



# *Iranian Journal of Numerical Analysis and Optimization*

Volume 15, Number 3

September 2025

Serial Number: 34

*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 15, Number 3, September 2025**

**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](mailto: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 15, Number 3, September 2025

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 Sci-  
ences, Sweden.

e-mail: Peter@Lohmander.com

**Lopez-Ruiz, Ricardo\***

(Complexity, nonlinear models)

University of Zaragoza, Spain.

e-mail: rilopez@unizar.es

**Mahdavi-Amiri, Nezam\***

(Optimization)

Sharif University of Technology, Iran.

e-mail: nezamm@sina.sharif.edu

**Mirzaei, Davoud\***

(Numerical Analysis)

University of Uppsala, Sweden.

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

**Omrani, Khaled\***

(Numerical Analysis)

University of Tunis El Manar, Tunisia.

khaled.omrani@issatso.rnu.tn

**Salehi Fathabadi, Hasan\***

(Operations Research )

University of Tehran, Iran.

e-mail: hsalehi@ut.ac.ir

**Soheili, Ali Reza\***

(Numerical Analysis)

Ferdowsi University of Mashhad, Iran.

e-mail: soheili@um.ac.ir

**Soleimani Damaneh, Majid\***

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

University of Tehran, Iran.

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

**Toutounian, Faezeh\***

(Numerical Analysis)

Ferdowsi University of Mashhad, Iran.

e-mail: toutouni@um.ac.ir

**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

<b>The analysis of the mathematical stability of a cholera disease model . . . . .</b>	<b>852</b>
I. Sahib, M. Baroudi, H. Gourram, B. Khajji, A. Labzai and M. Belam	
<b>Two-step inertial Tseng’s extragradient methods for a class of bilevel split variational inequalities . . . . .</b>	<b>877</b>
L.H.M. Van and T.V. Anh	
<b>Approximate symmetries of the perturbed KdV-KS equation</b>	<b>914</b>
A. Mohammadpouri, M.S. Hashemi, R. Abbasi and R. Abbasi	
<b>Convergence analysis of triangular and symmetric splitting method for fuzzy stochastic linear systems . . . . .</b>	<b>930</b>
B. Harika, D. Rajaiah, A. Shivaji, and L.P. Rajkumar	
<b>Mathematical modeling of COVID-19 spread with media coverage and optimal control analysis . . . . .</b>	<b>952</b>
G.P. Sahu and A.S. Thakur	
<b>Space-time localized scheme to solve some partial integro-differential equations . . . . .</b>	<b>993</b>
M. Hamaidi, M. Briki, A. Nouara and B. Hamdi	
<b>A study on the convergence and error bound of solutions to 2D mixed Volterra–Fredholm integral and integro-differential equations via high-order collocation method . . . . .</b>	<b>1012</b>
A.A. Shalangwa, M.R. Odekunle and S.O. Adeo	
<b>Cutting-edge spectral solutions for differential and integral equations utilizing Legendre’s derivatives . . . . .</b>	<b>1036</b>
A.M. Abbas, Y.H. Youssri, M. El-Kady and M. Abdelhakem	
<b>Mathematical modeling of Echinococcosis in humans, dogs and livestock with optimal control strategies . . . . .</b>	<b>1075</b>
I. Sannaky, M. Riouali, N. Ouldkhouia, I. El berrai, and K. Adnaoui	
<b>A new generalized model of cooperation of advertising companies based on differential games on networks . . . . .</b>	<b>1116</b>
M. Jashnesade, Z. Nikoeeinejad and G B. Loghmani	
<b>Mathematical modeling and optimal control strategies to limit cochineal infestation on cacti plants . . . . .</b>	<b>1145</b>
K. Sofiane and B. Omar	
<b>An efficient Dai-Kou-type method with image de-blurring application . . . . .</b>	<b>1171</b>
K. Ahmed, M.Y. Waziri, S. Murtala, A.S. Halilu, H. Abdullahi and Y.B. Musa	

<b>Combining the reproducing kernel method with Taylor series expansion to solve systems of nonlinear fractional Volterra integro-differential equations . . . . .</b>	<b>1210</b>
T. Amoozad, S. Abbasbandy, H. Sahihi, T. Allahviranloo	
<b>Convex-hull based two-phase algorithm to solve capacitated vehicle routing problem . . . . .</b>	<b>1241</b>
M. Afsharirad and A. Hashemi Borzabadi	
<b>Accurate ENO-like schemes for the model of fluid flows in a nozzle with variable cross-section . . . . .</b>	<b>1275</b>
D.H. Cuong and M.D. Thanh	



## The analysis of the mathematical stability of a cholera disease model

I. Sahib\*, M. Baroudi, H. Gourram, B. Khajji, A. Labzai and M. Belam

---

\*Corresponding author

Received 26 June 2024; revised 13 September 2024; accepted 20 February 2025

Issam Sahib

Laboratory LMACS, Sultan Moulay Slimane University, MATIC Research Team: Applied Mathematics and Information and Communication Technologie, Department of Mathematics and Computer Science, Khouribga Polydisciplinary Faculty, Morocco.. e-mail: [sahibissam@gmail.com](mailto:sahibissam@gmail.com)

Mohamed Baroudi

Laboratory LMACS, Sultan Moulay Slimane University, MATIC Research Team: Applied Mathematics and Information and Communication Technologie, Department of Mathematics and Computer Science, Khouribga Polydisciplinary Faculty, Morocco. e-mail: [m.mohamed.baroudi@gmail.com](mailto:m.mohamed.baroudi@gmail.com)

Hicham Gourram

Laboratory LMACS, Sultan Moulay Slimane University, MATIC Research Team: Applied Mathematics and Information and Communication Technologie, Department of Mathematics and Computer Science, Khouribga Polydisciplinary Faculty, Morocco. e-mail: [gourramhicham03@gmail.com](mailto:gourramhicham03@gmail.com)

Bouchaib Khajji

Laboratory of Analysis Modeling and Simulation, Department of Mathematics and Computer Science, Faculty of Sciences Ben M'sik, Hassan II University of Casablanca, Morocco. e-mail: [labzaiabdo1977@gmail.com](mailto:labzaiabdo1977@gmail.com)

Abderrahim Labzai

Laboratory of Analysis Modeling and Simulation, Department of Mathematics and Computer Science, Faculty of Sciences Ben M'sik, Hassan II University of Casablanca, Mo-

### Abstract

In this study, we develop a deterministic model for cholera transmission dynamics, incorporating vaccination campaigns, treatment of infected individuals, and water sanitation initiatives. A novel feature of our model is the inclusion of healthcare centers, which enhances the simulation of treatment dynamics, offering new insights into cholera management. The model's central metric is the basic reproduction number  $R_0$ , derived from the disease-free equilibrium (DFE) condition. Stability analysis shows that when  $R_0 \leq 1$ , the DFE is asymptotically stable, ensuring cholera eradication, while  $R_0 > 1$  leads to an endemic equilibrium. Sensitivity analysis highlights that vaccination, treatment, sanitation, and public awareness campaigns are critical for reducing  $R_0$ . The inclusion of healthcare centers further improves the model's effectiveness by ensuring timely treatment. Numerical simulations, validated using *MATLAB*, confirm that comprehensive public health strategies, including expanded vaccination campaigns and healthcare infrastructure, are essential for combating cholera outbreaks. This model underscores the importance of timely medical intervention in reducing infection rates and fatalities.

**AMS subject classifications (2020):** 49J15, 93C10, 92B05, 93A30.

**Keywords:** stability; sensitivity; optimal control; Disease cholera.

## 1 Introduction

Cholera remains a significant global public health threat, responsible for tens of thousands of deaths each year [22]. This highly contagious disease, caused

---

rocco. e-mail: khajjibouchaib@gmail.com

Mohamed Belam

Laboratory LMACS, Sultan Moulay Slimane University, MATIC Research Team: Applied Mathematics and Information and Communication Technologie, Department of Mathematics and Computer Science, Khouribga Polydisciplinary Faculty, Morocco. e-mail: m.belam@gmail.com

### How to cite this article

Sahib, I., Baroudi, M., Gourram, H., Khajji, B., Labzai, A. and Belam, M., The analysis of the mathematical stability of a cholera disease model. *Iran. J. Numer. Anal. Optim.*, 2025; 15(3): 852-876. <https://doi.org/10.22067/ijnao.2025.88685.1464>

by the bacterium *Vibrio cholerae*, spreads rapidly in communities through contaminated water and food sources. The transmission dynamics of cholera within a community are influenced by a complex interaction of social, environmental, and behavioral factors. In certain regions, the periodic recurrence of cholera outbreaks can be attributed in part to seasonal variations in contact rates, water quality, and sanitation levels [6, 15]. Understanding and accurately estimating the prevalence of *Vibrio cholerae* infections in endemic populations, as well as the correlation between the concentration of the bacteria and its virulence, is crucial to controlling the spread of the disease and mitigating its impact [3]. These seasonal fluctuations are particularly important in explaining the cyclical nature of cholera outbreaks [8].

Mathematical models have played a crucial role in capturing these dynamics. In 2001, scientists enhanced Capasso's model by incorporating the environmental component, specifically the concentration of *Vibrio cholerae* in the water supply, into a basic SIR (Susceptible-Infected-Recovered) model. This modification allowed for a better understanding of how the presence of the bacteria in water sources contributes to the incidence of cholera. The saturation effect of bacterial concentration was modeled using a logistic function, reflecting the nonlinear relationship between bacterial load and infection risk.

Further advancements in modeling cholera transmission were made by Hartley [10], who introduced a hyper-infectious stage for *Vibrio cholerae*. This addition, based on observations from laboratory settings, captured the highly transmissible nature of recently shed *Vibrio cholerae*, which is particularly potent immediately after being excreted by infected individuals [8]. This feature significantly enhances the pathogen's ability to spread during an outbreak, emphasizing the need for rapid and targeted interventions to control the transmission of this highly contagious form. Hartley's work underscored the importance of considering pathogen dynamics when designing intervention strategies. It is imperative to explore whether other common infectious diseases also exhibit such hyper-transmissible stages and, if so, to incorporate these stages into their respective prevention models to ensure a more comprehensive and effective response.

Nelson et al. [17] further refined these models by incorporating a more accurate representation of the pathogen's infectious dose, recognizing that the minimal dose required to cause an infection plays a critical role in determining how quickly an outbreak can spread. Interventions that focus on reducing exposure to the hyper-infectious form of *Vibrio cholerae*, such as improved sanitation and clean water distribution, are essential for preventing large-scale outbreaks. Moreover, evaluating other diseases for similar hyper-infectious conditions and integrating such findings into disease prevention models will allow for more targeted and efficient public health interventions [19].

