



Iranian Journal of Numerical Analysis and Optimization

Volume 16, Number 2

June 2026

Serial Number: 37

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 2, June 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 2, June 2026

Ferdowsi University of Mashhad - Iran

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: ghobjati@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: rlopez@unizar.es

Mahdavi-Amiri, Nezam*

(Optimization)

Sharif University of Technology, Iran.

e-mail: nezamm@sina.sharif.edu

Mirzaei, Davoud*

(Numerical Analysis)

University of Uppsala, Sweden.

e-mail: davoud.mirzaei@it.uu.se

Omrani, Khaled*

(Numerical Analysis)

University of Tunis El Manar, Tunisia.

khaled.omrani@issatso.rnu.tn

Salehi Fathabadi, Hasan*

(Operations Research)

University of Tehran, Iran.

e-mail: hsalehi@ut.ac.ir

Soheili, Ali Reza*

(Numerical Analysis)

Ferdowsi University of Mashhad, Iran.

e-mail: soheili@um.ac.ir

Soleimani Damaneh, Majid*

(Operations Research and Optimization, Finance, and
Machine Learning)

University of Tehran, Iran.

e-mail: m.soleimani.d@ut.ac.ir

Toutounian, Faezeh*

(Numerical Analysis)

Ferdowsi University of Mashhad, Iran.

e-mail: toutouni@um.ac.ir

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 solution approach to the multilevel linear fractional programming problems	339
A. Baghalnezhad and H. Sadeghi	
A mathematical model and optimal control analysis of merchant resistance to e-commerce adoption	356
H. Hassouni, S. Belhdid, O. Aarabate and O. Balatif	
Fractal-fractional dynamics of HIV/AIDS: Unveiling the synergistic impact of condoms, prophylaxis, and antiretroviral interventions for epidemic control	385
G.O. Acheneje, J. Amos, E. Abah, W. Atokolo, B.C. Agbata, Shyamsunder, A.K. Awasthi and B. Bolaji	
Finite element analysis of cross-diffusion three-species Shigesada–Kawasaki–Teramoto model	447
G.A. Al-Musawi, H.A. Challob and A.J. Harfash	
Characterizing maximum output sets in fractional-order discrete-time linear systems utilizing fractional feedback control	497
A. EL Bhih	
Spline-interpolation solution of a plane Stefan problem	525
P.N. Ivanshin	
Local fractional Khalouta transform: Properties and applications	539
A. Khalouta	
A nonmonotone line search method for solving constrained multiobjective optimization problems	563
M. Rashidi and E. Khorram	
Design and implementation of a graph-coloring algorithm for optimizing flight-level allocation in air traffic management in Iran	590
S.A.A. Mosavi, A. Erfanian and N. Sabeghi	
Fitted-mesh cubic spline methods for layer-dominated singularly perturbed parabolic problems with nonsmooth data	607
S.L. Cheru, G.F. Duressa and T.B. Mekonnen	
Optimal control and a system of quasi-variational inequalities	638
J. Ogorzaly	
Mathematical model for an infectious disease with quarantine	667
F. Iranzad, H.R. Marasi, H. Kheiri and A. Jabbari	
Synchronization and control of discrete incommensurate fractional-order and variable-order neural networks	683
S. Momani, I.M. Batiha, A. Hiousal and A. Ouannas	

Mathematical analysis of a fractional-order influenza model with resistant and nonresistant strains	703
I. Aamoum, A. Meskaf, Z. Yaagoub and E.M. Farah	
Tailed-Average SGD: A Polyak–Ruppert modification for optimal-rate constrained optimization	727
A. Fillali, S. Addoune and R. Zeghdane	



A solution approach to the multilevel linear fractional programming problems

A. Baghalnezhad and H. Sadeghi*

Abstract

In this paper, we consider the multilevel linear fractional programming problems over a bounded polytope set. We present a characterization of the optimum solution to the n -level linear fractional programming problem for case $n > 2$. Then, we propose an extension of the K th-best algorithm, for solving the n -level linear fractional programming problem with $n > 2$, and prove its convergence. Furthermore, we consider a previously published paper on such problems. It is shown that some results and proofs presented in that paper are incorrect by providing a counterexample. Finally, some numerical examples are presented, and the results are compared to those obtained from existing methods to show the accuracy and efficiency of the proposed algorithm.

AMS subject classifications (2020): Primary 90C26; Secondary 90C30, 90C32.

Keywords: Linear fractional programming; Multilevel programming; Vertex point; K th-best algorithm.

*Corresponding author

Received 22 February 2025; revised 4 July 2025; accepted 21 July 2025

Akram Baghalnezhad

Department of Mathematics, Shahid Chamran University of Ahvaz, Ahvaz, Iran. e-mail: a-baghalnezhad@stu.scu.ac.ir

Habibe Sadeghi

Department of Mathematics, Shahid Chamran University of Ahvaz, Ahvaz, Iran. e-mail: habibe.sadeghi@scu.ac.ir

How to cite this article

Baghalnezhad, A. and Sadeghi, H., A solution approach to the multilevel linear fractional programming problems. *Iran. J. Numer. Anal. Optim.*, 2026; 16(2): 339–355. <https://doi.org/10.22067/ijnao.2025.92296.1600>

1 Introduction

Multilevel programming is a tool for modeling decision-making problems in systems where decision-makers (DMs) are arranged hierarchically. At each level, in such systems, DMs attempt to optimize their own goals while bearing in mind the decisions of all higher levels. Multilevel programming, in cases with only two or three levels, is referred to as bilevel or trilevel programming, respectively. Various multilevel programming models have appeared in different fields, such as planning [8, 10], security, accident management [1, 6], supply chain management [20, 23], economics [11], and decentralized inventory [9]. On the other hand, linear fractional programming (LFP) is appropriate for modeling various real-world problems, including efficiency measurement and financial planning [21]. The LFP problem is a particular class of nonlinear programming problems wherein the objective function is a ratio of two linear functions, but all constraints are linear. The LFP has attracted conspicuous attention in the literature [16, 21]. In this paper, we study the multilevel linear fractional programming (MLLFP) problems in which the objective functions of all levels are linear fractional, and the common constraint region is a bounded polytope. Some researchers have also studied the MLLFP problems and proposed some solution approaches [12, 22]. Fathy et al. [7] converted fully intuitionistic fuzzy linear fractional multilevel programming problem into a multilevel multiobjective linear programming. Then ϵ -constraint method is used to reduce the resulting problem to a single objective linear programming problem [7]. In [3], the authors have proved the existence of a border-feasible vertex of the inducible region (the feasible region of the first-level DM which is implicitly defined by constraints and the lower-levels DMs), that solves the bilevel linear fractional programming (BLLFP) problem. In this study, we extend this result to the MLLFP problems, including $n > 2$ levels. We then develop the K th-best algorithm, one of the most well known methods for solving bilevel programming problems [5], to solve the MLLFP problems, and we discuss its convergence. Moreover, we consider a previously published article by Bhargava [2]. A counterexample is provided to demonstrate that some definitions and theorems in [2] are incorrect.

The structure of the paper is as follows. The mathematical formulation of the MLLFP problem under consideration, along with some relevant definitions, is presented in Section 2. The verification of some definitions and results given in [2] has been carried out in Section 3. In Section 4, some geometric properties of the optimum solution to the MLLFP problem have been proved. The K th-best algorithm to solve the MLLFP problems and the proof of its convergency are developed in Section 5. A number of numerical examples to show the feasibility of the proposed K th-best algorithm are presented in Section 6. Ultimately, the paper is finalized in Section 7.

2 Formulation of the problem

We address the MLLFP problem formulated as follows:

$$\begin{aligned}
 \min_{x_1 \in X_1} \quad & f_1(x_1, x_2, \dots, x_n) = \frac{\theta_1 + c_{11}x_1 + \dots + c_{1n}x_n}{\gamma_1 + d_{11}x_1 + \dots + d_{1n}x_n}, \\
 \min_{x_2 \in X_2} \quad & f_2(x_1, x_2, \dots, x_n) = \frac{\theta_2 + c_{21}x_1 + \dots + c_{2n}x_n}{\gamma_2 + d_{21}x_1 + \dots + d_{2n}x_n}, \\
 & \vdots \\
 \min_{x_n \in X_n} \quad & f_n(x_1, x_2, \dots, x_n) = \frac{\theta_n + c_{n1}x_1 + \dots + c_{nn}x_n}{\gamma_n + d_{n1}x_1 + \dots + d_{nn}x_n}, \\
 \text{s.t.} \quad & \sum_{i=1}^n A_i x_i \leq b,
 \end{aligned} \tag{1}$$

where n is a natural number, c_{ij}, d_{ij}, b , and A_i , for each $i, j \in \{1, 2, \dots, n\}$, are vectors and matrices of appropriate dimensions, respectively, θ_i , and γ_i are scalars, $n_1, \dots, n_n$ are natural numbers whose sum is equal to $\bar{n}$, $x_i \in X_i \subset \mathbb{R}_+^{n_i}$ is the variable under the control of the i th-level DM, and the linear fractional function $f_i : \mathbb{R}^{\bar{n}} \rightarrow \mathbb{R}$, is well-defined and continuous on the constraint region C , defined below. We make assumption $\gamma_i + d_{i1}x_1 + \dots + d_{in}x_n > 0$, for all $(x_1, \dots, x_n) \in C$. If this assumption does not hold, then the linear fractional objective function f_i can be equivalently expressed as $\frac{-(\theta_i + c_{i1}x_1 + \dots + c_{in}x_n)}{-(\gamma_i + d_{i1}x_1 + \dots + d_{in}x_n)}$.

Now, we express some key definitions and concepts associated with the MLLFP problem [19].

Definition 1. The constraint region of the problem (1) is defined as

$$C = \{(x_1, \dots, x_n) \in X_1 \times X_2 \cdots \times X_n : \sum_{i=1}^n A_i x_i \leq b\}.$$

Definition 2. The image of C onto $X_1 \times \dots \times X_i$, for $i = 2, \dots, n-1$ is defined as

$$\begin{aligned}
 C_{X_1 \times \dots \times X_i} = \{ & (x_1, \dots, x_i) : \text{there is } (x_{i+1}, \dots, x_n) \in X_{i+1} \times \dots \times X_n \\
 \text{s.t. } & (x_1, \dots, x_i, x_{i+1}, \dots, x_n) \in C\}.
 \end{aligned}$$

Definition 3. The image C onto $X_i \times \dots \times X_n$, for $i = 2, \dots, n-1$ is defined as

$$\begin{aligned}
 C_{X_i \times \dots \times X_n} = \{ & (x_i, \dots, x_n) : \text{there is } (x_1, \dots, x_{i-1}) \in X_1 \times \dots \times X_{i-1} \\
 \text{s.t. } & (x_1, \dots, x_i, \dots, x_n) \in C\}.
 \end{aligned}$$

Definition 4. The point-to-set mapping $P_n : C_{X_1 \times \dots \times X_{n-1}} \rightarrow C_{X_n}$ is defined as

$$\begin{aligned}
P_n(x_1, \dots, x_{n-1}) &= \{x_n \mid f_n(x_1, \dots, x_n) \\
&= \min\{f_n(x_1, \dots, x_{n-1}, \hat{x}_n) : (x_1, \dots, x_{n-1}, \hat{x}_n) \in C\}\}.
\end{aligned}$$

In actuality, P_n represents the logical reactions of the n th-level DM for the given $(x_1, \dots, x_{n-1})$.

Definition 5. The point-to-set mapping $P_k : C_{X_1 \times \dots \times X_{k-1}} \rightarrow C_{X_k \times \dots \times X_n}$ for $k = 2, \dots, n-1$, is defined as

$$\begin{aligned}
P_k(x_1, \dots, x_{k-1}) \\
&= \{(x_k, \dots, x_n) \mid f_k(x_1, \dots, x_k, \dots, x_n) = \min\{f_k(x_1, \dots, x_{k-1}, \hat{x}_k, \dots, \hat{x}_n) \\
&\quad \text{s.t. } (x_1, \dots, x_{k-1}, \hat{x}_k, \dots, \hat{x}_n) \in C \\
&\quad \text{and } (\hat{x}_{k+1}, \dots, \hat{x}_n) \in P_{k+1}(x_1, \dots, x_{k-1}, \hat{x}_k)\}\},
\end{aligned}$$

where P_k represents the logical reactions of K th-level DM for fixed values of $(x_1, \dots, x_{k-1})$.

Definition 6. The set of feasible solutions to the MLLFP problem (1) is termed inducible region (IR) and is defined as

$$IR := \{(x_1, \dots, x_n) \in C \mid (x_2, \dots, x_n) \in P_2(x_1)\}. \quad (2)$$

Thus, to obtain an optimum solution to the MLLFP problem (1), we can solve the following problem:

$$\min\{f_1(x_1, \dots, x_n) \mid (x_1, \dots, x_n) \in IR\}. \quad (3)$$

Obviously, an optimum solution to this problem is also optimum for the original problem (1).

We will also use the following assumptions to come up with the existence of an optimal solution to the problem (1), [19].

Assumptions

1. The constraint region C is a nonempty, closed, and bounded polytope.
2. The mapping $P_k(x_1, \dots, x_{k-1})$ assigns a unique value to each $(x_1, \dots, x_{k-1}) \in C_{X_1 \times \dots \times X_{k-1}}$ and for each $2 \leq k \leq n$.

In the following section, we analyze some definitions, results, and proofs in [2].

3 Examining the definitions, results and proofs in [2]

As the major drawback, the author has mentioned the IR set in several places in [2], while the IR set is not defined throughout the article. Some of the cases where the IR set is used in [2], are as follows:

1. On page 10, the statement “There exists an extreme point of IR , thus an extreme point of the polyhedron S , which is an optimal solution of the MLLFP problem”, without citing or proof.
- 2- In the last paragraph on page 14, the author express “Then, the extreme point in W which provides the best value of f_1 is selected to test whether it is a point of IR ”.

Actually, Bhargava [2] has defined $IR_1, \dots, IR_{n-1}$ regions, that IR_1 exhibits the feasible region of the BLLFP problem, IR_2 exhibits the feasible region of the trilevel linear fractional programming (TLLFP) problem, and finally, IR_{n-1} exhibits the feasible region of the MLLFP problem with n levels. However, the definition of IR_i is only correct for the two-level case, that is, IR_1 . The following TLLFP problem exemplifies this reality (note that any linear function can be considered as a linear fractional function with denominator one):

Example 1. Consider

$$\begin{aligned} \min_{x_1} \quad & f_1(x_1, x_2, x_3) = -x_1 + 2x_2 - 4x_3, \\ \min_{x_2} \quad & f_2(x_1, x_2, x_3) = 3x_2 - 2x_3, \\ \min_{x_3} \quad & f_3(x_1, x_2, x_3) = 2x_3 - x_2, \\ \text{s.t.} \quad & \begin{cases} x_1 + x_2 + x_3 \leq 2, \\ 0 \leq x_1, x_2, x_3 \leq 1. \end{cases} \end{aligned} \tag{4}$$

By [18], $IR = \{(x_1, 0, 0) : 0 \leq x_1 \leq 1\}$. According to the definition of IR_2 in [2], we have $IR_2 = \{(x_1, x_2, 0) : x_1 + x_2 \leq 2\}$, which is clearly different from the inducible region of the problem.

In [2, Theorem 1], the author claims that there is a border-feasible vertex (or boundary feasible extreme point) of C that solves the MLLFP problem. However, its proof is questionable. In the first paragraph of the proof of Theorem 1, the author asserted that “if $(\bar{x}_1, \bar{x}_2, \dots, \bar{x}_n)$ is not a boundary feasible point, every extreme point adjacent to $(\bar{x}_1, \bar{x}_2, \dots, \bar{x}_n)$ is in $(IR)_1$, and $f_1(\bar{x}_1, \bar{x}_2, \dots, \bar{x}_n) \leq f_1(x_1, x_2, \dots, x_n)$ for all point $(x_1, x_2, \dots, x_n)$ adjacent to $(\bar{x}_1, \bar{x}_2, \dots, \bar{x}_n)$.” Whereas according to [2, Definition 2], if $(\bar{x}_1, \bar{x}_2, \dots, \bar{x}_n)$ is not a boundary feasible point, then every extreme point adjacent to $(\bar{x}_1, \bar{x}_2, \dots, \bar{x}_n)$ is in $(IR)_{n-1}$. Also, the characterization of the optimal solution to the MLLFP problem is provided following Definition 2, and the defi-

inition of $(IR)_{n-1}$ on page 6, implies that $f_n(\bar{x}_1, \bar{x}_2, \dots, \bar{x}_n) \leq f_n(x_1, x_2, \dots, x_n)$ for all point $(x_1, x_2, \dots, x_n)$ adjacent to $(\bar{x}_1, \bar{x}_2, \dots, \bar{x}_n)$. Thus the inequality (5) in the proof of Theorem 1 is incorrect.

Bhargava [2] continued by introducing $(x_1^{[1]}, \dots, x_n^{[1]})$ as the optimum solution to the relaxed problem (5), (defined in the next section). In [2, Theorem 2], the author stated that there is an edge path in C from any arbitrary vertex of C to $(x_1^{[1]}, \dots, x_n^{[1]})$ that the value of $f_1, \dots, f_n$ is nonincreasing along it. This conclusion does not always hold. As a counterexample, consider the vertex $(0, 1, 0)$ in the above example. Here, $(x_1^{[1]}, x_2^{[1]}, x_3^{[1]}) = (1, 0, 1)$. In Figure 1, it is easily seen that each edge path from F to C must include one of the edges from A to C , D to C , B to C , or E to C . However, the function f_3 is not nonincreasing along any of these paths.

The author then proposed a K th-best algorithm to solve the MLLFP problem. Step 6 of the proposed algorithm [2, Page 15] begins with the phrase “on continuing the process” that is unclear for implementation. In what follows, we first present an extension of the K th-best algorithm presented in [3] to solve the TLLFP problem, and then for the MLLFP problems with more than three levels.

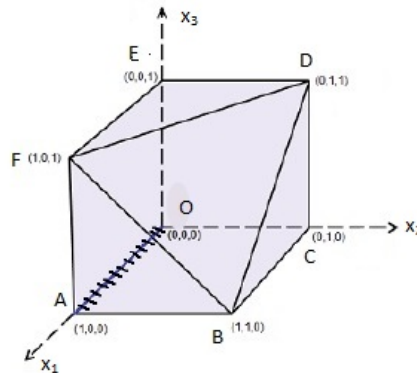


Figure 1: IR and the constraint region

4 Geometric characteristics

In this section, we first state several required definitions, and then we prove some main theorems on the optimum solution of the MLLFP problem (1).

Definition 7 (see [14]). Assume that g is a real-valued function defined on a convex subset X of R^n . Then

(1) g is quasi-concave on X if and only if $x^1, x^2 \in X, \lambda \in [0, 1]$ and $g(x^1) \leq g(x^2) \implies g(x^1) \leq$

$$g[(1 - \lambda)x^1 + \lambda x^2].$$

(2) g is semi-strictly quasi-concave on X if and only if $x^1, x^2 \in X, x^1 \neq x^2, \lambda \in (0, 1)$, and $g(x^1) < g(x^2) \implies g(x^1) < g[(1 - \lambda)x^1 + \lambda x^2]$.

(3) g is explicitly quasi-concave on X if and only if it is quasi-concave and semi-strictly quasi-concave on X .

The function g is explicitly quasi-convex if and only if $-g$ is explicitly quasi-concave.

(4) g is explicitly quasi-monotonic on X if and only if it is explicitly quasi-concave and explicitly quasi-convex on X .

In the MLLFP problem (1), the linear fractional functions $f_1, f_2, \dots, f_n$ are quasi-concave. Specifically, they are explicitly quasi-monotonic on C because $\gamma_i + d_{i1}x_1 + \dots + d_{in}x_n \neq 0$ on S , for all $i = 1, \dots, n$ (see [14, Theorem 3.53]). This establishes that the problem (1) belongs to a subclass of quasi-concave multilevel programming problems. In light of Assumption 2, it follows from [19] that

- the inducible region of the problem (1) is formed by the union of connected faces of the polytope C . Therefore, IR may not be a convex set.
- There exists a vertex of C (it also is a vertex of IR), which represents an optimum solution to the MLLFP problem (1).

Definition 8 (see [13]). A point $(x_1, \dots, x_n) \in IR$ is termed a border-feasible vertex of C if there is an edge E of C in such a way that $(x_1, \dots, x_n)$ is a vertex of E and another vertex of E does not lie in IR .

In the following theorem, we demonstrate that under some conditions, the optimum solution to the MLLFP problem (1) is a border-feasible vertex of IR . For this purpose, we introduce the following relaxed problem:

$$\begin{aligned} \min \quad & f_1(x_1, x_2, \dots, x_n) = \frac{\theta_1 + c_{11}x_1 + c_{12}x_2 + \dots + c_{1n}x_n}{\gamma_1 + d_{11}x_1 + d_{12}x_2 + \dots + d_{1n}x_n} \\ \text{s.t.} \quad & (x_1, x_2, \dots, x_n) \in C. \end{aligned} \quad (5)$$

It is a well-established result from [14] that there exists a vertex of C that solves the problem (5). Denote this solution by $(x_{11}^{[1]}, \dots, x_{1n}^{[1]})$. Note that f_1 is a quasi-concave function and that C is a nonempty, closed, and bounded polytope.

Theorem 1. If there exists a vertex of C outside IR that solves the problem (5), then an optimum solution to the MLLFP problem (1) lies on a border-feasible vertex of IR .

Proof. It is well known that an optimum solution to the MLLFP problem (1) lies on a vertex of C (see [19, Theorem 10]). We call it $(x_1^o, \dots, x_n^o)$. Obviously, $(x_1^o, \dots, x_n^o) \in IR$. If $(x_1^o, \dots, x_n^o)$ is

a border-feasible vertex of C , then the result follows, or else, all vertices adjacent to $(x_1^o, \dots, x_n^o)$ are in IR and for every vertex adjacent to $(x_1^o, \dots, x_n^o)$ in C , such as $(x_1, \dots, x_n)$, we have

$$f_1(x_1^o, \dots, x_n^o) \leq f_1(x_1, \dots, x_n). \quad (6)$$

We now establish that there exists a vertex of C , as $(x_1', \dots, x_n')$, adjacent to $(x_1^o, \dots, x_n^o)$ satisfying

$$f_1(x_1', \dots, x_n') = f_1(x_1^o, \dots, x_n^o). \quad (7)$$

To demonstrate this, one can examine the relaxed problem (5). It yields from (6) that $(x_1^o, \dots, x_n^o)$ is a local vertex-minimizer of f_1 in C . Because f_1 is quasi-concave, and explicitly quasi-convex on C , the point $(x_1^o, \dots, x_n^o)$ can be considered as a global minimum solution to the problem (5) (see [14, Theorem 5.13]), that is,

$$f_1(x_1^o, \dots, x_n^o) \leq f_1(x_1, \dots, x_n), \quad \text{for all } (x_1, \dots, x_n) \in C. \quad (8)$$

By hypothesis, a vertex of C , as $(x_1'', \dots, x_n'')$, outside IR exists that is the optimum solution to the problem (5). Therefore, $f_1(x_1^o, \dots, x_n^o) = f_1(x_1'', \dots, x_n'')$. Notably, $(x_1'', \dots, x_n'')$ is not adjacent to $(x_1^o, \dots, x_n^o)$ because $(x_1^o, \dots, x_n^o)$ is not a border-feasible vertex of C . Also, f_1 is continuous, quasi-convex, and explicitly quasi-concave on C . Thus, the convex hull of some vertices of C represents the set of optimum solutions to the problem (5) (see [14, Theorem 5.21]). So, the set of optimum solutions to the problem (5) is a polyhedron. Thus, there is an edge path in such polyhedron from $(x_1^o, \dots, x_n^o)$ to $(x_1'', \dots, x_n'')$. Then, there must be a vertex of C , as $(x_1', \dots, x_n')$, adjacent to $(x_1^o, \dots, x_n^o)$ remaining in the set of optimum solutions to the problem (5). Therefore, the validity of (7) is confirmed. If $(x_1', \dots, x_n')$ is a border-feasible vertex of C , then the result follows, or else, we examine the vertex $(x_1', \dots, x_n')$ instead of $(x_1^o, \dots, x_n^o)$ and continue the aforementioned process. We obtain a new vertex of C adjacent to $(x_1', \dots, x_n')$ for which (7) holds. If the new vertex is a border-feasible vertex of C , then the result follows. Else, by continuing the process, a border-feasible vertex of C will be obtained as an optimum solution to the MLLFP (1), after a finite number of stages (note that the number of vertices of the polytope C is finite). $\square$

Theorem 2. Let $(x_1, \dots, x_n)$ be a vertex of C . Then, an edge path in C from $(x_1, \dots, x_n)$ to $(x_{11}^{[1]}, \dots, x_{1n}^{[1]})$ exists so that along which the value of f_1 is nonincreasing.

Proof. See the proof of Theorem 6 in [3]. $\square$

5 The K th-best algorithm

In this section, we first present the K th-best algorithm for TLLFP problem, that is, for case $n = 3$ and then for the MLLFP problems with $n > 3$ levels. Consider the following TLLFP problem:

$$\begin{aligned}
 \min_{x_1 \in X_1} \quad & f_1(x_1, x_2, x_3) = \frac{\theta_1 + c_{11}x_1 + c_{12}x_2 + c_{13}x_3}{\gamma_1 + d_{11}x_1 + d_{12}x_2 + d_{13}x_3}, \\
 \min_{x_2 \in X_2} \quad & f_2(x_1, x_2, x_3) = \frac{\theta_2 + c_{21}x_1 + c_{22}x_2 + c_{23}x_3}{\gamma_2 + d_{21}x_1 + d_{22}x_2 + d_{23}x_3}, \\
 \min_{x_3 \in X_3} \quad & f_3(x_1, x_2, x_3) = \frac{\theta_3 + c_{31}x_1 + c_{32}x_2 + c_{33}x_3}{\gamma_3 + d_{31}x_1 + d_{32}x_2 + d_{33}x_3}, \\
 \text{s.t.} \quad & \sum_{i=1}^3 A_i x_i \leq b.
 \end{aligned} \tag{9}$$

5.1 The K th-best algorithm for the TLLFP problem

Step 1. Constitute the relaxed problem as follows:

$$\begin{aligned}
 \min \quad & f_1(x_1, x_2, x_3) \\
 \text{s.t.} \quad & (x_1, x_2, x_3) \in C.
 \end{aligned} \tag{10}$$

Obtain an optimum solution to it and call it $(x_{11}^{[1]}, x_{12}^{[1]}, x_{13}^{[1]})$. Set $i = 1$, $U = \emptyset$, $V = \{(x_{11}^{[1]}, x_{12}^{[1]}, x_{13}^{[1]})\}$, and go to Step 2.

Step 2. Fix $x_1 = x_{11}^{[i]}$ and $x_2 = x_{12}^{[i]}$, and obtain an optimum solution to the following problem:

$$\begin{aligned}
 \min \quad & f_3(x_{11}^{[i]}, x_{12}^{[i]}, x_3) \\
 \text{s.t.} \quad & (x_{11}^{[i]}, x_{12}^{[i]}, x_3) \in C.
 \end{aligned} \tag{11}$$

Call it $x_{33}^{[i]}$. If $x_{13}^{[i]} = x_{33}^{[i]}$, then go to Step 3. Else, go to Step 4.

- At this Step, it is tested whether point $(x_{11}^{[i]}, x_{12}^{[i]}, x_{13}^{[i]})$ belongs to the set IR of the bilevel problem, which consists of the second-level and third-level problems. If it is not, then Step 3 is not executed. Thus, the computational effort required to solve the TLLFP problem is reduced.

Step 3. Fix $x_1 = x_{11}^{[i]}$ and obtain an optimum solution to the following bilevel problem:

$$\begin{aligned}
& \min_{x_2} f_2(x_{11}^{[i]}, x_2, x_3), \\
& \min_{x_3} f_3(x_{11}^{[i]}, x_2, x_3) \\
& s.t. \quad (x_{11}^{[i]}, x_2, x_3) \in C.
\end{aligned} \tag{12}$$

Let $(x_{22}^{[i]}, x_{23}^{[i]})$ be an optimum solution to the bilevel problem (12). If $(x_{22}^{[i]}, x_{23}^{[i]}) = (x_{12}^{[i]}, x_{13}^{[i]})$, then $(x_{11}^{[i]}, x_{12}^{[i]}, x_{13}^{[i]})$ is an optimum solution to the problem (9) and stop. Else, go to Step 4.

Step 4. Let $V^{[i]}$ be the set of all vertices adjacent to $(x_{11}^{[i]}, x_{12}^{[i]}, x_{13}^{[i]})$ in C . Set $U = U \cup \{(x_{11}^{[i]}, x_{12}^{[i]}, x_{13}^{[i]})\}$, $V = (V \cup V^{[i]}) \setminus U$, and go to Step 5.

- One can use the presented algorithm in [15, Section 3.7.1] to obtain all adjacent vertices.

Step 5. $i \leftarrow i + 1$ and choose $(x_{11}^{[i]}, x_{12}^{[i]}, x_{13}^{[i]}) \in V$ such that

$$f_1(x_{11}^{[i]}, x_{12}^{[i]}, x_{13}^{[i]}) = \min_{(x_1, x_2, x_3) \in V} f_1(x_1, x_2, x_3),$$

and go to Step 2.

- To do this Step, one can sort the points of V in ascending order of f_1 objective value in a list from top to bottom and then select the vertex at the top of the list.

5.2 The K th-best algorithm for the MLLFP problem

Suppose that the K th-best algorithm is available to solve the MLLFP problems with $n - 1$ levels. The K th-best algorithm to solve the MLLFP problems with n levels is as follows:

The algorithm

Step 1. Find an optimum solution to the relaxed problem (5) and call it $(x_{11}^{[1]}, \dots, x_{1n}^{[1]})$. Set $i = 1$, $U = \emptyset$, $V = \{(x_{11}^{[1]}, \dots, x_{1n}^{[1]})\}$, and go to Step 2.

Step 2. Fix $x_1 = x_{11}^{[i]}$. Solve the following linear fractional $(n - 1)$ -level programming problem:

$$\begin{aligned}
& \min_{x_2} f_2(x_{11}^{[i]}, x_2, \dots, x_n), \\
& \vdots \\
& \min_{x_n} f_n(x_{11}^{[i]}, x_2, \dots, x_n), \\
& s.t. \quad (x_{11}^{[i]}, x_2, \dots, x_n) \in C.
\end{aligned} \tag{13}$$

Call its optimum solution $(x_{11}^{[i]}, x_{n-1,2}^{[i]}, \dots, x_{n-1,n}^{[i]})$. If $(x_{11}^{[i]}, \dots, x_{1n}^{[i]}) = (x_{11}^{[i]}, x_{n-1,2}^{[i]}, \dots, x_{n-1,n}^{[i]})$, then $(x_{11}^{[i]}, \dots, x_{1n}^{[i]})$ is an optimum solution to the problem (1) and stop. Else, go to Step 3.

Step 3. Let $V^{[i]}$ be the set of all vertices adjacent to $(x_{11}^{[i]}, x_{12}^{[i]}, \dots, x_{1n}^{[i]})$ in C . Set $U = U \cup \{(x_{11}^{[i]}, x_{12}^{[i]}, \dots, x_{1n}^{[i]})\}$, $V = (V \cup V^{[i]}) \setminus U$, and go to Step 4.

Step 4. $i \leftarrow i + 1$ and choose $(x_{11}^{[i]}, x_{12}^{[i]}, \dots, x_{1n}^{[i]}) \in V$ such that

$$f_1(x_{11}^{[i]}, x_{12}^{[i]}, \dots, x_{1n}^{[i]}) = \min_{(x_1, x_2, \dots, x_n) \in V} f_1(x_1, x_2, \dots, x_n),$$

and go to Step 2.

According to the above algorithm first, an optimum solution to the relaxed problem (5) is obtained, called $(x_{11}^{[1]}, \dots, x_{1n}^{[1]})$. Step 2 tests whether the obtained point belongs to the IR of problem (1) or not. If the answer is positive, then it is an optimum solution to the problem (1), and the algorithm stops. If not, it will be eliminated from V and the set of all vertices adjacent to $(x_{11}^{[1]}, \dots, x_{1n}^{[1]})$, that is, $V^{[1]}$ will be added to V . Then, the vertex in V that produces the second-best value of f_1 , is picked out for testing whether it belongs to IR or not. The process stops when it reaches the first point in the IR of the problem (1).

Now, let $(x_{11}^{[1]}, \dots, x_{1n}^{[1]}), (x_{11}^{[2]}, \dots, x_{1n}^{[2]}), \dots, (x_{11}^{[t]}, \dots, x_{1n}^{[t]})$ be the t -ordered vertex solutions of the problem (5), which are computed successively using the described algorithm, that is,

$$f_1(x_{11}^{[i]}, \dots, x_{1n}^{[i]}) \leq f_1(x_{11}^{[i+1]}, \dots, x_{1n}^{[i+1]}), \quad i = 1, \dots, t-1.$$

Clearly, $(x_{11}^{[k]}, \dots, x_{1n}^{[k]})$ is a global optimum solution for the MLLFP problem (1) if

$$k = \min_{i \in \{1, \dots, t\}} \{i \mid (x_{11}^{[i]}, \dots, x_{1n}^{[i]}) \in IR\}.$$

Hence, based on Theorem 2, the K th-best algorithm outputs a global optimum solution to the problem (1). However, if an optimum solution to the relaxed problem (5) lies out of IR , it can be inferred that the K th-best algorithm identifies a border-feasible vertex of IR as an optimum solution to the problem (1). The following theorem highlights the relationship between the vertices generated by the proposed algorithm.

Theorem 3. The $(i+1)$ st best vertex of C , $(x_{11}^{[i+1]}, \dots, x_{1n}^{[i+1]})$, is adjacent to one of earlier vertices $(x_{11}^{[1]}, \dots, x_{1n}^{[1]})$, $(x_{11}^{[2]}, \dots, x_{1n}^{[2]})$..., or $(x_{11}^{[i]}, \dots, x_{1n}^{[i]})$, $i < t$.

Proof. See the proof of Theorem 7 in [3]. □

Theorem 4. Assume that there is a vertex of C outside IR which solves the relaxed problem (5). Then, the K th-best algorithm gets a border-feasible vertex of IR as an optimum solution to the MLLFP problem (1).

Proof. It follows from Theorem 1 that an optimum solution to the MLLFP problem (1) takes place at a border-feasible vertex of IR . On the other hand, the Theorem 3 implies that the i th

best vertex of the relaxed problem (5) is adjacent to one of the 1st, 2nd, $\dots$, or $(i-1)$ st best vertex. Based on Theorem 2 and according to the termination condition, the K th-best algorithm gets a global optimum solution to the problem (1) at a border-feasible vertex of IR . $\square$

6 Numerical examples

In this section, a number of numerical examples are solved by the presented K th-best algorithm that show the efficiency of the algorithm.

Example 2. Consider

$$\begin{aligned} \min_{x_1} \quad & f_1 = -8x_1 - 4x_2 + 4x_3 - 40x_4 - 4x_5 \\ \min_{x_2} \quad & f_2 = \frac{-x_2 + 3x_3 + 4x_4 - x_5}{x_1 + x_2 + x_3 + x_4 + 2}, \\ \min_{(x_3, x_4, x_5)} \quad & f_3 = \frac{x_1 + x_2 + 2x_3 - x_4 + x_5 + 1}{2x_1 + x_3 + x_4 - 3x_5 + 6}, \\ \text{s.t.} \quad & \begin{cases} -x_3 + x_4 + x_5 \leq 1, \\ 2x_1 - x_3 + 2x_4 - 0.5x_5 \leq 1, \\ 2x_2 + 2x_3 - x_4 - 0.5x_5 \leq 1, \\ x_i \geq 0, \end{cases} \quad i = 1, 2, 3, 4, 5. \end{aligned}$$

First iteration of the K th-best algorithm:

Step 1. By solving the corresponding problem (10), the point $(0, 0, \frac{3}{2}, \frac{3}{2}, 1)$ is obtained as the optimum solution. The optimum value of objective function f_1 is -58.

Step 2. Fix $x_1 = x_2 = 0$ and solve the following problem:

$$\begin{aligned} \min \quad & f_3 = \frac{2x_3 - x_4 + x_5 + 1}{x_3 + x_4 - 3x_5 + 6}, \\ \text{s.t.} \quad & \begin{cases} -x_3 + x_4 + x_5 \leq 1, \\ -x_3 + 2x_4 - 0.5x_5 \leq 1, \\ 2x_3 - x_4 - 0.5x_5 \leq 1, \\ x_i \geq 0, \end{cases} \quad i = 3, 4, 5. \end{aligned}$$

The optimum solution is $(x_3, x_4, x_5) = (0, \frac{1}{2}, 0)$. Since $(0, \frac{1}{2}, 0) \neq (\frac{3}{2}, \frac{3}{2}, 1)$, we go to Step 4.

Step 4. We obtain $V^{[1]} = \{(\frac{3}{2}, 0, 1, 0, 2), (0, \frac{9}{10}, 0, \frac{3}{5}, \frac{2}{5}), (0, 0, 1, 1, 0), (0, 0, 1, 0, 2), (0, 0, 0, \frac{3}{5}, \frac{2}{5})\}$

and $U = \{(0, 0, \frac{3}{2}, \frac{3}{2}, 1)\}$, $V = V^{[1]}$.

Step 5. We set $i = 2$ and $(x_{11}^{[2]}, \dots, x_{15}^{[2]}) = (0, 0, 1, 1, 0)$ with $f_1 = -36$ and go to Step 2.

The output of the rest of iterations are displayed in Table 1. As seen in Table 1, in the third iteration of the algorithm, the point $(0, \frac{9}{10}, 0, \frac{3}{5}, \frac{2}{5})$ is obtained as optimum solution. The optimum value of (f_1, f_2, f_3) equals $(-\frac{146}{5}, \frac{11}{35}, \frac{17}{54}) = (-29.2, 0.314, 0.315)$.

Table 1: The output of each iteration in K th-best algorithm for Example 2

Iteration	$(x_{11}^{[i]}, \dots, x_{15}^{[i]})$	V
$i = 1$	$(0, 0, \frac{3}{2}, \frac{3}{2}, 1) \notin IR$ $f_1 = -58$	$(\frac{3}{2}, 0, 1, 0, 2)$ $f_1 = -16$
		$(0, \frac{9}{10}, 0, \frac{3}{5}, \frac{2}{5})$ $f_1 = -29.2$
		$(0, 0, 1, 1, 0)$ $f_1 = -36$
		$(0, 0, 1, 0, 2)$ $f_1 = -4$
		$(0, 0, 0, \frac{3}{5}, \frac{2}{5})$ $f_1 = -25.6$
$i = 2$	$(0, 0, 1, 1, 0) \notin IR$ $f_1 = -36$	$(\frac{3}{4}, 0, \frac{1}{2}, 0, 0)$ $f_1 = -4$
		$(0, \frac{3}{4}, 0, \frac{1}{2}, 0)$ $f_1 = -23$
		$(0, 0, \frac{1}{2}, 0, 0)$ $f_1 = 2$
		$(0, 0, 0, \frac{1}{2}, 0)$ $f_1 = -20$
$i = 3$	$(0, \frac{9}{10}, 0, \frac{3}{5}, \frac{2}{5}) \in IR$ $f_1 = -\frac{146}{5} = -29.2$	

Example 3. Consider

$$\begin{aligned}
 \max_{x_1, x_2} \quad & f_1 = \frac{7x_1 + 3x_2 - 4x_3 + 2x_4}{x_1 + x_2 + x_3 + 1}, \\
 \max_{x_3} \quad & f_2 = \frac{x_2 + 3x_3 + 4x_4}{x_1 + x_2 + x_3 + 2}, \\
 \max_{x_4} \quad & f_3 = \frac{2x_1 + x_2 + x_3 + x_4}{x_1 + x_2 + x_3 + 3}, \\
 s.t. \quad & \begin{cases} x_1 + x_2 + x_3 + x_4 \leq 5, & x_1 + x_2 - x_3 - x_4 \leq 2, \\ x_1 + x_2 + x_3 \geq 1, & x_1 - x_2 + x_3 + 2x_4 \leq 4, \\ x_1 + 2x_3 + 2x_4 \leq 3, & x_4 \leq 2, \\ x_i \geq 0, & i = 1, 2, 3, 4. \end{cases}
 \end{aligned}$$

By applying the proposed K th-best algorithm, the point $(\frac{7}{3}, 0, 0, \frac{1}{3})$ is obtained as the optimum solution in the first iteration. The optimum value of (f_1, f_2, f_3) equals $(\frac{51}{10}, \frac{11}{13}, \frac{15}{16})$. However, Lachwani and Poonia (2012) obtained the optimum solution $(x_1, x_2, x_3, x_4) = (1, 0, 0, 1)$ with the optimum value of $(f_1, f_2, f_3) = (\frac{9}{2}, \frac{4}{3}, \frac{3}{4})$. Also, Lachwani and Nehra [12] using a modified fuzzy goal programming approach obtained the point $(2.33, 0, 0, 2.37)$ as optimum solution, which is clearly infeasible because of the constraint $x_4 \leq 2$.

Table 2: The output of each iteration in K th-best algorithm for Example 4

Iteration	$(x_{11}^{[i]}, \dots, x_{15}^{[i]})$	V
$i = 1$	$(0, \frac{3}{4}, 0, 0, 1) \notin IR$ $f_1 = \frac{1}{52}$	$(\frac{3}{4}, \frac{3}{4}, 0, 0, 1), f_1 = \frac{4}{49}$ $(0, 0, 1, 0, 2), f_1 = \frac{1}{16}$ $(0, \frac{9}{10}, 0, \frac{3}{5}, \frac{2}{5}), f_1 = \frac{13}{106}$ $(0, \frac{1}{2}, 0, 0, 0), f_1 = \frac{1}{16}$ $(0, 0, 0, 0, 1), f_1 = \frac{1}{13}$
$i = 2$	$(0, \frac{1}{2}, 0, 0, 0) \notin IR$ $f_1 = \frac{1}{16}$	$(\frac{1}{2}, \frac{1}{2}, 0, 0, 0), f_1 = \frac{2}{15}$ $(0, 0, \frac{1}{2}, 0, 0), f_1 = \frac{1}{7}$ $(0, \frac{5}{16}, 0, \frac{1}{2}, 0), f_1 = \frac{109}{272}$ $(0, 0, 0, 0, 0), f_1 = \frac{1}{8}$
$i = 3$	$(0, 0, 1, 0, 2) \notin IR$ $f_1 = \frac{1}{16}$	$(\frac{3}{2}, 0, 1, 0, 2), f_1 = \frac{5}{29}$ $(0, 0, \frac{3}{2}, \frac{3}{2}, 1), f_1 = \frac{8}{23}$ $(0, 0, \frac{1}{2}, 0, 0), f_1 = \frac{1}{7}$ $(0, 0, 0, 0, 1), f_1 = \frac{1}{13}$
$i = 4$	$(0, 0, 0, 0, 1) \notin IR$ $f_1 = \frac{1}{13}$	$(0, 0, \frac{1}{2}, 0, 0), f_1 = \frac{1}{7}$ $(0, 0, 0, \frac{3}{5}, \frac{2}{5}), f_1 = \frac{11}{53}$ $(0, 0, 0, 0, 0), f_1 = \frac{1}{8}$
$i = 5$	$(\frac{3}{4}, \frac{3}{4}, 0, 0, 1) \in IR$ $f_1 = \frac{4}{49}$	

Example 4. Consider

$$\begin{aligned}
 \min_{(x_1, x_2)} \quad & f_1 = \frac{1 + x_1 - x_2 + 2x_4}{8 - x_1 - 2x_3 + x_4 + 5x_5}, \\
 \min_{(x_3, x_4)} \quad & f_2 = \frac{1 + x_1 + x_2 + 2x_3 - x_4 + x_5}{6 + 2x_1 + x_3 + x_4 - 3x_5}, \\
 \min_{x_5} \quad & f_3 = \frac{2 + x_1 + x_2 - x_3 + x_4 + x_5}{1 - x_1 + x_2 - x_3 + 2x_4 + x_5}, \\
 s.t. \quad & \begin{cases} -x_3 + x_4 + x_5 \leq 1, \\ 2x_1 - x_3 + 2x_4 - \frac{1}{2}x_5 \leq 1, \\ 2x_2 + 2x_3 - x_4 - \frac{1}{2}x_5 \leq 1, \\ x_i \geq 0, \end{cases} \quad i = 1, \dots, 5.
 \end{aligned}$$

The result of the proposed K th-best algorithm is shown below. As seen in Table 2, the optimum solution $(\frac{3}{4}, \frac{3}{4}, 0, 0, 1)$ is obtained in the fifth iteration with the optimum value of $(f_1, f_2, f_3) = (\frac{4}{49}, \frac{7}{9}, \frac{9}{4}) = (0.0816, 0.777, 1.119)$.

Example 5. Consider

$$\begin{aligned}
\min_{x_1} \quad & f_1 = \frac{-x_1 - 5x_2 - x_3 + x_4 - 4}{x_1 + 12x_2 + x_3 + 2x_4}, \\
\min_{x_2} \quad & f_2 = \frac{-8x_1 - 7x_2 - 2x_3 - x_4 - 4}{9x_1 + 11x_2 + x_3 + x_4 + 5}, \\
\min_{x_3} \quad & f_3 = \frac{-12x_1 - 3x_2 - x_3 + x_4 - 4}{12x_1 + 5x_2 + x_3 + 2x_4 + 3}, \\
\min_{x_4} \quad & f_4 = \frac{-4x_1 - 2x_2 - x_3 - x_4 - 4}{5x_1 + 3x_2 + x_3 + x_4 + 5}, \\
s.t. \quad & \begin{cases} x_1 + x_2 + x_3 - x_4 \leq 3, & x_1 + x_2 + x_3 - x_4 \geq 1, \\ x_1 + x_2 - x_3 - x_4 \leq 7, & -x_1 + x_2 + x_3 - x_4 \leq 1, \\ x_2 \geq 0.25, & x_3 \leq 1, & x_4 \leq 3, & x_i \geq 0, \end{cases} \quad i = 1, 2, 3, 4.
\end{aligned}$$

Using the K th-best algorithm, we obtain the optimum solution $(0, \frac{1}{4}, \frac{3}{4}, 0)$ in the first iteration. The optimum value of $(f_1, \dots, f_4)$ is equal to $(-1.6, -0.88, -1.1, -0.808)$. However, Pal, Kumar, and Sen [17] found the optimum solution $(\frac{1}{2}, \frac{1}{4}, 1, \frac{1}{4})$ with the optimum value of $(f_1, \dots, f_4) = (-1.33, -0.96, -0.89, -0.82)$.

7 Conclusions

In this study, an MLLFP problem has been considered wherein common constraint region is a closed, and bounded polytope. Under mild conditions, it is proved that an optimum solution to the MLLFP problem occurs at a border-feasible vertex of the inducible region. Also, a former published paper on the MLLFP problems has been considered, and some drawbacks in that paper have been stated. Moreover, the K th-best algorithm has been developed for solving the MLLFP problem in question, and its convergence is proved. The efficiency of the proposed algorithm has been demonstrated by solving several numerical examples and comparing the results with those obtained by some existing methods. The study of the MLLFP problems over an unbounded constraint region will be addressed in our future research.

Acknowledgements

Authors are grateful to there anonymous referees and editor for their constructive comments. In addition, the financial support provided by the Iran Small Industries and Industrial Parks Organization (ISIPO) in facilitating this research has been greatly appreciated.

Funding

This research received no external funding.

Conflicts of Interest

The authors declare no conflicts of interest.

References

- [1] Alguacil, N., Delgadillo, A. and Arroyo, J.M. *A trilevel programming approach for electric grid defense planning*, Comput. Oper. Res., 41 (2014), 282–290.
- [2] Bhargava, Sh. *Solving linear fractional multi-level programs*, Oper. Res. Decis., 1 (2014), 5–21.
- [3] Calvete, H.I. and Galé, C. *Solving linear fractional bilevel programs*, Oper. Res. Lett., 32 (2004), 143–151.
- [4] Cormen, T., Leiserson, C., Rivest, R. and Stein, C. *Introduction to algorithms*, 3rd ed., MIT Press, Cambridge, MA, 2009.
- [5] Dempe, S. and Zemkoho, A. *Bilevel Optimization: Advances and challenges*, Springer, Cham, Switzerland, 2020.
- [6] Fakhry, R., Hassini, E., Ezzeldin, M. and El-Dakhakhni, W. *Trilevel mixed-binary linear programming solution approaches and application in defending critical infrastructure*, Eur. J. Oper. Res., 298(3) (2022), 1114–1131.
- [7] Fathy, E., Ammar, E. and Helmy, M.A. *Fully intuitionistic fuzzy multilevel linear fractional programming problem*, Alexandria Eng. J., 77 (2023), 684–694.
- [8] Florensa, G., Garcia-Herreros, P., Mirsa, P. et al. *Capacity planning with competitive decision-makers: Trilevel MILP formulation, degeneracy, and solution approaches*, Eur. J. Oper. Res., 262 (2017), 449–463.
- [9] Han, J., Lu, J. and Zhang, G. *Tri-level decision-making for decentralized vendor-managed inventory*, Inf. Sci., 421 (2017), 85–103.

- [10] Henke, D., Lefebvre, H. and Schmidt, M. *On coupling constraints in linear bilevel optimization*, Optim. Lett., 19(3) (2025), 689–697.
- [11] Ke, G.Y. and Bookbinder, J.H. *Coordinating the discount policies for retailer, wholesaler, and less-than-truckload carrier under price-sensitive demand: A trilevel optimization approach*, Int. J. Prod. Econ., 196 (2018), 82–100.
- [12] Lachwani, K. and Nehra, S. *Modified FGP approach and MATLAB program for solving multi-level linear fractional programming problems*, J. Ind. Eng. Int., 11 (2015), 15–36.
- [13] Liu, Y.H. and Hart, S.M. *Characterizing an optimal solution to the linear bilevel programming problem*, Eur. J. Oper. Res., 73(1) (1994), 164–166.
- [14] Martos, B. *Nonlinear programming: Theory and methods*, North-Holland, Amsterdam, 1975.
- [15] Murty, K.G. *Linear programming*, John Wiley & Sons, New York, 1983.
- [16] Ozkok, B.A. *An iterative algorithm to solve a linear fractional programming problem*, Comput. Ind. Eng., 140 (2020), 1–15.
- [17] Pal, B., Kumar, M. and Sen, S. *Priority based fuzzy goal programming approach for fractional multilevel programming problems*, Int. Rev. Fuzzy Math., 6 (2011), 1–14.
- [18] Sadeghi, H. and Esmaili, M. *A modification of the trilevel Kth-best algorithm*, AIMS Math., 3(4) (2018), 524–538.
- [19] Sadeghi, H. and Esmaili, M. *On the quasi-concave multi-level programming problem*, Asia-Pac. J. Oper. Res., 39(3) (2022), 2150026.
- [20] Sana, S.S. *A production-inventory model of imperfect quality products in a three-layer supply chain*, Decis. Support Syst., 50 (2011), 539–547.
- [21] Stancu-Minasian, I.M. *A ninth bibliography of fractional programming*, Optim., 68 (2019), 2125–2169.
- [22] Wang, G., Jiang, B., Zhu, K. and Wan, Z. *Global convergent algorithm for the bilevel linear fractional programming based on modified convex simplex method*, J. Syst. Eng. Electron., 21(2) (2010), 239–243.
- [23] Xu, X., Meng, Z. and Shen, R. *A tri-level programming model based on conditional value-at-risk for three-stage supply chain management*, Comput. Ind. Eng., 66 (2013), 470–475.



A mathematical model and optimal control analysis of merchant resistance to e-commerce adoption

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

Abstract

In this study, we develop a mathematical model that captures the dynamics of merchants' attitudes toward the adoption of e-commerce for trading activities. The merchant population is classified into four compartments: Traditional merchants, nonadopting merchants, active e-commerce users, and merchants who have temporarily suspended their online activities. The proposed model provides a quantitative framework for analyzing the persistence of e-commerce nonadoption among certain merchant groups and for identifying effective strategies to promote

*Corresponding author

Received 5 October 2025; revised 2 November 2025; accepted 19 November 2025

Habib Hassouni

Laboratory of Fundamental Mathematics and their Applications, Department of Mathematics, Faculty of Sciences El Jadida, Chouaib Doukkali University, El Jadida, Morocco. e-mail: habibhassouni85@gmail.com

Salaheddine Belhdid

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

Oumaima Aarabate

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

Omar Balatif

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

How to cite this article

Hassouni, H., Belhdid, S., Aarabate, O. and Balatif, O., A mathematical model and optimal control analysis of merchant resistance to e-commerce adoption. *Iran. J. Numer. Anal. Optim.*, 2026; 16(2): 356-384. <https://doi.org/10.22067/ijnao.2025.95799.1743>

digital market participation. To this end, an optimal control problem is formulated with the objective of enhancing e-commerce adoption through suitable intervention policies. The Pontryagin maximum principle is applied to derive the necessary conditions for optimality, and numerical simulations implemented in MATLAB illustrate and validate the theoretical findings. Furthermore, a cost-effectiveness analysis is conducted to assess the efficiency of the proposed strategies, providing insight into the balance between implementation costs and adoption outcomes.

AMS subject classifications (2020): Primary 49J20; Secondary 34D20.

Keywords: Mathematical modeling, Optimal control theory, E-commerce, Pontryagin maximum.

1 Introduction

E-commerce constitutes a fundamental pillar for national economic growth and integration into the global marketplace. It enhances the visibility of domestic products internationally, facilitates access to diverse markets, and provides consumers with the convenience of purchasing goods and services from virtually any location at any time. Such digital transformation not only contributes to greater efficiency in the use of resources, time, and effort but also indirectly supports public health and environmental objectives by reducing physical contact, pollution, and traffic congestion.

E-commerce broadly encompasses all activities related to the buying, selling, and exchange of goods and services conducted through the Internet using various digital devices such as computers, smartphones, and wearable technologies. In recent years, its global expansion has been remarkable, driven by technological innovation and increased connectivity. However, the effective implementation of electronic commerce remains dependent on the availability of robust communication infrastructures and the provision of adequate training for the management and operation of online commercial platforms.

E-commerce has encountered three major developments. The first was e-commerce based on EDI (Electronic Data Interchange) [33]. EDI is the exchange of files between business partners from one computer to another in a basic format. Generally, companies will have their systems to translate this format into that used by the firm itself [15]. The recipient of that file sent by another partner will be processed immediately as an invoice or a purchase order. The format, for its part, must contain the type of information that the document has. This system made it possible to improve purchasing processes as well as relationships throughout the supply chain [34].

Therefore, Internet-based e-commerce constituted the second phase. The primary drawbacks of EDI were its expensive price, its inability to adjust to B2C transactions, and its network that was underdeveloped outside of major multinational corporations. The smallest companies desired to take advantage of technology advancements as well. Transnational companies also required information to be shared at increasing rates. Then, the United States lifted the restriction on employing the Internet for business purposes in the 1990's. Consequently, e-commerce has seen brighter days and its first step into the future in 2004 [24]. Thus, more people started to use the Internet, which resulted in greater profits for merchants and industries [38]. Finally, greater development was noticed after the Internet became more popular in every household [1].

Lots of research are being made on this subject and governments are beginning to develop policies encouraging businesses to go digital. The emergence of shopping platforms such as Amazon [19], eBay [25], and Alibaba [39] made the experience of online shopping easier and worldwide.

However, the developments in e-commerce depend on its advantages and limitations [37]. e-commerce presents many advantages, unlike traditional commerce shops, due to its availability anytime anywhere [14, 20, 29]. A person can now purchase products whenever and wherever they choose, thanks to e-commerce, even from the comfort of their own home. Additionally, they have access to additional information about the features of the products and can use their computer to compare prices across multiple retailers. On the other hand, e-commerce encounters some limitations and drawbacks, especially for the government regarding the taxation problem [43]. Some academics investigate that point, where it is difficult to track the taxes on online shops [2].

Regarding the big importance that e-commerce takes in our lives and how it affects people's behavior, many academics investigated the relationship between people and e-commerce. In [6, 30], the authors investigated the correlation between e-commerce and people's trust in it, given the differences between a virtual and a conventional marketplace and how customers' trust affects their consumption and vision toward e-commerce. In that same vision, some other researchers discussed and illustrated the needs of users in e-commerce platforms regarding their age, education, and income [28]. Therefore, the important attention people and governments give to e-commerce opened many opportunities for entrepreneurs. Thus, academics gave attention to this topic and studied the relationship between e-commerce and entrepreneurship [36, 17].

Noting the important role e-commerce took in people's lives and the correlation it has with our behavior and consumption aptitude, it is primordial to better understand the process of e-commerce buying and selling, and how merchants and entrepreneurs act and invest. Consequently, a proper and effective modeling approach must be used, such as the compartmental model approach, which was first introduced by Kermack and McKendrick [23] in 1927 in the Proceedings of the Royal Society of London, which is a seminal work that laid the groundwork

for the mathematical modeling of infectious diseases, principally named the SIR models. Inspired by the epidemiological modeling approach, a population can be divided into three main compartments susceptible, infected and recovered.

In recent years, fractional-order derivatives have gained considerable attention in the context of biological and epidemiological modeling, owing to their ability to capture memory and hereditary effects inherent in real-world processes. Their nonlocal properties allow for a more realistic description of biological dynamics, particularly those involving anomalous diffusion or delayed responses. For instance, fractional models have been employed to investigate complex infection and epidemic dynamics, such as multi-strain influenza with treatment effects [42], COVID-19 transmission incorporating Caputo-type fractional operators [21], and within-host viral infection models accounting for adaptive immunity [40, 35]. These studies highlight the growing importance of fractional calculus in enhancing the flexibility and realism of classical compartmental models. This modeling perspective can be extended beyond biological systems to other domains governed by human behavior and interaction, such as socio-economic and technological diffusion processes, thereby motivating the adoption of similar analytical approaches in the study of e-commerce dynamics.

Statement of problem.

The application of control theory to various problems has been explored by several authors, including the spread of rumours on social media. In [7], a study is presented where the author formulated a mathematical model and applied control theory to reduce the spread of rumours on social networks. Additionally, in [3], the author applied control theory to address nonregistration behavior in an effort to increase the number of people registered on the electoral roll. Further research on control theory can be found in [9, 22, 10, 26, 4].

Recent advances in compartmental modeling and optimal control have highlighted their utility in analyzing complex social and behavioral dynamics. For example, El Bhih et al. [8] developed a framework to study the spread of rumors and counter-rumors on social networks, formulating multiple equilibria and designing optimal interventions using Pontryagin's maximum principle to influence collective behavior. Similarly, studies on scam rumor propagation have employed deterministic models combined with cost-effectiveness analysis to evaluate strategies for mitigating false information on social platforms, emphasizing the efficiency of targeted interventions [5]. In the context of epidemiology, Laaroussi, El Bhih, and Rachik [27] demonstrated how optimal vaccination and treatment policies can be formulated under constrained resources for a multi-strain SEIR model of COVID-19, providing insight into balancing limited resources while achieving maximal impact. Furthermore, Yaagoub, El Bhih, and Allal [41] introduced a fractional-order infection model for computer virus propagation, illustrating how memory effects and fractional dynamics can improve the realism and predictive power of compartmental models.

Collectively, these studies underscore the versatility of compartmental and optimal control approaches for understanding the evolution of complex systems and designing effective intervention strategies. Inspired by these works, our study applies a similar framework to investigate merchants' adoption of e-commerce, where the dynamics of acceptance and refusal can be influenced through targeted control measures and assessed quantitatively for cost-effectiveness.

Building on these approaches, Hassouni, Belhdid, and Balatif [16] applied compartmental and optimal control modeling to study customer behavior toward e-commerce. Their model focused on the dynamics of refusal and adoption, analyzed stability and parameter sensitivity, and employed Pontryagin's maximum principle to optimize interventions. Numerical simulations confirmed the effectiveness of the proposed strategies, highlighting the relevance of control-based modeling for understanding behavioral dynamics in digital marketplaces. Inspired by this, our study extends the framework to examine merchant behavior, capturing interactions between adopters and nonadopters in commercial ecosystems and evaluating intervention strategies through cost-effectiveness analysis.

In our research, we formulate a new mathematical model in order to study the influence of merchants' behavior on e-commerce and their refusal to use e-commerce. We study the basic characteristics of the proposed *CAEP* model to discover the various aspects of the proposed model and to better understand the process of store owner behavior towards e-commerce. We also apply control theory to this model in order to increase the number of merchants who use electronic commerce, which will reflect positively on them, whether in terms of income or increasing the number of customers. The spread of the electronic neighbor will inevitably lead to an increase in the number of commercial markets available to citizens. This will reflect positively on the state's economy.

The originality of the proposed *CAEP* model lies in its focus on the dynamics of merchants who refuse to adopt e-commerce, in addition to those who already use it, thereby capturing interactions between adopters and nonadopters within the commercial ecosystem. Unlike previous studies that primarily examine general adoption patterns or consumer behavior, our model integrates an optimal control framework to identify effective strategies for promoting e-commerce adoption among reluctant merchants. Furthermore, the incorporation of a cost-effectiveness analysis provides quantitative guidance for policymakers and decision-makers, allowing them to evaluate trade-offs between intervention costs and adoption outcomes. Collectively, these features distinguish our work from existing literature and offer new insights into targeted interventions to enhance e-commerce penetration.

The paper is organized as follows. Section 2 introduces a mathematical model that analyzes merchants' behavior with respect to e-commerce. In Section 3, optimal control theory is applied to the model, and the positivity and boundedness of the solutions are established. Section 4

presents numerical simulations to assess the effectiveness of the proposed strategies. Finally, concluding remarks are provided in Section 5.

2 Formulation of the mathematical model

The study of epidemic spread has enabled researchers to adopt epidemiological models as a foundation for modeling the dynamics of e-commerce. Our research offers a qualitative extension to previous studies by introducing a novel mathematical model that describes merchant behavior in relation to electronic commerce.

In this section, we focus on the issue of merchants who refuse to use e-commerce for buying or selling their products. To address this, we propose the mathematical model *CAEP*, which aims to analyze the evolution of e-commerce usage among merchants and investigate the reasons behind its limited adoption by certain segments.

Furthermore, the model examines how merchants-whether or not they currently use e-commerce-may be influenced through direct interaction with those who resist adopting it. The proposed model consists of four compartments:

Compartment *C*: These are merchants who sell goods to customers or buy goods to sell. It also includes traders who produce goods and then sell them. Individuals in this compartment are considered to be nonusers of e-commerce, as they do not have the knowledge or skills required to set up an online shop. The number of merchants in this compartment increases with the number Λ_1 , which represents the number of new merchants, and the number of merchants in this compartment decreases with the number θ , which represents the rate of natural deaths or cessation of trade, and $\gamma_1 CA$, which is the proportion of the number of traders affected by the refusal to use e-commerce as a result of contacts with traders who refuse to use e-commerce, and $\gamma_2 CE$, which is the proportion of the number of traders affected by the refusal to use e-commerce as a result of contacts with traders who use e-commerce, and $\theta_1 C$, which is the proportion of the number of traders influenced to refuse to use e-commerce as a result of negative thoughts about e-commerce, and $\theta_2 C$, which is the proportion of the number of traders influenced to use e-commerce as a result of positive thoughts about e-commerce. Hence

$$\frac{dC(t)}{dt} = \Lambda_1 - \gamma_1 C(t)A(t) - \theta_1 C(t) - \theta_2 C(t) - \gamma_2 C(t)E(t) - \theta C(t). \quad (1)$$

Compartment *A*: These are merchants who refuse to use e-commerce to sell or buy their products because of negative experiences with e-commerce such as fraud, cybercrime, or because

of negative thoughts about e-commerce coming from the merchant who refuses to use e-commerce or a person who has negative thoughts about e-commerce. The number of this compartment is increased by Λ_2 , which represents the number of new merchants who know about e-commerce but do not want to use it, and $\gamma_1 CA$, which represents the proportion of the number of merchants who have been affected by the refusal to use e-commerce as a result of contacting merchants who refuse to use e-commerce, and $\gamma_4 AE$, which represents the proportion of the number of merchants who were influenced by the refusal to use e-commerce as a result of direct contact with merchants who refuse to use e-commerce, and $\theta_1 C$, which represents the proportion of the number of merchants who were influenced by the refusal to use e-commerce as a result of negative thoughts about e-commerce, and $\theta_2 E$ is the number of merchants who used e-commerce and were affected by the refusal to use e-commerce as a result of negative thoughts about e-commerce. The number for this compartment is reduced by θ , representing natural deaths or stopping trade, and $\gamma_3 AE$, which represents the proportion of the number of merchants who refused to use e-commerce and were influenced by e-commerce as a result of contact with merchants who use e-commerce, and $\theta_1 A$, which represents the proportion of the number of merchants who refused to use e-commerce and were influenced by e-commerce as a result of positive thoughts about e-commerce received from a customer or community member. Therefore

$$\begin{aligned} \frac{dA(t)}{dt} = & \Lambda_2 + \gamma_1 C(t)A(t) + \theta_1 C(t) + \gamma_4 A(t)E(t) + \theta_2 E(t) - \gamma_3 A(t)E(t) \\ & - \theta_1 A(t) - \theta A(t). \end{aligned} \quad (2)$$

Compartment E : These are merchants who use e-commerce to sell or buy their products because they know the pros and cons of e-commerce.

The number of this compartment is increased by Λ_3 , which represents the proportion of new merchants with e-commerce stores, and $\theta_2 C$, which represents the number of merchants who have been influenced by e-commerce as a result of positive thoughts about e-commerce coming from customers or people, $\gamma_2 CE$, which represents the number of merchants who have been influenced by e-commerce as a result of direct contact with merchants who use e-commerce, and $\gamma_3 AE$, which represents the number of merchants who declined to use e-commerce and were influenced by e-commerce as a result of contact with merchants who use e-commerce, and $\theta_1 A$, which represents the number of merchants who declined to use e-commerce and were influenced by e-commerce as a result of positive ideas about e-commerce received from the customer or a member of the community, and $\gamma_2 P$, which represents the number of merchants who return to using e-commerce after solving the problems they faced, such as maintenance. The number of members of this compartment decreases with the number θ , which represents natural deaths or stopping trade,

and $\gamma_4 AE$, which represents the proportion of the number of merchants who used e-commerce and were influenced by the refusal to use e-commerce as a result of contacting merchants who refuse to use e-commerce, and $\theta_2 E$, which represents the proportion of the number of merchants who have been affected by the refusal to use e-commerce as a result of negative thoughts about e-commerce, and $\gamma_1 E$, which represents the proportion of the number of merchants who have stopped using e-commerce due to problems they face, such as the difficulty of maintaining the site. Thus

$$\begin{aligned} \frac{dE(t)}{dt} = & \Lambda_3 + \theta_2 C(t) + \gamma_2 C(t)E(t) + \gamma_3 A(t)E(t) + \theta_1 A(t) + \gamma_2 P(t) \\ & - \gamma_4 A(t)E(t) - \theta_2 E(t) - \gamma_1 E(t) - \theta E(t). \end{aligned} \quad (3)$$

Compartment P: These are merchants who have temporarily stopped using e-commerce to sell or buy their products due to problems they have encountered on their websites, such as damage to the site or its maintenance, lack of human capacity to work on the site or, in some cases, the absence of the Internet. The number of merchants in this category increases with the number $\gamma_1 E$, which represents the percentage of merchants who stopped using e-commerce because of problems they encountered, such as difficulties in maintaining the site, and decreases with the number θ , which represents the percentage of merchants who died of natural causes or stopped trading, and with the number $\gamma_2 P$, which represents the percentage of merchants who returned to using e-commerce after solving the problems they encountered, such as maintenance. Then

$$\frac{dP(t)}{dt} = \gamma_1 E(t) - \gamma_2 P(t) - \theta P(t). \quad (4)$$

The chart below will show the trends of merchant flow between compartments. These directions will be represented by directed arrows to illustrate these movements; see the following Figure 1.

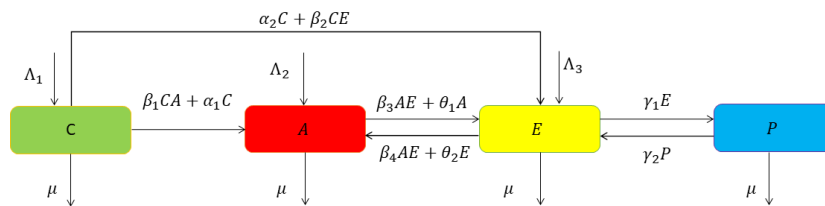


Figure 1: A diagram of the evolution of the transfer of e-commerce between merchants.

Based on the previously described compartments and the interactions between them, we now present the mathematical model that captures the dynamics of e-commerce usage rejection. The model is governed by the following system of differential equations:

$$\begin{cases} \frac{dC(t)}{dt} = \Lambda_1 - \gamma_1 C(t)A(t) - \theta_1 C(t) - \theta_2 C(t) - \gamma_2 C(t)E(t) - \theta C(t), \\ \frac{dA(t)}{dt} = \Lambda_2 + \gamma_1 C(t)A(t) + \theta_1 C(t) + \gamma_4 A(t)E(t) + \theta_2 E(t) - \gamma_3 A(t)E(t) \\ \quad \theta_1 A(t) - \theta A(t), \\ \frac{dE(t)}{dt} = \Lambda_3 + \theta_2 C(t) + \gamma_2 C(t)E(t) + \gamma_3 A(t)E(t) + \theta_1 A(t) + \gamma_2 P(t) \\ \quad \gamma_4 A(t)E(t) - \theta_2 E(t) - \gamma_1 E(t) - \theta E(t), \\ \frac{dP(t)}{dt} = \gamma_1 E(t) - \gamma_2 P(t) - \theta P(t). \end{cases} \quad (5)$$

With the initial conditions

$$C(0) \geq 0, \quad A(0) \geq 0, \quad E(0) \geq 0, \quad P(0) \geq 0,$$

the total number of merchants at time t is denoted as $N(t)$. Therefore, we have

$$N(t) = C(t) + A(t) + P(t) + E(t).$$

The parameters used in this model are summarized in Table 1.

2.1 Positivity of solutions

Since the model (5) tracks the behavior of merchants, it is important that all state variables and associated parameters remain positive. This will be verified by the following theorem.

Theorem 1. If $C(0) \geq 0$, $A(0) \geq 0$, $E(0) \geq 0$, and $P(0) \geq 0$, then the solutions $C(t)$, $A(t)$, $E(t)$, and $P(t)$ of system (5) will be positive for all $t \geq 0$.

Proof. From the first equation of system (5), we obtain the inequality

$$\frac{dC(t)}{dt} + (\theta + \theta_1 + \theta_2 + \gamma_1 A(t) + \gamma_2 E(t)) C(t) \geq \Lambda_1.$$

Rearranging this inequality, we have

$$\frac{dC(t)}{dt} + (\theta + \theta_1 + \theta_2 + \gamma_1 A(t) + \gamma_2 E(t)) C(t) \geq 0. \quad (6)$$

Table 1: Parameters definitions of the *CAEP* model.

Parameter	Description
Λ_1	The number of new merchants introduced to the market. (individual)
Λ_2	The number of new merchants not using e-commerce. (individual)
Λ_3	The number of new merchants using e-commerce. (individual)
γ_1	The rate at which a potential merchant adopts the idea of refusing e-commerce by contacting individuals from the against group. (individual/day)
γ_2	The rate at which a potential merchant adopts the idea of using e-commerce by contacting individuals from the e-commerce users' group. (individual/day)
γ_3	The rate at which a refusing merchant turns to a user of e-commerce by contacting e-commerce users. (individual/day)
γ_4	The rate at which an e-commerce user switches to not using it by contacting individuals from the against group. (individual/day)
θ_1	The rate of merchants who were affected by negative ideas about e-commerce. (individual/day)
θ_2	The rate of merchants who were influenced by positive ideas about e-commerce. (individual/day)
θ_1	The rate of merchants who refused to use e-commerce but were influenced by positive ideas about it. (individual/day)
θ_2	The rate of merchants who use e-commerce but are affected by negative thoughts about it. (individual/day)
γ_1	The rate of merchants who used e-commerce but temporarily stopped using it. (individual/day)
γ_2	The rate of merchants who temporarily stopped using e-commerce and returned to it. (individual/day)
θ	The rate of final departure from the compartment. (individual/day)

Multiplying both sides of inequality (6) by the factor

$$\exp \left(\int_0^t (\theta + \theta_1 + \theta_2 + \gamma_1 A(h) + \gamma_2 E(h)) dh \right),$$

we get

$$\begin{aligned} & \frac{dC(t)}{dt} \cdot \exp \left(\int_0^t (\theta + \theta_1 + \theta_2 + \gamma_1 A(h) + \gamma_2 E(h)) dh \right) \\ & + (\theta + \theta_1 + \theta_2 + \gamma_1 A(t) + \gamma_2 E(t)) \\ & \exp \left(\int_0^t (\theta + \theta_1 + \theta_2 + \gamma_1 A(h) + \gamma_2 E(h)) dh \right) \cdot C(t) \geq 0. \end{aligned}$$

This simplifies to

$$\frac{d}{dt} \left(C(t) \cdot \exp \left(\int_0^t (\theta + \theta_1 + \theta_2 + \gamma_1 A(h) + \gamma_2 E(h)) dh \right) \right) \geq 0, \quad \forall t \geq 0.$$

Integrating the inequality, we obtain

$$C(t) \geq C(0) \cdot \exp \left(\int_0^t -(\theta + \theta_1 + \theta_2 + \gamma_1 A(h) + \gamma_2 E(h)) \, dh \right).$$

Thus, $C(t)$ remains positive, as it is greater than or equal to the initial condition $C(0)$.

Similarly, we can prove that $A(t) \geq 0$, $E(t) \geq 0$, and $P(t) \geq 0$ for all $t \geq 0$.

□

2.2 Boundedness of solutions

In this subsection, we present a result which proves that all possible solutions of the system description model (5) are bounded.

Lemma 1. Let the region Ω be defined by

$$\Omega = \left\{ (C(t), A(t), E(t), P(t)) \in \mathbb{R}_+^4 \mid 0 \leq C(t) + A(t) + E(t) + P(t) \leq \frac{\Lambda_1 + \Lambda_2 + \Lambda_3}{\theta} \right\},$$

where the initial conditions are given as $C(0) \geq 0$, $A(0) \geq 0$, $E(0) \geq 0$, and $P(0) \geq 0$. Then the region Ω is positively invariant for the model (5).

Proof. We have the following differential inequality for the total number of merchants:

$$\frac{dN(t)}{dt} \leq \Lambda_1 + \Lambda_2 + \Lambda_3 - \theta N(t),$$

hence, solving this inequality gives

$$\frac{dN(t)}{dt} \leq N(0) \exp(-\theta t) + \left(\frac{\Lambda_1 + \Lambda_2 + \Lambda_3}{-\theta} \right) (\exp(-\theta t) - 1).$$

Thus, we obtain the asymptotic bound

$$\lim_{t \rightarrow +\infty} \sup N(t) = \frac{\Lambda_1 + \Lambda_2 + \Lambda_3}{\theta},$$

and consequently,

$$N(t) \leq \frac{\Lambda_1 + \Lambda_2 + \Lambda_3}{\theta}.$$

From this, we conclude that the set Ω is positively invariant for the system (5) [9].

□

2.3 Existence of solutions

In the following theorem, we prove the existence and uniqueness of solutions.

Theorem 2. The system (5) that fulfills the initial conditions $(C(0), A(0), E(0), P(0))$ has a unique solution.

Proof. Let

$$X(t) = \begin{pmatrix} C(t) \\ A(t) \\ E(t) \\ P(t) \end{pmatrix} \quad \text{and} \quad \Gamma(X) = \begin{pmatrix} \frac{dC(t)}{dt} \\ \frac{dA(t)}{dt} \\ \frac{dE(t)}{dt} \\ \frac{dP(t)}{dt} \end{pmatrix}.$$

Also,

$$\Gamma(X) = F_1 X + F_2(X), \quad (7)$$

in which

$$F_1 = \begin{bmatrix} -(\theta + \theta_1 + \theta_2) & 0 & 0 & 0 \\ \theta_1 & \theta_2 & -(\theta + \theta_1) & 0 \\ \theta_2 & \theta_1 - (\theta + \theta_2 + \gamma_1) & \gamma_2 & \\ 0 & 0 & \gamma_1 & -(\theta + \gamma_2) \end{bmatrix},$$

and

$$F_2 = \begin{bmatrix} \Lambda_1 - \gamma_1 CA - \gamma_2 CE \\ \Lambda_2 + \gamma_1 CA + \gamma_4 AE - \gamma_3 AE \\ \Lambda_3 + \gamma_2 CE + \gamma_3 AE - \gamma_4 AE \\ 0 \end{bmatrix}.$$

Then

$$\begin{aligned} & \|F_2(Y) - F_2(Z)\| \\ & \leq |\gamma_1 C_Y A_Y + \gamma_2 C_Y E_Y - \gamma_1 C_Z A_Z - \gamma_2 C_Z A_Z| \\ & \quad + |\gamma_1 C_Y A_Y + \gamma_4 A_Y E_Y - \gamma_3 A_Y E_Y - \gamma_1 C_Z A_Z - \gamma_4 A_Z E_Z + \gamma_3 A_Z E_Z| \\ & \quad + |\gamma_2 C_Y E_Y + \gamma_3 A_Y E_Y - \gamma_4 A_Y E_Y - \gamma_2 C_Z E_Z - \gamma_3 A_Z E_Z + \gamma_4 A_Z E_Z|, \\ & \leq \frac{\Lambda}{\theta} (2(\gamma_1 + \gamma_2) |A_Y - A_Z| + 6\gamma_1 |A_Y - A_Z| + (6\gamma_2 + 3\gamma_3 + 4\gamma_4) |E_Y - E_Z|), \\ & \leq C \|Y - Z\|. \end{aligned}$$

Thus, we have

$$M = \frac{\Lambda_1 + \Lambda_2 + \Lambda_3}{\theta} \cdot \max(2(\gamma_1 + \gamma_2), 6\gamma_1, (6\gamma_2 + 3\gamma_3 + 4\gamma_4)),$$

and also

$$\|\Gamma(Y) - \Gamma(Z)\| \leq N \|Y - Z\|,$$

where $N = \max(M, \|F_1\|) < \infty$. Thus, it follows that the function Γ is Lipschitz continuous, so the system admits a unique solution [12].

□

3 The optimal control problem

The failure of merchants to use e-commerce leads to a set of negative effects that hinder economic growth in terms of energy bills. Abstaining from e-commerce also contributes to increasing customers' daily expenses, as the cost of traveling to purchase a product is very high compared to purchasing via e-commerce. The weak use of e-commerce also leads to the spread of infectious diseases.

E-commerce offers many benefits, both for merchants who use it, and for customers who purchase products via e-commerce, as it allows e-store owners to display their goods and services to customers at any time of the day or night. They can also open an e-store without having a physical store and work from home.

- $v_1(t)$: represents the use of advertising across all visual and audio communication means and social media to introduce e-commerce, in addition to the steps that must be taken to create an e-commerce site in a way that is safe from cybercrimes. The entities responsible for building websites can also be displayed safely and at reasonable prices. It is also necessary to provide training for merchants on how to operate websites and how to display products for sale in a good way that meets the customer's needs.

- $v_2(t)$: indicates the provision of insurance for all merchants against all cybercrimes that they may be exposed to, such as fraud and theft, in addition to theft of personal information.

By adding the two controls $v_1(t)$ and $v_2(t)$ to the first system (5), we obtain the following control model:

$$\left\{ \begin{array}{l} \frac{dC(t)}{dt} = \Lambda_1 - \gamma_1 C(t) A(t) - \theta_1 C(t) - \theta_2 C(t) E(t) - \theta C(t) \\ \quad - v_1(t) C(t), \\ \frac{dA(t)}{dt} = \Lambda_2 + \gamma_1 C(t) A(t) + \theta_1 C(t) + \gamma_4 A(t) E(t) + \theta_2 E(t) \\ \quad - \gamma_3 A(t) E(t) - \theta_1 A(t) - \theta A(t) - v_2(t) A(t) - v_2(t) \theta_2 E(t) \\ \quad - \gamma_4 v_2(t) A(t) E(t), \\ \frac{dE(t)}{dt} = \Lambda_3 + \theta_2 C(t) + \gamma_2 C(t) E(t) + \gamma_3 A(t) E(t) + \theta_1 A(t) + \gamma_2 P(t) \\ \quad - \gamma_4 A(t) E(t) - \theta_2 E(t) - \gamma_1 E(t) - \theta E(t) + v_1(t) C(t) \\ \quad + v_2(t) A(t) + \theta_2 v_2(t) E(t) + \gamma_4 v_2(t) A(t) E(t), \\ \frac{dP(t)}{dt} = \gamma_1 E(t) - \gamma_2 P(t) - \theta P(t), \end{array} \right. \quad (8)$$

with the initial conditions $C(0) \geq 0$, $A(0) \geq 0$, $E(0) \geq 0$, and $P(0) \geq 0$.

3.1 The optimal control: Existence and characterization

In this subsection, our objective is to minimize the total number of merchants who refuse to use e-commerce and maximize the number of those who use it. The control variables $v_1(t)$ and $v_2(t)$ are introduced to minimize the optimal control model (5) subject to the objective functional defined as follows:

$$J(v_1, v_2) = A(t_f) - E(t_f) + \int_{t_0}^{t_f} \left[A(t) - E(t) + \frac{C_1}{2} (v_1(t))^2 + \frac{C_2}{2} (v_2(t))^2 \right] dt,$$

where $C_1 > 0$ and $C_2 > 0$ are the cost coefficients. These coefficients are chosen in order to weigh the relative importance of $v_1(t)$ and $v_2(t)$ at each time t , where t_f is the final time. In other words, we aim to find the optimal controls v_1^* and v_2^* such that

$$J(v_1^*, v_2^*) = \min_{(v_1, v_2) \in A_d} J(v_1, v_2).$$

Here, A_d is the set of admissible controls defined by

$$A_d = \{(v_1(t), v_2(t)) : 0 \leq v_1(t) \leq 1, 0 \leq v_2(t) \leq 1 \text{ for } t \in [t_0, t_f]\}. \quad (9)$$

3.2 Existence of optimal controls

In this subsection, we present the theorem which proves the existence of an optimal control (v_1^*, v_2^*) minimizing the cost function J .

Theorem 3. There exists an optimal control $(v_1^*, v_2^*) \in A_d$ such that

$$J(v_1^*, v_2^*) = \min_{(v_1, v_2) \in A_d} J(v_1, v_2).$$

Proof. To use the existence result in [11], we must check the following properties:

(K1): The set of controls and the corresponding state variables is nonempty.

(K2): The control set A_d is convex and closed.

(K3): The right-hand side of the state system is bounded by a linear function in the state and control variables.

(K4): The integral $L(A, E, v_1, v_2)$ of the objective functional is convex on A_d , and there exist constants $\kappa_1 > 0$, $\kappa_2 > 0$, and $\varepsilon > 1$ such that

$$L(A, E, v_1, v_2) \geq -\kappa_1 + \kappa_2 (|v_1|^2 + |v_2|^2)^{\frac{\varepsilon}{2}}.$$

The first condition (K_1) is verified using the result in [11]. The set A_d is convex and closed by definition, thus the condition (K_2) . Our state system is bilinear in v_1 and v_2 , moreover the solutions of the system are bounded as proved in model (5), hence the condition (K_3) . Also, we have the last needed condition (K_4) :

$$L(A, E, v_1, v_2) \geq -\kappa_1 + \kappa_2 (|v_1|^2 + |v_2|^2)^{\frac{\varepsilon}{2}}.$$

where

$$\kappa_1 = 2 \sup_{t \in [t_0, t_f]} (A(t), E(t)), \quad \kappa_2 = \inf \left(\frac{C_1}{2}, \frac{C_2}{2} \right), \varepsilon = 2,$$

since

$$C_1 > 0 \quad \text{and} \quad C_2 > 0.$$

We conclude that there exists an optimal control $(v_1^*, v_2^*) \in A_d$ such that

$$J(v_1^*, v_2^*) = \min_{(v_1, v_2) \in A_d} J(v_1, v_2).$$

□

3.3 Characterization of the optimal controls

In this subsection, we apply Pontryagin's maximum principle to address the problem of merchants and businesses that are capable of using e-commerce but choose not to. Our objective is to solve for the optimal control in system (8) such that the control variables $v_1(t)$ and $v_2(t)$ represent the optimal strategies, ultimately reducing the number of merchants and businesses that refrain from using e-commerce to sell their products.

To derive the optimal control conditions, we utilize Pontryagin's maximum principle. The Hamiltonian $\mathcal{H}(t)$ at time t is given by

$$\mathcal{H}(t) = A(t) - E(t) + \frac{C_1}{2}v_1(t)^2 + \frac{C_2}{2}v_2(t)^2 + \sum_{i=1}^4 \lambda_i h_i,$$

where h_i represents the right-hand side of the system of differential equations (8) for the i -th state variable.

Theorem 4. Given the optimal controls (v_1^*, v_2^*) and the corresponding solutions C^* , A^* , E^* , and P^* to the state system (8), there exist adjoint functions λ_1 , λ_2 , λ_3 , and λ_4 that satisfy the following system:

$$\left\{ \begin{array}{l} \frac{d\lambda_1}{dt} = -\frac{\partial \mathcal{H}}{\partial C} \\ \quad = -\lambda_1 [-\gamma_1 A(t) - \theta_1 - \theta_2 - \gamma_2 E(t) - \theta - v_1(t)] \\ \quad \quad - \lambda_2 [\gamma_1 A(t) + \theta_1] - \lambda_3 [\theta_2 + \gamma_2 E(t) + v_1(t)], \\ \frac{d\lambda_2}{dt} = -\frac{\partial \mathcal{H}}{\partial A} \\ \quad = -1 - \lambda_1 (-\gamma_1 C(t)) \\ \quad \quad - \lambda_2 [\gamma_1 C(t) + \gamma_4 E(t) - \gamma_3 E(t) - \theta_1 - \theta - v_2(t)] \\ \quad \quad - \lambda_2 [-\gamma_4 v_2(t) E(t)] \\ \quad \quad - \lambda_3 [\gamma_3 E(t) + \theta_1 - \gamma_4 E(t) + v_2(t) + \gamma_4 v_2(t) E(t)], \\ \frac{d\lambda_3}{dt} = -\frac{\partial \mathcal{H}}{\partial E} \\ \quad = 1 - \lambda_1 (-\gamma_2 C(t)) \\ \quad \quad - \lambda_2 [\gamma_4 A(t) + \theta_2 - \gamma_3 A(t) - v_2(t)\theta_2 - \gamma_4 v_2(t) A(t)] \\ \quad \quad - \lambda_3 [\gamma_2 C(t) + \gamma_3 A(t) - \gamma_4 A(t) - \theta_2 - \gamma_1 - \theta \\ \quad \quad \quad + \theta_2 v_2(t) + \gamma_4 v_2(t) A(t)] \\ \quad \quad - \lambda_4 (\gamma_1), \\ \frac{d\lambda_4}{dt} = -\frac{\partial \mathcal{H}}{\partial P} \\ \quad = -\lambda_3 (\gamma_2) - \lambda_4 (-\gamma_2 - \theta). \end{array} \right. \quad (10)$$

The transversality conditions at time t_f are given by

$$\begin{cases} \lambda_1(t_f) = 0, \\ \lambda_2(t_f) = 1, \\ \lambda_3(t_f) = -1, \\ \lambda_4(t_f) = 0. \end{cases} \quad (11)$$

Moreover, for $t \in [t_0, t_f]$, the optimal controls $v_1^*(t)$ and $v_2^*(t)$ are given by

$$\begin{cases} v_1^*(t) = \min \left(1, \max \left(0, \frac{1}{C_1} (\lambda_1 - \lambda_3) C(t) \right) \right), \\ v_2^*(t) = \min \left(1, \max \left(0, \frac{\lambda_2 - \lambda_3}{C_2} (A(t) + \theta_2 E(t) + \gamma_4 A(t) E(t)) \right) \right). \end{cases} \quad (12)$$

Proof. The Hamiltonian $\mathcal{H}$ is defined as

$$\mathcal{H}(t) = A(t) - E(t) + \frac{C_1}{2} v_1(t)^2 + \frac{C_2}{2} v_2(t)^2 + \sum_{i=1}^4 \lambda_i h_i,$$

where the h_i functions are given by

$$\left\{ \begin{array}{l} h_1 = \Lambda_1 - \gamma_1 C(t) A(t) - \theta_1 C(t) - \theta_2 C(t) - \gamma_2 C(t) E(t) - \theta C(t) \\ \quad - v_1(t) C(t), \\ h_2 = \Lambda_2 + \gamma_1 C(t) A(t) + \theta_1 C(t) + \gamma_4 A(t) E(t) + \theta_2 E(t) - \gamma_3 A(t) E(t) \\ \quad - \theta_1 A(t) - \theta A(t) - v_2(t) A(t) - v_2(t) \theta_2 E(t) - \gamma_4 v_2(t) A(t) E(t), \\ h_3 = \Lambda_3 + \theta_2 C(t) + \gamma_2 C(t) E(t) + \gamma_3 A(t) E(t) + \theta_1 A(t) + \gamma_2 P(t) \\ \quad - \gamma_4 A(t) E(t) - \theta_2 E(t) - \gamma_1 E(t) - \theta E(t) + v_1(t) C(t) + v_2(t) A(t) \\ \quad + \theta_2 v_2(t) E(t) + \gamma_4 v_2(t) A(t) E(t), \\ h_4 = \gamma_1 E(t) - \gamma_2 P(t) - \theta P(t). \end{array} \right.$$

By using Pontryagin's maximum principle [31], the adjoint equations and transversality conditions can be derived, leading to the system in equation (10) and the optimal controls expressed in equation (12).

The optimal controls $v_1^*(t)$ and $v_2^*(t)$ are determined by solving the optimality conditions. For v_1 and v_2 , we have

$$\frac{\partial \mathcal{H}}{\partial v_1} = C_1 v_1(t) - \lambda_1 C(t) + \lambda_3 C(t) = 0,$$

which gives

$$v_1 = \frac{C(t)}{C_1} (\lambda_1 - \lambda_3).$$

Similarly, for v_2 ,

$$\begin{aligned} \frac{\partial \mathcal{H}}{\partial v_2} &= C_2 v_2(t) + \lambda_2 (-A(t) - \theta_2 E(t) - \gamma_4 A(t) E(t)) \\ &\quad + \lambda_3 (A(t) + \theta_2 E(t) + \gamma_4 A(t) E(t)) \\ &= 0, \end{aligned}$$

which gives

$$v_2 = \frac{\lambda_2 - \lambda_3}{C_2} (A(t) + \theta_2 E(t) + \gamma_4 A(t) E(t)).$$

These expressions for v_1^* and v_2^* are bounded by the control limits, and hence we obtain the final form of the optimal controls in equation (12). $\square$

4 Numerical simulations

In this section, we perform numerical simulations to illustrate the strategies we have implemented. The goal is to reduce the number of merchants and companies that do not use e-commerce to sell their products, while increasing the number of merchants that adopt e-commerce. To solve the proposed *CAEP* model and the corresponding optimal control problem, numerical simulations were performed using MATLAB. The state equations were solved using a standard fourth-order Runge–Kutta method, which is widely used in the literature for compartmental and epidemiological models due to its stability and accuracy. The optimal control problem was addressed via the forward-backward sweep method, which iteratively solves the state system forward in time and the adjoint system backward in time until convergence is achieved. Parameter values and initial conditions were selected based on realistic assumptions and prior studies to reflect typical merchant behavior and e-commerce adoption dynamics. Figures presented in this section illustrate the temporal evolution of the merchant compartments under different control scenarios, and the results are interpreted to assess the effectiveness of the proposed strategies in promoting e-commerce adoption. This approach is consistent with methodologies employed in recent studies on optimal control in socio-economic and epidemiological systems [32].

The code is written and compiled in MATLAB using the following data to represent $N = 1,071,900$ companies in France in 2022 [18], noting that this number increased by +17%

compared to the previous year. During the same year, the number of online sales sites in France is expected to continue increasing. It is predicted that over 10000 new active sites will emerge, marking a 5% increase compared to the previous year. By the end of 2022, there were 207000 electronic stores marketing commercial activities [13].

To apply the numerical simulation to the application, we use the initial data for the number of people in each compartment as follows: $C(0) = 700000$, $A(0) = 159900$, $E(0) = 207000$, $P(0) = 50000$, $\Lambda_1 = 73000$, $\Lambda_2 = 4797$, $\Lambda_3 = 10000$, $\gamma_1 = 10^{-8}$, $\gamma_2 = 2 \times 10^{-9}$, $\gamma_3 = 3 \times 10^{-8}$, $\gamma_4 = 10^{-8}$, $\theta_1 = 7 \times 10^{-11}$, $\theta_2 = 6 \times 10^{-7}$, $\gamma_1 = 6 \times 10^{-8}$, $\gamma_2 = 5 \times 10^{-8}$, and $\theta = 0.078$, $\theta_1 = 0.0001$, $\theta_2 = 0.0002$.

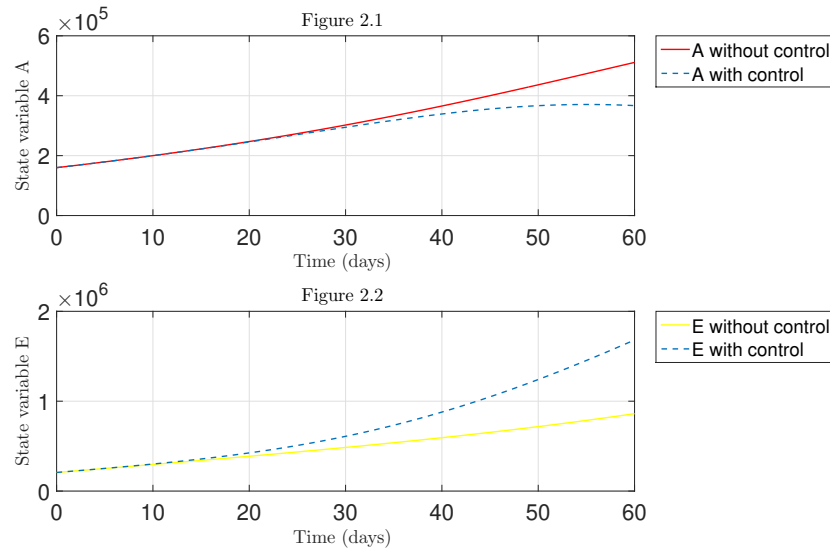
In this study, we have implemented two different control strategies, denoted $v_1(t)$ and $v_2(t)$, aimed at informing merchants and e-commerce companies about the protective measures available to them in case of exposure to cybercrime or fraud. To assess the specific impact of each control strategy, we have chosen to evaluate three scenarios:

4.1 First scenario: Motivating merchants to use e-commerce through advertising

In this subsection, the control strategy v_1 is applied to assist merchants and companies who may have limited experience in e-commerce and online store creation and management. The objective is to enable them to successfully create and operate an online store.

■ Figure 2.1 presents the dynamics of merchant behavior under the implementation of the first control strategy. Following the initiation of the strategy, a clear decline is observed in the number of merchants who initially refused to adopt e-commerce, represented in compartment A , beginning around day 20. By day 30, this number decreases by approximately 4.68%, indicating an early response of the nonadopting merchants to the applied intervention. This reduction appears to be correlated with a corresponding decrease in the population of nonadopting merchants in compartment C , suggesting that a portion of these individuals have been successfully influenced by the strategy to reconsider their adoption decisions.

Over the full 60-day simulation period, the cumulative effect of the first strategy results in a substantial 24% reduction in the number of merchants resistant to e-commerce adoption. This trend demonstrates the effectiveness of the control measures in shifting merchant behavior from nonadoption toward adoption. The results also underscore the importance of timely interventions: earlier implementation of the strategy leads to faster reductions in nonadopters, which can subsequently accelerate overall market penetration. Moreover, the gradual decline indicates

Figure 2: Dynamics with control v_1 .

that while the strategy is effective, behavioral change among merchants is a progressive process, influenced by factors such as peer adoption, perceived benefits, and exposure to incentives. Overall, the figure highlights the potential of targeted policy interventions to promote e-commerce adoption and improve the overall efficiency and competitiveness of the commercial ecosystem.

■ Figure 2.2 illustrates the impact of the first control strategy on the population of merchants actively using e-commerce, represented in compartment E . Following the implementation of the strategy, a noticeable increase in the number of adopting merchants is observed beginning around day 14. By day 30, the population of e-commerce users has risen by approximately 59.76%, indicating a rapid initial response to the intervention. This early increase suggests that the strategy effectively motivates merchants who were previously hesitant to transition toward adoption, likely due to perceived benefits or incentives embedded within the control measures.

Over the full 60-day simulation period, the cumulative effect of the strategy leads to a substantial overall increase in the total number of merchants using e-commerce. This trend highlights the effectiveness of the applied measures in promoting sustained adoption over time. Moreover, the gradual growth trajectory emphasizes that while the strategy yields significant short-term gains, continued monitoring and adaptive interventions may be necessary to maintain momentum and encourage long-term adherence. Overall, the figure demonstrates that targeted control strategies can successfully enhance e-commerce adoption, thereby improving market coverage, competitiveness, and economic efficiency.

4.2 Second scenario: Providing insurance against electronic crimes for merchants to encourage them to use electronic commerce.

In this strategy, we apply the control v_2 , which represents the display of the insurance element on compartment A .

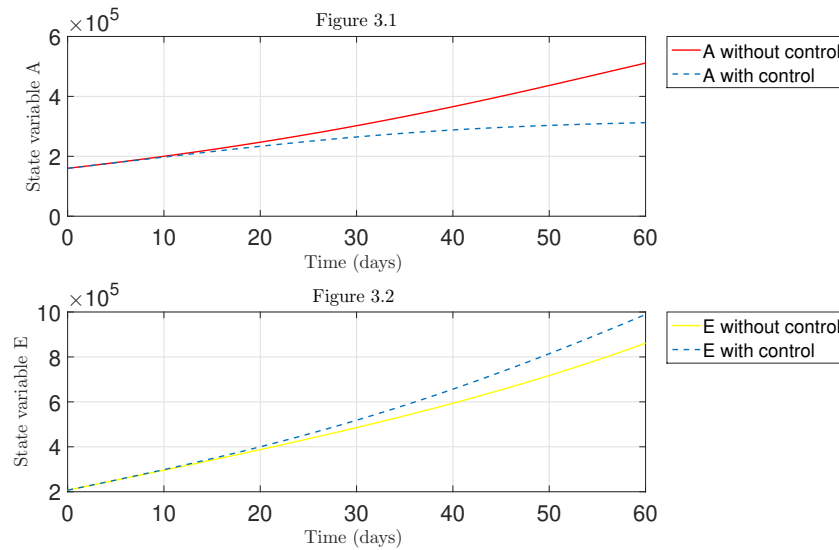


Figure 3: Dynamics with control v_2 .

■ Figure 3.1 illustrates the effect of the second control strategy on merchants resistant to e-commerce adoption in compartment A . The results show that the number of nonadopting merchants begins to decline from day 10, with a 23.66% reduction observed by day 30. By the end of the 60-day period, the total number of resistant merchants decreases by 48%, indicating that the second strategy is particularly effective in accelerating adoption. This rapid and substantial reduction suggests that targeted interventions addressing specific barriers can significantly influence merchant behavior, promoting faster integration into e-commerce. These findings provide valuable insights for policymakers and business stakeholders, highlighting the potential of well-designed strategies to reduce resistance, enhance market participation, and strengthen overall competitiveness in the commercial system.

■ Figure 3.2 shows the response of merchants adopting e-commerce in compartment E following the implementation of the second control strategy. The simulation indicates that adoption begins to rise noticeably from day 15, reaching a 15.85% increase by day 30. By the end

of the 60-day period, the total number of merchants using e-commerce has grown by 61.63%, demonstrating the strong effectiveness of this strategy. The progressive increase highlights how well-targeted interventions can gradually shift merchant behavior, encouraging hesitant individuals to embrace digital commerce. These results also emphasize the importance of timing and sustained application of control measures, as consistent efforts over the simulation period produce substantial adoption gains, thereby contributing to a more competitive and digitally integration.

4.3 Third scenario: Using advertising and insurance to encourage merchants to use electronic commerce.

In this strategy, we implement the controls v_1 and v_2 so that we can show merchants and companies who do not know how to use e-commerce to sell their products. How can they create online stores or how can they sell their products on some websites dedicated to e-commerce. We also provide insurance to merchants against electronic crimes and fraud.

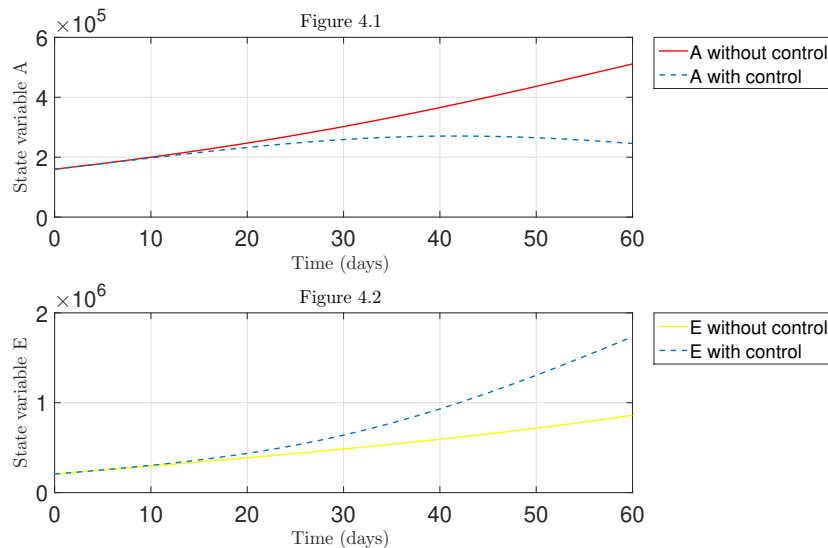


Figure 4: Dynamics with controls v_1 and v_2 .

Figures 4.1 and 4.2 illustrate the impact of the third control strategy on merchant behavior regarding e-commerce adoption. Following the implementation of this strategy, the number of merchants initially refusing to adopt e-commerce in compartment A begins to decline as early

as day 7, with a reduction of approximately 27.20% observed by day 30. Concurrently, the population of merchants actively using e-commerce in compartment E starts to increase from day 15, reaching a 74.63% rise by day 30. Over the full 60-day simulation period, the strategy produces a substantial overall decrease in nonadopting merchants alongside a significant increase in e-commerce adopters. This pattern demonstrates the effectiveness of the third strategy in both reducing resistance and promoting adoption, highlighting its capacity to accelerate behavioral shifts among merchants. The results also suggest that early interventions, coupled with sustained application, can produce rapid and lasting changes in merchant behavior. Overall, the third strategy appears to be the most effective among the three considered, significantly enhancing e-commerce uptake and contributing to greater market integration and competitiveness within the commercial ecosystem.

Finally, the results indicate that the combined application of the two optimal controls, v_1 and v_2 , substantially enhances the effectiveness of the proposed strategy. By simultaneously promoting e-commerce adoption and addressing resistance among hesitant merchants, the integration of these controls provides a more balanced and comprehensive intervention. This dual-control approach not only accelerates the transition of nonadopting merchants into active e-commerce users but also ensures a more efficient allocation of resources to influence behavioral change. The holistic nature of this strategy increases the likelihood of achieving sustained adoption, thereby generating broader economic benefits, improving market competitiveness, and strengthening the resilience of the commercial ecosystem. These findings underscore the potential of multi-faceted control policies in designing effective interventions for complex socio-economic systems.

4.4 Cost-effectiveness analysis (CEA)

In the context of e-commerce, implementing strategies such as promotion via advertising $v_1(t)$ and protection through cybercrime insurance $v_2(t)$ requires financial resources. Cost-effectiveness analysis (CEA) provides a quantitative framework to evaluate which strategies offer the best impact in terms of e-commerce adoption for the lowest cost.

A classic indicator in CEA is the average cost-effectiveness ratio (ACER), defined as

$$\text{ACER} = \frac{T_c}{T_a},$$

where T_c is the total cost of interventions and T_a is the total number of merchants converted to e-commerce thanks to the control strategies.

More precisely, the commercial impact measure T_a can be expressed as:

$$T_a = \int_0^T E_{\text{without control}}(t) dt - \int_0^T E_{\text{with control}}(t) dt,$$

which measures the reduction in the number of merchants refusing e-commerce due to the application of controls $v_1(t)$ and $v_2(t)$.

The total cost T_c associated with these strategies is given by

$$T_c = \int_0^T (c_1 v_1(t) C(t) + c_2 v_2(t) A(t)) dt,$$

where c_1 and c_2 are unit costs associated respectively with advertising/training and cybercrime insurance.

For example, $c_1 = 8$ could represent the average promotion cost per merchant in category C , while $c_2 = 10$ would represent the average insurance cost per merchant in category A .

Table 2 shows the numerical simulation results, sorted by increasing ACER.

Table 2: ACER sorted by increasing efficiency

Scenario	T_a (merchants converted)	T_c (total cost in \$)
Scenario 2	8.2979×10^7	5.8188×10^7
Scenario 1	8.7509×10^7	6.3199×10^7
Scenario 3	8.7580×10^7	6.3271×10^7

The corresponding ACER values are calculated as follows:

- **Scenario 2:**

$$\text{ACER} = \frac{5.8188 \times 10^7}{8.2979 \times 10^7} \approx 0.701237.$$

- **Scenario 1:**

$$\text{ACER} = \frac{6.3199 \times 10^7}{8.7509 \times 10^7} \approx 0.722205.$$

- **Scenario 3:**

$$\text{ACER} = \frac{6.3271 \times 10^7}{8.7580 \times 10^7} \approx 0.722441.$$

Although Scenario 2 presents the lowest ACER, it also offers the most moderate impact in terms of e-commerce adoption. Scenario 3, which combines promotion and insurance, achieves the highest total conversion despite a slightly higher ACER, making it the most robust overall strategy.

Therefore, it is recommended that economic decision-makers and platform managers adopt an integrated approach combining advertising and insurance to maximize the transition to e-commerce while controlling costs.

5 Conclusion

In this study, we developed a nonlinear mathematical model, *CAEP*, to describe the dynamics of merchants and companies that refrain from adopting e-commerce and to analyze its impact on both adopters and nonadopters. The findings highlight the significant role of e-commerce in enhancing economic efficiency, increasing merchants' revenues, and improving product visibility for consumers. The *CAEP* model was extended to an optimal control framework with two control variables aimed at promoting adoption and reducing resistance among merchants. The Pontryagin maximum principle was used to derive optimality conditions, and numerical simulations in MATLAB confirmed the effectiveness of the proposed strategies. An ACER provided insights into the trade-offs between implementation costs and adoption rates, identifying the most efficient strategies for decision-makers. Overall, the integration of mathematical modeling, optimal control, and economic evaluation offers a comprehensive understanding of e-commerce adoption dynamics. The results suggest that e-commerce has strong potential for future economic growth and job creation, particularly for merchants with limited access to physical retail spaces. Future research could extend this work by incorporating network effects and social influence on merchants' adoption decisions, introducing stochastic dynamics to account for market uncertainty and random fluctuations, and integrating merchant adoption models with consumer behavior models to capture feedback between supply and demand. These extensions would provide a more realistic and robust framework for designing effective strategies to promote e-commerce adoption.

Acknowledgements

Authors are grateful to there 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] Abbate, J. *Popularizing the internet*, Commun. Hist., Routledge (2018), 297–302.
- [2] Agrawal, D.R. and Fox, W.F. *Taxes in an e-commerce generation*, Int. Tax Public Finance 24 (2017), 903–926.
- [3] Balatif, O., Khajji, B. and Rachik, M. *Mathematical modeling, analysis, and optimal control of abstinence behavior of registration on the electoral lists*, Discrete Dyn. Nat. Soc. 2020 (2020), Art. ID 9738934.
- [4] Baroudi, M., Gourram, H., Labzai, A. and Belam, M. *Mathematical modeling and monkey-pox's optimal control strategy*, Commun. Math. Biol. Neurosci. 2023 (2023), Art. ID 110.
- [5] Belhdid, S., Balatif, O. and Khajji, B. *Optimal control and cost-effectiveness analysis of scam rumor propagation over social networks*, Results Control Optim. 16 (2024), 100441.
- [6] Corbitt, B.J., Thanasankit, T. and Yi, H. *Trust and e-commerce: a study of consumer perceptions*, Electron. Commer. Res. Appl. 2 (2003), 203–215.
- [7] El Bhih, A., Ghazzali, R., Ben Rhila, S., Rachik, M. and El Alami Laarroussi, A. *A discrete mathematical modeling and optimal control of the rumor propagation in online social network*, Discrete Dyn. Nat. Soc. 2020 (2020), 1–12.
- [8] El Bhih, A., Yaagoub, Z., Rachik, M., Allali, K. and Abdeljawad, T. *Controlling the dissemination of rumors and antirumors in social networks*, Eur. Phys. J. Plus 139 (2024), 1–23.
- [9] El Youssoufi, L., Moutamanni, H., Labzai, A., Balatif, O. and Rachik, M. *Optimal control for a discrete model of hepatitis C*, Commun. Math. Biol. Neurosci. 2020 (2020), Art. ID 82.
- [10] Elhia, M., Balatif, O., Boujallal, L. and Rachik, M. *Optimal control problem for a tuberculosis model*, Int. J. Optim. Control Theory Appl. 11 (2021), 75–91.

- [11] Fleming, W.H. and Rishel, R.W. *Deterministic and stochastic optimal control*, Springer, New York, 2012.
- [12] Gourram, H., Baroudi, M., Labzai, A. and Belam, M. *Mathematical modeling and optimal control strategy for the influenza (H5N1)*, Commun. Math. Biol. Neurosci. 2023 (2023), Art. ID 113.
- [13] Gouvernement, <https://www.gouvernement.fr>
- [14] Gupta, D.C., Bindal, M., Agarwal, N. and Khandelwal, K. *Traditional commerce v/s e-commerce and impact of demonetization on e-commerce*, Int. J. Eng. Manag. Res. 8 (2018), 136–142.
- [15] Hannum, K.M. and Downs, H.A. *Evaluation research in the EDI field*, Booyesen, LAE (2018), 60–84.
- [16] Hassouni, H., Belhdid, S. and Balatif, O. *Mathematical modeling and optimal control of customer's behavior toward e-commerce*, Iran. J. Numer. Anal. Optim. 14(4) (2024) 1168–1202.
- [17] Huang, L., Huang, Y., Huang, R., Xie, G. and Cai, W. *Factors influencing returning migrants' entrepreneurship intentions*, Sustainability 14 (2022), 3682.
- [18] Institut Natl. Statist. Études Écon. (INSEE), <https://www.insee.fr/fr/accueil>
- [19] Isckia, T. *Amazon's evolving ecosystem*, Canad. J. Admin. Sci. 26 (2009), 332–343.
- [20] Jang, S.Y. and Lee, C.S. *A Comparison of Customers' Perceived Value on Electronic Commerce and Traditional Commerce*, Asia-Pac. J. Inf. Syst. 10 (2000), 159–180.
- [21] Jose, S.A., Panigoro, H.S., Jirawattanapanit, A., Omede, B.I. and Yaagoub, Z. *Understanding COVID-19 propagation: A comprehensive mathematical model with Caputo fractional derivatives for Thailand*, Front. Appl. Math. Stat. 10 (2024), 1374721.
- [22] Kada, D., Khajji, B., Balatif, O., Rachik, M. and Labriji, E.H. *Optimal control approach of gaming disorder spread*, Discrete Dyn. Nat. Soc. 2021 (2021), 5584315.
- [23] Kermack, W.O. and McKendrick, A.G. *A contribution to the mathematical theory of epidemics*, Proc. Roy. Soc. Lond. Ser. A 115 (1927), 700–721.

- [24] Kim, K.K. and Prabhakar, B. *Initial trust and the adoption of B2C e-commerce: The case of internet banking*, ACM SIGMIS Database, 35 (2004), 50–64.
- [25] Krishnamurthy, S. *A comparative analysis of eBay and Amazon*, IGI Global, 2004, 29–44.
- [26] Laarabi, H., Rachik, M., El Kahlaoui, O. and Labriji, E. *Optimal vaccination strategies of an SIR model*, Univ. J. Appl. Math. 1 (2013), 185–191.
- [27] Laaroussi, A.E.A., El Bhih, A. and Rachik, M. *Optimal vaccination and treatment policies with constrained inequalities to study limited vaccination resources for a multistrain reaction-diffusion SEIR model of COVID-19*, Partial Differ. Equ. Appl. Math. 10 (2024), 100684.
- [28] Lightner, N.J. *What users want in e-commerce design: Effects of age, education and income*, Ergonomics 46 (2003), 153–168.
- [29] Maity, M. and Dass, M. *Consumer decision-making across modern and traditional channels: E-commerce, m-commerce, in-store*, Decision Support Systems 61 (2014), 34–46.
- [30] Olson, J.S. and Olson, G.M. *i2i trust in e-commerce*, Commun. ACM 43 (2000), 41–44.
- [31] Pontryagin, L.S., Boltyanskii, V.G., Gamkrelidze, R.V. and Mishchenko, E. F. *The mathematical theory of optimal processes*, Wiley, New York, 1962.
- [32] Purwani, S., Inayaturohmat, F. and Tresna, S.T. *COVID-19 epidemic model: study of numerical methods and solving optimal control problem through forward-backward sweep method*, Commun. Math. Biol. Neurosci. 2022 (2022), Art. ID 123.
- [33] Qin, Z., Chang, Y., Li, S. and Li, F. *E-commerce strategy*, Springer, 2014.
- [34] Rahman, S. and Qing, N. *Graduate students' perceptions of supply chain skills for supply chain managers*, Benchmarking 21 (2014), 276–299.
- [35] Sadki, M., Yaagoub, Z. and Allali, K. *Qualitative Analysis of a Fractional-Order for a Within-Host Infection Dynamics with Adaptive Immunity Using Caputo Derivative*, Iran. J. Sci. 49(3) (2025), 847–869.
- [36] Shemi, A.P. and Procter, C. *E-commerce and entrepreneurship in SMEs: case of myBot*, J. Small Bus. Enterprise Dev. 25 (2018), 501–520.
- [37] Taher, G. *E-commerce: advantages and limitations*, Int. J. Acad. Res. Account. Finance Manag. Sci. 11 (2021), 153–165.

- [38] Tan, Z. and Ouyang, W. *Diffusion and impacts of e-commerce in China*, Electron. Mark. 14 (2004), 25–35.
- [39] Wulf, J. *Alibaba Group*, Harv. Bus. Sch. Case 710 (2010), 436.
- [40] Yaagoub, Z. and Allali, K. *Fractional HCV infection model with adaptive immunity and treatment*, Math. Model. Comput. 10 (2023), 995–1013.
- [41] Yaagoub, Z., El Bhih, A. and Allali, K. *Global analysis of a fractional-order infection model for the propagation of computer viruses*, Model. Earth Syst. Environ. 11 (2025), 68.
- [42] Yaagoub, Z., Farah, E.M. and Ahmad, S. *Three-strain epidemic model for influenza virus involving fractional derivative and treatment*, J. Appl. Math. Comput. 71 (2025), 1247–1266.
- [43] Yapar, B.K., Bayrakdar, S. and Yapar, M. *The role of taxation problems on the development of e-commerce*, Procedia Soc. Behav. Sci. 195 (2015), 642–648.



Fractal-fractional dynamics of HIV/AIDS: Unveiling the synergistic impact of condoms, prophylaxis, and antiretroviral interventions for epidemic control

G.O. Acheneje^{,1}, J. Amos¹, E. Abah¹, W. Atokolo¹, B.C. Agbata², Shyamsunder*,^{,3}, A.K. Awasthi⁴ and B. Bolaji¹

Abstract

*Corresponding author

Received 17 September 2025; revised 6 November 2025; accepted 19 November 2025

Godwin Onuche Acheneje, (achenejegodwin@gmail.com)

Jeremiah Amos, (amosjeremiah1997@gmail.com)

Emmanuel Abah, (abahemmanuel1995@gmail.com)

William Atokolo, (williamsatokolo@gmail.com)

Benedict Celestine Agbata, (agbatacelestine92@gmail.com)

Shyamsunder, (skumawatmath@gmail.com)

A.K. Awasthi, (dramitawasthi@gmail.com)

Bolarinwa Bolaji, (bolarinwa.s.bolaji@gmail.com)

¹Department of Mathematical Sciences, Prince Abubakar Audu University, Anyigba, Nigeria, Laboratory of Mathematical Epidemiology, Prince Abubakar Audu University, Anyigba, Nigeria.

²Department of Mathematics and Statistics, Confluence University of Science and Technology, Osara, Kogi State, Nigeria.

³Department of Mathematics, SRM University Delhi-NCR, Sonapat-131029, Haryana, India.

⁴Department of Mathematics, School of Chemical Engineering and Physical Sciences, Lovely Professional University, Phagwara, Punjab, India.

How to cite this article

Acheneje, G.O., Amos, J., Abah, E., Atokolo, W., Agbata, B.C., Shyamsunder, Awasthi, A.K. and Bolaji, B. , Fractal-fractional dynamics of HIV/AIDS: Unveiling the synergistic impact of condoms, prophylaxis, and antiretroviral interventions for epidemic control. *Iran. J. Numer. Anal. Optim.*, 2026; 16(2): 385-446. <https://doi.org/10.22067/ijnao.2025.95479.1732>

HIV/AIDS remains a global health crisis, with over 39 million people living with HIV in 2023, mainly in sub-Saharan Africa. This study presents a novel **fractal-fractional order model** to capture memory effects and heterogeneous transmission patterns in HIV dynamics. Unlike classical integer-order models, it integrates fractional derivatives (for long-term dependencies) and fractal dimensions (for self-similar contact networks). We prove solution positivity, boundedness, and compute the basic reproduction number R_{0H} , with global asymptotic stability at the endemic equilibrium when $R_{0H} > 1$. Sensitivity analysis identifies mortality, AIDS progression, contact rate, and treatment uptake as key R_{0H} drivers. Fitted to South African data (2010–2023), the model estimates $R_{0H} \approx 0.95$ and projects a 28% decline in new infections by 2033 under current trends. Numerical simulations using a second-order Adams–Bashforth scheme show lower fractional orders induce realistic oscillations, outperforming integer-order models in data fit and forecasting. Results underscore that **synergistic scale-up of condoms, PrEP, and ART** is critical to achieving HIV elimination.

AMS subject classifications (2020): 92D30, 34A08, 26A33, 35R11, 37N25, 92C60, 65L05, 35B35

Keywords: Fractal-fractional model; HIV/AIDS transmission dynamics; Basic reproduction number; Sensitivity analysis; Adams–Bashforth scheme.

1 Introduction

Human immunodeficiency virus (HIV) and acquired immunodeficiency syndrome (AIDS) remain among the most significant global public health challenges of the modern era. HIV is a lentivirus that attacks the human immune system, specifically targeting CD4+ T helper cells, which are crucial components of the adaptive immune response [43]. When left untreated, HIV progressively weakens the immune system, leading to AIDS, a condition characterized by opportunistic infections and malignancies that ultimately prove fatal [23]. According to recent epidemiological data, approximately 38 million people worldwide are living with HIV, with sub-Saharan Africa bearing the heaviest burden of the epidemic [45]. The transmission of HIV occurs primarily through unprotected sexual contact, sharing of contaminated needles, mother-to-child transmission during pregnancy, childbirth, or breastfeeding, and transfusion of infected blood products [20]. Understanding these transmission pathways has been fundamental in developing comprehensive prevention strategies that form the cornerstone of global HIV control efforts. Implementing evidence-based prevention interventions has demonstrated remarkable success in reducing HIV incidence rates in various populations and geographical regions [14]. Condom use represents one of the most effective and widely accessible prevention methods for reducing sexual transmission of HIV. When used consistently and correctly, male latex condoms provide

approximately 80–95% protection against HIV transmission during heterosexual intercourse [48]. Female condoms have similarly demonstrated high efficacy rates, offering an additional option for individuals seeking barrier protection methods [25]. The promotion of condom use through comprehensive sexual education programs and widespread distribution campaigns has been associated with significant reductions in HIV incidence across diverse populations [39].

Pre-exposure prophylaxis (PrEP) has emerged as a revolutionary biomedical intervention for HIV prevention among high-risk populations. PrEP involves the daily administration of antiretroviral medications, typically a combination of tenofovir disoproxil fumarate and emtricitabine, to HIV-negative individuals who are at substantial risk of infection [27]. Clinical trials have demonstrated that consistent PrEP use can reduce HIV acquisition risk by up to 99% in sexually active adults [5]. The World Health Organization now recommends PrEP as part of comprehensive HIV prevention programs for individuals at high risk of HIV infection [49]. Post-exposure prophylaxis (PEP) serves as a critical emergency intervention following potential HIV exposure incidents. PEP involves the administration of a 28-day course of antiretroviral therapy (ART) initiated within 72 hours of exposure, with optimal effectiveness achieved when treatment begins within 2 hours [19]. Studies have shown that timely PEP administration can reduce HIV transmission risk by approximately 80%, making it an essential component of comprehensive HIV prevention strategies [18]. PEP protocols are particularly important for healthcare workers experiencing occupational exposures, individuals reporting sexual assault, and those experiencing high-risk sexual or injection drug use exposures [30]. ART has transformed HIV from a uniformly fatal diagnosis to a manageable chronic condition. Modern ART regimens effectively suppress viral replication, restore immune function, and dramatically improve quality of life and survival outcomes for people living with HIV [37]. Furthermore, the concept of treatment as prevention has gained significant recognition, with studies demonstrating that individuals achieving undetectable viral loads through consistent ART adherence cannot transmit HIV sexually, a principle known as “Undetectable = Untransmittable” (U=U) [40]. This breakthrough understanding has profound implications for individual health outcomes and population-level HIV prevention strategies [21].

Mathematical modeling has become increasingly important in understanding infectious disease dynamics and evaluating intervention strategies [12, 34, 33, 17, 1]. Multigroup formulations with stability analysis and optimal control have recently proven effective for regional COVID-19 management [22], motivating our fractal-fractional extension to capture memory-driven HIV persistence in sub-Saharan Africa. Traditional integer-order differential equation models have provided valuable insights into disease transmission patterns and control measures [6]. However, these conventional approaches often fail to capture the complex memory effects and hereditary properties inherent in biological systems, particularly in the context of chronic infections

like HIV [6]. The human immune system exhibits nonlocal temporal correlations and memory-dependent responses that are more accurately described by fractional calculus formulations [38]. Fractional-order differential equations incorporate noninteger derivatives that provide more flexible and realistic representations of biological phenomena characterized by memory and hereditary effects [29]. In HIV dynamics, fractional-order models can better capture the gradual immune system deterioration, the long-term effects of viral persistence, and the complex pharmacokinetic properties of antiretroviral medications [7]. These models have demonstrated superior fitting capabilities compared to classical integer-order models when applied to real-world HIV progression data [24]. Recent advances in fractional calculus have led to the development of fractal-fractional operators, which combine the benefits of fractal geometry and fractional calculus to model complex systems with both memory effects and fractal properties [9]. The fractal-fractional approach is particularly well-suited for modeling biological systems that exhibit self-similar patterns across multiple scales, such as the hierarchical structure of immune system responses and viral population dynamics [10]. These operators provide enhanced flexibility in capturing the intricate dynamics of host-pathogen interactions while maintaining mathematical rigor and computational feasibility [11].

The application of fractal-fractional methods to infectious disease modeling has gained considerable attention in recent literature. Studies have successfully employed these techniques to model various infectious diseases, including tuberculosis, malaria, and COVID-19, demonstrating improved accuracy in predicting disease trajectories and evaluating control strategies [41, 28, 4]. The fractal-fractional framework allows researchers to simultaneously incorporate multiple scales of biological organization, from molecular-level viral dynamics to population-level transmission patterns, providing a more comprehensive understanding of disease progression and intervention effectiveness [31].

In the specific context of HIV/AIDS modeling, integrating fractal-fractional operators offers unique advantages for capturing the multi-scale nature of infection dynamics. The fractal component can account for the spatial heterogeneity in viral distribution within tissues and the scale-invariant properties of immune cell networks, while the fractional component addresses the memory effects associated with immune system priming and antiretroviral drug accumulation [13]. This dual approach enables a more accurate representation of the complex feedback mechanisms governing HIV pathogenesis and treatment responses [35]. Developing robust numerical methods for solving fractal-fractional differential equations has facilitated their practical application in epidemiological modeling. Advanced computational techniques, including spectral methods, finite difference schemes, and machine learning-enhanced algorithms, have been developed to handle the increased mathematical complexity while maintaining computational efficiency [44]. These methodological advances have made fractal-fractional modeling accessible

to researchers investigating complex biological systems, opening new avenues for understanding disease dynamics and optimizing intervention strategies.

Based on the reviewed literature, there exists a significant gap in the mathematical modeling of HIV/AIDS transmission dynamics, particularly in exploring the combined impact of prophylaxis and antiretroviral therapy on reducing disease burden in human populations. This research addresses this gap by employing a novel fractal-fractional order modeling approach, offering a more nuanced analysis of the transmission dynamics compared to traditional integer-order methods.

The remaining part of this manuscript is outlined as follows: In Section 2, the HIV/AIDS integer-order model is presented; Section 3 presents the HIV/AIDS model in fractal-fractional order form. The major analysis of the model is carried out in Section 3. In Section 4, qualitative analysis of the fractal-fractional HIV/AIDS model are presented, and Section 5 contains the concluding remarks of the work.

2 Model formulation

2.1 Key assumptions in model formulation

1. The model presumes a uniformly mixed population where each individual has an equal likelihood of encountering an infected person, as described by the force of infection term λ_H [15].
2. It is assumed that PrEP, PEP, and ART exhibit partial effectiveness, quantified by the efficacy parameters ε_1 , ε_2 , and η , respectively [26].
3. The model posits that ART lowers the viral load, potentially allowing individuals to revert to a less infectious state (governed by parameter χ), while also reducing the likelihood of transmission (modified by parameter η) [20].

2.2 Model description

The human population (N) is divided into seven mutually exclusive groups (see Figure 1 for a visual overview and Table 1 for parameter definitions): Susceptible individuals (S_H), susceptible

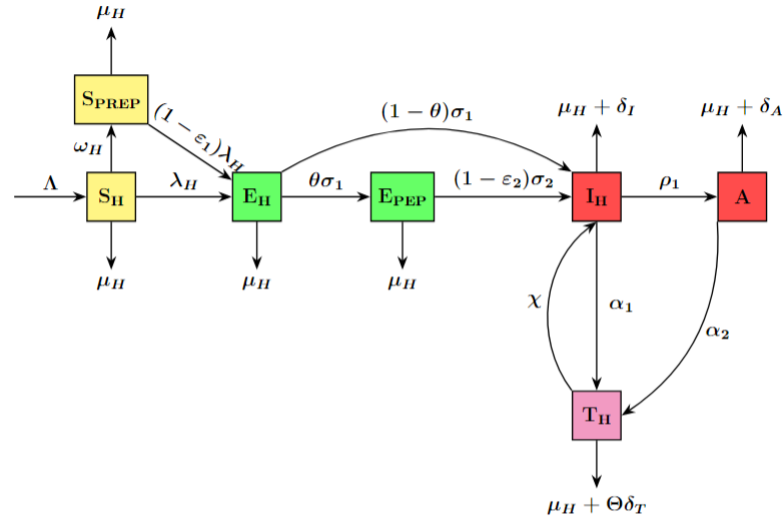


Figure 1: A diagram outlining the epidemiological compartments for HIV/AIDS transmission, depicting groups including susceptible individuals, those on PrEP, exposed individuals, those on PEP, people infected with HIV, those with HIV/AIDS, and various treatment stages [2]

individuals on PrEP (S_{PrEP}), exposed individuals (E_H), exposed individuals on PEP (E_{PEP}), individuals infected with HIV only (I_H), individuals with HIV/AIDS (A), and individuals with HIV/AIDS on ART (T_H).

As illustrated in the top-left of Figure 1, new individuals enter the susceptible population at a rate Λ . From S_H , individuals may transition to S_{PrEP} at rate ω_H (top branch) or become exposed (E_H) via the force of infection λ_H (central infection arrow). The force of infection depends on the transmission probability per contact with infectious individuals (I_H , A , or T_H), modulated by contact rate γ_H , condom adherence ς and efficacy Ψ , and number of sexual partners $\gamma > 1$. A natural mortality rate θ_H applies to all compartments (shown as downward arrows in Figure 1).

The parameter η ($0 < \eta < 1$) reduces transmission from the treated group (T_H) compared to I_H or A (indicated by a lighter infection arrow from T_H in Figure 1). The middle section of Figure 1 shows exposed individuals (E_H) progressing at rate σ_1 either to PEP (E_{PEP} , proportion θ) or directly to I_H (proportion $1 - \theta$). PrEP reduces infection risk by factor ε_1 , with failure leading to exposure at $(1 - \varepsilon_1)\lambda_H$. Similarly, PEP has efficacy ε_2 , with failure transitioning individuals to I_H at rate σ_2 .

The lower part of Figure 1 depicts disease progression and treatment pathways: HIV-infected individuals (I_H) either initiate ART at rate θ_1 or progress to AIDS (A) at rate ρ_1 . Individuals in A enter treatment at rate θ_2 . Treated individuals (T_H) experience reduced disease-induced

mortality ($\Theta\delta_T$) and may revert to I_H at rate χ due to treatment interruption or virological failure. Disease-induced mortality rates are δ_I for I_H , δ_A for A , and δ_T for T_H .

A key feature of the model is the reversible treatment pathway: Individuals in the ART group (T_H) transition back to the untreated HIV-infected group (I_H) at rate $\chi > 0$, representing treatment dropout, poor adherence, or virological rebound (see reversible arrow between I_H and T_H in Figure 1). This mechanism ensures the model reflects real-world challenges in sustaining viral suppression and prevents unrealistic assumptions of permanent treatment efficacy.

Table 1: Table of variables and parameters for the HIV/AIDS transmission model.

Variable (Parameter)	Description
S_H	Susceptible individuals
S_{PrEP}	Susceptible individuals on PrEP
E_H	Exposed individuals
E_{PEP}	Exposed individuals on PEP
I_H	Individuals infected with HIV only
A	Individuals with HIV/AIDS
T_H	Individuals with HIV/AIDS on ART
Λ	Rate of entry into the susceptible population
λ_H	Force of infection for HIV/AIDS
γ_H	Contact rate between susceptible and infectious individuals
ς	Condom use adherence rate ($0 < \varsigma < 1$)
Ψ	Condom efficacy ($0 < \Psi < 1$)
γ	Number of sexual partners per individual ($\gamma > 1$)
θ_H	Natural mortality rate across all groups
η	Parameter for reduced transmission in the treatment group ($0 < \eta < 1$)
ω_H	Rate of transition to the PrEP group
ε_1	Effectiveness of PrEP ($0 < \varepsilon_1 < 1$)
$(1 - \varepsilon_1)\lambda_H$	Failure rate of PrEP leading to exposure
σ_1	Progression rate of exposed individuals
θ	Proportion of exposed individuals using PEP
σ_2	Failure rate of PEP leading to infection
ε_2	Effectiveness of PEP ($0 < \varepsilon_2 < 1$)
θ_1	Rate of transition of HIV-infected individuals to ART
ρ_1	Rate of progression of HIV-infected individuals to the HIV/AIDS group
θ_2	Rate of transition of HIV/AIDS individuals to ART
δ_I	Disease-induced mortality rate for HIV-infected individuals
δ_A	Disease-induced mortality rate for HIV/AIDS individuals
δ_T	Disease-induced mortality rate for individuals in the treatment group
χ	Rate of reversion to the HIV-only group due to reduced viral load
Θ	Parameter for reduced disease-induced mortality in the treatment group

2.3 Model equations

Based on the model description, the system of differential equations governing the HIV/AIDS transmission dynamics in the human population is expressed as

$$\begin{aligned}
 \frac{dS_H}{dt} &= \Lambda - (\lambda_H + \omega_H + \theta_H) S_H, \\
 \frac{dS_{PrEP}}{dt} &= \omega_H S_H - ((1 - \varepsilon_1)\lambda_H + \theta_H) S_{PrEP}, \\
 \frac{dE_H}{dt} &= \lambda_H S_H + (1 - \varepsilon_1)\lambda_H S_{PrEP} - (\sigma_1 + \theta_H) E_H, \\
 \frac{dE_{PEP}}{dt} &= \theta \sigma_1 E_H - ((1 - \varepsilon_2)\sigma_2 + \theta_H) E_{PEP}, \\
 \frac{dI_H}{dt} &= (1 - \theta)\sigma_1 E_H + (1 - \varepsilon_2)\sigma_2 E_{PEP} - (\theta_1 + \rho_1 + \delta_I + \theta_H) I_H + \chi T_H, \\
 \frac{dA}{dt} &= \rho_1 I_H - (\theta_2 + \delta_A + \theta_H) A, \\
 \frac{dT_H}{dt} &= \theta_1 I_H + \theta_2 A - (\chi + \Theta \delta_T + \theta_H) T_H,
 \end{aligned} \tag{1}$$

where the total population is $N(t) = S_H + S_{PrEP} + E_H + E_{PEP} + I_H + A + T_H$, and the force of infection is defined as $\lambda_H = \frac{(1-\varsigma\Psi)\gamma\gamma_H(I_H+A+\eta T_H)}{N}$ for the HIV/AIDS model in (1).

2.4 Background materials

The manuscript relies on fractal-fractional operator definitions, which we outline first. Let τ represents the fractional order, and let ψ denotes the fractal dimension. One assumes a function $y(t)$ is continuous and possesses fractal differentiability on the interval (a, b) with respect to ψ . These definitions are detailed in references [8, 36].

Definition 1. [8] The Caputo fractional differential operator employs the Mittag-Leffler kernel to define the fractal-fractional derivative of a function $y(t)$ at order τ and fractal dimension ψ as follows:

$$FFH_{0,t}^{\tau,\psi} y(t) = \frac{AB(\tau)}{1-\tau} \int_a^t E_\tau \left[\frac{-\tau}{1-\tau} (t-s)^\tau \right] \frac{dy(s)}{ds^\psi} ds, \quad t \geq a \geq 0, \tag{2}$$

where $\tau, \psi \in (0, 1]$, and $AB(\tau) = 1 - \tau + \frac{\tau}{\Gamma(\tau)}$ is the normalization function satisfying $AB(0) = AB(1) = 1$. The fractal derivative is defined as

$$\frac{dx(t)}{dt^\psi} = \lim_{s \rightarrow t} \frac{x(s) - x(t)}{s^\psi - t^\psi}.$$

The Mittag-Leffler function is given by

$$E_\tau(z) = \sum_{k=0}^{\infty} \frac{z^k}{\Gamma(\tau k + 1)}, \quad z, \tau \in \mathbb{C}, \quad \tau > 0,$$

where $\mathbb{C}$ is the set of complex numbers.

Definition 2. [8, 36] The fractal-fractional integral with the Mittag-Leffler kernel in the Caputo sense, of order τ and dimension ψ , is defined as

$$\begin{aligned} FFH_{0,t}^{\tau,\psi} y(t) &= \frac{\psi(1-\tau)t^{\psi-1}y(t)}{AB(\tau)} \\ &+ \frac{\tau\psi}{AB(\tau)\Gamma(\tau)} \int_a^t (t-s)^{\tau-1} s^{\psi-1} y(s) ds, \quad t \geq a \geq 0. \end{aligned} \quad (3)$$

Definition 3. [8] For $x \in C((a, b), \mathbb{R})$, the fractal Laplace transform of order ψ is given by

$${}^F L_p^\psi x(t) = \int_0^\infty x(t) \exp(-pt) t^{\psi-1} dt. \quad (4)$$

Lemma 1. [36] Consider the fractal-fractional initial value problem:

$$FFH_{0,t}^{\tau,\psi} x(t) = G(t, x(t)), \quad t \in [0, T], \quad \tau, \psi \in (0, 1], \quad x(0) = x_0. \quad (5)$$

This can be equivalently expressed as

$$FFH_{0,t}^{\tau,\psi} x(t) = \psi t^{\psi-1} G(t, x(t)), \quad t \in [0, T], \quad \tau, \psi \in (0, 1], \quad x(0) = x_0. \quad (6)$$

The corresponding integral equation for the solution is

$$\begin{aligned} x(t) &= x(0) + \frac{\psi t^{\tau-1} (1-\tau) G(t, x(t))}{AB(\tau)} \\ &+ \frac{\psi \tau}{AB(\tau)\Gamma(\tau)} \int_0^t s^{\psi-1} (t-s)^{\tau-1} G(s, x(s)) ds. \end{aligned} \quad (7)$$

Lemma 2 (Banach fixed point theorem [3]). In a Banach space X , let $D \subset X, D \neq \emptyset$ be a closed subset. An operator $P : D \rightarrow D$ has a unique fixed point in D if it satisfies the contraction condition.

Lemma 3 (Krasnoselskii's fixed point theorem). [42] Let $D \subset Y$ be a nonempty, closed, and convex subset of a Banach space Y . For operators P and Q such that

1. $Pu + Qv \in D$ whenever $u, v \in D$,
2. Q is compact and continuous,
3. P is a contraction mapping,

there exists $k \in D$ such that $k = Pk + Qk$.

3 Analysis of HIV/AIDS model

The fractal-fractional derivative operator is applied to advance the HIV/AIDS model within the Mittag-Leffler kernel framework, with $\tau \in (0, 1]$ and $\psi \in [0, 1]$, as follows:

$$\begin{aligned}
 FFH_{0,t}^{\tau,\psi} S_H(t) &= G_1(t, S_H(t), S_{PREP}(t), E_H(t), E_{PEP}(t), I_H(t), A(t), T_H(t)), \\
 FFH_{0,t}^{\tau,\psi} S_{PREP}(t) &= G_2(t, S_H(t), S_{PREP}(t), E_H(t), E_{PEP}(t), I_H(t), A(t), T_H(t)), \\
 FFH_{0,t}^{\tau,\psi} E_H(t) &= G_3(t, S_H(t), S_{PREP}(t), E_H(t), E_{PEP}(t), I_H(t), A(t), T_H(t)), \\
 FFH_{0,t}^{\tau,\psi} E_{PEP}(t) &= G_4(t, S_H(t), S_{PREP}(t), E_H(t), E_{PEP}(t), I_H(t), A(t), T_H(t)), \\
 FFH_{0,t}^{\tau,\psi} I_H(t) &= G_5(t, S_H(t), S_{PREP}(t), E_H(t), E_{PEP}(t), I_H(t), A(t), T_H(t)), \\
 FFH_{0,t}^{\tau,\psi} A(t) &= G_6(t, S_H(t), S_{PREP}(t), E_H(t), E_{PEP}(t), I_H(t), A(t), T_H(t)), \\
 FFH_{0,t}^{\tau,\psi} T_H(t) &= G_7(t, S_H(t), S_{PREP}(t), E_H(t), E_{PEP}(t), I_H(t), A(t), T_H(t)),
 \end{aligned} \tag{8}$$

where

$$G_i = G_i(t, S_H(t), S_{PREP}(t), E_H(t), E_{PEP}(t), I_H(t), A(t), T_H(t)).$$

The functions G_i for $i = 1, 2, \dots, 7$ are defined as follows:

$$\begin{aligned}
 G_1 &= \Lambda - (\lambda_H + \omega_H + \theta_H) S_H, \\
 G_2 &= \omega_H S_H - ((1 - \varepsilon_1)\lambda_H + \theta_H) S_{PREP}, \\
 G_3 &= \lambda_H S_H + (1 - \varepsilon_1)\lambda_H S_{PREP} - (\sigma_1 + \theta_H) E_H, \\
 G_4 &= \theta\sigma_1 E_H - ((1 - \varepsilon_2)\sigma_2 + \theta_H) E_{PEP}, \\
 G_5 &= (1 - \theta)\sigma_1 E_H + (1 - \varepsilon_2)\sigma_2 E_{PEP} - (\theta_1 + \rho_1 + \delta_I + \theta_H) I_H + \chi T_H, \\
 G_6 &= \rho_1 I_H - (\theta_2 + \delta_A + \theta_H) A, \\
 G_7 &= \theta_1 I_H + \theta_2 A - (\chi + \Theta\delta_T + \theta_H) T_H.
 \end{aligned} \tag{9}$$

with initial conditions

$$S_{H0} \geq 0, S_{PREP0} \geq 0, E_{H0} \geq 0, E_{PEP0} \geq 0, I_{H0} \geq 0, A_0 \geq 0, T_{H0} \geq 0.$$

The fractal-fractional FF-HIV/AIDS model is labeled as (8). If $\tau = 1$, then the model reduces to an integer-order system as in (1). If $\psi = 1$, then it simplifies to a fractional-order HIV/AIDS model.

3.1 Positivity of the model solution

Define the nonnegative region

$$\mathbb{R}_+^7 = \left\{ G \in \mathbb{R}_+^7 : G \geq 0 \text{ and } (S_H, S_{PREP}, E_H, E_{PEP}, I_H, A, T_H)^T \right\},$$

where $(\cdot)^T$ denotes the vector transpose.

Theorem 1. G satisfies boundedness in $\mathbb{R}_+^7$ in addition to being nonnegative for the solution of the FF-HIV/AIDS model described by (8).

Proof. For $t \in (0, \infty)$, the existence and uniqueness of the FF-HIV/AIDS model (8) is obtained.

This result follows from the Lipschitz continuity of the right-hand side functions $\mathbf{G}(t, \mathbf{X})$ in the bounded region Ω_H , which ensures that the associated integral operator is a contraction mapping on a suitable Banach space of continuous functions. The Lipschitz condition holds because all transition rates (infection, progression, treatment, mortality) are either linear or bilinear in the state variables and are uniformly bounded due to population saturation and density-dependent transmission. Under these conditions, the Banach fixed point theorem guarantees a unique solution on any finite time interval, which can be extended globally by continuation.

Consequently, The nonnegative region $\mathbb{R}_+^7$ represents a positive invariant region according to the presented content. Our research used the FF-HIV/AIDS model described in (8) as the basis for our analysis:

$$\begin{aligned} FFH_{0,t}^{\tau,\psi} S_H(t) &= \Lambda \geq 0, \\ FFH_{0,t}^{\tau,\psi} S_{PREP}(t) &= \omega_H S_H \geq 0, \\ FFH_{0,t}^{\tau,\psi} E_H(t) &= \lambda_H S_H + (1 - \varepsilon_1) \lambda_H S_{PREP} \geq 0, \\ FFH_{0,t}^{\tau,\psi} E_{PEP}(t) &= \theta \sigma_1 E_H \geq 0, \\ FFH_{0,t}^{\tau,\psi} I_H(t) &= (1 - \theta) \sigma_1 E_H + (1 - \varepsilon_2) \sigma_2 E_{PEP} + \chi T_H \geq 0, \\ FFH_{0,t}^{\tau,\psi} A(t) &= \rho_1 I_H \geq 0, \end{aligned}$$

$$FFH_{0,t}^{\tau,\psi}T_H(t) = \theta_1 I_H + \theta_2 A \geq 0. \quad (10)$$

If $(S_H, S_{PREP}, E_H, E_{PEP}, I_H, A, T_H) \in \mathbb{R}_+^7$, then from the time interval from (10) solution G remains prevented from escaping through the hyper planes.

Consequently, $\mathbb{R}_+^7$ is a positively invariant and $N = S_H + S_{PREP} + E_H + E_{PEP} + I_H + A + T_H$.

So from the FF-HIV/AIDS model (8), we obtain

$$\left\{ FFH_{0,t}^{\tau,\psi}N(t) = \Lambda - \theta_H N - \delta_I I_H - \delta_A A - \Theta \delta_T T_H \right\}. \quad (11)$$

Using the fractal Laplace transform in (3), we conclude that $N(t) \leq \frac{\Lambda}{\theta_H}$ as $t \rightarrow \infty$.

So we now have the biological range of values of the FF-HIV/AIDS model (8) as follows:

$$\left\{ \begin{array}{l} (S_H, S_{PREP}, E_H, E_{PEP}, I_H, A, T_H) \in \mathbb{R}_+^7 \\ : S_H, S_{PREP}, E_H, E_{PEP}, I_H, A, T_H \geq 0 \end{array} \right\}$$

and $N(t) \leq \frac{\Lambda}{\theta_H}$.

This shows that the above model is biologically feasible.

The biological implication of these fractal-fractional theorems is that they ensure the model behaves realistically over time: No population compartment can become negative, and the total population remains bounded by a natural ceiling determined by birth rate (Λ) and natural death rate (θ_H). This reflects real-world demographics where populations grow but are limited by resources and mortality. The positive invariance property means that once individuals enter a disease state (e.g., exposed, infected, or under treatment), the model correctly tracks their progression without artificial loss or gain. Furthermore, the boundedness result implies long-term stability of the population size, which is crucial for sustainable public health planning and intervention forecasting—especially in high-burden regions like sub-Saharan Africa, where uncontrolled population growth in models could lead to misleading predictions of disease spread or intervention impact.

□

3.2 Basic reproduction number of the HIV/AIDS model

The basic reproduction number, R_{0H} , is computed using the next-generation matrix method [47], a standard approach for compartmental epidemic models as developed by Van Den Driessche and Watmough. This method involves separating the population into infected compartments and constructing the Jacobian matrices of new infections ($\mathcal{F}$) and transitions between infected states ($\mathcal{V}$) at the disease-free equilibrium. The spectral radius of the next-generation matrix $\mathcal{FV}^{-1}$

yields R_{0H} , representing the expected number of secondary HIV infections produced by a single infected individual in a fully susceptible population over their infectious period.

Therefore, $R_{0H} = \rho(FV^{-1})$, where F is defined as the matrix of infection and V is defined as the matrix of the transient terms.

$$F = \begin{bmatrix} 0 & 0 & \frac{(1-\varsigma\Psi)\gamma\gamma_H Z_2}{Z_1} & \frac{(1-\varsigma\Psi)\gamma\gamma_H Z_2}{Z_1} & \frac{\eta(1-\varsigma\Psi)\gamma\gamma_H Z_2}{Z_1} \\ 0 & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & 0 \end{bmatrix},$$

$$V = \begin{bmatrix} P_1 & 0 & 0 & 0 & 0 \\ -v_{21} & P_2 & 0 & 0 & 0 \\ -v_{31} & -v_{32} & P_3 & 0 & -\chi \\ 0 & 0 & -\rho_1 & P_4 & 0 \\ 0 & 0 & -\theta_1 & -\theta_2 & P_5 \end{bmatrix},$$

$$FV^{-1} = \begin{bmatrix} \widehat{F}_1 & \widehat{F}_2 & \widehat{F}_3 & \widehat{F}_4 & \widehat{F}_5 \\ 0 & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & 0 \end{bmatrix},$$

where

$$v_{21} = \theta\sigma_1, v_{31} = (1 - \theta)\sigma_1, v_{32} = (1 - \varepsilon_2)\sigma_2,$$

$$P_1 = (\sigma_1 + \theta_H), P_2 = ((1 - \varepsilon_2)\sigma_2 + \theta_H), P_3 = (\theta_1 + \rho_1 + \delta_I + \theta_H),$$

$$P_4 = (\theta_2 + \delta_A + \theta_H), P_5 = (\chi + \Theta\delta_T + \theta_H), Z_1 = (\omega_H + \theta_H), Z_2 = ((1 - \varepsilon_1)\omega_H + \theta_H).$$

$$\begin{aligned}\widehat{F}_1 &= \frac{\gamma_H (\psi \varsigma - 1) (P_2 v_{31} + v_{21} v_{32}) Z_2 ((\eta \theta_1 + P_5) P_4 + \rho_I (\eta \theta_2 + P_5)) \gamma}{P_1 ((-P_3 P_5 + \chi \theta_1) P_4 + \chi \theta_2 \rho_I) P_2 Z_1}, \\ \widehat{F}_2 &= \frac{\gamma \gamma_H ((\eta \theta_1 + P_5) P_4 + \rho_I (\eta \theta_2 + P_5)) v_{32} (\psi \varsigma - 1) Z_2}{P_2 ((-P_3 P_5 + \chi \theta_1) P_4 + \chi \theta_2 \rho_I) Z_1}, \\ \widehat{F}_3 &= \frac{\gamma (\psi \varsigma - 1) ((\eta \theta_1 + P_5) P_4 + \rho_I (\eta \theta_2 + P_5)) Z_2 \gamma_H}{Z_1 ((-P_3 P_5 + \chi \theta_1) P_4 + \chi \theta_2 \rho_I)}, \\ \widehat{F}_4 &= - \frac{\gamma (\psi \varsigma - 1) ((\theta_1 - \theta_2) \chi - P_3 (\eta \theta_2 + P_5)) Z_2 \gamma_H}{Z_1 ((P_4 \theta_1 + \theta_2 \rho_I) \chi - P_3 P_4 P_5)}, \\ \widehat{F}_5 &= \frac{((P_3 \eta + \chi) P_4 + \rho_I \chi) \gamma (\psi \varsigma - 1) Z_2 \gamma_H}{Z_1 ((-P_3 P_5 + \chi \theta_1) P_4 + \chi \theta_2 \rho_I)}.\end{aligned}$$

The dominant eigenvalue of FV^{-1} which is the basic reproduction number of the HIV/AIDS model is given as

$$R_{0H} = \frac{(1 - \varsigma \psi) \gamma_H \gamma ((\eta \theta_1 + P_5) P_4 + \rho_I (\eta \theta_2 + P_5)) (P_2 v_{31} + v_{21} v_{32}) Z_2}{Z_1 P_1 P_2 (P_4 (P_3 P_5 - \chi \theta_1) - \theta_2 \rho_I \chi)}. \quad (12)$$

4 Qualitative analysis of the fractal-fractional HIV/AIDS model

To demonstrate the qualitative results of the HIV/AIDS model (8), we need the specified Banach space. Let $\tau = [0, T]$ represent the time interval of all continuous real-valued functions denoted as $Y = C(\tau \times \mathbb{R}^7, \mathbb{R})$ equipped with the norm

$$\begin{aligned}\|x\| &= \|(S_H, S_{PREP}, E_H, E_{PEP}, I_H, A, T_H)\| \\ &= \|S_H\| + \|S_{PREP}\| + \|E_H\| + \|E_{PEP}\| + \|I_H\| + \|A\| + \|T_H\|, \\ \|S_H\| &= \sup_{t \in \tau} |S_H(t)| = A_{h1}, \\ \|S_{PREP}\| &= \sup_{t \in \tau} |S_{PREP}(t)| = A_{h2}, \\ \|E_H\| &= \sup_{t \in \tau} |E_H(t)| = A_{h3}, \\ \|E_{PEP}\| &= \sup_{t \in \tau} |E_{PEP}(t)| = A_{h4}, \\ \|I_H\| &= \sup_{t \in \tau} |I_H(t)| = A_{h5}, \\ \|A\| &= \sup_{t \in \tau} |A(t)| = A_{h6}, \\ \|T_H\| &= \sup_{t \in \tau} |T_H(t)| = A_{h7}.\end{aligned} \quad (13)$$

Here $(S_H, S_{PREP}, E_H, E_{PEP}, I_H, A, T_H) \in X$. Let $G \in X$ and $x \in C(\tau, \mathbb{R})$.

