\Phi_1\|_{\mathcal{B}(\mathbb{E})}$ and $\|\Phi_2\|_{\mathcal{B}(\mathbb{E})}$, it suffices to set $|a_1| = |a_2| = 1$ in the quantities Λ_1, Λ_2 (see the first example). This gives us

$$\|\Phi_1\|_{\mathcal{B}(\mathbb{E})} \leqslant \frac{31\alpha^3 + 126\alpha^2 + 126\alpha + 157}{30\Gamma(\alpha + 2)},$$
$$\|\Phi_2\|_{\mathcal{B}(\mathbb{E})} \leqslant \frac{21\alpha^3 + 65\alpha^2 + 65\alpha + 86}{20\Gamma(\alpha + 2)}.$$

By (22), we get

$$\delta \leqslant \Delta \leqslant \widetilde{\Delta} = \frac{1250\alpha^3 + 4470\alpha^2 + 4471\alpha + 5720}{\Gamma(\alpha + 2)}. \tag{34}$$

Thus, the problem $(S2)$ has at least one solution if

$$k := 8\beta e\|u_0\|_{\mathbb{E}} < \frac{1 - \sigma}{\widetilde{\Delta} - \frac{1}{\Gamma(\alpha)}} = \frac{\Gamma(\alpha)}{2\left(\Gamma(\alpha)\widetilde{\Delta} - 1\right)}.$$

If

$$k := 8\beta e\|u_0\|_{\mathbb{E}} < \frac{1 - \sigma}{\widetilde{\Delta}} = \frac{1}{2\widetilde{\Delta}} < \frac{1}{2\delta}, \tag{35}$$

then the problem $(S2)$ admits a unique solution y_0 and is Ulam–Hyers stable.

The function $\varphi(\tau) = 1 + \tau^2$, for all $\tau \in [0, 1]$ is Riemann integrable and

$$\int_0^1 \varphi(\theta)d\theta = \frac{4}{3} = \delta_\varphi \leqslant \frac{4}{3}\varphi(\tau), \ (0 \leqslant \tau \leqslant 1).$$

Thus, for k satisfying (35), the problem $(S2)$ is Ulam–Hyers–Rassias stable with respect to φ. Figures 8 and 9 illustrate the numerical behavior of the solution of Example 2 for several integer and noninteger fractional orders $\alpha \geqslant 3$. They show the influence of the fractional order on the dynamics of the solution and highlight the singular behavior near $\tau = 0$.

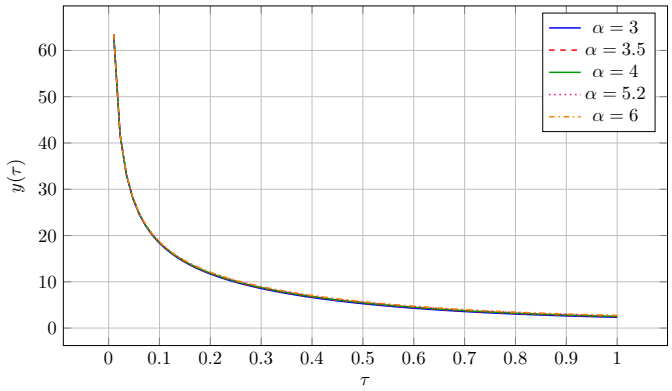


Figure 8: Numerical solution of example 2 for integer and noninteger fractional orders $\alpha \geqslant 3$.

Figure 8 illustrates the influence of the fractional order on the dynamics of the solution and highlights the effect of the singular fractional operator.

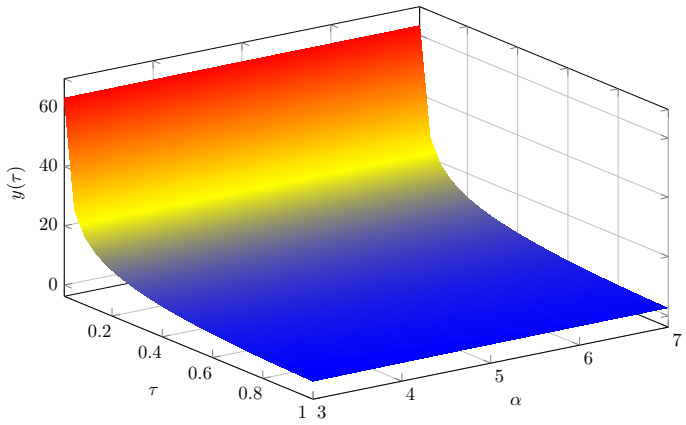


Figure 9: Three-dimensional numerical solution of Example 2 for integer and noninteger fractional orders $\alpha \geqslant 3$.

The surface illustrates the dependence of the solution $y(\tau)$ on both the variable τ and the fractional order α. Figure 9 highlights the effect induced by the fractional operator and the singular behavior near $\tau = 0$.

5 Conclusion

This article investigated a class of singular Caputo-fractional boundary value problems that involve two integral boundary conditions connected through linear operators. It employs Banach's fixed point theorem to establish a uniqueness criterion and Krasnosel'skii's fixed point theorem to obtain an existence criterion. Furthermore, the authors presented two Ulam-type stability results for the problems under consideration and illustrated the theoretical results by a concrete example. The study remains largely theoretical, and the authors propose future work to assess practical applicability by utilizing real data and performing numerical simulations.

Acknowledgment

This study is supported by the Laboratory of pure and applied Mathematics, Abdelhamid Ibn Badis University, Mostaganem, Algeria.

Funding

This research received no external funding.

Conflicts of Interest

The authors declare no conflicts of interest.

References

[1] Abdellouahab, N., Brahim, T. and Khaled, Z. *Existence and stability results of a nonlinear fractional integro-differential equation with integral boundary conditions*, Kragujev. J. Math. 46(5) (2022), 685–699.

[2] Ahmed, E., Hashish, A. and Rihan, F.A. *On fractional order cancer model*, Fract. Calc. Appl. Anal. 3(2) (2012), 1–6.

[3] Ali, A., Shah, K. and Li, Y. *Topological degree theory and Ulam's stability analysis of a boundary-value problem of fractional differential equation*, Frontiers in Functional Equations and Analytic Inequalities, Springer, Anastassiou G, Rassias J. (eds). (2019), 73–92.

[4] Area, I., Batarfi, H., Losada, J., Nieto, J.J., Shammakh, W. and Torres, Á. *On a fractional order Ebola epidemic model*, Adv. Diff. equ. 78(2015) (2015). 1–12.

[5] Arshad, S., Baleanu, D., Bu, W. and Tang, Y. *Effects of HIV infection on CD4$^+$ T-cell population based on a fractional-order model*, Adv. Diff. Equ. 92(2017) (2017), 1–14.

[6] Bagley, R.L. and Torvik, P.J. A *theoretical basis for the application of fractional calculus to viscoelasticity*, J. Rheology. 27(3) (1983), 201–210.

[7] Bagley, R.L. and Torvik, P.J. *Fractional calculus in the transient analysis of viscoelastically damped structures*, AIAA J. 23(6) (1985), 918–925.

[8] Baillie, R.T. *Long memory processes and fractional integration in econometrics*, J. Econometrics. 73 (1996), 5–59.

[9] Bochner, S. *Integration von Funktionen, deren Werte die Elemente eines Vektorraumes sind*, Fund. Math. 20 (1933), 262–276.

[10] Castro, L.P. and Silva, A.S. *Ulam–Hyers stability of four-point boundary value problem for Caputo fractional differential equations with a parameter*, Vladikavkaz Math. J. 24(4) (2022), 77–90.

[11] Chen, W.C. *Nonlinear dynamics and chaos in a fractional-order financial system*, Chaos Solit. 36(5) (2008), 1305–1314.

[12] El-Hady, E-S. and Öğreki S. *On Hyers-Ulam-Rassias stability of fractional differential equations with Caputo derivative*, J. Math. Comput. Sci. 22(4) (2021), 325–332.

[13] Ferdi, Y. *Some applications of fractional order calculus to design digital filters for biomedical signal processing*, J. Mech. Med. Biol. 12(2) (2012), 1–13.

[14] Gomez-Aguilar, J.F., López-López, M.G., Alvarado-Martínez, V.M., Baleanu, D. and Khan, H . *Chaos in a cancer model via fractional derivatives with exponential decay and Mittag-Lefler law*, Entropy. (2017), 1–19.

[15] Kheiryan, A. and Rezapour, S. *On Hyers-Ulam stability of two singular fractional integro-differential equation*, J. Adv. Math. Stud. 13(3) (2020), 339–349.

[16] Kilbas, A.A., Srivastava, H.M. and Trujillo, J.J. *Theory and applications of fractional differential equations*, North-Holland Mathematical Studies, Amsterdam: Elsevier Sciences B.V, 2006.

[17] Krasnosel'skii, M.A. *Two remarks on the method of successive approximations*, Uspekhi Mat. Nauk. 10 (1955), 123–127.

[18] Mikusński, J. *The Bochner integral*, Academic Press New York, San Fransisco, 1978.

[19] Ming, H., Wang, J.R. and Feckan, M. *The application of fractional calculus in Chinese economic growth models*, Mathematics, 7(8) (2019), 665.

[20] Morales-Delgado, V.F., Gómez-Aguilar, J.F., Saad, K. and Jiménez, R.F.E. *Application of the Caputo-Fabrizio and Atangana-Baleanu fractional derivatives to mathematical models of cancer chemotherapy effects*, Math. Meth. Appl. Sci. (2019), 1–27.

[21] Peng, S. and Wang, J.R. *Existence and Ulam–Hyers stability of ODEs involving two Caputo fractional derivatives*, Electron. J. Qual. Theory Differ. Equ. 52 (2015), 1–16.

[22] Rezapour, S. and Shabibi, M. *A singular fractional differential equation with Riemann–Liouville integral boundary condition*, J. Adv. Math. Stud. 8(4) (2015), 80–88.

[23] Rezapour, S., Mohammadi, H. and Samei, M.E. *SEIR epidemic model for COVID-19 transmission by Caputo derivative of fractional order*, Adv. Diff. Equ. 490(2020) (2020), 1–19.

[24] Rihan, F.A. *Numerical modeling of fractional-order biological systems*, Abstr. Appl. Anal. 2013 (2013), 1–11.

[25] Shabibi, M., Rezapour, S. and Vaezpour, S.M. *A singular fractional integro-differential equation*, UPB Sci. Bull. A: Appl. Math. Phys. 79(1) (2017), 109–118.

[26] Smart, D.R. *Fixed point theory*, Combridge Univ. Press, Combridge, 1974.

[27] Wang, J.R., Lv, L. and Zhou, Y. *Ulam stability and data dependance for fractional differential equations with Caputo derivative*, Electron. J. Qual. Theory Differ. Equ. (63) (2011), 1–10.

[28] Xu, H. *Analytical approximations for a population growth model with fractional order*, Commun. Nonlinear Sci. Numer. Simul. 14(5) (2009), 1978-–1983.

[29] Yosida, K. *Functional analysis*, Springer, 1965.

Iranian Journal of Numerical Analysis and Optimization
Vol. 16, No. 3, 2026, pp 1064–1082
https://doi.org/10.22067/ijnao.2026.96651.1785
https://ijnao.um.ac.ir/



Research Article



Inverse eigenvalue problems for composite banded tridiagonal matrices: Symmetric and nonsymmetric cases

M. Babaei Zarch

Abstract

We study an inverse eigenvalue problem associated with a special class of structured matrices in both symmetric and nonsymmetric forms. In the symmetric case, the problem involves two sets of eigenpairs corresponding to the original matrix and its leading principal submatrix of order $n-1$. For the nonsymmetric case, the minimum and maximum eigenvalues of each leading principal submatrix, together with an eigenpair associated with the maximum eigenvalue of the original matrix, are prescribed. By establishing recurrence relations between successive principal submatrices, we derive the necessary and sufficient conditions for the solvability of the problems. Moreover, explicit reconstruction algorithms are developed, and several numerical results are presented to illustrate the effectiveness of the proposed methods.

AMS subject classifications (2020): Primary 05C50; Secondary 65F18.

Keywords: Eigenvalue, Eigenpair, Leading principal submatrices, Inverse eigenvalue problem.

Received 23 November 2025; revised 13 February 2026; accepted 24 May 2026

Maryam Babaei Zarch
Faculty of Mathematical Sciences, Vali-e-Asr University of Rafsanjan, Kerman, Iran. e-mail:m.babaei@vru.ac.ir

How to cite this article
Babaei Zarch, M., Inverse eigenvalue problems for composite banded tridiagonal matrices: Symmetric and nonsymmetric cases. *Iran. J. Numer. Anal. Optim.*, 2026; 16(3): 1064-1082. https://doi.org/10.22067/ijnao.2026.96651.1785

1 Introduction

The inverse eigenvalue problem concerns the reconstruction of a matrix (or a class of matrices) from prescribed spectral data such as eigenvalues or eigenpairs. Because spectral properties often reflect essential physical and geometric characteristics of systems, inverse eigenvalue problems arise in various applications, including vibration analysis, structural dynamics, control theory, and signal processing. For this reason, they have been extensively studied in both theoretical and numerical contexts [1, 2].

A large portion of the existing literature focuses on inverse eigenvalue problems for structured matrices such as symmetric, tridiagonal and banded classes

[8, 11, 10, 5, 9, 7]. In this paper, we consider a class of composite banded tridiagonal matrices, which generalize classical tridiagonal matrices by combining two different structural patterns within a single matrix. This composite structure leads to recurrence relations that differ from the standard ones, thereby making the associated inverse eigenvalue problems more challenging. However, to the best of our knowledge, inverse eigenvalue problems for composite banded tridiagonal matrices of the present type have not been investigated in the existing literature. The main contributions of this work can be summarized as follows. First, we formulate two inverse eigenvalue problems for a new class of composite banded tridiagonal matrices in both symmetric and nonsymmetric settings. Second, we derive necessary and sufficient conditions for solvability and establish existence and uniqueness results. Finally, we develop explicit reconstruction algorithms and validate the proposed methods through numerical examples.

This paper is organized as follows. Section 2 presents the matrix structures and formulates the corresponding inverse eigenvalue problems. The symmetric case is analyzed in Section 3. Section 4 addresses the nonsymmetric case. In Section 5, numerical examples are provided to illustrate the effectiveness of the proposed methods, and concluding remarks are given in Section 6.

2 Problem formulation and preliminaries

In this section, we introduce the class of composite banded tridiagonal matrices and formally formulate the associated inverse eigenvalue problems. We also present the notation and preliminary results that will be used in the subsequent analysis. We consider an $n \times n$ composite banded tridiagonal matrix of the following form. In the symmetric case, the matrix A_n is given by

$$A_n = \begin{pmatrix} a_1 & b_1 & \dots & b_{m+1} & & & & \\ b_1 & a_2 & & & & & & \\ \vdots & & \ddots & & & & & \\ b_{m+1} & & & a_{m+2} & b_{m+2} & & & \\ & & & b_{m+2} & a_{m+3} & b_{m+3} & & \\ & & & & b_{m+3} & \ddots & \ddots & \\ & & & & & \ddots & \ddots & b_{n-1} \\ & & & & & & b_{n-1} & a_n \end{pmatrix}, \tag{1}$$

where $b_j \neq 0$ for $j = 1, \dots, n-1$.

In the nonsymmetric case, the matrix B_n is defined as

$$B_n = \begin{pmatrix} a_1 & c_1 & \dots & c_{m+1} & & & & \\ b_1 & a_2 & & & & & & \\ \vdots & & \ddots & & & & & \\ b_{m+1} & & & a_{m+2} & c_{m+2} & & & \\ & & & b_{m+2} & a_{m+3} & c_{m+3} & & \\ & & & & b_{m+3} & \ddots & \ddots & \\ & & & & & \ddots & \ddots & c_{n-1} \\ & & & & & & b_{n-1} & a_n \end{pmatrix}, \tag{2}$$

where $b_j c_j \neq 0$ for $j = 1, \dots, n-1$.

The matrix B_n of the form (2) is diagonally similar to the following symmetric matrix:

$$DB_nD^{-1} = \begin{pmatrix} a_1 & \sqrt{b_1c_1} & \cdots & \sqrt{b_{m+1}c_{m+1}} & & & & \\ \sqrt{b_1c_1} & a_2 & & & & & & \\ \vdots & & \ddots & & & & & \\ \sqrt{b_{m+1}c_{m+1}} & & & a_{m+2} & \sqrt{b_{m+2}c_{m+2}} & & & \\ & & & \sqrt{b_{m+2}c_{m+2}} & a_{m+3} & \sqrt{b_{m+3}c_{m+3}} & & \\ & & & & \sqrt{b_{m+3}c_{m+3}} & \ddots & \ddots & \\ & & & & & \ddots & \ddots & \sqrt{b_{n-1}c_{n-1}} \\ & & & & & & \sqrt{b_{n-1}c_{n-1}} & a_n \end{pmatrix}.$$

Also, $D = diag(1, d_2, \dots, d_n)$ such that

$$d_i = \begin{cases} \sqrt{\frac{c_{i-1}}{b_{i-1}}}, & i = 2, 3, \dots, m+1, \\ \sqrt{\prod_{j=m+1}^{i-1} \frac{c_j}{b_j}}, & i = m+2, \dots, n, \end{cases}$$

This implies that, under the transformation, the eigenvalues of the leading principal submatrices $B_j, j = 1, 2, \ldots, n$, of the matrix B_n remain invariant. The term composite refers to the fact that the matrix structure changes after the $(m+1)$th row and column. Specifically, the leading part of the matrix exhibits a bordered diagonal pattern, while the remaining part follows a classical tridiagonal structure. This hybrid form generalizes standard tridiagonal matrices.

Based on the above matrix structures, we consider the following inverse eigenvalue problems.

Problem 1 (IEP-S (Symmetric Case))**.** The real numbers $\lambda_{\max}^{(n)}$, $\lambda_{\max}^{(n-1)}$ and the real vectors $x = (x_1, \ldots, x_n)^{\mathrm{T}} \in \mathbb{R}^n$ and $x' = (x'_1, \ldots, x'_{n-1})^{\mathrm{T}} \in \mathbb{R}^{n-1}$ are given. Find an $n \times n$ symmetric matrix A_n of the form (1) such that $\lambda_{\max}^{(n)}$ and $\lambda_{\max}^{(n-1)}$ are, respectively, the maximum eigenvalues of A_n and its leading principal submatrix A_{n-1}, and $(\lambda_{\max}^{(n)}, x)$ and $(\lambda_{\max}^{(n-1)}, x')$ are eigenpairs of A_n and A_{n-1}, respectively.

Problem 2 (IEP-NS (Nonsymmetric Case))**.** The list of real numbers

$$\lambda_{\min}^{(j)},\ \lambda_{\max}^{(j)}, \quad j = 1, \ldots, n,$$

and the eigenpair $(\lambda_{\max}^{(n)}, x)$ are given. Find a matrix B_n of the form (2) such that $B_n x = \lambda_{\max}^{(n)} x$, and for each $j = 1, \ldots, n$, $\lambda_{\min}^{(j)}$ and $\lambda_{\max}^{(j)}$ are, respectively, the minimum and maximum eigenvalues of the leading principal submatrix B_j of B_n.

The considered inverse eigenvalue problems naturally arise in applications where partial spectral information of a structured system is available. For instance, in vibration analysis and structural dynamics, extremal eigenvalues of principal substructures may be measured experimentally, while only limited eigenvector information of the full system is accessible. In such situations, reconstructing the underlying matrix model from partial spectral data becomes essential. The composite banded structure considered here may represent systems composed of two interconnected subsystems with different internal configurations.

To prepare for the analysis in the subsequent sections, we first recall several known lemmas from the literature that will be used in the proofs of our results.

Lemma 1. [6] Let $\phi(\lambda)$ be a monic polynomial of degree n with all real zeroes. If $\lambda_{\min}^{(n)}$ and $\lambda_{\max}^{(n)}$ are, respectively, the minimum and the maximum zero of $\phi(\lambda)$, then

i. If $x < \lambda_{\min}^{(n)}$, we have $(-1)^n \phi(x) > 0$.

ii. If $x > \lambda_{\max}^{(n)}$, we have $\phi(x) > 0$.

Lemma 2. [3] Let A_n be an $n \times n$ matrix of the form (1) and let $\lambda_{\min}^{(j)}$, $\lambda_{\max}^{(j)}$ be the minimum and maximum eigenvalue of the leading principal submatrix A_j of A_n, $j = 1, 2, \ldots, n$. Then

$$\lambda_{\min}^{(n)} < \lambda_{\min}^{(n-1)} < \cdots < \lambda_{\min}^{(2)} < \lambda_{\min}^{(1)} < \lambda_{\max}^{(2)} < \cdots < \lambda_{\max}^{(n)} \tag{3}$$

and

$$\lambda_{\min}^{(j)} < a_k < \lambda_{\max}^{(j)}, \quad k = 1, 2, \ldots, j; \quad j = 2, \ldots, n. \tag{4}$$

Let $\{\phi_j(\lambda) = \det(\lambda I_j - A_j)\}_{j=1}^{n}$ be the sequence of characteristic polynomials of A_n. We obtain the recurrence relation $\phi_j(\lambda)$ of the matrix A_n in Lemma 3.

Lemma 3. [3] Let A_n be an $n \times n$ matrix of the form (1). Then the sequence of $\{\phi_j(\lambda)\}_{j=1}^{n}$ satisfies the following recurrence relations:

i. $\phi_1(\lambda) = (\lambda - a_1)$,

ii. $\phi_j(\lambda) = (\lambda - a_j)\phi_{j-1}(\lambda) - b_{j-1}^2 \prod_{k=2}^{j-1}(\lambda - a_k), \quad j = 2, \ldots, m+2$,

iii. $\phi_j(\lambda) = (\lambda - a_j)\phi_{j-1}(\lambda) - b_{j-1}^2 \phi_{j-2}(\lambda), \quad j = m+3, \ldots, n$,

where $\phi_0(\lambda) = 1$.

3 Formulation and solution of the IEP-S (Symmetric Case)

In this section, we present the main theoretical results of the paper. By using the preliminary lemmas established in section 2, we derive solvability conditions and reconstruction formulas for the inverse eigenvalue problem in the symmetric case.

By Lemma 2, we now derive explicit formulas for the components of the eigenvector associated with the maximum eigenvalue of the matrix and its leading principal submatrix. These relations play a key role in linking the given spectral data to the unknown matrix entries.

Lemma 4. Let $x = (x_1, x_2, \ldots, x_n)^T$ and $x' = (x'_1, x'_2, \ldots, x'_{n-1})^T$ be eigenvectors of A_n and A_{n-1} corresponding to the eigenvalues $\lambda_{\max}^{(n)}$ and $\lambda_{\max}^{(n-1)}$, respectively. Then $x_1 \neq 0$, $x'_1 \neq 0$, and the components of these eigenvectors can be expressed in terms of their first components as follows:

$$x_j = \frac{-b_{j-1}x_1}{a_j - \lambda_{\max}^{(n)}}, \qquad j = 2, \ldots, m+1, \tag{5}$$

$$x_j = \frac{(-1)^j \phi_{j-1}(\lambda_{\max}^{(n)})\, x_1}{\prod_{i=m+1}^{j-1} b_i \prod_{k=2}^{m+1} (a_k - \lambda_{\max}^{(n)})}, \qquad j = m+2, \ldots, n, \tag{6}$$

$$x'_j = \frac{-b_{j-1} x'_1}{a_j - \lambda_{\max}^{(n-1)}}, \qquad j = 2, \ldots, m+1, \tag{7}$$

$$x'_j = \frac{(-1)^j \phi_{j-1}(\lambda_{\max}^{(n-1)})\, x'_1}{\prod_{i=m+1}^{j-1} b_i \prod_{k=2}^{m+1} (a_k - \lambda_{\max}^{(n-1)})}, \qquad j = m+2, \ldots, n-1. \tag{8}$$

Proof. Let $(\lambda_{\max}^{(n)}, x)$ be an eigenpair of A_n. Then

$$A_n x = \lambda_{\max}^{(n)} x. \tag{9}$$

Comparing the first $(n-1)$ rows of (9) on both sides yields

$$(a_1 - \lambda_{\max}^{(n)}) x_1 + \sum_{j=2}^{m+2} b_{j-1} x_j = 0, \tag{10}$$

$$b_{j-1} x_1 + (a_j - \lambda_{\max}^{(n)}) x_j = 0, \qquad j = 2, \ldots, m+1, \tag{11}$$

$$b_{m+1} x_1 + (a_{m+2} - \lambda_{\max}^{(n)}) x_{m+2} + b_{m+2} x_{m+3} = 0, \tag{12}$$

$$b_{j-1} x_{j-1} + (a_j - \lambda_{\max}^{(n)}) x_j + b_j x_{j+1} = 0, \qquad j = m+3, \ldots, n-1. \tag{13}$$

By Lemma 2, $(a_j - \lambda_{\max}^{(n)}) \neq 0$ for $j = 2, \ldots, m+1$. Hence, from (11) we obtain

$$x_j = \frac{-b_{j-1} x_1}{a_j - \lambda_{\max}^{(n)}}, \qquad j = 2, \ldots, m+1,$$

which establishes (5).

Next, we determine the remaining components $x_{m+2}, \ldots, x_n$ recursively. From (10) we get

$$x_{m+2} = \frac{(\lambda_{\max}^{(n)} - a_1) x_1 - \sum_{i=2}^{m+1} b_{i-1} x_i}{b_{m+1}}.$$

Substituting the expressions for $x_2, \ldots, x_{m+1}$ from (5) yields

$$x_{m+2} = \frac{(-1)^{m+2} \phi_{m+1}(\lambda_{\max}^{(n)})\, x_1}{b_{m+1} \prod_{k=2}^{m+1} (a_k - \lambda_{\max}^{(n)})},$$

which verifies the base case of (6).

Assume now that (6) holds up to $j = n-1$. To prove it for $j = n$, consider (13) with $j = n-1$:

$$b_{n-2}x_{n-2} + (a_{n-1} - \lambda_{\max}^{(n)})x_{n-1} + b_{n-1}x_n = 0,$$

so that

$$x_n = \frac{-b_{n-2}x_{n-2} + (\lambda_{\max}^{(n)} - a_{n-1})x_{n-1}}{b_{n-1}}.$$

By the induction hypothesis,

$$x_{n-2} = \frac{(-1)^{n-2}\phi_{n-3}(\lambda_{\max}^{(n)})\, x_1}{\prod_{i=m+1}^{n-3} b_i \prod_{k=2}^{m+1}(a_k - \lambda_{\max}^{(n)})},$$

$$x_{n-1} = \frac{(-1)^{n-1}\phi_{n-2}(\lambda_{\max}^{(n)})\, x_1}{\prod_{i=m+1}^{n-2} b_i \prod_{k=2}^{m+1}(a_k - \lambda_{\max}^{(n)})}.$$

Substituting into the recurrence gives

$$x_n = \frac{(-1)^{n}\phi_{n-1}(\lambda_{\max}^{(n)})\, x_1}{\prod_{i=m+1}^{n-1} b_i \prod_{k=2}^{m+1}(a_k - \lambda_{\max}^{(n)})},$$

which proves (6).

Finally, since x is an eigenvector, $x \neq 0$. If $x_1 = 0$, then by relations (5)–(6) all components vanish, which is a contradiction. Hence $x_1 \neq 0$.

The same reasoning applies to the eigenpair $(\lambda_{\max}^{(n-1)}, x')$ of A_{n-1}: Replacing $\lambda_{\max}^{(n)}$ with $\lambda_{\max}^{(n-1)}$ and noting that indices now run up to $n-1$, we obtain the relations (7)–(8). This completes the proof. □

With the component formulas of the eigenvectors established in Lemma 4, we are now able to formulate the main result of this section. The following theorem provides conditions under which the inverse eigenvalue problem admits a unique solution and shows how the unknown entries of the matrix can be determined from the eigenvalues and eigenvector components.

Theorem 1. Problem 1 has a unique solution if the following conditions are satisfied:

a: $x_j \neq 0$, for $j = 1, 2, \ldots, n$ and $x'_j \neq 0$, for $j = 1, 2, \ldots, n-1$.

b: $E_j = \begin{vmatrix} x_1 & x_{m+2} \\ x'_1 & x'_{m+2} \end{vmatrix} \neq 0.$

c: $F_j = \begin{vmatrix} x_j & x_{j+1} \\ x'_j & x'_{j+1} \end{vmatrix} \neq 0,\ j = m+2, \ldots, n-2.$

Proof. Assume that $x_j \neq 0$ for $j = 1, 2, \ldots, n$ and $x'_j \neq 0$ for $j = 1, 2, \ldots, n-1$.

Since $(\lambda_{\max}^{(n)}, x)$ and $(\lambda_{\max}^{(n-1)}, x')$ are eigenpairs of the matrices A_n and A_{n-1}, respectively, it follows that for $j = 2, \ldots, m+1$ the following relations hold:

$$\begin{cases} b_{j-1}x_1 + (a_j - \lambda_{\max}^{(n)})x_j = 0, \\ b_{j-1}x'_1 + (a_j - \lambda_{\max}^{(n-1)})x'_j = 0. \end{cases}$$

Let D_j denote the determinant of the coefficient matrix of the above system. Then

$$D_j = x_1 x'_j - x'_1 x_j.$$

If $D_j \neq 0$, the system admits a unique solution for a_j and b_{j-1}, given by

$$b_{j-1} = \frac{(\lambda_{\max}^{(n)} - \lambda_{\max}^{(n-1)})x'_j x_j}{x_1 x'_j - x'_1 x_j},$$
$$a_j = \lambda_{\max}^{(n)} - \frac{b_{j-1}x_1}{x_j}.$$

According to Lemma 4, we have

$$D_j = x_1 \frac{-b_{j-1}x'_1}{(a_j - \lambda_{\max}^{(n-1)})} - x'_1 \frac{-b_{j-1}x_1}{(a_j - \lambda_{\max}^{(n)})} = \frac{-x_1 x'_1 b_{j-1}(\lambda_{\max}^{(n)} - \lambda_{\max}^{(n-1)})}{(a_j - \lambda_{\max}^{(n-1)})(a_j - \lambda_{\max}^{(n)})}.$$

By Lemma 2 and relation (3), it follows that

$$(\lambda_{\max}^{(n)} - \lambda_{\max}^{(n-1)}) \neq 0,$$

and thus $D_j \neq 0$. Since $(\lambda_{\max}^{(n)} - \lambda_{\max}^{(n-1)})x'_j x_j \neq 0$, we conclude that $b_{j-1} \neq 0$.

To determine a_1 and b_{m+1}, consider

$$\begin{cases} (a_1 - \lambda_{\max}^{(n)})x_1 + \sum_{j=2}^{m+2} b_{j-1}x_j = 0, \\ (a_1 - \lambda_{\max}^{(n-1)})x'_1 + \sum_{j=2}^{m+2} b_{j-1}x'_j = 0. \end{cases}$$

Since b_j for $j = 1, \ldots, m$ are known, this system can be solved uniquely for b_{m+1} and a_1 (as $E_j \neq 0$) as follows:

$$b_{m+1} = \frac{(\lambda_{\max}^{(n-1)} - \lambda_{\max}^{(n)}) \sum_{j=1}^{m+1} x_j x'_j}{E_j},$$

$$a_1 = \lambda_{\max}^{(n)} - \frac{\sum_{j=2}^{m+2} b_{j-1} x_j}{x_1}.$$

From Lemma 4, we obtain

$$x_j x'_j = \frac{b_{j-1}^2 x_1 x'_1}{(a_j - \lambda_{\max}^{(n)})(a_j - \lambda_{\max}^{(n-1)})}.$$

Moreover, by Lemma 2, relation (4), we have

$$\frac{b_{j-1}^2}{(a_j - \lambda_{\max}^{(n)})(a_j - \lambda_{\max}^{(n-1)})} > 0.$$

Therefore, the signs of $x_j x'_j$ and $x_1 x'_1$ coincide. Hence, $\sum_{j=1}^{m+1} x_j x'_j \neq 0$, and thus $b_{m+1} \neq 0$. To determine the coefficients a_{m+2} and b_{m+2}, consider the linear system

$$\begin{cases} b_{m+1} x_1 + (a_{m+2} - \lambda_{\max}^{(n)}) x_{m+2} + b_{m+2} x_{m+3} = 0, \\ b_{m+1} x'_1 + (a_{m+2} - \lambda_{\max}^{(n-1)}) x'_{m+2} + b_{m+2} x'_{m+3} = 0. \end{cases}$$

Since $F_{m+2} \neq 0$ by condition (c), the system has a unique solution. Therefore

$$b_{m+2} = \frac{(\lambda_{\max}^{(n-1)} - \lambda_{\max}^{(n)}) \sum_{j=1}^{m+2} x_j x'_j}{F_{m+2}},$$

$$a_{m+2} = \lambda_{\max}^{(n)} - \frac{b_{m+1} x_1 + b_{m+2} x_{m+3}}{x_{m+2}}.$$

Thus, both coefficients are uniquely determined once b_{m+1} is known.

By (13) for $j = m+3, \ldots, n-2$ we have

$$\begin{cases} b_{j-1} x_{j-1} + (a_j - \lambda_{\max}^{(n)}) x_j + b_j x_{j+1} = 0, \\ b_{j-1} x'_{j-1} + (a_j - \lambda_{\max}^{(n-1)}) x'_j + b_j x'_{j+1} = 0. \end{cases}$$

by condition (c), the system has a unique solution.

$$b_j = \frac{(\lambda_{\max}^{(n-1)} - \lambda_{\max}^{(n)}) \sum_{i=1}^{j} x_i x'_i}{F_j},$$

$$a_j = \lambda_{\max}^{(n)} - \frac{b_{j-1}x_{j-1} + b_j x_{j+1}}{x_j}.$$

Finally, to determine a_{n-1}, note that

$$b_{n-2}x'_{n-2} + (a_{n-1} - \lambda_{\max}^{(n-1)})x'_{n-1} = 0,$$

which implies

$$a_{n-1} = \lambda_{\max}^{(n-1)} - \frac{b_{n-2}x'_{n-2}}{x'_{n-1}}.$$

Using (13), we obtain

$$b_{n-1} = \frac{-(b_{n-2}x_{n-2} + (a_{n-1} - \lambda_{\max}^{(n)})x_{n-1})}{x_n} = \frac{(\lambda_{\max}^{(n)} - \lambda_{\max}^{(n-1)})\sum_{i=1}^{n-1} x_i x'_i}{x_n x'_{n-1}},$$

and consequently,

$$a_n = \lambda_{\max}^{(n)} - \frac{b_{n-1}x_{n-1}}{x_n}.$$

□

Theorem 1 shows that the solvability of Problem 1 depends only on the nonvanishing of certain eigenvector components and determinant conditions involving the given spectral data. These conditions ensure that each step of the reconstruction process is well defined and that the unknown matrix entries can be uniquely determined. In particular, the theorem guarantees that the prescribed eigenpairs contain sufficient information to recover the entire matrix.

4 Formulation and solution of the IEP-NS (Nonsymmetric case)

In this section, we study the inverse eigenvalue problem for the nonsymmetric matrix B_n of the form (2). To establish the main result, we first present some preliminary relations concerning the characteristic polynomials of the principal submatrices of B_n. These relations form the basis for deriving expressions of the eigenvectors associated with the dominant eigenvalue, which are essential for the subsequent analysis.

Lemma 5. Let B_n be an $n \times n$ matrix of the form (2). Then the sequence of $\{\phi_j(\lambda)\}_{j=1}^n$ satisfies the following recurrence relations:

i. $\phi_1(\lambda) = (\lambda - a_1)$,

ii. $\phi_j(\lambda) = (\lambda - a_j)\phi_{j-1}(\lambda) - b_{j-1}c_{j-1}\prod_{k=2}^{j-1}(\lambda - a_k), \quad j = 2, \ldots, m+2,$

iii. $\phi_j(\lambda) = (\lambda - a_j)\phi_{j-1}(\lambda) - b_{j-1}c_{j-1}\phi_{j-2}(\lambda), \quad j = m+3, \ldots, n,$

where $\phi_0(\lambda) = 1$.

Proof. The relations can be verified by expanding the determinant. □

Lemma 6. Let $x = (x_1, x_2, \ldots, x_n)^T$ denote an eigenvector of the matrix B_n corresponding to the eigenvalue $\lambda_{\max}^{(n)}$. Then $x_1 \neq 0$, and the components of this eigenvector can be expressed as

$$x_j = \frac{-b_{j-1}x_1}{a_j - \lambda_{\max}^{(n)}}, \qquad j = 2, \ldots, m+1, \tag{14}$$

$$x_j = \frac{(-1)^j \phi_{j-1}(\lambda_{\max}^{(n)})\, x_1}{\prod_{i=m+1}^{j-1} c_i \prod_{k=2}^{m+1}(a_k - \lambda_{\max}^{(n)})}, \qquad j = m+2, \ldots, n. \tag{15}$$

Proof. The proof follows the same reasoning as in Lemma 4, and is therefore omitted for brevity. □

To proceed toward the reconstruction of the matrix elements, we make use of the following result from [3], which establishes an interlacing property and explicit formulas for the entries of a symmetric matrix. This theorem will play a key role in proving our main result for the nonsymmetric case.

Theorem 2. [3] Let

$$\sigma = (\lambda_{\min}^{(1)}, \lambda_{\min}^{(2)}, \lambda_{\max}^{(2)}, \ldots, \lambda_{\min}^{(n)}, \lambda_{\max}^{(n)}),$$

be a list of real numbers. Then the list σ can be interlacingly realized by an $n \times n$ symmetric matrix A_n of the form (1) if and only if

$$\lambda_{\min}^{(n)} < \lambda_{\min}^{(n-1)} < \cdots < \lambda_{\min}^{(2)} < \lambda_{\min}^{(1)} < \lambda_{\max}^{(2)} < \cdots < \lambda_{\max}^{(n)}.$$

The entries of an $n \times n$ symmetric matrix A of the form (1) as follows:

$$a_1 = \lambda_{\min}^{(1)},$$

$$a_j = \frac{\lambda_{\min}^{(j)}\,\phi_{j-1}(\lambda_{\min}^{(j)})\prod_{i=2}^{j-1}(\lambda_{\max}^{(j)} - a_i) - \lambda_{\max}^{(j)}\,\phi_{j-1}(\lambda_{\max}^{(j)})\prod_{i=2}^{j-1}(\lambda_{\min}^{(j)} - a_i)}{\phi_{j-1}(\lambda_{\min}^{(j)})\prod_{i=2}^{j-1}(\lambda_{\max}^{(j)} - a_i) - \phi_{j-1}(\lambda_{\max}^{(j)})\prod_{i=2}^{j-1}(\lambda_{\min}^{(j)} - a_i)},$$

$$b_{j-1}^2 = \frac{(\lambda_{\max}^{(j)} - \lambda_{\min}^{(j)})\,\phi_{j-1}(\lambda_{\min}^{(j)})\,\phi_{j-1}(\lambda_{\max}^{(j)})}{\phi_{j-1}(\lambda_{\min}^{(j)})\prod_{i=2}^{j-1}(\lambda_{\max}^{(j)} - a_i) - \phi_{j-1}(\lambda_{\max}^{(j)})\prod_{i=2}^{j-1}(\lambda_{\min}^{(j)} - a_i)},$$

for any $j = 2, \ldots, m$ and

$$a_j = \frac{\lambda_{\min}^{(j)}\phi_{j-1}(\lambda_{\min}^{(j)})\phi_{j-2}(\lambda_{\max}^{(j)}) - \lambda_{\max}^{(j)}\phi_{j-1}(\lambda_{\max}^{(j)})\phi_{j-2}(\lambda_{\min}^{(j)})}{\phi_{j-1}(\lambda_{\min}^{(j)})\,\phi_{j-2}(\lambda_{\max}^{(j)}) - \phi_{j-1}(\lambda_{\max}^{(j)})\,\phi_{j-2}(\lambda_{\min}^{(j)})},$$

$$b_{j-1}^2 = \frac{(\lambda_{\max}^{(j)} - \lambda_{\min}^{(j)})\phi_{j-1}(\lambda_{\min}^{(j)})\,\phi_{j-1}(\lambda_{\max}^{(j)})}{\phi_{j-1}(\lambda_{\min}^{(j)})\,\phi_{j-2}(\lambda_{\max}^{(j)}) - \phi_{j-1}(\lambda_{\max}^{(j)})\,\phi_{j-2}(\lambda_{\min}^{(j)})},$$

for $j = m+1, \ldots, n$.

The next step is to extend these relations to the nonsymmetric setting. For this purpose, we introduce a set of auxiliary quantities that simplify the notation and facilitate the proof of the main theorem.

$$\begin{aligned}
M_j &= \lambda_{\min}^{(j)}\phi_{j-1}(\lambda_{\min}^{(j)})\prod_{i=2}^{j-1}(\lambda_{\max}^{(j)} - a_i),\\
N_j &= \lambda_{\max}^{(j)}\phi_{j-1}(\lambda_{\max}^{(j)})\prod_{i=2}^{j-1}(\lambda_{\min}^{(j)} - a_i),\\
P_j &= (\lambda_{\max}^{(j)} - \lambda_{\min}^{(j)})\phi_{j-1}(\lambda_{\min}^{(j)})\phi_{j-1}(\lambda_{\max}^{(j)}),\\
D_j &= \phi_{j-1}(\lambda_{\min}^{(j)})\prod_{i=2}^{j-1}(\lambda_{\max}^{(j)} - a_i) - \phi_{j-1}(\lambda_{\max}^{(j)})\prod_{i=2}^{j-1}(\lambda_{\min}^{(j)} - a_i),\\
S_j &= \lambda_{\min}^{(j)}\phi_{j-1}(\lambda_{\min}^{(j)})\phi_{j-2}(\lambda_{\max}^{(j)}),\\
T_j &= \lambda_{\max}^{(j)}\phi_{j-1}(\lambda_{\max}^{(j)})\phi_{j-2}(\lambda_{\min}^{(j)}),\\
D'_j &= \phi_{j-1}(\lambda_{\min}^{(j)})\phi_{j-2}(\lambda_{\max}^{(j)}) - \phi_{j-1}(\lambda_{\max}^{(j)})\phi_{j-2}(\lambda_{\min}^{(j)}).
\end{aligned}$$

We are now ready to state and prove the main theorem of this section, which establishes the existence and uniqueness of the solution to Problem 2 under the given spectral conditions.

Theorem 3. If the ordered spectral relation

$$\lambda_{\min}^{(n)} < \lambda_{\min}^{(n-1)} < \cdots < \lambda_{\min}^{(2)} < \lambda_{\min}^{(1)} < \lambda_{\max}^{(2)} < \cdots < \lambda_{\max}^{(n)}, \tag{16}$$

holds, and the component x_j of the corresponding eigenvector is nonzero (i.e., $x_j \neq 0$) for all $j = 1, 2, \ldots, n$, then Problem 2 admits a unique solution.