Raising public awareness through campaigns and educational initiatives has proven to be an effective strategy for controlling the spread of infectious diseases. By reducing the likelihood of contact transmission among vulnerable populations, awareness campaigns play a pivotal role in managing epidemics. In the digital age, the rapid dissemination of information through social media, coupled with increased global travel, has made awareness even more critical. These campaigns can significantly decrease the probability of transmission by educating the public on hygiene practices and the importance of early detection and treatment, ultimately improving the overall dynamics of epidemics [1, 16, 25].

The relationship between the spread of infectious diseases and human social behavior has been extensively studied in both theoretical and empirical research. Numerous mathematical models have been developed to explore these interactions, particularly in the context of cholera [4, 5, 9, 11, 13, 16, 20]. These models have helped public health officials design strategies to reduce the number of cholera cases and improve overall health outcomes in affected communities [18].

In this study, we delve deeply into the foundational mathematical components of cholera models, focusing on critical aspects such as the determination of equilibrium points and the calculation of the epidemic threshold, commonly referred to as the basic reproductive number  $R_0$ . The stability of these equilibrium points is rigorously analyzed, revealing the conditions under which the disease-free equilibrium (DFE) is globally asymptotically stable (GAS). Descartes' rule of signs is applied to derive global stability conditions, and

the local stability of the endemic equilibrium is evaluated using the center manifold theory [15]. These analyses provide valuable insights into the underlying dynamics of cholera transmission and the potential for controlling outbreaks through targeted interventions.

One critical aspect that has not been fully explored in many models is the role of healthcare infrastructure in managing cholera outbreaks. Integrating healthcare centers into the model is essential for accurately capturing the real-world dynamics of disease transmission and control. Healthcare centers are often the first line of defense during an outbreak, providing immediate treatment to those infected and acting as central points for public health interventions such as vaccination, water sanitation programs, and public awareness campaigns. By incorporating healthcare facilities into the model, we can simulate more realistic outbreak scenarios and assess the impact of different intervention strategies in a variety of contexts.

The inclusion of healthcare centers also allows for the optimization of resource allocation during an outbreak. For instance, the model can help determine the most efficient distribution of medical supplies and personnel across different regions, ensuring that healthcare resources are concentrated in areas where they will have the greatest impact. This addition not only enhances the accuracy and realism of the model but also provides public health officials with a powerful tool for decision-making in the midst of an outbreak. The importance of this integration cannot be overstated, as it bridges the gap between theoretical models and practical applications in public health policy.

In conclusion, this study builds upon existing cholera models by not only refining the mathematical understanding of cholera dynamics but also by emphasizing the critical role of healthcare infrastructure in controlling outbreaks. By combining rigorous mathematical analysis with practical considerations of public health intervention, this model provides a more comprehensive and applicable tool for managing cholera outbreaks and potentially other infectious diseases as well.

The paper is organized as follows: A mathematical model and its basic properties are presented in Section 2. The features of the model's local and global asymptotic stability are examined in Section 3. The sensitivity of the basic reproduction number concerning the model parameters is investigated

in Section 4. Numerical simulations and discussions are provided in Section 5. The paper is finally concluded in Section 6.

## 2 Fundamental properties and the mathematical model

### 2.1 A mathematical model

In the context of cholera, we introduce a continuous dynamics model of the SICR-B (Susceptible-Infectious-Centers-Recovered-Bacterial) type, which includes a category for bacterial concentration. The total population,  $N(t)$ , is divided into four classes: susceptible individuals  $S(t)$ , infected individuals  $I(t)$  exhibiting symptoms, individuals undergoing treatment in centers  $C(t)$ , and recovered individuals  $R(t)$ . The total population at time  $t$  is given by  $N(t) = S(t) + I(t) + C(t) + R(t)$ . The graphical representation of this model is shown in Figure 1.

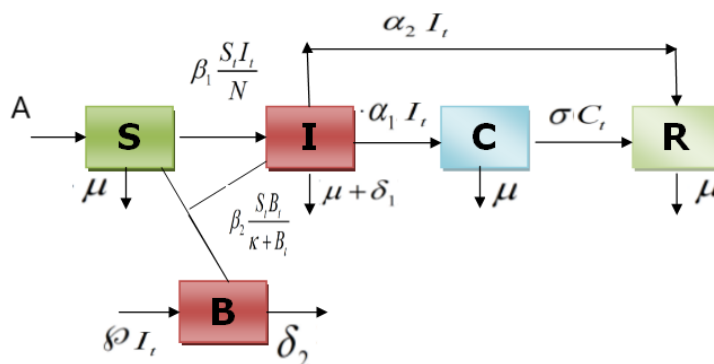


Figure 1: The dynamics among the five compartments SICR-B of cholera disease.

We study five nonlinear differential equations:

$$\begin{cases} \frac{dS(t)}{dt} = A - \mu S - \beta_1 \frac{SI}{N} - \beta_2 \frac{SB}{\kappa + B}, \\ \frac{dI(t)}{dt} = \beta_1 \frac{SI}{N} - I(\mu + \delta_1 + \alpha_1 + \alpha_2) + \beta_2 \frac{SB}{\kappa + B}, \\ \frac{dC(t)}{dt} = \alpha_1 I - (\sigma + \mu)C, \\ \frac{dR(t)}{dt} = \alpha_2 I + \sigma C - \mu R, \\ \frac{dB(t)}{dt} = \wp I - \delta_2 B. \end{cases} \quad (1)$$

The initial states are given as  $S(0) \geq 0$ ,  $I(0) \geq 0$ ,  $C(0) \geq 0$ ,  $R(0) \geq 0$ , and  $B(0) \geq 0$ . The total population  $N(t)$  at time  $t > 0$  is categorized into four classes: Susceptible individuals  $S(t)$ , infectious individuals  $I(t)$  showing symptoms, individuals undergoing treatment in centers  $C(t)$ , and recovered individuals  $R(t)$ .

Additionally, we introduce a class  $B(t)$  representing bacterial concentration at time  $t$ . We assume a positive recruitment rate  $A$  into the susceptible class  $S(t)$  and a positive natural death rate  $\mu$  for all time  $t$ . Susceptible individuals can contract cholera at a rate  $\beta_2 \frac{B(t)}{\kappa + B(t)}$ , where  $\beta_2 > 0$  is the ingestion rate of bacteria from contaminated sources,  $\kappa$  is the half-saturation constant of the bacteria population, and  $\frac{B(t)}{\kappa + B(t)}$  represents the probability of infection given exposure.

Infected individuals can opt for treatment in centers for a period, where they are isolated and receive appropriate medication at rates  $\alpha_1$  and  $\alpha_2$ . Recovery from treatment occurs at rate  $\sigma$ . Disease-related death rates for infected individuals undergoing treatment and those not in treatment are  $\delta_1$  and  $\mu$ , respectively.

Each infected individual contributes to an increase in bacterial concentration at rate  $\wp$ , while the bacterial concentration decreases due to mortality at rate  $\delta_2$ .

## 2.2 Fundamental characteristics of the model

### 2.2.1 Region of invariance

It is necessary to demonstrate that all solutions of system (1) starting from positive initial values will remain positive for all  $t > 0$ . This will be established through the following lemma.

**Lemma 1.** All admissible solutions  $S(t), I(t), C(t), R(t)$ , and  $B(t)$  of system (1) are bounded within the region  $\Omega = \Omega_N * \Omega_B$ , where

$$\begin{cases} \Omega_N = \left\{ (S, I, C, R) \in \mathbb{R}_+^4 : S + I + C + R \leq \frac{A}{\mu} \right\}, \\ \Omega_B = \left\{ B \in \mathbb{R}_+ : B \leq \frac{\wp}{\delta_2} \right\}. \end{cases} \quad (2)$$

*Proof.* From the equation of system (1)

$$\frac{dN(t)}{dt} = A - \mu N(t) - I\delta_1, \quad (3)$$

implies the following equation:

$$\frac{dN(t)}{dt} \leq A - \mu N(t). \quad (4)$$

Therefore, it is clear that

$$N(t) \leq \frac{A}{\mu}(1 - e^{-\mu t}) + N(0)e^{-\mu t}. \quad (5)$$

Since  $N(0)$  is the initial value of the total number of people,

$$\lim_{t \rightarrow +\infty} \text{Sup} N(t) \leq \frac{A}{\mu}. \quad (6)$$

Then

$$S(t) + I(t) + C(t) + R(t) \leq \frac{A}{\mu}. \quad (7)$$

Similarly,

$$\frac{dB(t)}{dt} = \wp I - \delta_2 B(t) \leq \wp - \delta_2 B(t), \quad (8)$$

$$B(t) \leq \frac{\wp}{\delta_2} + B(0)e^{-\delta_2 t}, \quad (9)$$

$$\lim_{t \rightarrow +\infty} \text{Sup} B(t) \leq \frac{\wp}{\delta_2}, \quad (10)$$

$$[B(t) \leq \frac{\rho}{\delta_2}. \quad (11)$$

For the analysis of model (1), we get the regions, which is given by the set  $\Omega = \Omega_N * \Omega_B$ , where

$$\begin{cases} \Omega_N = \left\{ (S, I, C, R) \in \mathbb{R}_+^4 : S + I + C + R \leq \frac{A}{\mu} \right\}, \\ \Omega_B = \left\{ B \in \mathbb{R}_+ : B \leq \frac{\rho}{\delta_2} \right\}, \end{cases} \quad (12)$$

which is a positively invariant set for (1). Therefore, it is only necessary to consider the dynamics of the system (1) in relation to the set of nonnegative solutions  $\Omega$ .  $\square$

### 2.2.2 Positivity of the model's solutions.

**Theorem 1.** If  $S(0) \geq 0$ ,  $I(0) \geq 0$ ,  $C(0) \geq 0$ ,  $R(0) \geq 0$ , and  $B(0) \geq 0$ , then the solutions of system equation (1),  $S(t)$ ,  $I(t)$ ,  $C(t)$ ,  $R(t)$ , and  $B(t)$  are positive for all  $t > 0$ .