Then, the HIV/AIDS model (1) can be expressed as

$$\begin{cases} FFH_{0,t}^{\tau,\psi} x(t) = G(t, x(t)), & t \in [0, T], \tau, \psi \in (0, 1], \\ x(0) = x_0, \end{cases}$$

where

$$x(t) = \begin{pmatrix} S_H(t) \\ S_{PREP}(t) \\ E_H(t) \\ E_{PEP}(t) \\ I_H(t) \\ A(t) \\ T_H(t) \end{pmatrix}, \quad x(0) = \begin{pmatrix} S_H(0) \\ S_{PREP}(0) \\ E_H(0) \\ E_{PEP}(0) \\ I_H(0) \\ A(0) \\ T_H(0) \end{pmatrix}, \quad G(t, x(t)) = \begin{pmatrix} G_1(t, S_H(t)) \\ G_2(t, S_{PREP}(t)) \\ G_3(t, E_H(t)) \\ G_4(t, E_{PEP}(t)) \\ G_5(t, I_H(t)) \\ G_6(t, A(t)) \\ G_7(t, T_H(t)) \end{pmatrix}. \quad (14)$$

The solution of problem (13) corresponds to this integral equation when G_i , $i = 1, 2, \dots, 7$, are defined as in (8):

$$x(t) = x(0) + \frac{\psi t^{\tau-1}(1-\tau)G(t, x(t))}{AB(\tau)} + \frac{\psi\tau}{AB(\tau)\Gamma(\tau)} \int_0^t s^{\psi-1}(t-s)^{\tau-1}G(s, x(s)) ds,$$

We simplify the calculations by introducing the following symbols in this work:

$$\begin{aligned} G_1(t, S_H(t)) &= G_1(t, S_H(t), S_{PREP}(t), E_H(t), E_{PEP}(t), I_H(t), A(t), T_H(t)), \\ G_2(t, S_{PREP}(t)) &= G_2(t, S_H(t), S_{PREP}(t), E_H(t), E_{PEP}(t), I_H(t), A(t), T_H(t)), \\ G_3(t, E_H(t)) &= G_3(t, S_H(t), S_{PREP}(t), E_H(t), E_{PEP}(t), I_H(t), A(t), T_H(t)), \\ G_4(t, E_{PEP}(t)) &= G_4(t, S_H(t), S_{PREP}(t), E_H(t), E_{PEP}(t), I_H(t), A(t), T_H(t)), \\ G_5(t, I_H(t)) &= G_5(t, S_H(t), S_{PREP}(t), E_H(t), E_{PEP}(t), I_H(t), A(t), T_H(t)), \\ G_6(t, A(t)) &= G_6(t, S_H(t), S_{PREP}(t), E_H(t), E_{PEP}(t), I_H(t), A(t), T_H(t)), \\ G_7(t, T_H(t)) &= G_7(t, S_H(t), S_{PREP}(t), E_H(t), E_{PEP}(t), I_H(t), A(t), T_H(t)). \end{aligned} \quad (15)$$

The FF-HIV/AIDS model (8) represents the following set of Volterra integral equations according to Lemma 2:

$$\begin{aligned} S_H(t) &= S_H(0) + \frac{\psi t^{\tau-1}(1-\tau)G_1(t, S_H(t))}{AB(\tau)} \\ &\quad + \frac{\psi\tau}{AB(\tau)\Gamma(\tau)} \int_0^t s^{\psi-1}(t-s)^{\tau-1}G_1(s, S_H(s)) ds, \end{aligned}$$

$$\begin{aligned}
S_{PREP}(t) &= S_{PREP}(0) + \frac{\psi t^{\tau-1}(1-\tau)G_2(t, S_{PREP}(t))}{AB(\tau)} \\
&\quad + \frac{\psi\tau}{AB(\tau)\Gamma(\tau)} \int_0^t s^{\psi-1}(t-s)^{\tau-1}G_2(s, S_{PREP}(s)) ds, \\
E_H(t) &= E_H(0) + \frac{\psi t^{\tau-1}(1-\tau)G_3(t, E_H(t))}{AB(\tau)} \\
&\quad + \frac{\psi\tau}{AB(\tau)\Gamma(\tau)} \int_0^t s^{\psi-1}(t-s)^{\tau-1}G_3(s, E_H(s)) ds, \\
E_{PEP}(t) &= E_{PEP}(0) + \frac{\psi t^{\tau-1}(1-\tau)G_4(t, E_{PEP}(t))}{AB(\tau)} \\
&\quad + \frac{\psi\tau}{AB(\tau)\Gamma(\tau)} \int_0^t s^{\psi-1}(t-s)^{\tau-1}G_4(s, E_{PEP}(s)) ds, \\
I_H(t) &= I_H(0) + \frac{\psi t^{\tau-1}(1-\tau)G_5(t, I_H(t))}{AB(\tau)} \\
&\quad + \frac{\psi\tau}{AB(\tau)\Gamma(\tau)} \int_0^t s^{\psi-1}(t-s)^{\tau-1}G_5(s, I_H(s)) ds, \\
A(t) &= A(0) + \frac{\psi t^{\tau-1}(1-\tau)G_6(t, A(t))}{AB(\tau)} \\
&\quad + \frac{\psi\tau}{AB(\tau)\Gamma(\tau)} \int_0^t s^{\psi-1}(t-s)^{\tau-1}G_6(s, A(s)) ds, \\
T_H(t) &= T_H(0) + \frac{\psi t^{\tau-1}(1-\tau)G_7(t, T_H(t))}{AB(\tau)} \\
&\quad + \frac{\psi\tau}{AB(\tau)\Gamma(\tau)} \int_0^t s^{\psi-1}(t-s)^{\tau-1}G_7(s, T_H(s)) ds.
\end{aligned} \tag{16}$$

In view of (16), we define an operator, $\iota : X \rightarrow X$, where

$$\begin{aligned}
\iota &= (\iota_1, \iota_2, \iota_3, \iota_4, \iota_5, \iota_6, \iota_7), \\
(\iota_1, S_H)(t) &= S_H(0) + \frac{\psi t^{\tau-1}(1-\tau)G_1(t, S_H(t))}{AB(\tau)} \\
&\quad + \frac{\psi\tau}{AB(\tau)\Gamma(\tau)} \int_0^t s^{\psi-1}(t-s)^{\tau-1}G_1(s, S_H(s)) ds, \\
(\iota_2, S_{PREP})(t) &= S_{PREP}(0) + \frac{\psi t^{\tau-1}(1-\tau)G_2(t, S_{PREP}(t))}{AB(\tau)} \\
&\quad + \frac{\psi\tau}{AB(\tau)\Gamma(\tau)} \int_0^t s^{\psi-1}(t-s)^{\tau-1}G_2(s, S_{PREP}(s)) ds, \\
(\iota_3, E_H)(t) &= E_H(0) + \frac{\psi t^{\tau-1}(1-\tau)G_3(t, E_H(t))}{AB(\tau)} \\
&\quad + \frac{\psi\tau}{AB(\tau)\Gamma(\tau)} \int_0^t s^{\psi-1}(t-s)^{\tau-1}G_3(s, E_H(s)) ds,
\end{aligned}$$

$$\begin{aligned}
 (\iota_4, E_{PEP})(t) &= E_{PEP}(0) + \frac{\psi t^{\tau-1}(1-\tau)G_4(t, E_{PEP}(t))}{AB(\tau)} \\
 &\quad + \frac{\psi\tau}{AB(\tau)\Gamma(\tau)} \int_0^t s^{\psi-1}(t-s)^{\tau-1}G_4(s, E_{PEP}(s)) ds, \\
 (\iota_5, I_H)(t) &= I_H(0) + \frac{\psi t^{\tau-1}(1-\tau)G_5(t, I_H(t))}{AB(\tau)} \\
 &\quad + \frac{\psi\tau}{AB(\tau)\Gamma(\tau)} \int_0^t s^{\psi-1}(t-s)^{\tau-1}G_5(s, I_H(s)) ds, \\
 (\iota_6, A)(t) &= A(0) + \frac{\psi t^{\tau-1}(1-\tau)G_6(t, A(t))}{AB(\tau)} \\
 &\quad + \frac{\psi\tau}{AB(\tau)\Gamma(\tau)} \int_0^t s^{\psi-1}(t-s)^{\tau-1}G_6(s, A(s)) ds, \\
 (\iota_7, T_H)(t) &= T_H(0) + \frac{\psi t^{\tau-1}(1-\tau)G_7(t, T_H(t))}{AB(\tau)} \\
 &\quad + \frac{\psi\tau}{AB(\tau)\Gamma(\tau)} \int_0^t s^{\psi-1}(t-s)^{\tau-1}G_7(s, T_H(s)) ds.
 \end{aligned} \tag{17}$$

We transform the examined model into a fixed point problem for fixed point theory investigation: $x = \iota x$. This will apply to create a fixed point theory. Our next section demonstrates the FF-HIV/AIDS model (8) achieves a single solution.

Theorem 2. Let $G \in Y$ satisfy the assumption B_1 . There exists a constant $L_{\max} = \max\{L_1, L_2, L_3, L_4, L_5, L_6, L_7\}$, such that

$$\begin{aligned}
 &\left| G(t, S_H(t), S_{PREP}(t), E_H(t), E_{PEP}(t), I_H(t), A(t), T_H(t)) \right. \\
 &\quad \left. - G(t, S_H^*(t), S_{PREP}^*(t), E_H^*(t), E_{PEP}^*(t), I_H^*(t), A^*(t), T_H^*(t)) \right| \\
 &\leq L_{\max} \left(|S_H - S_H^*| + |S_{PREP} - S_{PREP}^*| + |E_H - E_H^*| \right. \\
 &\quad \left. + |E_{PEP} - E_{PEP}^*| + |I_H - I_H^*| + |A - A^*| + |T_H - T_H^*| \right).
 \end{aligned}$$

For any

$S_H(t), S_{PREP}(t), E_H(t), E_{PEP}(t), I_H(t), A(t), T_H(t) \in X$ and $t \in \tau$, and $\left(\psi(1-\tau)T_{\min}^{\tau-1} + \frac{\tau T^{\tau+\psi-1}\Gamma(\psi+1)}{\Gamma(\tau+\psi)} \right) \frac{L_{\max}}{AB(\tau)} < 1$, then the FF-HIV/AIDS model (8) has a unique solution.

Proof. Let D_{Z_1} be a bounded, closed, and convex subset defined as

$$D_{Z_1} = \left\{ (S_H(t), S_{PREP}(t), E_H(t), E_{PEP}(t), I_H(t), A(t), T_H(t)) \in X : \right. \\
 \left. \|(S_H, S_{PREP}, E_H, E_{PEP}, I_H, A, T_H)\| \leq Z_1 \right\}, \text{ with radius} \\
 \frac{\sigma_{\max} + \left(\psi(1-\tau)T_{\min}^{\psi-1} + \frac{\tau T^{\tau+\psi-1}\Gamma(\psi+1)}{\Gamma(\tau+\psi)} \right) \frac{G_{\max}^*}{AB(\tau)}}{1 - \left(\psi(1-\tau)T_{\min}^{\psi-1} + \frac{\tau T^{\tau+\psi-1}\Gamma(\psi+1)}{\Gamma(\tau+\psi)} \right) \frac{L_{\max}^*}{AB(\tau)}}, \text{ where}$$

$\sigma_{\max} = \max \{S_{H0}(t), S_{PREP0}(t), E_{H0}(t), E_{PEP0}(t), I_{H0}(t), A_0(t), T_{H0}(t)\}$, and let $\sup_{t \in \tau} |G_i(s, 0)| = G_i^* < +\infty$, $i = 1, 2, 3, 4, 5, 6, 7$.

First, we show that $\iota D_{Z1} \subset D_{Z1}$. For any $(S_H(t), S_{PREP}(t), E_H(t), E_{PEP}(t), I_H(t), A(t), T_H(t)) \in D_{Z1}$, $t \in \tau$, we obtain

$$\begin{aligned} |(\iota_1 S_H)(t)| &\leq |S_H(0)| + \frac{\psi t^{\tau-1}(1-\tau)}{AB(\tau)} |G_1(t, S_H(t))| \\ &\quad + \frac{\tau\psi}{AB(\tau)\Gamma(\tau)} \int_0^t s^{\psi-1}(t-s)^{\tau-1} |G_1(s, S_H(s))| ds \\ &\leq |S_H(0)| + \frac{\psi t^{\tau-1}(1-\tau)}{AB(\tau)} [L_1 (|S_H(t)| + |S_{PREP}(t)| + |E_H(t)| \\ &\quad + |E_{PEP}(t)| + |I_H(t)| + |A(t)| + |T_H(t)|) + G_1^*] \end{aligned} \quad (18)$$

$$\begin{aligned} &\quad + \frac{\tau\psi}{AB(\tau)\Gamma(\tau)} \int_0^t s^{\psi-1}(t-s)^{\tau-1} [L_1 (|S_H(s)| + |S_{PREP}(s)| \\ &\quad + |E_H(s)| + |E_{PEP}(s)| + |I_H(s)| + |A(s)| + |T_H(s)|) + G_1^*] ds \\ &\leq S_{H0} + \left(\frac{\psi(1-\tau)T^{\tau-1}}{AB(\tau)} + \frac{\tau T^{\tau+\psi-1}\Gamma(\psi+1)}{AB(\tau)\Gamma(\tau+\psi)} \right) \\ &\quad \times [L_1 (\|S_H\| + \|S_{PREP}\| + \|E_H\| + \|E_{PEP}\| + \|I_H\| + \|A\| \\ &\quad + \|T_H\|) + G_1^*]. \end{aligned} \quad (19)$$

$$+ \|T_H\|) + G_1^*]. \quad (20)$$

Thus, $\|\iota_1 S_H\| \leq S_{H0} + \left(\frac{\psi(1-\tau)T_{\min}^{\psi-1}}{AB(\tau)} + \frac{\tau T^{\tau+\psi-1}\Gamma(\psi+1)}{AB(\tau)\Gamma(\tau+\psi)} \right) [L_1 Z_1 + G_1^*]$.

Similarly, for $i = 2, 3, \dots, 7$, we have

$$\begin{aligned} \|\iota_2 S_{PREP}\| &\leq S_{PREP0} + \left(\frac{\psi(1-\tau)T_{\min}^{\psi-1}}{AB(\tau)} + \frac{\tau T^{\tau+\psi-1}\Gamma(\psi+1)}{AB(\tau)\Gamma(\tau+\psi)} \right) [L_2 Z_2 + G_2^*], \\ \|\iota_3 E_H\| &\leq E_{H0} + \left(\frac{\psi(1-\tau)T_{\min}^{\psi-1}}{AB(\tau)} + \frac{\tau T^{\tau+\psi-1}\Gamma(\psi+1)}{AB(\tau)\Gamma(\tau+\psi)} \right) [L_3 Z_3 + G_3^*], \\ \|\iota_4 E_{PEP}\| &\leq E_{PEP0} + \left(\frac{\psi(1-\tau)T_{\min}^{\psi-1}}{AB(\tau)} + \frac{\tau T^{\tau+\psi-1}\Gamma(\psi+1)}{AB(\tau)\Gamma(\tau+\psi)} \right) [L_4 Z_4 + G_4^*], \\ \|\iota_5 I_H\| &\leq I_{H0} + \left(\frac{\psi(1-\tau)T_{\min}^{\psi-1}}{AB(\tau)} + \frac{\tau T^{\tau+\psi-1}\Gamma(\psi+1)}{AB(\tau)\Gamma(\tau+\psi)} \right) [L_5 Z_5 + G_5^*], \\ \|\iota_6 A\| &\leq A_0 + \left(\frac{\psi(1-\tau)T_{\min}^{\psi-1}}{AB(\tau)} + \frac{\tau T^{\tau+\psi-1}\Gamma(\psi+1)}{AB(\tau)\Gamma(\tau+\psi)} \right) [L_6 Z_6 + G_6^*], \\ \|\iota_7 T_H\| &\leq T_{H0} + \left(\frac{\psi(1-\tau)T_{\min}^{\psi-1}}{AB(\tau)} + \frac{\tau T^{\tau+\psi-1}\Gamma(\psi+1)}{AB(\tau)\Gamma(\tau+\psi)} \right) [L_7 Z_7 + G_7^*]. \end{aligned} \quad (21)$$

This implies $\iota D_{Z1} \subset D_{Z1}$.

Secondly, we show that ι is a contraction. Assume

$$(S_H(t), S_{PREP}(t), E_H(t), E_{PEP}(t), I_H(t), A(t), T_H(t)) \in D_{Z1},$$

and

$$(S_H^*(t), S_{PREP}^*(t), E_H^*(t), E_{PEP}^*(t), I_H^*(t), A^*(t), T_H^*(t)) \in D_{Z1}.$$

We have

$$\begin{aligned} & |(\iota_1 S_H)(t) - (\iota_1 S_H^*)(t)| \\ & \leq \frac{\psi t^{\psi-1}(1-\tau)}{AB(\tau)} |G_1(t, S_H(t)) - G_1(t, S_H^*(t))| \\ & \quad + \frac{\psi \tau}{AB(\tau)\Gamma(\tau)} \int_0^t s^{\psi-1}(t-s)^{\tau-1} |G_1(s, S_H(s)) - G_1(s, S_H^*(s))| ds, \\ & \leq \frac{\psi t^{\psi-1}(1-\tau)}{AB(\tau)} L_1 \Delta(t) + \frac{\psi \tau}{AB(\tau)\Gamma(\tau)} \int_0^t s^{\psi-1}(t-s)^{\tau-1} L_1 \Delta(s) ds, \\ & \leq \left(\frac{\psi(1-\tau)T_{\min}^{\psi-1}}{AB(\tau)} + \frac{\tau T^{\tau+\psi-1}\Gamma(\psi+1)}{\Gamma(\psi+\tau)} \right) \frac{L_1}{AB(\tau)} \|\Delta\|. \end{aligned} \quad (22)$$

Similarly, for $i = 2, 3, \dots, 7$, we obtain

$$\begin{aligned} & |(\iota_2 S_{PREP})(t) - (\iota_2 S_{PREP}^*)(t)| \\ & \leq \left(\frac{\psi(1-\tau)T_{\min}^{\psi-1}}{AB(\tau)} + \frac{\tau T^{\tau+\psi-1}\Gamma(\psi+1)}{\Gamma(\psi+\tau)} \right) \frac{L_2}{AB(\tau)} \sum_j \|X_j - X_j^*\|, \\ & |(\iota_3 E_H)(t) - (\iota_3 E_H^*)(t)| \\ & \leq \left(\frac{\psi(1-\tau)T_{\min}^{\psi-1}}{AB(\tau)} + \frac{\tau T^{\tau+\psi-1}\Gamma(\psi+1)}{\Gamma(\psi+\tau)} \right) \frac{L_3}{AB(\tau)} \sum_j \|X_j - X_j^*\|, \\ & |(\iota_4 E_{PEP})(t) - (\iota_4 E_{PEP}^*)(t)| \\ & \leq \left(\frac{\psi(1-\tau)T_{\min}^{\psi-1}}{AB(\tau)} + \frac{\tau T^{\tau+\psi-1}\Gamma(\psi+1)}{\Gamma(\psi+\tau)} \right) \frac{L_4}{AB(\tau)} \sum_j \|X_j - X_j^*\|, \\ & |(\iota_5 I_H)(t) - (\iota_5 I_H^*)(t)| \\ & \leq \left(\frac{\psi(1-\tau)T_{\min}^{\psi-1}}{AB(\tau)} + \frac{\tau T^{\tau+\psi-1}\Gamma(\psi+1)}{\Gamma(\psi+\tau)} \right) \frac{L_5}{AB(\tau)} \sum_j \|X_j - X_j^*\|, \\ & |(\iota_6 A)(t) - (\iota_6 A^*)(t)| \\ & \leq \left(\frac{\psi(1-\tau)T_{\min}^{\psi-1}}{AB(\tau)} + \frac{\tau T^{\tau+\psi-1}\Gamma(\psi+1)}{\Gamma(\psi+\tau)} \right) \frac{L_6}{AB(\tau)} \sum_j \|X_j - X_j^*\|, \end{aligned}$$

$$\begin{aligned}
& |(\iota_7 T_H)(t) - (\iota_7 T_H^*)(t)| \\
& \leq \left(\frac{\psi(1-\tau)T_{\min}^{\psi-1}}{AB(\tau)} + \frac{\tau T^{\tau+\psi-1}\Gamma(\psi+1)}{\Gamma(\psi+\tau)} \right) \frac{L_7}{AB(\tau)} \sum_j \|X_j - X_j^*\|, \tag{23}
\end{aligned}$$

where

$$\begin{aligned}
\sum_j \|X_j - X_j^*\| &= \|S_H - S_H^*\| + \|S_{PREP} - S_{PREP}^*\| \\
&\quad + \|E_H - E_H^*\| + \|E_{PEP} - E_{PEP}^*\| \\
&\quad + \|I_H - I_H^*\| + \|A - A^*\| + \|T_H - T_H^*\|.
\end{aligned}$$

Since $\iota = (\iota_1, \iota_2, \iota_3, \iota_4, \iota_5, \iota_6, \iota_7)$ and $L_{\max} > 0$, we have

$$\begin{aligned}
& \|\iota(t, S_H, S_{PREP}, E_H, E_{PEP}, I_H, A, T_H) \\
& - \iota(t, S_H^*, S_{PREP}^*, E_H^*, E_{PEP}^*, I_H^*, A^*, T_H^*)\| \\
& \leq \left(\frac{\psi(1-\tau)T_{\min}^{\psi-1}}{AB(\tau)} + \frac{\tau T^{\tau+\psi-1}\Gamma(\psi+1)}{\Gamma(\psi+\tau)} \right) \\
& \quad \cdot \frac{L_{\max}}{AB(\tau)} \left[\|S_H - S_H^*\| + \|S_{PREP} - S_{PREP}^*\| + \|E_H - E_H^*\| \right. \\
& \quad \left. + \|E_{PEP} - E_{PEP}^*\| + \|I_H - I_H^*\| + \|A - A^*\| + \|T_H - T_H^*\| \right].
\end{aligned}$$

By the condition, ι is a contraction. By Lemma 2, the FF-HIV/AIDS model (8) has a unique solution.

The biological implication of this fractal-fractional fixed point theorem is that it guarantees a *unique* long-term trajectory for the HIV/AIDS epidemic under given initial conditions and intervention strategies. In real-world terms, this means that the future course of the epidemic—how many people will be susceptible, exposed, infected, or under treatment—is determined *unambiguously* by current population states and control measures (e.g., PrEP uptake, ART coverage, condom use). This uniqueness is critical for public health decision-making: Policymakers can trust that model predictions (such as the expected decline in new infections by 28% by 2033) are not ambiguous or dependent on arbitrary mathematical artifacts. It ensures that intervention planning—whether scaling up prophylaxis, improving treatment access, or enhancing behavioral prevention—will have a consistent, predictable impact on disease dynamics, thereby supporting reliable resource allocation and epidemic control in high-prevalence settings like South Africa. $\square$

Theorem 3. Let $G \in X$ satisfy the assumptions (B_1) in Theorem 2 and (B_1) . There exist constants $g_{i,j} > 0$, $i = 0, 1, \dots, 6$, $j = 0, 1, \dots, 6$, such that

$$\begin{aligned}
& |G(S_H(t), S_{PREP}(t), E_H(t), E_{PEP}(t), I_H(t), A(t), T_H(t))| \\
& \leq g_{0j} + g_{1j} |S_H(t)| + g_{2j} |S_{PREP}(t)| + g_{3j} |E_H(t)| \\
& \quad + g_{4j} |E_{PEP}(t)| + g_{5j} |I_H(t)| + g_{6j} |A(t)| + g_{7j} |T_H(t)|.
\end{aligned}$$

If $\psi(1-\tau)T_{\min}^{\psi-1}L_{\max}AB(\tau) < 1$, then the HIV/AIDS model (1) has at least one solution.

Proof. We define a set

$$D_{Z2} = \left\{ (S_H(t), S_{PREP}(t), E_H(t), E_{PEP}(t), I_H(t), A(t), T_H(t)) \in X : \right. \\ \left. \| (S_H, S_{PREP}, E_H, E_{PEP}, I_H, A, T_H) \| \leq Z_2 \right\}.$$

Applying the model equations, we define two operators $(Q, P) : D_{Z2} \rightarrow X$, where $P = (P_1, P_2, P_3, P_4, P_5, P_6, P_7)$.

The operator Q is defined by

$$\begin{aligned}
(Q_1, S_H)(t) &= S_H(0) + \frac{\psi t^{\psi-1}(1-\tau)G_1(t, S_H(t))}{AB(\tau)}, \\
(Q_2, S_{PREP})(t) &= S_{PREP}(0) + \frac{\psi t^{\psi-1}(1-\tau)G_2(t, S_{PREP}(t))}{AB(\tau)}, \\
(Q_3, E_H)(t) &= E_H(0) + \frac{\psi t^{\psi-1}(1-\tau)G_3(t, E_H(t))}{AB(\tau)}, \\
(Q_4, E_{PEP})(t) &= E_{PEP}(0) + \frac{\psi t^{\psi-1}(1-\tau)G_4(t, E_{PEP}(t))}{AB(\tau)}, \\
(Q_5, I_H)(t) &= I_H(0) + \frac{\psi t^{\psi-1}(1-\tau)G_5(t, I_H(t))}{AB(\tau)}, \\
(Q_6, A)(t) &= A(0) + \frac{\psi t^{\psi-1}(1-\tau)G_6(t, A(t))}{AB(\tau)}, \\
(Q_7, T_H)(t) &= T_H(0) + \frac{\psi t^{\psi-1}(1-\tau)G_7(t, T_H(t))}{AB(\tau)}. \tag{24}
\end{aligned}$$

The operator P is defined by

$$\begin{aligned}
(P_1, S_H)(t) &= \frac{\psi\tau}{AB(\tau)\Gamma(\tau)} \int_0^t s^{\psi-1}(t-s)^{\tau-1}G_1(s, S_H(s)) ds, \\
(P_2, S_{PREP})(t) &= \frac{\psi\tau}{AB(\tau)\Gamma(\tau)} \int_0^t s^{\psi-1}(t-s)^{\tau-1}G_2(s, S_{PREP}(s)) ds, \\
(P_3, E_H)(t) &= \frac{\psi\tau}{AB(\tau)\Gamma(\tau)} \int_0^t s^{\psi-1}(t-s)^{\tau-1}G_3(s, E_H(s)) ds, \\
(P_4, E_{PEP})(t) &= \frac{\psi\tau}{AB(\tau)\Gamma(\tau)} \int_0^t s^{\psi-1}(t-s)^{\tau-1}G_4(s, E_{PEP}(s)) ds, \\
(P_5, I_H)(t) &= \frac{\psi\tau}{AB(\tau)\Gamma(\tau)} \int_0^t s^{\psi-1}(t-s)^{\tau-1}G_5(s, I_H(s)) ds,
\end{aligned}$$

$$\begin{aligned}
(P_6, A)(t) &= \frac{\psi\tau}{AB(\tau)\Gamma(\tau)} \int_0^t s^{\psi-1}(t-s)^{\tau-1} G_6(s, A(s)) ds, \\
(P_7, T_H)(t) &= \frac{\psi\tau}{AB(\tau)\Gamma(\tau)} \int_0^t s^{\psi-1}(t-s)^{\tau-1} G_7(s, T_H(s)) ds.
\end{aligned} \tag{25}$$

Using (B_1) in Theorem 2, for any

$$(S_H(t), S_{PREP}(t), E_H(t), E_{PEP}(t), I_H(t), A(t), T_H(t)) \in D_{Z2}$$

and

$$(S_H^*(t), S_{PREP}^*(t), E_H^*(t), E_{PEP}^*(t), I_H^*(t), A^*(t), T_H^*(t)) \in D_{Z2},$$

from the above, with $L_{\max}$, it follows that

$$\begin{aligned}
\|Q_1 S_H - Q_1 S_H^*\| &\leq \frac{\psi(1-\tau)T_{\min}^{\psi-1}}{AB(\tau)} \sup |G_1(t, S_H(t)) - G_1(t, S_H^*(t))|, \\
\|Q_1 S_H - Q_1 S_H^*\| &\leq \frac{\psi(1-\tau)T_{\min}^{\psi-1}}{AB(\tau)} L_1 (\|S_H - S_H^*\| + \|S_{PREP} - S_{PREP}^*\| + \|E_H - E_H^*\| \\
&\quad + \|E_{PEP} - E_{PEP}^*\| + \|I_H - I_H^*\| + \|A - A^*\| + \|T_H - T_H^*\|), \\
\|Q_2 S_{PREP} - Q_2 S_{PREP}^*\| &\leq \frac{\psi(1-\tau)T_{\min}^{\psi-1}}{AB(\tau)} L_2 (\|S_H - S_H^*\| + \|S_{PREP} - S_{PREP}^*\| + \|E_H - E_H^*\| \\
&\quad + \|E_{PEP} - E_{PEP}^*\| + \|I_H - I_H^*\| + \|A - A^*\| + \|T_H - T_H^*\|), \\
\|Q_3 E_H - Q_3 E_H^*\| &\leq \frac{\psi(1-\tau)T_{\min}^{\psi-1}}{AB(\tau)} L_3 (\|S_H - S_H^*\| + \|S_{PREP} - S_{PREP}^*\| + \|E_H - E_H^*\| \\
&\quad + \|E_{PEP} - E_{PEP}^*\| + \|I_H - I_H^*\| + \|A - A^*\| + \|T_H - T_H^*\|), \\
\|Q_4 E_{PEP} - Q_4 E_{PEP}^*\| &\leq \frac{\psi(1-\tau)T_{\min}^{\psi-1}}{AB(\tau)} L_4 (\|S_H - S_H^*\| + \|S_{PREP} - S_{PREP}^*\| + \|E_H - E_H^*\| \\
&\quad + \|E_{PEP} - E_{PEP}^*\| + \|I_H - I_H^*\| + \|A - A^*\| + \|T_H - T_H^*\|), \\
\|Q_5 I_H - Q_5 I_H^*\| &\leq \frac{\psi(1-\tau)T_{\min}^{\psi-1}}{AB(\tau)} L_5 (\|S_H - S_H^*\| + \|S_{PREP} - S_{PREP}^*\| + \|E_H - E_H^*\| \\
&\quad + \|E_{PEP} - E_{PEP}^*\| + \|I_H - I_H^*\| + \|A - A^*\| + \|T_H - T_H^*\|),
\end{aligned}$$

$$\begin{aligned}
& \|Q_6 A - Q_6 A^*\| \\
& \leq \frac{\psi(1-\tau)T_{\min}^{\psi-1}}{AB(\tau)} L_6 (\|S_H - S_H^*\| + \|S_{PREP} - S_{PREP}^*\| + \|E_H - E_H^*\| \\
& \quad + \|E_{PEP} - E_{PEP}^*\| + \|I_H - I_H^*\| + \|A - A^*\| + \|T_H - T_H^*\|), \\
& \|Q_7 T_H - Q_7 T_H^*\| \\
& \leq \frac{\psi(1-\tau)T_{\min}^{\psi-1}}{AB(\tau)} L_7 (\|S_H - S_H^*\| + \|S_{PREP} - S_{PREP}^*\| + \|E_H - E_H^*\| \\
& \quad + \|E_{PEP} - E_{PEP}^*\| + \|I_H - I_H^*\| + \|A - A^*\| + \|T_H - T_H^*\|). \tag{26}
\end{aligned}$$

Thus, P is bounded and equicontinuous, ensuring the HIV/AIDS model has at least one solution under the given conditions.

The biological implication of this fractal-fractional existence theorem is that, regardless of whether the disease dynamics are highly nonlinear or intervention effects are complex, *at least one realistic disease trajectory exists* that is consistent with the model's assumptions and observed data. In practical terms, this means that the epidemic will follow *some* biologically plausible path over time—whether it declines due to effective interventions (e.g., high ART coverage, PrEP use) or persists at endemic levels. This guarantees that the model is mathematically sound and capable of producing valid simulations, even when strong contraction conditions (required for uniqueness) are not met. For public health, this result underpins the reliability of numerical forecasts (such as the projected 28% reduction in new infections by 2033) and ensures that policy simulations—testing “what-if” scenarios like increased prophylaxis or treatment access—are grounded in a solution that respects real-world constraints and disease biology. $\square$

4.1 The endemic equilibrium point of the HIV/AIDS model

The endemic equilibrium point refers to a state where HIV/AIDS persists in the population at constant levels.

Theorem 4. The endemic equilibrium point of the HIV/AIDS model in (1) is stable whenever $R_{0H} > 1$ and unstable if $R_{0H} < 1$.

Proof. At the endemic equilibrium point,

$$\frac{dS_H}{dt} = \frac{dS_{PREP}}{dt} = \frac{dE}{dt} = \frac{dE_{PEP}}{dt} = \frac{dI}{dt} = \frac{dI}{dt} = \frac{dA}{dt} = \frac{dT}{dt} = 0.$$

Let the endemic equilibrium point of the model (1) be

$$\widehat{\varepsilon} = \left(S_H^*, S_{PREP}^*, E_H^*, E_{PEP}^*, I_H^*, A^*, T_H^* \right).$$

By setting the right-hand side of the HIV/AIDS model equations in (1) to zero(0) and solving for each compartments, we obtain

$$\begin{aligned} S_H^* &= \frac{\Lambda_H}{\lambda_H^* + Z_1}, \\ S_{PREP}^* &= \frac{\omega_H \Lambda_H}{(\lambda_H^* + Z_1)(-\lambda_H^* \varepsilon_1 + \theta_H + \lambda_H^*)}, \\ E_H^* &= \frac{\lambda_H^* \Lambda_H (-\lambda_H^* \varepsilon_1 + \theta_H + Z_2 + \lambda_H^*)}{(\lambda_H^* + Z_1)(-\lambda_H^* \varepsilon_1 + \theta_H + \lambda_H^*) P_1}, \\ E_{PEP}^* &= \frac{\lambda_H^* \Lambda_H (-\lambda_H^* \varepsilon_1 + \theta_H + Z_2 + \lambda_H^*) v_{21}}{(\lambda_H^* + Z_1)(-\lambda_H^* \varepsilon_1 + \theta_H + \lambda_H^*) P_1 P_2}, \\ I_H^* &= \frac{(-\lambda_H^* \varepsilon_1 + \theta_H + Z_2 + \lambda_H^*) \lambda_H^* P_4 (P_2 v_{31} + v_{21} v_{32}) \Lambda_H P_5}{(\lambda_H^* + Z_1)(-\lambda_H^* \varepsilon_1 + \theta_H + \lambda_H^*) P_1 P_2 (P_3 P_4 P_5 - P_4 \chi \theta_1 - \chi \theta_2 \rho_1)}, \\ A^* &= \frac{(-\lambda_H^* \varepsilon_1 + \theta_H + Z_2 + \lambda_H^*) \lambda_H^* (P_2 v_{31} + v_{21} v_{32}) \Lambda_H P_5 \rho_1}{(\lambda_H^* + Z_1)(-\lambda_H^* \varepsilon_1 + \theta_H + \lambda_H^*) P_1 P_2 (P_3 P_4 P_5 - P_4 \chi \theta_1 - \chi \theta_2 \rho_1)}, \\ T_H^* &= \frac{(P_4 \theta_1 + \theta_2 \rho_1)(-\lambda_H^* \varepsilon_1 + \theta_H + Z_2 + \lambda_H^*) \lambda_H^* (P_2 v_{31} + v_{21} v_{32}) \Lambda_H}{(\lambda_H^* + Z_1)(-\lambda_H^* \varepsilon_1 + \theta_H + \lambda_H^*) P_1 P_2 (P_3 P_4 P_5 - P_4 \chi \theta_1 - \chi \theta_2 \rho_1)}. \end{aligned} \quad (27)$$

Here

$$\begin{aligned} v_{21} &= \theta \sigma_1, v_{31} = (1 - \theta) \sigma_1, v_{32} = (1 - \varepsilon_2) \sigma_2, \\ P_1 &= (\sigma_1 + \theta_H), P_2 = ((1 - \varepsilon_2) \sigma_2 + \theta_H), P_3 = (\theta_1 + \rho_1 + \delta_I + \theta_H), \\ P_4 &= (\theta_2 + \delta_A + \theta_H), P_5 = (\chi + \Theta \delta_T + \theta_H), Z_1 = (\omega_H + \theta_H), \\ Z_2 &= ((1 - \varepsilon_1) \omega_H + \theta_H). \end{aligned}$$

Further simplification of the endemic equilibrium point of the model as presented in equation (27) yields the final endemic equilibrium point.

Substituting the endemic equilibrium point of the HIV/AIDS model $\widehat{\varepsilon}$ into its force of infection

$$\lambda_H^* = \frac{(1-\varsigma\Psi)\gamma\gamma_H \left(I_H^* + A^* + \eta T_H^* \right)}{N^*}, \text{ we obtain}$$

$$\lambda_H^* \left(\widehat{C}_1 \lambda_H^{*4} + \widehat{C}_2 \lambda_H^{*3} + \widehat{C}_3 \lambda_H^{*2} + \widehat{C}_4 \lambda_H^* + \widehat{C}_5 \right) = 0. \quad (28)$$

But at the endemic equilibrium ($\lambda_H \neq 0$), hence

$$\widehat{C}_1 \lambda_H^{*4} + \widehat{C}_2 \lambda_H^{*3} + \widehat{C}_3 \lambda_H^{*2} + \widehat{C}_4 \lambda_H^* + \widehat{C}_5 = 0, \quad (29)$$

where

$$\begin{aligned}\widehat{C}_1 = & \Lambda_H ((-P_3 P_5 + \chi \theta_1) P_4 + \chi \theta_2 \rho_1) P_1 P_2 (\varepsilon_1 - 1)^2 \\ & \times ((((-P_3 - v_{31}) P_5 + \theta_1 (\chi - v_{31})) P_2 \\ & + ((-P_3 - v_{32}) P_5 + \theta_1 (\chi - v_{32})) v_{21}) P_4 \\ & + ((-v_{31} P_5 + \theta_2 (\chi - v_{31})) P_2 + v_{21} (-v_{32} P_5 + \theta_2 (\chi - v_{32}))) \rho_1),\end{aligned}$$

$$\begin{aligned}\widehat{C}_2 = & -((-P_3 P_5 + \chi \theta_1) P_4 + \chi \theta_2 \rho_1) P_1 (\varepsilon_1 - 1) \Lambda_H \\ & \times [((\gamma v_{31} (\Psi \varsigma - 1) \gamma_H + v_{31} Z_1 + P_3 (P_1 + Z_1)) \varepsilon_1 - \gamma v_{31} (\Psi \varsigma - 1) \gamma_H \\ & + (-Z_2 - 2\theta_H - Z_1) v_{31} - P_3 (Z_2 + 2\theta_H + P_1 + Z_1)) P_5 \\ & + ((\gamma v_{31} \eta (\Psi \varsigma - 1) \gamma_H + v_{31} Z_1 - \chi (P_1 + Z_1)) \varepsilon_1 - \gamma v_{31} \eta (\Psi \varsigma - 1) \gamma_H \\ & + (-Z_2 - 2\theta_H - Z_1) v_{31} + \chi (Z_2 + 2\theta_H + P_1 + Z_1)) \theta_1) P_2 \\ & + v_{21} [((\gamma v_{32} (\Psi \varsigma - 1) \gamma_H + Z_1 (v_{32} + P_3)) \varepsilon_1 \\ & - \gamma v_{32} (\Psi \varsigma - 1) \gamma_H - (v_{32} + P_3) (Z_2 + 2\theta_H + Z_1)) P_5 \\ & + ((\gamma v_{32} \eta (\Psi \varsigma - 1) \gamma_H + Z_1 (v_{32} - \chi)) \varepsilon_1 \\ & - \gamma v_{32} \eta (\Psi \varsigma - 1) \gamma_H - (v_{32} - \chi) (Z_2 + 2\theta_H + Z_1)) \theta_1] P_4 \\ & + [((\gamma (\Psi \varsigma - 1) \gamma_H + Z_1) \varepsilon_1 \\ & + (-\Psi \varsigma + 1) \gamma \gamma_H - Z_2 - 2\theta_H - Z_1) v_{31} P_5 \\ & + \theta_2 ((\gamma v_{31} \eta (\Psi \varsigma - 1) \gamma_H + v_{31} Z_1 - \chi (P_1 + Z_1)) \varepsilon_1 \\ & - \gamma v_{31} \eta (\Psi \varsigma - 1) \gamma_H + (-Z_2 - 2\theta_H - Z_1) v_{31} \\ & + \chi (Z_2 + 2\theta_H + P_1 + Z_1))] P_2 + ((\gamma (\Psi \varsigma - 1) \gamma_H + Z_1) \varepsilon_1 \\ & + (-\Psi \varsigma + 1) \gamma \gamma_H - Z_2 - 2\theta_H - Z_1) v_{32} P_5 \\ & + ((\gamma v_{32} \eta (\Psi \varsigma - 1) \gamma_H + Z_1 (v_{32} - \chi)) \varepsilon_1 \\ & - \gamma v_{32} \eta (\Psi \varsigma - 1) \gamma_H - (v_{32} - \chi) (Z_2 + 2\theta_H + Z_1)) \theta_2) v_{21}) \rho_1] P_2,\end{aligned}$$

$$\widehat{C}_3 = \left[(\mathcal{A}_1 + \rho_1 \mathcal{A}_2 + \mathcal{A}_3 + \rho_1 \mathcal{A}_4) v_{21} \right] ((-P_3 P_5 + \chi \theta_1) P_4 + \chi \theta_2 \rho_1) P_1 \Lambda_H P_2,$$

where

$$\begin{aligned}\mathcal{A}_1 = & -(\varepsilon_1 - 1) \cdot \left[\gamma v_{31} (\varepsilon_1 - 1) (\eta \theta_1 + P_5) (\Psi \varsigma - 1) \gamma_H \right. \\ & \left. - P_1 (-P_3 P_5 + \chi \theta_1) \varepsilon_1 + \mathcal{X}_1 - \theta_1 \mathcal{Y}_1 \right] Z_1 \\ & + \gamma v_{31} (\varepsilon_1 - 1) (\eta \theta_1 + P_5) (Z_2 + 2\theta_H) (\Psi \varsigma - 1) \gamma_H\end{aligned}$$

$$-2P_1(-P_3P_5 + \chi\theta_1)(\theta_H + \frac{\omega_H}{2})\varepsilon_1 + \mathcal{X}_2 - \theta_1\mathcal{Y}_2)P_4,$$

$$\begin{aligned}\mathcal{A}_2 = & -(\varepsilon_1 - 1) \cdot \left[\gamma v_{31}(\varepsilon_1 - 1)(\eta\theta_2 + P_5)(\Psi_\varsigma - 1)\gamma_H \right. \\ & \left. - \theta_2\chi\varepsilon_1P_1 - v_{31}(Z_2 + 2\theta_H)P_5 - \theta_2\mathcal{Y}_1 \right] Z_1 \\ & + \gamma v_{31}(\varepsilon_1 - 1)(\eta\theta_2 + P_5)(Z_2 + 2\theta_H)(\Psi_\varsigma - 1)\gamma_H \\ & - 2\theta_2\chi P_1(\theta_H + \frac{\omega_H}{2})\varepsilon_1 - v_{31}\theta_H(Z_2 + \theta_H)P_5 - \theta_2\mathcal{Y}_2),\end{aligned}$$

$$\begin{aligned}\mathcal{A}_3 = & -(\varepsilon_1 - 1) \cdot \left[\gamma v_{32}(\varepsilon_1 - 1)(\eta\theta_1 + P_5)(\Psi_\varsigma - 1)\gamma_H \right. \\ & \left. - ((v_{32} + P_3)P_5 + \theta_1(v_{32} - \chi))(Z_2 + 2\theta_H) \right] Z_1 \\ & + \gamma v_{32}(\varepsilon_1 - 1)(\eta\theta_1 + P_5)(Z_2 + 2\theta_H)(\Psi_\varsigma - 1)\gamma_H \\ & - \theta_H(Z_2 + \theta_H)((v_{32} + P_3)P_5 + \theta_1(v_{32} - \chi))P_4,\end{aligned}$$

$$\begin{aligned}\mathcal{A}_4 = & -(\varepsilon_1 - 1) \cdot \left[\gamma v_{32}(\varepsilon_1 - 1)(\eta\theta_2 + P_5)(\Psi_\varsigma - 1)\gamma_H \right. \\ & \left. - (Z_2 + 2\theta_H)(v_{32}P_5 + \theta_2(v_{32} - \chi)) \right] Z_1 \\ & + \gamma v_{32}(\varepsilon_1 - 1)(\eta\theta_2 + P_5)(Z_2 + 2\theta_H)(\Psi_\varsigma - 1)\gamma_H \\ & - \theta_H(Z_2 + \theta_H)(v_{32}P_5 + \theta_2(v_{32} - \chi)),\end{aligned}$$

$$\mathcal{X}_1 = (-2v_{31} - 2P_3)\theta_H - Z_2v_{31} - P_3(Z_2 + P_1),$$

$$\mathcal{Y}_1 = (2v_{31} - 2\chi)\theta_H + Z_2v_{31} - \chi(Z_2 + P_1),$$

$$\mathcal{X}_2 = (-P_3 - v_{31})\theta_H^2 + (-Z_2v_{31} - P_3(Z_2 + 2P_1))\theta_H - \omega_H P_1P_3,$$

$$\mathcal{Y}_2 = (v_{31} - \chi)\theta_H^2 + (Z_2v_{31} - \chi(Z_2 + 2P_1))\theta_H - \chi\omega_H P_1,$$

$$\begin{aligned}\widehat{C}_4 = & - \left(\left[((v_{31} + P_3)P_5 + \theta_1(v_{31} - \chi))P_4 + \rho_1(v_{31}P_5 + \theta_2(v_{31} - \chi)) \right] P_2 \right. \\ & + v_{21} \left[((v_{32} + P_3)P_5 + \theta_1(v_{32} - \chi))P_4 + \rho_1(v_{32}P_5 + \theta_2(v_{32} - \chi)) \right] Z_1 \\ & + (P_2v_{31} + v_{21}v_{32})\gamma((\eta\theta_1 + P_5)P_4 + \rho_1(\eta\theta_2 + P_5))(\Psi_\varsigma - 1)\gamma_H \\ & - \left((-P_3P_5 + \chi\theta_1)P_4 + \chi\theta_2\rho_1 \right) P_1P_2 \theta_H^2 \\ & + \left[-2(P_2v_{31} + v_{21}v_{32})\gamma(\varepsilon_1 - 1)((\eta\theta_1 + P_5)P_4 + \rho_1(\eta\theta_2 + P_5))(\Psi_\varsigma - 1)\gamma_H \right. \\ & \left. + \left(((v_{31} + P_3)Z_2 - 2P_1P_3(\varepsilon_1 - 1))P_5 + \theta_1((v_{31} - \chi)Z_2 + 2\chi P_1(\varepsilon_1 - 1)) \right) P_4 \right]\end{aligned}$$

$$\begin{aligned}
& + \left(Z_2 v_{31} P_5 + \theta_2 ((v_{31} - \chi) Z_2 + 2\chi P_1 (\varepsilon_1 - 1)) \right) \rho_1 P_2 \\
& + Z_2 v_{21} \left(((v_{32} + P_3) P_5 + \theta_1 (v_{32} - \chi)) P_4 + \rho_1 (v_{32} P_5 + \theta_2 (v_{32} - \chi)) \right) \Big] Z_1 \\
& + (P_2 v_{31} + v_{21} v_{32}) \gamma Z_2 \left((\eta \theta_1 + P_5) P_4 + \rho_1 (\eta \theta_2 + P_5) \right) (\Psi \varsigma - 1) \gamma_H \\
& - \left((-P_3 P_5 + \chi \theta_1) P_4 + \chi \theta_2 \rho_1 \right) P_1 \omega_H P_2 \Big) \theta_H \\
& - (\varepsilon_1 - 1) Z_1 \left[(P_2 v_{31} + v_{21} v_{32}) \gamma Z_2 \left((\eta \theta_1 + P_5) P_4 + \rho_1 (\eta \theta_2 + P_5) \right) (\Psi \varsigma - 1) \gamma_H \right. \\
& \left. - \left((-P_3 P_5 + \chi \theta_1) P_4 + \chi \theta_2 \rho_1 \right) P_1 \omega_H P_2 \right] \left((-P_3 P_5 + \chi \theta_1) P_4 + \chi \theta_2 \rho_1 \right) P_1 \Lambda_H P_2, \\
\\
\widehat{C}_5 & = Z_1 \theta_H \Lambda_H P_1 P_2 (P_3 P_4 P_5 - P_4 \chi \theta_1 - \chi \theta_2 \rho_1) \\
& \times (((P_2 v_{31} + v_{21} v_{32}) ((\eta \theta_1 + P_5) P_4 + \rho_1 (\eta \theta_2 + P_5)) (\Psi \varsigma - 1) \gamma \gamma_H \\
& - P_1 ((-P_3 P_5 + \chi \theta_1) P_4 + \chi \theta_2 \rho_1) P_2) \theta_H \\
& - P_1 ((-P_3 P_5 + \chi \theta_1) P_4 + \chi \theta_2 \rho_1) P_2 \omega_H) \\
& + Z_1^2 \Lambda_H (\theta_H (P_1 P_2 (P_3 P_4 P_5 - P_4 \chi \theta_1 - \chi \theta_2 \rho_1)))^2 (1 - R_{0H}).
\end{aligned}$$

The components of the endemic equilibrium point (EEP) are determined by solving for λ_H^* from the polynomial equation (29) and substituting its positive values into the expressions given in (27). Moreover, based on the coefficients $\widehat{C}_i$ for $i = 1, 2, 3, \dots, 5$, it is evident that $\widehat{C}_1$ is always positive, while $\widehat{C}_5$ is positive when $R_{0H} < 1$ and negative when $R_{0H} > 1$. From this, we derive the following results:

Theorem 5. The HIV/AIDS model (1) has

1. four or two endemic equilibria if $\widehat{C}_2 < 0, \widehat{C}_3 > 0, \widehat{C}_4 < 0$ and $R_{0H} < 1$
2. four or two endemic equilibria if $\widehat{C}_2 < 0, \widehat{C}_3 < 0, \widehat{C}_4 > 0$ and $R_{0H} < 1$
3. no endemic equilibrium otherwise if $R_{0H} < 1$.

Items (1)–(3) of Theorem 5 indicate the potential occurrence of backward bifurcation in the HIV/AIDS model (1) when $R_{0H} < 1$. The presence of backward bifurcation implies that the conventional condition of having the basic reproduction number R_{0H} less than one, while necessary, is no longer sufficient to ensure the elimination of the disease from the population.

The biological implication of this fractal-fractional equilibrium analysis is profound for HIV control: When $R_{0H} > 1$, the disease settles into a stable, persistent state—meaning HIV will

remain in the population at a constant level unless interventions are dramatically strengthened. Conversely, when $R_{0H} < 1$, the disease-free state is stable, and elimination is possible—but the risk of *backward bifurcation* warns that simply reducing R_{0H} below 1 may not be enough. In real-world terms, this means that even with moderate success in interventions (e.g., partial PrEP coverage, inconsistent condom use, or incomplete ART access), HIV can persist at low but stable levels due to residual transmission chains, behavioral heterogeneity, or treatment dropout. This phenomenon explains why some regions struggle to eliminate HIV despite $R_{0H} < 1$. The model thus urges policymakers to push R_{0H} *significantly below 1* and sustain high intervention synergy (condoms + PrEP + ART) to avoid being trapped in a low-level endemic trap—critical insight for achieving the UNAIDS 95-95-95 targets and ending AIDS as a public health threat by 2030. $\square$

4.2 Global and asymptotic stability of HIV/AIDS at endemic equilibrium

The purpose of this work is to demonstrate the global asymptotic stability of HIV/AIDS endemic equilibrium points $\hat{e}$.

Lemma 4. If $R_{0H} > 1$, then additional conditions allow for the proof of global asymptotic stability at the endemic equilibrium for system (1).

Proof. Let the Lyapunov function $U = U(S_H, S_{PREP}, E_H, E_{PEP}, I_H, A, T_H)$ be defined as follows:

$$\begin{aligned}
 U = & U_1 \left(S_H - S_H^* - S_H^* \ln \left(\frac{S_H}{S_H^*} \right) \right) \\
 & + U_2 \left(S_{PREP} - S_{PREP}^* - S_{PREP}^* \ln \left(\frac{S_{PREP}}{S_{PREP}^*} \right) \right) \\
 & + U_3 \left(E_H - E_H^* - E_H^* \ln \left(\frac{E_H}{E_H^*} \right) \right) \\
 & + U_4 \left(E_{PEP} - E_{PEP}^* - E_{PEP}^* \ln \left(\frac{E_{PEP}}{E_{PEP}^*} \right) \right) \\
 & + U_5 \left(I_H - I_H^* - I_H^* \ln \left(\frac{I_H}{I_H^*} \right) \right) + U_6 \left(A - A^* - A^* \ln \left(\frac{A}{A^*} \right) \right) \\
 & + U_7 \left(T_H - T_H^* - T_H^* \ln \left(\frac{T_H}{T_H^*} \right) \right).
 \end{aligned} \tag{30}$$

By differentiating the Lyapunov function U with respect to time t along the positive system solutions of (30), we obtain

$$\begin{aligned}
\frac{dU}{dt} = & U_1 \left(1 - \frac{S_H^*}{S_H} \right) \frac{dS_H}{dt} + U_2 \left(1 - \frac{S_{PREP}^*}{S_{PREP}} \right) \frac{dS_{PREP}}{dt} \\
& + U_3 \left(1 - \frac{E_H^*}{E_H} \right) \frac{dE_H}{dt} + U_4 \left(1 - \frac{E_{PEP}^*}{E_{PEP}} \right) \frac{dE_{PEP}}{dt} \\
& + U_5 \left(1 - \frac{I_H^*}{I_H} \right) \frac{dI_H}{dt} + U_6 \left(1 - \frac{A^*}{A} \right) \frac{dA}{dt} \\
& + U_7 \left(1 - \frac{T_H^*}{T_H} \right) \frac{dT_H}{dt}.
\end{aligned} \tag{31}$$

Substituting the derivatives into the above, we have

$$\begin{aligned}
\frac{dU}{dt} = & U_1 \left(1 - \frac{S_H^*}{S_H} \right) (\Lambda - \lambda_H S_H - Z_1 S_H) \\
& + U_2 \left(1 - \frac{S_{PREP}^*}{S_{PREP}} \right) (\omega_H S_H - (1 - \varepsilon_1) \lambda_H S_{PREP} - Z_2 S_{PREP}) \\
& + U_3 \left(1 - \frac{E_H^*}{E_H} \right) (\lambda_H S_H + (1 - \varepsilon_1) \lambda_H S_{PREP} - P_1 E_H) \\
& + U_4 \left(1 - \frac{E_{PEP}^*}{E_{PEP}} \right) (v_{21} E_H - P_2 E_{PEP}) \\
& + U_5 \left(1 - \frac{I_H^*}{I_H} \right) (v_{31} E_H + v_{32} E_{PEP} - P_3 I_H + \chi T_H) \\
& + U_6 \left(1 - \frac{A^*}{A} \right) (\rho_1 I_H - P_4 A) \\
& + U_7 \left(1 - \frac{T_H^*}{T_H} \right) (\theta_1 I_H + \theta_2 A - P_5 T_H),
\end{aligned} \tag{32}$$

where the parameters are defined as follows:

$$\begin{aligned}
Z_1 &= \omega_H + \theta_H, Z_2 = (1 - \varepsilon_1) \lambda_H + \theta_H, P_1 = \sigma_1 + \theta_H, \\
P_2 &= (1 - \varepsilon_2) \sigma_2 + \theta_H, P_3 = \theta_1 + \rho_1 + \delta_I + \theta_H, P_4 = \theta_2 + \delta_A + \theta_H, \\
P_5 &= \chi + \Theta \delta_T + \theta_H.
\end{aligned} \tag{33}$$

At the endemic equilibrium point, we have

$$\begin{aligned}
 \Lambda &= \lambda_H S_H^* + Z_1 S_H^*, \\
 P_1 &= \frac{\lambda_H S_H^* + (1 - \varepsilon_1) \lambda_H S_{PREP}^*}{E_H^*}, \\
 P_2 &= \frac{v_{21} E_H^*}{E_{PEP}^*}, \\
 P_3 &= \frac{v_{31} E_H^* + v_{32} E_{PEP}^* + \chi T_H^*}{I_H^*}, \\
 P_4 &= \frac{\rho_1 I_H^*}{A^*}, \\
 P_5 &= \frac{\theta_1 I_H^* + \theta_2 A^*}{T_H^*}.
 \end{aligned} \tag{34}$$

Substituting $\Lambda, Z_1, P_1, P_2, P_3, P_4, P_5$ into the expression for $\frac{dU}{dt}$, we have

$$\begin{aligned}
 \frac{dU}{dt} = & U_1 \left(1 - \frac{S_H^*}{S_H} \right) (\lambda_H S_H^* + Z_1 S_H^* - (\lambda_H S_H + Z_1 S_H)) \\
 & + U_2 \left(1 - \frac{S_{PREP}^*}{S_{PREP}} \right) \left(\omega_H S_H - \frac{\omega_H S_H^*}{S_{PREP}^*} S_{PREP} \right) \\
 & + U_3 \left(1 - \frac{E_H^*}{E_H} \right) (\lambda_H S_H + (1 - \varepsilon_1) \lambda_H S_{PREP} \\
 & \quad - \frac{\lambda_H S_H^* + (1 - \varepsilon_1) \lambda_H S_{PREP}^*}{E_H^*} E_H) \\
 & + U_4 \left(1 - \frac{E_{PEP}^*}{E_{PEP}} \right) \left(v_{21} E_H - \frac{v_{21} E_H^*}{E_{PEP}^*} E_{PEP} \right) \\
 & + U_5 \left(1 - \frac{I_H^*}{I_H} \right) (v_{31} E_H + v_{32} E_{PEP} + \chi T_H \\
 & \quad - \frac{v_{31} E_H^* + v_{32} E_{PEP}^* + \chi T_H^*}{I_H^*} I_H) \\
 & + U_6 \left(1 - \frac{A^*}{A} \right) \left(\rho_1 I_H - \frac{\rho_1 I_H^*}{A^*} A \right) \\
 & + U_7 \left(1 - \frac{T_H^*}{T_H} \right) \left(\theta_1 I_H + \theta_2 A - \frac{\theta_1 I_H^* + \theta_2 A^*}{T_H^*} T_H \right).
 \end{aligned} \tag{35}$$

We obtain after performing expansion and collection of the positive and negative terms: $\frac{dU}{dt} = A - B$, where

$$\begin{aligned}
 A = & \lambda_H S_H^* U_1 + Z_1 S_H^* U_1 + \frac{\lambda_H S_H^{*2} U_1}{S_H} + \frac{Z_1 S_H^{*2} U_1}{S_H} \\
 & + \lambda_H S_H U_3 + (1 - \varepsilon_1) \lambda_H S_{PREP} U_3 + \lambda_H S_H^* U_3 + (1 - \varepsilon_1) \lambda_H S_{PREP}^* U_3,
 \end{aligned} \tag{36}$$

and

$$\begin{aligned}
 B = & \lambda_H S_H^* U_1 + Z_1 S_H^* U_1 + Z_1 S_H^* U_1 + \frac{\lambda_H S_H^{*2} U_1}{S_H} + \frac{Z_1 S_H^2 U_1}{S_H} \\
 & + \frac{\lambda_H S_H^{*2} E_H U_3}{E_H^*} + \frac{\lambda_H S_H E_H^* U_3}{E_H} + (1 - \varepsilon_1) \frac{\lambda_H S_{PREP} E_H^* U_3}{E_H} \\
 & + \frac{U_5 v_{31} E_H^* I_H}{I_H^*} + \frac{U_5 v_{31} E_H I_H^*}{I_H} + \frac{U_5 v_{32} E_{PEP}^* I_H}{I_H^*} + \frac{U_5 v_{32} E_{PEP} I_H^*}{I_H}.
 \end{aligned} \tag{37}$$

Hence, since $B < A$, we have $\frac{dU}{dt} \leq 0$ with $\frac{dU}{dt} = 0$ if and only if

$$S_H = S_H^*, S_{PREP} = S_{PREP}^*, E_H = E_H^*, E_{PEP} = E_{PEP}^*, I_H = I_H^*, A = A^*, T_H = T_H^*.$$

Therefore, the largest compact invariant set is a singleton set $\{E_H^*\}$ that contains the endemic equilibrium point of the system (1). The stability analysis demonstrates that the system (1) $\hat{e}$ is globally and asymptotically stable in accordance with La Salle's invariant principle as indicated in La Salle [32]:

$$\{(S_H^*, S_{PREP}^*, E_H^*, E_{PEP}^*, I_H^*, A^*, T_H^*) \in \Omega_H : \frac{dU}{dt} = 0\}.$$

The biological implication of this global asymptotic stability result is that, when $R_{0H} > 1$, the HIV epidemic will inevitably converge to a persistent, self-sustaining level—regardless of initial conditions (e.g., a small outbreak, a major surge, or partial intervention rollout). In real-world terms, this means that once transmission exceeds the critical threshold, HIV becomes endemic and irreversible without drastic, sustained intervention. The population will settle into a predictable, long-term state with fixed proportions of infected, treated, and at-risk individuals. This result is critical for public health planning: It confirms that half-measures will not work—PrEP, ART, and condom promotion must be scaled aggressively and maintained indefinitely to push R_{0H} below 1 and prevent convergence to this stable endemic trap. For policymakers, this underscores the urgency of achieving and sustaining the UNAIDS 95-95-95 targets; any lapse risks locking the epidemic into a costly, lifelong burden on healthcare systems and communities. $\square$

4.3 Numerical schemes for the fractal-fractional HIV/AIDS model

The second-order Adams–Bashforth algorithm will be used to create numerical solutions for the HIV/AIDS model (1).

We rewrite the integral equations, at $t = t_{n+1}$, this yields

$$\begin{aligned}
S_H(t_{n+1}) &= S_{H0} + \psi(1-\tau)t_n^{\psi-1}G_1(t_n, S_H(t_n)) \\
&\quad + \frac{\tau\psi}{AB(\tau)\Gamma(\tau)} \int_0^{t_{n+1}} (t_{n+1}-s)^{\tau-1} S_H^{\psi-1}(s, S_H(s)) ds, \\
S_{PREP}(t_{n+1}) &= S_{PREP0} + \psi(1-\tau)t_n^{\psi-1}G_2(t_n, S_{PREP}(t_n)) \\
&\quad + \frac{\tau\psi}{AB(\tau)\Gamma(\tau)} \int_0^{t_{n+1}} (t_{n+1}-s)^{\tau-1} S_{PREP}^{\psi-1}(s, S_{PREP}(s)) ds, \\
E_H(t_{n+1}) &= E_{H0} + \psi(1-\tau)t_n^{\psi-1}G_3(t_n, E_H(t_n)) \\
&\quad + \frac{\tau\psi}{AB(\tau)\Gamma(\tau)} \int_0^{t_{n+1}} (t_{n+1}-s)^{\tau-1} E_H^{\psi-1}(s, E_H(s)) ds, \\
E_{PEP}(t_{n+1}) &= E_{PEP0} + \psi(1-\tau)t_n^{\psi-1}G_4(t_n, E_{PEP}(t_n)) \\
&\quad + \frac{\tau\psi}{AB(\tau)\Gamma(\tau)} \int_0^{t_{n+1}} (t_{n+1}-s)^{\tau-1} E_{PEP}^{\psi-1}(s, E_{PEP}(s)) ds, \\
I_H(t_{n+1}) &= I_{H0} + \psi(1-\tau)t_n^{\psi-1}G_5(t_n, I_H(t_n)) \\
&\quad + \frac{\tau\psi}{AB(\tau)\Gamma(\tau)} \int_0^{t_{n+1}} (t_{n+1}-s)^{\tau-1} I_H^{\psi-1}(s, I_H(s)) ds, \\
A(t_{n+1}) &= A_0 + \psi(1-\tau)t_n^{\psi-1}G_6(t_n, A(t_n)) \\
&\quad + \frac{\tau\psi}{AB(\tau)\Gamma(\tau)} \int_0^{t_{n+1}} (t_{n+1}-s)^{\tau-1} A^{\psi-1}(s, A(s)) ds, \\
T_H(t_{n+1}) &= T_{H0} + \psi(1-\tau)t_n^{\psi-1}G_7(t_n, T_H(t_n)) \\
&\quad + \frac{\tau\psi}{AB(\tau)\Gamma(\tau)} \int_0^{t_{n+1}} (t_{n+1}-s)^{\tau-1} T_H^{\psi-1}(s, T_H(s)) ds,
\end{aligned} \tag{38}$$

So the numerical approximation of the integral is obtained as

$$\begin{aligned}
S_H(t_{n+1}) &= S_{H0} + \frac{\psi(1-\tau)t_n^{\psi-1}G_1(t_n, S_H(t_n))}{AB(\tau)} \\
&\quad + \frac{\tau\psi}{AB(\tau)\Gamma(\tau)} \sum_{j=1}^n \int_{t_j}^{t_{j+1}} (t_{j+1}-s)^{\tau-1} S_H^{\psi-1}(t_n, S_H^j(s)) ds, \\
S_{PREP}(t_{n+1}) &= S_{PREP0} + \frac{\psi(1-\tau)t_n^{\psi-1}G_2(t_n, S_{PREP}(t_n))}{AB(\tau)} \\
&\quad + \frac{\tau\psi}{AB(\tau)\Gamma(\tau)} \sum_{j=1}^n \int_{t_j}^{t_{j+1}} (t_{j+1}-s)^{\tau-1} S_{PREP}^{\psi-1}(t_n, S_{PREP}^j(s)) ds, \\
E_H(t_{n+1}) &= E_{H0} + \frac{\psi(1-\tau)t_n^{\psi-1}G_3(t_n, E_H(t_n))}{AB(\tau)} \\
&\quad + \frac{\tau\psi}{AB(\tau)\Gamma(\tau)} \sum_{j=1}^n \int_{t_j}^{t_{j+1}} (t_{j+1}-s)^{\tau-1} E_H^{\psi-1}(t_n, E_H^j(s)) ds,
\end{aligned}$$

$$\begin{aligned}
E_{PEP}(t_{n+1}) &= E_{PEP0} + \frac{\psi(1-\tau)t_n^{\psi-1}G_4(t_n, E_{PEP}(t_n))}{AB(\tau)} \\
&\quad + \frac{\tau\psi}{AB(\tau)\Gamma(\tau)} \sum_{j=1}^n \int_{t_j}^{t_{j+1}} (t_{j+1}-s)^{\tau-1} E_{PEP}^{\psi-1} G_4(t_n, E_{PEP}^j(s)) ds, \\
I_H(t_{n+1}) &= I_{H0} + \frac{\psi(1-\tau)t_n^{\psi-1}G_5(t_n, I_H(t_n))}{AB(\tau)} \\
&\quad + \frac{\tau\psi}{AB(\tau)\Gamma(\tau)} \sum_{j=1}^n \int_{t_j}^{t_{j+1}} (t_{j+1}-s)^{\tau-1} I_H^{\psi-1} G_5(t_n, I_H^j(s)) ds, \\
A(t_{n+1}) &= A_0 + \frac{\psi(1-\tau)t_n^{\psi-1}G_6(t_n, A(t_n))}{AB(\tau)} \\
&\quad + \frac{\tau\psi}{AB(\tau)\Gamma(\tau)} \sum_{j=1}^n \int_{t_j}^{t_{j+1}} (t_{j+1}-s)^{\tau-1} A^{\psi-1} G_6(t_n, A^j(s)) ds, \\
T_H(t_{n+1}) &= T_{H0} + \frac{\psi(1-\tau)t_n^{\psi-1}G_7(t_n, T_H(t_n))}{AB(\tau)} \\
&\quad + \frac{\tau\psi}{AB(\tau)\Gamma(\tau)} \sum_{j=1}^n \int_{t_j}^{t_{j+1}} (t_{j+1}-s)^{\tau-1} T_H^{\psi-1} G_7(t_n, T_H^j(s)) ds,
\end{aligned} \tag{39}$$

Two-step Lagrange's interpolation polynomials with $\Delta t = t_{j+1} - t_j$ function as an estimate for the integrand functions on $[t_j, t_{j+1}]$. Thus, we obtain

$$\begin{aligned}
S_H(t_{n+1}) &= S_{H0} + \frac{\psi(1-\tau)t_n^{\psi-1}G_1(t_n, S_H(t_n))}{AB(\tau)} \\
&\quad + \frac{\tau\psi}{AB(\tau)\Gamma(\tau)} \sum_{j=1}^n \int_{t_j}^{t_{j+1}} (t_{j+1}-s)^{\tau-1} S_H^{\psi-1} Z_j^{S_H}(s) ds, \\
S_{PREP}(t_{n+1}) &= S_{PREP0} + \frac{\psi(1-\tau)t_n^{\psi-1}G_2(t_n, S_{PREP}(t_n))}{AB(\tau)} \\
&\quad + \frac{\tau\psi}{AB(\tau)\Gamma(\tau)} \sum_{j=1}^n \int_{t_j}^{t_{j+1}} (t_{j+1}-s)^{\tau-1} S_{PREP}^{\psi-1} Z_j^{S_{PREP}}(s) ds, \\
E_H(t_{n+1}) &= E_{H0} + \frac{\psi(1-\tau)t_n^{\psi-1}G_3(t_n, E_H(t_n))}{AB(\tau)} \\
&\quad + \frac{\tau\psi}{AB(\tau)\Gamma(\tau)} \sum_{j=1}^n \int_{t_j}^{t_{j+1}} (t_{j+1}-s)^{\tau-1} E_H^{\psi-1} Z_j^{E_H}(s) ds, \\
E_{PEP}(t_{n+1}) &= E_{PEP0} + \frac{\psi(1-\tau)t_n^{\psi-1}G_4(t_n, E_{PEP}(t_n))}{AB(\tau)} \\
&\quad + \frac{\tau\psi}{AB(\tau)\Gamma(\tau)} \sum_{j=1}^n \int_{t_j}^{t_{j+1}} (t_{j+1}-s)^{\tau-1} E_{PEP}^{\psi-1} Z_j^{E_{PEP}}(s) ds,
\end{aligned}$$

$$\begin{aligned}
 I_H(t_{n+1}) &= I_{H0} + \frac{\psi(1-\tau)t_n^{\psi-1}G_5(t_n, I_H(t_n))}{AB(\tau)} \\
 &\quad + \frac{\tau\psi}{AB(\tau)\Gamma(\tau)} \sum_{j=1}^n \int_{t_j}^{t_{j+1}} (t_{j+1}-s)^{\tau-1} I_H^{\psi-1} Z_j^{I_H}(s) ds, \\
 A(t_{n+1}) &= A_0 + \frac{\psi(1-\tau)t_n^{\psi-1}G_6(t_n, A(t_n))}{AB(\tau)} \\
 &\quad + \frac{\tau\psi}{AB(\tau)\Gamma(\tau)} \sum_{j=1}^n \int_{t_j}^{t_{j+1}} (t_{j+1}-s)^{\tau-1} A^{\psi-1} Z_j^A(s) ds, \\
 T_H(t_{n+1}) &= T_{H0} + \frac{\psi(1-\tau)t_n^{\psi-1}G_7(t_n, T_H(t_n))}{AB(\tau)} \\
 &\quad + \frac{\tau\psi}{AB(\tau)\Gamma(\tau)} \sum_{j=1}^n \int_{t_j}^{t_{j+1}} (t_{j+1}-s)^{\tau-1} T_H^{\psi-1} Z_j^{T_H}(s) ds,
 \end{aligned} \tag{40}$$

where

$$\begin{aligned}
 Z_j^{S_H}(s) &= \frac{s-t_{j-1}}{t_j-t_{j-1}} t_j^{\psi-1} G_1(s, S_H^j(s)) - \frac{s-t_{j-1}}{t_j-t_{j-1}} t_{j-1}^{\psi-1} G_1(s, S_H^{j-1}(s)), \\
 Z_j^{S_{PREP}}(s) &= \frac{s-t_{j-1}}{t_j-t_{j-1}} t_j^{\psi-1} G_2(s, S_{PREP}^j(s)) \\
 &\quad - \frac{s-t_{j-1}}{t_j-t_{j-1}} t_{j-1}^{\psi-1} G_2(s, S_{PREP}^{j-1}(s)), \\
 Z_j^{E_H}(s) &= \frac{s-t_{j-1}}{t_j-t_{j-1}} t_j^{\psi-1} G_3(s, E_H^j(s)) - \frac{s-t_{j-1}}{t_j-t_{j-1}} t_{j-1}^{\psi-1} G_3(s, E_H^{j-1}(s)), \\
 Z_j^{E_{PEP}}(s) &= \frac{s-t_{j-1}}{t_j-t_{j-1}} t_j^{\psi-1} G_4(s, E_{PEP}^j(s)) - \frac{s-t_{j-1}}{t_j-t_{j-1}} t_{j-1}^{\psi-1} G_4(s, E_{PEP}^{j-1}(s)), \\
 Z_j^{I_H}(s) &= \frac{s-t_{j-1}}{t_j-t_{j-1}} t_j^{\psi-1} G_5(s, I_H^j(s)) - \frac{s-t_{j-1}}{t_j-t_{j-1}} t_{j-1}^{\psi-1} G_5(s, I_H^{j-1}(s)), \\
 Z_j^A(s) &= \frac{s-t_{j-1}}{t_j-t_{j-1}} t_j^{\psi-1} G_6(s, A^j(s)) - \frac{s-t_{j-1}}{t_j-t_{j-1}} t_{j-1}^{\psi-1} G_6(s, A^{j-1}(s)), \\
 Z_j^{T_H}(s) &= \frac{s-t_{j-1}}{t_j-t_{j-1}} t_j^{\psi-1} G_7(s, T_H^j(s)) - \frac{s-t_{j-1}}{t_j-t_{j-1}} t_{j-1}^{\psi-1} G_7(s, T_H^{j-1}(s)).
 \end{aligned} \tag{41}$$

We finally solve the schemes needed for numerical calculations of the HIV/AIDS model (1):

$$\begin{aligned}
 S_H(t_{n+1}) &= S_{H0} + \psi(1-\tau)t_n^{\psi-1}G_1(t_n, S_H(t_n)) \\
 &\quad + \frac{\psi(\Delta t)^\tau}{AB(\tau)\Gamma(\tau+2)} \sum_{j=1}^n \left[t_j^{\psi-1} G_1(s, S_H^j(s)) X_1(n, j) \right. \\
 &\quad \left. - t_{j-1}^{\psi-1} G_1(s, S_H^{j-1}(s)) X_2(n, j) \right], \\
 S_{PREP}(t_{n+1}) &= S_{PREP0} + \psi(1-\tau)t_n^{\psi-1}G_2(t_n, S_{PREP}(t_n))
 \end{aligned} \tag{42}$$

$$+ \frac{\psi(\Delta t)^\tau}{AB(\tau)\Gamma(\tau+2)} \sum_{j=1}^n \left[t_j^{\psi-1} G_2(s, S_{PREP}^j(s)) X_1(n, j) - t_{j-1}^{\psi-1} G_2(s, S_{PREP}^{j-1}(s)) X_2(n, j) \right], \quad (43)$$

$$E_H(t_{n+1}) = E_{H0} + \psi(1-\tau)t_n^{\psi-1}G_3(t_n, E_H(t_n)) + \frac{\psi(\Delta t)^\tau}{AB(\tau)\Gamma(\tau+2)} \sum_{j=1}^n \left[t_j^{\psi-1} G_3(s, E_H^j(s)) X_1(n, j) - t_{j-1}^{\psi-1} G_3(s, E_H^{j-1}(s)) X_2(n, j) \right], \quad (44)$$

$$E_{PEP}(t_{n+1}) = E_{PEP0} + \psi(1-\tau)t_n^{\psi-1}G_4(t_n, E_{PEP}(t_n)) + \frac{\psi(\Delta t)^\tau}{AB(\tau)\Gamma(\tau+2)} \sum_{j=1}^n \left[t_j^{\psi-1} G_4(s, E_{PEP}^j(s)) X_1(n, j) - t_{j-1}^{\psi-1} G_4(s, E_{PEP}^{j-1}(s)) X_2(n, j) \right], \quad (45)$$

$$I_H(t_{n+1}) = I_{H0} + \psi(1-\tau)t_n^{\psi-1}G_5(t_n, I_H(t_n)) + \frac{\psi(\Delta t)^\tau}{AB(\tau)\Gamma(\tau+2)} \sum_{j=1}^n \left[t_j^{\psi-1} G_5(s, I_H^j(s)) X_1(n, j) - t_{j-1}^{\psi-1} G_5(s, I_H^{j-1}(s)) X_2(n, j) \right], \quad (46)$$

$$A(t_{n+1}) = A_0 + \psi(1-\tau)t_n^{\psi-1}G_6(t_n, A(t_n)) + \frac{\psi(\Delta t)^\tau}{AB(\tau)\Gamma(\tau+2)} \sum_{j=1}^n \left[t_j^{\psi-1} G_6(s, A^j(s)) X_1(n, j) - t_{j-1}^{\psi-1} G_6(s, A^{j-1}(s)) X_2(n, j) \right], \quad (47)$$

$$T_H(t_{n+1}) = T_{H0} + \psi(1-\tau)t_n^{\psi-1}G_7(t_n, T_H(t_n)) + \frac{\psi(\Delta t)^\tau}{AB(\tau)\Gamma(\tau+2)} \sum_{j=1}^n \left[t_j^{\psi-1} G_7(s, T_H^j(s)) X_1(n, j) - t_{j-1}^{\psi-1} G_7(s, T_H^{j-1}(s)) X_2(n, j) \right], \quad (48)$$

$$- t_{j-1}^{\psi-1} G_7(s, T_H^{j-1}(s)) X_2(n, j) \Big], \quad (49)$$

where

$$\begin{aligned} X_1(n, j) &= (n+1-j)^{\tau+1} - (n-j)^\tau - (n-j)^\tau(n-j+2+2\tau), \\ X_2(n, j) &= (n+1-j)^{\tau+1} - (n-j)^\tau - (n-j)^\tau(n-j+1+\tau). \end{aligned} \quad (50)$$

The biological implication of this fractal-fractional numerical scheme is that it enables accurate, long-term simulation of HIV dynamics in populations where past exposure, treatment history, and behavioral memory significantly influence future infection risk—effects that standard models often ignore. By incorporating both fractional-order memory (capturing long-term immune or behavioral effects) and fractal dimensions (modeling heterogeneous contact patterns),

this method produces realistic epidemic curves that align with observed data in high-burden settings. For public health, this means trustworthy projections of new infections, treatment demand, and intervention impact over decades. Policymakers can now simulate “what-if” scenarios—such as scaling PrEP to 70% coverage or achieving 95% viral suppression—and obtain credible forecasts of whether elimination is feasible by 2030. This tool transforms abstract theory into actionable intelligence, guiding resource allocation, program design, and global HIV strategy with unprecedented precision.

4.4 Uncertainty and sensitivity analysis of the HIV/AIDS model

Variations are predicted because the model includes several parameters, and calculating their values for numerical simulations involves inherent errors. We used Latin Hypercube Sampling (LHS) to sample values for the parameters that affect the infected populations and the reproduction number, following the methodology in [16]. To evaluate the connection between parameter values and the sensitivity analysis final results, we calculated the Partial Rank Correlation Coefficient (PRCC). For every iteration of LHS, 1,000 simulations were run. The PRCC analysis determined the most critical parameters by using the basic reproduction number (R_{0H}) as the response function. These were the natural death rate (θ_H), the progression rate from HIV infection to the AIDS stage (ρ_1), the contact rate (γ_H), and the rate at which AIDS patients transition to treatment (θ_2).

As seen in Tables 2–3 and Figures 2–3, the sensitivity analysis reveals that the natural death rate (θ_H) has the strongest negative correlation (PRCC = -0.167, $p < 0.05$) with the reproduction number, indicating that higher natural mortality rates significantly decrease disease transmission potential. This parameter is the only one statistically significant at the $p < 0.05$ level. The progression rate from HIV infected individuals to the AIDS class (ρ_1) shows a positive correlation (PRCC = +0.136, $p = 0.054$), suggesting that a higher progression rate from the HIV-infected class to the AIDS infected class increases transmission potential, although this effect is marginally significant. Similarly, the contact rate (γ_H) demonstrates a positive correlation (PRCC = +0.133, $p = 0.061$), indicating that increased contact rates contribute to higher transmission. The progression rate from AIDS to treatment (θ_2) shows a negative correlation (PRCC = -0.128, $p = 0.072$), suggesting that faster initiation of treatment for AIDS patients may help reduce transmission. The disease-induced death rate for infected individuals (δ_I) shows a negative correlation (PRCC = -0.101, $p = 0.156$), while the proportion of exposed individuals

taking PEP (θ) also demonstrates a negative correlation (PRCC = -0.095, $p = 0.181$), suggesting that increased PEP uptake will contribute to reduced transmission. The parameter (γ) which accounts for the number of sexual parameters by an individual shows a positive correlation (PRCC = +0.089, $p = 0.211$) which suggest that an increase in the number of sexual partners by an individual will significantly increase the probability of contraction of HIV/AIDS as well as increasing the spread of the disease in the human population, while the compliance rate for condom usage (ς) shows a negative correlation (PRCC = -0.087, $p = 0.221$) suggesting that an increase in the rate of compliance to this parameter will substantially reduce the transmission of HIV/AIDS in the human population. The progression rate of HIV-infected individuals into antiretroviral treatment (θ_1) demonstrates a negative correlation (PRCC = -0.083, $p = 0.243$), and condom efficacy (Ψ) shows a negative correlation (PRCC = -0.082, $p = 0.251$), suggesting that harnessing these measures will reduce the burden of the disease in the population.

Parameters with lower PRCC values include the efficacy of PEP (ϵ_2), modification parameters for transmission in the treatment class (η), the rate at which treated individuals return to the HIV-infected class (χ), the failure rate of PEP (σ_2), proportion of individuals utilizing PrEP (Θ), efficacy of PrEP (ϵ_1), waning rate of protection (ω_H), disease-induced death rate for treated individuals (δ_T), recruitment rate into the susceptible class (Λ), progression rate of exposed individuals (σ_1), and disease-induced death rate for AIDS cases (δ_A). Based on these sensitivity analysis results, public health interventions should focus on strategies that influence mortality and demographic factors (θ_H), control progression rate from HIV infected individuals to the AIDS class (ρ_1), reduce contact rates (γ_H), and accelerate the initiation of treatment among AIDS patients (θ_2). While not statistically significant at conventional levels, interventions targeting the disease-induced death rate (δ_I) and PEP uptake (θ) may also contribute to controlling HIV/AIDS transmission in the population.

4.5 HIV/AIDS data fitting and forecasting analysis for South Africa

The HIV/AIDS data fitting and forecasting analysis for South Africa draws on historical statistics from 2010 to 2023 as presented in Table 4, sourced from UNAIDS global AIDS reports [46]. The dataset includes three key indicators: People living with HIV (PLHIV), new HIV infections, and ART coverage, measured as the percentage of PLHIV receiving treatment. These indicators

Table 2: PRCC Values (sorted by influence)

Rank	Parameter	PRCC	p-value
1	θ_H	-0.167302	0.017890
2	ρ_1	+0.136342	0.054219
3	γ_H	+0.132929	0.060593
4	θ_2	-0.127707	0.071528
5	δ_I	-0.100702	0.155952
6	θ	-0.095054	0.180612
7	γ	+0.088736	0.211477
8	ς	-0.086949	0.220853
9	θ_1	-0.082919	0.243083
10	Ψ	-0.081597	0.250704
11	ϵ_2	-0.079809	0.261274
12	η	-0.070033	0.324419
13	χ	-0.059325	0.404026
14	σ_2	+0.049524	0.486172
15	Θ	-0.048354	0.496546
16	ϵ_1	-0.032523	0.647536
17	ω_H	-0.025780	0.717089
18	δ_T	-0.024852	0.726851
19	Λ	+0.020503	0.773220
20	σ_1	+0.017072	0.810373
21	δ_A	-0.007588	0.915075

Table 3: PRCC Values for Top 10 Parameters.

Rank	Parameter	PRCC	p-value
1	θ_H	+0.167302	0.017890
2	ρ_1	+0.136342	0.054219
3	γ_H	+0.132929	0.060593
4	θ_2	-0.127707	0.071528
5	δ_I	-0.100702	0.155952
6	θ	-0.095054	0.180612
7	γ	+0.088736	0.211477
8	ς	-0.086949	0.220853
9	θ_1	-0.082919	0.243083
10	Ψ	-0.081597	0.250704

formed the empirical basis for a mathematical modeling approach aimed at estimating parameters and forecasting the future trajectory of the HIV epidemic in the country. The model applied is a compartmental structure capturing the dynamics of HIV transmission and treatment. It incorporates natural population growth and mortality (represented by birth and death rates, Λ and θ_H), transmission dynamics (via transmission coefficient γ_H), progression from HIV to AIDS, ART uptake and treatment efficacy, and different mortality outcomes depending on treatment

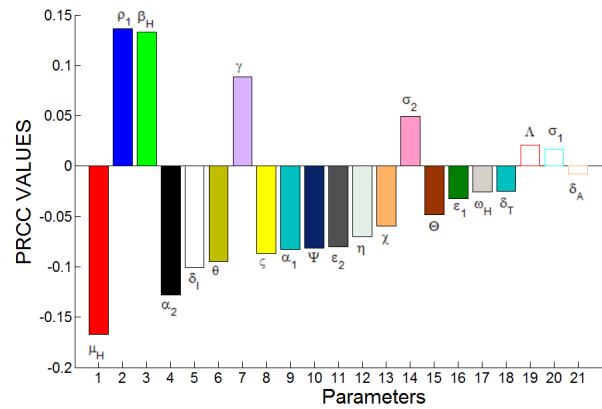


Figure 2: An extended PRCC analysis showing the correlation of all 21 model parameters with the HIV/AIDS basic reproduction number (R_{0H}). The visualization reveals the complex interplay of various factors affecting HIV transmission, including mortality rates, progression rates, and preventive interventions.

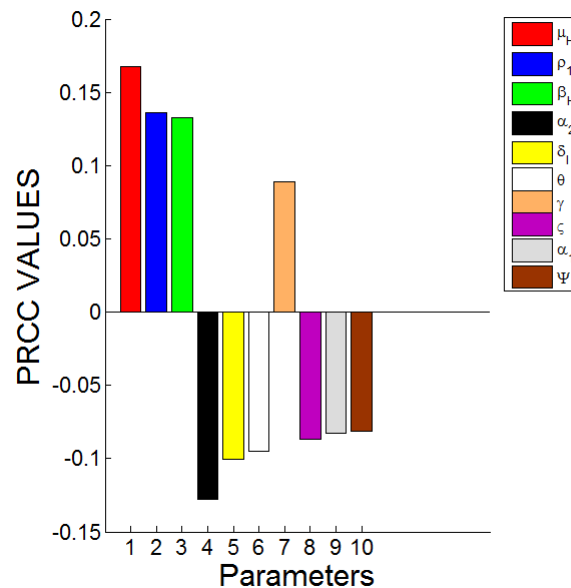


Figure 3: A bar graph displaying the sensitivity of the basic reproduction number (R_{0H}) to the top 10 model parameters. The graph highlights the most influential parameters, with notable positive and negative correlations, demonstrating the relative impact of different epidemiological factors on HIV transmission dynamics.

status. Parameter estimation was performed using MATLAB's `fmincon` optimization routine, which allows for constrained nonlinear optimization - ensuring that the model fit respects biological plausibility while accurately reproducing observed data trends.

The optimization process yielded 21 estimated parameters, with notable outcomes including a transmission coefficient of $\gamma_H = 0.0655$, along with multiple parameters describing treatment initiation and mortality differentials. The model aligns closely with observed data, as demonstrated in the graphical output (4a), where the model's red line accurately traces the trend in new HIV infections from 2018 to 2023. With this validated model, a forecast was generated covering the period from 2024 to 2033. The projections indicate a steady decline in new HIV infections, decreasing from approximately 200,213 in 2024 to 143,880 in 2033, marking a 28% reduction. Simultaneously, the number of individuals receiving ART is projected to grow, albeit at a slower rate, from 4.7 million in 2024 to 4.35 million in 2033. The number of untreated individuals with HIV or AIDS is expected to decline from 115,787 in 2024 to 101,785 in 2030, after which it stabilizes near 100,000. These trends suggest a substantial shift in the epidemic profile, with a growing dominance of the treated population, as illustrated by the yellow area in Figure 4b. Epidemiologically, the analysis underscores the success of South Africa's ART expansion, which has converted HIV into a manageable chronic condition for a majority of those infected. The epidemic has entered a mature control phase marked by high treatment coverage and falling incidence. Nevertheless, the persistence of approximately 100,000 untreated individuals highlights ongoing challenges in diagnosis, linkage to care, and adherence. From a public health perspective, the forecast aligns with global epidemic control targets but also points to the need for renewed efforts to reach marginalized populations.

Table 4: HIV Statistics for South Africa (2010-2023) [46].

Year	People living with HIV	New HIV infections	ART coverage (%)
2010	6,100,000	380,000	16
2011	6,300,000	370,000	19
2012	6,500,000	360,000	22
2013	6,700,000	350,000	29
2014	6,800,000	340,000	35
2015	7,000,000	330,000	45
2016	7,100,000	310,000	53
2017	7,200,000	300,000	61
2018	7,400,000	280,000	68
2019	7,500,000	240,000	71
2020	7,600,000	230,000	72
2021	7,700,000	210,000	75
2022	7,800,000	200,000	77
2023	7,900,000	190,000	79

Table 5: South Africa HIV/AIDS Indicators and Projection: New Infections, Untreated, and Treated PLHIV (2010–2033).

Year	New Infections	Untreated	Treated
2022	218272	128353	4843674
2023	208933	124745	4789972
2024	200213	115787	4744012
2025	191680	107609	4698742
2026	184949	105355	4658007
2027	177023	104067	4606564
2028	170070	103155	4558881
2029	165040	102502	4522845
2030	159648	101785	4482653
2031	153981	101003	4438531
2032	148124	100156	4390692
2033	143880	99518	4354472

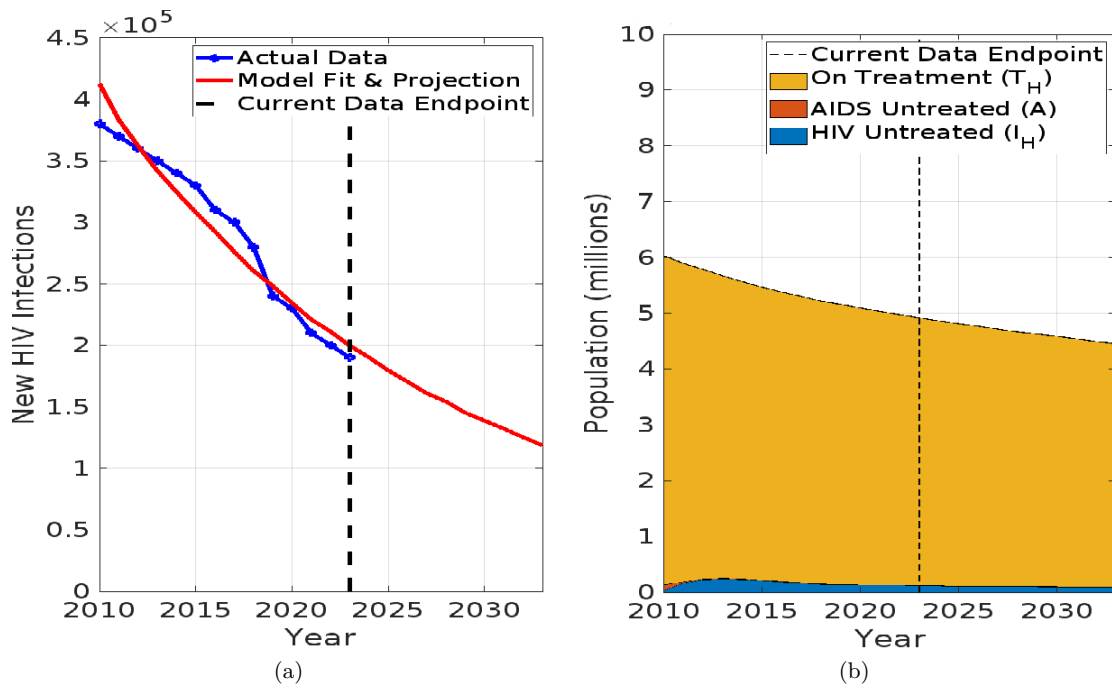


Figure 4: A time series plot (4a) comparing actual and projected new HIV infections in South Africa from 2010 to 2030. The graph illustrates the historical trend and the model's predictive projection, showing a consistent decline in new infections over time, and (4b) shows a stacked area chart depicting the changing population of HIV/AIDS patients in different states (on treatment, untreated HIV, and untreated AIDS) from 2010 to 2030. The visualization highlights shifting demographics of HIV/AIDS management in South Africa.

Table 6: Fitted parameter values for the HIV/AIDS model.

Parameter	Value	Range (per year)	Reference
Λ	0.0070	[0.0056, 0.0084]	Fitted
γ_H	0.0655	[0.0524, 0.0786]	Fitted
γ	1.4069	[1.1255, 1.6883]	Fitted
ς	0.9961	[0.7969, 1.1953]	Fitted
Ψ	0.1514	[0.1211, 0.1817]	Fitted
θ_H	0.0016	[0.0013, 0.0019]	Fitted
η	0.9959	[0.7967, 1.1951]	Fitted
ω_H	0.0413	[0.0330, 0.0496]	Fitted
ε_1	0.8463	[0.6770, 1.0156]	Fitted
σ_1	0.0119	[0.0095, 0.0143]	Fitted
θ	0.9573	[0.7658, 1.1488]	Fitted
σ_2	0.0834	[0.0667, 0.1001]	Fitted
ε_2	0.9370	[0.7496, 1.1244]	Fitted
θ_1	1.3025	[1.0420, 1.5630]	Fitted
ρ_1	0.0968	[0.0774, 0.1162]	Fitted
θ_2	0.1867	[0.1494, 0.2240]	Fitted
δ_I	0.3357	[0.2686, 0.4028]	Fitted
δ_A	1.9944	[1.5955, 2.3933]	Fitted
δ_T	0.0324	[0.0259, 0.0389]	Fitted
δ_V	0.0008	[0.0006, 0.0010]	Fitted
χ	0.0349	[0.0279, 0.0419]	Fitted
Θ	0.0022	[0.0018, 0.0026]	Fitted

4.6 Numerical simulations of HIV/AIDS model

Utilizing the parameter values obtained from fitting the HIV/AIDS model (1) to the South African HIV/AIDS data (Tables 4–5), as outlined in Table 6, we present the graphs of the numerical simulations of the model.

Figure 5a presents a two-dimensional plot with time (in days) on the x-axis and the number of susceptible individuals (S_H) on the y-axis, ranging from approximately 1.93 to 2.00 (in units of 10^6). The graph features multiple overlapping curves, each representing a different fractional order τ (from 0.55 to 0.95), with colors ranging from purple ($\tau = 0.55$) to yellow ($\tau = 0.95$). Initially, all curves start at a high value and exhibit a sharp decline, followed by oscillatory behavior that stabilizes over time, with lower τ values showing a slower decline and higher steady-state values compared to higher τ values. Figure 5a reflects the dynamics of the susceptible population under varying fractional orders τ , which model memory effects and nonlocal interactions in HIV transmission. A lower τ indicates a slower depletion of susceptible individuals due to prolonged memory effects, suggesting that interventions like awareness or behavioral changes have a delayed but sustained impact. As τ approaches 1, the rapid decline aligns with classi-

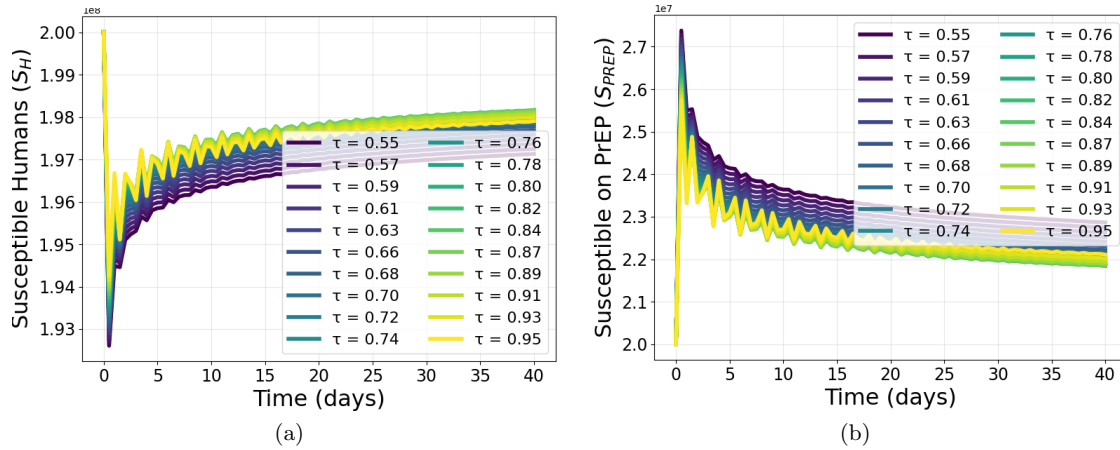


Figure 5: Impact of fractional order τ on susceptible populations.

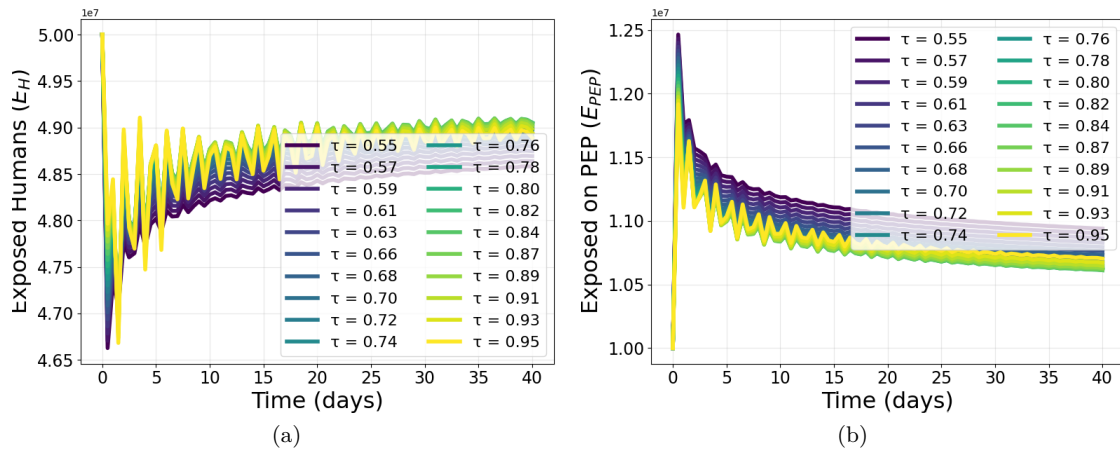


Figure 6: Fractional order effects on exposed populations.

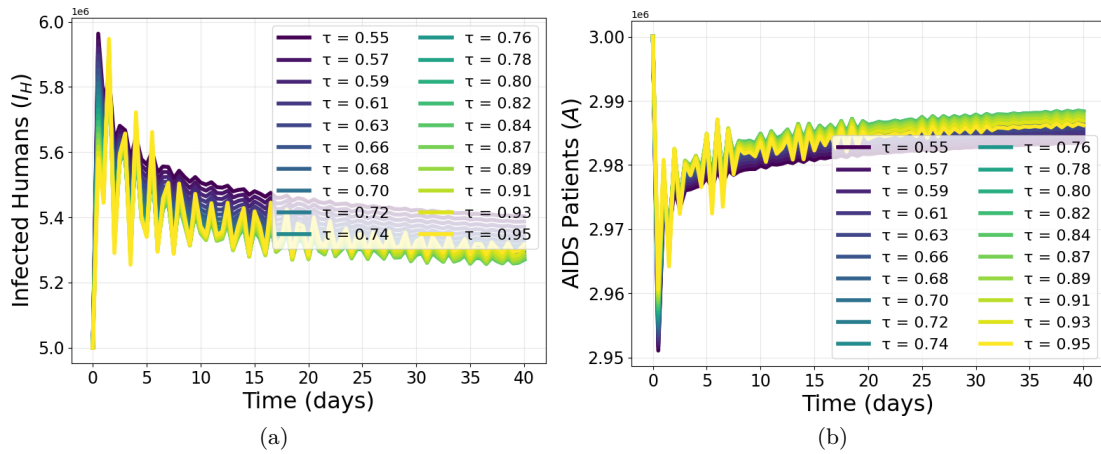


Figure 7: Disease progression dynamics under fractional order variations.

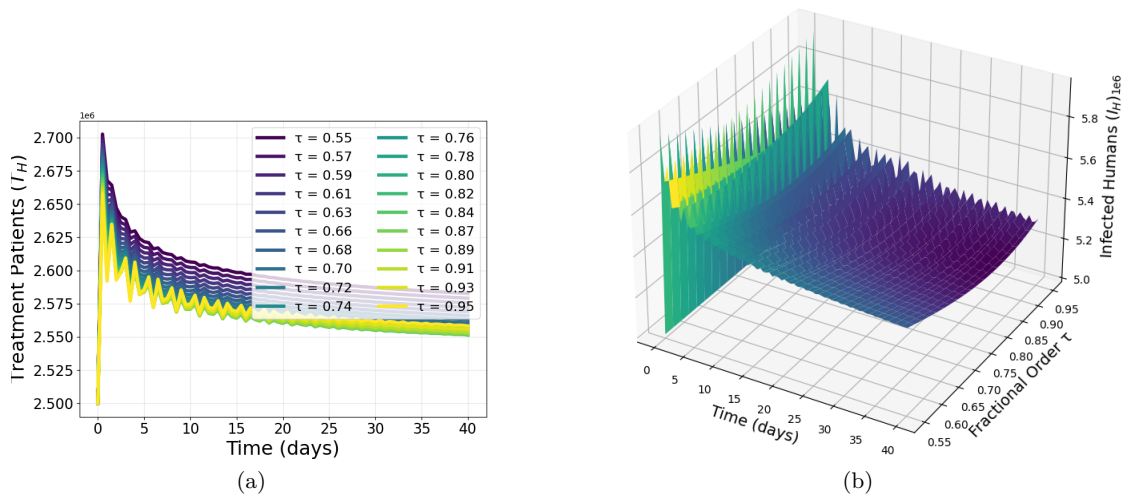


Figure 8: Treatment dynamics and three-dimensional infection analysis.

cal integer-order models, indicating a faster transition to exposure due to immediate infectious contacts. Figure 5b displays a two-dimensional plot with time (in days) on the x-axis and the number of susceptible individuals on PrEP (S_{PrEP}) on the y-axis, ranging from approximately 2.0 to 2.7 (in units of 10^7). Similar to Figure 5a, it features multiple curves for different τ values (0.55 to 0.95), with colors transitioning from purple to yellow. The curves show an initial peak followed by a gradual decline with oscillations, where lower τ values result in higher and more sustained S_{PrEP} levels over time. Figure 5b illustrates the impact of PrEP effectiveness across different fractional orders. Lower τ values suggest a prolonged protective effect of PrEP, reflect-

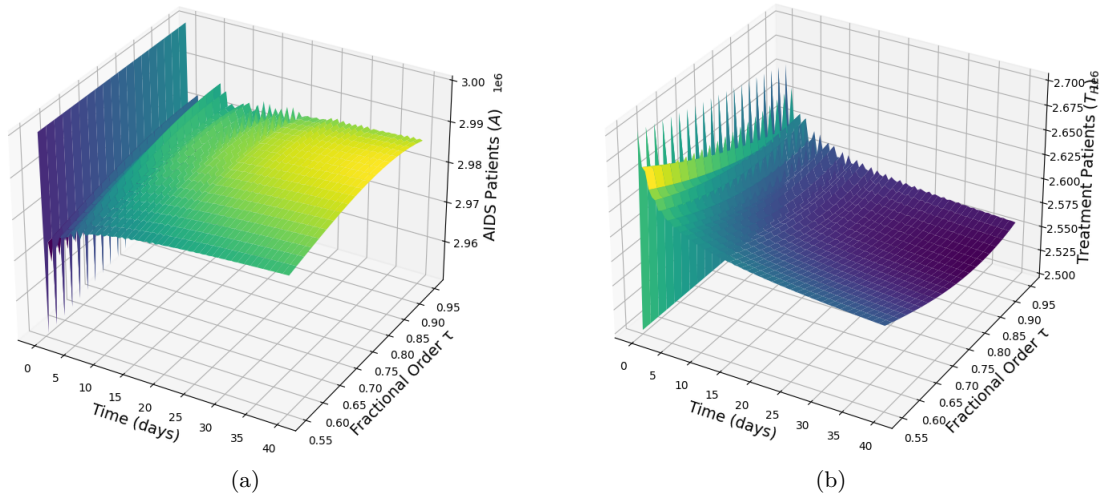


Figure 9: Three-dimensional surface plots of advanced disease stages and treatment.

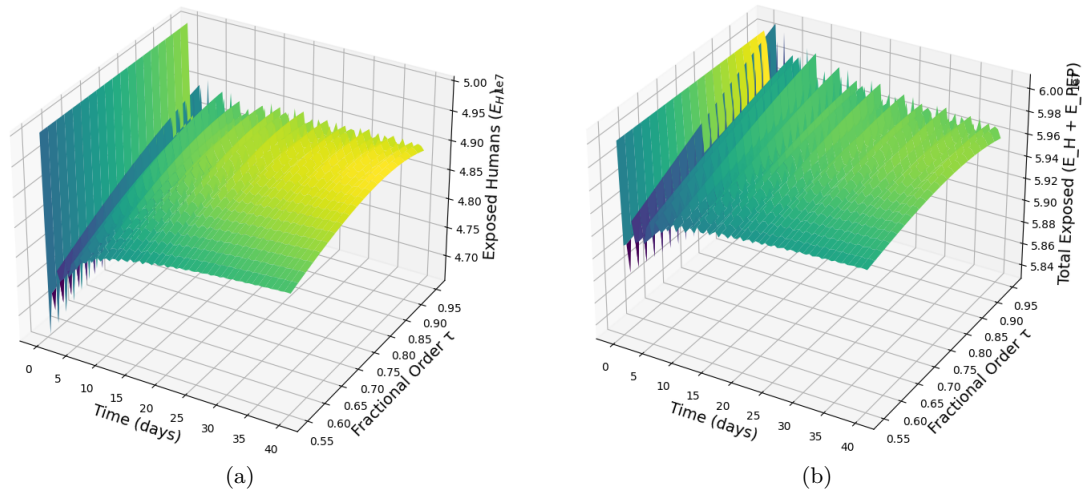


Figure 10: Combined exposure dynamics in three-dimensional analysis.

ing slower transitions to exposure due to memory in prophylactic adherence. This indicates that populations with stronger historical adherence to PrEP maintain higher susceptible numbers longer, highlighting the importance of consistent PrEP use in reducing HIV incidence. Figure 6a presents a two-dimensional plot with time (in days) on the x-axis and the number of exposed individuals (E_H) on the y-axis, ranging from approximately 4.65 to 5.00 (in units of 10^7). It includes multiple curves for τ values from 0.55 to 0.95, with colors from purple to yellow. The curves start high, decline sharply, and then stabilize with oscillations, where lower τ values lead

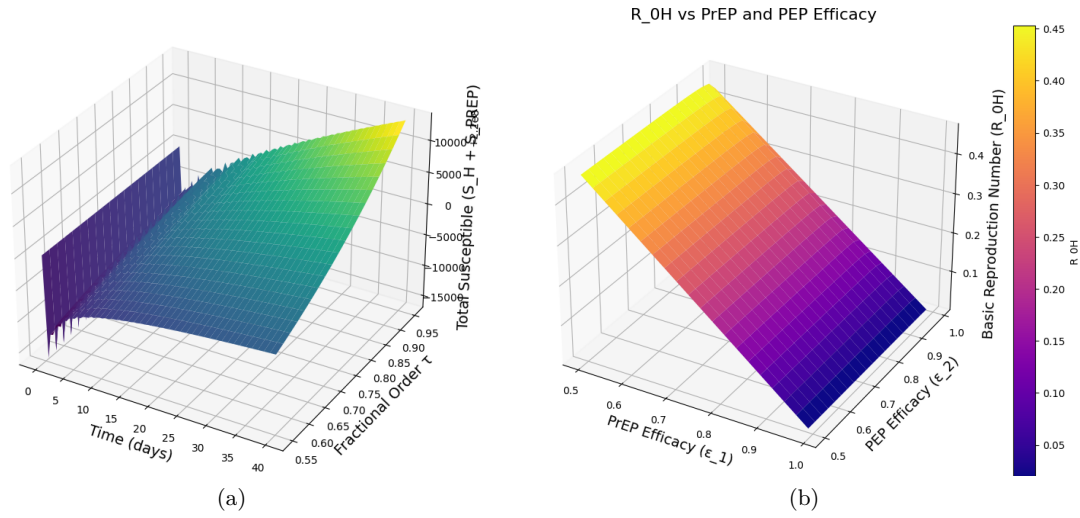


Figure 11: Susceptible population dynamics and prophylaxis efficacy analysis.

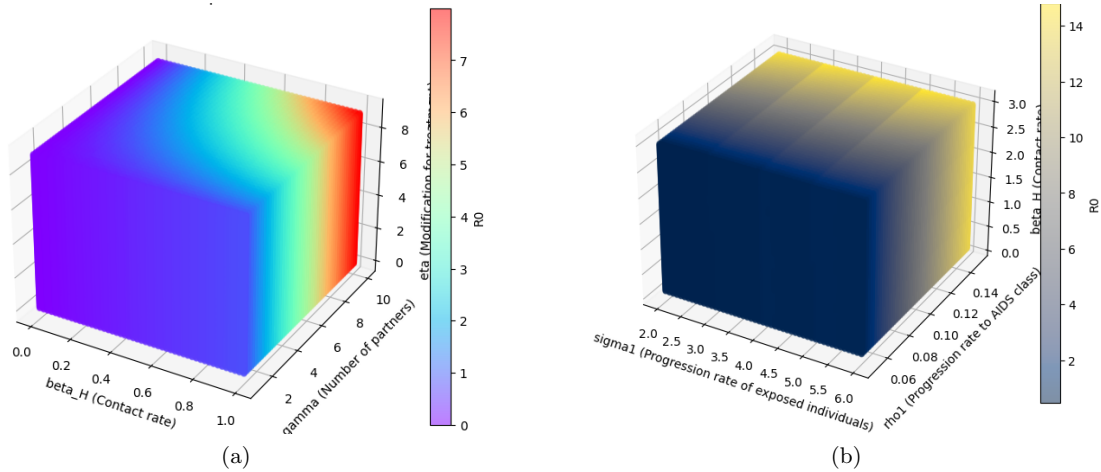


Figure 12: Parameter sensitivity analysis for transmission dynamics.

to higher steady-state exposed populations compared to higher τ values. Figure 6a shows how the exposed population evolves with fractional order τ , capturing the latency period before infection. Lower τ values indicate a slower progression from exposure to infection, possibly due to memory effects in immune response or delayed contact tracing. This suggests that interventions targeting early exposure stages may be more effective in populations with lower τ , where the exposed pool persists longer.

Figure 6b displays a two-dimensional plot with time (in days) on the x-axis and the number

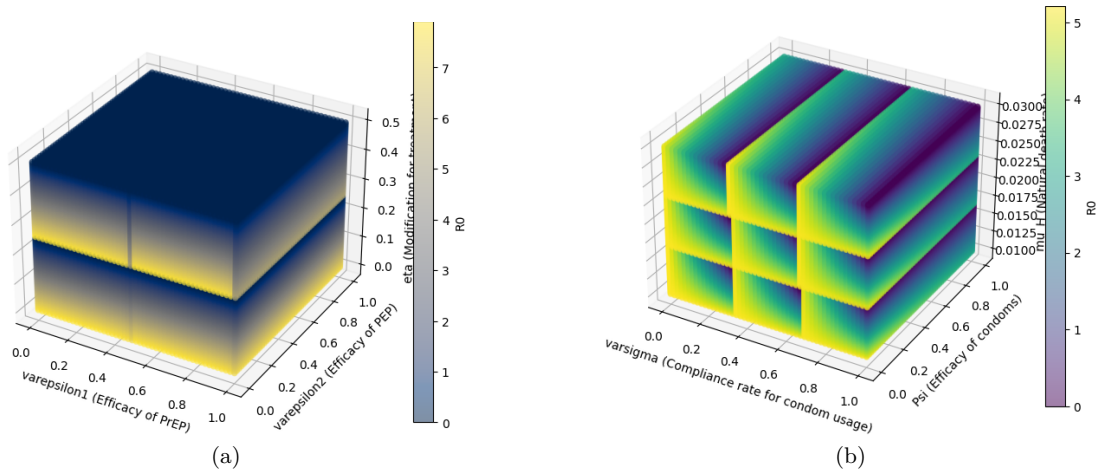


Figure 13: Intervention efficacy and behavioral parameter analysis.

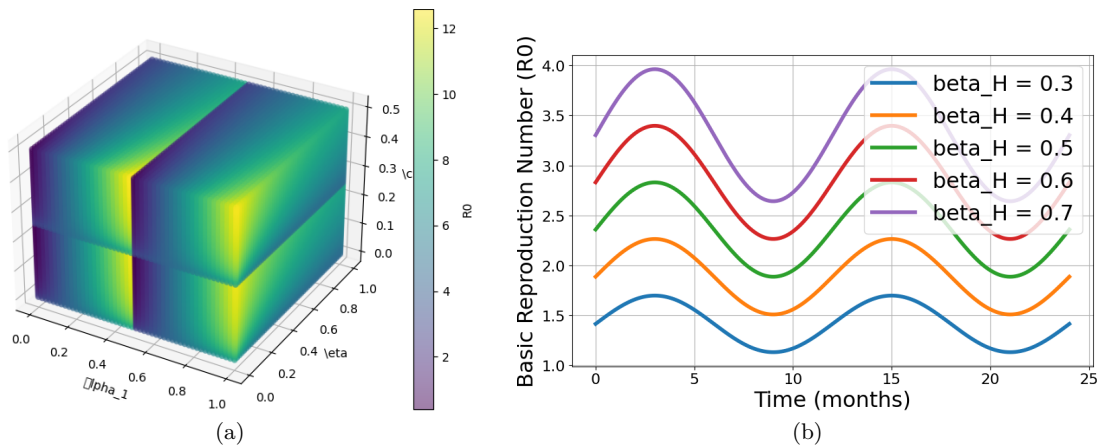


Figure 14: Treatment dynamics and temporal reproduction number analysis.

of exposed individuals on PEP (E_{PEP}) on the y-axis, ranging from approximately 1.00 to 1.25 (in units of 10^7). It features curves for τ values from 0.55 to 0.95, with colors from purple to yellow. The curves show an initial peak followed by a steady decline with minor oscillations, with lower τ values maintaining higher E_{PEP} levels over time. Figure 6b represents the dynamics of individuals using PEP during the exposed phase. Lower τ values suggest a prolonged presence of exposed individuals on PEP, reflecting memory effects in treatment adherence or efficacy. This indicates that PEP's protective role is more sustained in populations with lower τ , emphasizing the need for timely and consistent PEP administration to prevent progression to infection. Figure 7a presents a two-dimensional plot with time (in days) on the x-axis and the number of

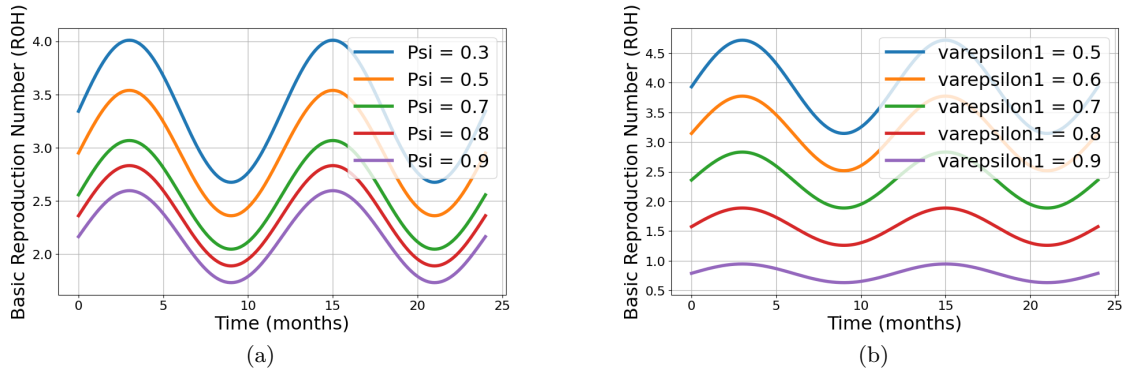


Figure 15: Temporal analysis of prevention intervention efficacy.

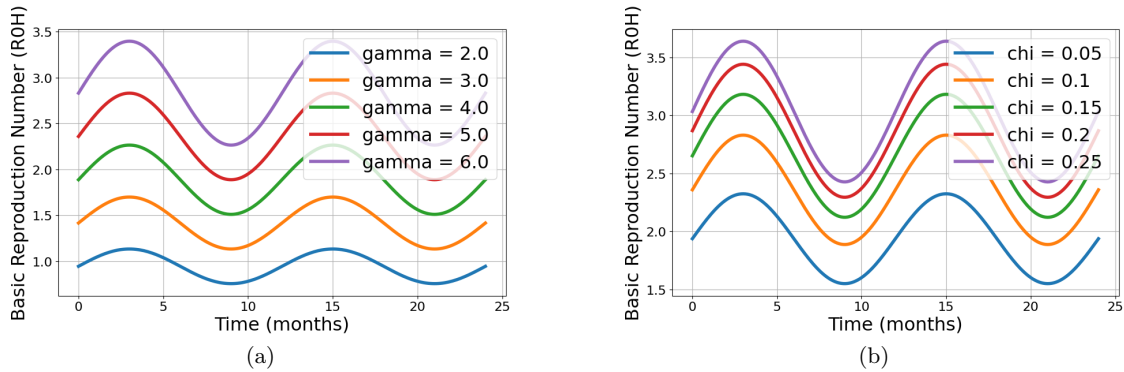


Figure 16: Sexual behavior and treatment reversion effects on transmission dynamics.

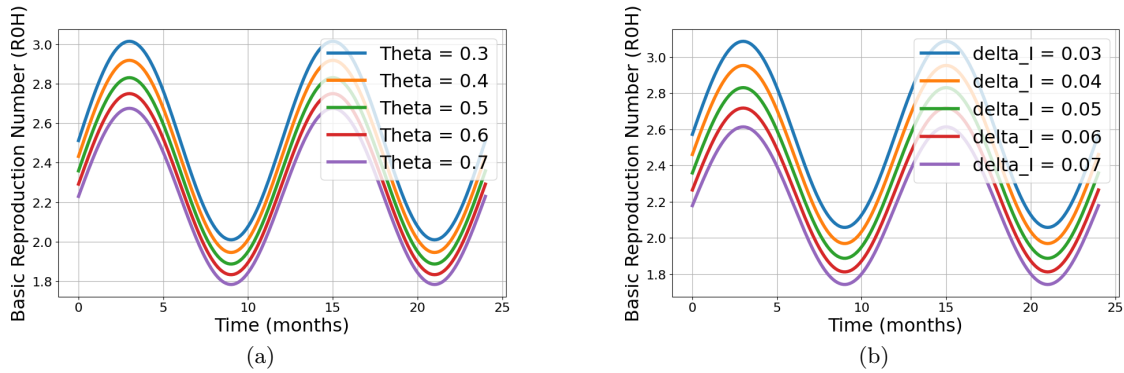


Figure 17: Mortality effects on reproduction number dynamics.

infected individuals (I_H) on the y-axis, ranging from approximately 5.0 to 6.0 (in units of 10^6). It

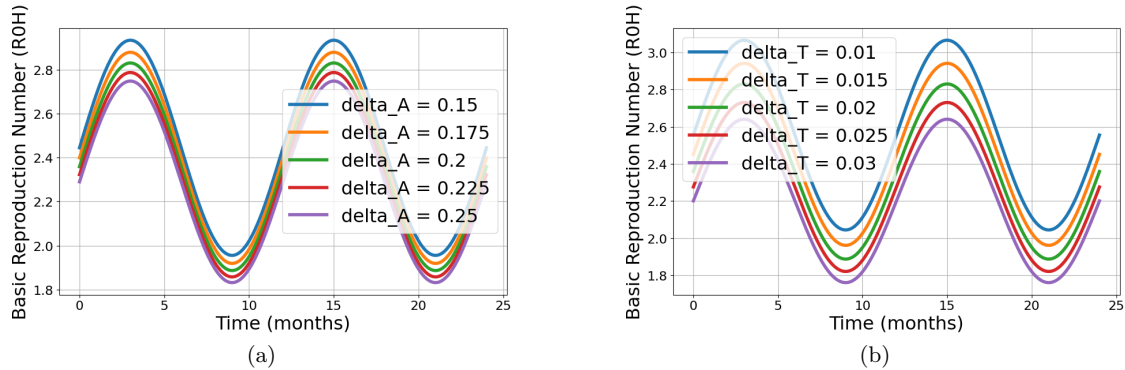


Figure 18: Stage-specific mortality impacts on transmission control.

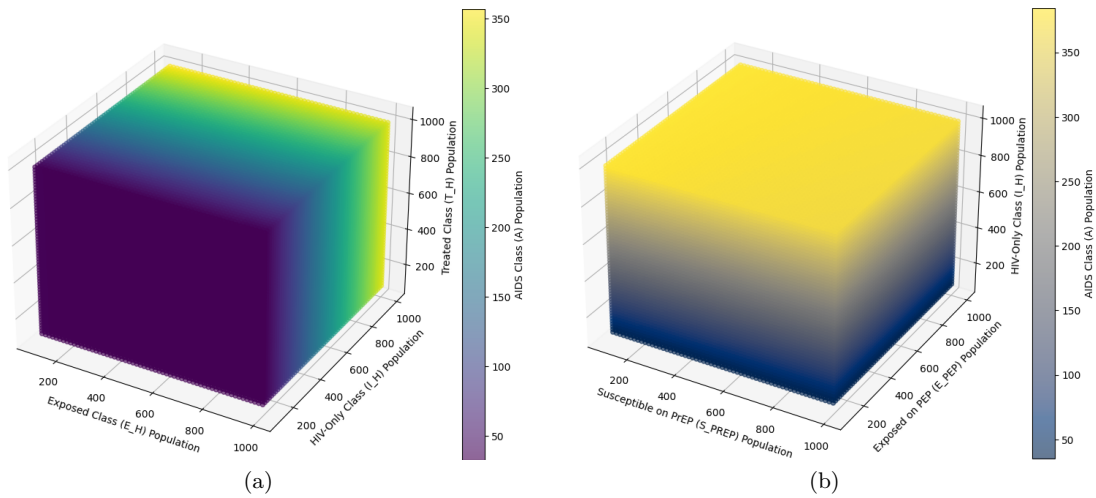


Figure 19: Multi-compartment interaction analysis and prophylaxis strategy evaluation.

includes curves for τ values from 0.55 to 0.95, with colors from purple to yellow. The curves show an initial peak followed by a decline with oscillatory stabilization, where lower τ values result in higher and more persistent infected populations. Figure 7a depicts the progression of HIV-only infected individuals, with lower τ values indicating slower disease progression due to memory effects in immune response or treatment uptake. This suggests that populations with lower τ may experience prolonged infectious periods, necessitating targeted interventions like ART to reduce transmission rates. Figure 7b displays a two-dimensional plot with time (in days) on the x-axis and the number of individuals with HIV/AIDS (A) on the y-axis, ranging from approximately 2.95 to 3.00 (in units of 10^6). It features curves for τ values from 0.55 to 0.95, with colors from purple to yellow. The curves show an initial increase followed by stabilization with oscillations,

where lower τ values lead to slightly higher AIDS cases over time. Figure 7b illustrates the progression to advanced HIV/AIDS, with lower τ values suggesting a slower transition from HIV to AIDS due to memory effects in disease progression or treatment delays. This indicates that populations with lower τ may have a prolonged pre-AIDS phase, highlighting the need for early ART to prevent progression to AIDS. Figure 8a presents a two-dimensional plot with time (in days) on the x-axis and the number of individuals on ART (T_H) on the y-axis, ranging from approximately 2.50 to 2.70 (in units of 10^6). It includes curves for τ values from 0.55 to 0.95, with colors from purple to yellow. The curves show an initial decline followed by stabilization with oscillations, where lower τ values maintain higher treatment numbers over time. Figure 8a reflects the dynamics of individuals on ART, with lower τ values indicating a sustained treatment population due to memory effects in therapy adherence or access. This suggests that consistent ART programs are crucial in populations with lower τ , where treatment effects persist longer, reducing mortality and transmission. Figure 8b displays a three-dimensional surface plot with time (in days) on the x-axis, fractional order τ on the y-axis (0.55 to 0.95), and the number of infected individuals (I_H) on the z-axis (5.0 to 5.8 in units of 10^6). The surface shows a peak at early times and lower τ , gradually flattening as time and τ increase, with a color gradient from purple to yellow indicating the density of infected cases. Biologically, Figure 8b provides a comprehensive view of how infection rates vary with time and fractional order, with lower τ values and early times showing higher infection peaks. This suggests that memory effects amplify initial infection spreads, underscoring the importance of early interventions to curb the epidemic in such scenarios.

Figure 9a presents a three-dimensional surface plot with time (in days) on the x-axis, fractional order τ on the y-axis (0.55 to 0.95), and the number of AIDS cases (A) on the z-axis (2.96 to 3.00 in units of 10^6). The surface shows a gradual rise and stabilization, with a color gradient from purple to yellow, where lower τ values maintain slightly higher AIDS levels over time. Figure 9a illustrates the temporal and fractional order dependence of AIDS progression, with lower τ values indicating a prolonged disease course. This highlights the need for sustained ART and support to manage advanced cases, especially in populations with significant memory effects. Figure 9b displays a three-dimensional surface plot with time (in days) on the x-axis, fractional order τ on the y-axis (0.55 to 0.95), and the number of individuals on treatment (T_H) on the z-axis (2.50 to 2.70 in units of 10^6). The surface shows an initial decline followed by stabilization, with a color gradient from purple to yellow, where lower τ values sustain higher treatment levels. Figure 9b reflects the interplay of time and fractional order on treatment dynamics, with lower τ values suggesting prolonged treatment efficacy due to memory effects. This indicates that consistent ART access is vital in populations with lower τ , where treatment benefits persist longer, improving survival rates.

Figure 10a is a three-dimensional surface plot with axes labeled as Time (days) ranging from 0 to 40, Fractional Order τ from 0.55 to 0.95, and Total Exposed Humans ($E_H + E_{PEP}$) scaled from approximately 4.70×10^7 to 5.00×10^7 . The surface exhibits a color gradient from purple (lower values) to yellow (higher values), showing a jagged, oscillatory elevation that increases sharply with time, particularly at lower τ , with pronounced peaks and valleys along the time axis, forming a ridged terrain that smooths slightly at higher τ . Figure 10a demonstrates the impact of fractional order τ on the dynamics of exposed individuals ($E_H + E_{PEP}$) over short-term periods, where lower τ (indicating stronger memory effects in the fractal-fractional model) leads to more volatile oscillations in exposure levels, reflecting delayed progression (σ_1, σ_2) and prophylaxis failures ($(1 - \varepsilon_1)\lambda_H, (1 - \varepsilon_2)\sigma_2$). This implies that incorporating noninteger derivatives captures realistic heterogeneities in HIV exposure, suggesting interventions like enhanced PEP (θ) could dampen surges in high-memory scenarios to prevent rapid accumulation of latently infected populations. Figure 10b presents a three-dimensional surface plot with time (days) from 0 to 40 on one axis, fractional order τ from 0.55 to 0.95 on another, and total exposed ($E_H + I_H + E_{PEP}$) – likely a summation including treated or related classes – ranging from 5.84×10^6 to 6.00×10^6 . Colored from dark blue/purple (low) to yellow (high), the surface rises with time in a jagged, oscillatory manner, with denser ridges at lower τ and a more gradual incline at higher values, exhibiting persistent fluctuations along both time and τ . Figure 10b illustrates how fractional order τ influences the combined exposed and infected compartments ($E_H + I_H + E_{PEP}$), with lower τ amplifying cyclic behaviors due to the Mittag-Leffler kernel's memory in transitions ($\sigma_1 \rightarrow I_H, \theta_1 \rightarrow T_H$), highlighting that fractal-fractional modeling reveals sustained epidemic waves in HIV progression. This signifies the biological advantage of higher τ (closer to integer-order) for stabilizing infection loads, underscoring the need for timely ART (θ_1, θ_2) to mitigate memory-driven rebounds in viral spread. Figure 11a depicts a three-dimensional surface plot with axes for Time (days) 0 to 40, Fractional Order τ 0.55 to 0.95, and Total Susceptible ($S_H + S_{PrEP}$) from -1.50×10^6 to 1.00×10^6 , though negative values may indicate modeling artifacts or scaling issues. The color transitions from purple (negative/low) to yellow (high), forming a declining surface with oscillatory ridges that deepen over time, more pronounced at lower τ , resembling a descending, wavy plane with irregular drops. Figure 11a explores the evolution of susceptible populations ($S_H + S_{PrEP}$) under fractional dynamics, where lower τ exacerbates depletions due to intensified force of infection λ_H and PrEP transitions (ω_H), potentially leading to unrealistic negatives as artifacts of numerical simulation. This biologically emphasizes that fractal memory effects accelerate susceptibility loss through exposure, advocating for boosted PrEP efficacy (ε_1) and entry rates (Λ) to sustain population resilience against cyclic HIV incursions. Figure 12a shows a three-dimensional surface plot with γ_H (Contact rate) from 0 to 1.0, γ (Number of partners) from 0 to 10, and R_0 modification factor from 0 to 8.

The surface uses a color gradient from purple (low) to red (high), rising steeply in a block-like manner with higher values at elevated γ_H and γ , forming a smooth, ascending plateau with minimal oscillations. Figure 12a highlights the synergistic effect of contact rate γ_H and partner number γ on the basic reproduction number R_{0H} modification, where increases in both amplify transmission potential via $\lambda_H = \frac{(1-\varsigma\Psi)\gamma\gamma_H(I_H+A+\eta T_H)}{N}$, leading to epidemic thresholds. This implies that behavioral interventions reducing γ or γ_H (e.g., education on safer practices) are crucial to cap R_{0H} below unity, preventing sustained HIV outbreaks in high-risk networks.

Figure 12b is a three-dimensional surface plot with σ_1 (Progression rate of exposed individuals) from 2.0 to 6.0, ρ_1 (Progression rate to AIDS) from 0.00 to 0.14, and R_0 from 0 to 3.0. Colored from blue (low) to yellow (high), the surface appears as a stepped, blocky incline with dotted or grid-like patterns, peaking at higher σ_1 and lower ρ_1 , with a relatively flat high region transitioning to lows. Biologically, Figure 12b examines how progression rates σ_1 (from E_H) and ρ_1 (from I_H to A) modulate R_{0H} , with faster σ_1 elevating reproduction by shortening latency and increasing infectious periods, while higher ρ_1 tempers it via quicker AIDS advancement and mortality (δ_A). This biologically suggests optimizing diagnostic timing to slow σ_1 and ρ_1 , thereby reducing overall transmissibility and averting progression to severe stages in HIV dynamics. Figure 13a displays a three-dimensional surface plot with ε_1 (Efficacy of PrEP) from 0 to 1.0, ε_2 (Efficacy of PEP)/ η (modification for treatment) from 0 to 0.8, and R_0 modification factor from 0 to 7. The color shifts from dark blue (low) to yellow (high), forming a layered, block-like structure with higher values at lower efficacies, exhibiting stepwise increases along the axes. Figure 13a reveals the interplay of PrEP efficacy ε_1 , PEP efficacy ε_2 , and treatment reduction η , where lower efficacies inflate R_{0H} by heightening failure rates ($(1-\varepsilon_1)\lambda_H$, $(1-\varepsilon_2)\sigma_2$) and sustaining infectivity in T_H . This indicates that enhancing prophylactic and therapeutic efficacies synergistically suppresses reproduction, biologically supporting integrated strategies to minimize exposure transitions and achieve epidemic control. Figure 13b illustrates a three-dimensional surface plot (representing four dimensions via color) with ς (Compliance rate for condom usage) from 0 to 1.0, Ψ (Efficacy of condoms) from 0 to 0.8, and θ_H (Natural mortality) from 0.0100 to 0.0275, with R_0 from 0 to 5 colored blue (low) to yellow (high). The surface forms a grid of blocky elevations, peaking at low ς/Ψ and high θ_H , with segmented layers. Figure 13b assesses how condom adherence ς , efficacy Ψ , and mortality θ_H affect R_{0H} , with poor condom metrics boosting transmission through $(1-\varsigma\Psi)$ in λ_H , while higher θ_H counterintuitively lowers R_{0H} by removing individuals faster. This biologically underscores promoting condom use to reduce effective contacts, outweighing mortality effects, for long-term HIV containment in populations. Figure 14b is a two-dimensional line plot with Basic Reproduction Number R_{0H} from 1.0 to 4.0 on the y-axis and Time (months) from 0 to 25 on the x-axis. Curves are color-coded for γ_H (Contact rate): $\gamma_H = 0.3$ (blue), 0.4 (orange), 0.5 (green), 0.6 (red), 0.7 (purple), showing sinusoidal

oscillations with a period of about 10 months, where higher γ_H yields greater amplitudes and peaks up to 3.8. Figure 14b shows the sensitivity of R_{0H} to contact rate γ_H , with increases intensifying oscillatory transmission due to elevated λ_H , signifying periodic epidemic resurgences. This highlights the biological imperative for contact-reducing measures (e.g., social distancing or education) to flatten curves and maintain R_{0H} below critical thresholds. Figure 15a presents a two-dimensional line plot with R_{0H} from 2.0 to 4.0 on the y-axis and Time (months) 0 to 25 on the x-axis. Parameter Ψ (Condom efficacy) varies: $\Psi = 0.3$ (blue), 0.5 (orange), 0.7 (green), 0.8 (red), 0.9 (purple), with oscillations of 10-month periods, lower Ψ corresponding to higher peaks and amplitudes. Figure 15a indicates that diminishing condom efficacy Ψ escalates R_{0H} fluctuations via weakened protection in $(1 - \varsigma\Psi)$, promoting cyclic infections. This biologically advocates for high-efficacy prophylactics to stabilize dynamics and curb HIV spread over time. Figure 15b depicts a two-dimensional line plot with R_{0H} from 0.5 to 4.5 on the y-axis and Time (months) 0 to 25 on the x-axis. Curves for ε_1 (PrEP effectiveness): $\varepsilon_1 = 0.5$ (blue), 0.6 (orange), 0.7 (green), 0.8 (red), 0.9 (purple), exhibiting sinusoidal waves with 10-month cycles, higher ε_1 linked to lower overall levels and damped amplitudes. Biologically, Figure 15b demonstrates that boosting PrEP effectiveness ε_1 mitigates R_{0H} oscillations by reducing failure to exposure $((1 - \varepsilon_1)\lambda_H)$, leading to smoother epidemic control. This signifies the biological value of advanced PrEP formulations in preventing periodic HIV upticks and achieving sustained suppression.

The Figure 16a depicts a two-dimensional line plot with the basic reproduction number R_{0H} on the y-axis ranging from approximately 1.0 to 3.5 and time in months on the x-axis spanning 0 to 25. Multiple colored curves represent different values of the parameter γ (number of sexual partners per individual), specifically $\gamma = 2.0$ (blue), $\gamma = 3.0$ (orange), $\gamma = 4.0$ (green), $\gamma = 5.0$ (red), and $\gamma = 6.0$ (purple). Each curve exhibits sinusoidal oscillations with a consistent period of about 10 months, starting at higher initial values for larger γ (around 3.2 for $\gamma = 6.0$) and decreasing in amplitude over time, though the oscillations persist without fully damping to equilibrium. Biologically, the Figure 16a illustrates the sensitivity of the basic reproduction number R_{0H} to variations in γ , highlighting how an increase in the average number of sexual partners amplifies the oscillatory behavior and elevates the overall transmission potential of HIV/AIDS. Higher γ values lead to greater force of infection $\lambda_H = \frac{(1-\varsigma\Psi)\gamma\gamma_H(I_H+A+\eta T_H)}{N}$, resulting in more pronounced peaks in R_{0H} , which signify periods of heightened epidemic risk; this underscores the importance of interventions reducing partner concurrency to mitigate cyclic surges in infection rates. Figure 16b shows a two-dimensional line plot with R_{0H} on the y-axis from about 1.5 to 3.5 and time in months on the x-axis from 0 to 25. Curves are color-coded for different values of χ (rate of reversion to the HIV-only group due to reduced viral load), including $\chi = 0.05$ (blue), $\chi = 0.1$ (orange), $\chi = 0.15$ (green), $\chi = 0.2$ (red), and $\chi = 0.25$ (purple). The lines display oscillatory patterns with a period of roughly 10 months, where higher χ corresponds to higher

amplitude and peak values, starting near 3.3 for $\chi = 0.25$ and showing slight damping over time. Figure 16b demonstrates that increasing the reversion rate χ from treatment T_H back to I_H exacerbates fluctuations in R_{0H} , as it allows more individuals to re-enter the infectious pool without sustained viral suppression, thereby prolonging epidemic waves; this emphasizes the need for improved treatment adherence to minimize such reversions and stabilize transmission dynamics. Figure 17a presents a two-dimensional line plot with R_{0H} scaled from 1.8 to 3.0 on the y-axis and time from 0 to 25 months on the x-axis. Parameter Θ (reduced disease-induced mortality in the treatment group) varies across curves: $\Theta = 0.3$ (blue), $\Theta = 0.4$ (orange), $\Theta = 0.5$ (green), $\Theta = 0.6$ (red), and $\Theta = 0.7$ (purple). Oscillations are evident with a period of about 10 months, and higher Θ values yield lower overall R_{0H} levels and reduced oscillation amplitudes, indicating a stabilizing effect. Biologically, Figure 17a reveals that enhancing Θ , which lowers mortality in treated individuals via $\Theta\delta_T$, dampens R_{0H} oscillations and reduces average transmission rates by prolonging survival in T_H with lower infectivity ($\eta < 1$), suggesting that therapies improving survival outcomes can help control epidemic cycles and lower long-term prevalence. Figure 17b is a two-dimensional line plot featuring R_{0H} from 1.8 to 3.0 on the y-axis and time from 0 to 25 months on the x-axis. It varies δ_I (disease-induced mortality rate for HIV-infected individuals): $\delta_I = 0.03$ (blue), $\delta_I = 0.04$ (orange), $\delta_I = 0.05$ (green), $\delta_I = 0.06$ (red), and $\delta_I = 0.07$ (purple). The curves oscillate sinusoidally with a 10-month period, where higher δ_I correlates with lower peak R_{0H} and milder fluctuations. Figure 17b indicates that elevating δ_I in the I_H compartment reduces R_{0H} by removing infectious individuals faster, thereby shortening infectious periods and attenuating epidemic oscillations; this highlights the dual-edged role of disease mortality in curbing spread, though ethically, interventions should focus on prevention rather than relying on higher death rates. Figure 18a displays a two-dimensional line plot with R_{0H} ranging from 1.8 to 2.8 on the y-axis and time from 0 to 25 months on the x-axis. Parameter δ_A (disease-induced mortality for HIV/AIDS individuals) is varied: $\delta_A = 0.15$ (blue), $\delta_A = 0.175$ (orange), $\delta_A = 0.2$ (green), $\delta_A = 0.225$ (red), and $\delta_A = 0.25$ (purple). Oscillatory behavior persists with a period of approximately 10 months, and increasing δ_A slightly lowers the average R_{0H} and oscillation intensity.

Biologically, Figure 18a shows that higher δ_A in the A compartment marginally suppresses R_{0H} fluctuations by accelerating removal of advanced-stage infectives, reducing their contribution to λ_H ; however, the effect is less pronounced than for earlier stages, implying that targeting progression rates like ρ_1 may be more effective for biological control. Figure 18b illustrates a two-dimensional line plot with R_{0H} from 1.8 to 3.0 on the y-axis and time from 0 to 25 months on the x-axis. It parameterizes δ_T (disease-induced mortality for treated individuals): $\delta_T = 0.01$ (blue), $\delta_T = 0.015$ (orange), $\delta_T = 0.02$ (green), $\delta_T = 0.025$ (red), and $\delta_T = 0.03$ (purple). The

lines exhibit 10-month periodic oscillations, with higher δ_T leading to elevated R_{0H} peaks due to shorter treatment durations. Figure 18b suggests that increasing δ_T amplifies R_{0H} oscillations by diminishing the benefits of treatment in T_H , as faster mortality limits the low-infectivity period (ηT_H); this biologically stresses the value of therapies that minimize δ_T to extend suppressive effects and stabilize epidemic trajectories. Figure 19a is a three-dimensional surface plot with axes labeled as Exposed Class (E_H) Population (0 to 1000), HIV-Only Class (I_H) Population (0 to 1000), and Treated Class (T_H/A) Population (0 to 1000), using a color gradient from purple (low, ~ 50) to yellow (high, ~ 350) representing AIDS Class Population. The surface forms a sloped, uneven block rising sharply along the Treated Class axis and more gradually along others, with higher AIDS populations in regions of low exposure and high treatment/HIV-only. Biologically, Figure 19a explores interactions among exposed (E_H), HIV-only (I_H), and treated/AIDS ($T_H + A$) compartments, showing that AIDS prevalence escalates when treatment coverage is high but exposure and untreated HIV are low, possibly due to prolonged survival leading to accumulation in later stages; this signifies the need for balanced interventions addressing early exposure (σ_1) and treatment transitions (θ_1, θ_2) to prevent AIDS buildup. Figure 19b depicts a three-dimensional surface plot with axes for Susceptible on PrEP (S_{PREP}) Population (0 to 1000), Exposed on PEP (E_{PEP}) Population (0 to 1000), and HIV-Only Class (I_H) Population (0 to 1000), colored from blue (low, ~ 50) to yellow (high, ~ 350) for AIDS Class Population. The surface appears as a steep, block-like incline, with AIDS population peaking at high S_{PREP} and low E_{PEP}/I_H , forming a nearly flat high plateau transitioning to lower values. Figure 19b (interpreting the color as a fourth dimension) highlights how prophylactic strategies interact with infection stages, where high PrEP coverage (S_{PREP}) correlates with elevated AIDS if PEP usage (E_{PEP}) and untreated HIV (I_H) are minimal, suggesting potential gaps in progression management post-prophylaxis failure ($(1 - \varepsilon_1)\lambda_H, (1 - \varepsilon_2)\sigma_2$); this biologically implies optimizing ε_1 and ε_2 to reduce downstream AIDS despite strong prevention.

4.7 Advantages of the fractal-fractional order method over classical integer-order models in HIV/AIDS modeling

The fractal-fractional order method offers significant advantages over classical integer-order differential equation models when analyzing complex, real-world HIV epidemics. First, it incorporates memory effects through the fractional derivative (order $\tau \in (0, 1]$), capturing how past infections, delayed treatment initiation, and behavioral risk history influence current transmission dynamics—critical in HIV where immune memory and cumulative exposure shape progression

and infectivity. Second, the fractal dimension (parameter ψ) models heterogeneous contact patterns in sexual or injection networks, reflecting clustered risk behaviors (e.g., in key populations) that integer-order models assume away via mean-field approximations. Third, this dual structure yields more accurate long-term predictions, particularly in simulating slow-decaying tails of incidence under partial intervention coverage—patterns observed in mature epidemics but missed by exponential decay in classical models. Fourth, the framework better reproduces empirical data from cohort studies showing non-Markovian progression and sub-exponential growth, enhancing model validation and credibility. Finally, from a public health perspective, it provides superior scenario analysis for combination prevention: It reveals how sustained PrEP, ART scale-up, and condom promotion interact over decades to push R_{0H} below unity and avoid backward bifurcation traps—insights essential for achieving and sustaining the UNAIDS 95-95-95 targets by 2030. While computationally intensive, these advantages make the fractal-fractional approach the preferred tool for strategic, evidence-based HIV elimination planning in heterogeneous, memory-dependent populations.

5 Conclusion

The fractal-fractional order model developed in this research offers a robust framework for analyzing HIV/AIDS transmission dynamics, significantly enhancing traditional integer-order models by incorporating fractal-fractional derivatives. These derivatives combine fractional calculus, which captures memory effects and long-term dependencies in disease progression, with fractal geometry, which models self-similar, heterogeneous transmission patterns across diverse population scales. This approach is particularly effective for HIV/AIDS, as it accounts for the nonlocal nature of infection dynamics and complex contact structures, providing a more accurate representation of real-world epidemic behavior. By integrating PrEP, PEP, and ART, the model demonstrates their partial efficacy in reducing the force of infection and overall disease burden. Analytical proofs confirm solution positivity, boundedness, and global asymptotic stability at the endemic equilibrium when $R_{0H} > 1$. Sensitivity analysis underscores the critical influence of natural mortality, progression rates, contact rates, and treatment initiation on transmission dynamics, while parameter fitting to South African data validates the model's predictive power, forecasting a substantial 28% decline in new infections by 2033 with sustained ART coverage. Numerical simulations further reveal that lower fractional orders intensify oscillatory behaviors in infected compartments, emphasizing the need for consistent prophylactic adherence and early treatment to prevent epidemic resurgence. Compared to integer-order models, the fractal-fractional framework better fits empirical data and offers deeper insights into multi-scale

dynamics, informing optimized resource allocation for epidemic control. Future extensions could incorporate spatial heterogeneity and emerging viral variants to enhance the model's applicability. Numerical simulations reveal that enhancing the synergistic effects of condoms, prophylaxis, and antiretroviral drugs will significantly reduce the burden of HIV/AIDS in the human population.

Recommendations for the Public

1. Sustain and expand ART access to ensure consistent treatment, particularly in high-prevalence regions like South Africa, by reducing cost, stigma, and logistical barriers.
2. Promote consistent PrEP and PEP use through community outreach and subsidized access to reduce infection rates in high-risk groups.
3. Enhance early testing and diagnosis through intensified HIV testing campaigns to prevent progression to AIDS and reduce transmission risk.
4. Reduce high-risk contact behaviors, such as multiple sexual partners, by promoting safe practices like condom use through targeted education.
5. Support long-term adherence to interventions to prevent epidemic resurgence, emphasizing consistent treatment and prophylaxis commitment.

5.1 Limitations of the Model

The fractal-fractional order approach, while powerful in capturing memory effects, behavioral heterogeneity, and long-term dynamics in HIV transmission, has several important limitations. First, the method introduces additional parameters (fractional order τ and fractal dimension ψ), which require careful calibration using longitudinal cohort or surveillance data—data that are often incomplete or unavailable in low-resource settings. Misestimation of these parameters can lead to biased projections of R_{0H} , equilibrium prevalence, or intervention impact. Second, the increased computational complexity of solving fractal-fractional systems—especially with the Adams–Bashforth scheme—demands significant processing power and time, limiting real-time forecasting during public health emergencies. Third, interpretability is reduced; while the model better reflects biological and social realism, the noninteger derivatives obscure direct causal links, making it harder for policymakers to translate results into intuitive action. Finally,

the assumption of uniform memory and fractal structure across populations may oversimplify local variations in risk behavior, treatment adherence, or viral evolution. These limitations suggest that while the fractal-fractional framework is valuable for strategic long-term planning and theoretical insight, it should be complemented with simpler integer-order models for rapid response and sensitivity analyses for parameter uncertainty in operational decision-making.

Areas for Future Research

1. Incorporate spatial heterogeneity into the HIV/AIDS model.
2. Model emerging HIV/AIDS variants.
3. Integrate stochastic effects into the model
4. Explore neural network enhancements.
5. Analyze time-dependent optimal control measures on the HIV/AIDS model.