Proof. To prove this theorem, we observe that Problem 2 is equivalent to solving the system of (17), (18), and (19) for the unknown parameters.

$$\phi_j(\lambda_{\max}^{(j)}) = 0, \quad j = 1, 2, \ldots, n, \tag{17}$$

$$\phi_j(\lambda_{\min}^{(j)}) = 0, \quad j = 1, 2, \ldots, n, \tag{18}$$

$$B_n x = \lambda_{\max}^{(n)} x. \tag{19}$$

It is immediate that $a_1 = \lambda_{\min}^{(1)}$. The following subsystem, defined by (17) and (18) for $j = 2, \ldots, m+2$.

$$\begin{cases} \phi_j(\lambda_{\max}^{(j)}) = a_j \phi_{j-1}(\lambda_{\max}^{(j)}) + b_{j-1}c_{j-1} \prod_{k=2}^{j-1}(\lambda_{\max}^{(j)} - a_k) \\ \qquad\qquad -\lambda_{\max}^{(j)} \phi_{j-1}(\lambda_{\max}^{(j)}) = 0, \\ \phi_j(\lambda_{\min}^{(j)}) = a_j \phi_{j-1}(\lambda_{\min}^{(j)}) + b_{j-1}c_{j-1} \prod_{k=2}^{j-1}(\lambda_{\min}^{(j)} - a_k) \\ \qquad\qquad -\lambda_{\min}^{(j)} \phi_{j-1}(\lambda_{\min}^{(j)}) = 0. \end{cases} \tag{20}$$

By Theorem 2 and (20), we can determine the coefficients a_j and the product $b_{j-1}c_{j-1}$ as follows:

$$a_j = \frac{M_j - N_j}{D_j},$$

$$b_{j-1}c_{j-1} = \frac{P_j}{D_j}.$$

Using the eigenvalue condition $B_n x = \lambda_{\max}^{(n)} x$, and substituting into (11), we obtain

$$b_{j-1} = (\lambda_{\max}^{(n)} - a_j)\frac{x_j}{x_1}.$$

Substituting the obtained expressions for a_j and b_{j-1} back into the first equation of system (20) yields

$$c_{j-1} = \frac{(\lambda_{\max}^{(j)} - a_j)\phi_{j-1}(\lambda_{\max}^{(j)})}{b_{j-1} \prod_{k=2}^{j-1}(\lambda_{\max}^{(j)} - a_k))}.$$

Furthermore, from (16) and Lemma 1, we confirm the positivity condition:

$$b_{j-1}c_{j-1} = \frac{(-1)^{j-1}P_j}{(-1)^{j-1}D_j} > 0.$$

Next, using Lemma 5, the conditions in (17) and (18) can be reformulated as the following system for $j = m+3, \ldots, n$:

$$\begin{cases} a_j \phi_{j-1}(\lambda_{\max}^{(j)}) + b_{j-1}c_{j-1}\phi_{j-2}(\lambda_{\max}^{(j)}) - \lambda_{\max}^{(j)}\phi_{j-1}(\lambda_{\max}^{(j)}) = 0, \\ a_j \phi_{j-1}(\lambda_{\min}^{(j)}) + b_{j-1}c_{j-1}\phi_{j-2}(\lambda_{\min}^{(j)}) - \lambda_{\min}^{(j)}\phi_{j-1}(\lambda_{\min}^{(j)}) = 0. \end{cases} \tag{21}$$

By Theorem 2 solving (21) for a_j and the product $b_{j-1}c_{j-1}$ yields

$$a_j = \frac{S_j - T_j}{D'_j},$$
$$b_{j-1}c_{j-1} = \frac{(\lambda_{\max}^{(j)} - \lambda_{\min}^{(j)})\phi_{j-1}(\lambda_{\min}^{(j)})\phi_{j-1}(\lambda_{\max}^{(j)})}{D'_j}.$$

Furthermore, b_{j-1} can be determined recursively from (15) as

$$c_{j-1} = \frac{(-1)^j \phi_{j-1}(\lambda_{\max}^{(n)})x_1}{\prod_{i=m+1}^{j-2} c_i \prod_{k=2}^{m+1}(a_k - \lambda_{\max}^{(n)})x_j}.$$

Finally, substituting the expressions for a_j and c_{j-1} into (17), we obtain

$$b_{j-1} = \frac{(\lambda_{\max}^{(j)} - a_j)\phi_{j-1}(\lambda_{\max}^{(j)})}{c_{j-1}\phi_{j-2}(\lambda_{\max}^{(j)})}.$$

The positivity of the resulting products follows from (16) and Lemma 1. Since all the required parameters are uniquely determined by the given spectral data and the components of the eigenvector, the solution to Problem 2 is unique. □

Theorem 3 establishes that the given extremal eigenvalues of the leading principal submatrices, together with a single eigenpair of the full matrix, uniquely determine the nonsymmetric matrix B_n. The interlacing condition plays a crucial role in ensuring the positivity of the reconstructed parameters and the consistency of the solution.

5 Algorithms and numerical examples

In this section, we present the solution algorithms for Problems 1 and 2, together with numerical examples.

Algorithm 1 To solve Problem 1

Require: $\lambda_{\max}^{(n)}$, $\lambda_{\max}^{(n-1)}$, ε, $x=(x_1,\ldots,x_n)$, $x'=(x'_1,\ldots,x'_{n-1})$
Ensure: A_n
  for $j=2$ **to** $m+1$ **do**
    $b_{j-1}=\frac{(\lambda_{\max}^{(n)}-\lambda_{\max}^{(n-1)})x'_jx_j}{x_1x'_j-x'_1x_j}$.
    $a_j=\lambda_{\max}^{(n)}-\frac{b_{j-1}x_1}{x_j}$.
  end for
  if $|x_1x'_{m+2}-x'_1x_{m+2}|<\varepsilon$ **then**
    Problem cannot be solved by this algorithm.
  end if
  $b_{m+1}=\frac{(\lambda_{\max}^{(n-1)}-\lambda_{\max}^{(n)})\sum_{j=1}^{m+1}x_jx'_j}{E_j}$
  $a_1=\lambda_{\max}^{(n)}-\frac{\sum_{j=2}^{m+2}b_{j-1}x_j}{x_1}$.
  if $|x_{m+2}x'_{m+3}-x'_{m+2}x_{m+3}|<\varepsilon$ **then**
    Problem cannot be solved by this algorithm.
  end if
  $b_{m+2}=\frac{(\lambda_{\max}^{(n-1)}-\lambda_{\max}^{(n)})\sum_{j=1}^{m+2}x_jx'_j}{F_j}$
  $a_{m+2}=\lambda_{\max}^{(n)}-\frac{b_{m+1}x_1+b_{m+2}x_{m+3}}{x_{m+2}}$.
  for $j=m+3$ **to** $n-2$ **do**
    if $|x_jx'_{j+1}-x'_jx_{j+1}|<\varepsilon$ **then**
      Problem cannot be solved by this algorithm.
    end if
    $b_j=\frac{(\lambda_{\max}^{(n-1)}-\lambda_{\max}^{(n)})\sum_{i=1}^{j}x_ix'_i}{F_j}$
    $a_j=\lambda_{\max}^{(n)}-\frac{b_{j-1}x_{j-1}+b_jx_{j+1}}{x_j}$.
  end for
  $a_{n-1}=\lambda_{\max}^{(n-1)}-\frac{b_{n-2}x'_{n-2}}{x'_{n-1}}$.
  $b_{n-1}=\frac{(\lambda_{\max}^{(n-1)}-\lambda_{\max}^{(n)})\sum_{i=1}^{n-1}x_ix'_i}{x_nx'_{n-1}}$,
  $a_n=\lambda_{\max}^{(n)}-\frac{b_{n-1}x_{n-1}}{x_n}$.

Example 1. Consider two distinct real numbers $\lambda_{\max}^{(7)}=9.8796$ and $\lambda_{\max}^{(6)}=7.1115$, together with the real vectors

$$\mathbf{x}=(0.0453,\ 0.0066,\ 0.0046,\ 0.0084,\ 0.2260,\ 0.7525,\ 0.6168)^T,$$

and

$$\mathbf{x}'=(0.21946,\ 0.0534,\ 0.0343,\ 0.0935,\ 0.4707,\ 0.8470)^T.$$

Determine a 7×7 matrix of the form (1) such that $(\lambda_{\max}^{(7)},\mathbf{x})$ forms an eigenpair of A_7 and $(\lambda_{\max}^{(6)},\mathbf{x}')$ is an eigenpair of A_6.

Solution: Using Algorithm 1, we obtain the following unique matrix:

$$A_7 = \begin{bmatrix} 4 & 1 & 0.8 & 0.9 & 1.1 & 0 & 0 \\ 1 & 3 & 0 & 0 & 0 & 0 & 0 \\ 0.8 & 0 & 2 & 0 & 0 & 0 & 0 \\ 0.9 & 0 & 0 & 5 & 0 & 0 & 0 \\ 1.1 & 0 & 0 & 0 & 3 & 2 & 0 \\ 0 & 0 & 0 & 0 & 2 & 6 & 4 \\ 0 & 0 & 0 & 0 & 0 & 4 & 5 \end{bmatrix}.$$

To verify the correctness of this result, the eigenvalues of A_6 and A_7 are recalculated from the above matrix. The obtained spectra are

$$\sigma(A_6) = \{\underline{7.1115},\ 5.7253,\ 4.2778,\ 2.6905,\ 2.0000,\ 1.1947\},$$
$$\sigma(A_7) = \{\underline{9.8796},\ 5.8562,\ 4.4413,\ 3.2988,\ 2.4780,\ 1.5593,\ 0.4865\}.$$

The underlined eigenvalues correspond to the prescribed maximum eigenvalues, confirming the consistency of the constructed matrix with the given spectral data.

Algorithm 2 To solve Problem 2

Require: Spectral data $\{\lambda_{\min}^{(j)}, \lambda_{\max}^{(j)}\}_{j=1}^{n}$, eigenvector $x = (x_1, \ldots, x_n)$ with $x_j \neq 0$, integer m.

for $j = 2$ **to** $m+2$ **do**

compute auxiliary scalars M_j, N_j, P_j, D_j from spectral data and $\{\phi_k\}$

$a_j \leftarrow \dfrac{M_j - N_j}{D_j}$.

$b_{j-1} \leftarrow (\lambda_{\max}^{(n)} - a_j)\dfrac{x_j}{x_1}$.

$c_{j-1} \leftarrow \dfrac{(\lambda_{\max}^{(j)} - a_j)\,\phi_{j-1}(\lambda_{\max}^{(j)})}{b_{j-1}\prod_{k=2}^{j-1}(\lambda_{\max}^{(j)} - a_k)}$.

end for

for $j = m+3$ **to** n **do**

$a_j \leftarrow \dfrac{S_j - T_j}{D'_j}$.

$c_{j-1} \leftarrow \dfrac{(-1)^j\,\phi_{j-1}(\lambda_{\max}^{(n)})\,x_1}{\left(\prod_{i=m+1}^{j-2} c_i\right)\left(\prod_{k=2}^{m+1}(a_k - \lambda_{\max}^{(n)})\right)x_j}$.

$b_{j-1} \leftarrow \frac{(\lambda_{\max}^{(j)} - a_j)\phi_{j-1}(\lambda_{\max}^{(j)})}{c_{j-1}\phi_{j-2}(\lambda_{\max}^{(j)})}$.

end for

Example 2. Given the 15 real numbers shown in Table 1, which displays the minimum and maximum eigenvalues of the leading principal submatrices B_j, and the real vector

$$x = (-0.0664, -0.0109, -0.0198, -0.2552, -0.4714, -0.6657, -0.4745, -0.1985)^T,$$

assume that the conditions of Theorem 3 are satisfied.

Table 1: minimum and maximum eigenvalues of leading principal submatrices B_j.

j	1	2	3	4	5	6	7	8
$\lambda_{\min}^{(j)}$	4	2.3170	1.4499	0.5659	0.4374	0.3939	0.2575	-1.4229
$\lambda_{\max}^{(j)}$	4	8.1830	10.8934	14.3663	16.7496	18.5961	19.6192	19.9541

Find a matrix B_8 of the form (2) such that $\lambda_{\min}^{(j)}$ and $\lambda_{\max}^{(j)}$ are the minimum and maximum eigenvalues of B_j, $j = 1, 2, \ldots, 8$, moreover $(\lambda_{\max}^{(8)}, x)$ is its eigenpair.

Solution: By applying Algorithm 2, we obtain the following matrix

$$B_8 = \begin{bmatrix} 4.0 & 3.2 & 2.8 & 3.8 & 0 & 0 & 0 & 0 \\ 2.2 & 6.5 & 0 & 0 & 0 & 0 & 0 & 0 \\ 3.2 & 0 & 9.2 & 0 & 0 & 0 & 0 & 0 \\ 3.8 & 0 & 0 & 12.5 & 3.5 & 0 & 0 & 0 \\ 0 & 0 & 0 & 4.8 & 11 & 4.5 & 0 & 0 \\ 0 & 0 & 0 & 0 & 6.2 & 12 & 5 & 0 \\ 0 & 0 & 0 & 0 & 0 & 7.8 & 6 & 7.2 \\ 0 & 0 & 0 & 0 & 0 & 0 & 5 & 8 \end{bmatrix}.$$

The eigenvalues of the leading principal submatrices are listed in the Table 2. The underlined values represent the minimum and maximum eigenvalues of these leading principal submatrices.

Table 2: Eigenvalues of the leading principal submatrices of B_8

Submatrix	**Eigenvalues**
B_1	$\{\underline{4.0000}\}$
B_2	$\{\underline{2.3170}, \underline{8.1830}\}$
B_3	$\{\underline{1.4499}, 7.3567, \underline{10.8934}\}$
B_4	$\{\underline{0.5659}, 7.1812, 10.0865, \underline{14.3663}\}$
B_5	$\{\underline{0.4374}, 6.9207, 8.1816, 10.9108, \underline{16.7496}\}$
B_6	$\{\underline{0.3939}, 5.1089, 7.2774, 10.2381, 13.5856, \underline{18.5961}\}$
B_7	$\{\underline{0.2575}, 1.1955, 6.8389, 7.7846, 10.5835, 14.9208, \underline{19.6192}\}$
B_8	$\{\underline{-1.4229}, 0.4466, 5.7080, 7.2957, 10.0053, 11.6269, 15.5863, \underline{19.9541}\}$

The underlined eigenvalues are the minimum and maximum eigenvalues, and

$$B_8 x = 19.9541x.$$

Although the presented examples are of moderate size, they illustrate all essential steps of the reconstruction procedure. Since the proposed algorithm is based on explicit recurrence relations and closed form expressions, its implementation does not depend on the matrix dimension in a structural sense. Therefore, the same reconstruction process applies to larger matrices without modification. In practice, the numerical computations remain stable whenever the solvability conditions of Theorem 1 and Theorem 3 are satisfied.

6 Conclusions

In this paper, we addressed two partially described inverse eigenvalue problems for a special class of structured matrices. By utilizing limited spectral data and establishing recurrence relations between successive principal submatrices, we derived necessary and sufficient solvability conditions and proposed explicit reconstruction algorithms. Numerical experiments confirmed the effectiveness of the developed methods, demonstrating that such inverse problems can be successfully solved even when only partial eigendata are available.

Funding

This work was supported by a grant from Vali-e-Asr university of Rafsanjan under Pata research contract.

Conflicts of Interest

The author declares no conflicts of interest.

References

[1] Arela-Pérez, S., Egaña, J., Pasten, G. and Pickmann-Soto, H., *Extremal realization spectra by two acyclic matrices whose graphs are caterpillars*, Linear Multilinear Algebra, 71(10) (2022), 1657–1680.

[2] Arela-Pérez, S., Lozano, C., Nina, H., Pickmann-Soto, H. and Rodriguez, J., *The new inverse eigenvalue problems for periodic and generalized periodic Jacobi matrices from their extremal spectral data*, Linear Algebra Appl. 659 (2023), 55–72.

[3] Chen, G.-L. and Xu, W.-R., *On inverse eigenvalue problems for two kinds of special banded matrices*, Filomat, 31(2) (2017), 371–385.

[4] Chu, M. T. and Golub, H., *Inverse eigenvalue problems: Theory, algorithms, and applications*, Numerical Mathematics and Scientific Computation, Oxford University Press, New York, 2005.

[5] Higgins, V. and Johnson, C., *Inverse spectral problems for collections of leading principal submatrices of tridiagonal matrices*, Linear Algebra Appl. 489 (2016), 104–122.

[6] Hogben, L., *Spectral graph theory and the inverse eigenvalue problem of a graph*, Electron. J. Linear Algebra 14 (2005), 12–31.

[7] Pickmann, H., Arela, S., Egaña, J. and Carrasco, D., *On the inverse eigenproblem for symmetric and nonsymmetric arrowhead matrices*, Proyecciones, 38 (2019), 811–828.

[8] Pickmann, H., Egana, J., Soto, R.L., *Extremal inverse eigenvalue problem for bordered diagonal matrices*, Linear Algebra Appl. 427 (2007), 256–271.

[9] Pickmann, H., Egaña, J. and Soto, R.L., *Extreme spectra realization by real symmetric tridiagonal and real symmetric arrow matrices*,Electron. J. Linear Algebra, 22 (2011), 780–795.

[10] Sharma, D. and Sarma, B., *Extremal inverse eigenvalue problem for irreducible acyclic matrices*, Appl. Math. Sci. Eng. 30 (2022), 192–209.

[11] Sharma, D. and Sen, M., *The minimax inverse eigenvalue problem for matrices whose graph is a generalized star of depth 2*, Linear Algebra Appl. 621 (2021), 334–344.

Iranian Journal of Numerical Analysis and Optimization
Vol. 16, No. 3, 2026, pp 1083–1106
https://doi.org/10.22067/ijnao.2026.97893.1848
https://ijnao.um.ac.ir/

CC BY NC

Research Article

Mathematical model on the effect of primary and secondary toxicants on a biological species: Occurrence of deformity

A.K. Agarwal*, and A. Agrawal

Abstract

A mathematical model is proposed and analyzed to study the effect of primary and secondary toxicants on a biological species. Here, we consider a situation that after-effect of the primary and secondary toxicants a subclass of biological species shows deformity as stagnating the reproduction capability. We analyzed the model using stability analysis and Hopf-bifurcation theory. The analysis of model shows, primary and secondary toxicants badly affects the biological species. These toxicants decrease the growth rate of species as well as causing deformity in a subclass. For high emission of primary toxicant, system becomes unstable, total density and deformed subclass density of species start oscillating around their equilibrium levels. This study helps us to plan best courses of action to make a healthier environment for living things such as plant trees, properly dispose the wastes, use pesticides and insecticides sparingly, use eco-friendly vehicles, and so on.

*Corresponding author

Received 16 February 2026; revised 2 May 2026; accepted 6 May 2026

Anuj Kumar Agarwal
Faculty of Mathematical and Statistical Sciences, Shri Ramswaroop Memorial University, Barabanki, 225003, Uttar Pradesh, India. e-mail: anuj.kumaragarwal20@gmail.com

Alok Agrawal
Department of Mathematics, Amity University, Lucknow, 226028, Uttar Pradesh, India. e-mail: akagrawal@lko.amity.edu

How to cite this article
Agarwal, A.K. and Agrawal, A., Mathematical model on the effect of primary and secondary toxicants on a biological species: Occurrence of deformity. *Iran. J. Numer. Anal. Optim.*, 2026; 16(3): 1083-1106. https://doi.org/10.22067/ijnao.2026.97893.1848

AMS subject classifications (2020): Primary 93A30; Secondary 37G15.

Keywords: Biological species; Primary and secondary toxicants; Deformity; Hopf-bifurcation.

1 Introduction

The term environmental contaminants refer to those potentially objectionable substances (or toxicants) that exist in the environment at enough level to harmful effect on the health of each type of biological species (i.e., human beings, animals and plants). The environmental contaminants are the result of various activities of human populations, such as fuel combustion, industrialisation, chemical manufacturing, domestic sewage, and so on, and some natural factors, such as volcanic eruptions, forest fires which contain dangerous gases, chemicals, and so on. Some toxicants that are not directly emitted in the environment, although they are formed in the environment by chemical and physical processes from primary toxicants (those toxicants that are directly emitted into the environment) also cause to increase the contaminants levels in the environment. These toxicants are known as secondary toxicants and may be even more toxic and harmful than primary toxicants. To illustrate, some gaseous toxicants (or primary toxicants) such as Sulfur Dioxide (SO_2), Nitric Oxides (NO_X), Carbon Monoxide (CO), Volatile Organic Compounds (VOC) including Hydrocarbons (C_XH_Y) are discharged from particular sources, such as burning of fuels, industrial activities, volcanic eruptions, and so on. The product of these primary toxicants produce secondary toxicants, such as ground level ozone — produced by product of Nitric Oxides (NO_X), Hydrocarbons (C_XH_Y) and Sunlight — and acid rain which is formed when SO_2 and NO_X react with Water (H_2O) and Oxygen(O_2). There are many other toxicants that are the product of primary toxicants, sunlight, and water, for example, Peroxy Acetyl Nitrates (PAN), Formaldehydes, Aerosols, and so on (see Figure 1).

When these toxicants enter the food chain of biological species, a range of diseases may occur in biological species including cancer, genetic disorder, lesions, deformities, stagnated capability of reproduction, and so on [3, 5, 11, 12, 13, 14, 16, 19].

Many researchers, economists and other authorities are busy to monitor and assess the effect of toxicants on biological species. Some of them use mathematical models to highlight the various natures of biological species in the real life fields such as ecology, ecotoxicology, epidemiology, and so on. With the help of modeling, researchers are able to predict the growth of biological species in toxic environment and understand diseases and deformities in biological species in a more meaningful manner. In particular, researchers have proposed, analyzed, and simulated mathematical models to provide important insights for the effect of toxicants on biological species

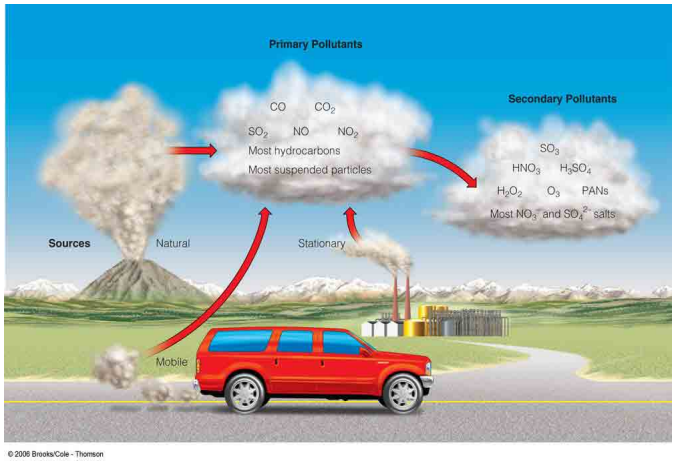


Figure 1: Primary and Secondary pollutants (or toxicants)
(Image Courtesy: sciencescene.com/Environmental%20Science/Graphics/1503.jpg)

[1, 4, 8, 17, 18]. Freedman and Shukla [4] have studied the effect of a single toxicant on biological species and predator: Prey case. They have shown that toxicant affected both the growth rate and carrying capacity. Shukla and Agrawal [17] have proposed a model by considering the case of allelopathy. In their model, the toxicant emitted into the environment by one species and decreased the growth rate and carrying capacity of other species. Shukla et al. [18] proposed and analyzed a model to study the growth of biomass resources (forests, agricultural, crops, etc.) with the effect of primary and secondary toxicants. They concluded that primary and secondary toxicants, both simultaneously affected the biomass resources. The biomass resources may extinct if the emission rate of primary and secondary toxicants are extremely high. Furthermore, a highly important observable fact when a subclass of biological species shows deformity and incapable in reproduction after the effect of a toxicant emitted from some external sources, was studied by Agrawal and Shukla [1]. Furthermore, Kumar et al. [8] proposed and analyzed a model with an assumption that toxicant emitted by biological species and a subclass of species is severely affected by this toxicant and suffer from deformity. They showed that the density of biological species decreased and density of the subclass of biological species increased when emission rate of toxicant increased. The system becomes unstable, and both densities (density of biological species and density of the subclass of biological species) show oscillations for higher emission rate of toxicant.

In this paper, we propose and analyze a mathematical model to study the effect of primary and secondary toxicant on a biological species of a subclass, which is severely affected and gets deformed. We also provide a numerical simulation of the model for the confirmation of our analytical results.

2 Mathematical model

We consider a biological species surviving in a polluted environment with population density $N(t)$ at time t. This polluted environment having primary and secondary toxicants with environmental concentrations $T(t)$ and $T_s(t)$ at time t. We assume that the primary toxicant is constantly discharged in the environment by some external sources. A part of this primary toxicant is converted into secondary toxicant. So, the rate at which primary toxicant is converted into secondary toxicant is directly proportional to the environmental concentration of primary toxicant. Let $U(t)$ and $U_s(t)$ be the concentrations of primary and secondary toxicants, respectively, taken up by the biological species $N(t)$ at time t. When these toxicants are being uptaken by the biological species, the growth rate of species $N(t)$ is decreasing as well as a subclass of species with population density $N_D(t)$ shows deformity as stagnating the reproduction capability. The remaining population density that is free from deformity is assumed as $N_A(t)$. Keeping these facts in mind, we propose the following mathematical model:

$$\begin{aligned}
\frac{dN_A}{dt} &= (b-d)N_A - (r_1U + r_2U_s)N_A - \frac{rN_AN}{K(T,T_s)},\\
\frac{dN_D}{dt} &= (r_1U + r_2U_s)N_A - \frac{rN_DN}{K(T,T_s)} - (\alpha+d)N_D,\\
\frac{dT}{dt} &= Q - \delta_1T - \gamma_1TN + \pi_1\nu_1NU - kT,\\
\frac{dT_s}{dt} &= \theta kT - \delta_2T_s - \gamma_2T_sN + \pi_2\nu_2NU_s,\\
\frac{dU}{dt} &= \gamma_1TN - \beta_1U - \nu_1NU,\\
\frac{dU_s}{dt} &= \gamma_2T_sN - \beta_2U_s - \nu_2NU_s.
\end{aligned} \tag{1}$$

$$N(t) = N_A(t) + N_D(t), \quad N_A(0), N_D(0), T(0), T_s(0) \geq 0, \quad U(0) \geq c_1T(0),$$
$$U_s(0) \geq c_2T_s(0) \quad 0 < \theta < 1, \quad c_i(0) > 0, \quad 0 \leq \pi_i \leq 1, \quad \text{for all} \quad i = 1, 2.$$

In the proposed model system (1), all the considered parameters are positive constants. The constants b, d, and r are the natural birth, death, and growth rate of biological species $N(t)$. The primary toxicant $T(t)$ is constantly emitted in the environment at the rate Q. The constant k represents the conversion rate at which primary toxicant is being converted into secondary toxicant. In the conversion process, it may happen that the entire part of primary toxicant (kT) is not fully converted into secondary toxicant, and other substances may be produced, which are harmless for biological species. For this kind of situation, we consider a constant $\theta(\leq 1)$

as a fraction of that part of primary toxicant, which is converted into secondary toxicant. For $\theta = 1$, the entire part of primary toxicant (kT) is converted into secondary toxicant. The primary and secondary toxicants are uptaken by the species at the rate γ_1 and γ_2, respectively. These primary and secondary toxicants are decreasing the population density at the rate r_1 and r_2, respectively. Also, α is the mortality rate of deformed population due to high toxicities in the environment. Moreover, δ_1 and δ_2 are the natural depletion rates of primary and secondary toxicants correspondingly. In addition, β_1 and β_2 are the natural depletion rates of uptake concentration of primary and secondary toxicants respectively. Since some members of species die out, the uptake concentrations $U(t), U_s(t)$ are depleted at the rates ν_1, ν_2 and fractions π_1, π_2 of these decay is re-entering into the environment, respectively. Also, $c_1, c_2 > 0$ are the proportionality constants to find out the initial uptake concentrations of primary and secondary toxicants. Furthermore, $K(T, T_s)$ represents the carrying capacity of the environment. It is a decreasing function of T and T_s, that is,

$$K(0,0) = K_0 > 0,\ K(T,T_s) > 0,\ \frac{\partial K(T,T_s)}{\partial T} < 0,\ \frac{\partial K(T,T_s)}{\partial T_s} < 0, \quad \text{for } T, T_s > 0, \tag{2}$$

where K_0 is the carrying capacity of the toxicants free environment.

To make our analysis easy, we reduce the model system (1) into following model system (3), which is free from $N_A(t)$ by using the fact that $N(t) = N_A(t) + N_D(t)$:

$$\begin{aligned}
\frac{dN}{dt} &= rN - \frac{rN^2}{K(T,T_s)} - (\alpha + b)N_D,\\
\frac{dN_D}{dt} &= (r_1U + r_2U_s)(N - N_D) - \frac{rN_DN}{K(T,T_s)} - (\alpha + d)N_D,\\
\frac{dT}{dt} &= Q - \delta_1 T - \gamma_1 TN + \pi_1\nu_1 NU - kT,\\
\frac{dT_s}{dt} &= \theta kT - \delta_2 T_s - \gamma_2 T_s N + \pi_2\nu_2 NU_s,\\
\frac{dU}{dt} &= \gamma_1 TN - \beta_1 U - \nu_1 NU,\\
\frac{dU_s}{dt} &= \gamma_2 T_s N - \beta_2 U_s - \nu_2 NU_s.
\end{aligned} \tag{3}$$

The region of attraction Ω that attracts all the positive solution of model system (3), is as follows:

$$\Omega = \{(N, N_D, T, T_s, U, U_s) :\ 0 \le N \le K_0,\ 0 \le N_D \le \frac{(r_1\delta_{m_2} + r_2\theta k)QK_0}{(r_1\delta_{m_2} + r_2\theta k)Q + (\alpha + d)\delta_{m_1}\delta_{m_2}},$$

$$0 \leq T+U \leq \frac{Q}{\delta_{m_1}},\ 0 \leq T_s+U_s \leq \frac{\theta k Q}{\delta_{m_1}\delta_{m_2}}\Bigg\},$$

where $\delta_{m_1} = \min(\delta_1 + k, \beta_1)$ and $\delta_{m_2} = \min(\delta_2, \beta_2)$.

The proof of region of attraction Ω is given in Appendix A.

3 Equilibrium points

After, equating the right-hand side of model system (3) is equal to zero, we obtain two nonnegative equilibrium points $E_1(0, 0, \frac{Q}{\delta_1+k}, \frac{\theta kQ}{\delta_2(\delta_1+k)}, 0, 0)$ and $E_2(N^*, N_D^*, T^*, T_s^*, U^*, U_s^*)$. Here, we omitted the existence of E_1 because it is obvious. The existence of E_2 is as follows:

The values $N^*, N_D^*, T^*, T_s^*, U^*$, and U_s^* are the positive solutions of the following nonlinear equations:

$$N = \frac{(r - r_1U - r_2U_s)K(T, T_s)}{r}, \tag{4a}$$

$$N_D = \frac{(r_1U + r_2U_s)NK(T, T_s)}{rN + (r_1U + r_2U_s + \alpha + d)K(T, T_s)}, \tag{4b}$$

$$T = \frac{Q(\beta_1 + \nu_1 N)}{f_1(N)} = g_1(N)\ (say), \tag{4c}$$

$$T_s = \frac{\theta k(\beta_2 + \nu_2 N)}{f_2(N)} \frac{Q(\beta_1 + \nu_1 N)}{f_1(N)} = g_2(N)\ (say), \tag{4d}$$

$$U = \frac{Q\gamma_1 N}{f_1(N)} = h_1(N)\ (say), \tag{4e}$$

$$U_s = \frac{\theta k\gamma_2 N}{f_2(N)} \frac{Q(\beta_1 + \nu_1 N)}{f_1(N)} = h_2(N)\ (say), \tag{4f}$$

where

$$f_1(N) = (\delta_1 + k)\beta_1 + \{\gamma_1\beta_1 + (\delta_1 + k)\nu_1\}N + \gamma_1\nu_1(1 - \pi_1)N^2, \tag{4g}$$

$$f_2(N) = \delta_2\beta_2 + (\gamma_2\beta_2 + \delta_2\nu_2)N + \gamma_2\nu_2(1 - \pi_2)N^2, \tag{4h}$$

From (4c) and (4d), we note that T is directly dependent on the parameter Q and T_s is directly proportional to the parameters Q, k, and θ. Since, carrying capacity $K(T, T_s)$ is a decreasing function of T and T_s. Hence, the carrying capacity of the environment decreases when the discharge rate of primary toxicant Q, the conversion rate of primary toxicant into secondary toxicant k or the fraction of that part of primary toxicant which is converted into secondary toxicant θ increase.

Let

$$F(N) = rN - (r - r_1h_1(N) - r_2h_2(N))K(g_1(N), g_2(N)). \tag{5}$$

At $N = 0$,

$$F(0) = -rK\left(\frac{Q}{(\delta_1 + k)}, \frac{\theta kQ}{(\delta_1 + k)\delta_2}\right) < 0. \tag{6}$$

At $N = K_0$,

$$F(K_0) = rK_0 - (r - r_1h_1(K_0) - r_2h_2(K_0))K(g_1(K_0), g_2(K_0)) > 0. \tag{7}$$

Equations (6) and (7) guarantee that $F(N) = 0$ has a solution in the interval $[0, K_0]$.

Also, the root N^* of $F(N) = 0$ is unique, if

$$\frac{dF}{dN} = \left[r + K(g_1(N), g_2(N))\left\{r_1\frac{dh_1}{dN} + r_2\frac{dh_2}{dN}\right\} - (r - r_1h_1(N) - r_2h_2(N))\left\{\frac{\partial K}{\partial T_1}\frac{dg_1}{dN} + \frac{\partial K}{\partial T_2}\frac{dg_2}{dN}\right\}\right] > 0. \tag{8}$$

Hence, $F(N) = 0$ must have a root N^* in the interval $[0, K_0]$ and the root N^* is unique if condition (8) holds. The values of other points can be found out using the value of N^* in (4).

4 Dynamical behavior

In this section, we show the dynamical behavior of model system (3) corresponding to the obtained equilibrium points E_1 and E_2.

4.1 Dynamical behavior corresponding to E_1

We evaluate the Jacobian matrices M_1(say) of model system (3) and its eigenvalues corresponding to the equilibrium point E_1.

We obtain the Jacobian matrix M_1 as follows:

$$M_1 = \begin{bmatrix} r & -(\alpha+b) & 0 & 0 & 0 & 0 \\ 0 & -(\alpha+d) & 0 & 0 & 0 & 0 \\ -\dfrac{\gamma_1 Q}{(\delta_1+k)} & 0 & -(\delta_1+k) & 0 & 0 & 0 \\ -\dfrac{\gamma_2 Q\theta k}{(\delta_1+k)\delta_2} & 0 & \theta k & -\delta_2 & 0 & 0 \\ \dfrac{\gamma_1 Q}{(\delta_1+k)} & 0 & 0 & 0 & -\beta_1 & 0 \\ \dfrac{\gamma_2 Q\theta k}{(\delta_1+k)\delta_2} & 0 & 0 & 0 & 0 & -\beta_2 \end{bmatrix}.$$

Here, the eigenvalues of the Jacobian matrix M_1 are $r, -(\alpha+d), -(\delta_1+k), -\delta_2, -\beta_1$, and $-\beta_2$, which shows that E_1 is a saddle point locally unstable manifold in $N-$ direction and locally stable manifold in $N_D - T - T_s - U - U_s$ space.

4.2 Dynamical behavior corresponding to E_2

Now, we linearize the model system (3) corresponding to the equilibrium point $E_2 = (N^*, N_D^*, T^*, T_s^*, U^*, U_s^*)$ by taking the following transformation:

$$\begin{aligned} N &= N^* + n, \quad N_D = N_D^* + n_d, \quad T = T^* + \tau, \\ T_s &= T_s^* + \tau_s, \quad U = U^* + u, \quad U_s = U_s^* + u_s. \end{aligned}$$

Here, n, n_d, τ, τ_s, u, and u_s are taken as small perturbations around E_2.

After using the above transformation, the model system (3) can be written in the terms of n, n_d, τ, τ_s, u, and u_s as follows:

$$\dot{X} = M_2 X + N_2, \tag{9}$$

where

$$X = \begin{bmatrix} n \\ n_d \\ \tau \\ \tau_s \\ u \\ u_s \end{bmatrix}, \quad M_2 = \begin{bmatrix} m_{11} & m_{12} & m_{13} & m_{14} & 0 & 0 \\ m_{21} & m_{22} & m_{23} & m_{24} & m_{25} & m_{26} \\ m_{31} & 0 & m_{33} & 0 & m_{35} & 0 \\ m_{41} & 0 & m_{43} & m_{44} & 0 & m_{46} \\ m_{51} & 0 & m_{53} & 0 & m_{55} & 0 \\ m_{61} & 0 & 0 & m_{64} & 0 & m_{66} \end{bmatrix}, \quad N_2 = \begin{bmatrix} n_1 \\ n_2 \\ n_3 \\ n_4 \\ n_5 \\ n_6 \end{bmatrix}.$$

Moreover,

$$m_{11} = -r\left\{\frac{2N^*}{K(T^*,T_s^*)} - 1\right\}, \qquad m_{12} = -(\alpha + b),$$
$$m_{13} = \frac{rN^{*2}}{K^2(T^*,T_s^*)}\left[\frac{\partial K}{\partial T}\right]_{E_2}, \qquad m_{14} = \frac{rN^{*2}}{K^2(T^*,T_s^*)}\left[\frac{\partial K}{\partial T_s}\right]_{E_2},$$
$$m_{21} = r_1U^* + r_2U_s^* - \frac{rN_D^*}{K(T^*,T_s^*)}, \qquad m_{22} = -(r_1U^* + r_2U_s^*)\frac{N^*}{N_D^*},$$
$$m_{23} = \frac{rN^*N_D^*}{K^2(T^*,T_s^*)}\left[\frac{\partial K}{\partial T}\right]_{E_2}, \qquad m_{24} = \frac{rN^*N_D^*}{K^2(T^*,T_s^*)}\left[\frac{\partial K}{\partial T_s}\right]_{E_2},$$
$$m_{25} = r_1(N^* - N_D^*), \qquad m_{26} = r_2(N^* - N_D^*),$$
$$m_{31} = [-\gamma_1T^* + \pi_1\nu_1U^*], \qquad m_{33} = -\frac{(Q + \pi_1\nu_1N^*U^*)}{T^*},$$
$$m_{35} = \pi_1\nu_1N^*, \qquad m_{41} = [-\gamma_2T_s^* + \pi_2\nu_2U_s^*], \qquad m_{43} = \theta k$$
$$m_{44} = -(\delta_2 + \gamma_2N^*), \qquad m_{46} = \pi_2\nu_2N^*, \qquad m_{51} = \gamma_1T^* - \nu_1U^*,$$
$$m_{53} = \gamma_1N^*, \qquad m_{55} = -(\beta_1 + \nu_1N^*), \qquad m_{61} = \gamma_2T_s^* - \nu_2U_s^*,$$
$$m_{64} = \gamma_2N^*, \qquad m_{66} = -(\beta_2 + \nu_2N^*),$$

$$n_1 = -\frac{r}{K(T^*,T_s^*)}.n^2 + \frac{r(2N^* + n)}{K^2(T^*,T_s^*)}\left[\left[\frac{\partial K}{\partial T}\right]_{E_2}.\tau + \left[\frac{\partial K}{\partial T_s}\right]_{E_2}.\tau_s\right].n,$$
$$n_2 = (r_1.u + r_2.u_s).(n - n_d) - \frac{r}{K(T^*,T_s^*)}.n.n_d$$
$$+ \frac{r(N^*.n_d + N_D^*.n + n.n_d)}{K^2(T^*,T_s^*)}\left[\left[\frac{\partial K}{\partial T}\right]_{E_2}.\tau + \left[\frac{\partial K}{\partial T_s}\right]_{E_2}.\tau_s\right],$$
$$n_3 = -\gamma_1.n.\tau + \pi_1\nu_1.n.u, \; n_4 = -\gamma_2.n.\tau_s + \pi_2\nu_2.n.u_s,$$
$$n_5 = \gamma_1.n.\tau - \nu_1.n.u, \; n_6 = \gamma_2.n.\tau_s - \nu_2.n.u_s.$$

In (9), M_2X and N_2 show the linear and nonlinear parts of the model system (3), respectively, and M_2 is a Jacobian matrix corresponding to the equilibrium point E_2. Thus, the characteristic equation of M_2 can be written as

$$p(x) = x^6 + c_1x^5 + c_2x^4 + c_3x^3 + c_4x^2 + c_5x + c_6. \tag{10}$$

For the coefficients c_1, c_2, c_3, c_4, c_5, and c_6, see **Appendix B**.

According to the Routh–Hurwitz criterion and characteristic equation (10), all the eigenvalues of Jacobian matrix M_2 are either negative or having negative real parts iff

$$c_j > 0, \quad (j = 1, 6), \tag{11a}$$
$$H_2 = c_1c_2 - c_3 > 0, \tag{11b}$$

$$H_3 = c_1c_2c_3 + c_1c_5 - c_3^2 - c_1^2c_4 > 0, \tag{11c}$$

$$H_4 = (c_1c_2 - c_3)(c_3c_4 - c_2c_5) + (c_1c_2 - c_3)c_1c_6 - (c_1c_4 - c_5)^2 > 0, \tag{11d}$$

$$\begin{aligned} H_5 = {} & c_3(c_1c_2 - c_3)(c_4c_5 - c_3c_6) + c_3c_5(c_2c_5 - c_1c_6) + c_1c_3c_6(c_1c_4 - 2c_5) \\ & - c_1(c_2c_5 - c_1c_6)^2 - c_5(c_1c_4 - c_5)^2 > 0. \end{aligned} \tag{11e}$$

So, we can state the following theorem for the locally asymptotically stability of the equilibrium point E_2.

Theorem 1. The equilibrium point E_2 of model system (3) is locally asymptotically stable, if conditions (11) are satisfied.

Hence, all the positive solutions of the model system (3) in the neighborhood of E_2, will become stable to E_2 if the conditions (11) are satisfied.

Existence of Hopf-bifurcation

In the model system (3), there is a possibility that Hopf-bifurcation [6, 9, 15] may exist corresponding to the equilibrium point E_2. So, we find the conditions for the existence of Hopf-bifurcation by treating Q (i.e., the discharge rate of primary toxicant, which is directly emitted in the environment from some sources) as a bifurcation parameter.

It is obvious that a Hopf-bifurcation may exist if all the eigenvalues of Jacobian matrix have negative real parts except a purely imaginary complex conjugate pair. In this case, the Jacobian matrix M_2 have six eigenvalues $x_j = R_j + iI_j$ $(j = 1, \ldots, 6)$ (say). So, the Hopf-bifurcation exist only when $R_1, R_2 = 0,\ I_1 = -I_2 \neq 0$ & $R_3, R_4, R_5, R_6 < 0$ at the critical value $Q = Q^*$(say).