*Proof.* Starting from the first equation of system (1), we obtain

$$\frac{dS(t)}{dt} = A - M(t)S(t). \quad (13)$$

Given that

$$M(t) = \mu + \beta_1 \frac{I(t)}{N} + \beta_2 \frac{B(t)}{\kappa + B(t)}, \quad (14)$$

We multiply (13) by  $\exp\left(\int_0^t M(s) ds\right)$ ; then we obtain

$$\frac{dS(t)}{dt} * \exp\left(\int_0^t M(s) ds\right) = [A - M(t) * S(t)] * \exp\left(\int_0^t M(s) ds\right), \quad (15)$$

$$\frac{dS(t)}{dt} * \exp\left(\int_0^t M(s) ds\right) + M(t) * S(t) * \exp\left(\int_0^t M(s) ds\right) = A * \exp\left(\int_0^t M(s) ds\right). \quad (16)$$

Therefore

$$\frac{d}{dt} [S(t) * \exp\left(\int_0^t M(s) ds\right)] = A * \exp\left(\int_0^t M(s) ds\right). \quad (17)$$

When we take the integral with respect to  $s$  from 0 to  $t$ , we obtain

$$S(t) * \exp\left(\int_0^t M(s) ds\right) - S(0) = A * \int_0^t \left(\exp\left(\int_0^w M(s) ds\right)\right) dw. \quad (18)$$

Multiplying (18) by  $\exp\left(-\int_0^t M(s) ds\right)$ , we obtain

$$\begin{aligned} S(t) - S(0) * \exp\left(-\int_0^t M(s) ds\right) \\ = A * \exp\left(-\int_0^t M(s) ds\right) * \int_0^t \left(\exp\left(\int_0^w M(s) ds\right)\right) dw. \end{aligned} \quad (19)$$

Then

$$\begin{aligned} S(t) = S(0) * \exp\left(-\int_0^t M(s) ds\right) \\ + A * \exp\left(-\int_0^t M(s) ds\right) * \int_0^t \left(\exp\left(\int_0^w M(s) ds\right)\right) dw \geq 0. \end{aligned}$$

Thus,  $S(t)$  is a positive solution. Similarly, based on the other equations in system (1), we obtain

$$I(t) \geq I(0) * \exp\left(-\int_0^t (\mu + \delta_1 + \alpha_1 + \alpha_2 - \beta_1 \frac{S(s)}{N}) ds\right) \geq 0, \quad (20)$$

$$\begin{cases} C(t) \geq C(0) \cdot \exp(-(\sigma + \mu)t) \geq 0, \\ R(t) \geq R(0) \cdot \exp(-(\alpha_2 + \mu)t) \geq 0, \\ B(t) \geq B(0) \cdot \exp(-\delta_2 t) \geq 0. \end{cases} \quad (21)$$

As a result, the proof is finished since we can see that for all  $t \geq 0$ , the solutions  $S(t)$ ,  $I(t)$ ,  $C(t)$ ,  $R(t)$ , and  $B(t)$  to the system (1) are positive. Since the variables  $C$  and  $R$  do not affect the first three equations in system (1), the dynamics of equation system (1) is equal to the dynamics of equation system:

$$\begin{cases} \frac{dS(t)}{dt} = A - \mu S - \beta_1 \frac{SI}{N} - \beta_2 \frac{SB}{\kappa + B}, \\ \frac{dI(t)}{dt} = \beta_1 \frac{SI}{N} - I(\mu + \delta_1 + \alpha_1 + \alpha_2) + \beta_2 \frac{SB}{\kappa + B}, \\ \frac{dB(t)}{dt} = \delta_1 I - \delta_2 B. \end{cases} \quad (22)$$

□

## 2.3 An examination of the model's sensitivity and stability.

### 2.3.1 Points of equilibrium:

There are two equilibrium points in this model: The DFE point, which occurs when cholera is absent, and the epidemic equilibrium point, which occurs when cholera is present. By setting the derivatives of the rate of change expressions in the (22) system to zero, these points can be determined.

In the absence of a virus ( $I = B = 0$ ), the Cholera DFE  $E_{eq}^0 = (\frac{A}{\mu}, 0, 0)$  is reached. When the disease exists ( $I \neq 0$  and  $B \neq 0$ ), the present equilibrium of the Cholera disease is reached, denoted by  $E_{eq}^* = (S^*; E^*; I^*)$ . To calculate the fundamental reproduction number  $R_0$ , we will apply the next-generation operator method.

**$R_0$  is the basic reproduction number.**

Diekmann et al. [2] defined the basic reproduction number ( $R_0$ ), which is an important indicator of the transmissibility of an infectious disease in epidemiology. In a population that is fully susceptible, it denotes the average number of secondary infections caused by one infected person. The mathematical method  $R_0$  is computed using the next-generation matrix approach [21].

The spectral radius of the product matrix  $FV^{-1}$  is denoted by the basic reproduction number  $R_0$ . In other words,  $R_0 = \rho(FV^{-1})$ , where the spectral radius is indicated by  $\rho$ .

We define  $F$  as a nonnegative matrix accounting for the new infective terms within the next-generation matrix approach framework. Comparably, both of the remaining transfer terms are represented by  $V$ , which is a non-singular  $M$ -matrix evaluated at the DFE. Then

$$F = \begin{pmatrix} \beta_1 \frac{A}{\mu N} & \beta_2 \frac{A}{\kappa \mu} \\ 0 & 0 \end{pmatrix}, \quad (23)$$

$$V = \begin{pmatrix} (\mu + \delta_1) + (\alpha_1 + \alpha_2) & 0 \\ -\varphi & \delta_2 \end{pmatrix}, \quad (24)$$

$$V^{-1} = \begin{pmatrix} \frac{1}{(\mu + \delta_1) + (\alpha_1 + \alpha_2)} & 0 \\ \frac{\varrho}{\delta_2[(\mu + \delta_1) + (\alpha_1 + \alpha_2)]} & \frac{1}{\delta_2} \end{pmatrix}, \quad (25)$$

$$F.V^{-1} = \begin{pmatrix} \beta_1 \frac{A}{\mu N[(\mu + \delta_1) + (\alpha_1 + \alpha_2)]} + \beta_2 \frac{A\varrho}{\delta_2 \kappa \mu[(\mu + \delta_1) + (\alpha_1 + \alpha_2)]} & \beta_2 \frac{A}{\kappa \mu \delta_2} \\ 0 & 0 \end{pmatrix}. \quad (26)$$

This suggests that  $R_0$ , the basic reproduction number, can be found by using the following relationships (where  $\rho$  is the spectral radius):

$$R_0 = \rho(F.V^{-1}), \quad (27)$$

$$R_0 = \beta_1 \frac{A}{\mu N(\mu + \delta_1 + \alpha_1 + \alpha_2)} + \beta_2 \frac{A\varrho}{\delta_2 \kappa \mu(\mu + \delta_1 + \alpha_1 + \alpha_2)}, \quad (28)$$

$$R_0 = \frac{A}{\mu(\mu + \delta_1 + \alpha_1 + \alpha_2)} \left[ \frac{\beta_1}{N} + \frac{\beta_2 \varrho}{\delta_2 \kappa} \right]. \quad (29)$$

The basic reproduction number, or  $R_0$ , measures the average number of newly infected people that are created in a population of susceptible people by a single infected person. Its value indicates the probability that an epidemic will occur [3, 21].

### 2.3.2 Analysis of local stability.

We are now going to look at equilibrium behavior and stability,  $E_{eq}^0$  and  $E_{eq}^*$ .

#### The state of DFE

This section looks at the Cholera DFE's local stability.

**Theorem 2.** The equilibrium  $E_{eq}^0 = (\frac{A}{\mu}, 0, 0)$  for the system (22) that is free of the Cholera disease is asymptotically stable when  $R_0 < 1$  and unstable when  $R_0 > 1$ .

*Proof.* At  $E_{eq}$ , the Jacobian matrix is provided by

$$J(E_{eq}) = \begin{pmatrix} -\mu - \beta_1 \frac{I}{N} - \beta_2 \frac{B}{\kappa + B} & -\beta_1 \frac{S}{N} & -\beta_2 \frac{S(\kappa + B) - SB}{(\kappa + B)^2} \\ \beta_1 \frac{I}{N} + \beta_2 \frac{B}{\kappa + B} & \beta_1 \frac{S}{N} - (\mu + \delta_1) - (\alpha_1 + \alpha_2) & \beta_2 \frac{S(\kappa + B) - SB}{(\kappa + B)^2} \\ 0 & \wp & -\delta_2 \end{pmatrix}. \quad (30)$$

For the DFE, the Jacobian matrix is provided by

$$J(E_{eq}^0) = \begin{pmatrix} -\mu & -\beta_1 \frac{A}{\mu N} & -\beta_2 \frac{A}{\kappa \mu} \\ 0 & \beta_1 \frac{A}{\mu N} - (\mu + \delta_1) - (\alpha_1 + \alpha_2) & \beta_2 \frac{A}{\kappa \mu} \\ 0 & \wp & -\delta_2 \end{pmatrix}. \quad (31)$$

This matrix's characteristic equation is  $\det(J(E_{eq}^0) - \lambda I_3) = 0$ , where  $I_3$  is an order three square identity matrix.

Consequently, we can observe that  $J(E_{eq}^0)$  has the following characteristic equations  $\phi(\lambda)$ :

$$\phi(\lambda) = (-\mu - \lambda) \left[ \left( \beta_1 \frac{A}{\mu N} - (\mu + \delta_1 + \alpha_1 + \alpha_2) - \lambda \right) (-\delta_2 - \lambda) - \wp \beta_2 \frac{A}{\kappa \mu} \right]. \quad (32)$$

The characteristic equation of  $J(E_{eq}^0)$  has the following eigenvalues:

$$\lambda_1 = -\mu,$$

and

$$\lambda^2 - \lambda \left[ -\delta_2 + \beta_1 \frac{A}{\mu N} - (\mu + \delta_1 + \alpha_1 + \alpha_2) \right] - \delta_2 \left( \beta_1 \frac{A}{\mu N} - (\mu + \delta_1 + \alpha_1 + \alpha_2) \right) - \wp \beta_2 \frac{A}{\kappa \mu} = 0. \quad (33)$$

One eigenvalue is obviously negative. The characteristic equation of the following submatrix  $J_1$  is now (33):

$$J_1 = \begin{pmatrix} \beta_1 \frac{A}{\mu N} - (\mu + \delta_1 + \alpha_1 + \alpha_2) & \beta_2 \frac{A}{\kappa \mu} \\ \wp & -\delta_2 \end{pmatrix}. \quad (34)$$

If the trace of  $J_1 < 0$  and the  $\det(J_1) > 0$ , then the eigenvalues are negative. The trace is

$$\begin{aligned}
tr(J_1) &= \beta_1 \frac{A}{\mu N} - (\mu + \delta_1 + \alpha_1 + \alpha_2) - \delta_2 \\
&= (\mu + \delta_1 + \alpha_1 + \alpha_2) \left[ -\beta_1 \frac{A}{\mu N(\mu + \delta_1 + \alpha_1 + \alpha_2)} + 1 \right] - \delta_2 \\
&= (\mu + \delta_1 + \alpha_1 + \alpha_2) \left[ -1 + (R_0 - \beta_2 \frac{A\wp}{\delta_2 \kappa \mu (\mu + \delta_1 + \alpha_1 + \alpha_2)}) \right] - \delta_2.
\end{aligned} \tag{35}$$