Funding

This research received no external funding.

Conflicts of Interest

The author declares no conflicts of interest.

References

- [1] Acheneje, G. O., Omale, D., Atokolo, W. and Bolaji, B., *Modeling the transmission dynamics of the co-infection of COVID-19 and monkeypox diseases with optimal control strategies and cost-benefit analysis*, Franklin Open, 8 (2024), 100130.
- [2] Acheneje, G. O., Omale, D., Atokolo, W., Celestine, A. B., Abah, E., Dervishi, R., Shyamsunder and Bolaji, B., *Mathematical modelling on the transmission dynamics of the co-infection of COVID-19 and monkeypox with treatment as a control strategy*, Discov. Public Health, 22(1) (2025), 535.

- [3] Agarwal, P., Jleli, M. and Samet, B., *Fixed point theory in metric spaces: recent advances and applications*, Springer, Singapore, 2018.
- [4] Ahmad, S., Ullah, A., Shah, K., Salahshour, S., Ahmadian, A. and Ciano, T., *Fuzzy fractional-order model of the novel coronavirus*, Adv. Differ. Equ., 1 (2020), 1–17.
- [5] Anderson, P. L., Glidden, D. V., Liu, A., Buchbinder, S., Lama, J. R., Guanira, J. V. and Grant, R. M., *Emtricitabine–tenofovir concentrations and pre-exposure prophylaxis efficacy in men who have sex with men*, Sci. Transl. Med., 4(151) (2012), 151ra125.
- [6] Anderson, R. M. and May, R. M., *Infectious diseases of humans: dynamics and control*, Oxford University Press, Oxford, 1991.
- [7] Arafa, A. A. M., Rida, S. Z. and Khalil, M., *Fractional modeling dynamics of HIV and $CD4^+$ T-cells during primary infection*, Nonlinear Biomed. Phys., 6(1) (2012), 1–7.
- [8] Atangana, A., *Fractal–fractional differentiation and integration: connecting fractal calculus and fractional calculus to predict complex systems*, Chaos Solitons Fractals, 102 (2017), 396–406.
- [9] Atangana, A., *Fractional operators with constant and variable order with application to geo-hydrology*, Academic Press, London, 2017.
- [10] Atangana, A. and Baleanu, D., *New fractional derivatives with nonlocal and non-singular kernel: theory and application to heat transfer model*, Therm. Sci., 20(2) (2016), 763–769.
- [11] Atangana, A. and Qureshi, S., *Modeling attractors of chaotic dynamical systems with fractal–fractional operators*, Chaos Solitons Fractals, 123 (2019), 320–337.
- [12] Atokolo, W., Aja, R. O., Omale, D., Ahman, Q. O., Acheneje, G. O. and Amos, J., *Fractional mathematical model for the transmission dynamics and control of Lassa fever*, Franklin Open, 7 (2024), 100110.
- [13] Baleanu, D., Jajarmi, A., Mohammadi, H. and Rezapour, S., *A new study on the mathematical modelling of human liver with Caputo–Fabrizio fractional derivative*, Chaos Solitons Fractals, 134 (2020), 109705.
- [14] Bekker, L. G., Johnson, L., Cowan, F., Overs, C., Besada, D., Hillier, S. and Cates Jr., W., *Combination HIV prevention for female sex workers: what is the evidence?*, Lancet, 385(9962) (2015), 72–87.

- [15] Bhunu, C. P. and Mushayabasa, S., *Modelling the transmission dynamics of HIV/AIDS and hepatitis B co-infection*, HIV AIDS Rev., 11(4) (2012), 118–124.
- [16] Blower, S. M. and Dowlatabadi, H., *Sensitivity and uncertainty analysis of complex models of disease transmission: an HIV model as an example*, Int. Stat. Rev., 62(2) (1994), 229–243.
- [17] Bolaji, B., Ibrahim, A., Ani, F., Omede, B. and Acheneje, G., *A model for the control of transmission dynamics of human monkeypox disease in Sub-Saharan Africa*, J. Niger. Soc. Phys. Sci., 6 (2024), 1800.
- [18] Bryant, J., Baxter, L. and Hird, S., *Non-occupational post-exposure prophylaxis for HIV: a systematic review*, Health Technol. Assess., 17(20) (2013), 1–195.
- [19] Cardo, D. M., Culver, D. H., Ciesielski, C. A., Srivastava, P. U., Marcus, R., Abiteboul, D. and McKibben, P. S., *A case-control study of HIV seroconversion in health care workers after percutaneous exposure*, N. Engl. J. Med., 337(21) (1997), 1485–1490.
- [20] Cohen, M. S., Chen, Y. Q., McCauley, M., Gamble, T., Hosseinipour, M. C., Kumarasamy, N. and Fleming, T. R., *Prevention of HIV-1 infection with early antiretroviral therapy*, N. Engl. J. Med., 365(6) (2011), 493–505.
- [21] Cohen, M. S., Chen, Y. Q., McCauley, M., Gamble, T., Hosseinipour, M. C., Kumarasamy, N. and Godbole, S. V., *Antiretroviral therapy for the prevention of HIV-1 transmission*, N. Engl. J. Med., 375(9) (2016), 830–839.
- [22] Davarian, N., Heydari, A., Hashemi Mehne, S. H. and Heydari, A. A., *The investigation of stability and optimal control of the multigroup COVID-19 epidemic model in Iran*, J. Math. Ext., 19(1) (2025), 1–20.
- [23] Deeks, S. G., Overbaugh, J., Phillips, A. and Buchbinder, S., *HIV infection*, Nat. Rev. Dis. Primers, 1(1) (2015), 1–22.
- [24] Ding, Y., Wang, Z. and Ye, H., *Optimal control of a fractional-order HIV-immune system with memory*, IEEE Trans. Control Syst. Technol., 20(3) (2012), 763–769.
- [25] French, P. P., Latka, M., Gollub, E. L., Rogers, C., Hoover, D. R. and Stein, Z. A., *Use-effectiveness of the female versus male condom in preventing sexually transmitted disease in women*, Sex. Transm. Dis., 30(5) (2003), 433–439.

- [26] Grant, R. M. and Glidden, D. V., *HIV moments and pre-exposure prophylaxis*, Lancet, 387(10027) (2016), 1507–1508.
- [27] Grant, R. M., Lama, J. R., Anderson, P. L., McMahan, V., Liu, A. Y., Vargas, L. and Glidden, D. V., *Preexposure chemoprophylaxis for HIV prevention in men who have sex with men*, N. Engl. J. Med., 363(27) (2010), 2587–2599.
- [28] Khan, M. A., Ullah, S. and Farhan, M., *The dynamics of Zika virus with Caputo fractional derivative*, AIMS Math., 4(1) (2020), 134–154.
- [29] Kilbas, A. A., Srivastava, H. M. and Trujillo, J. J., *Theory and applications of fractional differential equations*, Elsevier, Amsterdam, 2006.
- [30] Kuhar, D. T., Henderson, D. K., Struble, K. A., Heneine, W., Thomas, V., Cheever, L. W. and Panlilio, A. L., *Updated US public health service guidelines for the management of occupational exposures to HIV and recommendations for postexposure prophylaxis*, Infect. Control Hosp. Epidemiol., 34(9) (2013), 875–892.
- [31] Kumar, S., Kumar, R., Cattani, C. and Samet, B., *Chaotic behaviour of fractional predator–prey dynamical system*, Chaos Solitons Fractals, 135 (2020), 109811.
- [32] La Salle, J. P., *The stability of dynamical systems*, SIAM, Philadelphia, 1976.
- [33] Odiba, P., Acheneje, G. O. and Bolaji, B., *A compartmental deterministic epidemiological model with nonlinear differential equations for analyzing the co-infection dynamics between COVID-19, HIV and monkeypox diseases*, Healthc. Anal., 5 (2024), 100311.
- [34] Oname, A., Han, Q., Iyaniwura, S. A., Ebenezer, A., Bragazzi, N. L., Wang, X. and Woldegerima, W. A., *Understanding the impact of HIV on mpox transmission in the MSM population: a mathematical modeling study*, Infect. Dis. Model., 9(4) (2024), 1117–1137.
- [35] Owolabi, K. M. and Atangana, A., *Mathematical analysis and computational experiments for an epidemic system with nonlocal and nonsingular derivative*, Chaos Solitons Fractals, 126 (2019), 41–49.
- [36] Owolabi, K. M. and Atangana, A., *Fractal fractional-order derivative for HIV/AIDS model with Mittag-Leffler kernel*, Alexandria Eng. J., 61 (2022), 10965–10980.
- [37] Palella Jr., F. J., Delaney, K. M., Moorman, A. C., Loveless, M. O., Fuhrer, J., Satten, G. A. and Holmberg, S. D., *Declining morbidity and mortality among patients with advanced human immunodeficiency virus infection*, N. Engl. J. Med., 338(13) (1998), 853–860.

- [38] Perelson, A. S., *Modelling viral and immune system dynamics*, Nat. Rev. Immunol., 2(1) (2002), 28–36.
- [39] Potts, M., Halperin, D. T., Kirby, D., Swidler, A., Marseille, E., Klausner, J. D. and Gayle, H., *Reassessing HIV prevention*, Science, 320(5877) (2008), 749–750.
- [40] Rodger, A. J., Cambiano, V., Bruun, T., Vernazza, P., Collins, S., Degen, O. and Lundgren, J., *Sexual activity without condoms and risk of HIV transmission in serodifferent couples when the HIV-positive partner is using suppressive antiretroviral therapy*, JAMA, 316(2) (2016), 171–181.
- [41] Shah, K., Alqudah, M. A., Jarad, F. and Abdeljawad, T., *Semi-analytical study of Pine Wilt Disease model with convex rate under Caputo–Fabrizio fractional order derivative*, Chaos Solitons Fractals, 135 (2020), 109754.
- [42] Smart, D. R., *Fixed point theorems*, Cambridge University Press, Cambridge, 1974.
- [43] Sharp, P. M. and Hahn, B. H., *Origins of HIV and the AIDS pandemic*, Cold Spring Harb. Perspect. Med., 1(1) (2011), a006841.
- [44] Toufik, M. and Atangana, A., *New numerical approximation of fractional derivative with non-local and non-singular kernel: application to chaotic models*, Eur. Phys. J. Plus, 132(10) (2017), 444.
- [45] UNAIDS, *Global HIV & AIDS statistics: fact sheet*, Joint United Nations Programme on HIV/AIDS, 2021.
- [46] UNAIDS, *2024 UNAIDS global AIDS update summary for South Africa*, Joint United Nations Programme on HIV/AIDS, 2024.
- [47] Van Den Driessche, P. and Watmough, J., *Reproduction numbers and sub-threshold endemic equilibria for compartmental models of disease transmission*, Math. Biosci., 180(1–2) (2002), 29–48.
- [48] Weller, S. and Davis, K., *Condom effectiveness in reducing heterosexual HIV transmission*, Cochrane Database Syst. Rev., 1 (2002), CD003255.
- [49] World Health Organization, *Guideline on when to start antiretroviral therapy and on pre-exposure prophylaxis for HIV*, World Health Organization, Geneva, 2015.



Finite element analysis of cross-diffusion three-species Shigesada–Kawasaki–Teramoto model

G.A. Al-Musawi, H.A. Challob^{id} and A.J. Harfash*

Abstract

In this study, a three-species Shigesada–Kawasaki–Teramoto (SKT) model was studied numerically using the finite element method (FEM). The first step is to introduce the regularized problem to the original problem. After that, some bounds of the solution to the regularized problem are found in various spaces, which is considered the basic step to start the study. The regularized problem is approximated using FEM, and then the existence of the solution to the approximate problem is proven. Also, the approximate solutions' bounds will be found, and their spaces can be determined using these bounds. After that, the convergence of the approximate solutions to the exact solutions will be studied, and these findings will be followed by an examination of how the weak form of the approximate problem converges to the continuous weak form. Thus, the existence of the solutions of the SKT model will be proven. Finally,

*Corresponding author

Received 3 October 2025; revised 18 November 2025; accepted 21 November 2025

Ghufran A. Al-Musawi

Department of Mathematics, College of Education for Pure Sciences, University of Thi-Qar, Thi-Qar, Iraq.
e-mail: ghufranalumusawi@utq.edu.iq

Huda A. Challob

Department of Mathematics, College of Sciences, University of Basrah, Basrah, Iraq. e-mail:
huda.challob@uobasrah.edu.iq

Akil J. Harfash

Department of Mathematics, College of Sciences, University of Basrah, Basrah, Iraq. e-mail:
akil.harfash@uobasrah.edu.iq

How to cite this article

Al-Musawi, G.A., Challob, H.A. and Harfash, A.J., Finite element analysis of cross-diffusion three-species Shigesada–Kawasaki–Teramoto model. *Iran. J. Numer. Anal. Optim.*, 2026; 16(2): 447–496.
<https://doi.org/10.22067/ijnao.2025.95757.1742>

the algorithm for the approximate problem will be presented, and the numerical results will be introduced.

AMS subject classifications (2020): Primary 65N12, 65M60, 35D30.

Keywords: finite element, Shigesada–Kawasaki–Teramoto, convergence, weak solution.

1 Introduction

Partial differential equations (PDEs) play a fundamental role in modeling various physical phenomena in science and engineering. Solving PDEs analytically can be challenging or even impossible for many complex problems. In such cases, numerical methods come to the rescue. However, numerical solutions may not always exist or be unique due to the nature of the PDE and the boundary conditions involved. In such situations, a concept called “weak solutions” provides an alternative way to tackle the problem. A weak solution is a more relaxed notion of a solution that does not necessarily satisfy the PDE pointwise but instead fulfills it in a distributional or integral sense. The concept of weak solutions is particularly useful for dealing with PDEs with irregular data or complex geometries [37].

The existence of weak solutions can be proven by approximating the original PDE with a sequence of simpler problems and showing the convergence of the approximate solutions to a limiting solution that satisfies the PDE weakly. The first step is to convert the strong form of the PDE into its weak form. This typically involves multiplying the PDE by a suitable test function and integrating both sides over the domain. The next step involves choosing a numerical method to approximate the weak formulation. The finite element method (FEM) is commonly used for this purpose. In FEM, the domain is discretized into a finite number of elements, and the unknown functions are approximated using piecewise polynomial basis functions. Then, some bounds are found for the numerical solutions by performing a series of mathematical steps. Next, a sequence of numerical approximations is considered as the mesh size tends to zero. Then, the convergence analysis aims to show that as the mesh size tends to zero, the sequence of numerical approximations converges to a limiting function in a suitable norm. It can then be demonstrated that the limiting function satisfies the weak formulation of the original PDE. Finally, with the convergence and the weak formulation satisfied, it can be concluded that the limiting function is a weak solution to the original PDE. This method of proving the existence of a weak formulation of PDE has been used previously in several studies; see, for example,

[10, 9, 8, 11, 13, 12, 27, 28, 32, 30, 33, 31, 2, 3, 4, 5, 6]. Also, the FEM has been used and analyzed recently to solve various problems; see, for example, [29, 15, 38, 40, 7, 36].

A self- and cross-diffusion population model refers to a computational or mathematical framework that incorporates interactions within and between multiple populations. It allows for the examination of both the dynamics within a single population and the interplay between different populations. In such a model, individuals within a population interact with each other, influencing their behavior, characteristics, or attributes. Interactions between different groups of people involve sharing information, resources, and other important factors. This kind of analysis is used in many areas like public health, ecology, sociology, and economics to understand complex systems with multiple interacting groups. By examining how these groups connect and interact, researchers can learn more about their overall behavior, notice emerging trends, and see how these dynamics between populations affect them all.

The Shigesada–Kawasaki–Teramoto (SKT) model is a mathematical tool created to help us understand how different species interact within ecosystems. Developed by Shigesada, Kawasaki, and Teramoto [39], this model uses a series of equations to track how the populations of predators and their prey change over time. One key feature of the model is its assumption that, when there are no predators around, prey populations can grow quickly. Meanwhile, the number of predators relies heavily on how many prey are available. The SKT model also looks at various factors, like how often predators catch their prey, how effectively they convert that prey into their own body mass, and the mortality rates among both groups. It helps researchers explore the relationship between these populations, including situations where populations stabilize or fluctuate dramatically.

This model has been a valuable part of ecological studies, shedding light on topics like predator-prey relationships, population cycles, and how environmental changes impact these dynamics. In this paper, we take a closer look at the SKT model as it applies to three species, incorporating cross-diffusion effects for one of them. The model has the following form: (Q) Find $\{u, v, w\} \in \mathbb{R}^{\geq 0} \times \mathbb{R}^{\geq 0} \times \mathbb{R}^{\geq 0}$ such that

$$\partial_t u = \Delta(\theta_1 u + \theta_{11} u^2) + (a_1 - b_1 u - c_1 v - d_1 w)u, \quad \text{in } \mathfrak{D}, \quad t > 0, \quad (1)$$

$$\begin{aligned} \partial_t v = \Delta(\theta_2 v + \theta_{21} uv + \theta_{22} v^2 + \theta_{23} vw) + (a_2 - b_2 u - c_2 v - d_2 w)v, \\ \text{in } \mathfrak{D}, \quad t > 0, \end{aligned} \quad (2)$$

$$\partial_t w = \Delta(\theta_3 w + \theta_{33} w^2) + (a_3 - b_3 u - c_3 v - d_3 w)w, \quad \text{in } \mathfrak{D}, \quad t > 0, \quad (3)$$

$$\frac{\partial u}{\partial \nu} = \frac{\partial v}{\partial \nu} = \frac{\partial w}{\partial \nu} = 0, \quad \text{in } \partial\mathfrak{D}, \quad t > 0, \quad (4)$$

$$u(\mathbf{x}, 0) = u_0(\mathbf{x}), \quad v(\mathbf{x}, 0) = v_0(\mathbf{x}), \quad w(\mathbf{x}, 0) = w_0(\mathbf{x}), \quad \text{in } \mathfrak{D}. \quad (5)$$

In this model, we have a bounded and connected open area $\mathfrak{D}$ shaped like a polygon in three-dimensional space or less. The boundary of this area, referred to as the Lipschitz boundary, is denoted by the symbol $\partial\mathfrak{D}$. Additionally, ν represents the outward unit normal vector at the boundary. This setup helps in understanding the geometric and mathematical properties of the region we're studying. This model, proposed in [39], is used to describe spatial segregation processes in populations of interacting species. In this context, $u = u(\mathbf{x}, t)$, $v = v(\mathbf{x}, t)$, and $w = w(\mathbf{x}, t)$ represent the densities of three competing species. The parameters θ_1, θ_2 , and θ_3 represent diffusion coefficients, while θ_{11}, θ_{22} , and θ_{33} correspond to self-diffusion coefficients. The coefficients θ_{21} and θ_{23} describe cross-diffusion, which reflects the population fluxes of one species influenced by the presence of other species. The cross-diffusion coefficient can assume positive, negative, or zero values. Additionally, a_1, a_2 , and a_3 denote intrinsic growth rates, b_1, c_2 , and d_3 account for intraspecific competition, and c_1, d_1, b_2, d_2, b_3 , and c_3 represent inter-specific competition coefficients.

The existence and properties of weak global solutions for cross-diffusion population models are challenging and are a subject of ongoing research in mathematical biology and PDEs. The notion of weak solutions allows for the possibility of certain singularities or discontinuities in the solutions while satisfying the PDEs on a global scale. There is a wide range of research and literature on cross-diffusion population models, and several papers discuss the existence and behavior of weak global solutions; see, for example, [16, 17, 34, 21, 18].

Numerous research studies have delved into the numerical analysis of nonlinear cross-diffusion population models. Specifically, the attraction-repulsion chemotaxis model has been investigated using FEM in [25, 26]. This research provides a fully discrete FEM for the regularized problem and establishes the existence of approximate solutions. Convergence of this FEM, both in spatial and temporal domains, was examined separately. Furthermore, in [23, 24], finite element solutions for the Keller–Segel equations with additional cross-diffusion and logistic source terms were demonstrated to exist. The research also proved the convergence of the approximate solutions in both space and time. Another set of studies (see [32, 30, 33]) focused on a fully discrete FEM for different models: The Keller–Segel model, the two-species chemotaxis system with two chemicals, and the chemotaxis-growth model with indirect attractant production, respectively. Notably, these studies achieved simultaneous convergence of the numerical solutions concerning both time and space by modifying the regularized method. This marks a significant advancement

in the study of the convergence of numerical solutions for such problems. The present work aims to develop and generalize the ideas discussed in the aforementioned articles, with a focus on addressing the treatment of self-cross-diffusion terms, which are prevalent in the current system. Moreover, this work investigates the existence of weak solutions for the three-species SKT model with self- and cross-diffusion terms using the FEM.

This paper focuses on the FEM of the three-species SKT system. A significant contribution of this work is the verification of the existence of a weak global solution to this system utilizing the FEM. The following section introduces the key definitions that will be utilized throughout the paper. In Section 3, the regularized problem of (Q) is introduced, and certain bounds for the solutions of the regularized problem are established. Next, in Section 4.1, FEM for spatial discretization and the forward finite difference method for time discretization are employed to approximate the regularized problem. The existence of the solution to the approximate problem is proven in Section 4.2 through the use of the fixed point theorem. To determine the spaces to which the approximate solutions belong, various bounds for the solutions are explored in Section 5. The uniqueness of the approximated solutions is then shown in Section 6. Section 7 deals with the examination of the convergence of the approximate solutions to the exact solutions, along with the convergence of the weak form of the approximate problem (Q) . Finally, in Section 8, an algorithm for computing the numerical solutions is presented. This paper expands and develops upon the ideas studied in [10, 9, 8, 11, 13, 12].

2 Notation

Let us begin by presenting the following Sobolev theorem:

$$H^1(\mathfrak{D}) \xhookrightarrow{c} L^\iota(\mathfrak{D}) \hookrightarrow (H^1(\mathfrak{D}))' \text{ holds for } \iota \in \begin{cases} [1, \infty] & \text{if } d = 1, \\ [1, \infty) & \text{if } d = 2, \\ [1, 6] & \text{if } d = 3, \end{cases} \quad (6)$$

where we utilize the notation $\hookrightarrow$ to denote continuous embedding. Furthermore, it has been demonstrated through the Rellich–Kondrachov theorem, as seen in [14, p. 114] and [19, p. 8], that the embedding in (6) is compact. In the case when $d = 3$, the range for ι changes from $\iota \in [1, 6]$ to $\iota \in [1, 6)$. We will denote the compact embedding using the symbol $\xhookrightarrow{c}$. Also, the following relationship will be necessary:

$$\langle \vartheta, \eta \rangle = (\vartheta, \eta), \quad \text{for all } \vartheta \in L^2(\mathfrak{D}) \text{ and } \eta \in H^1(\mathfrak{D}). \quad (7)$$

We also recall the Gagliardo–Nirenberg inequality [1] as follows: Let $\lambda \in [1, \infty]$, $\tau \geq 1$, and $\vartheta \in W^{\tau, \lambda}(\mathfrak{D})$. Then, there exist a constant C and $\gamma = \frac{d}{\tau} \left(\frac{1}{\lambda} - \frac{1}{\iota} \right)$ such that

$$\|\vartheta\|_{0, \iota} \leq C \|\vartheta\|_{0, \lambda}^{1-\gamma} \|\vartheta\|_{\tau, \lambda}^{\gamma} \quad \text{holds for } \iota \in \begin{cases} [\lambda, \infty] & \text{if } \tau - \frac{d}{\lambda} > 0, \\ [\lambda, \infty) & \text{if } \tau - \frac{d}{\lambda} = 0, \\ [\lambda, -\frac{d}{\tau-d/\lambda}] & \text{if } \tau - \frac{d}{\lambda} < 0. \end{cases} \quad (8)$$

We require the Sobolev interpolation result [14]: For $\vartheta \in H^1(\mathfrak{D})$, there exist a constant C and $\gamma = \frac{2d(\iota-1)}{\iota(d+2)}$ such that the following inequality holds:

$$\|\vartheta\|_{0, \iota} \leq C \|\vartheta\|_{0, 1}^{1-\gamma} \|\vartheta\|_1^{\gamma} \quad \text{holds for } \iota \in \begin{cases} [1, \infty] & \text{if } d = 1, \\ [1, \infty) & \text{if } d = 2, \\ [1, 6] & \text{if } d = 3. \end{cases} \quad (9)$$

We also invoke the embedding compactness result stated in [35, p. 58]. According to this result, consider three Banach spaces: Y_0 , Y_1 , and Y_2 , where Y_0 and Y_2 are reflexive and $Y_0 \xhookrightarrow{c} Y_1 \hookrightarrow Y_2$. Also, let

$$\mathcal{H} = \{\vartheta : \vartheta \in L^{\iota_1}(0, T; Y_0), \frac{\partial \vartheta}{\partial t} \in L^{\iota_2}(0, T; Y_2)\},$$

where $T < \infty$ and $1 < \iota_1, \iota_2 < \infty$. Then

$$\mathcal{H} \xhookrightarrow{c} L^{\iota_1}(0, T; Y_1). \quad (10)$$

The frequently used Hölder's inequality can be stated as follows: For $1 \leq r_1, r_2 \leq \infty$ with $\frac{1}{r_1} + \frac{1}{r_2} = 1$, if $\vartheta_1 \in L^{r_1}(\mathfrak{D})$ and $\vartheta_2 \in L^{r_2}(\mathfrak{D})$, then $\vartheta_1 \vartheta_2 \in L^1(\mathfrak{D})$ and

$$\begin{aligned} \|\vartheta_1 \vartheta_2\|_{0, 1} &= \int_{\mathfrak{D}} |\vartheta_1 \vartheta_2| d\mathbf{x} \leq \left(\int_{\mathfrak{D}} |\vartheta_1|^{r_1} d\mathbf{x} \right)^{\frac{1}{r_1}} \left(\int_{\mathfrak{D}} |\vartheta_2|^{r_2} d\mathbf{x} \right)^{\frac{1}{r_2}} \\ &= \|\vartheta_1\|_{0, r_1} \|\vartheta_2\|_{0, r_2}. \end{aligned} \quad (11)$$

Applying the above inequality twice leads to the generalization result in the form:

$$\begin{aligned} \|\vartheta_1 \vartheta_2 \vartheta_3\|_{0, 1} &= \int_{\mathfrak{D}} |\vartheta_1 \vartheta_2 \vartheta_3| d\mathbf{x} \\ &\leq \left(\int_{\mathfrak{D}} |\vartheta_1|^{r_1} d\mathbf{x} \right)^{\frac{1}{r_1}} \left(\int_{\mathfrak{D}} |\vartheta_2|^{r_2} d\mathbf{x} \right)^{\frac{1}{r_2}} \left(\int_{\mathfrak{D}} |\vartheta_3|^{r_3} d\mathbf{x} \right)^{\frac{1}{r_3}} \\ &= \|\vartheta_1\|_{0, r_1} \|\vartheta_2\|_{0, r_2} \|\vartheta_3\|_{0, r_3}, \end{aligned} \quad (12)$$

for $1 \leq r_1, r_2, r_3 \leq \infty$ such that $\frac{1}{r_1} + \frac{1}{r_2} + \frac{1}{r_3} = 1$.

The following inequality is also needed:

$$(\mathfrak{r}_1 - \mathfrak{r}_2)^2 \geq \frac{\mathfrak{r}_1^2}{2} - \mathfrak{r}_2^2, \quad \text{for all } \mathfrak{r}_1, \mathfrak{r}_2 \in \mathbb{R}. \quad (13)$$

We need the following essential inequalities, which are valid for all $\chi \in \mathbb{R}$, which can be expressed as follows:

$$(1 - \chi) = [1 - \chi]_- + [1 - \chi]_+ \leq [1 - \chi]_+ \leq 1 - [\chi]_-, \quad (14)$$

$$(\chi - 1) = [\chi - 1]_- + [\chi - 1]_+ \geq [\chi - 1]_- \geq [\chi]_- - 1, \quad (15)$$

where $[\chi]_+ = \max\{\chi, 0\}$ and $[\chi]_- = \min\{\chi, 0\}$.

3 A regularized problem

To assure a well-defined entropy inequality in the multidimensional existence proof, an entropy function must be established. This step is critical in our finite element approximation of problem (Q). To create a well-defined entropy inequality, we construct a function $\mathfrak{G} \in C^\infty(\mathbb{R}^{>0})$ with the conditions $\gamma \mathfrak{G}''(\gamma) = 1$ and $\mathfrak{G}(1) = 0$. This function $\mathfrak{G} : \mathbb{R}^{\geq 0} \rightarrow \mathbb{R}^{\geq 0}$ is defined as follows:

$$\mathfrak{G}(\gamma) := (\ln \gamma - 1)\gamma + 1, \quad (16)$$

and hence,

$$\mathfrak{G}'(\gamma) = \ln \gamma, \quad \mathfrak{G}''(\gamma) = \frac{1}{\gamma}. \quad (17)$$

By multiplying (1), (2), and (3) by $\frac{(\theta_{21})^2}{\theta_{11}\theta_{22}}\mathfrak{G}'(u)$, $\mathfrak{G}'(v)$, and $\frac{(\theta_{23})^2}{\theta_{22}\theta_{33}}\mathfrak{G}'(w)$, respectively, integrating over $\mathfrak{D}$, considering (4), (5) and the fact that $\lambda \nabla \mathfrak{G}'(\lambda) = \nabla \lambda$, and utilizing Young's inequality, we obtain that

$$\begin{aligned} & \frac{d}{dt} \int_{\mathfrak{D}} \mathfrak{G}(v) d\mathbf{x} + \int_{\mathfrak{D}} \theta_2 |\nabla v|^2 v^{-1} d\mathbf{x} + \int_{\mathfrak{D}} \theta_{21} u v^{-1} |\nabla v|^2 d\mathbf{x} \\ & - \int_{\mathfrak{D}} \frac{(\theta_{21})^2}{2\theta_{22}} |\nabla u|^2 d\mathbf{x} + \int_{\mathfrak{D}} \theta_{22} |\nabla v|^2 d\mathbf{x} - \frac{(\theta_{23})^2}{2\theta_{22}} |\nabla w|^2 d\mathbf{x} \\ & + \int_{\mathfrak{D}} \theta_{23} w v^{-1} |\nabla v|^2 d\mathbf{x} \leq \int_{\mathfrak{D}} a_2 v \mathfrak{G}'(v) d\mathbf{x} - \int_{\mathfrak{D}} b_2 u v \mathfrak{G}'(v) d\mathbf{x} \\ & \quad - \int_{\mathfrak{D}} c_2 v^2 \mathfrak{G}'(v) d\mathbf{x} - \int_{\mathfrak{D}} d_2 w v \mathfrak{G}'(v) d\mathbf{x}, \end{aligned} \quad (18)$$

and

$$\begin{aligned}
& \frac{d}{dt} \int_{\Omega} \frac{(\theta_{23})^2}{\theta_{22}\theta_{33}} \mathfrak{G}(w) d\mathbf{x} + \int_{\Omega} \frac{\theta_3(\theta_{23})^2}{\theta_{22}\theta_{33}} w^{-1} |\nabla w|^2 d\mathbf{x} + \int_{\Omega} \frac{2(\theta_{23})^2}{\theta_{22}} |\nabla w|^2 d\mathbf{x} \\
&= \int_{\Omega} \frac{a_3(\theta_{23})^2}{\theta_{22}\theta_{33}} w \mathfrak{G}'(w) d\mathbf{x} - \int_{\Omega} \frac{b_3(\theta_{23})^2}{\theta_{22}\theta_{33}} uw \mathfrak{G}'(w) d\mathbf{x} \\
&\quad - \int_{\Omega} \frac{c_3(\theta_{23})^2}{\theta_{22}\theta_{33}} vw \mathfrak{G}'(w) d\mathbf{x} - \int_{\Omega} \frac{d_3(\theta_{23})^2}{\theta_{22}\theta_{33}} w^2 \mathfrak{G}'(w) d\mathbf{x}.
\end{aligned} \tag{19}$$

Combining the above equations and rearranging the terms results in

$$\begin{aligned}
& \frac{d}{dt} \int_{\Omega} \left[\frac{(\theta_{21})^2}{\theta_{11}\theta_{22}} \mathfrak{G}(u) + \mathfrak{G}(v) + \frac{(\theta_{23})^2}{\theta_{22}\theta_{33}} \mathfrak{G}(w) \right] d\mathbf{x} \\
&+ \int_{\Omega} \frac{\theta_1(\theta_{21})^2}{\theta_{11}\theta_{22}} |\nabla u|^2 u^{-1} d\mathbf{x} + \int_{\Omega} \frac{3(\theta_{21})^2}{2\theta_{22}} |\nabla u|^2 d\mathbf{x} \\
&+ \int_{\Omega} \theta_2 |\nabla v|^2 v^{-1} d\mathbf{x} + \int_{\Omega} \theta_{21} uv^{-1} |\nabla v|^2 d\mathbf{x} + \int_{\Omega} \theta_{22} |\nabla v|^2 d\mathbf{x} \\
&+ \int_{\Omega} \theta_{23} wv^{-1} |\nabla v|^2 d\mathbf{x} + \int_{\Omega} \frac{\theta_3(\theta_{23})^2}{\theta_{22}\theta_{33}} w^{-1} |\nabla w|^2 d\mathbf{x} + \int_{\Omega} \frac{3(\theta_{23})^2}{2\theta_{22}} |\nabla w|^2 d\mathbf{x} \\
&\leq \int_{\Omega} \frac{a_1(\theta_{21})^2}{\theta_{11}\theta_{22}} u \mathfrak{G}'(u) d\mathbf{x} - \int_{\Omega} \frac{b_1(\theta_{21})^2}{\theta_{11}\theta_{22}} u \mathfrak{G}'(u) d\mathbf{x} \\
&\quad - \int_{\Omega} \frac{c_1(\theta_{21})^2}{\theta_{11}\theta_{22}} vu \mathfrak{G}'(u) d\mathbf{x} - \int_{\Omega} \frac{d_1(\theta_{21})^2}{\theta_{11}\theta_{22}} wu \mathfrak{G}'(u) d\mathbf{x} + \int_{\Omega} a_2 v \mathfrak{G}'(v) d\mathbf{x} \\
&\quad - \int_{\Omega} b_2 uv \mathfrak{G}'(v) d\mathbf{x} - \int_{\Omega} c_2 v^2 \mathfrak{G}'(v) d\mathbf{x} - \int_{\Omega} d_2 wv \mathfrak{G}'(v) d\mathbf{x} \\
&\quad + \int_{\Omega} \frac{a_3(\theta_{23})^2}{\theta_{22}\theta_{33}} w \mathfrak{G}'(w) d\mathbf{x} - \int_{\Omega} \frac{b_3(\theta_{23})^2}{\theta_{22}\theta_{33}} uw \mathfrak{G}'(w) d\mathbf{x} \\
&\quad - \int_{\Omega} \frac{c_3(\theta_{23})^2}{\theta_{22}\theta_{33}} vw \mathfrak{G}'(w) d\mathbf{x} - \int_{\Omega} \frac{d_3(\theta_{23})^2}{\theta_{22}\theta_{33}} w^2 \mathfrak{G}'(w) d\mathbf{x}.
\end{aligned} \tag{20}$$

The bound (20) is strictly formal since we are unable to predict a priori that $u(\mathbf{x}, t) \in \mathbb{R}^{>0}$ for $\mathfrak{G}$ to be well-defined and the fourth term on the left to be positive. A regularization approach is required to properly establish this constraint and build a numerical approximation for problem (Q). We thus use an alternate regularization approach, where the function $\mathfrak{G} \in C^2(\mathbb{R}^{>0})$ is replaced by a regularized function $\mathfrak{G}_{\varrho} : \mathbb{R} \rightarrow \mathbb{R}^{\geq 0}$, where $\varrho \in (0, e^{-1})$, such that

$$\mathfrak{G}_{\varrho}(\gamma) := \begin{cases} \frac{\gamma^2 - \varrho^2}{2\varrho} + (\ln \varrho - 1)\gamma + 1, & \gamma \leq \varrho, \\ (\ln \gamma - 1)\gamma + 1, & \varrho \leq \gamma \leq \varrho^{-1}, \\ \frac{\gamma^2 - \varrho^2}{2\varrho^{-1}} + (\ln \varrho^{-1} - 1)\gamma + 1, & \varrho^{-1} \leq \gamma. \end{cases} \tag{21}$$

Therefore, $\mathfrak{G}_{\varrho} \in C^{2,1}(\mathbb{R})$, where the first two derivatives of $\mathfrak{G}_{\varrho}$ are provided as follows:

$$\mathfrak{G}'_{\varrho}(\gamma) := \begin{cases} \varrho^{-1}\gamma + \ln \varrho - 1, & \gamma \leq \varrho, \\ \ln \gamma, & \varrho \leq \gamma \leq \varrho^{-1}, \\ \varrho\gamma + \ln \varrho^{-1} - 1, & \varrho^{-1} \leq \gamma, \end{cases} \quad (22)$$

and

$$\mathfrak{G}''_{\varrho}(\gamma) := \begin{cases} \frac{1}{\varrho}, & \gamma \leq \varrho, \\ \frac{1}{\gamma}, & \varrho \leq \gamma \leq \varrho^{-1}, \\ \varrho, & \varrho^{-1} \leq \gamma, \end{cases} \quad (23)$$

respectively. The function $\xi_{\varrho} : \mathbb{R} \rightarrow [\varrho, \varrho^{-1}]$ is presented in the following form:

$$\xi_{\varrho}(\gamma) := [\mathfrak{G}''_{\varrho}(\gamma)]^{-1} := \begin{cases} \varrho, & \gamma \leq \varrho, \\ \gamma, & \varrho \leq \gamma \leq \varrho^{-1}, \\ \frac{1}{\varrho}, & \varrho^{-1} \leq \gamma. \end{cases} \quad (24)$$

In addition, we take into account the following features that are associated with the function $\mathfrak{G}_{\varrho}(\gamma)$ and $\xi_{\varrho}(\gamma)$:

- for all $\varrho \in (0, e^{-1})$ and for all $\gamma \leq 0$, it follows that

$$\mathfrak{G}_{\varrho}(\gamma) \geq \frac{\gamma^2}{2\varrho}. \quad (25)$$

- for all $\varrho \in (0, e^{-1})$ and for all $\gamma \geq 0$, we have that

$$\mathfrak{G}_{\varrho}(\gamma) \geq \frac{\varrho\gamma^2}{2} - 1. \quad (26)$$

- Additionally, it is straightforward to show that for all $\varrho \in (0, e^{-1})$ and for all $\gamma \in \mathbb{R}$, we have the following results:

$$\max\{\xi_{\varrho}(\gamma), \gamma\mathfrak{G}'_{\varrho}(\gamma)\} \leq 2\mathfrak{G}_{\varrho}(\gamma) + 1, \quad (27)$$

$$\xi_{\varrho}(\gamma)\mathfrak{G}'_{\varrho}(\gamma) \geq \gamma - 1. \quad (28)$$

Taylor's theorem implies that for any function $\mathfrak{G} \in C^2(\mathbb{R})$, we have the following result:

$$(\gamma - \hat{r})\mathfrak{G}'(\gamma) = \mathfrak{G}(\gamma) - \mathfrak{G}(\hat{r}) + \frac{(\gamma - \hat{r})^2}{2}\mathfrak{G}''(v), \quad (29)$$

for some v between γ and $\hat{r}$. Here is the analogical regularized version of problem (Q) for any $\varrho \in (0, e^{-1})$:

(Q_ρ) Find $\{u_\rho, v_\rho, w_\rho\} \in \mathbb{R}^{\geq 0} \times \mathbb{R}^{\geq 0} \times \mathbb{R}^{\geq 0}$ such that

$$\begin{aligned} \partial_t u_\rho = & \nabla \cdot (\theta_1 \nabla u_\rho + 2\theta_{11} \xi_\rho(u_\rho) \nabla u_\rho) + a_1 u_\rho - b_1 \xi_\rho(u_\rho) \xi_\rho(u_\rho) \\ & - c_1 \xi_\rho(v_\rho) \xi_\rho(u_\rho) - d_1 \xi_\rho(w_\rho) \xi_\rho(u_\rho), \quad x \in \mathfrak{D}, \quad t > 0, \end{aligned} \quad (30)$$

$$\begin{aligned} \partial_t v_\rho = & \nabla \cdot (\theta_2 \nabla v_\rho + \theta_{21} \xi_\rho(v_\rho) \nabla u_\rho + \theta_{21} \xi_\rho(u_\rho) \nabla v_\rho \\ & + 2\theta_{22} \xi_\rho(v_\rho) \nabla v_\rho + \theta_{23} \xi_\rho(v_\rho) \nabla w_\rho + \theta_{23} \xi_\rho(w_\rho) \nabla v_\rho) \\ & + (a_2 v_\rho - b_2 \xi_\rho(u_\rho) \xi_\rho(v_\rho) - c_2 \xi_\rho(v_\rho) \xi_\rho(v_\rho) \\ & - d_2 \xi_\rho(w_\rho) \xi_\rho(v_\rho)), \quad x \in \mathfrak{D}, \quad t > 0, \end{aligned} \quad (31)$$

$$\begin{aligned} \partial_t w_\rho = & \nabla \cdot (\theta_3 \nabla w_\rho + 2\theta_{33} \xi_\rho(w_\rho) \nabla w_\rho) + (a_3 w_\rho - b_3 \xi_\rho(u_\rho) \xi_\rho(w_\rho) \\ & - c_3 \xi_\rho(v_\rho) \xi_\rho(w_\rho) - d_3 \xi_\rho(w_\rho) \xi_\rho(w_\rho)), \quad x \in \mathfrak{D}, \quad t > 0, \end{aligned} \quad (32)$$

$$\nabla u_\rho \cdot \nu = \nabla v_\rho \cdot \nu = \nabla w_\rho \cdot \nu = 0 \quad \text{on } \partial\mathfrak{D} \times [0, T], \quad (33)$$

$$u_\rho(\cdot, 0) = u_{\rho 0}, \quad v_\rho(\cdot, 0) = v_{\rho 0}, \quad w_\rho(\cdot, 0) = w_{\rho 0}, \quad \text{in } \mathfrak{D}. \quad (34)$$

Lemma 1. Given the initial conditions $u^0(\mathbf{x}), v^0(\mathbf{x}), w^0(\mathbf{x})$, where $\mathfrak{G}_\rho(u^0(\mathbf{x})) \in L^1(\mathfrak{D})$ and $u^0(\mathbf{x}), v^0(\mathbf{x}) \in L^2(\mathfrak{D})$, and $w^0(\mathbf{x}) \in L^2(\mathfrak{D})$, we can establish the existence of a positive constant C (independent of the regularization parameter ρ) such that the solution of (Q_ρ) satisfies the following inequality:

$$\begin{aligned} & \sup_{0 \leq t \leq T} \int_{\mathfrak{D}} \left[u_\rho \frac{(\theta_{21})^2}{\theta_{11}\theta_{22}} \mathfrak{G}_\rho(u_\rho) + v_\rho \mathfrak{G}_\rho(v_\rho) + w_\rho \frac{(\theta_{23})^2}{\theta_{22}\theta_{33}} \mathfrak{G}_\rho(w_\rho) \right] d\mathbf{x} \\ & + \int_0^T \int_{\mathfrak{D}} \frac{\theta_1(\theta_{21})^2}{\theta_{11}\theta_{22}} \frac{|\nabla u_\rho|^2}{\xi_\rho(u_\rho)} d\mathbf{x} dt + \int_0^T \int_{\mathfrak{D}} \frac{3(\theta_{21})^2}{2\theta_{22}} |\nabla u_\rho|^2 d\mathbf{x} dt \\ & + \int_0^T \int_{\mathfrak{D}} \theta_2 \frac{|\nabla v_\rho|^2}{\xi_\rho(v_\rho)} d\mathbf{x} dt + \int_0^T \int_{\mathfrak{D}} \theta_{21} \xi_\rho(u_\rho) \frac{|\nabla v_\rho|^2}{\xi_\rho(v_\rho)} d\mathbf{x} dt \\ & + \int_0^T \int_{\mathfrak{D}} \theta_{22} |\nabla v_\rho|^2 d\mathbf{x} dt + \int_0^T \int_{\mathfrak{D}} \theta_{23} \xi_\rho(w_\rho) \frac{|\nabla v_\rho|^2}{\xi_\rho(v_\rho)} d\mathbf{x} dt \\ & + \int_0^T \int_{\mathfrak{D}} \frac{\theta_3(\theta_{23})^2}{\theta_{22}\theta_{33}} \frac{|\nabla w_\rho|^2}{\xi_\rho(w_\rho)} d\mathbf{x} dt + \int_0^T \int_{\mathfrak{D}} \frac{3(\theta_{23})^2}{2\theta_{22}} |\nabla w_\rho|^2 d\mathbf{x} dt \leq C. \end{aligned} \quad (35)$$

Furthermore, we have that

$$\sup_{0 \leq t \leq T} \int_{\mathfrak{D}} |[u_\rho]_-|^2 d\mathbf{x} \leq C\rho. \quad (36)$$

Proof. After multiplying (30) by $\mathfrak{G}'_\varrho(u_\varrho)$, integrating by parts over $\mathfrak{D}$, and rearranging the terms, the resulting expression is as follows:

$$\begin{aligned} & \frac{d}{dt} \int_{\mathfrak{D}} \mathfrak{G}_\varrho(u_\varrho) d\mathbf{x} + \int_{\mathfrak{D}} \theta_1 \frac{|\nabla u_\varrho|^2}{\xi_\varrho(u_\varrho)} d\mathbf{x} + \int_{\mathfrak{D}} 2\theta_{11} |\nabla u_\varrho|^2 d\mathbf{x} \\ &= \int_{\mathfrak{D}} a_1 u_\varrho \mathfrak{G}'_\varrho(u_\varrho) d\mathbf{x} - \int_{\mathfrak{D}} b_1 \xi_\varrho(u_\varrho) \xi_\varrho(u_\varrho) \mathfrak{G}'_\varrho(u_\varrho) d\mathbf{x} \\ & \quad - \int_{\mathfrak{D}} c_1 \xi_\varrho(v_\varrho) \xi_\varrho(u_\varrho) \mathfrak{G}'_\varrho(u_\varrho) d\mathbf{x} - \int_{\mathfrak{D}} d_1 \xi_\varrho(w_\varrho) \xi_\varrho(u_\varrho) \mathfrak{G}'_\varrho(u_\varrho) d\mathbf{x}, \end{aligned} \quad (37)$$

where we have used the fact that

$$\xi_\varrho(\lambda_\varrho) \nabla \mathfrak{G}'_\varrho(\lambda_\varrho) = \nabla \lambda_\varrho. \quad (38)$$

By observing (24), (27), (28), (14), the Young's inequality, and (25), we can deduce that

$$\begin{aligned} & \int_{\mathfrak{D}} a_1 u_\varrho \mathfrak{G}'_\varrho(u_\varrho) d\mathbf{x} - \int_{\mathfrak{D}} b_1 \xi_\varrho(u_\varrho) \xi_\varrho(u_\varrho) \mathfrak{G}'_\varrho(u_\varrho) d\mathbf{x} \\ & - \int_{\mathfrak{D}} c_1 \xi_\varrho(v_\varrho) \xi_\varrho(u_\varrho) \mathfrak{G}'_\varrho(u_\varrho) d\mathbf{x} - \int_{\mathfrak{D}} d_1 \xi_\varrho(w_\varrho) \xi_\varrho(u_\varrho) \mathfrak{G}'_\varrho(u_\varrho) d\mathbf{x} \\ & \leq \int_{\mathfrak{D}} [(2a_1 + 4b_1 + c_1 + d_1) \mathfrak{G}_\varrho(u_\varrho) + C + 3c_1 \mathfrak{G}_\varrho(v_\varrho) + 3d_1 \mathfrak{G}_\varrho(w_\varrho)] d\mathbf{x}. \end{aligned} \quad (39)$$

By substituting (39) into (37), we obtain that

$$\begin{aligned} & \frac{d}{dt} \int_{\mathfrak{D}} \mathfrak{G}_\varrho(u_\varrho) d\mathbf{x} + \int_{\mathfrak{D}} \theta_1 \frac{|\nabla u_\varrho|^2}{\xi_\varrho(u_\varrho)} d\mathbf{x} + \int_{\mathfrak{D}} 2\theta_{11} |\nabla u_\varrho|^2 d\mathbf{x} \\ & \leq \int_{\mathfrak{D}} [(2a_1 + 4b_1 + c_1 + d_1) \mathfrak{G}_\varrho(u_\varrho) + C + 3c_1 \mathfrak{G}_\varrho(v_\varrho) + 3d_1 \mathfrak{G}_\varrho(w_\varrho)] d\mathbf{x}. \end{aligned} \quad (40)$$

Similarly, by multiplying (31) by $\mathfrak{G}'_\varrho(v_\varrho)$ and (32) by $\mathfrak{G}'_\varrho(w_\varrho)$, integrating over $\mathfrak{D}$, and applying integration by parts, along with the utilization of (38) and Young's inequality, we can derive the following results:

$$\begin{aligned} & \frac{d}{dt} \int_{\mathfrak{D}} \mathfrak{G}_\varrho(v_\varrho) d\mathbf{x} + \int_{\mathfrak{D}} \theta_2 \frac{|\nabla v_\varrho|^2}{\xi_\varrho(v_\varrho)} d\mathbf{x} - \int_{\mathfrak{D}} \frac{(\theta_{21})^2}{2\theta_{22}} |\nabla u_\varrho|^2 d\mathbf{x} + \int_{\mathfrak{D}} \theta_{21} \xi_\varrho(u_\varrho) \frac{|\nabla v_\varrho|^2}{\xi_\varrho(v_\varrho)} d\mathbf{x} \\ & + \int_{\mathfrak{D}} \theta_{22} |\nabla v_\varrho|^2 d\mathbf{x} - \int_{\mathfrak{D}} \frac{(\theta_{23})^2}{2\theta_{22}} |\nabla w_\varrho|^2 d\mathbf{x} + \int_{\mathfrak{D}} \theta_{23} \xi_\varrho(w_\varrho) \frac{|\nabla v_\varrho|^2}{\xi_\varrho(v_\varrho)} d\mathbf{x} \\ & \leq \int_{\mathfrak{D}} [(2a_2 + 4c_2 + b_2 + d_2) \mathfrak{G}_\varrho(v_\varrho) + C + 3b_2 \mathfrak{G}_\varrho(u_\varrho) + 3d_2 \mathfrak{G}_\varrho(w_\varrho)] d\mathbf{x}, \end{aligned} \quad (41)$$

and

$$\begin{aligned} & \frac{d}{dt} \int_{\mathfrak{D}} \frac{(\theta_{23})^2}{\theta_{22}\theta_{33}} \mathfrak{G}_{\varrho}(w_{\varrho}) d\mathbf{x} + \int_{\mathfrak{D}} \theta_3 \frac{|\nabla w_{\varrho}|^2}{\xi_{\varrho}(w_{\varrho})} d\mathbf{x} + \int_{\mathfrak{D}} 2\theta_{33} |\nabla w_{\varrho}|^2 d\mathbf{x} \\ & \leq \int_{\mathfrak{D}} [(2a_3 + 4d_3 + c_3 + b_3) \mathfrak{G}_{\varrho}(w_{\varrho}) + C + 3c_3 \mathfrak{G}_{\varrho}(v_{\varrho}) + 3b_3 \mathfrak{G}_{\varrho}(u_{\varrho})] d\mathbf{x}. \end{aligned} \quad (42)$$

By multiplying (40) and (43) by $\frac{(\theta_{21})^2}{\theta_{11}\theta_{22}}$ and $\frac{(\theta_{23})^2}{\theta_{22}\theta_{33}}$ respectively, and then summing the results with (41) and rearranging the terms, we obtain that

$$\begin{aligned} & \frac{d}{dt} \int_{\mathfrak{D}} \left[u_{\varrho} \frac{(\theta_{21})^2}{\theta_{11}\theta_{22}} \mathfrak{G}_{\varrho}(u_{\varrho}) + v_{\varrho} \mathfrak{G}_{\varrho}(v_{\varrho}) + w_{\varrho} \frac{(\theta_{23})^2}{\theta_{22}\theta_{33}} \mathfrak{G}_{\varrho}(w_{\varrho}) \right] d\mathbf{x} \\ & + \int_{\mathfrak{D}} \frac{\theta_1(\theta_{21})^2}{\theta_{11}\theta_{22}} \frac{|\nabla u_{\varrho}|^2}{\xi_{\varrho}(u_{\varrho})} d\mathbf{x} + \int_{\mathfrak{D}} \frac{3(\theta_{21})^2}{2\theta_{22}} |\nabla u_{\varrho}|^2 d\mathbf{x} + \int_{\mathfrak{D}} \theta_2 \frac{|\nabla v_{\varrho}|^2}{\xi_{\varrho}(v_{\varrho})} d\mathbf{x} \\ & + \int_{\mathfrak{D}} \theta_{21} \xi_{\varrho}(u_{\varrho}) \frac{|\nabla v_{\varrho}|^2}{\xi_{\varrho}(v_{\varrho})} d\mathbf{x} + \int_{\mathfrak{D}} \theta_{22} |\nabla v_{\varrho}|^2 d\mathbf{x} + \int_{\mathfrak{D}} \theta_{23} \xi_{\varrho}(w_{\varrho}) \frac{|\nabla v_{\varrho}|^2}{\xi_{\varrho}(v_{\varrho})} d\mathbf{x} \\ & + \int_{\mathfrak{D}} \frac{\theta_3(\theta_{23})^2}{\theta_{22}\theta_{33}} \frac{|\nabla w_{\varrho}|^2}{\xi_{\varrho}(w_{\varrho})} d\mathbf{x} + \int_{\mathfrak{D}} \frac{3(\theta_{23})^2}{2\theta_{22}} |\nabla w_{\varrho}|^2 d\mathbf{x} \\ & \leq C \left(\int_{\mathfrak{D}} \left\{ \frac{(\theta_{21})^2}{\theta_{11}\theta_{22}} \mathfrak{G}_{\varrho}(u_{\varrho}) + \mathfrak{G}_{\varrho}(v_{\varrho}) + \frac{(\theta_{23})^2}{\theta_{22}\theta_{33}} \mathfrak{G}_{\varrho}(w_{\varrho}) \right\} d\mathbf{x} + 1 \right). \end{aligned} \quad (43)$$

By utilizing the Grönwall lemma on (43), we can obtain the desired outcome mentioned in (35). Moreover, by considering (35) and (25), we can directly derive the result stated in (36). $\square$

4 A finite element method

4.1 Notation

(A) Consider $\mathfrak{D}$ as a polygonal domain in $\mathbb{R}^2$ or a polyhedral domain in $\mathbb{R}^3$. Let $\mathcal{T}^h$ be a quasi-uniform and right-angled partition of $\mathfrak{D}$ into open simplices $\tilde{\tau}$ with diameters $h_{\tilde{\tau}}$ and h denoting the maximum diameter among the simplices. We have $\overline{\mathfrak{D}} = \cup_{\tilde{\tau} \in \mathcal{T}^h} \overline{\tilde{\tau}}$. Let $(0, T)$ be partitioned into time steps δt_n , where $0 = t_0 < t_1 < \dots < t_{N-1} < t_N = T$, and δt represents the maximum time step size. Assume X^h as a subset of $H^1(\mathfrak{D})$, the finite element space consisting of continuous piecewise linear functions.

$$X^h := \{\omega \in C(\overline{\mathfrak{D}}) : \omega|_{\tilde{\tau}} \text{ is linear, for all } \tilde{\tau} \in \mathcal{T}^h\}.$$

We introduce the set $\{\eta_j\}_{j=0}^J$ as the standard basis functions for X^h , where X^h is a finite element space. These basis functions satisfy the property $\eta_j(p_i) = \delta_{ij}$ for $i, j = 0, \dots, J$. Here, $\mathcal{N}^h := \{p_j\}_{j=0}^J$ denotes the set of nodes of the partitioning $\mathcal{T}^h$. Moreover, we present that

$$\begin{aligned} X_{\geq 0}^h &:= \{\omega \in X^h : \omega(p_j) \geq 0, \ j = 0, \dots, J\} \\ &\subset H_{\geq 0}^1 := \{\Sigma \in H^1(\mathfrak{D}) : \Sigma \geq 0 \text{ a.e.} \in \mathfrak{D}\}. \end{aligned}$$

Define $\pi^h : C(\overline{\mathfrak{D}}) \rightarrow X^h$ as the Lagrange interpolation operator, also known as the piecewise linear interpolant, which has the form

$$\pi^h \Sigma(p_j) := \Sigma(p_j), \quad \text{for } j = 0, \dots, J.$$

We can define a discrete L^2 inner (semi-inner) product on $X^h(C(\overline{\mathfrak{D}}))$ as follows:

$$(\vartheta_1, \vartheta_2)^h := \int_{\mathfrak{D}} \pi^h(\vartheta_1(x)\vartheta_2(x)) d\mathbf{x} = \sum_{j=0}^J \widehat{M}_{jj} \vartheta_1(p_j) \vartheta_2(p_j), \quad (44)$$

where $\widehat{M}_{jj} = (\eta_j, \eta_j)^h = (1, \eta_j) > 0$. Taking (44) into consideration, it follows that

$$(\vartheta_1, \vartheta_2)^h = (\pi^h \vartheta_1, \vartheta_2)^h = (\pi^h \vartheta_1, \pi^h \vartheta_2)^h, \quad \text{for all } \vartheta_1, \vartheta_2 \in C(\overline{\mathfrak{D}}). \quad (45)$$

Here are some well-known results regarding the finite element space X^h and its induced discrete semi-norm on $C(\overline{\mathfrak{D}})$ and norm on X^h . The discrete semi-norm, denoted as $|\cdot|_h$, is defined as $|\cdot|_h := [(\cdot, \cdot)^h]^{1/2}$. It is widely recognized that $|\cdot|_h$ is equivalent to the norm $|\cdot|_0$, defined as $|\cdot|_0 := [(\cdot, \cdot)]^{1/2}$, in the following manner:

$$\|\omega\|_0^2 \leq |\omega|_h^2 \leq (d+2)\|\omega\|_0^2, \quad \text{for all } \omega \in X^h. \quad (46)$$

As a consequence of the quasi-uniform condition (see [19, Theorem 3.2.6]), we include extra inverse inequalities for future reference.

$$|\vartheta|_{1,\varsigma, \text{"}\tau\text{"}} \leq Ch_{\text{"}\tau\text{"}}^{-1} |\vartheta|_{0,\varsigma, \text{"}\tau\text{"}}, \quad 1 \leq \varsigma \leq \infty, \quad (47)$$

$$|\vartheta|_{k,\varsigma_1, \text{"}\tau\text{"}} \leq Ch_{\text{"}\tau\text{"}}^{-d(\frac{1}{\varsigma_2} - \frac{1}{\varsigma_1})} |\vartheta|_{k,\varsigma_2, \text{"}\tau\text{"}}, \quad 1 \leq \varsigma_2 \leq \varsigma_1 \leq \infty, \ k \in \{0, 1\}. \quad (48)$$

The interpolation results that we will need to use are as follows: For all $\vartheta \in W^{1,\gamma}(\mathfrak{D})$, $\gamma \in [2, \infty]$ if $d = 1$ and $\gamma \in (d, \infty]$ if $d = 2$ or 3 :

$$|(I - \pi^h)\vartheta|_{k,\gamma} \leq Ch^{1-k}|\vartheta|_{1,\gamma}, \quad \gamma \in \{0, 1\}, \quad (49)$$

$$\lim_{h \rightarrow 0} |(I - \pi^h)\vartheta|_{1,\gamma} = 0. \quad (50)$$

For all $\mathfrak{S}_1, \mathfrak{S}_2 \in X^h$, we will also require the following result (e.g., [20]):

$$|(\mathfrak{S}_1, \mathfrak{S}_2) - (\mathfrak{S}_1, \mathfrak{S}_2)^h| \leq Ch^{1+m}|\mathfrak{S}_1|_{m,n_1}|\mathfrak{S}_2|_{1,n_2}, \quad (51)$$

for $m \in \{0, 1\}$ and $1 \leq n_1, n_2 \leq \infty$ with $\frac{1}{n_1} + \frac{1}{n_2} = 1$.

The generalized version of (51) for future use is provided below. For any $\mathfrak{S}_1, \mathfrak{S}_2, \mathfrak{S}_3 \in X^h$, the following result holds:

$$|(\mathfrak{S}_1, \mathfrak{S}_2, \mathfrak{S}_3) - (\mathfrak{S}_1, \mathfrak{S}_2, \mathfrak{S}_3)^h| \leq Ch^2|\mathfrak{S}_1|_{1,n_1}|\mathfrak{S}_2|_{1,n_2}|\mathfrak{S}_3|_{1,n_3}, \quad (52)$$

where $1 \leq n_1, n_2, n_3 \leq \infty$ with $\frac{1}{n_1} + \frac{1}{n_2} + \frac{1}{n_3} = 1$.

We introduce, for any $\varrho \in (0, e^{-1})$, $\Xi_\varrho : X^h \rightarrow [L^\infty(\mathfrak{D})]^{d \times d}$ such that for all $\omega \in X^h$ and a.e. in $\mathfrak{D}$, we have that

$$\Xi_\varrho(\omega) \text{ is symmetric and positive definite,} \quad (53)$$

$$\Xi_\varrho(\omega) \nabla \pi^h[\mathfrak{G}'_\varrho(\omega)] = \nabla \omega. \quad (54)$$

In the simple case of $d = 1$, we construct Ξ_ϱ . Let $\omega \in X^h$ and $\tilde{\tau} \in \mathcal{T}^h$, with vertices p_j and p_k . We define the following:

$$\Xi_\varrho(\omega)|_{\tilde{\tau}} := \begin{cases} \frac{\omega(p_k) - \omega(p_j)}{\mathfrak{G}'_\varrho(\omega(p_k)) - \mathfrak{G}'_\varrho(\omega(p_j))} = \frac{1}{\mathfrak{G}''_\varrho(\omega(\hat{\zeta}))}, & \text{for } \hat{\zeta} \in \tilde{\tau} \text{ if } \omega(p_k) \neq \omega(p_j), \\ \frac{1}{\mathfrak{G}''_\varrho(\omega(p_k))} & \text{if } \omega(p_k) = \omega(p_j). \end{cases} \quad (55)$$

Since $\mathfrak{G}''_\varepsilon(s) > 0$ and $\sum_{j=0}^J \nabla \eta_j = 0$, it can be easily seen that the piecewise constant function Ξ_ε satisfies the conditions (53) and (54). Following [22], we extend the above construction to $d = 2$ or 3 . Let $\{e_i\}_{i=1}^d$ be the orthonormal vectors in $\mathbb{R}^d$, such that the j -th component of e_i is δ_{ij} , $i, j = 1 \rightarrow d$. Given non-zero constants $\theta_i, i = 1 \rightarrow d$, let $\hat{\tau}(\{\theta_i\}_{i=1}^d)$ be a reference simplex in $\mathbb{R}^d$ with vertices $\{\hat{p}_i\}_{i=1}^d$, where $\hat{p}_0$ is the origin and $\hat{p}_i = \hat{p}_{i-1} + \theta_i e_i, i = 1 \rightarrow d$. Given a $\tilde{\tau} \in \mathcal{T}^h$ with vertices $\{p_{j_i}\}_{i=0}^d$, such that p_{j_0} is not a right-angled vertex, then there exists a rotation/reflection matrix $R_{\tilde{\tau}}$ and non-zero constants $\{\hat{p}_i\}_{i=1}^d$ such that the mapping $R_{\tilde{\tau}} : \hat{x} \in \mathbb{R}^d \rightarrow p_{j_0} + R_{\tilde{\tau}} \hat{x} \in \mathbb{R}^d$ maps the vertex $\hat{p}_i$ to p_{j_i} , and hence $\hat{\tau} \equiv \hat{\tau}(\{\theta_i\}_{i=1}^d)$ to $\tilde{\tau}$. For all $\tilde{\tau} \in \mathcal{T}^h$ and $\omega \in V^h$, we set

$$\Xi_\varrho(\omega)|_{\tilde{\tau}} := R_{\tilde{\tau}} \widehat{\Xi_\varrho(\omega)}|_{\hat{\tau}} R_{\tilde{\tau}}^T, \quad (56)$$

where $\widehat{\Xi_\varrho(\omega)}|_{\hat{\tau}}$ is the $d \times d$ diagonal matrix with diagonal entries, $k = 1, \dots, d$,

$$[\widehat{\Xi}_\varrho(\widehat{\omega})|_{\widehat{\tau}}]_{kk} := \begin{cases} \frac{\widehat{\omega}(\widehat{p}_k) - \widehat{\omega}(\widehat{p}_j)}{\widehat{\mathfrak{S}}'_\varrho(\widehat{\omega}(\widehat{p}_k)) - \widehat{\mathfrak{S}}'_\varrho(\widehat{\omega}(\widehat{p}_j))} = \frac{\omega(p_{j_k}) - \omega(p_{j_0})}{\widehat{\mathfrak{S}}'_\varrho(\omega(p_{j_k})) - \widehat{\mathfrak{S}}'_\varrho(\omega(p_{j_0}))} \\ = \frac{1}{\widehat{\mathfrak{S}}''_\varrho(\omega(\widehat{\zeta}))} & \text{for some } \widehat{\zeta} \text{ between } p_{j_k} \text{ and } p_{j_0} \\ & \text{if } \omega(p_{j_k}) \neq \omega(p_{j_0}), \\ \frac{1}{\widehat{\mathfrak{S}}''_\varrho(\widehat{\omega}(\widehat{p}_0))} = \frac{1}{\widehat{\mathfrak{S}}''_\varrho(\omega(p_{j_0}))} & \text{if } \omega(p_{j_k}) = \omega(p_{j_0}). \end{cases} \quad (57)$$

For any given $\varrho \in (0, e^{-1})$, we consider the following fully discrete FEM of (Q_ϱ) :
 $(Q_\varrho^{h,\delta t})$ For $n \geq 1$ find $\{\widehat{\Phi}_\varrho^n, \widehat{\Psi}_\varrho^n, \widehat{\Theta}_\varrho^n\} \in [X^h]^3$ such that for all $\omega \in X^h$,

$$\begin{aligned} & \left(\frac{\widehat{\Phi}_\varrho^n - \widehat{\Phi}_\varrho^{n-1}}{\delta t}, \omega \right)^h + (\theta_1 \nabla \widehat{\Phi}_\varrho^n + 2\theta_{11} \Xi_\varrho(\widehat{\Phi}_\varrho^n) \nabla \widehat{\Phi}_\varrho^n, \nabla \omega) \\ & = (a_1 \widehat{\Phi}_\varrho^n - b_1 \xi_\varrho(\widehat{\Phi}_\varrho^{n-1}) \xi_\varrho(\widehat{\Phi}_\varrho^n) - c_1 \xi_\varrho(\widehat{\Psi}_\varrho^{n-1}) \xi_\varrho(\widehat{\Phi}_\varrho^n) \\ & \quad - d_1 \xi_\varrho(\widehat{\Theta}_\varrho^{n-1}) \xi_\varrho(\widehat{\Phi}_\varrho^n), \omega)^h, \end{aligned} \quad (58)$$

$$\begin{aligned} & \left(\frac{\widehat{\Psi}_\varrho^n - \widehat{\Psi}_\varrho^{n-1}}{\delta t}, \omega \right)^h + (\theta_2 \nabla \widehat{\Psi}_\varrho^n + \theta_{21} \Xi_\varrho(\widehat{\Psi}_\varrho^n) \nabla \widehat{\Phi}_\varrho^n + \theta_{21} \Xi_\varrho(\widehat{\Phi}_\varrho^n) \nabla \widehat{\Psi}_\varrho^n \\ & + 2\theta_{22} \Xi_\varrho(\widehat{\Psi}_\varrho^n) \nabla \widehat{\Psi}_\varrho^n + \theta_{23} \Xi_\varrho(\widehat{\Psi}_\varrho^n) \nabla \widehat{\Theta}_\varrho^n + \theta_{23} \Xi_\varrho(\widehat{\Theta}_\varrho^n) \nabla \widehat{\Psi}_\varrho^n, \nabla \omega) \\ & = (a_2 \widehat{\Psi}_\varrho^n - b_2 \xi_\varrho(\widehat{\Phi}_\varrho^{n-1}) \xi_\varrho(\widehat{\Psi}_\varrho^n) - c_2 \xi_\varrho(\widehat{\Psi}_\varrho^{n-1}) \xi_\varrho(\widehat{\Psi}_\varrho^n) \\ & \quad - d_2 \xi_\varrho(\widehat{\Theta}_\varrho^{n-1}) \xi_\varrho(\widehat{\Psi}_\varrho^n), \omega)^h, \end{aligned} \quad (59)$$

and

$$\begin{aligned} & \left(\frac{\widehat{\Theta}_\varrho^n - \widehat{\Theta}_\varrho^{n-1}}{\delta t}, \omega \right)^h + (\theta_3 \nabla \widehat{\Theta}_\varrho^n + 2\theta_{33} \Xi_\varrho(\widehat{\Theta}_\varrho^n) \nabla \widehat{\Theta}_\varrho^n, \nabla \omega) \\ & = (a_3 \widehat{\Theta}_\varrho^n - b_3 \xi_\varrho(\widehat{\Phi}_\varrho^{n-1}) \xi_\varrho(\widehat{\Theta}_\varrho^n) - c_3 \xi_\varrho(\widehat{\Psi}_\varrho^{n-1}) \xi_\varrho(\widehat{\Theta}_\varrho^n) - d_3 \xi_\varrho(\widehat{\Theta}_\varrho^{n-1}) \xi_\varrho(\widehat{\Theta}_\varrho^n), \omega)^h, \end{aligned} \quad (60)$$

where $\widehat{\Phi}_\varrho^0, \widehat{\Psi}_\varrho^0, \widehat{\Theta}_\varrho^0 \in X^h$ are given approximation of $u^0, v^0, w^0 \in C(\mathfrak{D})$, respectively.

Lemma 2. Assuming the validity of assumptions **(A)**, the function $\Xi_\varrho : X^h \rightarrow [L^\infty(\mathfrak{D})]^{d \times d}$ satisfies the following result for *a.e.* in $\mathfrak{D}$ and for all $\varrho \in (0, e^{-1})$:

$$\varrho v^T v \leq v^T \Xi_\varrho(\omega) v \leq \frac{1}{\varrho} v^T v, \quad \text{for all } v \in \mathbb{R}^d, \quad \text{for all } \omega \in X^h. \quad (61)$$

Lemma 3. The function $\Xi_\varrho : X^h \rightarrow [L^\infty(\mathfrak{D})]^{d \times d}$ is continuous for any $\varrho \in (0, e^{-1})$. This continuity is defined in the sense that

$$\begin{aligned} & \|(\Xi_\varrho(\mathfrak{S}_1) - \Xi_\varrho(\mathfrak{S}_2))|_{\widehat{\tau}}\| \leq \frac{2}{\varrho} \|\mathfrak{S}_1 - \mathfrak{S}_2\|_{0,\infty}, \\ & \text{for all } \mathfrak{S}_1, \mathfrak{S}_2 \in X^h \text{ and for all } \widehat{\tau} \in \mathcal{T}^h. \end{aligned} \quad (62)$$

Lemma 4. Assume the assumptions (A) are valid. Then, the function ξ_ϱ satisfies Let us suppose that the assumptions (A) hold true. Under this assumption, the function ξ_ϱ fulfills the following results

$$\|\nabla \pi^h[\xi_\varrho(\omega)]\|_0^2 \leq (\nabla \omega, \nabla \pi^h[\xi_\varrho(\omega)]), \quad \text{for all } \omega \in X^h. \quad (63)$$

The objective of this document is to establish the existence of solutions to problem (Q) by taking the limit $\varrho \rightarrow 0$ in the discrete problem. In order to accomplish this, bounds on the approximate solution $(Q_M^{\delta t})$ have been derived, which are independent of ϱ . It is important to note that the goal is to simultaneously approach the limits $\varrho \rightarrow 0$ and $\delta t \rightarrow 0$ in the discrete model, where the relationship between ϱ and δt is given by the condition $\delta t = o(\varrho)$ as $\varrho \rightarrow 0$.

4.2 Existence of the approximations

Ensuring the correctness and dependability of computational models requires investigating the existence of numerical solutions for PDEs using FEM. The existence of a numerical solution establishes a strong mathematical basis and validates that the FEM formulation is well-posed and able to approximate the system's actual physical behavior. Numerical results could be unstable or physically meaningless without this kind of verification. In order to create reliable simulations in the fields of engineering, physics, and applied sciences, it is important to analyze the existence and stability of FEM solutions.

To demonstrate the existence of $\hat{\Phi}_\varrho^n$, $\hat{\Psi}_\varrho^n$ and $\hat{\Theta}_\varrho^n$, $n \geq 1$ to the system (58), (59) and (60), for given $\hat{\Phi}_\varrho^{n-1}$, $\hat{\Psi}_\varrho^{n-1}$ and $\hat{\Theta}_\varrho^{n-1}$, it is useful to define the functions $M_u : X^h \times X^h \rightarrow X^h$, $M_v : X^h \times X^h \rightarrow X^h$ and $M_w : X^h \times X^h \rightarrow X^h$ so that for all $\omega \in X^h$, we have that

$$\begin{aligned} (M_u(\hat{\Phi}, \hat{\Psi}, \hat{\Theta}), \omega)^h &:= (\hat{\Phi} - \hat{\Phi}_\varrho^{n-1}, \omega)^h + \delta t(\theta_1 \nabla \hat{\Phi} + 2\theta_{11} \Xi_\varrho(\hat{\Phi}) \nabla \hat{\Phi}, \nabla \omega) \\ &\quad - \delta t(a_1 \hat{\Phi} - b_1 \xi_\varrho(\hat{\Phi}_\varrho^{n-1}) \xi_\varrho(\hat{\Phi}) - c_1 \xi_\varrho(\hat{\Psi}_\varrho^{n-1}) \xi_\varrho(\hat{\Phi}) \\ &\quad - d_1 \xi_\varrho(\hat{\Theta}_\varrho^{n-1}) \xi_\varrho(\hat{\Phi}), \omega)^h, \end{aligned} \quad (64)$$

$$\begin{aligned} (M_v(\hat{\Phi}, \hat{\Psi}, \hat{\Theta}), \omega)^h &:= (\hat{\Psi} - \hat{\Psi}_\varrho^{n-1}, \omega)^h + \delta t(\theta_2 \nabla \hat{\Psi} + \theta_{21} \Xi_\varrho(\hat{\Psi}) \nabla \hat{\Phi} \\ &\quad + \theta_{21} \Xi_\varrho(\hat{\Phi}) \nabla \hat{\Psi} + 2\theta_{22} \Xi_\varrho(\hat{\Psi}) \nabla \hat{\Psi} \\ &\quad + \theta_{23} \Xi_\varrho(\hat{\Psi}) \nabla \hat{\Theta} + \theta_{23} \Xi_\varrho(\hat{\Theta}) \nabla \hat{\Psi}, \nabla \omega) \\ &\quad - \delta t(a_2 \hat{\Psi} - b_2 \xi_\varrho(\hat{\Phi}_\varrho^{n-1}) \xi_\varrho(\hat{\Psi}) - c_2 \xi_\varrho(\hat{\Psi}_\varrho^{n-1}) \xi_\varrho(\hat{\Psi}) \\ &\quad - d_2 \xi_\varrho(\hat{\Theta}_\varrho^{n-1}) \xi_\varrho(\hat{\Psi}), \omega)^h, \end{aligned} \quad (65)$$

and

$$\begin{aligned} (M_w(\widehat{\Phi}, \widehat{\Psi}, \widehat{\Theta}), \omega)^h := & (\widehat{\Theta} - \widehat{\Theta}_\varrho^{n-1}, \omega)^h + \delta t(\theta_3 \nabla \widehat{\Theta} + 2\theta_{33} \Xi_\varrho(\widehat{\Theta}) \nabla \widehat{\Theta}, \nabla \omega) \\ & - \delta t(a_3 \widehat{\Theta} - b_3 \xi_\varrho(\widehat{\Phi}_\varrho^{n-1}) \xi_\varrho(\widehat{\Theta})) \\ & - c_3 \xi_\varrho(\widehat{\Psi}_\varrho^{n-1}) \xi_\varrho(\widehat{\Theta}) - d_3 \xi_\varrho(\widehat{\Theta}_\varrho^{n-1}) \xi_\varrho(\widehat{\Theta}), \omega)^h. \end{aligned} \quad (66)$$

Firstly, it is worth noting that the continuous piecewise linear functions M_u , M_v and M_w can be uniquely determined based on their values at the node points. This uniqueness can be observed by setting $\omega \equiv \eta_j$, for $j = 0, \dots, J$ in (64), which subsequently leads to solvable square matrix systems

$$\widehat{M} M_u(\widehat{\Phi}, \widehat{\Psi}, \widehat{\Theta}) = \mathcal{S}_1,$$

$$\widehat{M} M_v(\widehat{\Phi}, \widehat{\Psi}, \widehat{\Theta}) = \mathcal{S}_2,$$

$$\widehat{M} M_w(\widehat{\Phi}, \widehat{\Psi}, \widehat{\Theta}) = \mathcal{S}_3,$$

where $\widehat{M}$ is the lumped mass matrix and $\mathcal{S}_1$, $\mathcal{S}_2$ and $\mathcal{S}_3$ are given vectors in terms of the nodal values of $\widehat{\Phi}$, $\widehat{\Psi}$, $\widehat{\Theta}$, $\widehat{\Phi}_\varrho^{n-1}$, $\widehat{\Psi}_\varrho^{n-1}$ and $\widehat{\Theta}_\varrho^{n-1}$ thus, the functions M_u , M_v and M_w are well defined.

The problem $(Q_\varrho^{h,\delta t})$ can be restated based on (64), (65), and (66), in the following form: For given $\{\widehat{\Phi}_\varrho^0, \widehat{\Psi}_\varrho^0, \widehat{\Theta}_\varrho^0\} \in X^h \times X^h \times X^h$, find $\{\widehat{\Phi}_\varrho^n, \widehat{\Psi}_\varrho^n, \widehat{\Theta}_\varrho^n\} \in X^h \times X^h \times X^h$, $n \geq 1$, such that

$$M_u(\widehat{\Phi}, \widehat{\Psi}, \widehat{\Theta}) = 0, \quad M_v(\widehat{\Phi}, \widehat{\Psi}, \widehat{\Theta}) = 0, \quad M_w(\widehat{\Phi}, \widehat{\Psi}, \widehat{\Theta}) = 0. \quad (67)$$

Lemma 5. The functions $M_u, M_v, M_w : [X^h]_R^3 \rightarrow X^h$, are continuous for any $R > 0$, where

$$[X^h]_R^3 = \left\{ \{\mathfrak{S}_1, \mathfrak{S}_2, \mathfrak{S}_3\} \in X^h \times X^h \times X^h : |\mathfrak{S}_1|_h^3 + |\mathfrak{S}_3|_h^2 + |\mathfrak{S}_3|_h^3 \leq R^2 \right\}.$$

Proof. By using a method similar to that used in [27, 28, 32], the aforementioned theorems can be demonstrated. $\square$

Theorem 1. Assume the assumptions (A) are valid. Suppose that

$\{\widehat{\Phi}_\varrho^{n-1}, \widehat{\Psi}_\varrho^{n-1}, \widehat{\Theta}_\varrho^{n-1}\} \in X^h \times X^h \times X^h$ be given a solution to the $(n-1)$ -th step of $(Q_\varrho^{h,\delta t})$, for some $n = 1, 2, \dots, N$, then for all $\varrho \in (0, e^{-1})$, for all $h > 0$, and for all $\delta t \leq \frac{1}{2\theta_i(2+a_i)}$, for all $i = 1, 2$ there exists a solution $\{\widehat{\Phi}_\varrho^n, \widehat{\Psi}_\varrho^n, \widehat{\Theta}_\varrho^n\} \in [X^h]_R^3$ to the n -th step of $(Q_\varrho^{h,\delta t})$.

Proof. The proofs of the above theorems can be established using a methodology analogous to that employed in [30, 33, 31]. $\square$

To process the initial data in the fully-discrete approximation, we introduce the discrete L^2 -projection $P^h : L^2(\mathfrak{D}) \rightarrow X^h$, where Σ is given by

$$(P^h \Sigma, \omega)^h = (\Sigma, \omega), \quad \text{for all } \omega \in X^h. \quad (68)$$

The above projection fulfills the following key outcomes [11]:

$$\|P^h \Sigma\|_{0,\infty} \leq \|\Sigma\|_{0,\infty}, \quad \text{for all } \Sigma \in L^\infty(\mathfrak{D}), \quad (69)$$

$$\|(I - P^h)\Sigma\|_{m,s} \leq Ch^{1-m} \|\Sigma\|_{1,s}, \quad \text{for all } \Sigma \in W^{1,s}(\mathfrak{D}), \quad (70)$$

$$\text{for any } s \in [2, \infty] \quad \text{and } m \in \{0, 1\}. \quad (71)$$

We define the “inverse Laplacian” operator $\mathfrak{G}_q$ for any $q \in (1, 2]$ as the mapping $\mathfrak{G}_q : (W^{1,q'}(\mathfrak{D}))' \rightarrow W^{1,q}(\mathfrak{D})$ in the form

$$(\nabla \mathfrak{G}_q \Pi, \nabla \Sigma) + (\mathfrak{G}_q \Pi, \Sigma) = \langle \Pi, \Sigma \rangle_{q'}, \quad \text{for all } \Sigma \in W^{1,q'}(\mathfrak{D}), \quad (72)$$

where $\frac{1}{q} + \frac{1}{q'} = 1$ and $\langle \cdot, \cdot \rangle_{q'}$ denotes the duality pairing between $(W^{1,q'}(\mathfrak{D}))'$ and $W^{1,q'}(\mathfrak{D})$ such that

$$\langle \Pi, \Sigma \rangle_{q'} = (\Pi, \Sigma), \quad \text{for all } \Pi \in L^2(\mathfrak{D}), \Sigma \in W^{1,q'}(\mathfrak{D}). \quad (73)$$

The operator $\mathfrak{G}$ is well-posed, which can be shown by the generalized Lax-Milgram theorem, which guarantees the existence of a positive constant C satisfying the following inequality:

$$\|\mathfrak{G}_q \Pi\|_{1,q} \leq C \|\Pi\|_{(W^{1,q'}(\mathfrak{D}))'}, \quad \text{for all } \Pi \in (W^{1,q'}(\mathfrak{D}))'. \quad (74)$$

We also introduce the operator $\mathfrak{G} : (H^1(\mathfrak{D}))' \rightarrow H^1(\mathfrak{D})$ in the form

$$(\nabla \mathfrak{G} \Pi, \nabla \Sigma) + (\mathfrak{G} \Pi, \Sigma) = \langle \Pi, \Sigma \rangle, \quad \text{for all } \Sigma \in H^1(\mathfrak{D}), \quad (75)$$

where $\langle \cdot, \cdot \rangle$ represents the duality pairing between $(H^1(\mathfrak{D}))'$ and $H^1(\mathfrak{D})$, so that

$$\langle \Pi, \Sigma \rangle = (\Pi, \Sigma), \quad \text{for all } \Pi \in L^2(\mathfrak{D}), \Sigma \in H^1(\mathfrak{D}). \quad (76)$$

We can also conclude, for $q \in (1, 2]$, that

$$\|\omega\|_{0,q} \leq Ch^{-1} \|\mathfrak{G}_q \omega\|_{1,q}, \quad \text{for all } \omega \in X^h. \quad (77)$$

5 Stability bounds

To guarantee the accuracy, stability, and convergence of the results, bounds for the numerical solution using FEM in various spaces must be established. These bounds ensure that the approximations meet the governing equations and boundary conditions across a range of spatial domains, help regulate the behavior of the solution, and stop non-physical oscillations. The numerical model may yield erratic or imprecise results in the absence of appropriate bounds, particularly in complex geometries or heterogeneous media.

We develop a discrete version of the a priori estimates referred to in Lemma 4 in the next section of this paper. The solutions $\hat{\Phi}_\varrho^n, \hat{\Psi}_\varrho^n, \hat{\Theta}_\varrho^n$ are also shown to have a number of uniform bounds that are independent of the parameters ϱ , h , and δt . Because they guarantee the stability of different norms of the numerical solutions, these constraints are extremely significant. Our main objective is to show that the issue (Q) has weak solutions. The next stage in our proof is to determine the limits $\delta t \rightarrow 0^+$ and $h, \varrho \rightarrow 0$ in the system $(Q_\varrho^{h,\delta t})$. To achieve this, we establish a connection between the cut-off parameter $\varrho < 1$ and the time step δt , ensuring that $\delta t = o(\varrho)$ as $\varrho \rightarrow 0$.

Lemma 6. Given $\hat{\Phi}_\varrho^{n-1}, \hat{\Psi}_\varrho^{n-1}, \hat{\Theta}_\varrho^{n-1} \in X^h \times X^h \times X^h$ for $n = 1, \dots, N$, there exists a solution $\hat{\Phi}_\varrho^n, \hat{\Psi}_\varrho^n, \hat{\Theta}_\varrho^n \in X^h \times X^h \times X^h$ to the n -th step of $(Q_\varrho^{h,\delta t})$ for all $\varrho \in (0, e^{-1})$, $h > 0$, and $\delta t > 0$, such that

$$\begin{aligned}
& \frac{(\theta_{21})^2}{\theta_{22}\theta_{11}} [1 - 2\delta ta_1 - \delta tb_1 - \delta tc_1 - \delta td_1] (\mathfrak{G}_\varrho(\hat{\Phi}_\varrho^n), 1)^h \\
& + (1 - 2a_2\delta t - b_2\delta t - c_2\delta t - d_2\delta t) (\mathfrak{G}_\varrho(\hat{\Psi}_\varrho^n), 1)^h \\
& + \frac{(\theta_{23})^2}{\theta_{22}\theta_{33}} [1 - 2\delta ta_3 - \delta tb_3 - \delta tc_3 - \delta td_3] (\mathfrak{G}_\varrho(\hat{\Theta}_\varrho^n), 1)^h \\
& + \frac{\delta t\theta_1(\theta_{21})^2}{\theta_{22}\theta_{11}} \int_{\Omega} \frac{|\nabla \hat{\Phi}_\varrho^n|^2}{\xi_\varrho(\hat{\Phi}_\varrho^n)} d\mathbf{x} \\
& + \frac{3\delta t(\theta_{21})^2}{2\theta_{22}} |\hat{\Phi}_\varrho^n|_1^2 + \delta t\theta_2 \int_{\Omega} \frac{|\nabla \hat{\Psi}_\varrho^n|^2}{\xi_\varrho(\hat{\Psi}_\varrho^n)} d\mathbf{x} \\
& + \delta t\theta_{21} \int_{\Omega} |\nabla \hat{\Psi}_\varrho^n|^2 \frac{\Xi_\varrho(\hat{\Phi}_\varrho^n)}{\Xi_\varrho(\hat{\Psi}_\varrho^n)} d\mathbf{x} + \delta t\theta_{22} |\hat{\Psi}_\varrho^n|_1^2 \\
& + \delta t\theta_{23} \int_{\Omega} |\nabla \hat{\Psi}_\varrho^n|^2 \frac{\Xi_\varrho(\hat{\Theta}_\varrho^n)}{\Xi_\varrho(\hat{\Psi}_\varrho^n)} d\mathbf{x} + \frac{\delta t\theta_3(\theta_{23})^2}{\theta_{22}\theta_{33}} \int_{\Omega} \frac{|\nabla \hat{\Theta}_\varrho^n|^2}{\xi_\varrho(\hat{\Theta}_\varrho^n)} d\mathbf{x} + \frac{3\delta t(\theta_{23})^2}{2\theta_{22}} |\hat{\Theta}_\varrho^n|_1^2 \\
& \leq \left[\frac{(\theta_{21})^2}{\theta_{22}\theta_{11}} + \frac{3\delta tb_1(\theta_{21})^2}{\theta_{22}\theta_{11}} + 3b_2\delta t + \frac{3b_3\delta t(\theta_{23})^2}{\theta_{22}\theta_{33}} \right] (\mathfrak{G}_\varrho(\hat{\Phi}_\varrho^{n-1}), 1)^h
\end{aligned}$$

$$\begin{aligned}
& + \left[\frac{3c_1\delta t(\theta_{21})^2}{\theta_{22}\theta_{11}} + 1 + 3c_2\delta t + \frac{3c_3\delta t(\theta_{23})^2}{\theta_{22}\theta_{33}} \right] (\mathfrak{G}_\varrho(\widehat{\Psi}_\varrho^{n-1}), 1)^h \\
& + \left[\frac{3d_1\delta t(\theta_{21})^2}{\theta_{22}\theta_{11}} + 3d_2\delta t + \frac{(\theta_{23})^2}{\theta_{22}\theta_{33}} + \frac{3\delta t d_3(\theta_{23})^2}{\theta_{22}\theta_{33}} \right] (\mathfrak{G}_\varrho(\widehat{\Theta}_\varrho^{n-1}), 1)^h \\
& + \delta t C(|\mathfrak{D}|, a_1, a_2, a_3, \theta_{21}, \theta_{22}, \theta_{11}, \theta_{23}, \theta_{22}, \theta_{33}).
\end{aligned} \tag{78}$$

Proof. Using $\omega \equiv \delta t \pi^h [\mathfrak{G}'_\varrho(\widehat{\Phi}_\varrho^n)]$ as a test function in (58), and considering (45) and (54), we obtain that

$$\begin{aligned}
& (\widehat{\Phi}_\varrho^n - \widehat{\Phi}_\varrho^{n-1}, \mathfrak{G}'_\varrho(\widehat{\Phi}_\varrho^n))^h + \delta t \theta_1 \int_{\mathfrak{D}} \frac{|\nabla \widehat{\Phi}_\varrho^n|^2}{\xi_\varrho(\widehat{\Phi}_\varrho^n)} d\mathbf{x} + 2\delta t \theta_{11} |\widehat{\Phi}_\varrho^n|_1^2 \\
& = \delta t (a_1 \widehat{\Phi}_\varrho^n - b_1 \xi_\varrho(\widehat{\Phi}_\varrho^{n-1}) \xi_\varrho(\widehat{\Phi}_\varrho^n) - c_1 \xi_\varrho(\widehat{\Psi}_\varrho^{n-1}) \xi_\varrho(\widehat{\Phi}_\varrho^n) \\
& \quad - d_1 \xi_\varrho(\widehat{\Theta}_\varrho^{n-1}) \xi_\varrho(\widehat{\Phi}_\varrho^n), \mathfrak{G}'_\varrho(\widehat{\Phi}_\varrho^n))^h.
\end{aligned} \tag{79}$$

By utilizing (27), (28), (14), Young's inequality, and the fact that $\mathfrak{G}_\varrho(\cdot) \geq 0$, we can deduce that

$$\begin{aligned}
& \delta t (a_1 \widehat{\Phi}_\varrho^n - b_1 \xi_\varrho(\widehat{\Phi}_\varrho^n) \xi_\varrho(\widehat{\Phi}_\varrho^{n-1}) - c_1 \xi_\varrho(\widehat{\Phi}_\varrho^n) \xi_\varrho(\widehat{\Psi}_\varrho^{n-1}) - d_1 \xi_\varrho(\widehat{\Phi}_\varrho^n) \xi_\varrho(\widehat{\Theta}_\varrho^{n-1}), \mathfrak{G}'_\varrho(\widehat{\Phi}_\varrho^n))^h \\
& \leq 2a_1 \delta t (\mathfrak{G}_\varrho(\widehat{\Phi}_\varrho^n), 1)^h + b_1 \delta t (\mathfrak{G}_\varrho(\widehat{\Phi}_\varrho^n), 1)^h + c_1 \delta t (\mathfrak{G}_\varrho(\widehat{\Phi}_\varrho^n), 1)^h + d_1 \delta t (\mathfrak{G}_\varrho(\widehat{\Phi}_\varrho^n), 1)^h \\
& \quad + 3b_1 \delta t (\mathfrak{G}_\varrho(\widehat{\Phi}_\varrho^{n-1}), 1)^h + 3c_1 \delta t (\mathfrak{G}_\varrho(\widehat{\Psi}_\varrho^{n-1}), 1)^h \\
& \quad + 3d_1 \delta t (\mathfrak{G}_\varrho(\widehat{\Theta}_\varrho^{n-1}), 1)^h + \delta t C(|\mathfrak{D}|, a_1).
\end{aligned} \tag{80}$$

By substituting (80) into (79), we obtain the following expression:

$$\begin{aligned}
& (\widehat{\Phi}_\varrho^n - \widehat{\Phi}_\varrho^{n-1}, \mathfrak{G}'_\varrho(\widehat{\Phi}_\varrho^n))^h + \delta t \theta_1 \int_{\mathfrak{D}} \frac{|\nabla \widehat{\Phi}_\varrho^n|^2}{\xi_\varrho(\widehat{\Phi}_\varrho^n)} d\mathbf{x} + 2\delta t \theta_{11} |\widehat{\Phi}_\varrho^n|_1^2 \\
& \leq 2a_1 \delta t (\mathfrak{G}_\varrho(\widehat{\Phi}_\varrho^n), 1)^h + b_1 \delta t (\mathfrak{G}_\varrho(\widehat{\Phi}_\varrho^n), 1)^h + c_1 \delta t (\mathfrak{G}_\varrho(\widehat{\Phi}_\varrho^n), 1)^h + d_1 \delta t (\mathfrak{G}_\varrho(\widehat{\Phi}_\varrho^n), 1)^h \\
& \quad + 3b_1 \delta t (\mathfrak{G}_\varrho(\widehat{\Phi}_\varrho^{n-1}), 1)^h + 3c_1 \delta t (\mathfrak{G}_\varrho(\widehat{\Psi}_\varrho^{n-1}), 1)^h \\
& \quad + 3d_1 \delta t (\mathfrak{G}_\varrho(\widehat{\Theta}_\varrho^{n-1}), 1)^h + \delta t C(|\mathfrak{D}|, a_1).
\end{aligned} \tag{81}$$

Therefore, it can be deduced from the second inequality in (81) that

$$\begin{aligned}
& [1 - 2\delta t a_1 - \delta t b_1 - \delta t c_1 - \delta t d_1] (\mathfrak{G}_\varrho(\widehat{\Phi}_\varrho^n), 1)^h + \delta t \theta_1 \int_{\mathfrak{D}} \frac{|\nabla \widehat{\Phi}_\varrho^n|^2}{\xi_\varrho(\widehat{\Phi}_\varrho^n)} d\mathbf{x} + 2\delta t \theta_{11} |\widehat{\Phi}_\varrho^n|_1^2 \\
& \leq [1 + 3\delta t b_1] (\mathfrak{G}_\varrho(\widehat{\Phi}_\varrho^{n-1}), 1)^h + 3c_1 \delta t (\mathfrak{G}_\varrho(\widehat{\Psi}_\varrho^{n-1}), 1)^h + 3d_1 \delta t (\mathfrak{G}_\varrho(\widehat{\Theta}_\varrho^{n-1}), 1)^h \\
& \quad + \delta t C(|\mathfrak{D}|, a_1).
\end{aligned} \tag{82}$$

Using $\omega \equiv \delta t \pi^h [\mathfrak{G}'_\varrho(\widehat{\Psi}_\varrho^n)]$ and $\omega \equiv \delta t \pi^h [\mathfrak{G}'_\varrho(\widehat{\Theta}_\varrho^n)]$ as a test function in (59) and (60), respectively, and then by repeating the techniques utilized to construct (5), it follows that

$$\begin{aligned}
 & (1 - 2a_2\delta t - b_2\delta t - c_2\delta t - d_2\delta t)(\mathfrak{G}_\varrho(\widehat{\Psi}_\varrho^n), 1)^h + \delta t \theta_2 \int_{\mathfrak{D}} \frac{|\nabla \widehat{\Psi}_\varrho^n|^2}{\xi_\varrho(\widehat{\Psi}_\varrho^n)} d\mathbf{x} \\
 & - \frac{\delta t (\theta_{21})^2}{2\theta_{22}} |\widehat{\Phi}_\varrho^n|_1^2 + \delta t \theta_{21} \int_{\mathfrak{D}} |\nabla \widehat{\Psi}_\varrho^n|^2 \frac{\Xi_\varrho(\widehat{\Phi}_\varrho^n)}{\Xi_\varrho(\widehat{\Psi}_\varrho^n)} d\mathbf{x} + \delta t \theta_{22} |\widehat{\Psi}_\varrho^n|_1^2 \\
 & \leq - \frac{\delta t (\theta_{23})^2}{2\theta_{22}} |\widehat{\Theta}_\varrho^n|_1^2 + \delta t \theta_{23} \int_{\mathfrak{D}} |\nabla \widehat{\Psi}_\varrho^n|^2 \frac{\Xi_\varrho(\widehat{\Theta}_\varrho^n)}{\Xi_\varrho(\widehat{\Psi}_\varrho^n)} d\mathbf{x} + 3b_2\delta t (\mathfrak{G}_\varrho(\widehat{\Phi}_\varrho^{n-1}), 1)^h \\
 & + (1 + 3c_2)\delta t (\mathfrak{G}_\varrho(\widehat{\Psi}_\varrho^{n-1}), 1)^h + 3d_2\delta t (\mathfrak{G}_\varrho(\widehat{\Theta}_\varrho^{n-1}), 1)^h + \delta t C(|\mathfrak{D}|, a_2),
 \end{aligned} \tag{83}$$

and

$$\begin{aligned}
 & [1 - 2\delta t a_3 - \delta t b_3 - \delta t c_3 - \delta t d_3](\mathfrak{G}_\varrho(\widehat{\Theta}_\varrho^n), 1)^h + \delta t \theta_3 \int_{\mathfrak{D}} \frac{|\nabla \widehat{\Theta}_\varrho^n|^2}{\xi_\varrho(\widehat{\Theta}_\varrho^n)} d\mathbf{x} + 2\delta t \theta_{33} |\widehat{\Theta}_\varrho^n|_1^2 \\
 & \leq 3b_3\delta t (\mathfrak{G}_\varrho(\widehat{\Phi}_\varrho^{n-1}), 1)^h + 3c_3\delta t (\mathfrak{G}_\varrho(\widehat{\Psi}_\varrho^{n-1}), 1)^h + [1 + 3\delta t d_3](\mathfrak{G}_\varrho(\widehat{\Theta}_\varrho^{n-1}), 1)^h \\
 & + \delta t C(|\mathfrak{D}|, a_3).
 \end{aligned} \tag{84}$$

By multiplying (5) by $\frac{(\theta_{21})^2}{\theta_{22}\theta_{11}}$ and (5) by $\frac{(\theta_{23})^2}{\theta_{22}\theta_{33}}$, and then combining the results with (83), we obtain the final result (78). $\square$

Lemma 7. Under the assumptions of Lemma 6, suppose $u^0, v^0, w^0 \in L^\infty(\mathfrak{D})$, where $u^0, v^0, w^0 \geq 0$ for almost every $x \in \mathfrak{D}$. Let $\widehat{\Phi}_\varrho^0 \equiv P^h u^0$, $\widehat{\Psi}_\varrho^0 \equiv P^h v^0$, $\widehat{\Theta}_\varrho^0 \equiv P^h w^0$ or $\widehat{\Phi}_\varrho^0 \equiv \pi^h u^0$, $\widehat{\Psi}_\varrho^0 \equiv \pi^h v^0$, $\widehat{\Theta}_\varrho^0 \equiv \pi^h w^0$ if $u^0, v^0, w^0 \in C(\mathfrak{D})$. Then, for any $\varrho \in (0, e^{-1})$, $h > 0$, and $\delta t \leq \frac{1-\delta}{2\widehat{a}+\widehat{b}+\widehat{c}+\widehat{d}}$, where $\delta \in (0, 1)$, $\widehat{a} = \max\{a_1, a_2, a_3\}$, $\widehat{b} = \max\{b_1, b_2, b_3\}$, $\widehat{c} = \max\{c_1, c_2, c_3\}$, $\widehat{d} = \max\{d_1, d_2, d_3\}$, the problem $(Q_\varrho^{h,\delta t})$ has a solution $\widehat{\Phi}_\varrho^n, \widehat{\Psi}_\varrho^n, \widehat{\Theta}_\varrho^n$, $n = 1, \dots, N$, satisfying

$$\begin{aligned}
 & \max_{n=1, \dots, N} \left[\left(\frac{(\theta_{21})^2}{\theta_{22}\theta_{11}} \mathfrak{G}_\varrho(\widehat{\Phi}_\varrho^n), 1 \right)^h + (\mathfrak{G}_\varrho(\widehat{\Psi}_\varrho^n), 1)^h \right. \\
 & + \left(\frac{(\theta_{23})^2}{\theta_{22}\theta_{33}} \mathfrak{G}_\varrho(\widehat{\Theta}_\varrho^n), 1 \right)^h + \varrho^{-1} \|\pi^h[\widehat{\Phi}_\varrho^n] - \|\|_0^2 + \varrho^{-1} \|\pi^h[\widehat{\Psi}_\varrho^n] - \|\|_0^2 \\
 & + \varrho^{-1} \|\pi^h[\widehat{\Theta}_\varrho^n] - \|\|_0^2 + \|\widehat{\Phi}_\varrho^n\|_{0,1}^2 + \|\widehat{\Psi}_\varrho^n\|_{0,1}^2 + \|\widehat{\Theta}_\varrho^n\|_{0,1}^2 \Big] \\
 & + \sum_{n=1}^N \delta t \left[\frac{3(\theta_{21})^2}{2\theta_{22}} |\widehat{\Phi}_\varrho^n|_1^2 + \theta_{22} |\widehat{\Psi}_\varrho^n|_1^2 + \frac{3(\theta_{23})^2}{2\theta_{22}} |\widehat{\Theta}_\varrho^n|_1^2 \right] \leq C.
 \end{aligned} \tag{85}$$

Proof. Based on (78) and our assumptions regarding δt , we can conclude that for $n = 1, \dots, N$ and $i = 1, 2, 3$, the following holds:

$$\begin{aligned}
& [1 - 2\delta t\hat{a} - \delta t\hat{b} - \delta t\hat{c} - \delta t\hat{d}] \left(\frac{(\theta_{21})^2}{\theta_{22}\theta_{11}} \mathfrak{G}_\rho(\hat{\Phi}_\rho^n) + \mathfrak{G}_\rho(\hat{\Psi}_\rho^n) + \frac{(\theta_{23})^2}{\theta_{22}\theta_{33}} \mathfrak{G}_\rho(\hat{\Theta}_\rho^n), 1 \right)^h \\
& + \frac{\delta t\theta_{11}(\theta_{21})^2}{\theta_{22}\theta_{11}} \int_{\mathcal{D}} \frac{|\nabla \hat{\Phi}_\rho^n|^2}{\xi_\rho(\hat{\Phi}_\rho^n)} d\mathbf{x} + \frac{3\delta t(\theta_{21})^2}{2\theta_{22}} |\hat{\Phi}_\rho^n|_1^2 + \delta t\theta_2 \int_{\mathcal{D}} \frac{|\nabla \hat{\Psi}_\rho^n|^2}{\xi_\rho(\hat{\Psi}_\rho^n)} d\mathbf{x} \\
& + \delta t\theta_{21} \int_{\mathcal{D}} |\nabla \hat{\Psi}_\rho^n|^2 \frac{\Xi_\rho(\hat{\Phi}_\rho^n)}{\Xi_\rho(\hat{\Psi}_\rho^n)} d\mathbf{x} + \delta t\theta_{22} |\hat{\Psi}_\rho^n|_1^2 + \delta t\theta_{23} \int_{\mathcal{D}} |\nabla \hat{\Psi}_\rho^n|^2 \frac{\Xi_\rho(\hat{\Theta}_\rho^n)}{\Xi_\rho(\hat{\Psi}_\rho^n)} d\mathbf{x} \\
& + \frac{\delta t\theta_3(\theta_{23})^2}{\theta_{22}\theta_{33}} \int_{\mathcal{D}} \frac{|\nabla \hat{\Theta}_\rho^n|^2}{\xi_\rho(\hat{\Theta}_\rho^n)} d\mathbf{x} + \frac{3\delta t(\theta_{23})^2}{2\theta_{22}} |\hat{\Theta}_\rho^n|_1^2 \\
& \leq \left[\frac{(\theta_{21})^2}{\theta_{22}\theta_{11}} + \frac{3\delta t b_1(\theta_{21})^2}{\theta_{22}\theta_{11}} + 3b_2\delta t + \frac{3b_3\delta t(\theta_{23})^2}{\theta_{22}\theta_{33}} \right] (\mathfrak{G}_\rho(\hat{\Phi}_\rho^{n-1}), 1)^h \\
& + \left[\frac{3c_1\delta t(\theta_{21})^2}{\theta_{22}\theta_{11}} + 1 + 3c_2\delta t + \frac{3c_3\delta t(\theta_{23})^2}{\theta_{22}\theta_{33}} \right] (\mathfrak{G}_\rho(\hat{\Psi}_\rho^{n-1}), 1)^h \\
& + \left[\frac{3d_1\delta t(\theta_{21})^2}{\theta_{22}\theta_{11}} + 3d_2\delta t + \frac{(\theta_{23})^2}{\theta_{22}\theta_{33}} + \frac{3\delta t d_3(\theta_{23})^2}{\theta_{22}\theta_{33}} \right] (\mathfrak{G}_\rho(\hat{\Theta}_\rho^{n-1}), 1)^h + C. \tag{86}
\end{aligned}$$

Based on our assumptions regarding the initial data u^0 , v^0 , w^0 , and (21), we can conclude that

$$(\mathfrak{G}_\rho(\hat{\Phi}_\rho^0), 1)^h \leq C, \quad (\mathfrak{G}_\rho(\hat{\Psi}_\rho^0), 1)^h \leq C \quad \text{and} \quad (\mathfrak{G}_\rho(\hat{\Theta}_\rho^0), 1)^h \leq C. \tag{87}$$

Therefore, it can be inferred from (86) and (87) that

$$\begin{aligned}
& \left(\frac{(\theta_{21})^2}{\theta_{22}\theta_{11}} \mathfrak{G}_\rho(\hat{\Phi}_\rho^n) + \mathfrak{G}_\rho(\hat{\Psi}_\rho^n) + \frac{(\theta_{23})^2}{\theta_{22}\theta_{33}} \mathfrak{G}_\rho(\hat{\Theta}_\rho^n), 1 \right)^h \\
& \leq C \left(\frac{1}{1 - (2\hat{a} + \hat{b} + \hat{c} + \hat{d})\delta t} \right) \left[\frac{(\theta_{21})^2}{\theta_{22}\theta_{11}} (\mathfrak{G}_\rho(\hat{\Phi}_\rho^{n-1}), 1)^h \right. \\
& \quad \left. + (\mathfrak{G}_\rho(\hat{\Psi}_\rho^{n-1}), 1)^h + \frac{(\theta_{23})^2}{\theta_{22}\theta_{33}} (\mathfrak{G}_\rho(\hat{\Theta}_\rho^{n-1}), 1)^h + 1 \right] \\
& \leq C \left(1 + \frac{(2\hat{a} + \hat{b} + \hat{c} + \hat{d})\delta t}{1 - (2\hat{a} + \hat{b} + \hat{c} + \hat{d})\delta t} \right) \left[\frac{(\theta_{21})^2}{\theta_{22}\theta_{11}} (\mathfrak{G}_\rho(\hat{\Phi}_\rho^{n-1}), 1)^h \right. \\
& \quad \left. + (\mathfrak{G}_\rho(\hat{\Psi}_\rho^{n-1}), 1)^h + \frac{(\theta_{23})^2}{\theta_{22}\theta_{33}} (\mathfrak{G}_\rho(\hat{\Theta}_\rho^{n-1}), 1)^h + 1 \right] \\
& \leq C \left(1 + \frac{(2\hat{a} + \hat{b} + \hat{c} + \hat{d})\delta t}{\delta} \right) \left[\frac{(\theta_{21})^2}{\theta_{22}\theta_{11}} (\mathfrak{G}_\rho(\hat{\Phi}_\rho^{n-1}), 1)^h \right. \\
& \quad \left. + (\mathfrak{G}_\rho(\hat{\Psi}_\rho^{n-1}), 1)^h + \frac{(\theta_{23})^2}{\theta_{22}\theta_{33}} (\mathfrak{G}_\rho(\hat{\Theta}_\rho^{n-1}), 1)^h + 1 \right]
\end{aligned}$$

$$\begin{aligned}
&\leq C e^{\frac{(2\hat{a}+\hat{b}+\hat{c}+\hat{d})\delta t}{\delta}} \left[\frac{(\theta_{21})^2}{\theta_{22}\theta_{11}} (\mathfrak{G}_\varrho(\hat{\Phi}_\varrho^{n-1}), 1)^h \right. \\
&\quad \left. + (\mathfrak{G}_\varrho(\hat{\Psi}_\varrho^{n-1}), 1)^h + \frac{(\theta_{23})^2}{\theta_{22}\theta_{33}} (\mathfrak{G}_\varrho(\hat{\Theta}_\varrho^{n-1}), 1)^h + 1 \right] \\
&\vdots \\
&\leq C e^{T \frac{(2\hat{a}+\hat{b}+\hat{c}+\hat{d})}{\delta}} \left[\frac{(\theta_{21})^2}{\theta_{22}\theta_{11}} (\mathfrak{G}_\varrho(\hat{\Phi}_\varrho^0), 1)^h + (\mathfrak{G}_\varrho(\hat{\Psi}_\varrho^0), 1)^h \right. \\
&\quad \left. + \frac{(\theta_{23})^2}{\theta_{22}\theta_{33}} (\mathfrak{G}_\varrho(\hat{\Theta}_\varrho^0), 1)^h + 1 \right] \leq C, \quad \text{for all } n \geq 1, \quad T = n \delta t. \tag{88}
\end{aligned}$$

By setting $\omega \equiv 1$ in (58) and considering the positivity of $\xi_\varrho(\gamma)$ and the given assumptions on δt , we obtain, for $n = 1, \dots, N$, the following results:

$$(\hat{\Phi}_\varrho^n, 1)^h \leq \frac{1}{1 - a_1 \delta t} (\hat{\Phi}_\varrho^{n-1}, 1)^h \leq (1 + \frac{a_1 \delta t}{\delta_1}) (\hat{\Phi}_\varrho^{n-1}, 1)^h \leq e^{\frac{a_1 \delta t}{\delta_1}} (\hat{\Phi}_\varrho^{n-1}, 1)^h. \tag{89}$$

From (89), the definition of the interpolation π^h , (69), and the assumption on u^0 , it can be inferred that

$$\max_{n=1, \dots, N} (\hat{\Phi}_\varrho^n, 1)^h \leq e^{\frac{a_1 T}{\delta_1}} (\hat{\Phi}_\varrho^0, 1)^h \leq |\mathfrak{D}| e^{\frac{a_1 T}{\delta_1}} \|u^0\|_{0, \infty} \leq C. \tag{90}$$

For $n = 1, \dots, N$, using the given equation $|s| = s - 2[s]_-$, along with (90) and Young's inequality, we can derive the following:

$$\begin{aligned}
\|\hat{\Phi}_\varrho^n\|_{0,1} &= (|\hat{\Phi}_\varrho^n|, 1) \leq (\pi^h |\hat{\Phi}_\varrho^n|, 1) \leq (\hat{\Phi}_\varrho^n - 2\pi^h [\hat{\Phi}_\varrho^n]_-, 1) \\
&= (\hat{\Phi}_\varrho^n, 1)^h - 2(\pi^h [\hat{\Phi}_\varrho^n]_-, 1) \leq C(1 + \|\pi[\hat{\Phi}_\varrho^n]_-\|_0^2). \tag{91}
\end{aligned}$$

From (46), (45), (25), and (88), it can be deduced, after recalling the expressions $\gamma = [\gamma]_+ + [\gamma]_-$ and $\mathfrak{G}_\varrho(\gamma) \geq 0$ for $n = 1, \dots, N$, that

$$\|\pi^h [\hat{\Phi}_\varrho^n]_-\|_0^2 \leq |\pi^h [\hat{\Phi}_\varrho^n]_-|_h^2 = ([\hat{\Phi}_\varrho^n]_-^2, 1)^h \leq 2 \varrho (\mathfrak{G}_\varrho(\hat{\Phi}_\varrho^n), 1)^h \leq C \varrho. \tag{92}$$

The other bounds, namely $\|\pi^h [\hat{\Psi}_\varrho^n]_-\|_0^2$, $\|\pi^h [\hat{\Theta}_\varrho^n]_-\|_0^2$, $\|\hat{\Psi}_\varrho^n\|_{0,1}$, and $\|\hat{\Theta}_\varrho^n\|_{0,1}$, can be obtained using the same methodology that has used to establish the above bounds. The bounds (1-9) in (85) have now been derived, and the last three bounds in (85) are obtained by summing (78) over n , with the assistance of (87). $\square$

Theorem 2. Under the assumptions of Lemma 7, the solution $\hat{\Phi}_\varrho^n, \hat{\Psi}_\varrho^n, \hat{\Theta}_\varrho^n$ to problem $(Q_\varrho^{h, \delta t})$ meets

$$\begin{aligned}
& \sum_{n=1}^N \delta t \left[\|\widehat{\Phi}_\varrho^n\|_{0,\sigma}^\sigma + \|\xi_\varrho(\widehat{\Phi}_\varrho^n)\|_{0,\sigma}^\sigma + \|\pi^h \xi_\varrho(\widehat{\Phi}_\varrho^n)\|_{0,\sigma}^\sigma + \|\Xi_\varrho(\widehat{\Phi}_\varrho^n)\|_{0,\sigma}^\sigma + \|\widehat{\Psi}_\varrho^n\|_{0,\sigma}^\sigma \right. \\
& + \|\xi_\varrho(\widehat{\Psi}_\varrho^n)\|_{0,\sigma}^\sigma + \|\pi^h \xi_\varrho(\widehat{\Psi}_\varrho^n)\|_{0,\sigma}^\sigma + \|\Xi_\varrho(\widehat{\Psi}_\varrho^n)\|_{0,\sigma}^\sigma + \|\widehat{\Theta}_\varrho^n\|_{0,\sigma}^\sigma + \|\xi_\varrho(\widehat{\Theta}_\varrho^n)\|_{0,\sigma}^\sigma \\
& + \|\pi^h \xi_\varrho(\widehat{\Theta}_\varrho^n)\|_{0,\sigma}^\sigma + \|\Xi_\varrho(\widehat{\Theta}_\varrho^n)\|_{0,\sigma}^\sigma \left. \right] + \sum_{n=1}^N \delta t \left[\left\| \frac{\widehat{\Phi}_\varrho^n - \widehat{\Phi}_\varrho^{n-1}}{\delta t} \right\|_{(W^{1,q'}(\mathcal{D}))'}^2 \right. \\
& + \left\| \frac{\widehat{\Psi}_\varrho^n - \widehat{\Psi}_\varrho^{n-1}}{\delta t} \right\|_{(W^{1,q'}(\mathcal{D}))'}^2 + \left\| \frac{\widehat{\Theta}_\varrho^n - \widehat{\Theta}_\varrho^{n-1}}{\delta t} \right\|_{(W^{1,q'}(\mathcal{D}))'}^2 \left. \right] \\
& + \sum_{n=1}^N \delta t \left[\left\| \mathfrak{G}_q \left[\frac{\widehat{\Phi}_\varrho^n - \widehat{\Phi}_\varrho^{n-1}}{\delta t} \right] \right\|_{1,q}^q + \left\| \mathfrak{G}_q \left[\frac{\widehat{\Psi}_\varrho^n - \widehat{\Psi}_\varrho^{n-1}}{\delta t} \right] \right\|_{1,q}^q \right. \\
& + \left. \left\| \mathfrak{G}_q \left[\frac{\widehat{\Theta}_\varrho^n - \widehat{\Theta}_\varrho^{n-1}}{\delta t} \right] \right\|_{1,q}^q \right] \leq C, \tag{93}
\end{aligned}$$

where $\sigma = \frac{2(d+1)}{d}$, $q = \frac{2(d+1)}{2d+1}$, $q' = 2(d+1)$.

Proof. For $n = 1, \dots, N$, the relation (9) and the seventh bound in (85) imply that

$$\|\widehat{\Phi}_\varrho^n\|_{0,\sigma}^\sigma \leq \|\widehat{\Phi}_\varrho^n\|_{0,1}^{\sigma-2} \|\widehat{\Phi}_\varrho^n\|_1^2 \leq C \|\widehat{\Phi}_\varrho^n\|_1^2, \tag{94}$$

where $\sigma(\frac{2d(\sigma-1)}{\sigma(d+2)}) = 2$; that is $\sigma = \frac{2(d+1)}{d}$.

Given the inequality $\xi_\varrho(s) \leq |s| + \varrho$, equation (94) implies, for $n = 1, \dots, N$, the following equation:

$$\|\xi_\varrho(\widehat{\Phi}_\varrho^n)\|_{0,\sigma}^\sigma \leq C + \|\widehat{\Phi}_\varrho^n\|_{0,\sigma}^\sigma \leq C(1 + \|\widehat{\Phi}_\varrho^n\|_1^2). \tag{95}$$

Based on $\xi_\varrho(s) \leq [s]_+ + \varrho = s - [s]_- + \varrho$, the Young's inequality, and the fourth and seventh bounds in (85), we can conclude, for $n = 1, \dots, N$, that

$$\begin{aligned}
\|\pi^h \xi_\varrho(\widehat{\Phi}_\varrho^n)\|_{0,1} & \leq \int_{\mathcal{D}} (\widehat{\Phi}_\varrho^n - \pi^h [\widehat{\Phi}_\varrho^n]_- + \varrho) d\mathbf{x} \\
& \leq C(1 + \|\widehat{\Phi}_\varrho^n\|_{0,1} + \|\pi^h [\widehat{\Phi}_\varrho^n]_-\|_0^2) \leq C,
\end{aligned} \tag{96}$$

and

$$\|\pi^h \xi_\varrho(\widehat{\Phi}_\varrho^n)\|_0^2 \leq C(1 + \|\widehat{\Phi}_\varrho^n\|_0^2 + \|\pi^h [\widehat{\Phi}_\varrho^n]_-\|_0^2) \leq C(1 + \|\widehat{\Phi}_\varrho^n\|_0^2). \tag{97}$$

For $n = 1, \dots, N$, equations (9), (96), (97), and (63) yield the following result:

$$\|\pi^h \xi_\varrho(\widehat{\Phi}_\varrho^n)\|_{0,\sigma}^\sigma \leq C \|\pi^h \xi_\varrho(\widehat{\Phi}_\varrho^n)\|_{0,1}^{\sigma-2} \|\pi^h \xi_\varrho(\widehat{\Phi}_\varrho^n)\|_1^2 \leq C(1 + \|\widehat{\Phi}_\varrho^n\|_1^2). \tag{98}$$

The bounds $1 \rightarrow 3$ in (93) can be derived from (85) and noting (94)–(98).

The fourth bound in (93) can be derived from the third bound in (93) by utilizing (56), (57), (24), and (48) in the form

$$\begin{aligned} \|\Xi_\varrho(\widehat{\Phi}_\varrho^n)\|_{0,\sigma}^\sigma &:= \int_\Omega \|\Xi_\varrho(\widehat{\Phi}_\varrho^n)\|^\sigma d\mathbf{x} = \sum_{\tilde{\tau} \in \mathcal{T}^h} \int_{\tilde{\tau}} \|\Xi_\varrho(\widehat{\Phi}_\varrho^n)\|^\sigma d\mathbf{x} \\ &\leq C \sum_{\tilde{\tau} \in \mathcal{T}^h} h_{\tilde{\tau}}^d \|\pi^h \xi_\varrho(\widehat{\Phi}_\varrho^n)\|_{0,\infty,\tilde{\tau}}^\sigma \leq C \sum_{\tilde{\tau} \in \mathcal{T}^h} \|\pi^h \xi_\varrho(\widehat{\Phi}_\varrho^n)\|_{0,\sigma,\tilde{\tau}}^\sigma \\ &\leq C \|\pi^h \xi_\varrho(\widehat{\Phi}_\varrho^n)\|_{0,\sigma}^\sigma. \end{aligned} \quad (99)$$

Based on the definition of π^h and (69), as well as the assumption of u^0 , v^0 , and w^0 , it follows that

$$\|\widehat{\Phi}_\varrho^0\|_0 + \|\widehat{\Psi}_\varrho^0\|_0 + \|\widehat{\Theta}_\varrho^0\|_0 \leq C(\|u^0\|_{0,\infty} + \|v^0\|_{0,\infty} + \|w^0\|_{0,\infty}) \leq C. \quad (100)$$

From (70), we can conclude that

$$\begin{aligned} (\Xi_\varrho(\widehat{\Phi}_\varrho^n) \nabla \widehat{\Phi}_\varrho^n, \nabla P^h \Sigma) &\leq \|\Xi_\varrho(\widehat{\Phi}_\varrho^n)\|_{0,\sigma} |\widehat{\Phi}_\varrho^n|_1 |P^h \Sigma|_{1,q'} \\ &\leq C \|\Xi_\varrho(\widehat{\Phi}_\varrho^n)\|_{0,\sigma} |\widehat{\Phi}_\varrho^n|_1 [|(I - P^h) \Sigma|_{1,q'} + |\Sigma|_{1,q'}] \\ &\leq C \|\Xi_\varrho(\widehat{\Phi}_\varrho^n)\|_{0,\sigma} |\widehat{\Phi}_\varrho^n|_1 \|\Sigma\|_{1,q'}, \end{aligned} \quad (101)$$

where $\frac{1}{2} + \frac{1}{\sigma} + \frac{1}{q'} = 1$, that is $\frac{1}{q} + \frac{1}{q'} = 1$ and $q = \frac{2(d+1)}{2d+1}$. From (45), (46), (97), and (70), we obtain the following results for any $\Sigma \in W^{1,q'}(\Omega)$ and $n = 1, \dots, N$:

$$\begin{aligned} &(\xi_\varrho(\widehat{\Phi}_\varrho^n) \xi_\varrho(\widehat{\Phi}_\varrho^{n-1}), P^h \Sigma)^h \\ &\leq C \|\pi^h \xi_\varrho(\widehat{\Phi}_\varrho^n)\|_{0,\sigma} \|\pi^h \xi_\varrho(\widehat{\Phi}_\varrho^{n-1})\|_0 \|P^h \Sigma\|_{0,q'} \\ &\leq C \|\pi^h \xi_\varrho(\widehat{\Phi}_\varrho^n)\|_{0,\sigma} (1 + \|\widehat{\Phi}_\varrho^{n-1}\|_0) [\|(I - P^h) \Sigma\|_{0,q'} + \|\Sigma\|_{0,q'}] \\ &\leq C \|\pi^h \xi_\varrho(\widehat{\Phi}_\varrho^n)\|_{0,\sigma} (1 + \|\widehat{\Phi}_\varrho^{n-1}\|_0) \|\Sigma\|_{1,q'}. \end{aligned} \quad (102)$$

Similarly to (102), we can derive that

$$(\xi_\varrho(\widehat{\Phi}_\varrho^n) \xi_\varrho(\widehat{\Psi}_\varrho^{n-1}), P^h \Sigma)^h \leq C \|\pi^h \xi_\varrho(\widehat{\Phi}_\varrho^n)\|_{0,\sigma} (1 + \|\widehat{\Psi}_\varrho^{n-1}\|_0) \|\Sigma\|_{1,q'}, \quad (103)$$

and

$$(\xi_\varrho(\widehat{\Phi}_\varrho^n) \xi_\varrho(\widehat{\Theta}_\varrho^{n-1}), P^h \Sigma)^h \leq C \|\pi^h \xi_\varrho(\widehat{\Phi}_\varrho^n)\|_{0,\sigma} (1 + \|\widehat{\Theta}_\varrho^{n-1}\|_0) \|\Sigma\|_{1,q'}. \quad (104)$$

Using (76), (68), (58), (70), (101), (102), (103), and (104), we can determine that for any $\Sigma \in W^{1,q'}(\mathfrak{D})$ and $n = 1, \dots, N$, the following holds:

$$\begin{aligned}
 \left\langle \frac{\widehat{\Phi}_\varrho^n - \widehat{\Phi}_\varrho^{n-1}}{\delta t}, \Sigma \right\rangle_{q'} &= \left(\frac{\widehat{\Phi}_\varrho^n - \widehat{\Phi}_\varrho^{n-1}}{\delta t}, \Sigma \right) = \left(\frac{\widehat{\Phi}_\varrho^n - \widehat{\Phi}_\varrho^{n-1}}{\delta t}, P^h \Sigma \right)^h \\
 &= (a_1 \widehat{\Phi}_\varrho^n - b_1 \xi_\varrho(\widehat{\Phi}_\varrho^n) \xi_\varrho(\widehat{\Phi}_\varrho^{n-1}) - c_1 \xi_\varrho(\widehat{\Phi}_\varrho^n) \xi_\varrho(\widehat{\Psi}_\varrho^{n-1}) \\
 &\quad - d_1 \xi_\varrho(\widehat{\Phi}_\varrho^n) \xi_\varrho(\widehat{\Theta}_\varrho^{n-1}), P^h \Sigma)^h \\
 &\quad - (\nabla \widehat{\Phi}_\varrho^n + \Xi_\varrho(\widehat{\Phi}_\varrho^n) \nabla \widehat{\Phi}_\varrho^n, \nabla P^h \Sigma) \\
 &\leq C(\|\widehat{\Phi}_\varrho^n\|_1 + \|\widehat{\Phi}_\varrho^{n-1}\|_0 + \|\widehat{\Psi}_\varrho^{n-1}\|_0 + \|\widehat{\Theta}_\varrho^{n-1}\|_0 + 1) \\
 &\quad \times (\|\pi^h \xi_\varrho(\widehat{\Phi}_\varrho^n)\|_{0,\sigma} + \|\Xi_\varrho(\widehat{\Phi}_\varrho^n)\|_{0,\sigma} + 1) \|\Sigma\|_{1,q'}, \tag{105}
 \end{aligned}$$

where we have utilized the continuous embedding of $W^{1,q'}(\mathfrak{D}) \hookrightarrow L^2(\mathfrak{D})$ and $W^{1,q'}(\mathfrak{D}) \hookrightarrow H^1(\mathfrak{D})$. Hence, we can conclude that

$$\left\| \frac{\widehat{\Phi}_\varrho^n - \widehat{\Phi}_\varrho^{n-1}}{\delta t} \right\|_{(W^{1,q'}(\mathfrak{D}))'} \leq C A_n B_n, \tag{106}$$

where

$$\begin{aligned}
 A_n &= (\|\widehat{\Phi}_\varrho^n\|_1 + \|\widehat{\Phi}_\varrho^{n-1}\|_0 + \|\widehat{\Psi}_\varrho^{n-1}\|_0 + \|\widehat{\Theta}_\varrho^{n-1}\|_0 + 1), \\
 B_n &= (\|\pi^h \xi_\varrho(\widehat{\Phi}_\varrho^n)\|_{0,\sigma} + \|\Xi_\varrho(\widehat{\Phi}_\varrho^n)\|_{0,\sigma} + 1).
 \end{aligned}$$

Based on (106), the Cauchy–Schwarz inequality, the third and fourth bounds in (93), and the assumption on the discretization and (100), we can conclude that

$$\begin{aligned}
 \sum_{n=1}^N \delta t_n \left\| \frac{\widehat{\Phi}_\varrho^n - \widehat{\Phi}_\varrho^{n-1}}{\delta t_n} \right\|_{(W^{1,q'}(\mathfrak{D}))'}^q &\leq C \sum_{n=1}^N \delta t_n A_n^q B_n^q \\
 &\leq C \left[\sum_{n=1}^N \delta t_n A_n^2 \right]^{\frac{q}{2}} \left[\sum_{n=1}^N \delta t_n B_n^\sigma \right]^{\frac{q}{\sigma}} \leq C, \tag{107}
 \end{aligned}$$

where $\frac{1}{q} + \frac{1}{q'} = 1$ and $q = \frac{2(d+1)}{2d+1}$. The final three results in (93) can be obtained from the thirteenth, fourteenth, and fifteenth results in (93), respectively, by considering the implications of (74). $\square$

6 Uniqueness

To guarantee the stability and dependability of computational models, the uniqueness of numerical solutions for PDEs utilizing FEM must be investigated. Establishing uniqueness ensures that the numerical approximation depicts a single, physically significant answer as opposed to several potential possibilities. In scientific and technical applications, where precise predictions of physical behavior rely on consistent and well-posed mathematical formulations, this aids in the validation of simulations.

Theorem 3. Assuming the conditions of Theorem 2 are met, and $\widehat{\Phi}_\varrho^n, \widehat{\Psi}_\varrho^n, \widehat{\Theta}_\varrho^n$, $n = 1, \dots, N$, is a solution of $(Q_\varrho^{h,\delta t})$. If $\delta t \in (0, \varpi)$ and

$$\varpi < \max\left\{\frac{1}{e_1 + e_4 + e_9}, \frac{1}{e_2 + e_5 + e_7 + e_8}, \frac{1}{e_3 + e_6 + e_{10}}\right\}, \quad (108)$$

where the values of the parameters $e_i, i = 1, \dots, 6$, were stated in the proof, then the solution $\{\widehat{\Phi}_\varrho^n, \widehat{\Psi}_\varrho^n, \widehat{\Theta}_\varrho^n\}$, $n = 1, \dots, N$, is unique.

Proof. The above theorems are proved by employing an approach similar to that used in [32, 30, 2]. □

7 Existence of a weak solution

Since many physical and engineering issues involve functions that might not have classical derivatives, it is crucial to investigate whether a weak solution exists for PDEs. These issues can be stated and examined within a more comprehensive functional framework thanks to the weak formulation, which guarantees that solutions are found even in cases when classical ones are not.

We commence by employing the following definitions:

$$\widehat{\Phi}_\varrho(t) := \left(\frac{t - t_{n-1}}{\delta t_n}\right)\widehat{\Phi}_\varrho^n + \left(\frac{t_n - t}{\delta t_n}\right)\widehat{\Phi}_\varrho^{n-1}, \quad t \in [t_{n-1}, t_n], \quad n \geq 1, \quad (109)$$

$$\widehat{\Psi}_\varrho(t) := \left(\frac{t - t_{n-1}}{\delta t_n}\right)\widehat{\Psi}_\varrho^n + \left(\frac{t_n - t}{\delta t_n}\right)\widehat{\Psi}_\varrho^{n-1}, \quad t \in [t_{n-1}, t_n], \quad n \geq 1, \quad (110)$$

$$\widehat{\Theta}_\varrho(t) := \left(\frac{t - t_{n-1}}{\delta t_n}\right)\widehat{\Theta}_\varrho^n + \left(\frac{t_n - t}{\delta t_n}\right)\widehat{\Theta}_\varrho^{n-1}, \quad t \in [t_{n-1}, t_n], \quad n \geq 1, \quad (111)$$

and

$$\widehat{\Phi}_\varrho^+(t) := \widehat{\Phi}_\varrho^n, \quad \widehat{\Phi}_\varrho^-(t) := \widehat{\Phi}_\varrho^{n-1}, \quad t \in (t_{n-1}, t_n], \quad n \geq 1, \quad (112)$$

$$\widehat{\Psi}_{\varrho}^{+}(t) := \widehat{\Psi}_{\varrho}^n, \quad \widehat{\Psi}_{\varrho}^{-}(t) := \widehat{\Psi}_{\varrho}^{n-1}, \quad t \in (t_{n-1}, t_n], \quad n \geq 1, \quad (113)$$

$$\widehat{\Theta}_{\varrho}^{+}(t) := \widehat{\Theta}_{\varrho}^n, \quad \widehat{\Theta}_{\varrho}^{-}(t) := \widehat{\Theta}_{\varrho}^{n-1}, \quad t \in (t_{n-1}, t_n], \quad n \geq 1. \quad (114)$$

By considering (109), (110), (112), and (113), we can conclude that

$$\frac{\partial \widehat{\Phi}_{\varrho}}{\partial t} = \frac{\widehat{\Phi}_{\varrho}^{+} - \widehat{\Phi}_{\varrho}^{-}}{\delta t} = \frac{\widehat{\Phi}_{\varrho}^{+} - \widehat{\Phi}_{\varrho}}{t_n - t} = \frac{\widehat{\Phi}_{\varrho}^{-} - \widehat{\Phi}_{\varrho}}{t - t_{n-1}}, \quad t \in (t_{n-1}, t_n), \quad n \geq 1, \quad (115)$$

$$\frac{\partial \widehat{\Psi}_{\varrho}}{\partial t} = \frac{\widehat{\Psi}_{\varrho}^{+} - \widehat{\Psi}_{\varrho}^{-}}{\delta t} = \frac{\widehat{\Psi}_{\varrho}^{+} - \widehat{\Psi}_{\varrho}}{t_n - t} = \frac{\widehat{\Psi}_{\varrho}^{-} - \widehat{\Psi}_{\varrho}}{t - t_{n-1}}, \quad t \in (t_{n-1}, t_n), \quad n \geq 1, \quad (116)$$

$$\frac{\partial \widehat{\Theta}_{\varrho}}{\partial t} = \frac{\widehat{\Theta}_{\varrho}^{+} - \widehat{\Theta}_{\varrho}^{-}}{\delta t} = \frac{\widehat{\Theta}_{\varrho}^{+} - \widehat{\Theta}_{\varrho}}{t_n - t} = \frac{\widehat{\Theta}_{\varrho}^{-} - \widehat{\Theta}_{\varrho}}{t - t_{n-1}}, \quad t \in (t_{n-1}, t_n), \quad n \geq 1. \quad (117)$$

Using the definitions provided above, we can rewrite the problem $(Q_{\varrho}^{h,\delta t})$ as follows: Find $\{\widehat{\Phi}_{\varrho}, \widehat{\Psi}_{\varrho}, \widehat{\Theta}_{\varrho}\} \in C([0, T]; X^h) \times C([0, T]; X^h \times C([0, T]; X^h))$ so that for all $\omega \in L^2(0, T; X^h)$,

$$\begin{aligned} & \int_0^T \left[\left(\frac{\partial \widehat{\Phi}_{\varrho}}{\partial t}, \omega \right)^h + (\theta_1 \nabla \widehat{\Phi}_{\varrho}^{+} + 2\theta_{11} \Xi_{\varrho}(\widehat{\Phi}_{\varrho}^{+}) \nabla \widehat{\Phi}_{\varrho}^{+}, \nabla \omega) \right] dt \\ &= \int_0^T \left[(a_1 \widehat{\Phi}_{\varrho}^{+} - b_1 \xi_{\varrho}(\widehat{\Phi}_{\varrho}^{+}) \xi_{\varrho}(\widehat{\Phi}_{\varrho}^{-}) - c_1 \xi_{\varrho}(\widehat{\Phi}_{\varrho}^{+}) \xi_{\varrho}(\widehat{\Psi}_{\varrho}^{-}) - d_1 \xi_{\varrho}(\widehat{\Phi}_{\varrho}^{+}) \xi_{\varrho}(\widehat{\Theta}_{\varrho}^{-}), \omega)^h \right] dt, \end{aligned} \quad (118)$$

$$\begin{aligned} & \int_0^T \left[\left(\frac{\partial \widehat{\Psi}_{\varrho}}{\partial t}, \omega \right)^h + (\theta_2 \nabla \widehat{\Psi}_{\varrho}^{+} + \theta_{21} \Xi_{\varrho}(\widehat{\Psi}_{\varrho}^{+}) \nabla \widehat{\Phi}_{\varrho}^{+} + \theta_{21} \Xi_{\varrho}(\widehat{\Phi}_{\varrho}^{+}) \nabla \widehat{\Psi}_{\varrho}^{+} \right. \\ & \quad \left. + 2\theta_{22} \Xi_{\varrho}(\widehat{\Psi}_{\varrho}^{+}) \nabla \widehat{\Psi}_{\varrho}^{+} + \theta_{23} \Xi_{\varrho}(\widehat{\Psi}_{\varrho}^{+}) \nabla \widehat{\Theta}_{\varrho}^{+} + \theta_{23} \Xi_{\varrho}(\widehat{\Theta}_{\varrho}^{+}) \nabla \widehat{\Psi}_{\varrho}^{+}, \nabla \omega) \right] dt \\ &= \int_0^T \left[(a_2 \widehat{\Psi}_{\varrho}^{+} - b_2 \xi_{\varrho}(\widehat{\Psi}_{\varrho}^{+}) \xi_{\varrho}(\widehat{\Phi}_{\varrho}^{-}) - c_2 \xi_{\varrho}(\widehat{\Psi}_{\varrho}^{+}) \xi_{\varrho}(\widehat{\Psi}_{\varrho}^{-}) - d_2 \xi_{\varrho}(\widehat{\Psi}_{\varrho}^{+}) \xi_{\varrho}(\widehat{\Theta}_{\varrho}^{-}), \omega)^h \right] dt, \end{aligned} \quad (119)$$

and

$$\begin{aligned} & \int_0^T \left[\left(\frac{\partial \widehat{\Theta}_{\varrho}}{\partial t}, \omega \right)^h + (\theta_3 \nabla \widehat{\Theta}_{\varrho}^{+} + 2\theta_{33} \Xi_{\varrho}(\widehat{\Theta}_{\varrho}^{+}) \nabla \widehat{\Theta}_{\varrho}^{+}, \nabla \omega) \right] dt \\ &= \int_0^T \left[(a_3 \widehat{\Theta}_{\varrho}^{+} - b_3 \xi_{\varrho}(\widehat{\Theta}_{\varrho}^{+}) \xi_{\varrho}(\widehat{\Phi}_{\varrho}^{-}) - c_3 \xi_{\varrho}(\widehat{\Theta}_{\varrho}^{+}) \xi_{\varrho}(\widehat{\Psi}_{\varrho}^{-}) - d_3 \xi_{\varrho}(\widehat{\Theta}_{\varrho}^{+}) \xi_{\varrho}(\widehat{\Theta}_{\varrho}^{-}), \omega)^h \right] dt. \end{aligned} \quad (120)$$

Theorem 4. Let the assumption (A) be valid and given $u^0, v^0, w^0 \in H_{\geq 0}^1(\mathfrak{D}) \cap L^{\infty}(\mathfrak{D})$, we consider the values of $\delta t, \varrho$, and h that satisfy the following conditions:

- (i) either $\widehat{\Phi}_{\varrho}^0 \equiv P^h u^0$, $\widehat{\Psi}_{\varrho}^0 \equiv P^h v^0$, $\widehat{\Theta}_{\varrho}^0 \equiv P^h w^0$ or $\widehat{\Phi}_{\varrho}^0 \equiv \pi^h u^0$, $\widehat{\Psi}_{\varrho}^0 \equiv \pi^h v^0$, $\widehat{\Theta}_{\varrho}^0 \equiv \pi^h w^0$, if either $d = 1$ or $u^0, v^0, w^0 \in W^{1,r}(\mathfrak{D})$ with $r > d$,

$$(ii) \quad \delta t \leq \frac{1-\delta}{2\hat{a}+\hat{b}+\hat{c}+\hat{d}},$$

$$(iii) \quad \delta t \rightarrow 0, \quad \varrho \rightarrow 0 \text{ as } h \rightarrow 0.$$

There is a subsequence of $\widehat{\Phi}_\varrho, \widehat{\Psi}_\varrho, \widehat{\Theta}_\varrho$ that satisfies (118), (119), and (120), along with corresponding functions

$$u \in L^2(0, T; H^1(\mathfrak{D})) \cap L^\sigma(\mathfrak{D}_T) \cap W^{1,q}(0, T; (W^{1,q'}(\mathfrak{D})))', \quad (121)$$

$$v \in L^2(0, T; H^1(\mathfrak{D})) \cap L^\sigma(\mathfrak{D}_T) \cap W^{1,q}(0, T; (W^{1,q'}(\mathfrak{D})))', \quad (122)$$

also

$$w \in L^2(0, T; H^1(\mathfrak{D})) \cap L^\sigma(\mathfrak{D}_T) \cap W^{1,q}(0, T; (W^{1,q'}(\mathfrak{D})))', \quad (123)$$

with

$$u(x, t), v(x, t), w(x, t) \geq 0, \quad a.e. \text{ on } \mathfrak{D},$$

and

$$u(., 0) = u^0, \quad v(., 0) = v^0, \quad w(., 0) = w^0, \quad \text{in } (W^{1,q'}(\mathfrak{D})))'. \quad (124)$$

As δt , ϱ , and h tend to zero, we can deduce that

$$\widehat{\Phi}_\varrho, \widehat{\Phi}_\varrho^\pm \rightharpoonup u, \quad \widehat{\Psi}_\varrho, \widehat{\Psi}_\varrho^\pm \rightharpoonup v, \quad \widehat{\Theta}_\varrho, \widehat{\Theta}_\varrho^\pm \rightharpoonup w, \quad \text{in } L^2(0, T; H^1(\mathfrak{D})) \cap L^\sigma(\mathfrak{D}_T), \quad (125)$$

$$\widehat{\Phi}_\varrho, \widehat{\Phi}_\varrho^\pm \rightharpoonup^* u, \quad \widehat{\Psi}_\varrho, \widehat{\Psi}_\varrho^\pm \rightharpoonup^* v, \quad \widehat{\Theta}_\varrho, \widehat{\Theta}_\varrho^\pm \rightharpoonup^* w, \quad \text{in } L^\infty(0, T; L^2(\mathfrak{D})), \quad (126)$$

$$\frac{\partial \widehat{\Phi}_\varrho}{\partial t} \rightharpoonup \frac{\partial u}{\partial t}, \quad \frac{\partial \widehat{\Psi}_\varrho}{\partial t} \rightharpoonup \frac{\partial v}{\partial t}, \quad \frac{\partial \widehat{\Theta}_\varrho}{\partial t} \rightharpoonup \frac{\partial w}{\partial t}, \quad \text{in } L^q(0, T; (W^{1,q'}(\mathfrak{D})))', \quad (127)$$

$$\widehat{\Phi}_\varrho, \widehat{\Phi}_\varrho^\pm \rightarrow u, \quad \widehat{\Psi}_\varrho, \widehat{\Psi}_\varrho^\pm \rightarrow v, \quad \widehat{\Theta}_\varrho, \widehat{\Theta}_\varrho^\pm \rightarrow w, \quad \text{in } L^2(0, T; L^{s_1}(\mathfrak{D})), \quad (128)$$

$$\xi_\varrho(\widehat{\Phi}_\varrho^\pm) \rightarrow u, \quad \xi_\varrho(\widehat{\Psi}_\varrho^\pm) \rightarrow v, \quad \xi_\varrho(\widehat{\Theta}_\varrho^\pm) \rightarrow w, \quad \text{in } L^2(0, T; L^{s_1}(\mathfrak{D})), \quad (129)$$

$$\pi^h \xi_\varrho(\widehat{\Phi}_\varrho^\pm) \rightarrow u, \quad \pi^h \xi_\varrho(\widehat{\Psi}_\varrho^\pm) \rightarrow v, \quad \pi^h \xi_\varrho(\widehat{\Theta}_\varrho^\pm) \rightarrow w, \quad \text{in } L^2(0, T; L^{s_1}(\mathfrak{D})), \quad (130)$$

$$\Xi_\varrho(\widehat{\Phi}_\varrho^\pm) \rightarrow u \mathcal{I} \Xi_\varrho(\widehat{\Psi}_\varrho^\pm) \rightarrow v \mathcal{I} \Xi_\varrho(\widehat{\Theta}_\varrho^\pm) \rightarrow w \mathcal{I} \text{ in } L^2(0, T; L^{s_1}(\mathfrak{D})), \quad (131)$$

for any

$$s_1 \in \begin{cases} [2, \infty] & \text{if } d = 1, \\ [2, \infty) & \text{if } d = 2, \\ [2, 6) & \text{if } d = 3. \end{cases}$$

Proof. Under the assumptions (i)-(ii) and utilizing (85), (93), (24), (56), (57), and (109)–(117), it can be proved that the following results, which do not depend on ϱ , h , or δt , are hold:

$$\begin{aligned} & \|\widehat{\Phi}_\varrho^\pm\|_{L^2(0,T;H^1(\mathfrak{D}))} + \|\widehat{\Phi}_\varrho^\pm\|_{L^\sigma(\mathfrak{D}_T)} + \|\widehat{\Phi}_\varrho^\pm\|_{L^\infty(0,T;L^1(\mathfrak{D}))} \\ & + \varrho^{\frac{-1}{2}} \|\pi^h[\widehat{\Phi}_\varrho^\pm]_-\|_{L^\infty(0,T;L^2(\mathfrak{D}))} + \left\| \frac{\partial \widehat{\Phi}_\varrho}{\partial t} \right\|_{L^q(0,T;(W^{1,q'}(\mathfrak{D})))'} \\ & + \|\mathfrak{G} \frac{\partial \widehat{\Phi}_\varrho}{\partial t}\|_{L^q(0,T;W^{1,q}(\mathfrak{D}))} + \|\xi_\varrho(\widehat{\Phi}_\varrho^\pm)\|_{L^\sigma(\mathfrak{D}_T)} \\ & + \|\pi^h \xi_\varrho(\widehat{\Phi}_\varrho^\pm)\|_{L^\sigma(\mathfrak{D}_T)} + \|\Xi_\varrho(\widehat{\Phi}_\varrho^\pm)\|_{L^\sigma(\mathfrak{D}_T)} \leq C, \end{aligned} \quad (132)$$

$$\begin{aligned} & \|\widehat{\Psi}_\varrho^\pm\|_{L^2(0,T;H^1(\mathfrak{D}))} + \|\widehat{\Psi}_\varrho^\pm\|_{L^\sigma(\mathfrak{D}_T)} + \|\widehat{\Psi}_\varrho^\pm\|_{L^\infty(0,T;L^1(\mathfrak{D}))} \\ & + \varrho^{\frac{-1}{2}} \|\pi^h[\widehat{\Psi}_\varrho^\pm]_-\|_{L^\infty(0,T;L^2(\mathfrak{D}))} + \left\| \frac{\partial \widehat{\Psi}_\varrho}{\partial t} \right\|_{L^q(0,T;(W^{1,q'}(\mathfrak{D})))'} \\ & + \|\mathfrak{G} \frac{\partial \widehat{\Psi}_\varrho}{\partial t}\|_{L^q(0,T;W^{1,q}(\mathfrak{D}))} + \|\xi_\varrho(\widehat{\Psi}_\varrho^\pm)\|_{L^\sigma(\mathfrak{D}_T)} \\ & + \|\pi^h \xi_\varrho(\widehat{\Psi}_\varrho^\pm)\|_{L^\sigma(\mathfrak{D}_T)} + \|\Xi_\varrho(\widehat{\Psi}_\varrho^\pm)\|_{L^\sigma(\mathfrak{D}_T)} \leq C, \end{aligned} \quad (133)$$

and

$$\begin{aligned} & \|\widehat{\Theta}_\varrho^\pm\|_{L^2(0,T;H^1(\mathfrak{D}))} + \|\widehat{\Theta}_\varrho^\pm\|_{L^\sigma(\mathfrak{D}_T)} + \|\widehat{\Theta}_\varrho^\pm\|_{L^\infty(0,T;L^1(\mathfrak{D}))} \\ & + \varrho^{\frac{-1}{2}} \|\pi^h[\widehat{\Theta}_\varrho^\pm]_-\|_{L^\infty(0,T;L^2(\mathfrak{D}))} + \left\| \frac{\partial \widehat{\Theta}_\varrho}{\partial t} \right\|_{L^q(0,T;(W^{1,q'}(\mathfrak{D})))'} \\ & + \|\mathfrak{G} \frac{\partial \widehat{\Theta}_\varrho}{\partial t}\|_{L^q(0,T;W^{1,q}(\mathfrak{D}))} + \|\xi_\varrho(\widehat{\Theta}_\varrho^\pm)\|_{L^\sigma(\mathfrak{D}_T)} \\ & + \|\pi^h \xi_\varrho(\widehat{\Theta}_\varrho^\pm)\|_{L^\sigma(\mathfrak{D}_T)} + \|\Xi_\varrho(\widehat{\Theta}_\varrho^\pm)\|_{L^\sigma(\mathfrak{D}_T)} \leq C, \end{aligned} \quad (134)$$

where $(\pm)$ is an adopted abbreviation for “with” and “without” the superscripts “+” and “-”. Based on (115) and the fifth bound in (132), we observe that

$$\begin{aligned} & \|\widehat{\Phi}_\varrho^\pm - \widehat{\Phi}_\varrho\|_{L^q(0,T;(W^{1,q'}(\mathfrak{D})))'}^q \\ & = \int_0^T \|\widehat{\Phi}_\varrho^\pm - \widehat{\Phi}_\varrho\|_{(W^{1,q'}(\mathfrak{D}))'}^q dt \\ & = \sum_{n=1}^N \int_{t_{n-1}}^{t_n} \|\widehat{\Phi}_\varrho^\pm - \widehat{\Phi}_\varrho\|_{(W^{1,q'}(\mathfrak{D}))'}^q dt = \sum_{n=1}^N \int_{t_{n-1}}^{t_n} |t - t_n^\pm|^q \left\| \frac{\partial \widehat{\Phi}_\varrho}{\partial t} \right\|_{(W^{1,q'}(\mathfrak{D}))'}^q dt \\ & \leq \sum_{n=1}^N (\delta t)^q \int_{t_{n-1}}^{t_n} \left\| \frac{\partial \widehat{\Phi}_\varrho}{\partial t} \right\|_{(W^{1,q'}(\mathfrak{D}))'}^q dt \leq (\delta t)^q \int_0^T \left\| \frac{\partial \widehat{\Phi}_\varrho}{\partial t} \right\|_{(W^{1,q'}(\mathfrak{D}))'}^q dt \end{aligned}$$

$$\leq (\delta t)^q \left\| \frac{\partial \widehat{\Phi}_\varrho}{\partial t} \right\|_{L^q(0,T;(W^{1,q'}(\mathfrak{D}))')}^q \leq C(\delta t)^q. \quad (135)$$

Similarly to (135), we can deduce, from (135), (116), (117), (133), and (134), that

$$\|\widehat{\Psi}_\varrho^\pm - \widehat{\Psi}_\varrho\|_{L^q(0,T;(W^{1,q'}(\mathfrak{D}))')}^q \leq (\delta t)^q, \quad (136)$$

and

$$\|\widehat{\Theta}_\varrho^\pm - \widehat{\Theta}_\varrho\|_{L^q(0,T;(W^{1,q'}(\mathfrak{D}))')}^q \leq (\delta t)^q. \quad (137)$$

Before completing the proof, it is necessary to state the following facts. The space $L^\infty(0,T;L^2(\mathfrak{D}))$ is the dual space of $L^1(0,T;L^2(\mathfrak{D}))$, which is a separable Banach space but not reflexive. On the other hand, the spaces $L^2(0,T;H^1(\mathfrak{D}))$ and $L^\sigma(\mathfrak{D}_T)$ are reflexive Banach spaces. By employing compactness arguments and utilizing the bounds (132), we can extract subsequences denoted as $\{\widehat{\Phi}_\varrho^\pm\}_h$ and $\{\widehat{\Phi}_\varrho\}_h$. These subsequences satisfy the results (125) and (126) as $h \rightarrow 0$. Furthermore, since $\{\frac{\partial \widehat{\Phi}_\varrho}{\partial t}\}_h \in L^q(0,T;(W^{1,q'}(\mathfrak{D}))')$, which are reflexive Banach spaces, by the weak compactness theorem, there exists a subsequence $\{\frac{\partial \widehat{\Phi}_\varrho}{\partial t}\}_h \in L^q(0,T;(W^{1,q'}(\mathfrak{D}))')$ and a function $\widetilde{\eta} \in L^q(0,T;(W^{1,q'}(\mathfrak{D}))')$ such that

$$\frac{\partial \widehat{\Phi}_\varrho}{\partial t} \rightharpoonup \widetilde{\eta} \quad \text{in } L^q(0,T;(W^{1,q'}(\mathfrak{D}))').$$

Using a well-known method outlined in [37], page 204, it can be readily demonstrated that $\widetilde{\eta} = \frac{\partial \widehat{\Phi}}{\partial t}$. Consequently, the result (127) is obtained.