According to the Liu's criterion [10], the model system (3) undergoes a Hopf-bifurcation, if the following conditions hold at the critical value $Q = Q^* > 0$:

$$[c_j]_{Q=Q^*} > 0 \quad \text{for } (j = 1, \ldots, 6), \tag{12a}$$

$$[H_2]_{Q=Q^*} = [c_1c_2 - c_3]_{Q=Q^*} > 0, \tag{12b}$$

$$[H_3]_{Q=Q^*} = [c_1c_2c_3 + c_1c_5 - c_3^2 - c_1^2c_4]_{Q=Q^*} > 0, \tag{12c}$$

$$\begin{aligned} [H_4]_{Q=Q^*} = {} & [(c_1c_2 - c_3)(c_3c_4 - c_2c_5) + (c_1c_2 - c_3)c_1c_6 - (c_1c_4 - c_5)^2]_{Q=Q^*} \\ & > 0, \end{aligned} \tag{12d}$$

$$\begin{aligned} [H_5]_{Q=Q^*} = {} & [c_3(c_1c_2 - c_3)(c_4c_5 - c_3c_6) + c_3c_5(c_2c_5 - c_1c_6) + c_1c_3c_6(c_1c_4 - 2c_5) \\ & - c_1(c_2c_5 - c_1c_6)^2 - c_5(c_1c_4 - c_5)^2]_{Q=Q^*} = 0, \end{aligned} \tag{12e}$$

$$\left[\frac{dR_j}{dQ}\right]_{Q=Q^*} \neq 0, \quad \text{for } j = 1, 2. \tag{12f}$$

With the help of *Existence of Hopf–bifurcation in a 6-dimensional system* [7], the last condition (12f) can be written in the terms of coefficients of characteristic equation (10) as follows:

$$\left[\frac{dR}{dQ}\right]_{Q=Q^*} = \left[\frac{1}{B}\frac{dH_5}{dQ}\right]_{Q=Q^*} \neq 0, \tag{13}$$

where H_5 is defined as (11e) and

$$B = 2\{c_1(c_1c_4 - c_5) - c_3(c_1c_2 - c_3)\} \begin{vmatrix} -c_1 & 1 & 1 & 0 & 0 \\ c_2 - 2A & -c_1 & 0 & 1 & 0 \\ c_1A - c_3 & c_2 - A & A & 0 & 1 \\ A^2 - c_2A + c_4 & c_1A - c_3 & 0 & A & 0 \\ 0 & A^2 - c_2A + c_4 & 0 & 0 & A \end{vmatrix}. \tag{14}$$

Also, $A = \dfrac{c_3c_5 + c_1^2c_6 - c_1c_2c_5}{c_1^2c_4 - c_1c_5 - c_1c_2c_3 + c_3^2}$.

So, we can state the following theorem for the existence of Hopf-bifurcation.

Theorem 2. The model (3) undergoes a Hopf-bifurcation corresponding to the equilibrium point E_2, if Q crosses the critical value $Q^* > 0$ such that

1. $[c_j]_{Q=Q^*} > 0 \quad \text{for } (j = 1, \ldots, 6)$,
2. $[H_2]_{Q=Q^*} = [c_1c_2 - c_3]_{Q=Q^*} > 0$,
3. $[H_3]_{Q=Q^*} = [c_1c_2c_3 + c_1c_5 - c_3^2 - c_1^2c_4]_{Q=Q^*} > 0$,
4. $[H_4]_{Q=Q^*} = [(c_1c_2 - c_3)(c_3c_4 - c_2c_5) + (c_1c_2 - c_3)c_1c_6 - (c_1c_4 - c_5)^2]_{Q=Q^*} > 0$,
5. $[H_5]_{Q=Q^*} = [c_3(c_1c_2 - c_3)(c_4c_5 - c_3c_6) + c_3c_5(c_2c_5 - c_1c_6)$
   $+c_1c_3c_6(c_1c_4 - 2c_5) - c_1(c_2c_5 - c_1c_6)^2 - c_5(c_1c_4 - c_5)^2]_{Q=Q^*} = 0$,
6. $\left[\dfrac{dR}{dQ}\right]_{Q=Q^*} = \left[\dfrac{1}{B}\dfrac{dH_5}{dQ}\right]_{Q=Q^*} \neq 0$,

where B is defined as (14).

Hence, if Theorem 2 holds, there exist Hopf-bifurcation at the critical value Q^*. The model system (3) has bifurcating periodic solutions starting from Hopf-bifurcation. So, it is essential to know the nature of bifurcating periodic solutions.

Nature of bifurcating periodic solutions

The direction and stability of bifurcating periodic solutions can be examined by the sign of these important coefficients μ_2, β_2, τ_2 given by Hassard, Kazarinoff, and Wan [6]. To evaluate these coefficients μ_2, β_2, τ_2 for the parameter Q, we derive some explicit formulas with the help of normal form and center manifold theory [6, 9]. To transform the model system (3) in normal form, we assume all the eigenvalues of Jacobian matrix M_2 as $\pm iv$, $-J_1$, $-J_2$, $-J_3$ and $-J_4$.

Let $X = PY$; then the normal form is

$$\dot{Y} = JY + F, \quad Y = col.(y_1, y_2, y_3, y_4, y_5, y_6), \tag{15}$$

where

$$\begin{aligned} J =& P^{-1}M_2P = \begin{bmatrix} 0 & -v & 0 & 0 & 0 & 0 \\ v & 0 & 0 & 0 & 0 & 0 \\ 0 & 0 & -J_1 & 0 & 0 & 0 \\ 0 & 0 & 0 & -J_2 & 0 & 0 \\ 0 & 0 & 0 & 0 & -J_3 & 0 \\ 0 & 0 & 0 & 0 & 0 & -J_4 \end{bmatrix} \\ \text{and} \quad F =& P^{-1}f = \begin{bmatrix} F_1(y_1, y_2, y_3, y_4, y_5, y_6) \\ F_2(y_1, y_2, y_3, y_4, y_5, y_6) \\ F_3(y_1, y_2, y_3, y_4, y_5, y_6) \\ F_4(y_1, y_2, y_3, y_4, y_5, y_6) \\ F_5(y_1, y_2, y_3, y_4, y_5, y_6) \\ F_6(y_1, y_2, y_3, y_4, y_5, y_6) \end{bmatrix}. \end{aligned} \tag{16}$$

Here P is a transformed matrix defined as

$$P = \begin{bmatrix} P_{11} & P_{12} & P_{13} & P_{14} & P_{15} & P_{16} \\ P_{21} & P_{22} & P_{23} & P_{24} & P_{25} & P_{26} \\ P_{31} & P_{32} & P_{33} & P_{34} & P_{35} & P_{36} \\ P_{41} & P_{42} & P_{43} & P_{44} & P_{45} & P_{46} \\ P_{51} & P_{52} & P_{53} & P_{54} & P_{55} & P_{56} \\ P_{61} & P_{62} & P_{63} & P_{64} & P_{65} & P_{66} \end{bmatrix},$$

where

$$P_{11} = m_{12}(L_3L_6 - L_4L_7), \quad P_{12} = m_{12}(L_3L_7 + L_4L_6), \quad P_{13} = m_{12}K_{31}K_{41},$$
$$P_{14} = m_{12}K_{32}K_{42}, \quad P_{15} = m_{12}K_{33}K_{43}, \quad P_{16} = m_{12}K_{34}K_{44}, \quad P_{21} = -L_{13},$$
$$P_{22} = -L_{14}, \quad P_{23} = K_{71}, \quad P_{24} = K_{72}, \quad P_{25} = K_{73}, \quad P_{26} = K_{74},$$
$$P_{31} = m_{12}(L_1L_6 + m_{31}vL_7), \quad P_{32} = m_{12}(L_1L_7 - m_{31}vL_6)$$
$$P_{33} = m_{12}K_{11}K_{41}, \quad P_{34} = m_{12}K_{12}K_{42}, \quad P_{35} = m_{12}K_{13}K_{43},$$
$$P_{36} = m_{12}K_{14}K_{44}, \quad P_{41} = m_{12}L_9, \quad P_{42} = m_{12}L_{10}, \quad P_{43} = m_{12}K_{61},$$
$$P_{44} = m_{12}K_{62}, \quad P_{45} = m_{12}K_{63}, \quad P_{46} = m_{12}K_{64},$$
$$P_{51} = m_{12}(L_2L_6 + m_{51}vL_7) \quad P_{52} = m_{12}(L_2L_7 - m_{51}vL_6),$$
$$P_{53} = m_{12}K_{21}K_{41}, \quad P_{54} = m_{12}K_{22}K_{42}, \quad P_{55} = m_{12}K_{23}K_{43},$$
$$P_{56} = m_{12}K_{24}K_{44}, \quad P_{61} = m_{12}L_{11}, \quad P_{62} = m_{12}L_{12}, \quad P_{63} = m_{12}K_{51},$$
$$P_{64} = m_{12}K_{52}, \quad P_{65} = m_{12}K_{53}, \quad P_{66} = m_{12}K_{54}.$$

Also,

$$L_1 = m_{31}m_{55} - m_{35}m_{51}, \quad L_2 = m_{33}m_{51} - m_{31}m_{53},$$
$$L_3 = m_{35}m_{53} - m_{33}m_{55} + v^2, \quad L_4 = (m_{33} + m_{55})v,$$
$$L_5 = m_{41}m_{66} - m_{46}m_{61}, \quad L_6 = m_{46}m_{64} - m_{44}m_{66} + v^2,$$
$$L_7 = (m_{44} + m_{66})v, \quad L_8 = m_{44}m_{61} - m_{41}m_{64},$$
$$L_9 = L_3L_5 + m_{43}m_{66}L_1 + m_{41}vL_4 - m_{43}m_{31}v^2,$$
$$L_{10} = L_4L_5 - m_{41}vL_3 - m_{43}vL_1 - m_{31}m_{43}m_{66}v,$$
$$L_{11} = L_3L_8 + m_{61}vL_4 - m_{43}m_{64}L_1,$$
$$L_{12} = L_4L_8 - m_{61}vL_3 + m_{31}m_{43}m_{64}v,$$
$$L_{13} = m_{11}(L_3L_6 - L_4L_7) + v(L_3L_7 + L_4L_6) + m_{13}L_1 + m_{14}L_9,$$
$$L_{14} = m_{11}(L_3L_7 + L_4L_6) - v(L_3L_6 - L_4L_7) + m_{14}L_{10} - m_{13}m_{31}v.$$

For $j = 1, 2, 3, 4,$

$$K_{1j} = m_{31}m_{55} - m_{35}m_{51} + m_{31}J_j, \quad K_{2j} = m_{33}m_{51} - m_{53}m_{31} + m_{51}J_j,$$
$$K_{3j} = m_{35}m_{53} - m_{33}m_{55} - (m_{33} + m_{55})J_j - J_j^2,$$
$$K_{4j} = m_{64}m_{46} - m_{44}m_{66} - (m_{44} + m_{66})J_j - J_j^2,$$
$$K_{5j} = (m_{61}m_{44} - m_{64}m_{41} + m_{61}J_j)K_{3j} - m_{64}m_{43}K_{1j},$$
$$K_{6j} = (m_{66} + J_j)(m_{41}K_{3j} + m_{43}K_{1j}) - m_{61}m_{46}K_{3j},$$
$$K_{7j} = -\{(m_{11} + J_j)K_{3j}K_{4j} + m_{13}K_{1j}K_{4j} + m_{14}K_{6j}\},$$

and

$$f = \begin{bmatrix} f_1(y_1, y_2, y_3, y_4, y_5, y_6) \\ f_2(y_1, y_2, y_3, y_4, y_5, y_6) \\ f_3(y_1, y_2, y_3, y_4, y_5, y_6) \\ f_4(y_1, y_2, y_3, y_4, y_5, y_6) \\ f_5(y_1, y_2, y_3, y_4, y_5, y_6) \\ f_6(y_1, y_2, y_3, y_4, y_5, y_6) \end{bmatrix},$$

where

$$\begin{aligned}
f_1(y_1, y_2, y_3, y_4, y_5, y_6) =& -\frac{r}{K(T^*, T_s^*)}.n^2 \\
&+ \frac{r(2N^* + n)}{K^2(T^*, T_s^*)} \left[\frac{\partial K}{\partial T}(T^*, T_s^*).\tau + \frac{\partial K}{\partial T_s}(T^*, T_s^*).\tau_s\right].n, \\
f_2(y_1, y_2, y_3, y_4, y_5, y_6) =& (r_1.u + r_2.u_s).(n - n_d) - \frac{r}{K(T^*, T_s^*)}.n.n_d \\
&+ \frac{r(N^*.n_d + N_D^*.n + n.n_d)}{K^2(T^*, T_s^*)} \\
&\times \left[\frac{\partial K}{\partial T}(T^*, T_s^*).\tau + \frac{\partial K}{\partial T_s}(T^*, T_s^*).\tau_s\right], \\
f_3(y_1, y_2, y_3, y_4, y_5, y_6) =& -\gamma_1.n.\tau + \pi_1\nu_1.n.u, \\
f_4(y_1, y_2, y_3, y_4, y_5, y_6) =& -\gamma_2.n.\tau_s + \pi_2\nu_2.n.u_s, \\
f_5(y_1, y_2, y_3, y_4, y_5, y_6) =& \gamma_1.n.\tau - \nu_1.n.u, \\
f_6(y_1, y_2, y_3, y_4, y_5, y_6) =& \gamma_2.n.\tau_s - \nu_2.n.u_s.
\end{aligned}$$

Here,

$$n = \sum_{j=1}^{6} P_{1j}y_j,\ n_d = \sum_{j=1}^{6} P_{2j}y_j,\ \tau = \sum_{j=1}^{6} P_{3j}y_j,$$
$$\tau_s = \sum_{j=1}^{6} P_{4j}y_j,\ u = \sum_{j=1}^{6} P_{5j}y_j,\ u_s = \sum_{j=1}^{6} P_{6j}y_j.$$

Now, we calculate the following quantities at critical value of parameter $Q = Q^*$ and $(y_1, y_2, y_3, y_4, y_5, y_6) = (0, 0, 0, 0, 0, 0)$:

$$
\begin{aligned}
g_{11} =& \frac{1}{4}\left\{\left(\frac{\partial^2 F_1}{\partial y_1^2}+\frac{\partial^2 F_1}{\partial y_2^2}\right)+i\left(\frac{\partial^2 F_2}{\partial y_1^2}+\frac{\partial^2 F_2}{\partial y_2^2}\right)\right\},\\
g_{02} =& \frac{1}{4}\left\{\left(\frac{\partial^2 F_1}{\partial y_1^2}-\frac{\partial^2 F_1}{\partial y_2^2}-2\frac{\partial^2 F_2}{\partial y_1 \partial y_2}\right)+i\left(\frac{\partial^2 F_2}{\partial y_1^2}-\frac{\partial^2 F_2}{\partial y_2^2}+2\frac{\partial^2 F_1}{\partial y_1 \partial y_2}\right)\right\},\\
g_{20} =& \frac{1}{4}\left\{\left(\frac{\partial^2 F_1}{\partial y_1^2}-\frac{\partial^2 F_1}{\partial y_2^2}+2\frac{\partial^2 F_2}{\partial y_1 \partial y_2}\right)+i\left(\frac{\partial^2 F_2}{\partial y_1^2}-\frac{\partial^2 F_2}{\partial y_2^2}-2\frac{\partial^2 F_1}{\partial y_1 \partial y_2}\right)\right\},\\
g_{21} =& G_{21}+\sum_{j=1}^{4}(2G_{110}^{j}w_{11}^{j}+G_{101}^{j}w_{20}^{j}),
\end{aligned}
$$

where

$$
\begin{aligned}
G_{21} =& \frac{1}{8}\left\{\left(\frac{\partial^3 F_1}{\partial y_1^3}+\frac{\partial^3 F_2}{\partial y_2^3}+\frac{\partial^3 F_2}{\partial y_1^2 \partial y_2}+\frac{\partial^3 F_1}{\partial y_1 \partial y_2^2}\right)\right.\\
&\left.+i\left(\frac{\partial^3 F_2}{\partial y_1^3}-\frac{\partial^3 F_1}{\partial y_2^3}-\frac{\partial^3 F_1}{\partial y_1^2 \partial y_2}+\frac{\partial^3 F_2}{\partial y_1 \partial y_2^2}\right)\right\}.
\end{aligned}
$$

For $j = 1, 2, 3, 4,$

$$
\begin{aligned}
G_{110}^{j} =& \frac{1}{2}\left\{\left(\frac{\partial^2 F_1}{\partial y_1 \partial y_{(j+2)}}+\frac{\partial^2 F_2}{\partial y_2 \partial y_{(j+2)}}\right)+i\left(\frac{\partial^2 F_2}{\partial y_1 \partial y_{(j+2)}}-\frac{\partial^2 F_1}{\partial y_2 \partial y_{(j+2)}}\right)\right\},\\
G_{101}^{j} =& \frac{1}{2}\left\{\left(\frac{\partial^2 F_1}{\partial y_1 \partial y_{(j+2)}}-\frac{\partial^2 F_2}{\partial y_2 \partial y_{(j+2)}}\right)+i\left(\frac{\partial^2 F_2}{\partial y_1 \partial y_{(j+2)}}+\frac{\partial^2 F_1}{\partial y_2 \partial y_{(j+2)}}\right)\right\},\\
h_{11}^{j} =& \frac{1}{4}\left(\frac{\partial^2 F_{(j+2)}}{\partial y_1^2}+\frac{\partial^2 F_{(j+2)}}{\partial y_2^2}\right),\\
h_{20}^{j} =& \frac{1}{4}\left(\frac{\partial^2 F_{(j+2)}}{\partial y_1^2}-\frac{\partial^2 F_{(j+2)}}{\partial y_2^2}-2i\frac{\partial^2 F_{(j+2)}}{\partial y_1 \partial y_2}\right),\\
w_{11}^{k} =& \frac{h_{11}^{j}}{J_k}, \quad w_{20}^{j}=\frac{h_{20}^{j}}{J_j+2iv}.
\end{aligned}
$$

Now, we evaluate the coefficients μ_2, β_2 and τ_2 using the following formulas:

$$
\begin{aligned}
C_1(0) &= \frac{i}{2v}\left[g_{20}g_{11}-2|g_{11}|^2-\frac{1}{3}|g_{02}|^2\right]+\frac{1}{2}g_{21}, \quad \mu_2 = -\frac{\mathrm{Re}\ C_10}{\mathrm{Re}\ x_1'},\\
\beta_2 &= 2\mathrm{Re}\ C_1(0), \qquad \tau_2 = -\frac{(\mathrm{Im}\ C_1(0)+\mu_2 \mathrm{Im}\ x_1')}{v}.
\end{aligned}
\tag{17}
$$

By observing the sign of the coefficients μ_2, β_2 and τ_2, we can state following theorem for the nature of bifurcating periodic solutions corresponding to the parameter Q.

Theorem 3. If $\mu_2 > 0$ (or $\mu_2 < 0$), the model system (3) shows a supercritical (or subcritical) Hopf-bifurcation and the bifurcating periodic solutions exist for $Q > Q^*$ (or $Q < Q^*$), if $\beta_2 < 0$ (or $\beta_2 > 0$), the bifurcating periodic solutions are stable (or unstable), if $\tau_2 > 0$ (or $\tau_2 < 0$), the period of bifurcating solutions increases (or decreases).

5 Numerical simulation

To clarify the analytical results of model system (3), which are found in the previous sections, numerical simulation is given with the help of MATCONT [2]. For that, we assume the carrying capacity function as

$$K(T, T_s) = K_0 - \frac{b_{11}T}{1 + b_{12}T} - \frac{b_{21}T_s}{1 + b_{22}T_s}, \tag{18a}$$

and a set of parameters as

$$\begin{aligned} &b = 0.2,\ d = 0.001,\ r_1 = 0.6,\ r_2 = 0.8,\ Q = 0.005, \alpha = 0.0008,\ \delta_1 = 0.08,\\ &\delta_2 = 0.005,\ \theta = 0.2,\ k = 0.15,\ \gamma_1 = 0.001,\ \gamma_2 = 0.0016,\ \pi_1 = 0.04,\\ &\pi_2 = 0.09,\ \nu_1 = 0.005,\ \nu_2 = 0.0004,\ \beta_1 = 0.025,\ \beta_2 = 0.02,\ K_0 = 10.0,\\ &b_{11} = 0.2,\ b_{12} = 1.0,\ b_{21} = 0.1,\ b_{22} = 2.0. \end{aligned} \tag{18b}$$

For this set of the parameters, the equilibrium point $E_2(N^*, N_D^*, T^*, T_s^*, U^*, U_s^*)$ attains the value as $N^* = 9.10873,\ N_D^* = 0.79867,\ T^* = 0.02093,\ T_s^* = 0.03242,\ U^* = 0.00270,\ U_s^* = 0.01998$. From (8), $\frac{dF}{dN} = 0.19893 > 0$ shows that E_2 is unique. The local stability conditions (11) corresponding to E_2 are also satisfied.

In Figure 2, we show the changes in total density and density of deformed subclass of biological species with respect to the value of parameter θ upto 1 (keep the remaining parameters same as (18b)). Consequently, Figure 2 shows that when the fraction of the part of primary toxicant (kT), which is converted into secondary toxicant, increases, the total density decreases and the density of deformed subclass increases.

Figure 3 shows that the total density of species decreases and that the density of deformed subclass of species increases as conversion rate of primary toxicant into secondary toxicant k increase upto 1 (the other parameter same as (18b)).

Figure 4 shows the variation in total and deformed subclass densities corresponding to the discharge rate of primary toxicant Q. As the discharge rate Q increases, the total density of species N decreases and deformed subclass density N_D firstly increases then decreases with N. For large value $Q(= 0.200)$ both densities N and N_D start oscillating. Hence, there exists a Hopf-bifurcation corresponding to the parameter Q.

Figure 5 confirms that Hopf-bifurcation exists with respect to the parameter Q at critical value $Q^* = 0.13628$. This figure shows the real and imaginary parts of eigenvalues of Jacobian matrix M_2 for the parameter Q. The real parts of all eigenvalues (i.e., $R_i < 0,\ i = 1, \ldots, 6$) are negative for $Q < Q^*(= 0.13628)$. At $Q = Q^*$ two eigenvalues become purely imaginary (i.e.,

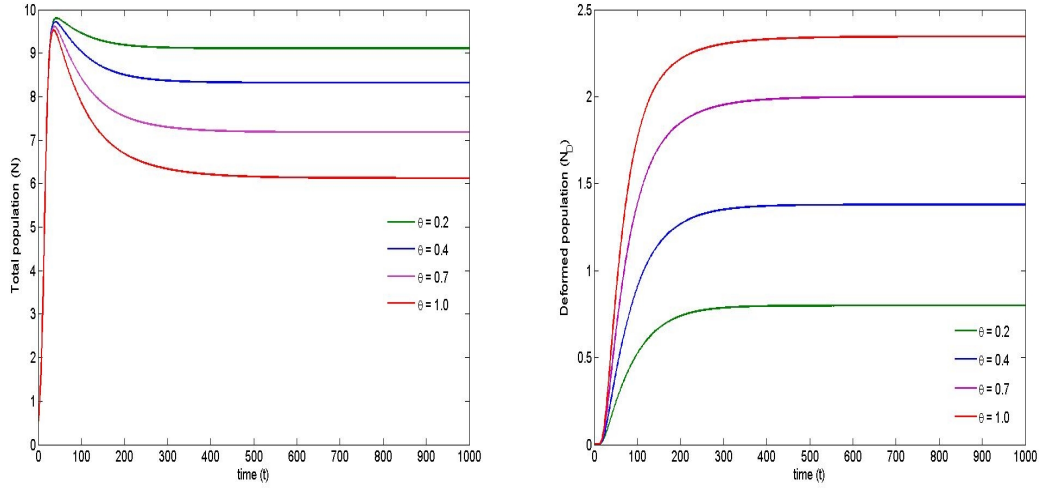


Figure 2: Time series graph for total and deformed subclass densities of species respect to the parameter θ.

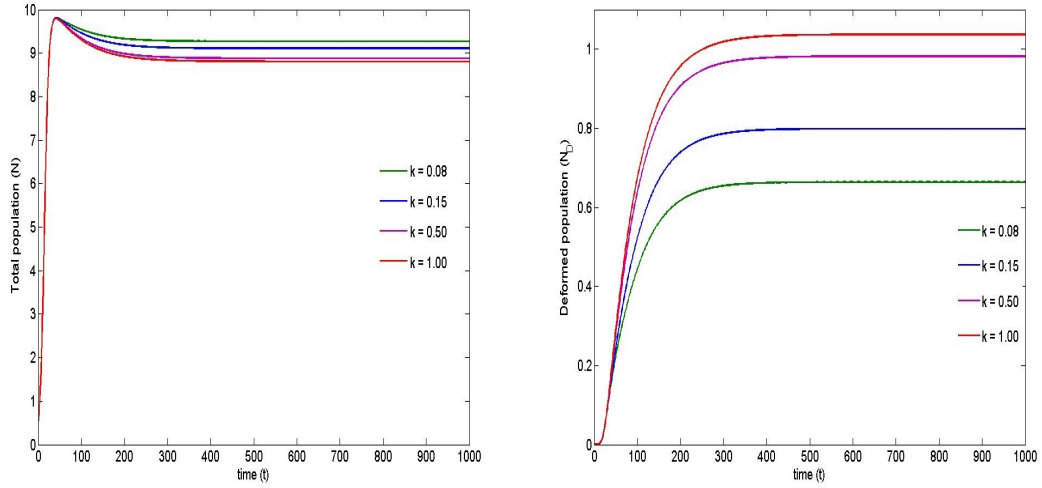


Figure 3: Time series graph for total and deformed subclass densities of species respect to the parameter k.

$R_1 = R_2 = 0$ and $I_1 = -I_2 \neq 0$) and other four eigenvalues have negative real parts (i.e., $R_i < 0$, $i = 3, \ldots, 6$).

Figure 6 shows the behavior of both densities N and N_D and also nature of bifurcating periodic solutions with respect to the discharge rate of primary toxicant Q. Both densities N and N_D become stable at equilibrium level for $Q < Q^*$. When Q crosses the critical value $Q^*(= 0.13628)$, the equilibrium point losses its stability and a supercritical Hopf-bifurcation oc-

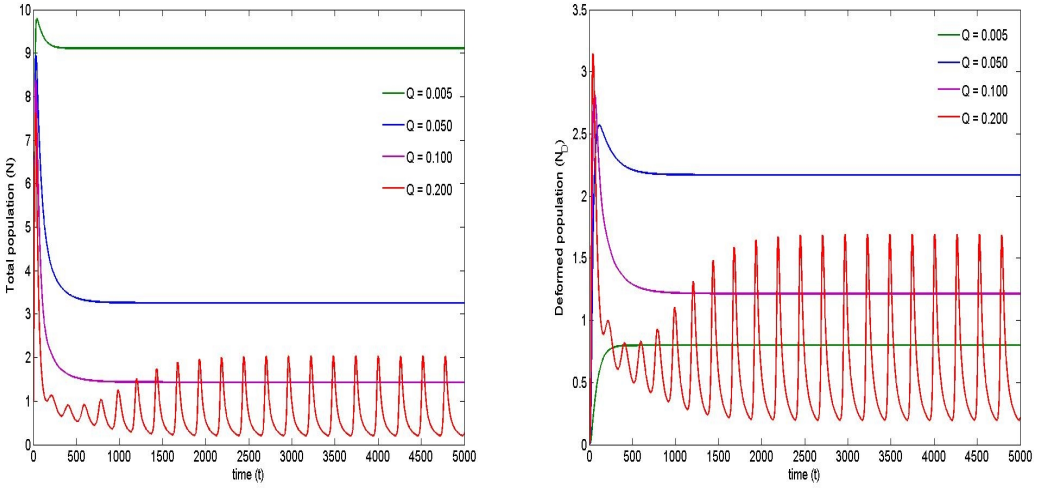


Figure 4: Time series graph for total and deformed population respect to the parameter Q.

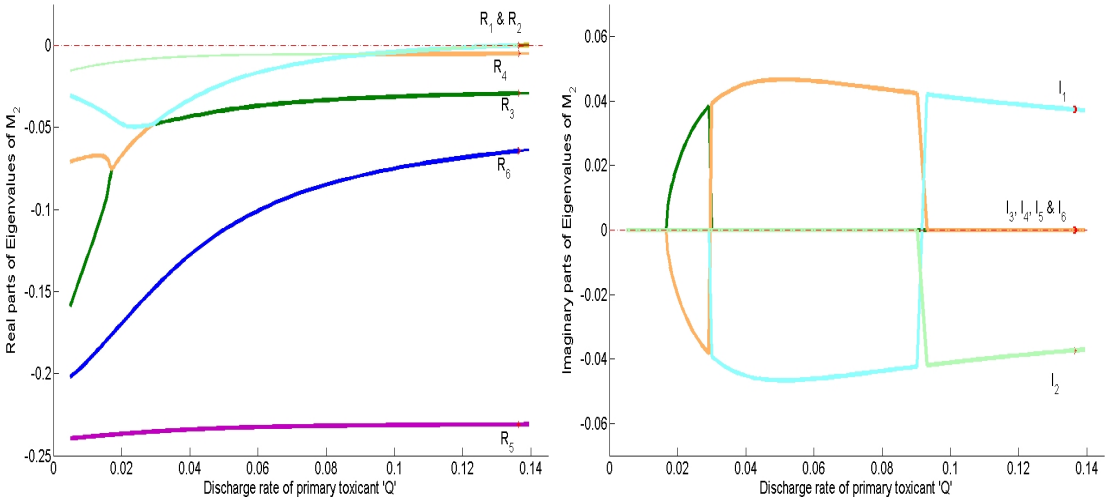


Figure 5: Real and Imaginary parts of eigenvalues of Jacobian matrix M_2 respect to the emission rate Q.

curs (since $\mu_2 = 1.8418 * 10^{-12} > 0$). Both densities start oscillating around their equilibrium level with stable bifurcating periodic orbits (since $\beta_2 = -1.2169 * 10^{-14} < 0$).

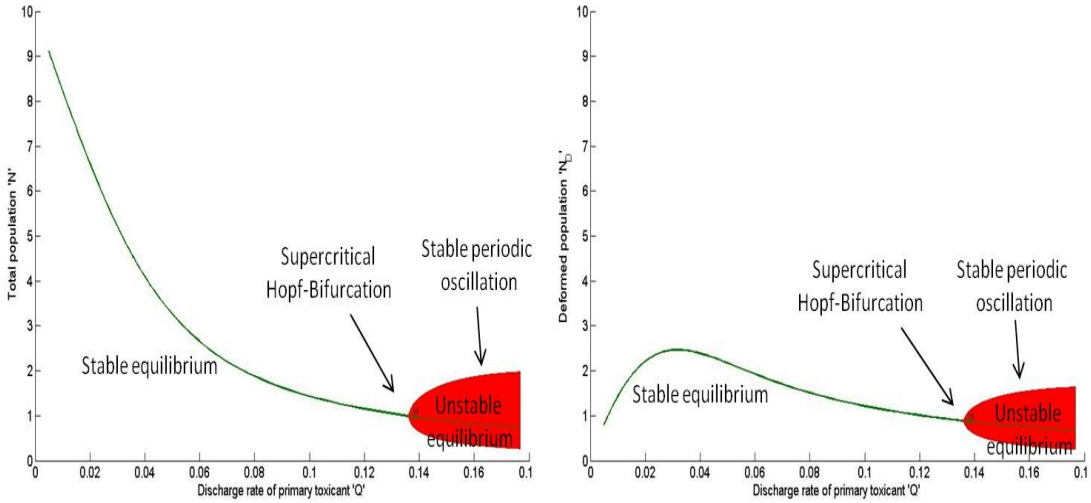


Figure 6: Dynamic behavior of densities of total and deformed populations for emission rate Q.

6 Conclusion

This paper is based on effect of primary and secondary toxicants on a biologically species. We have proposed and analyzed a mathematical model for the case in which after-effect of primary and secondary toxicants, a subclass of biological species shows deformity such as stagnating the reproduction capability. The stability analysis of model shows that total density decreases and deformed subclass density increases as the conversion rate of primary toxicant into secondary toxicant increases. If the discharge rate of primary toxicant into the environment increases, then total density of species decreases and deformed subclass density firstly increases then decreases with total density. If discharge rate crosses the critical value, then both densities start oscillations around their equilibrium and system shows a supercritical Hopf-bifurcation. All the bifurcating periodic solutions of model system are stable in nature.

A Proof of the region of attraction Ω

Proof. From the model system (3), we have

$$\frac{dN}{dt} \leq rN - \frac{rN^2}{K_0} = r\left(1 - \frac{N}{K_0}\right)N.$$

Thus, $\limsup_{t\to\infty} N(t) \leq K_0$. Also, we have

$$\begin{aligned}\frac{dT}{dt}+\frac{dU}{dt} &= Q-(\delta_1+k)T-\beta_1 U-(1-\pi_1)\nu_1 NU\\ &\leq Q-\delta_{m_1}(T+U),\end{aligned}$$

where $\delta_{m_1}=\min(\delta_1+k,\beta_1)$. Thus, $\limsup_{t\to\infty}(T(t)+U(t))\leq \frac{Q}{\delta_{m_1}}$.
Similarly,

$$\begin{aligned}\frac{dT_s}{dt}+\frac{dU_s}{dt} &= \theta kT-\delta_2 T_s-\beta_2 U_s-(1-\pi_2)\nu_2 NU_s\\ &\leq \theta k\frac{Q}{\delta_{m_1}}-\delta_{m_2}(T_s+U_s),\end{aligned}$$

where $\delta_{m_2}=\min(\delta_2,\beta_2)$. Thus, $\limsup_{t\to\infty}(T_s(t)+U_s(t))\leq \frac{\theta kQ}{\delta_{m_1}\delta_{m_2}}$.

From the second equation of model (3), we have

$$\begin{aligned}\frac{dN_D}{dt} &= (r_1U+r_2U_s)(N-N_D)-\frac{rN_DN}{K(T,T_s)}-(\alpha+d)N_D\\ &\leq (r_1\delta_{m_2}+r_2\theta k)\frac{Q}{\delta_{m_1}\delta_{m_2}}(K_0-N_D)-(\alpha+d)N_D.\end{aligned}$$

Thus, $\limsup_{t\to\infty}N_D(t)\leq \frac{(r_1\delta_{m_2}+r_2\theta k)QK_0}{[(r_1\delta_{m_2}+r_2\theta k)Q+(\alpha+d)\delta_{m_1}\delta_{m_2}]}$, proving the region of attraction Ω. □

B Coefficients of characteristic equation (10)

$$\begin{aligned}c_1 &= -(m_{11}+m_{22}+m_{33}+m_{44}+m_{55}+m_{66}),\\ c_2 &= (m_{11}+m_{22})(m_{55}+m_{66})+m_{22}(m_{33}+m_{44})\\ &\quad +(m_{33}+m_{55})(m_{44}+m_{66})\\ &\quad +\begin{vmatrix}m_{11} & m_{12}\\ m_{21} & m_{22}\end{vmatrix}+\begin{vmatrix}m_{11} & m_{13}\\ m_{31} & m_{33}\end{vmatrix}+\begin{vmatrix}m_{11} & m_{14}\\ m_{41} & m_{44}\end{vmatrix}+\begin{vmatrix}m_{33} & m_{35}\\ m_{53} & m_{55}\end{vmatrix}+\begin{vmatrix}m_{44} & m_{46}\\ m_{64} & m_{66}\end{vmatrix},\\ c_3 &= -\{m_{11}m_{55}m_{66}+m_{12}(m_{25}m_{51}+m_{26}m_{61})\\ &\quad +m_{22}(m_{33}+m_{55})(m_{44}+m_{66})\}+m_{13}\begin{vmatrix}m_{31} & m_{35}\\ m_{51} & m_{55}\end{vmatrix}\\ &\quad +m_{14}\left\{\begin{vmatrix}m_{41} & m_{46}\\ m_{61} & m_{66}\end{vmatrix}-\begin{vmatrix}m_{31} & m_{33}\\ m_{41} & m_{43}\end{vmatrix}\right\}\\ &\quad -m_{31}\begin{vmatrix}m_{12} & m_{13}\\ m_{22} & m_{23}\end{vmatrix}-m_{41}\begin{vmatrix}m_{12} & m_{14}\\ m_{22} & m_{24}\end{vmatrix}\end{aligned}$$

$$
\begin{aligned}
&- m_{55}\begin{vmatrix} m_{11} & m_{14} \\ m_{41} & m_{44} \end{vmatrix} - (m_{44}+m_{66})\begin{vmatrix} m_{11} & m_{13} \\ m_{31} & m_{33} \end{vmatrix} \\
&- (m_{11}+m_{22}+m_{44}+m_{66})\begin{vmatrix} m_{33} & m_{35} \\ m_{53} & m_{55} \end{vmatrix} \\
&- (m_{11}+m_{22}+m_{33}+m_{55})\begin{vmatrix} m_{44} & m_{46} \\ m_{64} & m_{66} \end{vmatrix} \\
&- (m_{33}+m_{44}+m_{55}+m_{66})\begin{vmatrix} m_{11} & m_{12} \\ m_{21} & m_{22} \end{vmatrix},
\end{aligned}
$$

$$
\begin{aligned}
c_4 = m_{12}\Bigg\{ & m_{26}m_{61}(m_{33}+m_{55}) + m_{25}m_{51}(m_{44}+m_{66}) + m_{31}\begin{vmatrix} m_{23} & m_{25} \\ m_{53} & m_{55} \end{vmatrix} \\
&+ m_{41}\begin{vmatrix} m_{24} & m_{26} \\ m_{64} & m_{66} \end{vmatrix} - m_{51}\begin{vmatrix} m_{23} & m_{25} \\ m_{33} & m_{35} \end{vmatrix} - m_{61}\begin{vmatrix} m_{24} & m_{26} \\ m_{44} & m_{46} \end{vmatrix} \Bigg\} \\
&+ m_{31}(m_{44}+m_{66})\begin{vmatrix} m_{12} & m_{13} \\ m_{22} & m_{23} \end{vmatrix} + m_{41}(m_{33}+m_{55})\begin{vmatrix} m_{12} & m_{14} \\ m_{22} & m_{24} \end{vmatrix} \\
&- m_{13}(m_{22}+m_{44}+m_{66})\begin{vmatrix} m_{31} & m_{35} \\ m_{51} & m_{55} \end{vmatrix} \\
&- m_{14}(m_{22}+m_{33}+m_{55})\begin{vmatrix} m_{41} & m_{46} \\ m_{61} & m_{66} \end{vmatrix} \\
&+ m_{43}\Bigg\{ m_{14}m_{31}m_{66} + m_{14}\begin{vmatrix} m_{31} & m_{35} \\ m_{51} & m_{55} \end{vmatrix} - m_{31}\begin{vmatrix} m_{12} & m_{14} \\ m_{22} & m_{24} \end{vmatrix} \Bigg\} \\
&+ \Bigg\{ m_{11}m_{55} + m_{22}(m_{33}+m_{55}) + \begin{vmatrix} m_{11} & m_{13} \\ m_{31} & m_{33} \end{vmatrix} \Bigg\}\begin{vmatrix} m_{44} & m_{46} \\ m_{64} & m_{66} \end{vmatrix} \\
&+ \Bigg\{ (m_{33}+m_{55})(m_{44}+m_{66}) + \begin{vmatrix} m_{33} & m_{35} \\ m_{53} & m_{55} \end{vmatrix} + \begin{vmatrix} m_{44} & m_{46} \\ m_{64} & m_{66} \end{vmatrix} \Bigg\}\begin{vmatrix} m_{11} & m_{12} \\ m_{21} & m_{22} \end{vmatrix} \\
&+ \Bigg\{ m_{11}m_{66} + m_{22}(m_{44}+m_{66}) + \begin{vmatrix} m_{11} & m_{14} \\ m_{41} & m_{44} \end{vmatrix} + \begin{vmatrix} m_{44} & m_{46} \\ m_{64} & m_{66} \end{vmatrix} \Bigg\}\begin{vmatrix} m_{33} & m_{35} \\ m_{53} & m_{55} \end{vmatrix},
\end{aligned}
$$

$$
\begin{aligned}
c_5 = m_{31}m_{43}\Bigg\{ & m_{66}\begin{vmatrix} m_{12} & m_{14} \\ m_{22} & m_{24} \end{vmatrix} - m_{12}m_{26}m_{64} \Bigg\} \\
&+ m_{43}\Bigg\{ \begin{vmatrix} m_{12} & m_{14} \\ m_{22} & m_{24} \end{vmatrix} - m_{14}m_{66} \Bigg\}\begin{vmatrix} m_{31} & m_{35} \\ m_{51} & m_{55} \end{vmatrix} \\
&+ (m_{33}+m_{55})\Bigg\{ m_{14}m_{22}\begin{vmatrix} m_{41} & m_{46} \\ m_{61} & m_{66} \end{vmatrix} + m_{61}m_{12}\begin{vmatrix} m_{24} & m_{26} \\ m_{44} & m_{46} \end{vmatrix}
\end{aligned}
$$

$$
\begin{aligned}
&- m_{12}m_{41}\begin{vmatrix} m_{24} & m_{26} \\ m_{64} & m_{66} \end{vmatrix} \Bigg\} \\
&+ (m_{44}+m_{66})\Bigg\{ m_{12}m_{51}\begin{vmatrix} m_{23} & m_{25} \\ m_{33} & m_{35} \end{vmatrix} - m_{12}m_{31}\begin{vmatrix} m_{23} & m_{25} \\ m_{53} & m_{55} \end{vmatrix} \Bigg\} \\
&+ m_{13}\Bigg\{ m_{22}(m_{44}+m_{66}) + \begin{vmatrix} m_{44} & m_{46} \\ m_{64} & m_{66} \end{vmatrix} \Bigg\} \begin{vmatrix} m_{31} & m_{35} \\ m_{51} & m_{55} \end{vmatrix} \\
&- \Bigg\{ m_{12}m_{25}m_{51} + m_{31}\begin{vmatrix} m_{12} & m_{13} \\ m_{22} & m_{23} \end{vmatrix} \Bigg\} \begin{vmatrix} m_{44} & m_{46} \\ m_{64} & m_{66} \end{vmatrix} \\
&- \Bigg\{ (m_{44}+m_{66})\begin{vmatrix} m_{33} & m_{35} \\ m_{53} & m_{55} \end{vmatrix} + (m_{33}+m_{55})\begin{vmatrix} m_{44} & m_{46} \\ m_{64} & m_{66} \end{vmatrix} \Bigg\} \begin{vmatrix} m_{11} & m_{12} \\ m_{21} & m_{22} \end{vmatrix} \\
&- \Bigg\{ (m_{11}+m_{22})\begin{vmatrix} m_{44} & m_{46} \\ m_{64} & m_{66} \end{vmatrix} + m_{12}m_{26}m_{61} + m_{41}\begin{vmatrix} m_{12} & m_{14} \\ m_{22} & m_{24} \end{vmatrix} \\
&- m_{14}\begin{vmatrix} m_{41} & m_{46} \\ m_{61} & m_{66} \end{vmatrix} \Bigg\} \begin{vmatrix} m_{33} & m_{35} \\ m_{53} & m_{55} \end{vmatrix}
\end{aligned}
$$

$$
\begin{aligned}
c_6 = {} & \begin{vmatrix} m_{11} & m_{12} \\ m_{21} & m_{22} \end{vmatrix} \begin{vmatrix} m_{33} & m_{35} \\ m_{53} & m_{55} \end{vmatrix} \begin{vmatrix} m_{44} & m_{46} \\ m_{64} & m_{66} \end{vmatrix} + \begin{vmatrix} m_{12} & m_{14} \\ m_{22} & m_{24} \end{vmatrix} \begin{vmatrix} m_{33} & m_{35} \\ m_{53} & m_{55} \end{vmatrix} \begin{vmatrix} m_{41} & m_{46} \\ m_{61} & m_{66} \end{vmatrix} \\
&+ \begin{vmatrix} m_{31} & m_{35} \\ m_{51} & m_{55} \end{vmatrix} \begin{vmatrix} m_{12} & m_{13} \\ m_{22} & m_{23} \end{vmatrix} \begin{vmatrix} m_{44} & m_{46} \\ m_{64} & m_{66} \end{vmatrix} - m_{12}m_{26} \begin{vmatrix} m_{41} & m_{44} \\ m_{61} & m_{64} \end{vmatrix} \begin{vmatrix} m_{33} & m_{35} \\ m_{53} & m_{55} \end{vmatrix} \\
&- m_{12}m_{25} \begin{vmatrix} m_{31} & m_{33} \\ m_{51} & m_{53} \end{vmatrix} \begin{vmatrix} m_{44} & m_{46} \\ m_{64} & m_{66} \end{vmatrix} \\
&+ \Bigg\{ m_{12}m_{26}m_{43}m_{64} - m_{43}m_{66}\begin{vmatrix} m_{12} & m_{14} \\ m_{22} & m_{24} \end{vmatrix} \Bigg\} \begin{vmatrix} m_{31} & m_{35} \\ m_{51} & m_{55} \end{vmatrix},
\end{aligned}
$$

where $\begin{vmatrix} m_1 & m_2 \\ m_3 & m_4 \end{vmatrix} = m_1m_4 - m_2m_3$.