Trace of  $J_1 < 0$  if  $R_0 < 1$ , and

$$\det(J_1) = -\delta_2 \beta_1 \frac{A}{\mu N} + \delta_2 (\mu + \delta_1 + \alpha_1 + \alpha_2) - \wp \beta_2 \frac{A}{\kappa \mu} > 0. \tag{36}$$

That is,

$$\begin{aligned}
&\delta_2 (\mu + \delta_1 + \alpha_1 + \alpha_2) \left[ 1 - \beta_1 \frac{A}{\mu N[(\mu + \delta_1) + (\alpha_1 + \alpha_2)]} \right. \\
&\quad \left. - \beta_2 \frac{A\wp}{\delta_2 \kappa \mu[(\mu + \delta_1) + (\alpha_1 + \alpha_2)]} \right] > 0,
\end{aligned} \tag{37}$$

$$\delta_2 (\mu + \delta_1 + \alpha_1 + \alpha_2) \left[ 1 - \frac{A}{\mu(\mu + \delta_1 + \alpha_1 + \alpha_2)} \left( \frac{\beta_1}{N} + \frac{\beta_2 \wp}{\delta_2 \kappa} \right) \right] > 0, \tag{38}$$

$$\delta_2 (\mu + \delta_1 + \alpha_1 + \alpha_2) [1 - R_0] > 0, \tag{39}$$

$$1 - R_0 > 0, \tag{40}$$

if

$$1 > R_0. \tag{41}$$

Consequently, given that each of the characteristic equation's eigenvalues (33) possess a negative real part, it is demonstrated that  $E_{eq}^0$  has a locally asymptotically stable value.  $\square$

### 3 Global stability

#### 3.1 The global stability of a cholera DFE

It is essential to show that the DFE of the model (22), as defined, is GAS in order to ensure that the eradication of cholera infection is unaffected by population sizes. We will use a concept presented in [18] to demonstrate this.

**Lemma 2.** [12] Let us express system (22) in the following form:

$$\begin{aligned}\frac{dX}{dt} &= N(X, Z), \\ \frac{dZ}{dt} &= M(X, Z), M(X, 0) = 0.\end{aligned}\quad (42)$$

In this case, the components associated with the number of uninfected people are represented by  $X \in \mathbb{R}^m$ , and the components associated with the number of infected people, including latent, infectious, and so on, are represented by  $Z \in \mathbb{R}^n$ .

The DFE state of system (22) is denoted by  $E_{eq}^0 = (X^*, 0)$ ,  $X^* = (\frac{A}{\mu})$ .

Furthermore, let us assume the following conditions,  $H_1$  and  $H_2$ :

( $H_1$ ) :  $\frac{dX}{dt} = N(X, 0)$ . Hence  $X^*$  is GAS.

( $H_2$ ) :  $M(X, Z) = AZ - \widehat{M}(X, Z)$ ,  $\widehat{M}(X, Z) \succeq 0$  for  $(X, Z) \in \Omega$ .

Where  $\Omega$  denotes the region where the model is biologically meaningful, the Jacobian  $A = \frac{\partial M}{\partial Z}(X^*, 0)$  is an M-matrix, and the off-diagonal elements of  $A$  are nonnegative.

If  $R_0 < 1$ , then the DFE state,  $E_{eq}^0 = (X^*, 0)$ , is globally stable.

**Theorem 3.** If  $R_0 < 1$ , then the DFE state of the model (22) is GAS.

*Proof.* All we have to do is demonstrate that when  $R_0 < 1$ , conditions ( $H_1$ ) and ( $H_2$ ) hold.

Given that  $X = (S)$ ,  $M = (I, B)$ , and  $X^* = (\frac{A}{\mu})$  in our system (1), then

$$M(X, Z) = \begin{pmatrix} \beta_1 \frac{SI}{N} - I(\mu + \delta_1 + \alpha_1 + \alpha_2) + \beta_2 \frac{SB}{\kappa + B} \\ \wp I - \delta_2 B \end{pmatrix}, \quad (43)$$

and

$$A = \begin{pmatrix} \beta_1 \frac{A}{\mu N} - (\mu + \delta_1 + \alpha_1 + \alpha_2) & \beta_1 \frac{A}{\kappa \mu} \\ \wp & -\delta_2 \end{pmatrix}. \quad (44)$$

Undoubtedly, this is an M-matrix. Meanwhile,

$$\widehat{M}(X, Z) = \begin{pmatrix} \beta_2 \frac{B^2}{\kappa(\kappa + B)} S \\ 0 \end{pmatrix}. \quad (45)$$

It is clear that for  $\widehat{M}(X, Z) \geq 0$ , the conditions ( $H_1$ ) and ( $H_2$ ) have been satisfied, and as a result,  $E_{eq}^0$  is GAS since  $0 \leq S \leq \frac{A}{\mu}$ .  $\square$

### 3.2 The equilibrium of the cholera disease is examined for global stability.

The following is the ultimate outcome of the global stability of  $E_{eq}^* = (S^*, I^*, B^*)$  in this section.

**Theorem 4.** The current equilibrium point of the cholera epidemic,  $E_{eq}^*$ , is GAS if  $R_0 > 1$ .

*Proof.* When the model (22) is solved steadily, the result is

$$B^* = \frac{\wp}{\delta_2} I^*, \quad (46)$$

$$S^* = \frac{N(\delta_2 \kappa + \wp I^*)(\mu + \delta_1 + \alpha_1 + \alpha_2)}{\beta_1(\delta_2 \kappa + \wp I^*) + \beta_2 N \wp}. \quad (47)$$

The following results from substituting (35) and (36) into system (22)'s first equation:

$$a_1 I^{*2} + a_2 I^* + a_3 = 0, \quad (48)$$

where

$$\begin{cases} a_1 = -(\mu + \delta_2 + \alpha_1 + \alpha_2)\beta_1 \wp, \\ a_2 = A\beta_2 \wp - N(\mu + \delta_2 + \alpha_1 + \alpha_2)(\mu \wp + \delta_2 \kappa \frac{\beta_1}{N} + \beta_2 \wp), \\ a_3 = (\mu + \delta_2 + \alpha_1 + \alpha_2)\mu N \delta_2 \kappa [R_0 - 1]. \end{cases} \quad (49)$$

If (48) has real, positive roots, then the system (22) is in endemic equilibrium. We apply Descartes' rule of signs to ascertain whether positive roots exist [15]. It follows that the model has a unique endemic equilibrium whenever  $R_0 > 1$  since the sign of  $a_1$  is negative and the sign of  $a_2$  is positive.  $\square$

## 4 Sensitivity analysis of $R_0$

Sensitivity analysis is an effective method for assessing how modifications to parameter values impact the robustness of a model. It assists in determining the important parameters affecting the reproduction number  $R_0$ , particularly when taking into account assumptions about parameter values and data col-

lection uncertainties. Using the methodology described in Chitnis et al. [5], we calculate the  $R_0$  normalized forward sensitivity indices.

Let

$$T_u^{R_0} = \frac{\partial R_0}{\partial u} * \frac{u}{R_0}. \quad (50)$$

Table 1 provides the sensitivity index of  $R_0$  with respect to the parameter  $u$ .

Table 1 demonstrates that the threshold  $R_0$  is correspondingly sensitive to

Table 1: Sensitivity indices of  $R_0$

Parameter	Description	Sensitivity indices
$A$	New populations are added to the model at a constant rate.	1.0000
$\mu$	The natural death rate	1.0000
$\beta_1$	Transmission rate from human to human	1.0000
$\beta_2$	Transmission rate from environment to human	1.0000
$\alpha_1 ; \alpha_2$	Recovery rate from cholera	-1.5504
$\kappa$	Concentration of Vibrio cholera	-1.0000
$\delta_1$	The death rate induced by the cholera	-0.1008
$\delta_2$	Bacteria death rate	-0.9706
$\wp$	Shedding rate of bacteria by infectious population	1.0000

variations in the parameter values  $A$ ,  $\beta_1$ ,  $\beta_2$ , and  $\wp$ . This suggests that the models will have an increase or decrease in  $R_0$  when the values of each of the parameters in this instance increase or decrease. Conversely, the threshold  $R_0$  is inversely proportional to the variation in  $\mu$ ,  $\alpha_1$ ,  $\alpha_2$ ,  $\delta_1$ , and  $\kappa$ . In this case, a rise or fall in the values of each parameter results in an equivalent rise or fall in  $R_0$ .

## 5 Numerical simulations

We provide numerical solutions to model (Figure 1) for a variety of parameter values in this section. In order to solve system (1), Gumel, Shivakumar, and Sahai [9] created the Gauss-Sade-like implicit finite-difference method (GSS1 method), which is described in [12].

### The fundamental data values:

The model's parameters are displayed in Table 2. The sources are also cited. First, we graphically depict the cholera DFE  $E_{eq}^0$ . Our initial values and parameters are the same as those in Table 2  $R_0 < 1$ .

Using the different values of the initial variables  $S(0)$ ,  $I(0)$ ,  $C(0)$ , and  $R(0)$ , the following observations were obtained from these Figures 2–9. The number of possible individuals increases and gets closer to  $S(0) = 250$ .

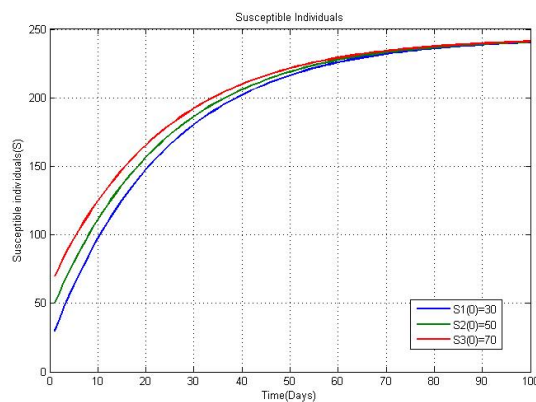


Figure 2: Susceptible individuals.

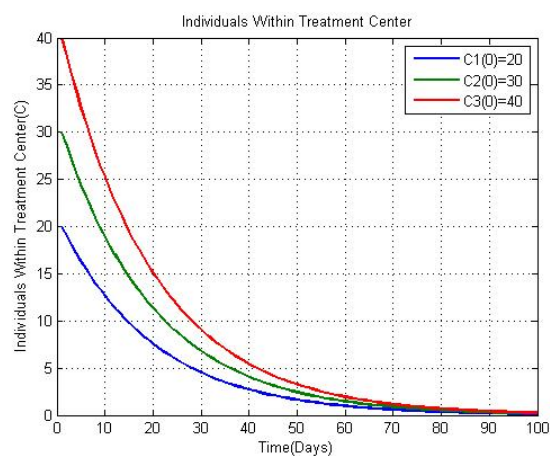


Figure 3: Individuals within treatment center.

The number of carriers and symptomatic infected individuals rises initially, then falls until it almost reaches zero.