From (125) and (126), it follows that $\widehat{\Phi}$ belongs to $L^2(0,T;H^1(\mathfrak{D}))$, $L^\sigma(\mathfrak{D}_T)$, and $L^\infty(0,T;L^2(\mathfrak{D}))$. To prove (121), we need to demonstrate that $\widehat{\Phi}$ also belongs to $W^{1,q}(0,T;(W^{1,q'}(\mathfrak{D}))')$. Using the embedding $L^2(0,T;H^1(\mathfrak{D})) \hookrightarrow L^q(0,T;(W^{1,q'}(\mathfrak{D}))')$, we conclude that $\widehat{\Phi} \in L^q(0,T;(W^{1,q'}(\mathfrak{D}))')$. Furthermore, from (127), we have $\frac{\partial \widehat{\Phi}}{\partial t} \in L^q(0,T;(W^{1,q'}(\mathfrak{D}))')$, therefore, it follows that

$$\|\widehat{\Phi}\|_{W^{1,q}(0,T;(W^{1,q'}(\mathfrak{D}))')} = \|\widehat{\Phi}\|_{L^q(0,T;(W^{1,q'}(\mathfrak{D}))')} + \left\| \frac{\partial \widehat{\Phi}}{\partial t} \right\|_{L^q(0,T;(W^{1,q'}(\mathfrak{D}))')} \leq C.$$

Therefore, (121) has been demonstrated. We can establish (122) by employing a similar approach as in (121).

Based on the given compact embedding result

$$\begin{aligned} \mathcal{H}_u &= \left\{ \Sigma, \Sigma \in L^2(0,T;H^1(\mathfrak{D})), \frac{\partial \Sigma}{\partial t} \in L^q(0,T;(W^{1,q'}(\mathfrak{D}))') \right\} \\ &\stackrel{c}{\hookrightarrow} L^2(0,T;L^{s_1}(\mathfrak{D})), \end{aligned}$$

we have, as $\widehat{\Phi}_\varrho \in L^2(0, T; H^1(\mathfrak{D}))$ and $\frac{\partial \widehat{\Phi}_\varrho}{\partial t} \in L^q(0, T; (W^{1,q'}(\mathfrak{D})))'$, that $\widehat{\Phi}_\varrho \in \mathcal{H}_e$, then We may extract a subsequence, still assigned $\widehat{\Phi}_\varrho$, that supports the convergence outcome (128).

By utilizing the strong convergence of $\widehat{\Phi}_\varrho$ to u in $L^2(0, T; L^{s_1}(\mathfrak{D}))$, along with the fourth bound in (132), we can extract a subsequence, denoted as $\widehat{\Phi}_\varrho$ as well, such that as $h \rightarrow 0$, the following condition holds

$$\widehat{\Phi}_\varrho \rightarrow u \quad \text{and} \quad \pi^h[\widehat{\Phi}_\varrho]_- \rightarrow 0 \quad \text{a.e. in } \mathfrak{D} \times (0, T). \quad (138)$$

The definition of π^h implies that

$$\widehat{\Phi}_\varrho = \pi^h[\widehat{\Phi}_\varrho]_+ + \pi^h[\widehat{\Phi}_\varrho]_-. \quad (139)$$

From (138) and (139), we can conclude that $u \geq 0$ almost everywhere in $\mathfrak{D}$.

To display the outcomes (129)–(131), it is important to initially observe that

$$\begin{aligned} & \|\xi_\varrho(\widehat{\Phi}_\varrho^\pm) - u\|_{L^2(0, T; L^{s_1}(\mathfrak{D}))} \\ & \leq \|\xi_\varrho(\widehat{\Phi}_\varrho^\pm) - \xi_\varrho(u)\|_{L^2(0, T; L^{s_1}(\mathfrak{D}))} + \|\xi_\varrho(u) - u\|_{L^2(0, T; L^{s_1}(\mathfrak{D}))}. \end{aligned} \quad (140)$$

By employing the fact that $u \in L^2(0, T; H^1(\mathfrak{D})) \hookrightarrow L^2(0, T; L^{s_1}(\mathfrak{D}))$, along with (24), the non-negativity of $\widehat{\Phi}$, assumption (iii), and applying the dominated convergence theorem, we can reach the conclusion that

$$\|\xi_\varrho(u) - u\|_{L^2(0, T; L^{s_1}(\mathfrak{D}))} \rightarrow 0 \quad \text{as } h \rightarrow 0. \quad (141)$$

The fact that ξ_ϱ exhibits Lipschitz continuity, along with the relationship stated in (128), leads to the inference that

$$\|\xi_\varrho(\widehat{\Phi}_\varrho^\pm) - \xi_\varrho(u)\|_{L^2(0, T; L^{s_1}(\mathfrak{D}))} \leq \|\widehat{\Phi}_\varrho^\pm - u\|_{L^2(0, T; L^{s_1}(\mathfrak{D}))} \rightarrow 0 \quad \text{as } h \rightarrow 0. \quad (142)$$

By employing (49), (24), (48), and the first bound stated in (132), it is possible to deduce that

$$\begin{aligned} & \|(I - \pi^h)\xi_\varrho(\widehat{\Phi}_\varrho^\pm)\|_{L^2(0, T; L^{s_1}(\mathfrak{D}))} + \|\Xi_\varrho(\widehat{\Phi}_\varrho^\pm) - u\|_{L^2(0, T; L^{s_1}(\mathfrak{D}))} \\ & \leq Ch^{1-d(\frac{1}{2}-\frac{1}{s_1})} \|\xi_\varrho(\widehat{\Phi}_\varrho^\pm) - u\|_{L^2(0, T; H^1(\mathfrak{D}))}. \end{aligned} \quad (143)$$

Combining (140), (141), (142), and (7) yields the convergence results (129)–(131). The convergence results for $\widehat{\Psi}_\varrho^\pm$ and $\widehat{\Theta}_\varrho^\pm$ can be proven using the same approach as the one followed to prove the convergence results for $\widehat{\Phi}_\varrho^\pm$.

To complete this proof, we need to show that the solution u, v, w , based on the initial approximations, satisfies (124). This can be deduced from the error estimates (49) and (70), as well as the given assumptions on u^0 , where we have that

$$\|u^0 - P^h u^0\|_0 \leq Ch|u^0|_1 \leq Ch,$$

and

$$\|u^0 - \pi^h u^0\|_0 \leq \begin{cases} Ch|u^0|_1 \leq Ch, & \text{for } d = 1, \\ Ch|u^0|_{1,r} \leq Ch, & \text{for } d = 1 \text{ or } d = 2, \end{cases} \quad (144)$$

and similarly for v^0 and w^0 . These results exhibit strong convergence as $h \rightarrow 0$, in the form

$$\widehat{\Phi}_\varrho^0 \rightarrow u^0, \quad \widehat{\Psi}_\varrho^0 \rightarrow v^0, \quad \widehat{\Theta}_\varrho^0 \rightarrow w^0, \quad \text{in } L^2(\mathfrak{D}). \quad (145)$$

By (129), we can conclude that

$$\widehat{\Phi}_\varrho(t) \rightarrow u(t), \quad \widehat{\Psi}_\varrho(t) \rightarrow v(t), \quad \widehat{\Theta}_\varrho(t) \rightarrow w(t), \quad \text{in } L^2(\mathfrak{D}), \quad \text{as } \varrho \rightarrow 0. \quad (146)$$

We observe that (145), and (146) alone do not suffice to prove the result in (124). This is because if $t = 0$ belongs to the null-set of the “almost everywhere” condition for (146), it is possible that $u(0) \neq u^0$, $v(0) \neq v^0$, and $w(0) \neq w^0$ (as discussed in [37], Section 7.4.4). Additionally, we rely on other properties of the solutions $\widehat{\Phi}_\varrho, \widehat{\Psi}_\varrho, \widehat{\Theta}_\varrho$ and the function u, v, w to establish the validity of (124). Firstly, we note that

$$\widehat{\Phi}_\varrho, u, \widehat{\Psi}_\varrho, v, \widehat{\Theta}_\varrho, w \in W^{1,q}(0, T; (W^{1,q'}(\mathfrak{D})))'.$$

Consequently, we have that

$$\widehat{\Phi}_\varrho, u, \widehat{\Psi}_\varrho, v, \widehat{\Theta}_\varrho, w \in C([0, T]; (W^{1,q'}(\mathfrak{D})))'; \quad (147)$$

see [37, Proposition 7.1 and Theorem 7.2], respectively. By noting the (145)–(147), the desired result (124) follows directly. $\square$

Lemma 8. Assuming the conditions of Theorem 4 are satisfied, it follows, as $h \rightarrow 0$, that

$$\pi^h \xi_\varrho(\widehat{\Phi}_\varrho^+) \pi^h \xi_\varrho(\widehat{\Phi}_\varrho^-) \rightarrow u^2 \quad \text{in } L^q(\mathfrak{D}_T), \quad (148)$$

$$\pi^h \xi_\varrho(\widehat{\Psi}_\varrho^+) \pi^h \xi_\varrho(\widehat{\Psi}_\varrho^-) \rightarrow v^2 \quad \text{in } L^q(\mathfrak{D}_T), \quad (149)$$

$$\pi^h \xi_\varrho(\widehat{\Theta}_\varrho^+) \pi^h \xi_\varrho(\widehat{\Theta}_\varrho^-) \rightarrow w^2 \quad \text{in } L^q(\mathfrak{D}_T), \quad (150)$$

$$\pi^h \xi_\varrho(\widehat{\Phi}_\varrho^+) \pi^h \xi_\varrho(\widehat{\Psi}_\varrho^-) \rightarrow uv \quad \text{in } L^q(\mathfrak{D}_T), \quad (151)$$

$$\pi^h \xi_\varrho(\widehat{\Theta}_\varrho^+) \pi^h \xi_\varrho(\widehat{\Phi}_\varrho^-) \rightarrow wu \quad \text{in } L^q(\mathfrak{D}_T), \quad (152)$$

$$\pi^h \xi_\varrho(\widehat{\Theta}_\varrho^+) \pi^h \xi_\varrho(\widehat{\Psi}_\varrho^-) \rightarrow wv \quad \text{in } L^q(\mathfrak{D}_T), \quad (153)$$

where $q = \frac{2(d+1)}{2d+1}$.

Proof. By utilizing Hölder's inequality, together with (132) and (121), we can deduce that

$$\begin{aligned} & \|\pi^h \xi_\varrho(\widehat{\Phi}_\varrho^+) \pi^h \xi_\varrho(\widehat{\Phi}_\varrho^-) - u^2\|_{L^q(\mathfrak{D}_T)} \leq \|\pi^h \xi_\varrho(\widehat{\Phi}_\varrho^-) - u\|_{L^2(\mathfrak{D}_T)} \|\pi^h \xi_\varrho(\widehat{\Phi}_\varrho^+)\|_{L^\sigma(\mathfrak{D}_T)} \\ & + \|\pi^h \xi_\varrho(\widehat{\Phi}_\varrho^+) - u\|_{L^2(\mathfrak{D}_T)} \|u\|_{L^\sigma(\mathfrak{D}_T)} \leq C(\|\pi^h \xi_\varrho(\widehat{\Phi}_\varrho^-) - u\|_{L^2(\mathfrak{D}_T)} \\ & + \|\pi^h \xi_\varrho(\widehat{\Phi}_\varrho^+) - u\|_{L^2(\mathfrak{D}_T)}). \end{aligned} \quad (154)$$

Hence, the outcome (148) can be deduced from (154), (128), and (130), while taking into consideration that $L^2(0, T; L^{s_1}(\mathfrak{D}))$ is embedded in $L^2(\mathfrak{D}_T)$. In a similar manner, the results (149)–(151) can be established by employing analogous methods as those utilized in establishing the proof for (148). $\square$

Theorem 5. Let the conditions of Theorem 4 be fulfilled, there exists a subsequence of $\{\widehat{\Phi}_\varrho, \widehat{\Psi}_\varrho, \widehat{\Theta}_\varrho\}$ that satisfies (118), (119), and (120). Consequently, the functions $\{u, v, w\}$ represent a global weak solution in the specified context:

$$\begin{aligned} & \int_0^T \left[\left\langle \frac{\partial u}{\partial t}, \Sigma \right\rangle + (\theta_1 \nabla u + 2\theta_{11} u \nabla u, \nabla \Sigma) \right] dt \\ & = \int_0^T \left[(a_1 u - b_1 u^2 - c_1 uv - d_1 uw, \Sigma) \right] dt, \quad \text{for all } \Sigma \in L^{q'}(0, T; W^{1,q'}(\mathfrak{D})), \end{aligned} \quad (155)$$

$$\begin{aligned} & \int_0^T \left[\left\langle \frac{\partial v}{\partial t}, \Sigma \right\rangle + (\theta_2 \nabla v + \theta_{21} v \nabla u + \theta_{21} u \nabla v + 2\theta_{22} v \nabla v + \theta_{23} v \nabla w + \theta_{23} w \nabla v, \nabla \Sigma) \right] dt \\ & = \int_0^T \left[(a_2 v - b_2 vu - c_2 v^2 - d_2 vw, \Sigma)^h \right] dt, \quad \text{for all } \Sigma \in L^{q'}(0, T; W^{1,q'}(\mathfrak{D})), \end{aligned} \quad (156)$$

and

$$\begin{aligned} & \int_0^T \left[\left\langle \frac{\partial w}{\partial t}, \Sigma \right\rangle + (\theta_3 \nabla w + 2\theta_{33} w \nabla w, \nabla \Sigma) \right] dt \\ & = \int_0^T \left[(a_3 w - b_3 wu - c_3 wv - d_3 w^2, \Sigma) \right] dt, \quad \text{for all } \Sigma \in L^{q'}(0, T; W^{1,q'}(\mathfrak{D})), \end{aligned} \quad (157)$$

where $q' = 2(d + 1)$.

Proof. A proof of the aforementioned theorems can be obtained by applying a procedure similar to the one described in [32, 30, 2]. $\square$

8 Numerical results

This section concentrates on numerical results concerning the three-species SKT model. An iterative approach has been developed to solve the nonlinear system of equations arising from problem $(Q_\rho^{h,\delta t})$. Moreover, numerical error analyses have been presented and examined for the SKT model with Dirichlet non-homogeneous boundary conditions in one, two, and three dimensions. Additionally, numerical simulations have been performed for the one- and two-dimensional cases. The numerical simulations were performed using MATLAB.

8.1 Numerical algorithm

At each time step, we suggest the following practical algorithm for solving the nonlinear algebraic system that emerges from problem $(Q_\rho^{h,\delta t})$: $(Q_{\rho,k}^{h,\delta t})$: Given $\{\hat{\Phi}_\rho^{n,0}, \hat{\Psi}_\rho^{n,0}\} \in X^h \times X^h$ for $k \geq 1$ find $\{\hat{\Phi}_\rho^{n,k}, \hat{\Psi}_\rho^{n,k}\} \in X^h \times X^h$ such that for all $\omega \in X^h$

$$\begin{aligned} & \left(\frac{\hat{\Phi}_\rho^{n,k} - \hat{\Phi}_\rho^{n-1}}{\delta t}, \omega \right)^h + (\theta_1 \nabla \hat{\Phi}_\rho^{n,k} + 2\theta_{11} \Xi_\rho(\hat{\Phi}_\rho^{n,k-1}) \nabla \hat{\Phi}_\rho^{n,k}, \nabla \omega) \\ & = (a_1 \hat{\Phi}_\rho^{n,k} - b_1 \xi_\rho(\hat{\Phi}_\rho^{n-1}) \xi_\rho(\hat{\Phi}_\rho^{n,k-1}) - c_1 \xi_\rho(\hat{\Psi}_\rho^{n-1}) \xi_\rho(\hat{\Phi}_\rho^{n,k-1}) \\ & \quad - d_1 \xi_\rho(\hat{\Theta}_\rho^{n-1}) \xi_\rho(\hat{\Phi}_\rho^{n,k-1}), \omega)^h, \end{aligned} \quad (158)$$

$$\begin{aligned} & \left(\frac{\hat{\Psi}_\rho^{n,k} - \hat{\Psi}_\rho^{n-1}}{\delta t}, \omega \right)^h + (\theta_2 \nabla \hat{\Psi}_\rho^{n,k} + \theta_{21} \Xi_\rho(\hat{\Psi}_\rho^{n,k-1}) \nabla \hat{\Phi}_\rho^{n,k} \\ & + \theta_{21} \Xi_\rho(\hat{\Phi}_\rho^{n,k-1}) \nabla \hat{\Psi}_\rho^{n,k} + 2\theta_{22} \Xi_\rho(\hat{\Psi}_\rho^{n,k-1}) \nabla \hat{\Psi}_\rho^{n,k} \\ & + \theta_{23} \Xi_\rho(\hat{\Psi}_\rho^{n,k-1}) \nabla \hat{\Theta}_\rho^{n,k} + \theta_{23} \Xi_\rho(\hat{\Theta}_\rho^{n,k-1}) \nabla \hat{\Psi}_\rho^{n,k}, \nabla \omega) \\ & = (a_2 \hat{\Psi}_\rho^{n,k} - b_2 \xi_\rho(\hat{\Phi}_\rho^{n-1}) \xi_\rho(\hat{\Psi}_\rho^{n,k-1}) - c_2 \xi_\rho(\hat{\Psi}_\rho^{n-1}) \xi_\rho(\hat{\Psi}_\rho^{n,k-1}) \\ & \quad - d_2 \xi_\rho(\hat{\Theta}_\rho^{n-1}) \xi_\rho(\hat{\Psi}_\rho^{n,k-1}), \omega)^h, \end{aligned} \quad (159)$$

$$\begin{aligned}
& \left(\frac{\hat{\Theta}_{\varrho}^{n,k} - \hat{\Theta}_{\varrho}^{n-1}}{\delta t}, \omega \right)^h + (\theta_3 \nabla \hat{\Theta}_{\varrho}^{n,k} + 2\theta_{33} \Xi_{\varrho}(\hat{\Theta}_{\varrho}^{n,k-1}) \nabla \hat{\Theta}_{\varrho}^{n,k}, \nabla \omega) \\
& = (a_3 \hat{\Theta}_{\varrho}^{n,k} - b_3 \xi_{\varrho}(\hat{\Phi}_{\varrho}^{n-1}) \xi_{\varrho}(\hat{\Theta}_{\varrho}^{n,k-1}) - c_3 \xi_{\varrho}(\hat{\Psi}_{\varrho}^{n-1}) \xi_{\varrho}(\hat{\Theta}_{\varrho}^{n,k-1}) \\
& \quad - d_3 \xi_{\varrho}(\hat{\Theta}_{\varrho}^{n-1}) \xi_{\varrho}(\hat{\Theta}_{\varrho}^{n,k-1}), \omega)^h.
\end{aligned} \tag{160}$$

Let us start with $\hat{\Phi}_{\varrho}^0 = \pi^h u^0$, $\hat{\Psi}_{\varrho}^0 = \pi^h v^0$ and $\hat{\Theta}_{\varrho}^0 = \pi^h W^0$. For $n \geq 1$, we define $\hat{\Phi}_{\varrho}^{n,0} = \hat{\Phi}_{\varrho}^{n-1}$, $\hat{\Psi}_{\varrho}^{n,0} = \hat{\Psi}_{\varrho}^{n-1}$ and $\hat{\Theta}_{\varrho}^{n,0} = \hat{\Theta}_{\varrho}^{n-1}$. The standard method for solving (158) - (160) involves testing the equations with $\eta_j, j = 0, \dots, J$, to obtain a $(2J+2)^3$ linear equations with $(2J+2)^3$ unknowns. $\hat{\Phi}_{\varrho}^{n,k}$, $\hat{\Psi}_{\varrho}^{n,k}$ and $\hat{\Theta}_{\varrho}^{n,k}$. This resulting system can be solved using the Gauss-Seidel iterative method. In our numerical results, we set $TOL = 10^{-8}$ and used the stopping criteria:

$$\max \left\{ |\hat{\Phi}_{\varrho}^{n,k} - \hat{\Phi}_{\varrho}^{n,k-1}|_{0,\infty}, |\hat{\Psi}_{\varrho}^{n,k} - \hat{\Psi}_{\varrho}^{n,k-1}|_{0,\infty}, |\hat{\Theta}_{\varrho}^{n,k} - \hat{\Theta}_{\varrho}^{n,k-1}|_{0,\infty} \right\} < TOL. \tag{161}$$

For a given k that satisfies (161), we define $\hat{\Phi}_{\varrho}^n = \hat{\Phi}_{\varrho}^{n,k}$, $\hat{\Psi}_{\varrho}^n = \hat{\Psi}_{\varrho}^{n,k}$ and $\hat{\Theta}_{\varrho}^n = \hat{\Theta}_{\varrho}^{n,k}$.

Although we could not confirm the convergence of $\{\hat{\Phi}_{\varrho}^{n,k}, \hat{\Psi}_{\varrho}^{n,k}, \hat{\Theta}_{\varrho}^{n,k}\}_{k=1}^{\infty}$ to $\{\hat{\Phi}_{\varrho}^n, \hat{\Psi}_{\varrho}^n, \hat{\Theta}_{\varrho}^n\}$ for a fixed value of n , strong convergence characteristics were observed in practice. The iterative method consistently achieved good convergence, often requiring only a few steps at each stage to meet the termination criteria.

8.2 Error computations

To quantify the error, we can introduce source terms $f_1(\mathbf{x}, t)$, $f_2(\mathbf{x}, t)$ and $f_3(\mathbf{x}, t)$ to the problem (Q). This modification transforms (1)–(3) into the following equations:

(Q) Find $\{u, v, w\}$ such that

$$\partial_t u = \Delta(\theta_1 u + \theta_{11} u^2) + (a_1 - b_1 u - c_1 v - d_1 w)u + f_1(\mathbf{x}, t), \quad \text{in } \mathfrak{D}, \quad t > 0, \tag{162}$$

$$\partial_t v = \Delta(\theta_2 v + \theta_{21} uv + \theta_{22} v^2 + \theta_{23} vw) + (a_2 - b_2 u - c_2 v - d_2 w)v + f_2(\mathbf{x}, t) \quad \text{in } \mathfrak{D}, \quad t > 0,$$

$$\partial_t w = \Delta(\theta_3 w + \theta_{33} w^2) + (a_3 - b_3 u - c_3 v - d_3 w)w + f_3(\mathbf{x}, t), \quad \text{in } \mathfrak{D}, \quad t > 0, \tag{163}$$

We consider the following discrete FEM of $(Q_{\varrho}^{h,\delta t})$:

Given $\hat{\Phi}_{\varrho}^{n,0}, \hat{\Psi}_{\varrho}^{n,0} \in X^h \times X^h$, we seek $\hat{\Phi}_{\varrho}^{n,k}, \hat{\Psi}_{\varrho}^{n,k} \in X^h \times X^h$ for $k \geq 1$ such that for all $\omega \in X^h$, we have that

$$\left(\frac{\hat{\Phi}_{\varrho}^{n,k} - \hat{\Phi}_{\varrho}^{n-1}}{\delta t}, \omega \right)^h + (\theta_1 \nabla \hat{\Phi}_{\varrho}^{n,k} + 2\theta_{11} \Xi_{\varrho}(\hat{\Phi}_{\varrho}^{n,k-1}) \nabla \hat{\Phi}_{\varrho}^{n,k}, \nabla \omega)$$

$$\begin{aligned}
& -(a_1 \widehat{\Phi}_\varrho^{n,k} - b_1 \xi_\varrho(\widehat{\Phi}_\varrho^{n-1}) \xi_\varrho(\widehat{\Phi}_\varrho^{n,k-1}) - c_1 \xi_\varrho(\widehat{\Psi}_\varrho^{n-1}) \xi_\varrho(\widehat{\Phi}_\varrho^{n,k-1}) \\
& - d_1 \xi_\varrho(\widehat{\Theta}_\varrho^{n-1}) \xi_\varrho(\widehat{\Phi}_\varrho^{n,k-1}), \omega)^h = (f_1(\mathbf{x}, t_n), \omega)^h,
\end{aligned} \tag{164}$$

$$\begin{aligned}
& \left(\frac{\widehat{\Psi}_\varrho^{n,k} - \widehat{\Psi}_\varrho^{n-1}}{\delta t}, \omega \right)^h + (\theta_2 \nabla \widehat{\Psi}_\varrho^{n,k} + \theta_{21} \Xi_\varrho(\widehat{\Psi}_\varrho^{n,k-1}) \nabla \widehat{\Phi}_\varrho^{n,k} \\
& + \theta_{21} \Xi_\varrho(\widehat{\Phi}_\varrho^{n,k-1}) \nabla \widehat{\Psi}_\varrho^{n,k} + 2\theta_{22} \Xi_\varrho(\widehat{\Psi}_\varrho^{n,k-1}) \nabla \widehat{\Psi}_\varrho^{n,k} + \theta_{23} \Xi_\varrho(\widehat{\Psi}_\varrho^{n,k-1}) \nabla \widehat{\Theta}_\varrho^{n,k} \\
& + \theta_{23} \Xi_\varrho(\widehat{\Theta}_\varrho^{n,k-1}) \nabla \widehat{\Psi}_\varrho^{n,k}, \nabla \omega) - (a_2 \widehat{\Psi}_\varrho^{n,k} - b_2 \xi_\varrho(\widehat{\Phi}_\varrho^{n-1}) \xi_\varrho(\widehat{\Psi}_\varrho^{n,k-1}) \\
& - c_2 \xi_\varrho(\widehat{\Psi}_\varrho^{n-1}) \xi_\varrho(\widehat{\Psi}_\varrho^{n,k-1}) - d_2 \xi_\varrho(\widehat{\Theta}_\varrho^{n-1}) \xi_\varrho(\widehat{\Psi}_\varrho^{n,k-1}), \omega)^h = (f_2(\mathbf{x}, t_n), \omega)^h,
\end{aligned} \tag{165}$$

$$\begin{aligned}
& \left(\frac{\widehat{\Theta}_\varrho^{n,k} - \widehat{\Theta}_\varrho^{n-1}}{\delta t}, \omega \right)^h + (\theta_3 \nabla \widehat{\Theta}_\varrho^{n,k} + 2\theta_{33} \Xi_\varrho(\widehat{\Theta}_\varrho^{n,k-1}) \nabla \widehat{\Theta}_\varrho^{n,k}, \nabla \omega) \\
& - (a_3 \widehat{\Theta}_\varrho^{n,k} - b_3 \xi_\varrho(\widehat{\Phi}_\varrho^{n-1}) \xi_\varrho(\widehat{\Theta}_\varrho^{n,k-1}) - c_3 \xi_\varrho(\widehat{\Psi}_\varrho^{n-1}) \xi_\varrho(\widehat{\Theta}_\varrho^{n,k-1}) \\
& - d_3 \xi_\varrho(\widehat{\Theta}_\varrho^{n-1}) \xi_\varrho(\widehat{\Theta}_\varrho^{n,k-1}), \omega)^h = (f_3(\mathbf{x}, t_n), \omega)^h.
\end{aligned} \tag{166}$$

8.2.1 One-dimensional accuracy

To assess the accuracy of our approximation, we solve (162)–(163) with non-homogeneous Dirichlet boundary conditions. Let's illustrate this with a numerical example. We evaluate the error of the approximate solution when all parameters of the problem are equal to 1 and $\mathfrak{D} = [0, 1]$. The initial conditions, boundary conditions, and source terms $f(x, t)$ and $g(x, t)$ can be obtained from the corresponding exact solution. We divide the interval $\mathfrak{D} = [0, 1]$ into J equal sub-intervals, with $h = 1/J$ as the element mesh size, and $\delta t = 10^{-6}$ as the time step. We consider the following exact solution:

$$u = \exp(-t + x), \quad v = \exp(-t + x^2), \quad w = \exp(-t + x^3).$$

The error norms in L^2, L^∞ for a set of simulations were shown in Table 1 and the order of convergence of L^2, L^∞ -error are introduced in Table 2.

8.2.2 Two-dimensional accuracy

Here is an example illustrating the error of approximate solutions for the two-dimensional case. We focus on investigating the Dirichlet boundary conditions for (162)–(163). In this numerical

Table 1: Discrete L^2, L^∞ -error for one-dimensional case at $T = 1$ with $\theta_i = \theta_{ii} = \theta_{21} = \theta_{23} = a_i = b_i = c_i = d_i = 1, i = 1, 2, 3$.

J	$\ u - \widehat{\Phi}_\rho^n\ $		$\ v - \widehat{\Psi}_\rho^n\ $		$\ w - \widehat{\Theta}_\rho^n\ $	
	L^2	L^∞	L^2	L^∞	L^2	L^∞
10	6.6E-05	2.7E-04	1.4E-03	5.3E-03	3.2E-03	1.3E-02
20	1.2E-05	7.0E-05	2.3E-04	1.3E-03	5.5E-04	3.3E-03
25	6.6E-06	4.5E-05	1.3E-04	8.6E-04	3.1E-04	2.1E-03
40	2.0E-06	1.7E-05	4.0E-05	3.4E-04	9.6E-05	8.4E-04
50	1.1E-06	1.1E-05	2.3E-05	2.2E-04	5.5E-05	5.4E-04
100	1.8E-07	2.6E-06	4.0E-06	5.4E-05	9.6E-06	1.3E-04

Table 2: The order of convergence of L^2, L^∞ -error for one-dimensional case at $T = 1$ with $\theta_i = \theta_{ii} = \theta_{21} = \theta_{23} = a_i = b_i = c_i = d_i = 1, i = 1, 2, 3$.

J	$\ u - \widehat{\Phi}_\rho^n\ $		$\ v - \widehat{\Psi}_\rho^n\ $		$\ w - \widehat{\Theta}_\rho^n\ $	
	L^2	L^∞	L^2	L^∞	L^2	L^∞
20	2.459	1.948	2.606	2.027	2.541	1.978
25	2.679	1.980	2.557	1.852	2.569	2.026
40	2.540	2.071	2.508	1.974	2.494	1.950
50	2.679	1.951	2.480	1.951	2.496	1.980
100	2.611	2.081	2.524	2.026	2.518	2.054

example, we consider a domain $\mathfrak{D} = [0, 1] \times [0, 1]$ and a time interval $[0, 1]$. The problem parameter values were fixed to be 1. We utilize exact solutions to determine the initial and boundary conditions, as well as the source terms $f(x, y, t)$ and $g(x, y, t)$. For the numerical computation, we choose a time step of $\delta t = 10^{-6}$. The exact solutions are provided below:

$$u = \exp(-t + x + y), \quad v = \exp(-t + x^2 + y^2), \quad w = \exp(-t + x^3 + y^3).$$

Table 3 presents the errors in the L^2 and L^∞ norms. The order of convergence of L^2 and L^∞ norms are shown in Table 4.

8.2.3 Three-dimensional accuracy

This section introduces an example that demonstrates the error of approximate solutions for a three-dimensional case. The specific focus of investigation is on the Dirichlet boundary conditions applied to (162)–(163). The domain is selected as $\mathfrak{D} = [0, 1] \times [0, 1] \times [0, 1]$, and the time interval is $[0, 1]$. The problem's parameter values are set as follows: $\theta_i = 0.1, \theta_{ii} = 0.001, \theta_{21} = \theta_{23} = 0.01, a_i = b_i = c_i = d_i = 1$, where $i = 1, 2, 3$. The exact solutions are utilized to determine the

Table 3: Discrete L^2, L^∞ -error for two-dimensional case at $T = 1$ with $\theta_i = \theta_{ii} = \theta_{21} = \theta_{23} = a_i = b_i = c_i = d_i = 1, i = 1, 2, 3$.

J	$\ u - \widehat{\Phi}_\rho^n\ $		$\ v - \widehat{\Psi}_\rho^n\ $		$\ w - \widehat{\Theta}_\rho^n\ $	
	L^2	L^∞	L^2	L^∞	L^2	L^∞
10	7.1E-06	9.4E-05	6.3E-04	9.4E-03	1.4E-03	2.3E-02
20	7.3E-07	2.3E-05	7.2E-05	2.4E-03	1.6E-04	5.9E-03
25	3.6E-07	1.5E-05	3.6E-05	1.5E-03	8.3E-05	3.8E-03
40	9.7E-08	6.1E-06	8.6E-06	6.1E-04	2.0E-05	1.5E-03
50	5.3E-08	3.9E-06	4.4E-06	3.9E-04	1.0E-05	9.7E-04

Table 4: The order of convergence of L^2, L^∞ -error for two-dimensional case at $T = 1$ with $\theta_i = \theta_{ii} = \theta_{21} = \theta_{23} = a_i = b_i = c_i = d_i = 1, i = 1, 2, 3$.

J	$\ u - \widehat{\Phi}_\rho^n\ $		$\ v - \widehat{\Psi}_\rho^n\ $		$\ w - \widehat{\Theta}_\rho^n\ $	
	L^2	L^∞	L^2	L^∞	L^2	L^∞
20	3.282	2.031	3.129	1.970	3.129	1.963
25	3.168	1.916	3.106	2.106	2.941	1.972
40	2.790	1.914	3.046	1.914	3.028	1.978
50	2.970	2.005	3.003	2.005	3.018	1.954

initial and boundary conditions, as well as the source terms $f(x, y, z, t)$ and $g(x, y, z, t)$. The value of the time step is chosen such that $\delta t = 10^{-4}$. The exact solutions are

$$u = \exp(-t + x + y + z), \quad v = \exp(-t + x^2 + y^2 + z^2), \quad w = \exp(-t + x^3 + y^3 + z^3).$$

Table 5 displays the errors calculated using the L^2 , and L^∞ norms and the order of convergence of L^2 , and L^∞ norms are presented in Table 6.

Here we will make notes regarding the results of Tables 1-6. We note that the highest error values were obtained by the solution v , and the reason is that the second equation, through which

Table 5: Discrete L^2, L^∞ -error for three-dimensional case at $T = 1$ with $\theta_i = 0.1, \theta_{ii} = 0.001, \theta_{21} = \theta_{23} = 0.01, a_i = b_i = c_i = d_i = 1, i = 1, 2, 3$.

J	$\ u - \widehat{\Phi}_\rho^n\ $		$\ v - \widehat{\Psi}_\rho^n\ $		$\ w - \widehat{\Theta}_\rho^n\ $	
	L^2	L^∞	L^2	L^∞	L^2	L^∞
10	5.5E-05	3.4E-03	1.1E-04	7.5E-03	1.2E-04	7.7E-03
20	4.2E-06	9.1E-04	8.4E-06	2.1E-03	9.6E-06	2.2E-03
25	1.9E-06	5.8E-04	3.8E-06	1.4E-03	4.3E-06	1.5E-03
40	3.2E-07	2.2E-04	7.1E-07	5.3E-04	8.1E-07	5.9E-04

Table 6: The order of convergence of L^2, L^∞ -error for three-dimensional case at $T = 1$ with $\theta_i = 0.1, \theta_{ii} = 0.001, \theta_{21} = \theta_{23} = 0.01, a_i = b_i = c_i = d_i = 1, i = 1, 2, 3$.

J	$\ u - \hat{\Phi}_\rho^n\ $		$\ v - \hat{\Psi}_\rho^n\ $		$\ w - \hat{\Theta}_\rho^n\ $	
	L^2	L^∞	L^2	L^∞	L^2	L^∞
20	3.711	1.902	3.711	1.837	3.644	1.807
25	3.555	2.019	3.555	1.817	3.599	1.716
40	3.790	2.063	3.569	2.067	3.552	1.985

the solution v is calculated, includes two additional cross-diffusion terms. This generates a higher approximation error than in the first and third equations. The lowest error value was recorded for the solution u , although the first equation, through which the solution u is calculated, is similar to the third equation, through which the solution w is calculated. The reason for this is that the form of the analytical solution associated with the solution w has a higher non-linearity than what is the case in the solution u . For the results that represent the error value of the three-dimensional problem, which were presented in Table 5, we were not able to choose the value of the time step $\delta t = 10^{-6}$ because this requires large calculations. Therefore, we chose the value of the diffusion parameters to be in the form $\theta_i = 0.1, \theta_{ii} = 0.001, \theta_{21} = \theta_{23} = 0.01$, where $i = 1, 2, 3$. This selection of diffusion parameter values makes the numerical method stable when choosing the time step $\delta t = 10^{-4}$.

8.3 One-dimensional simulations

In the subsequent numerical results, we performed simulations in one space dimension, utilizing a uniform partition of the domain $\mathfrak{D} = [0, 5]$ consisting of 256 subintervals ($J = 256$). The time step size employed was $\delta t = 10^{-5}$. The initial conditions for the simulations were set as follows: $u(x, 0) = 0.5 + 0.5 \cos(\frac{4\pi x}{5})$, $v(x, 0) = 0.5 - 0.5 \cos(\frac{4\pi x}{5})$, and $w(x, 0) = 0.5 + 0.3 \cos(\frac{4\pi x}{5})$. The parameters used in these experiments are fixed as $\theta_{21} = \theta_{23} = a_i = 1$ for $i = 1, 2, 3$, and $b_2 = b_3 = c_1 = c_3 = d_1 = d_2 = 1$, while $b_1 = c_2 = d_3 = 0$. However, we varied the values of θ_{ii} and θ_i for each experiment. Throughout all the experiments, solutions were graphically represented at various carefully chosen time points to showcase the evolution of the interacting cells as time (t) increases. The selected time points were $T = 0, 0.01, 0.1, 0.5$, and 1, enabling a clear visualization of the system's behavior at different stages of time progression.

In the first experiment, we set $\theta_i = \theta_{ii} = 1$ for $i = 1, 2, 3$, and the numerical solutions of $(Q_{\rho,k}^{h,\delta t})$ are displayed in Figures 1 (a), (b), and (c). A stable state was noticed in the solution. In

the subsequent experiment, we performed a similar investigation, but this time, we altered the values of θ_i . Precisely, we assigned $\theta_i = 10$ and $\theta_{ii} = 1$ for $i = 1, 2, 3$. The outcomes are depicted in Figures 1 (d), (e), and (f). The general pattern remained in line with our earlier observations. An essential discovery in these figures is that the equilibrium solutions were attained more rapidly compared to the first experiment. Next, we conducted a similar experiment, but this time with $\theta_i = 0.1$ and $\theta_{ii} = 1$ for $i = 1, 2, 3$. The results are depicted in Figures 1 (g), (h), and (i). The general behavior of the system remained consistent with the observations made in previous cases. However, we noticed that the stationary solutions were reached at a later time compared to the previous experiments. In the fourth experiment, we selected $\theta_i = 1$ and $\theta_{ii} = 0$ for $i = 1, 2, 3$. The numerical results for various time points are shown in Figures 2 (a), (b), and (c). As in the third experiment, it was observed that reaching the stationary solution took longer compared to the first and second experiments.

8.4 Two-dimensional simulations

To showcase the development of spatial arrangements in the three-species SKT model, we examine a two-dimensional region denoted by $\mathfrak{D} = (0, 5)^2$, which is divided into a mesh containing 80000 triangles. The model incorporates the subsequent parameter values: $\theta_{21} = 0.1, \theta_{23} = 0.2, a_1 = b_1 = 1, a_2 = 0.2, a_3 = 0.15, b_2 = 0.0037, b_3 = 0.0033, c_1 = 0.001, c_2 = 0.25, c_3 = d_2 = 0.0025, d_1 = 0.002$ and $d_3 = 0.1667$. In the experiments, the values of θ_i and θ_{ii} , $i = 1, 2, 3$, are varied to observe their impact on the solutions. The system's stable equilibrium for the chosen parameter values is given by $(u^*, v^*, w^*) = (0.9749, 0.7769, 0.8689)$. The initial conditions are as follows:

$$u(x, 0) = 0.9749 + 0.5 \sin(0.8\pi x) \cos(0.8\pi y),$$

$$v(x, 0) = 0.7769 + 0.2 \cos(0.8\pi x) \sin(0.8\pi y),$$

$$w(x, 0) = 0.8689 + 0.3 \sin(0.8\pi x) \sin(0.8\pi y).$$

The results of all experiments were presented at the following times: $T = 0.5, 1, 2, 3, 6, 9, 12$ and 15 . The values of the self-diffusion coefficients were chosen in Figures 3–6 such that $\theta_{ii} = 0.001$, $i = 1, 2, 3$. The values of diffusion coefficients were selected in Figures 3, 4, 5, and 6 such that $\theta_i = 0.005, 0.001, 0.0005$, and 0.0001 , $i = 1, 2, 3$, respectively. In Figure 3, $\theta_i = 0.005$, it was observed that the solution tended to reach the equilibrium point, as the solution was very close to reaching it at the last time. Then, with the decrease in the values of θ_i , the solution was chosen to the equilibrium point as well, but it became slower. This indicates the significant stabilizing

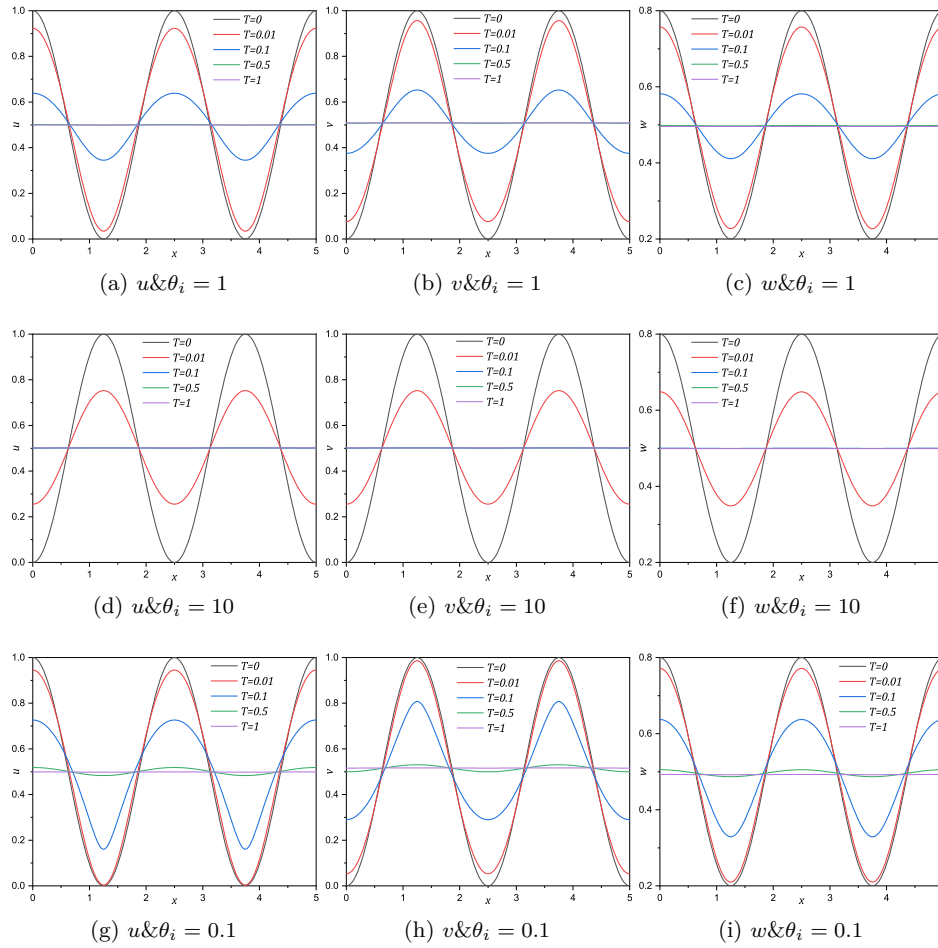


Figure 1: Numerical solutions for the one-dimensional case at different time points for $\theta_{ii} = \theta_{21} = \theta_{23} = a_i = 1$, $i = 1, 2, 3$, $b_2 = b_3 = c_1 = c_3 = d_1 = d_2 = 1$ and $b_1 = c_2 = d_3 = 0$.

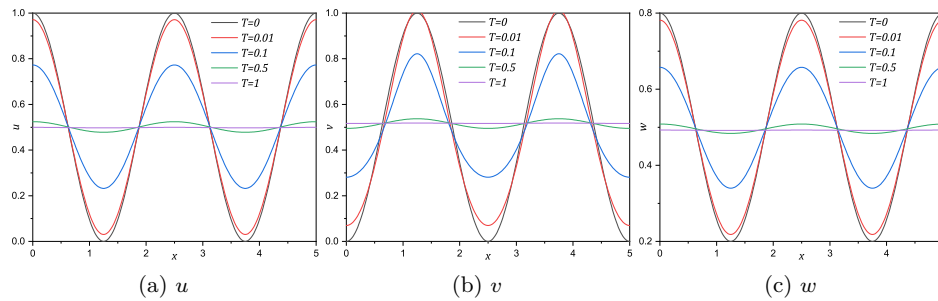


Figure 2: Numerical solutions for the one-dimensional case at different time points for $\theta_{ii} = 0$, $\theta_i = \theta_{21} = \theta_{23} = a_i = 1$, $i = 1, 2, 3$, $b_2 = b_3 = c_1 = c_3 = d_1 = d_2 = 1$ and $b_1 = c_2 = d_3 = 0$.

effect of the diffusion terms. In the last experiment, Figure 7, we fixed the values of diffusion coefficients at $\theta_i = 0.0001$, $i = 1, 2, 3$, and increased the values of self-diffusion coefficients to $\theta_{ii} = 0.01$, $i = 1, 2, 3$ to confirm the effect of the self-diffusion terms on the stability of the system. In this figure, it is observed that the solution moves faster than when the values of the self-diffusion coefficients are $\theta_{ii} = 0.001$, $i = 1, 2, 3$. This confirms the significant stabilizing effect of the self-diffusion terms.

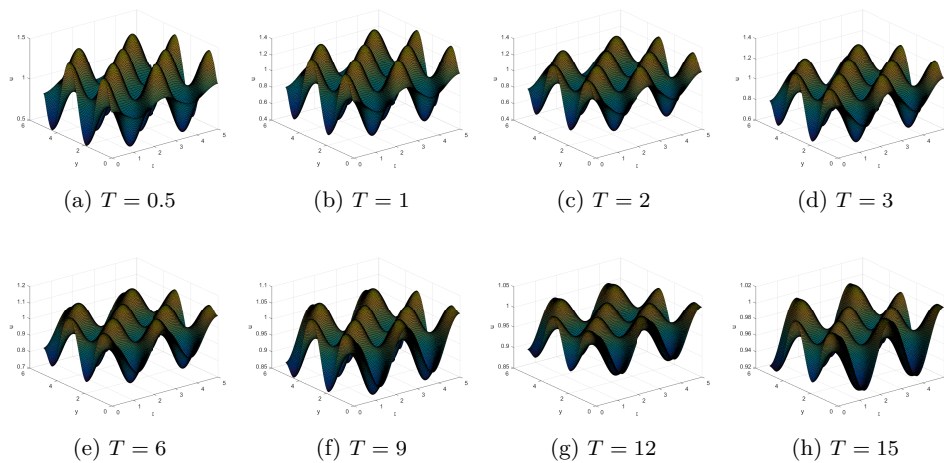


Figure 3: Numerical solution u for the two-dimensional case at different time points for $\theta_i = 0.005$ and $\theta_{ii} = 0.001$, $i = 1, 2, 3$.

9 Conclusions

This article looks into how we can use FEMs to understand the three-species SKT model better. It starts by outlining the main challenges of the model and explains a method to regularize it, ensuring some mathematical proofs to support this. Then, it describes an FEM to get approximate solutions for this adjusted model. The study shows how these approximations link back to the original model, proving that solutions for the SKT model do exist. Finally, it introduces a practical algorithm for solving the approximated problem and presents numerical results. This whole process helps us gain insight into the behavior of the SKT model in real-world scenarios.

It has been found that using FEM is a reliable way to manage complicated boundary conditions and varied spaces. The numerical results showed interesting patterns formed by cross-diffusion, which lined up with what theory suggested about how species can segregate. Overall,

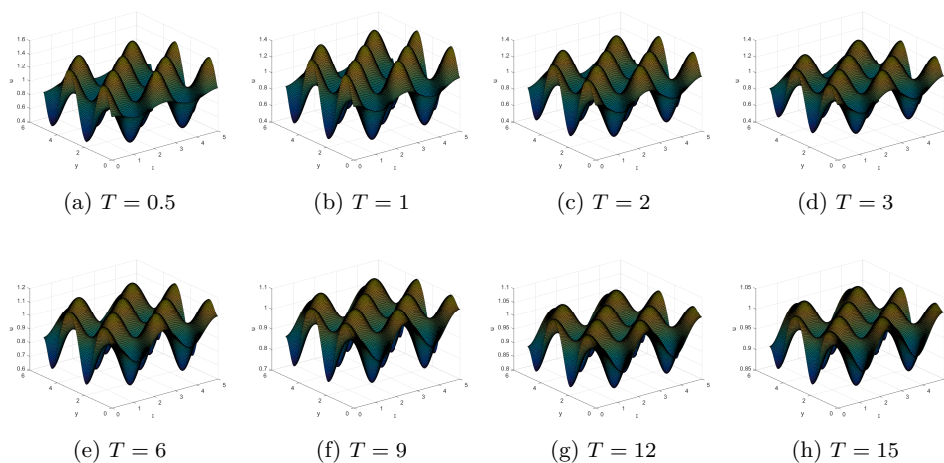


Figure 4: Numerical solution u for the two-dimensional case at different time points for $\theta_i = 0.001$ and $\theta_{ii} = 0.001$, $i = 1, 2, 3$.

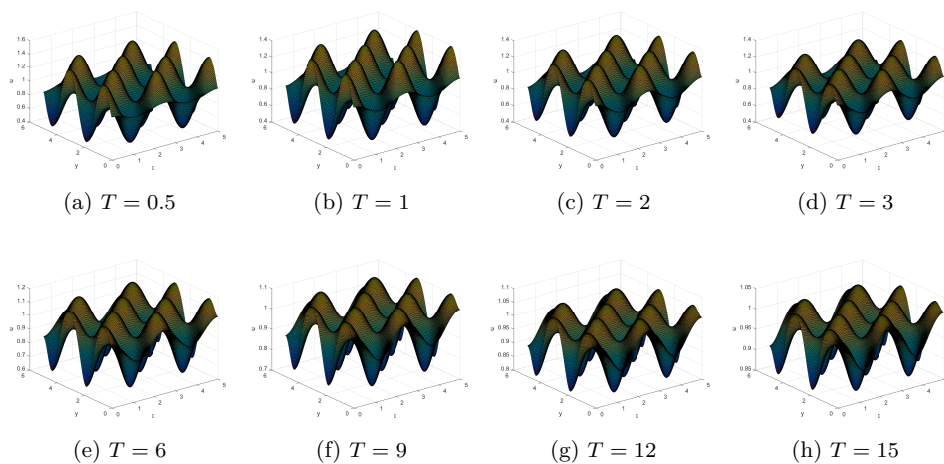


Figure 5: Numerical solution u for the two-dimensional case at different time points for $\theta_i = 0.0005$ and $\theta_{ii} = 0.001$, $i = 1, 2, 3$.

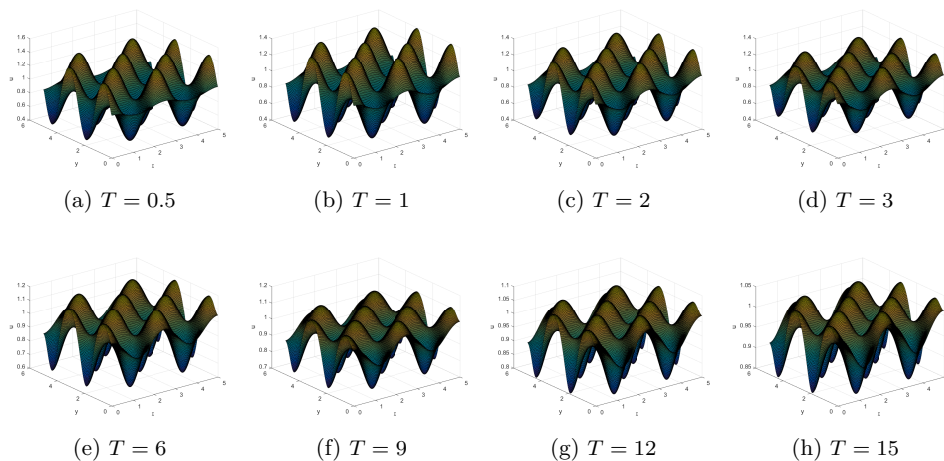


Figure 6: Numerical solution u for the two-dimensional case at different time points for $\theta_i = 0.0001$ and $\theta_{ii} = 0.001$, $i = 1, 2, 3$.

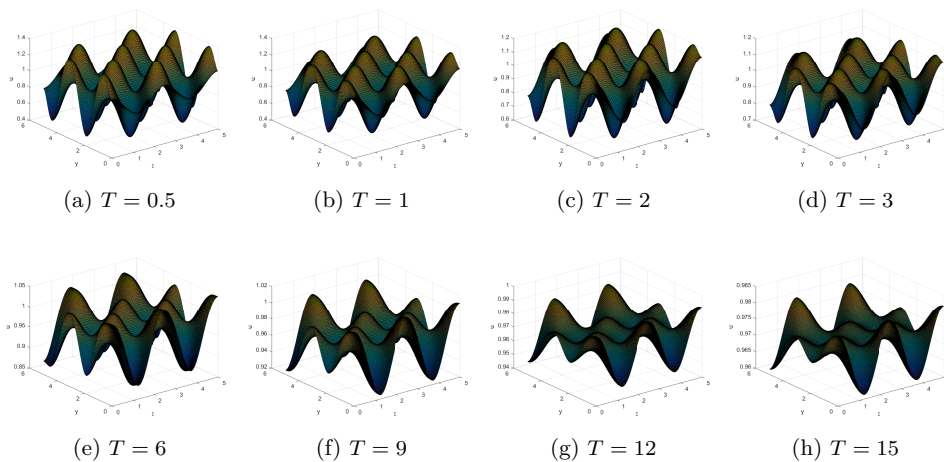


Figure 7: Numerical solution u for the two-dimensional case at different time points for $\theta_i = 0.0001$ and $\theta_{ii} = 0.01$, $i = 1, 2, 3$.

FEM is a great tool for studying ecological and biological systems based on cross-diffusion models, paving the way for more advanced future research.

The SKT cross-diffusion model is a prototype nonlinear parabolic system that depicts intricate diffusion-driven interactions across many species, making its analysis mathematically significant. Any significant numerical or analytical investigation requires that the model be mathematically well-posed, which is ensured by proving the existence, uniqueness, and regularity of weak and strong solutions. Numerical simulations have a strong theoretical basis thanks to the creation and validation of finite element approximations and rigorous proofs of convergence from the discrete weak formulation to the continuous weak solution. These findings advance the mathematical understanding of nonlinear PDEs in general, and cross-diffusion and coupled reaction terms in particular, which are known to pose analytical difficulties because of their nonlinearity and coupling structure.

From a computational perspective, numerical simulations of multicomponent diffusion systems gain trust when the convergence, stability, and correctness of the FEM are demonstrated. The effectiveness and dependability of the numerical method for real-world calculations are validated by the examination of error estimates in one, two, and three spatial dimensions. These results serve as standards for evaluating alternative numerical approaches used in complicated cross-diffusion systems, in addition to supporting the use of numerical approximations for theoretical research. Through simulations, numerical results and patterns can be shown, bridging the gap between mathematical theory and real-world phenomena and offering further insight into the qualitative behavior of the system.

The SKT model is essential to comprehending the spatial segregation and coexistence of competing biological species from a biological perspective. The cross-diffusion terms, which indicate avoidance behavior or territorial competition, explain why individuals of one species tend to migrate away from areas where substantial numbers of another species are present. These processes play a key role in the explanation of how spatial patterns like coexistence zones, segregation, and clustering develop in ecosystems. This kind of research allows for accurate population dynamics simulation by mathematically and numerically testing the model, offering predictive insight into how competition and diffusion mechanisms create ecological structures in genuine ecosystems.

This research is very interdisciplinary since it integrates biological modeling, numerical analysis, and mathematical rigor. It encourages the creation of quantitative tools for environmental scientists, ecologists, and biophysicists who want to simulate spatial population interactions in a range of environmental settings. Furthermore, the approach and findings can be applied to other systems, such as chemical reaction networks, tumor growth modeling, or bacterial colony formation, that are controlled by comparable reaction diffusion cross-diffusion structures. By

strengthening the link between abstract analysis and applied sciences, the study advances both theoretical mathematics and applied biology modeling.

Acknowledgments

We would like to thank two anonymous reviewers for their informative remarks and helpful criticism, which have significantly improved this work.

Funding

This research received no external funding.

Conflicts of Interest

The authors declare no conflicts of interest.

References

- [1] Adams, R.A. and Fournier, J.J., *Sobolev spaces*, Elsevier, 2003.
- [2] Al-Juaifri, G.A. and Harfash, A.J., *Finite element analysis of nonlinear reaction–diffusion system of Fitzhugh–Nagumo type with Robin boundary conditions*, Math. Comput. Simul., 203, (2023), 486–517.
- [3] Al-Juaifri, G.A. and Harfash, A.J., *Numerical analysis of the brusselator model with Robin boundary conditions*, SeMA J., 82, (2025), 351–397.
- [4] Al-Musawi, G.A. and Harfash, A.J., *Finite element analysis of extended Fisher-Kolmogorov equation with Neumann boundary conditions*, Appl. Numer. Math., 201, (2024), 41–71.
- [5] Al-Musawi, G.A. and Harfash, A.J., *Finite element approximation of coupled Cahn-Hilliard equations with a logarithmic potential and nondegenerate mobility*, J. Math. Model., 13(1), (2025), 49–68.

- [6] Al-Musawi, G.A. and Harfash, A.J., *Numerical approximation of the two-species Shigesada–Kawasaki–Teramoto model*, J. Math. Sci., (2025), 1–38.
- [7] Baharlouei, S., Mokhtari, R., and Chegini, N., *Solving two-dimensional coupled Burgers equations via a stable hybridized discontinuous Galerkin method*, Iran. J. Numer. Anal. Optim., 13(3), (2023), 397–425.
- [8] Barrett, J.W. and Blowey, J.F., *Finite element approximation of a nonlinear cross-diffusion population model*, Numer. Math., 98(2), (2004), 195–221.
- [9] Barrett, J.W., Garcke, H., and Nürnberg, R., *Finite element approximation of surfactant spreading on a thin film*, SIAM J. Numer. Anal., 41(4), (2003), 1427–1464.
- [10] Barrett, J.W. and Nürnberg, R., *Finite-element approximation of a nonlinear degenerate parabolic system describing bacterial pattern formation*, Interfaces and Free Boundaries, 4(3), (2002), 277–307.
- [11] Barrett, J.W. and Nürnberg, R., *Convergence of a finite-element approximation of surfactant spreading on a thin film in the presence of van der waals forces*, IMA J. Numer. Anal., 24(2), (2004), 323–363.
- [12] Barrett, J.W., Nürnberg, R., and Warner, M.R., *Finite element approximation of soluble surfactant spreading on a thin film*, SIAM J. Numer. Anal., 44(3), (2006), 1218–1247.
- [13] Barrett, J.W., Schwab, C., and Süli, E., *Existence of global weak solutions for some polymeric flow models*, Math. Models Methods Appl. Sci., 15(06), (2005), 939–983.
- [14] Cazenave, T., *Semilinear Schrödinger Equations*, vol. 10, American Mathematical Soc, 2003.
- [15] Chand, A. and Mohapatra, J., *An analytical and numerical approach for the $(1+1)$ -dimensional nonlinear Kolmogorov–Petrovskii–Piskunov equation*, Iran. J. Numer. Anal. Optim., 15(1), (2025), 79–98.
- [16] Chen, L. and Jüngel, A., *Analysis of a multidimensional parabolic population model with strong cross-diffusion*, SIAM J. Math. Anal., 36(1), (2004), 301–322.
- [17] Chen, L. and Jüngel, A., *Analysis of a parabolic cross-diffusion population model without self-diffusion*, J. Diff. Equ., 224(1), (2006), 39–59.
- [18] Chen, X., Daus, E.S., and Jüngel, A., *Global existence analysis of cross-diffusion population systems for multiple species*, Arch. Rational Mech. Anal., 227, (2018), 715–747.

- [19] Ciarlet, P.G., *The finite element method for elliptic problems*, SIAM, 2002.
- [20] Ciavaldini, J., *Analyse numerique d un problème de stefan à deux phases par une methode d éléments finis*, SIAM J. Numer. Anal., 12(3), (1975), 464–487.
- [21] Galiano, G. and Velasco, J., *Well-posedness of a cross-diffusion population model with non-local diffusion*, SIAM J. Math. Anal., 51(4), (2019), 2884–2902.
- [22] Grün, G. and Rumpf, M., *Nonnegativity preserving convergent schemes for the thin film equation*, Numer. Math., 87(1), (2000), 113–152.
- [23] Hashim, M.H. and Harfash, A.J., *Finite element analysis of a Keller–Segel model with additional cross-diffusion and logistic source. part I: Space convergence*, Comput. Math. Appl., 89(1), (2021), 44–56.
- [24] Hashim, M.H. and Harfash, A.J., *Finite element analysis of a Keller–Segel model with additional cross-diffusion and logistic source. part II: Time convergence and numerical simulation*, Comput. Math. Appl., 109(1), (2022), 216–234.
- [25] Hashim, M.H. and Harfash, A.J., *Finite element analysis of attraction-repulsion chemotaxis system. Part I: Space convergence*, Commun. Appl. Math. Comput., 4(3), (2022), 1011–1056.
- [26] Hashim, M.H. and Harfash, A.J., *Finite element analysis of attraction-repulsion chemotaxis system. Part II: Time convergence, error analysis and numerical results*, Commun. Appl. Math. Comput., 4(3), (2022), 1057–1104.
- [27] Hashim, M.H. and Harfash, A.J., *Finite element analysis for microscale heat equation with Neumann boundary conditions*, Iran. J. Numer. Anal. Optim., 14(3), (2024), 796–832.
- [28] Hashim, M.H. and Harfash, A.J., *The convergence of the finite element approximation to the weak solution of attraction-repulsion chemotaxis system with additional self-diffusion term*, J. Math. Sci., 289(2), (2025), 273–322.
- [29] Hashim, M.H., Hassan, S.M., Hameed, A.A., and Harfash, A.J., *Numerical approximations for damped wave equation with Neumann boundary conditions*, Math. Comput. Sci., 6(1), (2025), 1–18.
- [30] Hassan, S.M. and Harfash, A.J., *Finite element analysis of a two-species chemotaxis system with two chemicals*, Appl. Numer. Math., 182, (2022), 148–175.

- [31] Hassan, S.M. and Harfash, A.J., *Finite element analysis of the two-competing-species Keller–Segel chemotaxis model*, Comput. Math. Model., 33(4), (2022), 443–471.
- [32] Hassan, S.M. and Harfash, A.J., *Finite element approximation of a Keller–Segel model with additional self-and cross-diffusion terms and a logistic source*, Commun. Nonlinear Sci. Numer. Simul., 104, (2022), 106063.
- [33] Hassan, S.M. and Harfash, A.J., *Finite element analysis of chemotaxis-growth model with indirect attractant production and logistic source*, Int. J. Comput. Math., 100(4), (2023), 745–774.
- [34] Jüngel, A. and Zamponi, N., *Qualitative behavior of solutions to cross-diffusion systems from population dynamics*, J. Math. Anal. Appl., 440(2), (2016), 794–809.
- [35] Lions, J.L., *Quelques méthodes de résolution des problèmes aux limites non linéaires*, Paris: Dunod, 1969.
- [36] Ranjan, K.R. and Gowrisankar, S., *Discontinuous Galerkin approach for two-parametric convection-diffusion equation with discontinuous source term*, Iran. J. Numer. Anal. Optim., 14(2), (2024), 347–366.
- [37] Robinson, J.C., *Infinite-dimensional dynamical systems: an introduction to dissipative parabolic PDEs and the theory of global attractors*, vol. 28, Cambridge University Press, 2001.
- [38] Samy, H., Adel, W., Hanafy, I., and Ramadan, M., *A Petrov–Galerkin approach for the numerical analysis of soliton and multi-soliton solutions of the Kudryashov–Sinelshchikov equation*, Iran. J. Numer. Anal. Optim., 14(4), (2024), 1309–1335.
- [39] Shigesada, N., Kawasaki, K., and Teramoto, E., *Spatial segregation of interacting species*, J. Theor. Biol., 79(1), (1979), 83–99.
- [40] Youssri, Y. and Atta, A., *Modal spectral Tchebyshev Petrov–Galerkin stratagem for the time-fractional nonlinear Burgers’ equation*, Iran. J. Numer. Anal. Optim., 14(1) (2024), 172–199.



Characterizing maximum output sets in fractional-order discrete-time linear systems utilizing fractional feedback control

A. EL Bhih*, 

Abstract

This paper examines a linear controlled discrete-time fractional-order system, as defined by

$$\begin{cases} \Delta^\gamma x_{k+1} = Ax_k + B\Delta^\gamma u_k, \\ x_0 = x_0. \end{cases}$$

The corresponding output, denoted as $y_k = Cx_k$, is intended to be stable; that is, $\lim_{k \rightarrow \infty} y_k = 0$.

More precisely, we suppose the existence of a fractional feedback control denoted as $\Delta^\gamma u_k = \Delta^\gamma Lx_k$, where the matrices A , B , C , and L are suitably chosen. The characterization of the maximal output set $\Upsilon(\mathcal{R})$ is investigated by defining the fractional derivative in the Grunwald–Letnikov sense. Specifically, $\Upsilon(\mathcal{R}) = \{x_0 \in \mathbb{R}^n / y_k \in \mathcal{R}, \text{ for all } k \geq 0\}$, where $\mathcal{R} \subset \mathbb{R}^n$ represents a constraint set. It can be shown, through the use of stability and observability hypotheses, that a finite number of equations can derive $\Upsilon(\mathcal{R})$. The maximal output set is determined using an advanced algorithmic approach. In order to clarify the theoretical results, relevant algorithms and numerical simulations have been included.

*Corresponding author

Received 31 July 2025; revised 20 November 2025; accepted 27 November 2025

Amine EL Bhih

Laboratory of Analysis Modeling and Simulation, Department of Mathematics and Computer Science, Faculty of Sciences Ben M'Sik, Hassan II University Casablanca, BP 7955, Sidi Othman, Casablanca, Morocco
Multidisciplinary Research and Innovation Laboratory (LPRI), Moroccan School of Engineering Sciences (EMSI), Casablanca, Morocco. e-mail:elbhihamine@gmail.com

How to cite this article

EL Bhih, A., Characterizing maximum output sets in fractional-order discrete-time linear systems utilizing fractional feedback control. *Iran. J. Numer. Anal. Optim.*, 2026; 16(2): 497–524.
<https://doi.org/10.22067/ijnao.2025.94705.1693>

AMS subject classifications (2020): Primary 93C05, 93B07, 93D20; Secondary 49K24, 52A20, 93B15.

Keywords: Fractional order systems, Stability, Observability, Asymptotically Stable, Discrete-time system.

1 Introduction

Fractional calculus extends the arithmetic of integers by introducing derivatives of noninteger orders. This approach allows for a more precise representation of real systems in engineering and scientific domains. Debnath [4] highlighted the advantages of fractional-order models for capturing the nuanced behavior of systems characterized by long-lived transients and inherited properties, thereby enhancing the accuracy of their description.

Modeling physical phenomena often requires capturing intricate details of preceding histories, which extends beyond the consideration of instant time alone. In this regard, fractional calculus has proven to be invaluable. It is worth noting that recently, there has been substantial interest across a spectrum of disciplines in exploring applications involving fractional derivatives (of noninteger order). Mathematicians and researchers have dedicated their efforts to depicting authentic processes in various fields, such as mechanics, electricity, biology, chemistry, and control systems, by employing models of fractional order. These models can be considered within the context of either continuous or discrete-time representations.

In recent literature, there has been an increasing focus on discrete fractional systems. Notable works in this area include those by Dzieliński and Sierociuk [6], Kaczorek [18, 19], and Ferreira and Torres [12]. Significant advancements in the theory of fractional calculus are outlined by Kilbas, Srivastava, and Trujillo [20] and Oldham and Spanier [24]. Guermah, Djennoune, and Bettayeb [14] extended certain fundamental results on the controllability and observability of discrete-time fractional-order linear systems.

Maximum output sets (MOSs) are important in various practical applications, including stability analysis, stabilizability, and control of constrained systems (refer to [14, 28]). Despite extensive research on this concept, as demonstrated by in-depth investigations in [15, 26] and associated references, its examination within the realm of fractional discrete systems is currently limited. Therefore, further investigation and consideration of this domain is recommended within our scientific community.

The MOS is an important element in various constrained control methodologies, particularly in the implementation of diverse protocols for constraint management. It is defined as the set encompassing all initial states, ensuring that trajectories originating from this set consistently

adhere to the prescribed constraints on outputs at all times. The MOS plays a vital role in navigating constrained control scenarios.

Related Work

The purpose of this paper is to investigate the maximal output admissible set within a specific class of fractional linear systems, utilizing the Grunwald–Letnikov fractional derivative. Valuable insights into the properties and characterizations of the maximal output admissible set in the context of linear systems have been provided by Gilbert and Tan [13]. Moreover, the authors present an algorithm that can be applied to both discrete and continuous-time linear time-invariant systems. This research is intended for scientists and researchers who are looking for a detailed comprehension of the intricate dynamics inherent in fractional linear systems. The maximal output set has recently been characterized for observability and asymptotically stable systems in the domain of linear discrete-time systems (see [27, 3, 2]). This conceptual advancement significantly deepens our understanding of the analysis of constrained control systems. Additionally, it has practical utility in the field of control system design techniques, as highlighted in [21]. Recent attention has been given to exploring the established maximal output set for a specific class of semilinear systems and bilinear systems, as noted in [10, 1], respectively. The focus of these studies is to characterize the MOS for such systems, using a stability hypothesis. The authors have introduced a novel sufficient condition to ensure the asymptotic stability of the system. The maximal set was characterized using a finite number of inequalities. Furthermore, an algorithmic approach was proposed to determine the admissibility index k_0^* . The effectiveness of this methodology was demonstrated through several numerical examples. In [23], the authors explored infinite-dimensional linear systems and provides a characterization of the set of gain operators that can make the system insensitive to uncertain initial states. Further insights into the exploration of the maximal output set can be gleaned from additional references in [8, 9]. The work of Hirata and Ohta [16] in the realm of nonlinear systems is worthy of recognition. Their study delineates both necessary and sufficient conditions for meeting prescribed state and control constraints within a distinct class of nonlinear systems. Additionally, the authors propose a precise methodology for determining the maximal output admissible set. The significance of fractional-order derivatives in the modeling and analysis of biological phenomena has been emphasized in several recent studies [31, 29, 30, 17, 11, 22].

Statement problem

The objective of this study is to explore fractional-order discrete-time linear systems under fractional feedback control, as represented by the system dynamics expressed in (1). The focus is on understanding the properties of the MOS associated with such systems.

The study aims to contribute novel insights into this area, motivated by the lack of comprehensive research on the MOS for fractional-order systems. We propose a methodology to construct the MOS using a finite set of inequalities informed by stability and observability considerations.

Additionally, we aim to establish new sufficient conditions that ensure the finite determination of the MOS. These conditions are crucial for clarifying the system's behavior and may inform the development of more effective control strategies. It is worth noting that this study solely focuses on characterizing the MOS and does not consider the existence of fractional feedback control.

The central objective of this study is to investigate fractional-order discrete-time linear systems employing fractional feedback control, characterized by

$$\begin{cases} \Delta^\gamma x_{k+1} = Ax_k + B\Delta^\gamma u_k, \\ x_0 = x_0. \end{cases} \quad (1)$$

The associated output is $y_k = Cx_k$.

More precisely, we suppose the existence of a fractional feedback control

$$\Delta^\gamma u_k = L\Delta^\gamma x_k. \quad (2)$$

The paper unfolds in a structured manner to comprehensively address the research objectives. In Section 2, we provide a thorough examination of the fundamental definition of fractional derivatives within the Grunwald–Letnikov framework, along with an in-depth analysis of the discrete-time model discussed in [5]. Section 3 is dedicated to the characterization of the maximal output set, accompanied by the establishment of sufficient conditions to ensure its finite determination. Building upon this, Section 4 introduces conditions aimed at representing the maximal output set using a finite number of inequalities. In Section 5, we propose a conceptual algorithm for determining the output admissibility index. The subsequent section presents numerical examples to illustrate and validate the theoretical findings. Finally, the concluding section synthesizes the research outcomes and provides a concise summary of the study's contributions.

2 Analysis of dynamic models through fractional calculus

This section presents basic concepts related to fractional calculus that are utilized throughout the paper. In this section of the paper, the Grünwald–Letnikov method [24, 25] defines the discrete fractional derivative, while also defining a set of admissible outputs.

Definition 1. The Letnikov and Grunwald fractional difference operator, denoted by Δ^γ , acting on the function $x(\cdot)$ at the discrete time instance k within the interval $[0, +\infty[$, is defined as

$$\Delta^\gamma x(k) = \frac{1}{h^\gamma} \sum_{j=0}^k (-1)^j \binom{\gamma}{j} x(k-j). \quad (3)$$

The derivative approximation is calculated for a sampling period of unity, where the order of the derivative is between 0 and 1, and k is the number of samples.

The term $\binom{\gamma}{j}$ in Definition 1 can be obtained from the following relation:

$$\binom{\gamma}{j} = \begin{cases} 1 & \text{for } j = 0, \\ \frac{\gamma(\gamma-1)\cdots(\gamma-j+1)}{j!} & \text{for } j > 0. \end{cases} \quad (4)$$

The first-order difference for $x(k+1)$ defined as

$$\begin{aligned} \Delta^1 x(k+1) &= x(k+1) - x(k) \\ &= Ax_k + BL\Delta^1 x_k - x_k \\ &= (A - I_n)x_k + BL(x_k - x_{k-1}) \\ &= (A - I_n)x_k + BL(x_k - x_{k-1}) \\ &= (A + BL - I_n)x_k - BLx_{k-1} \\ &= A_d x_k - BLx_{k-1}, \end{aligned}$$

with $A_d = A + BL - I_n$.

We have

$$\Delta^\gamma x_{k+1} = x_{k+1} + \sum_{j=1}^{k+1} (-1)^j \binom{\gamma}{j} x(k-j+1)$$

and

$$\Delta^\gamma x_k = x_k + \sum_{j=0}^k (-1)^j \binom{\gamma}{j} x(k-j). \quad (5)$$

Then the fractional feedback control is given by

$$\Delta^\gamma u_k = L \sum_{j=0}^k (-1)^j \binom{\gamma}{j} x(k-j). \quad (6)$$

We replace $\Delta^\gamma x_k$ and $\Delta^\gamma u_k$ by its value, system (6) can be rewritten by

$$\begin{aligned} x_{k+1} &= A_d x_k - \sum_{j=1}^{k+1} (-1)^j \binom{\gamma}{j} x(k-j+1) + BL \sum_{j=1}^k (-1)^j \binom{\gamma}{j} x(k-j) \\ &= \left(A_d - \binom{\gamma}{1} \right) x_k - \sum_{j=1}^{k+1} (-1)^j \binom{\gamma}{j} x(k-j+1) \\ &\quad + BL \sum_{j=1}^k (-1)^j \binom{\gamma}{j} x(k-j) \\ &= \tilde{A} x_k - \sum_{j=2}^{k+1} (-1)^j \binom{\gamma}{j} x(k-j+1) + BL \sum_{j=1}^k (-1)^j \binom{\gamma}{j} x(k-j). \end{aligned} \quad (7)$$

Set $a_j = (-1)^j \binom{\gamma}{j}$ and $b_j = (-1)^j \binom{\gamma}{j}$. Equation (7) can be rewritten as

$$x_{k+1} = \tilde{A} x_k - \sum_{j=2}^{k+1} a_j x(k-j+1) + BL \sum_{j=1}^k b_j x(k-j).$$

Let us now write

$$A_0 = A_d - a_1 I_n \quad \text{and} \quad B_0 = b_1 I_n,$$

and further, for all $j \geq 0$,

$$A_j = -a_{j+1} I_n \quad \text{and} \quad B_j = b_{j+1} I_n,$$

which leads

$$\begin{cases} x_{k+1} = \sum_{j=0}^k A_j x_{i-j} + BL \sum_{j=0}^k B_j x_{i-j} \\ \quad = \sum_{j=0}^k (A_j + BL B_j) x_{i-j}, \\ y_k = C x_k, \end{cases}$$

and u_k is the control law (feedback control) defined by

$$\Delta^\gamma u_k = L x_k = L \sum_{j=0}^k B_j x_{k-j},$$

where

$$\begin{aligned} A_0 &= A + \binom{\gamma}{1} I_n \quad \text{and} \\ A_j &= -(-1)^{j+1} \text{diag} \left\{ \binom{\gamma}{j+1}, \dots, \binom{\gamma}{j+1} \right\}, \quad \text{for all } j \geq 1. \\ B_j &= (-1)^j \text{diag} \left\{ \binom{\gamma}{j}, \dots, \binom{\gamma}{j} \right\}, \quad \text{for all } j \geq 0. \end{aligned} \quad (8)$$

Then, putting $\tilde{A}_{B_j} = A_j + BLB_j$, we can write

$$x_{k+1} = \sum_{j=0}^k \tilde{A}_{B_j} x_{k-j}.$$

Thus, successively, we build the recurrence beginning with

$$x(1) = \tilde{A}_{B_0} x_0 = (A_0 + BLB_0) x_0 = \left(A + \binom{\gamma}{1} I_n + BL \right) x_0,$$

$$x(2) = \tilde{A}_{B_1} x_1 = (A_0 + BLB_0) x_1 + (A_1 + BLB_1) x_0, \dots$$

This leads to the following relation between the two column sequences:

$$\begin{bmatrix} x(1) \\ x(2) \\ x(3) \\ \vdots \\ x(k+1) \\ \vdots \end{bmatrix} = \begin{bmatrix} A_0 + BLB_0 & 0 & 0 & 0 \cdots \\ A_1 + BLB_1 & A_0 + BLB_0 & 0 & 0 \cdots \\ A_2 + BLB_2 & A_1 + BLB_1 & A_0 + BLB_0 & 0 \cdots \\ \vdots & \vdots & \vdots & \vdots \vdots \\ \vdots & \vdots & \vdots & \vdots \vdots \\ \vdots & \vdots & \vdots & \vdots \vdots \end{bmatrix} \begin{bmatrix} x(0) \\ x(1) \\ x(2) \\ \vdots \\ x(k) \\ \vdots \end{bmatrix}.$$

The state space model below represents the discrete-time fractional-order system:

$$\begin{cases} x_{k+1} = \sum_{j=0}^k \tilde{A}_{B_j} x_{k-j}, \\ x(0) = x_0 \in \mathbb{R}^n, \\ y_k = Cx_k. \end{cases} \quad (9)$$

Remark 1. (i) It can be inferred from (9) that the fractional system exhibits a resemblance to a system characterized by an increasing number of delays.

(ii) Equations (8) – (9) suggest that the coefficients of A_j and B_j become negative when γ and γ lie in the interval $(0, 1)$, and their magnitudes decrease rapidly as j increases.

For practical applicability, it is necessary to limit the number of considered samples to a predefined quantity l , which is referred to as the memory length [7].

In this case the system (9) is rewritten as

$$\begin{cases} x_{k+1} = \sum_{j=0}^l \tilde{A}_{B_j} x_{k-j}, \\ x(0) = x_0 \in \mathbb{R}^n, \\ y_k = Cx_k. \end{cases} \quad (10)$$

Let us define the transition matrix form N_{B_k} such that

$$N_{B_k} = \begin{cases} I_{n_d} & \text{for } k = 0, \\ \sum_{j=0}^l \tilde{A}_{B_j} N_{B((k-j-1))} & \text{for } k \geq 1. \end{cases}$$

The general solution of (10) can be written as follows:

$$x(k) = N_{B_k} x_0, \quad \text{for all } k \geq 0.$$

Remark 2. If $\gamma = 1$, then

$$x_k = (A + I + BL)^k x_0.$$

Definition 2. Let $\mathcal{R} \subset \mathbb{R}^p$ and let $x_0 \in \mathbb{R}^n$. The initial state x_0 is said $\mathcal{R}$ -output admissible if

$$y_k \in \mathcal{R}, \quad \text{for all } k \in \mathbb{N}.$$

The set of all such initial states is described by

$$\Upsilon(\mathcal{R}) = \{x_0 \in \mathbb{R}^n / Cx_0 \in \mathcal{R}, CN_{B_k}x_0 \in \mathcal{R}, \text{ for all } k \geq 1\},$$

or

$$\Upsilon(\mathcal{R}) = \left\{ x_0 \in \mathbb{R}^n / Cx_0 \in \mathcal{R}, C \sum_{j=0}^l \tilde{A}_{B_j} N_{B(k-j-1)} x_0 \in \mathcal{R}, \text{ for all } k \geq 1 \right\}.$$

This manuscript presents a new approach to the identification of the set of initial states that satisfy point-in-time criteria for the resultant output function, subject to certain conditions on the matrix B_k .

3 A characterization approach to $\Upsilon(\mathcal{R})$: Output admissible set analysis

The main goal of this section is to characterize, under certain assumptions, the set $\Upsilon(\mathcal{R})$ of all initial states that lead to trajectories satisfying the constraint set. We establish the finite determinate of $\Upsilon(\mathcal{R})$ and, as a consequence, devise an algorithmic procedure for its computation. To achieve this, we introduce the sets $\Upsilon_k(\mathcal{R})$ for each $k \geq 0$ and provide a detailed description of their properties as follows:

$$\Upsilon_k(\mathcal{R}) = \{x_0 \in \mathbb{R}^n / Cx_0 \in \mathcal{R}, CN_{B_k}x_0 \in \mathcal{R}, \text{ for all } k \in \{1, \dots, i\}\}.$$

Definition 3. Assuming the existence of an integer k^* for which $\Upsilon_{k^*}(\mathcal{R})$ equals $\Upsilon(\mathcal{R})$, the set $\Upsilon(\mathcal{R})$ is deemed to be finitely determined.

Definition 4. The dynamical system defined by (10) is acknowledged as asymptotically stable provided that for every $k \geq 1$

$$\lim_{k \rightarrow \infty} y_k = 0.$$

Consequently, the system defined by (10) achieves asymptotic stability if and only if

$$\|N_{B_k}\| \leq 1, \quad \text{for } k \geq 1.$$

By imposing specific conditions on N_{B_k} , for $k \geq 0$ and $\mathcal{R}$, we inherently introduce corresponding constraints on $\Upsilon(\mathcal{R})$. The following result presupposes that the interior of $\mathcal{R}$ is considered; a condition that is typically met in practical applications and has favorable consequences (see [13] for further insights).

Proposition 1. Should the system described by (10) achieve asymptotic stability, along with $0 \in \overset{\circ}{\mathcal{R}}$ and the norm of matrix C being bounded by 1. Then

$$0 \in \overset{\circ}{\Upsilon}(\mathcal{R}).$$

Remark 3. In the following proof, we denote the open ball centered at 0 with radius η as $\mathcal{B}_\eta(0)$. This notation represents the set of all points within a distance η from the origin.

Proof. Suppose that 0 belongs to the interior of $\overset{\circ}{\mathcal{R}}$. Consequently, there exists $\eta > 0$ such that the ball $\mathcal{B}_\eta(0)$ is contained within $\Upsilon(\mathcal{R})$.

If the system described by Equation (10) is asymptotically stable, then $\|N_{B_k}\| \leq 1$, for all $k \geq 0$.

Let z be an element of $\mathcal{B}_\eta(0)$ and let $k \geq 0$. We observe

$$\begin{aligned}\|CN_{B_k}z\| &\leq \|C\| \|N_{B_k}\| \|z\| \\ &\leq \|C\| \|N_{B_k}\| \eta \\ &\leq \eta.\end{aligned}$$

This implies that $CN_{B_k}z$ lies within $\mathcal{B}_\eta(0)$ for every z in $\mathcal{B}_\eta(0)$ and for all $k \geq 0$. Consequently, we have

$$CN_{B_k}z \in \mathcal{R}, \quad \text{for all } z \in \mathcal{B}_\eta(0), \text{ for all } k \geq 0.$$

This inclusion yields

$$\mathcal{B}_\eta(0) \subset \Upsilon(\mathcal{R}).$$

Thus, we conclude that $0 \in \mathring{\Upsilon}(\mathcal{R})$. □

Proposition 2. If $\mathcal{R}$ is convex, symmetry, and closed, then the set $\Upsilon(\mathcal{R})$ is also

- (i) Convex,
- (ii) Symmetric,
- (iii) Closed.

Proof. (i) Let $z_0^1, z_0^2 \in \Upsilon(\mathcal{R})$, $\theta \in]0, 1[$, and $k \geq 0$. We show that $\theta z_0^1 + (1 - \theta) z_0^2 \in \Upsilon(\mathcal{R})$ and $z_0^1, z_0^2 \in \Upsilon(\mathcal{R})$ implies that $z_0^1, z_0^2 \in \mathbb{R}^n$ such that $CN_{B_k}z_0^1 \in \mathcal{R}$ and $CN_{B_k}z_0^2 \in \mathcal{R}$.

We have

$$CN_{B_k}(\theta z_0^1 + (1 - \theta) z_0^2) = \theta CN_{B_k}z_0^1 + (1 - \theta) CN_{B_k}z_0^2.$$

We assume that $\mathcal{R}$ is convex, and that $CN_{B_k}(\theta z_0^1 + (1 - \theta) z_0^2) \in \mathcal{R}$ and $\theta z_0^1 + (1 - \theta) z_0^2 \in \Upsilon(\mathcal{R})$.

Then $\Upsilon(\mathcal{R})$ is convex.

(ii) From the definition of $\Upsilon(\mathcal{R})$, it is clear.

(iii) We define the function G_k by

$$\begin{aligned}G_k : \mathbb{R}^n &\rightarrow \mathbb{R}^p \\ z &\rightarrow CN_{B_k}z.\end{aligned}$$

Then $(G_k)_{k \geq 0}$ are continuous.

Let $G_k^{-1}(\mathcal{R}) = \{z \in \mathbb{R}^n / CN_{B_k}z \in \mathcal{R}\}$. Then

$$\Upsilon(\mathcal{R}) = \bigcap_{k \geq 0} G_k^{-1}(\mathcal{R}).$$

Since $(G_k)_{k \geq 0}$ are continuous and $\mathcal{R}$ is closed, $G_k^{-1}(\mathcal{R})$, $k \geq 0$ are closed.

Hence, $\Upsilon(\mathcal{R})$ is proven to be closed. Moreover, in the case where $\mathcal{R}$ exhibits symmetry, it follows that $\Upsilon(\mathcal{R})$ also inherits this property of symmetry.

Thus, the demonstration is concluded. □

The imposition of specific requirements on both N_{B_k} (where $k > 0$) and $\mathcal{R}$ naturally extends to analogous conditions for $\Upsilon(\mathcal{R})$. The ensuing outcome presupposes the inclusion of the origin, 0, within the interior of $\mathcal{R}$, denoted as $\overset{\circ}{\mathcal{R}}$. This assumption is not only consistent with most practical applications but also yields favorable outcomes, as elaborated in reference [13].

Proposition 3. Under the condition that $\Upsilon(\mathcal{R})$ is determined to be finite, there exists a natural number k such that $\Upsilon_k(\mathcal{R}) = \Upsilon_{k+1}(\mathcal{R})$.

Proof. Let us assume that $\Upsilon(\mathcal{R})$ is determined to be finite. Then, there exists $(k \in \mathbb{N})$, such that $\Upsilon(\mathcal{R}) = \Upsilon_k(\mathcal{R})$. On the other hand, $\Upsilon_k(\mathcal{R}) = \Upsilon(\mathcal{R}) \subset \Upsilon_k(\mathcal{R})$. Since $\{\Upsilon_k(\mathcal{R})\}_{k \geq 0}$ is decreasing sequence then $\Upsilon_{k+1}(\mathcal{R}) \subset \Upsilon_k(\mathcal{R})$.

As a result, we have $\Upsilon_k(\mathcal{R}) = \Upsilon_{k+1}(\mathcal{R})$, for some $k \geq 0$, thereby concluding the proof. □

We now introduce an intriguing result that enables us to determine the set $\Upsilon(\mathcal{R})$ through a finite number of inequalities. This, in turn, paves the way for the development of an algorithmic approach to calculate the admissibility index k^* .

In the following result, the set is taken as $\mathcal{R} = \{y \in \mathbb{R}^p / \|y\| \leq \xi\}$.

In our pursuit of a solution, we explore two distinct scenarios:

a) In the first scenario, we assume that $\dim \mathcal{R} = n$, signifying that both the observation space and the state space exhibit identical dimensions.

b) Under the second scenario, we assume that $\dim \mathcal{R} < n$, indicating a discrepancy in dimensions between the observation space and the state space.

Considering the first scenario where $\dim \mathcal{R} = n$, the matrix C takes the form of an $n \times n$ matrix.

Proposition 4. Assume that the following conditions are met:

Condition (i) $\sum_{j=0}^l \|\tilde{A}_{B_j}\| \leq 1$ and $\Upsilon_k(\mathcal{R}) = \Upsilon_{k+1}(\mathcal{R})$ for some k ,

Condition (ii) C commutes with $\tilde{A}_{B_j}$ for all $0 \leq j \leq L$.

Then, it follows that $\Upsilon(\mathcal{R})$ is finitely determined.

Proof. It is evident that $\Upsilon(\mathcal{R}) \subset \Upsilon_k(\mathcal{R})$.

Assume $x_0 \in \Upsilon_k(\mathcal{R})$. Then $z_k = CN_{B_k}x_0 \in \mathcal{R}$ for all $k \in \{0, 1, 2, \dots, i+1\}$, implying $\|z_k\| \leq \xi$ for all $k \leq i+1$.

For $k = i+2$, we have

$$\begin{aligned} \|z_k\| &= \|CN_{B_k}x_0\| \\ &= \left\| C \sum_{j=0}^l \tilde{A}_{B_j} N_{B((k-j-1))} x_0 \right\| \\ &= \left\| C \sum_{j=0}^l \tilde{A}_{B_j} N_{B((k-j-1))} x_0 \right\| \\ &= \left\| \sum_{j=0}^l \tilde{A}_{B_j} CN_{B((k-j-1))} x_0 \right\| \\ &\leq \sum_{j=0}^l \left\| \tilde{A}_{B_j} \right\| \|CN_{B((k-j-1))} x_0\| \\ &\leq \xi \sum_{j=0}^l \left\| \tilde{A}_{B_j} \right\|. \end{aligned}$$

Given that $\|CN_{B((k-j-1))} x_0\| \leq \xi$ holds for all $j \in \{0, \dots, L\}$, we employ the inequality $\sum_{j=0}^l \left\| \tilde{A}_{B_j} \right\| < 1$ to deduce that $\|CN_{B((k+2))} x_0\| \leq \xi$.

Through iterative reasoning, we establish that $\|CN_{B((0k+j))} x_0\| \leq \xi$ for all $j \geq 2$.

This implies $\|CN_{B_i} x_0\| \leq \xi$ for all $i \geq k+2$, consequently leading to $\|CN_{B_i} x_0\| \leq \xi$ for all $i \geq 0$.

As a result, $x_0 \in \Upsilon(\mathcal{R})$, thereby confirming the finiteness of $\Upsilon(\mathcal{R})$. □

In the second scenario where $\dim \mathcal{R} = p < n$, considering that the matrix $C \in \mathcal{L}(\mathbb{R}^n, \mathbb{R}^p)$, let us define $\tilde{C}$ and $\tilde{\mathcal{R}}$ as follows:

$$\tilde{C} = \begin{pmatrix} C \\ 0 \end{pmatrix} \in \mathcal{L}(\mathbb{R}^n, \mathbb{R}^n), \quad \tilde{\mathcal{R}} = \mathcal{R} \times \{0_{\mathbb{R}^{n-p}}\} \subset \mathbb{R}^n.$$

Now, let us explore the implications of the new observation $\tilde{y}_i = \tilde{C}x_i$. Upon careful examination, it is evident that for every integer i ,

$$y_i \in \mathcal{R} \iff \tilde{y}_i \in \tilde{\mathcal{R}}.$$

Remark 4. It is interesting to note that the output admissibility of an initial state, denoted by x_0 , regarding C and $\mathcal{R}$, is equivalent to its output admissibility regarding $\tilde{C}$ and $\tilde{\mathcal{R}}$.

The findings from the initial case inform the derivation of the subsequent statement, given that $\dim \tilde{\mathcal{R}} = n$.

Proposition 5. Assume that the following conditions are met:

Condition (i) $\sum_{j=0}^l \|\tilde{A}_{B_j}\| \leq 1$ and $\Upsilon_k(\mathcal{R}) = \Upsilon_{k+1}(\mathcal{R})$ for some k ,

Condition (ii) $\tilde{C}$ commutes with $\tilde{A}_{B_j}$ for all $0 \leq j \leq L$.

Then, it follows that $\Upsilon(\mathcal{R})$ is finitely determined.

In section 4, we will present an algorithmic method designed to determine the minimum integer value, denoted as i^* , where $\Upsilon(\mathcal{R}) = \Upsilon_{k^*}(\mathcal{R})$.

4 Algorithmic determination

Drawing directly from the implications of Propositions 3 and 4, this section outlines a systematic approach to ascertain k^* , the admissibility index, and consequently, the set $\Upsilon(\mathcal{R})$.

Let $\mathbb{R}^p$ be endowed with the following norm:

$$\|x\|_\infty = \max_{1 \leq i \leq p} |x_i|, \quad \text{for all } x \in \mathbb{R}^p.$$

Define $\mathcal{R}$ as $\mathcal{R} = \{y \in \mathbb{R}^p | h_j(y) \leq 0, j = 1, 2, \dots, 2p\}$, where $h_j : \mathbb{R}^p \rightarrow \mathbb{R}$ are given functions. Such sets hold particular relevance in practical scenarios.

In this instance, $\Upsilon_k(\mathcal{R})$ is described for every integer i as

$$\Upsilon_k(\mathcal{R}) = \{x_0 \in \mathbb{R}^n | h_j(CN_{B_i}x) \leq 0, j = 0, \dots, 2p, i = 0, \dots, k\}.$$

On the other hand,

$$\begin{aligned} \Upsilon_{k+1}(\mathcal{R}) &= \{x_0 \in \Upsilon_k(\mathcal{R}) / CN_{B_{k+1}}x_0 \in \mathcal{R}\} \\ &= \{x_0 \in \Upsilon_k(\mathcal{R}) / h_j(CN_{B_{k+1}}x_0) \leq 0 \text{ for } j = 1, \dots, 2p\}. \end{aligned}$$

Now, since $\Upsilon_{k+1}(\mathcal{R}) \subset \Upsilon_k(\mathcal{R})$ for every k , then

$$\begin{aligned} \Upsilon_{k+1}(\mathcal{R}) = \Upsilon_k(\mathcal{R}) &\iff \Upsilon_k(\mathcal{R}) \subset \Upsilon_{k+1}(\mathcal{R}) \\ &\iff \text{for all } x_0 \in \Upsilon_k(\mathcal{R}), \quad h_j(CN_{B_{k+1}}x_0) \leq 0, \end{aligned}$$

$$\begin{aligned}
& \text{for all } j = \{1, \dots, 2p\} \\
& \iff \sup_{\substack{x_0 \in \mathbb{R}^n, h_k(CN_{B_l}x_0) \leq 0 \\ \text{for all } k=\{1, \dots, 2p\} \text{ for all } l \in \{0, \dots, i\}}} h_j(CN_{B_{k+1}}x_0) \leq 0, \\
& \text{for all } j = \{1, \dots, 2p\}.
\end{aligned}$$

As a result, the equivalence $\Upsilon_{k+1}(\mathcal{R}) = \Upsilon_k(\mathcal{R})$ generates a collection of mathematical programming problems.

Algorithm 1 Determination of k^*

Require: $n, p, l \in \mathbb{N}^*, C, N_{B_i}, \xi > 0$

$k \leftarrow 0$

For $j = 1$ **to** p **Maximize** $J_j(x) = |(CN_{B_{k+1}}x)_j| - \xi$

Subject to the constraints:

$$\begin{cases} |(CN_{B_{k+1}}x)_j| - \xi \\ \text{for all } k = \{1, \dots, 2p\} \quad \text{for all } l \in \{0, \dots, i\} \end{cases} \quad J_j^* \leftarrow \max \{J_j(x)\}$$

If $J_j^* \leq 0$ **for all** $j = 1, \dots, p$, **then** $k^* \leftarrow 0$

To showcase our findings, we will provide concrete numerical examples in section 6.

5 Criteria ensuring finite determination of $\Upsilon_k(\mathcal{R})$

The theorem presented below represents the pivotal result in our endeavor, focusing on the notion that having straightforward conditions to attain the finite determination of the set $\Upsilon(\mathcal{R})$ is highly preferable.

Theorem 1. If $\|N_{B_k}\| \leq \gamma_k$, for all $k \geq 0$, with $\gamma_k \rightarrow 0$ when $k \rightarrow \infty$, then $\Upsilon(\mathcal{R})$ is finitely determined.

Proof. We have $\|N_{B_k}\| \leq \gamma_k \implies \|CN_{B_k}x\| \leq \gamma_k \|C\| \|x\|$, $\gamma_k \rightarrow 0$, when $k \rightarrow \infty$. Then there exists an integer i_0 such that $\|CN_{B_k}x\| \leq \varepsilon$, for all $k \geq i_0$, for all $x \in \mathbb{R}^n$.

Hence $\|CN_{B_{(i_0+1)}}x\| \leq \varepsilon$, for all $x \in \Upsilon_{k_0}(\mathcal{R})$, and we have $\Upsilon(\mathcal{R}) \subset \Upsilon_{k_0+1}(\mathcal{R}) \subset \Upsilon_{k_0}(\mathcal{R})$.

Consequently, $\Upsilon(\mathcal{R}) = \Upsilon_{k_0+1}(\mathcal{R}) = \Upsilon_{k_0}(\mathcal{R})$.

Then $\Upsilon(\mathcal{R})$ is finitely determined. □

6 Numerical examples

In order to clarify our outcomes, we showcase numerical instances, utilizing a carefully crafted algorithm. This approach allows us to precisely define the set $\Upsilon(\mathcal{R})$ through a finite set of inequalities, providing substantial support for our underlying hypothesis $\sum_{j=0}^l \|\tilde{A}_{B_j}\| \leq 1$ holds in all instances presented. We strategically choose the matrix $A_0 = A + BL + \gamma I_n$ and $\tilde{A}_{B_0} = A_0 + BLB_0$.

We have

$$\sum_{j=0}^l (-1)^j \binom{\gamma}{j} = \frac{\Gamma(l+1-\gamma)}{\Gamma(1-\gamma) \Gamma(l+1)},$$

$$\sum_{j=0}^l (-1)^j \binom{\gamma}{j} = \frac{\Gamma(l+1-\gamma)}{\Gamma(1-\gamma) \Gamma(l+1)},$$

and the fact

$$\sum_{j=0}^l \|\tilde{A}_{B_j}\| = \sum_{j=0}^l \|A_j + BLB_j\| = \|A_0 + BLB_0\| + \sum_{j=1}^l \|A_j + BLB_j\|.$$

Example 1. Consider the following system:

$$\begin{cases} x(k+1) = \sum_{j=0}^l \tilde{A}_{B_j} x_{k-j}, \\ x_0 = x(0) \in \mathbb{R}^n. \end{cases}$$

The matrices A_j and B_j are defined as $A_j = -a_{j+1}I_n$ and $B_j = b_{j+1}I_n$, where $a_j = (-1)^j \binom{\gamma}{j}$ and $b_j = (-1)^j \binom{\gamma}{j}$.

The parameters γ , γ , ε , n , and the matrices A , B , L , and C are specified as follows: $\gamma = 0.25$, $\varepsilon = 0.5$, $\gamma = 0.25$, $n = 3$, $C = \begin{pmatrix} 1 & -1 \end{pmatrix}$, $A = \begin{pmatrix} \frac{1}{54} & \frac{1}{16} \\ -\frac{1}{32} & \frac{1}{12} \end{pmatrix}$, $B = \begin{pmatrix} \frac{1}{12} \\ \frac{1}{18} \end{pmatrix}$, and $L = \begin{pmatrix} \frac{1}{2} & \frac{1}{2} \end{pmatrix}$.

Then

$$\begin{aligned} N_{B_0} &= I_2 = \begin{pmatrix} 1 & 0 \\ 0 & 1 \end{pmatrix}, \\ N_{B_1} &= A_{B_0} = A_0 + BLB_0 = A + BL + I_2 + BLB_0 = A + I_2 + BL(I_2 + B_0) \\ &= \begin{pmatrix} \frac{1}{54} & \frac{1}{16} \\ -\frac{1}{32} & \frac{1}{12} \end{pmatrix} + \begin{pmatrix} 1 & 0 \\ 0 & 1 \end{pmatrix} + \begin{pmatrix} \frac{1}{12} \\ \frac{1}{18} \end{pmatrix} \begin{pmatrix} \frac{1}{2} & \frac{1}{2} \end{pmatrix} \left(\begin{pmatrix} 1 & 0 \\ 0 & 1 \end{pmatrix} - 0.25 \begin{pmatrix} 1 & 0 \\ 0 & 1 \end{pmatrix} \right) \\ &= \begin{pmatrix} 1.0498 & 0.09375 \\ -1.0417 \times 10^{-2} & 1.1042 \end{pmatrix}, \end{aligned}$$

$$\begin{aligned}
N_{B_2} &= A_{B_0}^2 + A_{B_1} = (A_0 + BLB_0)^2 + A_1 + BLB_1 \\
&= \begin{pmatrix} 1.0498 & 0.09375 \\ -1.0417 \times 10^{-2} & 1.1042 \end{pmatrix}^2 - 0.09375 \begin{pmatrix} 1 & 0 \\ 0 & 1 \end{pmatrix} \\
&\quad - 0.25 \begin{pmatrix} \frac{1}{12} \\ \frac{1}{18} \end{pmatrix} \begin{pmatrix} \frac{1}{2} & \frac{1}{2} \end{pmatrix} \begin{pmatrix} 1 & 0 \\ 0 & 1 \end{pmatrix} \\
&= \begin{pmatrix} 0.99694 & 0.19152 \\ -2.9383 \times 10^{-2} & 1.1176 \end{pmatrix}, \\
N_{B_3} &= A_{B_0}^3 + A_{B_1}^2 + A_{B_2} = (A_0 + BLB_0)^3 + (A_1 + BLB_1)^2 + (A_2 + BLB_2) \\
&= \begin{pmatrix} 1.0498 & 0.09375 \\ -1.0417 \times 10^{-2} & 1.1042 \end{pmatrix}^3 \\
&\quad + \left(-0.09375 \begin{pmatrix} 1 & 0 \\ 0 & 1 \end{pmatrix} - 0.25 \begin{pmatrix} \frac{1}{12} \\ \frac{1}{18} \end{pmatrix} \begin{pmatrix} \frac{1}{2} & \frac{1}{2} \end{pmatrix} \begin{pmatrix} 1 & 0 \\ 0 & 1 \end{pmatrix} \right)^2 \\
&\quad + \left(-5.4688 \times 10^{-2} \begin{pmatrix} 1 & 0 \\ 0 & 1 \end{pmatrix} - 0.09375 \begin{pmatrix} \frac{1}{12} \\ \frac{1}{18} \end{pmatrix} \begin{pmatrix} \frac{1}{2} & \frac{1}{2} \end{pmatrix} \begin{pmatrix} 1 & 0 \\ 0 & 1 \end{pmatrix} \right) \\
&= \begin{pmatrix} 1.1314 & 0.3497 \\ -2.0582 \times 10^{-2} & 1.3129 \end{pmatrix}, \\
N_{B_4} &= A_{B_0}^4 + A_{B_1}^3 + A_{B_2}^2 + A_{B_3} = (A_0 + BLB_0)^4 + (A_1 + BLB_1)^3 \\
&\quad + (A_2 + BLB_2)^2 + (A_3 + BLB_3) \\
&= \begin{pmatrix} 1.0498 & 0.09375 \\ -1.0417 \times 10^{-2} & 1.1042 \end{pmatrix}^4 \\
&\quad + \left(-0.09375 \begin{pmatrix} 1 & 0 \\ 0 & 1 \end{pmatrix} - 0.25 \begin{pmatrix} \frac{1}{12} \\ \frac{1}{18} \end{pmatrix} \begin{pmatrix} \frac{1}{2} & \frac{1}{2} \end{pmatrix} \begin{pmatrix} 1 & 0 \\ 0 & 1 \end{pmatrix} \right)^3 \\
&\quad + \left(-5.4688 \times 10^{-2} \begin{pmatrix} 1 & 0 \\ 0 & 1 \end{pmatrix} - 0.09375 \begin{pmatrix} \frac{1}{12} \\ \frac{1}{18} \end{pmatrix} \begin{pmatrix} \frac{1}{2} & \frac{1}{2} \end{pmatrix} \begin{pmatrix} 1 & 0 \\ 0 & 1 \end{pmatrix} \right)^2 \\
&\quad + \left(3.7598 \times 10^{-2} \begin{pmatrix} 1 & 0 \\ 0 & 1 \end{pmatrix} - 3.7598 \times 10^{-2} \begin{pmatrix} \frac{1}{12} \\ \frac{1}{18} \end{pmatrix} \begin{pmatrix} \frac{1}{2} & \frac{1}{2} \end{pmatrix} \begin{pmatrix} 1 & 0 \\ 0 & 1 \end{pmatrix} \right) \\
&= \begin{pmatrix} 1.2141 & 0.47237 \\ -4.9375 \times 10^{-2} & 1.4845 \end{pmatrix},
\end{aligned}$$

$$CN_{B_0} \begin{pmatrix} x \\ y \end{pmatrix} = \begin{pmatrix} 1 & -1 \end{pmatrix} \begin{pmatrix} 1 & 0 \\ 0 & 1 \end{pmatrix} \begin{pmatrix} x \\ y \end{pmatrix} = x - y,$$

$$\begin{aligned}
CN_{B_1} \begin{pmatrix} x \\ y \end{pmatrix} &= \begin{pmatrix} 1 & -1 \end{pmatrix} \begin{pmatrix} 1.0498 & 0.09375 \\ -1.0417 \times 10^{-2} & 1.1042 \end{pmatrix} \begin{pmatrix} x \\ y \end{pmatrix} \\
&= 1.0602x - 1.0105y, \\
CN_{B_2} \begin{pmatrix} x \\ y \end{pmatrix} &= \begin{pmatrix} 1 & -1 \end{pmatrix} \begin{pmatrix} 0.99694 & 0.19152 \\ -2.9383 \times 10^{-2} & 1.1176 \end{pmatrix} \begin{pmatrix} x \\ y \end{pmatrix} \\
&= 1.0263x - 0.92608y, \\
CN_{B_3} \begin{pmatrix} x \\ y \end{pmatrix} &= \begin{pmatrix} 1 & -1 \end{pmatrix} \begin{pmatrix} 1.1314 & 0.3497 \\ -2.0582 \times 10^{-2} & 1.3129 \end{pmatrix} \begin{pmatrix} x \\ y \end{pmatrix} \\
&= 1.1520x - 0.9632y, \\
CN_{B_4} \begin{pmatrix} x \\ y \end{pmatrix} &= \begin{pmatrix} 1 & -1 \end{pmatrix} \begin{pmatrix} 1.2141 & 0.47237 \\ -4.9375 \times 10^{-2} & 1.4845 \end{pmatrix} \begin{pmatrix} x \\ y \end{pmatrix} \\
&= 1.2635x - 1.0121y.
\end{aligned}$$

Algorithm:

$k = 0$,

$i = 1$,

Maximize $J_1(x) = 1.0602x - 1.0105y - 0.1$,

S.C: $x - y - 0.1 \leq 0$ and $-x + y - 0.1 \leq 0$,

$k = 0$,

$i = 2$,

Maximize $J_2(x) = -1.0602x + 1.0105y - 0.1$,

S.C: $x - y - 0.1 \leq 0.1$ and $-x + y - 0.1 \leq 0$,

We have $J_1^*, J_2^* \geq 0$. Then go to the next step.

$k = 1$,

$i = 1$,

Maximize $J_1(x) = 0.16595x - 6.8985 \times 10^{-2}y - 0.1$,

S.C: $1.0602x - 1.0105y - 0.1 \leq 0$ and $-1.0602x + 1.0105y - 0.1 \leq 0$,

$x - y - 0.1 \leq 0$ and $-x + y - 0.1 \leq 0$,

$k = 1$,

$i = 2$,

Maximize $J_2(x) = -0.16595x + 6.8985 \times 10^{-2}y - 0.1$,

S.C: $1.0602x - 1.0105y - 0.1 \leq 0$ and $-1.0602x + 1.0105y - 0.1 \leq 0$,

$x - y - 0.1 \leq 0$ and $-x + y - 0.1 \leq 0$,

We have $J_1^*, J_2^* \geq 0$. Then go to the next step.

$k = 2$,

$i = 1$,

Maximize $J_1(x) = 1.1520x - 0.9632y - 0.1$,

S.C: $0.16595x + 6.8985 \times 10^{-2}y - 0.1 \leq 0$ and $-0.16595x + 6.8985 \times 10^{-2}y - 0.1 \leq 0$,

$1.0602x - 1.0105y - 0.1 \leq 0$ and $-1.0602x + 1.0105y - 0.1 \leq 0$,

$x - y - 0.1 \leq 0$ and $-x + y - 0.1 \leq 0$,

$k = 2$,

$i = 2$,

Maximize $J_2(x) = -1.1520x + 0.9632y - 0.1$,

S.C : $0.16595x + 6.8985 \times 10^{-2}y - 0.1 \leq 0$ and $-0.16595x + 6.8985 \times 10^{-2}y - 0.1 \leq 0$,

$1.0602x - 1.0105y - 0.1 \leq 0$ and $-1.0602x + 1.0105y - 0.1 \leq 0$,

$x - y - 0.1 \leq 0$ and $-x + y - 0.1 \leq 0$,

We have $J_1^*, J_2^* \geq 0$. Then go to the next step.

$k = 3$,

$i = 1$,

Maximize $J_1(x) = 1.2635x - 1.0121y - 0.1$,

S.C : $1.1520x - 0.9632y - 0.1 \leq 0$ and $-1.1520x + 0.9632y - 0.1 \leq 0$,

$0.16595x + 6.8985 \times 10^{-2}y - 0.1 \leq 0$ and $-0.16595x + 6.8985 \times 10^{-2}y - 0.1 \leq 0$,

$1.0602x - 1.0105y - 0.1 \leq 0$ and $-1.0602x + 1.0105y - 0.1 \leq 0$,

$x - y - 0.1 \leq 0$ and $-x + y - 0.1 \leq 0$,

$k = 3$,

$i = 1$,

Maximize $J_2(x) = -1.2635x + 1.0121y - 0.1$,

S.C: $1.1520x - 0.9632y - 0.1 \leq 0$ and $-1.1520x + 0.9632y - 0.1 \leq 0$,

$0.16595x + 6.8985 \times 10^{-2}y - 0.1 \leq 0$ and $-0.16595x + 6.8985 \times 10^{-2}y - 0.1 \leq 0$,

$1.0602x - 1.0105y - 0.1 \leq 0$ and $-1.0602x + 1.0105y - 0.1 \leq 0$,

$x - y - 0.1 \leq 0$ and $-x + y - 0.1 \leq 0$,

Since $J_1^*, J_2^* \leq 0$, we stop.

Then we use Algorithm II to prove that $k^* = 3$, and we have

$$\mathcal{R} = \{x \in \mathbb{R} \mid |x| \leq 1\} = [-1, 1],$$

$$\Upsilon_3(\mathcal{R}) = \left\{ \begin{pmatrix} x \\ y \end{pmatrix} \left| \begin{array}{l} |x - y| \leq 0.1, |1.0602x - 1.0105y| \leq 0.1, \\ |1.0263x - 0.92608y| \leq 0.1, \\ |1.1520x - 0.9632y| \leq 0.1 \end{array} \right. \right\}.$$

Example 2. Consider the following system:

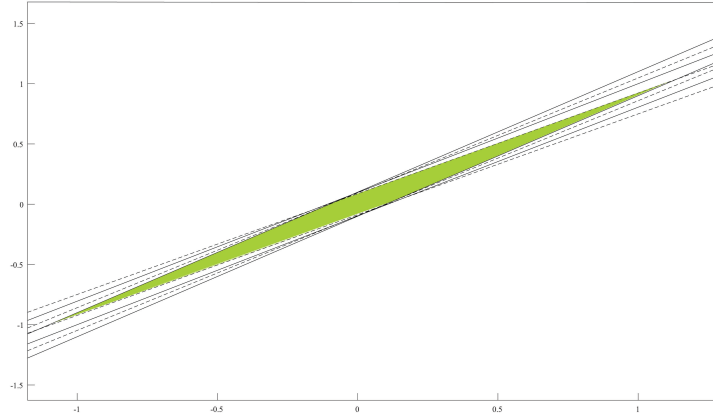


Figure 1: The shaded area represents the set $\Upsilon_3(\mathcal{R})$ as illustrated in Example 1 with $\gamma = 0.25$

$$\begin{cases} x(k+1) = \sum_{j=0}^l \tilde{A}_{B_j} x_{k-j}, \\ x_0 = x(0) \in \mathbb{R}^n. \end{cases}$$

We define the parameters γ , γ , ε , n , and the matrices A , B , L , and C as follows: $\gamma = 0.25$, $\gamma = 0.25$, $\varepsilon = 0.4$, $n = 3$, $C = \begin{pmatrix} 1 & -1 \end{pmatrix}$, $A = \begin{pmatrix} \frac{35}{12} & -\frac{13}{7} \\ \frac{3}{32} & -\frac{11}{7} \end{pmatrix}$, $B = \begin{pmatrix} 3 & 2 \\ 0 & 1 \end{pmatrix}$, and $L = \begin{pmatrix} -1 & 0 \\ 0 & 1 \end{pmatrix}$.

Then

$$\begin{aligned} N_{B_0} &= I_2 = \begin{pmatrix} 1 & 0 \\ 0 & 1 \end{pmatrix}, \\ N_{B_1} &= A_{B_0} = A_0 + BLB_0 \\ &= A + BL + I_2 + BLB_0 = A + I_2 + BL(I_2 + B_0) \\ &= \begin{pmatrix} \frac{35}{12} & -\frac{13}{7} \\ \frac{3}{32} & -\frac{11}{7} \end{pmatrix} + \begin{pmatrix} 1 & 0 \\ 0 & 1 \end{pmatrix} + \begin{pmatrix} 3 & 2 \\ 0 & 1 \end{pmatrix} \begin{pmatrix} -1 & 0 \\ 0 & 1 \end{pmatrix} \left(\begin{pmatrix} 1 & 0 \\ 0 & 1 \end{pmatrix} - 0.25 \begin{pmatrix} 1 & 0 \\ 0 & 1 \end{pmatrix} \right) \\ &= \begin{pmatrix} 1.6667 & -0.35714 \\ \frac{3}{32} & 0.17857 \end{pmatrix}, \\ A_{B_1} &= A_1 + BLB_1 = 0.09375 \begin{pmatrix} 1 & 0 \\ 0 & 1 \end{pmatrix} - 0.09375 \begin{pmatrix} 3 & 2 \\ 0 & 1 \end{pmatrix} \begin{pmatrix} -1 & 0 \\ 0 & 1 \end{pmatrix} \begin{pmatrix} 1 & 0 \\ 0 & 1 \end{pmatrix} \\ &= \begin{pmatrix} 0.375 & -0.1875 \\ 0 & 0 \end{pmatrix}, \\ N_{B_2} &= A_{B_0}^2 + A_{B_1} = (A_0 + BLB_0)^2 + A_1 + BLB_1 \end{aligned}$$

$$\begin{aligned}
&= \begin{pmatrix} 1.6667 & -0.35714 \\ \frac{3}{32} & 0.17857 \end{pmatrix}^2 + \begin{pmatrix} 0.375 & -0.1875 \\ 0 & 0 \end{pmatrix} \\
&= \begin{pmatrix} 3.1194 & -0.84652 \\ 0.17299 & -1.5946 \times 10^{-3} \end{pmatrix}, \\
N_{B_3} &= A_{B_0}^3 + A_{B_1}^2 + A_{B_2} = (A_0 + BLB_0)^3 + (A_1 + BLB_1)^2 + (A_2 + BLB_2) \\
&= \begin{pmatrix} 1.6667 & -0.35714 \\ \frac{3}{32} & 0.17857 \end{pmatrix}^3 + \begin{pmatrix} 0.375 & -0.1875 \\ 0 & 0 \end{pmatrix}^2 \\
&\quad + \left(\begin{pmatrix} 5.4688 \times 10^{-2} & 0 \\ 0 & 5.4688 \times 10^{-2} \end{pmatrix} \right) \\
&\quad + \left(\begin{pmatrix} 3 & 2 \\ 0 & 1 \end{pmatrix} \begin{pmatrix} -1 & 0 \\ 0 & 1 \end{pmatrix} \begin{pmatrix} -5.4688 \times 10^{-2} & 0 \\ 0 & -5.4688 \times 10^{-2} \end{pmatrix} \right) \\
&= \begin{pmatrix} 4.8717 & -1.2775 \\ 0.28818 & -6.2068 \times 10^{-2} \end{pmatrix}
\end{aligned}$$

$$\begin{aligned}
CN_{B_0} \begin{pmatrix} x \\ y \end{pmatrix} &= \begin{pmatrix} 1 & -1 \end{pmatrix} \begin{pmatrix} 1 & 0 \\ 0 & 1 \end{pmatrix} \begin{pmatrix} x \\ y \end{pmatrix} = x - y \\
CN_{B_1} \begin{pmatrix} x \\ y \end{pmatrix} &= \begin{pmatrix} 1 & -1 \end{pmatrix} \begin{pmatrix} 1.6667 & -0.35714 \\ \frac{3}{32} & 0.17857 \end{pmatrix} \begin{pmatrix} x \\ y \end{pmatrix} = 1.5730x - 0.53571y \\
CN_{B_2} \begin{pmatrix} x \\ y \end{pmatrix} &= \begin{pmatrix} 1 & -1 \end{pmatrix} \begin{pmatrix} 3.1194 & -0.84652 \\ 0.17299 & -1.5946 \times 10^{-3} \end{pmatrix} \begin{pmatrix} x \\ y \end{pmatrix} \\
&= 2.9464x - 0.84493y \\
CN_{B_3} \begin{pmatrix} x \\ y \end{pmatrix} &= \begin{pmatrix} 1 & -1 \end{pmatrix} \begin{pmatrix} 4.8717 & -1.2775 \\ 0.28818 & -6.2068 \times 10^{-2} \end{pmatrix} \begin{pmatrix} x \\ y \end{pmatrix} \\
&= 4.5835x - 1.2154y
\end{aligned}$$

Then we use Algorithm II to prove that $k^* = 1$ and we have

$$\Upsilon_1(\mathcal{R}) = \left\{ \begin{pmatrix} x \\ y \end{pmatrix} \mid |x - y| \leq 0.4, \quad |1.5730x - 0.53571y| \leq 0.4 \right\}.$$

Example 3. Consider the following system:

$$\begin{cases} x(k+1) = \sum_{j=0}^L \tilde{A}_{B_j} x_{k-j}, \\ x_0 = x(0) \in \mathbb{R}^n. \end{cases}$$

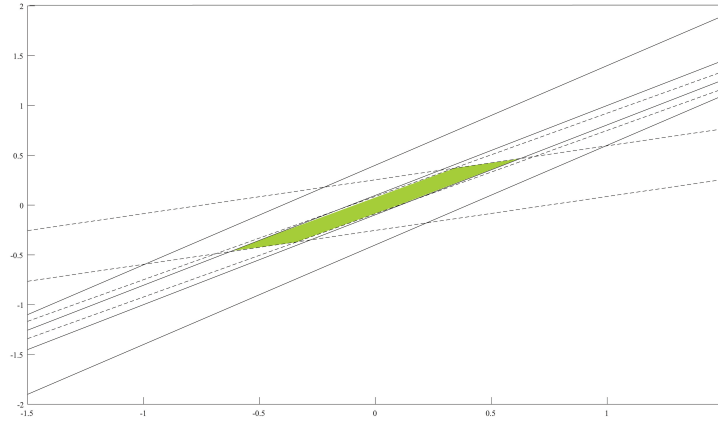


Figure 2: The shaded area represents the set $\Upsilon_1(\mathcal{R})$ as illustrated in Example 2 with $\gamma = 0.25$

Let the parameters $\gamma, \gamma, \varepsilon, n$ and the matrix A, B, L and C be defined by

$$\gamma = 0.1, \quad \varepsilon = 0.3, n = 2, \quad \gamma = 0.1, \quad C = \begin{pmatrix} 2 & -1 \end{pmatrix}, \quad A = \begin{pmatrix} \frac{1}{48} & \frac{1}{16} \\ 0 & \frac{1}{40} \end{pmatrix}, \quad B = \begin{pmatrix} \frac{1}{16} \\ \frac{1}{10} \end{pmatrix}, \quad L = \begin{pmatrix} \frac{1}{2} & \frac{1}{2} \end{pmatrix}.$$

Then

$$\begin{aligned} N_{B_0} &= I_2 = \begin{pmatrix} 1 & 0 \\ 0 & 1 \end{pmatrix}, \\ N_{B_1} &= N_{B_1} = A_{B_0} = A_0 + BLB_0 \\ &= A + BL + I_2 + BLB_0 = A + I_2 + BL(I_2 + B_0) \\ &= \begin{pmatrix} \frac{1}{48} & \frac{1}{16} \\ 0 & \frac{1}{40} \end{pmatrix} + \begin{pmatrix} 1 & 0 \\ 0 & 1 \end{pmatrix} + \begin{pmatrix} \frac{1}{16} \\ \frac{1}{10} \end{pmatrix} \begin{pmatrix} \frac{1}{2} & \frac{1}{2} \end{pmatrix} \left(\begin{pmatrix} 1 & 0 \\ 0 & 1 \end{pmatrix} - 0.25 \begin{pmatrix} 1 & 0 \\ 0 & 1 \end{pmatrix} \right) \\ &= \begin{pmatrix} 1.0443 & 8.5938 \times 10^{-2} \\ 0.0375 & 1.0625 \end{pmatrix}, \\ N_{B_2} &= A_{B_0}^2 + A_{B_1} = (A_0 + BLB_0)^2 + A_1 + BLB_1 \\ &= \begin{pmatrix} 1.0443 & 8.5938 \times 10^{-2} \\ 0.0375 & 1.0625 \end{pmatrix}^2 + 0.08 \begin{pmatrix} 1 & 0 \\ 0 & 1 \end{pmatrix} \\ &\quad - 0.2 \begin{pmatrix} \frac{1}{16} \\ \frac{1}{10} \end{pmatrix} \begin{pmatrix} \frac{1}{2} & \frac{1}{2} \end{pmatrix} \begin{pmatrix} 1 & 0 \\ 0 & 1 \end{pmatrix} \\ &= \begin{pmatrix} 1.1675 & 0.1748 \\ 6.9005 \times 10^{-2} & 1.2021 \end{pmatrix}, \\ N_{B_3} &= A_{B_0}^3 + A_{B_1}^2 + A_{B_2} \end{aligned}$$

$$\begin{aligned}
&= (A_0 + BLB_0)^3 + (A_1 + BLB_1)^2 + (A_2 + BLB_2) \\
&= \begin{pmatrix} 1.0443 & 8.5938 \times 10^{-2} \\ 0.0375 & 1.0625 \end{pmatrix}^3 \\
&\quad + \left(0.08 \begin{pmatrix} 1 & 0 \\ 0 & 1 \end{pmatrix} - 0.2 \begin{pmatrix} \frac{1}{16} \\ \frac{1}{10} \end{pmatrix} \begin{pmatrix} \frac{1}{2} & \frac{1}{2} \end{pmatrix} \begin{pmatrix} 1 & 0 \\ 0 & 1 \end{pmatrix} \right)^2 \\
&\quad + \left(-0.08 \begin{pmatrix} 1 & 0 \\ 0 & 1 \end{pmatrix} + \begin{pmatrix} \frac{1}{16} \\ \frac{1}{10} \end{pmatrix} \begin{pmatrix} \frac{1}{2} & \frac{1}{2} \end{pmatrix} 0.08 \begin{pmatrix} 1 & 0 \\ 0 & 1 \end{pmatrix} \right) \\
&= \begin{pmatrix} 1.077 & 0.28797 \\ 0.12752 & 1.1386 \end{pmatrix}, \\
CN_{B_0} \begin{pmatrix} x \\ y \end{pmatrix} &= \begin{pmatrix} 2 & -1 \end{pmatrix} \begin{pmatrix} 1 & 0 \\ 0 & 1 \end{pmatrix} \begin{pmatrix} x \\ y \end{pmatrix} = 2x - y, \\
CN_{B_1} \begin{pmatrix} x \\ y \end{pmatrix} &= \begin{pmatrix} 2 & -1 \end{pmatrix} \begin{pmatrix} 1.0443 & 8.5938 \times 10^{-2} \\ 0.0375 & 1.0625 \end{pmatrix} \begin{pmatrix} x \\ y \end{pmatrix} \\
&= 2.0511x - 0.89062y, \\
CN_{B_3} \begin{pmatrix} x \\ y \end{pmatrix} &= \begin{pmatrix} 2 & -1 \end{pmatrix} \begin{pmatrix} 1.1675 & 0.1748 \\ 6.9005 \times 10^{-2} & 1.2021 \end{pmatrix} \begin{pmatrix} x \\ y \end{pmatrix} \\
&= 2.2660x - 0.8525y, \\
CN_{B_4} \begin{pmatrix} x \\ y \end{pmatrix} &= \begin{pmatrix} 2 & -1 \end{pmatrix} \begin{pmatrix} 1.077 & 0.28797 \\ 0.12752 & 1.1386 \end{pmatrix} \begin{pmatrix} x \\ y \end{pmatrix} \\
&= 2.0265x - 0.56266y.
\end{aligned}$$

Then we use Algorithm II to prove that $k^* = 2$, and we have

$$\begin{aligned}
\Upsilon_2(\mathcal{R}) &= \left\{ \begin{pmatrix} x \\ y \end{pmatrix} \mid |2x - y| \leq 0.3, \quad |2.0511x - 0.89062y| \leq 0.3, \right. \\
&\quad \left. |2.2660x - 0.8525y| \leq 0.3 \right\}.
\end{aligned}$$

Example 4. Let the parameters $\gamma, \gamma, \varepsilon, n$ and the matrix $\tilde{A}, B, L$ and C be defined by

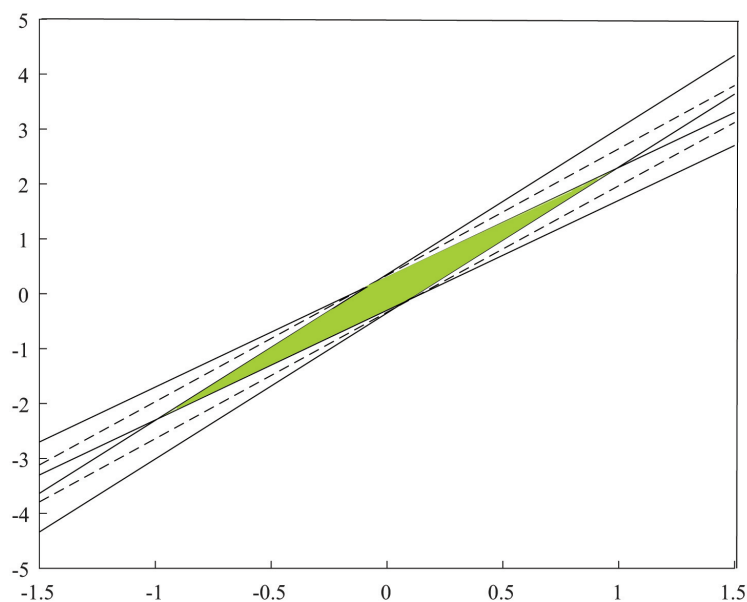


Figure 3: The shaded area represents the set $\Upsilon_2(\mathcal{R})$ as illustrated in Example 3 with $\gamma = 0.1$

$$\gamma = 0.7, \varepsilon = 0.5, n = 3, \quad C = \begin{pmatrix} 1 & 1 & 1 \end{pmatrix}, \quad \tilde{A} = \begin{pmatrix} \frac{1}{12} & \frac{1}{9} & \frac{1}{7} \\ \frac{1}{15} & \frac{1}{13} & \frac{1}{11} \\ \frac{1}{7} & \frac{1}{15} & \frac{1}{14} \end{pmatrix}, \quad B = \begin{pmatrix} \frac{1}{10} \\ \frac{1}{5} \\ \frac{1}{8} \end{pmatrix}, \quad L = \begin{pmatrix} \frac{1}{3} & \frac{1}{2} & \frac{1}{4} \end{pmatrix}.$$

Also, $CN_{B_0} \begin{pmatrix} x \\ y \\ z \end{pmatrix}$, $CN_{B_1} \begin{pmatrix} x \\ y \\ z \end{pmatrix}$, $CN_{B_2} \begin{pmatrix} x \\ y \\ z \end{pmatrix}$, $CN_{B_3} \begin{pmatrix} x \\ y \\ z \end{pmatrix}$ are given by

$$CN_{B_0} \begin{pmatrix} x \\ y \\ z \end{pmatrix} = \begin{pmatrix} 1 & 1 & 1 \end{pmatrix} \begin{pmatrix} 1 & 0 & 0 \\ 0 & 1 & 0 \\ 0 & 0 & 1 \end{pmatrix} \begin{pmatrix} x \\ y \\ z \end{pmatrix} = x + y + z,$$

$$CN_{B_1} \begin{pmatrix} x \\ y \\ z \end{pmatrix} = \begin{pmatrix} 1 & 1 & 1 \end{pmatrix} \begin{pmatrix} 0.85 & \frac{19}{90} & \frac{27}{140} \\ \frac{1}{5} & 0.97692 & \frac{21}{110} \\ \frac{19}{84} & \frac{23}{120} & 0.83393 \end{pmatrix} \begin{pmatrix} x \\ y \\ z \end{pmatrix} \\ = 1.2762x + 1.3797y + 1.2177z,$$

$$CN_{B_2} \begin{pmatrix} x \\ y \\ z \end{pmatrix} = \begin{pmatrix} 1 & 1 & 1 \end{pmatrix} \begin{pmatrix} 0.70684 & 0.42790 & 0.36769 \\ 0.41557 & 0.93869 & 0.38953 \\ 0.42360 & 0.40139 & 0.67393 \end{pmatrix} \begin{pmatrix} x \\ y \\ z \end{pmatrix}$$

$$\begin{aligned}
&= 1.546x + 1.7680y + 1.4312z, \\
CN_{B_3} \begin{pmatrix} x \\ y \\ z \end{pmatrix} &= \begin{pmatrix} 1 & 1 & 1 \end{pmatrix} \begin{pmatrix} 0.90853 & 0.65023 & 0.53937 \\ 0.63646 & 1.2192 & 0.59321 \\ 0.60801 & 0.61878 & 0.85754 \end{pmatrix} \begin{pmatrix} x \\ y \\ z \end{pmatrix} \\
&= 2.153x + 2.4882y + 1.9901z.
\end{aligned}$$

Algorithm II is employed to demonstrate that $k^* = 2$ and to identify the MOS $\Upsilon_2(\mathcal{R})$ as follows:

$$\Upsilon_2(\mathcal{R}) = \left\{ \begin{pmatrix} x \\ y \\ z \end{pmatrix} \left| \begin{array}{l} |x + y + z| \leq 0.5, \\ |1.2762x + 1.3797y + 1.2177z| \leq 0.5, \\ |1.546x + 1.7680y + 1.4312z| \leq 0.5, \end{array} \right. \right\}.$$

Comment

In Examples 1 and 2, the admissibility index k_0 and the MOS $\Upsilon(\mathcal{R})$ for vectors satisfying certain constraints were determined using the simplex method. This choice was made in order to solve the maximization problems inherent in Algorithm 2. The constraints characterizing the sets $\Upsilon(\mathcal{R})$ in Figures 1–3 have been visualized to provide a graphical insight into $\Upsilon(\mathcal{R})$, as illustrated in Examples 1–3.

7 Conclusion

This paper addressed a novel problem concerning the maximum output admissible set in discrete-time fractional control systems utilizing the Grünwald–Letnikov fractional derivative. It introduced a new sufficient condition to ensure system stability and characterizes the maximal output admissible set, denoted as $\Upsilon(\mathcal{R})$. The study demonstrated that, under specific conditions, finite determinations of $\Upsilon(\mathcal{R})$ can be achieved. Additionally, it proposes an effective algorithmic approach to determine the admissibility index k^* and establish the set using a finite number of inequalities. The application of the algorithm to controlled scenarios is exemplified through numerical examples, effectively showcasing the theoretical results in both two and three dimensions. As a natural continuation of this work, future investigations will address the extension of the proposed fractional-order control methodology to nonlinear dynamics as well as to infinite-dimensional systems, where additional theoretical challenges and practical implications arise.

Conflict of interest

The authors declare that they have no known competing financial interests or personal relationships that could have appeared to

Availability of data and materials

Data is available upon request.

Acknowledgements

The research reported in this paper was supported by the Moroccan Systems Theory Network.

Funding

This research received no external funding.

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] Benfatah, Y., El Bhih, A., Rachik, M. and Tridane, A. *On the maximal output admissible set for a class of bilinear discrete-time systems*, Int. J. Control, Autom. and Syst., 19 (2021), 3551–3568.
- [2] Buslowicz, M. *Robust stability of positive discrete-time linear systems of fractional order*, Bull. Polish Academy Sci.: Tech. Sci., 58 (2010), 567–572.
- [3] Caponetto, R., Dongola, G., Fortuna, L. and Petras, I. *Fractional order systems: modelling and control applications*, World Scientific, 2010.

- [4] Debnath, L. *Recent applications of fractional calculus to science and engineering*, IJMMS, 54 (2003), 3413–3442.
- [5] Dzielinski, A. and Sierociuk, D. *Adaptive feedback control of fractional order discrete state-space systems*, CIMCA–IAWTIC’06, (2005), 804–809.
- [6] Dzielinski, A. and Sierociuk, D. *Stability of discrete fractional order state-space systems*, J. Vib. Control, 14 (2008), 1543–1556.
- [7] Dzielinski, A. and Sierociuk, D. *Stability of discrete fractional order state-space systems*, J. Vib. Control, 14 (2008), 1543–1556.
- [8] El Bhih, A., Benfatah, Y., Ben Rhila, S., Rachik, M. and Laaroussi, A.E.A. *A spatiotemporal prey–predator discrete model and optimal controls for environmental sustainability in multifishing areas of Morocco*, Discrete Dyn. Nature Soc., (2020), Article ID 2780651, 1–18.
- [9] El Bhih, A., Benfatah, Y., Kouidere, A. and Rachik, M. *A discrete mathematical modeling of transmission of COVID-19 pandemic using optimal control*, Commun. Math. Biol. Neurosci., (2020), Article ID 75, 1–23.
- [10] El Bhih, A., Benfatah, Y. and Rachik, M. *Exact determination of maximal output admissible set for a class of semilinear discrete systems*, Arch. Control Sci., 30 (2020), 523–552.
- [11] El Bhih, A., Yaagoub, Z., Rachik, M., Allali, K. and Abdeljawad, T. *Controlling the dissemination of rumors and antirumors in social networks*, European Phys. J. Plus, 139 (2024), 1–23.
- [12] Ferreira, R.A.C. and Torres, D.F.M. *Fractional h -difference equations arising from the calculus of variations*, Appl. Anal. Discrete Math., 5 (2011), 110–121.
- [13] Gilbert, E.G. and Tan, K.T. *Linear systems with state and control constraints: the theory and application of maximal output admissible sets*, IEEE Trans. Auto. Control, 36 (1991), 1008–1020.
- [14] Guermah, S., Djennoune, S. and Bettayeb, M. *Controllability and observability of linear discrete-time fractional-order systems*, Int. J. Appl. Math. Comput. Sci., 18 (2008), 213–222.
- [15] Gutman, P.O. and Hagander, P. *A new design of constrained controllers for linear systems*, IEEE Trans. Auto. Control, 30 (1985), 22–23.

- [16] Hirata, K. and Ohta, Y. *Exact determinations of the maximal output admissible set for a class of nonlinear systems*, Proceedings of the 44th IEEE Conference on Decision and Control, (2005), 8276–8281.
- [17] Jose, S.A., Panigoro, H.S., Jirawattanapanit, A., Omede, B.I. and Yaagoub, Z. *Understanding COVID-19 propagation: a mathematical model with Caputo fractional derivatives*, Front. Appl. Math. Stat., 10 (2024).
- [18] Kaczorek, T. *Reachability and controllability to zero of cone fractional discrete-time systems*, Arch. Control Sci., 17 (2007), 357–367.
- [19] Kaczorek, T. *Fractional positive continuous-time linear systems and their reachability*, Int. J. Appl. Math. Comput. Sci., 18 (2008), 223–228.
- [20] Kilbas, A.A., Srivastava, H.M. and Trujillo, J.J. *Theory and application of fractional differential equations*, North Holland Mathematics Studies, Elsevier, 2006.
- [21] Kolmanovsky, I. and Gilbert, E.G. *Multimode regulators for systems with state control constraints and disturbance inputs*, Control Using Logic-Based Switching, 222 (1997), 104–117.
- [22] Laaroussi, A.E.A., El Bhil, A. and Rachik, M. *Optimal vaccination and treatment policies with constrained inequalities for a multistrain reaction-diffusion SEIR model of COVID-19*, Partial Differ. Equ. Appl. Math., 10 (2024), Article 100684.
- [23] Larrache, A., Lhous, M., Ben Rhila, S., Rachik, M. and Tridane, A. *An output sensitivity problem for a class of linear distributed systems with uncertain initial state*, Arch. Control Sci., 30 (2020), 139–155.
- [24] Oldham, K.B. and Spanier, J. *The fractional calculus*, Academic Press, 1974.
- [25] Podlubny I. *Fractional differential equations*, Vol. 198. Academic Press, 1999.
- [26] Rachik, M., Abdelhak, A. and Karakchou, J. *Discrete systems with delays in state, control and observation: maximal output sets with state and control constraints*, Optimization, 42 (1997), 169–183.
- [27] Rachik, M. and Lhous, M. *An observer-based control of linear systems with uncertain parameters*, Archives of Control Sciences, 26 (2016), 565–576.

- [28] Rachik, M., Lhous, M. and Tridane, A. *On the maximal output admissible set for a class of nonlinear discrete systems*, Syst. Anal. Model. Simul., 42 (2002), 1639–1658.
- [29] Yaagoub, Z. and Allali, K. *Fractional HCV infection model with adaptive immunity and treatment*, Math. Model. Comput., 10 (2023), 995–1013.
- [30] Yaagoub, Z., El Bhih, A. and Allali, K. *Global analysis of a fractional-order infection model for computer viruses*, Model. Earth Syst. Envir., 11 (2025), 68.
- [31] Yaagoub, Z., Farah, E.M. and Ahmad, S. *Three-strain epidemic model for influenza virus involving fractional derivative and treatment*, J. Appl. Math. Comput., 71 (2025), 1247–1266.



Spline-interpolation solution of a plane Stefan problem

P.N. Ivanshin*

Abstract

We construct a spline-interpolation solution of the two-dimensional Stefan problem. We reduce it to the set of boundary value problems for analytic functions and integral equations. Our construction of an approximate solution is based on the introduction of analytic functions and the application of approximate solutions of boundary-value problems. We modify the initial problem in order to reduce it to a set of boundary value problems. The first of these problems is the Cauchy problem for the Laplace equation. The second problem is the Riemann problem for analytic function, We also give estimates of the approximate solution error and some examples of the solution.

AMS subject classifications (2020): Primary 35J25; Secondary 30E10.

Keywords: Spline; analytic function; approximate solution; Cauchy problem.

1 Introduction

The task of studying the conditions for the formation and dynamics of the liquid-solid phase transition has wide practical applications, as it is directly related to issues such as storing thermal

*Corresponding author

Received 18 July 2025; revised 17 November 2025; accepted 24 December 2025

Pyotr N. Ivanshin

Kazan National Research Technical University, K. Marksa st., 10, Kazan, Russia. e-mail: pivanshi@yandex.ru

How to cite this article

Ivanshin, P.N., Spline-interpolation solution of a plane Stefan problem. *Iran. J. Numer. Anal. Optim.*, 2026; 16(2): 525-538. <https://doi.org/10.22067/ijnao.2025.94508.1683>

energy, controlling temperature and humidity, and ensuring the durability and strength of the systems under study. In addition, the energy stored during melting can be released during the solidification process, which enables a wide range of energy applications.

Possible applications span various technologies, such as in the steel industry, soil freezing, and thermal control systems for spacecraft, as well as ground-based industrial and civil facilities.

The solution to this problem has practical applications in the construction industry, where the pressing issue of determining the current position of the freezing front in external enclosing structures of buildings and construction elements arises. The position of the freezing front is directly linked to the durability of the enclosing structures. In fact, in the zone where the freezing front is moving, extremely unfavorable operating conditions for the material of the enclosing structures are created due to the potential alternation of freezing and thawing. This can gradually lead to a decrease in strength, and ultimately, to the destruction of the structure.

The Stefan problem for the temperature field $T(t, x, y)$, is formulated as follows: Consider a medium occupying a region Ω consisting of two phases in the corresponding subdomains of Ω : Phase 1 in the subdomain $D_1(t)$ and phase 2 in $D_2(t)$. These subdomains have the common dynamic boundary $\Gamma'(t)$. Let the two phases have thermal diffusivities θ_1 and θ_2 . Here $\frac{\partial T}{\partial t} = \nabla \cdot (\theta_1 \nabla T)$ in the region D_1 and $\frac{\partial T}{\partial t} = \nabla \cdot (\theta_2 \nabla T)$ in the region D_2 .

Assume that we know the boundary $\Gamma'(0)$ between the zones $D_1(0)$ and $D_2(0)$. So given the initial value of $T(0, x, y)$ at Ω and the boundary data $T(t, x, y)|_{(x,y) \in \partial\Omega = \Gamma \cup \Gamma'}$ it should be possible to reconstruct both the function $T(t, x, y)$, $t > 0$, $(x, y) \in \Omega$ and the boundary $\Gamma'(t)$ between two subdomains D_1 and D_2 at any time $t > 0$. The Stefan condition determines the evolution of the curve $\Gamma'(t)$ by giving an equation governing the velocity V of the free boundary in the normal direction ν , specifically $V = \theta_1 \partial_\nu T_1 - \theta_2 \partial_\nu T_2$. By boundary values of T_1 we mean the limit of the gradient as (x, y) approaches $\Gamma'(t)$ from the region $D_1(t)$, and by boundary values of T_2 we mean the limit of the gradient as (x, y) approaches $\Gamma'(t)$ from the region $D_2(t)$.

There exists a number of numerical methods of solution of this problem [1, 3, 5, 14, 16]. Most of these approaches are variations of finite-difference and finite-element methods [17, 18]. The authors usually consider different types of meshes. Here, we try to exclude the mesh at least from the spatial coordinates.

Here, we apply the Cauchy integral formula to the solution of the Stefan problem. We first modify the initial problem in order to reduce it to a set of certain boundary value problems. The first of these problems is the Cauchy problem for Laplace equation. So we again see the connection of the Stefan problem with the subject of incorrectly over-determined problems [4]. This is the crucial part of our construction. We show the solvability of this problem and present certain examples of the solution. We apply methods similar to our constructions of [8]. The

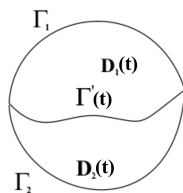


Figure 1: Domain D bounded by the curves Γ and Γ'

second problem is the usual Schwartz problem of reconstruction of the boundary imaginary part of the function analytic in a simply-connected domain by the real part of this function.

Our construction of an approximate solution is based on the introduction of analytic functions and the application of approximate solutions of boundary-value problems. We do not apply here the common methods of solution of the integral equations corresponding to so-called weak solutions [9, 10, 11, 12] of finite-element methods [5, 14, 16]. Our solution is thus an approximate-analytic one.

2 Approximate solution construction

Here we construct an approximate solution of a two-phase Stefan problem for the constant thermal diffusivities θ_j , $j = 1, 2$. Consider the domain $D_j(t)$, $j = 1, 2$, bounded by the simple curves Γ_j and $\Gamma'(t)$, so $\partial D_j(t) = \Gamma_j \cup \Gamma'(t)$ (see Figure 1). Assume that the curve Γ_j is fixed, and that the curve $\Gamma'(t)$ is the free part of the boundary $\partial D_j(t)$. Position of $\Gamma'(t)$ is known for $t = 0$.

The first step in our construction is to fix the subdomain $D_1 = D_1(0)$ and to construct an approximate solution $T^1(t, x, y)$ in $[0, 1] \times D_1$. Recall the main relation

$$\frac{\partial T}{\partial t} = \nabla \cdot (\theta_1 \nabla T) \quad (1)$$

in this domain.

Consider the modified version of the problem: Let the function T take values at $T(0, x(s), y(s))$ and $T(1, x(s), y(s))$, with $(x(s), y(s))$ belonging to Γ_1 , denoted by $U_0(s)$ and $U_1(s)$, respectively. Similarly, let the values of the normal derivative of T , that is, $\frac{\partial T(t, x, y)}{\partial n}$, at $t = 0$ and $t = 1$, for points $(x(s), y(s)) \in \Gamma_1$, be denoted by $V_0(s)$ and $V_1(s)$, respectively.

These conditions allow us to formulate a linearized solution, specifically a linear approximation in terms of t , given by $T^1(t, z, \bar{z}) = T_0(z, \bar{z}) + tT_1(z, \bar{z})$ for $t \in [0, 1]$. According to equation (1), the function $T_1(z, \bar{z})$ is harmonic with respect to z and $\bar{z}$ in the domain D_1 . To determine $T_1(z, \bar{z})$, we are faced with an ill-posed problem, where the function $T_1(z, \bar{z})$ is specified only on a part of the boundary.

The first boundary value problem that we need to solve is as follows: Given the value of the holomorphic function $T_1(z, \bar{z})$ on the boundary curve Γ_1 of the domain D_1 , it should be possible to extend $T_1(z, \bar{z})$ to the boundary curve $\Gamma'(0)$.

Then we apply the initial boundary conditions on ∂D_1 and the reconstructed values of $T_1(z, \bar{z})$ and find the 2-harmonic function $T_0(z, \bar{z})$ by the formula $T_0(z, \bar{z}) = \text{Re}[\int \int T_1(z, \bar{z}) d\bar{z} dz] + h(z) + \overline{h(z)}$, where $h(z)$ is the analytic in D function reconstructed from the initial boundary conditions. So we construct the linear on t spline $T^1(t, x, y)$ for the function $T(t, x, y)$ in the subdomain D_1 .

The second step in our construction is to repeat the procedure of the first step in order to reconstruct the spline $T^2(t, x, y)$, $(t, x, y) \in [0, 1] \times D_2(0)$, for the function $T(t, x, y)$ in the subdomain $D_2(t)$. Both of these steps provide us with the values of $T(t, x, y)$ on $[0, 1] \times D$.