Acknowledgements

The first author is heartily thankful to Prof.(Dr.) Joydip Dhar, Engineering Sciences, Atal Bihari Vajpayee - Indian Institute of Information Technology and Management, Gwalior, INDIA, for his help available all the time.

Funding

This research received no external funding.

Conflicts of Interest

The authors declare no conflicts of interest.

References

[1] Agrawal, A.K. and Shukla, J.B. *Effect of a toxicant on a biological population causing severe symptoms on a subclass*, South Pacific J. Pure Appl. Math. 1 (2012) 12–27.

[2] Dhooge, A., Govaerts, W. and Kuznetsov, Y.A. *MATCONT, a MATLAB package for numerical bifurcation analysis of ODEs*, ACM Trans. Math. Softw. (TOMS), 29 (2003), 141–164.

[3] Dickman, M. and Rygiel, G. *Chironomid larval deformity frequencies, mortality, and diversity in heavy - metal contaminated sediments of a Canadian riverine wetland*, Environ. Inter. 22 (1996), 693–703.

[4] Freedman, H.I. and Shukla, J.B. *Models for the effect of toxicant in single species and predator - prey systems*, J. Math. Biol., 30 (1991), 15–30.

[5] Fry, D.M. *Reproductive effects in birds exposed to pesticides and industrial chemicals*, Environ. Health Perspect. 103 (1995), 165.

[6] Hassard, B., Kazarinoff, N. and Wan, Y. *Theory and application of Hopf bifurcation*, London Mathematical Society Lecture Note Series, Cambridge University Press, Cambridge, UK, 41, 1981.

[7] Kumar, A. and Agrawal, A. *Existence of hopf-bifurcation in a 6- dimensional system*, Commun. Appl. Anal. 21 (2017) 119–126.

[8] Kumar, A., Agrawal, A.K., Hasan, A. and Misra, A.K. *Modeling the effect of toxicant on the deformity in a subclass of a biological species*, Model. Earth Syst. Environ., 2 (2016) 40.

[9] Kuznetsov, Y. *Elements of applied bifurcation theory*, Springer-Verlag, New York, 2004.

[10] Liu, W. *Criterion of Hopf Bifurcation without using Eigenvalues*, J. Math. Anal. App., 182 (1994), 250–256.

[11] Miguel, B.L., Donald, J.B., Ravinder, S.S. and Fernando, G.F. *Induction of Morphological Deformities and Moulting Alterations in Litopenaeus vannamei (Boone) Juveniles Exposed to the Triazole - Derivative Fungicide Tilt*, Arch. Environ. Contam. Toxicol, 51 (2006), 69–78.

[12] Patil, V.K. and David, M. *Behavioral and morphological endpoints: as an early response to sublethal malathion intoxication in the freshwater fish, Labeo rohita*, Drug Chem. Toxicol. 33 (2010), 160–165.

[13] Ronit, H.K. and Eldad, B.C. *The effect of colchicine treatment on sperm production and function: a review*, Hum. Reprod., 13 (1998), 360-362.

[14] Saquib, M., Ahmad, A. and Ansari, K. *Morphological and physiological responses of Croton Bonplandianum baill to air pollution*, Ecoprint: Inter. J. Ecol. 17 (2010), 35–41.

[15] Seydel, R. *Practical bifurcation and stability analysis*, Springer Science & Business Media, 2009.

[16] Sharara, F.I., Seifer, D.B. and Flaws, J.A. *Environmental toxicants and female reproduction*, Fertil. steril. 70 (1998), 613–622.

[17] Shukla, J.B. and Agrawal, A.K. *Some mathematical models in ecotoxicology; Effects of toxicants on biological species*, Sadhana, 24 (1999), 25–40.

[18] Shukla, J.B., Agrawal, A.K., Sinha, P. and Dubey, B. *Modeling effects of primary and secondary toxicants on renewable resources*, Nat. Resour. Model. 16 (2003), 99–120.

[19] Veeramachaneni, D.N.R., Palmer, J.S. and Amann, R.P. *Long term effects on male reproduction of early exposure to common chemical contaminants in drinking water*, Hum. Reprod., 16 (2001), 979–987.

Iranian Journal of Numerical Analysis and Optimization
Vol. 16, No. 3, 2026, pp 1107–1132
https://doi.org/10.22067/ijnao.2026.97170.1815
https://ijnao.um.ac.ir/

CC BY NC

Research Article

Analysis of 2D Burger's equation using numerical and semi-analytic techniques

A. Kaur, A. Pasrija and S. Arora*

Abstract

The effect of advective and diffusive components is studied using the 2D Burgers' equation. The diffusive term is influenced by the Reynolds number. Spatial discretization is performed using radial basis collocation, while the temporal direction is discretized using 4th Runge–Kutta method. To analyze the error behavior, the closed-form solution of 2D Burgers' equation is derived analytically using a generalized bivariate homotopy perturbation approach. A comprehensive discussion and comparison of both numerical and semi-analytical solutions are presented in terms of absolute error, as well as $\|e\|_\infty$ and $\|e\|_2$ error norms. The effect of Reynolds number on the diffusion component is illustrated graphically. Algorithms for both numerical and analytical techniques have been implemented in MATLAB R2023.

AMS subject classifications (2020): Primary 65D12, 65N12; Secondary 35B25, 65N35.

*Corresponding author

Received 24 December 2025; revised 21 April 2026; accepted 29 April 2026

Amandeep Kaur
Department of Mathematics, Punjabi University, Patiala-147002, Punjab, India. e-mail: amandeepk_rs23@pbi.ac.in

Atul Pasrija
Department of Computer Applications, CGC, Landran, Mohali-140307, Punjab, India. e-mail: atulpasrija94678@gmail.com

Shelly Arora
Department of Mathematics, Punjabi University, Patiala-147002, Punjab, India. e-mail: aroshelly@gmail.com

How to cite this article
Kaur, A., Pasrija, A. and Arora, S., Analysis of 2D Burger's equation using numerical and semi-analytic techniques. *Iran. J. Numer. Anal. Optim.*, 2026; 16(3): 1107-1132. https://doi.org/10.22067/ijnao.2026.97170.1815

Keywords: Radial basis functions, Runge–Kutta method, GB transform, Burger's equation.

1 Introduction

The transport phenomenon in advection-diffusion equation is best represented by Navier–Stokes equation. It represents the transfer of heat and energy from one medium to another. Burger's equation is a special case of Navier Stoke's equation. This equation has numerous applications in fluid dynamics, physics, solid state physics, aeronautics, turbulence theory, and so on. Bateman [6] developed Burger's equation from Navier–Stokes by excluding the pressure gradient term and later, in 1948, Burger has redefined it. In the present scenario, the application of Burger's equation is not confined to fluid dynamics, also includes traffic flow [11], cosmology [44], gas dynamics [9], acoustics [29], continuous stochastic processes [10], electromagnetic waves [27] and many more.

The advective and diffusive components are represented by nondissipative hyperbolic component and the dissipative parabolic component, respectively.

Consider the two-dimensional (2D) Burgers' equation defined on a bounded domain Ω as follows:

$$\vartheta_t + \alpha\vartheta\nabla\vartheta - \frac{1}{Re}\nabla^2\vartheta = f(X,t),\ X \in \Omega \subset \mathbb{R}^2, \quad t \geq 0, \tag{1}$$

where ∇ is the gradient operator and ∇^2 is the Laplacian operator, Re represents the Reynolds number, α is any parameter such that $\alpha \geq 0$, and $f(X,t)$ are a continuous function defined on the domain Ω and for $t \geq 0$. The initial condition is defined as follows:

$$\vartheta(X,0) = \psi(X), \qquad X \in \Omega, \tag{2}$$

and the boundary conditions are defined as

$$\mathcal{B}\vartheta(X,t) = \phi(X,t), \qquad X \in \partial\Omega,\ t \geq 0, \tag{3}$$

where the function ϑ describes advection and diffusion quantities such as heat, mass, energy , and so on, over the spatial domain Ω at time $t \geq 0$, and $\mathcal{B}$ represents the boundary operator on the boundary of the spatial domain. The nonlinear term with ∇ operator accounts for advection and Re is the Reynolds number, which is inversely proportional to the viscosity parameter that measures the rate of diffusion. Moreover, $\psi(X)$ is the bounded function defined on the domain Ω and $\phi(X,t)$ are continuous functions defined on the boundary $\partial\Omega$.

The linear advection-diffusion equation has been studied in [35] through the framework of finite elements. The coupling of nonstandard finite difference and Haar wavelets with quasi-linearization for a class of generalized Burger's equations has been discussed in [26]. Polynomial differential quadrature method with Runge–Kutta (RK)4 was implemented in [31] to obtain the numerical results for one-dimensional D and 2D Burgers' equation systems. Nonuniform grid friendly localized differential quadrature method was proposed in [4] to solve viscous Burger's equation in one and two dimensions. Fully implicit finite difference scheme was discussed in [5, 21] for 2D Burger's equation.

Nevertheless, scholars have contributed with several analytic methods to determine the explicit solutions of linear and nonlinear advection-diffusion equation such as [33, 32, 1]. There are several other numerical techniques followed by investigators to solve Burger's equation such as the sine and cosine method [22], inverse scattering technique [17], the simplest equation technique [36], extended simplest equation technique [8], Kudryashov method [23], Backlund transformation technique [19], exp-function technique [30], Hirota's method [43], homogeneous balance technique [20], bifurcation method [28], Darboux transformation [45], G'/ G-expansion technique [34], Lie group analysis [42] and weighted finite difference method [37] and so on.

In the present study, the semi-analytic and numerical perspectives have been explored by generalized bivariate homotopy perturbation method (GB-HPM) and 4th order RK method, respectively, with mesh-less approach based on radial basis functions (RBFs).

2 Radial basis functions

Conventional mesh generates methods such as finite difference, finite element and spectral methods suffer from the complexity of mesh-generation for multidimensional problems. Alleviating the labor of mesh-generation, mesh-less methods are developed with enough flexibility to extend their applicability to irregular geometries, scattered and multivariate data. RBFs with collocation method have become an important tool to study the multidimensional problems.

RBFs were first introduced by Hardy [18] to study the topographical mapping. Later on the origination of Kansa's work on collocation technique with RBFs revolutionized the perspective of the study of partial differential equations using RBFs [24]. These functions basically, represent the distance between the node point $\mathbf{X_i}$ and the data point $\mathbf{X}$ and are independent of the space dimension and isotropy. In simple words RBFs are univariate function whose value depend upon distance between the point and the center point of the function. An RBF can be defined as follows:

A function $\Phi : \mathcal{R}^d \rightarrow \mathcal{R}$ is said to be RBF if there exists a radial distance r such that $r = \|X - X_k\|$ which implies $r = \sqrt{(\xi - \xi_k)^2 + (\zeta - \zeta_k)^2}$ and $\Phi(X) = \Phi(r)$, where $\|.\|$ denotes Euclidean norm.

By the time, various RBFs have been developed to study the partial differential equations such as Gaussian (GA), multiquadric (MQ), inverse multiquadric (IMQ), and so on. The list of commonly used RBFs is given in Table 1. From this table, it is observed that an unknown parameter ν is introduced which is a free parameter, commonly known as the shape parameter.

Table 1: List of RBFs

RBFs	Definition
GA	$\Phi(\mathbf{r}) = e^{-(\nu \mathbf{r})^2}$
MQ	$\Phi(\mathbf{r}) = \sqrt{1 + (\nu \mathbf{r})^2}$
IMQ	$\Phi(\mathbf{r}) = \frac{1}{\sqrt{1+(\nu \mathbf{r})^2}}$
Inverse quadric	$\Phi(\mathbf{r}) = \frac{1}{1+(\nu \mathbf{r})^2}$
Thin plate spline	$\Phi(\mathbf{r}) = \mathbf{r}^2 ln(\mathbf{r})$
Monomial	$\Phi(\mathbf{r}) = \mathbf{r}^{2k-1}; k \in N$

Shape parameter

Along with the choice of RBFs, the shape parameter possesses great influence on the accuracy and condition number of the coefficient matrix of RBFs [40]. Infinitely smooth RBFs such as MQ and GA manifest exponential convergence as the shape parameter approaches infinity. However, it further leads to high condition number associated with ill-conditioning of the system. The open problem of selecting optimal shape parameter has engrossed several studies with generation of different criterion to investigate the good choice of shape parameter. Trial and error method is most common in practice to pick the best shape parameter that has least error. When exact solution is not given for 2D problems following strategies have been adopted to choose the shape parameter.

Hardy [18] suggested the shape parameter for IMQ RBFs as

$$\nu = \frac{1}{0.815d}, \qquad d = \frac{1}{N}\sum_{j=1}^{N} d_j,$$

where N is the number of data points and d_j represents the distance to the nearest data point in neighbor.

Franke [16] formulated shape parameter for 2D systems as

$$\nu = \frac{0.8\sqrt{N}}{D},$$

where D is the diameter of the smallest circle which includes all data points.

Fasshauer [14] elaborated the good choice for shape parameter as

$$\nu = \frac{2}{\sqrt{N}}.$$

In [14], these strategies are remarked as optional for very specific examples and certainly cannot be considered as general instructions. Carlson and Foley [13] stated the shape parameter as problem dependent. Borrowing idea from method of cross validation practiced in statistics literature, Rippa [39] proposed special case of leave-one-out cross validation to circumvent the issue of selecting optimal shape parameter in context of interpolating scattered data with RBFs. Moreover, Fasshauer and Zhang [15] reviewed LOOCV and introduced its extension for pseudo-spectral methods for solving partial differential equations. Biazar and Hosami [7] discussed the strategy to find an interval for the optimal shape parameter rather than constant value. The leading stumbling block of global collocation is the severe ill-conditioning of the system matrix. Using a greater value for shape parameter leads to the ill-conditioned interpolation matrix, with a condition number grows exponentially large. Addressing this issue, in [25], authors have remarked on the tolerance of RBFs for a large condition number of matrix upto $\mathcal{O}(10^{16})$ within machine double precision. The ill-conditioning has a significant impact on the results.This behavior is called the uncertainty principle. The present study relies on the algorithm drawn in Algorithm 1.

Algorithm 1 Algorithm for Shape Parameter

Input: $\mathcal{CD}_{\min}, \mathcal{CD}_{\max}, \nu_{increment}$
$\mathcal{CD} = 1$
while $\mathcal{CD} > \mathcal{CD}_{\max}$ and $\mathcal{CD} < \mathcal{CD}_{\min}$ **do**
  form required matrix say ;
  use P,Q,R = $svd(\ )$;
  $\mathcal{CD} = Q(1,1)/Q(end, end)$;
  if $\mathcal{CD} < \mathcal{CD}_{\min}$ **then** $\nu = \nu$ - $\nu_{increment}$;
  else if $\mathcal{CD} > \mathcal{CD}_{\max}$ **then** $\nu = \nu + \nu_{increment}$;
  end if
end while
Output: ν

3 Implementation of RBFs with collocation method

Collocation technique is one of the weighted residual method in which the residual is set equal to zero at collocation points. Therefore, in RBF collocation method, the node points are taken to be the collocation points. The unconditional positive definite feature of GA RBFs guarantees the solvability of the resulting system which ensures the implementation of GA as base functions to implement the collocation technique. In the present study, the GA RBF Φ_k has been proposed to define the trial function. The GA RBFs are defined as:

$$\Phi_k(r) = \exp(-\nu^2 r^2), \qquad k = 1, 2, \ldots, N, \tag{4}$$

where $r = \sqrt{(\xi - \xi_k)^2 + (\zeta - \zeta_k)^2}$, $k = 1, 2, \ldots, N$.

Let $\bar{\vartheta}$ be the approximation of $\vartheta(r, t)$ in terms of GA RBFs Φ with a given set of node points $\{X_1, X_2, \ldots, X_N\} \subset \Omega$ can be expressed as

$$\bar{\vartheta} = \sum_{k=1}^{N} \Phi_k(r) C_k(t), \tag{5}$$

where C_k's are unknown coefficients and N is the number of collocation points which coincide with the node points. The node points r are given by $r = \|X - X_k\| = \sqrt{(\xi - \xi_k) + (\zeta - \zeta_k)^2}$.

The adjustment of node points is the main task to implement the collocation technique. Let N_{int} be the node points in the interior of the domain and let N_b be the node points on the boundary such that the total number of node points is $N = N_{int} + N_b$. The uniform distribution of node points is shown in Figure 1 for $N = 400$ and the flowchart for shape parameter is given in Figure 2. The trial function $\vartheta(r, t)$ at time t_p and the jth node point can be written as

$$\bar{\vartheta}_j^p = \sum_{k=1}^{N} \Phi_k(r_j) C_k^p, \qquad j = 1, 2, \ldots, N. \tag{6}$$

In the matrix form, (6) can be written as

$$\vartheta(r_j, t_p) = \ . \ , \tag{7}$$

where $_{\mathbf{j,k}} = \left[\Phi_k(r_j)\right]_{N \times N}$, $1 \le j, k \le N$ is the matrix of RBFs and $\ = [C_1^p, C_2^p, \ldots, C_N^p]'$ denotes the vector of expansion coefficients. The RBF matrix is defined below:

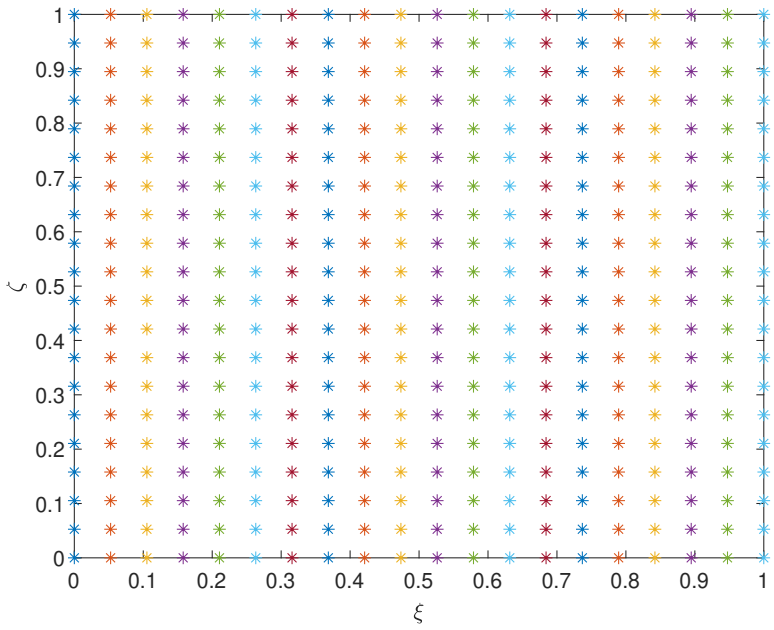


Figure 1: Uniform point distribution for N = 400.

$$\begin{bmatrix} \Phi_1(r_1) & \Phi_2(r_1) & \Phi_3(r_1) & \dots & \Phi_N(r_1) \\ \Phi_1(r_2) & \Phi_2(r_2) & \Phi_3(r_2) & \dots & \Phi_N(r_2) \\ \Phi_1(r_3) & \Phi_2(r_3) & \Phi_3(r_3) & \dots & \Phi_N(r_3) \\ \vdots & \vdots & \vdots & \ddots & \vdots \\ \Phi_1(r_N) & \Phi_2(r_N) & \Phi_3(r_N) & \dots & \Phi_N(r_N) \end{bmatrix},$$

where $\Phi_k(r_j) = \exp(-\nu^2 r_j^2)$ and $r_j = \sqrt{(\xi_j - \xi_k)^2 + (\zeta_j - \zeta_k)^2}$.

4 Temporal discretization using 4th order Runge–Kutta Method

Implementing (6) on (1) and (3) at t_p, the following system of equations is obtained:

$$\sum_{k=1}^{N} \Phi_{jk} \frac{dC_k}{dt} = \frac{1}{Re} \sum_{k=1}^{N} \nabla^2(\Phi_{jk}) C_k - \alpha \sum_{k=1}^{N} \Phi_{jk} C_k(t) \sum_{k=1}^{N} \nabla(\Phi_{jk}) C_k + f(X_j, t), \quad j = 1, 2, \dots, N_{int}. \tag{8}$$

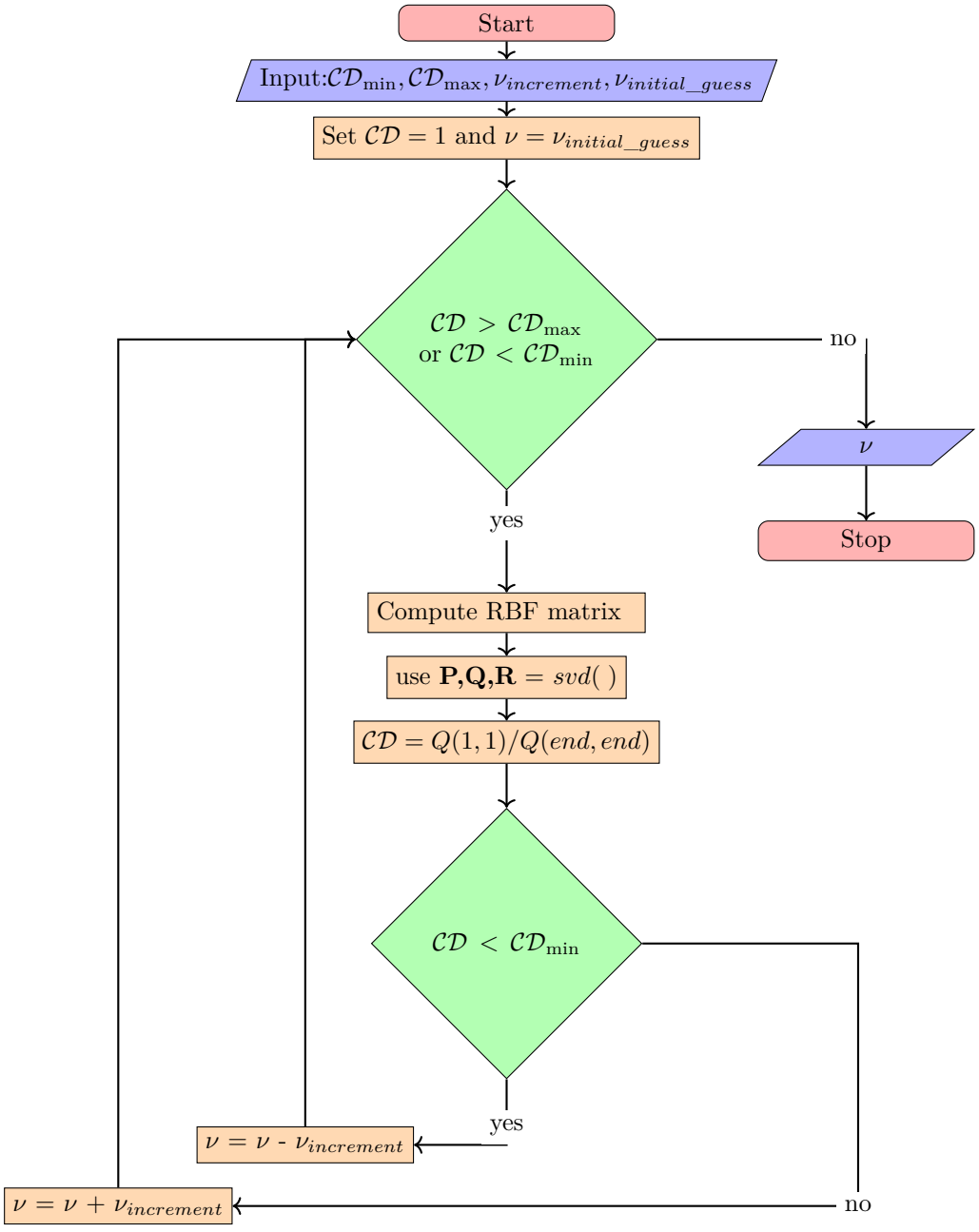


Figure 2: Flowchart of pseudo code for optimal shape parameter

The boundary conditions can be imposed as

$$\mathcal{B}\sum_{k=1}^{N}\Phi_{mk}C_k = \phi(r_m,t), \quad m=1,2,\ldots,N_b. \tag{9}$$

In matrix form, (8) and (9) can be written as

$$\frac{\partial}{\partial t} = \frac{1}{Re}(\nabla^2\ ) \ -\alpha(\ \ )(\nabla(\ )\ ) + f(X_j,t). \tag{10}$$

By replacing appropriate rows corresponding to the indexing of boundary points in (10), imposed boundary conditions and setting the equations in differential form the 4th order RK method is applied.

To control the step-size (δt), the algorithm of RK method of order 4 is presented with flowchart. Hence, present study follows 4th order Runge–Kutta method given as

$$\begin{aligned}
\mathcal{S}_1 =& \delta t\mathcal{F}(t_p,\vartheta_j^p),\\
\mathcal{S}_2 =& \delta t\mathcal{F}(t_p+\frac{\delta t}{2},\vartheta_j^p+\frac{\mathcal{S}_1}{2}),\\
\mathcal{S}_3 =& \delta t\mathcal{F}(t_p+\frac{\delta t}{2},\vartheta_j^p+\frac{\mathcal{S}_2}{2}),\\
\mathcal{S}_4 =& \delta t\mathcal{F}(t_p+\delta t,\vartheta_j^p+\mathcal{S}_3),\\
\vartheta_j^{p+1} =& \vartheta_j^p+\frac{\delta t}{6}(\mathcal{S}_1+2\mathcal{S}_2+2\mathcal{S}_3+\mathcal{S}_4).
\end{aligned} \tag{11}$$

The algorithm and flowchart of RBF with 4th order RK method is given in Algorithm 2 and Figure 3, respectively.

Algorithm 2 Algorithm for 4th order Runge–Kutta method

Input: $t_0, t_f, \delta t$
% t_0 = initial time, t_f = finial time, δt = step size,
$T = t_0 : \delta t : t_f$
for j = 1:length of T **do**

$$\begin{aligned}
\mathcal{S}_1 =& \delta t\mathcal{F}(t_p,\vartheta_j^p),\\
\mathcal{S}_2 =& \delta t\mathcal{F}(t_p+\frac{\delta t}{2},\vartheta_j^p+\frac{\mathcal{S}_1}{2}),\\
\mathcal{S}_3 =& \delta t\mathcal{F}(t_p+\frac{\delta t}{2},\vartheta_j^p+\frac{\mathcal{S}_2}{2}),\\
\mathcal{S}_4 =& \delta t\mathcal{F}(t_p+\delta t,\vartheta_j^p+\mathcal{S}_3),\\
\vartheta_j^{p+1} =& \vartheta_j^p+\frac{\delta t}{6}(\mathcal{S}_1+2\mathcal{S}_2+2\mathcal{S}_3+\mathcal{S}_4)
\end{aligned}$$

end for
Output: ϑ_j^{p+1}

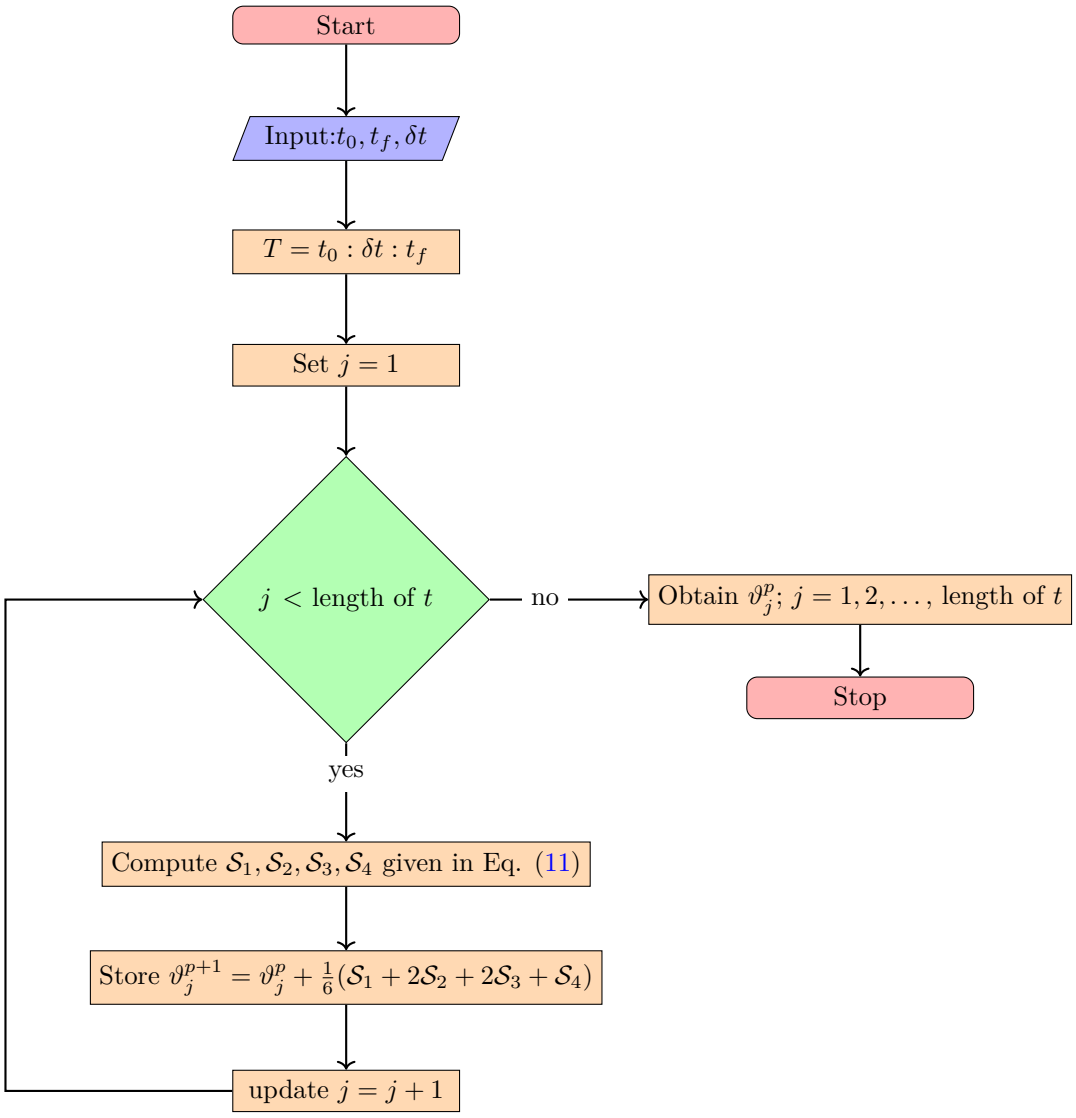


Figure 3: Flowchart of pseudo code for 4th order RK method

5 Generalized bivariate homotopy perturbation method

The analytical approach produces an exact solution that reflects the qualitative behavior of the model problem. Inefficiency of existing analytic approaches for complex problems shifts the preference to numerical approaches which deliver approximate quantitative data for selected points. The middle course of both approaches, as mentioned before, is a semi-analytic approach. It is based on the methodology of the expansion method and assumes results in the form of a series and evaluates the subsequent component of the series with the prior one. The hybrid approach of GB-HPM, inherits the merits of both analytic and approximation methodologies. It yields the exact solution through the closed form of the approximate series solution. Some of the primary results of the GB transform are as follows:

Definition 1. [2] The GB transform of function $\vartheta(X,t)$ with respect to t is given by

$$\mathcal{G}_m\left(\vartheta(X,t)\right) = \mathcal{B}_m(X,s,\gamma) = \frac{s}{\gamma^m}\int_0^\infty t^{m-1}\exp\left(-\frac{s}{\gamma}t\right)\vartheta(X,t)\,dt, \tag{12}$$

where m is the order and s and γ are transformation variable.

Definition 2. [2] If $\mathcal{B}_m(s,\gamma)$ is the GB transform of function $\vartheta(X,t)$, then the inverse GB transform of $\mathcal{B}_m(X,s,\gamma)$ is $\vartheta(X,t)$ such that

$$\mathcal{G}_m^{-1}\left(\mathcal{B}_m(X,s,\gamma)\right) = \vartheta(X,t). \tag{13}$$

Theorem 1. [2] Let $\vartheta(X,t), \vartheta'(X,t), \ldots, \vartheta^{m-1}(X,t)$ be continuous and of exponential order on $[0,\infty)$ while $\vartheta^n(X,t)$ is piece-wise continuous on $[0,\infty)$ with respect to t. Then

$$\begin{aligned}\mathcal{G}_m[\vartheta^n(X,t)] =& (-1)^{m-1} \\ &\times \frac{s}{\gamma}\frac{\partial^{m-1}}{\partial s^{m-1}}\left[\left(\frac{s}{\gamma}\right)^{n-1}\mathcal{G}_1[\vartheta(X,t)] - \sum_{k=0}^{n-1}\left(\frac{s}{\gamma}\right)^{n-(k+1)}\vartheta^k(X,0)\right].\end{aligned}$$

The pseudo code of the GB-HPM is defined in Algorithm given in Figure 4. Applying GB transform of order one on (1) yields

$$\mathcal{G}_1[\vartheta(X,t)] = \vartheta(X,0)\frac{\gamma}{s}\mathcal{G}_1\left[\alpha\vartheta\nabla\vartheta - \frac{1}{Re}\nabla^2\vartheta\right]. \tag{14}$$

Operating inverse of GB transform on (14) gives

$$\vartheta(X,t) = \vartheta(X,0) - \mathcal{G}_1^{-1}\left\{\frac{\gamma}{s}\mathcal{G}_1\left[\alpha\vartheta\nabla\vartheta - \frac{1}{Re}\nabla^2\vartheta\right]\right\}. \tag{15}$$

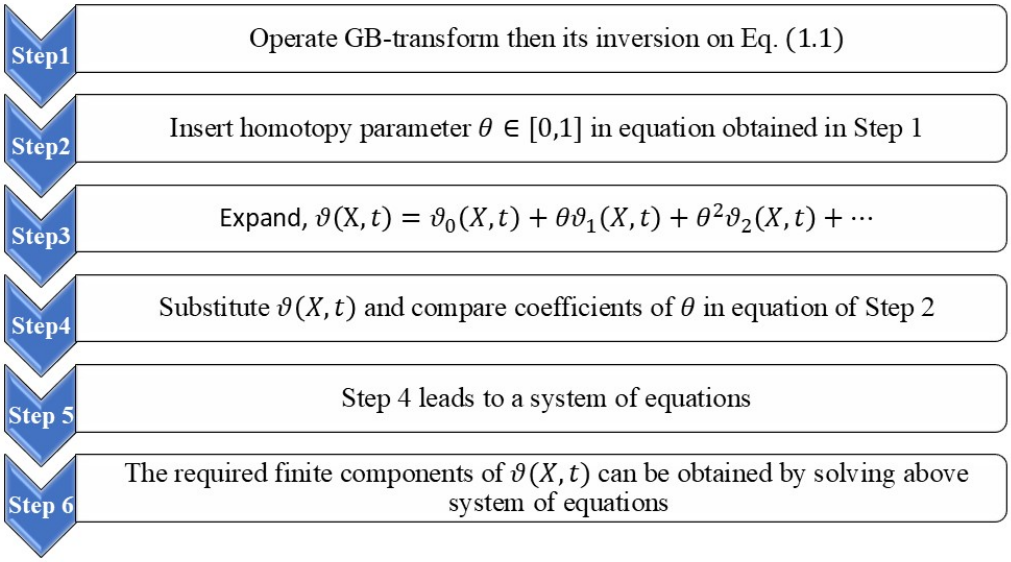


Figure 4: Algorithm for GB-HPM.

To form homotopic equation, insert homotopy parameter $\theta \in [0,1]$ (15) as follows:

$$\vartheta(X,t) = \vartheta(X,0) - \theta\mathcal{G}_1^{-1}\left\{\frac{\gamma}{s}\mathcal{G}_1\left[\alpha\vartheta\nabla\vartheta - \frac{1}{Re}\nabla^2\vartheta\right]\right\}. \tag{16}$$

Assume that the solution of (16) can be expanded as

$$\vartheta(X,t) = \vartheta_0(X,t) + \theta\vartheta_1(X,t) + \theta^2\vartheta_2(X,t) + \cdots . \tag{17}$$

Substituting (17) in (16) and comparing the coefficients of θ^n (for $n = 0,1,2,\ldots$) result the following system of equations:

$$\begin{aligned}
&\vartheta_0(X,t) = \vartheta(X,0),\\
&\vartheta_1(X,t) = \mathcal{G}_1^{-1}\left\{\frac{\gamma}{s}\mathcal{G}_1\left[\alpha\vartheta_0\nabla\vartheta_0 - \frac{1}{Re}\nabla^2\vartheta_0\right]\right\},\\
&\vartheta_2(X,t) = \mathcal{G}_1^{-1}\left\{\frac{\gamma}{s}\mathcal{G}_1\left[\alpha(\vartheta_0\nabla\vartheta_1 + \vartheta_1\nabla\vartheta_0) - \frac{1}{Re}\nabla^2\vartheta_1\right]\right\},
\end{aligned} \tag{18}$$

$$\vdots$$

$$\vartheta_n(X,t) = \mathcal{G}_1^{-1}\left\{\frac{\gamma}{s}\mathcal{G}_1\left[\sum_{k=0}^{n}\alpha\vartheta_k\nabla\vartheta_{n-k-1} - \frac{1}{Re}\nabla^2\vartheta_{n-1}\right]\right\}.$$

Computing the component of above system of equations yields the required series solution (17). For more details of GB-HPM, readers are advised to read [2, 3]. The semi-analytic solutions of different problems under different conditions are presented in the succeeding section.

6 Implementation of numerical and analytic techniques

The implementation of any technique whether it is analytic or numerical is a vital part of the study to ensure its application and the verification under different conditions. In this section, the numerical tests for proposed semi-analytic and numerical procedures are presented. The comparison of numerical and analytic techniques is presented in terms of error analysis defined below:

$$\begin{aligned} Abs &= |\vartheta_j^p - \tilde{\vartheta}_j^p|,\\ \|e\|_\infty &= \max_j |\vartheta_j^p - \tilde{\vartheta}_j^p|,\\ \|e\|_2 &= \sqrt{\frac{1}{N}\sum_{j=1}^{N}\left(\vartheta_j^p - \tilde{\vartheta}_j^p\right)^2}, \end{aligned}$$

where ϑ_j^p and $\tilde{\vartheta}_j^p$ are exact and approximate solutions, respectively, computed at $X_j \in \Omega$ and $t_p \in [0,T]$.

Example 1. Consider the unsteady viscous 2D Burger's equation:

$$\frac{1}{Re}(\vartheta_{\xi\xi} + \vartheta_{\zeta\zeta}) = \vartheta_t + \alpha\vartheta(\vartheta_\xi + \vartheta_\zeta), \qquad 0 < \xi, \zeta < 1, t > 0. \tag{19}$$

For the specific case $\alpha = 1$, the analytic solution is mentioned in [31, 38]as follows:

$$\vartheta(\xi,\zeta,t) = \frac{1}{\exp\left(\frac{Re(\xi+\zeta-t)}{2}\right)}, \tag{20}$$

The initial and boundary conditions are obtained from (20).

According to the description of the of GB-HPM given in Section 2.1, the component of semi-analytic series solutions are obtained as

$$
\begin{aligned}
\vartheta_0(\xi,\zeta,t) &= 1/(\exp((Re(\xi+\zeta))/2)+1),\\
\vartheta_1(\xi,\zeta,t) &= Ret\exp((Re(\xi+\zeta))/2)\Big(2\alpha+\frac{\exp((Re(\xi+\zeta)/2)-1))}{2(\exp((Re(\xi+\zeta))/2)+1)^3}\Big),\\
\vartheta_2(\xi,\zeta,t) &= (Re^2t^2\exp((Re(\xi+\zeta))/2)(4\alpha-11\exp(Re(\xi+\zeta))\\
&\quad+11\exp((Re(\xi+\zeta))/2)+\exp((3Re(\xi+\zeta))/2)\\
&\quad+12\alpha\exp(Re(\xi+\zeta))-24\alpha\exp((Re(\xi+\zeta))/2)\\
&\quad+12\alpha^2\frac{\exp((Re(\xi+\zeta))/2)-4\alpha^2-1))}{(8(\exp((Re(\xi+\zeta))/2)+1)^5)},\\
&\vdots
\end{aligned}
$$

Furthermore, for $n \geq 2$, desired number of terms of series solution can be obtained to enhance its accuracy. The numerical values listed in Table 2 are calculated at various time levels t for different numbers of collocation points N and step size δt. Table 3 demonstrates better accuracy as value of N increases for high Reynolds numbers that have been not considered yet and it displays superior outcomes than available in literature for $Re = 100$. It is observed from these tables that even Reynolds number increases, the error norms remain stable and does not fluctuate or diverge, which shows the applicability of technique on the proposed problem.

Table 2: Error norms of RBF-RK solutions of Example 1 for $\alpha = 1$ and $Re = 10$ at various time levels.