The quantity of recovered cases declines until it almost reaches zero.

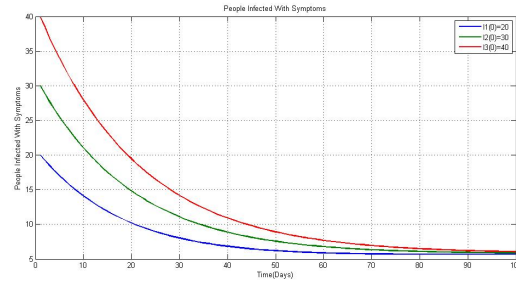


Figure 4: People Infected with symptoms.

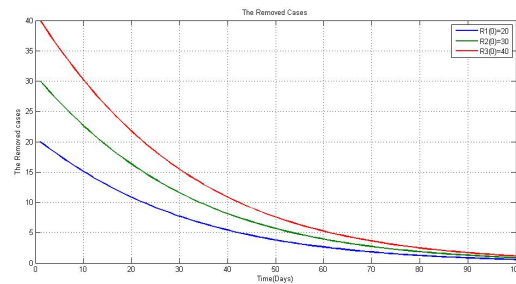


Figure 5: The Removed cases.

We possess an equilibrium for cholera disease with  $R_0 > 1$ . As per Theorem 3, the cholera disease equilibrium  $E_{eq}^*$  of the system (1) is GAS. Furthermore, we begin with a graphic representation of the current equilibrium of the cholera disease  $E_{eq}^*$  and apply the same parameters and initial values in Table 2,  $R_0 > 1$ .

The total number of possible individuals rises initially, then somewhat declines and gets close to  $S^* = 42$ . The percentage of infected cases that show no symptoms or only minor symptoms initially declines quickly before slightly increasing.

The patient population at the treatment center is advancing towards the threshold of (22).

The number of carriers of the bacteria and symptomatic infected individuals converge at  $I^* = 24$ .

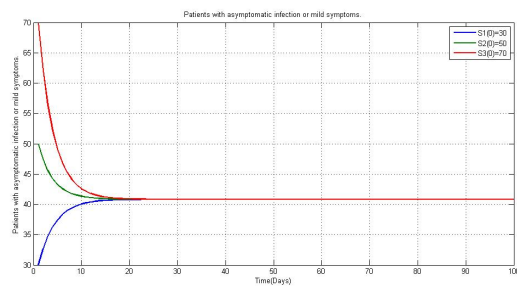


Figure 6: Patients with asymptomatic infection or mild symptoms.

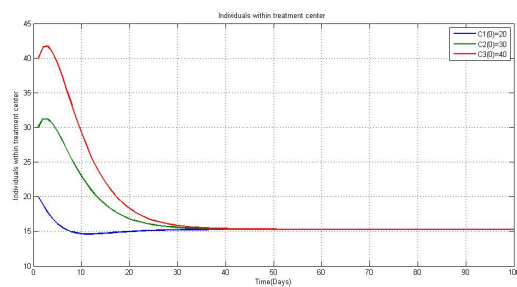


Figure 7: Individuals within treatment center.

Table 2: Baseline parameter values for system (22)

Parameter	Baseline value	Reference
$A$	10	Assumed
$\mu$	0.025	[7]
$\beta_1$	0.02	[10]
$\beta_2$	0.02	[10]
$\alpha_1 ; \alpha_2$	0.214	[7]
$\kappa$	$10^4 \text{ cell/day}$	Assumed
$\delta_1$	0.013	[23]
$\delta_2$	0.33	[14]
$\varphi$	10 cell/day	[23]

### Discussion of result:

A further qualitative examination of the model indicates that its solutions are both positively invariant and bound. For the study of cholera infection,

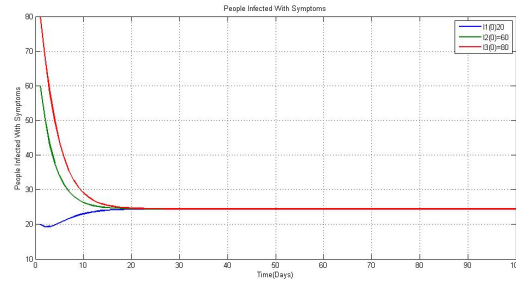


Figure 8: People Infected With Symptoms.

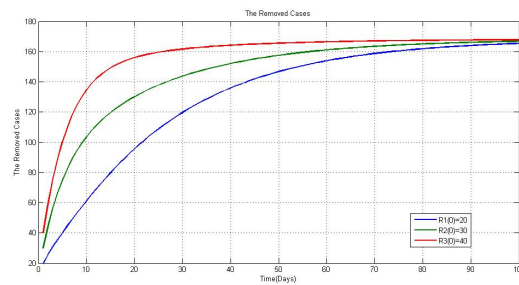


Figure 9: People Removed cases.

the basic reproduction number is required, as follows:

$$R_0 = \beta_1 \frac{A}{\mu N(\mu + \delta_1 + \alpha_1 + \alpha_2)} + \beta_2 \frac{A\varphi}{\delta_2 \kappa \mu(\mu + \delta_1 + \alpha_1 + \alpha_2)}. \quad (51)$$

The calculation was performed utilizing the next-generation methodology, serving as a benchmark to anticipate outbreaks and evaluate control measures. The stability analysis of the DFE was also conducted employing the linearization method, with  $R_0$  as the pivotal parameter. When the basic reproduction number is less than one, Theorem 2 and Lemma 2 indicate that the DFE is asymptotically stable both locally and globally. This means that cholera can be eliminated from the population if the initial population sizes are within the DFE's basin of attraction,  $E_0$ . Furthermore, Lemma 2 demonstrates that the DFE is GAS, implying that cholera can be eliminated regardless of the initial population size.

## 6 Conclusion

In this work, we formulated and presented a continuous SICR-B mathematical model of cholera disease that describes the dynamics of individuals infected with this disease. We found that

$$R_0 = \beta_1 \frac{A}{\mu N(\mu + \delta_1 + \alpha_1 + \alpha_2)} + \beta_2 \frac{A\phi}{\delta_2 \kappa \mu(\mu + \delta_1 + \alpha_1 + \alpha_2)}$$

is the basic reproduction number of the system, which is a crucial indicator in understanding the system's dynamics and the progression of the disease. These results help us identify the key factors influencing disease spread and control. We also performed a sensitivity analysis of the model parameters to determine which parameters have the most significant impact on the reproduction number  $R_0$ . This study highlights the importance of identifying the factors that contribute the most to the spread of the disease, thereby guiding policymakers in optimizing prevention and treatment strategies.

We applied the stability analysis theory for nonlinear systems to analyze the mathematical model of cholera and to study the local and global stability of the disease. The results show that the local asymptotic stability of the DFE  $E_{eq}^0$  can be achieved if  $R_0 \leq 1$ , meaning the disease will eventually die out over time. On the other hand, if  $R_0 > 1$ , then cholera reaches an equilibrium state  $E_{eq}^*$  and remains locally stable, indicating the persistence of the disease.

These results significantly contribute to achieving the overall objectives of the study by improving the understanding of cholera transmission dynamics and providing deeper insights into how to control the disease. Highlighting the most influential factors affecting the basic reproduction number allows for the development of more effective preventive strategies to reduce cholera spread.

Looking ahead, we aim to explore the use of fractional derivatives within a spatiotemporal framework to deepen our understanding of cholera transmission dynamics. This approach is expected to capture complex spatial and temporal patterns, thereby enhancing the accuracy of disease forecasts and control strategies.

## Acknowledgements

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

## Conflicts of interest

The authors declare no conflicts of interest.

## References

- [1] Agarwal, R.P. *Difference equations and inequalities: Theory, methods, and applications*, CRC Press, 2000.
- [2] Akinsulie, O.C., Adesola, R.O., Aliyu, V.A., Oladapo, I.P. and Hamzat, A. *Epidemiology and Transmission Dynamics of Viral Encephalitis in West Africa*. Infect Dis Rep. 2023 Sep 5;15(5):504-517. doi: 10.3390/idr15050050.
- [3] Bani-Yaghoub, M., Gautam, R., Shuai, Z., Van Den Driessche, P. and Ivanek, R. *Reproduction numbers for infections with free-living pathogens growing in the environment*, J. Biol. Dyn. 6 (2) (2012) 923–940.
- [4] Bouaine, A., Rachik, M. and Hattaf, K. *Optimization strategies are applied to discrete epidemic models with specific nonlinear incidence rates*, Int. J. Math. Appl. 4 (2016) 73–80.
- [5] Chitnis, N., Hyman, J.M. and Cushing, J.M. *Determining important parameters in the spread of malaria through the sensitivity analysis of a mathematical model*, Bull. Math. Biol. 70 (2008) 1272–1296.
- [6] Codeço, C.T. *Endemic and epidemic dynamics of cholera: The role of the aquatic reservoir*, BMC Infect. Dis. 1 (2001) 1–14.
- [7] Falaye, A.J., Awonusi, F., Eseyin, O. and Eboigbe, G. *The weak form market efficiency and the Nigerian stock exchange*, Afro Asian Journal of Social Sciences, 9 (4) (2018).

- [8] Ferguson, N. *Capturing human behaviour*. Nature 446, (7137) (2007) 733–733.
- [9] Gumel, A.B., Shivakumar, P.N. and Sahai, B.M. *A mathematical model for the dynamics of HIV-1 during the typical course of infection*, Non-linear Anal. Theory Methods Appl. 47 (3) (2001) 1773–1783.
- [10] Hartley, T.W. *Public perception and participation in water reuse*, Desalination 187 (1-3) (2006): 115–126.
- [11] Hu, Z., Teng, Z. and Jiang, H. *Stability analysis in a class of discrete SIRS epidemic models*, Nonlinear Anal. Real World Appl. 13 (5) (2012) 2017–2033.
- [12] Karrakchou, J., Rachik, M. and Gourari, S. *Optimal control and infectiology: application to an HIV/AIDS model*, Appl.Math. Comput. 177 (2) (2006) 807–818.
- [13] Kot, M. *Elements of mathematical ecology*, Cambridge University Press, 2001.
- [14] Kumar, P., Mishra, D.K., Deshmukh, D.G., Jain, M., Zade, A.M., Ingole, K.V., Goel, A.K. and Yadava, P.K. *Vibrio cholerae O1 Ogawa El Tor strains with the *ctxB7* allele driving cholera outbreaks in south-western India in 2012*, Infect. Genet. Evol. 25 (2014) 93–96.
- [15] Liu, L., Wong, Y.S. and Lee, B.H.K. *Application of the Centre Manifold Theory in Non-Linear Aeroelasticity*, Journal of Sound and Vibration. 234 (4) (2000) 641–659.
- [16] Mwasa, A. and Tchuente, J.M. *Mathematical Analysis of a Cholera Model with Public Health Interventions*. Biosystems 105 (3) (2011) 190–200.
- [17] Nelson, E.J., Harris, J.B., Glenn Morris Jr, J., Calderwood, S.B. and Camilli, A. *Cholera transmission: The host, pathogen and bacteriophage dynamic*, Nat. Rev. Microbiol. 7(10) (2009) 693–702.