Finally, we find $\Gamma'(1)$ through the relation on the velocity $v = \theta_1 \partial_\nu T^1(1, x(s), y(s)) - \theta_2 \partial_\nu T^2(1, x(s), y(s))$, $(x(s), y(s)) \in \Gamma'(0)$, of the points of this curve. Since the approximate solution is linear on t for any $(x(s), y(s)) \in \Gamma'(0)$ the shift simply equals $\theta_1 \partial_\nu T^1(1, x(s), y(s)) - \theta_2 \partial_\nu T^2(1, x(s), y(s))$. Note that in general $T^1(1, x(s), y(s)) \neq T^2(1, x(s), y(s))$, $s \in [S_1, S]$. Indeed the temperature $T(1, x(s), y(s))|_{(x(s), y(s)) \in \Gamma'(1)} \equiv \text{const}$. So we again know the boundary data necessary for the construction of the spline for $t \in [1, 2]$. We then reconstruct the spline for $t \in [1, 2]$, and so on.

Naturally, the linear solution is not smooth on the variable t . In order to overcome this we construct a solution as a polynomial on t spline in each slice $[N-1, N] \times D_i$, $i = 1, 2$. Solution then takes the form $T^1(t, z, \bar{z}) = \sum_{j=0}^{2K-1} t^j T_j(z, \bar{z})$, here K is the order of differentiability of the spline. Again, the function $T_{2K-1}(z, \bar{z})$ is harmonic on $z, \bar{z}$. The function $T_{2K-2}(z, \bar{z}) = \text{Re}[\int \int T_{2K-1}(z, \bar{z}) d\bar{z} dz] + h_1(z) + \overline{h_1(z)}$, for an analytic function $h(z)$ is biharmonic on $z, \bar{z}$. Finally, the function $T_0(z, \bar{z})$ is a $2K$ -harmonic function. We start our reconstruction of the solution from the function $T_{2K-1}(z, \bar{z})$ and move down to $T_0(z, \bar{z})$ in a way similar to that for the linear spline.

3 Analytic foundation of solution

In order to solve the boundary value problem for $T_1(z, \bar{z})$, we apply the techniques developed in [7, 8].

Let the union of curves $\Gamma_1 \cup \Gamma'(0)$ be parameterized by the arc length $s \in [0, S]$. Assume that $s \in [0, S_1]$ corresponds to the curve Γ_1 . The function $f(z(s)) = \psi(s) + i\psi(s)$ represents the boundary values of the function $f(z)$ analytic in the domain $D_1(0)$ if and only if the following holds true [6, 13]:

$$f(z(s)) = \frac{1}{\pi i} \int_0^S \frac{f(z(t))z'(t)}{z(t) - z(s)} dt, \quad s \in \partial D_1(0) = z(s), \quad s \in [0, S]. \quad (2)$$

Assume that $f(z(s))|_{\Gamma_1} = \psi(s) + i\psi(s)$, $s \in [0, S_1]$, and $f(z(s))|_{\Gamma'(0)} = \tilde{\psi}(s) + i\tilde{\chi}(s)$, $s \in [S_1, S]$. Here $\psi(s) = U_1(s) - U_0(s)$, $\psi(s) = \int_0^s V_1(t) - V_0(t) dt$, $s \in [0, S_1]$. In this case we arrive to the integral equation similar to that of [8]:

$$\begin{aligned} \psi(s) + i\psi(s) &= \frac{1}{\pi i} \int_{\Gamma_1} \frac{(\psi(t) + i\psi(t))z'(t)}{z(t) - z(s)} dt \\ &+ \frac{1}{\pi i} \int_{\Gamma'(0)} \frac{(\tilde{\psi}(t) + i\tilde{\chi}(t))z'(t)}{z(t) - z(s)} dt, \quad s \in [0, S_1], \end{aligned} \quad (3)$$

$$\begin{aligned} \tilde{\psi}(s) + i\tilde{\chi}(s) &= \frac{1}{\pi i} \int_{\Gamma_1} \frac{(\psi(t) + i\psi(t))z'(t)}{z(t) - z(s)} dt \\ &+ \frac{1}{\pi i} \int_{\Gamma'(0)} \frac{(\tilde{\psi}(t) + i\tilde{\chi}(t))z'(t)}{z(t) - z(s)} dt, \quad s \in [S_1, S]. \end{aligned} \quad (4)$$

Lemma 1. Solutions $\tilde{\psi}(s)$, $\tilde{\chi}(s)$ of integral equation (3), (4) minimize the functions $\|f(z) - \frac{1}{\pi i} \int \frac{f(z(s))ds}{z(t) - z(s)}\|_{2, \Gamma_1}$, $\|f(z) - \frac{1}{\pi i} \int \frac{f(z(s))ds}{z(t) - z(s)}\|_{2, \Gamma'(0)}$, respectively, on the set of all analytic in $D_1(0)$ functions that equal $\psi(s) + i\psi(s)$ on $\partial D_1(0)$.

Proof. Consider the variation equation $\int_{\partial D_1} |f(z(s))|^2 ds$. Here it turns into the relation

$$\int_{\partial D_1(0)} \left\{ \left(\operatorname{Re}[f(z(t))] - \int_{\Gamma_1} \operatorname{Re}\left[\frac{(\psi(s) + i\psi(s))z'(s)}{z(t) - z(s)}\right] ds + \int_{\Gamma'(0)} \operatorname{Re}\left[\frac{f(s)z'(s)}{z(t) - z(s)}\right] ds \right)^2 \right.$$

$$+ \left(\operatorname{Im}[f(z(t))] - \int_{\Gamma_1} \operatorname{Im} \left[\frac{(\psi(s) + i\psi(s))z'(s)}{z(t) - z(s)} \right] ds + \int_{\Gamma'(0)} \operatorname{Im} \left[\frac{f(z(s))z'(s)}{z(t) - z(s)} \right] ds \right)^2 dt.$$

This relation splits into the sum of equations relative to (3) and (4).

We then simply variate this relation with respect to $f(z(s))$ and achieve relations (3) and (4). □

Hence, we solve the integral equation (4) and find the function closest to an analytic one on $D_1(0)$ that equals $\psi(s) + i\psi(s)$ on Γ_1 . So the real problem is to find the conditions under which equation (4) is resolvable.

Theorem 1. Assume that the given curve Γ_1 is a simple curve that does not intersect the simple curve $\Gamma'(0)$. Assume also that the functions $U_j(z(s), \overline{z(s)})|_{z(s) \in \Gamma_1}$, $V_j(z(s), \overline{z(s)})|_{z(s) \in \Gamma_1}$, $j = 0, 1$, meet the Hölder condition. Then we can reconstruct $T_1(z(s), \overline{z(s)})|_{z(s) \in \Gamma'(0)}$ as solution of equation (4).

Proof. Let us solve the second integral equation.

Any curve connecting the endpoints of the curve Γ_1 is longer than the straight line connecting these points. Consider then the case of the line segment.

Assume that we approximate the function $\tilde{\psi}(s) + i\tilde{\chi}(s)$ by a piece-wise linear function, that is, a linear spline. The approximate calculation of the singular principal value Cauchy integral over the part of the curve relative to Γ_1 is as follows:

$$\frac{1}{\pi i} \int_{\Gamma_1} \frac{f(e^{i\theta}) i e^{i\theta} d\theta}{e^{i\theta} - e^{it_k}},$$

at the intermediate node $t_k = \pi k/(N)$, $k = 1, \dots, N$, $p = 0, 2, \dots, 2N$, can be performed by application of the uniform fragmentation of the intervals $[\varphi_{j-1}, \varphi_j]$, $j = 1, \dots, N$. So the matrix of the system corresponding to the straight segment equals

$$\begin{pmatrix} 0 & \frac{1}{2} & \frac{1}{3} & \dots & \frac{1}{n} \\ -\frac{1}{2} & 0 & \frac{1}{2} & \dots & \frac{1}{n-1} \\ -\frac{1}{3} & -\frac{1}{2} & 0 & \dots & \frac{1}{n-1} \\ \vdots & \vdots & \vdots & \ddots & \vdots \\ -\frac{1}{n} & -\frac{1}{n-1} & \dots & -\frac{1}{2} & 0 \end{pmatrix}.$$

The norm of the Hilbert matrix then majorizes the norm of this one. We estimated the norm of the matrix in [8].

For any curve different from the straight segment the matrix of the system equals

$$\begin{pmatrix} 0 & \frac{1}{2l_{1,2}} & \frac{1}{3l_{1,3}} & \cdots & \frac{1}{nl_{1,n}} \\ -\frac{1}{2l_{1,2}} & 0 & \frac{1}{2l_{2,3}} & \cdots & \frac{1}{(n-1)l_{2,n-1}} \\ -\frac{1}{3l_{1,3}} & -\frac{1}{2l_{2,3}} & 0 & \cdots & \frac{1}{(n-1)l_{3,n-1}} \\ \vdots & \vdots & \vdots & \ddots & \vdots \\ -\frac{1}{nl_{1,n}} & -\frac{1}{(n-1)l_{2,n}} & \cdots & -\frac{1}{2l_{n-1,n}} & 0 \end{pmatrix}.$$

Here $l_{i,j}$ is the length of the segment connecting the i th and the j th node divided by the length of the line segment connecting these nodes. Since $l_{i,j} \geq 1$ for all $i \neq j$ the norm of the matrix for any curve is less than or equal to π .

So we can replace either of the singular integral equations (3) or (4) with respect to the values of the unknown functions at the intermediate nodes with the linear system with the block matrix of order N . We investigate the solvability of this finite system for sufficiently large N . Let N tend to infinity. The finite system tends to the infinite one with the matrix $I - \frac{1}{\pi}M$, where $M = [a_{kj}]_{k,j}^\infty$, and

$$a_{kj} = \begin{cases} -\ln\left(1 + \frac{2}{2k-2j-1}\right) & \text{if } j < k, \\ \ln\left(1 + \frac{2}{2j-2k-1}\right) & \text{if } j > k, \\ 0 & \text{if } j = k. \end{cases}$$

Here, I is the infinite identity matrix. The L_2 -norm of this matrix can be compared to that of the Hilbert matrix.

If M^* is the matrix conjugate of M , then the L_2 -norm of the product $M \cdot M^*$ is strictly less than that of the product $H \cdot H^*$, where $H = \left[\frac{1}{j-k}\right]_{j,k=1, j \neq k}^\infty$. To show this, it suffices to consider the series representation of the natural logarithm, keeping in mind that $\ln(1+x) \leq x$, and establish the estimates of the matrix M elements by those of H .

Therefore, due to the well-known results [15, Theorem 1], since the L_2 -norm of the matrix H equals π , the norm of the matrix M does not exceed π . Thus, the infinite matrix $I - \frac{1}{\pi}M$ is invertible, and the corresponding finite system with respect to $f(e^{it_k})$, $k = 1, \dots, N$, is also resolvable for sufficiently large N .

Finally, we conclude that singular integral equations (3) and (4) can be resolved approximately. The convergence to the exact solution depends first on the accuracy of the approximation of the initial boundary data. Since the boundary data is approximated by the linear spline, the error equals $O(\max(\omega(\psi, \Delta x)\Delta x, \omega(\psi, \Delta x)\Delta x))$ [2].

The second obstacle in our approximation is the inversion of the matrix M of our system. Assume that the maximal subinterval equals Δx . Since the elements of the matrix M differ from the exact values by $O(\Delta x)$, we have $O(\max\{\Delta x, \Delta\psi\|M^{-1}\|, \Delta\psi\|M^{-1}\|\})$ as the final estimate. $\square$

Assume that the time interval step is equal to Δt . Then we have the following result.

Corollary 1. Spline-interpolation solution error estimate is $O(\max\{\Delta x, \omega(\psi, \Delta t)\Delta t, \omega(\psi, \Delta t)\Delta t, \|M^{-1}\|\omega(\psi, \Delta x)\Delta x, \|M^{-1}\|\omega(\psi, \Delta x)\Delta x\})$.

For the variable t we apply the standard convergence estimate of $O(\Delta t^{2K})$ [2].

Note. These constructions hold even in the case of approximation of the function $\tilde{\psi}(s) + i\tilde{\chi}(s)$ with the linear spline. The approximate calculation of the singular principal value Cauchy integral over the part of the curve relative to $\Gamma'(0)$

$$\frac{1}{\pi i} \int_{\Gamma'(0)} \frac{f(e^{i\theta})ie^{i\theta}d\theta}{e^{i\theta} - e^{it_k}},$$

at the intermediate node $t_k = \pi k/(N) - \pi/(2N)$, $k = 1, \dots, N$, can be performed by application of the uniform fragmentation of the interval $[\pi, 2\pi]$. So we assume that the subintervals of both Γ_1 and $\Gamma'(0)$ are equally spaced and uniformly divided. We apply the following approximation formula:

$$\begin{aligned} & \frac{1}{\pi i} \int_0^\pi \frac{f(e^{i\theta})ie^{i\theta}d\theta}{e^{i\theta} - e^{it_k}} \\ & \approx \frac{1}{\pi i} \sum_{j=1, j \neq k}^N f\left(\frac{\pi j}{N} - \frac{\pi}{2N}\right) \ln \frac{e^{\pi i(2(j-k)+1)/(2N)} - 1}{e^{\pi i(2(j-k-1)+1)/(2N)} - 1} \\ & \quad + \frac{1}{\pi i} f\left(\frac{\pi k}{N} - \frac{\pi}{2N}\right) \ln \frac{e^{\pi i/(2N)} - 1}{e^{-\pi i/(2N)} - 1} \\ & = \frac{1}{\pi i} \sum_{j=1}^{k-1} f\left(\frac{j}{N} - \frac{\pi}{2N}\right) \left[\frac{\pi i}{2N} + \ln \frac{\sin((2(k-j)-1)\pi/(4N))}{\sin((2(k-j+1)-1)\pi/(4N))} \right] \\ & \quad + f\left(\frac{k\pi}{N} - \frac{\pi}{2N}\right) \frac{1}{2N} + \frac{1}{\pi i} \sum_{j=k+1}^N f\left(\frac{j\pi}{N} - \frac{\pi}{2N}\right) \left[\frac{\pi i}{2N} \right. \\ & \quad \left. + \ln \frac{\sin((2(j-k)+1)\pi/(4N))}{\sin((2(j-k-1)+1)\pi/(4N))} \right]. \end{aligned}$$

There exist singularities in the neighborhood of the meeting points of the curves Γ_1 and $\Gamma'(0)$. The matrix of the relative finite algebraic linear system should be modified so that the approximate solution at these points is continuous.

The second boundary value problem in our construction is the reconstruction of the imaginary part of the analytic function $f(z)$ such that

$$T_0(z, \bar{z}) = \iint T_1(z, \bar{z}) dz d\bar{z} + f(z) + \overline{f(z)},$$

by the value of its real part on the curve $\partial D_1(0) = \Gamma_1 \cup \Gamma'(0)$.

We again consider the necessary and sufficient condition on the function $f(z)$ analyticity:

$$f(z(s)) = \frac{1}{\pi i} \int_0^S \frac{f(z(t))z'(t)}{z(t) - z(s)} dt, \quad s \in [0, S].$$

Assume that $f(z(s)) = U_0(z(s)) + iv(z(s))$, where $z(s) \in \partial D_1$. Then, by the values of $U_0(z(s))$, $s \in [0, S]$, it should be possible to find $v(s) = v(z(s))$. We reduce this problem to the solution of the linear equation system for the set of coefficients with $\sin(ks)$, $\cos(ks)$, for $k = 1, \dots, N$, as follows:

$$\begin{pmatrix} AA & AB \\ BA & BB \end{pmatrix} \begin{pmatrix} \theta \\ \gamma \end{pmatrix} = \begin{pmatrix} F \\ G \end{pmatrix}, \quad (5)$$

here

$$\theta = \begin{pmatrix} \theta_1 \\ \vdots \\ \theta_M \end{pmatrix}, \quad \gamma = \begin{pmatrix} \gamma_1 \\ \vdots \\ \gamma_M \end{pmatrix}.$$

The vectors

$$F = \begin{pmatrix} f_1 \\ \vdots \\ f_M \end{pmatrix}, \quad G = \begin{pmatrix} g_1 \\ \vdots \\ g_M \end{pmatrix}.$$

in the right-hand side of this system consist of

$$f_l = \left(-\sum_j \text{Im} z_j^l + \sum_k \text{Im} \hat{z}_k^{-l} \right) \frac{1}{l} + \frac{1}{\pi^2} \int_0^{2\pi} u(\tau) d\tau \int_0^{2\pi} L(\tau, t) \cos(lt) dt,$$

$$g_l = \left(\sum_j \text{Re} z_j^l + \sum_k \text{Re} \hat{z}_k^{-l} \right) \frac{1}{l} + \frac{1}{\pi^2} \int_0^{2\pi} u(\tau) d\tau \int_0^{2\pi} L(\tau, t) \sin(lt) dt,$$

for $l = 1, \dots, M$. Here

$$\begin{aligned} K(\tau, t) &= \operatorname{Im} \left[\ln \frac{z(\tau) - z(t)}{e^{i\tau} - e^{it}} \right]'_{\tau} \\ &= \operatorname{Im} \left[\ln \left(\sum_{k=1}^n c_k e^{ikt} \sum_{l=0}^{k-1} e^{il(\tau-t)} - \sum_{j=1}^m c_{-j} e^{-ij\tau} \sum_{l=0}^{j-1} e^{il(\tau-t)} \right) \right]'_{\tau}, \end{aligned} \quad (6)$$

and

$$\begin{aligned} L(\tau, t) &= \operatorname{Re} \left[\ln \frac{z(\tau) - z(t)}{e^{i\tau} - e^{it}} \right]'_{\tau} \\ &= \operatorname{Re} \left[\ln \left(\sum_{k=1}^n c_k e^{ikt} \sum_{l=0}^{k-1} e^{il(\tau-t)} - \sum_{j=1}^m c_{-j} e^{-ij\tau} \sum_{l=0}^{j-1} e^{il(\tau-t)} \right) \right]'_{\tau}. \end{aligned} \quad (7)$$

Block-matrices of M

$$AA, AB, BA, BB$$

consist of

$$\begin{aligned} AA &= \left(\delta_{ln} - \frac{1}{\pi^2} \int_0^{2\pi} \cos(n\tau) d\tau \int_0^{2\pi} K(\tau, t) \cos(lt) dt \right)_{l,n=1}^M, \\ AB &= \left(-\frac{1}{\pi^2} \int_0^{2\pi} \sin(n\tau) d\tau \int_0^{2\pi} K(\tau, t) \cos(lt) dt \right)_{l,n=1}^M, \\ BA &= \left(-\frac{1}{\pi^2} \int_0^{2\pi} \cos(n\tau) d\tau \int_0^{2\pi} K(\tau, t) \sin(lt) dt \right)_{l,n=1}^M, \\ BB &= \left(\delta_{ln} - \frac{1}{\pi^2} \int_0^{2\pi} \sin(n\tau) d\tau \int_0^{2\pi} K(\tau, t) \sin(lt) dt \right)_{l,n=1}^M, \end{aligned}$$

here δ_{ln} is the Kronecker delta-function. The convergence rate of this part of the solution is $O(\frac{1}{M^2})$, and M equals the size of the block in the matrix of the system. Note that for contours with singularities or for boundaries of thin domains this estimate is not really effective and requires certain additional work [7].

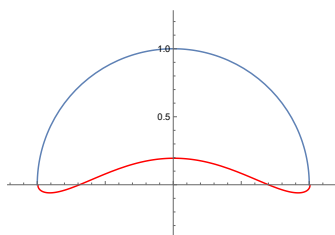
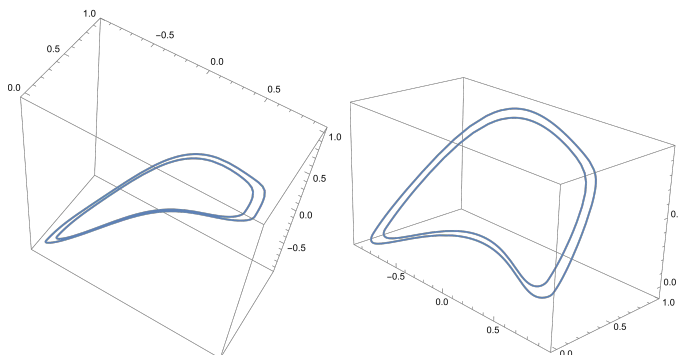


Figure 2: Domain of Example 1

Figure 3: Real and imaginary parts of the function $T_1(z, \bar{z})$

4 Examples

Here we present the reconstruction of the functions $T_1(z, \bar{z})$, $T_0(z, \bar{z})$, that is, the principle part of the approximate solution construction.

Example 1. Consider the upper half of the unit circle as Γ_1 and the curve given by the equation $(\cos(t), 0.1 \sin(3t) - 0.1 \sin(t))$, $t \in [\pi, 2\pi]$ as $\Gamma'(0)$ (see Figure 2). Note that the domain does not contain the coordinate origin. So the reconstructed function may have a pole at 0. The boundary values of T_1 are given as $\cos(t) + i(\sin(t) + 0.15 \sin(3t) + 0.2 \sin(4t))$.

The real and imaginary parts of the reconstructed function $T_1(x, y)$ are given in Figure 3. The surface of the reconstructed function T_0 is at Figure 4.

Let us now consider the example with more acute angle between the initial free and the fixed boundaries.

Example 2. Consider the upper half of the unit circle as Γ_1 and the curve given by the equation $(\cos(t), 0.1 \sin(t) + 0.1 \sin(2t))$, $t \in [\pi, 2\pi]$ as $\Gamma'(0)$ (see Figure 5). The boundary values of T_1 are given as $\cos(t) + i(\sin(t) + 0.15 \sin(3t) + 0.2 \sin(4t))$

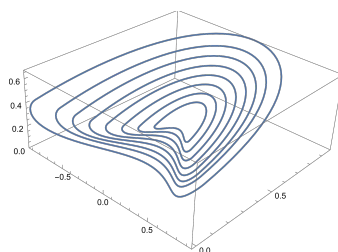
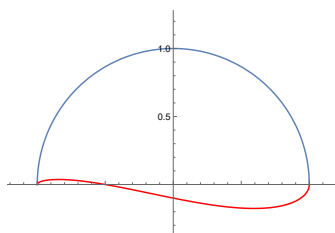
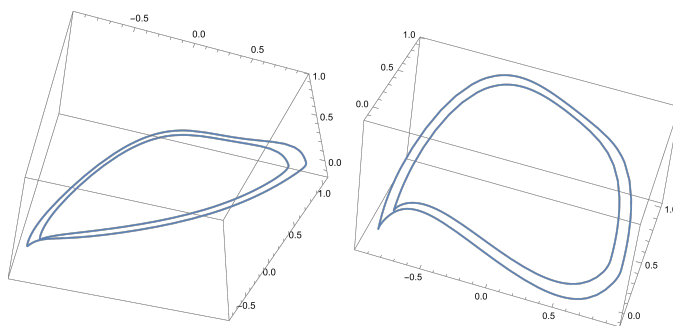
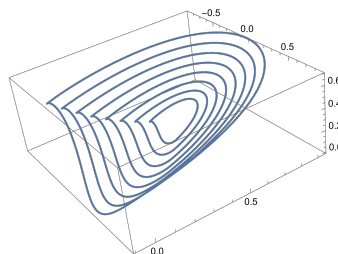
Figure 4: Real part of the function $T_0(z, \bar{z})$ 

Figure 5: Domain of Example 2

Figure 6: Real and imaginary parts of the function $T_1(z, \bar{z})$ Figure 7: Real part of the function $T_0(z, \bar{z})$

The real and imaginary parts of the reconstructed function $T_1(x, y)$ are given on Figure 6. The surface of the reconstructed function T_0 is at Figure 7.

Funding

This research received no external funding.

Conflicts of Interest

The author declares no conflicts of interest.

References

- [1] Apushkinskaya, D.E. and Lazareva, G.G. *Algorithm for the numerical solution of the Stefan problem and its application to calculations of the temperature of Tungsten under impulse action*, Contemp. Math. Fundam. Dir., 67(3) (2021), 442–454.
- [2] de Boor, C. *A practical guide to splines*. Springer, 2001.
- [3] Gurkov, Y.V. and Petrova, A.G. *Numerical solution of the stefan problem with a variable phase-transition temperature*, J. Appl. Mech. Tech. Phys., 37 (1996), 547–553.
- [4] Kinderlehrer, D. and Stampacchia, G. *An introduction to variational inequalities and their applications*. SIAM, 2000.
- [5] Meirmanov, A.M. *The Stefan problem*, Berlin, New York, De Gruyter, 1992.
- [6] Gakhov, F.D. *Boundary value problems*, Pergamon Press, Oxford, 1966.
- [7] Ivanshin, P.N. and Shirokova, E.A. *The approximate conformal mapping of a disk onto domain with an acute angle*, Int. J. Appl. Comput. Math. 9(5), (2023), 54
- [8] Shirokova, E.A. and Ivanshin, P.N. *On Cauchy problem solution for a harmonic function in a simply connected domain*, Issues Anal., 12(30) (2023), 87–96.
- [9] Khasanov, A.B. and Tursunov, F.R. *Of Cauchy problem for Laplace equation*, Ufa Math. J. 11(4) (2019), 91–107.

- [10] Khasanov, A.B. and Tursunov, F.R. *On Cauchy problem for the three-dimensional Laplace equation*, Rus. Math., 65(2) (2021), 49–64.
- [11] Lavrent'ev, M.M., Romanov, V.G. and Shishatskiy, S.P. *Ill-posed problems of mathematical physics and analysis*, Transl. of Math. Monographs, Amer. Math. Soc., 1986.
- [12] Markushevich, A.I. *Theory of analytic functions*, GTTL, Moscow, 1950.
- [13] Muskhelishvili, N. I. *Singular integral equations*, Springer, 1958.
- [14] Saini, S., Das, P. and Kumar, S. *Computational cost reduction for coupled system of multiple scale reaction diffusion problems with mixed type boundary conditions having boundary layers*, Rev. Real Acad. Cienc. Exactas Fis. Nat. Ser. A-Mat. 117(2) (2023) 66.
- [15] Schur, I. *Bemerkungen zur Theorie der Beschraenkten Bilinearformen mit unendlich vielen Veraenderlichen*, J. Reine Angew. Math. 140 (1911) 1–28.
- [16] Shiromani, R., Shanthi, V. and Das, P. *A higher order hybrid-numerical approximation for a class of singularly perturbed two-dimensional convection-diffusion elliptic problem with non-smooth convection and source terms*, Comput. Math. Appl., 142 (2023), 9–30.
- [17] Shlapunov, A.A. *Iterations of self-adjoint operators and their applications to elliptic systems*, Math. Nachrichten, 218 (2000), 165–174.
- [18] Tarkhanov, N. *The Cauchy problem for solutions of elliptic equations*, Akademie-Verlag, Berlin, 1995.



Local fractional Khalouta transform: Properties and applications

A. Khalouta*

Abstract

In the current paper, we propose a new hybrid method for solving local fractional differential equations. The proposed method is based on the coupling of the Khalouta transform with the local fractional derivative, and we call the local fractional Khalouta transform method. Also, we provide important results and properties of this method. For justification and verification of the present method, some applications of linear local fractional differential equations are presented. The results of the solved examples show the effectiveness of this method. Moreover, the present method can be applied to handle the solutions of various problems with local fractional derivatives.

AMS subject classifications (2020): 44A05, 44A20, 26A33, 34A08.

Keywords: Khalouta transform; Local fractional derivative; Local fractional differential equation; Local fractional Volterra integro-differential equation; Exact solution.

*Corresponding author

Received 28 April 2025; revised 30 November 2025; accepted 24 December 2025

Ali Khalouta

Laboratory of Fundamental and Numerical Mathematics, Department of Mathematics, Faculty of Sciences, Setif 1 University-Ferhat ABBAS, Algeria.

e-mail: nadjibkh@yahoo.fr ; ali.khalouta@univ-setif.dz

How to cite this article

Khalouta, A., Local fractional Khalouta transform: Properties and applications. *Iran. J. Numer. Anal. Optim.*, 2026; 16(2): 539-562. <https://doi.org/10.22067/ijnao.2025.93216.1634>

1 Introduction

In recent years, fractional calculus has attracted increasing attention due to its applications in various fields. The history of this theory dates back to the 17th century, when in 1695 Leibniz [16] described the derivative of order $\theta = \frac{1}{2}$ in his letter to L'Hospital. Since then, the new theory has become very attractive to mathematicians as well as physicists, biologists, engineers, and economists. Fractional calculus has various applications in biophysics, quantum mechanics, wave theory, polymers, continuum mechanics, Lie theory, field theory, spectroscopy, group theory, and other applications in science and engineering [7, 18, 21, 22].

There are many concepts of fractional derivatives defined and developed in the literature, which are not compatible with each other, among them is, Riemann–Liouville fractional derivative [23], Caputo fractional derivative [15], conformable fractional derivative [1], Hadamard fractional derivative [14], Atangana–Baleanu fractional derivative [2], and local fractional derivative [25].

Today, fractional differential equations involving the local fractional derivative constitute an important area of modeling natural phenomena, and researchers are paying considerable attention to them due to their numerous applications in science and engineering. New developments in this field have emerged in disciplines such as mathematics, physics, chemistry, biology, medicine, economics, and engineering, highlighting the growing importance of this type of equation in the analysis of real-world problems. For this reason, researchers in mathematics and physics have focused on proposing various methods for solving local fractional differential equations [3, 4, 5, 9].

There are many computational methods for dealing with local fractional differential equations, such as the local fractional decomposition method [10], the local fractional variational iteration method [8], the local fractional differential transform method [6], the local fractional reduced differential transform method [17], the local fractional homotopy perturbation method [27], the local fractional homotopy analysis method [20], and others.

Recently, the author [12, 13] proposed a new integral transform called Khalouta transform, which is a generalization of many well known integral transforms such as Laplace–Carson transform, Sumudu transform, ZZ transform, ZMA transform, Elzaki transform, Aboodh transform, Natural transform, and Shehu transform. Thus, the main objective of the present paper is to propose a new hybrid method based on the coupling of the Khalouta transform with the local fractional derivative for solving local fractional differential equations.

The proposed method offers great advantages of straightforward applicability, computational efficiency and high accuracy. Its strength lies in its ability to combine two powerful concepts for obtaining exact solutions of linear local fractional differential equations. Unlike other meth-

ods, the proposed method requires neither linearization nor discretization, nor small or large parameters in the problem.

The structure of this paper is as follows. In Section 2, we give the basic definitions and properties of the local fractional calculus. In Section 3, we present important results regarding the local fractional Khalouta transform. In Section 4, we illustrate the relation between the local fractional Khalouta transform and other known local fractional integral transforms. In Section 5, we extend this new integral transform and apply it to solve linear differential equations using the local fractional derivative. Finally, in Section 6, we present our conclusions.

2 Local fractional calculus

This section gives the basic concepts of local fractional calculus, and in particular the local fractional derivative and the local fractional integral, which have been utilized in this work. For more details, see [25, 26].

Definition 1. If we have the relation

$$|u(t) - u(t_0)| < \varepsilon^\theta, \quad 0 < \theta \leq 1,$$

where there is $|t - t_0| < \delta$ for $\delta, \varepsilon \in \mathbb{R}_+^*$, then the function $u(t)$ is called local fractional continuous at $t = t_0$.

We say that the function $u(t)$ is local fractional continuous on the interval (a, b) if it is local fractional continuous at any point t_0 of (a, b) , and we denote it by $u(t) \in C_\theta(a, b)$, $0 < \theta \leq 1$.

Definition 2. Consider the function $u(t) \in C_\theta(a, b)$. The local fractional integral of $u(t)$ of order θ in the interval $[a, b]$ is defined by

$$\begin{aligned} {}_a I_b^{(\theta)} u(t) &= \frac{1}{\Gamma(1 + \theta)} \int_a^b u(\tau) (d\tau)^\theta \\ &= \frac{1}{\Gamma(1 + \theta)} \lim_{\Delta\tau \rightarrow 0} \sum_{i=0}^{N-1} u(\tau_i) (\Delta\tau_i)^\theta, \end{aligned}$$

where $\Delta\tau_i = \tau_{i+1} - \tau_i$ with $\Delta\tau = \max\{\Delta\tau_0, \Delta\tau_1, \Delta\tau_2, \Delta\tau_3, \dots\}$ and $[\tau_{i+1}, \tau_i]$, $i = 0, \dots, N - 1$, $\tau_0 = a$, $\tau_N = b$ is a partition of the interval $[a, b]$.

Definition 3. Consider the function $u(t) \in C_\theta(a, b)$. The local fractional derivative of $u(t)$ of order θ with $0 < \theta \leq 1$ in the interval $[a, b]$ is defined by

$$D^{(\theta)}u(t) = \frac{d^\theta u(t_0)}{dt^\theta} = u^{(\theta)}(t) = \lim_{t \rightarrow t_0} \frac{\Delta^\theta(u(t) - u(t_0))}{(t - t_0)^\theta},$$

where

$$\Delta^\theta(u(t) - u(t_0)) \simeq \Gamma(1 + \theta) [u(t) - u(t_0)].$$

Moreover, the local fractional derivative of higher order is defined as

$$D^{(n\theta)}u(t) = \frac{d^{n\theta}u(t_0)}{dt^{n\theta}} = u^{(n\theta)}(t) = \underbrace{D^{(\theta)}D^{(\theta)} \dots D^{(\theta)}}_{n\text{-times}}u(t).$$

Definition 4. The local fractional partial differential operator of order θ with $0 < \theta \leq 1$ is given by

$$D_t^{(\theta)}u(x, t) = \frac{\partial^\theta u(x, t_0)}{\partial t^\theta} = u^{(\theta)}(x, t_0) = \lim_{t \rightarrow t_0} \frac{\Delta^\theta(u(x, t) - u(x, t_0))}{(t - t_0)^\theta},$$

where

$$\Delta^\theta(u(x, t) - u(x, t_0)) \simeq \Gamma(1 + \theta) [u(x, t) - u(x, t_0)].$$

Moreover, the local fractional partial derivative of higher order is defined as

$$D_t^{(n\theta)}u(x, t) = \frac{\partial^{n\theta}u(x, t_0)}{\partial t^{n\theta}} = u^{(n\theta)}(x, t) = \underbrace{D_t^{(\theta)}D_t^{(\theta)} \dots D_t^{(\theta)}}_{n\text{-times}}u(x, t).$$

Definition 5. The local fractional Mc-Laurin's series of the Mittag-Leffler function is given by

$$E_\theta(z^\theta) = \sum_{n=0}^{\infty} \frac{z^{n\theta}}{\Gamma(1 + n\theta)}, \quad z \in \mathbb{R}, 0 < \theta \leq 1.$$

3 Main results

This section presents our main results concerning the definition of local fractional Khalouta transform denoted in this paper by ${}^{LF}\mathbb{KH}_\theta[\cdot]$ and discusses some interesting properties of this transform.

Definition 6. Let

$$\frac{1}{\Gamma(1 + \theta)} \int_0^{+\infty} |u(t)| (dt)^\theta < A < \infty.$$

The local fractional Khalouta transform of the function $u(t)$ of order θ is defined as by the following integral:

$$\begin{aligned}
{}^{LF}\mathbb{KH}_\theta[u(t)] &= \mathcal{K}_\theta(s, \gamma, \eta) = \frac{1}{\Gamma(1+\theta)} \frac{s^\theta}{\gamma^\theta \eta^\theta} \int_0^\infty E_\theta\left(-\frac{s^\theta t^\theta}{\gamma^\theta \eta^\theta}\right) u(t) (dt)^\theta \\
&= \lim_{\sigma \rightarrow \infty} \frac{1}{\Gamma(1+\theta)} \frac{s^\theta}{\gamma^\theta \eta^\theta} \int_0^\sigma E_\theta\left(-\frac{s^\theta t^\theta}{\gamma^\theta \eta^\theta}\right) u(t) (dt)^\theta,
\end{aligned} \tag{1}$$

where s^θ, γ^θ , and η^θ with $0 < \theta \leq 1$, are the Khalouta transform variables, σ is a real number, and the integral is taken along the line $t = \sigma$.

Definition 7. The inverse local fractional Khalouta transform of the function $u(t)$ of order θ is defined by

$${}^{LF}\mathbb{KH}_\theta^{-1}[\mathcal{K}_\theta(s, \eta, \gamma)] = u(t).$$

Now, we give some interesting properties of the local fractional Khalouta transform.

Theorem 1. Let ${}^{LF}\mathbb{KH}_\theta[u(t)] = \mathcal{K}_\theta(s, \gamma, \eta)$ and ${}^{LF}\mathbb{KH}_\theta[v(t)] = \mathcal{H}_\theta(s, \gamma, \eta)$. Then, we have

$${}^{LF}\mathbb{KH}_\theta[\lambda u(t) \pm \theta v(t)] = \lambda \mathcal{K}_\theta(s, \gamma, \eta) \pm \theta \mathcal{H}_\theta(s, \gamma, \eta),$$

where λ and θ are constants.

Proof. Using the definition of the local fractional Khalouta transform, we have

$$\begin{aligned}
\mathbb{KH}[\lambda u(t) \pm \theta v(t)] &= \frac{1}{\Gamma(1+\theta)} \frac{s^\theta}{\gamma^\theta \eta^\theta} \int_0^\infty E_\theta\left(-\frac{s^\theta t^\theta}{\gamma^\theta \eta^\theta}\right) [\lambda u(t) \pm \theta v(t)] (dt)^\theta \\
&= \frac{1}{\Gamma(1+\theta)} \frac{s^\theta}{\gamma^\theta \eta^\theta} \int_0^\infty \left[\lambda E_\theta\left(-\frac{s^\theta t^\theta}{\gamma^\theta \eta^\theta}\right) u(t) \right. \\
&\quad \left. \pm \theta E_\theta\left(-\frac{s^\theta t^\theta}{\gamma^\theta \eta^\theta}\right) v(t) \right] (dt)^\theta \\
&= \lambda \frac{1}{\Gamma(1+\theta)} \frac{s^\theta}{\gamma^\theta \eta^\theta} \int_0^\infty E_\theta\left(-\frac{s^\theta t^\theta}{\gamma^\theta \eta^\theta}\right) u(t) (dt)^\theta \\
&\quad \pm \theta \frac{1}{\Gamma(1+\theta)} \frac{s^\theta}{\gamma^\theta \eta^\theta} \int_0^\infty E_\theta\left(-\frac{s^\theta t^\theta}{\gamma^\theta \eta^\theta}\right) v(t) (dt)^\theta \\
&= \lambda \mathcal{K}_\theta(s, \gamma, \eta) \pm \theta \mathcal{H}_\theta(s, \gamma, \eta).
\end{aligned}$$

□

Theorem 2. Let ${}^{LF}\mathbb{KH}_\theta[u(t)] = \mathcal{K}_\theta(s, \gamma, \eta)$ where $0 < \theta \leq 1$. Then, the local fractional Khalouta transforms of the local fractional derivatives are given by

$${}^{LF}\mathbb{KH}_\theta\left[D^{(\theta)}u(t)\right] = \left(\frac{s}{\gamma\eta}\right)^\theta \mathcal{K}_\theta(s, \gamma, \eta) - \left(\frac{s}{\gamma\eta}\right)^\theta u(0), \tag{2}$$

and

$${}^{LF}\mathbb{KH}_\theta \left[D^{(n\theta)} u(t) \right] = \left(\frac{s}{\gamma\eta} \right)^{n\theta} \mathcal{K}_\theta(s, \gamma, \eta) - \sum_{k=0}^{n-1} \left(\frac{s}{\gamma\eta} \right)^{(n-k)\theta} u^{(k\theta)}(0). \quad (3)$$

Proof. We proof the formula (2). From the definition of local fractional Khalouta transform, we have

$${}^{LF}\mathbb{KH}_\theta \left[D^{(\theta)} u(t) \right] = \frac{1}{\Gamma(1+\theta)} \frac{s^\theta}{\gamma^\theta \eta^\theta} \int_0^\infty E_\theta \left(-\frac{s^\theta t^\theta}{\gamma^\theta \eta^\theta} \right) D^{(\theta)} u(t) (dt)^\theta.$$

Using the fractional integration by parts formula [11], we get the following equality:

$$\begin{aligned} {}^{LF}\mathbb{KH}_\theta \left[D^{(\theta)} u(t) \right] &= \frac{1}{\Gamma(1+\theta)} \frac{s^\theta}{\gamma^\theta \eta^\theta} \left(\lim_{\sigma \rightarrow \infty} \Gamma(1+\theta) \left[E_\theta \left(-\frac{s^\theta t^\theta}{\gamma^\theta \eta^\theta} \right) u(t) \right]_0^\sigma \right. \\ &\quad \left. + \frac{s^\theta}{\gamma^\theta \eta^\theta} \lim_{\sigma \rightarrow \infty} \int_0^\sigma E_\theta \left(-\frac{s^\theta t^\theta}{\gamma^\theta \eta^\theta} \right) u(t) (dt)^\theta \right) \\ &= \frac{1}{\Gamma(1+\theta)} \frac{s^\theta}{\gamma^\theta \eta^\theta} \\ &\quad \times \left(\begin{aligned} &-\Gamma(1+\theta) u(0) \\ &+ \frac{s^\theta}{\gamma^\theta \eta^\theta} \lim_{\sigma \rightarrow \infty} \int_0^\sigma E_\theta \left(-\frac{s^\theta t^\theta}{\gamma^\theta \eta^\theta} \right) u(t) (dt)^\theta \end{aligned} \right) \\ &= -\frac{s^\theta}{\gamma^\theta \eta^\theta} u(0) \\ &\quad + \frac{s^\theta}{\gamma^\theta \eta^\theta} \left(\frac{1}{\Gamma(1+\theta)} \frac{s^\theta}{\gamma^\theta \eta^\theta} \int_0^\infty E_\theta \left(-\frac{s^\theta t^\theta}{\gamma^\theta \eta^\theta} \right) u(t) (dt)^\theta \right) \\ &= -\frac{s^\theta}{\gamma^\theta \eta^\theta} u(0) + \frac{s^\theta}{\gamma^\theta \eta^\theta} \mathcal{K}_\theta(s, \gamma, \eta) \\ &= \left(\frac{s}{\gamma\eta} \right)^\theta \mathcal{K}_\theta(s, \gamma, \eta) - \left(\frac{s}{\gamma\eta} \right)^\theta u(0). \end{aligned}$$

To prove the validity of the formula (3), we use mathematical induction.

For $n = 1$ and from the formula (3), we have

$${}^{LF}\mathbb{KH}_\theta \left[D^{(\theta)} u(t) \right] = \left(\frac{s}{\gamma\eta} \right)^\theta \mathcal{K}_\theta(s, \gamma, \eta) - \left(\frac{s}{\gamma\eta} \right)^\theta u(0).$$

So, from (2), we note that the formula is true when $n = 1$.

Assume that the formula (3) is true for n . We get

$${}^{LF}\mathbb{KH}_\theta \left[D^{(n\theta)} u(t) \right] = \left(\frac{s}{\gamma\eta} \right)^{n\theta} \mathcal{K}_\theta(s, \gamma, \eta) - \sum_{k=0}^{n-1} \left(\frac{s}{\gamma\eta} \right)^{(n-k)\theta} u^{(k\theta)}(0). \quad (4)$$

We show that (3) is true for $n + 1$. Let $D^{(n\theta)} u(t) = U(t)$, where ${}^{LF}\mathbb{KH}_\theta [U(t)] = \mathcal{U}_\theta(s, \gamma, \eta)$ and from (2) and (4), we have

$$\begin{aligned}
{}^{LF}\mathbb{KH}_\theta \left[D^{((n+1)\theta)} u(t) \right] &= {}^{LF}\mathbb{KH}_\theta \left[D^{(\theta)} U(t) \right] \\
&= \left(\frac{s}{\gamma\eta} \right)^\theta \mathcal{U}_\theta(s, \gamma, \eta) - \left(\frac{s}{\gamma\eta} \right)^\theta U(0) \\
&= \left(\frac{s}{\gamma\eta} \right)^{(n+1)\theta} \mathcal{K}_\theta(s, \gamma, \eta) \\
&\quad - \sum_{k=0}^{n-1} \left(\frac{s}{\gamma\eta} \right)^{(n+1-k)\theta} u^{(k\theta)}(0) \\
&\quad - \left(\frac{s}{\gamma\eta} \right)^\theta D^{(n\theta)} u(0), \\
&= \left(\frac{s}{\gamma\eta} \right)^{(n+1)\theta} \mathcal{K}_\theta(s, \gamma, \eta) \\
&\quad - \sum_{k=0}^n \left(\frac{s}{\gamma\eta} \right)^{(n+1-k)\theta} u^{(k\theta)}(0).
\end{aligned}$$

So, the formula (3) is true for $n + 1$.

Thus, by the principle of mathematical induction for all $n > 1$, the formula (3) is true. $\square$

Theorem 3. Let ${}^{LF}\mathbb{KH}_\theta [u(t)] = \mathcal{K}_\theta(s, \gamma, \eta)$. Then, the local fractional Khalouta transform of the local fractional integral is given by

$${}^{LF}\mathbb{KH}_\theta \left[{}_0I_t^{(\theta)} u(t) \right] = \left(\frac{\gamma\eta}{s} \right)^\theta \mathcal{K}_\theta(s, \gamma, \eta), \quad 0 < \theta \leq 1.$$

Proof. Let

$$K(t) = {}_0I_t^{(\theta)} u(t). \quad (5)$$

Operating the local fractional derivative on (5), we attain

$$D^{(\theta)} K(t) = D^{(\theta)} {}_0I_t^{(\theta)} u(t) = u(t), \quad (6)$$

and

$$K(0) = 0.$$

Operating the local fractional Khalouta transform on (6), we attain

$${}^{LF}\mathbb{KH}_\theta \left[D^{(\theta)} K(t) \right] = {}^{LF}\mathbb{KH}_\theta [u(t)],$$

which implies

$$\left(\frac{s}{\gamma\eta} \right)^\theta ({}^{LF}\mathbb{KH}_\theta [K(t)]) = \mathcal{K}_\theta(s, \gamma, \eta).$$

Thus, we get

$${}^{LF}\mathbb{KH}_\theta \left[{}_0I_t^{(\theta)} u(t) \right] = \left(\frac{\gamma\eta}{s} \right)^\theta \mathcal{K}_\theta(s, \gamma, \eta).$$

□

Theorem 4. Let ${}^{LF}\mathbb{KH}_\theta [u(t)] = \mathcal{K}_\theta(s, \gamma, \eta)$ and ${}^{LF}\mathbb{KH}_\theta [v(t)] = \mathcal{H}_\theta(s, \gamma, \eta)$. Then, the local fractional Khalouta transform of their convolution is given by

$${}^{LF}\mathbb{KH}_\theta [(u(t) * v(t))_\theta] = \left(\frac{\gamma\eta}{s} \right)^\theta \mathcal{K}_\theta(s, \gamma, \eta) \mathcal{H}_\theta(s, \gamma, \eta),$$

where

$$(u(t) * v(t))_\theta = \frac{1}{\Gamma(1+\theta)} \int_0^\infty u(\tau) v(t-\tau) (d\tau)^\theta.$$

Proof. Using the definition of the local fractional Khalouta transform in (1), we arrive at

$$\begin{aligned} & {}^{LF}\mathbb{KH}_\theta [(u(t) * v(t))_\theta] \\ &= \frac{1}{\Gamma(1+\theta)} \frac{s^\theta}{\gamma^\theta \eta^\theta} \int_0^\infty E_\theta \left(-\frac{s^\theta t^\theta}{\gamma^\theta \eta^\theta} \right) (u(t) * v(t))_\theta (dt)^\theta \\ &= \frac{1}{\Gamma(1+\theta)} \frac{s^\theta}{\gamma^\theta \eta^\theta} \\ &\quad \times \int_0^\infty \left(\begin{aligned} & E_\theta \left(-\frac{s^\theta t^\theta}{\gamma^\theta \eta^\theta} \right) \\ & \times \left(\frac{1}{\Gamma(1+\theta)} \int_0^\infty u(\tau) v(t-\tau) (d\tau)^\theta \right) \end{aligned} \right) (dt)^\theta \\ &= \frac{1}{\Gamma(1+\theta)} \frac{1}{\Gamma(1+\theta)} \frac{s^\theta}{\gamma^\theta \eta^\theta} \\ &\quad \times \int_0^\infty \int_0^\infty \left(E_\theta \left(-\frac{s^\theta t^\theta}{\gamma^\theta \eta^\theta} \right) u(\tau) v(t-\tau) (d\tau)^\theta \right) (dt)^\theta. \end{aligned} \tag{7}$$

Substituting $x = t - \tau$ and $dx = dt$ in (7), we attain

$$\begin{aligned} & {}^{LF}\mathbb{KH}_\theta [(u(t) * v(t))_\theta] \\ &= \frac{1}{\Gamma(1+\theta)} \frac{1}{\Gamma(1+\theta)} \frac{s^\theta}{\gamma^\theta \eta^\theta} \\ &\quad \times \int_0^\infty \int_0^\infty E_\theta \left(-\frac{s^\theta (x+\tau)^\theta}{\gamma^\theta \eta^\theta} \right) u(\tau) v(x) (d\tau)^\theta (dx)^\theta \\ &= \frac{1}{\Gamma(1+\theta)} \frac{1}{\Gamma(1+\theta)} \frac{s^\theta}{\gamma^\theta \eta^\theta} \\ &\quad \times \int_0^\infty \int_0^\infty \left(\begin{aligned} & E_\theta \left(-\frac{s^\theta x^\theta}{\gamma^\theta \eta^\theta} \right) E_\theta \left(-\frac{s^\theta \tau^\theta}{\gamma^\theta \eta^\theta} \right) \\ & \times u(\tau) v(x) \end{aligned} \right) (d\tau)^\theta (dx)^\theta \end{aligned}$$

$$\begin{aligned}
&= \frac{\gamma^\theta \eta^\theta}{s^\theta} \left(\frac{1}{\Gamma(1+\theta)} \frac{s^\theta}{\gamma^\theta \eta^\theta} \int_0^\infty E_\theta \left(-\frac{s^\theta \tau^\theta}{\gamma^\theta \eta^\theta} \right) u(\tau) (d\tau)^\theta \right) \\
&\quad \times \left(\frac{1}{\Gamma(1+\theta)} \frac{s^\theta}{\gamma^\theta \eta^\theta} \int_0^\infty E_\theta \left(-\frac{s^\theta x^\theta}{\gamma^\theta \eta^\theta} \right) v(x) (dx)^\theta \right) \\
&= \left(\frac{\gamma \eta}{s} \right)^\theta \mathcal{K}_\theta(s, \gamma, \eta) \mathcal{H}_\theta(s, \gamma, \eta).
\end{aligned}$$

□

Now, we present the results related to the local fractional Khalouta transform of some special functions.

Theorem 5. Let $u(t) = 1$. Then, its local fractional Khalouta transform is given by

$${}^{LF}\mathbb{KH}_\theta[1] = 1.$$

Proof. By using (1), we attain

$$\begin{aligned}
{}^{LF}\mathbb{KH}_\theta[1] &= \frac{1}{\Gamma(1+\theta)} \frac{s^\theta}{\gamma^\theta \eta^\theta} \int_0^\infty E_\theta \left(-\frac{s^\theta t^\theta}{\gamma^\theta \eta^\theta} \right) (dt)^\theta \\
&= \frac{s^\theta}{\gamma^\theta \eta^\theta} \lim_{\sigma \rightarrow \infty} \left[-\frac{\gamma^\theta \eta^\theta}{s^\theta} E_\theta \left(-\frac{s^\theta t^\theta}{\gamma^\theta \eta^\theta} \right) \right]_0^\sigma \\
&= 1.
\end{aligned}$$

□

Theorem 6. Let $u(t) = \frac{t^\theta}{\Gamma(1+\theta)}$, $0 < \theta \leq 1$. Then, its local fractional Khalouta transform is given by

$${}^{LF}\mathbb{KH}_\theta \left[\frac{t^\theta}{\Gamma(1+\theta)} \right] = \frac{\gamma^\theta \eta^\theta}{s^\theta}.$$

Proof. By using (1) and the fractional integration by parts formula [11], we attain

$$\begin{aligned}
&{}^{LF}\mathbb{KH}_\theta \left[\frac{t^\theta}{\Gamma(1+\theta)} \right] \\
&= \frac{1}{\Gamma(1+\theta)} \frac{s^\theta}{\gamma^\theta \eta^\theta} \int_0^\infty E_\theta \left(-\frac{s^\theta t^\theta}{\gamma^\theta \eta^\theta} \right) \frac{t^\theta}{\Gamma(1+\theta)} (dt)^\theta \\
&= \frac{1}{\Gamma(1+\theta)} \frac{s^\theta}{\gamma^\theta \eta^\theta} \int_0^\infty \left(\left(-\frac{\gamma^\theta \eta^\theta}{s^\theta} E_\theta \left(-\frac{s^\theta t^\theta}{\gamma^\theta \eta^\theta} \right) \right)^{(\theta)} \right. \\
&\quad \left. \times \frac{t^\theta}{\Gamma(1+\theta)} \right) (dt)^\theta \\
&= \frac{s^\theta}{\gamma^\theta \eta^\theta} \left(\lim_{\sigma \rightarrow \infty} \left[-\frac{\gamma^\theta \eta^\theta}{s^\theta} E_\theta \left(-\frac{s^\theta t^\theta}{\gamma^\theta \eta^\theta} \right) \frac{t^\theta}{\Gamma(1+\theta)} \right]_0^\sigma \right)
\end{aligned}$$

$$\begin{aligned}
& + \frac{1}{\Gamma(1+\theta)} \frac{\gamma^\theta \eta^\theta}{s^\theta} \int_0^\infty E_\theta \left(-\frac{s^\theta t^\theta}{\gamma^\theta \eta^\theta} \right) (dt)^\theta \\
& = \frac{s^\theta}{\gamma^\theta \eta^\theta} \left(\frac{\gamma^{2\theta} \eta^{2\theta}}{s^{2\theta}} \lim_{\sigma \rightarrow \infty} \left[-E_\theta \left(-\frac{s^\theta t^\theta}{\gamma^\theta \eta^\theta} \right) \right]_0^\sigma \right) \\
& = \frac{\gamma^\theta \eta^\theta}{s^\theta}.
\end{aligned}$$

□

Theorem 7. Let $u(t) = E_\theta(at^\theta)$, $0 < \theta \leq 1$. Then, its local fractional Khalouta transform is given by

$${}^{LF}\mathbb{K}\mathbb{H}_\theta [E_\theta(at^\theta)] = \frac{s^\theta}{s^\theta - a\gamma^\theta \eta^\theta}.$$

Proof. By using (1), we attain

$$\begin{aligned}
& {}^{LF}\mathbb{K}\mathbb{H}_\theta [E_\theta(at^\theta)] \\
& = \frac{1}{\Gamma(1+\theta)} \frac{s^\theta}{\gamma^\theta \eta^\theta} \int_0^\infty E_\theta \left(-\frac{s^\theta t^\theta}{\gamma^\theta \eta^\theta} \right) E_\theta(at^\theta) (dt)^\theta \\
& = \frac{1}{\Gamma(1+\theta)} \frac{s^\theta}{\gamma^\theta \eta^\theta} \int_0^\infty E_\theta \left(-\left(\frac{s^\theta - a\gamma^\theta \eta^\theta}{\gamma^\theta \eta^\theta} \right) t^\theta \right) (dt)^\theta \\
& = \frac{s^\theta}{\gamma^\theta \eta^\theta} \left(\lim_{\sigma \rightarrow \infty} \left[-\frac{\gamma^\theta \eta^\theta}{s^\theta - a\gamma^\theta \eta^\theta} E_\theta \left(-\left(\frac{s^\theta - a\gamma^\theta \eta^\theta}{\gamma^\theta \eta^\theta} \right) t^\theta \right) \right]_0^\sigma \right) \\
& = \frac{s^\theta}{s^\theta - a\gamma^\theta \eta^\theta}.
\end{aligned}$$

□

Theorem 8. Let $u(t) = \frac{t^\theta}{\Gamma(1+\theta)} E_\theta(at^\theta)$, $0 < \theta \leq 1$. Then, its local fractional Khalouta transform is given by

$${}^{LF}\mathbb{K}\mathbb{H}_\theta \left[\frac{t^\theta}{\Gamma(1+\theta)} E_\theta(at^\theta) \right] = \frac{s^\theta \gamma^\theta \eta^\theta}{(s^\theta - a\gamma^\theta \eta^\theta)^2}.$$

Proof. By using (1) and the fractional integration by parts formula [11], we attain

$$\begin{aligned}
& {}^{LF}\mathbb{K}\mathbb{H}_\theta \left[\frac{t^\theta}{\Gamma(1+\theta)} E_\theta(at^\theta) \right] \\
& = \frac{1}{\Gamma(1+\theta)} \frac{s^\theta}{\gamma^\theta \eta^\theta} \\
& \quad \times \int_0^\infty E_\theta \left(-\frac{s^\theta t^\theta}{\gamma^\theta \eta^\theta} \right) \frac{t^\theta}{\Gamma(1+\theta)} E_\theta(at^\theta) (dt)^\theta
\end{aligned}$$

$$\begin{aligned}
&= \frac{1}{\Gamma(1+\theta)} \frac{s^\theta}{\gamma^\theta \eta^\theta} \\
&\quad \times \int_0^\infty E_\theta \left(- \left(\frac{s^\theta - a\gamma^\theta \eta^\theta}{\gamma^\theta \eta^\theta} \right) t^\theta \right) \frac{t^\theta}{\Gamma(1+\theta)} (dt)^\theta \\
&= \frac{1}{\Gamma(1+\theta)} \frac{s^\theta}{\gamma^\theta \eta^\theta} \\
&\quad \times \int_0^\infty \left(\left(\begin{array}{c} -\frac{\gamma^\theta \eta^\theta}{s^\theta - a\gamma^\theta \eta^\theta} \\ \times E_\theta \left(- \left(\frac{s^\theta - a\gamma^\theta \eta^\theta}{\gamma^\theta \eta^\theta} \right) t^\theta \right) \\ \times \frac{t^\theta}{\Gamma(1+\theta)} \end{array} \right)^{(\theta)} \right) (dt)^\theta \\
&= \frac{s^\theta}{\gamma^\theta \eta^\theta} \left(\lim_{\sigma \rightarrow \infty} \left[\begin{array}{c} -\frac{\gamma^\theta \eta^\theta}{s^\theta - a\gamma^\theta \eta^\theta} \\ \times E_\theta \left(- \left(\frac{s^\theta - a\gamma^\theta \eta^\theta}{\gamma^\theta \eta^\theta} \right) t^\theta \right) \\ \frac{t^\theta}{\Gamma(1+\theta)} \end{array} \right]^\sigma \right)_0 \\
&\quad + \frac{1}{\Gamma(1+\theta)} \frac{\gamma^\theta \eta^\theta}{s^\theta - a\gamma^\theta \eta^\theta} \\
&\quad \times \int_0^\infty E_\theta \left(- \left(\frac{s^\theta - a\gamma^\theta \eta^\theta}{\gamma^\theta \eta^\theta} \right) t^\theta \right) (dt)^\theta \\
&= \frac{s^\theta}{\gamma^\theta \eta^\theta} \left(\begin{array}{c} \frac{\gamma^{2\theta} \eta^{2\theta}}{(s^\theta - a\gamma^\theta \eta^\theta)^2} \\ \times \lim_{\sigma \rightarrow \infty} \left[-E_\theta \left(- \left(\frac{s^\theta - a\gamma^\theta \eta^\theta}{\gamma^\theta \eta^\theta} \right) t^\theta \right) \right]^\sigma \end{array} \right)_0 \\
&= \frac{s^\theta \gamma^\theta \eta^\theta}{(s^\theta - a\gamma^\theta \eta^\theta)^2}.
\end{aligned}$$

□

Theorem 9. Let $u(t) = \sin_\theta(at^\theta)$, $0 < \theta \leq 1$. Then, its local fractional Khalouta transform is given by

$${}^{LF}\mathbb{KH}_\theta [\sin_\theta(at^\theta)] = \frac{as^\theta \gamma^\theta \eta^\theta}{s^{2\theta} + a^2 \gamma^{2\theta} \eta^{2\theta}}.$$

Proof. By using (1) and the fractional integration by parts formula [11], we attain

$$\begin{aligned}
&{}^{LF}\mathbb{KH}_\theta [\sin_\theta(at^\theta)] \\
&= \frac{1}{\Gamma(1+\theta)} \frac{s^\theta}{\gamma^\theta \eta^\theta} \int_0^\infty E_\theta \left(- \frac{s^\theta t^\theta}{\gamma^\theta \eta^\theta} \right) \sin_\theta(at^\theta) (dt)^\theta \\
&= \frac{1}{\Gamma(1+\theta)} \frac{s^\theta}{\gamma^\theta \eta^\theta}
\end{aligned}$$

$$\begin{aligned}
& \times \int_0^\infty E_\theta \left(-\frac{s^\theta t^\theta}{\gamma^\theta \eta^\theta} \right) \left(\frac{E_\theta (ai^\theta t^\theta) - E_\theta (-ai^\theta t^\theta)}{2i^\theta} \right) (dt)^\theta \\
&= \frac{1}{2i^\theta} \frac{1}{\Gamma(1+\theta)} \frac{s^\theta}{\gamma^\theta \eta^\theta} \\
& \times \left(\int_0^\infty E_\theta \left(-\left(\frac{s^\theta - ai^\theta \gamma^\theta \eta^\theta}{\gamma^\theta \eta^\theta} \right) t^\theta \right) (dt)^\theta \right. \\
& \quad \left. - \int_0^\infty E_\theta \left(-\left(\frac{s^\theta + ai^\theta \gamma^\theta \eta^\theta}{\gamma^\theta \eta^\theta} \right) t^\theta \right) (dt)^\theta \right) \\
&= \frac{1}{2i^\theta} \frac{s^\theta}{\gamma^\theta \eta^\theta} \\
& \left(-\frac{\gamma^\theta \eta^\theta}{s^\theta - ai^\theta \gamma^\theta \eta^\theta} \lim_{\sigma \rightarrow \infty} \left[E_\theta \left(-\left(\frac{s^\theta - ai^\theta \gamma^\theta \eta^\theta}{\gamma^\theta \eta^\theta} \right) t^\theta \right) \right]_0^\sigma \right. \\
& \quad \left. + \frac{\gamma^\theta \eta^\theta}{s^\theta + ai^\theta \gamma^\theta \eta^\theta} \lim_{\sigma \rightarrow \infty} \left[E_\theta \left(-\left(\frac{s^\theta + ai^\theta \gamma^\theta \eta^\theta}{\gamma^\theta \eta^\theta} \right) t^\theta \right) \right]_0^\sigma \right) \\
&= \frac{1}{2i^\theta} \frac{s^\theta}{\gamma^\theta \eta^\theta} \left(\frac{\gamma^\theta \eta^\theta}{s^\theta - ai^\theta \gamma^\theta \eta^\theta} - \frac{\gamma^\theta \eta^\theta}{s^\theta + ai^\theta \gamma^\theta \eta^\theta} \right) \\
&= \frac{as^\theta \gamma^\theta \eta^\theta}{s^{2\theta} + a^2 \gamma^{2\theta} \eta^{2\theta}}.
\end{aligned}$$

□

Theorem 10. Let $u(t) = \sinh_\theta(at^\theta)$, $0 < \theta \leq 1$. Then, its local fractional Khalouta transform is given by

$${}^{LF}\mathbb{KH}_\theta [\sinh_\theta(at^\theta)] = \frac{as^\theta \gamma^\theta \eta^\theta}{s^{2\theta} - a^2 \gamma^{2\theta} \eta^{2\theta}}.$$

Proof. To prove this theorem, we follow the same previous steps, knowing that

$$\sinh_\theta(at^\theta) = \frac{E_\theta(at^\theta) - E_\theta(-at^\theta)}{2}.$$

□

Theorem 11. Let $u(t) = \cos_\theta(at^\theta)$, $0 < \theta \leq 1$. Then, its local fractional Khalouta transform is given by

$${}^{LF}\mathbb{KH}_\theta [\cos_\theta(at^\theta)] = \frac{s^{2\theta}}{s^{2\theta} + a^2 \gamma^{2\theta} \eta^{2\theta}}.$$

Proof. By using (1) and the fractional integration by parts formula [11], we attain

$$\begin{aligned}
& {}^{LF}\mathbb{KH}_\theta [\cos_\theta(at^\theta)] \\
&= \frac{1}{\Gamma(1+\theta)} \frac{s^\theta}{\gamma^\theta \eta^\theta} \int_0^\infty E_\theta \left(-\frac{s^\theta t^\theta}{\gamma^\theta \eta^\theta} \right) \cos_\theta(at^\theta) (dt)^\theta
\end{aligned}$$

$$\begin{aligned}
&= \frac{1}{\Gamma(1+\theta)} \frac{s^\theta}{\gamma^\theta \eta^\theta} \\
&\quad \times \int_0^\infty E_\theta \left(-\frac{s^\theta t^\theta}{\gamma^\theta \eta^\theta} \right) \left(\frac{E_\theta (ai^\theta t^\theta) + E_\theta (-ai^\theta t^\theta)}{2} \right) (dt)^\theta \\
&= \frac{1}{2} \frac{1}{\Gamma(1+\theta)} \frac{s^\theta}{\gamma^\theta \eta^\theta} \\
&\quad \times \left(\int_0^\infty E_\theta \left(-\left(\frac{s^\theta - ai^\theta \gamma^\theta \eta^\theta}{\gamma^\theta \eta^\theta} \right) t^\theta \right) (dt)^\theta \right. \\
&\quad \left. + \int_0^\infty E_\theta \left(-\left(\frac{s^\theta + ai^\theta \gamma^\theta \eta^\theta}{\gamma^\theta \eta^\theta} \right) t^\theta \right) (dt)^\theta \right) \\
&= \frac{1}{2} \frac{s^\theta}{\gamma^\theta \eta^\theta} \\
&\quad \times \left(-\frac{\gamma^\theta \eta^\theta}{s^\theta - ai^\theta \gamma^\theta \eta^\theta} \lim_{\sigma \rightarrow \infty} \left[E_\theta \left(-\left(\frac{s^\theta - ai^\theta \gamma^\theta \eta^\theta}{\gamma^\theta \eta^\theta} \right) t^\theta \right) \right]_0^\sigma \right. \\
&\quad \left. - \frac{\gamma^\theta \eta^\theta}{s^\theta + ai^\theta \gamma^\theta \eta^\theta} \lim_{\sigma \rightarrow \infty} \left[E_\theta \left(-\left(\frac{s^\theta + ai^\theta \gamma^\theta \eta^\theta}{\gamma^\theta \eta^\theta} \right) t^\theta \right) \right]_0^\sigma \right) \\
&= \frac{1}{2} \frac{s^\theta}{\gamma^\theta \eta^\theta} \left(\frac{\gamma^\theta \eta^\theta}{s^\theta - ai^\theta \gamma^\theta \eta^\theta} + \frac{\gamma^\theta \eta^\theta}{s^\theta + ai^\theta \gamma^\theta \eta^\theta} \right) \\
&= \frac{s^{2\theta}}{s^{2\theta} + a\gamma^{2\theta}\eta^{2\theta}}.
\end{aligned}$$

□

Theorem 12. Let $u(t) = \cosh_\theta(at^\theta)$, $0 < \theta \leq 1$. Then, its local fractional Khalouta transform is given by

$${}^{LF}\mathbb{K}\mathbb{H}_\theta [\cosh_\theta(at^\theta)] = \frac{s^{2\theta}}{s^{2\theta} - a^2\gamma^{2\theta}\eta^{2\theta}}.$$

Proof. To prove this theorem, we follow the same previous steps, knowing that

$$\cosh_\theta(at^\theta) = \frac{E_\theta(at^\theta) + E_\theta(-at^\theta)}{2}.$$

□

4 Relation with other known local fractional integral transforms

This section illustrates the relation between the local fractional Khalouta transform and other known local fractional integral transforms.

- **Local fractional Khalouta–Laplace relation:**

Let $\mathcal{K}_\theta(s, \gamma, \eta)$ be the local fractional Khalouta transform of the function $u(t)$. If $\gamma = \eta = 1$, then

$$\begin{aligned}\mathcal{K}_\theta(s, 1, 1) &= s^\theta \left(\frac{1}{\Gamma(1+\theta)} \int_0^\infty E_\theta(-s^\theta t^\theta) u(t) (dt)^\theta \right) \\ &= s^\theta \mathcal{L}_\theta(s),\end{aligned}$$

where $\mathcal{L}_\theta(s)$ denotes the local fractional Laplace transform of the function $u(t)$ [28].

- **Local fractional Khalouta–Sumudu relation:**

Let $\mathcal{K}_\theta(s, \gamma, \eta)$ be the local fractional Khalouta transform of the function $u(t)$. If $s = \gamma = 1$, then

$$\begin{aligned}\mathcal{K}_\theta(1, 1, \eta) &= \frac{1}{\Gamma(1+\theta)} \frac{1}{\eta^\theta} \int_0^\infty E_\theta\left(-\frac{t^\theta}{\eta^\theta}\right) u(t) (dt)^\theta \\ &= \mathcal{S}_\theta(\eta),\end{aligned}$$

where $\mathcal{S}_\theta(\eta)$ denotes the local fractional Sumudu transform of the function $u(t)$ [24].

- **Local fractional Khalouta–Elzaki relation:**

Let $\mathcal{K}_\theta(s, \gamma, \eta)$ be the local fractional Khalouta transform of the function $u(t)$. If $s = \eta = 1$, then

$$\begin{aligned}\mathcal{K}_\theta(1, \gamma, 1) &= \frac{1}{\gamma^{2\theta}} \left(\frac{1}{\Gamma(1+\theta)} \gamma^\theta \int_0^\infty E_\theta\left(-\frac{t^\theta}{\gamma^\theta}\right) u(t) (dt)^\theta \right) \\ &= \frac{1}{\gamma^{2\theta}} \mathcal{T}_\theta(\eta),\end{aligned}$$

where $\mathcal{T}_\theta(\eta)$ denotes the local fractional Elzaki transform of the function $u(t)$ [30].

- **Local fractional Khalouta–Aboodh relation:**

Let $\mathcal{K}_\theta(s, \gamma, \eta)$ be the local fractional Khalouta transform of the function $u(t)$. If $\gamma = \eta = 1$, then

$$\begin{aligned}\mathcal{K}_\theta(s, 1, 1) &= s^{2\theta} \left(\frac{1}{\Gamma(1+\theta)} \frac{1}{s^\theta} \int_0^\infty E_\theta(-s^\theta t^\theta) u(t) (dt)^\theta \right) \\ &= s^{2\theta} \mathcal{A}_\theta(s),\end{aligned}$$

where $\mathcal{A}_\theta(\eta)$ denotes the local fractional Aboodh transform of the function $u(t)$ [29].

• **Local fractional Khalouta–Natural relation:**

Let $\mathcal{K}_\theta(s, \gamma, \eta)$ be the local fractional Khalouta transform of the function $u(t)$. If $\gamma = 1$, then

$$\begin{aligned}\mathcal{K}_\theta(s, 1, \eta) &= s^\theta \left(\frac{1}{\Gamma(1+\theta)} \frac{1}{\eta^\theta} \int_0^\infty E_\theta \left(-\frac{s^\theta t^\theta}{\eta^\theta} \right) u(t) (dt)^\theta \right) \\ &= s^\theta \mathcal{N}_\theta(s, \eta),\end{aligned}$$

where $\mathcal{N}_\theta(s, \eta)$ denotes the local fractional natural transform of the function $u(t)$ [19].

5 Applications

This section demonstrates the applicability, efficiency, and high accuracy of the local fractional Khalouta transform method to some linear local fractional differential equations.

Example 1. Let us examine the local fractional differential equation

$$D^{(\theta)}u(t) + u(t) = \frac{t^\theta}{\Gamma(1+\theta)}, \quad 0 < \theta \leq 1, \quad (8)$$

within the initial condition

$$u(0) = -1. \quad (9)$$

Operating the local fractional Khalouta transform on (8) and utilizing Theorem 1, we attain

$${}^{LF}\mathbb{K}\mathbb{H}_\theta \left[D^{(\theta)}u(t) \right] + {}^{LF}\mathbb{K}\mathbb{H}_\theta [u(t)] = {}^{LF}\mathbb{K}\mathbb{H}_\theta \left[\frac{t^\theta}{\Gamma(1+\theta)} \right]. \quad (10)$$

Implementing Theorems 2 and 6, (10) becomes

$$\left(\frac{s}{\gamma\eta} \right)^\theta \mathcal{K}_\theta(s, \gamma, \eta) - \left(\frac{s}{\gamma\eta} \right)^\theta u(0) + \mathcal{K}_\theta(s, \gamma, \eta) = \frac{\gamma^\theta \eta^\theta}{s^\theta}. \quad (11)$$

Making use of the initial condition (9), (11) can be written as

$$\left(\frac{s}{\gamma\eta} \right)^\theta \mathcal{K}_\theta(s, \gamma, \eta) + \left(\frac{s}{\gamma\eta} \right)^\theta + \mathcal{K}_\theta(s, \gamma, \eta) = \frac{\gamma^\theta \eta^\theta}{s^\theta}.$$

Then, we have

$$\mathcal{K}_\theta(s, \gamma, \eta) = \frac{\gamma^{2\theta} \eta^{2\theta} - s^{2\theta}}{s^\theta (s^\theta + \gamma^\theta \eta^\theta)}$$

$$= \frac{\gamma^\theta \eta^\theta}{s^\theta} - 1. \quad (12)$$

Performing the inverse local fractional Khalouta transform on (12), we achieve

$$u(t) = \frac{t^\theta}{\Gamma(1+\theta)} - 1. \quad (13)$$

The result (13) indicates the exact solution to (8).

Example 2. Let us examine the local fractional differential equation

$$D^{(2\theta)}u(t) - u(t) = -\sin_\theta(t^\theta), \quad 0 < \theta \leq 1, \quad (14)$$

within the initial conditions

$$u(0) = 0, \quad D^{(\theta)}u(0) = -1. \quad (15)$$

Operating the local fractional Khalouta transform on (14) and utilizing Theorem 1, we attain

$${}^{LF}\mathbb{K}\mathbb{H}_\theta \left[D^{(2\theta)}u(t) \right] - {}^{LF}\mathbb{K}\mathbb{H}_\theta [u(t)] = {}^{LF}\mathbb{K}\mathbb{H}_\theta [-\sin_\theta(t^\theta)]. \quad (16)$$

Implementing Theorems 2 and 9, (16) becomes

$$\begin{aligned} & \left(\frac{s}{\gamma\eta} \right)^{2\theta} \mathcal{K}_\theta(s, \gamma, \eta) - \left(\frac{s}{\gamma\eta} \right)^{2\theta} u(0) - \left(\frac{s}{\gamma\eta} \right)^\theta u^{(\theta)}(0) - \mathcal{K}_\theta(s, \gamma, \eta) \\ &= -\frac{s^\theta \gamma^\theta \eta^\theta}{s^{2\theta} + \gamma^{2\theta} \eta^{2\theta}}. \end{aligned} \quad (17)$$

Making use of the initial conditions (15), (17) can be written as

$$\left(\frac{s}{\gamma\eta} \right)^{2\theta} \mathcal{K}_\theta(s, \gamma, \eta) + \left(\frac{s}{\gamma\eta} \right)^\theta - \mathcal{K}_\theta(s, \gamma, \eta) = -\frac{s^\theta \gamma^\theta \eta^\theta}{s^{2\theta} + \gamma^{2\theta} \eta^{2\theta}}.$$

Then, we have

$$\begin{aligned} \mathcal{K}_\theta(s, \gamma, \eta) &= \frac{-s^{3\theta} \gamma^\theta \eta^\theta - 2s^\theta \gamma^{3\theta} \eta^{3\theta}}{(s^{2\theta} + \gamma^{2\theta} \eta^{2\theta})(s^{2\theta} - \gamma^{2\theta} \eta^{2\theta})} \\ &= \frac{3}{4} \frac{s^\theta}{s^\theta + \gamma^\theta \eta^\theta} - \frac{3}{4} \frac{s^\theta}{s^\theta - \gamma^\theta \eta^\theta} + \frac{1}{2} \frac{s^\theta \gamma^\theta \eta^\theta}{s^{2\theta} + \gamma^{2\theta} \eta^{2\theta}}. \end{aligned} \quad (18)$$

Performing the inverse local fractional Khalouta transform on (18), we achieve

$$u(t) = \frac{3}{4} E_\theta(-t^\theta) - \frac{3}{4} E_\theta(t^\theta) + \frac{1}{2} \sin_\theta(t^\theta). \quad (19)$$

The result (19) indicates the exact solution to (14).

Example 3. Let us examine the local fractional differential equation

$$D^{(2\theta)}u(t) + D^{(\theta)}u(t) - u(t) = E_{\theta}(-2t^{\theta}), \quad 0 < \theta \leq 1, \quad (20)$$

within the initial conditions

$$u(0) = 1, \quad D^{(\theta)}u(0) = -2. \quad (21)$$

Operating the local fractional Khalouta transform on (20) and utilizing Theorem 1, we attain

$${}^{LF}\mathbb{KH}_{\theta} \left[D^{(2\theta)}u(t) \right] + {}^{LF}\mathbb{KH}_{\theta} \left[D^{(\theta)}u(t) \right] - {}^{LF}\mathbb{KH}_{\theta} [u(t)] = {}^{LF}\mathbb{KH}_{\theta} [E_{\theta}(-2t^{\theta})]. \quad (22)$$

Implementing Theorems 2 and 7, (22) becomes

$$\begin{aligned} \frac{s^{\theta}}{s^{\theta} + 2\gamma^{\theta}\eta^{\theta}} &= \left(\frac{s}{\gamma\eta} \right)^{2\theta} \mathcal{K}_{\theta}(s, \gamma, \eta) - \left(\frac{s}{\gamma\eta} \right)^{2\theta} u(0) - \left(\frac{s}{\gamma\eta} \right)^{\theta} u^{(\theta)}(0) \\ &\quad + \left(\frac{s}{\gamma\eta} \right)^{\theta} \mathcal{K}_{\theta}(s, \gamma, \eta) - \left(\frac{s}{\gamma\eta} \right)^{\theta} u(0) - \mathcal{K}_{\theta}(s, \gamma, \eta). \end{aligned} \quad (23)$$

Making use of the initial conditions (21), (23) can be written as

$$\begin{aligned} \left(\frac{s}{\gamma\eta} \right)^{2\theta} \mathcal{K}_{\theta}(s, \gamma, \eta) - \left(\frac{s}{\gamma\eta} \right)^{2\theta} + \left(\frac{s}{\gamma\eta} \right)^{\theta} + \left(\frac{s}{\gamma\eta} \right)^{\theta} \mathcal{K}_{\theta}(s, \gamma, \eta) - \mathcal{K}_{\theta}(s, \gamma, \eta) \\ = \frac{s^{\theta}}{s^{\theta} + 2\gamma^{\theta}\eta^{\theta}}. \end{aligned}$$

Then, we have

$$\begin{aligned} \mathcal{K}_{\theta}(s, \gamma, \eta) &= \frac{\gamma^{2\theta}\eta^{2\theta}}{s^{2\theta} + s^{\theta}\gamma^{\theta}\eta^{\theta} - \gamma^{2\theta}\eta^{2\theta}} \left(\frac{s^{\theta}}{s^{\theta} + 2\gamma^{\theta}\eta^{\theta}} + \frac{s^{2\theta} - s^{\theta}\gamma^{\theta}\eta^{\theta}}{\gamma^{2\theta}\eta^{2\theta}} \right) \\ &= \frac{\gamma^{2\theta}\eta^{2\theta}}{s^{2\theta} + s^{\theta}\gamma^{\theta}\eta^{\theta} - \gamma^{2\theta}\eta^{2\theta}} \left(\frac{s^{\theta}(s^{2\theta} + s^{\theta}\gamma^{\theta}\eta^{\theta} - \gamma^{2\theta}\eta^{2\theta})}{(s^{\theta} + 2\gamma^{\theta}\eta^{\theta})\gamma^{2\theta}\eta^{2\theta}} \right) \\ &= \frac{s^{\theta}}{s^{\theta} + 2\gamma^{\theta}\eta^{\theta}}. \end{aligned} \quad (24)$$

Performing the inverse local fractional Khalouta transform on (24), we achieve

$$u(t) = E_{\theta}(-2t^{\theta}). \quad (25)$$

The result (25) indicates the exact solution to (20).

Example 4. Let us examine the local fractional differential equation

$$D^{(4\theta)}u(t) - u(t) = 0, \quad 0 < \theta \leq 1, \quad (26)$$

within the initial conditions

$$u(0) = 0, \quad D^{(\theta)}u(0) = -1, \quad D^{(2\theta)}u(0) = 0, \quad D^{(3\theta)}u(0) = 0. \quad (27)$$

Operating the local fractional Khalouta transform on (26) and utilizing Theorem 1, we attain

$${}^{LF}\mathbb{K}\mathbb{H}_\theta \left[D^{(4\theta)}u(t) \right] - {}^{LF}\mathbb{K}\mathbb{H}_\theta [u(t)] = 0. \quad (28)$$

Implementing Theorem 2, (28) becomes

$$\begin{aligned} 0 = & \left(\frac{s}{\gamma\eta} \right)^{4\theta} \mathcal{K}_\theta(s, \gamma, \eta) - \left(\frac{s}{\gamma\eta} \right)^{4\theta} u(0) - \left(\frac{s}{\gamma\eta} \right)^{3\theta} u^{(\theta)}(0) \\ & - \left(\frac{s}{\gamma\eta} \right)^{2\theta} u^{(2\theta)}(0) - \left(\frac{s}{\gamma\eta} \right)^{\theta} u^{(3\theta)}(0) - \mathcal{K}_\theta(s, \gamma, \eta). \end{aligned} \quad (29)$$

Making use of the initial conditions (27), (29) can be written as

$$\left(\frac{s}{\gamma\eta} \right)^{4\theta} \mathcal{K}_\theta(s, \gamma, \eta) + \left(\frac{s}{\gamma\eta} \right)^{3\theta} - \mathcal{K}_\theta(s, \gamma, \eta) = 0.$$

Then, we have

$$\begin{aligned} \mathcal{K}_\theta(s, \gamma, \eta) &= -\frac{s^{3\theta}\gamma^\theta\eta^\theta}{s^{4\theta} - \gamma^{4\theta}\eta^{4\theta}} \\ &= \frac{1}{4} \frac{s^\theta}{s^\theta + \gamma^\theta\eta^\theta} - \frac{1}{4} \frac{s^\theta}{s^\theta - \gamma^\theta\eta^\theta} - \frac{1}{2} \frac{s^\theta\gamma^\theta\eta^\theta}{s^{2\theta} + \gamma^{2\theta}\eta^{2\theta}}. \end{aligned} \quad (30)$$

Performing the inverse local fractional Khalouta transform on (30), we achieve

$$u(t) = \frac{1}{4} E_\theta(-t^\theta) - \frac{1}{4} E_\theta(t^\theta) - \frac{1}{2} \sin_\theta(t^\theta). \quad (31)$$

The result (31) indicates the exact solution to (26).

Example 5. Let us examine the local fractional Volterra integro-differential equation

$$D^{(\theta)}u(t) = 1 + \frac{1}{\Gamma(1+\theta)} \int_0^t u(\tau) (d\tau)^\theta, \quad 0 < \theta \leq 1, \quad (32)$$

within the initial condition

$$u(0) = 0. \quad (33)$$

Operating the local fractional Khalouta transform on (32) and utilizing Theorem 1, we attain

$${}^{LF}\mathbb{KH}_\theta \left[D^{(\theta)} u(t) \right] = {}^{LF}\mathbb{KH}_\theta [1] + {}^{LF}\mathbb{KH}_\theta \left[\frac{1}{\Gamma(1+\theta)} \int_0^t u(\tau) (d\tau)^\theta \right]. \quad (34)$$

Then

$${}^{LF}\mathbb{KH}_\theta \left[D^{(\theta)} u(t) \right] = {}^{LF}\mathbb{KH}_\theta [1] + {}^{LF}\mathbb{KH}_\theta [(u(t) * 1)_\theta]. \quad (35)$$

Implementing Theorems 2, 4, and 5, we arrive at

$$\left(\frac{s}{\gamma\eta} \right)^\theta \mathcal{K}_\theta(s, \gamma, \eta) - \left(\frac{s}{\gamma\eta} \right)^\theta u(0) = 1 + \left(\frac{\gamma\eta}{s} \right)^\theta \mathcal{K}_\theta(s, \gamma, \eta). \quad (36)$$

Making use of the initial condition (33), (36) can be written as

$$\left(\frac{s}{\gamma\eta} \right)^\theta \mathcal{K}_\theta(s, \gamma, \eta) = 1 + \left(\frac{\gamma\eta}{s} \right)^\theta \mathcal{K}_\theta(s, \gamma, \eta).$$

Then, we have

$$\mathcal{K}_\theta(s, \gamma, \eta) = \frac{s^\theta \gamma^\theta \eta^\theta}{s^{2\theta} - \gamma^{2\theta} \eta^{2\theta}}. \quad (37)$$

Performing the inverse local fractional Khalouta transform on (37), we achieve

$$u(t) = \sinh_\theta(t^\theta). \quad (38)$$

The result (38) indicates the exact solution to (32).

Example 6. Let us examine the local fractional Volterra integro-differential equation

$$D^{(2\theta)} u(t) = 1 + \frac{1}{\Gamma(1+\theta)} \int_0^t \frac{(t-\tau)^\theta}{\Gamma(1+\theta)} u(\tau) (d\tau)^\theta, \quad 0 < \theta \leq 1, \quad (39)$$

within the initial conditions

$$u(0) = 1, \quad D^{(\theta)} u(0) = 0. \quad (40)$$

Operating the local fractional Khalouta transform on (39) and utilizing Theorem 1, we attain

$${}^{LF}\mathbb{KH}_\theta \left[D^{(2\theta)} u(t) \right] = {}^{LF}\mathbb{KH}_\theta [1] + {}^{LF}\mathbb{KH}_\theta \left[\frac{1}{\Gamma(1+\theta)} \int_0^t \frac{(t-\tau)^\theta}{\Gamma(1+\theta)} u(\tau) (d\tau)^\theta \right].$$

Then

$${}^{LF}\mathbb{KH}_\theta \left[D^{(2\theta)} u(t) \right] = {}^{LF}\mathbb{KH}_\theta [1] + {}^{LF}\mathbb{KH}_\theta \left[\left(\frac{t^\theta}{\Gamma(1+\theta)} * u(t) \right)_\theta \right]. \quad (41)$$

Using Theorems 2, 4, 5 and 6, (41) becomes

$$\begin{aligned} & \left(\frac{s}{\gamma\eta} \right)^{2\theta} \mathcal{K}_\theta(s, \gamma, \eta) - \left(\frac{s}{\gamma\eta} \right)^{2\theta} u(0) - \left(\frac{s}{\gamma\eta} \right)^\theta u^{(\theta)}(0) \\ &= 1 + \left(\frac{\gamma\eta}{s} \right)^\theta \left(\frac{\gamma\eta}{s} \right)^\theta \mathcal{K}_\theta(s, \gamma, \eta). \end{aligned} \quad (42)$$

Making use of the initial conditions (40), (42) can be written as

$$\left(\frac{s}{\gamma\eta} \right)^{2\theta} \mathcal{K}_\theta(s, \gamma, \eta) - \left(\frac{s}{\gamma\eta} \right)^{2\theta} = 1 + \left(\frac{\gamma\eta}{s} \right)^{2\theta} \mathcal{K}_\theta(s, \gamma, \eta).$$

Then, we have

$$\begin{aligned} \mathcal{K}_\theta(s, \gamma, \eta) &= \left(\frac{s^{2\theta} \gamma^{2\theta} \eta^{2\theta}}{s^{4\theta} - \gamma^{4\theta} \eta^{4\theta}} \right) \left(\frac{\gamma^{2\theta} \eta^{2\theta} + s^{2\theta}}{\gamma^{2\theta} \eta^{2\theta}} \right) \\ &= \frac{s^{2\theta}}{s^{2\theta} - \gamma^{2\theta} \eta^{2\theta}}. \end{aligned} \quad (43)$$

Performing the inverse local fractional Khalouta transform on (43), we achieve

$$u(t) = \cosh_\theta(t^\theta). \quad (44)$$

The result (44) indicates the exact solution to (39).

6 Conclusions

In this paper, we proposed the local fractional Khalouta transform method, based on the coupling of two powerful concepts: The Khalouta transform and the local fractional derivative. Several illustrative examples of local linear fractional differential equations have been provided in order to demonstrate the advantages of the proposed method. The results obtained in the examples presented have shown that this method is capable of solving any linear local integral equation.

These results open up avenues for further research and applications in related fields of science and mathematics. Future research directions could include:

- Proposing novel numerical methods that combine the local fractional Khalouta transform method with semi-analytical methods, such as the Adomian decomposition method, the ho-

motopy perturbation method, and the variational iteration method, to solve nonlinear local fractional differential equations.

- Investigating the stability and convergence analysis of proposed methods.
- Extending the applicability of these methods to more complex problems, such as nonlinear local fractional partial differential equations and nonlinear local fractional integro-differential equations.

Acknowledgements

The author thanks the anonymous reviewers and the editor for their constructive comments.

Funding

- This research received no external funding.

Conflicts of Interest

- The author declares no conflicts of interest.

References

- [1] Abdeljawad, T. *On conformable fractional calculus*, J. Comput. Appl. Math. 279 (2015), 57–66.
- [2] Atangana, A. and Koca, I. *Chaos in a simple nonlinear system with Atangana-Baleanu derivatives with fractional order*, Chaos, Soliton. Fract. 89 (2016) 447–454.
- [3] Baleanu, D. and Jassim, H.K. *A modification fractional homotopy perturbation method for solving Helmholtz and coupled Helmholtz equations on Cantor sets*, Fractal Fract. 3(30) (2019), 1–8.
- [4] Baleanu, D. and Jassim, H.K. *Approximate solutions of the damped wave equation and dissipative wave equation in fractal strings*, Fractal Fract. 3(26) (2019), 1–12.

- [5] Baleanu, D., Jassim, H.K. and Al Qurashi, M. *Solving Helmholtz equation with local fractional derivative operators*, Fractal Fract. 3(43) (2019), 1–13.
- [6] Baleanu, D., Jassim, H.K. and Al Qurashic, M. *Approximate analytical solutions of Goursat problem within local fractional operators*, J. Nonlinear Sci. Appl. 9 (2016), 4829–4837.
- [7] Hilfer, R. *Applications of fractional calculus in physics*, World Scientific, River Edge, New Jersey, 2000.
- [8] Jafari, H. and Jassim, H.K. *Application of the Local Fractional Variational Iteration Method to Solve System of Coupled Partial Differential Equations Involving Local Fractional Operator*, Appl. Math. Inf. Sci. Lett. 5(2) (2017), 45–50.
- [9] Jafari, H., Jassim, H.K. and Baleanu, D. *On the existence and uniqueness of solutions for local differential equations*, Entropy. 18(420) (2016), 1–9.
- [10] Jafari, H., Jassim, H.K., Tchier, F. and Baleanu, D. *On the approximate solutions of local fractional differential equations with local fractional operator*, Entropy. 18 (2016), 1–12.
- [11] Jumarie, G. *Table of some basic fractional calculus formulae derived from a modified Riemann-Liouville derivative for non-differentiable functions*, Appl. Math. Lett. 22(3) (2009), 378–385.
- [12] Khalouta, A. *A new exponential type kernel integral transform: Khalouta transform and its applications*, Math. Montisnigri. 57 (2023), 5–23.
- [13] Khalouta, A. *New results of Khalouta transform method for non-constant coefficients ordinary differential equations*, J. Interdiscip. Math. 28(7) (2025), 2555–2566.
- [14] Kilbas, A.A. *Hadamard-type fractional calculus*, J. Korean Math. Soc. 3(6) (2001), 1191–1204.
- [15] Kilbas, A.A, Srivastava, H.M. and Trujillo, J.J. *Theory and application of fractional differential equations*, Elsevier, Amsterdam, 2006.
- [16] Leibniz, G.W. *Letter from Hanover, Germany to G.F.A. L'Hospital, September 30, 1695*, Leibniz Mathematische Schriften. Olms-Verlag, Hildesheim, Germany, 1962. 301–302, First published in 1849.

- [17] Liu, J-G., Yang, X-J. Feng, Y-Y. and Cui, P. *On the $(N + 1)$ -dimensional local fractional reduced differential transform method and its applications*, Math. Methods Appl. Sci. 43(5) (2020), 8856-8866.
- [18] Mainardi, F. *Fractional Calculus and Waves in Linear Viscoelasticity: An Introduction to Mathematical Models*, Imperial College Press, London, 2010.
- [19] Maitama, S. *Local fractional natural homotopy perturbation method for solving partial differential equations with local fractional derivative*, Progr. Fract. Differ. Appl. 4(3) (2018), 219–228.
- [20] Maitama, S. and Zhao, W. *Local fractional homotopy analysis method for solving non-differentiable problems on Cantor sets*, Adv. Differ. Equ. 2019(127) (2019), 1–22.
- [21] Miller, K.S. and Ross, B. *An introduction to the fractional calculus and fractional differential equations*, Wiley, New York, 1993.
- [22] Odziejewicz, T. *Generalized fractional isoperimetric problem of several variables*, Discrete Contin. Dyn. Syst. Ser. B. 19(8) (2014), 2617–2629.
- [23] Podlubny, I. *Fractional Differential Equations*, Academic Press, New York, 1999.
- [24] Srivastava, H.M., Golmankhaneh, A.K. Baleanu, D. and Yang, X-J. *Local fractional Sumudu transform with application to IVPs on Cantor sets*, Abstr. Appl. Anal. 2014 (2014), 1–7.
- [25] Yang, X.J. *Local Fractional Functional Analysis and Its Applications*, Asian Academic: Hong Kong, China, 2011.
- [26] Yang, X.J., *Advanced local fractional calculus and its applications*, World Science: New York, NY, USA, 2012.
- [27] Yang, Y-J. and Wang, S-Q. *A local fractional homotopy perturbation method for solving the local fractional Korteweg-de Vries equations with non-homogeneous term*, Therm. Sci. 23(00) (2019), 216–216.
- [28] Zhao, C.G., Yang, A.M., Jafari, H. and Haghbin, A. *The Yang-Laplace transform for solving the IVPs with local fractional derivative*, Abstr. Appl. Anal. 2014 (2014), 1–5.
- [29] Ziane, Dj., Belgacema, R. and Bokhari, A. *Local fractional Aboodh transform and its applications to solve linear local fractional differential equations*, Adv. Theory Nonlinear Anal. Appl. 6(2) (2022), 217–228.

- [30] Ziane, Dj. and Hamdi Cherif, M. *Local fractional Elzaki transform and its application to local fractional differential equations*, J. New Results Sci. 10(3) (2021), 19–33.



A nonmonotone line search method for solving constrained multiobjective optimization problems

M. Rashidi and E. Khorram*

Abstract

In this paper, we propose a globally convergent sequential quadratic programming method for solving constrained multiobjective optimization problems with inequality constraints. At each iteration, a feasible descent direction is computed by solving an auxiliary quadratic programming subproblem constructed from linear approximations of the objective and constraint functions. Constraint violations are addressed using a nonsmooth exact penalty function. Moreover, the algorithm incorporates a nonmonotone max-type line-search strategy, which improves practical performance by allowing temporary increases in the penalty function value and thereby promoting more effective exploration of the search space. Under mild regularity assumptions, we prove that the sequence generated by the method converges to a weakly or strongly critical point. Numerical experiments on standard benchmark problems demonstrate the effectiveness and robustness of the proposed approach, both in terms of computational performance criteria and the quality of the approximated Pareto front.

AMS subject classifications (2020): 90C30, 49M05, 97N40 90B50.

*Corresponding author

Received 30 July 2025; revised 27 December 2025; accepted 30 December 2025

Mehri Rashidi

Faculty of Mathematics and Computer Sciences, Amirkabir University of Technology, Tehran, Iran. e-mail: mrashidi@aut.ac.ir

Esmale Khorram

Faculty of Mathematics and Computer Sciences, Amirkabir University of Technology, Tehran, Iran. e-mail: eskh@aut.ac.ir

How to cite this article

Rashidi, M. and Khorram, E., A nonmonotone line search method for solving constrained multi-objective optimization problems. *Iran. J. Numer. Anal. Optim.*, 2026; 16(2): 563–589. <https://doi.org/10.22067/ijnao.2025.94685.1691>

Keywords: SQP method, Nonmonotone line search, Penalty method, Mangasarian–Fromovitz constraint qualification, Critical point.

1 Introduction

Multiobjective optimization is essential for addressing complex problems in areas such as game theory, economics, and engineering, where several conflicting objectives must be considered simultaneously [28]. The goal is to identify solutions that provide the best possible compromises among multiple criteria over a given decision space. Because optimizing all objectives simultaneously is generally impossible due to inherent trade-offs, the classical notion of optimality is replaced by efficiency. A point is said to be *efficient* (Pareto optimal) if no objective can be improved without deteriorating at least one other objective, and the collection of such points forms the *nondominated* (Pareto) set. Accurate approximation of this set, as well as its image in the objective space known as the Pareto front, is crucial for understanding the trade-off structure and has motivated extensive research on algorithms for approximating nondominated fronts in multiobjective optimization problems (MOPs) [16, 35].

Scalarization is a classical approach for solving MOPs, where a vector objective is reduced to a single-objective formulation, typically through weighted sums, ε -constraint schemes, or related constructions [19, 20, 32]. Although conceptually appealing, scalarization often requires the choice of user-defined parameters that reflect preferences or trade-off weights, which may be difficult to specify reliably and can significantly affect the obtained solutions. Moreover, depending on the scalarization model and the problem structure, the resulting subproblems may exhibit undesirable behavior such as ill-conditioning or even unboundedness [21].

These limitations have motivated the development of preference-free, direction-based iterative methods that work directly in the multiobjective setting. Starting from a random or feasible point, such methods typically compute a common descent direction by solving an auxiliary subproblem, and then employ a line-search procedure to select a suitable step size, aiming to improve all objectives in the Pareto sense [17, 18, 22]. Within this framework, a variety of algorithmic ideas have been explored, including Newton and quasi-Newton type strategies [21, 30, 40], sequential quadratic programming (SQP)-type methods for constrained MOPs [2, 23, 24, 37], and projection-based schemes on convex feasible sets [6, 43]. In parallel, nonmonotone globalization mechanisms have been increasingly used to enhance robustness in single objective settings, building on nonmonotone line-search principles [11, 27, 41] and extending them to nonconvex and unconstrained multiobjective algorithms [25, 34].

Recent developments in nonmonotone optimization, particularly within trust-region and line-search frameworks, have led to substantial improvements for both constrained and unconstrained problems. In contrast to classical monotone schemes that enforce a decrease in the merit function at every iteration, nonmonotone strategies permit occasional increases, which can improve robustness and promote more effective exploration of the optimization landscape, especially in nonconvex settings [27, 41]. In the context of nonmonotone methods for unconstrained MOPs, several recent contributions have advanced both theory and algorithmic design. Mahdavi-Amiri and Salehi Sadaghiani [30] proposed a nonmonotone quasi-Newton scheme with superlinear convergence, while Ghalavand, Khorram, and Morovati [25] developed adaptive nonmonotone trust-region algorithms to enhance robustness. More recently, Upadhayay et al. [39] introduced a nonmonotone conditional gradient framework for MOPs, and Upadhayay, Ghosh, and Kumar [40] investigated nonmonotone Wolfe-type quasi-Newton methods, providing effective globalization strategies in the unconstrained setting. Despite these advances, constrained MOPs remain particularly challenging, and comparatively few studies have focused on algorithms with strong theoretical guarantees and reliable practical performance in this setting. Notable recent contributions include conditional gradient type methods [8], accelerated proximal gradient schemes [38], away-step Frank–Wolfe strategies for constrained MOPs [26], SQP-type approaches combined with gradient sampling for nonsmooth constrained MOPs [37], and proximal gradient methods designed for convex MOPs [42].

For nonmonotone methods in constrained MOPs, recent contributions include Zhao and Yao [43], who proposed a nonmonotone projected gradient framework; Carrizo, Fazzio, and Schuverdt [6], who developed a nonmonotone projected gradient method on convex sets; Pinheiro and Grapiglia [34], who introduced a universal nonmonotone line-search scheme for nonconvex MOPs with convex constraints; and Chen et al. [7], who presented a nonmonotone proximal gradient algorithm for nonsmooth MOPs, with extensions to robust multiobjective optimization. While nonmonotone line search strategies for unconstrained optimization are well established, their extension to constrained MOPs remains comparatively limited. The main difficulty is the presence of constraints, which typically requires merit or penalty functions, and designing a nonmonotone acceptance criterion that preserves both feasibility progress and convergence guarantees is technically delicate. In particular, nonmonotone line search applied directly to an exact penalty function is still rare. To the best of our knowledge, the most recent work explicitly pursuing this direction is Rashidi, Khorram, and Soleimani-Damaneh [36]. This approach integrates an exact penalty formulation with an SQP-type framework, a nonmonotone line-search mechanism, and rapid infeasibility detection. Moreover, an average-type nonmonotone criterion is employed, and up to two inequality-constrained subproblems may be solved at each itera-

tion. Penalty updates are controlled through trust-region ideas, yielding a hybrid strategy that combines features of both trust-region and line-search globalization.

Ansary and Panda [2] proposed an SQP method for inequality-constrained MOPs. Their scheme computes a descent direction from an auxiliary quadratic subproblem based on linear approximations of the objective and constraint functions, and it handles constraint violations through a nonsmooth exact penalty function.

Motivated by the SQP framework of Ansary and Panda [2] and by the use of nonmonotone globalization mechanisms for penalty-based constrained MOPs in [36], we develop a new globally convergent SQP method for inequality-constrained MOPs. The proposed algorithm preserves the main SQP structure of [2] but introduces a different penalty-function formulation together with a distinct strategy for selecting and updating the penalty parameter. These modifications require a technical justification, both to establish global convergence and to justify the observed numerical behavior. In addition, we incorporate a nonmonotone max-type line search rule, which allows controlled temporary increases of the penalty function and can improve robustness by preventing overly conservative steps. Under mild regularity assumptions, we prove that the sequence generated by the method converges to a weakly or strongly critical point. Numerical results on standard benchmarks confirm the effectiveness and robustness of the proposed approach.

The remainder of the paper is organized as follows. Section 2 introduces the basic definitions and preliminaries. Section 3 presents the proposed method. The convergence analysis is developed in Section 4. Section 5 reports the numerical implementation details and computational results. Finally, Section 6 concludes the paper.

2 Preliminaries

We consider the following constrained MOP:

$$\begin{aligned} & \min f(x) \\ & \text{s.t. } g(x) \leq 0, \end{aligned} \tag{1}$$

where $f(x) = (f_1(x), f_2(x), \dots, f_m(x))^T$ and $g(x) = (g_1(x), g_2(x), \dots, g_r(x))^T$. Here, each $f_i : \mathbb{R}^n \rightarrow \mathbb{R}$ for $i = 1, \dots, m$, and each $g_j : \mathbb{R}^n \rightarrow \mathbb{R}$ for $j = 1, \dots, r$. The functions $f(\cdot)$ and $g(\cdot)$ are continuously differentiable over the feasible set $X \subseteq \mathbb{R}^n$, which is defined as:

$$X = \{x \in \mathbb{R}^n \mid g_j(x) \leq 0, \text{ for } j = 1, 2, \dots, r\}.$$

Thus, the optimization problem can be expressed equivalently as

$$\begin{aligned} & \min f(x) \\ & \text{subject to } x \in X. \end{aligned}$$

In MOP, the objective space is a partial order set, while in single-objective optimization, it is a total order set. Let $\mathbb{R}_{\geq}^m$ and $\mathbb{R}_{>}^m$ be defined as

$$\mathbb{R}_{\geq}^m = \{\mathbf{x} \in \mathbb{R}^m : x_k \geq 0, \text{ for all } k\}, \quad \mathbb{R}_{>}^m = \{\mathbf{x} \in \mathbb{R}^m : x_k > 0, \text{ for all } k\}.$$

These sets describe relationships between vector-valued functions $f(x)$ and $f(y)$ based on component-wise comparisons:

$$f(x) \leq f(y) \Leftrightarrow f(y) - f(x) \in \mathbb{R}_{\geq}^m, \quad f(x) < f(y) \Leftrightarrow f(y) - f(x) \in \mathbb{R}_{>}^m.$$

A point $x \in X$ is called a weakly efficient point of problem (1) if there is no $y \in X$ such that $f(y) < f(x)$, in which case $f(x)$ is referred to as a weakly nondominated point [19]. A point $x \in X$ is called an efficient point of problem (1) if there is no $y \in X$ such that $f(y) \leq f(x)$ and $f(y) \neq f(x)$, and $f(x)$ is then referred to as a nondominated point [19]. Furthermore, $\bar{x}$ is a *local weakly efficient* solution of (1) if there exists $\varepsilon > 0$ such that there is no $x \in X \cap B(\bar{x}, \varepsilon)$ satisfying $f(x) < f(\bar{x})$. In this case, $f(\bar{x})$ is called a *local weakly nondominated* point.

Penalty methods convert the constrained MOP (1) into an unconstrained optimization problem by incorporating constraint violations into a merit (penalty) function. In this paper, we adopt a penalization framework and minimize a shifted exact penalty function [5, 36]. The proposed penalty function employs a violation measure of the constraint that is structurally similar to that used by Ansary and Panda [2]. Moreover,

$$\psi_{\rho}^i(x) := \rho \left(f_i(x) - \underline{f}_i \right) + v(x), \quad i = 1, \dots, m, \quad (2)$$

in which $\rho \geq 0$ is the penalty parameter, and

$$v(x) = \max\{0, g_j(x) \mid j = 1, \dots, r\},$$

is the violation measure of the constraint. Here, $v(x) = 0$ characterizes feasibility. Moreover, for each i , let $\underline{f}_i := \inf\{f_i(x) : x \in X\}$, which is finite under the boundedness assumption stated in subsection 4.1. Equivalently, $\underline{f}_i$ may be viewed as the infimum of f_i over the smallest convex set containing the iterates $\{x_k\}$ generated by the algorithm. The bound $\underline{f}_i$ is introduced only for theoretical purposes, namely to ensure that the line search procedure is well defined. Its existence

follows from assumptions (A1)–(A2). In practice, we do not require $\underline{f}_i$ (or any approximation of it), and the proposed method is not sensitive to its value [36].

Define the *active set*

$$\mathcal{A}(x) = \{j : g_j(x) = v(x)\},$$

collect all constraints attaining the worst violation. Then the true directional derivative of v in any $d \in \mathbb{R}^n$ is

$$v'(x; d) = \max_{j \in \mathcal{A}(x)} \nabla g_j(x)^T d,$$

which is generally discontinuous as $\mathcal{A}(x)$ changes.

That is, although each g_j is continuously differentiable, the directional derivative $v'(x; d)$ is, in general, *discontinuous with respect to the base point x* . This discontinuity is caused by changes in the active set $\mathcal{A}(x)$ when the maximizer in

$$v(x) = \max\{0, g_j(x) \mid j = 1, \dots, m\}$$

switches from one constraint to another. For fixed x , however, $v'(x; d)$ is continuous (indeed, piecewise linear) as a function of the direction d . For more clarification, we show this behavior in the following example:

Example 1. Let

$$g_1(x) = x, \quad g_2(x) = -x, \quad v(x) = \max\{0, g_1(x), g_2(x)\} = |x|.$$

Then

$$\mathcal{A}(x) = \begin{cases} \{1\}, & x > 0, \\ \{2\}, & x < 0, \\ \{1, 2\}, & x = 0, \end{cases}$$

which changes at $x = 0$. Take the direction $d = 1$. For $x > 0$ we have $v'(x; 1) = \nabla g_1(x)^T 1 = 1$; for $x < 0$, $v'(x; 1) = \nabla g_2(x)^T 1 = -1$; and at $x = 0$,

$$v'(0; 1) = \max\{\nabla g_1(0)^T 1, \nabla g_2(0)^T 1\} = \max\{1, -1\} = 1.$$

Thus,

$$v'(x; 1) = \begin{cases} 1, & x \geq 0, \\ -1, & x < 0, \end{cases}$$

which is clearly discontinuous at $x = 0$. The discontinuity is caused by a switch of the maximizer from g_2 to g_1 , that is, by a change in the active set $\mathcal{A}(x)$. Thus, we employ the continuous approximation proposed in [2, 4]:

$$v^*(x; d) = \max \left\{ g_j(x) + \nabla g_j(x)^T d, 0 \mid j \in \mathcal{A}(x) \right\} - v(x).$$

Therefore, we approximate the directional derivative of ψ_ρ^i by the following continuous term:

$$\chi_\rho^i(x; d) = \rho \nabla f_i(x)^T d + v^*(x; d), \quad i = 1, \dots, m. \quad (3)$$

The following theorem provides necessary optimality conditions for a (weakly) efficient point of problem (1).

Theorem 1 ([32]). Suppose that f_i , $i = 1, \dots, m$, and g_j , $j = 1, \dots, r$, are continuously differentiable at $x \in X$. If x is a weakly efficient point of problem (1), then there exist multipliers $\lambda \in \mathbb{R}_{\geq 0}^m$ and $\theta \in \mathbb{R}_{\geq 0}^r$ such that $(\lambda, \theta) \neq (0, 0)$ and

$$\begin{cases} \sum_{i=1}^m \lambda_i \nabla f_i(x) + \sum_{j=1}^r \theta_j \nabla g_j(x) = 0, \\ \theta_j g_j(x) = 0, \quad j = 1, \dots, r. \end{cases} \quad (4)$$

The Fritz–John conditions in Theorem 1 are necessary for weak efficiency. In contrast, the Karush–Kuhn–Tucker conditions are necessary under an appropriate constraint qualification [32]. Therefore, in our convergence analysis, we assume that the Mangasarian–Fromovitz constraint qualification (MFCQ) holds at the limit points under consideration.

Definition 1 (Mangasarian–Fromovitz constraint qualification (MFCQ)). [31] The MFCQ is said to hold at a point $x \in \mathbb{R}^n$ if there exists a vector $z \in \mathbb{R}^n$ such that

$$\nabla g_j(x)^T z < 0, \quad \text{for all } j \in \mathcal{A}(x).$$

Definition 2. [2] A feasible point $x \in X$ of the problem (1) is said to be

1. a *strongly critical point* if there exist multipliers $\lambda \in \mathbb{R}_+^m \setminus \{0\}$ and $\theta \in \mathbb{R}_+^r$ satisfying the conditions (4).
2. a *weakly critical point* if there exists an infeasible sequence $\{x_k\}$ with $x_k \rightarrow x$ such that

$$\lim_{k \rightarrow \infty} \frac{\max_{d \in D(x_k)} \max_{j=1, \dots, r} \{g_j(x_k) + \nabla g_j(x_k)^T d, 0\}}{v(x_k)} = 1,$$

where $D(x_k) = \{d \in \mathbb{R}^n : \nabla f_i(x_k)^T d \leq 0, i = 1, \dots, m\}$.

Remark 1. The above definition of a *weakly critical point* is proposed because the penalty-SQP scheme may generate *infeasible* iterates, so we need a stationarity concept that remains meaningful outside the feasible set. Geometrically, the condition says that there is no direction d that improves (or at least does not worsen) all objectives while simultaneously decreasing violation measure of the constraint $v(x)$ according to the linearized constraints. Hence, at a weakly critical point the algorithm cannot make simultaneous progress toward feasibility and objective decrease, and such points naturally describe possible limit behavior of infeasible iterates.

More precisely, at an iterate x_k the set $D(x_k) = \{d \in \mathbb{R}^n : \nabla f_i(x_k)^T d \leq 0, i = 1, \dots, m\}$ is a cone of *nonascending* directions for all objectives. For a given direction d , the quantity $\max_j \max\{g_j(x_k) + \nabla g_j(x_k)^T d, 0\}$ is the worst predicted violation under the linearized constraints. Therefore, the numerator $\max_{d \in D(x_k)} \max_j \max\{g_j(x_k) + \nabla g_j(x_k)^T d, 0\}$ represents the best achievable (i.e., smallest possible) worst-case linearized infeasibility among directions that do not worsen the objectives, while the denominator $v(x_k)$ is the current infeasibility. The limit being equal to 1 means that, asymptotically, even the best direction in $D(x_k)$ cannot yield a meaningful relative decrease in infeasibility: The predicted violation remains essentially the same as the current one.

3 Description of the algorithm

This section presents our nonmonotone multiobjective method in the SQP framework (NMOSQP). Our method is developed within the SQP framework of [2]; however, we introduce two essential modifications : (i) a new formulation of the (shifted) exact penalty function together with a different strategy for selecting and updating the penalty parameter, and (ii) a nonmonotone max-type line search rule applied to the penalty function. These changes require very additional technical justification, both in the global convergence analysis and in the numerical investigation.

More specifically, as in [2], NMOSQP computes a common improving direction by solving an auxiliary quadratic subproblem based on linear models of the objective and constraint functions, and then determines the step length by a line search procedure. The first main distinction from [2] lies in the treatment of the penalty parameter, while [2] employs a particular penalty-function structure and update rule, we adopt a different penalty-parameter update mechanism tailored to our nonmonotone line search. The second main distinction concerns globalization. Unlike [2], which uses a monotone line search, NMOSQP employs a nonmonotone max-type line search on the penalty function.

In [36], a descent direction is computed by minimizing a quadratic model of an exact penalty function. In each iteration, the method may require solving up to two inequality-constrained subproblems, while updates of the penalty parameter are governed by steering rules. In [37], an SQP-type method for constrained MOPs with nonconvex, nonsmooth, and locally Lipschitz objectives and constraints is proposed. The approach uses gradient sampling to handle nonsmoothness and employs an exact penalty merit function; its subproblem can be viewed as a nonsmooth counterpart of the one in [36].

We keep all linearized inequality constraints in subproblem (5), rather than restricting to the currently active set as in [33]. In nonconvex constrained MOPs the active set may change abruptly from one iterate to the next, and constraints that are inactive at x_k can become active after a step. Including all constraints improves robustness of the computed direction and helps prevent steps that move toward nearly active constraints and trigger excessive backtracking in the line search. Hence

$$\begin{aligned} \min_{d,z} \quad & z + \frac{1}{2}\|d\|^2, \\ \text{subject to:} \quad & (\nabla f_i(x))^T d \leq z, \quad \text{for all } i = 1, \dots, m, \\ & g_j(x) + (\nabla g_j(x))^T d \leq z, \quad \text{for all } j = 1, \dots, r, \\ & (d, z) \in \mathbb{R}^n \times \mathbb{R}. \end{aligned} \tag{5}$$

In this subproblem, z is as an upper bound for all linearized terms, and both z and d are variables. The improving direction is defined as the d -component of a minimizer of subproblem (5). The smooth quadratic subproblem (5) plays an crucial role in the analysis and implementation of the NMOSQP method. The satisfaction of MFCQ for the problem (1) ensures that the subproblem (5) also meets the MFCQ condition [2].

Inspired by the approach proposed in [2], the penalty parameter ρ_k is dynamically updated at each iteration to ensure that the direction d_k remains a descent direction for all penalized functions $\psi_{\rho_k}^i(x_k)$, $i = 1, \dots, m$. If d_k is already a descent direction for all $\psi_{\rho_k}^i$, more precisely, if $\chi_{\rho_k}^i(x_k; d_k) < 0$, $i = 1, \dots, m$, then the penalty parameter remains unchanged, that is, $\rho_{k+1} = \rho_k$. Otherwise, it is updated as

$$\rho_{k+1} = \min \left\{ \theta_\rho \rho_k, \min_{i=1, \dots, m} \frac{v^*(x_k; d_k) + \frac{1}{2} d_k^T d_k}{-\nabla f_i(x_k)^T d_k} \right\}, \tag{6}$$

where $v^*(x_k; d_k)$ is the smooth approximation of the directional derivative of the violation function $v(x)$, and the ratio ensures that d_k becomes a descent direction for each penalized objective $\psi_{\rho_k}^i$.

3.1 Nonmonotone line search

After obtaining the descent direction $d(x)$, the NMOSQP algorithm employs a backtracking line search to compute a suitable step length $\theta > 0$ along this direction. The monotone line search is guided by the Armijo rule, which ensures that the step length satisfies the sufficient decrease condition:

$$\psi_{\rho_{k+1}}^i(x_k + \theta_k d_k) \leq \psi_{\rho_{k+1}}^i(x_k) + \theta_k \gamma \chi_{\rho_{k+1}}^i(x_k; d_k), \quad (7)$$

for $i = 1, \dots, m$, where $x_k \in X$ is not a critical point of problem (1), and $\gamma \in (0, 1)$ is a constant. This condition guarantees sufficient improvement in the objective functions, thereby promoting progress in the optimization process.

Nonmonotone line search strategies improve optimization performance by allowing temporary increases in the objective function. This adaptability proves especially effective in escaping narrow or curved valleys and in navigating complex, nonconvex regions of the search space [1, 11]. Such methods typically offer faster convergence and enhanced robustness, making them particularly suitable for solving large-scale and challenging optimization problems. In this paper, we propose a novel nonmonotone line search method for solving constrained MOPs. Our approach extends the classical nonmonotone rule of Grippo, Lampariello, and Lucidi [27], which generalizes the traditional Armijo rule by permitting occasional increases in the function value. This flexibility helps the algorithm escape narrow valleys or avoid excessively small steps in challenging nonlinear problems. Therefore,

$$\psi_{\rho_{k+1}}^i(x_k + \theta_k d_k) \leq \psi_{\rho_k, \mathfrak{r}}^i(x_k) + \theta_k \gamma \chi_{\rho_{k+1}}^i(x_k; d_k), \quad (8)$$

for $i = 1, \dots, m$, where

$$\psi_{\rho_k, \mathfrak{r}}^i(x_k) := \max_{0 \leq j \leq m(k)} \left\{ \psi_{\rho_{k-j}}^i(x_{k-j}) \right\}, \quad \text{for all } k \in \mathbb{N} \cup \{0\}, \quad (9)$$

$m(0) = 0$, $0 \leq m(k) \leq \min\{m(k-1) + 1, N\}$, and $N \geq 0$ is an integer. Let $\mathfrak{r}(k)$ be integer such that $k - m(k) \leq \mathfrak{r}(k) \leq k$ and $\psi_{\rho_k, \mathfrak{r}}^i(x_k) = \psi_{\rho_{\mathfrak{r}(k)}}^i(x_{\mathfrak{r}(k)})$. In Lemma 3, we prove that for any $\eta \in (0, 1)$, inequality (8) is satisfied for sufficiently small $\theta > 0$. Now, we can propose NMOSQP algorithm.

3.2 NMOSQP: A Nonmonotone SQP-based algorithm for MOP

1. Choose $x_0 \in \mathbb{R}^n$, some scalars $r \in (0, 1)$, $\gamma \in (0, 1)$, the initial penalty parameter $\rho > 0$, and an error tolerance ϵ . Set $k := 0$.
2. Solve the quadratic program (5) to find the descent direction (z_k, d_k) with Lagrange multipliers λ_k, θ_k .
3. If $\|d_k\| < \epsilon$, then stop; otherwise, go to Step 4.
4. If $v(x_k) = 0$ or $\chi_{i, \rho_k}(x_k; d_k) < 0$, for all i , then let $\rho_{k+1} = \rho_k$. Otherwise, update ρ_{k+1} using the rule in (6).
5. Compute the step length θ_k as the first number in the sequence $\{1, r, r^2, \dots\}$, such that for all $i = 1, \dots, m$, satisfying (8).
6. Update $x_{k+1} = x_k + \theta_k d_k$. Set $k := k + 1$ and go to Step 2.

The stopping criteria and line search process in Steps 3 and 5 in the NMOSQP algorithm rely on Lemmas 1 and 3.

4 Convergence analysis

4.1 Assumptions

The following assumptions are steadily held during our convergence analysis.

- **(A1)**: The functions f_i , $i = 1, 2, \dots, p$, and g_j , $j = 1, 2, \dots, m$, are twice continuously differentiable over the feasible set $X \subseteq \mathbb{R}^n$ and $\nabla f_j(x)$ and $\nabla g_i(x)$ are Lipschitz continuous for $i = 1, \dots, m$ and $j = 1, \dots, r$ with Lipschitz constant L .
- **(A2)**: The sequences $\{x_k\}$ and $\{(z_k, d_k)\}$, which correspond to the sequence of solutions for the subproblem (5), are bounded.

Lemma 1 establishes a connection between the algorithm's stopping criteria and the critical points of problem (1). It shows that $d(x) \neq 0$ implies x is not a critical point and it provides a descent direction of the penalty.

Lemma 1. [2] Let (z_k, d_k) be the solution of the subproblem (5). Then the following statements hold:

(i) The value of z_k satisfies

$$z_k \leq v(x_k) - \frac{1}{2}d_k^T d_k. \quad (10)$$

(ii) If $d_k = 0$ and MFCQ holds at x_k , then x_k is a strongly critical point of the problem (1).

(iii) If $d_k \neq 0$, then for sufficiently small ρ_k , then the vector d_k is a descent direction of the penalty function $\psi_{\rho_k}^i(x_k)$ at x_k .

Proof. Since the subproblem (5) is identical to that considered in [2], the proof follows similarly from [2, Lemma 1] and is therefore omitted here. $\square$

Remark 2. The proof strategy in lemma 3 follows a standard line of argument used in related single-objective analyses (see, e.g., [41]); nevertheless, we present it here in a form tailored to our exact penalty function and the constrained MOP setting.

Lemma 2. During iteration k of the NMOSQP algorithm, we have

$$\psi_{\rho_{k+1}}^i(x_{k+1}) \leq \psi_{\rho_{k+1}, \tau}^i(x_{k+1}) \leq \psi_{\rho_k, \tau}^i(x_k), \quad (11)$$

for all $k = 0, 1, 2, \dots$. Moreover, the sequence $\psi_{\rho_k, \tau}^i(x_k)$ is convergent.

Proof. Let x_k be not a critical point of (1). The first inequality is obvious from (9). From (9) and $m(k+1) \leq m(k) + 1$ we have

$$\begin{aligned} \psi_{\rho_{k+1}, \tau}^i(x_{k+1}) &= \max_{0 \leq j \leq m(k+1)} \left\{ \psi_{\rho_{k+1-j}}^i(x_{k+1-j}) \right\} \\ &\leq \max_{0 \leq j \leq m(k)+1} \left\{ \psi_{\rho_{k+1-j}}^i(x_{k+1-j}) \right\} \\ &= \max \left\{ \max_{0 \leq j \leq m(k)} \left\{ \psi_{\rho_{k-j}}^i(x_{k-j}) \right\}, \psi_{\rho_{k+1}}^i(x_{k+1}) \right\} \\ &= \max \left\{ \psi_{\rho_k, \tau}^i(x_k), \psi_{\rho_{k+1}}^i(x_{k+1}) \right\} \\ &= \psi_{\rho_k, \tau}^i(x_k). \end{aligned} \quad (12) \quad (13)$$

From Step 4 of the NMOSQP algorithm, we have $\chi_{\rho_{k+1}}^i(x_k; d_k) < 0$ for all $i = 1, \dots, m$. Therefore, the last equality is concluded from (8). This means the sequence $\psi_{\rho_k, \tau}^i(x_k)$ is nonincreasing for all $i = 1, \dots, m$. Consequently, we obtain

$$0 \leq \psi_{\rho_{k+n}}^i(x_{k+n}) \leq \psi_{\rho_{k+n}, \tau}^i(x_{k+n}) \leq \cdots \leq \psi_{\rho_k, \tau}^i(x_k). \quad (14)$$

Thus, the sequence $\{\psi_{\rho_k, \tau}^i(x_k)\}$ is monotonically non-increasing and bounded below by zero, which implies that it must converge. $\square$

Lemma 3. Suppose that (z_k, d_k) is an optimal solution for subproblem (5) at iteration k of the NMOSQP algorithm. If x_k is not a weakly or strongly critical point of problem (1), then there exists θ_k that satisfies the nonmonotone line search (8).

Proof. Since x_k is not a weakly or strongly critical point of problem (1), from Lemma 1 we have $d(x_k) \neq 0$. Under Assumption (A1), it follows from [4, Lemma 2.1(3)] that there exists a constant $L > 0$ such that for $i = 1, \dots, m$,

$$\psi_{\rho_{k+1}}^i(x_k + \theta d_k) \leq \psi_{\rho_{k+1}}^i(x_k) + \theta \chi_{\rho_{k+1}}^i(x_k; d_k) + \frac{1}{2}(\rho_{k+1} + 1)L\theta^2\|d_k\|^2, \quad (15)$$

for $\theta \in [0, 1]$, and let $\gamma \in (0, 1)$ be a constant as specified in Step 1 of NMOSQP algorithm.

Since $\rho_{k+1} \leq \rho_k$, and $(f_i(x_k) - \underline{f}) \geq 0$, from (2), we have

$$\rho_{k+1}(f_i(x_k) - \underline{f}) + v(x_k) \leq \rho_k(f_i(x_k) - \underline{f}) + v(x_k), \quad i = 1, \dots, m, \quad (16)$$

which means

$$\psi_{\rho_{k+1}}^i(x_k) \leq \psi_{\rho_k}^i(x_k). \quad (17)$$

Then, by applying inequality (15), we have

$$\begin{aligned} & \psi_{\rho_{k+1}}^i(x_k + \theta d_k) - \psi_{\rho_k}^i(x_k) - \gamma \theta \chi_{\rho_{k+1}}^i(x_k; d_k) \\ & \leq (1 - \gamma) \theta \chi_{\rho_{k+1}}^i(x_k; d_k) + \frac{1}{2}(\rho_{k+1} + 1)L\theta^2\|d_k\|^2, \end{aligned} \quad (18)$$

for $i = 1, \dots, m$. Step 4 of the NMOSQP algorithm implies

$$\chi_{\rho_{k+1}}^i(x_k; d_k) < 0, \quad i = 1, \dots, m.$$

Since $d_k \neq 0$, from (18), it follows that the right-hand side becomes negative for sufficiently small values of $\theta > 0$, ensuring that the monotone Armijo's rule (7) is satisfied.

Given that $\psi_{\rho_k}^i(x_k) \leq \psi_{\rho_k, \tau}^i(x_k)$, condition (7) implies the validity of (8). Consequently, there exists a scalar $\theta_k > 0$ for which inequality (8) holds. $\square$

Lemma 4. Let (z_k, d_k) be the solution to the subproblem (5). If $x_k \rightarrow x^*$ as $k \rightarrow \infty$, then the sequence $\{(z_k, d_k)\}$ converges to (z^*, d^*) , where (z^*, d^*) solves subproblem (5) at x^* . In

particular, if $d_k \rightarrow 0$ and the MFCQ holds at every x_k , then x^* is a strong critical point of the problem (1).

Proof. Since the subproblem (5) is identical to that considered in [2], the proof follows directly from [2, Lemma 3] and is therefore omitted. $\square$

Lemma 5. Suppose that $\rho_k = \rho > 0$ for all sufficiently large k . Assume that $\{x_k\}_{k \in K}$ is a convergent subsequence. Then $d_k \rightarrow 0$ as $k \rightarrow \infty$ and $k \in K$.

Proof. Without loss of generality, assume that $\rho_k = \rho > 0$ for all $k \in K$. If possible suppose that there exist an infinite subset $K' \subset K$ and a positive constant η such that

$$\|d_k\| \geq \eta, \quad k \in K'. \quad (19)$$

First, we will show that there exists $\delta > 0$ such that $\theta_k \geq \delta$ holds for every k , where θ_k is the step length obtained in Step 5 of NMOSQP algorithm. From this step either $\theta_k = 1$ or $\theta_k = r^{k_1}$ holds for some $k_1 \in \mathbb{N}$. If $\theta_k = r^{k_1}$ holds then there exists $i \in \{1, \dots, m\}$ satisfying,

$$\psi_\rho^i(x_k + r^{k_1-1}d_k) - \psi_{\rho, \tau}^i(x_k) > r^{k_1-1}\gamma\chi_\rho^i(x_k; d_k).$$

Then from (18),

$$r^{k_1-1}(1 - \gamma)\chi_\rho^i(x_k; d_k) + \frac{1}{2}(1 + \rho)Lr^{2(k_1-1)}\|d_k\|^2 > 0.$$

Thus,

$$\frac{1}{2}(1 + \rho)Lr^{2(k_1-1)}\|d_k\|^2 \geq -r^{k_1-1}(1 - \gamma)\chi_\rho^i(x_k; d_k). \quad (20)$$

If $\exists j \in \{1, \dots, r\}$ such that $g_j(x_k) + \nabla g_j(x_k)^T d_k > 0$, then from Lemma 1 and subproblem (5), we have

$$\max_{j=1, \dots, r} \{g_j(x_k) + \nabla g_j(x_k)^T d_k, 0\} - v(x_k) \leq v(x_k) - \frac{1}{2}d_k^T d_k - v(x_k) = -\frac{1}{2}d_k^T d_k. \quad (21)$$

If $g_j(x_k) + \nabla g_j(x_k)^T d_k \leq 0$, for all $j = 1, \dots, r$, then

$$\max_{j=1, \dots, r} \{g_j(x_k) + \nabla g_j(x_k)^T d_k, 0\} - v(x_k) = -v(x_k). \quad (22)$$

Thus, by the definition of v^* and $\mathcal{A}(x) \subseteq \{1, \dots, r\}$, we have

$$v^*(x_k; d_k) \leq \max_{j=1, \dots, r} \{g_j(x_k) + \nabla g_j(x_k)^T d_k, 0\} - v(x_k) \leq \max\{-\frac{1}{2}d_k^T d_k, -v(x_k)\}. \quad (23)$$

The last inequality follows from (5), (21), and (22).

Since $(z_k, d_k) = (0, 0)$ is feasible for subproblem (5), we have $z_k + \frac{1}{2}d_k^T d_k \leq 0$. Then, (3) and (5) imply

$$\begin{aligned}\chi_\rho^i(x_k; d_k) &\leq \rho z_k + v^*(x_k; d_k) \\ &\leq \rho\left(-\frac{1}{2}d_k^T d_k\right) + \max\left\{-\frac{1}{2}d_k^T d_k, -v(x_k)\right\} \\ &\leq -\frac{\rho}{2}d_k^T d_k.\end{aligned}$$

From (20) we have

$$\frac{1}{2}(1+\rho)Lr^{k_1-1}\|d_k\|^2 \geq \frac{\rho}{2}(1-\gamma)\|d_k\|^2.$$

This implies $r^{k_1} \geq \frac{r(1-\gamma)\rho}{(1+\rho)L}$. Choose $\delta = \min\left\{1, \frac{r(1-\gamma)\rho}{(1+\rho)L}\right\}$. Then $\theta_k \geq \delta$ holds for every k . Now, from (7) and (20), we have

$$\begin{aligned}\psi_\rho^i(x_{k+1}) - \psi_\rho^i(x_0) &= \sum_{l=0}^k [\psi_\rho^i(x_{l+1}) - \psi_\rho^i(x_l)] < -\frac{\gamma\rho}{2} \sum_{l=0}^k \theta_l \|d_l\|^2 \\ &\leq -\frac{1}{2}(k+1)\delta\gamma\rho\eta^2, \quad \text{for all } k \in K' .\end{aligned}$$

This implies $\psi_\rho^i(x_k + \theta_k d_k) \rightarrow -\infty$ as $k \rightarrow \infty$ and $k \in K'$ (since $\theta_0\gamma\rho\eta^2 > 0$). This contradicts the assumption that $\{x_k\}, \{(z_k, d_k)\}$ are bounded as ψ_ρ^i is a continuous function. So there exists no any $K' \subset K$ and $\eta > 0$ such that (19) holds. $\square$

Lemma 6. If $\rho_k \rightarrow 0$, then $\lim_{k \rightarrow \infty} v(x_k) = 0$.

Proof. Suppose that $\lim_{k \rightarrow \infty} v(x_k)$ exists if $\rho_k \rightarrow 0$. By the definition of $v(x_k)$, we have $\lim_{k \rightarrow \infty} v(x_k) \geq 0$. Suppose $\lim_{k \rightarrow \infty} v(x_k) > 0$. Then there exists a constant $c_1 > 0$ such that for k large enough $v(x_k) > c_1$. We consider two cases separately.

Case 1: There exists a positive constant $c_2 > 0$ such that, for k large enough, $\|d_k\| > c_2$. Thus, from (23) and under assumptions (A1) and (A2), the following inequality holds for k large enough:

$$v^*(x_k; d_k) \leq \max\left\{-\frac{1}{2}d_k^T d_k, -v(x_k)\right\} \leq -c_3,$$

where $c_3 = \min\{\frac{1}{2}c_2^2, c_1\} > 0$. Then, from (3), (5), and Lemma 1, we have

$$\begin{aligned}\chi_{\rho_k}^i(x_k; d_k) &\leq \rho_k z_k + v^*(x_k; d_k) \\ &\leq \rho_k(v(x_k) - \frac{1}{2}d_k^T d_k) + \max\left\{-\frac{1}{2}d_k^T d_k, -v(x_k)\right\}.\end{aligned}$$

Thus, if $\rho_k \rightarrow 0$, then $\chi_{\rho_k}^i(x_k; d_k) \leq -c_3$. From Step 4 of the NMOSQP algorithm, this would imply that $\rho_{k+1} = \rho_k$ for all sufficiently large k , which contradicts the assumption that $\rho_k \rightarrow 0$.

Case 2: Case 1 does not hold. Thus there exists an infinite index set K such that $\{d_k\}_{k \in K}$ converges to zero. Since K is an infinite index set, by Assumption (A2), $\{x_k\}_{k \in K}$ has at least one accumulation point. Let $K' \subseteq K$ such that $\{x_k\}_{k \in K'}$ converges to $\bar{x}$. By Lemma 4, $\bar{x}$ is a strong stationary point of (1). Hence, $v(\bar{x}) = 0$, which implies that $v(x_k) < c_1$ holds for $k \in K'$ large enough. This contradicts $v(x_k) > c_1$. $\square$

Theorem 2. Let $\{x_k\}$ be a sequence generated by NMOSQP algorithm, and let the MFCQ be satisfied at every x_k . Then any accumulation point of $\{x_k\}$ is a critical point (either strongly or weakly critical point) of problem (1).

Proof. (i) Convergence to a strongly critical point:

Let K be an infinite index set such that $x_k \rightarrow x^*$ as $k \rightarrow \infty$ and $k \in K$. Let $\{(z_k, d_k)\}$ be the solution of subproblem (5). If $d_k \rightarrow 0$ as $k \rightarrow \infty$, then by Lemma 4, x^* is a strongly critical point of P .

(ii) Convergence to a weakly critical point:

If there exists a constant $c_0 > 0$ such that $\|d_k\| \geq c_0$ for large $k \in K$, then from Lemma 5, $\rho_k \rightarrow 0$ as $k \rightarrow \infty$. Since the sequences $\{x_k\}, \{(z_k, d_k)\}$ are bounded, so from Lemma 6, $\lim_{k \rightarrow \infty} v(x_k) = 0$. Hence from (10) and subproblem (5),

$$\nabla f_j(x_k)^T d_k \leq z_k < 0, \quad \text{for all } i = 1, \dots, m.$$

This implies, $d_k \in D(x_k)$ for large k . We follow the proof by contradiction. Suppose that x_k is not a weakly critical point. From Definition 2, there exists a constant $\eta_2 > 0$ such that for sufficiently large k ,

$$\max_{d \in D(x_k)} \max_{j=1, \dots, r} \{g_j(x_k) + \nabla g_j(x_k)^T d; 0\} \leq v(x_k) - \eta_2.$$

Suppose that $\hat{d}_k$ maximizes $\max_{j=1, \dots, r} \{g_j(x_k) + \nabla g_j(x_k)^T d; 0\}$. From (23) we have

$$v^*(x_k; d_k) \leq \max\{-\frac{1}{2} d_k^T d_k, -v(x_k)\} \leq \max_{d \in D(x_k)} \max\{-\frac{1}{2} d^T d, -v(x_k)\} \leq -\eta_2.$$

Then, from (3), (5), and Lemma 1, we have

$$\begin{aligned} \chi_{\rho_k}^i(x_k; d_k) &\leq \rho_k z_k + v^*(x_k; d_k) \\ &\leq \rho_k(v(x_k) - \frac{1}{2} d_k^T d_k) + \max\{-\frac{1}{2} d_k^T d_k, -v(x_k)\}. \end{aligned}$$

This means that if $\rho_k \rightarrow 0$, then $\chi_{\rho_k}^i(x_k; d_k) \leq -\eta_2$. From Step 4 of the NMOSQP algorithm, it follows that $\rho_{k+1} = \rho_k$ for all sufficiently large k , which contradicts that $\rho_k \rightarrow 0$. As a result,

the sequence $\{x_k\}$ converges to a weakly critical point. Hence, any accumulation point of $\{x_k\}$ is either a strongly critical point or a weakly critical point. $\square$

5 Numerical results

In this section, we describe the MATLAB R2021a implementation of NMOSQP and assess its performance on a diverse collection of standard test problems from the MOPs literature. The test set covers a broad range of features, including convex and nonconvex instances, biobjective and three-objective models, and both box-constrained and generally constrained MOPs. Table 1 summarizes the test problems: The first column reports the problem name; n denotes the decision-space dimension (set from 2 to 200); and n_f and n_g represent the numbers of objective functions and constraints, respectively. A dash in the n_g column of Table 1 indicates that the problem has no explicit constraints. The bound vectors x_L and x_U are also reported, and the last column provides the corresponding references.

Table 1: Details of box-constrained and general constrained test problems.

Name	n	n_f	n_g	x_L	x_U	Reference
ABC-comp	2	2	3	[0, 0]	[10, 10]	[28]
BNH	2	2	2	[0, 0]	[5, 3]	[12]
DTLZ0	3	3	2	[0, 0, 0]	[2, 1, 1]	[14]
Eichfelder2	2	2	1	[0, 0]	[100, 100]	[20]
Eichfelder6	3	3	1	[0, 0, 0]	[100, 100, 100]	[20]
Eichfelder7	3	3	1	[0, 0, 1.2]	$[\pi, 100, 100]$	[20]
FDSa	5	3	—	$-[2, \dots, 2]$	$[2, \dots, 2]$	[14]
FDSb	10	3	—	$-[2, \dots, 2]$	$[2, \dots, 2]$	[14]
FDSc	50	3	—	$-[2, \dots, 2]$	$[2, \dots, 2]$	[14]
FDSd	100	3	—	$-[2, \dots, 2]$	$[2, \dots, 2]$	[14]
Hanne2	2	2	1	[0, 0]	[10, 10]	[9]
Hanne4	2	2	1	[0, 0]	[10, 10]	[9]
HLL1	2	2	1	[0, 0]	[100, 100]	[28]
JOS1a	50	2	—	$-[2, \dots, 2]$	$[2, \dots, 2]$	[12]
JOS1c	200	2	—	$-[2, \dots, 2]$	$[2, \dots, 2]$	[20]
JOS1h	200	2	—	$-100[1, \dots, 1]$	$100[1, \dots, 1]$	[14]
P11	10	3	3	$-[2, \dots, 2]$	$[2, \dots, 2]$	[36]
P12	50	3	3	$-[2, \dots, 2]$	$[2, \dots, 2]$	[36]
P13	200	3	3	$-[2, \dots, 2]$	$[2, \dots, 2]$	[36]
TNK	2	2	2	[0, 0]	$[\pi, \pi]$	[13]

To ensure that the sequence generated by NMOSQP remains inside the box constraints $x_L \leq x \leq x_U$, these constraints are added to subproblem (5), which leads to the following subproblem:

$$\begin{aligned} \min \quad & q(v, z) = z + \frac{1}{2} \|v\|^2, \\ \text{s.t.} \quad & \begin{cases} (\nabla f_i(x))^T v \leq z, & \text{for } i = 1, 2, \dots, m, \\ g_j(x) + (\nabla g_j(x))^T v \leq z, & \text{for } j = 1, \dots, r, \\ x_L - x \leq v \leq x_U - x, \\ (v, z) \in \mathbb{R}^n \times \mathbb{R}. \end{cases} \end{aligned} \quad (24)$$

In our implementation, we utilize subproblem (24) instead of (5) to ensure practical applicability. The stopping criterion is defined as $\|d_k\| \leq \varepsilon$, where $\varepsilon = 10^{-4}$. The Armijo rule parameter is set to $\gamma = 0.1$. To handle potential nonconvergence in nonconvex problems, we limit the maximum number of iterations to 500. We employ $N = 10$ in the nonmonotone term (9). At each iteration, we solve subproblem (24) using MATLAB's `quadprog` function.

Both MOSQP and NMOSQP start from arbitrary initial points, which we generate using MATLAB's `randn` function. To ensure a comprehensive evaluation, we generate 100 initial points for each method. Employing a multi-start strategy [21], we apply each algorithm from all initial points to approximate the nondominated frontier.

To comprehensively compare NMOSQP and MOSQP, we conduct two experiments, each focusing on a different performance aspect. In the first experiment, we examine the capability of NMOSQP to approximate the Pareto front and compare its results with those obtained by MOSQP as well as with the reference (true) Pareto front. Figure 1 shows the nondominated fronts generated by NMOSQP and MOSQP together with the reference front. For the Hanne2 problem, NMOSQP does not cover the entire front; however, it produces a dense and well-distributed approximation in the middle region, which is noticeably superior to MOSQP. In contrast, MOSQP performs poorly on Hanne2 and even generates dominated points. For the ABC-comp and BNH problems, both methods exhibit similar and satisfactory behavior relative to the reference front. In the TNK problem, NMOSQP generates a larger nondominated portion of the front than MOSQP; nevertheless, the two approximations are complementary, and their union covers a broader part of the true front, yielding a more well-distributed overall approximation.

In the second experiment, we use the Dolan and Moré performance profiles [15] to compare NMOSQP and MOSQP across the specified performance criteria.

The performance profile is a more reliable visualization tool for comparing the results obtained by different solvers, especially when the number of problems and solvers increases. Follow-

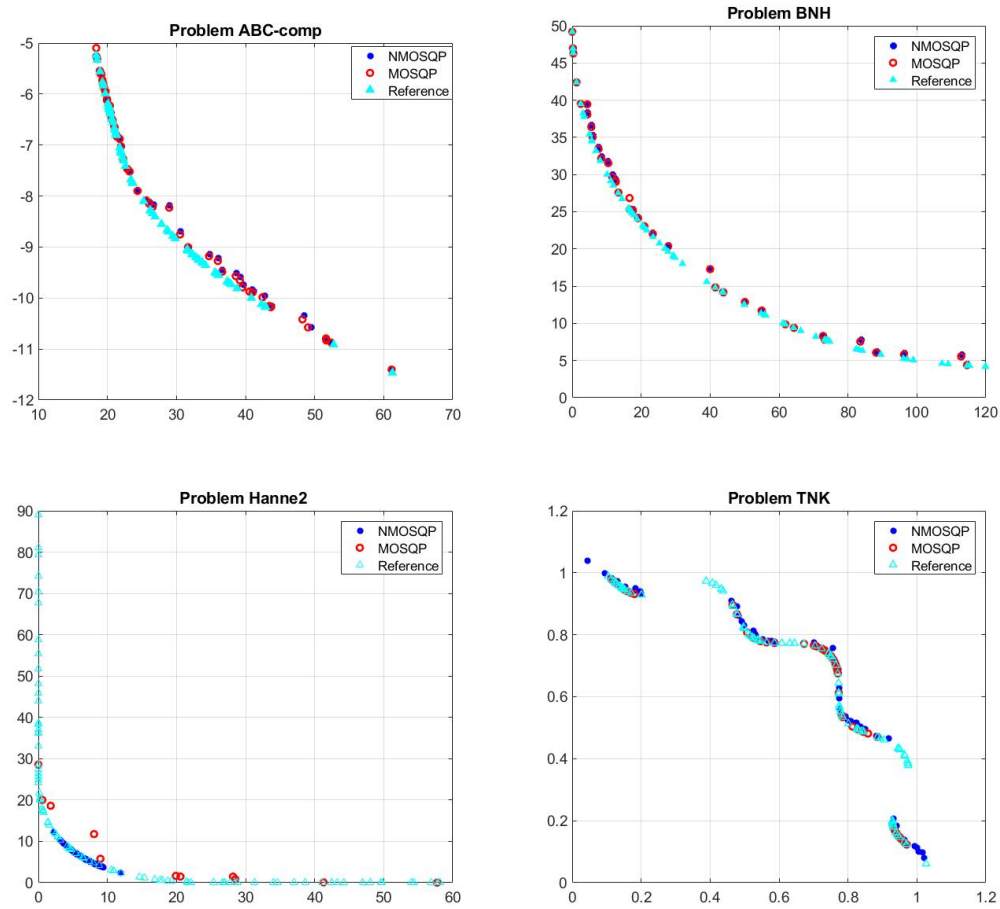


Figure 1: Nondominated fronts generated by NMOSQP and MOSQP on some test problems.

ing is a simplified description of the performance profile of a specific criterion for a solver associated with a problem [15]. Consider $t_{p,s}$ as the performance of the solver $s \in S$ in solving the problem $p \in P$. Accordingly, the performance ratio $r_{p,s}$ is defined as $r_{p,s} := t_{p,s} / \min\{t_{p,s} : s \in S\}$. Based on this definition, the probability of the performance ratio of the solver $s \in S$ to be within the interval $[1, \tau)$ is denoted by $\rho_s(\tau)$ and defined as follows:

$$\rho_s(\tau) := \frac{1}{|P|} |p \in P : r_{p,s} \leq \tau|,$$

where ρ_s is the cumulative distribution function for the performance ratio [15]. The performance profiles depict the function $\rho_s(\tau)$ for each solver $s \in S$. According to the above explanation

about $\rho_s(\tau)$, the value of $\rho_s(1)$ is the probability that the solver s will overcome the remaining ones. Thus, $\rho_s(1)$ shows the win percentage of solver s compared to the other solvers.

Our evaluation is based on two complementary aspects: (i) computational effort and (ii) quality of the produced Pareto front approximation. The criteria related to computational effort (and hence the speed of convergence) include the number of iterations, the numbers of objective and constraint function evaluations, the numbers of gradient evaluations for both objectives and constraints, and the wall-clock time. For multiobjective methods, however, it is equally important to assess the quality of the approximated Pareto front. Therefore, in addition to the above measures, we report the number of generated nondominated points and three widely used multiobjective indicators: The epsilon indicator [29], the spread [3, 10], and the hyper-volume (HV) [44]. The epsilon indicator provides a measure of how far the computed front is from the reference Pareto front, the spread evaluates how evenly the points are distributed along the front, and HV quantifies the portion of the objective space dominated (covered) by the obtained nondominated set (larger HV indicates a better approximation).

In this study, the HV indicator is computed using a common reference point selected separately for each test problem. For a given problem, we first form the union of all solutions generated by all algorithms over all runs. Then, for each objective, we determine the best observed value (the smallest one, since all problems are minimization) and the worst observed value (the largest one); the difference between these values defines the observed objective range. The HV reference point is chosen slightly worse than the worst observed value by adding 5% of this range to each objective component. This choice ensures that the reference point is dominated by all observed solutions and, consequently, makes the HV values fair and comparable across algorithms. With this setting, larger HV values indicate better performance. Finally, because Dolan and Moré performance profiles are typically constructed for measures where smaller values are better, we transform HV and the number of nondominated points (both larger-is-better criteria) into smaller-is-better quantities before plotting, so that algorithms achieving larger HV or producing more nondominated points are correctly identified as superior in the performance profiles.

Figure 2 compares the two methods with respect to the computational criteria. It contains four subplots reporting the number of iterations, wall-clock time, the number of objective function evaluations, and the number of constraint function evaluations.

Figure 3 summarizes the quality of the obtained Pareto front approximations by reporting the number of generated nondominated points for both methods, epsilon indicator, spread, and the HV metrics.