			$t = 0.05$		$t = 0.25$		$t = 0.5$	
Method	N	δt	$\|e\|_\infty$	$\|e\|_2$	$\|e\|_\infty$	$\|e\|_2$	$\|e\|_\infty$	$\|e\|_2$
RBF-RK	100	0.005	2.2342e-05	6.4207e-06	1.5028e-04	3.0529e-05	1.3028e-04	3.9193e-05
	400	0.001	3.2139e-06	4.0586e-07	1.3798e-05	1.5346e-06	9.3718e-06	3.0963e-06
FIFD [33]	900	0.005	9.3799e-05	1.6417e-05	2.3733e-04	5.2981e-05	-	-
MTCB-DQM [1]	3600	0.0001	6.9200e-05	1.1600e-05	5.3000e-05	1.5100e-05	5.2600e-05	1.8800e-05

Table 3: Error norms of RBF-RK solutions of Example 1 at $\alpha = 1$ and $t = 0.05$ with $\delta t = 0.001$ for various Re.

		$Re = 100$		$Re = 200$	
Method	N	$\|e\|_\infty$	$\|e\|_2$	$\|e\|_\infty$	$\|e\|_2$
RBF-RK	25	1.2067e-07	2.8168e-08	7.5064e-08	1.7538e-08
	36	5.2821e-08	2.0907e-08	1.7069e-08	1.1214e-08
BBFE [32]	25	3.2143e-06	3.4403e-06	-	-
	441	2.0547e-06	2.3549e-06	-	-

		$Re = 500$		$Re = 700$	
		$\|e\|_\infty$	$\|e\|_2$	$\|e\|_\infty$	$\|e\|_2$
RBF-RK	25	3.3011e-08	7.7190e-09	2.3994e-08	5.6116e-09
	36	7.4839e-09	4.9135e-09	5.4365e-09	3.5689e-09

In Tables 4 and 5, numerical values are tabulated by considering initial five terms of GB-HPM series solution. Semi-analytic GB-HPM is distinguished at small time level $t = 0.01$. However,

it faces challenges as time marched to $t = 0.5$. The numerical RBF-RK with $N = 400$ and $\delta t = 0.001$ performs quite well where absolute error ranges from 10^{-5} to 10^{-8}.

Table 4: Numerical results and absolute errors of solutions of Example 1 at $\alpha = 1$ and $Re = 10$.

t	ξ	ζ	GB-HPM	RBF-RK	Abs$_{GB-HPM}$	Abs$_{RBF-RK}$
0.01	0.1	0.1	2.78884e-01	2.7888e-01	1.26684e-10	4.0228e-07
	0.5	0.1	4.97365e-02	4.9733e-02	1.38650e-11	3.1659e-06
	0.9	0.1	7.03358e-03	7.0314e-03	1.40396e-11	2.2335e-06
	0.3	0.3	4.97365e-02	4.9734e-02	1.38650e-11	2.5540e-06
	0.7	0.3	7.03358e-03	7.0303e-03	1.40396e-11	3.3208e-06
	0.1	0.5	4.97365e-02	4.9735e-02	1.38650e-11	1.0479e-06
	0.5	0.5	7.03358e-03	7.0318e-03	1.40396e-11	1.7792e-06
	0.9	0.5	9.57717e-04	9.5640e-04	2.32464e-12	1.3206e-06
	0.1	0.9	7.03358e-03	7.0334e-03	1.40396e-11	2.0050e-07
	0.9	0.9	1.29720e-04	1.2978e-04	3.22787e-13	5.6619e-08
0.5	0.1	0.1	7.51387e-01	8.1757e-01	6.61867e-02	2.7296e-07
	0.5	0.1	4.04386e-01	3.7755e-01	2.68453e-02	5.0835e-06
	0.9	0.1	7.02560e-02	7.5855e-02	5.60212e-03	2.8064e-06
	0.3	0.3	4.04386e-01	3.7754e-01	2.68453e-02	2.5281e-06
	0.7	0.3	7.02560e-02	7.5860e-02	5.60212e-03	2.3182e-06
	0.1	0.5	4.04386e-01	3.7754e-01	2.68453e-02	9.5033e-08
	0.5	0.5	7.02560e-02	7.5862e-02	5.60212e-03	4.0642e-06
	0.9	0.5	9.84591e-03	1.0989e-02	1.14103e-03	2.3675e-06
	0.1	0.9	7.02560e-02	7.5857e-02	5.60212e-03	8.8786e-07
	0.9	0.9	1.33883e-03	1.5056e-03	1.62345e-04	4.3955e-06

Table 5: Numerical results of Example 1 at $t = 0.01$ and $Re = 10$ for different α.

		$\alpha = 0$		$\alpha = 2$	
ξ	ζ	GB-HPM	RBF-RK	GB-HPM	RBF-RK
0.1	0.1	2.7334E-01	2.7347e-01	2.8474e-01	2.8453e-01
0.5	0.1	4.9493E-02	4.9491e-02	4.9984e-02	4.9985e-02
0.9	0.1	7.0284E-03	7.0188e-03	7.0388e-03	7.0440e-03
0.3	0.3	4.9493E-02	4.9484e-02	4.9984e-02	4.9993e-02
0.7	0.3	7.0284E-03	7.0359e-03	7.0388e-03	7.0305e-03
0.1	0.5	4.9493E-02	4.9492e-02	4.9984e-03	4.9986e-02
0.5	0.5	7.0284E-03	7.0276e-03	7.0388e-03	7.0413e-03
0.9	0.5	9.5762E-04	9.5081e-04	9.5781e-04	9.6304e-04
0.1	0.9	7.0284E-03	7.0205e-03	7.0388e-03	7.0454e-03
0.9	0.9	1.2972E-04	1.3110e-04	1.2972e-04	1.2936e-04

The 3D behavior of numerical values and absolute error is shown in Figures 5 and 6, respectively for $Re = 1$ with $N = 441$. It is observed that numerical values lie between -0.01 to 0.01. The absolute error is found to be of order 10^{-6}, which again implies the applicability of the proposed technique.

Figures 7 and 8 show the behavior of the numerical solution and the absolute error, respectively, for $N = 400$ and $\delta t = 0.001$ with $Re = 10$ and $\alpha = 1$ at $t = 1$. Maximum error is observed

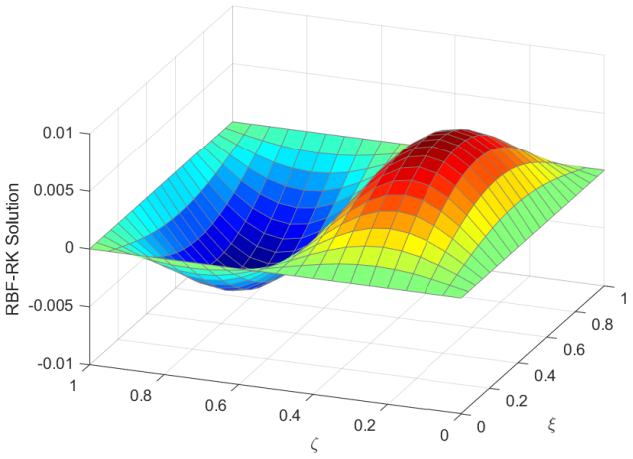


Figure 5: Visualization of numerical solution for Example 1 when $Re = 1$, $\alpha = 1$, $t = 0.1$, $\delta t = 0.001$ and $N = 441$.

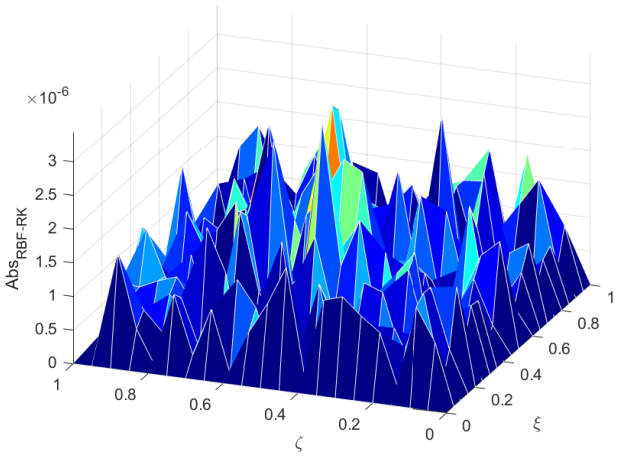


Figure 6: Visualization of absolute error for Example 1 when $Re = 1$, $\alpha = 1$, $t = 0.1$, $\delta t = 0.001$ and $N = 441$.

of order 10^{-5}, which verifies the application of proposed technique. Figures 9–11 imply the active intervention of advection coefficient α in controlling direction of flow. For $\alpha = 0$, the solution profile exhibits multi-directional flow. For $\alpha = 1$, the fluid velocity is directed along single direction, however, for $\alpha = 2$, the fluid flow seems to gradually narrow and concentrate near the origin.

Example 2. Consider the unsteady viscous 2D Burger's equation (19) with initial and boundary data given as follows:

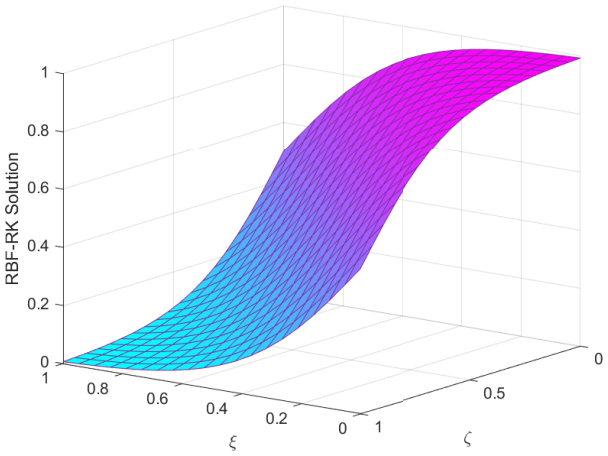


Figure 7: Visualization of numerical solution for Example 1 when $Re = 10$, $\alpha = 1$, $t = 1$, $\delta t = 0.001$ and $N = 400$.

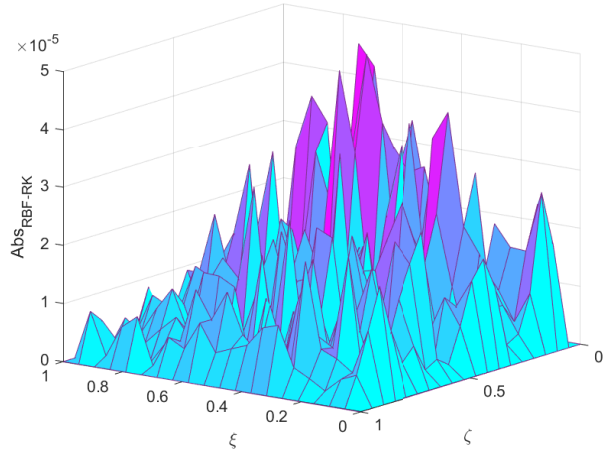


Figure 8: Visualization of absolute error for Example 1 when $Re = 10$, $\alpha = 1$, $t = 1$, $\delta t = 0.001$ and $N = 400$.

$$\vartheta(\xi, \zeta, 0) = \sin(\pi\xi)\sin(2\pi\zeta), \tag{21}$$

where $\Omega = [0, 1] \times [0, 1]$ is the space domain. For the specific case $Re = 1$ and $\alpha = 0$, the analytic solution is mentioned in [32, 38] as follows:

$$\vartheta(\xi, \zeta, t) = \exp(-5t\pi^2)\sin(\pi\xi)\sin(2\pi\zeta). \tag{22}$$

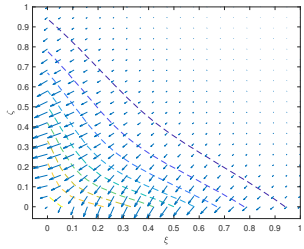


Figure 9: RBF-RK contour profiles for Example 1 when $Re = 1$, $t = 0.5$, $\delta t = 0.001$, $\alpha = 0$ and $N = 400$.

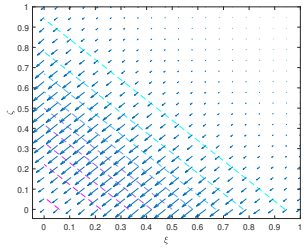


Figure 10: RBF-RK contour profiles for Example 1 when $Re = 1$, $t = 0.5$, $\delta t = 0.001$, $\alpha = 1$ and $N = 400$.

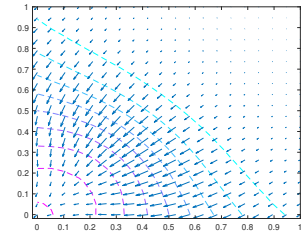


Figure 11: RBF-RK contour profiles for Example 1 when $Re = 1$, $t = 0.5$, $\delta t = 0.001$, $\alpha = 2$ and $N = 400$.

Using initial data given in (21) as initial component in procedure of GB-HPM yields further components of series solution as

$$\begin{aligned}
\vartheta_0(\xi,\zeta,t) =& \sin(\pi\xi)\sin(2\pi\zeta),\\
\vartheta_1(\xi,\zeta,t) =& -(t\sin(\pi\xi)\sin(2\pi\zeta)(5\pi^2 + Re\alpha\pi\cos(\pi\xi)\sin(2\pi\zeta)\\
&+ 2Re\alpha\pi\cos(2\pi\zeta)\sin(\pi\xi)))/Re,\\
\vartheta_2(\xi,\zeta,t) =& (t^2(25\pi^4\sin(\pi\xi)\sin(2\pi\zeta) - Re^2\alpha^2\pi^2\sin(\pi\xi)\sin(2\pi\zeta)^3\\
&- 4Re^2\alpha^2\pi^2\sin(\pi\xi)^3\sin(2\pi\zeta)\\
&- 8Re\alpha\pi^3\cos(\pi\xi)\sin(\pi\xi)(2\cos(2\pi\zeta)^2 - 1)\\
&- 4Re\alpha\pi^3\cos(2\pi\zeta)\sin(2\pi\zeta)(2\cos(\pi x)^2 - 1)
\end{aligned}$$

$$
\begin{aligned}
&+ 14Re\alpha\pi^3 \cos(\pi\xi)\sin(\pi x)\sin(2\pi\zeta)^2 \\
&+ 52Re\alpha\pi^3 \cos(2\pi\zeta)\sin(\pi\xi)^2\sin(2\pi\zeta) \\
&+ 3Re^2\alpha^2\pi^2 \cos(\pi\xi)^2\sin(\pi\xi)\sin(2\pi\zeta)^3 \\
&+ 12Re^2\alpha^2\pi^2 \cos(2\pi\zeta)^2\sin(\pi\xi)^3\sin(2\pi\zeta) \\
&+ 12Re^2\alpha^2\pi^2 \cos(\pi\xi)\cos(2\pi\zeta)\sin(\pi\xi)^2\sin(2\pi\zeta)^2))/(2Re^2), \\
&\vdots
\end{aligned}
$$

Following the same process for $n \geq 2$, the desired number of terms of series solution can be obtained. In the present case, numerical values are computed from the initial five terms of GB-HPM series solution.

For numerical procedures, with the initial condition of a sine waveform, the boundary conditions are set to be periodic as given in (21) and (22). In Table 6, tabulated error norms corresponding to the available exact solution for $\alpha = 0$ and $Re = 1$ are computed by employing different values of N and δt across space-time domain. Accuracy level enhances as step size δt becomes smaller. In Table 7, comparison analysis in terms of absolute error demonstrates the excellence of semi-analytic GB-HPM for very small time $t = 0.001$ and $t = 0.02$. In Table 8, the numerical results are presented for different values of Re and $\alpha = 1$. The agreement between GB-HPM and RBF-RK method signifies the application of proposed techniques on a given 2D Burger's equation.

Table 6: Error norms of RBF-RK solutions of Example 2 for $\alpha = 0$ and $Re = 1$ at various time levels.

			$t = 0.01$		$t = 0.05$		$t = 0.1$	
Method	N	δt	$\|e\|_\infty$	$\|e\|_2$	$\|e\|_\infty$	$\|e\|_2$	$\|e\|_\infty$	$\|e\|_2$
RBF-RK	100	0.001	3.2465e-05	9.9245e-06	1.1422e-05	4.2433e-06	2.6308e-06	8.3001e-07
		0.0001	3.2455e-05	9.9211e-06	1.1415e-05	4.2390e-06	2.6293e-06	8.2940e-07
	441	0.001	1.4088e-06	4.1234e-07	3.1881e-06	8.8184e-07	3.4405e-06	9.4979e-07
		0.0001	1.4088e-06	4.1399e-07	3.1908e-06	8.8215e-07	3.4409e-06	9.4979e-07

However, GB-HPM experiences challenges for $t \geq 0.02$. Therefore, for the aim of small time propagation of physical phenomena such as simulation of earthquake, tsunamis and cosmic dynamics [12, 41, 44], GB-HPM can be an ideal choice. Indeed for further analysis one needs to employ suggested numerical procedures which maintain the standard of accuracy for comparatively large time. In the scenario of lacking an exact solution, mutual comparison of proposed approximate solutions verifies the credibility of the obtained numerical results by the proposed schemes.

Table 7: Numerical results and absolute errors of solutions of Example 2 for $\alpha = 0$ and $Re = 1$.

t	ξ	ζ	GB-HPM	RBF-RK	Abs_{GB-HPM}	Abs_{RBF-RK}
0.001	0.1	0.1	1.7289e-01	1.7290e-01	4.3935e-10	6.8636e-06
	0.5	0.1	5.5948e01	5.5949e-01	1.4218e-09	3.5509e-06
	0.9	0.1	1.7289e-01	1.7289e-01	4.3935e-10	1.0765e-06
	0.3	0.3	7.3237e-01	7.3238e-01	1.8611e-09	6.4155e-06
	0.1	0.9	7.3237e-01	7.3237e-01	1.8611e-09	1.5544e-08
	0.7	0.3	-1.7289e-01	-1.7288e-01	4.3935e-10	6.6740e-06
	0.9	0.9	-1.7289e-09	-1.7289e-01	4.3935e-10	3.3662e-06
0.02	0.1	0.1	6.8911e-02	6.7699e-02	1.2135e-03	2.0509e-06
	0.5	0.1	2.2300e-01	2.1908e-01	3.9271e-03	2.8359e-06
	0.9	0.1	6.8911e-02	6.7699e-02	1.2135e-03	2.1765e-06
	0.3	0.3	2.9191e-01	2.8677e-01	5.1406e-03	2.8133e-06
	0.7	0.3	2.9191e-01	2.8677e-01	5.1406e-03	2.8780e-06
	0.1	0.9	-6.8911e-02	-6.7696e-02	1.2135e-03	1.0952e-06
	0.9	0.9	-6.8911e-02	-6.7696e-02	1.2135e-03	9.0260e-07

Table 8: Numerical results of Example 2 at $\alpha = 1$ and $t = 0.001$ for different Re.

		$Re = 10$		$Re = 100$	
ξ	ζ	GB-HPM	RBF-RK	GB-HPM	RBF-RK
0.1	0.1	1.8014e-01	1.8014e-01	1.8094e-01	1.8095e-01
0.5	0.1	5.8195e-01	5.8196e-01	5.8452e-01	5.8453e-01
0.9	0.1	1.8077e-01	1.8078e-01	1.8158e-01	1.8158e-01
0.3	0.3	7.6548e-01	7.6548e-01	7.6889e-01	7.6889e-01
0.7	0.3	7.6816e-01	7.6816e-01	7.7160e-01	7.7160e-01
0.1	0.9	-1.8077e-01	-1.8078e-01	-1.8158e-01	-1.8158e-01
0.9	0.9	-1.8014e-01	-1.8014e-01	-1.8094e-01	-1.8094e-01

In Figures 12 to 15, RBF-RK solution profiles are shown for $N = 441$ and $\delta t = 0.001$, with the Reynolds number Re varying from 1 to 50. It displays the configuration of sinusoidal waveform. The sinusoidal waveform is generally found in sine curved channels, arteries, alternating current and many more.

7 Conclusion

The present study made an original contribution by conducting a comparative examination of the approximate solutions of two schema classes, namely semi-analytic GB-HPM and numerical collocation methods based on RBFs. Few initial terms of semi-analytic series show adequate convergence to the exact solution and radial basis collocation methods prove their competency along with their complementary meshfree nature. The potential of proposed schemes are compared using absolute, $\|e\|_\infty$ and $\|e\|_2$ error norms for short, moderate and long-time propagation.

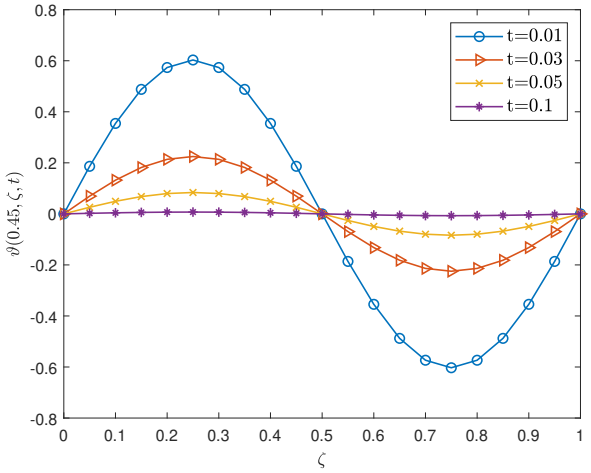


Figure 12: RBF-RK solution profiles of Example 2 when $Re = 1, \alpha = 0$, $\delta t = 0.001$ and $N = 441$.

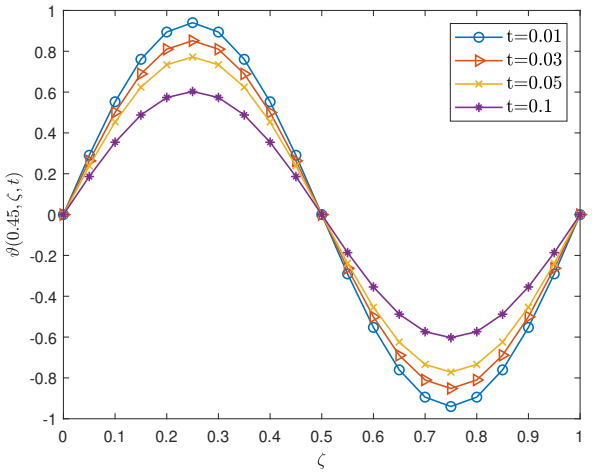


Figure 13: RBF-RK solution profiles of Example 2 when $Re = 10, \alpha = 0$, $\delta t = 0.001$ and $N = 441$.

The GB-HPM exhibits promising accuracy for short and moderate-time propagation but fails to maintain its standard for long-time computations. However at cost of some more computational efforts suggested numerical scheme has performed pretty well at all time levels.

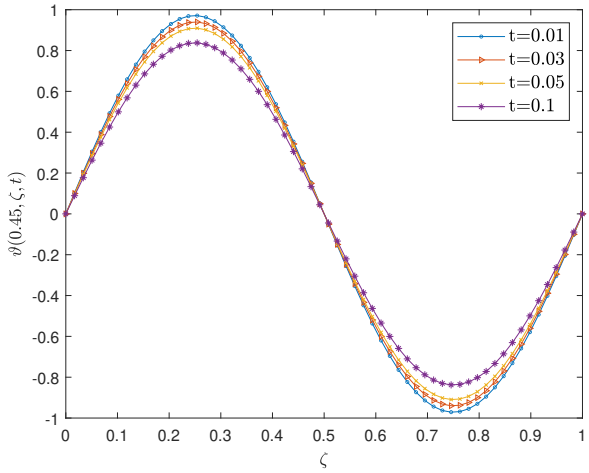


Figure 14: RBF-RK solution profiles of Example 2 when $Re = 30, \alpha = 0$, $\delta t = 0.001$ and $N = 441$.

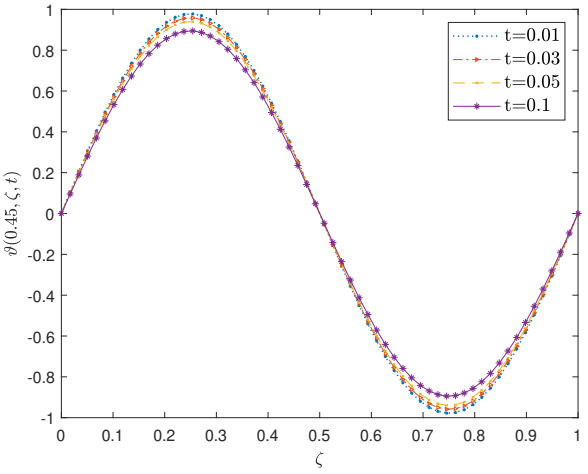


Figure 15: RBF-RK solution profiles of Example 2 when $Re = 50, \alpha = 0$, $\delta t = 0.001$ and $N = 441$.

Funding

Dr. Shelly Arora is thankful to DST-ANRF for providing grant via SPG/2022/001269 and Ms. Amandeep Kaur is thankful to CSIR for providing Junior Research Fellowship F. No. 09/0140(17151)/2023-EMR-I.

Conflicts of Interest

All authors jointly worked on the results and they read and approved the final manuscript and also declare that they have no competing interests.

References

[1] Arora, G., and Joshi, V. *A computational approach using modified trigonometric cubic B-spline for numerical solution of Burger's equation in one and two dimensions*, Alex. Eng. J. 57(2) (2018), 1087–1098.

[2] Arora, S., Dhaliwal, S.S. Ma, W.X., and Pasrija, A. *Analysis of fractional order Schrodinger equation with singular and nonsingular kernel derivatives via novel hybrid scheme*, J. Appl. Anal. Comput. 15(2) (2025), 1039–1067.

[3] Arora, S., and Pasrija, A. *A novel integral transform operator and its applications,* Iran. J. Numer. Anal. Optim. 13(3) (2023), 553–575.

[4] Babu, A., Han, B., and Asharaf, N., *Numerical solution of the viscous Burger's equation using localized differential quadrature method*, Partial Differ. Equ. Appl. Math. 4 (2021), 100044.

[5] Bahadır, A.R. *A fully implicit finite-difference scheme for two-dimensional Burger's equations*, Appl. Math. Comput. 137(1) (2003), 131–137.

[6] Bateman, H. *Some recent researches on the motion of fluids*, Mon. Weather Rev. 43(4) (1915), 163–170.

[7] Biazar, J., and Hosami, M. *An interval for the shape parameter in radial basis function approximation*, Appl. Math. Comput. 315 (2017), 131–149.

[8] Bilige, S., and Chaolu, T. *An extended simplest equation method and its application to several forms of the fifth-order KdV equation*, Appl. Math. Comput. 216(11) (2010), 3146–3153.

[9] Bonkile, M.P., Awasthi, A., Lakshmi, C., Mukundan, V., and Aswin, V.S. *A systematic literature review of Burger's equation with recent advances*, Pramana, 90 (2018), 1-21.

[10] Breuer, H.P., and Petruccione, F. *Burger's model of turbulence as a stochastic process*, J. Phys. A. 25(11) (1992), L661.

[11] Burger, M., Göttlich, S. and Jung,T. *Derivation of second order traffic flow models with time delays*, Netw. Heterog. Media. 14(2) (2019), 265–288.

[12] Carcione, J.M., Poletto, F., Farina, B., and Craglietto, A. *Simulation of seismic waves at the Earth's crust (brittle–ductile transition) based on the Burger's model*, Solid Earth, 5(2) (2014), 1001–1010.

[13] Carlson, R.E., and Foley, T.A. *The parameter R2 in multiquadric interpolation*, Comput. Math. Appl. 21(9) (1991), 29–42.

[14] Fasshauer, G.E. *Newton iteration with multiquadrics for the solution of nonlinear PDEs*, Comput. Math. Appl. 43(3-5) (2002), 423–438.

[15] Fasshauer, G.E., and Zhang, J.G. *On choosing "optimal" shape parameters for RBF approximation*, Numer. Algorithms, 45 (2007), 345–368.

[16] Franke, R., *Scattered data interpolation: tests of some methods*, Math. Comput. 38:157 (1982), 181–200.

[17] Gardner, C.S., Greene, J.M., Kruskal, M.D., and Miura, R.M. *Method for solving the Korteweg-deVries equation*, Phys. Rev. Lett. 19(19) (1967), 1095.

[18] Hardy, R.L. *Theory and applications of the multiquadric-biharmonic method 20 years of discovery 1968–1988*, Comput. Math. Appl. 19(8-9) (1990), 163–208.

[19] Hirota, R., and Satsuma, J. *Nonlinear evolution equations generated from the Bäcklund transformation for the Boussinesq equation*, Prog. Theor. Phys. 57(3) (1977), 797–807.

[20] Jafari, H., Tajadodi, H., and Baleanu, D. *Application of a homogeneous balance method to exact solutions of nonlinear fractional evolution equations*, J. Comput. Nonlinear Dyn. 9(2) (2014), 021019.

[21] Jain, P.C., and Holla, D.N. *Numerical solutions of coupled Burger's equation*, Int. J. Non-Linear Mech. 13(4) (1978), 213–222.

[22] Jawad, A.J.A.M. *The sine-cosine function method for the exact solutions of nonlinear partial differential equations*, Sys. 2(3) (2012) 1.

[23] Kadkhoda, N., and Jafari, H., *Kudryashov method for exact solutions of isothermal magnetostatic atmospheres*, Iran. J. Numer. Anal. Optim. 6(1) (2016), 43–53.

[24] Kansa, E.J., *Multiquadrics—A scattered data approximation scheme with applications to computational fluid-dynamics—I surface approximations and partial derivative estimates*, Comput. Math. Appl. 19(8-9) (1990), 127–145.

[25] Kansa, E.J., and Holoborodko, P. *On the ill-conditioned nature of C^∞ RBF strong collocation*, Eng. Anal. Bound. Elem. 78 (2017), 26–30.

[26] Kumar, N., Verma, A.K., and Agarwal, R.P. *Two-Dimensional Uniform and Non-Uniform Haar Wavelet Collocation Approach for a Class of Nonlinear PDEs*, Comput. 11(10) (2023), 189.

[27] Li, M., Nikan, O., Qiu, W., and Xu, D. *An efficient localized meshless collocation method for the two-dimensional Burger's-type equation arising in fluid turbulent flows*, Eng. Anal. Bound. Elem. 144 (2022), 44–54.

[28] Liu, Z., and Yang, C. *The application of bifurcation method to a higher-order KdV equation*, J. Math. Anal. Appl. 275(1) (2002), 1–12.

[29] Lombard, B., and Matignon, D. *Diffusive approximation of a time-fractional Burger's equation in nonlinear acoustics*, SIAM J. Appl. Math. 76(5) (2016), 1765–1791.

[30] Ma, W.X., Huang, T., and Zhang, Y. *A multiple exp-function method for nonlinear differential equations and its application*, Phys. Scr. 82(6) (2010), 065003.

[31] Mittal, R.C., and Jiwari, R. *A differential quadrature method for numerical solutions of Burger's-type equations*, Int. J. Numer. Methods Heat Fluid Flow. 22(7) (2012), 880–895.

[32] Mittal, R.C., and Tripathi, A. *Numerical solutions of two-dimensional Burger's equations using modified Bi-cubic B-spline finite elements*, Eng. Comput. 32(5) (2015), 1275–1306.

[33] Mohamed, N.A. *Solving one-and two-dimensional unsteady Burger's equation using fully implicit finite difference schemes*, Arab J. Basic Appl. Sci. 26(1) (2019), 254–268.

[34] Mohanty, S.K., Kumar, S., Dev, A.N., Deka, M.K., Churikov, D.V., and Kravchenko, O.V. *An efficient technique of G'/ G–expansion method for modified KdV and Burger's equations with variable coefficients*, Result. Phys. 37 (2022), 105504.

[35] Mojtabi, A., and Deville, M.O. *One-dimensional linear advection–diffusion equation: Analytical and finite element solutions*, Comput. Fluids. 107 (2015), 189–195.

[36] Nofal, T.A. *Simple equation method for nonlinear partial differential equations and its applications*, Int. J. Egypt. Math. Soc. 24(2) (2016), 204–209.

[37] Ozer, S. and Kutluay, S. *An efficient and reliable numerical study for the 2D nonlinear coupled Burgers equation simulating many physical phenomena.*Int. J. Mod. Phys. C. 37(1) (2026), 2550051.

[38] Radwan, S.F. *Comparison of higher-order accurate schemes for solving the two-dimensional unsteady Burger's equation*, J. Comput. Appl. Math. 174(2) (2005), 383–397.

[39] Rippa, S., *An algorithm for selecting a good value for the parameter c in radial basis function interpolation*, Adv. Comput. Math., 11 (1999), 193–210.

[40] Sarra, S.A., and Kansa, E.J. *Multiquadric radial basis function approximation methods for the numerical solution of partial differential equations*, Adv. Comput. Mech. 2(2) (2009), 220.

[41] Satake, K. *Tsunamis*, In International geophysics, vol. 81, pp. 437-451. Academic press, 2002.

[42] Wacławczyk, M., and Oberlack, M. *Application of the extended Lie group analysis to the Hopf functional formulation of the Burger's equation*, J. Math. Phys. 54(7) (2013).

[43] Wazwaz, A.M. *Multiple-soliton solutions for the KP equation by Hirota's bilinear method and by the tanh-coth method*, Appl. Math. Comput. 190(1) (2007), 633–640.

[44] Webb, G.M. and McKenzie, J.F. *Burger's equation and cosmic ray shocks*, J. Plasma Phys. 31(3) (1984), 337–368.

[45] Xu, S., He, J., and Wang, L. *The Darboux transformation of the derivative nonlinear Schrödinger equation*, J. Phys. A: Math. Theor. 44(3)0 (2011), 305203.

Iranian Journal of Numerical Analysis and Optimization
Vol. 16, No. 3, 2026, pp 1133–1160
https://doi.org/10.22067/ijnao.2026.96880.1797
https://ijnao.um.ac.ir/

CC BY NC

Research Article

Quasilinearized Genocchi-Polynomial Collocation for Efficient Simulation of a Nonlinear SIRV Anthrax Transmission Model

N. Mohamed, M. El-Gamel and W. Adel*

Abstract

Mathematical models of anthrax transmission often involve strong nonlinearities that challenge standard solvers and can limit robust inference. We present a four-compartment (SIRV) animal anthrax model and an efficient numerical framework that combines quasilinearization with a shifted Genocchi-polynomial collocation scheme (Genocchi-QLM). We provide convergence and error analyses for the approximations on a finite interval and construct a residual-error diagnostic to certify solution quality. Across representative parameter sets, Genocchi-QLM achieves rapid convergence and close agreement with a benchmark Runge–Kutta implementation (`ode45`), while maintaining numerical stability over long horizons. We further

*Corresponding author

Received 7 December 2025; revised 21 April 2026; accepted 28 April 2026

Nesreen Mohamed
Department of Basic Science, Faculty of Engineering, Horus University, New Damietta 34517, Egypt. e-mail: nesreen.msedik@gmail.com

Mohamed El-Gamel
Department of Mathematics and Engineering Physics, Faculty of Engineering, Mansoura University, Mansoura 35516, Egypt. e-mail: gamel_eg@yahoo.com

Waleed Adel
Nanoelectronics Integrated Systems Center, Nile University, Giza, 12588, Egypt. Department of Mathematics and Engineering Physics, Faculty of Engineering, Mansoura University, Mansoura, 35516, Egypt. e-mail: waleed.ouf@ufe.edu.eg

How to cite this article
Mohamed, N., El-Gamel, M. and Adel, W., Quasilinearized Genocchi-Polynomial Collocation for Efficient Simulation of a Nonlinear SIRV Anthrax Transmission Model. *Iran. J. Numer. Anal. Optim.*, 2026; 16(3): 1133-1160. https://doi.org/10.22067/ijnao.2026.96880.1797

assess practical relevance by comparing simulated infected-case trajectories to reported outbreak patterns from endemic settings, observing qualitative agreement in rise-and-decay phases consistent with recovery, mortality, vaccination, and waning-immunity mechanisms. Sensitivity exploration highlights how vaccination rate and immunity waning jointly shape persistence. Because the method converts a nonlinear initial-value problem into a sequence of linear systems with a sparse structure, it is readily extensible to higher-dimensional or controlled models and can be adapted to fractional or stochastic variants. These results indicate that Genocchi-QLM is a fast, accurate, and scalable tool for solving nonlinear epidemiological systems and for supporting data-informed analysis of anthrax control strategies.

AMS subject classifications (2020): 92D30; 65L05; 65L07; 65M70; 42C10.

Keywords: Anthrax Disease Model; Boundary Value Problems; Genocchi Quasilinearization Method; Genocchi Polynomials; Numerical Methods.

1 Introduction

Bacillus anthracis is the bacterium that causes anthrax, an acute and often fatal zoonotic disease that has been recognized for centuries as a major threat to both animal and human health [25, 30]. Both domestic and wild animals are affected by the disease, with herbivorous species such as cattle, horses, sheep, goats, and antelopes being the most vulnerable due to their frequent contact with contaminated soil, water, or vegetation. These animals have very low natural resistance to anthrax, making it one of the most dangerous infectious diseases in the world [28, 36]. Bacillus anthracis bacterium is a spore-forming, rod-shaped organism that can survive in soil for decades under harsh environmental conditions; allowing repeated outbreaks in endemic regions. Infection in animals occurs in a variety of ways, such as by consuming water or grass contaminated with spores, inhaling spores while grazing, or coming into contact with carcasses of diseased animals [10].

Humans are also susceptible to anthrax and can contract the disease through direct or indirect contact with infected animals or their byproducts, including hides, wool, and meat [13]. Anthrax transmission occurs mainly through four well-recognized pathways [34]. Cutaneous anthrax enters the body through cuts or abrasions in the skin and is the most common form; although it is less deadly, untreated cases can still progress to systemic infection. Gastrointestinal anthrax develops after the ingestion of raw or undercooked meat from infected animals and is associated with severe abdominal pain, vomiting, and diarrhea, sometimes progressing rapidly to systemic illness. Pulmonary anthrax, also known as inhalation anthrax, results from breathing in

airborne spores and is the most dangerous form; even with treatment, it often has a high fatality rate due to rapid progression to severe respiratory distress and septicemia. Injection anthrax, the most recently identified form, occurs in people who inject contaminated drugs, leading to severe soft-tissue infections and sepsis.

Clinically, anthrax symptoms vary according to the route of infection but often share common features, including fever, fatigue, and rapid deterioration in severe cases. Cutaneous anthrax typically presents as a painless ulcer with a characteristic black eschar. Inhalation anthrax often begins with mild flu-like symptoms but quickly advances to respiratory failure if untreated. Gastrointestinal anthrax may cause severe hemorrhagic gastroenteritis, while injection anthrax is marked by extensive tissue damage and systemic toxicity. Globally, anthrax remains a significant public health and veterinary concern, particularly in Sub-Saharan Africa, parts of Asia, and some regions of North America where the bacteria are endemic in the soil. In recent decades, outbreaks in animal populations have been documented in several countries. In South Africa's Kruger National Park, for instance, the number of roan antelope plummeted from 450 to just 45 due to recurring anthrax outbreaks [11, 6]. In Tanzania, outbreaks have caused the deaths of 10 cattle, 26 goats, 3 sheep, and 131 wild animals, in addition to 21 humans suspected to the infection in a single season [22]. Human outbreaks have also been reported; in the Kilimanjaro region, 36 individuals were hospitalized after consuming anthrax-contaminated beef [26]. Tragically, in Arusha, eight of the 134 infected individuals succumbed to the disease, and in Dar es Salaam, six of the 22 infected patients died [26]. In Northern Canada, nine separate anthrax epidemics in bison populations were reported between 1962 and 1993, underscoring the persistent environmental reservoirs of the pathogen [7]. These examples highlight the enduring global presence of anthrax and the critical need for surveillance, rapid diagnosis, and effective vaccination programs to control its spread. Therefore, mathematical modeling plays a crucial role in predicting the spread of the disease and in understanding its complex dynamics, making it an essential tool for effective analysis and control strategies. The main modes of transmission of the anthrax in humans and animals can be seen in Figure 1.

Over the years, a number of mathematical models have been presented to explain how anthrax spreads among animals. Mushayabasa [24] created the first basic model of the dynamics of anthrax disease transmission, which included three main compartments: Pathogens, Contamination, and Susceptible. However, in this initial model, the role of infectious animals as a major contributor to the spread of anthrax was not addressed. Building on this foundation, Sabir and Simbawa [33] applied Bayesian regularization to develop a stochastic machine learning approach that demonstrated the numerical performance of the anthrax disease system in animal populations. In addition, Raza and Abdella [31] further examined the nonlinear dynamics of anthrax transmission by incorporating delay effects to account for the incubation period mech-

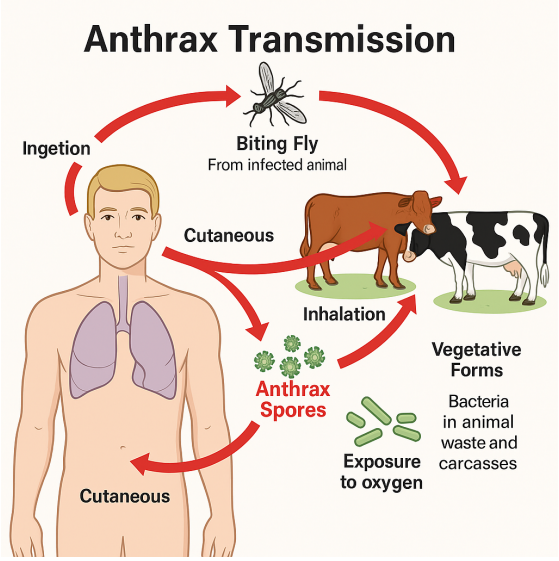


Figure 1: Modes of transmission of anthrax.

anism, providing deeper insights into the stability and progression of the disease. Also, Alfwzan et al. [2] introduced an unconventional computational methodology aimed at improving the accuracy of model predictions, evaluating potential control strategies, and identifying critical parameters that drive the dynamics of stochastic anthrax epidemics. Bylaiah et al. [3] utilized advanced machine learning techniques to create a predictive model for assessing anthrax risk, demonstrating how variations in environmental factors such as precipitation levels can influence disease outbreaks.

On the other hand, Helikumi, Lolika, and Mushayabasa [12] highlighted the significance of the basic reproduction number and its implications in simulating and understanding a wide range of infectious diseases, including anthrax. Mackey and Kribs [23] investigated the modeling of disease spread among zebra populations, providing valuable insights into the spatial and temporal patterns of anthrax transmission in wildlife. A proposition of a novel framework to evaluate the effectiveness of public health strategies for anthrax prevention, integrating decision-making processes related to health security have been proposed by Chen et al. [4]. Stella et al. [35] developed an SIR-based model to describe the transmission dynamics of anthrax within animal communities. Croicu [5] proposed the IACN model, which incorporated compartments representing infected animals, carcasses, and environmental reservoirs to better capture the ecological complexity of herbivore anthrax transmission. Osman, Makinde, and Theuri [28] provided an enhanced mathematical analysis of anthrax dynamics, offering qualitative insights and valuable implications for disease forecasting and control measures. Van den Driessche [37] emphasized

the importance of mathematical modeling as a critical tool for understanding the spread and control of infectious diseases, including anthrax.