- [18] Okolo, P.N., Magaji, A.S., Joshua, I. and Useini, P.F. *Mathematical modelling and analysis of cholera disease dynamics with control* FUDMA J. Sci. 4 (4) (2020) 363–381.
- [19] Samanta, S., Rana, S., Sharma, A., Misra, A.K. and Chattopadhyay, J. *Effect of awareness programs by media on the epidemic outbreaks: A mathematical model*, Appl. Math. Comput. 219 (12) (2013) 6965–6977.
- [20] Sebastian, E. and Victor, P. *Optimal control strategy of a discrete-time svir epidemic model with immigration of infectives*, Int. J. Pure Appl. Math. 113 (8) (2017) 55–63.
- [21] Tahir, M., Inayat Ali Shah, S., Zaman, G. and Muhammad, S. *Ebola virus epidemic disease its modeling and stability analysis required abstain strategies*, Cogent Biol. 4 (1) (2018) 1488511.
- [22] 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.
- [23] Vashist, A., Verma, J., Narendrakumar, L. and Das, B. *Molecular Insights into Genomic Islands and Evolution of Vibrio cholerae*, In Microbial Genomic Islands in Adaptation and Pathogenicity: Springer, 2023.
- [24] Wang, J. and Modnak, C. *Modeling cholera dynamics with controls*, Can. Appl. Math. Q. 19 (3) (2011) 255–273.
- [25] Watson, A.P., Armstrong, A.Q., White, G.H. and Thran, B.H. *Health-based ingestion exposure guidelines for Vibrio cholerae: Technical basis for water reuse applications*, Sci. Total Environ. 613 (2018) 379–387.



## Two-step inertial Tseng's extragradient methods for a class of bilevel split variational inequalities

L.H.M. Van and T.V. Anh\*

### Abstract

This work presents a two-step inertial Tseng's extragradient method with a self-adaptive step size for solving a bilevel split variational inequality problem (BSVIP) in Hilbert spaces. This algorithm only requires two projections per iteration, enhancing its practicality. We establish a strong convergence theorem for the method, showing that it effectively tackles the BSVIP without necessitating prior knowledge of the Lipschitz or strongly monotone constants associated with the mappings. Additionally, the implementation of this method removes the need to compute or estimate the

Received 28 September 2024; revised 19 March 2025; accepted 6 April 2025

Le Huynh My Van

Department of Mathematics and Physics, University of Information Technology, Ho Chi Minh City, Vietnam; Department of Applied Mathematics, Faculty of Applied Science, Ho Chi Minh City University of Technology (HCMUT), 268 Ly Thuong Kiet Street, District 10, Ho Chi Minh City, Vietnam; Vietnam National University Ho Chi Minh City, Linh Trung Ward, Thu Duc City, Ho Chi Minh City, Vietnam. e-mail: vanlhm@uit.edu.vn; lhmvan.sdh231@hcmut.edu.vn

Tran Viet Anh

Department of Scientific Fundamentals, Posts and Telecommunications Institute of Technology, Hanoi, Vietnam. e-mail: tranvietanh@outlook.com; tvanh@ptit.edu.vn

### How to cite this article

Van, L.H.M. and Anh, T.V., Two-step inertial Tseng's extragradient methods for a class of bilevel split variational inequalities. *Iran. J. Numer. Anal. Optim.*, 2025; 15(3): 877-913. <https://doi.org/10.22067/ijnao.2025.90001.1528>

norm of the given operator, a task that can often be challenging in practical situations. We also explore specific cases to demonstrate the versatility of the method. Finally, we present an application of the split minimum norm problem in production and consumption systems and provide several numerical experiments to validate the practical implementability of the proposed algorithms.

**AMS subject classifications (2020):** Primary 49M37; Secondary 90C26, 90C33.

**Keywords:** Bilevel split variational inequality problem; Bilevel variational inequality problem; Split feasibility problem; Tseng's extragradient method.

## 1 Introduction

Let  $C$  and  $Q$  be nonempty closed convex subsets of the real Hilbert spaces  $\mathcal{H}_1$  and  $\mathcal{H}_2$ , respectively. Let  $A : \mathcal{H}_1 \rightarrow \mathcal{H}_2$  be a bounded linear operator. Define the mappings  $F_1 : \mathcal{H}_1 \rightarrow \mathcal{H}_1$  and  $F_2 : \mathcal{H}_2 \rightarrow \mathcal{H}_2$  on  $\mathcal{H}_1$  and  $\mathcal{H}_2$ , respectively. The split variational inequality problem (SVIP), initially proposed by Censor, Gibali, and Reich [15], can be expressed as follows:

$$\text{Find } x^* \in C : \langle F_1(x^*), x - x^* \rangle \geq 0, \text{ for all } x \in C \quad (1)$$

such that

$$y^* = Ax^* \in Q : \langle F_2(y^*), y - y^* \rangle \geq 0, \text{ for all } y \in Q. \quad (2)$$

When  $F_1 = 0$  and  $F_2 = 0$ , the SVIP reduces to a special case known as the split feasibility problem (SFP), which is formulated as

$$\text{Find } x^* \in C \text{ such that } Ax^* \in Q. \quad (3)$$

This problem was first introduced by Censor and Elfving [13] as a model for inverse problems in finite-dimensional Hilbert spaces. Recently, its applicability has been extended to fields such as intensity-modulated radiation therapy [12, 16, 14] and other practical scenarios. For additional details on the SFP, refer to [1, 3, 5, 14, 7, 8, 11, 10, 9, 23, 24, 34, 36, 43] and the sources cited within those references.

In this paper, our primary objective is to solve a variational inequality problem (VIP) defined over the solution set of the SVIP. Specifically, we aim to address the following problem:

$$\text{Find } x^* \in \Omega_{\text{SVIP}} \text{ such that } \langle F(x^*), x - x^* \rangle \geq 0, \text{ for all } x \in \Omega_{\text{SVIP}}, \quad (4)$$

where  $F : \mathcal{H}_1 \rightarrow \mathcal{H}_1$  is  $\eta$ -strongly monotone and  $L$ -Lipschitz continuous on  $\mathcal{H}_1$ , and  $\Omega_{\text{SVIP}}$  represents the solution set of the SVIP defined by equations (1) and (2). Problem (4) is referred to as the bilevel split variational inequality problem (BSVIP) in [1]. Suppose that  $\mathcal{H}_1 = \mathcal{H}$ ,  $F : \mathcal{H} \rightarrow \mathcal{H}$  is strongly monotone and Lipschitz continuous on  $\mathcal{H}$ , that  $F_1 = G : \mathcal{H} \rightarrow \mathcal{H}$  is a mapping on  $\mathcal{H}$ , that  $F_2 = 0$ , and that  $Q = \mathcal{H}_2$ . Then the BSVIP (4) simplifies to the following bilevel VIP:

$$\text{Find } x^* \in \text{Sol}(C, G) \text{ such that } \langle F(x^*), y - x^* \rangle \geq 0, \text{ for all } y \in \text{Sol}(C, G), \quad (5)$$

where  $\text{Sol}(C, G)$  represents the set of all solutions to the VIP given by

$$\text{Find } y^* \in C \text{ such that } \langle G(y^*), z - y^* \rangle \geq 0, \text{ for all } z \in C. \quad (6)$$

Bilevel VIPs (5)–(6) encompass various types of bilevel optimization problems [20, 37, 6], minimum norm problems related to the solution set of variational inequalities [42, 44], and other variational inequalities [28, 29, 21, 19, 22]. In recent years, numerous approaches have been developed to solve the BVIP (5)–(6) in both finite and infinite-dimensional spaces. For a comprehensive overview, see [2, 4, 38] and the references therein.

One of the most famous methods for solving VIPs is the extragradient method, first proposed by Korpelevich [30] for saddle problems. However, the extragradient method may be costly, since it requires two projections at each step. To improve this, Tseng [39] introduced an alternative extragradient method that reduces the number of projections required. Instead of performing two projections, Tseng's method requires only one projection onto  $C$  per iteration. Tseng's extragradient method is described as follows:

$$\begin{cases} x^0 \in \mathcal{H}, \\ y^n = P_C(x^n - \lambda F_1(x^n)), \\ x^{n+1} = y^n - \lambda(F_1(y^n) - F_1(x^n)), \end{cases} \quad (7)$$

where  $F_1$  is  $L_1$ -Lipschitz continuous and  $\lambda \in \left(0, \frac{1}{L_1}\right)$ .

Inspired the Tseng's extragradient method for solving VIPs, Huy et al. [25] introduced the modified Tseng's extragradient method for solving the BSVIP (4), where  $F : \mathcal{H}_1 \rightarrow \mathcal{H}_1$  is  $\eta$ -strongly monotone and  $L$ -Lipschitz continuous on  $\mathcal{H}_1$ ,  $F_1 : \mathcal{H}_1 \rightarrow \mathcal{H}_1$  and  $F_2 : \mathcal{H}_2 \rightarrow \mathcal{H}_2$  are pseudomonotone and Lipschitz continuous mappings. Specifically, they proposed the following algorithm

$$\left\{ \begin{array}{l} x^0 \in \mathcal{H}_1, \\ u^n = A(x^n), \\ v^n = P_Q(u^n - \mu_n F_2(u^n)), \\ w^n = v^n - \mu_n (F_2(v^n) - F_2(u^n)), \\ \mu_{n+1} = \begin{cases} \min \left\{ \frac{\mu \|u^n - v^n\|}{\|F_2(u^n) - F_2(v^n)\|}, \mu_n \right\} & \text{if } F_2(u^n) \neq F_2(v^n), \\ \mu_n & \text{if } F_2(u^n) = F_2(v^n), \end{cases} \\ y^n = x^n + \delta_n A^*(w^n - u^n), \\ \delta_n = \begin{cases} \frac{\|w^n - u^n\|^2}{2\|A^*(w^n - u^n)\|^2} & \text{if } A^*(w^n - u^n) \neq 0, \\ 0 & \text{if } A^*(w^n - u^n) = 0. \end{cases} \\ z^n = P_C(y^n - \lambda_n F_1(y^n)), \\ t^n = z^n - \lambda_n (F_1(z^n) - F_1(y^n)), \\ \lambda_{n+1} = \begin{cases} \min \left\{ \frac{\lambda \|y^n - z^n\|}{\|F_1(y^n) - F_1(z^n)\|}, \lambda_n \right\} & \text{if } F_1(y^n) \neq F_1(z^n), \\ \lambda_n & \text{if } F_1(y^n) = F_1(z^n), \end{cases} \\ x^{n+1} = t^n - \varepsilon_n F(t^n), \end{array} \right. \quad (8)$$

where  $\mu_0 > 0$ ,  $\lambda_0 > 0$ ,  $\mu \in (0, 1)$ ,  $\lambda \in (0, 1)$ ,  $\{\varepsilon_n\} \subset (0, 1)$ ,  $\lim_{n \rightarrow \infty} \varepsilon_n = 0$ , and  $\sum_{n=0}^{\infty} \varepsilon_n = \infty$ . The author demonstrated that the sequence  $\{x^n\}$  produced by the algorithm (8) converges strongly to the unique solution of the BSVIP (4), provided that the solution set of the SVIP (1)–(2) is nonempty.