We provide a detailed comparison of NMOSQP and MOSQP using the Dolan and Moré performance profiles across the selected computational and quality criteria. For the *number*

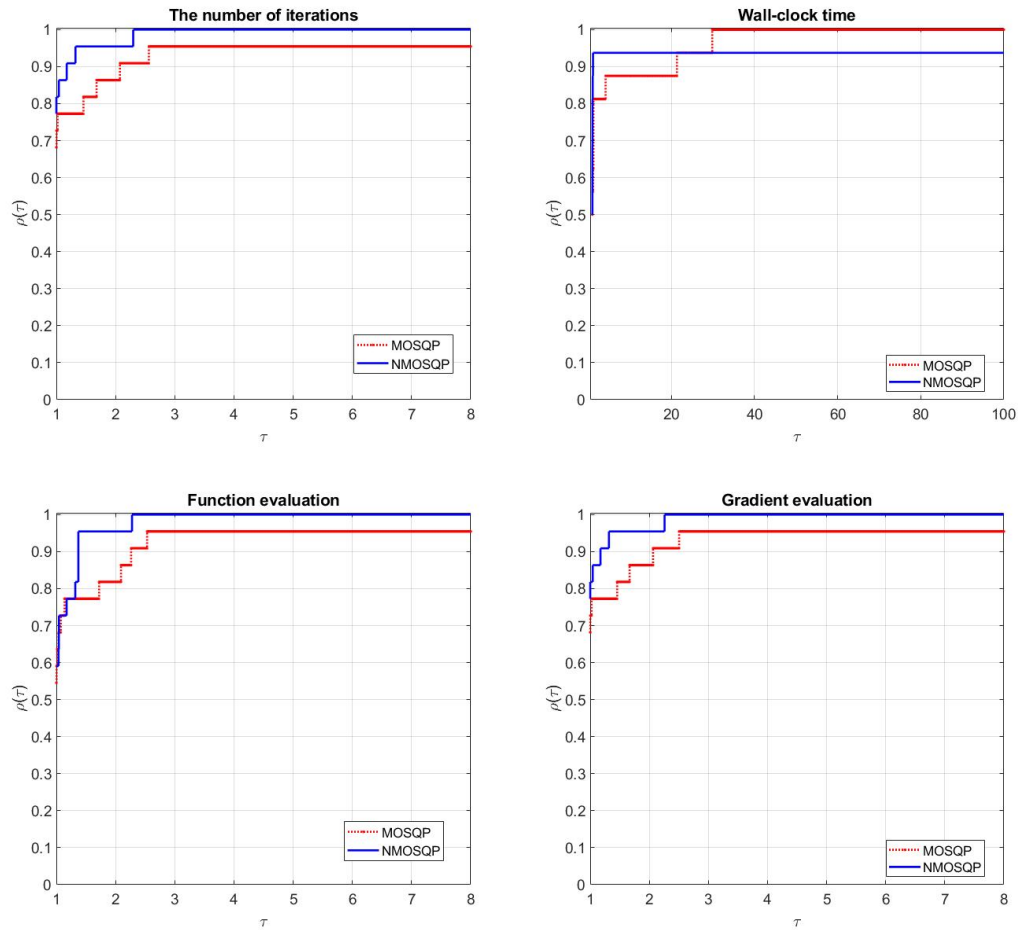


Figure 2: Iterations, Function evaluations, Gradient evaluations, and Wall-clock time performance profiles.

of iterations and gradient evaluations, NMOSQP clearly dominates and is the best performer across all test instances. Regarding function evaluations, the two methods are competitive on roughly 80% of the test instances; nevertheless, NMOSQP attains the best performance more frequently and shows an overall advantage. A similar behavior is observed for wall-clock time: Although the methods compete in about 80% of the instances, NMOSQP achieves the best performance on approximately 95% of the test problems, indicating a clear improvement in practical efficiency. Turning to the quality-related performance profiles, NMOSQP and MOSQP are largely competitive in terms of front generation and the epsilon indicator. For the spread metric, the methods still compete in about 85% of the instances, but NMOSQP more often ranks first and thus provides a better overall distribution of points along the front. Finally, the

HV profile shows a significant superiority of NMOSQP over MOSQP, indicating that NMOSQP produces nondominated sets that dominate a larger portion of the objective space and hence yield higher-quality Pareto-front approximations.

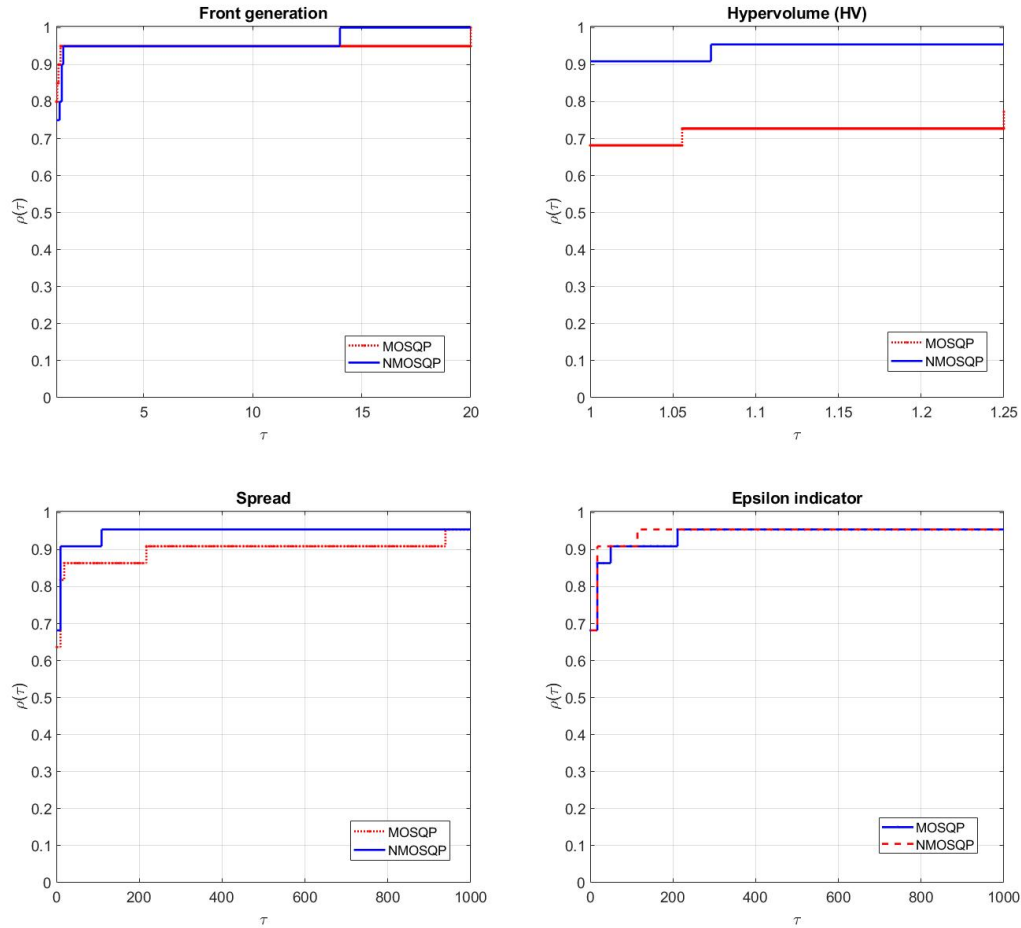


Figure 3: Number of nondominated points, Hypervolume(HV), Spread, and Epsilon indicator performance profiles.

6 Conclusion

This paper proposed a nonmonotone penalty method in the SQP framework for solving general inequality-constrained MOPs. In contrast to standard monotone penalty-based schemes, which enforce a sufficient decrease of the penalty function at every iteration, the proposed approach

employs a nonmonotone max-type line search that allows controlled temporary increases of the penalty function. This additional flexibility can prevent overly conservative steps and improve practical efficiency on difficult constrained instances. We compare the proposed nonmonotone method (NMOSQP) with the monotone baseline (MOSQP). Dolan and Moré performance profiles indicate that NMOSQP is consistently faster and more efficient across the considered test set. Moreover, under mild regularity assumptions, we establish global convergence results showing that, for any starting point, the sequence generated by NMOSQP converges to a weakly or strongly critical point. Finally, as a direction for future research, it is of interest to investigate alternative nonmonotone line search rules within the proposed penalty-SQP framework to further enhance performance.

Acknowledgements

The authors are grateful to the anonymous referees and the Editor for their constructive comments.

Funding

This research received no external funding.

Conflicts of Interest

The authors declare no conflicts of interest.

References

- [1] Amini, K. and Rashidi, M. On determining radius in nonmonotone trust-region approaches, *J. Math. Modeling*, 11(3) (2023), 507–526.
- [2] Ansary, M.A.T. and Panda, G. A sequential quadratic programming method for constrained multiobjective optimization problems, *J. Appl. Math. Comput.* 64(1) (2020), 379–397.
- [3] Bandyopadhyay, S., Pal, S.K. and Aruna, B. Multiobjective GAs, quantitative indices, and pattern classification, *IEEE Trans. Syst. Man Cybern. B Cybern.* 34(5) (2004), 2088–2099.

- [4] Bazaraa, M.S. and Goode, J.J. An algorithm for solving linearly constrained minimax problems, *Eur. J. Oper. Res.* 11(2) (1982), 158–166.
- [5] Burke, J.V., Curtis, F.E. and Wang, H. A sequential quadratic optimization algorithm with rapid infeasibility detection, *SIAM J. Optim.* 24(2) (2014), 839–872.
- [6] Carrizo, G.A., Fazzio, N.S. and Schuverdt, M.L. A nonmonotone projected gradient method for multiobjective problems on convex sets, *J. Oper. Res. Soc. China*, 12(2) (2024), 410–427.
- [7] Chen, J. Liu, J. Qin, X. and Yao, J.-C. A nonmonotone proximal gradient algorithm for solving nonsmooth multiobjective optimization problems with an extending application to robust multiobjective optimization, *J. Comput. Appl. Math.* 460 (2025).
- [8] Chen, W. Yang, X. and Zhao, Y. Conditional gradient method for vector optimization, *Comput. Optim. Appl.* 85(3) (2023), 857–896.
- [9] Collette, Y. and Siarry, P. Multiobjective optimization: Principles and case studies, Springer Science & Business Media, 2003.
- [10] Custódio, A.L., Madeira, J.F., Vaz, A.I. and Vicente, L. N. Direct multisearch for multiobjective optimization, *SIAM J. Optim.* 21 (2011), 1109–1140.
- [11] Dai, Y.-H. On the Nonmonotone Line Search, *J. Optim. Theory Appl.* 112 (2002), 315–330.
- [12] Deb, K. Multiobjective optimization using evolutionary algorithms, John Wiley & Sons, 2001.
- [13] Deb, K., Pratap, A. and Meyarivan, T. Constrained test problems for multiobjective evolutionary optimization, in *Evolutionary Multi-Criterion Optimization*, Springer, Berlin, Heidelberg, 2001, 284–298.
- [14] Deb, K., Thiele, L., Laumanns, M. and Zitzler, E. Scalable test problems for evolutionary multiobjective optimization, Springer, London, 2005.
- [15] Dolan, E.D. and Moré, J.J. Benchmarking optimization software with performance profiles, *Math. Program.* 91 (2002), 201–213.
- [16] Dolatnezhadsomarin, A. and Khorram, E. Two efficient algorithms for constructing almost even approximations of the Pareto front in multiobjective optimization problems, *Eng. Optim.* 51(4) (2019), 567–589.

- [17] Drummond, L.M.G. A projected gradient method for vector optimization problems, *Comput. Optim. Appl.* 28 (2004), 5–29.
- [18] Drummond, L.M.G. and Svaiter, B.F. A steepest descent method for vector optimization, *J. Comput. Appl. Math.* 175 (2005), 395–414.
- [19] Ehrgott, M. *Multicriteria optimization*, Springer, Berlin, 2005.
- [20] Eichfelder, G. *Adaptive scalarization methods in multiobjective optimization*, Springer, Berlin, 2008.
- [21] Fliege, J., Drummond, L.M.G. and Svaiter, B.F. Newton’s method for multiobjective optimization, *SIAM J. Optim.* 20(2) (2009), 602–626.
- [22] Fliege, J. and Svaiter, B.F. Steepest descent methods for multicriteria optimization, *Math. Methods Oper. Res.* 51 (2000), 479–494.
- [23] Fliege, J. and Vaz, A.I.F. A method for constrained multiobjective optimization based on SQP techniques, *SIAM J. Optim.* 26(4) (2016), 2091–2119.
- [24] Gebken, B., Peitz, S. and Dellnitz, M. A descent method for equality and inequality constrained multiobjective optimization problems, in *Numerical and Evolutionary Optimization – NEO 2017*, Springer International Publishing, 2018, 29–61.
- [25] Ghalavand, N., Khorram, E. and Morovati, V. Two adaptive nonmonotone trust-region algorithms for solving multiobjective optimization problems, *Optimization*, 73(9) (2023), 2953–2985.
- [26] Gonçalves, D.S., Gonçalves, M.L.N. and Melo, J.G. An away-step Frank–Wolfe algorithm for constrained multiobjective optimization, *Comput. Optim. Appl.* 88(3) (2024), 759–781.
- [27] Grippo, L., Lampariello, F. and Lucidi, S. A nonmonotone line search technique for Newton’s method, *SIAM J. Numer. Anal.* 23 (1986), 707–716.
- [28] Hwang, C.-L. and Masud, A.S.M. *Multiple objective decision making—Methods and applications*, *Lecture Notes in Economics and Mathematical Systems*, vol. 164, Springer, Berlin, 1979.
- [29] Knowles, J., Thiele, L. and Zitzler, E. A Tutorial on the Performance Assessment of Stochastic Multiobjective Optimizers, TIK Report 214, Computer Engineering and Networks Laboratory, ETH Zurich, 2006.

- [30] Mahdavi-Amiri, N. and Salehi Sadaghiani, F. A superlinearly convergent nonmonotone quasi-Newton method for unconstrained multiobjective optimization, *Optim. Methods Softw.* 35(6) (2020), 1223–1247.
- [31] Mangasarian, O.L. and Fromovitz, S. The Fritz John necessary optimality conditions in the presence of equality and inequality constraints, *J. Math. Anal. Appl.* 17(1) (1967), 37–47.
- [32] Miettinen, K. Nonlinear multiobjective optimization, *International Series in Operations Research and Management Science*, vol. 12, Springer Science and Business Media, 2012.
- [33] Morovati, V. and Pourkarimi, L. Extension of Zoutendijk method for solving constrained multiobjective optimization problems, *Eur. J. Oper. Res.* 273(1) (2019), 44–57.
- [34] Pinheiro, M.E. and Grapiglia, G.N. Universal nonmonotone line search method for nonconvex multiobjective optimization problems with convex constraints, *Comput. Appl. Math.* 44(1) (2025), 56.
- [35] Pirouz, B. and Khorram, E. A computational approach based on the ε -constraint method in multiobjective optimization problems, *Adv. Appl. Stat.* 49(6) (2016), 453–483.
- [36] Rashidi, M., Khorram, E. and Soleimani-Damaneh, M. An exact penalty method with nonmonotone line search and rapid infeasibility detection for constrained multiobjective optimization: Application in supervised machine learning, *Comput. Oper. Res.* 188 (2026), 107351.
- [37] Rashidi, M. and Soleimani-damaneh, M. MultiSQP-GS: a sequential quadratic programming algorithm via gradient sampling for nonsmooth constrained multiobjective optimization, *Comput. Optim. Appl.* 89 (2024), 729–767.
- [38] Tanabe, H., Fukuda, E.H., and Yamashita, N. An accelerated proximal gradient method for multiobjective optimization, *Comput. Optim. Appl.* 86(2) (2023), 421–455.
- [39] Upadhayay, A., Ghosh, D., Jauny, Yao, J.C. and Zhao, X. A nonmonotone conditional gradient method for multiobjective optimization problems, *Soft Comput.* 28(17-18) (2024), 9609–9630.
- [40] Upadhayay, A., Ghosh, D. and Kumar, K. Nonmonotone Wolfe-type quasi-Newton methods for multiobjective optimization problems, *Optimization*, (2025), 1–33.

- [41] Zhang, H.C. and Hager, W.W. A nonmonotone line search technique for unconstrained optimization, *SIAM J. Optim.* 14(4) (2004), 1043–1056.
- [42] Zhao, X., Raushan, R., Ghosh, D., Yao, J.-C. and Qi, M. Proximal gradient method for convex multiobjective optimization problems without Lipschitz continuous gradients, *Comput. Optim. Appl.* 91(1) (2025), 27–66.
- [43] Zhao, X. and Yao, J.-C. Linear convergence of a nonmonotone projected gradient method for multiobjective optimization, *J. Global Optim.* 82(3) (2022), 577–594.
- [44] Zitzler, E., Thiele, L., Laumanns, M., Fonseca, C.M. and da Fonseca, V.G. Performance assessment of multiobjective optimizers: an analysis and review, *IEEE Trans. Evol. Comput.* 7(2) (2003), 117–132.



Design and implementation of a graph-coloring algorithm for optimizing flight-level allocation in air traffic management in Iran

S.A.A. Mosavi, A. Erfanian and N. Sabeghi*

Abstract

The rapid growth of air traffic demand highlights the necessity of efficient and reliable methods for air traffic flow management (ATFM). In Iran, current flight level allocation is predominantly performed manually by human operators, which is prone to errors, lacks scalability, and does not guarantee optimal use of available airspace resources. To address this limitation, this study proposes a novel optimization framework based on graph coloring techniques for the allocation of flight levels.

The airspace is modeled as a graph, where each flow corresponds to a node and potential conflicts are represented as edges. The problem is then formulated as an optimization model

*Corresponding author

Received 11 October 2025; revised 27 December 2025; accepted 5 January 2026

Seyed Ali Asghar Mosavi

Department of Pure Mathematics, Faculty of Mathematical Sciences, Ferdowsi University of Mashhad, Mashhad, Iran. e-mail: saam9473@gmail.com

Ahmad Erfania

Department of Pure Mathematics, Faculty of Mathematical Sciences, Ferdowsi University of Mashhad, Mashhad, Iran. e-mail: erfanian@um.ac.ir

Narjes Sabeghi

Department of Mathematics, Faculty of Basic Sciences, Velayat University, Iranshahr, Iran. e-mail: n.sabeghi@velayat.ac.ir

How to cite this article

Mosavi, S.A.A., Erfanian, A. and Sabeghi, N., Design and implementation of a graph-coloring algorithm for optimizing flight-level allocation in air traffic management in Iran. *Iran. J. Numer. Anal. Optim.*, 2026; 16(2): 590-606. <https://doi.org/10.22067/ijnao.2026.95911.1749>

with the goal of minimizing the number of distinct flight levels while ensuring safety constraints. A hybrid algorithm is developed that combines the DSatur heuristic for generating an initial solution with a constraint programming model enhanced by maximal clique detection for refinement and optimization.

The approach is applied to real operational data from Tehran's Mehrabad and Mashhad Hasheminejhad Airports during peak hours. Experimental results demonstrate that the proposed method reduces the number of required flight levels compared to DSatur alone, ensuring both improved airspace efficiency and conflict-free operations.

AMS subject classifications (2020): Primary 05C90, 90C59; Secondary 05C15.

Keywords: Air traffic flow management (ATFM); Graph coloring techniques; DSatur heuristic; Constraint programming (CP).

1 Introduction

With the rapid expansion of cultural, commercial, and tourism interactions at both the national and international levels, the air transport industry, as one of the most important connectivity infrastructures, plays a vital role in facilitating fast and safe movement. The dramatic reduction in travel time compared with other modes of transportation, and the ability to connect cities, countries, and even continents, have further increased the importance of this industry. The continuous rise in demand for air travel makes precise planning and the use of scientific solutions unavoidable in order to maintain safety, optimize resource utilization, and reduce operational risks.

At present, air traffic management in Iran relies heavily on human experience and decision-making. This dependence, while creating limits on the scalability of operations, can increase the likelihood of errors. The lack of intelligent systems based on mathematical models has made the need to develop novel methods, capable of automatically performing optimal allocation of flight resources with minimal human intervention, even more urgent.

One effective approach in this area is the use of graph modeling and combinatorial optimization algorithms, particularly graph-coloring algorithms. In this modeling, each route or section of airspace is represented as a vertex (node), and conflicts or concurrency constraints between flights are represented as edges. The problem of assigning flight levels and scheduling flights can be viewed as equivalent to the graph-coloring problem, in which colors denote safe altitudes (flight levels) or permissible time slots for flights. The objective is to minimize the number of colors (flight resources) while satisfying all safety and operational constraints.

To achieve this goal, real data obtained from radar control stations are used as system inputs. These data are prepared for modeling through the following steps:

1. Defining the problem domain and collecting raw data,
2. Preprocessing and cleaning the data,
3. Extracting patterns and mathematical relationships from the data,
4. Evaluating and applying the extracted patterns in the proposed system.

In this study, by employing a multi-objective mathematical programming model, an algorithm for graph coloring is presented that can simultaneously eliminate flight conflicts, ensure optimal utilization of resources, and enhance flight safety within the country's airspace. The structure of this paper is as follows: Section 2 reviews related research on the application of graph coloring in air traffic management. Section 3 introduces the dataset used, along with preprocessing steps and the construction of the flight graph model. Section 4 presents the proposed algorithm and details of its implementation. Section 5 discusses experimental results and performance analysis of the proposed approach in real-world scenarios. Finally, Section 6 provides conclusions and suggestions for future research.

2 Literature review

Graph coloring, as a mathematical modeling tool, has found remarkable applications in recent years in the field of Air traffic flow management (ATFM). In this section, two key studies in this area are reviewed.

Barnier and Brisset [1] proposed an innovative model for ATFM in France, based on the design of direct route networks and the static allocation of flight levels to intersecting routes. By reformulating the problem as a graph-coloring problem and leveraging a combination of constraint programming (CP) methods and greedy algorithms, the model aimed to completely eliminate flight conflicts. The key novelty of this study lies in the use of a greedy algorithm to identify large cliques in the intersection graph and apply distinct global constraints to them. Moreover, by designing a clique-based search strategy and breaking color symmetry, the model was able to obtain optimal solutions in a short time for most real-world cases. Findings showed that for flows with more than five flights, a maximum of 12 flight levels is required, and the approach outperforms the DSatur (Degree of saturation) algorithm in large instances. In addition, the model can be adapted to the current structure of European airspace (altitudes from 20,000 to

40,000 feet) and potentially scaled up to the European level. However, the authors emphasized that further development and consideration of additional operational constraints are necessary for practical application.

In [4], the problem of flight level allocation was modeled as a factorized optimization problem that simultaneously addresses two conflicting objectives: Reducing fuel consumption (airlines' interest) and maintaining air traffic safety (public interest). In this study, using threshold-based spectral graph coloring, each flight path was represented as a vertex and each path intersection as an edge, while colors denoted flight levels (FL10 to FL400). The interference matrix defined the safety cost based on the distance between flight levels, such that separations below the safety threshold imposed infinite interference, while larger separations incurred decreasing interference values. The fuel consumption model, defined for Airbus A320 data, included both the cost of climbing to the optimal altitude and the cruise cost as a function of distance and specific range. To solve the problem, a hybrid approach combining Simulated annealing (SA) and Hill Climbing (HC) algorithms was applied: SA allowed controlled acceptance of worse solutions to escape local optima, while HC accepted only better solutions. Results indicate that enforcing safety constraints leads to a 5% increase in fuel consumption but significantly improves safety separation between flights. The authors suggested that in the future, the model could be extended to real-time operations with dynamic positioning for managing both aircraft and UAV traffic, as well as applied to areas such as route-sharing optimization in urban transportation.

Taken together, these two studies demonstrate that graph-coloring-based approaches, particularly those using advanced or hybrid algorithms, can serve as a foundation for the design of modern air traffic management systems, while also highlighting the need for adaptability to each country's operational constraints and conditions.

A review of prior studies shows that although graph-coloring approaches, whether in the form of static models focused on eliminating flight conflicts or multiobjective models balancing safety and fuel efficiency, have yielded valuable results, several fundamental limitations remain:

1. Dependence on European airspace data and structure: Existing models are largely based on European flight data and route structures (notably France and Spain), which limits their direct applicability in other countries due to differences in route design, operational restrictions, and flight regulations.
2. Incomplete alignment with Iran's operational conditions: Iran's airspace, due to its geopolitical location and the high volume of international overflights, has features not considered in prior models, including differences in overflight density, temporal flight patterns, and specific security considerations.

3. Limited scalability under variable traffic conditions: Many existing approaches either function statically or lose efficiency in high-traffic scenarios. Iran's air traffic management requires a system capable of maintaining optimal flight-level allocation even under dynamic, real-time traffic fluctuations.
4. Incomplete utilization of real-time data: Although some studies mention real-time extensions, few operational models have been developed that can be deployed in control centers using live data with continuously updated flight graphs.

In light of these gaps, the present study aims to design and implement a multi-objective graph-coloring algorithm tailored to the characteristics and constraints of Iranian airspace, capable of operating under dynamic traffic conditions with real-time data. By using real radar control station data, this approach seeks not only to ensure flight safety but also to maximize the efficiency of air routes, serving as a foundation for the development of intelligent air traffic management systems in Iran.

3 Research methodology

In this study, the ATFM problem is modeled as a *graph coloring* problem. In this framework, the airspace or flight network is represented as a graph $G = (V, E)$, which is defined as follows:

- Each vertex $v \in V$ represents a flight, an airport, a way point, or a specific sector of airspace.
- Each edge $e \in E$ denotes a potential conflict between two vertices; for example, two flights entering the same airspace sector at the same time or altitude.

In the coloring process, each vertex is assigned a color corresponding to an operational resource (e.g., flight level, time slot, or route), such that no two adjacent vertices (conflicting flights) share the same color.

The application of graph coloring to ATFM leads to the optimal allocation of resources, preventing congestion and ensuring safety. Three key types of allocation problems can be identified:

1. **Time Slot Assignment:** conflicting flights are assigned to different time windows to avoid runway or sector congestion.
2. **Flight Level Assignment:** intersecting routes are assigned to different flight levels to maintain the minimum vertical separation.

3. **Route Assignment:** alternative routes are selected to reduce interference in high-density sectors.

3.1 Illustrative example

Example 1. Suppose that ten flights must pass through a congested sector within one hour, while the sector's capacity is only three flights every 15 minutes.

- Vertices: flights
- Edges: pairs of flights that are in the sector within the same 15-minute interval
- Colors: time slots (e.g., Color A: 0–15 minutes, Color B: 16–30 minutes, etc.)

By solving the graph-coloring problem, flights are assigned to different time slots such that no conflicting flights share the same color, thus reducing congestion.

3.2 Computational challenges

Several computational challenges are inherent to this problem:

- Due to its NP-Hard nature, heuristic and metaheuristic algorithms are required for large-scale networks.
- Unexpected changes caused by weather conditions or delays demand real-time recoloring (dynamic conditions).
- Simultaneous consideration of time, altitude, and route increases problem complexity (multi-resource constraints).

Air traffic control centers are responsible for ensuring safe aircraft separation. Ideally, all flights would travel on direct routes. However, if all flights used direct horizontal routes only, human control would become infeasible due to excessive intersections. A practical solution is to apply *vertical separation* (flight levels) to routes with horizontal intersections: Flights with the same origin-destination pair share the same direct path but use different flight levels to avoid conflicts. Thus, only flights with overlapping schedules require vertical separation.

Inspired by [1], this study adopts the concept of *flows*, where all flights with the same origin-destination are grouped into a single flow. The size of each flow (n_i) equals the number of aircraft within that flow. For each flow, the start and end times (based on the first and last flight) are calculated. Hence, for a given day, the model input is defined by the set of flows.

If the problem size becomes large, to control computational complexity, then only flows with more than a minimum threshold $n_{\min}$ can be considered. For instance, commercial scheduled flights usually form large flows that require higher flight levels (to save fuel), whereas private flights (Instrument flight rule (IFR)) typically form small flows operating at lower levels with fewer conflicts. Therefore, the model can be solved only for large flows (e.g., more than 5–10 flights), while small flows can be assigned by default below the main flight levels. In this research, however, we consider all flows ($n_{\min} = 1$).

If two flows intersect spatially and their time intervals overlap, then they cannot be assigned to the same flight level. Formally

$$\text{for all } i, j \quad (1 \leq i < j), \quad L_i \neq L_j \quad \text{if} \quad \text{intersect}(r_i, r_j) \wedge \text{overlap}(t_i, t_j). \quad (1)$$

3.3 Experimental basis

The experimental basis of this research consists of real-world data from 52 flight flows at Tehran Mehrabad International Airport and Mashhad Hasheminejhad International Airport during weekly peak hours (Table 1). These data include detailed information on entry and exit times in controlled sectors, the size of each flow, and traversed routes, obtained from the national air traffic control center.

3.4 Problem modeling of ATFM with graphs

In this study, ATFM during the peak traffic period at Tehran Mehrabad and Mashhad Shahid Hasheminejhad Airports is formulated as a graph coloring problem. The considered airspace is divided into 11 flight corridors, each comprising several flows (flights with specific origin/destination pairs). Each vertex of the graph represents a flow, while each edge denotes the existence of a temporal or route conflict between two flows within the same corridor. Each corridor is thus modeled as an independent clique (a complete subgraph), since the flows within a corridor

Table 1: Flight flows at Tehran Mehrabad and Mashhad Shahi Hasheminejhad International Airport during peak hours.

Start	End	Corridor	Origin-Destination	n_i	Flow
0500	0700	1	Tehran-Bandar Abbas	5	f_1
0500	0700	1	Tehran-Zahedan	3	f_2
0500	0700	1	Tehran-Zabol	1	f_3
0500	0700	1	Tehran-Iranshahr	1	f_4
0500	0700	1	Tehran-Yazd	1	f_5
0500	0700	1	Tehran-Qeshm	2	f_6
0500	0700	1	Tehran-Kerman	2	f_7
0500	0700	1	Tehran-Chabahar	2	f_8
0500	0700	1	Tehran-Sirjan	1	f_9
0500	0700	2	Tehran-Asaluyeh	6	f_{10}
0500	0700	2	Tehran-Shiraz	5	f_{11}
0500	0700	2	Tehran-Kish	5	f_{12}
0500	0700	2	Tehran-Isfahan	2	f_{13}
0500	0700	2	Tehran-Bushehr	3	f_{14}
0500	0700	2	Tehran-Dubai	1	f_{15}
0500	0700	3	Tehran-Ahvaz	2	f_{16}
0500	0700	3	Tehran-Abadan	1	f_{17}
0500	0700	3	Tehran-Muscat	1	f_{18}
0500	0700	4	Tehran-Tabriz	5	f_{19}
0500	0700	4	Tehran-Urmia	1	f_{20}
0500	0700	5	Tehran-Ardabil	2	f_{21}
0500	0700	5	Tehran-Rasht	1	f_{22}
0500	0700	6	Tehran-Mashhad	11	f_{23}
0500	0700	7	Tehran-Kermanshah	2	f_{24}
0500	0700	7	Tehran-Ilam	1	f_{25}
0500	0700	7	Tehran-Najaf	3	f_{26}
0500	0700	7	Tehran-Baghdad	2	f_{27}
0500	0700	8	Mashhad-Ahvaz	1	f_{28}
0500	0700	8	Mashhad-Abadan	1	f_{29}
0500	0700	8	Mashhad-Yazd	1	f_{30}
0500	0700	8	Mashhad-Shiraz	1	f_{31}
0500	0700	8	Mashhad-Bushehr	1	f_{32}
0500	0700	8	Mashhad-Isfahan	1	f_{33}
0500	0700	9	Mashhad-Bandar Abbas	2	f_{34}
0500	0700	9	Mashhad-Kerman	1	f_{35}
0500	0700	9	Mashhad-Kish	2	f_{36}
0500	0700	9	Mashhad-Asaluyeh	1	f_{37}
0500	0700	9	Mashhad-Qeshm	1	f_{38}
0500	0700	9	Mashhad-Dubai	2	f_{39}
0500	0700	10	Mashhad-Zahedan	1	f_{40}
0500	0700	10	Mashhad-Zabol	1	f_{41}
0500	0700	10	Mashhad-Chabahar	2	f_{42}
0500	0700	10	Mashhad-Iranshahr	1	f_{43}
0500	0700	11	Mashhad-Ardabil	1	f_{44}
0500	0700	11	Mashhad-Urmia	1	f_{45}
0500	0700	11	Mashhad-Tabriz	2	f_{46}
0500	0700	11	Mashhad-Kermanshah	1	f_{47}
0500	0700	11	Mashhad-Ilam	1	f_{48}
0500	0700	11	Mashhad-Rasht	2	f_{49}
0500	0700	11	Mashhad-Tehran	6	f_{50}
0500	0700	11	Mashhad-Baghdad	2	f_{51}
0500	0700	11	Mashhad-Najaf	3	f_{52}

are mutually conflicting, while flows belonging to different corridors do not interfere with one another. Figure 1 shows the graph constructed from the data presented in Table 1.

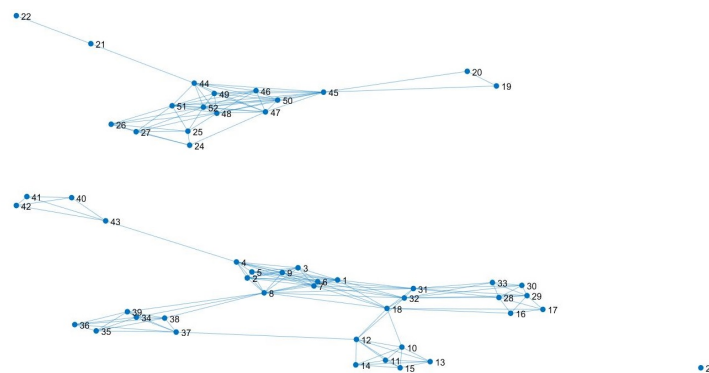


Figure 1: Graph corresponding to the data in Table 1.

The initial model, based on relation (1), includes only local constraints between each pair of flows. However, in graphs resulting from intersecting routes, large complete subgraphs (cliques) are frequently observed. In such cases, instead of formulating multiple pairwise constraints, one can impose an all-different constraint over the clique members, which is both simpler and more effective in detecting inconsistencies. Since finding the maximum clique or all maximal cliques is an NP-hard problem, greedy or approximate algorithms are typically employed. The algorithm adopted in this work (inspired by Bomze et al. [2]) starts from each node and incrementally builds a clique by selecting, at each step, the node with the largest neighborhood intersection. This procedure generates only maximal cliques, after which duplicate cliques or trivial ones (of size 2) are eliminated. The theoretical complexity of the algorithm is $O(n^4)$ in the worst case (for complete graphs). Nevertheless, in larger real-world instances, the graphs are sparse and the effective complexity growth is significantly lower [2].

4 Proposed algorithm

The proposed algorithm, employing a combination of greedy strategies for initial assignment and CP for final optimization, is designed to achieve resource allocation at an operational scale using real-world data.

4.1 Initial solution generation

In this paper, the DSatur algorithm is employed to assign flight levels to the flights listed in Table 1. DSatur is a graph coloring algorithm introduced by Daniel Brélaz in 1979 [3]. Similar to greedy coloring algorithms, DSatur colors the vertices of a graph one by one, adding a new color whenever necessary. After coloring a new vertex, the algorithm identifies the uncolored vertex with the largest number of distinct colors already assigned to its neighbors, and this vertex is selected for the next step. Brélaz [3] defined this number as the saturation degree of a vertex, from which the algorithm derives its name.

Although DSatur is a heuristic graph coloring algorithm, it provides exact results for bipartite, cyclic, and wheel graphs [3]. It is also known in the literature as LF saturation [5].

To generate a quick initial solution, the algorithm proceeds as follows:

1. Select the vertex with the highest saturation degree (the largest number of distinct colors among its neighbors).
2. In case of a tie, choose the vertex with the highest degree (the most incident edges).
3. Assign the smallest feasible color not used by its neighbors.
4. Update the saturation degrees of the uncolored neighbors.
5. Repeat until the graph is fully colored.

4.2 CP Model

In the proposed approach, the output of DSatur is used as an upper bound for the number of colors in a CP model. Then

- Large cliques in the graph are identified.
- Global coloring constraints are imposed on these cliques.
- The objective is to minimize the total number of colors (flight levels).

Algorithm 1 Clique-based graph coloring with all different constraints**Require:** F (set of flows),

- 1: For each flow $i \in F$: route r_i , time window t_i ,
- 2: $n_{\min}$ (minimum flow size to enter the model),
- 3: $LEVELS = \{1, \dots, N\}$ (initial domain of colors (flight levels)).

Ensure: Mapping $L : F \rightarrow \mathbb{N}$ minimizing number of distinct colors.

- 4: $F \leftarrow \{i \in F \mid n_i \geq n_{\min}\}$ (remove small flows).
- 5: $N \leftarrow |F|$ (upper bound on colors).
- 6: $V \leftarrow F$ (one vertex per flow).
- 7: **for** All $i, j \in F$, $i < j$ **do**
- 8: **if** $\text{intersect}(r_i, r_j)$ **and** $\text{overlap}(t_i, t_j)$ **then**
- 9: add edge (i, j) to E
- 10: **end if**
- 11: **end for**
- 12: Detect maximal cliques by Greedy algorithm ($\mathcal{C}$).
- 13: Remove from $\mathcal{C}$ any clique with $|C| < 3$ or duplicates.
- 14: $k_{\max} \leftarrow \max_{C \in \mathcal{C}} |C|$ (lower bound).
- 15: **for** All $i \in V$ **do**
- 16: $\text{Domain}(i) \leftarrow \{1, \dots, N\}$.
- 17: **end for**
- 18: **for** All $C \in \mathcal{C}$ **do**
- 19: impose $\text{ALLDIFFERENT}(\{L_i : i \in C\})$.
- 20: **end for**
- 21: **for** All $(i, j) \in E$ not covered by a clique constraint **do**
- 22: impose $L_i \neq L_j$.
- 23: **end for**
- 24: Sort the cliques in descending order of size.
- 25: **if** there is a remaining clique **then**
- 26: A variable from that clique is selected, otherwise a variable with the smallest domain is selected.
- 27: **if** No colors in the variable's domain have been used previously **then**
- 28: the smallest unused color is assigned deterministically, otherwise the algorithm proceeds with branching.
- 29: **end if**
- 30: **end if**

4.3 Solution procedure

Considering the explanations in the previous sections, the steps for solving the problem are as follows:

1. Identify flows and corridors from the real-world data (Table 1).
2. Construct the graph based on the interference matrix.

3. Detect cliques (each corridor forms an independent clique).
4. Execute DSatur to obtain an initial estimate of the number of colors.
5. Apply CP combined with clique constraints to optimize and reduce the number of colors.
6. Assign the final colors (Table 2).
7. Validate the results against operational constraints.

The “Clique-based graph coloring with all different constraints” algorithm was implemented in MATLAB 2020 and applied to the data presented in Table 1. The corresponding output is shown in Table 2. It can be observed that, for the flows in Table 1, five flight levels are sufficient to ensure vertical separation and maintain flight safety during the peak hour under consideration.

Table 2: Flow color assignment after executing the algorithm

Corridor	Flow	Flow Color	Corridor	Flow	Flow Color
1	f_1	1	7	f_{27}	4
1	f_2	2	8	f_{28}	3
1	f_3	3	8	f_{29}	4
1	f_4	4	8	f_{30}	1
1	f_5	5	8	f_{31}	5
1	f_6	6	8	f_{32}	9
1	f_7	7	8	f_{33}	2
1	f_8	8	9	f_{34}	3
1	f_9	9	9	f_{35}	1
2	f_{10}	1	9	f_{36}	2
2	f_{11}	2	9	f_{37}	4
2	f_{12}	3	9	f_{38}	5
2	f_{13}	4	9	f_{39}	6
2	f_{14}	5	10	f_{40}	1
2	f_{15}	6	10	f_{41}	2
3	f_{16}	1	10	f_{42}	3
3	f_{17}	2	10	f_{43}	5
3	f_{18}	4	11	f_{44}	2
4	f_{19}	1	11	f_{45}	3
4	f_{20}	2	11	f_{46}	1
5	f_{21}	1	11	f_{47}	4
5	f_{22}	2	11	f_{48}	5
6	f_{23}	1	11	f_{49}	6
7	f_{24}	1	11	f_{50}	7
7	f_{25}	2	11	f_{51}	8
7	f_{26}	3	11	f_{52}	9

4.4 Results and analysis

To evaluate the performance of the proposed algorithm compared to DSatur, a simple flight network scenario was designed as Example 2.

Example 2. Consider a scenario with eight flight routes that intersect at several points. Each route is modeled as a vertex in the graph, and each intersection between two routes is represented as an edge. In total, this network includes eight flight routes and 24 intersections. The corresponding adjacency matrix is given by

$$A = \begin{bmatrix} 0 & 1 & 0 & 1 & 0 & 1 & 0 & 0 \\ 1 & 0 & 0 & 0 & 1 & 1 & 1 & 0 \\ 0 & 0 & 0 & 1 & 0 & 0 & 1 & 1 \\ 1 & 0 & 1 & 0 & 1 & 0 & 0 & 1 \\ 0 & 1 & 0 & 1 & 0 & 1 & 0 & 0 \\ 1 & 1 & 0 & 0 & 1 & 0 & 0 & 1 \\ 0 & 1 & 1 & 0 & 0 & 0 & 0 & 1 \\ 0 & 0 & 1 & 1 & 0 & 1 & 1 & 0 \end{bmatrix}.$$

The graph corresponding to this hypothetical network is illustrated in Figure 2.

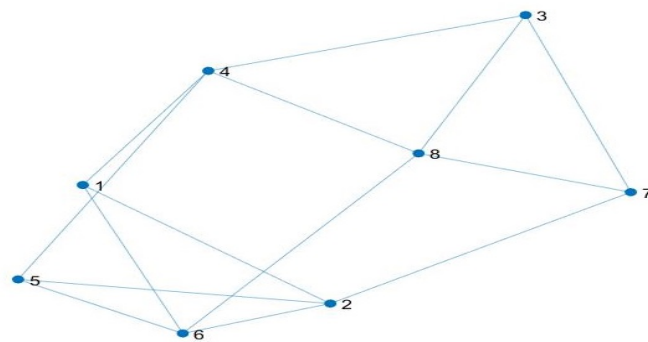


Figure 2: Graph corresponding to the hypothetical flight network in Example 2.

The objective is to assign the minimum number of distinct flight levels (colors) to the routes, such that no two intersecting routes share the same flight level.

Our proposed algorithm (based on complete search techniques combined with maximal clique guidance) is able to solve the problem optimally using only three flight levels. In other words, three flight levels are sufficient to eliminate all conflicts among the routes.

In contrast, applying the DSatur algorithm to the same dataset—although significantly faster—resulted in the use of four flight levels. This implies that DSatur, in this case, reached a suboptimal solution by introducing an unnecessary additional flight level into the network.

This example clearly demonstrates that in real-world applications such as air traffic network management, relying solely on greedy algorithms like DSatur can lead to resource waste (flight levels). In contrast, our proposed algorithm, by guaranteeing optimality, is able to provide a more efficient allocation of flight levels.

The analysis of results shows that the proposed CP + Clique approach achieved a reduction in the number of required flight levels compared to DSatur. This reduction can lead to

- more efficient utilization of airspace capacity,
- reduction of delays,
- fuel savings and lower operational costs,
- improved flight safety.

Moreover, modeling each corridor as a complete graph (due to the absence of inter-corridor conflicts) simplified the model and increased the accuracy of the coloring allocation.

5 Conclusion and recommendations

To the best of our knowledge, this research represents the first scientific study in Iran that directly employs real operational data from two major Airports (Mehrabad and Hasheminejhad) during peak traffic hours to model and solve the ATFM problem using graph coloring algorithms.

Based on the collected data, it was observed that, at present, route allocation for maintaining flight safety and handling high-demand flights within the same time window is performed manually. This manual approach has limitations; for example, the Tehran–Mashhad and Mashhad–Tehran flights are treated differently, and similar cases exist that require manual adjustments.

The proposed methods in this study can overcome such limitations and enable more efficient utilization of alternative routes. The released capacity can then be used for other purposes such as humanitarian services, military operations, political programs, and other special applications.

Strengths

- Utilization of real-world data,
- Potential generalization to other major and secondary airports in the country,
- Optimal combination of greedy algorithms with constraint-based optimization methods.

Weaknesses

- Limitations due to security and military restrictions in accessing full airspace data,
- Lack of weather simulation and dynamic flight conditions in the current model.

Recommendations

- Development of an intelligent ATFM software tool for real-time allocation of flight levels,
- Collaboration with security authorities to optimize restricted airspace zones,
- Incorporating the effects of weather conditions and dynamic modeling.
- Considering the constraints on the number of available flight levels, future research should focus on developing methods to reduce the required number of levels through strategies such as
 1. Flight time adjustment (rescheduling, delaying, or advancing),
 2. Route modification (re-routing),
 3. Increasing resources/capacity (e.g., adding runways or allowing higher concurrency subject to safety separation),
 4. Vertex splitting or cost-based decomposition (e.g., sequestering a service or dividing a flight into smaller time slots).

Figure 3 presents the potential future research directions for optimizing flight level allocation.

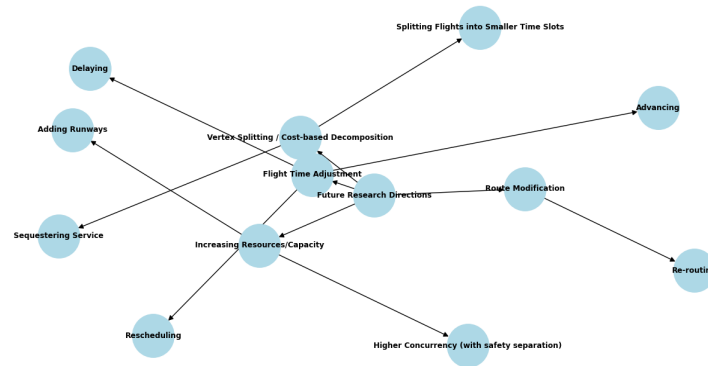


Figure 3: Suggested future research directions for optimizing flight level allocation

Acknowledgements

The research of the second author, Ahmad Erfanian, was supported in part by the Ferdowsi University of Mashhad. The research of the third author, Narjes Sabeghi, was supported in part by the Velayat University.

Funding

This research received no external funding.

Conflicts of Interest

The authors declare no conflicts of interest.

References

- [1] Barnier, N., and Brisset, P. *Graph coloring for air traffic flow management*, Ann. Oper. Res., 130(1) (2004), 163–178.
- [2] Bomze, I.M., Budinich, M., Pardalos, P.M., and Pelillo, M. *The maximum clique problem*, In Handbook of Combinatorial Optimization: Supplement Volume A (pp. 1–74). Boston, MA: Springer US, 1999.

- [3] Brélaz, D. *New methods to color the vertices of a graph*, Commun. ACM., 22(4) (1979), 251–256.
- [4] Gimenez-Guzman, J.M., Martínez-Moraian, A., Reyes-Bardales, R.D., Orden, D., and Marsa-Maestre, I. *Flight level assignment using graph coloring*, Appl. Sci. 10(18) (2020), 6157.
- [5] Yekezare, N., Zohrehbandian, M., Maghasedi, M., and Bonomo-Braberman, F. *Optimality of DSatur algorithm on chordal graphs*, Oper. Res. Lett., 57 (2024), 107185.



Fitted-mesh cubic spline methods for layer-dominated singularly perturbed parabolic problems with nonsmooth data

S.L. Cheru*, , G.F. Duressa  and T.B. Mekonnen 

Abstract

This paper introduces a numerical framework for addressing a two-parameter singularly perturbed parabolic problems characterized by discontinuities in both the convection coefficients and the source terms. The presence of these discontinuities alongside small perturbation parameters creates distinct boundary and interior layers, which present considerable obstacles to obtaining precise numerical approximations. To address these difficulties, a layer-adapted(Shishkin-type) spatial mesh is constructed a priori, providing enhanced resolution in regions where steep solution gradients occur. The suggested method uses the implicit Euler scheme for time integration and a cubic spline in tension for spatial discretization on a

*Corresponding author

Received 21 November 2025; revised 27 January 2026; accepted 29 January 2026

Shegaye Lema Cheru

Department of Mathematics, College of Natural and Computational Sciences, Wollega University, Ethiopia.
e-mail: shegayel@wollegauniversity.edu.et

Gemechis File Duressa

Department of Mathematics, College of Natural and Computational Sciences, Jimma University, Ethiopia.
e-mail: gammeef@gmail.com

Tariku Birabasa Mekonnen

Department of Mathematics, College of Natural and Computational Sciences, Wollega University, Ethiopia.
e-mail: seenaa29@gmail.com

How to cite this article

Cheru, S.L., Duressa, G.F. and Mekonnen, T.B., Fitted-mesh cubic spline methods for layer-dominated singularly perturbed parabolic problems with nonsmooth data. *Iran. J. Numer. Anal. Optim.*, 2026; 16(2): 607-637. <https://doi.org/10.22067/ijnao.2026.96617.1784>

Shishkin-type mesh. A thorough analysis shows that the method achieves first-order in time and second-order up to a logarithmic factor in space when using the maximum norm, regardless of how big the perturbation parameters are. The method is especially good for modeling real-world processes like heat and mass transfer in layered media, pollutant dispersion in mixed environments, and viscous flow through composite channels, where there are often discontinuities and strong layer behavior.

AMS subject classifications (2020): 35G05, 35K20, 35K51, 65M06, 65M50.

Keywords: Singular perturbations; Two parameters; Discontinuous data; Uniform convergence; Boundary layer; Interior layer

1 Introduction

In fluid dynamics, getting the boundary and interior layer phenomena (see [14, 24]) is still a big problem for numerical modeling of complicated transport processes. These layers, usually caused by strong convection, diffusion imbalance, or reactive effects, have a big effect on how accurate and stable computational results are in real-world situations. In this research, we conduct a comprehensive numerical analysis of two-parameter singularly perturbed parabolic problems (TP-SPPPs) that represent these complex behaviors. The main goal is to build a strong and parameter uniform numerical framework that can accurately resolve sharp gradients and stay stable across the whole space-time domain. We now recall the following TP-SPPP:

$$\left\{ \begin{array}{l} \mathcal{L}_{\varepsilon, \theta} w(x, t) := \left(\varepsilon \frac{\partial^2 w}{\partial x^2} + \theta a \frac{\partial w}{\partial x} - bw - c \frac{\partial w}{\partial t} = f \right) (x, t), (x, t) \in \mathfrak{D}^- \cup \mathfrak{D}^+, \\ w(0, t) = \mathcal{B}(t) \quad \text{in } \mathfrak{D}_l, \quad w(1, t) = \mathcal{R}(t) \quad \text{in } \mathfrak{D}_r, \\ w(x, 0) = \mathcal{I}(x) \quad \text{in } \mathfrak{D}_b, \\ |[a](s)| \leq C, \quad |[f](s)| \leq C, \\ a(x, t) \leq -\theta_1 < 0, (x, t) \in \mathfrak{D}^-, a(x, t) \geq \theta_2 > 0, (x, t) \in \mathfrak{D}^+, \\ b(x, t) \geq \gamma > 0, \quad c(x, t) \geq c > 0, \end{array} \right. \quad (1)$$

where $0 < \varepsilon \ll 1$ and $0 \leq \theta \leq 1$ are perturbation parameters, with domains $\mathfrak{D} = \Omega \times \Lambda$, $\mathfrak{D}^\pm = \Omega^\pm \times \Lambda$, and $\mathbb{M} = \overline{\mathfrak{D}} \setminus \mathfrak{D}$; here $\Omega = (0, 1)$, $\Omega^- = (0, s)$, $\Omega^+ = (s, 1)$. The boundaries are $\mathfrak{D}_l = \{(0, t) : 0 \leq t \leq T\}$, $\mathfrak{D}_r = \{(1, t) : 0 \leq t \leq T\}$, and $\mathfrak{D}_b = \mathfrak{D}_b^- \cup \mathfrak{D}_b^+$, where $\mathfrak{D}_b^- = [0, s] \times \{t = 0\}$, $\mathfrak{D}_b^+ = [s, 1] \times \{t = 0\}$. Moreover, $\mathbb{M} = \mathfrak{D}_b \cup \mathfrak{D}_l \cup \mathfrak{D}_r$ and $\mathbb{M}^\pm = \{(s, t) : 0 \leq t \leq T\}$.

The coefficients $a(x, t)$ and $f(x, t)$ are assumed to be smooth in $(\mathfrak{D}^- \cup \mathfrak{D}^+) \cup \mathfrak{D}$, whereas $b(x, t)$ and $c(x, t)$ remain continuous in $\mathfrak{D}$. The inequalities $||[a](s)|| \leq C$ and $||[f](s)|| \leq C$ in (1) are used to bound the magnitude of the jump discontinuities of $a(x, t)$ and $f(x, t)$ at interior discontinuity point s . At the interface $x = s \in \Omega$, both $a(x, t)$ and $f(x, t)$ possess a jump discontinuity, denoted by $[\psi](s, t) = \psi(s^+, t) - \psi(s^-, t)$, where $\psi(s^\pm, t) = \lim_{x \rightarrow s^\pm 0} \psi(x, t)$. A valid solution persists even if $a(x, t)$ changes sign across the interface, indicating opposite convection directions on each side. The solution to (1) exhibits boundary layers caused by ε, θ and an interior layer around $x = s$ resulting from discontinuities in $a(x, t)$ and $f(x, t)$. The structure and thickness of these layers depend on the ratio between the diffusion parameter and the square of the convection coefficient, as well as the sign pattern of $a(x, t)$. As a result, the thickness of the boundary layers changes, and symmetric interior layers form near $x = s$. This kind of behavior often happens in systems with convection-dominated transport and reactive flow, which shows how important it is to use parameter-uniform numerical schemes [25, 22].

The problem admits two notable limiting cases: The reaction-diffusion case ($\theta = 0$) and the convection-diffusion case ($\theta = 1$), both of which can emerge as particular forms of a general TP-SPPP model. When $\theta = 0$, boundary layers develop symmetrically on both ends of the domain, each with an approximate thickness of $O(\sqrt{\varepsilon})$. In contrast, when $\theta = 1$, the boundary layer forms near either the left or right boundary, depending primarily on the sign of the convection coefficient, and has a characteristic width of $O(\varepsilon)$. If the problem involves a discontinuous source term while $a(x, t) > 0$ (or $a(x, t) < 0$), then the solution exhibits weak interior layers for any small ε . However, when the convection coefficient itself is discontinuous such as $a(x, t) < 0$ for $x < s$ and $a(x, t) > 0$ for $x > s$, strong interior layers appear within the domain. A detailed analysis of this particular case ($\theta = 1$) can be found in [19], where the sign change of the convection coefficient is shown to shift the location of the layers within the solution.

Several researchers have contributed to the numerical analysis of TP-SPPPs, both with and without delay, employing a variety of numerical approaches (see, for instance, [9, 7, 10, 13, 16, 18, 5]). Despite these advancements, the numerical treatment of TP-SPPPs involving discontinuous convection coefficients and source terms remains at a relatively early stage of development. Chandru, Das, and Ramos [1] proposed a nearly first-order uniformly convergent scheme for (1), effectively addressing small diffusion and convection parameters with discontinuities in both convection and source terms, by combining an implicit Euler time discretization with an upwind spatial scheme. Further improvement was made by Chandru et al. [2], who designed a nearly first-order parameter-uniform scheme incorporating an adaptive mesh to enhance spatial resolution near layer regions. Kumar and Kumari [15] developed a related approach using the Crank–Nicolson method for temporal discretization on a uniform mesh and an upwind scheme on a Shishkin mesh for spatial discretization. More recently, Singh, Choudhary, and Kumar

[26] introduced a spline-based numerical method for TP-SPPPs with discontinuous convection and source terms, employing the Crank–Nicolson scheme in time coupled with trigonometric spline approximations in space. The results of these investigations overall illustrate continuous advancement in the development of resilient, parameter-consistent algorithms proficient in effectively addressing boundary and interior layers in scenarios involving discontinuous data.

In real-world scenarios like heat and mass transfer in composite materials, groundwater flow through stratified media, and pollutant transport with sudden source variations, TP-SPPPs with discontinuous convection and source terms frequently occur. Strong boundary and interior layers are induced by these discontinuities, making precise computation and analysis difficult. In order to solve this, a hybrid numerical framework is created that combines a cubic spline approximation in space over a Shishkin mesh with the implicit Euler method in time. Even in the face of extreme discontinuities, stability, accuracy, and parameter-uniform convergence are guaranteed by the adaptive mesh’s ability to capture steep gradients [23, 2, 19, 6].

Fitted mesh techniques for TP-SPPPs with nonsmooth or discontinuous data are still rarely studied, despite their practical importance. The majority of current research ignores strongly perturbed cases in favor of smooth or piecewise continuous coefficients. The suggested scheme uses a Shishkin-type uniform mesh in conjunction with a cubic spline in tension, which provides high accuracy and robustness, in contrast to traditional layer-adapted approaches that depend on intricate mesh adjustments. This approach fills a significant gap in the existing numerical literature by effectively resolving sharp boundary and interior layers while preserving computational simplicity.

2 Properties of the continuous solution

This section elaborates on the derivative bounds of (1), which are essential for the ensuing numerical analysis. It is assumed in this work that the solution of (1) is uniformly bounded with respect to the perturbation parameter. Farrell *et al.* [11] previously studied the parabolic case corresponding to $\theta = 1$, while O’Riordan and Shishkin [21] studied the steady-state formulation with $\theta = 0$. We now introduce the following lemma, which, under the given assumptions, ensures the existence of a solution to (1):

Lemma 1 (Existence of solution). The solution $w(x, t)$ of (1) satisfies $w \in C^0(\overline{\mathfrak{D}}) \cap C^1(\mathfrak{D}) \cap C^2(\mathfrak{D}^- \cup \mathfrak{D}^+)$.

Proof. Let w_1 and w_2 denote the solutions of SPPs [15, 26, 3, 1]:

$$\begin{aligned}\varepsilon w_{1xx} + \theta a_1 w_{1x} - b w_1 - c w_{1t} &= f, & (x, t) \in \mathfrak{D}^-, \\ \varepsilon w_{2xx} + \theta a_2 w_{2x} - b w_2 - c w_{2t} &= f, & (x, t) \in \mathfrak{D}^+, \end{aligned}$$

where $a_1, a_2 \in C^2(\mathfrak{D})$ satisfy $a_1(x, t) < 0$ in $\mathfrak{D}^-$ and $a_2(x, t) > 0$ in $\mathfrak{D}^+$. The composite function

$$w(x, t) = \begin{cases} w_1(x, t) + \Lambda_1 \Xi_2(x, t), & (x, t) \in \mathfrak{D}^-, \\ w_2(x, t) + \Lambda_2 \Xi_1(x, t), & (x, t) \in \mathfrak{D}^+, \end{cases}$$

satisfies (1), where Ξ_1, Ξ_2 are solutions of homogeneous auxiliary problems associated with (1). The constants Λ_1 and Λ_2 are obtained by enforcing continuity of w and w_x at $x = s$:

$$w(s^-, t) = w(s^+, t), \quad w_x(s^-, t) = w_x(s^+, t).$$

These yield a linear system whose coefficient matrix has a nonzero determinant, due to the monotonic nature of Ξ_1 and Ξ_2 . Hence, unique constants Λ_1, Λ_2 exist, ensuring that $w(x, t)$ and $w_x(x, t)$ remain continuous across the interface $x = s$. $\square$

Lemma 2 (Continuity at the interface). At the discontinuity point $x = s$, the solution $w(x, t)$ and its spatial derivative are continuous, that is,

$$w(s^-, t) = w(s^+, t), \quad w'(s^-, t) = w'(s^+, t).$$

Proof. A detailed proof is provided in [6]. $\square$

The operator $\mathcal{L}_{\varepsilon, \theta}$ of (1) satisfies maximum principle on $\overline{\mathfrak{D}}$ as follows.

Lemma 3 (Comparison Principle). Let $w \in C^0(\overline{\mathfrak{D}}) \cap C^1(\mathfrak{D}) \cap C^2(\mathfrak{D}^- \cup \mathfrak{D}^+)$ satisfy

$$w(0, t) \geq 0, \quad w(1, t) \geq 0, \quad w(x, 0) \geq 0, \quad [w'](s, t) \geq 0,$$

and $\mathcal{L}_{\varepsilon, \theta} w(x, t) \geq 0$ for all $(x, t) \in \mathfrak{D}$. Then $w(x, t) \leq 0$ for every $(x, t) \in \overline{\mathfrak{D}}$.

Proof. Define an auxiliary function

$$w(x, t) = e^{-\theta \theta |x-s|/(2\varepsilon)} \mathcal{V}(x, t), \quad \theta = \min\{\theta_1, \theta_2\}.$$

Substituting this into (1) yields, for $x \in \mathfrak{D}^-$,

$$\mathcal{L}_{\varepsilon, \theta} w = e^{-\theta_1 \theta (s-x)/(2\varepsilon)} \left(\varepsilon \mathcal{V}_{xx} + \theta(\theta_1 + a_1) \mathcal{V}_x + \left[\frac{\theta_1 \theta^2}{2\varepsilon} \left(\frac{\theta_1}{2} + a_1 \right) - b \right] \mathcal{V} - c \mathcal{V}_t \right),$$

and for $x \in \mathfrak{D}^+$,

$$\mathcal{L}_{\varepsilon, \theta} w = e^{-\theta_2 \theta (x-s)/(2\varepsilon)} \left(\varepsilon \mathcal{V}_{xx} + \theta(a_2 - \theta_2) \mathcal{V}_x + \left[\frac{\theta_2 \theta^2}{2\varepsilon} \left(\frac{\theta_2}{2} - a_2 \right) - b \right] \mathcal{V} - c \mathcal{V}_t \right).$$

Let $(\chi, \tau) \in \overline{\mathfrak{D}}$ be the point where $\mathcal{V}$ attains its maximum. If $\mathcal{V}(\chi, \tau) \leq 0$, then the claim is immediate. Otherwise, assume $\mathcal{V}(\chi, \tau) > 0$. Applying the maximum principle, (χ, τ) must lie in $\mathfrak{D}^-$, $\mathfrak{D}^+$, or at $x = s$. For interior points, $\mathcal{V}_x = \mathcal{V}_t = 0$, $\mathcal{V}_{xx} < 0$, and since $a_1 \leq \theta_1 < 0$ and $a_2 \geq \theta_2 > 0$, it follows that $\mathcal{L}_{\varepsilon, \theta} w(\chi, \tau) < 0$, contradicting the assumption $\mathcal{L}_{\varepsilon, \theta} w \geq 0$. At $x = s$, the continuity of w and the conditions $[\mathcal{V}] = 0$ and $[\mathcal{V}_x] \leq 0$ imply $[w_x](s, \tau) \leq 0$, again a contradiction. Hence, $w(x, t) \leq 0$ in $\overline{\mathfrak{D}}$. $\square$

Lemma (3) immediately establishes the corresponding stability result, thereby ensuring the uniqueness of the solution.

Lemma 4. Let $w(x, t)$ be the solution of (1). Then

$$\|w\|_{\overline{\mathfrak{D}}} \leq \mathbb{C} \max \{ \|\mathcal{B}\|_{\mathfrak{D}_l}, \|\mathcal{R}\|_{\mathfrak{D}_r}, \|\mathcal{I}\|_{\mathfrak{D}_b} \} + \frac{1}{\mathcal{S}} \|\mathcal{L}_{\varepsilon, \theta} w\|_{(\mathfrak{D}^- \cup \mathfrak{D}^+)}, \quad (2)$$

where $\mathcal{S} = \min \left\{ \frac{\theta_1}{s}, \frac{\theta_2}{1-s} \right\}$.

Proof. The proof is derived by applying Lemma (3) in conjunction with appropriately constructed barrier functions (refer to [15, 26, 4, 6]). $\square$

Lemma 5. Let $\theta = \min \{ \theta_1, \theta_2 \}$ and $\mathfrak{B} = \min_{\mathfrak{D}^- \cup \mathfrak{D}^+} \left\{ \frac{b(x, t)}{a(x, t)} \right\}$. For $1 \leq \kappa + 2\iota \leq 3$. Then $w(x, t)$ of (1) and its derivatives satisfy the following conditions:

1. If $\theta\theta^2 \leq \mathfrak{B}\varepsilon$, then

$$\left\| \frac{\partial^{\kappa+\iota} w}{\partial x^\kappa \partial t^\iota} \right\|_{\mathfrak{D}^- \cup \mathfrak{D}^+} \leq \frac{C}{(\sqrt{\varepsilon})^\kappa} \max \left\{ \|w\|_{\overline{\mathfrak{D}}}, \sum_{i+2j=0}^2 (\sqrt{\varepsilon})^i \left\| \frac{\partial^{i+j} w}{\partial x^i \partial t^j} \right\|, \sum_{i=0}^4 \left[(\sqrt{\varepsilon})^i \left\| \frac{d^i \mathcal{I}}{dx^i} \right\|_{\mathfrak{D}_b} + \left\| \frac{d^i \mathcal{B}}{dt^i} \right\|_{\mathfrak{D}_l} + \left\| \frac{d^i \mathcal{R}}{dt^i} \right\|_{\mathfrak{D}_r} \right] \right\},$$

2. If $\theta\theta^2 \geq \mathfrak{B}\varepsilon$, then

$$\begin{aligned} \left\| \frac{\partial^{\kappa+\iota} w}{\partial x^\kappa \partial t^\iota} \right\|_{\mathfrak{D}^- \cup \mathfrak{D}^+} &\leq C \left(\frac{\theta}{\varepsilon} \right)^\kappa \left(\frac{\theta^2}{\varepsilon} \right)^\iota \max \left\{ \|w\|_{\overline{\mathfrak{D}}}, \right. \\ &\quad \sum_{i+2j=0}^2 \frac{\varepsilon^{i+j+1}}{\theta^{i+2j+2}} \left\| \frac{\partial^{i+j} w}{\partial x^i \partial t^j} \right\|, \\ &\quad \left. \sum_{i=0}^4 \left[\left(\frac{\varepsilon}{\theta} \right)^i \left\| \frac{d^i \mathcal{I}}{dx^i} \right\|_{\mathfrak{D}_b} + \left\| \frac{d^i \mathcal{B}}{dt^i} \right\|_{\mathfrak{D}_t} + \left\| \frac{d^i \mathcal{R}}{dt^i} \right\|_{\mathfrak{D}_r} \right] \right\}, \end{aligned}$$

where C depends only on the coefficients and their derivatives.

Proof. For a detailed proof, the reader is referred to [15, 3]. □

Lemma 6. The solution $w(x, t)$ of (1) satisfies the bounds for second order time derivative

$$\|w_{tt}\|_{\mathfrak{D}} \leq \begin{cases} C & \text{if } \theta\theta^2 \leq \mathfrak{B}\varepsilon, \\ C\theta^{-4}\varepsilon^{-2} & \text{if } \theta\theta^2 \geq \mathfrak{B}\varepsilon. \end{cases}$$

Proof. This result stems from Lemma (5) and the methods presented by Gracia, O’Riordan, and Pickett [12] and O’Riordan, Pickett, and Shishkin [20]. □

3 Numerical illustrations

This section begins by semi-discretizing the problem in time on a uniform mesh, followed by approximating the spatial variable using *cubic splines in compression* over a piecewise uniform mesh, providing a stable and accurate numerical scheme.

3.1 Time variable discretization

To discretize in time, a uniform mesh is introduced over the interval $[0, T]$ with step size Δt , dividing the interval into $M = T/\Delta t$ equal subintervals. The mesh is denoted by $\Omega^M = \{t_j = j\Delta t : j = 0, 1, \dots, M, \Delta t = \frac{T}{M}\}$. Time integration is performed using the *implicit Euler method*, yielding the semi-discrete form of problem (1) on Ω^M :

$$\begin{aligned}
& -c^{j+1}(x)D_t^-W^{j+1}(x) + \varepsilon (W^{j+1})_{xx}(x) + \theta a^{j+1}(x) (W^{j+1})_x(x) \\
& -b^{j+1}(x)W^{j+1}(x) = f^{j+1}(x), \quad x \in \Omega^s, \quad 0 \leq j \leq M-1, \\
& W^{j+1}(0) = \mathcal{B}(t_{j+1}), \quad W^{j+1}(1) = \mathcal{R}(t_{j+1}), \\
& W^0(x) = \mathcal{I}(x), \quad x \in \overline{\mathfrak{D}},
\end{aligned}$$

where $W^{j+1}(x)$ approximates $w(x, t_{j+1})$, and $D_t^-W^{j+1}(x) = \frac{W^{j+1}(x) - W^j(x)}{\Delta t}$ is the backward difference operator. Rewrite the above equation as

$$\begin{cases} \mathcal{L}_{\varepsilon, \theta}^M W^{j+1}(x) = \mathcal{G}^{j+1}(x), & x \in \Omega^s, \quad 0 \leq j \leq M-1, \\ W^{j+1}(0) = \mathcal{B}(t_{j+1}), \quad W^{j+1}(1) = \mathcal{R}(t_{j+1}), & 0 \leq j \leq M-1, \\ W^0(x) = \mathcal{I}(x), & x \in \overline{\mathfrak{D}}, \end{cases} \quad (3)$$

where the operator $\mathcal{L}_{\varepsilon, \theta}^M$ is defined as

$$\mathcal{L}_{\varepsilon, \theta}^M \equiv \varepsilon \frac{d^2}{dx^2} + \theta a^{j+1}(x) \frac{d}{dx} - \left(b^{j+1}(x) + \frac{c^{j+1}(x)}{\Delta t} \right) I, \quad \text{and}$$

$$\mathcal{G}^{j+1}(x) = f^{j+1}(x) - \frac{c^{j+1}(x)}{\Delta t} W^j(x).$$

The local truncation error (LTE) in the temporal semi-discretization of (3) is defined as $\mathcal{E}_{j+1} = W^{j+1}(x) - \overline{U}(x)$, where $\overline{U}(x)$ solves

$$\begin{cases} \mathcal{L}_{\varepsilon, \theta} \overline{U}(x) = \varpi(x, t_{j+1}) - \frac{c(x, t_{j+1})}{\Delta t} W^j(x), & x \in \Omega^s, \quad 0 \leq j \leq M-1, \\ \overline{U}(0) = \mathcal{B}(t_{j+1}), \quad \overline{U}(1) = \mathcal{R}(t_{j+1}), & 0 \leq j \leq M-1, \\ \overline{U}(x) = \mathcal{I}(x), & x \in \overline{\mathfrak{D}}. \end{cases}$$

The following lemma provides bounds for the LTE and the cumulative global error, following the approach in [15].

Lemma 7. Let $\mathcal{E}_{j+1}$ denote the LTE and let $\mathbb{E}_j = \sum_{\iota=1}^j \mathcal{E}_{\iota}$ be the global error up to time t_j . Then,

$$\|\mathcal{E}_{j+1}\| \leq C\Delta t^2, \quad \|\mathbb{E}_j\| \leq C\Delta t, \quad 0 \leq j \leq M, \quad (4)$$

where C is a positive constant independent of ε and θ .

Proof. The first inequality in (4) follows from [26, 3]. For the global error

$$\|\mathbb{E}_j\| = \left\| \sum_{\iota=1}^j \mathcal{E}_\iota \right\| \leq \sum_{\iota=1}^j \|\mathcal{E}_\iota\| \leq Cj\Delta t^2 = C\Delta t(j\Delta t) \leq C\Delta tT \leq C\Delta t,$$

showing first-order convergence in time. $\square$

The semidiscrete solution exhibits boundary and interior layer behavior characterized by the roots of the associated characteristic equation

$$\varepsilon \mathbf{r}^2(x) + \theta a^{j+1}(x) \mathbf{r}(x) - \zeta^{j+1}(x) = 0, \quad \zeta^{j+1}(x) = b^{j+1}(x) + \frac{c^{j+1}(x)}{\Delta t}.$$

The roots are

$$\begin{aligned} \mathbf{r}_0(x) &= -\frac{\theta a^{j+1}(x)}{2\varepsilon} - \sqrt{\left(\frac{\theta a^{j+1}(x)}{2\varepsilon}\right)^2 + \frac{\zeta^{j+1}(x)}{\varepsilon}} < 0, \\ \mathbf{r}_1(x) &= -\frac{\theta a^{j+1}(x)}{2\varepsilon} + \sqrt{\left(\frac{\theta a^{j+1}(x)}{2\varepsilon}\right)^2 + \frac{\zeta^{j+1}(x)}{\varepsilon}} > 0, \end{aligned}$$

with decay/growth rates in the layer regions given by [15, 26, 3, 1] $\wp_0 = -\max_{x \in \bar{\Omega}} \mathbf{r}_0(x)$ and $\wp_1 = \min_{x \in \bar{\Omega}} \mathbf{r}_1(x)$. Let $\aleph = \min_{\mathfrak{D}^- \cup \mathfrak{D}^+} \left\{ \frac{b(x,t)}{a(x,t)} \right\}$ and $\theta = \min\{\theta_1, \theta_2\}$. Two cases arise:

- If $\theta\theta^2 \leq \aleph\varepsilon$, then $\mathbf{r}_0, \mathbf{r}_1 = O(\varepsilon^{-1/2})$,
- If $\theta\theta^2 \geq \aleph\varepsilon$, then $\mathbf{r}_0 = O(\theta/\varepsilon)$ and $\mathbf{r}_1 = O(\theta^{-1})$.

These expressions explicitly characterize the solution's layer behavior in the semi-discrete problem.

3.2 Space variable discretization

Extensive studies on SPPs indicate that a uniform mesh cannot achieve parameter uniform convergence for (1) due to sharp layer regions. The problem features boundary layers near both ends of the spatial domain and an interior layer at the discontinuity. To accurately resolve these steep gradients, we adopt a priori Shishkin mesh, denoted by $\Omega^N = \{x_i\}_{i=0}^N$, which subdivides $[0, 1]$ into six strategically refined subintervals:

$$[0, 1] = [0, \delta_1] \cup [\delta_1, s - \delta_2] \cup [s - \delta_2, s] \cup [s, s + \delta_3] \cup [s + \delta_3, 1 - \delta_4] \cup [1 - \delta_4, 1].$$

Here, $\delta_1, \delta_2, \delta_3$, and δ_4 are transition parameters that determine the location and thickness of the layers. The transition parameters are judiciously chosen to maintain an optimal balance between

the effects of boundary and interior layers, ensuring both stability and accuracy of the numerical approximation throughout the computational domain. They are defined as follows:

$$\begin{aligned}\delta_1 &= \min \left\{ \frac{s}{4}, \frac{2}{\wp_0} \ln N \right\}, & \delta_2 &= \min \left\{ \frac{s}{4}, \frac{2}{\wp_1} \ln N \right\}, \\ \delta_3 &= \min \left\{ \frac{1-s}{4}, \frac{2}{\wp_1} \ln N \right\}, & \delta_4 &= \min \left\{ \frac{1-s}{4}, \frac{2}{\wp_0} \ln N \right\},\end{aligned}$$

where N is number of mesh intervals in $[0, 1]$ and $N \geq 8$. The values of $\wp_0$ and $\wp_1$ are given as [15, 26]

$$\wp_0 = \begin{cases} \frac{1}{2} \sqrt{\frac{\theta \aleph}{\varepsilon}}, & \text{if } \theta \theta^2 \leq \aleph \varepsilon, \\ \frac{\theta \theta}{2\varepsilon}, & \text{if } \theta \theta^2 \geq \aleph \varepsilon, \end{cases} \quad \wp_1 = \begin{cases} \frac{1}{2} \sqrt{\frac{\theta \aleph}{\varepsilon}}, & \text{if } \theta \theta^2 \leq \aleph \varepsilon, \\ \frac{\aleph}{2\theta}, & \text{if } \theta \theta^2 \geq \aleph \varepsilon. \end{cases}$$

Each subintervals $[0, \delta_1]$, $[s - \delta_2, s]$, $[s, s + \delta_3]$, and $[1 - \delta_4, 1]$ contains $N/8$ mesh points and $[\delta_1, s - \delta_2]$ and $[s + \delta_3, 1 - \delta_4]$ contain $N/4$ mesh points. The mesh spacing is given as

$$h_i = x_i - x_{i-1} = \begin{cases} \frac{8\delta_1}{N}, & \text{for } i = 1, 2, \dots, \frac{N}{8}, \\ \frac{4(s - \delta_1 - \delta_2)}{N}, & \text{for } \frac{N}{8} + 1, \frac{N}{8} + 2, \dots, \frac{3N}{8}, \\ \frac{8\delta_2}{N}, & \text{for } \frac{3N}{8} + 1, \frac{3N}{8} + 2, \dots, \frac{5N}{8}, \\ \frac{8\delta_3}{N}, & \text{for } \frac{5N}{8} + 1, \frac{5N}{8} + 2, \dots, \frac{7N}{8}, \\ \frac{4(1 - s - \delta_3 - \delta_4)}{N}, & \text{for } \frac{7N}{8} + 1, \frac{7N}{8} + 2, \dots, N, \\ \frac{8\delta_4}{N}, & \end{cases}$$

and the mesh points are given by

$$x_i = \begin{cases} \frac{8\delta_1}{N}i, & \text{for } i = 0, 1, \dots, \frac{N}{8}, \\ \delta_1 + \frac{4(s - \delta_1 - \delta_2)}{N} \left(i - \frac{N}{8}\right), & \text{for } \frac{N}{8} + 1, \frac{N}{8} + 2, \dots, \frac{3N}{8}, \\ s - \delta_2 + \frac{8\delta_2}{N} \left(i - \frac{3N}{8}\right), & \text{for } \frac{3N}{8} + 1, \frac{3N}{8} + 2, \dots, \frac{5N}{8}, \\ s + \frac{8\delta_3}{N} \left(i - \frac{5N}{8}\right), & \text{for } \frac{5N}{8} + 1, \frac{5N}{8} + 2, \dots, \frac{7N}{8}, \\ s + \delta_3 + \frac{4(1 - s - \delta_3 - \delta_4)}{N} \left(i - \frac{7N}{8}\right), & \text{for } \frac{7N}{8} + 1, \frac{7N}{8} + 2, \dots, N, \\ 1 - \delta_4 + \frac{8\delta_4}{N} \left(i - \frac{7N}{8}\right), & \end{cases} \quad (5)$$

The interior mesh points are denoted by

$$\Omega_x^N = \left\{ x_i : 1 \leq i \leq \frac{N}{2} - 1 \right\} \cup \left\{ x_i : \frac{N}{2} + 1 \leq i \leq N - 1 \right\}.$$

We choose $x_{\frac{N}{2}} = s$ and $\overline{\Omega}_x^N = \{x_i\}_0^N \cup \{s\}$.

3.3 Cubic spline in tension method

A function $S(x, t_{j+1})$ interpolates the solution $W(x_i, t_{j+1}) = W_i^{j+1}$ and passes through the points $(x_i, t_{j+1}, W_i^{j+1})$ and $(x_{i+1}, t_{j+1}, W_{i+1}^{j+1})$ such that

$$\begin{aligned} S(x, t_{j+1}) = & -\frac{h_i^2}{\theta^2 \sin(\theta)} \left[\sin\left(\frac{\theta(x-x_i)}{h_i}\right) Y_{i+1}^{j+1} + \sin\left(\frac{\theta(x_{i+1}-x)}{h_i}\right) Y_i^{j+1} \right] \\ & + \frac{h_i^2}{\theta^2} \left[\frac{x-x_i}{h_i} \left(Y_{i+1}^{j+1} + \frac{h_i^2}{\theta^2} W_{i+1}^{j+1} \right) \right. \\ & \left. + \frac{x_{i+1}-x}{h_i} \left(Y_i^{j+1} - \frac{\theta^2}{h_i^2} W_i^{j+1} \right) \right], \end{aligned} \quad (6)$$

where $\theta \geq 0$ is a real tension parameter that governs the stiffness of the spline by controlling the influence of curvature in the interpolant $S(x, t_{j+1})$. Moreover, $Y_i^{j+1} = (W'')_i^{j+1}$ and $Y_{i+1}^{j+1} = (W'')_{i+1}^{j+1}$ denote the corresponding second-order derivative values at the spatial nodes. The first derivatives of $S(x, t_{j+1})$ from the right and left at $x = x_i$ are given, respectively, by

$$\begin{aligned} S'(x_i^+, t_{j+1}) &= \frac{W_{i+1}^{j+1} - W_i^{j+1}}{h_i} \\ &\quad + \frac{h_i}{\theta^2} \left[(1 - \theta \csc(\theta)) Y_{i+1}^{j+1} - (1 - \theta \cot(\theta)) Y_i^{j+1} \right], \\ S'(x_i^-, t_{j+1}) &= \frac{W_i^{j+1} - W_{i-1}^{j+1}}{h_{i-1}} \\ &\quad + \frac{h_{i-1}}{\theta^2} \left[(1 - \theta \cot(\theta)) Y_i^{j+1} - (1 - \theta \csc(\theta)) Y_{i-1}^{j+1} \right]. \end{aligned}$$

Then, from the continuity condition, we have $S'(x_i^+, t_{j+1}) = S'(x_i^-, t_{j+1})$ and obtain

$$\begin{aligned} & \lambda_1 h_{i-1} Y_{i-1}^{j+1} + \lambda_2 (h_{i-1} + h_i) Y_i^{j+1} + \lambda_1 h_i Y_{i+1}^{j+1} \\ &= \frac{W_{i+1}^{j+1} - W_i^{j+1}}{h_i} - \frac{W_i^{j+1} - W_{i-1}^{j+1}}{h_{i-1}}, \end{aligned} \quad (7)$$

where θ is chosen in a problem-dependent manner, which is $\theta = h_i \sqrt{\frac{b+c/\Delta t}{\varepsilon}}$, $\lambda_1 = \frac{\theta \csc(\theta) - 1}{\theta^2}$, and $\lambda_2 = \frac{1 - \theta \cot(\theta)}{\theta^2}$. Application of the L'Hopitals rule gives $\lambda_1 + \lambda_2 = \frac{1}{2}$. Using Taylor expansion of $\sin \theta$ and $\cos \theta$, it can be shown that as $\theta \rightarrow 0$, the coefficients satisfy $\lambda_1 \rightarrow \frac{1}{6}$ and $\lambda_2 \rightarrow \frac{2}{3}$. Consequently, the cubic spline in tension reduces to the classical cubic spline formulation,

ensuring consistency of the proposed method. At $x = x_{i,\pm 1}$ using the notation $Y_i^{j+1} = (W'')_i^{j+1}$, from (3) we can have

$$\begin{aligned} Y_{i-1}^{j+1} &= \frac{1}{\varepsilon} \left\{ -\theta a_{i-1}^{j+1} (W')_{i-1}^{j+1} + \left(b_{i-1}^{j+1} + \frac{c_{i-1}^{j+1}}{\Delta t} \right) W_{i-1}^{j+1} + f_{i-1}^{j+1} - \frac{c_{i-1}^j}{\Delta t} W_{i-1}^{j+1} \right\}, \\ Y_i^{j+1} &= \frac{1}{\varepsilon} \left\{ -\theta a_i^{j+1} (W')_i^{j+1} + \left(b_i^{j+1} + \frac{c_i^{j+1}}{\Delta t} \right) W_i^{j+1} + f_i^{j+1} - \frac{c_i^j}{\Delta t} W_i^{j+1} \right\}, \\ Y_{i+1}^{j+1} &= \frac{1}{\varepsilon} \left\{ -\theta a_{i+1}^{j+1} (W')_{i+1}^{j+1} + \left(b_{i+1}^{j+1} + \frac{c_{i+1}^{j+1}}{\Delta t} \right) W_{i+1}^{j+1} + f_{i+1}^{j+1} - \frac{c_{i+1}^j}{\Delta t} W_{i+1}^{j+1} \right\}. \end{aligned} \quad (8)$$

Using Taylor's series approximations for W about $x_k, k = i \pm 1$, we have

$$W^{j+1}(x_{i-1}) \approx W^{j+1}(x_i) - h_{i-1} \frac{dW^{j+1}(x_i)}{dx} + \frac{h_{i-1}^2}{2} \frac{d^2W^{j+1}(x_i)}{dx^2}, \quad (9)$$

$$W^{j+1}(x_{i+1}) \approx W^{j+1}(x_i) + h_i \frac{dW^{j+1}(x_i)}{dx} + \frac{h_i^2}{2} \frac{d^2W^{j+1}(x_i)}{dx^2}. \quad (10)$$

Multiplying (9) by $\frac{h_i^2}{h_{i-1}^2}$ and subtracting the result from (10), we derive an approximate expression for $(W')_i^{j+1}$:

$$\begin{aligned} \frac{dW^{j+1}(x_i)}{dx} &\approx \frac{1}{h_i h_{i-1} (h_{i-1} + h_i)} (h_{i-1}^2 W^{j+1}(x_{i-1})) \\ &\quad + \frac{1}{h_i h_{i-1} (h_{i-1} + h_i)} ((h_i^2 - h_{i-1}^2) W^{j+1}(x_i) + h_{i-1}^2 W^{j+1}(x_{i+1})). \end{aligned} \quad (11)$$

Similarly, multiplying (9) by $\frac{h_i}{h_{i-1}}$ and then adding it to (10), we get an approximation for $(W'')_i^{j+1}$:

$$\begin{aligned} \frac{d^2W^{j+1}(x_i)}{dx^2} &\approx \frac{2}{h_i h_{i-1} (h_{i-1} + h_i)} (h_i W^{j+1}(x_{i-1}) + (h_i - h_{i-1}) W^{j+1}(x_i)) \\ &\quad + \frac{2}{h_i h_{i-1} (h_{i-1} + h_i)} (h_{i-1} W^{j+1}(x_{i+1})). \end{aligned} \quad (12)$$

Inserting (11) and (12) in

$$\begin{aligned} \frac{dW^{j+1}(x_{i+1})}{dx} &\approx \frac{dW^{j+1}(x_i)}{dx} + h_i \frac{d^2W^{j+1}(x_{i+1})}{dx^2} \quad \text{and} \\ \frac{dW^{j+1}(x_{i-1})}{dx} &\approx \frac{dW^{j+1}(x_i)}{dx} + h_{i-1} \frac{d^2W^{j+1}(x_{i+1})}{dx^2}, \end{aligned}$$

we obtain

$$\begin{aligned}
 (W')_{i-1}^{j+1} &= \frac{1}{h_i h_{i-1} (h_{i-1} + h_i)} \left\{ - (h_i^2 + 2h_{i-1} h_i) W_{i-1}^{j+1} + (h_{i-1} + h_i)^2 W_i^{j+1} \right. \\
 &\quad \left. - h_{i-1}^2 W_{i+1}^{j+1} \right\}, \\
 (W')_i^{j+1} &= \frac{1}{h_i h_{i-1} (h_{i-1} + h_i)} \left\{ -h_i^2 W_{i-1}^{j+1} + (h_i^2 - h_{i-1}^2) W_i^{j+1} \right. \\
 &\quad \left. + h_{i-1}^2 W_{i+1}^{j+1} \right\}, \\
 (W')_{i+1}^{j+1} &= \frac{1}{h_i h_{i-1} (h_{i-1} + h_i)} \left\{ h_i^2 W_{i-1}^{j+1} - (h_{i-1} + h_i)^2 W_i^{j+1} \right. \\
 &\quad \left. + (h_{i-1}^2 + 2h_{i-1} h_i) W_{i+1}^{j+1} \right\}.
 \end{aligned} \tag{13}$$

Employing the cubic spline in tension formulation proposed by Kumar and Ravi Kanth [17] gives

$$\begin{aligned}
 &\lambda_1 h_{i-1} Y_{i-1}^{j+1} + \lambda_2 (h_i + h_{i-1}) Y_i^{j+1} + \lambda_1 h_i Y_{i+1}^{j+1} \\
 &= \frac{W_{i+1}^{j+1} - W_i^{j+1}}{h_i} - \frac{W_i^{j+1} - W_{i-1}^{j+1}}{h_{i-1}}.
 \end{aligned} \tag{14}$$

Inserting (8) and (13) into (14) yields a combined expression integrating the spline formulation with the finite difference framework:

$$\mathcal{L}_{\varepsilon, \theta}^{N, M} W_i^{j+1} = r_i^- W_{i-1}^{j+1} + r_i^c W_i^{j+1} + r_i^+ W_{i+1}^{j+1} = F_i^{j+1}, \tag{15}$$

where $\mathcal{L}_{\varepsilon, \theta}^{N, M}$ represents the overall discrete operator that encapsulates the full numerical approximation of the governing problem, and the matrix coefficients r are specified by

$$\left\{ \begin{aligned}
 r_i^- &= \theta \left[\lambda_1 \left(\frac{h_i + 2h_{i-1}}{h_{i-1} + h_i} \right) a_{i-1}^{j+1} + \lambda_2 \left(\frac{h_i}{h_{i-1}} \right) a_i^{j+1} \right. \\
 &\quad \left. - \lambda_1 \left(\frac{h_i^2}{h_{i-1} (h_{i-1} + h_i)} \right) a_{i+1}^{j+1} \right] + \lambda_1 h_{i-1} \left(b_{i-1}^{j+1} + \frac{c_{i-1}^{j+1}}{\Delta t} \right) - \frac{\varepsilon}{h_{i-1}}, \\
 r_i^c &= -\theta \left[\lambda_1 \left(\frac{h_{i-1} + h_i}{h_i} \right) a_i^{j+1} + \lambda_2 \left(\frac{h_i^2 - h_{i-1}^2}{h_i h_{i-1}} \right) a_i^{j+1} \right. \\
 &\quad \left. - \lambda_1 \left(\frac{h_{i-1} + h_i}{h_{i-1}} \right) a_{i+1}^{j+1} \right] + \lambda_2 (h_i + h_{i-1}) \left(b_i^{j+1} + \frac{c_i^{j+1}}{\Delta t} \right) + \varepsilon \left(\frac{h_i + h_{i-1}}{h_{i-1} h_i} \right), \\
 r_i^+ &= \theta \left[\lambda_1 \left(\frac{h_{i-1}^2}{h_i (h_{i-1} + h_i)} \right) a_{i-1}^{j+1} - \lambda_2 \left(\frac{h_{i-1}}{h_i} \right) a_i^{j+1} \right. \\
 &\quad \left. - \lambda_1 \left(\frac{h_{i-1} + 2h_i}{(h_{i-1} + h_i)} \right) a_{i+1}^{j+1} \right] + \lambda_1 h_i \left(b_{i+1}^{j+1} + \frac{c_{i+1}^{j+1}}{\Delta t} \right) - \frac{\varepsilon}{h_i}, \\
 F_i^{j+1} &= \lambda_1 h_{i-1} f_{i-1}^{j+1} + \lambda_2 (h_{i-1} + h_i) f_i^{j+1} + \lambda_1 h_i f_{i+1}^{j+1} - \lambda_1 h_{i-1} \frac{c_{i-1}^j}{\Delta t} W_{i-1}^j \\
 &\quad - \lambda_2 (h_{i-1} + h_i) \frac{c_i^j}{\Delta t} W_i^j - \lambda_1 h_i \frac{c_{i+1}^j}{\Delta t} W_{i+1}^j.
 \end{aligned} \right.$$

3.4 Discrete scheme at the discontinuity point $x = s$.

Let the spatial mesh be divided such that

$$\Omega_x^N = \{x_i : 1 \leq i \leq \frac{N}{2} - 1\} \cup \{x_i : \frac{N}{2} + 1 \leq i \leq N - 1\},$$

and choose the interface point $x_{\frac{N}{2}} = s$. The extended mesh including the discontinuity is denoted by $\bar{\Omega}_x^N = \{x_i\}_0^N \cup \{s\}$. Define the local mesh sizes near the interface as $h_L = x_{\frac{N}{2}} - x_{\frac{N}{2}-1}$, $h_R = x_{\frac{N}{2}+1} - x_{\frac{N}{2}}$. Then, at the discontinuity node $x = s = x_{\frac{N}{2}}$, the fully discrete cubic spline in tension scheme can be written as

$$r_s^- W_{\frac{N}{2}-1}^{j+1} + r_s^c W_{\frac{N}{2}}^{j+1} + r_s^+ W_{\frac{N}{2}+1}^{j+1} = F_s^{j+1}.$$

The coefficients r_s^- , r_s^c , r_s^+ , and the right-hand side F_s^{j+1} are obtained by substituting $i = \frac{N}{2}$, $h_{i-1} = h_L$, and $h_i = h_R$ into the general expressions as

$$\begin{aligned} r_s^- &= -\frac{\varepsilon}{h_L} + \theta \left[\lambda_1 \left(\frac{h_R + 2h_L}{h_L + h_R} \right) a_{\frac{N}{2}-1}^{j+1} + \lambda_2 \left(\frac{h_R}{h_L} \right) a_{\frac{N}{2}}^{j+1} \right. \\ &\quad \left. - \lambda_1 \left(\frac{h_R^2}{h_L(h_L + h_R)} \right) a_{\frac{N}{2}+1}^{j+1} \right] + \lambda_1 h_L \left(b_{\frac{N}{2}-1}^{j+1} + \frac{c_{\frac{N}{2}-1}^{j+1}}{\Delta t} \right), \\ r_s^c &= \varepsilon \left(\frac{h_R + h_L}{h_L h_R} \right) - \theta \left[\lambda_1 \left(\frac{h_L + h_R}{h_R} \right) a_{\frac{N}{2}}^{j+1} + \lambda_2 \left(\frac{h_R^2 - h_L^2}{h_R h_L} \right) a_{\frac{N}{2}}^{j+1} \right. \\ &\quad \left. - \lambda_1 \left(\frac{h_L + h_R}{h_L} \right) a_{\frac{N}{2}+1}^{j+1} \right] + \lambda_2 (h_R + h_L) \left(b_{\frac{N}{2}}^{j+1} + \frac{c_{\frac{N}{2}}^{j+1}}{\Delta t} \right), \\ r_s^+ &= -\frac{\varepsilon}{h_R} + \theta \left[\lambda_1 \left(\frac{h_L^2}{h_R(h_L + h_R)} \right) a_{\frac{N}{2}-1}^{j+1} - \lambda_2 \left(\frac{h_L}{h_R} \right) a_{\frac{N}{2}}^{j+1} \right. \\ &\quad \left. - \lambda_1 \left(\frac{h_L + 2h_R}{h_L + h_R} \right) a_{\frac{N}{2}+1}^{j+1} \right] + \lambda_1 h_R \left(b_{\frac{N}{2}+1}^{j+1} + \frac{c_{\frac{N}{2}+1}^{j+1}}{\Delta t} \right). \end{aligned}$$

The corresponding right-hand side term is given by

$$\begin{aligned} F_s^{j+1} &= \lambda_1 h_L f_{\frac{N}{2}-1}^{j+1} + \lambda_2 (h_L + h_R) f_{\frac{N}{2}}^{j+1} + \lambda_1 h_R f_{\frac{N}{2}+1}^{j+1} \\ &\quad - \lambda_1 h_L \frac{c_{\frac{N}{2}-1}^j}{\Delta t} W_{\frac{N}{2}-1}^j - \lambda_2 (h_L + h_R) \frac{c_{\frac{N}{2}}^j}{\Delta t} W_{\frac{N}{2}}^j - \lambda_1 h_R \frac{c_{\frac{N}{2}+1}^j}{\Delta t} W_{\frac{N}{2}+1}^j. \end{aligned}$$

For sufficiently small values of h , the resulting matrix remains nonsingular, satisfying the condition $|r_i^c| \geq |r_i^-| + |r_i^+|$, which ensures that the matrix is diagonally dominant. Consequently, it qualifies as an M -matrix, possessing a positive inverse. Therefore, the corresponding system of discrete equations admits a unique solution, which can be obtained through the matrix inverse method, subject to prescribed boundary conditions $W^{j+1}(0)$ and $W^{j+1}(1)$.

4 Convergence analysis

In this section, we analyze the stability and ε -uniform convergence of the fully discrete scheme (15). To this end, we first establish a *discrete comparison principle* on the mesh $\mathfrak{D}^{N,M}$, which ensures that the numerical solution remains bounded and free of spurious oscillations. This principle forms the cornerstone for proving ε -uniform stability, guaranteeing that the convergence of the scheme is robust even as ε and θ becomes small.

Lemma 8 (Discrete comparison Principle). Suppose that W_i^{j+1} satisfies

$$r_i^- W_{i-1}^{j+1} + r_i^c W_i^{j+1} + r_i^+ W_{i+1}^{j+1} \geq 0,$$

with $r_i^- \geq 0, r_i^+ \geq 0, r_i^c \geq r_i^- + r_i^+$ (i.e., diagonally dominant). Then if $W_i^0 \geq 0$ and boundary data $W_0^{j+1}, W_N^{j+1} \geq 0$, it follows that $W_i^{j+1} \geq 0$, for all i, j .

The LTE at (x_i, t_{j+1}) is defined by substituting the exact solution into the discrete scheme:

$$\begin{aligned} T_i &= \mathcal{L}_{\varepsilon, \theta}^{N,M} W(x_i, t_{j+1}) - F_i^{j+1}, \\ &= r_i^- W^{j+1}(x_{i-1}) + r_i^c W^{j+1}(x_i) + r_i^+ W^{j+1}(x_{i+1}) \\ &\quad - \lambda_1 h_{i-1} \mathcal{G}^{j+1}(x_{i-1}) - \lambda_2 (h_i + h_{i-1}) \mathcal{G}^{j+1}(x_i) - \lambda_1 h_i \mathcal{G}^{j+1}(x_{i+1}). \end{aligned} \quad (16)$$

By substituting the expressions of $\mathcal{G}^{j+1}(x_{i-1})$, $\mathcal{G}^{j+1}(x_i)$, and $\mathcal{G}^{j+1}(x_{i+1})$ from (3) into (16), we derive the following relation:

$$\begin{aligned} T_i &= r_i^- W^{j+1}(x_{i-1}) + r_i^c W^{j+1}(x_i) + r_i^+ W^{j+1}(x_{i+1}) \\ &\quad - \lambda_1 h_{i-1} \left\{ \varepsilon \frac{d^2 W^{j+1}(x_{i-1})}{dx^2} + \theta a_{i-1} \frac{dW^{j+1}(x_{i-1})}{dx} - Q_{i-1} W^{j+1}(x_{i-1}) \right\} \\ &\quad - \lambda_2 (h_i + h_{i-1}) \left\{ \varepsilon \frac{d^2 W^{j+1}(x_i)}{dx^2} + \theta a_i \frac{dW^{j+1}(x_i)}{dx} - Q_i W^{j+1}(x_i) \right\} \\ &\quad - \lambda_1 h_i \left\{ \varepsilon \frac{d^2 W^{j+1}(x_{i+1})}{dx^2} + \theta a_{i+1} \frac{dW^{j+1}(x_{i+1})}{dx} - Q_{i+1} W^{j+1}(x_{i+1}) \right\}. \end{aligned} \quad (17)$$

Using Taylor's series approximation for $W^{j+1}(x_{i-1}), W^{j+1}(x_{i+1})$ and their derivatives in the spatial variables, we have

$$\begin{aligned} W^{j+1}(x_{i+1}) \approx & W^{j+1}(x_i) + h_i \frac{dW^{j+1}(x_i)}{dx} + \frac{h_i^2}{2} \frac{d^2W^{j+1}(x_i)}{dx^2} \\ & + \frac{h_i^3}{6} \frac{d^3W^{j+1}(x_i)}{dx^3} + \frac{h_i^4}{24} \frac{d^4W^{j+1}(x_i)}{dx^4} + \dots \end{aligned} \quad (18)$$

$$\begin{aligned} W^{j+1}(x_{i-1}) \approx & W^{j+1}(x_i) - h_{i-1} \frac{dW^{j+1}(x_i)}{dx} + \frac{h_{i-1}^2}{2} \frac{d^2W^{j+1}(x_i)}{dx^2} \\ & - \frac{h_{i-1}^3}{6} \frac{d^3W^{j+1}(x_i)}{dx^3} + \frac{h_{i-1}^4}{24} \frac{d^4W^{j+1}(x_i)}{dx^4} + \dots \end{aligned} \quad (19)$$

$$\begin{aligned} \frac{dW^{j+1}(x_{i-1})}{dx} \approx & \frac{dW^{j+1}(x_i)}{dx} - h_{i-1} \frac{d^2W^{j+1}(x_i)}{dx^2} + \frac{h_{i-1}^2}{2} \frac{d^3W^{j+1}(x_i)}{dx^3} \\ & - \frac{h_{i-1}^3}{6} \frac{d^4W^{j+1}(x_i)}{dx^4} + \dots \end{aligned} \quad (20)$$

$$\begin{aligned} \frac{dW^{j+1}(x_{i+1})}{dx} \approx & \frac{dW^{j+1}(x_i)}{dx} + h_i \frac{d^2W^{j+1}(x_i)}{dx^2} + \frac{h_i^2}{2} \frac{d^3W^{j+1}(x_i)}{dx^3} \\ & + \frac{h_i^3}{6} \frac{d^4W^{j+1}(x_i)}{dx^4} + \dots \end{aligned} \quad (21)$$

$$\frac{d^2W^{j+1}(x_{i+1})}{dx^2} \approx \frac{d^2W^{j+1}(x_i)}{dx^2} + h_i \frac{d^3W^{j+1}(x_i)}{dx^3} + \frac{h_i^2}{2} \frac{d^4W^{j+1}(x_i)}{dx^4} + \dots \quad (22)$$

$$\frac{d^2W^{j+1}(x_{i-1})}{dx^2} \approx \frac{d^2W^{j+1}(x_i)}{dx^2} - h_i \frac{d^3W^{j+1}(x_i)}{dx^3} + \frac{h_i^2}{2} \frac{d^4W^{j+1}(x_i)}{dx^4} + \dots \quad (23)$$

By inserting the expressions from (18)-(23) into (17), we have

$$\begin{aligned} T_i = & T_{1,i} W^{j+1}(x_i) + T_{2,i} \frac{dW^{j+1}(x_{i+1})}{dx} + T_{3,i} \frac{d^2W^{j+1}(x_{i+1})}{dx^2} \\ & + T_{4,i} \frac{d^3W^{j+1}(x_{i+1})}{dx^3} + T_{5,i} \frac{d^4W^{j+1}(x_{i+1})}{dx^4} + \dots, \end{aligned} \quad (24)$$

where

$$\begin{aligned}
T_{1,i} &= r_i^- + r_i^c + r_i^+ + \lambda_1 h_{i-1} Q_{i-1} + \lambda_2 (h_{i-1} + h_i) Q_i + \lambda_1 h_i Q_{i+1}, \\
T_{2,i} &= -h_{i-1} r_i^- + h_i r_i^+ - \lambda_1 \theta h_{i-1} a_{i-1} - \lambda_1 h_{i-1}^2 Q_{i-1} \\
&\quad - \lambda_2 \theta (h_i + h_{i-1}) a_i - \lambda_1 \theta h_i a_{i+1} + \lambda_1 h_i^2 Q_{i+1} \\
T_{3,i} &= \frac{h_{i-1}^2}{2} r_i^- + \frac{h_i^2}{2} r_i^+ - \varepsilon \lambda_1 h_{i-1} + \lambda_1 \theta h_{i-1}^2 a_{i-1} + \lambda_1 \frac{h_{i-1}^3}{2} Q_{i-1} \\
&\quad - \varepsilon \lambda_2 (h_i + h_{i-1}) - \varepsilon \lambda_1 h_i - \lambda_1 \theta h_i^2 a_{i+1} + \lambda_1 \frac{h_i^3}{2} Q_{i+1}, \\
T_{4,i} &= -\frac{h_{i-1}^3}{6} r_i^- + \frac{h_i^3}{6} r_i^+ + \varepsilon \lambda_1 h_{i-1}^2 - \lambda_1 \theta \frac{h_{i-1}^2}{2} a_{i-1} - \lambda_1 \frac{h_{i-1}^4}{6} Q_{i-1} \\
&\quad - \varepsilon \lambda_1 h_i^2 - \lambda_1 \theta \frac{h_i^3}{2} a_{i+1} + \lambda_1 \frac{h_i^4}{6} Q_{i+1}, \\
T_{5,i} &= \frac{h_{i-1}^4}{24} r_i^- + \frac{h_i^4}{24} r_i^+ - \varepsilon \lambda_1 \frac{h_{i-1}^3}{2} + \lambda_1 \theta \frac{h_{i-1}^4}{6} a_{i-1} + \lambda_1 \frac{h_{i-1}^5}{24} Q_{i-1} \\
&\quad - \varepsilon \lambda_1 \frac{h_i^2}{2} - \lambda_1 \theta \frac{h_i^4}{6} a_{i+1} + \lambda_1 \frac{h_i^5}{24} Q_{i+1}.
\end{aligned}$$

By simplifying the expressions and retaining only the leading terms in the coefficient expansions, we assume that the coefficients satisfy the boundedness conditions $|a(x)| \leq A$ and $|Q(x)| \leq Q_{\max}$, while the mesh sizes meet $h_i, h_{i-1} \leq h_{\max}$. Furthermore, let the weighting parameters fulfill the constraint $\lambda_1 + \lambda_2 = \frac{1}{2}$. When the coefficients in (15) of the three-point stencil satisfy the *moment consistency* conditions $r_i^- + r_i^c + r_i^+ = 0$ and $-h_{i-1} r_i^- + h_i r_i^+ = 0$, the scheme is said to be *stencil-consistent*. Under these assumptions, (24) at x_i, t_{j+1} satisfies

$$T_i = \sum_{n=1}^5 T_{n,i} \partial_x^{n-1} W(x_i, t_{j+1}) + O(h_i^5),$$

where the coefficients $T_{n,i}$ depend on $h_i, h_{i-1}, \varepsilon, \theta, a(x), Q(x)$, and the spline parameters $r_i^\pm, r_i^c$. The moment consistency conditions of the cubic spline in tension imply $T_{1,i} = T_{2,i} = T_{3,i} = 0$, so that the leading contribution of the LTE is

$$|T_i| \leq C(|T_{4,i}| \|W_{xxx}\| + |T_{5,i}| \|W_{xxxx}\|). \quad (25)$$

Lemma 9 (Solution decomposition and layer bounds). The exact solution $w(x, t)$ of problem (1) admits the decomposition:

$$w(x, t) = v(x, t) + u_l(x, t) + u_i(x, t) + u_r(x, t), \quad (26)$$

where v is the regular component, u_l and u_r are the left and right boundary layer components, respectively, and u_i denotes the interior layer component at $x = s$. Moreover, for $0 \leq m \leq 4$,

there exists a positive constant C , independent of ε and θ , such that

$$\begin{aligned} |v^{(m)}(x, t)| &\leq C, & |u_l^{(m)}(x, t)| &\leq C\varphi_0^m e^{-\varphi_0 x}, \\ |u_r^{(m)}(x, t)| &\leq C\varphi_0^m e^{-\varphi_0(1-x)}, \\ |u_i^{(m)}(x, t)| &\leq C\varphi_1^m e^{-\varphi_1|x-s|}. \end{aligned}$$

Proof. The result follows from standard singular perturbation theory for two-parameter convection-diffusion-reaction problems with discontinuous coefficients. The analysis relies on the sign assumptions imposed on $a(x, t)$, the positivity of the discrete operator coefficients, and the boundedness of the jump conditions across the interior point $x = s$ (see [1, 23, 2]). $\square$

Lemma 10 (LTE bound). For all interior mesh points $x_i \in \Omega_x^N$, the LTE satisfies

$$|T_i| \leq \begin{cases} CN^{-2} (\ln N)^2, & x_i \in [0, \delta_1] \cup [s - \delta_2, s] \cup [s, s + \delta_3] \cup [1 - \delta_4, 1], \\ CN^{-2}, & x_i \in [\delta_1, s - \delta_2] \cup [s + \delta_3, 1 - \delta_4]. \end{cases} \quad (27)$$

Proof. By combining the derivative bounds given in Lemma 9 with the representation in (25), the truncation error can be estimated separately over each mesh subregion.

- (i) Boundary layer regions: For $x_i \in [0, \delta_1] \cup [1 - \delta_4, 1]$, the mesh width satisfies $h_i = O\left(\frac{\ln N}{N\varphi_0}\right)$ and the fourth derivatives of the left and right layer components are bounded by $|u_l^{(4)}|, |u_r^{(4)}| \leq C\varphi_0^4$, which yields $|T_i| \leq C\left(\frac{\ln N}{N}\right)^2 \leq CN^{-2} (\ln N)^2$.
- (ii) Interior layer regions: For $x_i \in [s - \delta_2, s] \cup [s, s + \delta_3]$, the mesh width satisfies $h_i = O\left(\frac{\ln N}{N\varphi_1}\right)$, and the fourth derivatives of the interior layer is bounded by $|u_i^{(4)}| \leq C\varphi_1^4$, which yields $|T_i| \leq C\left(\frac{\ln N}{N}\right)^2 \leq CN^{-2} (\ln N)^2$.
- (iii) Regular(smooth) regions: For $x_i \in [\delta_1, s - \delta_2] \cup [s + \delta_3, 1 - \delta_4]$ derivatives are bounded and $h_i = O(N^{-1})$, yielding $|T_i| \leq CN^{-2}$.

By aggregating the estimates obtained over all mesh subregions, a uniform spatial truncation error bound is established. $\square$

Lemma 11. Let W_i^{j+1} be the numerical solution obtained using the cubic spline in tension method on the six-piece Shishkin mesh. Then, for all $\varepsilon, \theta \in (0, 1]$, the spatial LTE satisfies

$$\|W(x_i, t_{j+1}) - W^{j+1}\| \leq CN^{-2} (\ln N)^2,$$

where C is independent of ε, θ and N .

The desired result for the complete discrete scheme follows immediately by combining the triangle inequality with Lemmas 7, 10, and 11.

Theorem 1. Let $w(x, t), (x, t) \in \overline{\mathfrak{D}}$ be the solution of (1) and let $W(x_i, t_{j+1}), (x_i, t_{j+1}) \in \Omega^{M_t} \times \Omega^{N_x}$ be the solution of (15). Then, the error estimate for the fully discrete scheme is given by

$$\sup_{0 < \varepsilon, \theta \leq 1} \max_{i,j} \|w(x_i, t_{j+1}) - W(x_i, t_{j+1})\| \leq C \left(N^{-2} (\ln N)^2 + \Delta t \right).$$

Therefore, the proposed numerical scheme exhibits uniform convergence, ensuring reliable performance across both smooth and singular solution regimes. The spatial approximation attains second-order accuracy up to a logarithmic factor, whereas the temporal discretization is first-order accurate, thereby validating the consistency and stability of the method.

5 Numerical examples and discussions

Several representative test problems are solved using the proposed numerical scheme to validate the theoretical convergence results. Since the exact solutions are not available, the maximum pointwise error for each (ε, θ) pair is estimated via the double-mesh approach [9, 5, 15]:

$$\mathbb{E}_{\varepsilon, \theta}^{N, M} = \max_{0 \leq i \leq N, 0 \leq j \leq M} |W_{i,j}^{N, M} - W_{i,j}^{2N, 2M}|,$$

where $W_{i,j}^{N, M}$ denotes the numerical solution computed on a mesh with N and M intervals in space and time, respectively. The refined solution $W_{i,j}^{2N, 2M}$ is obtained by introducing midpoints in each previous interval. The corresponding numerical convergence rate is defined as

$$\mathbb{R}_{\varepsilon, \theta}^{N, M} = \frac{\log \left(\mathbb{E}_{\varepsilon, \theta}^{N, M} \right) - \log \left(\mathbb{E}_{\varepsilon, \theta}^{2N, 2M} \right)}{\log(2)}.$$

To illustrate the layer behavior induced by variations in $a(x, t)$ and $f(x, t)$, we consider two test cases.

Example 1. Let $\Omega = (0, 1)$ with $s = 0.5$ [23, 2, 6]. The problem is

$$\begin{cases} \varepsilon \frac{\partial^2 w}{\partial x^2} + \theta a(x, t) \frac{\partial w}{\partial x} - (1 + e^x)w - \frac{\partial w}{\partial t} = f(x, t), & (x, t) \in \mathfrak{D}^- \cup \mathfrak{D}^+, \\ w(0, t) = 0, \quad w(1, t) = 0, & t > 0, \\ w(x, 0) = 0, & x \in \overline{\mathfrak{D}}, \end{cases}$$

where

$$a(x, t) = \begin{cases} -(1 + x - x^2), & (x, t) \in \mathfrak{D}^-, \\ 1 + x - x^2, & (x, t) \in \mathfrak{D}^+, \end{cases}$$

$$f(x, t) = \begin{cases} -2t(1 + x^2), & (x, t) \in \mathfrak{D}^-, \\ 2t(1 + x^2), & (x, t) \in \mathfrak{D}^+. \end{cases}$$

Example 2. Consider (1) with coefficients [15]:

$$a(x, t) = \begin{cases} -(1 + e^{-xt}), & 0 \leq x \leq s, \\ 2 + x + t, & s < x \leq 1, \end{cases}$$

$$f(x, t) = \begin{cases} (e^{t^2} - 1)(1 + xt), & 0 \leq x \leq s, \\ -t^2(2 + x), & s < x \leq 1, \end{cases}$$

with $b(x, t) = 2 + xt$, $c(x, t) = 1$, boundary and initial conditions $w(0, t) = w(1, t) = 0$, $w(x, 0) = 0$.

Examples 1 and 2 exhibit pronounced boundary and interior layers due to ε, θ and nonsmooth data. Such behavior often arises in practical models involving heat conduction in composite materials, fluid flow through porous media, or chemical diffusion with abrupt concentration changes. When a uniform mesh is used in the x -direction, the sharp gradients within these layers are poorly resolved, leading to parameter-dependent inaccuracies. To address this, a Shishkin mesh is employed to refine the grid near the layer regions, ensuring accurate layer capture and uniform convergence. The corresponding graphs clearly illustrate how the adaptive mesh effectively resolves the steep solution variations across the domain.

Numerical results are summarized in Tables 1, 2, and 5 for various choices of ε , θ , N , and M . The data demonstrate that the maximum pointwise error decreases as the mesh is refined, while the error remains essentially insensitive to further reductions in the perturbation parameters, confirming parameter-robust behavior. In Table 1, we fix $\theta = 2^{-4}$ and $s = 0.5$ and vary ε from 1 to 2^{-20} to report $\mathbb{E}_{\varepsilon, \theta}^{N, M}$ and $\mathbb{P}_{\varepsilon, \theta}^{N, M}$ for Example 1; these values indicate nearly first-order, parameter-uniform convergence. Table 2 presents a complementary study with ε fixed and varying θ , again showing close to first-order uniform convergence. For Example 2, Table 5 displays consistent uniform convergence across tested parameter sets. Comparative results in Tables 1, 3, 4, 5, and 6 against the methods of [15, 26, 8, 2] further confirm that the proposed scheme attains superior accuracy. In Tables 4 and 6, we present numerical results for $\mathbb{E}_{\varepsilon, \theta}^{N, M}, \mathbb{P}_{\varepsilon, \theta}^{N, M}$, taking different values of ε, θ , and s , and these tables confirm our theoretical

Table 1: Computed $\mathbb{E}_{\varepsilon,\theta}^{N,M}$ and $\mathbb{R}_{\varepsilon,\theta}^{N,M}$ when $\theta = 2^{-4}$ for Example 1

ε	$N = 32$	$N = 64$	$N = 128$	$N = 256$	$N = 512$
$\downarrow$	$M = 32$	$M = 64$	$M = 128$	$M = 256$	$M = 512$
Present Method					
2^{-0}	$1.3300e-02$	$9.5221e-03$	$5.7104e-03$	$3.1247e-03$	$1.6541e-03$
2^{-2}	$2.1453e-02$	$1.2117e-02$	$6.5107e-03$	$3.4218e-03$	$1.7729e-03$
2^{-4}	$2.4427e-02$	$1.3234e-02$	$6.9725e-03$	$3.6067e-03$	$1.8453e-03$
2^{-6}	$2.5976e-02$	$1.3938e-02$	$7.2532e-03$	$3.7168e-03$	$1.8876e-03$
2^{-8}	$2.7178e-02$	$1.4351e-02$	$7.4179e-03$	$3.7802e-03$	$1.9117e-03$
2^{-10}	$2.7718e-02$	$1.4590e-02$	$7.5105e-03$	$3.8156e-03$	$1.9251e-03$
2^{-12}	$2.7993e-02$	$1.4716e-02$	$7.5618e-03$	$3.8353e-03$	$1.9325e-03$
2^{-14}	$2.8433e-02$	$1.4784e-02$	$7.5884e-03$	$3.8459e-03$	$1.9365e-03$
2^{-16}	$2.8913e-02$	$1.4816e-02$	$7.6016e-03$	$3.8514e-03$	$1.9386e-03$
2^{-18}	$2.9071e-02$	$1.4823e-02$	$7.6079e-03$	$3.8540e-03$	$1.9396e-03$
2^{-20}	$2.9113e-02$	$1.4825e-02$	$7.6091e-03$	$3.8550e-03$	$1.9401e-03$
$\mathbb{E}_{\varepsilon,\theta}^{N,M}$	$2.9113e-02$	$1.4825e-02$	$7.6091e-03$	$3.8550e-03$	$1.9401e-03$
$\mathbb{R}_{\varepsilon,\theta}^{N,M}$	0.974	0.962	0.981	0.991	—
Result in [2](Uniform S-mesh)					
$\mathbb{E}_{\varepsilon,\theta}^{N,M}$		$8.27430e-02$	$5.27314e-02$	$3.03482e-02$	$1.64782e-02$
$\mathbb{R}_{\varepsilon,\theta}^{N,M}$		0.650	0.797	0.881	—

Table 2: Computed $\mathbb{E}_{\varepsilon,\theta}^{N,M}$ and $\mathbb{P}_{\varepsilon,\theta}^{N,M}$ when $\varepsilon = 2^{-8}$ for Example 1

θ	$N = 32$	$N = 64$	$N = 128$	$N = 256$	$N = 512$
$\downarrow$	$M = 32$	$M = 64$	$M = 128$	$M = 256$	$M = 512$
2^{-0}	$2.5997e-02$	$1.4063e-02$	$7.3439e-03$	$3.7619e-03$	$1.9074e-03$
2^{-2}	$2.6886e-02$	$1.4309e-02$	$7.4054e-03$	$3.7771e-03$	$1.9109e-03$
2^{-4}	$2.7178e-02$	$1.4351e-02$	$7.4179e-03$	$3.7802e-03$	$1.9117e-03$
2^{-6}	$2.7210e-02$	$1.4370e-02$	$7.4207e-03$	$3.7809e-03$	$1.9119e-03$
2^{-8}	$2.7211e-02$	$1.4373e-02$	$7.4210e-03$	$3.7809e-03$	$1.9119e-03$
2^{-10}	$2.7210e-02$	$1.4374e-02$	$7.4211e-03$	$3.7810e-03$	$1.9119e-03$
$\vdots$	$\vdots$	$\vdots$	$\vdots$	$\vdots$	$\vdots$
2^{-20}	$2.7210e-02$	$1.4374e-02$	$7.4211e-03$	$3.7810e-03$	$1.9119e-03$
$\mathbb{E}_{\varepsilon,\theta}^{N,M}$	$2.7211e-02$	$1.4374e-02$	$7.4211e-03$	$3.7810e-03$	$1.9119e-03$
$\mathbb{R}_{\varepsilon,\theta}^{N,M}$	0.921	0.954	0.973	0.984	—

claims. Finally, both examples exhibit a pronounced solution transition near $x = 0.5$, a direct consequence of the discontinuity, which is effectively resolved by the Shishkin mesh and is evident in the accompanying solution plots. The computed solution profiles are illustrated as surface-

Table 3: Comparison of $\mathbb{E}_{N,M}^{\varepsilon,\theta}$ for Example 1 with $\theta = 2^{-4}$

ε	$N = 64$			$N = 128$			$N = 256$			$N = 512$		
	In [2]	In [15]	PM	In [2]	In [15]	PM	In [2]	In [15]	PM	In [2]	In [15]	PM
2^{-6}	$4.28e-03$	$3.04e-03$	$1.39e-02$	$1.87e-03$	$1.55e-03$	$7.25e-03$	$8.63e-04$	$7.83e-04$	$3.72e-03$	$4.14e-04$	$3.94e-04$	$1.89e-03$
2^{-10}	$3.64e-02$	$1.86e-02$	$1.46e-02$	$1.84e-02$	$1.23e-02$	$7.51e-03$	$8.96e-03$	$7.52e-03$	$3.82e-03$	$5.14e-03$	$4.46e-03$	$1.93e-03$
2^{-14}	$7.63e-02$	$2.98e-02$	$1.48e-02$	$4.69e-02$	$1.80e-02$	$7.59e-03$	$2.62e-02$	$1.08e-02$	$3.85e-03$	$1.39e-02$	$6.35e-03$	$1.94e-03$
2^{-18}	$8.23e-02$	$3.13e-02$	$1.48e-02$	$5.23e-02$	$1.93e-02$	$7.61e-03$	$3.00e-02$	$1.18e-02$	$3.85e-03$	$1.63e-02$	$6.94e-03$	$1.94e-03$
2^{-22}	$8.27e-02$	$3.14e-02$	$1.48e-02$	$5.27e-02$	$1.94e-02$	$7.61e-03$	$3.03e-02$	$1.19e-02$	$3.86e-03$	$1.65e-02$	$6.98e-03$	$1.94e-03$

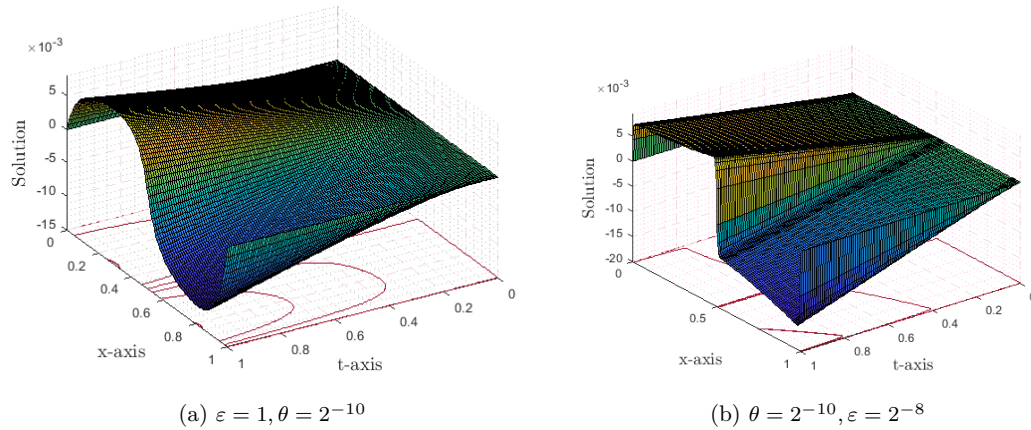
Table 4: Computed $\mathbb{E}_{\varepsilon,\theta}^{N,M}$ and $\mathbb{P}_{\varepsilon,\theta}^{N,M}$ for different ε, θ and s for Example 1.

Number of intervals in space and time						
		$N = 16$	$N = 32$	$N = 64$	$N = 128$	$N = 256$
s	(ε, θ)	$M = 10$	$M = 20$	$M = 40$	$M = 80$	$M = 160$
Present Approach						
0.2	$(2^{-6}, 2^{-18})$	$7.1577e-02$	$3.9198e-02$	$2.1401e-02$	$1.133e-02$	$5.8576e-03$
		0.869	0.873	0.918	0.952	—
	$(2^{-12}, 2^{-28})$	$7.7493e-02$	$4.2412e-02$	$2.2892e-02$	$1.1917e-02$	$6.0871e-03$
		0.870	0.890	0.942	0.969	—
Results in [26]						
	$(2^{-6}, 2^{-18})$	$7.8576e-02$	$6.9196e-02$	$4.7451e-02$	$2.8080e-02$	$1.5321e-02$
		0.183	0.544	0.757	0.874	—
	$(2^{-12}, 2^{-28})$	$1.1880e-01$	$7.3299e-02$	$4.1310e-02$	$2.2011e-02$	$1.1373e-02$
		0.697	0.827	0.908	0.953	—
Present Approach						
0.8	$(2^{-6}, 2^{-18})$	$6.1000e-02$	$3.8654e-02$	$2.1423e-02$	$1.1328e-02$	$5.8584e-03$
		0.658	0.851	0.919	0.951	—
	$(2^{-12}, 2^{-28})$	$7.7482e-02$	$4.2405e-02$	$2.2891e-02$	$1.1917e-02$	$6.0871e-03$
		0.870	0.890	0.942	0.969	—
Results in [26]						
	$(2^{-6}, 2^{-18})$	$7.8860e-02$	$4.4466e-02$	$2.3437e-02$	$1.2013e-02$	$6.0799e-03$
		0.827	0.924	0.964	0.983	—
	$(2^{-12}, 2^{-28})$	$1.4226e-01$	$7.7775e-02$	$4.0788e-02$	$2.0889e-02$	$1.0571e-02$
		0.871	0.931	0.965	0.983	—

contour and mesh plots in Figures 1, 3, 4, 6, and 7 for Examples 1 and 2, respectively. These figures clearly demonstrate the steep gradients of the numerical solution near the discontinuity points ($x = s$), particularly as the perturbation parameters ε and θ decrease. Such behavior is a direct indication of the presence of boundary and interior layers, a common feature in singularly perturbed problems involving discontinuous data. The sensitivity of the solution to the perturbation parameters is highlighted by the narrowing of these layers with decreasing ε or θ (see Figures 1a and 4a), underscoring the need for fitted mesh approaches for precise numerical resolution. Furthermore, Figures 3 and 6, which correspond to Examples 1 and 2, show surface plots for different discontinuity positions s . The line plots at various time levels

Table 5: Computed $\mathbb{E}_{\varepsilon,\theta}^{N,M}$ and $\mathbb{P}_{\varepsilon,\theta}^{N,M}$ when $\varepsilon = 2^{-8}$ for Example 2

θ	$N = 32$	$N = 64$	$N = 128$	$N = 256$	$N = 512$
$\downarrow$	$M = 32$	$M = 64$	$M = 128$	$M = 256$	$M = 512$
1	$2.2303e-02$	$1.1082e-02$	$5.6730e-03$	$2.8751e-03$	$1.4485e-03$
2^{-2}	$2.2303e-02$	$1.1071e-02$	$5.6738e-03$	$2.8754e-03$	$1.4485e-03$
2^{-4}	$2.1148e-02$	$1.1066e-02$	$5.6735e-03$	$2.8754e-03$	$1.4485e-03$
2^{-6}	$2.1087e-02$	$1.1063e-02$	$5.6732e-03$	$2.8754e-03$	$1.4485e-03$
2^{-8}	$2.1095e-02$	$1.1064e-02$	$5.6730e-03$	$2.8754e-03$	$1.4485e-03$
2^{-10}	$2.1097e-02$	$1.1065e-02$	$5.6731e-03$	$2.8754e-03$	$1.4485e-03$
$\vdots$	$\vdots$	$\vdots$	$\vdots$	$\vdots$	$\vdots$
2^{-22}	$2.1097e-02$	$1.1065e-02$	$5.6731e-03$	$2.8754e-03$	$1.4485e-03$
2^{-24}	$2.1097e-022$	$1.1065e-02$	$5.6731e-03$	$2.8754e-03$	$1.4485e-03$
$\mathbb{E}^{N,M}$	$2.2303e-02$	$1.1082e-02$	$5.6738e-03$	$2.8754e-03$	$1.4485e-03$
$\mathbb{R}^{N,M}$	1.01	0.966	0.981	0.989	—
Result in [1]					
$\mathbb{E}^{N,M}$		$1.33874e-02$	$1.16318e-02$	$8.67966e-03$	$5.89059e-03$
$\mathbb{P}^{N,M}$		0.202	0.422	0.559	—

Figure 1: The computed solution profiles for Example 1 using $N = M = 128$.

with $N = M = 128$ (Figures 2 and 5) show that the boundary and interior layer regions, where the solution profile varies quickly, are where the maximum numerical error is found. The plots also support the theoretical predictions about how layer thickness depends on perturbation magnitudes by showing that decreasing ε and θ results in sharper layer transitions. The results are in line with previous research presented in [21], which found that, under the condition that the source term $f(x, t)$ stays continuous, interior layers are mainly caused by discontinuities in the

Table 6: Computed $\mathbf{E}_{\varepsilon,\theta}^{N,M}$ and $\mathbf{P}_{\varepsilon,\theta}^{N,M}$ for different ε, θ and s for Example 2.

Number of intervals in space and time						
		$N = 16$	$N = 32$	$N = 64$	$N = 128$	$N = 256$
s	(ε, θ)	$M = 10$	$M = 20$	$M = 40$	$M = 80$	$M = 160$
Present Approach						
0.2	$(2^{-6}, 2^{-18})$	$5.6359e-02$	$3.1368e-02$	$1.6956e-02$	$8.8405e-03$	$4.5256e-03$
		0.869	0.873	0.918	0.952	—
	$(2^{-12}, 2^{-28})$	$5.8369e-02$	$3.2425e-02$	$1.7407e-02$	$9.0228e-03$	$4.5945e-03$
		0.848	0.897	0.948	0.974	—
	Results in [26]					
	$(2^{-6}, 2^{-18})$	$6.7015e-02$	$6.1676e-02$	$4.3579e-02$	$2.6251e-02$	$1.4462e-02$
0.120		0.501	0.731	0.860	—	
$(2^{-12}, 2^{-28})$	$3.7093e-02$	$2.8247e-02$	$1.8060e-02$	$9.9259e-03$	$5.2073e-03$	
	0.393	0.645	0.864	0.931	—	
Present Approach						
0.8	$(2^{-6}, 2^{-18})$	$5.2613e-02$	$3.1062e-02$	$1.7031e-02$	$8.9668e-03$	$4.6131e-03$
		0.760	0.867	0.926	0.959	—
	$(2^{-12}, 2^{-28})$	$5.8369e-02$	$3.2582e-02$	$1.7746e-02$	$9.2488e-03$	$4.7213e-03$
		0.841	0.877	0.940	0.970	—
	Results in [26]					
	$(2^{-6}, 2^{-18})$	$5.6605e-02$	$4.4466e-02$	$1.7632e-02$	$9.0974e-03$	$4.6191e-03$
0.779		0.904	0.955	0.978	—	
$(2^{-12}, 2^{-28})$	$3.9828e-02$	$3.1977e-02$	$1.9895e-02$	$1.0191e-02$	$5.1934e-03$	
	0.317	0.685	0.965	0.973	—	

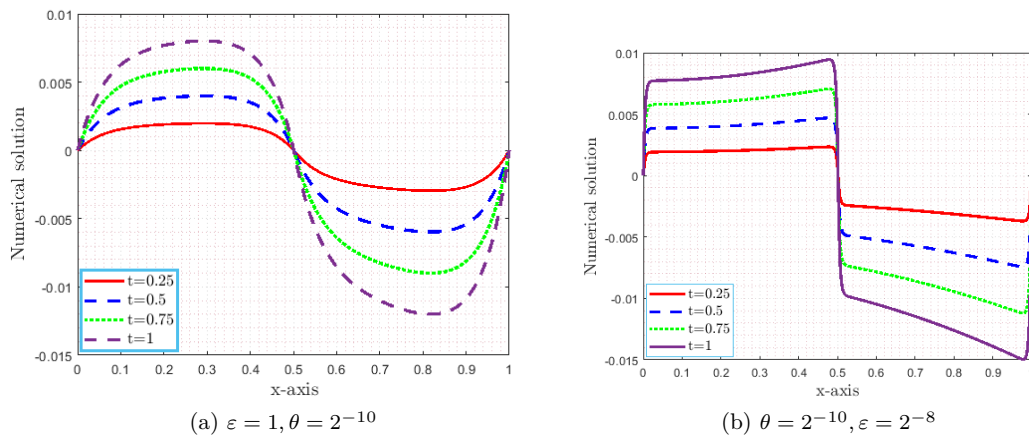


Figure 2: Line plots illustrating Example 1 for several time increments.

convection coefficient. Our comprehension of the interaction between discontinuous coefficients and solution layer structures is improved by this numerical evidence. As can be seen in Figure 7, the resulting solution typically only shows boundary layers if the discontinuity only occurs in

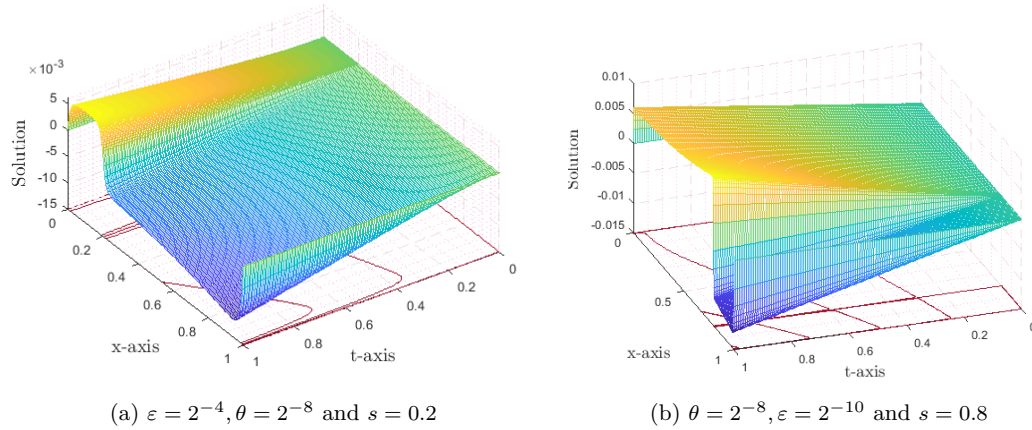


Figure 3: Computed solution profiles for Example 1 with $N = 256$ and $M = 160$, illustrating the influence of varying s values.

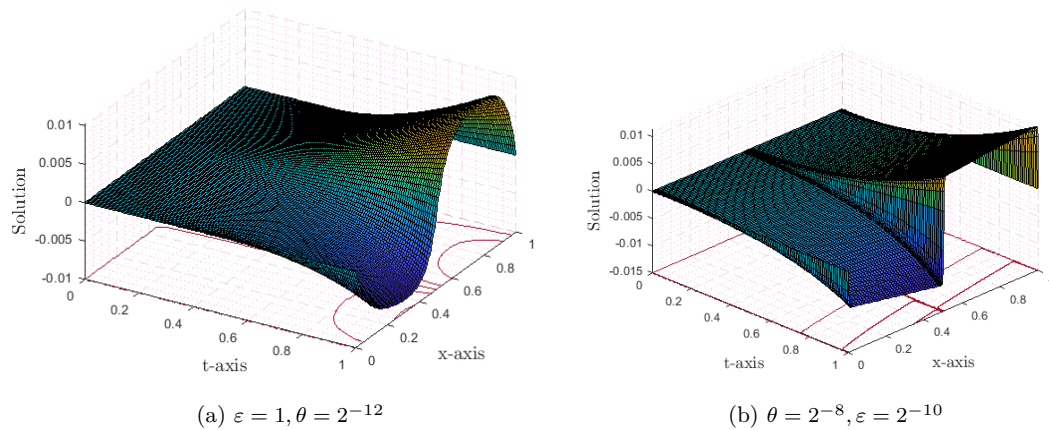


Figure 4: The computed solution profiles for Example 2 using $N = M = 128$.

the convection coefficient (with a smooth $f(x, t)$). To further investigate this behavior, the discontinuous source term in Example 2 was replaced by the smooth functions $f(x, t) = (e^{t^2} - 1)(1 + xt)$ and $f(x, t) = -t^2(2 + x)$. The corresponding numerical results are presented in Figures 7a and 7b, respectively. The findings demonstrate that when the source term is smooth, no distinct interior layer forms, even though the discontinuity in the convection coefficient causes a shift in the layer position. This finding indicates that the emergence and intensity of interior layers are significantly influenced by discontinuities in the source term. Additionally, Figures 8a and 8b for Examples 1 and 2, respectively, show error distributions on a log-log scale to verify the theoretical convergence behavior of the proposed approach. These plots support the robustness

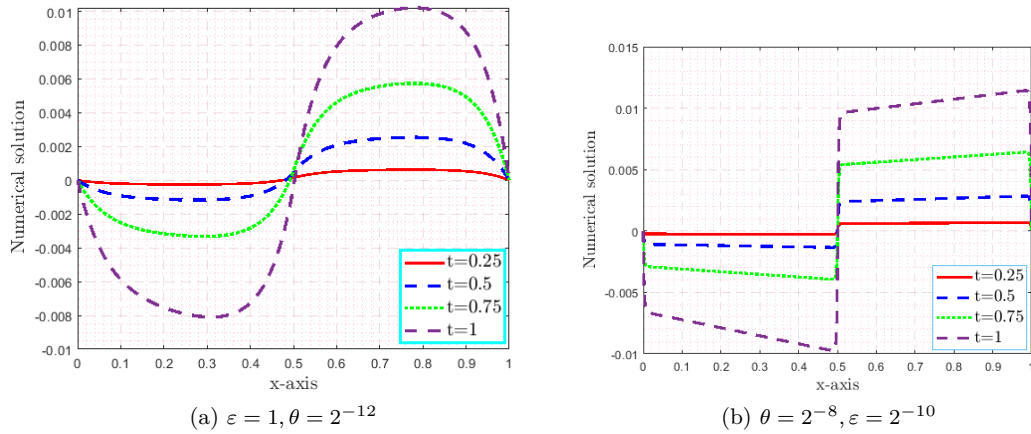
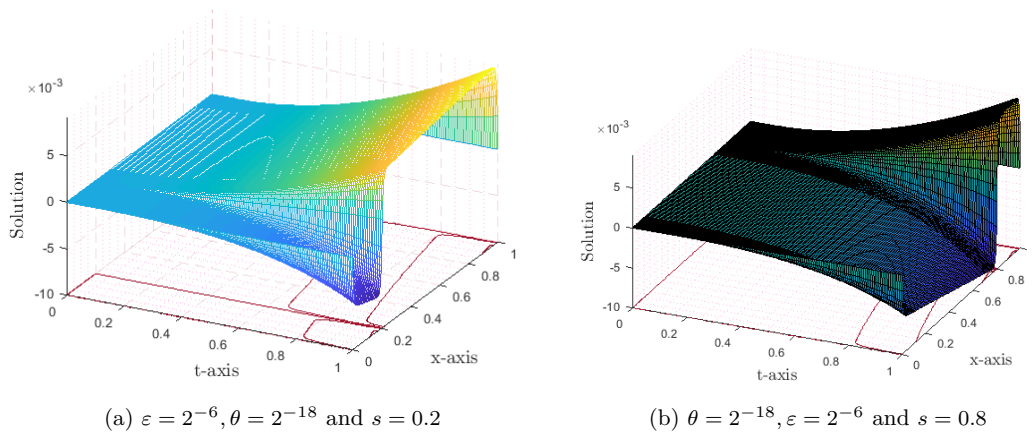


Figure 5: Line plots illustrating Example 2 for several time increments.

Figure 6: Computed solution profiles for Example 2 with $N = 256$ and $M = 160$, illustrating the influence of varying s values.

and (ε, θ) -uniform accuracy of the numerical scheme by confirming that it achieves the anticipated rate of convergence. These results have important ramifications for real-world applications in fluid transport, convection-diffusion processes, and semiconductor modeling, where abrupt gradients and discontinuities frequently occur naturally.

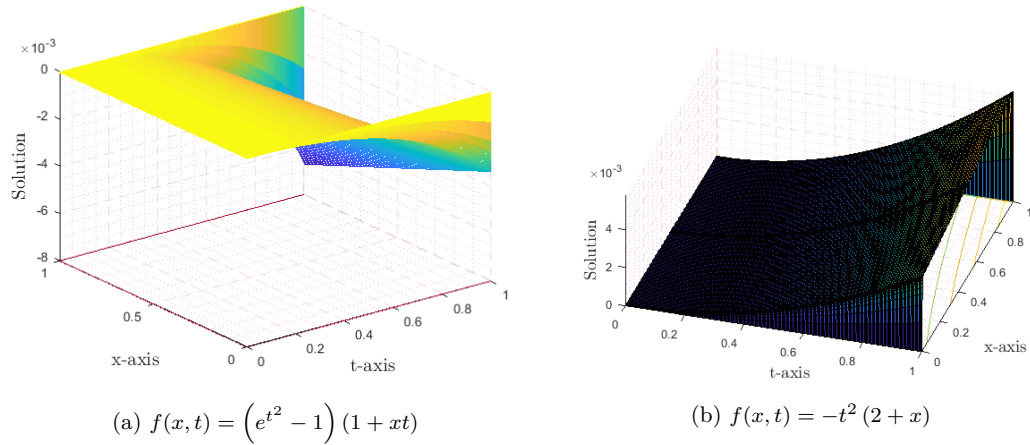


Figure 7: Computed solution profiles for Example 2 with $N = M = 256$, $\varepsilon = 2^{-12}$, $\theta = 2^{-8}$, considering multiple continuous source terms.

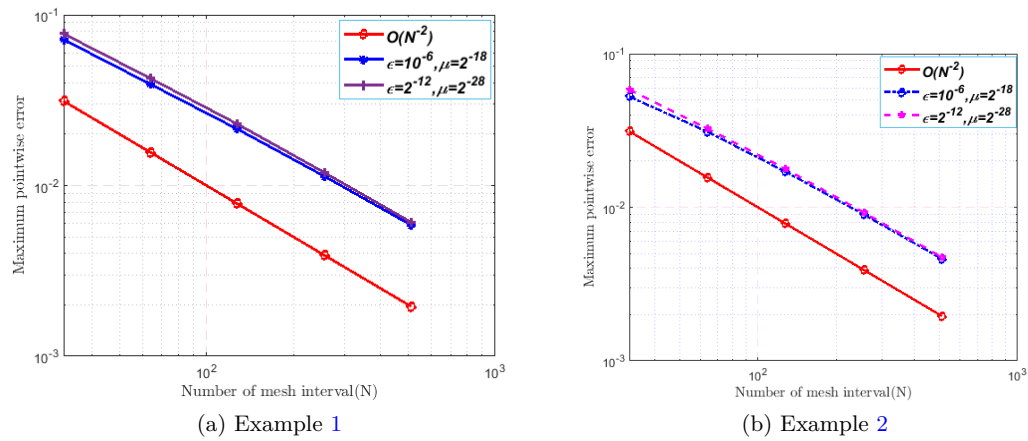


Figure 8: Loglog plot of Maximum pointwise errors.

6 Conclusion

An implicit numerical framework for solving a TP-SPPPs with discontinuities in the source and convection terms was presented in this work. It is based on a predefined Shishkin mesh. These issues were distinguished from situations involving smooth data by the complex solution behavior introduced by the existence of both boundary and interior layers. The suggested approach combined a cubic spline in tension for spatial approximation on a fitted mesh with the implicit Euler method for temporal discretization. Convergence analysis and extensive numerical experiments confirmed that the method achieves first-order accuracy in time and second-order accuracy up to

a logarithmic factor in space and is uniformly stable and robust. As the perturbation parameters ε and θ decrease, the graphic results clearly show the emergence of parabolic boundary layers along the lateral edges and interior layers around $x = s$. The boundary subregions and steep gradients near the discontinuity location s showed how well the scheme captures layer behavior even under extreme perturbations. The reliability of the proposed approach was demonstrated by comparative studies, which yield accuracy that is on par with or better than that of existing methods. The method's resilience in handling discontinuous singular perturbation problems was demonstrated by both theoretical validation and empirical results.

Acknowledgements

Authors are grateful to there 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] Chandru, M., Das, P. and Ramos, H. *Numerical treatment of two-parameter singularly perturbed parabolic convection diffusion problems with non-smooth data*, *Math. Methods Appl. Sci.* 41(14) (2018), 5359–5387.
- [2] Chandru, M., Prabha, T., Das, P. and Shanthi, V. *A numerical method for solving boundary and interior layers dominated parabolic problems with discontinuous convection coefficient and source terms*, *Differ. Equ. Dyn. Syst.* 27(1) (2019), 91–112.
- [3] Chandru, M., Prabha, T. and Shanthi, V. *A parameter robust higher order numerical method for singularly perturbed two parameter problems with non-smooth data*, *J. Comput. Appl. Math.* 309 (2017), 11–27.

- [4] Cheru, S. L., Duressa, G. F. and Mekonnen, T. B. *Robust numerical approach for singularly perturbed parabolic problems with discontinuous convection coefficients and source terms*, *Comput. Methods Differ. Equ.* (2025) (in press).
- [5] Cheru, S. L., Duressa, G. F. and Mekonnen, T. B. *Gaussian quadrature method with exponential fitting factor for two-parameter singularly perturbed parabolic problem*, *BMC Res. Notes* 17(1) (2024), 1–7.
- [6] Cheru, S. L., Duressa, G. F. and Mekonnen, T. B. *Robust numerical methods for two-parameter singularly perturbed parabolic problems with non-smooth data*, *Sci. Afr.* (2025), e03064.
- [7] Cheru, S. L., Duressa, G. F. and Mekonnen, T. B. *Numerical integration method for two-parameter singularly perturbed time delay parabolic problem*, *Front. Appl. Math. Stat.* 10(15) (2024) 1414899.
- [8] Clavero, C., Gracia, J. L., Shishkin, G. I. and Shishkina, L. P. *An efficient numerical scheme for 1D parabolic singularly perturbed problems with an interior and boundary layers*, *J. Comput. Appl. Math.* 318 (2017), 634–645.
- [9] Daba, I. T., Melesse, W. G., Gelu, F. W. and Kebede, G. D. *An efficient numerical approach for singularly perturbed time delayed parabolic problems with two-parameters*, *BMC Res. Notes* 17(1) (2024) 158.
- [10] Das, P. and Mehrmann, V. *Numerical solution of singularly perturbed convection–diffusion–reaction problems with two small parameters*, *BIT Numer. Math.* 56 (2016), 51–76.
- [11] Farrell, P. A., Hegarty, A. F., Miller, J. J., O’Riordan, E. and Shishkin, G. I. *Global maximum norm parameter-uniform numerical method for a singularly perturbed convection–diffusion problem with discontinuous convection coefficient*, *Math. Comput. Model.* 40(11–12) (2004), 1375–1392.
- [12] Gracia, J. L., O’Riordan, E. and Pickett, M. L. *A parameter robust second order numerical method for a singularly perturbed two-parameter problem*, *Appl. Numer. Math.* 56(7) (2006), 962–980.
- [13] Kadalbajoo, M. K. and Jha, A. *Exponentially fitted cubic spline for two-parameter singularly perturbed boundary value problems*, *Int. J. Comput. Math.* 89(6) (2012), 836–850.

- [14] Kaushik, A. and Sharma, M. *An optimal asymptotic-numerical method for convection dominated systems having exponential boundary layers*, J. Differ. Equ. Appl., 22(9) (2016) 1307–1324.
- [15] Kumar, D. and Kumari, P. *Uniformly convergent scheme for two-parameter singularly perturbed problems with non-smooth data*, Numer. Methods Partial Differ. Equ. 37(1) (2021), 796–817.
- [16] Kumar, D., Yadaw, A. S. and Kadalbajoo, M. K. *A parameter-uniform method for two parameters singularly perturbed BVPs via asymptotic expansion*, Appl. Math. Inf. Sci. 7(4) (2013), 1525.
- [17] Kumar, P. M. M. and Ravi Kanth, A. S. V. *Computational study for a class of time-dependent singularly perturbed parabolic partial differential equation through tension spline*, Comput. Appl. Math. 39(3) (2020), 233.
- [18] Linss, T. *A posteriori error estimation for a singularly perturbed problem with two small parameters*, Int. J. Numer. Anal. Model. 7(3) (2010) 491–506.
- [19] Mukherjee, K. and Natesan, S. *An efficient numerical scheme for singularly perturbed parabolic problems with interior layer*, Neural Parallel Sci. Comput. 16(3) (2008), 405.
- [20] O’Riordan, E., Pickett, M. and Shishkin, G. *Parameter-uniform finite difference schemes for singularly perturbed parabolic diffusion–convection–reaction problems*, Math. Comp. 75(255) (2006), 1135–1154.
- [21] O’Riordan, E. and Shishkin, G. I. *Singularly perturbed parabolic problems with non-smooth data*, J. Comput. Appl. Math. 166(1) (2004), 233–245.
- [22] Prabha, T., Chandru, M. and Shanthi, V. *Hybrid difference scheme for singularly perturbed parabolic problem with boundary and interior layers*, Appl. Math. Comput. 314 (2017), 237–256.
- [23] Prabha, T., Chandru, M., Shanthi, V. and Ramos, H. *Discrete approximation for a two-parameter singularly perturbed boundary value problem having discontinuity in convection coefficient and source term*, J. Comput. Appl. Math. 359 (2019), 102–118.
- [24] Schlichting, H. and Gersten, K. *Boundary-layer theory*. springer; 2016.

- [25] Shanthi, V., Ramanujam, N. and Natesan, S. *Fitted mesh method for singularly perturbed reaction-convection-diffusion problems with boundary and interior layers*, *J. Appl. Math. Comput.* 22 (2006), 49–65.
- [26] Singh, S., Choudhary, R. and Kumar, D. *An efficient numerical technique for two-parameter singularly perturbed problems having discontinuity in convection coefficient and source term*, *Comput. Appl. Math.* 42(1) (2023), 62.



Optimal control and a system of quasi-variational inequalities

J. Ogorzaly*

Abstract

We consider the system of two quasi-variational inequalities. First, we show a convergence result with respect the set of constraints. Next, we formulate optimal control problem for which we prove the existence of optimal solution, together with convergence of optimal control pairs. Finally, we apply our results in the study of contact problem with thermal effect and friction.

AMS subject classifications (2020): 35J88, 35J87, 47J20, 49J40, 74B20, 74M10

Keywords: Quasi-variational inequality; Mosco convergence; Optimal control; Thermal effect.

1 Introduction

The theory of optimal control has its origins in the calculus of variations in the 17th century, and it was developed to find the optimal ways to control a dynamical system. The main interest of this theory is the existence and, if it is possible, the uniqueness of optimal state-control pair,

*Corresponding author

Received 8 March 2025; revised 31 October 2025; accepted 27 November 2025

Justyna Ogorzaly

Institute of Mathematics and Cryptology, Military University of Technology, ul. gen. S. Kaliskiego 2, 00-908 Warsaw, Poland, e-mail: justyna.ogorzaly@gmail.com

How to cite this article

Ogorzaly, J., Optimal control and a system of quasi-variational inequalities. *Iran. J. Numer. Anal. Optim.*, 2026; 16(2): 638–666. <https://doi.org/10.22067/ijnao.2025.92539.1611>

as well as the derivation of necessary (or necessary and sufficient) conditions of optimality. The key development of this theory had place in 20th century by works of Bellman, Potryagin, and Kalman. Nowadays, the theory of optimal control has many applications in such areas as, for example, aerospace, process control, robotics, bioengineering, and economics.

In the literature, there are some results related to the optimal control of quasi-variational inequalities of different types. Here, we briefly present a few examples of such works. Sofonea and Tarzia [23] considered an optimal control problem governed by an elliptic quasi-variational inequality with unilateral constraints. The authors defined a new optimal control problem by perturbing the state inequality and the cost functional. Then, they provided sufficient conditions that guarantee the convergence of solutions. The abstract results were used in the study of two models, namely, the static contact problem with normal compliance and unilateral constraint, and a stationary heat transfer problem with unilateral constraints. Khan and Sama [10] studied the optimal control of quasi-variational inequalities with multivalued monotone and pseudomonotone maps. Moreover, they analyzed the convergence behavior of the optimal control problem when the data defining the underlying quasi-variational inequality is contaminated by a certain level of noise. The optimal control problem presented in this paper subsumes some important optimal control problems as particular cases, for example, the optimal control problems for single-valued quasi-variational inequalities, the set-valued/single-valued variational inequalities.