Pantha, Day, and Lenhart [29] explored optimal control strategies within their mathematical framework, identifying approaches to effectively manage anthrax outbreaks through targeted interventions. Roy, van den Driessche, and Yakubu [32] developed a model to investigate how the growth and persistence of Bacillus anthracis spores in the environment affect fluctuations in susceptible animal populations. In addition, Kar and Batabyal [20] incorporated optimal control theory into their analysis of a nonlinear SIR epidemic model, emphasizing the role of vaccination as a key component in reducing the burden of the disease. Oh et al. [27] presented an analysis of a mathematical model specifically focused on gastrointestinal anthrax, shedding light on the transmission mechanisms and potential intervention strategies for this form of the disease. Finally, Karginov et al. [21] examined therapeutic approaches, modeling the combined effects of ciprofloxacin and antibody treatments in controlling bacterial growth and improving recovery outcomes.

Building on previous studies that develop models to analyze the transmission dynamics of anthrax, this work introduces a numerical approach that leverages the synergy between the Genocchi polynomials framework and the quasilinearization method (QLM). Earlier models in the literature have offered valuable insights into anthrax epidemiology using deterministic, stochastic, or optimal control frameworks. However, these models often rely on traditional numerical solvers and do not fully exploit advanced approximation techniques capable of handling nonlinearities with high accuracy and stability. Motivated by these gaps, our study presents a refined and computationally efficient framework that not only enhances the accuracy of solutions but also provides a robust foundation for sensitivity and control analysis of anthrax disease dynamics.

The main contributions and novelty of this work can be summarized as follows:

1. A complete analyze of an extended four-compartment anthrax transmission model, incorporating the interactions between susceptible, infected, recovered, and vaccinated animal populations, while capturing key biological processes such as disease-induced mortality, immunity waning, and demographic effects.

2. A hybrid Genocchi-quasilinearization method (Genocchi-QLM) is proposed, which is a numerical scheme that transforms the nonlinear system into a sequence of linear systems efficiently approximated by shifted Genocchi polynomials. To the best of our knowledge, this approach has not been previously applied to anthrax epidemiological models.

3. A rigorous convergence and error analysis is conducted, theoretically establishing the uniform convergence of the method and validating its accuracy through a residual error function technique.

4. We perform an extensive comparative validation to `ode45`, demonstrating the accuracy, efficiency, and stability of the Genocchi-QLM in capturing the dynamics of the model.

5. Real-world validation of the results confirms the ability of the method to provide highly accurate and reliable outcomes for the simulations.

6. Through comprehensive numerical experiments, we investigate the sensitivity of the system to key parameters, providing insights into the impact of vaccination strategies, recovery rates, and immunity waning on disease persistence and control.

The organization of this paper is as follows: In section 2, we present the mathematical model of anthrax disease in animals and define the underlying assumptions. Section 2 provides a review of Genocchi polynomials and then the shifted Genocchi polynomials with the error analysis. In addition to discussing the QLM for the anthrax model, section 4 presents the Genocchi collocation method, which is based on the quasilinearization approach with residual error. Numerical simulation results are presented in section 5, emphasizing the impact of significant parameters on disease control and transmission. At last, a conclusion to the paper is presented in section 6.

2 Model formulation

In this section, we need to construct a nonlinear mathematical model of the anthrax disease model in animals. To formulate the proposed model, let us subdivide the total number of population $N(t)$ at time t into four mutually exclusive classes such as: $S(t)$, which represents the number of susceptible animals that are at risk of contracting anthrax, while $I(t)$ denotes the infected animals exhibiting clinical symptoms of the disease. The $R(t)$ compartment accounts for recovered animals that have survived infection and acquired temporary or partial immunity. The final compartment, $V(t)$, represents the vaccinated animal population, which has gained immunity through preventive vaccination programs. Thus, the final model can be considered as follows [33, 9]:

$$\begin{aligned}
\frac{dS(t)}{dt} &= \eta - \alpha I(t)S(t) - (\beta+\gamma)S(t) + \varepsilon V(t) + \theta R(t),\\
\frac{dI(t)}{dt} &= \alpha I(t)S(t) - (\beta+\mu+\lambda)I(t),\\
\frac{dR(t)}{dt} &= \lambda I(t) - (\beta+\theta)R(t),\\
\frac{dV(t)}{dt} &= \gamma S(t) - (\beta+\varepsilon)V(t),
\end{aligned} \tag{1}$$

with the initial conditions

$$S(0) = S_0, \quad I(0) = I_0, \quad R(0) = R_0, \quad V(0) = V_0. \tag{2}$$

The description of the parameters of the model is given in Table 1, while Figure 2 represents the flowchart diagram of the anthrax disease model. In the next section, we will provide the fundamentals of the Genocchi polynomials.

Table 1: Declaration of parameters for the anthrax disease model.

Parameters	**Description**
η	Rate of recruitment
α	Rate of contact
β	Rate of natural death
γ	Rate of vaccination
θ	The waning rate of recovery
ε	The waning immunity rate of vaccinated animals
μ	Rate of induced-disease death
λ	The recovery rate of animals

3 Fundamentals of Genocchi polynomials

In this section, we first demonstrate the overview of Genocchi polynomials with significant properties on the interval $[0, 1]$. Then, we shift the interval to general form $[0, b]$ to get the modified Genocchi polynomials and present the error analysis.

3.1 Fundamental features of Genocchi polynomials

The Genocchi polynomials and the Genocchi numbers can be defined by the generating functions as follows [15, 8]:

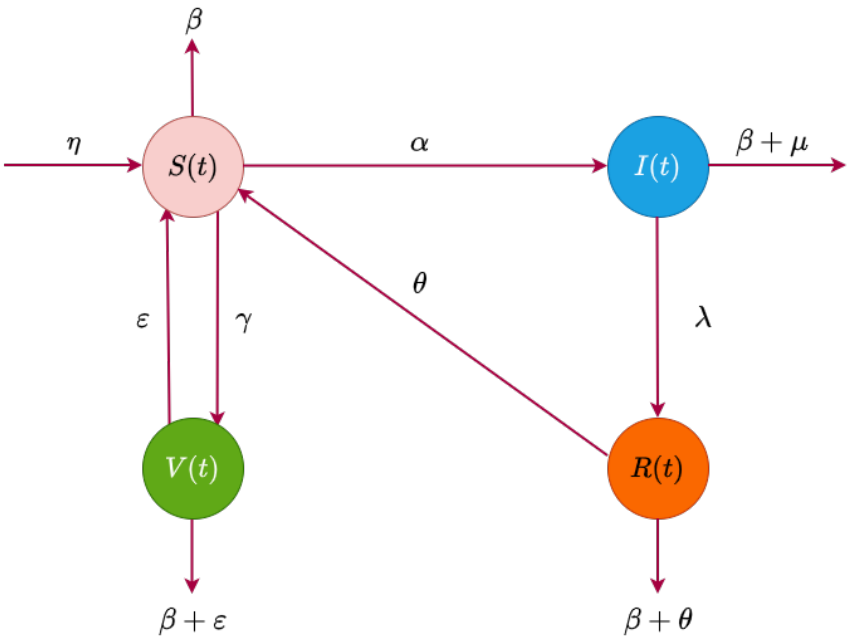


Figure 2: The flowchart for the anthrax disease model.

$$\frac{2xe^{x\tau}}{e^x+1}=\sum_{q=0}^{\infty}G_q(\tau)\frac{x^q}{q!},\quad (|x|<\pi), \tag{3}$$

$$\frac{2x}{e^x+1}=\sum_{q=0}^{\infty}G_q\frac{x^q}{q!},\quad (|x|<\pi). \tag{4}$$

Definition 1. The Genocchi polynomials $G_q(\tau)$ of degree q on the interval $[0,1]$ are defined as follows:

$$G_q(\tau)=\sum_{k=0}^{q}\binom{q}{k}G_k\,\tau^{q-k}, \tag{5}$$

where G_k represents the Genocchi numbers, which are connected with Bernoulli numbers in the following way:

$$G_q=2(1-2^q)B_q,\quad q>1. \tag{6}$$

The following equations are some of the many interesting characteristics of the Genocchi polynomials:

$$\frac{d}{d\tau}G_q(\tau)=q\,G_{q-1}(\tau),\quad q\geq 1, \tag{7}$$

$$G_q(\tau+1)+G_q(\tau)=2\,q\,\tau^{q-1},\quad q\geq 1, \tag{8}$$

$$G_q(1)+G_q(0)=0,\quad q>1. \tag{9}$$

Let us represent the $(Q+1)$ vector of Genocchi polynomials by

$$\boldsymbol{G}_Q(\tau)=\left[G_1(\tau)\;G_2(\tau)\;\ldots\;G_{Q+1}(\tau)\right]. \tag{10}$$

We can deduce from this representation the following lemma.

Lemma 1. The mathematical matrix representation for $\boldsymbol{G}_Q(\tau)$ satisfies

$$\boldsymbol{G}_Q(\tau)=\boldsymbol{\Delta}_Q(\tau)\boldsymbol{M}, \tag{11}$$

where

$$\boldsymbol{\Delta}_Q(\tau)=\left[1\;\tau\;\tau^2\;\ldots\;\tau^Q\right],$$

and

$$\boldsymbol{M}=\begin{bmatrix} \binom{1}{1}G_1 & \binom{2}{2}G_2 & \binom{3}{3}G_3 & \ldots & \binom{Q}{Q}G_Q & \binom{Q+1}{Q+1}G_{Q+1}\\ 0 & \binom{2}{1}G_1 & \binom{3}{2}G_2 & \ldots & \binom{Q}{Q-1}G_{Q-1} & \binom{Q+1}{Q}G_Q\\ 0 & 0 & \binom{3}{1}G_1 & \ldots & \binom{Q}{Q-2}G_{Q-2} & \binom{Q+1}{Q-1}G_{Q-1}\\ \vdots & \vdots & \vdots & \ddots & \vdots & \vdots\\ 0 & 0 & 0 & \ldots & \binom{Q}{1}G_1 & \binom{Q+1}{2}G_2\\ 0 & 0 & 0 & \ldots & \ldots & \binom{Q+1}{1}G_1 \end{bmatrix}_{(Q+1)\times(Q+1)}. \tag{12}$$

3.2 Modified Genocchi polynomials on the interval $[0, b]$

The approximate solution of function for interval $[0,1]$ in terms of Genocchi polynomials is expressed as

$$\Psi_Q(\tau)=\sum_{k=1}^{Q+1}\varphi_k G_k(\tau),\qquad \tau\in[0,1]. \tag{13}$$

To obtain the shifted version of Genocchi polynomials, we replace τ by t/b in the matrix $\boldsymbol{G}_Q(\tau)$.

Definition 2. The shifted Genocchi polynomials on $[0,b]$ is defined via relation

$$\boldsymbol{G}_Q(t)=\boldsymbol{\Delta}_Q(t)\boldsymbol{M},\qquad t\in[0,b], \tag{14}$$

where

$$\boldsymbol{\Delta}_Q(t) = \left[1 \; t/b \; (t/b)^2 \; \ldots \; (t/b)^Q\right], \tag{15}$$

so the approximate solution of the function on the interval $[0, b]$ is given by

$$\Psi_Q(t) = \sum_{k=1}^{Q+1} \varphi_k G_k(t), \qquad t \in [0, b]. \tag{16}$$

3.3 Error analysis on the interval $[0, b]$

In this section, we expand the error analysis of the Genocchi polynomials approximation from the unit interval $[0, 1]$ to a broad finite interval $[0, b]$, where $b > 0$. Assume first that the approximated function $g(t)$ is an arbitrary element of $H = L^2[0, b]$ and

$$\Pi = span\{G_1(t), G_2(t), \ldots, G_{Q+1}(t)\} \subset H,$$

where $\{G_k(t)\}_{k=1}^{Q+1}$ denotes the shifted Genocchi polynomials on $[0, b]$. We can state that $g(t)$ has a unique best approximation $g^*(t)$ in the space Π such that for all $\Psi(t) \in \Pi$,

$$\|g(t) - g^*(t)\|_2 \leq \|g(t) - \Psi(t)\|_2 . \tag{17}$$

Since $g^*(t) \in \Pi$, there exists a unique set of coefficients $\{\varphi_k\}_{k=1}^{Q+1}$ such that

$$g(t) \cong g^*(t) = \sum_{k=1}^{Q+1} \varphi_k G_k(t) = \boldsymbol{G}(t)\boldsymbol{\Phi},$$

where $\boldsymbol{\Phi} = \left[\varphi_1 \; \varphi_2 \; \ldots \; \varphi_{Q+1}\right]^T$, $\boldsymbol{G}(t) = \left[G_1(t) \; G_2(t) \; \ldots \; G_{Q+1}(t)\right]$. The following lemma is required in order to determine the values of the coefficients.

Lemma 2. Assume that $g \in C^Q[0, b]$ can be approximated by the Genocchi series $\sum_{k=1}^{Q+1} \varphi_k G_k(t)$. Then the unknown coefficients $\{\varphi_k\}_{k=1}^{Q+1}$ satisfy

$$\varphi_k = \frac{1}{2(k!)} \left(g^{(k-1)}(0) + g^{(k-1)}(b)\right).$$

Proof. The result follows from the properties of Genocchi polynomials in (8) and (9) for the interval $[0, b]$. Please refer to [14] for proof. □

Lemma 3. Suppose that $g(t) \in C^{(k+1)}[0,b]$ with $k \leq Q$ and that $\Pi = \text{span}\{G_1(t), G_2(t), \ldots, G_{Q+1}(t)\}$. If $\boldsymbol{G}(t)\boldsymbol{\Phi}$ is the best approximation of $g(t)$ from Π in the L^2-norm, then

$$\|\xi_{Q+1}(g)\| \leq \frac{b^{\frac{2k+3}{2}}\,\Theta}{(k+1)!\,\sqrt{2k+3}},$$

where

$$\xi_{Q+1}(g) = g(t) - \boldsymbol{G}(t)\boldsymbol{\Phi}, \quad \Theta = \max_{t\in[0,b]} |g^{(k+1)}(t)|.$$

Proof. In order to observe this, we set

$$\Psi_1(t) = g(0) + tg'(0) + \frac{t^2}{2!}g''(0) + \cdots + \frac{t^k}{k!}g^{(k)}(0).$$

It is evident from Taylor's expansion that

$$|g(t) - \Psi_1(t)| \leq \left|g^{(k+1)}(\xi_t)\right| \frac{(t)^{k+1}}{(k+1)!},$$

where $\xi_t \in [0,b]$. Since $\boldsymbol{G}(t)\boldsymbol{\Phi}$ is the best approximation of $g(t)$ out of Π and $\Psi_1(t) \in \Pi$, then from (17), we have

$$\begin{aligned}
\|g(t) - \boldsymbol{G}(t)\boldsymbol{\Phi}\|_{L_2}^2 \leq \|g(t) - \Psi_1(t)\|_{L_2}^2 &= \int_0^b |g(t) - \Psi_1(t)|^2 dt \\
&\leq \int_0^b |g^{(k+1)}(\xi_t)|^2 \left(\frac{t^{k+1}}{(k+1)!}\right)^2 dt \\
&\leq \frac{b^{2k+3}\,\Theta^2}{((k+1)!)^2\,(2k+3)}.
\end{aligned}$$

When we take square roots of both sides, we get

$$\|g(t) - \boldsymbol{G}(t)\boldsymbol{\Phi}\| \leq \frac{b^{\frac{2k+3}{2}}\,\Theta}{(k+1)!\,\sqrt{2k+3}};$$

this complete the proof. □

4 Quasilinearization technique

In this section, we will cover the main concept and steps of the recommended QLM. The goal of the QLM is to convert the original nonlinear model (1) into a set of linear equations. The Genocchi collocation method will next be applied to the linearized models that have been produced.

Here, we briefly outline the basic concepts that allow the QLM to solve our model problems; refer to [18, 16, 38, 1, 17] for more in-depth discussions and applications. Let us examine the nonlinear form (1), which represents the standard formulation of our model. To solve the system of equations (1), we need to first restate the anthrax disease model as

$$\dot{\boldsymbol{\Upsilon}}(t) = \boldsymbol{H}(t, \boldsymbol{\Upsilon}(t)), \tag{18}$$

where

$$\dot{\boldsymbol{\Upsilon}}(t) = \begin{bmatrix} \dot{S}(t) \\ \dot{I}(t) \\ \dot{R}(t) \\ \dot{V}(t) \end{bmatrix}, \quad \boldsymbol{\Upsilon}(t) = \begin{bmatrix} S(t) \\ I(t) \\ R(t) \\ V(t) \end{bmatrix},$$

$$\boldsymbol{H}(t, \boldsymbol{\Upsilon}(t)) = \begin{bmatrix} \eta - \alpha I(t)S(t) - (\beta + \gamma)S(t) + \theta R(t) + \varepsilon V(t) \\ \alpha I(t)S(t) - (\beta + \mu + \lambda)I(t) \\ \lambda I(t) - (\beta + \varepsilon)R(t) \\ \gamma S(t) - (\beta + \theta)V(t) \end{bmatrix}.$$

Next, select an initial approximation $\boldsymbol{\Upsilon}_0(t)$ for $\boldsymbol{\Upsilon}(t)$. Then, the QLM iteration for (18) is defined as follows:

$$\dot{\boldsymbol{\Upsilon}}_{\psi+1}(t) = \boldsymbol{H}(t, \boldsymbol{\Upsilon}_\psi(t)) + (\boldsymbol{\Upsilon}_{\psi+1}(t) - \boldsymbol{\Upsilon}_\psi(t))\boldsymbol{H}_{\boldsymbol{\Upsilon}}(t, \boldsymbol{\Upsilon}_\psi(t)), \quad \psi = 0, 1, \ldots, \tag{19}$$

with the initial conditions as mentioned in (2). The functional derivatives of $\boldsymbol{H}(t, \boldsymbol{\Upsilon}(t))$ are demonstrated by the (4×4) Jacobian matrix of anthrax system (18). The following linearized model form is produced after the QLM approach is applied to the nonlinear model (18) as follows:

$$\dot{\boldsymbol{\Upsilon}}_{\psi+1}(t) + \boldsymbol{E}_\psi(t)\, \boldsymbol{\Upsilon}_{\psi+1}(t) = \boldsymbol{\Sigma}_\psi(t), \qquad \psi = 0, 1, \ldots, \tag{20}$$

where

$$\boldsymbol{E}_\psi(t) = \begin{bmatrix} \alpha I_\psi(t) + \beta + \gamma & \alpha S_\psi(t) & -\theta & -\varepsilon \\ -\alpha I_\psi(t) & -\alpha S_\psi(t) + \beta + \mu + \lambda & 0 & 0 \\ 0 & -\lambda & \beta + \theta & 0 \\ -\gamma & 0 & 0 & \beta + \varepsilon \end{bmatrix},$$

$$\boldsymbol{\Sigma}_\psi(t) = \begin{bmatrix} \eta + \alpha S_\psi(t) I_\psi(t) \\ -\alpha S_\psi(t) I_\psi(t) \\ 0 \\ 0 \end{bmatrix},$$

with the initial conditions given by

$$\boldsymbol{\Upsilon}_{\psi+1}(0) = \begin{bmatrix} S_{\psi+1}(0) \\ I_{\psi+1}(0) \\ R_{\psi+1}(0) \\ V_{\psi+1}(0) \end{bmatrix} = \begin{bmatrix} S_0 \\ I_0 \\ R_0 \\ V_0 \end{bmatrix}. \tag{21}$$

Therefore, we solve the quasilinear model (20) using the Genocchi collocation method instead of solving the original model (1). The last method is called Genocchi-QLM.

4.1 Genocchi-QLM

Here, we aim to solve the system of linearized initial-value problems in (20) and (21) on the interval $[0, b]$ using Genocchi-QLM. To continue, we suppose that the truncated Genocchi polynomials series form (16) can be used to express the solutions of model (20). At iteration $\psi \geq 0$, these approximations can be written as follows:

$$\begin{aligned} S_Q^{\psi+1}(t) &= \sum_{k=1}^{Q+1} \varphi_{k,1}^{\psi} G_k(t) = \boldsymbol{G}_Q(t)\boldsymbol{\Phi}_{Q,1}^{\psi}, \\ I_Q^{\psi+1}(t) &= \sum_{k=1}^{Q+1} \varphi_{k,2}^{\psi} G_k(t) = \boldsymbol{G}_Q(t)\boldsymbol{\Phi}_{Q,2}^{\psi}, \\ R_Q^{\psi+1}(t) &= \sum_{k=1}^{Q+1} \varphi_{k,3}^{\psi} G_k(t) = \boldsymbol{G}_Q(t)\boldsymbol{\Phi}_{Q,3}^{\psi}, \\ V_Q^{\psi+1}(t) &= \sum_{k=1}^{Q+1} \varphi_{k,4}^{\psi} G_k(t) = \boldsymbol{G}_Q(t)\boldsymbol{\Phi}_{Q,4}^{\psi}, \end{aligned} \tag{22}$$

where $\boldsymbol{\Phi}_{Q,\upsilon}^{\psi}$ represents the vectors of unknown coefficients associated with each approximate solution and is given by

$$\boldsymbol{\Phi}_{Q,\upsilon}^{\psi} = \left[\varphi_{1,\upsilon}^{\psi}\ \varphi_{2,\upsilon}^{\psi}\ \ldots\ \varphi_{Q+1,\upsilon}^{\psi}\right]^T, \quad \upsilon = 1, 2, 3, 4.$$

The representation of $\boldsymbol{G}_Q(t)$ in (14) enables us to reconstruct the relations in (22) as follows:

$$\begin{aligned} S_Q^{\psi+1}(t) &= \boldsymbol{\Delta}_Q(t)\, \boldsymbol{M}\, \boldsymbol{\Phi}_{Q,1}^{\psi}, \qquad I_Q^{\psi+1}(t) = \boldsymbol{\Delta}_Q(t)\, \boldsymbol{M}\, \boldsymbol{\Phi}_{Q,2}^{\psi}, \\ R_Q^{\psi+1}(t) &= \boldsymbol{\Delta}_Q(t)\, \boldsymbol{M}\, \boldsymbol{\Phi}_{Q,3}^{\psi}, \qquad V_Q^{\psi+1}(t) = \boldsymbol{\Delta}_Q(t)\, \boldsymbol{M}\, \boldsymbol{\Phi}_{Q,4}^{\psi}. \end{aligned} \tag{23}$$

Next, we deduce the first derivative of approximate solutions in (23) by the following relations:

$$\begin{aligned} \dot{S}_Q^{\psi+1}(t) &= \boldsymbol{\Delta}_Q(t)\, \boldsymbol{D}\, \boldsymbol{M}\, \boldsymbol{\Phi}_{Q,1}^{\psi}, \qquad \dot{I}_Q^{\psi+1}(t) = \boldsymbol{\Delta}_Q(t)\, \boldsymbol{D}\, \boldsymbol{M}\, \boldsymbol{\Phi}_{Q,2}^{\psi}, \\ \dot{R}_Q^{\psi+1}(t) &= \boldsymbol{\Delta}_Q(t)\, \boldsymbol{D}\, \boldsymbol{M}\, \boldsymbol{\Phi}_{Q,3}^{\psi}, \qquad \dot{V}_Q^{\psi+1}(t) = \boldsymbol{\Delta}_Q(t)\, \boldsymbol{D}\, \boldsymbol{M}\, \boldsymbol{\Phi}_{Q,4}^{\psi}, \end{aligned} \tag{24}$$

where $\boldsymbol{D}$ is the differentiation matrix of Genocchi polynomials and is defined as follows:

$$\boldsymbol{D} = \frac{1}{b} \begin{bmatrix} 0 & 2 & 0 & \cdots & 0 \\ 0 & 0 & 3 & \cdots & 0 \\ \vdots & \vdots & \vdots & \ddots & \vdots \\ 0 & 0 & 0 & \cdots & Q+1 \\ 0 & 0 & 0 & \cdots & 0 \end{bmatrix}_{(Q+1)\times(Q+1)}. \tag{25}$$

Now, we are interested in the matrix representations of $\boldsymbol{\Upsilon}_{\psi+1}(t)$ and $\dot{\boldsymbol{\Upsilon}}_{\psi+1}(t)$ in (20). So we have the following matrix representations:

$$\begin{aligned} \bar{\boldsymbol{\Delta}}_Q(t) &= \begin{bmatrix} \boldsymbol{\Delta}_Q(t) & \mathbf{0} & \mathbf{0} & \mathbf{0} \\ \mathbf{0} & \boldsymbol{\Delta}_Q(t) & \mathbf{0} & \mathbf{0} \\ \mathbf{0} & \mathbf{0} & \boldsymbol{\Delta}_Q(t) & \mathbf{0} \\ \mathbf{0} & \mathbf{0} & \mathbf{0} & \boldsymbol{\Delta}_Q(t) \end{bmatrix}, \quad \bar{\boldsymbol{\Phi}}_Q^{\psi} = \begin{bmatrix} \boldsymbol{\Phi}_{Q,1}^{\psi} \\ \boldsymbol{\Phi}_{Q,2}^{\psi} \\ \boldsymbol{\Phi}_{Q,3}^{\psi} \\ \boldsymbol{\Phi}_{Q,4}^{\psi} \end{bmatrix}, \\ \bar{\boldsymbol{M}} &= \begin{bmatrix} \boldsymbol{M} & \mathbf{0} & \mathbf{0} & \mathbf{0} \\ \mathbf{0} & \boldsymbol{M} & \mathbf{0} & \mathbf{0} \\ \mathbf{0} & \mathbf{0} & \boldsymbol{M} & \mathbf{0} \\ \mathbf{0} & \mathbf{0} & \mathbf{0} & \boldsymbol{M} \end{bmatrix}, \quad \bar{\boldsymbol{D}} = \begin{bmatrix} \boldsymbol{D} & \mathbf{0} & \mathbf{0} & \mathbf{0} \\ \mathbf{0} & \boldsymbol{D} & \mathbf{0} & \mathbf{0} \\ \mathbf{0} & \mathbf{0} & \boldsymbol{D} & \mathbf{0} \\ \mathbf{0} & \mathbf{0} & \mathbf{0} & \boldsymbol{D} \end{bmatrix}. \end{aligned} \tag{26}$$

Then, the approximate solutions $\boldsymbol{\Upsilon}_{\psi+1}(t)$ and $\dot{\boldsymbol{\Upsilon}}_{\psi+1}(t)$ can be expressed in matrix forms as follows:

$$\boldsymbol{\Upsilon}_{\psi+1}(t) = \bar{\boldsymbol{\Delta}}_Q(t)\, \bar{\boldsymbol{M}}\, \bar{\boldsymbol{\Phi}}_Q^{\psi}, \qquad \dot{\boldsymbol{\Upsilon}}_{\psi+1}(t) = \bar{\boldsymbol{\Delta}}_Q(t)\, \bar{\boldsymbol{D}}\, \bar{\boldsymbol{M}}\, \bar{\boldsymbol{\Phi}}_Q^{\psi}. \tag{27}$$

By using the collocation points defined as

$$t_i = \frac{i\, b}{Q}, \qquad i = 0, 1, \ldots, Q, \tag{28}$$

in linear model (20), we obtain

$$\dot{\boldsymbol{\Upsilon}}_{\psi+1}(t_i) + \boldsymbol{E}_{\psi}(t_i)\, \boldsymbol{\Upsilon}_{\psi+1}(t_i) = \boldsymbol{\Sigma}_{\psi}(t_i), \qquad i = 0, 1, \ldots, Q, \tag{29}$$

which can be simplified as follows:

$$\breve{\dot{\boldsymbol{\Upsilon}}}_{\psi+1} + \breve{\boldsymbol{E}}_{\psi}\, \breve{\boldsymbol{\Upsilon}}_{\psi+1} = \breve{\boldsymbol{\Sigma}}_{\psi}, \tag{30}$$

where

$$\breve{\dot{\boldsymbol{\Upsilon}}}_{\psi+1} = \begin{bmatrix} \dot{\boldsymbol{\Upsilon}}_{\psi+1}(t_0) \\ \dot{\boldsymbol{\Upsilon}}_{\psi+1}(t_1) \\ \vdots \\ \dot{\boldsymbol{\Upsilon}}_{\psi+1}(t_Q) \end{bmatrix}, \quad \breve{\boldsymbol{E}}_{\psi} = \begin{bmatrix} \boldsymbol{E}_{\psi}(t_0) & \mathbf{0} & \cdots & \mathbf{0} \\ \mathbf{0} & \boldsymbol{E}_{\psi}(t_1) & \cdots & \mathbf{0} \\ \vdots & \vdots & \ddots & \vdots \\ \mathbf{0} & \mathbf{0} & \cdots & \boldsymbol{E}_{\psi}(t_Q) \end{bmatrix},$$

$$\breve{\boldsymbol{\Upsilon}}_{\psi+1} = \begin{bmatrix} \boldsymbol{\Upsilon}_{\psi+1}(t_0) \\ \boldsymbol{\Upsilon}_{\psi+1}(t_1) \\ \vdots \\ \boldsymbol{\Upsilon}_{\psi+1}(t_Q) \end{bmatrix}, \quad \breve{\boldsymbol{\Sigma}}_{\psi} = \begin{bmatrix} \boldsymbol{\Sigma}_{\psi}(t_0) \\ \boldsymbol{\Sigma}_{\psi}(t_1) \\ \vdots \\ \boldsymbol{\Sigma}_{\psi}(t_Q) \end{bmatrix}.$$

For the matrix representation of $\breve{\dot{\boldsymbol{\Upsilon}}}_{\psi+1}$ and $\breve{\boldsymbol{\Upsilon}}_{\psi+1}$, we need the following lemma.

Lemma 4. The matrices $\breve{\boldsymbol{\Upsilon}}_{\psi+1}$ and $\breve{\dot{\boldsymbol{\Upsilon}}}_{\psi+1}$, produced from using collocation points (28) in $\boldsymbol{\Upsilon}_{\psi+1}(t_i)$ and $\dot{\boldsymbol{\Upsilon}}_{\psi+1}(t_i)$, can be constructed as follows:

$$\breve{\boldsymbol{\Upsilon}}_{\psi+1} = \breve{\boldsymbol{\Delta}}_{Q}\, \bar{\boldsymbol{M}}\, \bar{\boldsymbol{\Phi}}_{Q}^{\psi}, \qquad \breve{\dot{\boldsymbol{\Upsilon}}}_{\psi+1} = \breve{\boldsymbol{\Delta}}_{Q}\, \bar{\boldsymbol{D}}\, \bar{\boldsymbol{M}}\, \bar{\boldsymbol{\Phi}}_{Q}^{\psi}, \tag{31}$$

where

$$\breve{\boldsymbol{\Delta}}_{Q} = \begin{bmatrix} \bar{\boldsymbol{\Delta}}_{Q}(t_0) & \mathbf{0} & \cdots & \mathbf{0} \\ \mathbf{0} & \bar{\boldsymbol{\Delta}}_{Q}(t_1) & \cdots & \mathbf{0} \\ \vdots & \vdots & \ddots & \vdots \\ \mathbf{0} & \mathbf{0} & \cdots & \bar{\boldsymbol{\Delta}}_{Q}(t_Q) \end{bmatrix}. \tag{32}$$

By using (31) into (29), we obtain the matrix form of the linear model as follows:

$$\breve{\boldsymbol{\Delta}}_{Q}\, \bar{\boldsymbol{D}}\, \bar{\boldsymbol{M}}\, \bar{\boldsymbol{\Phi}}_{Q}^{\psi} + \breve{\boldsymbol{E}}_{\psi}\, \breve{\boldsymbol{\Delta}}_{Q}\, \bar{\boldsymbol{M}}\, \bar{\boldsymbol{\Phi}}_{Q}^{\psi} = \breve{\boldsymbol{\Sigma}}_{\psi}, \tag{33}$$

which can be rewritten in the form:

$$\boldsymbol{\Xi}_{\psi+1}\, \bar{\boldsymbol{\Phi}}_{Q}^{\psi} = \breve{\boldsymbol{\Sigma}}_{\psi}, \qquad or \qquad [\boldsymbol{\Xi}_{\psi+1}; \breve{\boldsymbol{\Sigma}}_{\psi}], \tag{34}$$

where

$$\boldsymbol{\Xi}_{\psi+1} = \breve{\boldsymbol{\Delta}}_{Q}\, \bar{\boldsymbol{D}}\, \bar{\boldsymbol{M}} + \breve{\boldsymbol{E}}_{\psi}\, \breve{\boldsymbol{\Delta}}_{Q}\, \bar{\boldsymbol{M}}.$$

At last, the matrix representation of the initial conditions, given in (2), is produced when tending $t \to 0$ in (27) as follows:

$$\bar{\boldsymbol{\Delta}}_Q(0)\,\bar{\boldsymbol{M}}\,\bar{\boldsymbol{\Phi}}_Q^{\psi} = \boldsymbol{B}_0^{\psi+1}, \tag{35}$$

where $\boldsymbol{B}_0^{\psi+1}$ is given in (21). After approximate the initial solution of model $\boldsymbol{\Upsilon}_0(t)$, we replace the first four rows of (34) with (35) to obtain the new augmented matrix

$$[\mathring{\bar{\boldsymbol{\Xi}}}_{\psi+1}; \mathring{\bar{\boldsymbol{\Sigma}}}_{\psi}]. \tag{36}$$

By solving this system of linear equations, we obtain the unknown coefficients $\boldsymbol{\Phi}_Q^{\psi}$, which are used to gain the four approximations of the anthrax disease model in (1).

4.2 Residual error functions

In this subsection, we further validate the results by calculating the residual error for the model (1). Validating the accuracy of the provided collocation technique is essential in this case. Let $Res_Q^{\psi+1}(t)$ represent the residual error function. As a result, we obtain

$$\begin{aligned}
Res_{S,Q}^{\psi+1}(t) &= |\dot{S}_Q^{\psi+1}(t) - \eta + \alpha I_Q^{\psi+1}(t) S_Q^{\psi+1}(t) + (\beta+\gamma) S_Q^{\psi+1}(t) - \theta R_Q^{\psi+1}(t) \\
&\quad - \varepsilon V_Q^{\psi+1}(t)| \cong 0, \\
Res_{I,Q}^{\psi+1}(t) &= |\dot{I}_Q^{\psi+1}(t) - \alpha I_Q^{\psi+1}(t) S_Q^{\psi+1}(t) + (\beta+\mu+\lambda) I_Q^{\psi+1}(t)| \cong 0, \\
Res_{R,Q}^{\psi+1}(t) &= |\dot{R}_Q^{\psi+1}(t) - \lambda I_Q^{\psi+1}(t) + (\beta+\theta) R_Q^{\psi+1}(t)| \cong 0, \\
Res_{V,Q}^{\psi+1}(t) &= |\dot{V}_Q^{\psi+1}(t) - \gamma S_Q^{\psi+1}(t) + (\beta+\varepsilon) V_Q^{\psi+1}(t)| \cong 0,
\end{aligned} \tag{37}$$

where $\psi = 0, 1, \ldots, 5$. Indeed, at the collocation points specified in (28), the residual error function $Res_Q^{\psi+1}(t)$ becomes zero. As a result, we anticipate that as Q increases, the residual error $Res_Q^{\psi+1}(t)$ converges to zero, ensuring that the approximate solution becomes increasingly accurate.

5 Numerical simulations and validation

5.1 Numerical results

In this subsection, we present the numerical results for the nonlinear anthrax disease model in animals, considering two cases. The suggested Genocchi-QLM is used to run the simulations, proving the method's effectiveness and dependability. To ensure the accuracy and consistency of the results, MATLAB 2023 is used for all calculations and simulations.

Example 1. Consider the following values of parameters in the first case of the anthrax disease model [33]:

$$\eta = 200, \quad \alpha = 0.0001, \quad \gamma = 0.1, \quad \beta = 0.001,$$
$$\theta = 0.02, \quad \varepsilon = 0.02, \quad \mu = 0.1, \quad \lambda = 0.01,$$

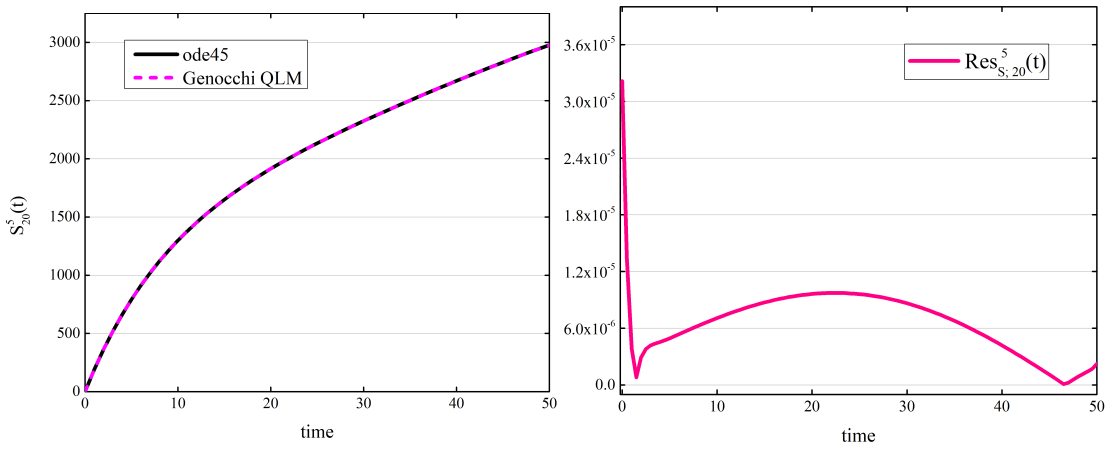
where the initial conditions are

$$S(0) = 0.2, \quad I(0) = 0.3, \quad R(0) = 0.4, \quad V(0) = 0.5.$$

In the iterated Genocchi-QLM procedure, by setting $\psi = 5$ and using $\boldsymbol{\Upsilon}_0(t) = 0$ as the initial approximation for the linearized system (20) with $Q = 20$ and a simulation time of $b = 50$, the dynamics of the approximate solutions for the susceptible, infected, recovered, and vaccinated populations are illustrated in Figures 3a, 4a, 5a, and 6a, respectively, with CPU time 20.648 sec. For validation purposes, the results obtained from the classical `ode45` solver in MATLAB are also presented for comparison. As observed, the numerical solutions generated by the Genocchi-QLM method exhibit an excellent agreement with those produced by `ode45`, confirming the reliability and accuracy of the proposed approach. Furthermore, the corresponding residual errors for each population class —susceptible, infected, recovered, and vaccinated— are depicted in Figures 3b, 4b, 5b, and 6b, respectively, demonstrating the high accuracy and convergence of the Genocchi-QLM technique across all compartments.

From a physical perspective, the results reflect the expected epidemiological behavior of anthrax transmission. The susceptible population $S(t)$ initially rises due to constant recruitment ($\eta = 200$) before stabilizing as infection, recovery, and vaccination interact to balance the dynamics. The infected population $I(t)$ steadily declines over time, driven by recovery ($\lambda = 0.01$) and disease-induced mortality ($\mu = 0.1$), highlighting the importance of control interventions to reduce infection levels. The recovered population $R(t)$ grows moderately, reflecting successful recoveries, while waning immunity ($\theta = 0.02$) allows some individuals to transition back into the susceptible pool, sustaining a low-level risk of reinfection. Similarly, the vaccinated population

$V(t)$ increases rapidly in the early phase due to a relatively high vaccination rate ($\gamma = 0.1$) before stabilizing, with waning immunity ($\varepsilon = 0.02$) gradually reintroducing a portion of vaccinated individuals into susceptibility. These trends collectively underscore the role of effective vaccination strategies and the need for continuous monitoring to manage disease persistence in animal populations.



(a) $S^5_{20}(t)$ and `ode45` (b) $Res^5_{S,20}(t)$

Figure 3: The approximate and `ode45` solutions of $S(t)$ for $Q = 20$ (3a) and the residual error (3b) for $\psi = 5$ in Example 1.

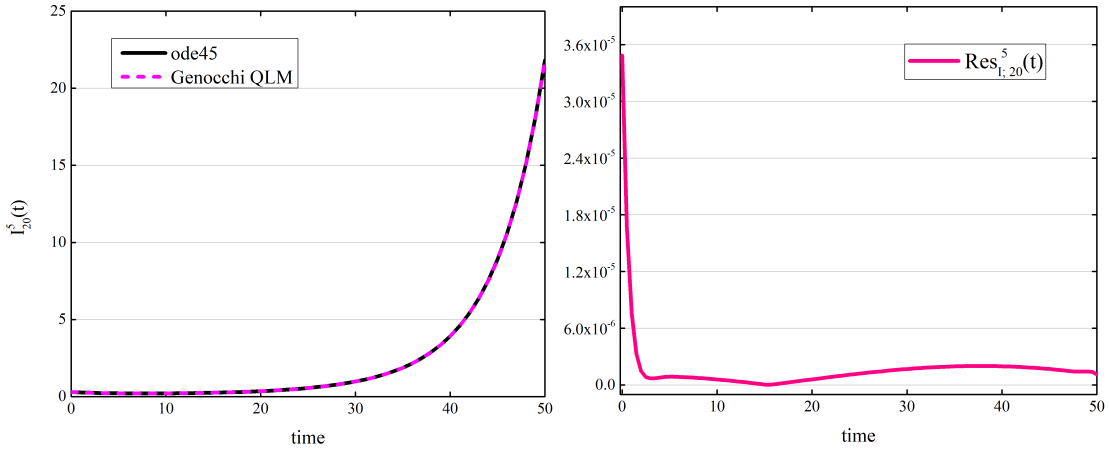
Example 2. In the next case of the anthrax disease model, consider the following values of parameters [9]:

$$\eta = 200, \quad \alpha = 0.0001, \quad \gamma = 0.1, \quad \beta = 0.001,$$
$$\theta = 0.02, \quad \varepsilon = 0.003, \quad \mu = 0.1, \quad \lambda = 0.01,$$

where the initial conditions are

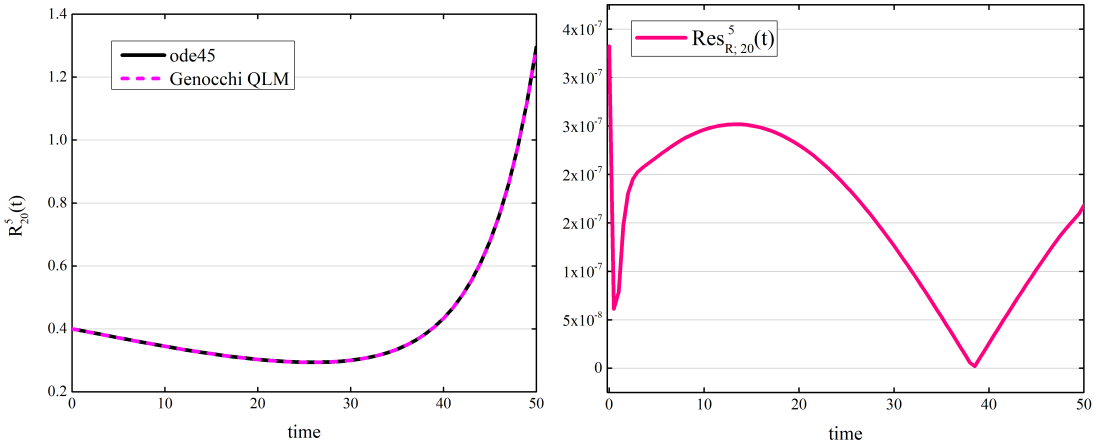
$$S(0) = 2000, \quad I(0) = 100, \quad R(0) = 300, \quad V(0) = 500.$$

Figures 7a, 8a, 9a, and 10a, respectively, show the results of approximate solutions for $S(t)$, $I(t)$, $R(t)$, and $V(t)$ obtained via the Genocchi-QLM method for $Q = 20$, simulation time $b = 50$, and iterated QLM with $\psi = 5$. Additionally, the results obtained from the classical `ode45` solver are presented for validation. As observed, the approximate solutions generated by the Genocchi-QLM approach show excellent agreement with those of `ode45`, confirming the accuracy and



(a) $I^5_{20}(t)$ and `ode45` (b) $Res^5_{I,20}(t)$

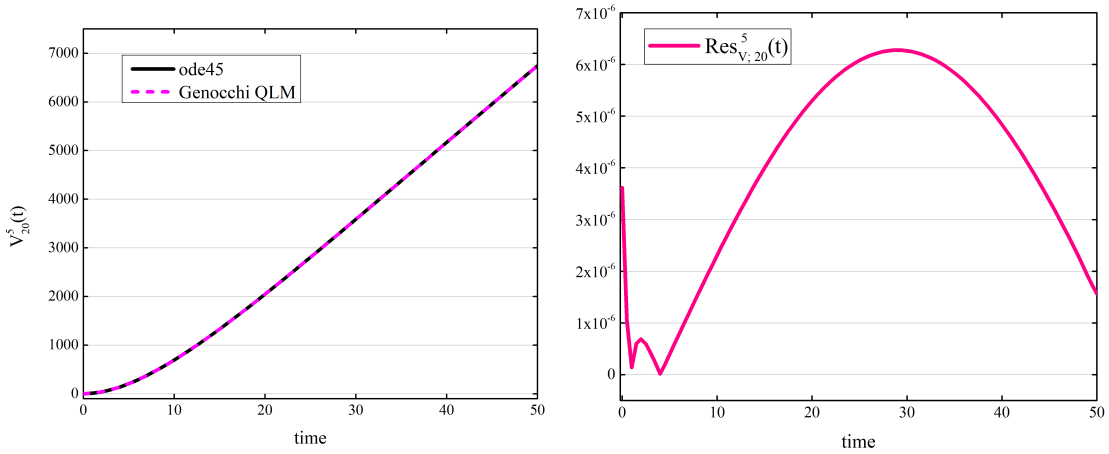
Figure 4: The approximate and `ode45` solutions of $I(t)$ for $Q = 20$ (4a) and the residual error (4b) for $\psi = 5$ in Example 1.