To enhance the convergence rate of algorithms, inertial acceleration is commonly utilized. Originally introduced by Polyak [33] in 1964 for solving smooth convex minimization problems, the inertial algorithm distinguishes

itself by leveraging the previous two iterates to generate the next one. Numerous researchers have explored and implemented the inertial scheme to accelerate algorithmic convergence (see [32, 40] and references therein). These studies primarily employ a single inertial parameter to achieve acceleration. However, recent works by some authors [26, 27] have investigated multi-step inertial algorithms, demonstrating that incorporating multi-step inertial terms, such as the two-step inertial term, further enhances algorithmic speed.

In this paper, drawing inspiration from the aforementioned studies, we introduce a novel iterative scheme that combines the two-step inertial technique with a modified Tseng's extragradient method, as employed by Huy et al. [25], to solve the BSVIP in (4). We demonstrate that the sequence produced by our method converges strongly to the unique solution of (4), with the stepsize determined at each iteration. Consequently, our approach does not necessitate prior knowledge of the Lipschitz or strong monotonicity constants for the mappings involved. Additionally, the implementation of this method eliminates the need to compute or estimate the norm of the bounded linear operator.

The structure of the paper is organized as follows. Section 2 presents essential definitions and lemmas that will be utilized in section 3, where we outline the algorithm and demonstrate its strong convergence. We conclude this section by exploring various applications of our results to the bilevel VIPs, the simple bilevel optimization problem and VIPs with the SF constraints. Lastly, we apply the split minimum norm problem (SMNP) to production and consumption systems and conduct numerical experiments to evaluate the effectiveness of the proposed algorithms.

## 2 Preliminaries

In the following discussion, we denote the strong convergence of a sequence  $\{x^n\}$  to  $x$  in a real Hilbert space  $\mathcal{H}$  as  $x^n \rightarrow x$  and the weak convergence as  $x^n \rightharpoonup x$ . Recall that for a nonempty closed convex subset  $C$  of  $\mathcal{H}$ , the metric projection  $P_C$  is a mapping from  $\mathcal{H}$  to  $C$ . For each  $x \in \mathcal{H}$ ,  $P_C(x)$  is defined as the unique point in  $C$  that minimizes the distance to  $x$ , satisfying the condition:

$$\|x - P_C(x)\| = \min\{\|x - y\| : y \in C\}.$$

Let us also recall some well-known definitions which will be used in this paper.

**Definition 1.** ([35]) Consider two Hilbert spaces, denoted as  $\mathcal{H}_1$  and  $\mathcal{H}_2$ . Let  $A : \mathcal{H}_1 \rightarrow \mathcal{H}_2$  be a bounded linear operator. The adjoint of this operator, represented as  $A^* : \mathcal{H}_2 \rightarrow \mathcal{H}_1$ , is characterized by the following relationship:

$$\langle A(x), y \rangle = \langle x, A^*(y) \rangle, \text{ for all } x \in \mathcal{H}_1, \text{ for all } y \in \mathcal{H}_2.$$

The adjoint operator of a bounded linear operator  $A$  between Hilbert spaces  $\mathcal{H}_1$  and  $\mathcal{H}_2$  is both well-defined and unique. Moreover, the adjoint operator  $A^*$  is also a bounded linear operator, satisfying the property that  $\|A^*\| = \|A\|$ .

**Definition 2.** ([17, 29])

A mapping  $F : \mathcal{H} \rightarrow \mathcal{H}$  is said to be

(i)  $\eta$ -strongly monotone on  $\mathcal{H}$  if there exists  $\eta > 0$  such that

$$\langle F(x) - F(y), x - y \rangle \geq \eta \|x - y\|^2, \text{ for all } x, y \in \mathcal{H};$$

(ii)  $L$ -Lipschitz continuous on  $\mathcal{H}$  if there exists  $L > 0$  such that

$$\|F(x) - F(y)\| \leq L \|x - y\|, \text{ for all } x, y \in \mathcal{H};$$

(iii) monotone on  $\mathcal{H}$  if

$$\langle F(x) - F(y), x - y \rangle \geq 0, \text{ for all } x, y \in \mathcal{H};$$

(iv) pseudomonotone on  $C$  if

$$\langle F(y), x - y \rangle \geq 0 \Rightarrow \langle F(x), x - y \rangle \geq 0, \text{ for all } x, y \in C.$$

To demonstrate the convergence of the proposed algorithm, we will require the following lemmas.

**Lemma 1.** ([25]) Let  $C$  be a nonempty closed convex subset of a real Hilbert space  $\mathcal{H}$ . Let  $F : \mathcal{H} \rightarrow \mathcal{H}$  be pseudomonotone on  $C$  and  $L$ -Lipschitz continuous on  $\mathcal{H}$  such that the solution set  $\text{Sol}(C, F)$  of the  $VIP(C, F)$  is

nonempty. Let let  $x \in \mathcal{H}$ , and let  $\mu \in (0, 1)$ ,  $\lambda > 0$ , and define

$$\begin{aligned} y &= P_C(x - \lambda F(x)), \\ z &= y - \lambda(F(y) - F(x)), \\ \gamma &= \begin{cases} \min \left\{ \frac{\mu \|x - y\|}{\|F(x) - F(y)\|}, \lambda \right\} & \text{if } F(x) \neq F(y), \\ \lambda & \text{if } F(x) = F(y). \end{cases} \end{aligned}$$

Then for all  $x^* \in \text{Sol}(C, F)$

$$\|z - x^*\|^2 \leq \|x - x^*\|^2 - \left(1 - \mu^2 \frac{\lambda^2}{\gamma^2}\right) \|x - y\|^2.$$

**Lemma 2.** ([25]) Let  $C$  be a nonempty closed convex subset of a real Hilbert space  $\mathcal{H}$ . Let  $F : \mathcal{H} \rightarrow \mathcal{H}$  be a mapping such that  $\limsup_{n \rightarrow \infty} \langle F(x^n), z - y^n \rangle \leq \langle F(\bar{x}), z - \bar{y} \rangle$  for every sequences  $\{x^n\}, \{y^n\}$  in  $\mathcal{H}$  converging weakly to  $\bar{x}$  and  $\bar{y}$ , respectively. Assume that  $\lambda_n \geq a > 0$  for all  $n$ ,  $\{x^n\}$  is a sequence in  $\mathcal{H}$  satisfying  $x^n \rightharpoonup \bar{x}$  and  $\lim_{n \rightarrow \infty} \|x^n - y^n\| = 0$ , where  $y^n = P_C(x^n - \lambda_n F(x^n))$  for all  $n$ . Then  $\bar{x} \in \text{Sol}(C, F)$ .

**Lemma 3.** ([31, Remark 4.4]) Let  $\{a_n\}$  be a sequence of nonnegative real numbers. Suppose that for any integer  $m$ , there exists an integer  $p$  such that  $p \geq m$  and  $a_p \leq a_{p+1}$ . Let  $n_0$  be an integer such that  $a_{n_0} \leq a_{n_0+1}$  and define, for all integer  $n \geq n_0$ , by

$$\tau(n) = \max\{k \in \mathbb{N} : n_0 \leq k \leq n, a_k \leq a_{k+1}\}.$$

Then  $\{\tau(n)\}_{n \geq n_0}$  is a nondecreasing sequence satisfying  $\lim_{n \rightarrow \infty} \tau(n) = \infty$  and the following inequalities hold true:

$$a_{\tau(n)} \leq a_{\tau(n)+1}, a_n \leq a_{\tau(n)+1}, \quad \text{for all } n \geq n_0.$$

**Lemma 4.** ([41]) Let  $\{a_n\}$  be a sequence of nonnegative real numbers, let  $\{\varepsilon_n\}$  be a sequence in  $(0, 1)$  such that  $\sum_{n=0}^{\infty} \varepsilon_n = \infty$ , and let  $\{b_n\}$  be a sequence of real numbers with  $\limsup_{n \rightarrow \infty} b_n \leq 0$ . Suppose that

$$a_{n+1} \leq (1 - \varepsilon_n)a_n + \varepsilon_n b_n, \quad \text{for all } n \geq 0.$$

Then  $\lim_{n \rightarrow \infty} a_n = 0$ .

### 3 The algorithm and convergence analysis

In this section, we propose a strong convergence algorithm for solving BSVIP by using two-step inertial Tseng's extragradient methods with self-adaptive step size. We impose the following assumptions concerning the mappings  $F$ ,  $F_1$ , and  $F_2$  related to the BSVIP.

**Assumption 1.** ([1, 25]) Let the following hold:

- $A_1)$   $F : \mathcal{H}_1 \longrightarrow \mathcal{H}_1$  is  $\eta$ -strongly monotone and  $L$ -Lipschitz continuous on  $\mathcal{H}_1$ .
- $A_2)$   $F_1 : \mathcal{H}_1 \longrightarrow \mathcal{H}_1$  is pseudomonotone on  $C$  and  $L_1$ -Lipschitz continuous on  $\mathcal{H}_1$ .
- $A_3)$   $\limsup_{n \rightarrow \infty} \langle F_1(x^n), y - y^n \rangle \leq \langle F_1(\bar{x}), y - \bar{y} \rangle$  holds for any sequences  $\{x^n\}$  and  $\{y^n\}$  in  $\mathcal{H}_1$  that converge weakly to  $\bar{x}$  and  $\bar{y}$ , respectively.
- $A_4)$   $F_2 : \mathcal{H}_2 \longrightarrow \mathcal{H}_2$  is pseudomonotone on  $Q$  and  $L_2$ -Lipschitz continuous on  $\mathcal{H}_2$ .
- $A_5)$   $\limsup_{n \rightarrow \infty} \langle F_2(u^n), v - v^n \rangle \leq \langle F_2(\bar{u}), v - \bar{v} \rangle$  holds for any sequences  $\{u^n\}$  and  $\{v^n\}$  in  $\mathcal{H}_2$  that converge weakly to  $\bar{u}$  and  $\bar{v}$ , respectively.