The paper [21] was motivated by mathematical models that describe the time-dependent unilateral contact of a deformable body with a foundation. Taking into account appropriate mechanical assumptions on the constitutive law and the interface conditions, these models lead to a weak formulation in the form of a system that couples an ordinary differential equation with a variational or quasi-variational inequality. The authors proved the continuous dependence of the solution with respect to data, as well as the existence of at least one optimal pair for an associated control problem. Finally, they illustrated the abstract results in the study of a free boundary problem which describes the equilibrium of a viscoelastic body in frictionless contact with a foundation made of a rigid body covered by a rigid-elastic layer. The applications of optimal control theory to variational inequalities can be found in [4, 8, 14, 19, 24, 25, 26].

The system of variational inequalities extends the concept of a variational inequality to multiple interrelated inequalities involving multiple variables. These kind of systems naturally arise in the models where a few processes interact under constraints, and they affect on each another (see, e.g., [16]). The systems of variational inequalities are used to the model an equilibrium problem, for example, the Nash equilibrium problem in the game theory [1]. They are finding applications in fields such as traffic, spatial price equilibrium, mechanics, signal processing, networks, and optimization. This paper extends existing results for optimal control of elliptic quasi-variational inequality to optimal control of the system of elliptic quasi-variational inequalities.

ities. Note that Problem 1 contains several special cases: coupled variational inequalities such that K_1 and K_2 do not vary with the solution (u, θ) and $C_1 = K_1$, $C_2 = K_2$ (see [13]), an elliptic variational inequality of the first kind ($A_2 = 0$, $N_1 = 0$, $N_2 = 0$, $\varphi_2 = 0$, K_1 is independent of u , $g = 0$) and of the second kind ($K_1 = V$); see [9, 12]. Our work inspires future research into optimal control problems for the systems of quasi-variational inequalities of different kinds. To the best of our knowledge, these problems have not been considered in the literature until now.

The paper is organized as follows. In Section 2, we present some notations, and in Section 3, we show the existence and uniqueness result for the system of quasi-variational inequalities as well as the boundedness of the solution. In Section 4, we will show convergence result for the system of two quasi-variational inequalities with respect the set of constraints. Section 5 is devoted to a class of optimal control problem associated to the system of quasi-variational inequalities and existence of optimal pairs. In Section 6, we prove the convergence of the optimal pairs. Finally, in Section 7, we will use abstract results in the study of contact problem with thermal effect.

2 Preliminary

In this section, we collect a basic material that is useful in our paper (see, e.g., [5, 6, 7, 11, 18]). For a Banach space $(X, \|\cdot\|_X)$, we denote by $(X^*, \|\cdot\|_{X^*})$ its dual space and by $\langle \cdot, \cdot \rangle_{X^* \times X}$ the duality pairing between X^* and X . For a Hilbert space Y , we denote the inner product by $(\cdot, \cdot)_Y$. The notation X_w stands for a Banach space X endowed with the weak topology. For a multivalued operator $T: X \rightarrow 2^{X^*}$, its graph $\text{Gr}(T)$ is defined by $\text{Gr}(T) = \{(x, x^*) \in X \times X^* \mid x^* \in Tx\}$. An operator $T: X \rightarrow 2^{X^*}$ is called monotone when $\langle u^* - v^*, u - v \rangle_{X^* \times X} \geq 0$ for all $(u, u^*), (v, v^*) \in \text{Gr}(T)$. Let $\mathbb{R}_+ = [0, +\infty)$. We say that T is u_0 -coercive if for $u_0 \in X$, there exists a function $\theta: \mathbb{R}_+ \rightarrow \mathbb{R}$ with $\lim_{r \rightarrow +\infty} \theta(r) = +\infty$ such that for all $(u, u^*) \in \text{Gr}(T)$, we have $\langle u^*, u - u_0 \rangle_{X^* \times X} \geq \theta(\|u\|_X) \|u\|_X$. An operator T is bounded if T maps bounded sets of X into bounded sets of X^* . A single-valued operator $A: X \rightarrow X^*$ is called pseudomonotone if it is bounded and $u_n \rightarrow u$ in X_w with $\limsup \langle Au_n, u_n - u \rangle_{X^* \times X} \leq 0$ implies $\langle Au, u - v \rangle_{X^* \times X} \leq \liminf \langle Au_n, u_n - v \rangle_{X^* \times X}$ for all $v \in X$. If X is a reflexive Banach space, then $A: X \rightarrow X^*$ is pseudomonotone, if and only if it is bounded and $u_n \rightarrow u$ in X_w with $\limsup \langle Au_n, u_n - u \rangle_{X^* \times X} \leq 0$ entail $\lim \langle Au_n, u_n - u \rangle_{X^* \times X} = 0$ and $Au_n \rightarrow Au$ in X_w^* . A map $K: E \rightarrow F$ between Banach spaces E and F is completely continuous if for any sequence $\{x_n\} \subset E$ converging weakly to $x \in E$, we have $K(x_n) \rightarrow K(x)$ in F . Let Z be a normed space, and let $C, \{C_n\}$ be closed and convex subsets of Z . The sequence $\{C_n\}$ Mosco converges to C , denoted by $C_n \xrightarrow{M} C$ as $n \rightarrow \infty$, if we have

- for any $z_n \in C_n$ with $z_n \rightarrow z$ in Z_w , up to a subsequence, we have $z \in C$,
- for any $z \in C$, there exists $z_n \in C_n$ with $z_n \rightarrow z$ in Z .

3 Solvability of a system of quasi-variational inequalities

Let V and E be reflexive Banach, and let Y_1, Y_2 be the real Hilbert spaces. Let $C_1 \subset V$ and $C_2 \subset E$ be given subsets, and let $A_1: V \rightarrow V^*$, $N_1: E \rightarrow V^*$, $N_2: V \rightarrow E^*$, $A_2: E \rightarrow E^*$ be given operators. We have also given functions $\varphi_1: V \times E \times V \rightarrow \mathbb{R}$, $\varphi_2: V \times E \times E \rightarrow \mathbb{R}$, set-valued maps $K_1: C_1 \times C_2 \rightarrow 2^{C_1} \setminus \{\emptyset\}$, $K_2: C_1 \times C_2 \rightarrow 2^{C_2} \setminus \{\emptyset\}$, functions $f \in Y_1, g \in Y_2$, and operators $\pi_1: V \rightarrow Y_1$ and $\pi_2: E \rightarrow Y_2$.

The problem is as follows.

Problem 1. Find $(u, \theta) \in C_1 \times C_2$ such that $u \in K_1(u, \theta)$, $\theta \in K_2(u, \theta)$ and for all $v \in K_1(u, \theta), \eta \in K_2(u, \theta)$ hold

$$\begin{aligned} \langle A_1 u + N_1 \theta, v - u \rangle_{V^* \times V} + \varphi_1(u, \theta, v) - \varphi_1(u, \theta, u) &\geq (f, \pi_1 v - \pi_2 u)_{Y_1}, \\ \langle A_2 \theta + N_2 u, \eta - \theta \rangle_{E^* \times E} + \varphi_2(u, \theta, \eta) - \varphi_2(u, \theta, \theta) &\geq (g, \pi_2 \eta - \pi_2 \theta)_{Y_2}. \end{aligned} \quad (1)$$

The hypotheses on the data of Problem 1 are the following:

$$C_1 \subset V, C_2 \subset E \quad \text{are nonempty, closed, and convex sets,} \quad (2)$$

$K_1: C_1 \times C_2 \rightarrow 2^{C_1} \setminus \{\emptyset\}$ is such that

- K_1 has closed convex values in C_1 ,
 - K_1 is weakly Mosco continuous, that is, for any $\{(u_n, \theta_n)\} \in C_1 \times C_2$ such that $(u_n, \theta_n) \rightarrow (u, \theta)$ in $(V \times E)_w$, one has $K_1(u_n, \theta_n) \xrightarrow{M} K_1(u, \theta)$,
 - there is a bounded set $K_0 \subset C_1$ such that $K_1(u, \theta) \cap K_0 \neq \emptyset$ for all $(u, \theta) \in C_1 \times C_2$,
- (3)

$K_2: C_1 \times C_2 \rightarrow 2^{C_2} \setminus \{\emptyset\}$ is such that

- K_2 has closed convex values in C_2 ,
 - K_2 is weakly Mosco continuous, that is, for any $\{(u_n, \theta_n)\} \in C_1 \times C_2$ such that $(u_n, \theta_n) \rightarrow (u, \theta)$ in $(V \times E)_w$, one has $K_2(u_n, \theta_n) \xrightarrow{M} K_2(u, \theta)$,
 - there is a bounded set $K'_0 \subset C_2$ such that $K_2(u, \theta) \cap K'_0 \neq \emptyset$ for all $(u, \theta) \in C_1 \times C_2$,
- (4)

$A_1: V \longrightarrow V^*$ and $A_2: E \longrightarrow E^*$ are such that

- (a) $\|A_1 v\|_{V^*} \leq a_1 + a_2 \|v\|_V$ for all $v \in V$ with $a_1, a_2 \geq 0$,
- (b) $\|A_2 \eta\|_{E^*} \leq b_1 + b_2 \|\eta\|_E$ for all $\eta \in E$ with $b_1, b_2 \geq 0$,
- (c) A_1, A_2 are pseudomonotone operators,

(5)

$N_1: E \longrightarrow V^*, N_2: V \longrightarrow E^*$ are such that

- (a) N_1, N_2 are completely continuous,
- (b) $\|N_1 \theta\|_{V^*} \leq m_1(1 + \|\theta\|_E)$ for all $\theta \in E$, $m_1 > 0$,
- (c) $\|N_2 v\|_{E^*} \leq m_2(1 + \|v\|_E)$ for all $v \in V$, $m_2 > 0$,

(6)

$f \in Y_1, \quad g \in Y_2,$

(7)

$\pi_1: V \rightarrow Y_1, \quad \pi_2: E \rightarrow Y_2$ are linear continuous operators such that

- $\|\pi_1 v\|_{Y_1} \leq d_1 \|v\|_V$ for every $v \in V$ with $d_1 > 0$,
- $\|\pi_2 \theta\|_{Y_2} \leq d_2 \|\theta\|_E$ for every $\theta \in E$ with $d_2 > 0$,

(8)

$\varphi_1: V \times E \times V \rightarrow \mathbb{R}$ is such that

- (a) $\varphi_1(w, \theta, \cdot)$ is convex and lower semicontinuous for all $w \in V, \theta \in E$,
- (b) there exists a function $c_{\varphi_1}: \mathbb{R}_+ \times \mathbb{R}_+ \rightarrow \mathbb{R}_+$ bounded on bounded sets such that $\varphi_1(w, \theta, v_1) - \varphi_1(w, \theta, v_2) \leq c_{\varphi_1}(\|w\|_V, \|\theta\|_E) \|v_1 - v_2\|_V$ for all $w, v_1, v_2 \in V, \theta \in E$,
- (c) for all $\{w_n\} \subset V, \{\theta_n\} \subset E, \{v_n\} \subset V, \{u_n\} \subset V$ such that $w_n \rightarrow w$ in $V, \theta_n \rightarrow \theta$ in $E, v_n \rightarrow v$ in $V, u_n \rightarrow u$ in V for some $w, u, v \in V, \theta \in E$, it holds

$$\limsup (\varphi_1(w_n, \theta_n, v_n) - \varphi_1(w_n, \theta_n, u_n)) \leq \varphi_1(w, \theta, v) - \varphi_1(w, \theta, u),$$

(9)

$\varphi_2: V \times E \times E \rightarrow \mathbb{R}$ is such that

- (a) $\varphi_2(w, \theta, \cdot)$ is convex and lower semicontinuous for all $w \in V, \theta \in E$,
- (b) there exists a function $c_{\varphi_2}: \mathbb{R}_+ \times \mathbb{R}_+ \rightarrow \mathbb{R}_+$ bounded on bounded sets such that $\varphi_2(w, \theta, \eta_1) - \varphi_2(w, \theta, \eta_2) \leq c_{\varphi_2}(\|w\|_V, \|\theta\|_E) \|\eta_1 - \eta_2\|_E$ for all $w \in V, \theta, \eta_1, \eta_2 \in E$,
- (c) for all $\{w_n\} \subset V, \{\theta_n\} \subset E, \{\eta_n\} \subset E, \{\zeta_n\} \subset E$ such that $w_n \rightarrow w$ in $V, \theta_n \rightarrow \theta$ in $E, \eta_n \rightarrow \eta$ in $E, \zeta_n \rightarrow \zeta$ in E for some $w \in V, \theta, \eta, \zeta \in E$, it holds

$$\limsup (\varphi_2(w_n, \theta_n, \eta_n) - \varphi_2(w_n, \theta_n, \zeta_n)) \leq \varphi_2(w, \theta, \eta) - \varphi_2(w, \theta, \zeta).$$

(10)

Moreover, we assume that

- (a) there exist $u_0 \in K_0$ and a function $\theta_1: \mathbb{R}_+ \rightarrow \mathbb{R}$ with $\theta_1(r) \rightarrow +\infty$ as $r \rightarrow +\infty$ such that for all $v \in V$, it holds $\langle A_1 v, v - u_0 \rangle \geq \theta_1(\|v\|)\|v\|$,
- (b) there exist $u_1 \in K'_0$ and a function $\theta_2: \mathbb{R}_+ \rightarrow \mathbb{R}$ with $\theta_2(r) \rightarrow +\infty$ as $r \rightarrow +\infty$ such that for all $\eta \in E$, it holds
$$\langle A_2 \eta, \eta - u_1 \rangle_{E^* \times E} \geq \theta_2(\|\eta\|_E)\|\eta\|_E, \quad (11)$$

$$V \ni v \mapsto A_1 v \in 2^{V^*} \text{ and } E \ni \eta \mapsto A_2 \eta \in 2^{E^*} \text{ are a monotone map.} \quad (12)$$

By hypothesis (11) and (5), it is clear (see [16, Section 3.4]) that A_1 and A_2 are pseudomonotone and bounded as a multivalued operators from V to 2^{V^*} and E to 2^{E^*} , respectively.

Theorem 1. Under the assumptions (2)–(12), Problem 1 has a solution.

Proof. The proof of Theorem 1 is a simplified version of the proof in [17, Theorem 10]. However, for convenience of the reader, we briefly recall the main steps of this proof. The important tool in the proof is the following version of Kakutani–Ky Fan fixed point theorem in a reflexive Banach space (see, e.g., [6, Corollary 1.7.42]). $\square$

Theorem 2 (Kakutani-Ky Fan). Let Y be a reflexive Banach space, and let $D \subseteq Y$ be a nonempty, bounded, closed, and convex set. Suppose that $\Lambda: D \rightarrow 2^D$ is a set-valued map such that the values of Λ are nonempty and convex sets, and $\text{Gr } \Lambda$ is sequentially closed in $Y_w \times Y_w$. Then Λ has a fixed point: $y^* \in D$ such that $y^* \in \Lambda y^*$.

Remark 1. In Theorem 2, the sequential closedness of the graph of Λ in $Y_w \times Y_w$ implies that the values of Λ are closed subsets of Y_w . By hypothesis, the values are convex sets. Hence and Mazur's theorem, we deduce that the values of Λ are also closed sets in Y .

Let $\langle \bar{f}, v \rangle_{V^* \times V} = \langle f, \pi_1 v \rangle_{Y_1}$ and $\langle \bar{g}, \eta \rangle_{E^* \times E} = \langle g, \pi_2 \eta \rangle_{Y_2}$ for every $v \in V, \eta \in E$. Moreover, let $(w, \xi) \in C_1 \times C_2 \subset V \times E$ be fixed. Consider the following auxiliary problem:

Problem 2. Find $(u, \theta) \in K_1(w, \xi) \times K_2(w, \xi)$ such that

$$\begin{aligned} \langle A_1 u + N_1 \xi - \bar{f}, v - u \rangle_{V^* \times V} + \varphi(w, \xi, v) - \varphi(w, \xi, u) &\geq 0 \\ \text{for all } v &\in K_1(w, \xi), \\ \langle A_2 \theta + N_2 w - \bar{g}, \eta - \theta \rangle_{E^* \times E} + \psi(w, \xi, \eta) - \psi(w, \xi, \theta) &\geq 0 \\ \text{for all } \eta &\in K_2(w, \xi). \end{aligned} \quad (13)$$

We define a set-valued map $S: C_1 \times C_2 \rightarrow 2^{C_1 \times C_2}$ by

$$S(w, \xi) := \{(u, \theta) \in K_1(w, \xi) \times K_2(w, \xi) \mid (u, \theta) \text{ is a solution to Problem 2}\}. \quad (14)$$

The system (13) is decoupled, therefore, we shall examine each inequality separately. Let $D := \{(w, \xi) \in C_1 \times C_2 \mid \|w\|_V \leq r, \|\xi\|_E \leq r\}$ with a suitable positive constant r . It is obvious that D is a nonempty, closed, convex, and bounded subset of $V \times E$. Let $\Lambda: D \rightarrow 2^D$ be given by $\Lambda(w, \xi) := S(w, \xi)$ for $(w, \xi) \in D$. In order to apply Theorem 2 with a reflexive Banach space $Y := V \times E$, we shall verify

- for each $(w, \xi) \in C_1 \times C_2$, $S(w, \xi)$ defined by (14) is nonempty and convex,
- $\Lambda(D) \subset D$, that is, there exists $r > 0$ for which Λ maps D into itself,
- $\text{Gr } \Lambda$ is sequentially closed in $(V \times E)_w \times (V \times E)_w$.

The details of the proofs of these properties are similar as in [17, Theorem 10], so we skip them here. For the second inequality in Problem 1, we use the same arguments as above, and we obtain that $\theta \in K_2(w, \xi)$ satisfies the limit inequality. Finally, we infer that $(u, \theta) \in S(w, \xi)$, which completes the proof of the closedness of the graph of Λ .

Having verified the hypotheses of Theorem 2, we deduce from it that there exists a fixed point $(u^*, \theta^*) \in D$ of Λ , that is, $(u^*, \theta^*) \in \Lambda(u^*, \theta^*) = S(u^*, \theta^*)$. In conclusion, $(u^*, \theta^*) \in K_1(u^*, \theta^*) \times K_2(u^*, \theta^*)$ is a solution to Problem 1.

In order to prove the uniqueness of the solution of Problem 1, we need the following additional assumptions:

$$\left. \begin{array}{l} \text{(a) } \langle A_1 u_1 - A_1 u_2, u_1 - u_2 \rangle_{V^* \times V} \geq m_{A_1} \|u_1 - u_2\|_V^2 \text{ with } m_{A_1} > 0, \\ \qquad \qquad \qquad \text{for all } u_1, u_2 \in V, \\ \text{(b) } \langle N_2 \theta_1 - N_2 \theta_2, \theta_1 - \theta_2 \rangle_{E^* \times E} \geq m_{N_2} \|\theta_1 - \theta_2\|_E^2 \text{ with } m_{N_2} > 0, \\ \qquad \qquad \qquad \text{for all } \theta_1, \theta_2 \in E, \\ \text{(c) } \|N_1 \theta_1 - N_1 \theta_2\|_E \leq L_{N_1} \|\theta_1 - \theta_2\|_E \text{ with } L_{N_1} > 0, \text{ for all } \theta_1, \theta_2 \in E, \\ \text{(d) } \|A_2 u_1 - A_2 u_2\|_V \leq L_{A_2} \|u_1 - u_2\|_V \text{ with } L_{A_2} > 0, \text{ for all } u_1, u_2 \in V, \\ \text{(e) } \varphi_1(u_1, \theta_1, u_2) - \varphi_1(u_1, \theta_1, u_1) + \varphi_1(u_2, \theta_2, u_1) - \varphi_1(u_2, \theta_2, u_2) \\ \qquad \leq \theta_{\varphi_1} (\|u_1 - u_2\|_V + \|\theta_1 - \theta_2\|_E) \|u_1 - u_2\|_V \\ \qquad \text{with } \theta_{\varphi_1} > 0, \text{ for all } u_1, u_2 \in V \text{ and } \theta_1, \theta_2 \in E, \\ \text{(f) } \varphi_2(u_1, \theta_1, \theta_2) - \varphi_2(u_1, \theta_1, \theta_1) + \varphi_2(u_2, \theta_2, \theta_1) - \varphi_2(u_2, \theta_2, \theta_2) \\ \qquad \leq \theta_{\varphi_2} (\|u_1 - u_2\|_V + \|\theta_1 - \theta_2\|_E) \|\theta_1 - \theta_2\|_E \\ \qquad \text{with } \theta_{\varphi_2} > 0, \text{ for all } u_1, u_2 \in V \text{ and } \theta_1, \theta_2 \in E. \end{array} \right\} \quad (15)$$

Theorem 3. If (2)–(12), (15) hold and

$$m_{A_1} > \theta_{\varphi_1}, \quad m_{N_2} > \theta_{\varphi_2}, \quad \frac{(L_{N_1} + \theta_{\varphi_1})(L_{A_2} + \theta_{\varphi_2})}{(m_{A_1} - \theta_{\varphi_1})(m_{N_2} - \theta_{\varphi_2})} < 1, \quad (16)$$

then Problem 1 has a unique solution.

Proof. Let $(u_1, \theta_1), (u_2, \theta_2) \in C_1 \times C_2$ be such that $u_1, u_2 \in K_1(u, \theta)$ and $\theta_1, \theta_2 \in K_2(u, \theta)$. We take the first inequality in (1) with $v = u_2$ and $v = u_1$, respectively. Hence, we have

$$\begin{aligned} & \langle A_1 u_1 + N_1 \theta_1, u_2 - u_1 \rangle_{V^* \times V} + \varphi_1(u_1, \theta_1, u_2) - \varphi_1(u_1, \theta_1, u_1) \\ & \geq (f, \pi_1 u_2 - \pi_1 u_1)_{Y_1}, \\ & \langle A_1 u_2 + N_1 \theta_2, u_1 - u_2 \rangle_{V^* \times V} + \varphi_1(u_2, \theta_2, u_1) - \varphi_1(u_2, \theta_2, u_2) \\ & \geq (f, \pi_2 u_1 - \pi_2 u_2)_{Y_1}. \end{aligned}$$

Adding the inequalities above and using (15)(a), (c), (e), we obtain

$$(m_{A_1} - \theta_{\varphi_1}) \|u_1 - u_2\|_V \leq (L_{N_1} + \theta_{\varphi_1}) \|\theta_1 - \theta_2\|_E. \quad (17)$$

Taking $\eta = \theta_1$ and $\eta = \theta_2$ in the second inequality in (1) and using (15)(b), (d), (f), we have

$$(m_{N_2} - \theta_{\varphi_2}) \|\theta_1 - \theta_2\|_E \leq (L_{A_2} + \theta_{\varphi_2}) \|u_1 - u_2\|_V. \quad (18)$$

From (17) and (18), we obtain

$$\|u_1 - u_2\|_V \leq \frac{(L_{N_1} + \theta_{\varphi_1})(L_{A_2} + \theta_{\varphi_2})}{(m_{A_1} - \theta_{\varphi_1})(m_{N_2} - \theta_{\varphi_2})} \|u_1 - u_2\|_V. \quad (19)$$

Combining inequality (19) with the smallness condition (16), we conclude that $u_1 = u_2$. Then, from the inequality (18) we deduce that $\theta_1 = \theta_2$. Consequently, Problem 1 has a unique solution. $\square$

Remark 2. Removing the smallness assumption (16) remains a task for future research, since it is made for mathematical reasons, and does not seem to relate to any inherent physical constraints of the contact problem considered in Section 7.

Finally, we present the result we will use in the convergence for optimal pairs. The assumption, which we need in this step, are as follows:

- (a) $\mathbf{0}_V \in K_1(u, \theta)$, $\mathbf{0}_E \in K_2(u, \theta)$,
- (b) $\langle A_2\theta_1 - A_2\theta_2, \theta_1 - \theta_2 \rangle_{E^* \times E} \geq m_{A_2} \|\theta_1 - \theta_2\|_E^2$ with $m_{A_2} > 0$
for all $\theta_1, \theta_2 \in E$,
- (c) $\varphi_1: V \times E \times V \rightarrow \mathbb{R}$ is such that for all $u, v \in V$, $\theta \in E$, and for $\lambda > 0$,
we have $\varphi_1(u, \theta, \lambda v) = \lambda \varphi_1(u, \theta, v)$, $\varphi_1(v, \theta, v) \geq 0$.
- By analogy we make assumptions for the function φ_2 .

(20)

The result is as follows.

Lemma 1. Let $f \in Y_1$ and $g \in Y_2$ be the given functions, let the assumptions (6)(b)-(c), (15)(a), (8) hold, and let

$$m_{A_1} m_{A_2} > m_2 m_1. \quad (21)$$

Then the solutions $u = u(f, g)$ and $\theta = \theta(f, g)$ of the system (1) satisfy the bounds

$$\|u\|_V \leq c_1 \quad \text{and} \quad \|\theta\|_E \leq c_2, \quad (22)$$

respectively, where c_1, c_2 are constants.

Proof. First, we use (20)(a), and we take $v = \mathbf{0}_V \in K_1(u, \theta)$ in the first inequality in the system (1). Using assumptions (6)(b), (15)(a), (8), and (20), we infer that

$$\|u\|_V \leq \frac{1}{m_{A_1}} (\|A_1 \mathbf{0}_V\|_{V^*} + m_1(1 + \|\theta\|_E) + d_1 \|f\|_{Y_1}). \quad (23)$$

Next, we use (20)(a) and we take $\mathbf{0}_E \in K_2(u, \theta)$ in the second inequality in (1). Using (6)(c), (8) and (20) in the obtained inequality, we get

$$\|\theta\|_E \leq \frac{1}{m_{A_2}} (\|A_2 \mathbf{0}_E\|_{E^*} + m_2(1 + \|u\|_V) + d_2 \|g\|_{Y_2}). \quad (24)$$

Combining (23), (24), and (21), we deduce (22). $\square$

4 A convergence result for the system of quasi-variational inequalities

The convergence result presented in this section is important in the study of optimal control problems associated with the system (1). We assume that the assumptions (2)–(12), (15)(a),

(16) hold and that

(a) $\{K_{1n}(u, \theta)\}_{n \in \mathbb{N}}$ converges to $K_1(u, \theta)$ in the sense of Mosco as $n \rightarrow \infty$.

The similar assumption is for the sets $K_2(u, \theta)$ and $\{K_{2n}(u, \theta)\}_{n \in \mathbb{N}}$.

- (b) $\langle A_2\theta - A_2\eta, \theta - \eta \rangle_{E^* \times E} \geq m_{A_2} \|\theta - \eta\|_E^2$, with $m_{A_2} > 0$, for all $\eta, \theta \in E$,
 (c) $f_n \rightharpoonup f$ in Y_1 as $n \rightarrow \infty$,
 (d) $g_n \rightharpoonup g$ in Y_2 as $n \rightarrow \infty$,
 (e) for all sequence $\{v_n\} \subset V$ such that $v_n \rightharpoonup v$ in V ,
 we have $\pi_1 v_n \rightarrow \pi_1 v$ in Y_1 ,
 (f) for all sequence $\{\eta_n\} \subset E$ such that $\eta_n \rightharpoonup \eta$ in E ,
 we have $\pi_2 \eta_n \rightarrow \pi_2 \eta$ in Y_2 .

(25)

Problem 3. Find $(u_n, \theta_n) \in C_1 \times C_2$ such that $u_n \in K_{1n}(u_n, \theta_n)$, $\theta_n \in K_{2n}(u_n, \theta_n)$ and

$$\begin{aligned} & \langle A_1 u_n + N_1 \theta_n, v - u_n \rangle_{V^* \times V} + \varphi_1(u_n, \theta_n, v) - \varphi_1(u_n, \theta_n, u_n) \\ & \geq (f_n, \pi_1 v - \pi_1 u_n)_{Y_1} \quad \text{for all } v \in K_{1n}(u_n, \theta_n), \\ & \langle A_2 \theta_n + N_2 u_n, \eta - \theta_n \rangle_{E^* \times E} + \varphi_2(u_n, \theta_n, \eta) - \varphi_2(u_n, \theta_n, \theta_n) \\ & \geq (g_n, \pi_2 \eta - \pi_2 \theta_n)_{Y_2} \quad \text{for all } \eta \in K_{2n}(u_n, \theta_n). \end{aligned}$$

(26)

We observe that from Theorem 3, for each $n \in \mathbb{N}$, there exists a unique solution (u_n, θ_n) of Problem 3. The main result in this section is as follows.

Theorem 4. Assume that (2)–(12), (15), (16), and (25) hold. Then the sequence $\{u_n\}$ converges strongly to u in V , and the sequence $\{\theta_n\}$ converges strongly to θ in E .

The proof of Theorem 4 is divided into two steps. The first step of the proof is the following.

Lemma 2. There exist the elements $\hat{u} \in K_{1n}(\hat{u}, \hat{\theta})$ and $\hat{\theta} \in K_{2n}(\hat{u}, \hat{\theta})$ and the subsequences of $\{u_n\}$ and $\{\theta_n\}$, still denoted by $\{u_n\}$, $\{\theta_n\}$ respectively, such that, as $n \rightarrow \infty$, we have

$$\begin{aligned} u_n & \rightharpoonup \hat{u} \text{ in } V, \\ \theta_n & \rightharpoonup \hat{\theta} \text{ in } E. \end{aligned}$$

Proof. First, we will show the boundedness of $\{u_n\}$ in V and $\{\theta_n\}$ in E . Let $v \in K_1(u, \theta)$ and $\eta \in K_2(u, \theta)$ be given. We assume (3), and we consider the sequences $\{v_n\}$, $\{\eta_n\}$ such that $v_n \in K_{1n}(u_n, \theta_n)$, $\eta_n \in K_{2n}(u_n, \theta_n)$ for all $n \in \mathbb{N}$ and

$$v_n \longrightarrow v \text{ in } V, \quad \eta_n \longrightarrow \eta \text{ in } E. \quad (27)$$

Let $n \in \mathbb{N}$. Taking the first inequality in (26) with $v_n \in K_{1n}(u_n, \theta_n)$ and $\eta_n \in K_{2n}(u_n, \theta_n)$, we obtain

$$\begin{aligned} & \langle A_1 u_n + N_1 \eta_n, v_n - u_n \rangle_{V^* \times V} + \varphi_1(u_n, \eta_n, v_n) - \varphi_1(u_n, \eta_n, u_n) \\ & \geq (f_n, \pi_1 v_n - \pi_1 u_n)_{Y_1}, \\ & \langle A_1 u_n, -v_n + u_n \rangle_{V^* \times V} \leq \langle N_1 \eta_n, v_n - u_n \rangle_{V^* \times V} + \varphi_1(u_n, \eta_n, v_n) \\ & \quad - \varphi_1(u_n, \eta_n, u_n) + (f_n, \pi_1 u_n - \pi_1 v_n)_{Y_1}. \end{aligned}$$

Using (15)(a), we get

$$\begin{aligned} \langle A_1 u_n, u_n - v_n \rangle_{V^* \times V} &= \langle A_1 u_n - A_1 v_n, u_n - v_n \rangle_{V^* \times V} \\ & \quad + \langle A_1 v_n, u_n - v_n \rangle_{V^* \times V} \\ & \geq m_{A_1} \|u_n - v_n\|_V^2 + \langle A_1 v_n, u_n - v_n \rangle_{V^* \times V}. \end{aligned}$$

From (9)(b), we obtain

$$\varphi_1(u_n, \eta_n, v_n) - \varphi_1(u_n, \eta_n, u_n) \leq c_{\varphi_1}(\|u_n\|_V, \|\eta_n\|_E) \|v_n - u_n\|_V,$$

and from (6)(b), we have

$$|\langle N_1 \eta_n, v_n - u_n \rangle_{V^* \times V}| \leq \|N_1 \eta_n\|_{V^*} \|v_n - u_n\|_V \leq m_1(1 + \|\eta_n\|_E) \|v_n - u_n\|_V.$$

Moreover, from (8), we conclude that

$$|(f_n, \pi_1 v_n - \pi_1 u_n)_{Y_1}| \leq \|f_n\|_{Y_1} \|\pi_1 v_n - \pi_1 u_n\|_V \leq \|f_n\|_{Y_1} d_1 \|v_n - u_n\|_V,$$

hence

$$\begin{aligned} m_{A_1} \|u_n - v_n\|_V^2 &\leq \|A_1 v_n\|_{V^*} \|v_n - u_n\|_V + m_1(1 + \|\eta_n\|_E) \|v_n - u_n\|_V \\ & \quad + c_{\varphi_1}(\|u_n\|_V, \|\eta_n\|_E) \|v_n - u_n\|_V + \|f_n\|_{Y_1} d_1 \|v_n - u_n\|_V. \end{aligned}$$

From the inequality above and (5)(a), we get

$$\begin{aligned} m_{A_1} \|u_n - v_n\|_V &\leq a_1 + a_2 \|v_n\|_V + m_1(1 + \|\eta_n\|_E) + c_{\varphi_1}(\|u_n\|_V, \|\eta_n\|_E) \\ & \quad + d_1 \|f_n\|_{Y_1}. \end{aligned} \quad (28)$$

The relations (27) and (25)(c) imply that the sequences $\{v_n\}$, $\{\eta_n\}$ and $\{f_n\}$ are bounded in V , E and Y_1 , respectively. Therefore, from the inequality (28) we deduce that there is a constant $c > 0$ independent on n such that $\|u_n - v_n\|_V \leq c$. So, $\{u_n\}$ is a bounded sequence in V , and from reflexivity of V , passing to a subsequence, if necessary, we deduce that

$$u_n \rightharpoonup \hat{u} \text{ in } V \text{ as } n \rightarrow \infty. \quad (29)$$

Similarly, we can show that $\|\theta_n - \eta_n\|_E \leq c$, and from this, $\{\theta_n\}$ is a bounded sequence in E . Also, the reflexivity of E leads to the conclusion that

$$\theta_n \rightharpoonup \hat{\theta} \text{ in } E \text{ as } n \rightarrow \infty. \quad (30)$$

Here $\hat{u} \in V$ and $\hat{\theta} \in E$. Since (29), (30) hold, we can use the Mosco convergence (25)(a) to obtain

$$\hat{u} \in K_1(u_n, \theta_n) \text{ and } \hat{\theta} \in K_2(u_n, \theta_n). \quad (31)$$

Now, we can use to (31) the assumption that the sets $K_1(u_n, \theta_n)$ and $K_2(u_n, \theta_n)$ are weakly Mosco continuous (see (3)(b), (4)(b)), to obtain $\hat{u} \in K_1(\hat{u}, \hat{\theta})$ and $\hat{\theta} \in K_2(\hat{u}, \hat{\theta})$. $\square$

Lemma 3. The elements $\hat{u} \in K_1(\hat{u}, \hat{\theta})$ and $\hat{\theta} \in K_2(\hat{u}, \hat{\theta})$ are the solutions to Problem 1.

Proof. From the condition (25)(a) of the Mosco convergence, for each $v, \bar{u} \in K_1(u, \theta)$, we can find the sequences $\{v_n\}$, $\{\bar{u}_n\} \subset K_{1n}(u_n, \theta_n)$ such that $v_n \rightarrow v$, $\bar{u}_n \rightarrow \hat{u}$ in V . From the first inequality in (26) with $v = \bar{u}_n$, we get

$$\begin{aligned} \langle A_1 u_n + N_1 \theta_n, \bar{u}_n - u_n \rangle_{V^* \times V} + \varphi_1(u_n, \theta_n, \bar{u}_n) - \varphi_1(u_n, \theta_n, u_n) \\ \geq (f_n, \pi_1 \bar{u}_n - \pi_1 u_n)_{Y_1}. \end{aligned} \quad (32)$$

From (32), we see that

$$\begin{aligned} \langle A_1 u_n, u_n - \bar{u}_n \rangle_{V^* \times V} \leq \langle N_1 \theta_n, \bar{u}_n - u_n \rangle_{V^* \times V} + \varphi_1(u_n, \theta_n, \bar{u}_n) \\ - \varphi_1(u_n, \theta_n, u_n) + (f_n, \pi_1 u_n - \pi_1 \bar{u}_n)_{Y_1} \end{aligned}$$

and moreover, we have

$$\begin{aligned} \langle A_1 u_n, u_n - \hat{u} \rangle_{V^* \times V} &= \langle A_1 u_n, (u_n - \bar{u}_n) + (\bar{u}_n - \hat{u}) \rangle_{V^* \times V} \\ &= \langle A_1 u_n, u_n - \bar{u}_n \rangle_{V^* \times V} + \langle A_1 u_n, \bar{u}_n - \hat{u} \rangle_{V^* \times V}. \end{aligned}$$

Thus

$$\begin{aligned} \limsup \langle A_1 u_n, u_n - \widehat{u} \rangle_{V^* \times V} &\leq \limsup (\langle N_1 \theta_n, \bar{u}_n - u_n \rangle_{V^* \times V} + \varphi_1(u_n, \theta_n, \bar{u}_n) \\ &\quad - \varphi_1(u_n, \theta_n, u_n) + (f_n, \pi_1 u_n - \pi_2 \bar{u}_n)_{Y_1} \\ &\quad + \langle A_1 u_n, \bar{u}_n - \widehat{u} \rangle_{V^* \times V}). \end{aligned}$$

Hence $\limsup \langle A_1 u_n, u_n - \widehat{u} \rangle_{V^* \times V} \leq 0$. From the fact that $u_n \rightharpoonup \widehat{u}$ in V and the pseudomonotonicity of A_1 , we obtain

$$\langle A_1 \widehat{u}, \widehat{u} - v \rangle_{V^* \times V} \leq \liminf \langle A_1 u_n, u_n - v \rangle_{V^* \times V} \text{ for all } v \in V. \quad (33)$$

On the other hand, we choose $v = v_n \in K_{1n}(u_n, \theta_n)$ in the first inequality of (26). From (32), we see that

$$\liminf \langle A_1 u_n, u_n - v \rangle_{V^* \times V} \leq \limsup \langle A_1 u_n, u_n - v \rangle_{V^* \times V}. \quad (34)$$

Using (34) into the inequality

$$\begin{aligned} &\limsup \langle A_1 u_n, u_n - v \rangle_{V^* \times V} \\ &= \limsup (\langle A_1 u_n, (u_n - v_n) + (v_n - v) \rangle_{V^* \times V}) \\ &= \limsup (\langle A_1 u_n, u_n - v_n \rangle_{V^* \times V} + \langle A_1 u_n, v_n - v \rangle_{V^* \times V}) \\ &\leq \limsup (\langle N_1 \theta_n, v_n - u_n \rangle_{V^* \times V} + \varphi_1(u_n, \theta_n, v_n) - \varphi(u_n, \theta_n, u_n) \\ &\quad + (f_n, \pi_1 u_n - \pi_2 v_n)_{Y_1} + \langle A_1 u_n, v_n - v \rangle_{V^* \times V}), \end{aligned}$$

and from (33), (34), (29), (27), and (25)(c),(e), we conclude

$$\begin{aligned} \langle A_1 \widehat{u}, \widehat{u} - v \rangle_{V^* \times V} &\leq \langle N_1 \widehat{\theta}, v - \widehat{u} \rangle_{V^* \times V} + \varphi_1(\widehat{u}, \widehat{\theta}, v) - \varphi_1(\widehat{u}, \widehat{\theta}, \widehat{u}) \\ &\quad + (\widehat{f}, \pi_1 \widehat{u} - \pi_1 v)_{Y_1} \quad \text{for all } v \in K_1(\widehat{u}, \widehat{\theta}). \end{aligned} \quad (35)$$

So, $\widehat{u}$ satisfies the first inequality in (26). Using similar arguments, we can prove that $\widehat{\theta} \in K_2(\widehat{u}, \widehat{\theta})$ satisfies the second inequality in (26). $\square$

Proof of Theorem 4. From the fact that Problem 1 has a unique solution (u, θ) and Lemma 3, we deduce that $\widehat{u} = u$ and $\widehat{\theta} = \theta$. Hence, every subsequence of $\{u_n\}$, $\{\theta_n\}$, which is convergent weakly in V and E , respectively, has the same limit. So, the whole sequences $\{u_n\}$ converges weakly in V to u as $n \rightarrow \infty$ and $\{\theta_n\}$ converges weakly in E to θ as $n \rightarrow \infty$.

We now prove that $u_n \rightarrow u$ in V and $\theta_n \rightarrow \theta$ in E as $n \rightarrow \infty$. We take $v = \hat{u} \in K_1(\hat{u}, \hat{\theta})$ in (35), and we use $\hat{u} = u$. Then

$$0 \leq \liminf \langle A_1 u_n, u_n - u \rangle_{V^* \times V} \leq \limsup \langle A_1 u_n, u_n - u \rangle_{V^* \times V} \leq 0,$$

which shows that $\langle A_1 u_n, u_n - u \rangle_{V^* \times V} \rightarrow 0$ as $n \rightarrow \infty$. From this, (15)(a), taking $\hat{u} = u$ and $\langle A_1 u_n, u_n - u \rangle_{V^* \times V} = \langle A_1 u_n - A_1 u, u_n - u \rangle_{V^* \times V} + \langle A_1 u, u_n - u \rangle_{V^* \times V}$, we obtain

$$\begin{aligned} \langle A_1 u_n, u_n - u \rangle_{V^* \times V} - \langle A_1 u, u_n - u \rangle_{V^* \times V} &= \langle A_1 u_n - A_1 u, u_n - u \rangle_{V^* \times V} \\ &\geq m_{A_1} \|u_n - u\|_V^2. \end{aligned}$$

The left-hand side of the equality above convergence to zero, as $n \rightarrow \infty$, hence $u_n \rightarrow u$ in V . Similarly, we prove that $\theta_n \rightarrow \theta$ in E . $\square$

5 Optimal control problem

In this section, we will consider the optimal control problem associated to (1). We will control the solution of the system (1) with $f \in Y_1$ and $g \in Y_2$. We define the set of admissible pair

$$\mathcal{V} = \{(u, \theta) \in C_1 \times C_2 \text{ and } (f, g) \in Y_1 \times Y_2 \text{ such that (1) holds}\}. \quad (36)$$

We deduce from here that the pairs (u, θ) and (f, g) belong to $\mathcal{V}$ if and only if $f \in Y_1$, $g \in Y_2$ and (u, θ) is the solution of the system (1), that is, $u = u(f, g)$, $\theta = \theta(f, g)$. Let $\mathcal{L}: V \times E \times Y_1 \times Y_2 \rightarrow \mathbb{R}$ be a cost functional. Thus, the optimal control problem is as follows.

Problem 4. Find $(u^*, \theta^*), (f^*, g^*) \in \mathcal{V}$ such that

$$\mathcal{L}(u^*, \theta^*, f^*, g^*) = \min_{(u, \theta), (f, g) \in \mathcal{V}} \mathcal{L}(u, \theta, f, g).$$

We assume that for all $(u, \theta) \in V \times E$ and $(f, g) \in Y_1 \times Y_2$, we have

$$\mathcal{L}(u, \theta, f, g) = U(u, \theta) + F(f, g), \quad (37)$$

where

$$\left. \begin{array}{l} U: V \times E \longrightarrow \mathbb{R} \text{ is continuous, bounded, nonnegative, that is,} \\ \text{(a) } v_n \longrightarrow v \text{ in } V, \theta_n \longrightarrow \theta \text{ in } E \implies U(v_n, \theta_n) \longrightarrow U(v, \theta), \\ \text{(b) } U \text{ maps bounded sets in } V \times E \text{ into bounded sets in } \mathbb{R}, \\ \text{(c) } U(v, \theta) \geq 0 \quad \text{for all } v \in V, \theta \in E, \end{array} \right\} \quad (38)$$

$$\left. \begin{array}{l} F: Y_1 \times Y_2 \longrightarrow \mathbb{R} \text{ is weakly lower semicontinuous, coercive, nonnegative,} \\ \text{that is,} \\ \text{(a) } f_n \rightharpoonup f \text{ in } Y_1 \text{ and } g_n \rightharpoonup g \text{ in } Y_2 \implies \liminf F(f_n, g_n) \geq F(f, g), \\ \text{(b) } F(f, g) \geq 0 \text{ for all } (f, g) \in Y_1 \times Y_2, \\ \text{(c) } \|f_n\|_{Y_1} \rightarrow +\infty, \|g_n\|_{Y_2} \rightarrow +\infty \implies F(f_n, g_n) \rightarrow +\infty. \end{array} \right\} \quad (39)$$

Theorem 5. Assume that (2)–(12), (15), (16), (38)–(39) hold. Then, there exists a solution $(u^*, \theta^*), (f^*, g^*) \in \mathcal{V}$ of Problem 4.

Proof. Let

$$\xi = \inf_{(u, \theta), (f, g) \in \mathcal{V}} \mathcal{L}(u, \theta, f, g) \in \mathbb{R} \quad (40)$$

and let $\{(u_n, \theta_n)\}, \{(f_n, g_n)\} \subset \mathcal{V}$ be the minimizing sequences for the functional $\mathcal{L}$, that is,

$$\lim \mathcal{L}(u_n, \theta_n, f_n, g_n) = \xi. \quad (41)$$

Our goal is to show that the sequences $\{f_n\}, \{g_n\}$ are bounded in Y_1 and Y_2 , respectively. We will argue by contradiction. We assume that $\{f_n\}, \{g_n\}$ are not bounded in Y_1 and Y_2 , respectively. Passing to the subsequences, denoted by $\{f_n\}, \{g_n\}$, we have, as $n \rightarrow \infty$, that

$$\|f_n\|_{Y_1} \rightarrow +\infty, \quad \|g_n\|_{Y_2} \rightarrow +\infty. \quad (42)$$

Now, from (37), (38)(c), we get

$$\mathcal{L}(u_n, \theta_n, f_n, g_n) \geq F(f_n, g_n).$$

This, (39)(c) and (41) gives that $\xi = \lim \mathcal{L}(u_n, f_n, \theta_n, g_n) = +\infty$ which is in contradiction with (40), so, $\{f_n\}, \{g_n\}$ are bounded in Y_1, Y_2 , respectively. Therefore, there exist $f^* \in Y_1$ and $g^* \in Y_2$ such that, passing to the subsequences, as $n \rightarrow \infty$, we get

$$f \rightharpoonup f^* \text{ in } Y_1 \quad \text{and} \quad g \rightharpoonup g^* \text{ in } Y_2. \quad (43)$$

Let $u^* = u(f^*, g^*)$ and let $\theta^* = \theta(f^*, g^*)$ be the solution of the system (1) for $f = f^*, g = g^*$. From (36) we have

$$(u^*, \theta^*), (f^*, g^*) \in \mathcal{V}. \quad (44)$$

Applying Theorem 4 and (43), it follows that, as $n \rightarrow \infty$, we have

$$u_n \rightarrow u^* \text{ in } V \quad \text{and} \quad \theta_n \rightarrow \theta^* \text{ in } E. \quad (45)$$

We observe that the assumptions (38)(a) and (39)(a) guarantee the weakly lower semicontinuity of the functional $\mathcal{L}$. This and the convergences (43) and (45) imply that

$$\liminf \mathcal{L}(u_n, \theta_n, f_n, g_n) \geq \mathcal{L}(u^*, \theta^*, f^*, g^*). \quad (46)$$

From (41) and (46) we see $\xi \geq \mathcal{L}(u^*, \theta^*, f^*, g^*)$. Additionally, from (40) and (44), we obtain that $\xi \leq \mathcal{L}(u^*, \theta^*, f^*, g^*)$. The last two inequalities together with (44) imply that the equality from the Problem 4 holds. $\square$

6 Convergence for optimal pairs

In this section, we study the dependence of the optimal controls pairs of Problem 4 with respect f and g . We assume that (2)–(12), (15), (16), and (25) hold. We consider the following perturbation of problem (1).

Problem 5. Given $f \in Y_1$ and $g \in Y_2$. Find $(u_n, \theta_n) \in C_1 \times C_2$ such that $u_n \in K_{1n}(u_n, \theta_n)$, $\theta_n \in K_{2n}(u_n, \theta_n)$ and

$$\begin{aligned} & \langle A_1 u_n + N_1 \theta_n, v - u_n \rangle_{V^* \times V} + \varphi_1(u_n, \theta_n, v) - \varphi_1(u_n, \theta_n, u_n) \\ & \geq \langle f, \pi_1 v - \pi_2 u_n \rangle_{Y_1} \quad \text{for all } v \in K_{1n}(u_n, \theta_n), \\ & \langle A_2 \theta_n + N_2 u_n, \eta - \theta_n \rangle_{E^* \times E} + \varphi_2(u_n, \theta_n, \eta) - \varphi_1(u_n, \theta_n, \theta_n) \\ & \geq \langle g, \pi_2 \eta - \pi_1 \theta_n \rangle_{Y_2} \quad \text{for all } \eta \in K_{2n}(u_n, \theta_n). \end{aligned} \quad (47)$$

From Theorem 3, we conclude that for each $f \in Y_1$ and $g \in Y_2$, there exists a unique solution $u_n = u(f, g)$, $\theta_n = \theta(f, g)$ to the system (47). Moreover, the solutions satisfy the estimations as in Lemma 4. We define the set of admissible pairs for (47) by

$$\mathcal{V}^n = \{(u_n, \theta_n), (f, g) \text{ such that (47) holds}\}. \quad (48)$$

The optimal control problem associated with Problem 5 is as follows.

Problem 6. Find $(u_n^*, \theta_n^*), (f_n^*, g_n^*) \in \mathcal{V}^n$ such that

$$\mathcal{L}(u_n^*, \theta_n^*, f_n^*, g_n^*) = \min_{(u_n, \theta_n), (f_n, g_n) \in \mathcal{V}^n} \mathcal{L}(u_n, \theta_n, f_n, g_n).$$

From assumptions (37)–(39) and Theorem 5, for each $n \in \mathbb{N}$, there exists a solution $(u_n^*, \theta_n^*), (f_n^*, g_n^*) \in \mathcal{V}^n$ of Problem 6. The result is as follows.

Theorem 6. Let $\{(u_n^*, \theta_n^*)\}, \{(f_n^*, g_n^*)\}$ be the sequences of solutions of Problem 6. Then, there exist the subsequences of the sequences $\{(u_n^*, \theta_n^*)\}, \{(f_n^*, g_n^*)\}$, denoted in the same way, and a solution $(u^*, \theta^*), (f^*, g^*)$ of Problem 6, such that

$$u_n \rightarrow u^* \text{ in } V, \quad \theta_n \rightarrow \theta^* \text{ in } E, \quad f_n^* \rightharpoonup f^* \text{ in } Y_1, \quad g_n^* \rightharpoonup g^* \text{ in } Y_2. \quad (49)$$

Proof. Let $n \in \mathbb{N}$.

Claim 1. The sequences $\{f_n^*\}, \{g_n^*\}$ are bounded in Y_1 and Y_2 , respectively.

Assume that $\{f_n^*\}, \{g_n^*\}$ are not bounded in Y_1 and Y_2 , respectively. Thus, for $n \rightarrow \infty$, we have

$$\|f_n^*\|_{Y_1} \rightarrow +\infty, \quad \|g_n^*\|_{Y_2} \rightarrow +\infty. \quad (50)$$

Using the assumptions (37) and (38)(c), we infer that

$$\mathcal{L}(u_n^*, \theta_n^*, f_n^*, g_n^*) \geq F(f_n^*, g_n^*). \quad (51)$$

Next, we pass to the limit in (51), as $n \rightarrow \infty$. Combining (50) and (39)(c), we deduce that the following equality holds:

$$\mathcal{L}(u_n^*, \theta_n^*, f_n^*, g_n^*) = +\infty. \quad (52)$$

Since $(u_n^*, \theta_n^*), (f_n^*, g_n^*)$ are solutions of Problem 6, we have

$$\mathcal{L}(u_n^*, \theta_n^*, f_n^*, g_n^*) \leq \mathcal{L}(u_n, \theta_n, f_n, g_n) \quad \text{for all } (u_n, \theta_n), (f_n, g_n) \in \mathcal{V}^n. \quad (53)$$

Let $f_0 \in Y_1$ and $g_0 \in Y_2$ be fixed, and let (u_n^0, θ_n^0) be the solution of Problem 6 for $f = f_0, g = g_0$, that is, $u_n^0 = u(f_0, g_0)$ and $\theta_n^0 = \theta(f_0, g_0)$. Then $(u_n^0, \theta_n^0), (f_0, g_0) \in \mathcal{V}^n$. Therefore, from (38)(c), (39)(b), and (22), we have for every $n > 0$ that

$$\mathcal{L}(u_n^*, \theta_n^*, f_n^*, g_n^*) \leq U(u_n^0, \theta_n^0) + F(f_0, g_0) \leq C, \quad (54)$$

where $C > 0$ does not depend on n . Relations (52) and (54) lead to contradiction which concludes the proof of the claim.

The sequences $\{f_n^*\}$, $\{g_n^*\}$ are bounded in Y_1 and Y_2 , respectively, so, we can find subsequences, denoted in the same way, and the elements f^* , g^* such that, as $n \rightarrow \infty$, we have

$$f_n^* \rightharpoonup f^* \text{ in } Y_1, \quad g_n^* \rightharpoonup g^* \text{ in } Y_2. \quad (55)$$

We denote by (u^*, θ^*) the solution of Problem 1 for $f = f^*$ and $g = g^*$, that is, $u^* = u(f^*, g^*)$, $\theta^* = \theta(f^*, g^*)$, then

$$(u^*, \theta^*), \quad (f^*, g^*) \in \mathcal{V}^n. \quad (56)$$

Moreover, from Theorem 4, as $n \rightarrow \infty$, we obtain

$$u_n^* \rightarrow u^* \text{ in } V, \quad \theta_n^* \rightarrow \theta^* \text{ in } E. \quad (57)$$

Claim 2. The pairs (u^*, θ^*) and (f^*, g^*) are solution to Problem 1.

In the proof of this claim, we will use the conditions (55), (57) and the assumptions (38)(a), (39)(a), that guarantee the weakly lower semicontinuity of the functional $\mathcal{L}$. Therefore

$$\liminf_{n \rightarrow \infty} \mathcal{L}(u_n^*, \theta_n^*, f_n^*, g_n^*) \geq \mathcal{L}(u^*, \theta^*, f^*, g^*). \quad (58)$$

Let the solution (u_0^*, θ_0^*) , (f_0^*, g_0^*) of Problem 4 be fixed. For each $n \in \mathbb{N}$, we denote by $(\widehat{u}_{0n}, \widehat{\theta}_{0n})$ the solution to Problem 6 for $f_n = f_0^*$ and $g_n = g_0^*$, that is, $\widehat{u}_{0n} = u(f_0^*, g_0^*)$, $\widehat{\theta}_{0n} = \theta(f_0^*, g_0^*)$. Hence, $(\widehat{u}_{0n}, \widehat{\theta}_{0n})$, $(f_0^*, g_0^*) \in \mathcal{V}^n$. By the optimality of pairs (u_n^*, θ_n^*) , (f_n^*, g_n^*) , we conclude that for every $n > 0$ the following inequality holds $\mathcal{L}(u_n^*, \theta_n^*, f_n^*, g_n^*) \leq \mathcal{L}(\widehat{u}_{0n}, \widehat{\theta}_{0n}, f_0^*, g_0^*)$. Passing to the upper limit in the last inequality, we infer that

$$\limsup \mathcal{L}(u_n^*, \theta_n^*, f_n^*, g_n^*) \leq \limsup \mathcal{L}(\widehat{u}_{0n}, \widehat{\theta}_{0n}, f_0^*, g_0^*). \quad (59)$$

From the fact that (u_0^*, θ_0^*) is the solution of Problem 1 for $f = f_0^*$, $g = g_0^*$, and $(\widehat{u}_{0n}, \widehat{\theta}_{0n})$ is the solution of Problem 6 for $f_n = f_0^*$, $g_n = g_0^*$, that is,

$$\begin{aligned} \widehat{u}_0^* &= u(f_0^*, g_0^*), & \widehat{\theta}_0^* &= \theta(f_0^*, g_0^*), \\ \widehat{u}_{0n} &= u(f_{0n}^*, g_{0n}^*), & \widehat{\theta}_{0n} &= \theta(f_{0n}^*, g_{0n}^*), \end{aligned}$$

we are able to use Theorem 4 and deduce that, as $n \rightarrow +\infty$, we have $\widehat{u}_{0n} \rightarrow u_0^*$ in V , $\widehat{\theta}_{0n} \rightarrow \theta_0^*$ in E . From here and the continuity of the functional $(u, \theta) \mapsto \mathcal{L}(u, \theta, f_0^*, g_0^*) : V \times E \times Y_1 \times Y_2 \rightarrow \mathbb{R}$, we get

$$\lim \mathcal{L}(\widehat{u}_{0n}, \widehat{\theta}_{0n}, f_0^*, g_0^*) = \mathcal{L}(u_0^*, \theta_0^*, f_0^*, g_0^*). \quad (60)$$

combining (58)–(60), we see that

$$\mathcal{L}(u^*, \theta^*, f^*, g^*) \leq \mathcal{L}(u_0^*, \theta_0^*, f_0^*, g_0^*). \quad (61)$$

On the other hand, since $(u_0^*, \theta_0^*), (f_0^*, g_0^*)$ is a solution to Problem 4, inclusion (56) implies that

$$\mathcal{L}(u_0^*, \theta_0^*, f_0^*, g_0^*) \leq \mathcal{L}(u^*, \theta^*, f^*, g^*). \quad (62)$$

Using the inequalities (61) and (62), we obtain

$$\mathcal{L}(u^*, \theta^*, f^*, g^*) = \mathcal{L}(u_0^*, \theta_0^*, f_0^*, g_0^*).$$

Hence from (56), we infer that

$$(u^*, \theta^*), (f^*, g^*) \text{ is a solution of Problem 4.} \quad (63)$$

Our theorem is a consequence of (55), (57), and (63). $\square$

7 Thermoelastic frictional contact problem

In this section, we use abstract results from Sections 4–6 to study thermoelastic frictional contact problem. The model presented in this Section is simpler version of the model presented in [17]. It is worth to mention, that in the paper we consider the optimal control problem that has not been the subject of research in [17].

The physical setting of the contact problem is as follows. We consider an elastic body which occupies a bounded domain $\Omega \subset \mathbb{R}^d$, where $d = 2, 3$, with Lipschitz boundary Γ . By ν we denote the unit normal vector to Γ . The boundary Γ is divided into mutually disjoint, measurable parts $\Gamma_1, \Gamma_2, \Gamma_3$, with $\text{meas}(\Gamma_1) > 0$. The body is clamped on Γ_1 , and a volume force of density f_0 and a heat source q_0 act in Ω . Surface tractions of density f_2 act on the part Γ_2 . The part Γ_3 is a contact surface between the body and the foundation. We denote by (V, H_1, V^*) and (E, H_2, E^*)

the evolution triples, where $H_1 = L^2(\Omega; \mathbb{R}^d)$ and $H_2 = L^2(\Omega)$. We define also the following sets

$$C_1 = \{v \in V \mid v_\nu \leq g_1 \text{ on } \Gamma_3\}, \quad C_2 = \{\eta \in E \mid \eta \leq \theta_F \text{ on } \Gamma_3\}, \quad (64)$$

and the set-valued maps $K_1: C_1 \times C_2 \rightarrow 2^{C_1}$ and $K_2: C_1 \times C_2 \rightarrow 2^{C_2}$ by

$$\begin{aligned} K_1(u, \theta) &= \{v \in C_1 \mid r_1(v) \leq m_1(u, \theta)\}, \\ K_2(u, \theta) &= \{\eta \in C_2 \mid r_2(\eta) \leq m_2(u, \theta)\} \end{aligned} \quad (65)$$

for $(u, \theta) \in C_1 \times C_2$. The examples of the sets $K_1(u, \theta)$ and $K_2(u, \theta)$ are provided in [17].

We introduce the following notations. By $\varepsilon(u) = \varepsilon_{ij}(u) = \frac{1}{2}(u_{i,j} + u_{j,i})$ we denote the linearized strain tensor. Moreover, by $\text{Div}\sigma = (\sigma_{ij,j})$ and $\text{div}q = (q_{i,i})$ we denote the divergence operators, respectively, for tensor and vector valued functions. Finally, the canonical inner products on $\mathbb{R}^d$ and $\mathbb{S}^d$ we denote by " $\cdot$ " and " $:$ ", respectively. The thermoelastic contact problem is the following.

Problem 7. Find a displacement field $u: \Omega \rightarrow \mathbb{R}^d$, a stress field $\sigma: \Omega \rightarrow \mathbb{S}^d$, a temperature $\theta: \Omega \rightarrow \mathbb{R}$, and a heat flux field $q: \Omega \rightarrow \mathbb{R}^d$ such that $r_1(u) \leq m_1(u, \theta)$, $r_2(\theta) \leq m_2(u, \theta)$ and

- a) $\sigma = \mathcal{A}\varepsilon(u) + \mathcal{M}\theta$ in Ω ,
- b) $q = -\mathcal{K}(\nabla\theta)$ in Ω ,
- c) $-\text{Div}\sigma = f_0$, $\text{div}q = q_0$ in Ω ,
- d) $u = 0$ on Γ_1 ,
- e) $\sigma\nu = f_2$ on Γ_2 ,
- f) $\theta = 0$ on $\Gamma_1 \cup \Gamma_2$,
- g) $-q \cdot \nu \leq 0$, $\theta - \theta_F \leq 0$, $(q \cdot \nu)(\theta - \theta_F) = 0$ on Γ_3 ,
- h) $u_\nu \leq g_1$, $\sigma_\nu \leq 0$, $\sigma_\nu(u_\nu - g_1) = 0$ on Γ_3 ,
- i) $-\sigma_\tau = p_\nu(u_\nu)h_\tau(\theta - \theta_F) \frac{u_\tau}{\|u_\tau\|}$ on Γ_3 ,
- j) $q \cdot \nu = h_1(u_\nu)k_e(\theta - \theta_F)$ on Γ_3 .

We present the physical interpretation of Problem 7, and refer to [2, 3, 9, 16, 17, 20, 22] for details. In this model, the thermoelastic constitutive law with nonlinear elasticity $\mathcal{A}$ and thermal expansion operators $\mathcal{M}$ is described by (a). The law is justified from the material strength point of view, since the properties of the elastic materials are often effected by temperature. The condition (b) presents the Fourier law of heat conduction in which $\mathcal{K}$ is the thermal conductivity, and q is the heat flux vector. The equilibrium equations for the stress and heat flux fields are presented in (c). Moreover, f_0 and q_0 stand for the densities of volume forces and of the applied volume heat sources. The boundary conditions (d) and (e) present the homogeneous Dirichlet and Neumann conditions, respectively, and the condition (f) shows that the temperature vanishes on $\Gamma_1 \cup \Gamma_2$. Condition (g) represents the Signorini-type condition between the relative temperature $\theta - \theta_F$ and the heat flux $q \cdot \nu$. It means that on Γ_3 the heat flux is supposed to be unilaterally directed from the foundation to the body, and then the body temperature does not exceed the foundation temperature. The Signorini-type contact condition (h) models the contact with a foundation made of a rigid body covered by a layer made of deformable material of the thickness g_1 (see [22, Sections 8.4 and 8.5]). The relation (i) is the normal compliance contact condition coupled with the Coulomb friction law (see [9, Section 5.4]), in which p_ν is prescribed function such that $p(r) = 0$ which vanishes for $r \leq 0$ and the coefficient of friction h_τ depends on the relative temperature $\theta - \theta_F$, where $\theta_F \geq 0$ is the temperature of the foundation. The heat exchange between the temperature and the flux is modeled by condition (j), where the heat exchange coefficient h_1 depends on the normal displacement. The inequalities $r_1(u) \leq m_1(u, \theta)$ and $r_2(\theta) \leq m_2(u, \theta)$ may accommodate various additional implicit constraints in which we are interested in the model (see [17, Example 18]).

We introduce the following assumptions for the operators and functions.

The elasticity operator $\mathcal{A}: \Omega \times \mathbb{S}^d \rightarrow \mathbb{S}^d$ is such that

$$\left. \begin{array}{l} \text{(a) } \mathcal{A}(\cdot, \varepsilon) \text{ is measurable on } \Omega \text{ for all } \varepsilon \in \mathbb{S}^d, \\ \text{(b) } \mathcal{A}(x, \cdot) \text{ is continuous on } \mathbb{S}^d \text{ for a.e. } x \in \Omega, \\ \text{(c) } \|\mathcal{A}(x, \varepsilon)\|_{\mathbb{S}^d} \leq a_0(x) + \bar{a}_1 \|\varepsilon\|_{\mathbb{S}^d} \text{ for all } \varepsilon \in \mathbb{S}^d, \text{ a.e. } x \in \Omega \\ \text{with } a_0 \in L^2(\Omega), a_0 \geq 0 \text{ and } \bar{a}_1 > 0, \\ \text{(d) } (\mathcal{A}(x, \varepsilon_1) - \mathcal{A}(x, \varepsilon_2)) : (\varepsilon_1 - \varepsilon_2) \geq m_{\mathcal{A}} \|\varepsilon_1 - \varepsilon_2\|_{\mathbb{S}^d}^2 \text{ for all } \varepsilon_1, \varepsilon_2 \in \mathbb{S}^d \\ \text{and a.e. } x \in \Omega \text{ with } m_{\mathcal{A}} > 0. \end{array} \right\} \quad (66)$$

The thermal expansion operator $\mathcal{M}: \Omega \times \mathbb{R} \rightarrow \mathbb{S}^d$ is such that

$$\left. \begin{array}{l} \text{(a) } \mathcal{M}(\cdot, s) \text{ is measurable on } \Omega \text{ for all } s \in \mathbb{R}, \\ \text{(b) } \mathcal{M}(x, \cdot) \text{ is continuous on } \mathbb{R} \text{ for a.e. } x \in \Omega, \\ \text{(c) } \|\mathcal{M}(x, r)\|_{\mathbb{S}^d} \leq c_M(1 + |r|) \text{ for all } r \in \mathbb{R}, \text{ a.e. } x \in \Omega, \text{ with } c_M > 0. \end{array} \right\} \quad (67)$$

The thermal conductivity operator $\mathcal{K}: \Omega \times \mathbb{R}^d \rightarrow \mathbb{R}^d$ is such that

$$\left. \begin{array}{l} \text{(a) } \mathcal{K}(\cdot, \xi) \text{ is measurable on } \Omega \text{ for all } \xi \in \mathbb{R}^d, \\ \text{(b) } \mathcal{K}(x, \cdot) \text{ is continuous on } \mathbb{R}^d \text{ for a.e. } x \in \Omega, \\ \text{(c) } \|\mathcal{K}(x, \xi)\|_{\mathbb{R}^d} \leq k_0(x) + k_1\|\xi\|_{\mathbb{R}^d} \text{ for all } \xi \in \mathbb{R}^d, \text{ a.e. } x \in \Omega \\ \text{with } k_0 \in L^2(\Omega), k_0 \geq 0, k_1 > 0, \\ \text{(d) } (\mathcal{K}(x, \xi_1) - \mathcal{K}(x, \xi_2)) \cdot (\xi_1 - \xi_2) \geq m_{\mathcal{K}}\|\xi_1 - \xi_2\|_{\mathbb{R}^d}^2 \text{ for all } \xi_1, \xi_2 \in \mathbb{R}^d, \\ \text{a.e. } x \in \Omega \text{ with } m_{\mathcal{K}} > 0. \end{array} \right\} \quad (68)$$

The coefficient of heat exchange $h_1: \Gamma_3 \times \mathbb{R} \rightarrow \mathbb{R}_+$ and the function $h_\tau: \Gamma_C \times \mathbb{R} \rightarrow \mathbb{R}_+$, are such that for $i = 1, \tau$ we have

$$\left. \begin{array}{l} \text{(a) } h_i(\cdot, r) \text{ is measurable on } \Gamma_3 \text{ for all } r \in \mathbb{R}, \\ \text{(b) } |h_i(x, r_1) - h_i(x, r_2)| \leq L_{h_i}|r_1 - r_2| \text{ for all } r_1, r_2 \in \mathbb{R}, \text{ a.e. } x \in \Gamma_3 \\ \text{with } L_{h_i} > 0, \\ \text{(c) } h_i(x, r) \leq \bar{h}_i \text{ for all } r \in \mathbb{R}, \text{ a.e. } x \in \Gamma_3 \text{ with } \bar{h}_i > 0. \end{array} \right\} \quad (69)$$

The normal compliance function $p_\nu: \Gamma_3 \times \mathbb{R} \rightarrow \mathbb{R}_+$ is such that

$$\left. \begin{array}{l} \text{(a) } p_\nu(\cdot, r) \text{ is measurable on } \Gamma_3 \text{ for all } r \in \mathbb{R}, \\ \text{(b) } |p_\nu(x, r_1) - p_\nu(x, r_2)| \leq L_{p_\nu}|r_1 - r_2| \text{ for all } r_1, r_2 \in \mathbb{R}, \text{ a.e. } x \in \Gamma_3 \\ \text{with } L_{p_\nu} > 0, \\ \text{(c) } p_\nu(x, r) \leq \bar{p} \text{ for all } r \in \mathbb{R}, \text{ a.e. } x \in \Gamma_3 \text{ with } \bar{p} > 0. \end{array} \right\} \quad (70)$$

The coefficient $k_e: \Gamma_3 \times \mathbb{R} \rightarrow \mathbb{R}_+$ is such that

$$\left. \begin{array}{l} \text{(a) } k_e(\cdot, r) \text{ is measurable on } \Gamma_3 \text{ for all } r \in \mathbb{R}, \\ \text{(b) } |k_e(x, r_1) - k_e(x, r_2)| \leq L_k|r_1 - r_2| \text{ for all } r_1, r_2 \in \mathbb{R}, \text{ a.e. } x \in \Gamma_3 \\ \text{with } L_k > 0, \\ \text{(c) } k_e(x, r) \leq \bar{k} \text{ for all } r \in \mathbb{R}, \text{ a.e. } x \in \Gamma_3 \text{ with } \bar{k} > 0. \end{array} \right\} \quad (71)$$

$r_1: V \rightarrow \mathbb{R}$, $r_2: E \rightarrow \mathbb{R}$, $m_1, m_2: H_1 \times H_2 \rightarrow \mathbb{R}$ are functions such that

$$\left. \begin{array}{l} \text{(a) } r_1, r_2 \text{ are positively homogeneous, subadditive,} \\ \text{and lower semicontinuous,} \\ \text{(b) } m_1 \text{ is continuous, } m_1^0 := \inf\{m_1(u, \theta) \mid (u, \theta) \in H_1 \times H_2\} > 0 \\ \text{and } r_1(0) \leq m_1^0, \\ \text{(c) } m_2 \text{ is continuous, } m_2^0 := \inf\{m_2(u, \theta) \mid (u, \theta) \in H_1 \times H_2\} > 0 \\ \text{and } r_2(0) \leq m_2^0, \end{array} \right\} \quad (72)$$

$$f_0 \in L^2(\Omega; \mathbb{R}^d), \quad f_N \in L^2(\Gamma_N; \mathbb{R}^d), \quad q_0 \in L^2(\Omega), \quad g_1 \in L^2(\Gamma_3), \quad g_1 \geq 0, \quad \theta_F \geq 0. \quad (73)$$

Remark 3. In the proof of the uniqueness of solution, we will need the following assumptions.

$$\begin{array}{l} \text{(a) } \|\mathcal{M}(x, r_1) - \mathcal{M}(x, r_2)\|_{\mathbb{S}^d} \leq c_{M_1} |r_1 - r_2| \text{ with } c_{M_1} > 0 \\ \text{for all } r_1, r_2 \in \mathbb{R}, \\ \text{(b) } \|\mathcal{K}(x, \xi_1) - \mathcal{K}(x, \xi_2)\|_{\mathbb{R}^d} \leq c_K \|\xi_1 - \xi_2\|_{\mathbb{R}^d} \text{ with } c_K > 0, \\ \text{for all } \xi_1, \xi_2 \in \mathbb{R}^d. \end{array} \quad (74)$$

The variational formulation of Problem 7 leads to the following system of inequalities.

Problem 8. Find $(u, \theta) \in C_1 \times C_2$ with $u \in K_1(u, \theta)$, $\theta \in K_2(u, \theta)$ such that for all $v \in K_1(u, \theta)$, $\eta \in K_2(u, \theta)$, we have

$$\begin{aligned} & \int_{\Omega} (\mathcal{A}\varepsilon(u) + \mathcal{M}\theta) \cdot (\varepsilon(v) - \varepsilon(u)) \, dx \\ & + \int_{\Gamma_C} p_{\nu}(u_{\nu}) h_{\tau}(\theta - \theta_F)(\|v_{\tau}\| - \|u_{\tau}\|) \, d\Gamma \\ & \geq \int_{\Omega} f_0 \cdot (v - u) \, dx + \int_{\Gamma_N} f_2 \cdot (v - u) \, d\Gamma, \\ & \int_{\Omega} \mathcal{K}(\nabla\theta) \cdot (\nabla\eta - \nabla\theta) \, dx + \int_{\Gamma_C} h_1(u_{\nu}) k_e(\theta - \theta_F)(\eta - \theta) \, d\Gamma \\ & \geq \int_{\Omega} q_0 \cdot (\eta - \theta) \, dx. \end{aligned}$$

The main result in this section is as follows.

Theorem 7. Under the assumptions (66)–(74) Problem 8 has a unique solution.

Proof. The idea is to associate with Problem 8 an auxiliary system of inequalities, and apply Theorem 3. To this goal, we introduce the operators $A_1: V \rightarrow V^*$, $A_2: E \rightarrow E^*$, $N_1: E \rightarrow V^*$, and $N_2: V \rightarrow E^*$ defined by

$$\langle A_1 u, v \rangle = \int_{\Omega} \mathcal{A} \varepsilon(u) : \varepsilon(v) dx \quad \text{for } u, v \in V, \quad (75)$$

$$\langle A_2 \theta, \eta \rangle_{E^* \times E} = \int_{\Omega} \mathcal{K}(\nabla \theta) \cdot \nabla \eta dx \quad \text{for } \theta, \eta \in E, \quad (76)$$

$$\langle N_1 \theta, v \rangle = \int_{\Omega} \mathcal{M} \theta : \varepsilon(v) dx \quad \text{for } \theta \in E, v \in V \quad (77)$$

and $N_2 \equiv 0$. Furthermore, we define functions $\varphi_1: V \times E \times V \rightarrow \mathbb{R}$, and $\varphi_2: V \times E \times E \rightarrow \mathbb{R}$ by

$$\varphi_1(w, \theta, v) = \int_{\Gamma_C} p_{\nu}(w_{\nu}) h_{\tau}(\theta - \theta_F) \|v_{\tau}\| d\Gamma \quad \text{for } w, v \in V, \theta \in E, \quad (78)$$

$$\varphi_2(w, \theta, \eta) = \int_{\Gamma_C} h(w_{\nu}) k_e(\theta - \theta_F) \eta d\Gamma. \quad (79)$$

Finally, we introduce $f \in Y_1 = L^2(\Omega; \mathbb{R}^d) \times L^2(\Gamma_2; \mathbb{R}^d)$ and $g \in Y_2 = L^2(\Omega)$ defined by

$$\begin{aligned} (f, \pi_1 v)_{Y_1} &= \langle f_0, v \rangle_{L^2(\Omega; \mathbb{R}^d)} + \langle f_2, v \rangle_{L^2(\Gamma_2; \mathbb{R}^d)}, \quad \text{for } v \in V \\ (g, \pi_2 \eta)_{Y_2} &= \langle q_0, \eta \rangle_{L^2(\Omega)} \quad \text{for } \eta \in E, \end{aligned} \quad (80)$$

where

$$\begin{aligned} \pi_1: V &\rightarrow Y_1 & \pi_1 v &= (v, v|_{\Gamma_2}) \quad \forall v \in V, \\ \pi_2: E &\rightarrow Y_2 & \pi_2 \eta &= \eta \quad \text{for all } \eta \in E. \end{aligned}$$

Using (75)–(80), we can rewrite Problem 8 in the following way.

Problem 9. Find $(u, \theta) \in C_1 \times C_2$ with $u \in K_1(u, \theta)$, $\theta \in K_2(u, \theta)$ where

$$\left\{ \begin{array}{ll} \langle A_1 u + N_1 \theta, v - u \rangle + \varphi_1(u, \theta, v) - \varphi_1(u, \theta, u) \geq (f, v)_{Y_1} & \text{for all } v \in K_1(u, \theta), \\ \langle A_2 \theta, \eta - \theta \rangle_{E^* \times E} + \varphi_2(u, \theta, \eta) - \varphi_2(u, \theta, \theta) \geq (g, \eta)_{Y_2} & \text{for all } \eta \in K_2(u, \theta). \end{array} \right.$$

We see that for $N_2 = 0$ Problem 3 reduce to Problem 9. We are ready to apply Theorem 1 and show that Problem 9 has a solution. To this end, we shall verify hypotheses (5)–(10), (11) and (12). The proofs of these assumptions are analogous to [17, Chapter 4]. Now, we will verify (3) while condition (4) can be demonstrated by analogous arguments. Assumption (73) implies that $0 \in C_1$. Assumption (72)(b) implies that $r_1(0) \leq m_1(u, \theta)$ for all $(u, \theta) \in C_1 \times C_2$ so $0 \in K_1(u, \theta)$. Thus (3)(c) holds with $K_0 = \{0\}$. By (72)(a), it follows that $r_1: V \rightarrow \mathbb{R}$ is a convex function, thus, the convexity of $K_1(u, \theta)$ for each $(u, \theta) \in C_1 \times C_2$ is obvious. The closedness of

$K_1(u, \theta)$ for $(u, \theta) \in C_1 \times C_2$ is a consequence of the lower semicontinuity of r_1 and the continuity of embedding of V to H_1 . So, (3)(a) is satisfied. The proof that $K_2(u_n, \theta_n) \xrightarrow{M} K_2(u, \theta)$ for all $\{(u_n, \theta_n)\} \subset C_1 \times C_2$, such that $(u_n, \theta_n) \rightarrow (u, \theta)$ in $(V \times E)_w$, is similar to [15, Theorem 4.1], so it is omitted. In order to prove the uniqueness of the solution of Problem 8, we will use Theorem 3. Therefore, we should verify the assumptions (15) and (16). We observe that from (66)(d) it follows that (15)(a) with $m_{A_1} = m_A$. The condition (15)(b) holds with $m_{N_2} = 0$. Moreover, from (74)(a), we conclude that (15)(c) holds with $L_{N_1} = c_{M_1}$ and from (74)(b), we obtain that (15)(d) holds with $L_{A_2} = c_K$. Finally, from (69)(b), (c), (70)(b), (c), (71)(b), (c) and the Holder inequality, we see that

$$\begin{aligned} & \int_{\Gamma_3} p_\nu(u_{1\nu}) h_\tau(\theta_1 - \theta_F)(\|u_{2\tau}\| - \|u_{1\tau}\|) d\Gamma \\ & + \int_{\Gamma_3} p_\nu(u_{2\nu}) h_\tau(\theta_2 - \theta_F)(\|u_{1\tau}\| - \|u_{2\tau}\|) d\Gamma \\ & \leq \int_{\Gamma_3} \left[|p_\nu(u_{1\nu}) - p_{2\nu}(u_{2\nu})| |h(\theta_1 - \theta_F)| \right. \\ & \quad \left. + |p_\nu(u_{2\nu})| |h_\tau(\theta_1 - \theta_F) - h_\tau(\theta_2 - \theta_F)| \left| \|u_{2\tau}\| - \|u_{1\tau}\| \right| \right] d\Gamma \\ & \leq (L_{p_\nu} \bar{h}_\tau |u_{1\nu} - u_{2\nu}| + \bar{p} L_{h_\tau} |\theta_1 - \theta_2|) \left| \|u_{2\tau}\| - \|u_{1\tau}\| \right| d\Gamma \\ & \leq (L_{p_\nu} \bar{h}_\tau \|\gamma_1\|^2 \|u_1 - u_2\|_V + \bar{p} L_{h_\tau} \|\gamma_1\| \|\gamma_2\| \|\theta_1 - \theta_2\|_E) \|u_2 - u_1\|_V, \end{aligned}$$

and, in the similar way as above, we obtain

$$\begin{aligned} & \int_{\Gamma_3} h(u_{1\nu}) k_e(\theta_1 - \theta_F)(\theta_2 - \theta_1) d\Gamma + \int_{\Gamma_3} h(u_{2\nu}) k_e(\theta_2 - \theta_F)(\theta_1 - \theta_2) d\Gamma \\ & \leq \int_{\Gamma_3} \left[(h(u_{1\nu}) - h(u_{2\nu})) k_e(\theta_1 - \theta_F) \right. \\ & \quad \left. + h(u_{2\nu}) (k_e(\theta_1 - \theta_F) - k_e(\theta_2 - \theta_F)) \right] (\theta_2 - \theta_1) d\Gamma \\ & \leq (\bar{k} L_{h_1} \|\gamma_1\| \|\gamma_2\| \|u_1 - u_2\|_V + \bar{h}_1 L_k \|\gamma_2\|^2 \|\theta_1 - \theta_2\|_E) \|\theta_2 - \theta_1\|_E. \end{aligned}$$

So, the condition (15)(e) and (f) hold with $\theta_{\varphi_1} = \max\{\|\gamma_1\|^2 L_{p_\nu} \bar{h}_\tau, \|\gamma_1\| \|\gamma_2\| L_{h_\tau} \bar{p}\}$ and $\theta_{\varphi_2} = \max\{\|\gamma_1\| \|\gamma_2\| L_{h_1} \bar{k}, \|\gamma_2\|^2 L_{k_e} \bar{h}_1\}$, respectively. Here $\gamma_1: V \rightarrow L^2(\Gamma_3; \mathbb{R}^d)$ and $\gamma_2: E \rightarrow L^2(\Gamma_3)$ we understand as the trace operators. Proceeding in the similar way as in the proof of Theorem 3, we note that an analogous condition to (16) holds too. $\square$

Next, in Problem 9, we can take the sets

$$\begin{aligned} K_{1n}(u, \theta) &= \{v \in C_{1n} \mid r_1(v) \leq m_1(u, \theta)\}, \quad C_{1n} = \{v \in V \mid v_\nu \leq g_{1n} \text{ on } \Gamma_3\}, \\ K_{2n}(u, \theta) &= \{\eta \in C_{2n} \mid r_2(\eta) \leq m_2(u, \theta)\}, \quad C_{2n} = \{\eta \in E \mid \eta \leq \theta_{Fn} \text{ on } \Gamma_3\}. \end{aligned}$$

We are in position to use Theorem 4 in order to deduce that $u_n \rightarrow u$ in V and $\theta_n \rightarrow \theta$ in E .

We end this section with example of optimal control problem for which the results presented in sections 5 and 6 hold. Let f_0, g_1, θ_F be given. We define the set of admissible pair for Problem 8 as follows:

$$\begin{aligned} \bar{\mathcal{V}}_{ad} = \{ & (u, \theta) \in C_1 \times C_2 \text{ and } (f_2, q_0) \in L^2(\Gamma_3; \mathbb{R}^d) \times L^2(\Omega) \\ & \text{such that Problem 8 holds} \} \end{aligned}$$

and we choose the cost functional

$$\begin{aligned} \mathcal{L}(u, \theta, f_2, q_0) = & \theta_1 \|u_\nu - z_0\|_{L^2(\Gamma_3)} + \theta_2 \|\theta - \theta_0\|_{L^2(\Gamma_3)} + \gamma_1 \|f_2\|_{L^2(\Gamma_3; \mathbb{R}^d)} \\ & + \gamma_2 \|q_0\|_{L^2(\Omega)}, \end{aligned} \quad (81)$$

where $\theta_0, z_0 \in L^2(\Gamma_3)$ and $\theta_1, \theta_2, \gamma_1, \gamma_2 > 0$. Then, the problem which we consider can be formulated as follows.

Problem 10. Find $(u^*, \theta^*), (f_2^*, q_0^*) \in \bar{\mathcal{V}}_{ad}$ such that

$$\mathcal{L}(u^*, \theta^*, f_2^*, q_0^*) = \min_{(u, \theta), (f_2, q_0) \in \bar{\mathcal{V}}_{ad}} \mathcal{L}(u, \theta, f_2, q_0).$$

The mechanical interpretation of the Problem 10 is as follows. We are looking for a surface pressure f_2 acting on Γ_2 and heat source q_0 acting on Ω such that a normal displacement and a temperature of the body is as close as possible to the desired displacement z_0 and desired temperature θ_0 . Moreover, this choice has to fulfill a minimum expenditure condition which is taken into account by the last term in (81). We have the following result in the study of Problem 10.

Theorem 8. Let (66)–(73) hold, and moreover, let $f \in V^*, q_0 \in E^*, z_0, \theta_0 \in L^2(\Gamma_C)$, and let $\theta_1, \theta_2, \gamma_1, \gamma_2 > 0$. Then there exists a solution $(u^*, \theta^*), (f_2^*, q_0^*) \in \bar{\mathcal{V}}_{ad}$ to Problem 10.

Theorem 8 is a direct consequence of Theorem 5, and we note that Theorem 6 provides a convergence result for a sequence of optimal pairs of Problem 10. The question about conditions of optimality will be the future direction of study.

Acknowledgements

Author is grateful to anonymous referees and editor for their constructive remarks.

Funding

This research received no external funding.

Conflicts of Interest

The author declares no conflicts of interest.

References

- [1] Aubin, J.P. *Mathematical methods of game and economic theory*, vol. 7, North-Holland, Amsterdam, The Netherlands, 1979.
- [2] Bai, Y., Migórski, S., Nguyen, V. H. and Peng, J. *Existence of solution to a new class of coupled variational-hemivariational inequalities*, J. Nonlinear Var. Anal. 6 (5) (2022), 499–516.
- [3] Baiz, O., Benaissa, H., Bouchantouf, R. and Moutawakil, D. *Optimization problems for a thermoelastic frictional contact problem*, Math. Model. Anal. 26 (2021), 444–468.
- [4] Barbu, V. *Optimal control of variational inequalities*, Pitman, Boston, 1984.
- [5] Browder, F.E. *Nonlinear monotone operators and convex sets in Banach spaces*, Bull. Amer. Math. Soc. 71 (1965), 780–785.
- [6] Denkowski, Z., Migórski, S. and Papageorgiou, N.S. *An Introduction to Nonlinear Analysis: Applications*, Kluwer Academic Publishers, Boston, Dordrecht, London, New York, 2003.
- [7] Denkowski, Z., Migórski, S. and Papageorgiou, N.S. *An introduction to nonlinear analysis: Theory*, Kluwer Academic Publishers, Boston, Dordrecht, London, New York, 2003.
- [8] Friedman, A. *Optimal control for variational inequalities*, SIAM J. Control Optim. 24 (1986), 439–451.
- [9] Han, W. and Sofonea, M. *Quasistatic contact problems in viscoelasticity and viscoplasticity*, Studies in Advanced Mathematics 30, American Mathematical Society, Providence, RI International Press, Somerville, MA, 2002.

- [10] Khan, A. and Sama, M. *Optimal control of multivalued quasi variational inequalities*, Nonlinear Anal. 75 (2012), 1419–1428.
- [11] Lions, J.L. *Quelques methodes de resolution des problemes aux limites non lineaires*, Dunod, Paris, 1969.
- [12] Lions, J.L. and Stampacchia, G. *Variational inequalities*, Comm. Pure Appl. Math. 20 (1967), 493–519.
- [13] Liu, J., Yang, X., Zeng, S.D. and Zhao, Y. *Coupled variational inequalities: existence, stability and optimal control*, J. Optim. Theory Appl. 193 (2022), 877–909.
- [14] Mignot, F. and Puel, J.P. *Optimal control in some variational inequalities*, SIAM J. Control Optim. 22 (1984), 466–476.
- [15] Migórski, S. and Dudek, S. *A new class of variational-hemivariational inequalities for steady Oseen flow with unilateral and frictional type boundary conditions*, Z. Angew. Math. Mech. 100 (2020), 1–23.
- [16] Migórski, S., Ochal, A. and Sofonea, M. *Nonlinear inclusions and hemivariational inequalities. Models and analysis of contact problems*, Advances in Mechanics and Mathematics 26, Springer, New York, 2013.
- [17] Migórski, S., Ogorzały, J. and Dudek, S. *A new general class of systems of elliptic quasi-variational-hemivariational inequalities*, Commun. Nonlinear Sci. Numer. Simul. 121 (2023), 1–17.
- [18] Mosco, U. *Convergence of convex sets and of solutions of variational inequalities* Adv. Math. 3 (1969), 510–585.
- [19] Neittaanmaki, P., Sprekels, J. and Tiba, D. *Optimization of elliptic systems: Theory and applications*, Springer Monogr. Math., Springer, New York, 2006.
- [20] Shillor, M., Sofonea, M. and Telega, J.J. *Models and analysis of quasistatic contact*, Lect. Notes Phys. 655, Springer, Berlin, Heidelberg, 2004.
- [21] Sofonea, M., Bollait, J. and Tarzia, D. *Optimal control of differential quasivariational inequalities with applications in contact mechanics*, J. Math. Anal. Appl. 493 (2021), 1–23.
- [22] Sofonea, M. and Matei, A. *Mathematical models in contact mechanics*, London Mathematical Society Lecture Note Series 398, Cambridge University Press, 2012.

- [23] Sofonea, M. and Tarzia, D. *Convergence Results for Optimal Control Problems Governed by Elliptic Quasivariational Inequalities*, Numer. Funct. Anal. Optim. 41 (2020), 1326–1351.
- [24] Sofonea, M., Xiao, Y.B. and Couderc, M. *Optimization problems for elastic contact models with unilateral constraints*, Z. Angew. Math. Phys. 70(1) (2019), 1–17.
- [25] Tiba, D. *Optimal control of nonsmooth distributed parameter systems*, Springer, Berlin, 1990.
- [26] Tiba, D. *Lectures on the optimal control of elliptic equations*, Lecture Notes, vol.32, University of Jyväskylä, Jyväskylä, 1995.



Mathematical model for an infectious disease with quarantine

F. Iranzad, H.R. Marasi, H. Kheiri*, and A. Jabbari

Abstract

This article presents a novel compartmental mathematical model for an infectious disease that uniquely distinguishes susceptible individuals into quarantined and nonquarantined groups, while explicitly capturing dynamic transitions between these compartments. Unlike traditional models, ours integrates the combined effects of quarantine effectiveness and population mobility on disease transmission dynamics. We rigorously define the biologically feasible region and prove its positive invariance, ensuring model consistency. Using the next-generation matrix method, we derive a precise expression for the basic reproduction number ($\mathcal{R}_0$). Our

*Corresponding author

Received 3 June 2025; revised 8 September 2025; accepted 20 November 2025

Fatemeh Iranzad

Faculty of Mathematics, Statistics, and Computer Science, University of Tabriz, Tabriz, Iran. e-mail: fateme-hiranzad1994@gmail.com

Hamid Reza Marasi

Faculty of Mathematics, Statistics, and Computer Science, University of Tabriz, Tabriz, Iran. e-mail: hamidreza.marasi@gmail.com

Hossein Kheiri

Faculty of Mathematics, Statistics, and Computer Science, University of Tabriz, Tabriz, Iran. e-mail: h-kheiri@tabrizu.ac.ir

Azizeh Jabbari

Faculty of Mathematics, Statistics, and Computer Science, University of Tabriz, Tabriz, Iran. e-mail: azizeh.jabbari@gmail.com

How to cite this article

Iranzad, F., Marasi, H.R., Kheiri, H. and Jabbari, A., Mathematical model for an infectious disease with quarantine. *Iran. J. Numer. Anal. Optim.*, 2026; 16(2): 667–682. <https://doi.org/10.22067/ijnao.2025.93898.1661>

analytical contributions include establishing both local and global stability of the disease-free equilibrium for $\mathcal{R}_0 < 1$, and, notably, developing a novel Lyapunov function to prove the global asymptotic stability of the endemic equilibrium when $\mathcal{R}_0 > 1$. Through comprehensive numerical simulations, we validate the theoretical results and provide new insights into the critical role of quarantine measures in epidemic control. These findings offer valuable guidance for designing more effective public health intervention strategies.

AMS subject classifications (2020): 37Dxx; 00A71

Keywords: Infectious diseases, Local stability, Global stability, Reproduction number, Lyapunov function.

1 Introduction

Historically, infectious diseases have been a major source of morbidity and mortality among human populations, often devastating entire communities. Beyond individual health implications, these diseases significantly disrupt community health systems and severely impact economic and social infrastructures. From the bubonic plague to recent pandemics such as COVID-19, outbreaks have profoundly shaped human history and public health responses.

Over the last decade, with the rise of global epidemics, quarantine has gained renewed importance as an effective nonpharmaceutical intervention for infectious disease control. Quarantine aims to isolate individuals who may have been exposed to the pathogen, thereby reducing transmission potential [1]. This intervention not only curtails the spread of disease but also alleviates strain on healthcare facilities by modulating the flow of infectious and susceptible individuals.

Mathematical modeling serves as a fundamental tool to understand infectious disease transmission dynamics and evaluate control strategies [6]. While prior models have addressed interventions such as treatment and contact tracing [8, 5], many do not explicitly differentiate between quarantined and nonquarantined susceptibles or incorporate the role of hospitalization distinctly within the population compartments.

In this paper, we propose a novel compartmental model that subdivides the total population into six epidemiological classes: nonquarantined susceptible, quarantined susceptible, latent, infectious, hospitalized, and recovered individuals. Our model explicitly incorporates the dynamics of quarantine not only for susceptibles but also distinguishes hospitalized individuals as a separate compartment, capturing their unique role in disease progression and control. The force of infection is modeled by effective contacts with both infectious and hospitalized individuals, reflecting realistic transmission pathways.

The originality of this work lies in the detailed representation of quarantine processes, including transfer rates between susceptible and quarantined classes, and an explicit hospitalized class, which is often aggregated or omitted in existing models. This framework allows a more nuanced analysis of quarantine's impact on reducing infection spread and healthcare burden. Rigorous qualitative analyses, including stability and threshold dynamics, are conducted to provide insights into the disease behavior and control under quarantine measures.

By integrating these aspects, our model enhances the understanding of how quarantine and hospitalization jointly influence infectious disease dynamics. The insights gained could inform public health strategies tailored to optimize quarantine implementation and hospital resource management in the context of controlling infectious diseases.

2 Model formulation

We denote the total population at time t by $N(t)$, which is subdivided into six disjoint classes of nonquarantined susceptible $S(t)$, quarantined susceptible $S_q(t)$, latent $L(t)$, infectious $I(t)$, hospitalized $H(t)$, and recovered $R(t)$ individuals. So, it can be written

$$N(t) = S(t) + S_q(t) + L(t) + I(t) + H(t) + R(t).$$

The population of susceptible individuals grows through the recruitment of the new population (at the rate of Λ). Also, quarantined susceptible may leave their class with rate of ψ and move into susceptible individuals. It is reduced by infected following efficient contacts with infected individuals, at rate of λ , given by

$$\lambda = \frac{\gamma(\eta_I I + \eta_H H)}{N(t)}, \quad (1)$$

where γ is effective contact rate, and η_I and η_H are modifier parameters for infectious and hospitalized. This individual is reduced by transferring from susceptible individuals to quarantined susceptible individuals with rate of γ , and natural death (with the rate of θ). Then, the dynamic of susceptible is obtained as follows:

$$\frac{dS}{dt} = \Lambda - \lambda S - (\theta + \gamma)S + \psi S_q.$$

In a similar way, other dynamic equations can be obtained. We have

$$\frac{dS}{dt} = \Lambda - \lambda S - (\theta + \gamma)S + \psi S_q,$$

Table 1: Description of parameters in the model (2)

Variable	Description
S	Population of nonquarantined susceptible individuals
S_q	Population of quarantined susceptible individuals
L	Population of latent individuals
I	Population of infected symptomatic individuals
H	Population of hospitalized individuals
R	Population of effectively-treated individuals
Parameter	Description
Λ	Recruitment rate
η_I	Modification parameter for infectious
η_H	Modification parameter for hospitalized
γ	Effective contact rate
θ	Per-capita natural mortality rate
δ_I	Per-capita death rate due to infectious
δ_H	Per-capita death rate due to hospitalized
θ	Per-capita rate of infected people hospitalized
ψ	Transfer rate of quarantined susceptible to nonquarantined susceptible
γ	Transfer rate of nonquarantined susceptible to quarantined susceptible
r	Fixed number between 0 and 1
σ	Endogenous re-activation rate of latent individuals
ψ_I	Effective treatment rate of infectious individuals
ψ_H	Effective treatment rate of hospitalized individuals

$$\begin{aligned}
\frac{dS_q}{dt} &= -r\lambda S_q + \gamma S - (\psi + \theta)S_q, \\
\frac{dL}{dt} &= \lambda S + r\lambda S_q - (\theta + \sigma)L, \\
\frac{dI}{dt} &= \sigma L - (\theta + \psi_I + \theta + \delta_I)I, \\
\frac{dH}{dt} &= \theta I - (\psi_H + \theta + \delta_H)H, \\
\frac{dR}{dt} &= \psi_I I + \psi_H H - \theta R.
\end{aligned} \tag{2}$$

The parameters used in the model are given in Table 1, and the graphical representation of the model is shown in Figure 1.

3 Basic properties of the model

We establish that system (2) is biologically well-posed, demonstrating that its solutions remain nonnegative for all time.

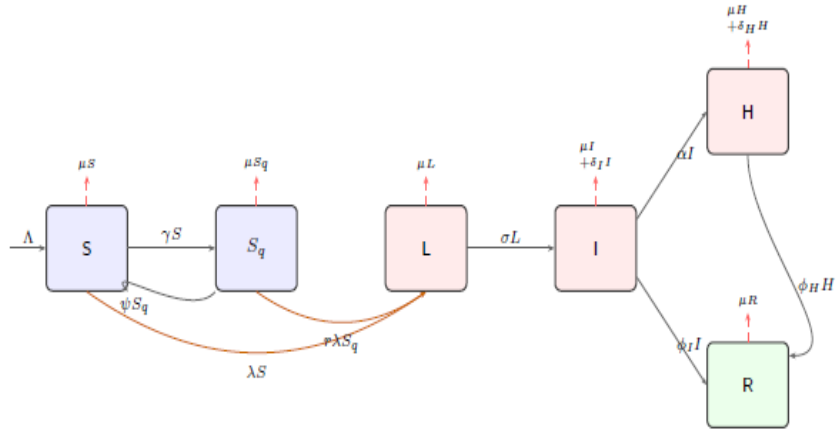


Figure 1: Susceptible (S, S_q), Latent (L), Infectious (I), Hospitalized (H), and Recovered (R). Transmission flows, orange arrows show infection, solid black arrows show transitions, and dashed red arrows show mortality.

Theorem 1. The domain

$$\Omega_+ = \{(S, S_q, L, I, H, R) \in \mathbb{R}_+^6 \mid S > 0, S_q, L, I, H, R \geq 0\}$$

remains positively invariant under the dynamics of system (2).

Proof. The results in [9] guarantee the existence and uniqueness of solutions to (2) on the interval $(0, \infty)$. Restricting the system to the set Ω_+ , it follows that

$$\begin{aligned} \frac{dS}{dt} \Big|_{S=0} &= \Lambda + \psi S_q \geq 0, \\ \frac{dS_q}{dt} \Big|_{S_q=0} &= \gamma S \geq 0, \\ \frac{dL}{dt} \Big|_{L=0} &= \lambda S + r \lambda S_q \geq 0, \\ \frac{dI}{dt} \Big|_{I=0} &= \sigma L \geq 0, \\ \frac{dH}{dt} \Big|_{H=0} &= \theta I \geq 0, \\ \frac{dR}{dt} \Big|_{R=0} &= \psi_I I + \psi_H H \geq 0. \end{aligned} \tag{3}$$

Now, if $(S(0), S_q(0), L(0), I(0), H(0), R(0)) \in \Omega_+$, by (3), then the solution (S, S_q, L, I, H, R) cannot leave the hyper-planes $S = 0$, $S_q = 0$, $L = 0$, $I = 0$, $H = 0$, or $R = 0$. The vector field is either tangent to each hyperplane or directed inward relative to Ω_+ . Consequently, all solutions remain confined within Ω_+ , proving its positive invariance. $\square$

Theorem 2. The set Ω_+ is positively invariant, and all trajectories with initial conditions in $\mathbb{R}_+^6$ eventually enter Ω_+ under the dynamics of (2).

Proof. Summing all the right-hand side expressions of model (2) yields

$$\begin{aligned}\frac{dN}{dt} &= \Lambda - \theta(S + S_q + L + I + H + R) - \delta_I I - \delta_H H \\ &\leq \Lambda - \theta(S + S_q + L + I + H + R) = \Lambda - \theta N,\end{aligned}$$

which can be written as

$$N(t) \leq \frac{\Lambda}{\theta}(1 - e^{-\theta t}) + N(0)e^{-\theta t}.$$

By choosing initial values from Ω_+ , it results that $N(t) \leq \frac{\Lambda}{\theta}$. Hence, the set Ω_+ is positively invariant. $\square$

3.1 Disease-free equilibrium

The disease-free equilibrium (DFE) of system (2) is obtained by setting all right-hand side terms to zero and solving for the state variables under the condition that all infection-related components are zero:

$$\varepsilon_0 = (S^0, S_q^0, L^0, I^0, H^0, R^0) = \left(\frac{\Lambda(\theta + \psi)}{\theta(\theta + \psi + \gamma)}, \frac{\Lambda\gamma}{\theta(\theta + \psi + \gamma)}, 0, 0, 0, 0 \right). \quad (4)$$

Applying the next-generation matrix method [3, 13], we derive the basic reproduction number $\mathcal{R}_0$ as the spectral radius of the matrix product $\mathcal{F}\mathcal{V}^{-1}$, which governs the local stability of the DFE ε_0 . Here, $\mathcal{F}$ represents the new infection terms and $\mathcal{V}$ the transition terms for system (2), defined as follows:

$$\mathcal{F} = \begin{bmatrix} 0 & \frac{\gamma\eta_I[(\theta+\psi)+r\gamma]}{(\theta+\psi+\gamma)} & \frac{\gamma\eta_H[(\theta+\psi)+r\gamma]}{(\theta+\psi+\gamma)} & 0 \\ 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 \end{bmatrix}, \quad (5)$$

$$\mathcal{V} = \begin{bmatrix} \theta + \sigma & 0 & 0 & 0 \\ -\sigma & \theta + \psi_I + \theta + \delta_I & 0 & 0 \\ 0 & -\theta & \psi_H + \theta + \delta_H & 0 \\ 0 & -\psi_I & -\psi_H & \theta \end{bmatrix}. \quad (6)$$

The basic reproduction number $\mathcal{R}_0$ for system (2) is given by

$$\mathcal{R}_0 = \rho(\mathcal{F}\mathcal{V}^{-1}) = \frac{\gamma\sigma(\theta + \psi + r\gamma)(\eta_H\theta + \eta_I k_3)}{k_1 k_2 k_3(\theta + \psi + \gamma)},$$

where

$$\begin{aligned} k_1 &= \sigma + \theta, \\ k_2 &= \theta + \psi_I + \theta + \delta_I, \\ k_3 &= \psi_H + \theta + \delta_H. \end{aligned} \quad (7)$$

For more details see [3].

Theorem 3. The DFE of system (2) is locally asymptotically stable (L.A.S) when $\mathcal{R}_0 < 1$, and unstable when $\mathcal{R}_0 > 1$.

Proof. The Jacobian matrix of system (2) evaluated at the DFE point ε_0 is

$$\mathcal{J}_{\varepsilon_0} = \begin{bmatrix} -(\theta + \gamma) & \psi & 0 & 0 & 0 & 0 \\ \gamma & -(\theta + \psi) & 0 & 0 & 0 & 0 \\ -\frac{\gamma(\theta + \psi)}{\theta + \psi + \gamma} - \frac{\gamma r \gamma}{\theta + \psi + \gamma} & -(\theta + \sigma) & \frac{\gamma \theta \eta_I [(\theta + \psi) + r \gamma]}{\theta(\theta + \psi + \gamma)} & \frac{\gamma \theta \eta_H [(\theta + \psi) + r \gamma]}{\theta(\theta + \psi + \gamma)} & 0 & 0 \\ 0 & 0 & \sigma & -(\theta + \psi_I + \theta + \delta_I) & 0 & 0 \\ 0 & 0 & 0 & \theta & -(\theta + \psi_H + \delta_H) & 0 \\ 0 & 0 & 0 & \psi_I & \psi_H & -\theta \end{bmatrix}.$$

This matrix can be decomposed as

$$\mathcal{J}_{\varepsilon_0} = \begin{bmatrix} A & 0 \\ B & \mathcal{F} - \mathcal{V} \end{bmatrix},$$

where

$$A = \begin{bmatrix} -(\theta + \gamma) & \psi \\ \gamma & -(\theta + \psi) \end{bmatrix}, \quad B = \begin{bmatrix} -\frac{\gamma(\theta + \psi)}{\theta + \psi + \gamma} - \frac{\gamma r \gamma}{\theta + \psi + \gamma} & \\ 0 & 0 \\ 0 & 0 \\ 0 & 0 \end{bmatrix}.$$

The DFE is L.S.A if and only if all eigenvalues of $\mathcal{J}_{\varepsilon_0}$ have negative real parts. Since $\mathcal{J}_{\varepsilon_0}$ is block lower-triangular, its eigenvalues consist of those of A and those of $\mathcal{F} - \mathcal{V}$. The eigenvalues of A are explicitly

$$-\theta \quad \text{and} \quad -(\theta + \psi + \gamma),$$

both of which are negative. Hence, stability depends on the spectral properties of $\mathcal{F} - \mathcal{V}$. The conditions reduce to

$$\lambda(\mathcal{J}_{\varepsilon_0}) < 0 \iff \lambda(\mathcal{F} - \mathcal{V}) < 0 \iff \rho(\mathcal{F}\mathcal{V}^{-1}) < 1,$$

where $\rho(\cdot)$ denotes the spectral radius and the invertibility of $\mathcal{V}$ is assumed.

Since the basic reproduction number $\mathcal{R}_0$ is defined by $\mathcal{R}_0 = \rho(\mathcal{F}\mathcal{V}^{-1})$, we conclude the threshold behavior:

- If $\mathcal{R}_0 < 1$, then the DFE is L.A.S.
- If $\mathcal{R}_0 > 1$, then the DFE is unstable.

□

To analyze the global stability of the DFE, we rigorously apply [12, Theorem 2.1], which provides sufficient conditions for the global asymptotic stability of an equilibrium in epidemiological models.

We first rewrite system (2) in the form

$$x' = (\mathcal{F} - \mathcal{V})x,$$

where $x = [L, I, H, R]^T$ denotes the infected state variables, $\mathcal{F}$ represents the matrix of new infection terms, and $\mathcal{V}$ aggregates transition and removal terms as defined in (5) and (6), respectively. We verify the required conditions for the method:

1. By the model construction and equations (5) and (6), the matrix $\mathcal{F}$ satisfies $\mathcal{F} \geq 0$ (componentwise nonnegative), and the matrix $\mathcal{V}$ is a nonsingular M -matrix with inverse $\mathcal{V}^{-1} \geq 0$, explicitly given by

$$\mathcal{V}^{-1} = \begin{bmatrix} \frac{1}{k_1} & 0 & 0 & 0 \\ \frac{\sigma}{k_1 k_2} & \frac{1}{k_2} & 0 & 0 \\ \frac{\theta \sigma}{k_1 k_2 k_3} & \frac{\theta}{k_2 k_3} & \frac{1}{k_3} & 0 \\ \frac{\sigma(\theta \psi_H + \delta_H \psi_I + \theta \psi_I + \psi_H \psi_I)}{\theta k_1 k_2 k_3} & \frac{\theta \psi_H + \delta_H \psi_I + \theta \psi_I + \psi_H \psi_I}{\theta k_2 k_3} & \frac{\psi_H}{\theta k_3} & \frac{1}{\theta} \end{bmatrix}.$$

2. The next-generation matrix $\mathcal{K} = \mathcal{V}^{-1}\mathcal{F}$ has a nonnegative left eigenvector $\omega^T = [0, \eta_I, \eta_H, 0] \geq 0$ corresponding to the spectral radius $\mathcal{R}_0$, that is,

$$\omega^T \mathcal{K} = \mathcal{R}_0 \omega^T.$$

3. Define the Lyapunov function

$$Q(x) = \omega^T \mathcal{V}^{-1}x,$$

which is nonnegative for $x \geq 0$.

4. Differentiating Q along solutions of the system gives

$$\begin{aligned} Q' &= \omega^T \mathcal{V}^{-1}x' = \omega^T \mathcal{V}^{-1}((\mathcal{F} - \mathcal{V})x) \\ &= \omega^T (\mathcal{K} - I)x = (\mathcal{R}_0 - 1)\omega^T x, \end{aligned}$$

where we used the eigenvector relation $\omega^T \mathcal{K} = \mathcal{R}_0 \omega^T$.

Since $\omega^T x \geq 0$, it follows that, if $\mathcal{R}_0 < 1$, then $Q' \leq 0$, with equality only at $x = 0$, that is, the DFE. By LaSalle's Invariance Principle, the largest invariant set where $Q' = 0$ is the singleton $\{0\}$.

Therefore, all trajectories of the infectious subsystem converge to the DFE, that means the DFE is globally asymptotically stable when $\mathcal{R}_0 < 1$.

Theorem 4. System (2) at the DFE is globally asymptotically stable if and only if $\mathcal{R}_0 < 1$.

3.2 Endemic equilibrium point

The endemic equilibrium point (EEP), denoted by

$$\varepsilon^* = (S^*, S_q^*, L^*, I^*, H^*, R^*),$$

of system (2) is found by solving the steady-state conditions

$$\dot{S} = \dot{S}_q = \dot{L} = \dot{I} = \dot{H} = \dot{R} = 0.$$

The equilibrium values are given by

$$\begin{aligned}
 S^* &= \frac{\Lambda(r\lambda^* + \theta + \psi)}{(\theta + r\lambda^* + \psi)(\theta + \lambda^* + \gamma) - \psi\gamma}, \\
 S_q^* &= \frac{\Lambda\gamma}{(\theta + r\lambda^* + \psi)(\theta + \lambda^* + \gamma) - \psi\gamma}, \\
 L^* &= \frac{\Lambda\lambda^*[\Lambda(r(\lambda^* + \gamma) + \theta + \psi)]}{k_1[(\theta + r\lambda^* + \psi)(\theta + \lambda^* + \gamma) - \psi\gamma]}, \\
 I^* &= \frac{\sigma\Lambda\lambda^*[\Lambda(r(\lambda^* + \gamma) + \theta + \psi)]}{k_1k_2[(\theta + r\lambda^* + \psi)(\theta + \lambda^* + \gamma) - \psi\gamma]}, \\
 H^* &= \frac{\theta\sigma\Lambda\lambda^*[\Lambda(r(\lambda^* + \gamma) + \theta + \psi)]}{k_1k_2k_3[(\theta + r\lambda^* + \psi)(\theta + \lambda^* + \gamma) - \psi\gamma]}, \\
 R^* &= \frac{\sigma\Lambda\lambda^*[\Lambda(r(\lambda^* + \gamma) + \theta + \psi)][\psi_Ik_3 + \psi_H\theta]}{\theta k_1k_2k_3[(\theta + r\lambda^* + \psi)(\theta + \lambda^* + \gamma) - \psi\gamma]}.
 \end{aligned} \tag{8}$$

The endemic steady-state value λ^* , defined in (1), satisfies the quadratic equation

$$a_2(\lambda^*)^2 + a_1\lambda^* + a_0 = 0,$$

with coefficients

$$\begin{aligned}
 a_2 &= k_1k_2k_3 > 0, \\
 a_1 &= k_1k_2k_3[r(\theta + \gamma) + (\theta + \psi)] - \gamma\theta r[k_3(k_2 + \eta_I\sigma) + \eta_H\theta\sigma], \\
 a_0 &= \theta(1 - \mathcal{R}_0)(\theta + \psi + \gamma)k_1k_2k_3,
 \end{aligned}$$

where k_1, k_2, k_3 are defined in (7). Using the relation $\lambda_1\lambda_2 = \frac{a_0}{a_2}$, the discriminant analysis of this quadratic shows that a positive EEP exists if and only if $\mathcal{R}_0 > 1$.

Theorem 5. The EEP of system (2), given in (8), is globally asymptotically stable in the interior of the feasible region when $\mathcal{R}_0 > 1$.

Proof. To prove the globally asymptotically stable of EEP, when $\mathcal{R}_0 > 1$, we construct a Lyapunov function V

$$\begin{aligned}
 V &= \left(S - S^* - S^* \ln \frac{S}{S^*} \right) + \left(S_q - S_q^* - S_q^* \ln \frac{S_q}{S_q^*} \right) \\
 &\quad + \left(L - L^* - L^* \ln \frac{L}{L^*} \right) + \frac{k_1}{\sigma} \left(I - I^* - I^* \ln \frac{I}{I^*} \right) \\
 &\quad + \frac{k_1}{\sigma} \cdot \frac{\eta_H\theta}{\eta_{\text{eff}}k_3} \left(H - H^* - H^* \ln \frac{H}{H^*} \right),
 \end{aligned}$$

where $\eta_{eff} = \eta_I + \eta_H \frac{\theta}{k_3}$. By introducing $g(u) = u - 1 - \ln u$, it is clear that $g(u) \geq 0$ for $u > 0$, with equality only at $g(1) = 0$. Since $g(u)$ is convex and has a global minimum at $u = 1$. The Lyapunov function V can be expressed in terms of the function $g(u)$ as follows:

$$V = g\left(\frac{S}{S^*}\right) S^* + g\left(\frac{S_q}{S_q^*}\right) S_q^* + g\left(\frac{L}{L^*}\right) L^* \\ + \frac{k_1}{\sigma} g\left(\frac{I}{I^*}\right) I^* + \frac{k_1 \eta_H \theta}{\sigma \eta_{eff} k_3} g\left(\frac{H}{H^*}\right) H^*.$$

It is clear that $V(X) \geq 0$ for $X \geq 0$, with equality only at $V(\varepsilon^*) = 0$.

Now, we analyze the time derivative of the composite Lyapunov function $V = V_1 + V_2 + V_3 + V_4 + V_5$ to establish global stability of the EEP. Each component corresponds to a specific compartment in our epidemic model. For the susceptible compartment, we examine

$$\frac{dV_1}{dt} = \left(1 - \frac{S^*}{S}\right) \frac{dS}{dt}.$$

Substituting the $\frac{dS}{dt}$ from (2), we have

$$\frac{dV_1}{dt} = \left(1 - \frac{S^*}{S}\right) [\Lambda - \lambda S - (\theta + \gamma)S + \psi S_q].$$

Applying the equilibrium condition $\Lambda - \lambda^* S^* - (\theta + \gamma)S^* + \psi S_q^* = 0$, we can get

$$\frac{dV_1}{dt} = \left(1 - \frac{S^*}{S}\right) [\lambda^* S^* - \lambda S + (\theta + \gamma)(S^* - S) + \psi(S_q - S_q^*)].$$

Expanding this expression yields

$$\frac{dV_1}{dt} = (\lambda^* S^* - \lambda S) + \psi(S_q - S_q^*) \\ - \frac{S^*}{S} [\lambda^* S^* - \lambda S + \psi(S_q - S_q^*)] - (\theta + \gamma) \frac{(S - S^*)^2}{S}.$$

Similarly, for the other components, it can be calculated

$$\frac{dV_2}{dt} = -(\psi + \theta) \frac{(S_q - S_q^*)^2}{S_q} + \gamma(S - S^*) \left(1 - \frac{S_q^*}{S_q}\right) \\ + r\lambda^*(S_q^* - S_q) \left(1 - \frac{S_q^*}{S_q}\right) + r(\lambda^* - \lambda)S_q \left(1 - \frac{S_q^*}{S_q}\right), \\ \frac{dV_3}{dt} = (\lambda S + r\lambda S_q) - k_1(L - L^*) \\ - \frac{L^*}{L} [\lambda S + r\lambda S_q - (\lambda^* S^* + r\lambda^* S_q^*) - k_1(L - L^*)],$$

Table 2: Key parameter values for model (2).

Parameter	Value	Source
Λ	11400	[11]
η_I	0.5	[11]
η_H	0.15	[11]
γ	Changes	Assumed
θ	0.00021	[4]
δ_I	0.00999	[2]
δ_H	0.0007	[2]
θ	0.0267	[2]
ψ	0.4947	[2]
γ	0.2422	[2]
r	0.37	Assumed
σ	0.25	[10]
ψ_I	0.27	[10]
ψ_H	0.5	[10]

$$\begin{aligned}\frac{dV_4}{dt} &= k_1 \left(1 - \frac{I^*}{I}\right) (L - L^*) - \frac{k_1 k_2 (I - I^*)^2}{\sigma I}, \\ \frac{dV_5}{dt} &= \frac{k_1 \eta_H \theta}{\sigma \eta_{\text{eff}} k_3} \left(1 - \frac{H^*}{H}\right) [\theta(I - I^*) - k_3(H - H^*)].\end{aligned}$$

After algebraic calculations and systematic cancellation of all cross-interaction terms, the complete time derivative becomes

$$\begin{aligned}\frac{dV}{dt} &= -(\theta + \gamma) \frac{(S - S^*)^2}{S} - (\psi + \theta) \frac{(S_q - S_q^*)^2}{S_q} - k_1 \frac{(L - L^*)^2}{L} \\ &\quad - \frac{k_1 k_2 (I - I^*)^2}{\sigma I} - \frac{k_1 \eta_H \theta (H - H^*)^2}{\sigma \eta_{\text{eff}} H}.\end{aligned}\tag{9}$$

Therefore $\frac{dV}{dt} \leq 0$ for all trajectories, with equality if and only if

$$(S, S_q, L, I, H) = (S^*, S_q^*, L^*, I^*, H^*).\tag{10}$$

By using the LaSalle Theorem [7], every bounded trajectory approaches the invariant set Ω_+ to EEP when $\mathcal{R}_0 > 1$. $\square$

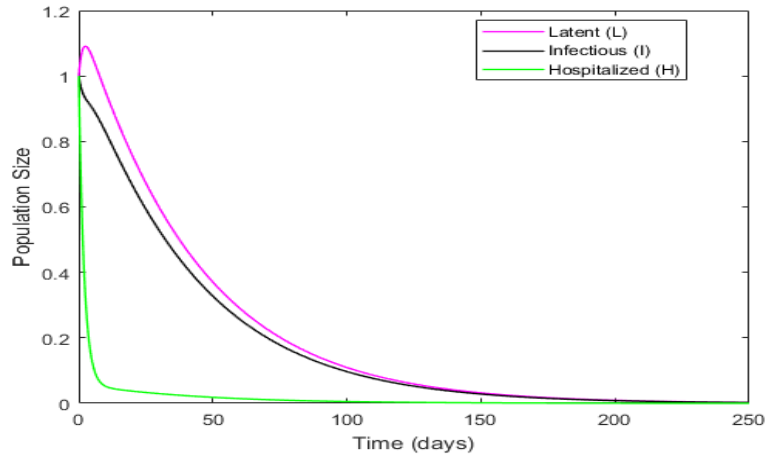


Figure 2: The population of latent, infectious and hospitalized individuals for $\mathcal{R}_0 < 1$.

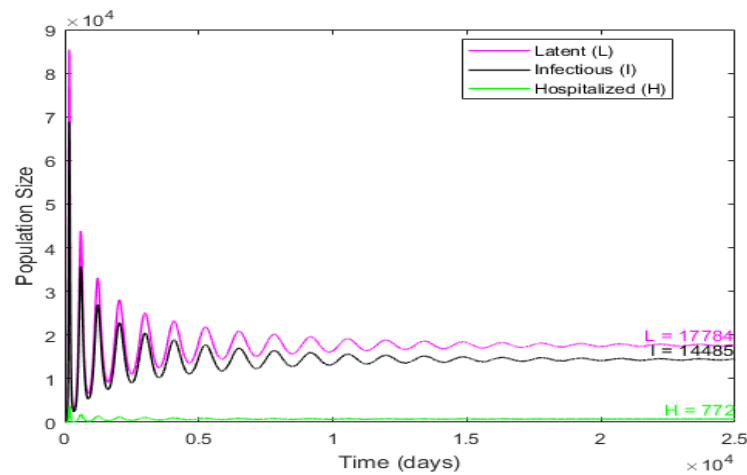


Figure 3: The population of latent, infectious and hospitalized individuals for $\mathcal{R}_0 > 1$.

4 Numerical simulations

The numerical simulations for system (2) are performed in MATLAB using the ODE45 solver. For Figure 1, the time interval $[0, 250]$ is discretized into $N = 1000$ uniform steps to ensure adequate resolution of the system dynamics. For Figure 2, the time interval $[0, 2.5 \times 10^4]$ is discretized into $N = 10^6$ uniform steps to achieve sufficient resolution of the system dynamics.

Using the parameter values listed in Table 2, we computed the basic reproduction number $\mathcal{R}_0$ for two transmission rates: $\gamma = 0.63$ and $\gamma = 1.23$. The corresponding reproduction numbers

were found to be $\mathcal{R}_0 = 0.8262$ and $\mathcal{R}_0 = 1.6131$, respectively. These values allow us to analyze the system's behavior under different epidemiological scenarios.

DFE Stability ($\mathcal{R}_0 < 1$)

For $\gamma = 0.63$ ($\mathcal{R}_0 = 0.8262 < 1$), the system exhibits convergence to the DFE. As illustrated in Figure 2, the population of latent individuals, infectious individuals, and hospitalized individuals consistently decline over time, eventually approaching zero. This numerical observation confirms the theoretical stability of the DFE when $\mathcal{R}_0 < 1$, indicating that the infection cannot sustain itself and will eventually die out under these conditions.

EEP Stability ($\mathcal{R}_0 > 1$)

Conversely, for $\gamma = 1.23$ ($\mathcal{R}_0 = 1.6131 > 1$), the system transitions to an EEP. Figure 3 demonstrates that the subpopulations of latent, infectious, and hospitalized individuals stabilize at nonzero equilibrium values over time. This persistent behavior suggests that the disease remains endemic in the population, with the EEP exhibiting asymptotic stability. The convergence of trajectories to fixed values further supports the existence of a stable endemic state when $\mathcal{R}_0 > 1$.

Theoretical and numerical consistency

The numerical simulations align precisely with our analytical stability analysis. The clear distinction in system dynamics decaying populations for $\mathcal{R}_0 < 1$ versus sustained endemic levels for $\mathcal{R}_0 > 1$ validates our theoretical predictions.

These findings have important implications for public health strategies. When $\mathcal{R}_0 < 1$, control measures need only to maintain conditions below this threshold to ensure disease eradication. However, if $\mathcal{R}_0 > 1$, then long-term management strategies must account for the stable endemic state, emphasizing the need for sustained interventions such as vaccination, improved treatment rates, or reduced transmission.

5 Conclusion

In this study, we developed and analyzed an infectious disease model that incorporates both quarantine and nonquarantine groups. Our analysis shows that when the basic reproduction number is below one, the DFE is globally stable, emphasizing the critical role of targeted public health measures in controlling outbreaks. Moreover, by utilizing a Lyapunov function approach, we demonstrated the global asymptotic stability of the endemic equilibrium, providing insight into the complex behavior of sustained infections. These theoretical results are further validated by numerical simulations, reinforcing the robustness of our findings and their potential implications for disease control strategies.

Funding

This research received no external funding.

Conflicts of Interest

The authors declare no conflicts of interest.

References

- [1] Aakash, M., Gunasundari, C., Athithan, S., Sharmila, N.B., Kumar, G.S. and Guefaifia, R. *Mathematical modeling of COVID-19 with the effects of quarantine and detection*, Partial Differ. Equ. Appl. Math., 9, (2024) 100609.
- [2] Al-Yahyai, M., Al-Musalhi, F., Elmojtaba, I. and Al-Salti, N. *Mathematical analysis of a COVID-19 model with different types of quarantine and isolation*, Math. Biosci. Eng., 20 (2022), 1344–1375.
- [3] Diekmann, O., Heesterbeek, J.A.P. and Metz, J.A.J. *On the definition and computation of the basic reproduction ratio $\mathcal{R}_0$ in models for infectious disease in heterogeneous population*, J. Math. Biol. 28 (1990), 365–382.

- [4] Ega, T.T. and Ngeleja, R.C. *Mathematical Model Formulation and Analysis for COVID-19 Transmission with Virus Transfer Media and Quarantine on Arrival*, Comput. Math. Methods, 1 (2022), 1–16.
- [5] Kheiri, H. and Jafari, M. *Fractional optimal control of an HIV/AIDS epidemic model with random testing and contact tracing*, J. Appl. Math. Comput., 60 (2019), 387–411.
- [6] Kiseleva, O., Yakovlev, S., Prytomanova, O. and Kuzenkov, O. *Mathematical modeling of regional infectious disease dynamics based on extended compartmental models*, Computation, 13, (2025) 187. 14.
- [7] LaSalle, J.P. *The stability of dynamical systems*, Society for Industrial and Applied Mathematics, Philadelphia, 1976.
- [8] Lotfi, M., Jabbari, A. and Kheiri, H. *A mathematical analysis of a tuberculosis epidemic model with two treatments and exogenous re-infection*, Int. J. Biomath., 13 (08) (2020), 2050082.
- [9] Lynch, S. *Dynamical systems with applications using MATLAB*, Springer Science & Business Media, 20
- [10] Ndairou, F., Area, I., Nieto, J.J. and Torres, D.F., *Mathematical modeling of COVID-19 transmission dynamics with a case study of Wuhan*, Chaos Solit. Fract., 135 (2020), 1–6.
- [11] Safdar, S., Ngonghala, C.N. and Gumel, A.B. *Mathematical assessment of the role of waning and boosting immunity against the BA.1 Omicron variant in the United States*, Math. Biosci. Eng., 20 (2023), 179–212.
- [12] Shuai Z. and van den Driessche, P. *Global stability of infectious disease models using Lyapunov function*, SIAM J. Appl. Math., 73 (2013), 1513–1532.
- [13] Van den Driessche, P. and Watmough, J. *Reproduction numbers and subthreshold endemic equilibria for compartmental models of disease transmission*, Math. Biosci., 180 (2002), 29–48.



Synchronization and control of discrete incommensurate fractional-order and variable-order neural networks

S. Momani, I.M. Batiha*, , A. Hioual and A. Ouannas

Abstract

The examination of dynamics and synchronization within discrete fractional neural networks has attracted notable interest in contemporary studies. Nevertheless, current research has primarily focused on commensurate discrete fractional-order neural networks. This study aims to

*Corresponding author

Received 15 August 2025; revised 19 December 2025; accepted 13 January 2026

Shaher Momani

Department of Mathematics, Faculty of Science, University of Jordan, Amman 11942, Jordan. Nonlinear Dynamics Research Center (NDRC), Ajman University, Ajman, UAE. e-mail: s.momani@ju.edu.jo

Iqbal M. Batiha

Nonlinear Dynamics Research Center (NDRC), Ajman University, Ajman, UAE.

Department of Mathematics, Al Zaytoonah University of Jordan, Amman 11733, Jordan. e-mail: i.batiha@zuj.edu.jo

Amel Hioual

Laboratory of Dynamical Systems and Control, University of Oum EL-Bouaghi, Oum El Bouaghi 04000, Algeria. e-mail: amelhioual4@gmail.com

Adel Ouannas

Department of Mathematics and Computer Science, University of Oum EL-Bouaghi, Oum El Bouaghi 04000, Algeria. e-mail: ouannas.adel@univ-oeb.dz

How to cite this article

Momani, S., Batiha, I.M., Hioual, A. and Ouannas, A., Synchronization and control of discrete incommensurate fractional-order and variable-order neural networks. *Iran. J. Numer. Anal. Optim.*, 2026; 16(2): 683-702. <https://doi.org/10.22067/ijnao.2026.94933.1707>

explore the synchronization of incommensurate discrete fractional neural networks, spanning both constant and variable orders. By employing linear feedback control techniques, we establish a satisfactory criterion to guarantee the synchronization of noncommensurate discrete fractional neural networks with constant orders. This condition is formulated in terms of linear matrix inequalities, offering a systematic approach to synchronization. Furthermore, under specific conditions, the Lyapunov functional is employed to analyze the synchronization of noncommensurate discrete fractional neural networks with variable orders. This synchronization condition is solely dependent on the system parameters, facilitating easy verification and implementation. In order to confirm the efficacy and practicality of the proposed methodologies, two numerical examples are selected and presented, demonstrating the successful synchronization of incommensurate discrete fractional neural networks.

AMS subject classifications (2020): Primary 39A12; Secondary 34D06, 34D20, 92B20, 93C55, 39A30.

Keywords: Discrete incommensurate fractional neural networks; Discrete incommensurate variable-order neural networks; Lyapunov functional; Complete synchronization.

1 Introduction

In recent years, the analysis of nonlinear dynamical systems has increasingly focused on synchronization phenomena, owing to their fundamental role in science and engineering applications such as secure communications, signal processing, neural computation, and coordinated control of complex networks. With the widespread use of digital devices and numerical platforms, discrete-time models have become particularly relevant, as they naturally arise in digital circuits, computer simulations, and networked control implementations. Consequently, the synchronization and control of discrete-time chaotic systems has emerged as an active research area, especially when memory effects and long-term dependencies are taken into account.

To accurately capture such memory effects, fractional-order modeling has been widely adopted. Fractional-order systems generalize classical integer-order models by introducing non-local operators that depend on the entire past evolution of the system. In the discrete-time setting, fractional difference operators provide a flexible mathematical framework for modeling systems with hereditary properties. Foundational developments in discrete fractional calculus, including Riemann-type and nabla-type fractional differences and their associated transforms, have been established in works such as [2, 12]. More recently, stability and asymptotic behavior of incommensurate nabla fractional difference systems have been investigated in [7], while nonlinear nabla variable-order fractional discrete systems and their applications to neural networks

were studied in [13]. These results form the mathematical basis for the analysis of fractional discrete neural networks.

Within this framework, a wide range of synchronization and control strategies has been proposed for discrete fractional-order chaotic systems. Active control techniques, which rely on real-time state reconstruction, have been successfully applied to achieve synchronization and secure communication in fractional-order discrete-time chaotic systems [8]. Impulsive control methods, based on intermittent control actions, were developed in [23] to synchronize fractional-order discrete chaotic systems with reduced control effort. Adaptive and sliding-mode-based strategies have also been investigated to improve robustness and precision, as demonstrated in fractional-order discrete-time synchronization problems arising in engineering applications [20]. In addition, linear state error feedback control remains a simple and effective approach for synchronizing chaotic discrete dynamical systems [36].

The incorporation of fractional difference operators into neural network models has revealed a rich variety of dynamical behaviors in discrete-time systems. Fractional-order discrete neural networks are capable of exhibiting chaos, bifurcations, multistability, and complex attractors, which are not observed in their integer-order counterparts. Chaos generation and applications to image encryption were explored in fractional-order discrete neural networks in [5], while incommensurate fractional discrete neural networks with complex chaotic behavior were studied in [1]. The dynamics of fractional-order neural networks with discrete and distributed delays were investigated in [33]. Moreover, extensive research has been devoted to the stability and control of fractional discrete neural networks, including new stability criteria [17], solvability and stability of variable-order fractional discrete neural networks [16, 3], stabilization and synchronization of nonlinear fractional discrete neural networks [15], finite-time stability results for variable-order fractional difference equations [11], and finite-time stability analysis of various types of fractional discrete neural networks [14, 24].

Among the many dynamical phenomena observed in fractional neural networks, synchronization plays a particularly important role. Synchronization enables coordinated behavior between coupled neural systems and is essential for applications in information transmission and secure communications. Synchronization of fractional-order discrete-time neural networks with time delays has been studied in [10], while complete synchronization of discrete-time fractional-order coupled neural networks was analyzed in [6]. Global Mittag-Leffler stability and synchronization of complex-valued discrete-time fractional-order neural networks were investigated in [37], and finite-time synchronization of fractional-order memristive recurrent neural networks was reported in [21]. More refined synchronization patterns have also been explored, including synchronization analysis for discrete fractional-order complex-valued neural networks [22], quasi-projective

synchronization of discrete-time fractional-order quaternion-valued neural networks [38], and lag synchronization of discrete fractional-order neural networks with time delays [10].

Despite these advances, much of the existing literature has focused on commensurate fractional-order systems, where all state variables share the same fractional order. However, in many realistic neural and biological systems, different neurons may possess distinct memory characteristics, leading naturally to incommensurate fractional-order models. In recent years, synchronization of incommensurate fractional-order neural networks has therefore attracted increasing attention. Predefined-time synchronization of incommensurate fractional-order competitive neural networks with time-varying delays was studied in [35]. Synchronization in incommensurate Riemann–Liouville fractional-order competitive neural networks with different time scales was investigated in [9]. Lag quasi-synchronization of incommensurate fractional-order memristor-based neural networks with nonidentical characteristics was addressed in [18], while a detailed dynamical analysis and digital circuit realization of incommensurate fractional-order Hopfield neural networks was presented in [34].

In parallel, variable-order fractional models have emerged as a powerful generalization of constant-order fractional systems, allowing the fractional order to vary with time and thereby providing a more accurate description of evolving memory effects. Variable-order fractional discrete neural networks have been analyzed in terms of existence, uniqueness, stability, and finite-time behavior in [16, 3, 11, 14]. These models further enrich the dynamics of neural networks and pose new challenges for synchronization and control design.

It is also worth noting that synchronization of chaotic and fractional-order systems has been extensively studied through generalized, hybrid, and projective synchronization frameworks. Comprehensive treatments of fractional-order control and synchronization can be found in the monograph [4]. Robust and hybrid synchronization methods for fractional and integer-order chaotic systems were developed in [26, 25, 27, 29], while generalized synchronization of discrete-time chaotic systems with different dimensions was investigated in [32]. Fractional-order discrete-time chaotic systems exhibiting chaos, stabilization, and synchronization were studied in [19], and chaos control and synchronization of fractional discrete-time systems with no equilibrium were reported in [31]. Applications of chaotic synchronization to secure communication schemes were further explored in [30], and inverse full-state hybrid function projective synchronization was investigated in [28]. These works provide an important theoretical and methodological foundation for the present study.

Motivated by the above discussion, this paper investigates the synchronization and control of discrete incommensurate fractional-order and variable-order neural networks. Using the Nabla Caputo fractional difference operator, we propose novel discrete neural network models that incorporate both heterogeneous (incommensurate) and time-varying (variable-order) memory

effects. Simple linear control laws are designed to achieve synchronization between drive and response networks. By constructing suitable Lyapunov functionals, new sufficient conditions are derived to guarantee synchronization for both constant-order and variable-order incommensurate neural networks

The remainder of this paper is organized as follows. Section 2 introduces preliminary results from discrete fractional calculus and variable-order fractional difference theory based on the nabla operator. Section 3 presents synchronization criteria for discrete incommensurate fractional-order neural networks with constant orders. Section 4 extends the results to discrete incommensurate variable-order neural networks. Section 5 provides numerical simulations to validate the theoretical findings. Finally, section 6 concludes the paper and outlines directions for future research.

2 Preliminaries

In the following discussion, we introduce fundamental definitions and key principles concerning discrete fractional calculus to lay the foundation for our investigation. Consider real numbers a and b , along with a positive constant h . We utilize the notation $\mathbb{N}_a = \{a, a+1, a+2, \dots\}$ to denote a set of natural numbers starting from a .

Definition 1. [2] For a function $x : \mathbb{N}_\theta \rightarrow \mathbb{R}$ with θ as the starting point, and for $\theta > 0$, the θ th order summation operator is defined as follows:

$${}_a\nabla^{-\theta}x(t) = \frac{1}{\Gamma(\theta)} \sum_{s=a+1}^t (t-\rho(s))^{\theta-1}x(sh), \quad \rho(s) = s-1, \quad t \in \mathbb{N}_a. \quad (1)$$

Additionally, we establish the polynomial factorial function with ascending powers as

$$t^\theta = \frac{\Gamma(t+\theta)}{\Gamma(t)}.$$

Definition 2. [2] Consider a function $x : \mathbb{N}_a \rightarrow \mathbb{R}$ defined on $\mathbb{N}_a$, where $\theta > 0$. The discrete Caputo nabla operator is expressed as

$${}_a^C\nabla^\theta x(t) = \frac{1}{\Gamma(1-\theta)} \sum_{s=a+1}^t (t-\rho(s))^{-\theta}\nabla x(s), \quad t \in \mathbb{N}_a. \quad (2)$$

Here,

$$\nabla x(t) = x(t) - x(t-1). \quad (3)$$

Let the nonlinear incommensurate fractional-order discrete-time system be described by the nabla operator:

$${}_a^C \nabla^\theta x(t) = Ax(t), \quad (4)$$

where $\theta = (\theta_1, \theta_2, \dots, \theta_n) \in (0, 1)$, $x(t) = (x_1(t), x_2(t), \dots, x_n(t))$.

Theorem 2.1. [7] If $\det(I - HA) \neq 0$, then the system (4) possesses a unique solution for $x_0 \in \mathbb{R}^n$. Moreover, if all the eigenvalues of the characteristic equation expressed as

$$\det \left(\text{diag} \left(\left(1 - \frac{1}{s}\right)^{\theta_1}, \left(1 - \frac{1}{s}\right)^{\theta_2}, \dots, \left(1 - \frac{1}{s}\right)^{\theta_n} \right) - \mathbf{A} \right) = 0 \quad (5)$$

reside within the unit disk, then the trivial solution of system (4) achieves asymptotic stability.

Lemma 2.1. Examine the subsequent fractional-order difference inequalities accompanied by initial conditions: $0 \leq V_i(0, x_i(0)) \leq W_i(0, y_i(0))$, $i = 1, 2, \dots$,

$$\begin{aligned} {}_a^C \nabla^{\theta_1} V_1(t, x_1(t)) &\leq {}_a^C \nabla^{\theta_1} W_1(t, y_1(t)), \\ {}_a^C \nabla^{\theta_2} V_2(t, x_2(t)) &\leq {}_a^C \nabla^{\theta_2} W_2(t, y_2(t)), \\ &\vdots \\ {}_a^C \nabla^{\theta_n} V_n(t, x_n(t)) &\leq {}_a^C \nabla^{\theta_n} W_n(t, y_n(t)). \end{aligned}$$

The subsequent inequalities are valid:

$$V_i(t, x_i(t)) \leq W_i(t, y_i(t)), \quad \text{for all } t > 0, \quad i = 1, 2, \dots, n, \quad (6)$$

where $V_i(t, x_i(t))$ and $W_i(t, x_i(t)) : [0, \infty) \times \mathbb{N}_a \rightarrow [0, \infty)$.

Definition 3. [13] Let $0 < \theta(t) \leq 1$ for all $t \in \mathbb{N}_a$. Suppose that we have a function $x : \mathbb{N}_a \rightarrow \mathbb{R}$. We define the left nabla fractional sum of order $\theta(t)$ as follows:

$${}_a \nabla^{-\theta(t)} x(t) = \frac{1}{\Gamma(\theta(t))} \sum_{s=a+1}^t (t - \rho(s))^{\overline{\theta(t)-1}} x(s), \quad t \in \mathbb{N}_{a+1}. \quad (7)$$

Definition 4. [13] Given a function $x : \mathbb{N}_{a+1} \rightarrow \mathbb{R}$ with $0 < \theta(t) < 1$, the Caputo nabla discrete difference with variable-order is expressed by

$${}_a^C \nabla^{\theta(t)} x(t) = \frac{1}{\Gamma(1 - \theta(t))} \sum_{s=a+1}^t (t - \rho(s))^{\overline{-\theta(t)}} \nabla x(s), \quad t \in \mathbb{N}_{a+1}. \quad (8)$$

Lemma 2.2. [13] Let $x : \mathbb{N}_{a+1} \rightarrow \mathbb{R}$; then $x(t)$ fulfills the following inequality:

$${}_a^C \nabla^{\theta(t)} x^2(t) \leq 2x(t) {}_a^C \nabla^{\theta(t)} x(t). \quad (9)$$

Let us explore the dynamical system described by the following nonlinear nabla noncommensurate discrete fractional equations:

$${}_a^C \nabla^{\theta_i(t)} x_i(t) = f_i(t, x_i(t)). \quad (10)$$

In the context of the nabla discrete-time fractional-order system, we consider $\theta(t) \in (0, 1)$, $x(t) \in D \subset \mathbb{R}^n$ representing the state space that includes the equilibrium point $x_e = 0$. The variable t is an element of $\mathbb{N}_{a+1}$, where $a \in \mathbb{R}$, and the function $f : \mathbb{N}_{a+1} \times D \rightarrow \mathbb{R}^n$ adheres to the Lipschitz continuity condition.

Theorem 2.2. [13] If we consider the system (10) defined under the Caputo definition, then the conditions for asymptotic stability can be established through the existence of class K functions $\gamma_1, \gamma_2, \gamma_3$, parameters $\gamma \in (0, \min \theta_i(t))$, $p_i > 0$, and Lyapunov functions $V_i(t, x_i(t)) : \mathbb{N}_a \times D \rightarrow \mathbb{R}$ for each $i = 1, 2, \dots, n$, satisfying

$$\gamma_1(\|x(t)\|) \leq \sum_{i=1}^n p_i {}_a^C \nabla^{\theta_i(t)-\gamma} V_i(t, x_i(t)) \leq \gamma_2(\|x(t)\|), \quad (11)$$

$$\sum_{i=1}^n p_i \frac{\partial V_i(t, x_i(t))}{\partial x_i(t)} f_i(t, x(t)) \leq -\gamma_3(\|x(t)\|). \quad (12)$$

In the context where $x(t) \in D$, $t \in \mathbb{N}_{a+1}$, $a \in \mathbb{R}$, and $V_i(t, x_i(t))$ is a differentiable and convex function concerning $x_i(t)$, the system (10) is deemed to be uniformly asymptotically stable at $x_e = 0$.

3 Synchronization of discrete incommensurate fractional-order neural networks

In the following analysis, we investigate the synchronization of incommensurate discrete fractional-order neural networks, characterized by

$${}_a^C \nabla^{\bar{\theta}} x(t) = -Cx(t) + Bf(t, x(t)). \quad (13)$$

In the provided equation, i belongs to the set of natural numbers $\mathbb{N} = 1, 2, \dots, n$, with t being greater than or equal to zero. The parameter $\bar{\theta} = [\theta_1, \theta_2, \dots, \theta_n]$ denotes the vector of orders, where n represents the total number of units in the neural network. At time t , the state vector $x(t) = (x_1(t), \dots, x_n(t))^T$ lies in $\mathbb{R}^n$, and $f(x(t)) = (f_1(x_1(t)), \dots, f_n(x_n(t)))^T$ signifies the activation function of the neurons. Additionally, $C = \text{diag}(c_1, \dots, c_n)$ signifies the diagonal matrix containing the constant rates at which each unit resets its potential individually, unaffected by external inputs or network activity. The weights governing the connection between the j th and i th neurons at time t are represented by $(a_{ij})_{n \times n}$, and $I = (I_1, I_2, \dots, I_n)^T$ denotes the external bias vector.

Before proceeding further, we state the following hypothesis:

(H₁): The activation functions f_j in the neural network exhibit Lipschitz continuity, ensuring the existence of positive constants l_j , where $j = 1, 2, \dots, n$. These constants are defined such that

$$|f_j(x) - f_j(y)| < l_j |x - y|, \quad \text{for all } x, y \in \mathbb{R}.$$

The system denoted as (13) serves as the driving system, while the slave system is defined by

$${}^C \nabla_a^{\bar{\theta}} y(t) = -Cy(t) + Bf(t, y(t)) + u(t). \quad (14)$$

Here, $y(t) \in \mathbb{R}^n$ represents the state vector of the slave system, with $u(t)$ representing the controller yet to be devised.

The discrepancy in states between the master and slave systems defines the synchronization error, articulated as

$$\mathbf{e}(t) = y(t) - x(t). \quad (15)$$

The error dynamics, deduced from (13) to (14), can be expressed in the following manner:

$${}^C \nabla_a^{\bar{\theta}} \mathbf{e}(t) = -C\mathbf{e}(t) + B(f(t, x(t)) - f(t, y(t))) + u(t). \quad (16)$$

Our objective is to devise an appropriate feedback control strategy,

$$u(t) = K\mathbf{e}(t), \quad (17)$$

where $\mathbf{K} = \text{diag}(\kappa_1, \kappa_2, \dots, \kappa_n)$ is a diagonal matrix, such that the subsequent error system takes the form:

$${}^C \nabla_a^{\bar{\theta}} \mathbf{e}(t) = -(C - K)\mathbf{e}(t) + B(f(t, x(t)) - f(t, y(t))). \quad (18)$$

Alternatively,

$${}^C\nabla_a^{\theta_i} \mathbf{e}_i(t) = -(c_i - k_i)\mathbf{e}_i(t) + \sum_{j=1}^n b_{ij}(f_j(t, x_j(t)) - f_j(t, y_j(t))), \quad i = 1, 2, \dots, n. \quad (19)$$

The asymptotic stability implies that the trajectory of the slave system governed by (14) tends towards the trajectory of the master system over time, initialized at $y(a)$, and has the capability to gradually converge towards the drive system defined by (13), initiated at $x(a)$

$$\lim_{t \rightarrow \infty} \|\mathbf{e}(t)\| = \lim_{t \rightarrow \infty} \|x(t) - y(t)\| = 0. \quad (20)$$

Theorem 3.1. If (H_1) is valid, and the subsequent condition is met:

$$k_i > c_i - \sum_{j=1}^n \frac{|b_{ij}|l_j + |b_{ji}|l_i}{2}, \quad (21)$$

then the master-slave systems (13)-(14) achieve synchronization.

Proof. Choose the auxiliary function:

$$V(t) = \sum_{i=1}^n V_i(t) = \frac{1}{2} \sum_{i=1}^n \mathbf{e}_i^2(t). \quad (22)$$

Employing Lemma 2.2 and computing the θ_i -order difference on $\mathcal{V}_i$, we derive

$$\begin{aligned} {}^C\nabla_a^{\theta_i} V_i(t) &\leq \mathbf{e}_i(t) {}^C\nabla_a^{\theta_i} \mathbf{e}_i(t), \\ &= \mathbf{e}_i(t) \left(-(c_i - k_i)\mathbf{e}_i(t) + \sum_{j=1}^n b_{ij}(f_j(t, x_j(t)) - f_j(t, y_j(t))) \right) \\ &\leq -(c_i - k_i)\mathbf{e}_i^2(t) + \sum_{j=1}^n |b_{ij}|l_j \mathbf{e}_i(t)\mathbf{e}_j(t) \\ &\leq -(c_i - k_i)\mathbf{e}_i^2(t) + \sum_{j=1}^n |b_{ij}|l_j \frac{\mathbf{e}_i^2(t) + \mathbf{e}_j^2(t)}{2} \\ &\leq -(c_i - k_i)\mathbf{e}_i^2(t) + \mathbf{e}_i^2(t) \sum_{j=1}^n \frac{|b_{ij}|l_j + |b_{ji}|l_i}{2} \\ &\leq - \left(c_i - k_i - \sum_{j=1}^n \frac{|b_{ij}|l_j + |b_{ji}|l_i}{2} \right) \mathbf{e}_i^2(t) \\ &= - \left(c_i - k_i - \sum_{j=1}^n \frac{|b_{ij}|l_j + |b_{ji}|l_i}{2} \right) V_i(t). \end{aligned} \quad (23)$$

Utilizing (23), we can formulate the corresponding comparative system:

$$\begin{bmatrix} {}^C\nabla_a^{\theta_1} W_1(t) \\ {}^C\nabla_a^{\theta_2} W_2(t) \\ \vdots \\ {}^C\nabla_a^{\theta_n} W_n(t) \end{bmatrix} = \begin{bmatrix} \epsilon_{11} & 0 & \cdots & 0 \\ 0 & \epsilon_{22} & \cdots & 0 \\ \vdots & \vdots & \ddots & \vdots \\ 0 & 0 & \cdots & \epsilon_{nn} \end{bmatrix} \begin{bmatrix} W_1(t) \\ W_2(t) \\ \vdots \\ W_n(t) \end{bmatrix}, \quad (24)$$

in which

$$\epsilon_{ii} = - \left(c_i - k_i - \sum_{j=1}^n \frac{|b_{ij}|l_j + |b_{ji}|l_i}{2} \right). \quad (25)$$

The system represented by (24) can be reformulated as

$${}^C\nabla_a^{\bar{\theta}} W(t) = AW(t). \quad (26)$$

Let $W(t) = (W_1(t), W_2(t), \dots, W_n(t))^T$ and let $A = \text{diag} \epsilon_{ii} n \times n$. Lemma 2.1 implies that the controlled system (26) is asymptotically stable if a certain condition is met, specifically if $W(t) \rightarrow 0$ when $k_i > c_i - \sum_{j=1}^n \frac{|b_{ij}|l_j + |b_{ji}|l_i}{2}$. By employing the comparison principle from Lemma 2.1, we deduce that $V(t) \leq W(t)$ and $V(t) \rightarrow 0$. Given $V(t) = \sum_{i=1}^n V_i(t) = \sum_{i=1}^n \epsilon_i^2(t)$, it follows that $\epsilon_i(t) \rightarrow 0$. Consequently, the synchronization error system (16) is also stable, thus concluding the proof. $\square$

4 Synchronization of incommensurate discrete variable-order neural network

Let us explore the synchronization of a discrete incommensurate variable-order neural network, denoted by (27). We introduce an innovative theorem to ensure synchronized dynamics between two chaotic neural networks using simple linear control laws. Consider the discrete incommensurate variable-order neural network defined by (27) as the master system:

$$\begin{cases} {}^C\nabla_a^{\theta_1(t)} x_1(t) = -c_1 x_1(t) + \sum_{j=1}^n b_{1j} f_j(t, x_j(t)), \\ {}^C\nabla_a^{\theta_2(t)} x_2(t) = -c_2 x_2(t) + \sum_{j=1}^n b_{2j} f_j(t, x_j(t)), \\ \vdots \\ {}^C\nabla_a^{\theta_n(t)} x_n(t) = -c_n x_n(t) + \sum_{j=1}^n b_{nj} f_j(t, x_j(t)). \end{cases} \quad (27)$$

Conversely, the controlled discrete incommensurate variable-order neural network represented by (28) is regarded as the slave system, with n synchronization controllers denoted as $L_1, L_2, \dots, L_n$, respectively:

$$\begin{cases} {}^C\nabla_a^{\theta_1(t)} y_1(t) = -c_1 y_1(t) + \sum_{j=1}^n b_{1j} f_j(t, y_j(t)) + L_1(t), \\ {}^C\nabla_a^{\theta_2(t)} y_2(t) = -c_2 y_2(t) + \sum_{j=1}^n b_{2j} f_j(t, y_j(t)) + L_2(t), \\ \vdots \\ {}^C\nabla_a^{\theta_n(t)} y_n(t) = -c_n y_n(t) + \sum_{j=1}^n b_{nj} f_j(t, y_j(t)) + L_n(t). \end{cases} \quad (28)$$

To achieve complete synchronization, the synchronization error is defined as follows:

$$\mathbf{e}_i(t) = x_i(t) - y_i(t), \quad i = 1, \dots, n. \quad (29)$$

The synchronization scheme is deemed successful if the following condition holds:

$$\lim_{s \rightarrow +\infty} |\mathbf{e}_i(t)| = 0, \quad \text{for } i = 1, \dots, n. \quad (30)$$

In order to emphasize the pivotal results of the proposed synchronization strategy, the ensuing theorem is introduced to guarantee the synchronization between the master and slave neural networks.