(a) $R^5_{20}(t)$ and `ode45` (b) $Res^5_{R,20}(t)$

Figure 5: The approximate and `ode45` solutions of $R(t)$ for $Q = 20$ (5a) and the residual error (5b) for $\psi = 5$ in Example 1.

stability of the proposed numerical technique. The residual errors for susceptible, infected, recovered, and vaccinated populations are shown in Figures 7b, 8b, 9b, and 10b with CPU time 20.113 sec, demonstrating the high precision and convergence behavior of the Genocchi-QLM method.



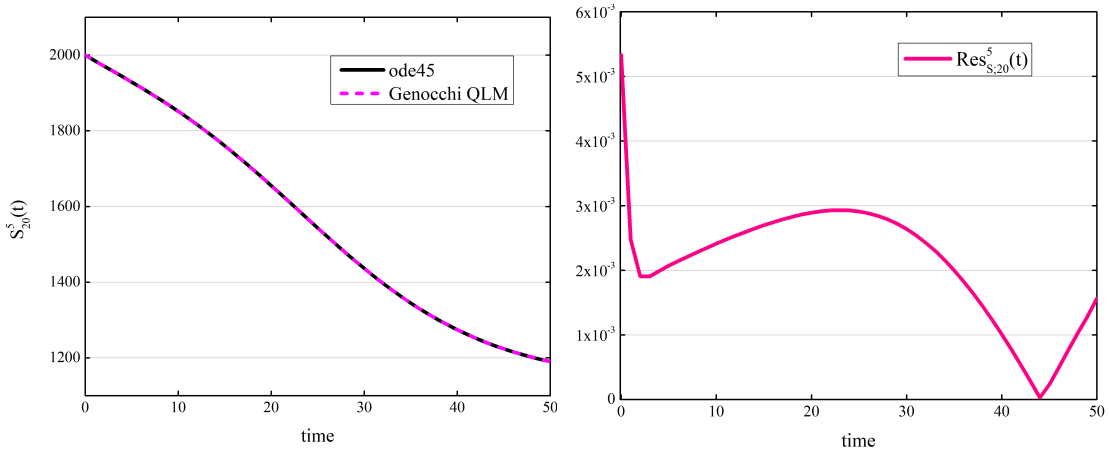
(a) $V_{20}^5(t)$ and `ode45` (b) $Res_{V,20}^5(t)$

Figure 6: The approximate and `ode45` solutions of $V(t)$ for $Q = 20$ (6a) and the residual error (6b) for $\psi = 5$ in Example 1.

From a physical perspective, the second example highlights more pronounced dynamic changes compared to the first case, reflecting the impact of parameter adjustments on disease transmission. The susceptible population $S(t)$ initially increases but stabilizes at a lower level, indicating stronger transmission or reduced vaccination coverage in this scenario. The infected population $I(t)$ displays a slower decline, suggesting either a reduced recovery rate or higher persistence of infection, which emphasizes the importance of timely interventions and robust treatment strategies. The recovered population $R(t)$ grows at a faster rate than in the first example, indicating that a significant portion of infected individuals eventually recover and gain temporary immunity, though waning immunity ($\theta = 0.02$) ensures a continuous flow back into the susceptible pool. The vaccinated population $V(t)$ also increases sharply at first, driven by a high vaccination rate ($\gamma = 0.1$), but gradually declines over time due to immunity loss ($\varepsilon = 0.003$). These dynamics collectively underline the importance of maintaining consistent vaccination efforts and integrating strategies that account for immunity waning to achieve sustainable control of anthrax outbreaks in animal populations.

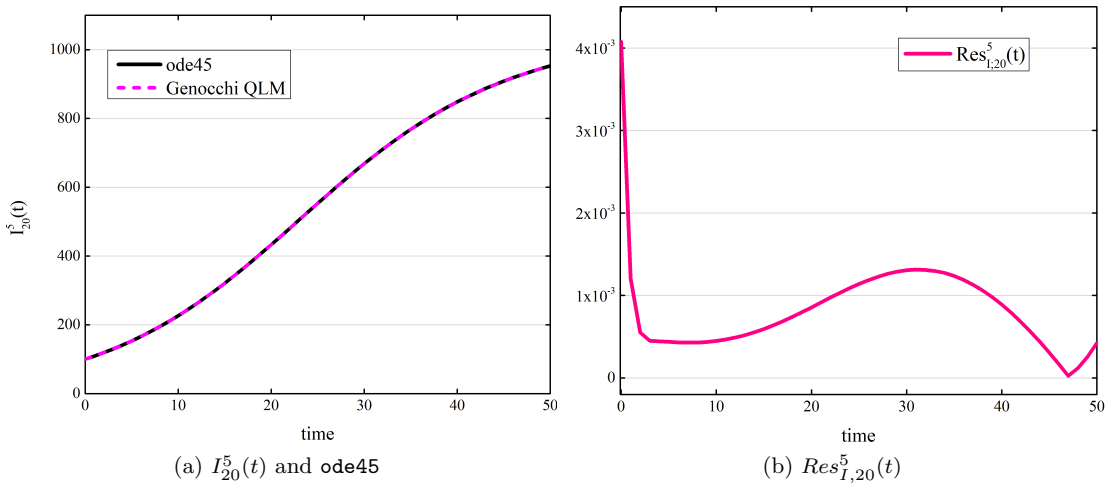
5.2 Real world validation

In this subsection, we validate the performance and applicability of the proposed Genocchi-QLM numerical framework by comparing the simulated results of the anthrax transmission model



(a) $S^5_{20}(t)$ and `ode45` (b) $Res^5_{S,20}(t)$

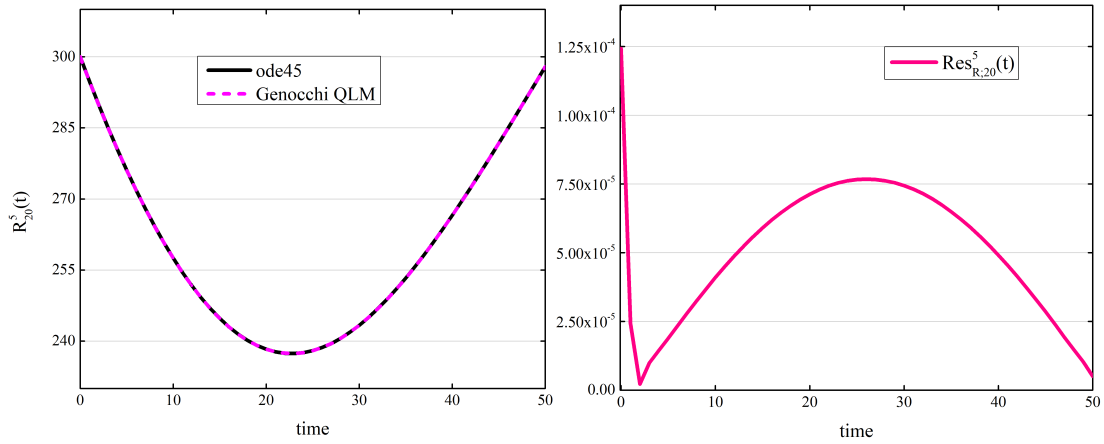
Figure 7: The approximate and `ode45` solutions of $S(t)$ for $Q = 20$ (7a) and the residual error (7b) for $\psi = 5$ in Example 2.



(a) $I^5_{20}(t)$ and `ode45` (b) $Res^5_{I,20}(t)$

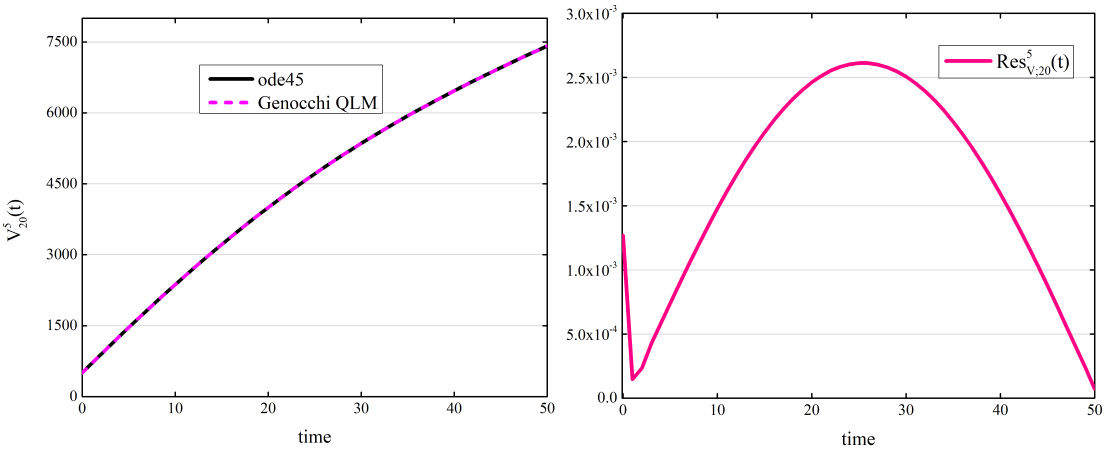
Figure 8: The approximate and `ode45` solutions of $I(t)$ for $Q = 20$ (8a) and the residual error (8b) for $\psi = 5$ in Example 2.

with real-world outbreak data. Figure 11 presents the reported number of anthrax cases and corresponding fatalities collected from official epidemiological surveillance data spanning multiple outbreak years form Minnesota and Kazakhstan [19]. These records, gathered from veterinary and public health databases, reflect the dynamics of anthrax outbreaks in animal populations under natural transmission conditions. To replicate the observed trends, we performed numerical simulations of the nonlinear four-compartment model over the time interval using the Genocchi-



(a) $R^5_{20}(t)$ and `ode45` (b) $Res^5_{R,20}(t)$

Figure 9: The approximate and `ode45` solutions of $R(t)$ for $Q = 20$ (9a) and the residual error (9b) for $\psi = 5$ in Example 2.



(a) $V^5_{20}(t)$ and `ode45` (b) $Res^5_{V,20}(t)$

Figure 10: The approximate and `ode45` solutions of $V(t)$ for $Q = 20$ (10a) and the residual error (10b) for $\psi = 5$ in Example 2.

QLM method with parameters both reported in the examples. As seen in Figure 11, there is a clear agreement between the simulated dynamics of the infected cases $I(t)$ and the actual case data. The model accurately reproduces the initial rapid rise in the number of infected cases during the early phases of each outbreak, followed by a gradual stabilization or decline due to recovery, mortality, and intervention measures such as vaccination campaigns and quarantine protocols. The validation results in Figure 11 highlight the practical applicability of the developed model

and numerical scheme. This level of agreement reinforces the potential of the Genocchi-QLM method to support decision-making in disease surveillance, risk assessment, and the design of effective intervention strategies in endemic regions [19].

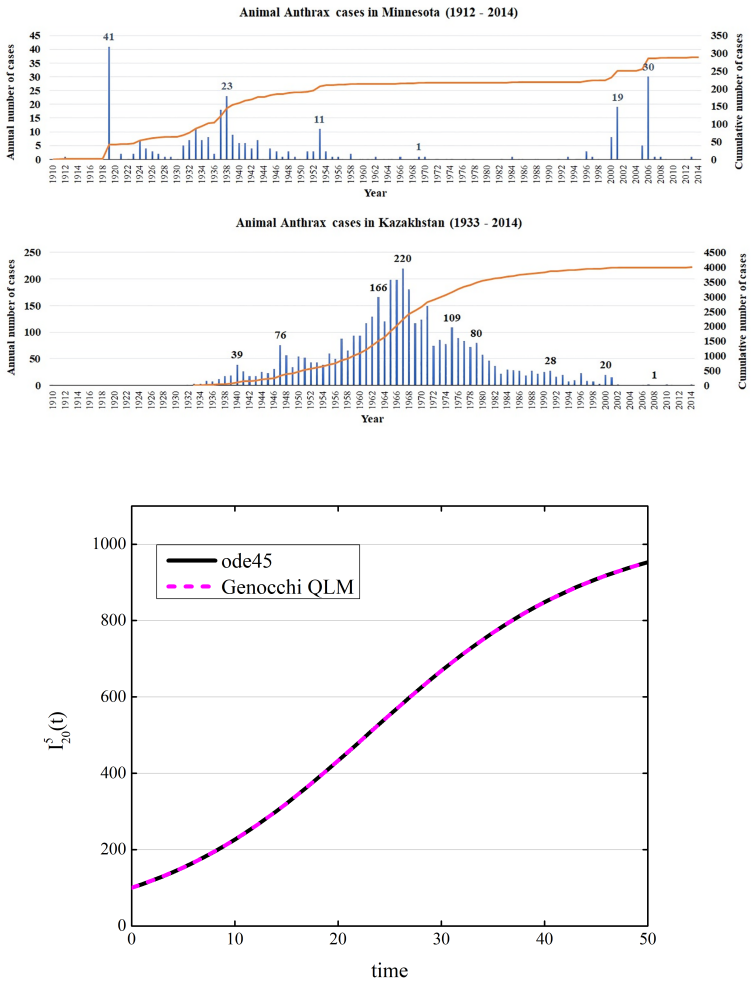


Figure 11: Comparison between the real-world outbreak data in Minnesota and Kazakhstan [19] and the simulated anthrax transmission dynamics obtained using the Genocchi-QLM method.

6 Conclusion

In this study, we developed and analyzed a nonlinear four-compartment anthrax transmission model that captures the interactions between susceptible, infected, recovered, and vaccinated animal populations. The Genocchi-QLM was introduced as a numerical approach, offering high accuracy, stability, and rapid convergence when compared with the classical `ode45` solver. The integration of shifted Genocchi polynomials with the quasilinearization technique allowed the construction of an efficient and scalable numerical framework, further validated through comprehensive error and convergence analyses. The validation of the proposed model against real-world anthrax outbreak data is discussed. The close agreement between the simulated results and empirical observations demonstrates the model's ability to reliably capture the transmission and control dynamics of anthrax in animal populations. This validation confirms that the proposed approach is not only mathematically sound but also practically relevant for epidemiological analysis. Looking forward, the Genocchi-QLM framework can be extended to incorporate fractional-order dynamics to account for memory effects, stochastic formulations to capture variability in transmission and recovery, and integration with larger datasets for predictive modeling and real-time risk assessment.

Acknowledgements

The authors would like to convey their thanks to the Editor and Reviewers for the helpful comments and suggestions which improved the work.

Funding

This research received no external funding.

Conflicts of Interest

The authors declare no conflicts of interest.

References

[1] Ahmed, S., Jahan, S. and Nisar, K.S. *Hybrid Fibonacci wavelet method to solve fractional-order logistic growth model,* Math. Methods Appl. Sci., 46 (2023), 16218–16231.

[2] Alfwzan, W.F., Abuasbe, K., Raza, A., Rafiq, M., Awadalla, M. and Almulla, M.A. *A non-standard computational method for stochastic anthrax epidemic model,* AIP Adv., 2023, 13.

[3] Bylaiah, S., Shedole, S., Suresh, K.P., Gowda, L., Shivananda, B., Shivamallu, C. and Patil, S.S. *Disease prediction model to assess the impact of changes in precipitation level on the risk of anthrax infectiousness among the livestock hosts in Karnataka,* India, Int. J. Spec. Educ, 37 (2022), 711–727.

[4] Chen, X., Bahl, P., de Silva, C., Heslop, D., Doolan, C., Lim, S. and MacIntyre, C.R. *Systematic review and evaluation of mathematical attack models of human inhalational anthrax for supporting public health decision making and response,* Prehosp. Disaster Med., 35 (2020), 412–419.

[5] Croicu, A.M. *An optimal control model to reduce and eradicate anthrax disease in herbivorous animals,* Bull. Math. Biol., 81 (2019), 235–255.

[6] De Vos, V. and Bryden, H. *Anthrax in the Kruger National Park: temporal and spatial patterns of disease occurrence,* Salisbury Med. Bull., 87 (1996), 26–30.

[7] Dragon, D. and Elkin, B. *An overview of early anthrax outbreaks in northern Canada: Field reports of the Health of Animals Branch, Agriculture Canada, 1962–71*, Arctic, 54 (2001), 32–40.

[8] Ebrahimzadeh, A. and Hashemizadeh, E. *Optimal control of non-linear volterra integral equations with weakly singular Kernels based on Genocchi polynomials and collocation method,* J. Nonlinear Math. Phys., 30, (2023) 1758–1773.

[9] Githire, G.T., Kimathi, G. and Wainaina, M. *Analysis of transmission dynamics of anthrax in animals: A modeling approach,* J. Sci. Res. Rep., 23, (2019) 2550127.

[10] Gutting, B., Channel, S., Berger, A., Gearhart, J., Andrews, G., Sherwood, R. and Nichols, T. *Mathematical modelling of transmission dynamics of anthrax in human and animal population,* Microbe, 3 (2008), 78–85.

[11] Harrington, R., Owen-Smith, N., Viljoen, P.C., Biggs, H.C., Mason, D.R. and Funston, P. *Establishing the causes of the roan antelope decline in the Kruger National Park,* South Africa, Biol. Conserv., 90 (1999), 69–78.

[12] Helikumi, M., Lolika, P.O. and Mushayabasa, S. *Implications of seasonal variations, host and vector migration on spatial spread of sleeping sickness: Insights from a mathematical model,* Informatics Med. Unlocked, 24 (2021), 100570.

[13] Idowu, O.K. and Loyinmi, A.C. *Qualitative analysis of the transmission dynamics and optimal control of Covid-19,* Educatum J. Sci. Math. Technol., 10 (2023), 54–70.

[14] Isah, A. and Ibrahim, S. *Shifted Genocchi polynomials operational matrix for solving fractional order system,* Eurasian J. Sci. Eng., 7 (2021), 74–83.

[15] Izadi, M., Ansari, K.J. and Srivastava, H.M. *A highly accurate and efficient Genocchi-based spectral technique applied to singular fractional order boundary value problems,* Math. Methods Appl. Sci., 2024, 48, 905–925.

[16] Izadi, M. and Atangana, A. *Computational analysis of a class of singular nonlinear fractional multi-order heat conduction model of the human head,* Sci. Rep., 14 (2024), 3466.

[17] Izadi, M. and Srivastava, H.M. *Applications of modified Bessel polynomials to solve a nonlinear chaotic fractional-order system in the Financial Market: Domain-splitting collocation techniques,* Computation, 11 (2023), 130.

[18] Izadi, M., Veeresha, P. and Adel, W. *Mathematical and computational analysis of a fractional-order drug abuse model with nonlinear incidence and logistic growth,* Math. Comput. Simul., 240 (2025), 451–472.

[19] Kanankege, K.S.T., Abdrakhmanov, K.S., Alvarez, J., Glaser, L., Bender, J.B., Mukhanbetkaliyev, Y.Y., Korennoy, F.I., Kadyrov, A.S., Abdrakhmanova, A.S. and Perez, A.M. *Comparison of spatiotemporal patterns of historic natural anthrax outbreaks in Minnesota and Kazakhstan,* PLoS One, 14 (2019), e0217144.

[20] Kar, T. and Batabyal, A. *Stability analysis and optimal control of an SIR epidemic model with vaccination,* Biosystems, 104 (2011), 127–135.

[21] Karginov, V.A., Robinson, T.M., Riemenschneider, J., Golding, B., Kennedy, M., Shiloach, J. and Alibek, K. *Treatment of anthrax infection with combination of ciprofloxacin and*

antibodies to protective antigen of Bacillus anthracis, FEMS Immunol. Med. Microbiol., 40 (2004), 71–74.

[22] Lembo, T., Hampson, K., Auty, H., Beesley, C.A., Bessell, P., Packer, C., Halliday, J., Fyumagwa, R., Hoare, R., Ernest, E., Mentzel, C., Mlengeya, T., Stamey, K., Wilkins, P.P. and Cleaveland, S. *Serologic surveillance of anthrax in the serengeti ecosystem, Tanzania, 1996-2009,* Emerg. Infect. Dis., 17 (2011), 387–394.

[23] Mackey, C. and Kribs, C. *Can scavengers save zebras from anthrax? A modeling study,* Infect. Dis. Model., 6 (2021), 56–74.

[24] Mushayabasa, S. *Global stability of an anthrax model with environmental decontamination and time delay,* Discrete Dyn. Nat. Soc., 2015 (2015), 573146.

[25] Mushayabasa, S., Marijani, T. and Masocha, M. *Dynamical analysis and control strategies in modeling anthrax,* Comput. Appl. Math., 36 (2017), 1333–1348.

[26] Mwakapeje, E.R., Høgset, S., Fyumagwa, R., Nonga, H.E., Mdegela, R.H. and Skjerve, E. *Anthrax outbreaks in the humans-livestock and wildlife interface areas of Northern Tanzania: a retrospective record review, 2006–2016,* BMC Public Health, 18 (2018), 106.

[27] Oh, S.Y., Chateau, A., Tomatsidou, A., Elli, D., Gula, H., Schneewind, O. and Missiakas, D. *Modeling gastrointestinal anthrax disease* Res. Microbiol., 174 (2023), 104026.

[28] Osman, S., Makinde, O.D. and Theuri, D.M. *Mathematical modelling of transmission dynamics of anthrax in human and animal population,* Math. Theory Model., 8 (2018) 47–67.

[29] Pantha, B., Day, J. and Lenhart, S. *Optimal control applied in an anthrax epizootic model,* J. Biol. Syst., 24 (2016), 495–517.

[30] Raymond, B., Wyres, K.L., Sheppard, S.K., Ellis, R.J., Bonsall, M.B. and Guttman, D.S. *Environmental factors determining the epidemiology and population genetic structure of the Bacillus cereus group in the field,* PLoS Pathog., 6 (2010), e1000905.

[31] Raza, A. and Abdella, K. *Analysis of the dynamics of anthrax epidemic model with delay,* Discover Appl. Sci., 6 (2024), 128.

[32] Saad-Roy, C.M., van den Driessche, P. and Yakubu, A-A. *A mathematical model of anthrax transmission in animal populations,* Bull. Math. Biol., 79 (2016), 303–324.

[33] Sabir, Z. and Simbawa, E.*A machine learning computational approach for the mathematical anthrax disease system in animals,* PLoS One, 20 (2025), e0320327.

[34] Savransky, V., Ionin, B. and Reece, J. *Current status and trends in prophylaxis and management of anthrax disease,* Pathogens, 9 (2020), 370.

[35] Stella, E., Mari, L., Gabrieli, J., Barbante, C. and Bertuzzo, E. *Permafrost dynamics and the risk of anthrax transmission: A modelling study,* Sci. Rep., 10 (2020), 16460.

[36] Turner, W.C., Imologhome, P., Havarua, Z., Kaaya, G.P., Mfune, E., Mpofu, I.D.T. and Getz, W.M. *Soil ingestion, nutrition and the seasonality of anthrax in herbivores of Etosha National Park,* Ecosphere, 4 (2013), 1–19.

[37] Van den Driessche, P. *Reproduction numbers of infectious disease models,* Infect. Dis. Model., 2 (2017), 288–303.

[38] Yüzbaşı, Ş. and Izadi, M. *Bessel-quasilinearization technique to solve the fractional-order HIV-1 infection of CD4+ T-cells considering the impact of antiviral drug treatment,* Appl. Math. Comput., 431 (2022), 127319.

Iranian Journal of Numerical Analysis and Optimization
Vol. 16, No. 3, 2026, pp 1161–1188
https://doi.org/10.22067/ijnao.2026.97142.1813
https://ijnao.um.ac.ir/

CC BY NC

Research Article

Analysis of the mathematical model's stability describing the prevalence of obesity and its role in diabetes

A. El Mansouri*, M. Belam and A. Labzai

Abstract

We propose a continuous-time dynamic epidemiological model that describes the prevalence of obesity in a population and examines the relationship between weight gain and the development of diabetes, particularly type 2 diabetes. We investigate both the local and global stability of the disease-free and endemic equilibrium points using the Routh–Hurwitz criteria and the construction of appropriate Lyapunov functions. Furthermore, a sensitivity analysis of the

*Corresponding author

Received 24 December 2025; revised 9 May 2026; accepted 10 May 2026

Abdelbar El Mansouri
Department of Mathematics and Computer Science, Laboratory of Advanced Research in Industrial and Logistic Engineering (LARILE), Applied Mathematics Team, ENSEM, Hassan II University of Casablanca, Morocco.
e-mail: abdelbar.elmansouri@gmail.com

Mohamed Belam
Department of Mathematics and Computer Science, Laboratory of Advanced Research in Industrial and Logistic Engineering (LARILE), Applied Mathematics Team, ENSEM, Hassan II University of Casablanca, Morocco.
e-mail: m.belam@gmail.com

Abderrahim Labzai
Department of Mathematics and Computer Science, Laboratory of Advanced Research in Industrial and Logistic Engineering (LARILE), Applied Mathematics Team, ENSEM, Hassan II University of Casablanca, Morocco.
e-mail: labzaiabdo1977@gmail.com

How to cite this article
El Mansouri, A., Belam, M. and Labzai, A., Analysis of the mathematical model's stability describing the prevalence of obesity and its role in diabetes. *Iran. J. Numer. Anal. Optim.*, 2026; 16(3): 1161-1188. https://doi.org/10.22067/ijnao.2026.97142.1813

basic reproduction number R_0 is carried out to determine the relative impact of each model parameter. Numerical simulations performed using MATLAB are presented to illustrate the dynamical behavior of the proposed model. The simulation results are consistent with and support the theoretical findings.

AMS subject classifications (2020): Primary 92D30; Secondary 34A34, 92C50, 37N25.

Keywords: Mathematical model; Stability; Sensitivity; Reproduction number; Asymptotically stable.

1 Introduction

An abnormal or excessive accumulation of body fat that has a detrimental effect on health is referred to as overweight or obesity [7]. This accumulation occurs when a person consumes more dietary energy than their body requires [41]. A common tool used to diagnose obesity is the body mass index (BMI), which is calculated using an individual's height and weight. According to the values of this index (expressed in kilograms per square meter), the World Health Organization (WHO) has established thresholds for overweight, obesity, and morbid obesity: From 25 to 30, from 30 to 40, and above 40, respectively [47].

Numerous factors, including genetic predisposition [43], unhealthy dietary habits, and sedentary lifestyles [37], may contribute to obesity. According to the WHO, approximately 13% of the global population is obese [47]. A report by this organization [48] indicates that obesity has reached nearly pandemic proportions in Europe, with 59% of adults and 8% of children under five years old being overweight. Moreover, one in three school-aged children is affected by this condition. The same study [48] also revealed a link between obesity and a variety of diseases, such as diabetes, cardiovascular diseases, and at least 13 different types of cancer, with an estimated 200,000 new cases each year. Obesity can also lead to disability and may result in premature death.

Diabetes is a chronic metabolic disease characterized by high blood glucose levels and is primarily associated with aging and obesity [2]. Approximately 422 million people worldwide live with diabetes, and about one in ten adults in the UK has type 2 diabetes (T2D) [32, 45]. According to the WHO [44], between 2000 and 2019 the number of premature deaths related to diabetes increased by 3%. In particular, obesity represents one of the most significant risk factors for T2D [35, 2]. A recent US Census report and estimates from the CDC indicate that about 11% of Americans have diabetes and approximately 14% are obese [28, 27].

Our study focuses exclusively on **T2D**, which is the most prevalent form globally, accounting for approximately 90–95% of all diabetes cases [46, 3]. Type 1 and gestational diabetes are much less frequent, and their etiology and risk factors differ significantly. Therefore, T2D is the most relevant for modeling the link with obesity.

A substantial body of clinical and epidemiological evidence has established a strong relationship between obesity and T2D. Excess body fat, particularly visceral adiposity, plays a central role in the development of insulin resistance, which is one of the primary mechanisms underlying T2D [17]. Adipose tissue is an active endocrine organ that secretes adipokines and pro-inflammatory cytokines such as TNF-α and IL-6. These mediators contribute to chronic low-grade inflammation and interfere with insulin signaling pathways in peripheral tissues.

Obesity is also closely associated with metabolic syndrome, a cluster of metabolic abnormalities including abdominal obesity, dyslipidemia, hypertension, and impaired glucose metabolism, which significantly increases the risk of developing T2D [14]. Behavioral and environmental factors, such as unhealthy diets, physical inactivity, and sedentary lifestyles, further exacerbate these metabolic disturbances and contribute to the growing prevalence of both obesity and diabetes worldwide. The main biological pathway linking obesity to T2D is summarized in Figure 1.

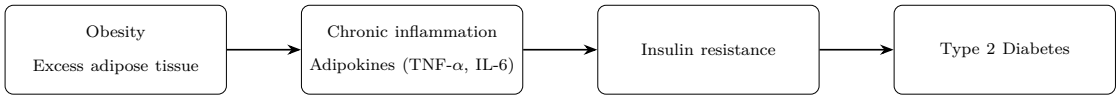


Figure 1: Conceptual biological pathway linking obesity to T2D. Excess adipose tissue promotes chronic low-grade inflammation and the secretion of adipokines and pro-inflammatory cytokines. These processes impair insulin signaling, leading to insulin resistance and ultimately the development of T2D.

Numerous academic studies have investigated the dynamics of various epidemiological phenomena in the context of combating different diseases (such as alcoholism, smoking, COVID-19, and monkeypox) [13, 25, 26, 9, 33, 12, 18, 22, 5, 21, 38, 39, 42, 15, 6, 24, 31, 8, 36, 11].

Since obesity can spread through social interactions and behavioral influence [4], the dynamics of this condition have been analyzed using several epidemiological models, for instance [16, 34, 19, 1, 9]. The authors in [16] investigated childhood obesity and, through numerical simulations, showed that eating habits are a major contributing factor. By performing a numerical covariate analysis of an SIS epidemic model, the authors in [34] demonstrated that preventive measures become more effective when adult obesity is considered. Similar results were obtained by the authors in [19] using an SIS difference equation model. However, these studies considered continuous control strategies, which may be inefficient or impractical in certain contexts. In [1], a deterministic model incorporating time-dependent control strategies was proposed, including dietary programs, healthy living campaigns, and therapeutic interventions. In our previous work

[9], we studied a discrete mathematical model for the prevalence of obesity and applied four control strategies to reduce its spread: Educational programs, dietary and exercise interventions, drug-based medical treatment, and surgical intervention. The optimal control strategies were determined using the discrete-time Pontryagin maximum principle.

In previous work [10], we also investigated the strong relationship between obesity and diabetes using a continuous mathematical model, in which optimal strategies were proposed to limit their spread.

Building upon previous models, our proposed continuous-time model synthesizes key elements from both works:

- From [9], we adopt the modeling framework and the control strategies for obesity prevalence.
- From [20], we incorporate components describing the dynamics of T2D in the population.

Importantly, our model extends these previous works [9], [20], by integrating the dynamics of obesity and T2D simultaneously in a unified continuous-time framework, allowing for joint sensitivity analysis and optimal control strategies that target both conditions. This novel contribution provides a more comprehensive understanding of the interaction between obesity and diabetes, which was not addressed in the previous studies.

In the present article, we introduce the following contributions:

- We employ the next-generation matrix method to determine the basic reproduction number (R_0).
- Local and global stability analyses are performed.
- A sensitivity analysis is conducted to determine the relative importance of the model parameters in the transmission dynamics of these two diseases.

The remainder of this paper is organized as follows: Section 2 presents the formulation of the mathematical model. Section 3 is devoted to the computation of the basic reproduction number. Section 4 deals with the stability analysis of the model. Section 5 presents the numerical simulations and discussion of the results. Section 6 outlines the perspectives and possible extensions of the model. Finally, Section 7 concludes the paper.

2 The mathematical model and its fundamental characteristics

2.1 The mathematical model

We propose an $SWRIEDC$ model that highlights the detrimental role of overweight in the development of diabetes, particularly T2D, and describes the dynamics of obesity and diabetes epidemics within a given population. The total population is divided into seven compartments. A schematic representation of the proposed model is illustrated in Figure 2.

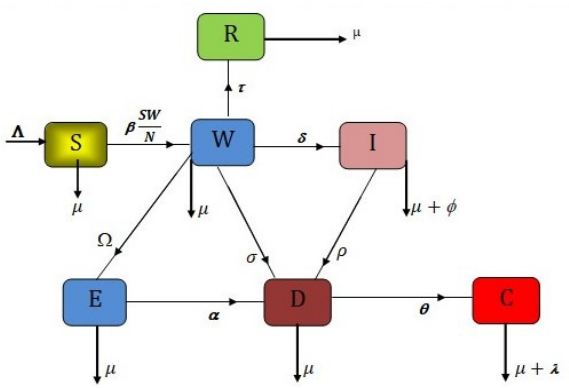


Figure 2: Flowchart of the proposed model.

Table 1 describes the different population classes considered at time t.

Table 1: Definitions of the state variables

Variable	Meaning
$S(t)$	Individuals susceptible to obesity
$W(t)$	Overweight or moderately obese individuals
$I(t)$	Obese individuals
$R(t)$	Recovered individuals
$E(t)$	Pre-diabetic individuals
$D(t)$	Diabetic individuals without complications
$C(t)$	Diabetic individuals with complications

Let

$$N(t) = S(t) + W(t) + I(t) + R(t) + E(t) + D(t) + C(t)$$

denote the total population at time t.

The dynamics of the model are described by the following system of seven nonlinear differential equations:

$$\begin{cases} \frac{dS(t)}{dt} = \Lambda - \mu S(t) - \beta \frac{S(t)}{N} W(t), \\ \frac{dW(t)}{dt} = \beta \frac{S(t)}{N} W(t) - (\tau + \sigma + \mu + \delta + \Omega) W(t), \\ \frac{dR(t)}{dt} = \tau W(t) - \mu R(t), \\ \frac{dI(t)}{dt} = \delta W(t) - (\phi + \rho + \mu) I(t), \\ \frac{dE(t)}{dt} = \Omega W(t) - (\mu + \alpha) E(t), \\ \frac{dD(t)}{dt} = \sigma W(t) + \rho I(t) + \alpha E(t) - (\mu + \theta) D(t), \\ \frac{dC(t)}{dt} = \theta D(t) - (\mu + \lambda) C(t), \end{cases} \tag{1}$$

subject to the initial conditions

$$S(0), W(0), R(0), I(0), E(0), D(0), C(0) > 0.$$

The model parameters are described in Table 2.

Table 2: Definitions and values of the parameters

Parameter	Definition	Value	Reference
Λ	Recruitment rate of individuals susceptible to obesity	4×10^6	Assumed
μ	Natural death rate	0.02	[9]
β	Rate at which susceptible individuals become overweight due to social influence	0.45	[9]
τ	Recovery rate of overweight individuals	0.07	Assumed
δ	Rate at which overweight individuals become obese	0.05	[9]
λ	Mortality rate due to diabetes complications	0.001	[20]
ϕ	Obesity-related mortality rate	0.03	Assumed
ρ	Rate at which obese individuals develop diabetes without complications	0.06	Assumed
Ω	Rate at which overweight individuals become pre-diabetic	0.03	Assumed
α	Rate at which pre-diabetic individuals develop diabetes without complications	0.06	[20]
σ	Rate at which overweight individuals develop diabetes without complications	0.025	Assumed
θ	Rate at which diabetic individuals develop complications	0.08	[20]

2.2 Fundamental properties of the model

2.2.1 Invariant region

It must be shown that all solutions of system (1) with positive initial conditions remain positive for all $t > 0$. This result is established by the following lemma.

Lemma 1. All feasible solutions $S(t), W(t), R(t), I(t), E(t), D(t)$, and $C(t)$ of system (1) are bounded in the region

$$\Delta = \left\{ (S, W, R, I, E, D, C) \in \mathbb{R}_+^7 : S + W + R + I + E + D + C \leq \frac{\Lambda}{\mu} \right\}. \tag{2}$$

Proof. From system (1), we have

$$\frac{dN}{dt} = \frac{dS}{dt} + \frac{dW}{dt} + \frac{dR}{dt} + \frac{dI}{dt} + \frac{dE}{dt} + \frac{dD}{dt} + \frac{dC}{dt},$$

which yields

$$\frac{dN(t)}{dt} = \Lambda - \mu N(t) - \phi I(t) - \lambda C(t).$$

Hence,

$$\frac{dN(t)}{dt} \leq \Lambda - \mu N(t).$$

Solving this differential inequality gives

$$N(t) \leq \frac{\Lambda}{\mu} + N(0)e^{-\mu t}.$$

Taking the limit as $t \to +\infty$ yields

$$\limsup_{t \to +\infty} N(t) \leq \frac{\Lambda}{\mu}.$$

Therefore,

$$S(t) + W(t) + R(t) + I(t) + E(t) + D(t) + C(t) \leq \frac{\Lambda}{\mu}.$$

Thus, the region Δ defined in (2) is positively invariant for system (1). Consequently, it is sufficient to study the dynamics of system (1) in the region Δ of nonnegative solutions.

□

2.2.2 Positivity of solutions

We establish the following theorem.

Theorem 1. If $S(0), W(0), R(0), I(0), E(0), D(0)$, and $C(0)$ are positive, then the solutions $S(t), W(t), R(t), I(t), E(t), D(t)$, and $C(t)$ of system (1) remain nonnegative for all $t > 0$.

Proof. From the first equation of system (1), we have

$$\frac{dS(t)}{dt} = \Lambda - A(t)S(t),$$

where

$$A(t) = \beta \frac{W(t)}{N} + \mu.$$

Multiplying both sides by $\exp\left(\int_0^t A(s)ds\right)$ gives

$$\frac{d}{dt}\left[S(t)\exp\left(\int_0^t A(s)ds\right)\right] = \Lambda \exp\left(\int_0^t A(s)ds\right).$$

Integrating from 0 to t yields

$$S(t) = S(0)e^{-\int_0^t A(s)ds} + \Lambda e^{-\int_0^t A(s)ds}\int_0^t e^{\int_0^w A(s)ds}dw \geq 0.$$

Thus, $S(t) \geq 0$.

Similarly, using the remaining equations of system (1), we obtain

$$W(t) \geq W(0)\exp\left(-\int_0^t (\beta\frac{S(s)}{N} - (\mu + \tau + \sigma + \delta + \Omega))ds\right) \geq 0, \tag{3}$$

$$R(t) \geq R(0)e^{-\mu t} \geq 0, \tag{4}$$

$$I(t) \geq I(0)e^{-(\mu+\phi+\rho)t} \geq 0, \tag{5}$$

$$E(t) \geq E(0)e^{-(\mu+\alpha)t} \geq 0, \tag{6}$$

$$D(t) \geq D(0)e^{-(\mu+\theta)t} \geq 0,$$

$$C(t) \geq C(0)e^{-(\mu+\lambda)t} \geq 0.$$

Therefore, all solutions of system (1) remain positive for all $t \geq 0$.

□

3 Model stability and sensitivity analysis

3.1 Disease-free equilibrium point

The disease-free equilibrium (DFE) of system (1), corresponding to the absence of obesity in the population, is easily obtained. At this equilibrium we have

$$W^0 = 0, \quad R^0 = 0, \quad I^0 = 0, \quad E^0 = 0, \quad D^0 = 0, \quad C^0 = 0.$$

From the first equation of system (1), we obtain

$$\frac{dS(t)}{dt}(S^0, 0, 0, 0, 0, 0, 0) = \Lambda - \mu S^0 = 0, \tag{7}$$

which yields

$$S^0 = \frac{\Lambda}{\mu}. \tag{8}$$

Therefore, the DFE point is given by

$$E^0_{eq} = \left(\frac{\Lambda}{\mu}, 0, 0, 0, 0, 0, 0\right). \tag{9}$$

3.2 The basic reproduction number (R_0)

The basic reproduction number is a threshold parameter that governs the overall dynamics of the disease. It represents the average number of secondary infections produced by a single infected individual during its entire infectious period in a fully susceptible population.

To compute R_0, we use the next generation matrix method [40, 29, 30].