Theorem 4.1. The master-slave systems represented by (27) and (28) achieve synchronization when the controlled variables $L_i, i = 1, \dots, n$, are chosen as follows:

$$L_i(t) = - \left(\sum_{j=1}^n \frac{|b_{ij}|l_j + |b_{ji}|l_i}{2} \right) \mathbf{e}_i(t), \quad i = 1, \dots, n. \quad (31)$$

Proof. To ensure the convergence of the synchronization errors defined in (29), we begin by applying the nabla Caputo-type variable-order differences to (29), resulting in

$$\begin{cases} {}^C\nabla_a^{\theta_1(t)} e_1(t) = -c_1 \mathbf{e}_1(t) + \sum_{j=1}^n b_{1j} (f_j(t, x_j(t)) - f_j(t, y_j(t))) + L_1(t), \\ {}^C\nabla_a^{\theta_2(t)} e_2(t) = -c_2 \mathbf{e}_2(t) + \sum_{j=1}^n b_{2j} (f_j(t, x_j(t)) - f_j(t, y_j(t))) + L_2(t), \\ \vdots \\ {}^C\nabla_a^{\theta_n(t)} e_n(t) = -c_n \mathbf{e}_n(t) + \sum_{j=1}^n b_{nj} (f_j(t, x_j(t)) - f_j(t, y_j(t))) + L_n(t). \end{cases} \quad (32)$$

Substituting the proposed control law (31) for $L_i, i = 1, \dots, n$, in system (32) yields the following updated fractional discrete system:

$$\left\{ \begin{array}{l} {}^C\nabla_a^{\theta_1(t)} \mathbf{e}_1(t) = -c_1 \mathbf{e}_1(t) + \sum_{j=1}^n b_{nj} (f_j(t, x_j(t)) - f_j(t, y_j(t))) \\ \quad - \left(\sum_{j=1}^n \frac{|b_{1j}|l_j + |b_{j1}|l_1}{2} \right) \mathbf{e}_1(t), \\ {}^C\nabla_a^{\theta_2(t)} \mathbf{e}_2(t) = -c_2 \mathbf{e}_2(t) + \sum_{j=1}^n b_{nj} (f_j(t, x_j(t)) - f_j(t, y_j(t))) \\ \quad - \left(\sum_{j=1}^n \frac{|b_{2j}|l_j + |b_{j2}|l_2}{2} \right) \mathbf{e}_2(t), \\ \vdots \\ {}^C\nabla_a^{\theta_n(t)} \mathbf{e}_n(t) = -c_n \mathbf{e}_n(t) + \sum_{j=1}^n b_{nj} (f_j(t, x_j(t)) - f_j(t, y_j(t))) \\ \quad - \left(\sum_{j=1}^n \frac{|b_{nj}|l_j + |b_{jn}|l_n}{2} \right) \mathbf{e}_n(t). \end{array} \right. \quad (33)$$

We proceed to examine the Lyapunov function, which incorporates the collective error state

$$V(t) = \frac{1}{2} \sum_{i=1}^n {}^C\nabla_a^{\theta_i(t)-\gamma} e_i^2(t). \quad (34)$$

This results in ${}^C\nabla_a^\gamma V(t) = \frac{1}{2} \sum_{i=1}^n {}^C\nabla_a^{\theta_i(t)} e_i^2(t)$. Furthermore, utilizing Lemma 2.2, we derive the following inequality:

$$\begin{aligned} {}^C\nabla_a^\gamma V(t) &\leq \sum_{i=1}^n e_i(t) {}^C\nabla_a^{\theta_i(t)} e_i(t) \\ &= \sum_{i=1}^n e_i(t) \left(-c_i \mathbf{e}_i(t) + \sum_{j=1}^n b_{ij} (f_j(t, x_j(t)) - f_j(t, y_j(t))) \right. \\ &\quad \left. - \left(\sum_{j=1}^n \frac{|b_{ij}|l_j + |b_{ji}|l_i}{2} \right) \mathbf{e}_i(t) \right) \\ &\leq \sum_{i=1}^n -c_i \mathbf{e}_i^2(t) + \sum_{j=1}^n |b_{ij}| \mathbf{e}_i(t) \mathbf{e}_j(t) - \left(\sum_{j=1}^n \frac{|b_{ij}|l_j + |b_{ji}|l_i}{2} \right) \mathbf{e}_i^2(t) \\ &\leq \sum_{i=1}^n -c_i \mathbf{e}_i^2(t) + \sum_{j=1}^n |b_{ij}| \frac{\mathbf{e}_i^2(t) + \mathbf{e}_j^2(t)}{2} - \left(\sum_{j=1}^n \frac{|b_{ij}|l_j + |b_{ji}|l_i}{2} \right) \mathbf{e}_i^2(t) \\ &\leq - \sum_{i=1}^n c_i \mathbf{e}_i^2(t) \leq -\gamma \sum_{i=1}^n \mathbf{e}_i^2(t) = -\gamma \|e(t)\| < 0. \end{aligned} \quad (35)$$

Additionally, based on Theorem 2.2, it can be affirmed that the dynamics of the error system (29) for discrete incommensurate variable-order networks converge towards zero, indicating stability. Consequently, the synchronization between the master system (27) and the slave sys-

tem (28) of the discrete incommensurate variable-order neural network is successfully attained through the utilization of linear control laws (31). Thus, the proof is duly concluded. $\square$

5 Numerical examples

In this section, we employ numerical simulations conducted in MATLAB to provide a comprehensive validation of the theoretical findings presented in the preceding theorem. Through detailed numerical experiments, we aim to offer a deeper understanding and insight into the synchronization behavior predicted by our theoretical framework.

Example 1. The master system is the incommensurate discrete fractional neural network with two neurons:

$$\begin{cases} {}^C\nabla_a^{\theta_1} x_1(t) &= -0.25x_1(t) + 0.1 \sin(x_1(t)) - 0.2 \sin(x_2(t)), \\ {}^C\nabla_a^{\theta_2} x_2(t) &= -0.3x_2(t) + 0.3 \sin(x_1(t)) + 0.4 \sin(x_2(t)). \end{cases} \quad (36)$$

The slave system is

$$\begin{cases} {}^C\nabla_a^{\theta_1} y_1(t) &= -1y_1(t) + 0.1 \sin(y_1(t)) - 0.2 \sin(y_2(t)) + k_1(t), \\ {}^C\nabla_a^{\theta_2} y_2(t) &= -1.5y_2(t) + 0.3 \sin(y_1(t)) + 0.4 \sin(y_2(t)) + k_2(t). \end{cases} \quad (37)$$

The synchronization state-feedback gain is determined as follows:

$$K = \begin{bmatrix} 0.2 & 0 \\ 0 & 0.1 \end{bmatrix}, \quad (38)$$

resulting in the matrix A defined as

$$A = \begin{bmatrix} 0.1 & 0 \\ 0 & 0.2 \end{bmatrix}. \quad (39)$$

Certainly, the matrix A fulfills the criteria delineated in both Theorem 2.1 and Theorem 2.2, signifying the asymptotic stability of the error system. This indicates the attainment of synchronization between the driving and responding systems. In our computational trials, we commence with the driving and responding systems initialized at $x(0) = (3, 1, 2)^T$ and $y(0) = (4, 2, 3)^T$, respectively. Figure 1 visually presents the evolution of state synchronization between these systems, while Figure 2 depicts the synchronization error.

Example 2. Let us regard the three-dimensional discrete incommensurate variable-order neural network expressed by (40) and (41) as the master-slave systems for our investigation:

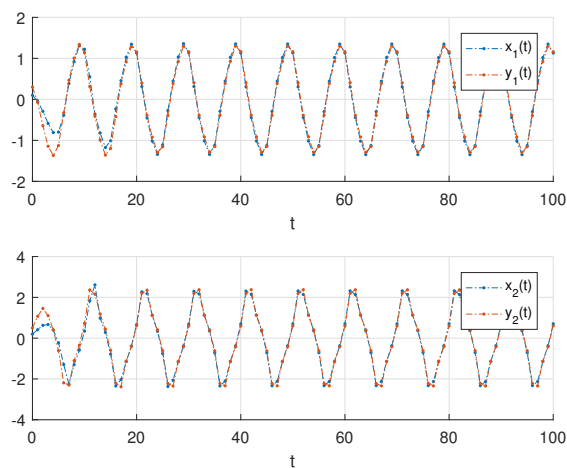


Figure 1: State trajectory of master-slave system (36)–(37).

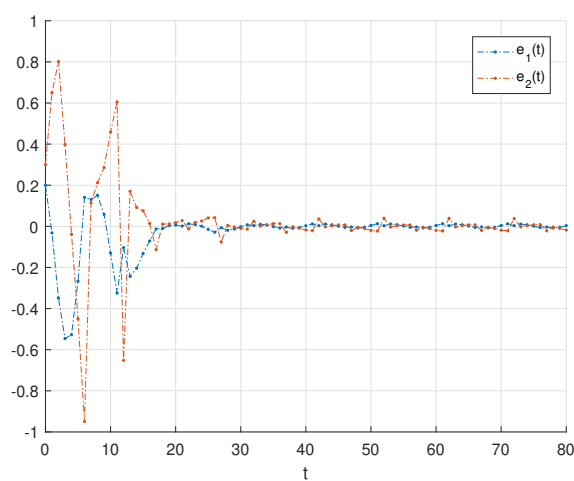


Figure 2: Time evolution of the error system.

$$\begin{cases} {}^C\nabla_a^{\theta_1(t)} x_1(t) = -x_1(t) - 0.5 \tanh(x_1(t)) - 0.7 \tanh(x_2(t)) \\ \quad - 0.5 \tanh(x_3(t)), \\ {}^C\nabla_a^{\theta_2(t)} x_2(t) = -x_2(t) - 0.6 \tanh(x_1(t)) - \tanh(x_2(t)) \\ \quad + 0.1 \tanh(x_3(t)), \\ {}^C\nabla_a^{\theta_3(t)} x_3(t) = -1.3x_3(t) - 0.6 \tanh(x_1(t)) + 0.5 \tanh(x_2(t)) \\ \quad - 0.3 \tanh(x_3(t)), \end{cases} \quad (40)$$

while,

$$\begin{cases} {}^C\nabla_a^{\theta_1(t)} y_1(t) = -y_1(t) - 0.5 \tanh(y_1(t)) - 0.7 \tanh(y_2(t)) \\ \quad - 0.5 \tanh(y_3(t)) + L_1(t), \\ {}^C\nabla_a^{\theta_2(t)} y_2(t) = -y_2(t) - 0.6 \tanh(y_1(t)) - \tanh(y_2(t)) \\ \quad + 0.1 \tanh(y_3(t)) + L_2(t), \\ {}^C\nabla_a^{\theta_3(t)} y_3(t) = -1.3y_3(t) - 0.6 \tanh(y_1(t)) + 0.5 \tanh(y_2(t)) \\ \quad - 0.3 \tanh(y_3(t)) + L_3(t). \end{cases} \quad (41)$$

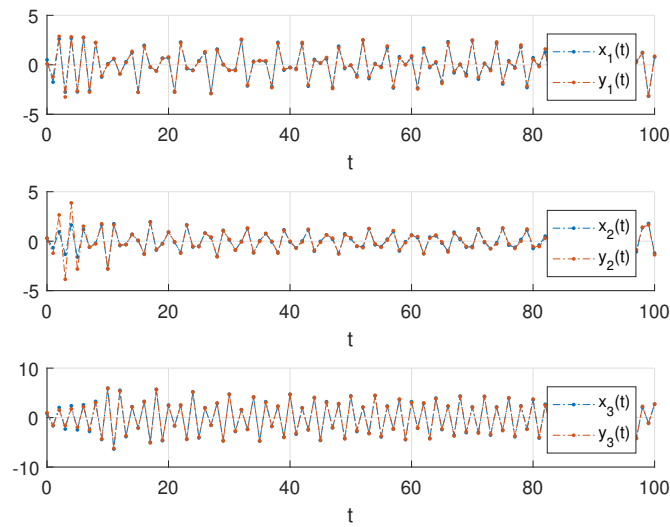


Figure 3: Trajectory of states in the master-slave systems (40)–(41).

The initial states, $\mathbf{e}_1(0)$, $\mathbf{e}_2(0)$, and $\mathbf{e}_3(0)$, are considered as the starting values. To simulate the states of the variable-order error system and synchronized systems (40) and (41), the predictor-corrector method is employed. The initial values are specified as $x(0) = (0.5, 0.3, 0.8)^T$ and $y(0) = (0.1, 0.3, 1)^T$ with

$$\begin{cases} \theta_1(t) = \left| \frac{1}{2} - e^{-4(t+1)} \right|, \\ \theta_2(t) = \frac{1}{5} |\log(t+3)|, \\ \theta_3(t) = \frac{1}{2} - \frac{e^{-\frac{10}{t+1}}}{t+1}. \end{cases} \quad (42)$$

In Figure 3, the numerical states of the discrete incommensurate systems (40)–(41) are depicted. Meanwhile, Figure 4 illustrates the time evolution of the error system. These visualizations

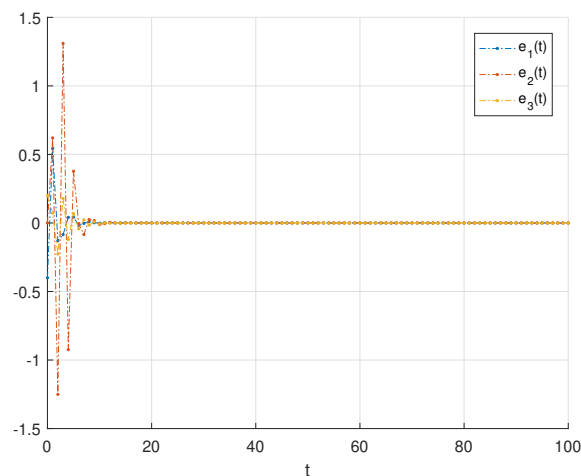


Figure 4: Temporal evolution of the error system.

distinctly demonstrate the synchronized dynamics achieved by the two chaotic fractional neural networks, confirming the effectiveness and simplicity of the linear control laws (31) in achieving synchronization.

6 Conclusion

In conclusion, this study has addressed the synchronization of incommensurate discrete fractional neural networks, which extends previous research focused on commensurate systems. By utilizing linear feedback control methods and establishing sufficient conditions for synchronization in terms of linear matrix inequalities, we provided a systematic approach applicable to networks with both constant and variable orders. The application of the Lyapunov functional under specific conditions further strengthened our analysis, offering insights into synchronization dynamics and facilitating easy verification and implementation. Through numerical examples, we demonstrated the effectiveness and feasibility of our proposed methods, underscoring their potential for practical applications in various fields reliant on network synchronization. Overall, our contributions advance the understanding and practical implementation of synchronization mechanisms in complex fractional-order neural networks.

Funding

This research received no external funding.

Conflicts of Interest

The authors declare no conflicts of interest.

References

- [1] Abbes, A., Ouannas, A., Shawagfeh, N. and Khennaoui, A.A. *Incommensurate fractional discrete neural network: chaos and complexity*, Eur. Phys. J. Plus, 137(2) (2022), 235.
- [2] Abdeljawad, T. *Dual identities in fractional difference calculus within Riemann*, Adv. Differ. Equ., 2013 (2013), 1–16.
- [3] Almatroud, O.A., Hioual, A., Ouannas, A., Sawalha, M.M., Alshammari, S. and Alshammari, M. *On variable-order fractional discrete neural networks: existence, uniqueness and stability*, Fractal Fract., 7(2) (2023), 118.
- [4] Azar, A.T., Vaidyanathan, S. and Ouannas, A. (Eds.). *Fractional order control and synchronization of chaotic systems* (Vol. 688). Springer, 2017.
- [5] Chen, L., Yin, H., Huang, T., Yuan, L., Zheng, S. and Yin, L. *Chaos in fractional-order discrete neural networks with application to image encryption*, Neural Netw., 125 (2020), 174–184.
- [6] Cui, X., Li, H.L., Zhang, L., Hu, C. and Bao, H. *Complete synchronization for discrete-time fractional-order coupled neural networks with time delays*, Chaos Solitons Fractals, 174 (2023), 113772.
- [7] Djenina, N., Ouannas, A., Oussaeif, T.E., Grassi, G., Batiha, I.M., Momani, S. and Albadarneh, R.B. *On the stability of incommensurate h -nabla fractional-order difference systems*, Fractal Fract., 6(3) (2022), 158.
- [8] Djenoune, S., Bettayeb, M. and Al-Saggaf, U.M. *Synchronization of fractional-order discrete-time chaotic systems by an exact delayed state reconstructor: Application to secure communication*, Int. J. Appl. Math. Comput. Sci., 29(1) (2019), 179–194.

- [9] Gu, Y., Wang, H. and Yu, Y. *Synchronization for incommensurate Riemann–Liouville fractional-order time-delayed competitive neural networks with different time scales and known or unknown parameters*, J. Comput. Nonlinear Dyn., 14(5) (2019), 051002.
- [10] Gu, Y., Wang, H. and Yu, Y. *Synchronization for fractional-order discrete-time neural networks with time delays*, Appl. Math. Comput., 372 (2020), 124995.
- [11] Hamadneh, T., Hioual, A., Alsayyed, O., Al-Khassawneh, Y.A., Al-Husban, A. and Ouannas, A. *Finite Time Stability Results for Neural Networks Described by Variable-Order Fractional Difference Equations*, Fractal Fract., 7(8) (2023), 616.
- [12] Hein, J., McCarthy, Z., Gaswick, N., McKain, B. and Speer, K. *Laplace transforms for the nabla-difference operator*, Pan Am. Math. J., 21(3) (2011), 79–96.
- [13] Hioual, A., Ouannas, A., Grassi, G. and Oussaeif, T.E. *Nonlinear nabla variable-order fractional discrete systems: Asymptotic stability and application to neural networks*, J. Comput. Appl. Math., 423 (2023), 114939.
- [14] Hioual, A., Ouannas, A., Momani, S. and Oussaeif, T.E. *Finite-Time Stability of ABC Type h -Fractional Discrete Neural Networks: Gronwall Inequality and Stability Criterion*, In *2023 International Conference on Fractional Differentiation and Its Applications (ICFDA)*, pp. 1–6. IEEE, 2023.
- [15] Hioual, A., Ouannas, A. and Oussaeif, T.E. *Nonlinear Fractional Discrete Neural Networks: Stability, Stabilization and Synchronization*, In *State Estimation and Stabilization of Nonlinear Systems: Theory and Applications*, pp. 169–204. Springer Nature Switzerland, Cham, 2023.
- [16] Hioual, A., Ouannas, A., Oussaeif, T.E., Grassi, G., Batiha, I.M. and Momani, S. *On variable-order fractional discrete neural networks: solvability and stability*, Fractal Fract., 6(2) (2022), 119.
- [17] Hioual, A., Oussaeif, T.E., Ouannas, A., Grassi, G., Batiha, I.M. and Momani, S. *New results for the stability of fractional-order discrete-time neural networks*, Alex. Eng. J., 61(12) (2022), 10359–10369.
- [18] Jia, J., Wang, F. and Zeng, Z. *Lag quasi-synchronization of incommensurate fractional-order memristor-based neural networks with nonidentical characteristics via quantized control: A vector fractional Halanay inequality approach*, J. Franklin Inst., 359(12) (2022), 6392–6437.

- [19] Khennaoui, A.A., Ouannas, A., Bendoukha, S., Grassi, G., Lozi, R.P. and Pham, V.T. *On fractional-order discrete-time systems: Chaos, stabilization and synchronization*, Chaos Solitons Fractals, 119 (2019), 150–162.
- [20] Kuang, Z., Gao, H. and Tomizuka, M. *Precise linear-motor synchronization control via cross-coupled second-order discrete-time fractional-order sliding mode*, IEEE/ASME Trans. Mechatronics, 26(1) (2020), 358–368.
- [21] Li, X., Fang, J.A., Zhang, W. and Li, H. *Finite-time synchronization of fractional-order memristive recurrent neural networks with discontinuous activation functions*, Neurocomputing, 316 (2018), 284–293.
- [22] Liu, X. and Yu, Y. *Synchronization analysis for discrete fractional-order complex-valued neural networks with time delays*, Neural Comput. Appl., 33 (2021), 10503–10514.
- [23] Megherbi, O., Hamiche, H., Djennoune, S. and Bettayeb, M. *A new contribution for the impulsive synchronization of fractional-order discrete-time chaotic systems*, Nonlinear Dyn., 90 (2017), 1519–1533.
- [24] Momani, S., Batiha, I.M., Hioual, A. and Ouannas, A. *Fractional Neural Networks: Finite time stability and its application to synchronization*, In *2023 International Conference on Fractional Differentiation and Its Applications (ICFDA)*, pp. 1–5. IEEE, 2023.
- [25] Ouannas, A., Azar, A.T. and Abu-Saris, R. *A new type of hybrid synchronization between arbitrary hyperchaotic maps*, Int. J. Mach. Learn. Cybern., 8(6) (2017), 1887–1894.
- [26] Ouannas, A., Azar, A.T. and Vaidyanathan, S. *A robust method for new fractional hybrid chaos synchronization*, Math. Methods Appl. Sci., 40(5) (2017), 1804–1812.
- [27] Ouannas, A., Azar, A.T. and Vaidyanathan, S. *New hybrid synchronisation schemes based on coexistence of various types of synchronisation between master-slave hyperchaotic systems*, Int. J. Comput. Appl. Technol., 55(2) (2017), 112–120.
- [28] Ouannas, A., Azar, A.T. and Ziar, T. *On inverse full state hybrid function projective synchronization for continuous-time chaotic dynamical systems with arbitrary dimensions*, Differ. Equ. Dyn. Syst., 28(4) (2020), 1045–1058.
- [29] Ouannas, A., Azar, A.T., Ziar, T. and Radwan, A.G. *Generalized synchronization of different dimensional integer-order and fractional order chaotic systems*, In *Fractional Order Control and Synchronization of Chaotic Systems*, pp. 671–697. Springer, Cham, 2017.

- [30] Ouannas, A., Karouma, A., Grassi, G., Pham, V.T. and Luong, V.S. *A novel secure communications scheme based on chaotic modulation, recursive encryption and chaotic masking*, Alex. Eng. J., 60(1) (2021), 1873–1884.
- [31] Ouannas, A., Khennaoui, A.A., Momani, S., Grassi, G. and Pham, V.T. *Chaos and control of a three-dimensional fractional order discrete-time system with no equilibrium and its synchronization*, AIP Adv., 10(4) (2020).
- [32] Ouannas, A. and Odibat, Z. *Generalized synchronization of different dimensional chaotic dynamical systems in discrete time*, Nonlinear Dyn., 81(1) (2015), 765–771.
- [33] Si, L., Xiao, M., Jiang, G., Cheng, Z., Song, Q. and Cao, J. *Dynamics of fractional-order neural networks with discrete and distributed delays*, IEEE Access, 8 (2019), 46071–46080.
- [34] Wang, M., Wang, Y. and Chu, R. *Dynamical Analysis of the Incommensurate Fractional-Order Hopfield Neural Network System and Its Digital Circuit Realization*, Fractal Fract., 7(6) (2023), 474.
- [35] Wang, S. and Jian, J. *Predefined-time synchronization of incommensurate fractional-order competitive neural networks with time-varying delays*, Chaos Solitons Fractals, 177 (2023), 114216.
- [36] Xin, B. and Wu, Z. *Projective synchronization of chaotic discrete dynamical systems via linear state error feedback control*, Entropy, 17(5) (2015), 2677–2687.
- [37] You, X., Song, Q. and Zhao, Z. *Global Mittag-Leffler stability and synchronization of discrete-time fractional-order complex-valued neural networks with time delay*, Neural Netw., 122 (2020), 382–394.
- [38] Zhao, M., Li, H.L., Zhang, L., Hu, C. and Jiang, H. *Quasi-projective synchronization of discrete-time fractional-order quaternion-valued neural networks*, J. Franklin Inst., 360(4) (2023), 3263–3279.



Mathematical analysis of a fractional-order influenza model with resistant and nonresistant strains

I. Aamoum, A. Meskaf, Z. Yaagoub*,^{id} and E.M. Farah

Abstract

There are numerous acute viral infections worldwide. Influenza is a respiratory illness caused by the influenza virus. In the twenty-first century, fractional-order derivative modeling of

*Corresponding author

Received 2 December 2025; revised 9 February 2026; accepted 12 February 2026

Ismail Aamoum

Experimental Laboratory of Innovation in Technology and Simulations, Chouaib Doukkali University, El Jadida, Morocco. e-mail: aamoumismail@gmail.com

Adil Meskaf

Experimental Laboratory of Innovation in Technology and Simulations, Chouaib Doukkali University, El Jadida, Morocco. e-mail: adilmeskaf@gmail.com

Department of Economics and Management (SEG), Faculty of Legal, Economic and Social Sciences, Chouaib Doukkali University, El Jadida, Morocco.

Zakaria Yaagoub

National School of Public Health, Ministry of Health and Social Protection, Rabat, Morocco. e-mail: zakaria.yaagoub@gmail.com

Department of Computer Science, Faculty of Sciences, Mohammed V University in Rabat, Morocco.

El Mehdi Farah

Laboratory of Advanced Research in Industrial and Logistics Engineering (LARILE), National Higher School of Electricity and Mechanics (ENSEM), Hassan II University of Casablanca, Morocco. e-mail: elme-hdi.farah.ed@gmail.com

How to cite this article

Aamoum, I., Meskaf, A., Yaagoub, Z. and Farah, E.M., Mathematical analysis of a fractional-order influenza model with resistant and nonresistant strains. *Iran. J. Numer. Anal. Optim.*, 2026; 16(2): 703–726. <https://doi.org/10.22067/ijnao.2026.96816.1791>

infectious diseases has gained significant popularity. In the literature, several fractional-order derivative models have been developed. In this research, we propose a fractional two-strain influenza epidemic model that incorporates both resistant and nonresistant strains. We prove the existence and uniqueness of a global positive solution. The equilibrium points and the basic reproduction number are derived. We study the global stability of the equilibrium points using Lyapunov functions and LaSalle's invariance principle. Finally, a series of numerical simulations is presented to illustrate the behavior of the model and to verify our theoretical results. The results show that although the order of the fractional derivatives does not affect the stability of the equilibrium states, it plays an important role in the speed at which the system converges to these equilibria.

AMS subject classifications (2020): Primary 34A08, 92D30; Secondary 34D20, 37N25, 26A33.

Keywords: Fractional order; Influenza; Basic reproduction number; Global stability; Two-strain.

1 Introduction

Influenza is a highly contagious viral infection that remains a significant public health concern worldwide. Caused primarily by influenza A and B viruses, it leads to seasonal epidemics and occasional pandemics that can have severe health and economic consequences [26]. According to the World Health Organization (WHO), annual influenza epidemics are responsible for approximately 3 to 5 million cases of severe illness and 290,000 to 650,000 respiratory-related deaths globally each year [31]. Influenza A viruses are further subdivided into subtypes based on the hemagglutinin (H) and neuraminidase (N) surface proteins, while antigenic drift and shift contribute to the virus's genetic variability and epidemiological complexity [30]. A critical challenge in influenza control is the emergence of multiple co-circulating strains, including those that develop resistance to antiviral medications such as neuraminidase inhibitors [13]. These resistant strains can compromise treatment efficacy and alter transmission dynamics, necessitating mathematical tools capable of capturing such complexities. Modeling the interactions between drug-resistant and drug-sensitive strains is essential for assessing intervention strategies, guiding vaccine formulation, and understanding long-term evolutionary pressures on the virus [7].

Classical compartmental models, such as the SIR (Susceptible–Infectious–Recovered) and SEIR (Susceptible–Exposed–Infectious–Recovered) frameworks, have been instrumental in studying infectious diseases, including influenza [17, 19, 6, 14, 20, 4, 5, 16, 35]. These models describe population dynamics through systems of ordinary differential equations (ODEs), assuming that

the rate of change in each compartment depends only on the current state and not on past states. However, real-world epidemiological processes often exhibit memory and hereditary properties. For example, immunity to influenza wanes gradually, contact rates may vary based on past exposure or social behavior, and virus–host interactions can show time-dependent adaptation [34]. To address these limitations, researchers have increasingly turned to fractional-order differential equations (FDEs), a generalization of classical calculus that allows for derivatives of noninteger order [25, 10]. The Caputo fractional derivative, in particular, enables the incorporation of initial conditions in a way that is consistent with physical and biological systems [18, 9, 32, 24]. Applications of fractional calculus in infectious disease modeling have demonstrated improved fitting to epidemic data compared to integer-order models. For example, sub-exponential growth patterns observed during the early phases of outbreaks of diseases such as COVID-19 and dengue fever have been effectively captured using fractional models [1, 28, 2, 15, 21, 3, 22]. In the context of influenza, incorporating fractional dynamics can help model the gradual loss of immunity, delays in antiviral response effectiveness, and heterogeneous contact structures, especially in the presence of multiple strains with varying transmission and resistance characteristics [23, 27, 12, 36]. In their study [8], Baba et al. developed and analyzed a deterministic two-strain influenza model with resistant and nonresistant strains. These strains are differentiated by their incidence rates, which are bilinear and saturated for the nonresistant strain, and saturated for the resistant strain. This work extends the previous model of Baba et al. by proposing and analyzing a new fractional-derivative model that incorporates Caputo derivatives to reflect the influence of past infection and intervention history on current disease progression.

This paper is organized into eight sections. In Section 2, we formulate a fractional two-strain influenza model incorporating two distinct incidence rates. In Section 3, we introduce some mathematical preliminaries that assist in the analysis of our model. Section 4 deals with the positivity, boundedness, and uniqueness of the solutions. Section 5 is devoted to the determination of the basic reproduction number, the identification of the equilibrium points, and their global stability analysis. Numerical simulations are presented in Section 6. Finally, Section 7 concludes the study.

2 Model formulation

The model represents the transmission dynamics of a disease involving two viral strains: a non-resistant strain, denoted by $I_N(t)$, and a resistant strain, denoted by $I_R(t)$. The system is formulated using Caputo fractional derivatives of order θ , allowing the incorporation of memory effects into the disease progression and treatment processes. The population is divided into

four compartments: susceptible individuals $S(t)$, individuals infected by the nonresistant strain $I_N(t)$, individuals infected by the resistant strain $I_R(t)$, and recovered individuals $R(t)$. Susceptible individuals are recruited at a constant rate b and may become infected through contact with either $I_N(t)$ or $I_R(t)$. Transmission by the nonresistant strain follows a standard bilinear form $aS(t)I_N(t)$, whereas transmission by the resistant strain is modeled by a saturated incidence function $\frac{\gamma S(t)I_R(t)}{1 + kI_R(t)}$, where the parameter k represents a saturation effect that limits the effective transmission of the resistant strain as the number of infected individuals increases. This saturation may arise from behavioral changes, partial immunity, limited contact opportunities, or treatment-related constraints. Infectious individuals in $I_N(t)$ may recover at rate θ , die naturally, or move to the recovered class. In contrast, individuals in $I_R(t)$ represent infections caused by a strain that resists treatment; hence, the recovery rate γ may be lower. The saturated transmission term for $I_R(t)$ also indicates that as the number of resistant infections increases, the effective transmission rate declines, which can simulate immunological or medical saturation effects. Recovered individuals gain immunity through recovery from either strain and exit the system through natural death. The fractional-order derivatives account for nonlocal, history-dependent effects, which are biologically relevant for diseases in which immune response, treatment impact, or behavioral patterns are influenced by past states. This model is particularly suited to studying the coexistence and competition between resistant and nonresistant strains, and how memory effects influence the speed and stability of convergence toward the disease-free or endemic equilibrium. Based on [8], the mathematical representation of these interactions is given by the following system of fractional differential equations

$$\begin{cases} D^\theta S(t) = b - dS(t) - aS(t)I_N(t) - \frac{\gamma S(t)I_R(t)}{1 + kI_R(t)}, \\ D^\theta I_N(t) = aS(t)I_N(t) - (d + \theta)I_N(t), \\ D^\theta I_R(t) = \frac{\gamma S(t)I_R(t)}{1 + kI_R(t)} - (d + \gamma)I_R(t), \\ D^\theta R(t) = \theta I_N(t) + \gamma I_R(t) - dR(t). \end{cases} \quad (1)$$

Where D^θ denotes the Caputo fractional derivative of order θ , with $0 < \theta \leq 1$. All parameters are described in Table 1. The two-strain diagram is shown in Figure 1. The initial conditions $S(0)$, $I_N(0)$, $I_R(0)$, and $R(0)$ of system (1) are assumed to be positive. In our model, we assume that individuals become susceptible only through birth or immigration, co-infection does not occur, and immunity is present.

Table 1: An explanation of the parameters of the model (1).

Parameter	Description
b	Recruitment into the susceptible class
d	Death rate
γ	Rate of infection by the resistant strain
a	Rate of infection by the nonresistant strain
γ	Rate at which individuals with the resistant strain are removed from the population
θ	Rate at which individuals with the nonresistant strain are removed from the population
k	Saturation parameter in the resistant-strain incidence rate

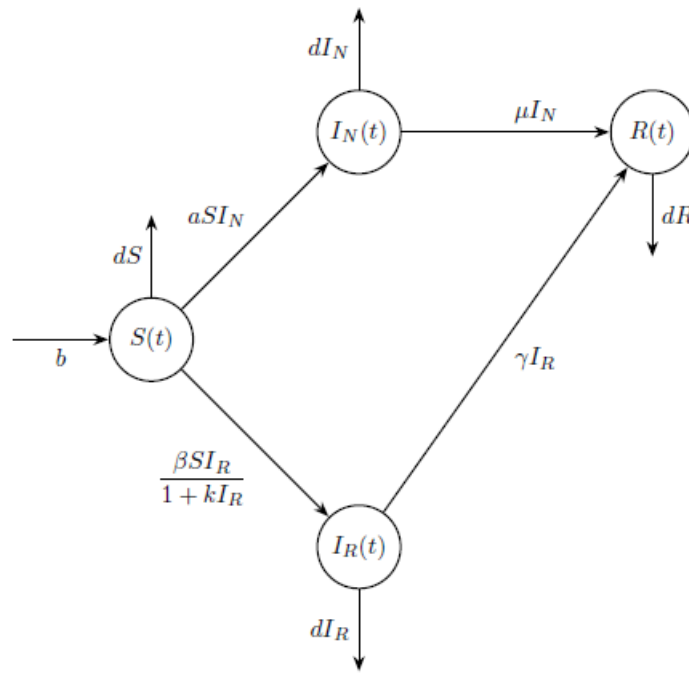


Figure 1: Compartmental diagram of the fractional-order model representing the transmission of nonresistant and resistant strains.

3 Preliminaries

We present in this section some theoretical results that will help us examine system (1) later in this study.

Definition 1. Consider the fractional integral of order $\theta > 0$ of a function $H : \mathbb{R}_+ \rightarrow \mathbb{R}$. It is defined as

$$I^\theta H(t) = \frac{1}{\Gamma(\theta)} \int_0^t (t-x)^{\theta-1} H(x) dx, \quad (2)$$

where the Gamma function $\Gamma(\theta)$ is defined by

$$\Gamma(\theta) = \int_0^\infty e^{-t} t^{\theta-1} dt.$$

Definition 2. For a function $H : \mathbb{R}_+ \rightarrow \mathbb{R}$, the Caputo fractional derivative of order $\theta > 0$ is defined by

$$D^\theta H(t) \equiv I^{n-\theta} D^n H(t), \quad (3)$$

where $D = \frac{d}{dt}$, n is the smallest integer greater than θ (i.e., $n-1 < \theta \leq n$), and $n \in \mathbb{N}$. In particular, when $0 < \theta \leq 1$, the fractional derivative can be expressed as

$$D^\theta H(t) = \frac{1}{\Gamma(1-\theta)} \int_0^t \frac{H'(x)}{(t-x)^\theta} dx. \quad (4)$$

Definition 3. The Mittag-Leffler function with parameter $\theta > 0$, denoted by E_θ , is defined as

$$E_\theta(t) = \sum_{i=0}^{\infty} \frac{t^i}{\Gamma(\theta i + 1)}. \quad (5)$$

Let $U : \mathbb{R}_+^n \rightarrow \mathbb{R}_+^n$, with $n \geq 1$. We consider the fractional-order system described by

$$D^\theta U(t) = G(U(t)), \quad (6)$$

with the initial condition $U(0) = U_0$, where $U_0 \in \mathbb{R}_+^n$ and $0 < \theta \leq 1$.

Lemma 1 ([8]). Let D be a compact subset of $\mathbb{R}_+^n$. Every solution of system (6) that starts in D remains in D for all $t \geq 0$, provided that there exists a continuously differentiable function $V(U) \in C^1(D)$ satisfying

$$D^\theta V(U) \leq 0. \quad (7)$$

Define E as the set of points in D where $D^\theta V = 0$, and let Ψ denote the largest invariant set contained in E . Then, every solution $U(t)$ starting in D asymptotically approaches Ψ as $t \rightarrow \infty$.

Lemma 2 ([8]). Suppose that $h(t) \in C[a, b]$ and $D^\theta h(t) \in C[a, b]$ for $0 < \theta \leq 1$. Then we have

$$h(t) = h(a) + \frac{1}{\Gamma(\theta)} D^\theta h(\nu)(t-a)^\theta, \quad a < \nu < t, \quad \forall t \in (a, b]. \quad (8)$$

Lemma 3 ([8]). Suppose that $h(t) \in C[a, b]$ and $D^\theta h(t) \in C[a, b]$ for $0 < \theta \leq 1$. If $D^\theta h(t) \geq 0$ for all $t \in [a, b]$, then $h(t)$ is nondecreasing on $[a, b]$. If $D^\theta h(t) \leq 0$ for all $t \in [a, b]$, then $h(t)$ is nonincreasing on $[a, b]$.

Lemma 4 ([8]). Let $0 < \theta \leq 1$ and let $x(t)$ be a positive and sufficiently smooth function. For any constant $x^* > 0$, define

$$\psi(x) = x - x^* - x^* \ln\left(\frac{x}{x^*}\right).$$

Then, along any positive solution $x(t) > 0$, the fractional derivative satisfies

$$D^\theta \psi(x(t)) \leq \left(1 - \frac{x^*}{x(t)}\right) D^\theta x(t).$$

Consequently, for $X(t) = (x_1(t), \dots, x_n(t))$ and

$$V(X) = \sum_{i=1}^n \left(x_i - x_i^* - x_i^* \ln\left(\frac{x_i}{x_i^*}\right) \right),$$

we have

$$D^\theta V(X(t)) \leq \sum_{i=1}^n \left(1 - \frac{x_i^*}{x_i(t)} \right) D^\theta x_i(t).$$

Here, D^θ denotes the Caputo fractional derivative.

4 Positivity, boundedness, and uniqueness of the solution

In this section, we show that, for any given initial data, system (1) admits a unique, positive, and bounded solution.

Theorem 1. For any positive initial conditions $S(0)$, $I_N(0)$, $I_R(0)$, and $R(0)$, system (1) admits a unique, positive, and bounded solution defined in $\mathbb{R}_+^4$ for all $t \geq 0$.

Proof. First, we demonstrate that the solutions of system (1) remain positive. Assuming that all model parameters are positive, we have

$$\begin{cases} D^\theta S|_{S=0} = b \geq 0, \\ D^\theta I_N|_{I_N=0} = 0 \geq 0, \\ D^\theta I_R|_{I_R=0} = 0 \geq 0, \\ D^\theta R|_{R=0} = \theta I_N + \gamma I_R \geq 0. \end{cases} \quad (9)$$

From Lemmas 2 and 3, it follows that the solutions of system (1) remain positive. We now show that the solutions of system (1) remain bounded by considering the total population given by

$$N(t) = S(t) + I_N(t) + I_R(t) + R(t). \quad (10)$$

Then,

$$D^\theta N(t) \leq b - dN(t), \quad (11)$$

Consequently, by applying the Laplace transform, we conclude that

$$N(t) \leq \left(-\frac{b}{d} + N(0)\right)E_\theta(-dt^\theta) + \frac{b}{d}. \quad (12)$$

where E_θ is the Mittag-Leffler function with parameter θ . Consequently, we conclude that the set

$$\mathcal{T} := \left\{ (S, I_N, I_R, R) \in \mathbb{R}_+^4 \mid S + I_N + I_R + R \leq \frac{b}{d} \right\}$$

is positively invariant for system (1). Hence, every solution of system (1) remains positive and bounded. $\square$

Theorem 2. For every initial condition $U(0) = U_0 \in \mathcal{T}$, system (1) admits a unique solution $U(t)$ defined for all $t \geq 0$.

Proof. Let $U(t) = (S(t), I_N(t), I_R(t), R(t))^\top$ and rewrite system (1) as

$${}^C D^\theta U(t) = G(U(t)), \quad 0 < \theta \leq 1, \quad (13)$$

where

$$G(U) = \begin{pmatrix} b - dS - aSI_N - \frac{\gamma SI_R}{1 + kI_R} \\ aSI_N - (d + \theta)I_N \\ \frac{\gamma SI_R}{1 + kI_R} - (d + \gamma)I_R \\ \theta I_N + \gamma I_R - dR \end{pmatrix}. \quad (14)$$

The mapping $G : \mathbb{R}_+^4 \rightarrow \mathbb{R}^4$ is continuously differentiable with respect to U on $\mathbb{R}_+^4$. Hence, G is locally Lipschitz on $\mathbb{R}_+^4$. Therefore, for any $U_0 \in \mathbb{R}_+^4$, there exists a unique local solution on $[0, T_{\max})$ [10]. Moreover, by Theorem 1, the set $\mathcal{T}$ is positively invariant, and all solutions starting in $\mathcal{T}$ remain bounded in $\mathcal{T}$ for as long as they exist. Assume, by contradiction, that the maximal existence time satisfies $T_{\max} < +\infty$. By the standard continuation criterion for Caputo fractional differential equations, a maximal solution can cease to exist in finite time only if it

becomes unbounded as $t \rightarrow T_{\max}^-$. This contradicts the boundedness of the solution in $\mathcal{T}$. Hence, $T_{\max} = +\infty$, and the solution exists globally and is unique. $\square$

5 Global stability of equilibria

In this section, we show that system (1) admits a disease-free equilibrium, a nonresistant equilibrium, a resistant equilibrium, and an endemic equilibrium. Moreover, we employ appropriate Lyapunov functions and LaSalle's invariance principle to investigate the global stability of these equilibrium points.

5.1 Basic reproduction number

Following the epidemiological literature [29], the basic reproduction number R_0 is defined as the expected number of secondary infections generated by a single infected individual introduced into a wholly susceptible population. We compute R_0 for system (1) using the next-generation matrix approach [33]. At the disease-free equilibrium,

$$E_0 = \left(\frac{b}{d}, 0, 0, 0 \right),$$

the matrices of new infection and transition terms are given by

$$F = \begin{pmatrix} \frac{ab}{d} & 0 \\ 0 & \frac{\gamma b}{d} \end{pmatrix}, \quad V = \begin{pmatrix} d + \theta & 0 \\ 0 & d + \gamma \end{pmatrix}.$$

The inverse of V is

$$V^{-1} = \begin{pmatrix} \frac{1}{d + \theta} & 0 \\ 0 & \frac{1}{d + \gamma} \end{pmatrix},$$

hence,

$$FV^{-1} = \begin{pmatrix} \frac{ab}{d(d + \theta)} & 0 \\ 0 & \frac{\gamma b}{d(d + \gamma)} \end{pmatrix}.$$

Therefore, the basic reproduction number is given by

$$R_0 = \rho(FV^{-1}) = \max\{R_N, R_R\},$$

where

$$R_N = \frac{ab}{d(d+\theta)}, \quad R_R = \frac{\gamma b}{d(d+\gamma)}.$$

5.2 Equilibrium points

In this subsection, we characterize the equilibrium structure of the proposed fractional-order epidemic model given in system (1). The system admits four steady states: the disease-free equilibrium E_0 , the nonresistant equilibrium E_1 , the resistant equilibrium E_2 , and the endemic (coexistence) equilibrium E_3 . We obtain the following equations by setting the right-hand side of system (1) equal to zero

$$\begin{cases} b - dS - aSI_N - \frac{\gamma SI_R}{1 + kI_R} = 0, \\ aSI_N - (d + \theta)I_N = 0, \\ \frac{\gamma SI_R}{1 + kI_R} - (d + \gamma)I_R = 0, \\ \theta I_N + \gamma I_R - dR = 0. \end{cases} \quad (15)$$

1. Disease-free equilibrium point

The disease-free equilibrium point denotes the state in which the population carries no infection and is given by

$$E_0 = \left(\frac{b}{d}, 0, 0, 0 \right). \quad (16)$$

For any set of positive parameters, this equilibrium point exists.

2. Nonresistant equilibrium point

The nonresistant equilibrium point represents a steady state where $I_N^* > 0$ and $I_R^* = 0$.

It is given by

$$E_1 = \left(\frac{d + \theta}{a}, \frac{ab - d(d + \theta)}{a(d + \theta)}, 0, \frac{\theta}{d} I_N^* \right). \quad (17)$$

The nonresistant equilibrium point E_1 exists if and only if $R_N = \frac{ab}{d(d + \theta)} > 1$.

3. Resistant equilibrium point

The resistant equilibrium point represents a steady state where $I_R^* > 0$, $I_N^* = 0$. It is represented by

$$E_2 = \left(\frac{d+\gamma}{\gamma} (1 + k I_R^*), 0, \frac{b\gamma - d(d+\gamma)}{(d+\gamma)(\gamma + dk)}, \frac{\gamma}{d} I_R^* \right). \quad (18)$$

The resistant equilibrium point E_2 exists if only if $R_R = \frac{b\gamma}{d(d+\gamma)} > 1$.

4. Endemic equilibrium point

The endemic equilibrium point E_3 corresponds to the case where $I_N^* > 0$ and $I_R^* > 0$, and it is given by

$$E_3 = \left(\frac{d+\theta}{a}, \frac{b + \frac{d+\gamma}{k} - \frac{d+\theta}{a} \left(d + \frac{\gamma}{k} \right)}{d+\theta}, \frac{1}{k} \left(\frac{\gamma(d+\theta)}{a(d+\gamma)} - 1 \right), \frac{\theta I_N^* + \gamma I_R^*}{d} \right). \quad (19)$$

The endemic equilibrium point E_3 exists if and only if

$$\frac{\gamma(d+\theta)}{a(d+\gamma)} > 1 \quad \text{and} \quad b + \frac{d+\gamma}{k} > \frac{d+\theta}{a} \left(d + \frac{\gamma}{k} \right). \quad (20)$$

These conditions are necessary and sufficient to ensure $S^* > 0$, $I_N^* > 0$ and $I_R^* > 0$. In particular, these inequalities imply that $R_N > 1$ and $R_R > 1$. Hence, the condition $R_R \geq R_N > 1$ should be interpreted as a sufficient but not equivalent condition for the existence of the coexistence equilibrium point.

5.3 Stability analysis

This subsection presents theorems on the global stability of the equilibrium points, derived via the Lyapunov method and LaSalle's invariance principle.

Theorem 3. E_0 is globally asymptotically stable if $\max\{R_N, R_R\} \leq 1$.

Proof. We introduce the following Lyapunov function

$$V_0 = (S - S_0 \ln S) + I_R + I_N. \quad (21)$$

Using Lemma 4, the θ derivative of V_0 satisfies

$$D^\theta V_0 \leq \left(1 - \frac{S_0}{S}\right) D^\theta S + D^\theta I_R + D^\theta I_N \quad (22)$$

$$\begin{aligned} D^\theta V_0 \leq \left(1 - \frac{S_0}{S}\right) & \left(b - dS - aSI_N - \frac{\gamma SI_R}{1 + kI_R}\right) \\ & + aSI_N - (d + \theta)I_N + \frac{\gamma SI_R}{1 + kI_R} - (d + \gamma)I_R \end{aligned} \quad (23)$$

$$\begin{aligned} D^\theta V_0 \leq dS_0 - dS - \frac{dS_0^2}{S} + (d + \theta)I_N & \left(\frac{aS_0}{d + \theta} - 1\right) \\ & + (d + \gamma)I_R \left(\frac{\gamma S_0}{(d + \gamma)(1 + kI_R)} - 1\right) \end{aligned} \quad (24)$$

$$\begin{aligned} D^\theta V_0 \leq dS_0 \left(2 - \frac{S}{S_0} - \frac{S_0}{S}\right) & + (d + \theta)I_N(R_N - 1) \\ & + (d + \gamma)I_R(R_R - 1). \end{aligned} \quad (25)$$

The arithmetic mean is greater than or equal to the geometric mean, we Have

$$2 - \frac{S}{S_0} - \frac{S_0}{S} \leq 0. \quad (26)$$

Then, $D^\theta V_0 \leq 0$ if $R_R \leq 1$ and $R_N \leq 1$. We denote A_0 The largest invariant set in (S, I_N, I_R) ; $D^\theta V_0 = 0$. Thus, $D^\theta V_0 = 0$ if and only if $S = S_0 = \frac{b}{d}$ and $I_N = I_R = 0$. Hence, the disease-free equilibrium is

$$A_0 = E_0 = \left(\frac{b}{d}, 0, 0, 0\right).$$

Since LaSalle's invariance principle that the E_0 is globally asymptotically stable if $R_R \leq 1$ and $R_N \leq 1$. $\square$

Theorem 4. Assume $R_N > 1$ and $R_R \leq 1$. Then the equilibrium E_1 of (1) is globally asymptotically stable.

Proof. Consider the Lyapunov function

$$V_1(S, I_N, I_R) = S - S_1 - S_1 \ln\left(\frac{S}{S_1}\right) + I_N - I_{N_1} - I_{N_1} \ln\left(\frac{I_N}{I_{N_1}}\right) + I_R. \quad (27)$$

Using Lemma 4, the θ derivative of V_1 satisfies

$$D^\theta V_1 \leq \left(1 - \frac{S_1}{S}\right) D_t^\theta S + \left(1 - \frac{I_{N_1}}{I_N}\right) D_t^\theta I_N + D_t^\theta I_R. \quad (28)$$

It is clear that,

$$aS_1 = d + \theta, \quad b - dS_1 - aS_1I_{N_1} = 0, \quad (29)$$

then

$$D^\theta V_1 \leq -d \frac{(S - S_1)^2}{S} - (d + \theta) \frac{(I_N - I_{N_1})^2}{I_N} + \left(\frac{\gamma S_1}{1 + kI_R} - (d + \gamma) \right) I_R. \quad (30)$$

Since $k > 0$ and $I_R \geq 0$, one has $1 + kI_R \geq 1$. Hence,

$$\frac{\gamma S_1}{1 + kI_R} - (d + \gamma) \leq \gamma S_1 - (d + \gamma) = (d + \gamma)(R_R - 1) \leq 0.$$

Therefore, the last term in (30) is nonpositive, and we conclude that

$$D^\theta V_1 \leq -d \frac{(S - S_1)^2}{S} - (d + \theta) \frac{(I_N - I_{N_1})^2}{I_N} + (d + \gamma)(R_R - 1) \leq 0, \quad (31)$$

with equality if and only if $S = S_1$, $I_N = I_{N_1}$, and $I_R = 0$.

By LaSalle's invariance principle, every solution with initial data in $\mathcal{T}$ converges to the largest invariant subset of $\{D^\theta V_1 = 0\}$, which is $\{(S_1, I_{N_1}, 0)\}$. Finally, since $N = S + I_N + I_R + R$ satisfies $D^\theta N = b - dN$, we have $N(t) \rightarrow S_1$, and thus

$$R(t) = N(t) - S(t) - I_N(t) - I_R(t) \rightarrow R_1.$$

Hence, the full state converges to E_1 , proving global asymptotic stability. $\square$

Theorem 5. Assume $R_R > 1$ and $R_N \leq 1$. Then the equilibrium E_2 of (1) is globally asymptotically stable.

Proof. Consider the Lyapunov function

$$V_2(S, I_N, I_R) = S - S_2 - S_2 \ln\left(\frac{S}{S_2}\right) + I_R - I_{R_2} - I_{R_2} \ln\left(\frac{I_R}{I_{R_2}}\right) + I_N. \quad (32)$$

Using Lemma 4, the θ derivative of V_2 satisfies

$$D^\theta V_2 \leq \left(1 - \frac{S_2}{S}\right) D^\theta S + D^\theta I_N + \left(1 - \frac{I_{R_2}}{I_R}\right) D^\theta I_R. \quad (33)$$

We have

$$\frac{\gamma S_2}{1 + kI_{R_2}} = d + \gamma, \quad b - dS_2 - \frac{\gamma S_2 I_{R_2}}{1 + kI_{R_2}} = 0, \quad (34)$$

So,

$$D^\theta V_2 \leq -d \frac{(S - S_2)^2}{S} - (d + \gamma) \frac{(I_R - I_{R_2})^2}{I_R} + (aS_2 - (d + \theta)) I_N. \quad (35)$$

We have $aS_2 - (d + \theta) \leq 0$, hence the last term is also nonpositive. Therefore $D^\theta V_2 \leq 0$, with equality if and only if $S = S_2$, $I_R = I_{R_2}$, and $I_N = 0$. By LaSalle's invariance principle, every solution starting in Ω converges to the largest invariant subset of $\{D^\theta V_2 = 0\}$, which reduces to $\{(S_2, 0, I_{R_2})\}$. Finally, since $N = S + I_N + I_R + R$ satisfies $D^\theta N = b - dN$, we have $N(t) \rightarrow S_2$, so $R(t) = N(t) - S(t) - I_N(t) - I_R(t) \rightarrow R_2$. Hence the full state converges to E_2 , which proves global asymptotic stability. $\square$

Theorem 6. Assume that the endemic equilibrium E_3 exists, i.e.,

$$\frac{\gamma(d + \theta)}{a(d + \gamma)} > 1 \quad \text{and} \quad b + \frac{d + \gamma}{k} > \frac{d + \theta}{a} \left(d + \frac{\gamma}{k}\right).$$

Then the endemic equilibrium E_3 is globally asymptotically stable.

Proof. Consider the Lyapunov function

$$\begin{aligned} V_3(S, I_R, I_N) = & (S - S_3 - S_3 \ln \frac{S}{S_3}) + (I_R - I_{R_3} - I_{R_3} \ln \frac{I_R}{I_{R_3}}) \\ & + (I_N - I_{N_3} - I_{N_3} \ln \frac{I_N}{I_{N_3}}), \end{aligned} \quad (36)$$

Using Lemma 4, the θ derivative of V_3 satisfies

$$D^\theta V_3 \leq \left(1 - \frac{S_3}{S}\right) D^\theta S + \left(1 - \frac{I_{R_3}}{I_R}\right) D^\theta I_R + \left(1 - \frac{I_{N_3}}{I_N}\right) D^\theta I_N. \quad (37)$$

Substituting the system gives

$$\begin{aligned} D^\theta V_3 \leq & \left(1 - \frac{S_3}{S}\right) \left(b - aSI_N - \frac{\gamma SI_R}{1 + kI_R} - dS\right) \\ & + \left(1 - \frac{I_{R_3}}{I_R}\right) \left(\frac{\gamma SI_R}{1 + kI_R} - (d + \gamma)I_R\right) \\ & + \left(1 - \frac{I_{N_3}}{I_N}\right) (aSI_N - (d + \theta)I_N). \end{aligned} \quad (38)$$

At the endemic equilibrium E_3 we have

$$aS_3 = d + \theta, \quad \frac{\gamma S_3}{1 + kI_{R_3}} = d + \gamma, \quad b = aS_3I_{N_3} + \frac{\gamma S_3I_{R_3}}{1 + kI_{R_3}} + dS_3.$$

We obtain

$$D^\theta V_3 \leq -d \frac{(S - S_3)^2}{S} - (d + \gamma) \frac{(I_R - I_{R_3})^2}{I_R} - (d + \theta) \frac{(I_N - I_{N_3})^2}{I_N} \leq 0. \quad (39)$$

So, $D^\theta V_3 = 0$ holds only at $S = S_3$, $I_R = I_{R_3}$ and $I_N = I_{N_3}$. Therefore, by LaSalle's invariance principle for fractional-order systems, E_3 is globally asymptotically stable whenever it exists, i.e. when $R_R \geq R_N > 1$. $\square$

6 Numerical simulations and discussion

In this section, we give some numerical simulations to illustrate our analytical results. For this end, we apply the numerical algorithm presented by Erturk et al. [11] to numerically solve the fractional-order model (1). Let us consider the following problem

$$\begin{cases} D^\theta U(t) = G(U(t)), \\ U(0) = U_0, \quad U_0 \in \mathbb{R}_+^4, \end{cases} \quad (40)$$

where

$$U(t) = (S(t), I_N(t), I_R(t), R(t))^\top \quad \text{and} \quad U_0 = (S(0), I_N(0), I_R(0), R(0))^\top$$

is the initial condition. The problem above is solved using the following numerical scheme

$$\begin{aligned} U(t_j) = & \frac{h^\theta}{\Gamma(\theta+2)} ((j-1)^{\theta+1} - (j-\theta-1)^{\theta+1}) G(U_0) \\ & + \frac{h^\theta}{\Gamma(\theta+2)} \sum_{i=1}^{j-1} [(j-i+1)^{\theta+1} - 2(j-i)^{\theta+1} + (j-i-1)^{\theta+1}] G(U(t_i)) \\ & + \frac{h^\theta}{\Gamma(\theta+2)} G(U(t_{j-1})) + \frac{h^\theta}{\Gamma(\theta+1)} G(U(t_{j-1})), \end{aligned} \quad (41)$$

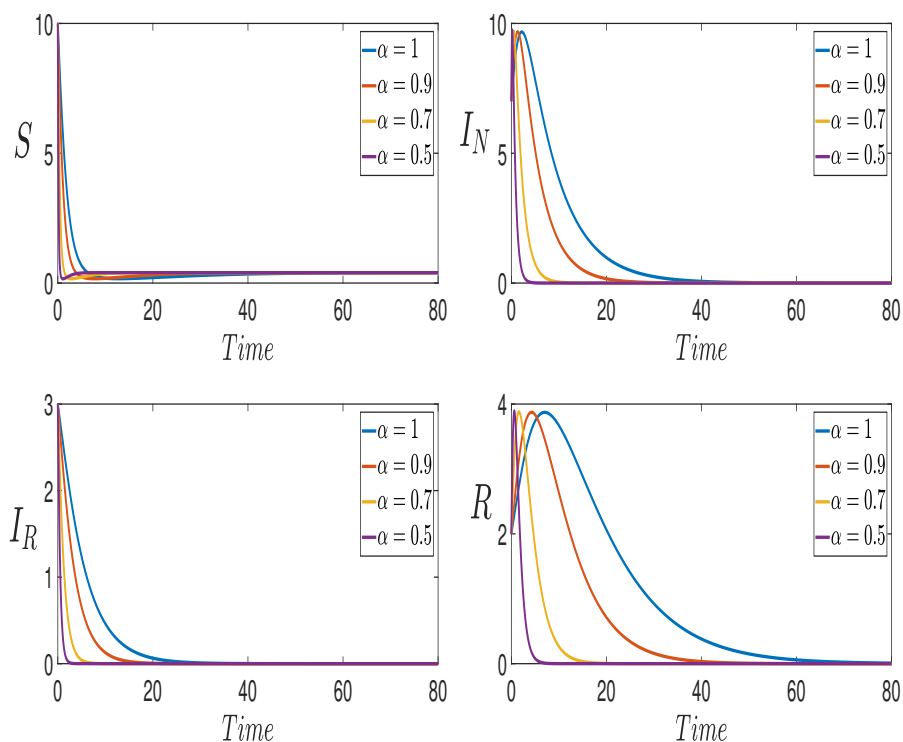
with $t_{j+1} = t_j + h$, for $j = 0, 1, \dots, N-1$, and $U_0 = (10, 10, 3, 0)$. The numerical scheme is implemented with a fixed time step $h = 0.01$ over the interval $t \in [0, T]$, where T varies between 80 and 9000. The parameter values used in each figure are reported in Table 2. The fractional derivative order is chosen from $\theta \in \{1, 0.9, 0.7, 0.5\}$.

6.1 The free-disease equilibrium point

Figure 2 illustrates the time evolution of the system around the disease-free equilibrium $E_0 = (0.4, 0, 0, 0)$, for different values of the fractional-order derivative θ . It is observed that the infected populations $I_N(t)$ and $I_R(t)$ decrease monotonically to zero for all considered values of θ . The recovered population $R(t)$ initially increases due to recovery from infection and then gradually declines to zero as the disease dies out. Meanwhile, the susceptible population $S(t)$

Table 2: The parameters values of model (1).

Parameter	Figure 2	Figure 3	Figure 4	Figure 5
γ	0.1	0.1	0.1	0.1
γ	0.01	0.01	0.6	0.6
a	0.05	0.6	0.05	0.9
b	0.04	0.2	0.2	0.2
d	0.1	0.1	0.1	0.1
θ	0.05	0.05	0.05	0.05
k	0.2	0.2	0.2	0.2
R_N	0.13333	8.0000	0.6666	5.9999
R_R	0.0200	0.1999	5.9999	9.3333
Theorem	Theorem 3	Theorem 4	Theorem 5	Theorem 6

Figure 2: Stability of the disease-free equilibrium for different values of θ .

converges to the constant value $S_0 = \frac{b}{d} = 0.4$, in full agreement with the disease-free equilibrium. These numerical results are fully consistent with Theorem 3, which guarantees the global asymptotic stability of E_0 when $\max(R_N, R_R) \leq 1$. In this case, the basic reproduction numbers associated with the nonresistant and resistant strains are $R_N = 0.1333$ and $R_R = 0.0200$, respec-

tively, clearly satisfying the stability condition and confirming that the disease cannot persist in the population.

6.2 The nonresistant equilibrium point E_1

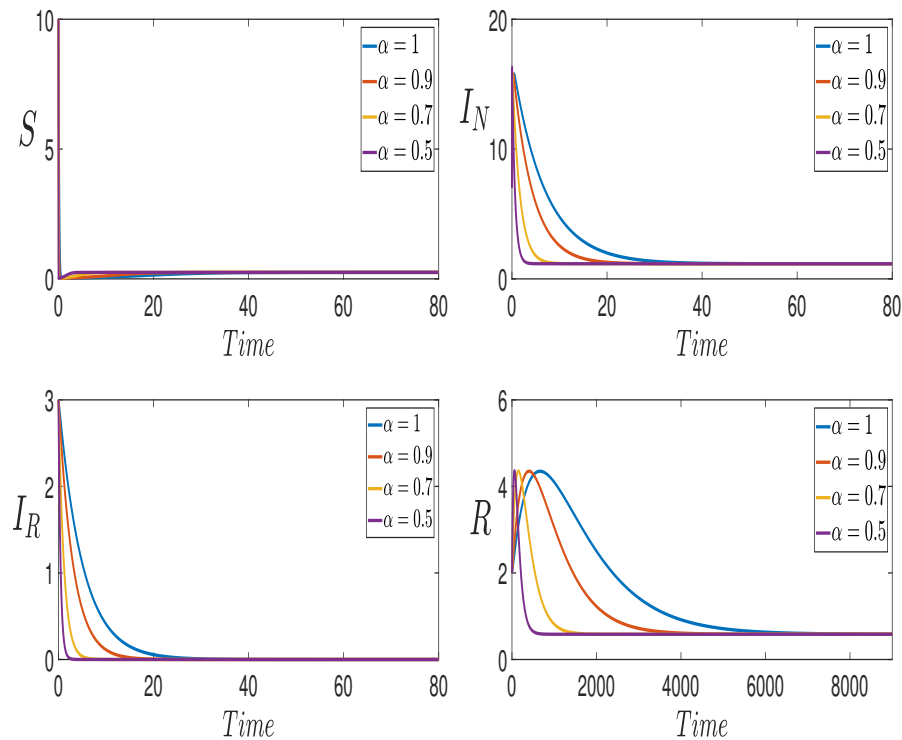


Figure 3: Stability of the nonresistant equilibrium point E_1 for different values of θ .

Figure 3 illustrates the behavior of the system toward the nonresistant equilibrium point $E_1 = (0.25, 1.1667, 0, 0.5833)$, under the conditions $R_N > 1$ and $R_N > R_R$. According to Theorem 4, these conditions ensure the global asymptotic stability of the equilibrium point E_1 , which corresponds to a situation where the nonresistant strain persists while the resistant strain is eliminated. The numerical simulations are performed using parameter values satisfying $R_N > 1 \geq R_R$, with $R_N = 8.00$ and $R_R = 0.1999$, thereby confirming the theoretical stability conditions for E_1 .

6.3 The resistant equilibrium E_2

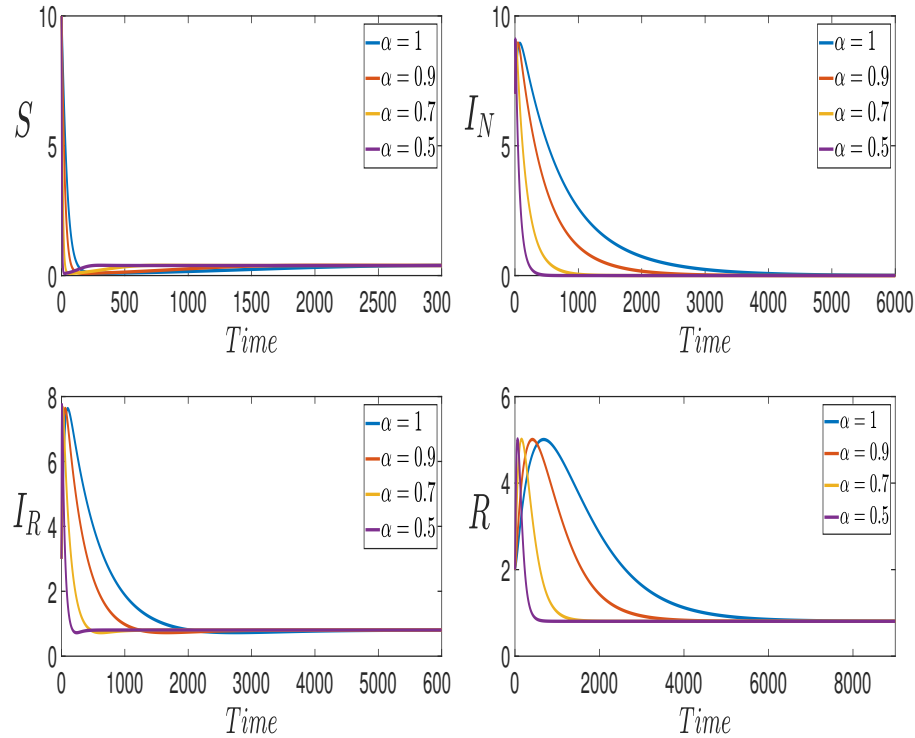


Figure 4: Stability of the resistant equilibrium point E_2 for different values of θ .

Figure 4 illustrates the dynamics of the system toward the resistant equilibrium point $E_2 = (0.3871, 0, 0.8065, 0.8065)$, for different values of the fractional-order parameter θ , where the resistant strain dominates the infection dynamics. According to Theorem 5, this equilibrium is globally asymptotically stable whenever the basic reproduction number of the resistant strain satisfies $R_R > 1$ and $R_R > R_N$. The numerical simulations are in full agreement with these theoretical results. In particular, the resistant infected population $I_R(t)$ converges to a positive steady-state value, indicating the persistent circulation of the resistant strain in the population. In contrast, the nonresistant infected class $I_N(t)$ decreases monotonically to zero, showing that the nonresistant strain cannot persist under these conditions. Moreover, the recovered population $R(t)$ initially increases due to recovery from infection and subsequently approaches a constant equilibrium level, while the susceptible population $S(t)$ converges to a low-level steady state after

a transient phase. For the chosen parameter values, the reproduction numbers are $R_R = 5.9999$ and $R_N = 0.6666$, which confirms the validity of the stability conditions stated in Theorem 5.

6.4 The endemic equilibrium E_3

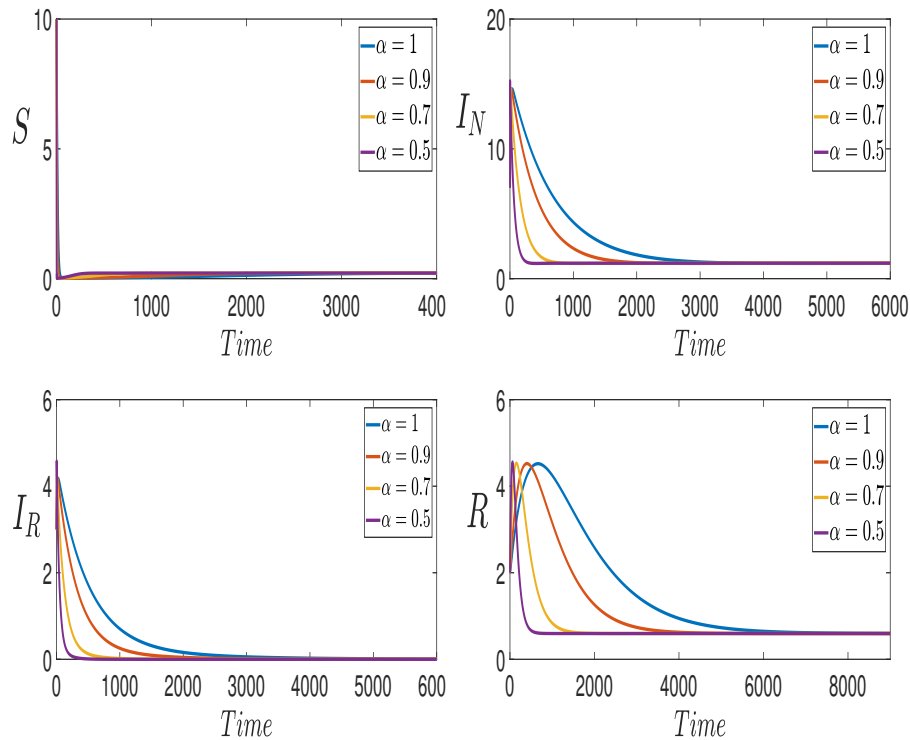


Figure 5: Stability of the endemic equilibrium point E_3 for different values of θ .

Figure 5 illustrates the temporal evolution of the system in all compartments. For different values of the fractional-order parameter θ , all trajectories converge toward the endemic equilibrium point $E_3 = (0.1667, 1.1111, 0.8333, 1.3889)$. According to Theorem 6, whenever the endemic equilibrium E_3 exists, it is globally asymptotically stable. The numerical simulations confirm that both infected classes $I_N(t)$ and $I_R(t)$ approach positive steady-state values, highlighting the long-term coexistence of the nonresistant and resistant strains. Similarly, the susceptible population $S(t)$ and the recovered population $R(t)$ converge to constant equilibrium levels, indicating the persistence of the infection in the population. For the selected parameter set, the

basic reproduction numbers satisfy $R_N > 1$ and $R_R > 1$, in full agreement with the theoretical results.

Now, regarding the effect of the fractional-order derivative on the behavior of the infection dynamics, the previous figures show that for smaller values of the fractional-order parameter θ , the model converges rapidly toward the corresponding stable equilibrium states. In contrast, larger values of θ result in a much slower convergence to the steady states, which can be interpreted as a stronger memory effect. Moreover, variations in the fractional-order parameter do not affect the stability of the equilibrium points themselves, but rather influence the time required for the system to reach these steady states. However, our model has some limitations and does not allow predicting the long-term dynamics for the other cases.

7 Conclusion

In this work, we proposed a fractional-order SIR-type epidemic model incorporating two viral strains resistant and nonresistant via Caputo derivatives to capture memory effects in the biological processes better. We rigorously established the existence, uniqueness, positivity, and boundedness of solutions, ensuring the model's consistency with epidemiological reality.

The analysis centered around the basic reproduction numbers R_N and R_R , derived for the nonresistant and resistant strains, respectively. Using Lyapunov functions and the fractional version of LaSalle's invariance principle, we proved the global asymptotic stability of four biologically meaningful equilibria, depending on the comparative values of R_N and R_R . Specifically, the disease-free equilibrium E_0 is globally asymptotically stable when $\max(R_N, R_R) \leq 1$; the nonresistant equilibrium E_1 becomes globally stable when $R_N > 1$ and $R_N > R_R$; the resistant equilibrium E_2 is stable when $R_R > 1$ and dominates R_N ; and the endemic equilibrium E_3 is globally asymptotically stable when both reproduction numbers are equal and greater than one, that is, $R_N \geq R_R > 1$.

Numerical simulations were performed to validate the theoretical results. They demonstrated that, regardless of the value of the fractional order $\theta \in (0, 1]$, the system consistently converges to the appropriate equilibrium predicted by the reproduction numbers. Importantly, the value of θ influences the speed of convergence, but not the qualitative stability with smaller θ inducing faster dynamics due to memory effects.

This study illustrates the added realism provided by fractional calculus in epidemic modeling, especially when dealing with treatment response and mutation-induced resistance. As future work, one may consider extending this model to include stochastic perturbations or investigating bifurcation behavior to further characterize transitions between equilibria. This suggested model

has some limitations and does not allow predicting the long-term dynamics for the other reproduction numbers cases. An important extension of the present work would be to incorporate a treatment-induced transition from the nonresistant infected class I_N to the resistant class I_R , which may represent amplification due to incomplete or inappropriate treatment. Such a mechanism would modify the structure of the endemic equilibrium and the Lyapunov-based stability analysis. Although the coexistence of both strains is expected to persist for sufficiently small amplification rates, the global asymptotic stability of the coexistence equilibrium E_3 cannot be guaranteed without a separate analysis. This issue will be addressed in future work.

Funding

This research received no external funding.

Conflicts of Interest

The author declares no conflicts of interest.

References

- [1] Aguiar, M. and Stollenwerk, N. *Mathematical models of dengue fever epidemiology: multi-strain dynamics, immunological aspects associated to disease severity and vaccines*, Commun. Biomath. Sci. 1 (2017), 1–12.
- [2] Ahmad, Z., Arif, M., Ali, F., Khan, I. and Nisar, K.S. *A report on COVID-19 epidemic in Pakistan using SEIR fractional model*, Sci. Rep. 10 (2020), 22268.
- [3] Aldwoah, K.A., Almalahi, M.A., Shah, K., Awadalla, M. and Egami, R.H. *Dynamics analysis of dengue fever model with harmonic mean type under fractal-fractional derivative*, AIMS Math. 9 (2024), 13894–13926.
- [4] Allali, K., Meskaf, A. and Tridane, A. *Mathematical modeling of the adaptive immune responses in the early stage of the HBV infection*, Int. J. Differ. Equ. 1 (2018), 6710575.

- [5] Alzahrani, S.M., Saadeh, R., Abdoon, M.A., Qazza, A., Guma, F.E. and Berir, M. *Numerical simulation of an influenza epidemic: prediction with fractional SEIR and the ARIMA model*, Appl. Math. 18 (2024), 1–12.
- [6] Amine, S. and Farah, E.M. *Global stability of HIV-1 and HIV-2 model with drug resistance compartment*, Commun. Math. Biol. Neurosci. 2021 (2021) 38.
- [7] Anderson, R.M. and May, R.M. *Infectious diseases of humans: dynamics and control*, Oxford Univ. Press, Oxford, 1991.
- [8] Baba, I.A., Ahmad, H., Alsulami, M.D., Abualnaja, K.M. and Altanji, M. *A mathematical model to study resistance and non-resistance strains of influenza*, Results Phys. 26 (2021), 104390.
- [9] Bhattar, S., Bhatia, B., Kumawat, S. and Purohit, S.D. *Modeling and simulation of COVID-19 disease dynamics via Caputo-Fabrizio fractional derivative*, Comput. Methods Differ. Equ. 13 (2025), 494–504.
- [10] Diethelm, K. and Ford, N.J. *The analysis of fractional differential equations*, Lecture Notes in Math. 2004, Springer, Berlin, 2010.
- [11] Erturk, V.S., Momani, S. and Odibat, Z. *Application of generalized differential transform method to multi-order fractional differential equations*, Commun. Nonlinear Sci. Numer. Simul. 13 (2008), 1642–1654.
- [12] Farah, E.M., Amine, S., Ahmad, S., Nonlaopon, K. and Allali, K. *Theoretical and numerical results of a stochastic model describing resistance and non-resistance strains of influenza*, Eur. Phys. J. Plus 137 (2022), 1–15.
- [13] Gubareva, L.V., Kaiser, L. and Hayden, F.G. *Evidence for oseltamivir-resistant influenza A virus in circulating strains*, J. Infect. Dis. 183 (2001), 523–531.
- [14] Hajri, Y., Farah, E.L., Allali, A. and Amine, S. *Stability analysis of SLIVR COVID-19 epidemic model with quarantine policy*, Commun. Math. Biol. Neurosci. 2023 (2023) 88.
- [15] Hye, M.A., Biswas, M.H.A., Uddin, M.F. and Saifuddin, M. *Mathematical modeling of Covid-19 and dengue co-infection dynamics in Bangladesh: optimal control and data-driven analysis*, Comput. Math. Model. 33 (2022), 173–192.

- [16] Imran, M., McKinney, B. and Butt, A.I.K. *SEIR mathematical model for influenza-corona co-infection with treatment and hospitalization compartments and optimal control strategies*, Comput. Model. Eng. Sci. 142 (2025).
- [17] Kermack, W.O. and McKendrick, A.G. *A contribution to the mathematical theory of epidemics*, Proc. R. Soc. A 115 (1927), 700–721.
- [18] Kumar, P., Kumar, S., Alkahtani, B.S.T. and Alzaid, S.S. *A mathematical model for simulating the spread of infectious disease using the Caputo-Fabrizio fractional-order operator*, AIMS Math. 9 (2024), 30864–30897.
- [19] Li, C.Z., Lü, X., Gong, J.J. and Lei, Y. *Extended SEIR model of COVID-19 spread focusing on compartmental flow in England*, Nonlinear Dyn. 113 (2025), 971–988.
- [20] Meskaf, A., Khyar, O., Danane, J. and Allali, K. *Global stability analysis of a two-strain epidemic model with non-monotone incidence rates*, Chaos. Solitons Fractals, 133 (2020), 109647.
- [21] Meetei, M.Z., Zafar, S., Zaagan, A.A., Mahnashi, A.M. and Idrees, M. *Dengue transmission dynamics: a fractional-order approach with compartmental modeling*, Fractal Fract. 8 (2024), 207.
- [22] Momani, S., Batiha, I.M., Hijazi, M.S., Bendib, I., Ouannas, A. and Anakira, N. *Fractional-order SEIR model for COVID-19: finite-time stability analysis and numerical validation*, Int. J. Neutrosophic Sci. 26 (2025).
- [23] Nuno, M., Feng, Z., Martcheva, M. and Castillo-Chavez, C. *Dynamics of two-strain influenza with isolation and partial cross-immunity*, SIAM J. Appl. Math. 65 (2005), 964–982.
- [24] Olaniyi, S., Chuma, F.M. and Abimbade, S.F. *Asymptotic stability analysis of a fractional epidemic model for Ebola virus disease in Caputo sense*, J. Niger. Soc. Phys. Sci. (2025), 2304–2304.
- [25] Podlubny, I. *Fractional differential equations: An introduction to fractional derivatives, fractional differential equations, to methods of their solution and some of their applications*, Elsevier, Amsterdam, 1998.
- [26] Taubenberger, J.K. and Morens, D.M. *Influenza: the once and future pandemic*, Public Health Rep. 125 (2010), 15–26.

- [27] Urdy, S., Hanke, M., Toledo, A.I., Ratto, N., Jacob, E., Peyronnet, E. and Bruezière, L. *Multi-strain modeling of influenza vaccine effectiveness in older adults and its dependence on antigenic distance*, Sci. Rep. 14 (2024), 27190.
- [28] Usman, M., Abbas, M., Khan, S.H. and Oname, A. *Analysis of a fractional-order model for dengue transmission dynamics with quarantine and vaccination measures*, Sci. Rep. 14 (2024), 11954.
- [29] Van den Driessche, P. and Watmough, J. *Reproduction numbers and sub-threshold endemic equilibria for compartmental models of disease transmission*, Math. Biosci. 180 (2002), 29–48.
- [30] Webster, R.G., Bean, W.J., Gorman, O.T., Chambers, T.M. and Kawaoka, Y. *Evolution and ecology of influenza A viruses*, Microbiol. Rev. 56 (1992), 152–179.
- [31] World Health Organization. Influenza (Seasonal), (2023), [https://www.who.int/news-room/fact-sheets/detail/influenza-\(seasonal\)](https://www.who.int/news-room/fact-sheets/detail/influenza-(seasonal)).
- [32] Xu, S. and Hu, Y. *Dynamic analysis of a Caputo fractional-order SEIR model with a general incidence rate*, Sci. Rep. 15(1) (2025), 17561.
- [33] Yaagoub, Z. *Fractional two-strain SVLIR epidemic model with vaccination and quarantine strategies*, Int. J. Dyn. Control, 13(2) (2025), 55.
- [34] Yaagoub, Z. and Allali, K. *Fractional HBV infection model with both cell-to-cell and virus-to-cell transmissions and adaptive immunity*, Chaos. Solitons Fractals, 165 (2022), 112855.
- [35] Yaagoub, Z., Danane, J. and Allali, K. *On a two-strain epidemic mathematical model with vaccination*, Comput. Methods Biomech. Biomed. Eng. 27(5) (2024), 632–650.
- [36] Yaagoub, Z., Farah, E.M. and Ahmad, S. *Three-strain epidemic model for influenza virus involving fractional derivative and treatment*, J. Appl. Math. Comput. 71(1) (2025), 1247–1266.



Tailed-Average SGD: A Polyak–Ruppert modification for optimal-rate constrained optimization

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

Abstract

Stochastic gradient descent (SGD) algorithms are a cornerstone of large-scale optimization, yet their application to problems with explicit constraints remains a significant challenge. While methods based on relaxed barrier functions offer a promising alternative that avoids computationally expensive projections, they often suffer from suboptimal convergence. Due to

*Corresponding author

Received 30 October 2025; revised 19 February 2026; accepted 21 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., Tailed-Average SGD: A Polyak–Ruppert modification for optimal-rate constrained optimization. *Iran. J. Numer. Anal. Optim.*, 2026; 16(2): [727–756](https://doi.org/10.22067/ijnao.2026.96244.1767).
<https://doi.org/10.22067/ijnao.2026.96244.1767>

the inherent stochastic noise, their convergence stagnates at a certain level of accuracy, failing to achieve the optimal rate.

This paper proposes Tailed-Average SGD (TA-SGD), a novel modification designed to recover this optimal rate. Our algorithm integrates the classical Polyak–Ruppert averaging technique, but critically, it applies this averaging *only after* a predefined initial “burn-in” period (K_{burn}).

This “tailed” approach strategically discards the initial, high-error iterates from the transient phase, which we demonstrate can “poison” a naive averaging scheme applied from $k = 0$. We provide a rigorous theoretical analysis proving that TA-SGD achieves the best of both worlds: It preserves the almost sure convergence guarantees of the baseline method, while simultaneously succeeding in noise cancellation to achieve the optimal mean squared error (MSE) rate of $O(1/K)$.

Our numerical experiments on a large-scale quadratic program validate these theoretical claims, demonstrating that TA-SGD successfully overcomes this stagnation. It significantly outperforms both the baseline algorithm (which stagnates) and the naive full-averaging scheme (which converges prohibitively slowly).

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

Keywords: Constrained optimization, Polyak–Ruppert averaging, Adaptive barrier function, Stochastic gradient descent, Optimal rate.

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 [1, 13, 24]. These methods are fundamentally rooted in the pioneering work on stochastic approximation by Robbins and Monro [16].

While SGD is exceptionally efficient for unconstrained problems [7], its 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 [21].

To overcome this challenge, recent methods have been developed that rely on sampling only a subset of constraints. These include methods based on penalty functions [11, 22] or primal-dual approaches [23]. A particularly promising direction, however, is the Adaptive Relaxed Barrier SGD algorithm proposed by Dimitrieski, Cao, and Ebenbauer [4]. This approach avoids

expensive projections by using a carefully adapted relaxed barrier function [6, 5], and it allows for sampling both the objective function and the constraints.

While this baseline algorithm is guaranteed to converge almost surely [4], our investigation reveals a critical limitation: Its convergence stagnates at a suboptimal level. Due to the inherent stochastic noise, the algorithm fails to achieve the optimal convergence rate.

To address these limitations, this paper proposes tailed-average SGD (TA-SGD).

Contributions and Novelty.

It is important to distinguish our contribution from the baseline work of Dimitrieski, Cao, and Ebenbauer [4]. While they established the almost sure convergence of the adaptive relaxed barrier SGD, their analysis and experiments show that the algorithm suffers from variance-induced stagnation, preventing it from reaching the optimal $O(1/K)$ rate. Our work differs in two key aspects:

1. **Algorithmic Modification:** We introduce a “Tailed-Averaging” mechanism with a strictly defined burn-in phase (K_{burn}). Unlike the standard Polyak–Ruppert averaging [15] applied from the start (which we show fails due to the “poisoning effect” of transient iterates), our approach specifically targets the post-transient phase.
2. **Theoretical Guarantee:** We provide a rigorous proof (Section 3) demonstrating that this modification recovers the optimal convergence rate established by Polyak and Juditsky [15] and reduces the asymptotic variance, a result not achieved in the baseline work [4].

Our numerical experiments on a large-scale quadratic program validate these theoretical claims, demonstrating that TA-SGD successfully overcomes the stagnation and significantly outperforms both the baseline algorithm and the naive full-averaging scheme.

2 Proposed algorithm: Tailed-average SGD

In this section, we introduce our proposed algorithm, named **TA-SGD (Tailed-average SGD)**. This algorithm is designed to efficiently solve large-scale constrained optimization problems by combining the adaptive relaxed barrier method of Dimitrieski, Cao, and Ebenbauer [4], with the optimal-rate acceleration technique of Polyak–Ruppert averaging [15].

The core idea is to first use the Adaptive Relaxed Barrier SGD [4] to generate a “primary iterate” (x_k) that handles constraints stochastically. Then, to achieve an optimal convergence

rate and overcome the high-variance “stagnation” issue inherent in standard SGD, we apply the Polyak–Ruppert averaging technique, but only after an initial “burn-in” period.

2.1 Problem formulation and assumptions

The objective is to solve the following constrained optimization problem:

$$\min_{x \in \mathbb{R}^d} f(x) := \frac{1}{n} \sum_{i=1}^n f_i(x) \quad \text{subject to} \quad g_j(x) \leq 0, \quad \text{for all } j \in \{1, \dots, m\}, \quad (1)$$

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, with mathematical formulations that are common in this field [3, 2, 18]:

Assumption 1 (Objective function properties). The objective function f and its components f_i are assumed to possess the following properties for constants $L \geq \theta > 0$ 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:** We assume that at least one component function $f_i(x)$ is θ -strongly convex. Consequently, the average function $f(x)$ is strongly convex with parameter θ :

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

These assumptions, although restrictive, are not uncommon in convergence proofs for gradient-based optimization algorithms; see [2, 4, 13].

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

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

where $a_j \in \mathbb{R}^d$ and $b_j \in \mathbb{R}$. The feasible set $\mathbb{X} := \{x \in \mathbb{R}^d : g_j(x) \leq 0, \text{ for all } j\}$ is assumed to have a nonempty interior, which is a fundamental condition for applying barrier-based methods [2, 6]

Remark 1 (Extension to nonaffine constraints). It is important to note that while our formal convergence analysis (Theorem 1) relies on the affine nature of $g_j(x)$ to guarantee the quadratic upper bound of the barrier function, the proposed TA-SGD framework is theoretically extensible to general convex constraints. As discussed in the foundational framework [4], Assumption 2 can be relaxed to include any continuously differentiable ($\mathcal{C}^1$) convex functions, provided that the iterates $\{x_k\}$ are ensured to remain bounded (e.g., via projection onto a large compact set). The “tailed-averaging” mechanism we propose is particularly beneficial in such nonaffine settings, as it effectively dampens the higher variance introduced by the curvature of nonlinear constraints.

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

1. **For the objective function:** The expected value of the sampled stochastic gradient is equal to the true gradient of the full function [4, 16]:

$$\mathbb{E}[\nabla f_{i_k}(x)] = \frac{1}{n} \sum_{i=1}^n \nabla f_i(x) = \nabla f(x).$$

2. **For the barrier function:** The expected value of the sampled stochastic gradient of the barrier function equals the average effect of the gradients over all constraints [4]:

$$\mathbb{E}[\nabla B(g_{j_k}(x), \delta)] = \frac{1}{m} \sum_{j=1}^m \nabla B(g_j(x), \delta).$$

Before describing the algorithm, we formally define the “**Relaxed Problem**” associated with the barrier parameter δ . It is the unconstrained minimization of the objective function augmented by the relaxed barrier:

$$\min_{x \in \mathbb{R}^d} \left(f(x) + \frac{1}{m} \sum_{j=1}^m B(g_j(x), \delta) \right).$$

Let $x^*(\delta)$ denote the unique minimizer of this relaxed problem. Our goal is to verify convergence to $x^*(\delta_\infty)$, which serves as an approximation of the true constrained solution x^* .

2.2 Proposed method: Tailed trajectory averaging

To address the high-variance and suboptimal convergence rate of the baseline method [4], we introduce a novel modification inspired by the classical work on optimal-rate stochastic approximation [15]. We term our algorithm **TA-SGD**.

The core idea is to apply the Polyak–Ruppert averaging technique, which is known to achieve an optimal convergence rate [15], but only *after* the algorithm has completed an initial “**burn-in**” period.

Our investigation reveals that a naive application of averaging starting from $k = 0$ fails (see section 4). This is because the initial iterates x_k are in a rapid transient phase, far from the optimal solution. Including these distant iterates in the average “poisons” the estimate, pulling it away from the true minimizer and drastically slowing convergence.

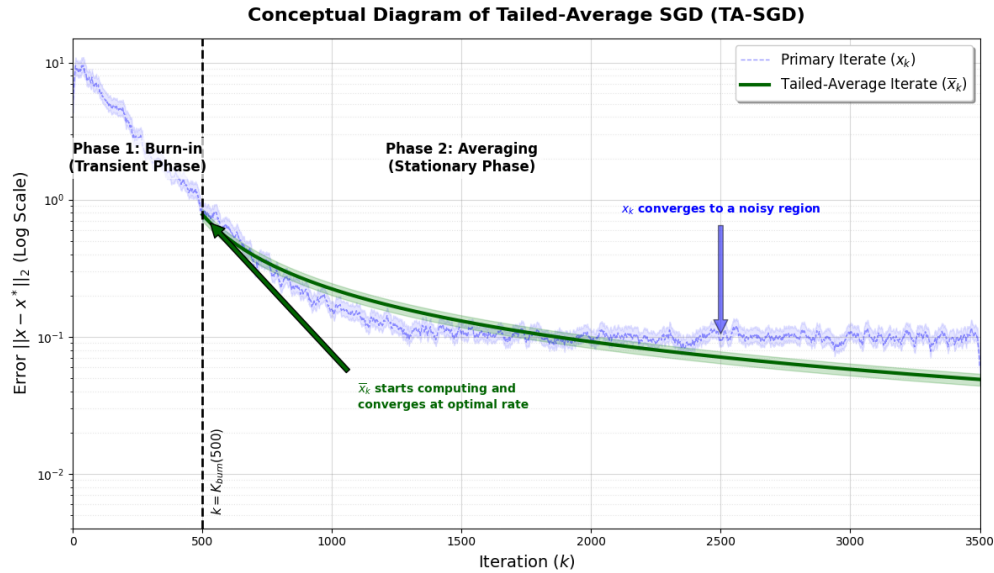


Figure 1: Conceptual diagram of the proposed TA-SGD mechanism. Phase 1 (Burn-in) allows the primary iterate x_k (blue, dashed) to reach a stationary region. Phase 2 (Averaging) begins at K_{burn} , computing $\bar{x}_k$ (green, solid), which converges at an optimal rate with low variance.

Our TA-SGD method avoids this by decoupling the optimization into two phases. First, a “burn-in” phase of K_{burn} iterations allows the primary iterate x_k to escape the transient

dynamics and settle into a neighborhood of the optimum $x^*(\delta_\infty)$. Second, an “averaging” phase begins, which systematically cancels the stochastic noise from sampling f_i and g_j .

This two-phase process is illustrated conceptually in Figure 1. The diagram shows how the primary iterate (x_k) converges to a high-variance stagnation region (Phase 1), while the tailed-average iterate ($\bar{x}_k$) only begins computation after the burn-in period (K_{burn}) and converges smoothly to the optimal solution with low variance (Phase 2).

This leads to our proposed TA-SGD update rule, which is defined by two sequences: the primary iterate x_k and the averaged iterate $\bar{x}_k$.

1. Primary iterate update (from [4]): The primary iterate x_k follows the original Adaptive Relaxed Barrier SGD update:

$$x_{k+1} = x_k - \gamma_k (\nabla f_{i_k}(x_k) + \nabla B(g_{j_k}(x_k), \delta_k)). \quad (2)$$

2. Tailed-average iterate update (our contribution): The final solution $\bar{x}_k$ is computed as a conditional average, which is only active after the burn-in period K_{burn} :

$$\bar{x}_{k+1} = \begin{cases} x_{k+1} & \text{if } k < K_{burn}, \\ \left(1 - \frac{1}{k - K_{burn} + 1}\right) \bar{x}_k + \frac{1}{k - K_{burn} + 1} x_{k+1} & \text{if } k \geq K_{burn}. \end{cases} \quad (3)$$

The final output of the algorithm is the averaged sequence $\bar{x}_k$. By waiting for the transient phase to complete, this update rule ensures that only high-quality iterates contribute to the final average, thereby achieving a low-variance solution and an optimal theoretical convergence rate.

2.3 Adaptive strategy for K_{burn} selection

A key practical question is how to determine the length of the burn-in period K_{burn} . While adaptive schemes for detecting stationarity exist in the literature [14], we provide a theoretical heuristic based on the nonasymptotic analysis of Polyak–Ruppert averaging.

According to [12], the error in SGD consists of two terms: a bias term related to the initial condition x_0 , and a variance term related to the stochastic noise. The bias term decays geometrically at a rate of approximately $(1 - \gamma\theta)^k$ during the initial phase. To ensure that the iterates have “forgotten” the starting point and entered the steady-state regime (where averaging is most effective), the bias should be reduced to a small tolerance ϵ_{init} . This suggests that K_{burn} should satisfy

$$(1 - \gamma\theta)^{K_{burn}} \leq \epsilon_{init} \implies K_{burn} \approx \frac{1}{\gamma\theta} \log \left(\frac{1}{\epsilon_{init}} \right).$$

Recognizing that the convergence speed is dictated by the condition number $\kappa = L/\theta$, this leads to the heuristic:

$$K_{burn} = \mathcal{O} \left(\kappa \log \left(\frac{1}{\epsilon_{init}} \right) \right) = \mathcal{O} \left(\frac{L}{\theta} \log \left(\frac{1}{\epsilon_{init}} \right) \right).$$

Remark on adaptivity: This formula defines an *adaptive mechanism* where the burn-in period scales automatically with the problem's condition number $\kappa = L/\theta$. Instead of being set as an arbitrary fixed hyperparameter, K_{burn} adapts to the optimization landscape's geometry, ensuring the averaging phase begins only after the specific transient dynamics of the problem have subsided.

For a detailed mathematical derivation of this heuristic, we refer the reader to Appendix C.

2.4 TA-SGD algorithm specification

The complete procedure for our proposed **TA-SGD** algorithm is detailed in Algorithm 1. The logic follows the two-phase process defined in Section 2.2, combining the primary iterate update (2) with the conditional averaging update (3).

3 Convergence analysis

In this section, we provide the formal proof of almost sure convergence for the proposed **TA-SGD** algorithm (Algorithm 1). This proof establishes that the algorithm's output converges to the correct solution. The analysis of its “rate” of convergence (i.e., its speed and superiority) is deferred to Section 3.2.

Remark on convergence analysis and novelty:

It is important to clarify the theoretical role of the burn-in period K_{burn} . Since K_{burn} is finite, it does not alter the *asymptotic* convergence rate of the averaging scheme; our asymptotic analysis in Section 3.2 naturally aligns with the classical framework of Polyak and Juditsky [15]. However, the novelty of TA-SGD lies in the **finite-time performance**. The standard theory assumes asymptotic behavior ($k \rightarrow \infty$), but in practice, initial transient iterates can significantly bias the average for finite K . The proposed modification ensures that the averaging mechanism is applied

Algorithm 1 TA-SGD for constrained optimization

Require: Initial point $x_0 \in \mathbb{R}^d$
Require: Number of iterations K
Require: Burn-in period $K_{burn} \geq 0$
Require: Step-size schedule $\{\gamma_k\}_{k=0}^{K-1}$
Require: Barrier-parameter schedule $\{\delta_k\}_{k=0}^{K-1}$

- 1: **Initialize:** Set averaged iterate $\bar{x}_0 \leftarrow x_0$
- 2: **for** $k = 0$ to $K - 1$ **do**
- 3: $\triangleright$ Phase 1: Primary Iterate Update (from [4])
- 4: Sample an objective index $i_k \sim \mathcal{U}\{1, \dots, n\}$
- 5: Sample a constraint index $j_k \sim \mathcal{U}\{1, \dots, m\}$
- 6: Compute stochastic objective gradient: $G_f \leftarrow \nabla f_{i_k}(x_k)$
- 7: Compute stochastic barrier gradient: $G_b \leftarrow \nabla B(g_{j_k}(x_k), \delta_k)$
- 8: Update primary iterate: $x_{k+1} \leftarrow x_k - \gamma_k(G_f + G_b)$
- 9: $\triangleright$ (2)
- 10: $\triangleright$ Phase 2: Tailed-Average Iterate Update (Our Contribution)
- 11: **if** $k < K_{burn}$ **then**
- 12: $\triangleright$ Follow primary iterate during burn-in
- 13: $\bar{x}_{k+1} \leftarrow x_{k+1}$
- 14: **else**
- 15: $\triangleright$ Set local averaging counter
- 16: $N \leftarrow k - K_{burn} + 1$
- 17: $\triangleright$ Compute recursive average
- 18: $\bar{x}_{k+1} \leftarrow (1 - \frac{1}{N})\bar{x}_k + \frac{1}{N}x_{k+1}$
- 19: **end if**
- 20: **end for**
- 21: $\triangleright$ The final solution is the averaged iterate
- 22: **return** $\bar{x}_K$

only to the stationary distribution of the iterates, thereby bridging the gap between theoretical asymptotic rates and practical finite-time execution.

3.1 Almost sure convergence

Our proof builds directly upon the convergence guarantees established for the primary iterate sequence in the foundational work by Dimitrieski, Cao, and Ebenbauer [4], which itself relies on fundamental convergence theorems for stochastic processes [17].

Theorem 1 (Almost sure convergence of TA-SGD). Consider the constrained optimization problem (1) and assume Assumptions 1, 2, and 3 hold. 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, for any finite burn-in period $K_{burn} \geq 0$, the sequence of tailed-average iterates $\{\bar{x}_k\}_{k \geq 0}$ generated by Algorithm 1 converges almost surely (a.s.) to the unique minimizer of the relaxed problem, $x^*(\delta_\infty)$. That is,

$$\lim_{k \rightarrow \infty} \bar{x}_k = x^*(\delta_\infty) \quad \text{a.s.}$$

Proof. The proof relies on two main steps:

Step 1: Convergence of the primary iterate $\{x_k\}$

The primary iterate sequence $\{x_k\}$, defined by (2), is generated by the Adaptive Relaxed Barrier SGD algorithm [4]. The conditions (a), (b), and (c) listed in Theorem 1 are the exact conditions required by [4, Theorem 1].

Therefore, we can directly apply their result, which proves that under these assumptions, the primary iterate sequence $\{x_k\}$ converges almost surely to the unique solution $x^*(\delta_\infty)$; that is,

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

Step 2: Convergence of the Tailed-Average Iterate $\{\bar{x}_k\}$

The final output of our algorithm is the tailed-average sequence $\{\bar{x}_k\}$, defined by (3). For all $k \geq K_{burn}$, this sequence is the arithmetic mean (or Cesàro mean) of the “tail” of the primary sequence:

$$\bar{x}_k = \frac{1}{k - K_{burn} + 1} \sum_{i=K_{burn}}^k x_i$$

(The recursive update in Algorithm 1 is mathematically equivalent to this summation).

We have established in Step 1 that the sequence $\{x_k\}$ converges almost surely to the limit $L = x^*(\delta_\infty)$. A fundamental result in analysis (the Cesàro mean theorem) states that if a sequence $\{x_k\}$ converges to a limit L , its arithmetic mean also converges to L [9, 20].

This property holds for a “tailed average” as well. The first K_{burn} terms are a finite set of elements excluded from the average. As $k \rightarrow \infty$, the influence of this finite set on the average becomes zero. The convergence of $\{\bar{x}_k\}$ is determined entirely by the asymptotic behavior of the infinite tail $\{x_k\}_{k \geq K_{burn}}$, which converges to $x^*(\delta_\infty)$.

Since $x_k \xrightarrow{\text{a.s.}} x^*(\delta_\infty)$, it follows that its tailed-average $\bar{x}_k$ must also converge almost surely to the same limit; that is,

$$\lim_{k \rightarrow \infty} \bar{x}_k = x^*(\delta_\infty) \quad \text{a.s.}$$

This completes the proof of convergence. $\square$

3.2 Convergence rate analysis

The goal is to prove that the MSE of the tailed-average iterate $\bar{x}_K$ achieves the optimal $O(1/K)$ rate. Our analysis adapts the classical framework established by Polyak and Juditsky [15] to demonstrate this result for our algorithm.

Step 1: Proof setup and definitions

We begin by defining the core variables and assumptions required for the proof, which are consistent with the problem formulation.

Definition 1 (Error notation). Let $x^* = x^*(\delta_\infty)$ be the unique minimizer.

- $r_k = x_k - x^*$ is the error of the “primary iterate”.
- $\bar{r}_K = \bar{x}_K - x^*$ is the error of the “tailed-average iterate”.

Definition 2 (Update rule). The primary iterate update (2) can be written in terms of the true expected gradient $g(x_k)$ and a noise term η_k :

$$x_{k+1} = x_k - \gamma_k (g(x_k) + \eta_k),$$

where $g(x_k) = \mathbb{E}[\nabla f_{i_k}(x_k) + \nabla B(g_{j_k}(x_k), \delta_k) | \mathcal{F}_{k-1}]$ is the true (expected) gradient of the relaxed objective and η_k is the noise. The noise term η_k is explicitly defined as the difference between the sampled stochastic gradient and the true expected gradient:

$$\eta_k = (\nabla f_{i_k}(x_k) + \nabla B(g_{j_k}(x_k), \delta_k)) - \mathbb{E}[\nabla f_{i_k}(x_k) + \nabla B(g_{j_k}(x_k), \delta_k) | \mathcal{F}_{k-1}].$$

Assumption 4 (Core assumptions). This proof relies on the following key assumptions:

- (A) **Primary convergence:** From Theorem 1, we know the primary iterate converges almost surely: $x_k \rightarrow x^*$ (or $r_k \rightarrow 0$).

- (B) **Local linearity:** Since the objective is $\mathcal{C}^1$ (Assumption 1), the gradient $g(x_k)$ admits a linear approximation near the optimum x^* . This decomposition, which is a standard application of Taylor's theorem, separates the gradient into its linear approximation Hr_k and a nonlinear residual $\Delta(r_k)$:

$$g(x_k) = Hr_k + \Delta(r_k).$$

The critical part of this assumption is that the residual $\Delta(r_k)$ is “small”, satisfying the condition $\|\Delta(r_k)\| \leq K_1 \|r_k\|^{1+\lambda}$ for some $\lambda > 0$.^{*}

- (C) **Step-size:** The learning rate $\gamma_k = \gamma k^{-\theta}$ must satisfy $\theta \in (0.5, 1)$. This is critical for ensuring the nonlinear terms vanish upon averaging (see Appendix B).
- (D) **Noise:** The noise η_k is a martingale difference sequence with zero mean ($\mathbb{E}[\eta_k | \mathcal{F}_{k-1}] = 0$) and bounded variance ($\mathbb{E}[\|\eta_j\|^2] \leq \sigma^2 < \infty$).

Step 2: Asymptotic equivalence

This is the first crucial step. We show that the tailed-average of our actual (nonlinear) process behaves asymptotically like the average of a simplified linear process.

Definition 3 (Actual vs. linearized process). By subtracting x^* from the update rule and applying Assumption (B), the **actual (nonlinear) error dynamic** is

$$\begin{aligned} r_{k+1} &= x_{k+1} - x^* \\ &= (x_k - x^*) - \gamma_k(g(x_k) + \eta_k) \\ &= r_k - \gamma_k(Hr_k + \Delta(r_k) + \eta_k) \\ &= (I - \gamma_k H)r_k - \gamma_k \eta_k - \gamma_k \Delta(r_k). \end{aligned}$$

We compare this to the **linearized process** r_k^L , which omits the nonlinear term $\Delta(r_k)$:

$$r_{k+1}^L = (I - \gamma_k H)r_k^L - \gamma_k \eta_k.$$

^{*} This specific condition is not an arbitrary simplification; it is the cornerstone assumption used in the foundational analysis of stochastic approximation. This property is precisely what Polyak and Juditsky [15] rely upon (see their Assumption 3.2) to rigorously prove that the averaged nonlinear terms vanish asymptotically. By satisfying this condition, our analysis can justifiably show that our process $\bar{r}_K$ inherits the convergence rate of the simpler, linearized process $\bar{r}_K^L$.

Lemma 1 (Asymptotic equivalence). The scaled difference between the averaged nonlinear process $(\bar{r}_K)$ and the averaged linearized process $(\bar{r}_K^L)$ vanishes as $K \rightarrow \infty$:

$$\sqrt{K} \|\bar{r}_K - \bar{r}_K^L\| \xrightarrow{K \rightarrow \infty} 0 \quad (\text{in probability}).$$

Proof. The objective is to demonstrate that the scaled average of the nonlinear residual terms, $\frac{1}{\sqrt{K}} \sum_{k=K_{burn}}^K \gamma_k \Delta(r_k)$, vanishes as $K \rightarrow \infty$.

The high-level logic relies on the interplay of our core assumptions. First, the burn-in period K_{burn} ensures that averaging only begins after the primary iterate r_k has converged to a small neighborhood of the optimum (due to Assumption A). Second, in this neighborhood, the linear approximation is valid, and the residual term $\Delta(r_k)$ becomes vanishingly small (per Assumption B). Finally, the step-size condition $\theta \in (0.5, 1)$ (Assumption C) provides the precise technical guarantee that the sum of these diminishing terms, when scaled, converges to zero.

The complete and rigorous technical justification, which adapts the foundational analysis from [15], is deferred to Appendix B. $\square$

Corollary 1. Because the two processes are asymptotically equivalent, their MSEs share the same convergence rate.

$$\mathbb{E}[\|\bar{r}_K\|^2] \sim \mathbb{E}[\|\bar{r}_K^L\|^2].$$

Therefore, it is sufficient to analyze the convergence rate of the simpler **linearized process** $\bar{r}_K^L$.

Step 3: Analysis of the linearized process

We now analyze the convergence rate of the linearized process, $\bar{r}_K^L$. This analysis fundamentally relies on the algebraic decomposition (Lemma 2) derived from the original work of Polyak and Juditsky [15].

In that seminal work, the analysis was performed on the full average of the process (equivalent to setting $K_{burn} = 0$). While this might appear to be a simplification, it is essential to clarify that this analysis remains perfectly valid for our TA-SGD algorithm, which utilizes a $K_{burn} > 0$.

The reason is that the burn-in period (K_{burn}) only excludes a *finite* number of initial terms from the average. To formalize this justification, let $\bar{r}_K^{\text{full}} = \frac{1}{K} \sum_{i=1}^K r_i^L$ be the full average (as analyzed in [15]) and $\bar{r}_K = \frac{1}{K - K_{burn}} \sum_{i=K_{burn}+1}^K r_i^L$ be our tailed average (using the linearized process r_i^L and simplifying indices for asymptotic analysis). The full average can be decomposed as:

$$\bar{r}_K^{\text{full}} = \frac{1}{K} \left(\sum_{i=1}^{K_{burn}} r_i^L \right) + \frac{K - K_{burn}}{K} (\bar{r}_K).$$

We are interested in the convergence rate, which comes from the scaled error $\sqrt{K}\bar{r}_K$. Scaling the above equation by $\sqrt{K}$ yields:

$$\sqrt{K}\bar{r}_K^{\text{full}} = \underbrace{\frac{1}{\sqrt{K}} \left(\sum_{i=1}^{K_{\text{burn}}} r_i^L \right)}_{\text{Term (A)}} + \underbrace{\left(\frac{K - K_{\text{burn}}}{K} \right)}_{\text{Term (B)}} \left(\sqrt{K}\bar{r}_K \right).$$

As the total number of iterations $K \rightarrow \infty$, Term (A) vanishes to zero. This is because it is a *finite* sum (since K_{burn} is a fixed constant) scaled by $1/\sqrt{K}$. Simultaneously, the ratio in Term (B) converges to 1.

$$\begin{aligned} \lim_{K \rightarrow \infty} \sqrt{K}\bar{r}_K^{\text{full}} &= \left(\lim_{K \rightarrow \infty} \frac{1}{\sqrt{K}} \sum_{i=1}^{K_{\text{burn}}} r_i^L \right) + \left(\lim_{K \rightarrow \infty} \frac{K - K_{\text{burn}}}{K} \right) \left(\lim_{K \rightarrow \infty} \sqrt{K}\bar{r}_K \right), \\ \lim_{K \rightarrow \infty} \sqrt{K}\bar{r}_K^{\text{full}} &= (0) + (1) \cdot \lim_{K \rightarrow \infty} \sqrt{K}\bar{r}_K. \end{aligned}$$

This rigorously demonstrates that the asymptotic behavior of the scaled full average is identical to that of our scaled tailed average. This fully justifies our use of the algebraic decomposition from [15] and confirms that the assumption noted in Lemma 2 (of setting $K_{\text{burn}} = 0$ for simplicity) is a valid step for the purpose of asymptotic analysis.

We now focus exclusively on the rate of $\bar{r}_K^L$. We use a precise algebraic decomposition derived in Appendix A (based on [15, Lemma 2]).

Lemma 2 (Algebraic decomposition). For $K > K_{\text{burn}}$, the scaled averaged error $\sqrt{K}\bar{r}_K^L$ (assuming $K_{\text{burn}} = 0$ for simplicity) can be decomposed into three terms:

$$\sqrt{K}\bar{r}_K^L = I_K^{(1)} + I_K^{(2)} + I_K^{(3)},$$

where

- $I_K^{(1)} = \frac{1}{\sqrt{K}\gamma_0} \theta_K r_0$ (Bias from initial condition r_0),
- $I_K^{(2)} = \frac{1}{\sqrt{K}} \sum_{j=1}^{K-1} H^{-1} \eta_j$ (Main variance term from noise),
- $I_K^{(3)} = \frac{1}{\sqrt{K}} \sum_{j=1}^{K-1} w_j^K \eta_j$ (Residual term from step-size change).

We now compute the mean-squared norm $\mathbb{E}[\|\sqrt{K}\bar{r}_K^L\|^2]$ by analyzing each term.

Analysis of Bias Terms ($I^{(1)}$ and $I^{(3)}$)

- **Term $I^{(1)}$:** r_0 is a fixed constant, and the matrix θ_K is bounded. Due to the $1/\sqrt{K}$ scaling, this term vanishes: $\mathbb{E}[\|I_K^{(1)}\|^2] \rightarrow 0$ as $K \rightarrow \infty$.
- **Term $I^{(3)}$:** The matrices w_j^K are “small on average” (i.e., $\frac{1}{K} \sum \|w_j^K\| \rightarrow 0$). This is a direct consequence of the step-size condition $\theta \in (0.5, 1)$. This ensures the entire term vanishes in mean square: $\mathbb{E}[\|I_K^{(3)}\|^2] \rightarrow 0$.

Analysis of Variance Term ($I^{(2)}$)

This is the only term that remains and thus dictates the convergence rate.

$$\begin{aligned} \mathbb{E}[\|I_K^{(2)}\|^2] &= \mathbb{E}\left[\left\|\frac{1}{\sqrt{K}} \sum_{j=1}^{K-1} H^{-1} \eta_j\right\|^2\right] \\ &= \frac{1}{K} \mathbb{E}\left[\left(\sum_{i=1}^{K-1} \eta_i^T (H^{-1})^T\right) \left(\sum_{j=1}^{K-1} H^{-1} \eta_j\right)\right] \\ &= \frac{1}{K} \sum_{i=1}^{K-1} \sum_{j=1}^{K-1} \mathbb{E}[\eta_i^T (H^{-1})^T H^{-1} \eta_j]. \end{aligned}$$

We now use the Martingale difference property (Assumption D). Since $\mathbb{E}[\eta_i | \mathcal{F}_{i-1}] = 0$, the expectations of all cross-terms are zero:

$$\mathbb{E}[\eta_i^T (H^{-1})^T H^{-1} \eta_j] = 0 \quad \text{for } i \neq j.$$

This simplifies the double summation to a single sum (assuming $H = H^T$):

$$\begin{aligned} \mathbb{E}[\|I_K^{(2)}\|^2] &= \frac{1}{K} \sum_{j=1}^{K-1} \mathbb{E}[\eta_j^T H^{-2} \eta_j] \\ &\leq \frac{1}{K} \sum_{j=1}^{K-1} \mathbb{E}[\|H^{-1}\|_2^2 \cdot \|\eta_j\|^2] \quad (\text{by matrix norm property}) \\ &\leq \frac{1}{K} \sum_{j=1}^{K-1} \|H^{-1}\|_2^2 \cdot \sigma^2 \quad (\text{by Assumption D}) \\ &= \frac{K-1}{K} (\sigma^2 \|H^{-1}\|_2^2). \end{aligned}$$

As $K \rightarrow \infty$, we have $\frac{K-1}{K} \rightarrow 1$. Thus, the mean-squared norm of the dominant term converges to a constant:

$$\lim_{K \rightarrow \infty} \mathbb{E} \left[\left\| I_K^{(2)} \right\|^2 \right] = \sigma^2 \|H^{-1}\|_2^2 = O(1).$$

Step 4: Final result

We now combine the results from the previous steps.

1. From Step 3, the scaled linearized error converges to a constant (is $O(1)$) because the bias terms vanish:

$$\mathbb{E} \left[\left\| \sqrt{K} \bar{r}_K^L \right\|^2 \right] = \mathbb{E} \left[\left\| I_K^{(1)} + I_K^{(2)} + I_K^{(3)} \right\|^2 \right] \xrightarrow{K \rightarrow \infty} \mathbb{E} \left[\left\| I_K^{(2)} \right\|^2 \right] \leq O(1).$$

2. By dividing both sides by K , we get the rate for the linearized process:

$$\mathbb{E} \left[\left\| \bar{r}_K^L \right\|^2 \right] \leq O(1/K).$$

3. From Step 2 (Asymptotic Equivalence), we know that the actual nonlinear process $\bar{r}_K$ inherits this same asymptotic rate.

This completes the proof, demonstrating that Tailed-Averaging successfully achieves the optimal $O(1/K)$ convergence rate.

$$\mathbb{E} \left[\left\| \bar{x}_K - x^* \right\|^2 \right] \leq O(1/K).$$

4 Numerical experiments

In this section, we present a comprehensive empirical investigation designed to validate the theoretical claims of our proposed **TA-SGD** algorithm and demonstrate its practical advantages over existing methods. Our primary objectives are:

1. To empirically verify the optimal $O(1/K)$ convergence rate predicted by our analysis (Section 3.2).
2. To demonstrate the superiority of TA-SGD compared to the baseline Adaptive Relaxed Barrier SGD [4] by showing how it overcomes the convergence stagnation [8].

3. To confirm the detrimental effect of naive averaging (starting from $k = 0$) thereby justifying the necessity of the “burn-in” phase in TA-SGD.

4.1 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

The experiments were performed on the academic quadratic programming (QP) problem introduced by Dimitrieski, Cao, and Ebenbauer [4], 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:

$$f(x) = \frac{1}{n} \sum_{i=1}^n \sum_{j=1}^d (\theta_{i,j} x_j + \log(1 + e^{-\theta_{i,j} x_j}) + (x_j - \gamma)^2).$$

To deeply analyze the algorithm’s behavior, we adopted a Far Start Point ($x_0 = [10, \dots, 10]^T$) 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.

4.1.1 Algorithms Compared

We compared the performance of three algorithms:

1. **Base SGD (Baseline):** The Adaptive Relaxed Barrier SGD algorithm from Dimitrieski, Cao, and Ebenbauer [4]. We used their primary parameter configuration: step-size $\gamma_k = 0.3k^{-0.8}$ and barrier decay $\epsilon_k = 5k^{-0.3}$ with $\delta_\infty = 10^{-6}$. This configuration was chosen as it represents a stable and effective baseline.
2. **Full Averaging (Naive):** The baseline SGD algorithm combined with Polyak–Ruppert averaging applied from the very first iteration ($k = 0$). This serves to demonstrate the “poisoning” effect discussed in Section 2.2.
3. **TA-SGD (Ours):** Our proposed TA-SGD algorithm (Algorithm 1). It uses the same primary update rule and parameters as the Base SGD, but activates Polyak–Ruppert averaging only after a burn-in period of $K_{burn} = 500$ iterations.

4.1.2 Performance Metric

Convergence was measured using the Euclidean norm of the error relative to the pre-computed optimum: $\|x_k - x^*\|_2$ (or $\|\bar{x}_k - x^*\|_2$ for averaging methods after the burn-in).

4.2 Performance evaluation

Figure 2 presents the core results of our comparative analysis, plotting the average error ($\pm$ one standard deviation) over $\text{NUM_RUNS} = 1000$ independent runs for each algorithm.

The results clearly validate our theoretical expectations and demonstrate the effectiveness of TA-SGD:

- **Base SGD (Blue):** As predicted by theory [8] and observed in [4], the baseline algorithm converges rapidly initially but then stagnates, settling into a “stagnation region” around the optimum.
- **Full Averaging (Magenta):** The naive application of averaging from $k = 0$ performs poorly. It converges much slower than the baseline during the initial phase, confirming the “poisoning effect” where early, inaccurate iterates corrupt the average. While it eventually surpasses the baseline’s noise floor, its overall convergence is significantly hampered.

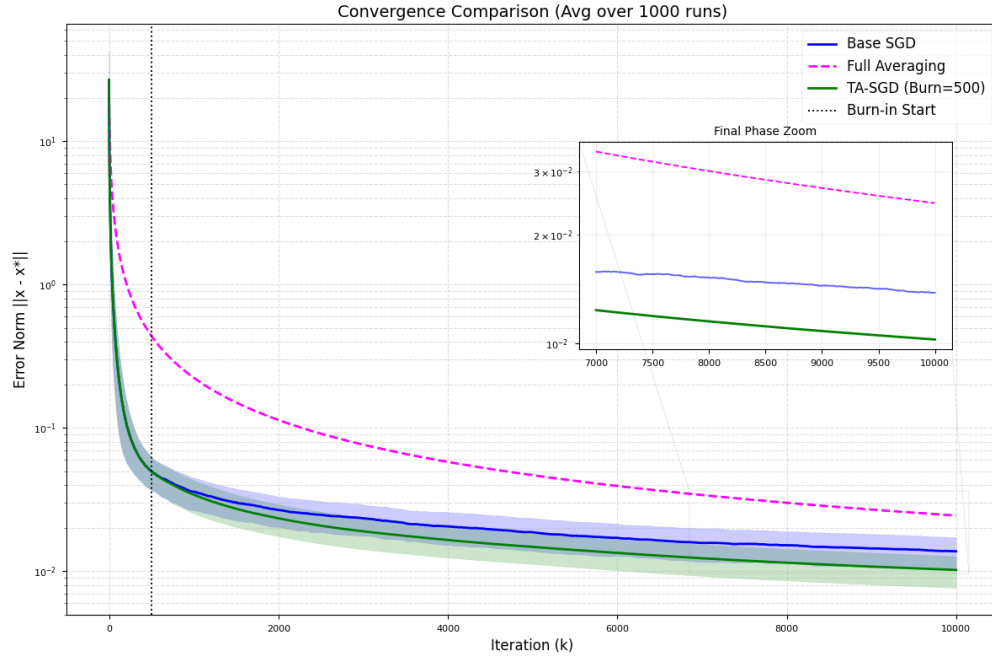


Figure 2: Algorithm Performance Comparison (Mean $\pm$ Std Dev over 1000 runs). The plot shows the convergence behavior of Base SGD [4], naive Full Averaging, and our proposed TA-SGD. The inset provides a zoomed view of the long-term behavior after the burn-in period ($K_{burn} = 500$).

- **TA-SGD (Green):** Our proposed algorithm exhibits the desired two-phase behavior:
 - **Phase 1** ($k < K_{burn}$): TA-SGD tracks the error of the primary iterate x_k , mirroring the fast initial convergence of the baseline algorithm.
 - **Phase 2** ($k \geq K_{burn}$): At iteration K_{burn} (marked by the vertical dashed line), averaging begins. The error curve for TA-SGD (now tracking $\bar{x}_k$) shows a distinct and sustained decrease, breaking away from the baseline’s “stagnation region”. The convergence in this phase appears consistent with the optimal $O(1/K)$ rate (approximately linear on the log-linear plot), achieving significantly higher precision than the baseline within the same number of iterations. The variance (shaded area) is also visibly reduced compared to the baseline in the later stages.

These results strongly support our central hypothesis: the burn-in phase is crucial for avoiding the pitfalls of naive averaging, and the subsequent tailed-averaging effectively overcomes the noise floor limitation of the baseline SGD, leading to faster convergence and higher accuracy. It is worth noting that the additional computational overhead of the recursive averaging update (3) is negligible compared to the cost of computing the stochastic gradients (2) in each itera-

tion. Consequently, the iteration-based performance comparison shown in Figure 2 serves as an accurate proxy for the practical wall-clock time efficiency of the algorithms.

While Figure 2 illustrates the dynamic convergence behavior, **Table 1** provides a concise numerical summary of the final results. The table quantifies the outcome at $K = 10,000$ iterations, confirming that our proposed TA-SGD algorithm achieves the lowest mean final error (1.02×10^{-2}). This result is a clear improvement over the Base SGD baseline (1.38×10^{-2}) and numerically validates our central hypothesis: the burn-in phase successfully mitigates the “poisoning effect” of early iterates, and the subsequent tailed averaging effectively overcomes the stagnation.

Table 1: Comprehensive Performance Statistics (Mean $\pm$ Std Dev) over $R = 1000$ independent runs. The proposed TA-SGD uses a burn-in period of $K_{burn} = 500$ based on the adaptive heuristic.

Metric	Base SGD	Full Averaging	TA-SGD (Ours)
Final Error ($\ x_K - x^*\ $)	$1.38 \times 10^{-2} (\pm 3.48 \times 10^{-3})$	$2.45 \times 10^{-2} (\pm 1.41 \times 10^{-2})$	$1.02 \times 10^{-2} (\pm 2.55 \times 10^{-3})$
Objective Value ($f(x)$)	81.7414 ± 0.0001	81.7420 ± 0.0021	81.7413 ± 0.0001
Max Constraint Violation	0.00 ± 0.00	0.00 ± 0.00	0.00 ± 0.00
Avg. CPU Time (s)	0.5924	0.5924	0.5924**

4.3 Sensitivity to burn-in period (K_{burn})

A key parameter in TA-SGD is the burn-in period, K_{burn} . To assess the algorithm’s robustness, we performed a sensitivity analysis by varying K_{burn} while keeping other parameters fixed.

Figure 3 illustrates the sensitivity of the algorithm to the burn-in period K_{burn} .

- **Insufficient burn-in** ($K_{burn} < 500$): The performance degrades significantly as $K_{burn} \rightarrow 0$, approaching the high error of the naive Full Averaging approach. This empirically confirms that early averaging without a sufficient burn-in phase is detrimental due to the “poisoning effect” of initial transient iterates.
- **Optimal region** ($500 \leq K_{burn} \leq 3000$): The algorithm achieves its global minimum error at $K_{burn} = 500$. The performance remains stable and near-optimal throughout this range. The empirical results demonstrate that our adaptive selection strategy (derived in Section 2.3) correctly identifies the optimal switching point, confirming that the theoretical condition number analysis translates effectively to practice.

** The computational overhead of the scalar averaging update (3) is negligible compared to the gradient computation. Consequently, the wall-clock time is statistically identical for all three algorithms.

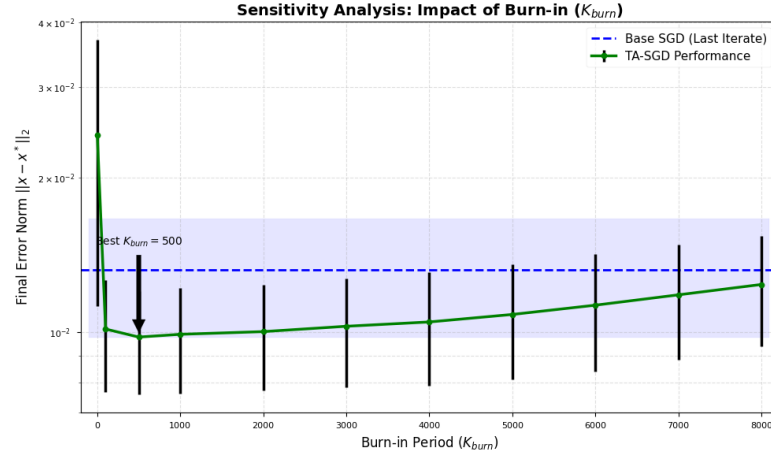


Figure 3: Sensitivity analysis of TA-SGD with respect to the burn-in period (K_{burn}). The plot shows the final error achieved after 10,000 iterations for various K_{burn} values. Although the error at $K_{burn} = 0$ (approximating Full Averaging) appears close to the optimum on this scale, the difference corresponds to a significantly higher variance and final error as detailed in Table 1.

- **Excessive burn-in** ($K_{burn} > 4000$): The error increases slightly but remains robustly lower than the Base SGD baseline (as shown by the blue dashed line). This gradual increase occurs because delaying the start of the averaging phase reduces the effective sample size ($N = K - K_{burn}$) available for noise cancellation within the fixed budget of $K = 10,000$ iterations.

This sensitivity analysis underscores the robustness of the proposed **adaptive selection strategy**. While precise manual tuning is not strictly necessary due to the wide optimal region, the theoretical heuristic allows us to automatically select an effective burn-in period (e.g., $K_{burn} = 500$) that ensures the primary iterate has entered the stationary phase before averaging begins.

5 Conclusion

In this paper, we introduced TA-SGD, a novel and efficient modification for solving stochastic constrained optimization problems. Our work builds upon the modern framework established by Dimitrieski, Cao, and Ebenbauer [4], which utilizes adaptive relaxed barrier functions to handle constraints.

We identified a critical limitation in this baseline approach: while it guarantees convergence, it suffers from convergence stagnation, settling into a high-variance “stagnation region” around

the optimum. This limitation prevents the algorithm from achieving the optimal convergence rate.

To overcome this, we drew inspiration from the seminal work of Polyak and Juditsky [15], on stochastic approximation. We proposed integrating their classical averaging technique, which is known to effectively cancel stochastic noise and accelerate convergence.

Our primary contribution is the “tailed-average” mechanism. We demonstrated, both theoretically and empirically, that a naive application of averaging from the first iteration fails, as the estimate is “poisoned” by the initial, high-error transient iterates. Our TA-SGD algorithm intelligently circumvents this by introducing a “burn-in” period (K_{burn}), allowing the primary iterate to first settle into a stationary region. Only then does averaging begin, ensuring that only high-quality iterates contribute to the final solution.

Our rigorous theoretical analysis established that this two-phase approach achieves the best of both worlds:

1. It preserves the strong almost sure convergence guarantees of the underlying relaxed barrier method [4].
2. It successfully breaks the noise floor to achieve the optimal $O(1/K)$ convergence rate established by Polyak and Juditsky [15].

Our numerical experiments provided a clear validation of these theoretical claims. The results demonstrated that TA-SGD decisively overcomes the stagnation of the baseline algorithm, achieving significantly faster convergence and a final solution of substantially higher precision.

Future Work

The promising results of our TA-SGD algorithm open several interesting avenues for future research. Based on our findings, we suggest the following directions:

- **Theoretical Extensions to General Convexity:** Our convergence rate analysis successfully established the optimal $O(1/K)$ rate, but it relies on the standard assumption of θ -strong convexity. A valuable theoretical extension would be to analyze the convergence rate of TA-SGD for general convex (i.e., nonstrongly convex) objective functions, following recent advances in analyzing SGD for nonconvex functions satisfying weaker conditions such as the Polyak-Lojasiewicz property [10]

- **Synergy with Nonuniform Sampling:** Our algorithm currently employs standard uniform sampling for both objectives and constraints. Future work could explore the synergy between Tailed-Averaging and nonuniform or importance sampling techniques. For instance, adaptively sampling constraints that are more frequently violated might accelerate the burn-in phase, allowing the algorithm to reach the stationary region (and thus begin averaging) more quickly.

A Justification of linear decomposition (Step 3)

This appendix justifies the decomposition $\sqrt{K}\bar{r}_K^L = I^{(1)} + I^{(2)} + I^{(3)}$ used in Step 3. This derivation follows the algebraic steps of Lemma 2 in the appendix of [15], which results in the precise decomposition shown in Equation (A10) of that work. The linearized process is $r_k^L = (I - \gamma_{k-1}H)r_{k-1}^L - \gamma_{k-1}\eta_{k-1}$. By unrolling this recurrence from r_0 , we get the explicit solution:

$$r_k^L = X_0^k r_0 - \sum_{j=0}^{k-1} X_{j+1}^k \gamma_j \eta_j,$$

where $X_j^k = \prod_{i=j}^{k-1} (I - \gamma_i H)$ is the state transition matrix.

The average $\bar{r}_K^L$ is $\frac{1}{K} \sum_{k=0}^{K-1} r_k^L$:

$$\bar{r}_K^L = \left(\frac{1}{K} \sum_{k=0}^{K-1} X_0^k \right) r_0 - \frac{1}{K} \sum_{k=0}^{K-1} \sum_{j=0}^{k-1} X_{j+1}^k \gamma_j \eta_j.$$

We apply summation by parts (changing the order of the double summation):

$$\bar{r}_K^L = \left(\frac{1}{K} \sum_{k=0}^{K-1} X_0^k \right) r_0 - \frac{1}{K} \sum_{j=0}^{K-1} \left(\sum_{k=j+1}^{K-1} X_{j+1}^k \gamma_j \right) \eta_j.$$

The core insight (from [15]) is that the inner sum (the linear accumulation) can be approximated by H^{-1} .^{***} We define the residual w_j^K :

$$\sum_{k=j+1}^{K-1} X_{j+1}^k \gamma_j = H^{-1} - w_j^K.$$

^{***} This approximation is derived from the geometric series sum. In the idealized case where the step-size is constant ($\gamma_k = \gamma$), the inner sum $\sum_{k=j+1}^{K-1} X_{j+1}^k \gamma_j$ simplifies to $\gamma \sum_{n=0}^{K-j-2} (I - \gamma H)^n$. As $K \rightarrow \infty$, this sum converges to $\gamma [I - (I - \gamma H)]^{-1} = \gamma [\gamma H]^{-1} = H^{-1}$. The term w_j^K defined next thus captures the residual error from this ideal approximation, which arises because the step-sizes γ_k are not constant.

Substituting this back and defining $\theta_K = \gamma_0 \sum X_0^k$, we have

$$\begin{aligned}\bar{r}_K^L &= \left(\frac{1}{\gamma_0 K} \theta_K \right) r_0 - \frac{1}{K} \sum_{j=0}^{K-1} (H^{-1} - w_j^K) \eta_j \\ &= \frac{1}{\gamma_0 K} \theta_K r_0 - \frac{1}{K} \sum_{j=0}^{K-1} H^{-1} \eta_j + \frac{1}{K} \sum_{j=0}^{K-1} w_j^K \eta_j.\end{aligned}$$

Multiplying the entire equation by $\sqrt{K}$ yields the desired decomposition:

$$\sqrt{K} \bar{r}_K^L = \underbrace{\frac{1}{\sqrt{K} \gamma_0} \theta_K r_0}_{I^{(1)}} + \underbrace{\frac{1}{\sqrt{K}} \sum_{j=0}^{K-1} H^{-1} \eta_j}_{I^{(2)}} + \underbrace{\frac{1}{\sqrt{K}} \sum_{j=0}^{K-1} w_j^K \eta_j}_{I^{(3)}}.$$

The condition $\theta \in (0.5, 1)$ (Assumption C) is crucial to prove that $\frac{1}{K} \sum \|w_j^K\| \rightarrow 0$, which ensures $I^{(3)}$ vanishes.[†]

B Justification of asymptotic equivalence (Step 2)

This appendix justifies why the nonlinear terms $\Delta(r_k)$ can be ignored when averaging. We need to show that $\sqrt{K} \|\bar{r}_K - \bar{r}_K^L\| \rightarrow 0$.

This difference is proportional to the scaled average of the nonlinear residuals:

$$\text{Difference} \propto \frac{1}{\sqrt{K}} \sum_{k=K_{burn}}^K \gamma_k \Delta(r_k).$$

To prove this scaled average vanishes, we apply a standard result from stochastic approximation (see [15]), which is a direct consequence of Kronecker's Lemma. This lemma effectively transforms our original problem (showing the scaled average converges to zero) into an equivalent, simpler task: proving that a related sum is finite. Using Assumption (B), $\|\Delta(r_k)\| \leq K_1 \|r_k\|^{1+\lambda}$, this task reduces to proving that the following sum S converges:

$$S = \sum_{k=1}^{\infty} \frac{\|r_k\|^{1+\lambda}}{\sqrt{k}} < \infty.$$

[†] This specific range represents a critical balance. If $\theta \leq 0.5$, the step-size γ_k decays too slowly, causing the approximation error w_j^K to be too large on average for its mean to vanish. Conversely, if $\theta \geq 1$, γ_k decays so quickly that its sum becomes finite ($\sum \gamma_k < \infty$), which violates the primary convergence condition (a) of Theorem 1. This “sweet spot” ensures the error term $I^{(3)}$ vanishes while still guaranteeing convergence.

To prove S converges, we test the convergence of its expectation, $\mathbb{E}[S]$. To do this, we first need the convergence rate of the primary (nonaveraged) iterate r_k .

While the formal proof requires a detailed Lyapunov analysis (as shown in [15]), this $O(k^{-\theta})$ rate is the expected standard convergence rate for an SGD iterate that uses a step-size of $\gamma_k \propto k^{-\theta}$.

From the analysis in [15], we can use this established bound:

$$\mathbb{E}[\|r_k\|^2] \propto \mathbb{E}[V_k] \leq K_c \gamma_k \leq K_c k^{-\theta} \quad (\text{for large } k).$$

We apply Jensen's Inequality using the function $f(x) = x^{(1+\lambda)/2}$. This function is convex for $x \geq 0$ because the exponent $p = (1 + \lambda)/2 \geq 1$ (since $\lambda \geq 1$). Applying the inequality yields:

$$\begin{aligned} \mathbb{E}[\|r_k\|^{1+\lambda}] &= \mathbb{E}[(\|r_k\|^2)^{(1+\lambda)/2}] \leq \left(\mathbb{E}[\|r_k\|^2]\right)^{(1+\lambda)/2} \\ &\leq (K_c \gamma_k k^{-\theta})^{(1+\lambda)/2}. \end{aligned}$$

Now, we test the convergence of $\mathbb{E}[S]$:

$$\begin{aligned} \mathbb{E}[S] &= \sum_{k=1}^{\infty} \frac{\mathbb{E}[\|r_k\|^{1+\lambda}]}{\sqrt{k}} \leq \sum_{k=1}^{\infty} \frac{(K_c \gamma_k k^{-\theta})^{(1+\lambda)/2}}{k^{1/2}} \\ &= C \sum_{k=1}^{\infty} \frac{k^{-\theta(1+\lambda)/2}}{k^{1/2}} \\ &= C \sum_{k=1}^{\infty} k^{-\left(\frac{\theta(1+\lambda)}{2} + \frac{1}{2}\right)}. \end{aligned}$$

This is a p -series, which converges if and only if the exponent $p > 1$:

$$\begin{aligned} p &= \frac{\theta(1+\lambda)}{2} + \frac{1}{2} > 1, \\ \frac{\theta(1+\lambda)}{2} &> \frac{1}{2} \implies \theta(1+\lambda) > 1. \end{aligned}$$

This condition, $\theta(1 + \lambda) > 1$, is not arbitrary; it is the exact requirement imposed by [15] in their Assumptions 3.4 and 4.7. As noted in their commentary following Assumption 3.4, for the standard $\mathcal{C}^2$ case where $\lambda = 1$, this implies $2\theta > 1$ or $\theta > 0.5$. This confirms that our Assumption (C) aligns with the conditions required by the original proof.

Since $\mathbb{E}[S] < \infty$, the sum S converges almost surely. By Kronecker's Lemma, the convergence of this sum implies that the scaled average vanishes. This rigorously justifies Step 2.

C Derivation of the adaptive burn-in heuristic

In this appendix, we provide the formal derivation for the heuristic estimate of K_{burn} presented in section 2.2. The goal is to determine the number of iterations k required for the influence of the initial condition x_0 to become negligible compared to the stochastic noise.

Under Assumption 1 (strong convexity with parameter θ and L -smoothness), the expected distance to the optimum at iteration k satisfies the standard recurrence relation [1]:

$$\mathbb{E}[\|x_k - x^*\|^2] \leq (1 - 2\gamma_k\theta)\mathbb{E}[\|x_{k-1} - x^*\|^2] + \gamma_k^2\sigma^2.$$

During the transient “burn-in” phase, the error is dominated by the initial distance $\|x_0 - x^*\|^2$ rather than the variance term σ^2 . Focusing on the contraction of this initial bias with a constant step-size approximation γ , we have:

$$\|x_k - x^*\|^2 \approx (1 - \gamma\theta)^k \|x_0 - x^*\|^2.$$

We seek the smallest k such that the initial error is reduced to a tolerance ϵ_{init} :

$$(1 - \gamma\theta)^k \leq \epsilon_{init}.$$

Taking the natural logarithm of both sides:

$$k \ln(1 - \gamma\theta) \leq \ln(\epsilon_{init}).$$

Using the Taylor expansion $\ln(1 - x) \approx -x$ for small $x = \gamma\theta$, we obtain:

$$k(-\gamma\theta) \lesssim -\ln\left(\frac{1}{\epsilon_{init}}\right) \implies k \gtrsim \frac{1}{\gamma\theta} \ln\left(\frac{1}{\epsilon_{init}}\right).$$

For stability in smooth optimization, the step size is typically bounded by $\gamma \leq \frac{1}{L}$. Substituting $\gamma \approx \frac{1}{L}$:

$$k \gtrsim \frac{L}{\theta} \ln\left(\frac{1}{\epsilon_{init}}\right).$$

Defining the condition number as $\kappa = L/\theta$, we recover the heuristic:

$$K_{burn} = \mathcal{O}\left(\kappa \log\left(\frac{1}{\epsilon_{init}}\right)\right).$$

This confirms that the required length of the burn-in period scales linearly with the condition number of the problem.

D Experimental details

This appendix provides the specific implementation and parameter details used to generate the numerical results in section 4, ensuring full reproducibility. All experiments were conducted in Python using the `NumPy` library for numerical operations and `Matplotlib` for visualization.

D.1 Problem generation

The experiments were performed on the quadratic programming (QP) problem introduced by Dimitrieski, Cao, and Ebenbauer [4], as described in Section 4.1. The problem was generated once using a master seed (`master_seed = 42`) with the following parameters:

- Dimension: $d = 50$
- Objective components: $n = 10$
- Number of constraints: $m = 10,000$
- Objective parameter: $\gamma = 2.365$.[‡]

The objective function coefficients $\theta \in \mathbb{R}^{n \times d}$ were sampled from $\mathcal{U}(0.1, 1.1)$. The constraint gradients $A \in \mathbb{R}^{m \times d}$ were generated by first defining a diagonal matrix Q with entries sampled from $\mathcal{U}(1, 1.5)$, and then sampling m points on the surface of the ellipsoid defined by Q to form the constraint matrix A .

D.2 Algorithm implementation and hyperparameters

We implemented the three algorithms (Base SGD, Full Averaging, and TA-SGD) as described in Section 4.1.

[‡] Following the benchmark design in [4], these parameters are calibrated such that the unconstrained global minimizer $x_f^* = \arg \min f(x)$ satisfies $\|x_f^*\|_2 = 15$. This guarantees that $x_f^* \notin \mathcal{E}$, ensuring that the constraints are active at the optimal solution.

- **Total iterations:** All algorithms were run for $K = 10,000$ iterations.
- **Starting point:** All runs were initialized at the “Far Start Point” $x_0 = [10, \dots, 10]^T \in \mathbb{R}^{50}$.
- **Schedules:** The step-size and barrier parameter schedules followed the configuration in [4]:

$$\gamma_k = 0.3 \cdot k^{-0.8} \quad \text{and} \quad \epsilon_k = 5.0 \cdot k^{-0.3}$$

with the final barrier parameter set to $\delta_\infty = 10^{-6}$.

D.3 Reproducibility and analysis

- **Reference solution (x^*):** The optimal solution $x^* = x^*(\delta_\infty)$ was pre-computed using a *deterministic full-gradient descent method*. Unlike the stochastic approximations, this solver utilized the full dataset at each iteration (computing the exact full gradient $\nabla f(x)$ and $\nabla B(x)$ without sampling). This deterministic process was executed for $K_{long} = 50,000$ iterations to ensure convergence to a high-precision ground truth, completely independent of the stochastic algorithms being evaluated.
- **Statistical analysis:** The results in Figure 2 and Table 1 were generated by averaging over $R = 1,000$ independent runs.
- **Seeding:** To ensure statistically independent yet reproducible trials, the main problem was generated with `master_seed = 42`, and each of the $R = 1,000$ runs (indexed by $r \in [0, \dots, 999]$) was seeded with `master_seed + r + 2`.
- **Metrics:** The “Error Norm” (Figure 2) corresponds to the mean of $\|x_k - x^*\|_2$ (or $\|\bar{x}_k - x^*\|_2$) at each iteration k across the R runs. The shaded region represents ± 1 standard deviation. The “Mean Final Error” (Table 1) is the mean and standard deviation of this error norm specifically at the final iteration, $k = 10,000$.

Funding

This research received no external funding.

Conflicts of Interest

The authors declare no conflicts of interest.

References

- [1] Bottou, L., Curtis, F.E. and Nocedal, J. *Optimization methods for large-scale machine learning*, SIAM Rev. 60(2) (2018), 223–311.
- [2] Boyd, S. and Vandenberghe, L. *Convex optimization*, Cambridge University Press, 2004.
- [3] Bubeck, S. *Convex optimization: Algorithms and complexity*, Found. Trends Mach. Learn. 8(3-4) (2015), 231–358.
- [4] Dimitrieski, N., Cao, J. and Ebenbauer, C. *Stochastic gradient descent for constrained optimization based on adaptive relaxed barrier functions*, Int. J. Robust. Nonlinear Control. (2025).
- [5] Feller, C. and Ebenbauer, C. *A stabilizing iteration scheme for model predictive control based on relaxed barrier functions*, Automatica 80 (2017), 328–339.
- [6] Fiacco, A.V. and McCormick, G.P. *Nonlinear programming: Sequential unconstrained minimization techniques*, John Wiley & Sons, 1968.
- [7] Garrigos, G. and Gower, R.M. *Handbook of convergence theorems for (stochastic) gradient methods*, arXiv preprint arXiv:2301.11235 (2023).
- [8] Gower, R.M., Loizou, N., Qian, X., Sailanbayev, A., Shulgin, E. and Richtárik, P. *SGD: General analysis and improved rates*, arXiv preprint arXiv:1901.09401 (2019).
- [9] Hardy, G.H. *Divergent series*, Oxford University Press, 1949.
- [10] Karandikar, R.L. and Vidyasagar, M. *Convergence rates for stochastic approximation: biased noise with unbounded variance, and applications*, arXiv preprint arXiv:2312.02828 (2024).
- [11] Li, M., Grigas, P. and Atamtürk, A. *New penalized stochastic gradient methods for linearly constrained strongly convex optimization*, SIAM J. Optim. 32(2) (2022), 859–887.

- [12] Moulines, E. and Bach, F. *Non-asymptotic analysis of stochastic approximation algorithms*, In Advances in Neural Information Processing Systems (NeurIPS), 451–459, 2011.
- [13] Nocedal, J. and Wright, S.J. Numerical optimization, 2nd ed., Springer, 2006.
- [14] Pflug, G.C. *Non-asymptotic confidence bounds for stochastic approximation algorithms with constant step size*, Monatsh. Math. 110(3-4) (1990), 297–314.
- [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, 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, 1–8, 2017.
- [20] Rudin, W. *Principles of mathematical analysis*, 3rd ed., McGraw-Hill, 1976.
- [21] Wang, M. and Bertsekas, D.P. *Incremental constraint projection methods for variational inequalities*, Math. Program. 150 (2015), 321–363.
- [22] Wang, X., Ma, S. and Yuan, Y. *Penalty methods with stochastic approximation for stochastic nonlinear programming*, arXiv preprint arXiv:1605.05609 (2016).
- [23] 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.
- [24] Zhang, T. *Solving large scale linear prediction problems using stochastic gradient descent algorithms*, In Proc. 21st Int. Conf. Mach. Learn. (ICML), 116, 2004.

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. 2, pp 339-756

A solution approach to the multilevel linear fractional programming problems	339
A. Baghalnezhad and H. Sadeghi	
A mathematical model and optimal control analysis of merchant resistance to e-commerce adoption	356
H. Hassouni, S. Belhdid, O. Aarabate and O. Balatif	
Fractal-fractional dynamics of HIV/AIDS: Unveiling the synergistic impact of condoms, prophylaxis, and antiretroviral interventions for epidemic control	385
G.O. Acheneje, J. Amos, E. Abah, W. Atokolo, B.C. Agbata, Shyamsunder, A.K. Awasthi and B. Bolaji	
Finite element analysis of cross-diffusion three-species Shigesada–Kawasaki–Teramoto model	447
G.A. Al-Musawi, H.A. Challob and A.J. Harfash	
Characterizing maximum output sets in fractional-order discrete-time linear systems utilizing fractional feedback control	497
A. EL Bhih	
Spline-interpolation solution of a plane Stefan problem	525
P.N. Ivanshin	
Local fractional Khalouta transform: Properties and applications	539
A. Khalouta	
A nonmonotone line search method for solving constrained multiobjective optimization problems	563
M. Rashidi and E. Khorram	
Design and implementation of a graph-coloring algorithm for optimizing flight-level allocation in air traffic management in Iran	590
S.A.A. Mosavi, A. Erfanian and N. Sabeghi	
Fitted-mesh cubic spline methods for layer-dominated singularly perturbed parabolic problems with nonsmooth data	607
S.L. Cheru, G.F. Duressa and T.B. Mekonnen	
Optimal control and a system of quasi-variational inequalities	638
J. Ogorzaly	
Mathematical model for an infectious disease with quarantine	667
F. Iranzad, H.R. Marasi, H. Kheiri and A. Jabbari	
Synchronization and control of discrete incommensurate fractional-order and variable-order neural networks	683
S. Momani, I.M. Batiha, A. Hiousal and A. Ouannas	

Mathematical analysis of a fractional-order influenza model with resistant and nonresistant strains	703
I. Aamoum, A. Meskaf, Z. Yaagoub and E.M. Farah	

Tailed-Average SGD: A Polyak–Ruppert modification for optimal-rate constrained optimization	727
A. Fillali, S. Addoune and R. Zeghdane	

web site: <https://ijnao.um.ac.ir>

Email: ijnao@um.ac.ir

ISSN-Print: 2423-6977

ISSN-Online: 2423-6969