Considering the infected compartments, system (1) can be rewritten as

$$\begin{cases} \frac{dW(t)}{dt} = \beta\frac{S(t)}{N}W(t) - (\tau+\sigma+\mu+\delta+\Omega)W(t) = \beta\frac{S(t)}{N}W(t) - K_1W(t), \\ \frac{dR(t)}{dt} = \tau W(t) - \mu R(t), \\ \frac{dI(t)}{dt} = \delta W(t) - (\phi+\rho+\mu)I(t) = \delta W(t) - K_2I(t), \\ \frac{dE(t)}{dt} = \Omega W(t) - (\mu+\alpha)E(t) = \Omega W(t) - K_3E(t), \\ \frac{dD(t)}{dt} = \sigma W(t) + \rho I(t) + \alpha E(t) - (\mu+\theta)D(t) = \sigma W(t) + \rho I(t) + \alpha E(t) - K_4D(t), \\ \frac{dC(t)}{dt} = \theta D(t) - (\mu+\lambda)C(t) = \theta D(t) - K_5C(t), \end{cases} \tag{10}$$

where

$$K_1 = \tau+\sigma+\mu+\delta+\Omega, \quad K_2 = \phi+\rho+\mu, \quad K_3 = \mu+\alpha, \quad K_4 = \mu+\theta, \quad K_5 = \mu+\lambda. \tag{11}$$

The system can be written in matrix form as

$$\frac{dX}{dt} = F(X) - V(X),$$

where F represents the new infection terms and V represents the transition terms.

At the DFE E^0_{eq} we have $S = \frac{\Lambda}{\mu}$. The matrices F and V are therefore given by

$$F = \begin{bmatrix} \beta\frac{\Lambda}{N\mu} & 0 & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & 0 & 0 \end{bmatrix}$$

and

$$V = \begin{bmatrix} K_1 & 0 & 0 & 0 & 0 & 0 \\ -\tau & \mu & 0 & 0 & 0 & 0 \\ -\delta & 0 & K_2 & 0 & 0 & 0 \\ -\Omega & 0 & 0 & K_3 & 0 & 0 \\ -\sigma & 0 & -\rho & -\alpha & K_4 & 0 \\ 0 & 0 & 0 & 0 & -\theta & K_5 \end{bmatrix}.$$

The next-generation matrix is given by FV^{-1}. The spectral radius of this matrix yields the basic reproduction number. At the DFE we have $N_0 = \frac{\Lambda}{\mu}$. Substituting this value into the expression of R_0 gives

$$R_0 = \frac{\beta\Lambda}{N_0\mu(\tau+\sigma+\mu+\delta+\Omega)} = \frac{\beta}{\tau+\sigma+\mu+\delta+\Omega} = \frac{\beta}{K_1}. \tag{12}$$

3.3 Endemic equilibrium point

The endemic equilibrium represents a steady state in which the disease persists in the population.

Let

$$E^*_{eq} = (S^*, W^*, R^*, I^*, E^*, D^*, C^*)$$

denote the endemic equilibrium point obtained by setting the right-hand side of system (1) equal to zero.

The equilibrium values are given by

$$\begin{cases} S^* &= \frac{NK_1}{\beta}, \\ W^* &= \frac{N\mu}{\beta}(R_0-1), \\ R^* &= \frac{N\tau}{\beta}(R_0-1), \\ I^* &= \frac{N\delta\mu}{\beta K_2}(R_0-1), \\ E^* &= \frac{N\Omega\mu}{\beta K_3}(R_0-1), \\ D^* &= \frac{N\mu}{\beta K_4}\left(\sigma+\frac{\delta\rho}{K_2}+\frac{\alpha\Omega}{K_3}\right)(R_0-1), \\ C^* &= \frac{\theta N\mu}{\beta K_4 K_5}\left(\sigma+\frac{\delta\rho}{K_2}+\frac{\alpha\Omega}{K_3}\right)(R_0-1). \end{cases} \tag{13}$$

3.4 Local stability analysis

We analyze the local stability of the equilibrium points E^0_{eq} and E^*_{eq}.

3.4.1 Local stability of the DFE

Theorem 2. The DFE E^0_{eq} is locally asymptotically stable if $R_0 < 1$ and unstable if $R_0 > 1$.

Proof. The Jacobian matrix of system (1) is given by

$$J = \begin{bmatrix} -\beta\frac{W}{N} - \mu & -\beta\frac{S}{N} & 0 & 0 & 0 & 0 & 0 \\ \beta\frac{W}{N} & \beta\frac{S}{N} - K_1 & 0 & 0 & 0 & 0 & 0 \\ 0 & \tau & -\mu & 0 & 0 & 0 & 0 \\ 0 & \delta & 0 & -K_2 & 0 & 0 & 0 \\ 0 & \Omega & 0 & 0 & -K_3 & 0 & 0 \\ 0 & \sigma & 0 & \rho & \alpha & -K_4 & 0 \\ 0 & 0 & 0 & 0 & 0 & \theta & -K_5 \end{bmatrix}.$$

Evaluating this matrix at E_{eq}^0 yields

$$J(E_{eq}^0) = \begin{bmatrix} -\mu & -\beta\frac{S^0}{N} & 0 & 0 & 0 & 0 & 0 \\ 0 & \beta\frac{S^0}{N} - K_1 & 0 & 0 & 0 & 0 & 0 \\ 0 & \tau & -\mu & 0 & 0 & 0 & 0 \\ 0 & \delta & 0 & -K_2 & 0 & 0 & 0 \\ 0 & \Omega & 0 & 0 & -K_3 & 0 & 0 \\ 0 & \sigma & 0 & \rho & \alpha & -K_4 & 0 \\ 0 & 0 & 0 & 0 & 0 & \theta & -K_5 \end{bmatrix}.$$

The eigenvalues are

$$-\mu, \quad K_1(R_0 - 1), \quad -K_2, \quad -K_3, \quad -K_4, \quad -K_5.$$

Thus, if $R_0 < 1$, all eigenvalues have negative real parts. Therefore, the DFE is locally asymptotically stable. Conversely, if $R_0 > 1$, the equilibrium is unstable.

□

3.4.2 Bifurcation behavior at $R_0 = 1$

The change in stability of the DFE at $R_0 = 1$ indicates the occurrence of a transcritical bifurcation. This behavior is consistent with the structure of system (1), where the transmission process is governed by the bilinear incidence term $\beta\frac{S}{N}W$.

In epidemic models with bilinear incidence and without nonlinear feedback mechanisms, the disease dynamics are typically characterized by a single threshold determined by the basic reproduction number R_0. In this case, the DFE is stable when $R_0 < 1$ and loses stability when $R_0 > 1$, while an endemic equilibrium emerges.

From the analytical expressions of the endemic equilibrium given in (13), all endemic state variables are proportional to $(R_0 - 1)$. Therefore, the endemic equilibrium exists only when $R_0 > 1$, which rules out the possibility of backward bifurcation in the present model.

3.4.3 Local stability of the endemic equilibrium

Theorem 3. For system (1), the endemic equilibrium E^*_{eq} is unstable if $R_0 \leq 1$ and locally asymptotically stable if $R_0 > 1$.

The proof of Theorem 3 is provided in Appendix A.

3.5 Global stability analysis

To establish the global asymptotic stability of system (1), we use Lyapunov's method. We first study the global stability of the DFE E^0_{eq}.

3.5.1 Global stability of the DFE

Theorem 4. The DFE E^0_{eq} is globally asymptotically stable if $R_0 \leq 1$ and unstable otherwise.

Proof. Consider the Lyapunov function

$$V(S, W) = \frac{1}{2}\left[(S - S^0) + W\right]^2 + \frac{N}{\beta}(\mu + K_1)W.$$

Differentiating with respect to time yields

$$\frac{dV}{dt} = -\mu(S - S^0)^2 - K_1 W^2 - \frac{N}{\beta}(\mu + K_1)K_1(1 - R_0)W.$$

Thus, $\frac{dV}{dt} \leq 0$ for $R_0 \leq 1$. Moreover, $\frac{dV}{dt} = 0$ if and only if $S = S^0$ and $W = 0$.

Furthermore, when $W = 0$, the remaining state variables satisfy linear differential equations with negative coefficients. Hence $R(t)$, $I(t)$, $E(t)$, $D(t)$ and $C(t)$ converge exponentially to zero. Therefore, the largest invariant set contained in $\{\dot{V} = 0\}$ reduces to the DFE E^0_{eq}.

According to the LaSalle's invariance principle [23], the equilibrium E^0_{eq} is globally asymptotically stable when $R_0 \leq 1$.

□

3.5.2 Global stability of the endemic equilibrium

Theorem 5. If $R_0 > 1$, then the endemic equilibrium E^*_{eq} of system (1) is globally asymptotically stable.

The proof of Theorem 5 is provided in Appendix B.

4 Sensitivity analysis of R_0

Understanding the relative importance of the parameters involved in the spread of obesity and diabetes is essential. In this context, we perform a sensitivity analysis of the basic reproduction number R_0 in order to evaluate how variations in the parameters of the $SWRIEDC$ model affect its value.

This analysis is carried out by computing the sensitivity indices of R_0 with respect to the model parameters. These indices allow us to determine the relative contribution of each parameter to the transmission dynamics of the disease.

Using the approach proposed by Chitnis, Hyman, and Cushing [5], the normalized forward sensitivity index of R_0 with respect to a parameter n is defined as

$$\Upsilon_n^{R_0} = \frac{\partial R_0}{\partial n} \times \frac{n}{R_0}. \tag{14}$$

If $\Upsilon_n^{R_0} < 0$, then the parameter contributes to disease control. Conversely, if $\Upsilon_n^{R_0} > 0$, the parameter contributes to the spread of the disease.

Recall that the basic reproduction number is given by

$$R_0 = \frac{\beta}{K_1}, \tag{15}$$

where

$$K_1 = \mu + \delta + \tau + \sigma + \Omega.$$

Table 3 presents the sensitivity indices corresponding to the parameters β, μ, δ, σ, τ, and Ω.

The result $\Upsilon_\beta^{R_0} = 1$ indicates that increasing (decreasing) the transmission rate β by a given percentage leads to an increase (decrease) of the same percentage in the basic reproduction number R_0.

On the other hand, the parameters μ, σ, δ, τ, and Ω have negative sensitivity indices, which means that increasing their values contributes to reducing the spread of the disease.

Table 3: Sensitivity indices of the basic reproduction number R_0 with respect to selected parameters

Parameter	Sensitivity index
β	$\Upsilon_{\beta}^{R_0} = 1 > 0$
μ	$\Upsilon_{\mu}^{R_0} = -1 - \frac{\mu}{K_1} < 0$
δ	$\Upsilon_{\delta}^{R_0} = -\frac{\delta}{K_1} < 0$
σ	$\Upsilon_{\sigma}^{R_0} = -\frac{\sigma}{K_1} < 0$
τ	$\Upsilon_{\tau}^{R_0} = -\frac{\tau}{K_1} < 0$
Ω	$\Upsilon_{\Omega}^{R_0} = -\frac{\Omega}{K_1} < 0$

Therefore, strengthening prevention and intervention strategies associated with these parameters may significantly reduce the transmission dynamics of obesity and diabetes.

5 Numerical simulations

In this section, we present numerical simulations of system (1) using different initial conditions. We assume that at each time t the total population satisfies

$$S(t) + W(t) + R(t) + I(t) + E(t) + D(t) + C(t) = N.$$

The parameter values used in the simulations are listed in Table 2. For illustration purposes, we consider

$$N = 1.16 \times 10^7.$$

Case $R_0 < 1$

For the DFE we have

$$E_{eq}^0 = (S^0, W^0, R^0, I^0, E^0, D^0, C^0) = (2 \times 10^8, 0, 0, 0, 0, 0, 0)$$

with

$$R_0 = 0.88 < 1.$$

Starting from different initial values of the state variables, Theorem 4 implies that the DFE E_{eq}^0 of system (1) is globally asymptotically stable in the feasible region Δ(2).

The following curves are obtained for different initial values $S_0, W_0, R_0, I_0, E_0, D_0$, and C_0 (Figures 3–9).

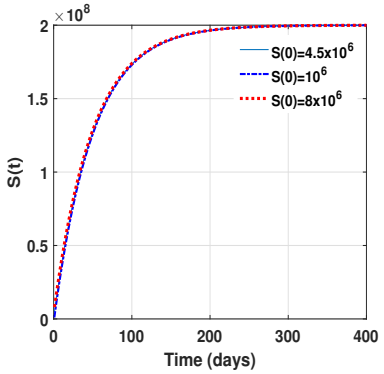


Figure 3: Evolution of $S(t)$ for $R_0 = 0.88$

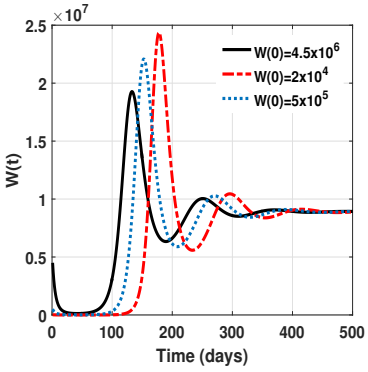


Figure 6: Evolution of $W(t)$ for $R_0 = 0.88$

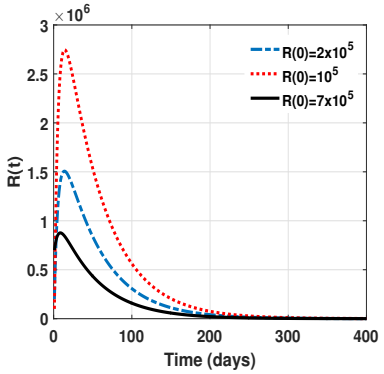


Figure 4: Evolution of $R(t)$ for $R_0 = 0.88$

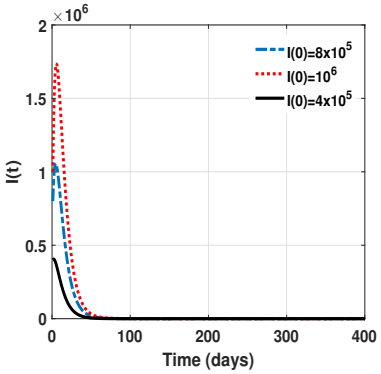


Figure 7: Evolution of $I(t)$ for $R_0 = 0.88$

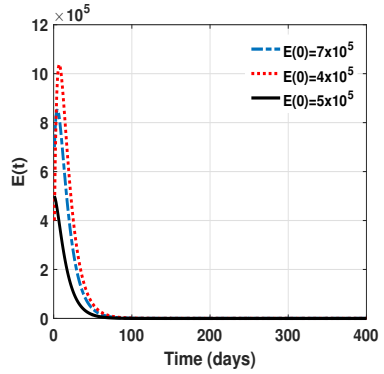


Figure 5: Evolution of $E(t)$ for $R_0 = 0.88$

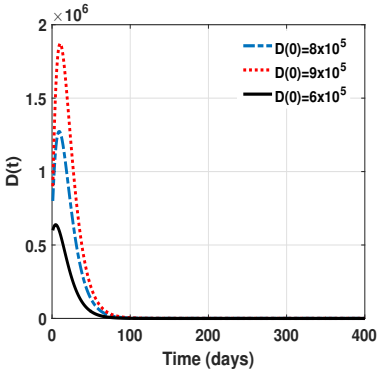


Figure 8: Evolution of $D(t)$ for $R_0 = 0.88$

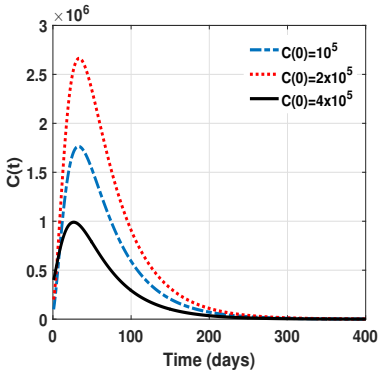


Figure 9: Evolution of $C(t)$ for $R_0 = 0.88$

Figures 3–9 show that the number of susceptible individuals increases and approaches approximately $S_0 = 2 \times 10^8$ (Figure 3). In addition, the number of individuals exposed to diabetes (E) decreases and approaches zero (Figure 5). Similarly, the number of overweight individuals (W), obese individuals (I), diabetic individuals without complications (D), and diabetic individuals with complications (C) gradually decreases and converges to zero (Figures 6–9).

The number of recovered individuals (R) initially increases, then decreases and eventually approaches zero (Figure 4). Therefore, when $R_0 = 0.88 < 1$, the solution converges to the DFE point

$$E^0_{eq} = (S_0, 0, 0, 0, 0, 0, 0).$$

Consequently, system (1) is globally asymptotically stable in this case.

Case $R_0 > 1$

Keeping the same parameter values given in Table 2, we obtain the endemic equilibrium

$$E^*_{eq} = (S^*, W^*, R^*, I^*, E^*, D^*, C^*)$$

with

$$(1.13 \times 10^8, 8.9 \times 10^6, 3.12 \times 10^7, 4 \times 10^6, 3.34 \times 10^6, 6.67 \times 10^6, 2.53 \times 10^7)$$

and

$$R_0 = 1.77 > 1.$$

According to Theorem 5, the endemic equilibrium E^*_{eq} of system (1) is globally asymptotically stable in the feasible region $\Delta(2)$. The corresponding numerical simulations are shown in Figures 10–16.

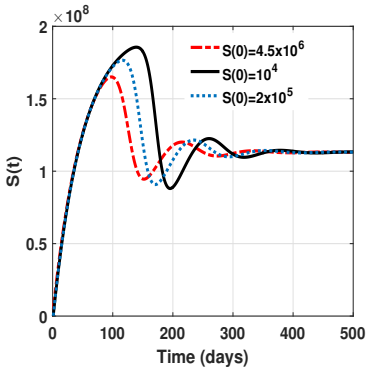


Figure 10: Evolution of $S(t)$ for $R_0 = 1.77$

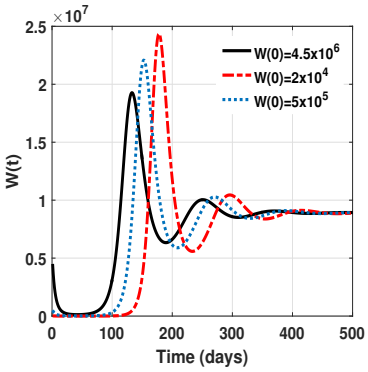


Figure 13: Evolution of $W(t)$ for $R_0 = 1.77$

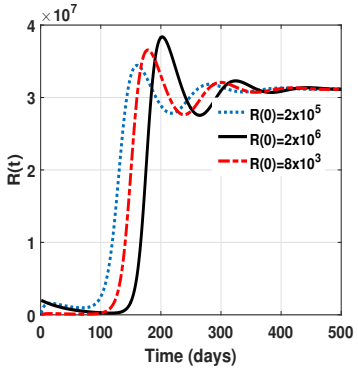


Figure 11: Evolution of $R(t)$ for $R_0 = 1.77$

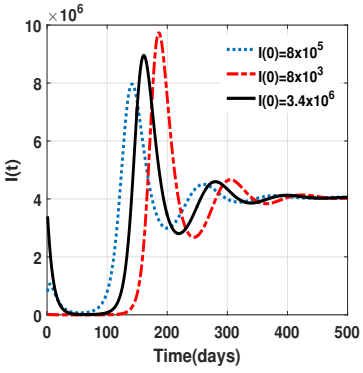


Figure 14: Evolution of $I(t)$ for $R_0 = 1.77$

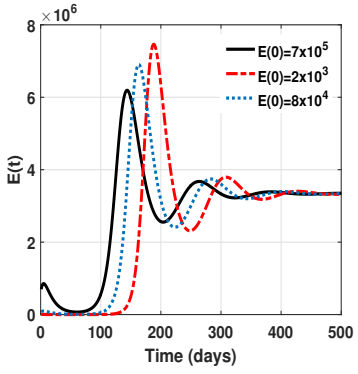


Figure 12: Evolution of $E(t)$ for $R_0 = 1.77$

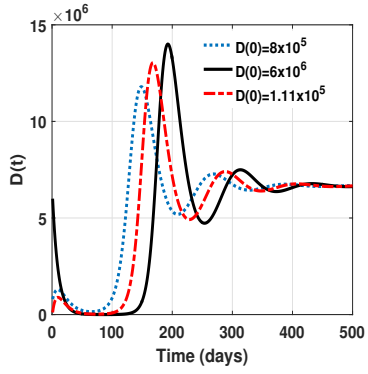


Figure 15: Evolution of $D(t)$ for $R_0 = 1.77$

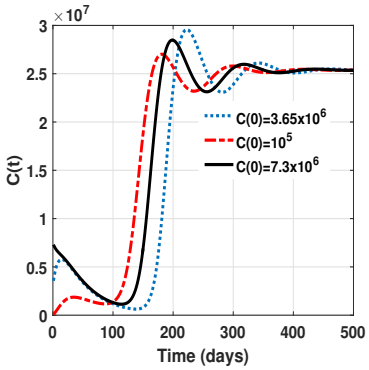


Figure 16: Evolution of $C(t)$ for $R_0 = 1.77$

From these figures we observe that the number of susceptible individuals initially increases and then gradually decreases until it stabilizes at approximately

$$S^* \approx 1 \times 10^8.$$

The number of overweight (W) and obese (I) individuals increases initially, reaches a peak after approximately five months, and then stabilizes after about one year at

$$W^* \approx 9 \times 10^6, \qquad I^* \approx 4 \times 10^6.$$

The number of individuals exposed to diabetes (E) and diabetic patients without complications (D) follows a similar trend: The number remains relatively low during the first 100 days, then increases rapidly before stabilizing after about one year at

$$E^* \approx 3 \times 10^6, \qquad D^* \approx 6.6 \times 10^6.$$

According to Figure 16, the number of diabetic individuals with complications (C) increases after approximately 100 days, reaches a maximum around day 200, and then gradually decreases until stabilizing at

$$C^* \approx 2.5 \times 10^7$$

around day 300.

Finally, the number of recovered individuals (R) initially increases rapidly to a maximum value and then decreases slightly before stabilizing after approximately one year at

$$R^* \approx 3 \times 10^7.$$

Therefore, when $R_0 > 1$, the solution converges to the endemic equilibrium

$$E^*_{eq}(S^*, W^*, R^*, I^*, E^*, D^*, C^*).$$

Hence, the proposed model is globally asymptotically stable in the endemic case.

6 Perspectives and future work

While the present model provides valuable insights into the interaction between obesity and T2D, several limitations should be noted:

- The model focuses on T2D and does not include type 1 or gestational diabetes.
- It does not account for the fact that some obese individuals may never develop diabetes, nor that some nonobese individuals may develop it.
- Age, demographic, and other individual-specific risk factors are not explicitly incorporated.
- Intermediate conditions such as metabolic syndrome are not included.

Addressing these limitations will be the focus of future research. Possible extensions include as follows:

- Expanding the model to include additional compartments representing different risk profiles.
- Incorporating age, gender, and other demographic factors to improve realism.
- Including metabolic syndrome and other intermediate risk factors for both obesity and T2D.
- Exploring interactions between T2D and other chronic diseases for more comprehensive public health insights.

These extensions aim to enhance the model's applicability and provide a more detailed understanding of population-level dynamics and targeted intervention strategies.

7 Conclusion

In this paper, we have studied a continuous mathematical model, $SWRIEDC$, for obesity-related diabetes, which describes the dynamics of individuals affected by these two diseases. We derived the basic reproduction number of the system (1) as

$$R_0 = \frac{\Lambda\beta}{\mu N(\mu+\delta+\tau+\sigma+\Omega)}.$$

To identify the parameters that most significantly influence R_0, we performed a sensitivity analysis. Using the theory of nonlinear system stability, we analyzed the local and global behavior of the model. The DFE E^0_{eq} is locally asymptotically stable if $R_0 \leq 1$, whereas the endemic equilibrium E^*_{eq} is locally asymptotically stable when $R_0 > 1$.

Furthermore, by constructing appropriate Lyapunov functions, we showed that E^0_{eq} is globally asymptotically stable for $R_0 \leq 1$, and E^*_{eq} is globally asymptotically stable for $R_0 > 1$. These results highlight the critical role of the reproduction number in controlling the spread of obesity-related diabetes and provide a theoretical foundation for designing effective prevention and intervention strategies.

A Proof of Theorem 3

Proof. Determine the local stability of the endemic balance point from the system of differential equations (1), the Routh Hurwitz criterion is used.

We first give the Jacobean matrix in E^*_{eq} as follows:

$$J(E^*_{eq}) = \begin{bmatrix} -\mu-\frac{\beta W^*}{N} & -\beta\frac{S^*}{N} & 0 & 0 & 0 & 0 & 0 \\ \frac{\beta W^*}{N} & \frac{\beta S^*}{N} - K_1 & 0 & 0 & 0 & 0 & 0 \\ 0 & \tau & -\mu & 0 & 0 & 0 & 0 \\ 0 & \delta & 0 & -K_2 & 0 & 0 & 0 \\ 0 & \Omega & 0 & 0 & -K_3 & 0 & 0 \\ 0 & \sigma & 0 & \rho & \alpha & -K_4 & 0 \\ 0 & 0 & 0 & 0 & 0 & \theta & -K_5 \end{bmatrix}. \tag{16}$$

It can be observed that the Jacobian matrix has a block triangular structure. In particular, the last five eigenvalues are directly obtained from the diagonal elements

$$\lambda_3 = -\mu, \quad \lambda_4 = -K_2, \quad \lambda_5 = -K_3, \quad \lambda_6 = -K_4, \quad \lambda_7 = -K_5.$$

Since all model parameters are positive, these eigenvalues are strictly negative.

Therefore, the local stability of the endemic equilibrium depends only on the eigenvalues associated with the first 2×2 block of the Jacobian matrix

$$J_2 = \begin{pmatrix} J_{11}^* & J_{12}^* \\ J_{21}^* & J_{22}^* \end{pmatrix}.$$

The characteristic equation associated with this block is

$$\lambda^2 - \mathrm{Tr}(J_2)\lambda + \det(J_2) = 0$$

where

$$\mathrm{Tr}(J_2) = J_{11}^* + J_{22}^*,$$

$$\det(J_2) = J_{11}^* J_{22}^* - J_{12}^* J_{21}^*.$$

Using the expressions of the equilibrium values S^* and W^*, it follows that

$$\mathrm{Tr}(J_2) < 0, \qquad \det(J_2) > 0$$

whenever $R_0 > 1$.

Therefore, both eigenvalues of J_2 have negative real parts.

Consequently, all eigenvalues of the Jacobian matrix have negative real parts when $R_0 > 1$.

Hence, the endemic equilibrium E_{eq}^* is locally asymptotically stable. □

B Proof of Theorem 5

Proof. Consider the following Lyapunov function:

$$\begin{aligned} V \ : \quad \psi \ \rightarrow \ \mathbb{R} \\ V\left(S^*, W^*, R^*, I^*, E^*, D^*, C^*\right) = S - S^* \ln(\frac{S}{S^*}) + W - W^* \ln(\frac{W}{W^*}) \\ + R - R^* \ln(\frac{R}{R^*}) + I - I^* \ln\left(\frac{I}{I^*}\right) \\ + E - E^* \ln\left(\frac{E}{E^*}\right) + D - D^* \ln\left(\frac{D}{D^*}\right) \\ + C - C^* \ln\left(\frac{C}{C^*}\right), \end{aligned} \tag{17}$$

where

$$\psi = \{(S^*, W^*, R^*, I^*, E^*, D^*, C^*) \in \psi; \\ S^*, W^*, R^*, I^*, E^*, D^*, \textit{and } C^* \textit{are all positive}\}. \tag{18}$$

The Lyapunov function's temporal derivative is given by

$$\frac{dV}{dt} = (1 - \frac{S^*}{S})\frac{ds}{dt} + (1 - \frac{W^*}{W})\frac{dW}{dt} + (1 - \frac{R^*}{R})\frac{dR}{dt} + (1 - \frac{I^*}{I})\frac{dI}{dt} + (1 - \frac{E^*}{E})\frac{dE}{dt} \\ +(1 - \frac{D^*}{D})\frac{dD}{dt} + (1 - \frac{C^*}{C})\frac{dC}{dt}. \tag{19}$$

Then,

$$\frac{dV}{dt} = (1 - \frac{S^*}{S})(\Lambda - \mu S - \frac{\beta SW}{N}) + (1 - \frac{W^*}{W})\left[\frac{\beta SW}{N} - K_1 W\right] \\ +(1 - \frac{R^*}{R})\left[\tau W - \mu R\right] + (1 - \frac{I^*}{I})\left[\delta W - K_2 I\right] \\ +(1 - \frac{E^*}{E})\left[\Omega W - K_3 E\right] + (1 - \frac{D^*}{D})(\sigma W + \rho I + \alpha E - K_4 D) \\ +(1 - \frac{C^*}{C})\left[\theta D - K_5 C\right]. \tag{20}$$

Using the inequality

$$x - x^* - x^* \ln\left(\frac{x}{x^*}\right) \geq 0,$$

for all positive variables, each term of the Lyapunov function is nonnegative and vanishes only at the equilibrium point.

Substituting the system equations into $\frac{dV}{dt}$ and using the equilibrium relations satisfied by $(S^*, W^*, R^*, I^*, E^*, D^*, C^*)$, we obtain

$$\frac{dV}{dt} \leq 0$$

for all states in the feasible region ψ.

Moreover,

$$\frac{dV}{dt} = 0$$

if and only if

$$S = S^*, \quad W = W^*, \quad R = R^*, \quad I = I^*, \quad E = E^*, \quad D = D^*, \quad C = C^*.$$

Hence the largest invariant set contained in

$$\left\{X \in \psi : \frac{dV}{dt} = 0\right\}$$

is the singleton $\{E^*_{eq}\}$.

Therefore, according to LaSalle's invariance principle [23], the endemic equilibrium E^*_{eq} is globally asymptotically stable in ψ whenever $R_0 > 1$.

□

Funding

This research received no external funding.

Conflicts of Interest

The author declares no conflicts of interest.

References

[1] Aldila, D., Rarasati, N., Nuraini, N. and Soewono, E. *Optimal control problem of treatment for obesity in a closed population*, Int. J. Math. Math. Sci. 2014 (1) (2014), 273037.

[2] Boles, A., Kandimalla, R. and Reddy, P.H. *Dynamics of diabetes and obesity: Epidemiological perspective*, Biochim. Biophys. Acta Mol. Basis Dis. 1863(5) (2017), 1026–1036.

[3] Centers for Disease Control and Prevention, *National Diabetes Statistics Report*, U.S. Department of Health and Human Services, 2022.

[4] Christakis, N.A. and Fowler, J.H. *The spread of obesity in a large social network over 32 years*, New Engl. J. Med. 357(4) (2007), 370–379.

[5] Chitnis, N., Hyman, J.M. and Cushing, J.M. *Determining important parameters in the spread of malaria through the sensitivity analysis of a mathematical model*, Bull. Math. Biol. 70 (2008), 1272–1296.

[6] Chiyaka, E.T. and Garira, W. *Mathematical analysis of the transmission dynamics of schistosomiasis in the human-snail hosts*, J. Biol. Syst. 17(3) (2009), 397–423.

[7] Després, J.P. *L'obésité abdominale, une maladie métabolique*, John Libbey Eurotext, 2007.

[8] Dubey, B., Patra, A., Srivastava, P.K. and Dubey, U.S. *Modeling and analysis of an SEIR model with different types of nonlinear treatment rates*, J. Biol. Syst. 21(3) (2013) p.1350023.

[9] El Mansouri, A., Labzai, A., Belam, M. and Rachik, M. *Mathematical modeling and optimal control strategy for the obesity epidemic*, 2022 (2022), Article ID 20.

[10] El Mansouri, A., Smouni, I., Khajji, B., Labzai, A., Belam, M. and Tidli, Y. *Optimal control in a mathematical model of the spread of the obesity epidemic and its impact on diabetes*, Commun. Math. Biol. Neurosci., 2023 (2023), Article ID 74.

[11] El Mansouri, A., Smouni, I., Khajji, B., Labzai, A. and Belam, M. *Mathematical modeling and optimal control strategy for the monkeypox epidemic*, Math.l Model. Comput. 10(3) (2023), 944–955.

[12] Essounaini, A., Labzai, A., Laarabi, H. and Rachik, M. *Mathematical modeling and optimal control strategy for a discrete time model of COVID-19 variants*, Commun. Math. Biol. Neurosci., 2022 (2022), Article ID 25.

[13] Gumel, A.B., Shivakumar, P.N. and Sahai, B.M. *A mathematical model for the dynamics of HIV-1 during the typical course of infection*, Nonlinear Anal. Theory. Method. Appl. 47(3) (2001), 1773–1783.

[14] Hotamisligil, G. S. *Inflammation and metabolic disorders*, Nature 444 (2006), 860–867.

[15] Hu, Z., Teng, Z. and Jiang, H. *Stability analysis in a class of discrete SIRS epidemic models*, Nonlinear Anal. Real World Appl. 13(5) (2012), 2017–2033.

[16] Jódar, L., Santonja, F.J. and González-Parra, G. *Modeling dynamics of infant obesity in the region of Valencia, Spain*, Comput. Math. with Appl. 56(3) (2008), 679–689.

[17] Kahn, S. E., Hull, R. L. and Utzschneider, K. M. *Mechanisms linking obesity to insulin resistance and type 2 diabetes*, Nature 444 (2006), 840–846.

[18] Khajji, B., Kada, D., Balatif, O. and Rachik, M. *A multi-region discrete time mathematical modeling of the dynamics of COVID-19 virus propagation using optimal control*, J. Appl.Math. Comput. 64 (2020), 255–281.

[19] Kim, M.S., Chu, C. and Kim, Y. *A note on obesity as epidemic in Korea*, Osong Public Health. Res. Perspect. 2 (2011), 135–140.

[20] Kouidere, A., Labzai, A., Ferjouchia, H., Balatif, O. and Rachik, M. *A new mathematical modeling with optimal control strategy for the dynamics of population of diabetics and its complications*, J. Appl. Math. 2020(1) (2020), 1943410.

[21] Kot, M. *Elements of Mathematical Ecology*, Cambridge University Press, 2001.

[22] Labzai, A., Kouidere, A., Balatif, O. and Rachik, M. *Stability analysis of mathematical model of COVID-19 disease spread in population*, Commun. Math. Biol. Neurosci., 2020 (2020), Article ID 41.

[23] La Salle, J.P. *The Stability of Dynamical Systems*, Society for Industrial and Applied Mathematics, 1976.

[24] Misra, A.K., Sharma, A. and Singh, V. *Effect of awareness programs in controlling the prevalence of an epidemic with time delay*, J. Biol. Syst. 19(2) (2011), 389–402.

[25] Mwasa, A. and Tchuenche, J.M. *Mathematical analysis of a cholera model with public health intervention*, Biosystems 105(3) (2011), 190–200.

[26] Naheed, A., Singh, M. and Lucy, D. *Numerical study of SARS epidemic model with the inclusion of diffusion in the system*, Appl. Math. Comput. 229 (2014), 480–498.

[27] National Center for Chronic Disease Prevention and Health Promotion (U.S.). Division of Diabetes Translation. *Maps in trends of diagnosed diabetes and obesity* (2015) https://stacks.cdc.gov/view/cdc/29563.

[28] Ogden, C.L., Carroll, M.D., Kit, B.K. and Flegal, K.M. *Prevalence of childhood and adult obesity in the United States, 2011–2012*, JAMA 311(8) (2014), 806–814.

[29] Osman, S., Otoo, D., Sebil, C. and Makinde, O.D. *Bifurcation, sensitivity and optimal control analysis of modelling anthrax–listeriosis co-dynamics*, Commun. Math. Biol. Neurosci. 2020 (2020) 98.

[30] Osman, S., Togbenon, H.A. and Otoo, D. *Modelling the dynamics of campylobacteriosis using nonstandard finite difference approach with optimal control*, Comput. Math. Methods Med. 2020(1) (2020), p.8843299.

[31] Pang, L., Liu, S., Zhang, X., Tian, T. and Zhao, Z. *Transmission dynamics and control strategies of COVID-19 in Wuhan, China*, J. Biol. Syst. 28(3) (2020), 543–560.

[32] Public Health England, *3.8 million people in England now have diabetes*, 2016.

[33] Sahib, I., Baroudi, M., Gourram, H., Khajji, B., Labzai, A. and Belam, M. *The analysis of the mathematical stability of a cholera disease model*, Iran. J. Numer. Anal. Optim. 15(3) (2025), 852–876.

[34] Santonja, F.J., Villanueva, R.J., Jódar, L. and González-Parra, G. *Mathematical modelling of social obesity epidemic in the region of Valencia, Spain*, Math. Comput. Model. Dyn. Syst. 16(1) (2010), 23–34.

[35] Smith, K.B. and Smith, M.S. *Obesity statistics*, Prim. Care - Clin. Off. 43(1) (2016), 121–135.

[36] Smouni, I., El Mansouri, A., Khajji, B., Labzai, A., Belam, M. and Tidli, Y. *Mathematical modeling and analysis of a monkeypox model*, Commun. Math. Biol. Neurosci. 2023 (2023) Article ID 85.

[37] Sherr, D. and Lipman, R.D. *Diabetes educators: skilled professionals for improving prediabetes outcomes*, Am. J. Prev. Med. 44(4) (2013), S390–S393.

[38] Sharma, S. and Samanta, G.P. *Drinking as an epidemic: a mathematical model with dynamic behaviour*, J. Appl. Math. Inform. 31(1-2) (2013), 1–25.

[39] Tahir, M., Inayat Ali Shah, S., Zaman, G. and Muhammad, S. *Ebola virus epidemic disease: modeling and stability analysis with required abstain strategies*, Cogent Biol. 4(1) (2018) 1488511.

[40] Tilahun, G.T., Makinde, O.D. and Malonza, D. *Modelling and optimal control of pneumonia disease with cost-effective strategies*, J. Biol. Dyn. 11 (2017), 400–426.

[41] Tremblay, A. and Chaput, J.P. *About unsuspected potential determinants of obesity*, Appl. Physiol. Nutr. Metab. 33(4) (2008), 791–796.

[42] Wang, X.Y., Hattaf, K., Huo, H.F. and Xiang, H. *Stability analysis of a delayed social epidemics model with general contact rate and its optimal control*, J. Ind. Manag. Optim. 12(4) (2016) 1267.

[43] Wang, S.S. and Brownell, K.D. *Public policy and obesity: the need to marry science with advocacy*, Psychiatric Clin. 28(1) (2005), 235–252.

[44] World Health Organization et al. *Diabetes* (2022).

[45] World Health Organization, *Global Report on Diabetes* (2016).

[46] World Health Organization, *Diabetes (Fact Sheet)* (2021).

[47] World Health Organization, *Obesity and overweight*, Newsroom (2021).

[48] World Health Organization, Regional Office for Europe, *WHO European Regional Obesity Report* (2022).

Aims and scope

Iranian Journal of Numerical Analysis and Optimization (IJNAO) is published by the Department of Applied Mathematics, Faculty of Mathematical Sciences, Ferdowsi University of Mashhad. Papers dealing with different aspects of numerical analysis and optimization, theories and their applications in engineering and industry are considered for publication.

Journal Policy

All submissions to IJNAO are first evaluated by the journal's Editor-in-Chief or one of the journal's Associate Editors for their appropriateness to the scope and objectives of IJNAO. If deemed appropriate, the paper is sent out for review using a single blind process. Manuscripts are reviewed simultaneously by reviewers who are experts in their respective fields. The first review of every manuscript is performed by at least two anonymous referees. Upon the receipt of the referee's reports, the paper is accepted, rejected, or sent back to the author(s) for revision. Revised papers are assigned to an Associate Editor who makes an evaluation of the acceptability of the revision. Based upon the Associate Editor's evaluation, the paper is accepted, rejected, or returned to the author(s) for another revision. The second revision is then evaluated by the Editor-in-Chief, possibly in consultation with the Associate Editor who handled the original paper and the first revision, for a usually final resolution.

The authors can track their submissions and the process of peer review via: http://ijnao.um.ac.ir

All manuscripts submitted to IJNAO are tracked by using "iThenticate" for possible plagiarism before acceptance.

Instruction for Authors

The Journal publishes all papers in the fields of numerical analysis and optimization. Articles must be written in English.
All submitted papers will be refereed and the authors may be asked to revise their manuscripts according to the referee's reports. The Editorial Board of the Journal keeps the right to accept or reject the papers for publication.
The papers with more than one authors, should determine the corresponding author. The e-mail address of the corresponding author must appear at the end of the manuscript or as a footnote of the first page.

It is strongly recommended to set up the manuscript by Latex or Tex, using the template provided in the web site of the Journal. Manuscripts should be typed double-spaced with wide margins to provide enough room for editorial remarks.

References should be arranged in alphabetical order by the surname of the first author as examples below:

[1] Brunner, H. *A survey of recent advances in the numerical treatment of Volterra integral and integro-differential equations*, J. Comput. Appl. Math. 8 (1982), 213-229.

[2] Stoer, J. and Bulirsch, R. *Introduction to Numerical Analysis*, Springer-verlag, New York, 2002.

Iranian Journal of Numerical Analysis and Optimization

CONTENTS Vol. 16, No. 3, pp 757-1188

A fitted numerical technique for a class of singularly perturbed delay differential equations 757
A. Mardi and H. Shankar Prasad

Adaptive-batch stochastic gradient descent for constrained optimization based on relaxed barrier functions 790
A. Fillali, S. Addoune and R. Zeghdane

Comparative semi-analytical solutions of Caputo time-fractional Fokker–Planck equations via Laplace transform-based decomposition and iterative methods 815
M. Hossain and M. Aman Ullah

A probabilistic approach to portfolio selection with Risk-Aversion coefficient estimated via logistic regression 852
L. Chin, A. Sukmana, E. Chendra and A. P. Wijaya

Tension spline-quasilinearization scheme for generalized nonlinear time-fractional convection-diffusion equations 876
A.S. Bekela, L.G. Enyadene and M.M. Woldaregay

Robust stochastic gradient descent for linearly constrained problems via adaptive barrier amplification 922
A. Fillali, S. Addoune and R. Zeghdane

A new numerical computational method for solving nonhomogeneous time-fractional nonlinear parabolic problems in unbounded domains 952
S.A. Ghanizadeh and M.H. Akrami

A compact common-weight DEA model for efficiency measurement 991
A. Hadya K, T. Radha Ramanan, M.C. Lineesh and C.M. Sushama

Barycentric rational interpolation method for fractional-order differential equation model of HIV infection of $CD4^+$T-cells 1009
Kh. Divsar and R. Katani

Existence and Ulam–Hyers stability for a singular Caputo-fractional problem with integral boundary conditions 1031
A.E-S. Bouziane and B. Bendoukha

Inverse eigenvalue problems for composite banded tridiagonal matrices: Symmetric and nonsymmetric cases 1064
M. Babaei Zarch

Mathematical model on the effect of primary and secondary toxicants on a biological species: Occurrence of deformity 1083
A.K. Agarwal and A. Agrawal

Analysis of 2D Burger's equation using numerical and semi-analytic techniques 1107
A. Kaur, A. Pasrija and S. Arora

Quasilinearized Genocchi-Polynomial Collocation for Efficient Simulation of a Nonlinear SIRV Anthrax Transmission Model 1133
N. Mohamed, M. El-Gamel and W. Adel

Analysis of the mathematical model's stability describing the prevalence of obesity and its role in diabetes 1161
A. El Mansouri, M. Belam and A. Labzai

web site: **https://ijnao.um.ac.ir**
Email: **ijnao@um.ac.ir**
ISSN-Print: **2423-6977**
ISSN-Online: **2423-6969**