One can see that in finite-dimensional spaces, the conditions  $A_3$  and  $A_5$  automatically result from the Lipschitz continuity of  $F_1$  and  $F_2$ .

**Remark 1.** In Algorithm 1, we introduce a two-step inertial version of Tseng's extragradient method. The inertial update is applied in Step 2, where we replace  $x^n$  with  $y^n = x^n + \alpha_n(x^n - x^{n-1}) + \beta_n(x^{n-2} - x^{n-1})$  for the next step. Starting from Step 3, our algorithm closely follows [25, Algorithm 3.1], as described in (8). The key differences between our approach and [25, Algorithm 3.1] lie in the order of applying the modified Tseng's extragradient method in the two spaces, as well as the inclusion of the two-step inertial update. In [25, Algorithm 3.1], the authors first transform to space  $\mathcal{H}_2$ , apply the modified Tseng's extragradient method to the mapping  $F_2$ ,

**Algorithm 1**

**Step 0.** Choose  $\mu_0 > 0$ ,  $\lambda_0 > 0$ ,  $\mu \in (0, 1)$ ,  $\lambda \in (0, 1)$ ,  $\{\rho_n\} \subset [a, b] \subset (0, 1)$ ,  $\{\gamma_n\} \subset [0, \infty)$ ,  $\{\xi_n\} \subset [0, \infty)$ ,  $\{\eta_n\} \subset (0, \infty)$ ,  $\{\varepsilon_n\} \subset (0, 1)$  such that  $\lim_{n \rightarrow \infty} \frac{\eta_n}{\varepsilon_n} = 0$ ,

$$\lim_{n \rightarrow \infty} \varepsilon_n = 0, \quad \sum_{n=0}^{\infty} \varepsilon_n = \infty.$$

**Step 1.** Let  $x^{-2}, x^{-1}, x^0 \in \mathcal{H}_1$ . Set  $n := 0$ .

**Step 2.** Compute  $y^n = x^n + \alpha_n(x^n - x^{n-1}) + \beta_n(x^{n-2} - x^{n-1})$ , where

$$\alpha_n = \begin{cases} \min \left\{ \frac{\eta_n}{\|x^n - x^{n-1}\|}, \gamma_n \right\} & \text{if } x^n \neq x^{n-1}, \\ \gamma_n & \text{if } x^n = x^{n-1}, \end{cases}$$

and

$$\beta_n = \begin{cases} \min \left\{ \frac{\eta_n}{\|x^{n-2} - x^{n-1}\|}, \xi_n \right\} & \text{if } x^{n-2} \neq x^{n-1}, \\ \xi_n & \text{if } x^{n-2} = x^{n-1}. \end{cases}$$

**Step 3.** Compute

$$z^n = P_C(y^n - \lambda_n F_1(y^n)),$$

$$t^n = z^n - \lambda_n(F_1(z^n) - F_1(y^n)),$$

where

$$\lambda_{n+1} = \begin{cases} \min \left\{ \frac{\lambda \|y^n - z^n\|}{\|F_1(y^n) - F_1(z^n)\|}, \lambda_n \right\} & \text{if } F_1(y^n) \neq F_1(z^n), \\ \lambda_n & \text{if } F_1(y^n) = F_1(z^n). \end{cases}$$

**Step 4.** Compute  $u^n = A(t^n)$  and

$$v^n = P_Q(u^n - \mu_n F_2(u^n)),$$

$$w^n = v^n - \mu_n(F_2(v^n) - F_2(u^n)),$$

where

$$\mu_{n+1} = \begin{cases} \min \left\{ \frac{\mu \|u^n - v^n\|}{\|F_2(u^n) - F_2(v^n)\|}, \mu_n \right\} & \text{if } F_2(u^n) \neq F_2(v^n), \\ \mu_n & \text{if } F_2(u^n) = F_2(v^n). \end{cases}$$

**Step 5.** Compute

$$s^n = t^n + \delta_n A^*(w^n - u^n),$$

where the stepsize  $\delta_n$  is chosen in such a way that

$$\delta_n = \begin{cases} \frac{\rho_n \|w^n - u^n\|^2}{\|A^*(w^n - u^n)\|^2} & \text{if } A^*(w^n - u^n) \neq 0, \\ 0 & \text{if } A^*(w^n - u^n) = 0. \end{cases}$$

**Step 6.** Compute

$$x^{n+1} = s^n - \varepsilon_n F(s^n).$$

**Step 7.** Set  $n := n + 1$ , and go to **Step 2**.

return to space  $\mathcal{H}_1$ , and then apply the method again to the mapping  $F_1$ . In contrast, our algorithm first applies this modified extragradient method to the mapping  $F_1$  in space  $\mathcal{H}_1$  (Step 3), then transforms to space  $\mathcal{H}_2$  and applies it to  $F_2$  (Step 4), before returning to space  $\mathcal{H}_1$  in Step 5. Notably, before applying the method to  $F_1$  in space  $\mathcal{H}_1$ , we use the two-step inertial update  $y^n = x^n + \alpha_n(x^n - x^{n-1}) + \beta_n(x^{n-2} - x^{n-1})$  instead of  $x^n$ .

The following lemma is part of the proof of [25, Algorithm 3.1], but we have made a slight modification to better suit the new proof.

**Lemma 5.** Assume that the conditions  $(A_1) - (A_5)$  are satisfied and that  $\Omega_{\text{SVIP}} \neq \emptyset$ . Let  $\mu, \lambda$  as in Algorithm 1, let  $\varepsilon \in \left(0, \frac{2\eta}{L^2}\right)$ , and let the sequences  $\{\mu^n\}$  and  $\{\lambda^n\}$  be generated by Algorithm 1. We show that there exists  $n_0 \in \mathbb{N}$  such that

$$1 - \mu^2 \frac{\mu_n^2}{\mu_{n+1}^2} > \frac{1 - \mu^2}{2} > 0, \quad 1 - \lambda^2 \frac{\lambda_n^2}{\lambda_{n+1}^2} > \frac{1 - \lambda^2}{2} > 0, \quad \varepsilon_n < \varepsilon, \quad \text{for all } n \geq n_0.$$

*Proof.* With  $F_2$  being  $L_2$ -Lipschitz continuous on  $\mathcal{H}_2$ , it follows that  $\|F_2(u^n) - F_2(v^n)\| \leq L_2 \|u^n - v^n\|$ . Consequently, employing induction, we have  $\mu_n \geq \min\left(\frac{\mu}{L_2}, \mu_0\right) > 0$  for all  $n \geq 0$ . The definition of  $\mu_{n+1}$  implies  $\mu_{n+1} \leq \mu_n$  for all  $n \geq 0$ . Combining this with  $\mu_n \geq \min\left(\frac{\mu}{L_2}, \mu_0\right) > 0$  for all  $n \geq 0$ , we infer the existence of the limit of the sequence  $\{\mu_n\}$ . Let us denote  $\lim_{n \rightarrow \infty} \mu_n = \mu^*$ . It is evident that  $\mu^* \geq \min\left(\frac{\mu}{L_2}, \mu_0\right) > 0$ .

Using the same reasoning as before, we find that

$$\lambda_0 \geq \lambda_n \geq \min\left(\frac{\lambda}{L_1}, \lambda_0\right) > 0, \quad \text{for all } n \geq 0$$

and

$$\lim_{n \rightarrow \infty} \lambda_n = \lambda^* \geq \min\left(\frac{\lambda}{L_1}, \lambda_0\right) > 0.$$

From  $\lim_{n \rightarrow \infty} \mu_n = \mu^* > 0$  and  $\lim_{n \rightarrow \infty} \lambda_n = \lambda^* > 0$ , we get  $\lim_{n \rightarrow \infty} \left(1 - \mu^2 \frac{\mu_n^2}{\mu_{n+1}^2}\right) = 1 - \mu^2 > 0$ ,  $\lim_{n \rightarrow \infty} \left(1 - \lambda^2 \frac{\lambda_n^2}{\lambda_{n+1}^2}\right) = 1 - \lambda^2 > 0$ . Since  $\lim_{n \rightarrow \infty} \varepsilon_n = 0$ , there exists  $n_0 \in \mathbb{N}$  such that

$$1 - \mu^2 \frac{\mu_n^2}{\mu_{n+1}^2} > \frac{1 - \mu^2}{2} > 0, \quad 1 - \lambda^2 \frac{\lambda_n^2}{\lambda_{n+1}^2} > \frac{1 - \lambda^2}{2} > 0, \quad \varepsilon_n < \varepsilon, \quad \text{for all } n \geq n_0.$$

□

**Lemma 6.** Let  $\{t^n\}$ ,  $\{u^n\}$ ,  $\{w^n\}$  and  $\{s^n\}$  be the sequences generated by Algorithm 1. Then, for all  $n \geq n_0$ , where  $n_0$  is given in Lemma 5, the following inequalities hold:

$$0 \leq \frac{a^2}{(\|A\| + 1)^2} \|w^n - u^n\|^2 \leq \|s^n - t^n\|^2 \leq \frac{b}{1-b} (\|t^n - x^*\|^2 - \|s^n - x^*\|^2),$$

where  $x^*$  is the unique solution to the problem (4).

*Proof.* As  $\Omega_{\text{SVIP}}$  is nonempty, problem (4) has a unique solution denoted by  $x^*$ . Specifically,  $x^* \in \Omega_{\text{SVIP}}$ , implying that it satisfies  $x^* \in \text{Sol}(C, F_1)$  and  $Ax^* \in \text{Sol}(Q, F_2)$ . According to Lemma 1, for all  $n \geq 0$ , we have

$$\|w^n - Ax^*\|^2 \leq \|u^n - Ax^*\|^2 - \left(1 - \mu^2 \frac{\mu_n^2}{\mu_{n+1}^2}\right) \|u^n - v^n\|^2, \quad (9)$$

$$\|t^n - x^*\|^2 \leq \|y^n - x^*\|^2 - \left(1 - \lambda^2 \frac{\lambda_n^2}{\lambda_{n+1}^2}\right) \|y^n - z^n\|^2. \quad (10)$$

From Lemma 5, (9) and (10), we get

$$\|w^n - Ax^*\| \leq \|u^n - Ax^*\|, \quad \text{for all } n \geq n_0, \quad (11)$$

$$\|t^n - x^*\| \leq \|y^n - x^*\|, \quad \text{for all } n \geq n_0. \quad (12)$$

From (11), since  $u^n = A(t^n)$ , we obtain, for all  $n \geq n_0$

$$\begin{aligned} 2\langle t^n - x^*, A^*(w^n - u^n) \rangle &= 2\langle A(t^n - x^*), w^n - u^n \rangle \\ &= 2\langle u^n - Ax^*, w^n - u^n \rangle \\ &= 2[\langle w^n - Ax^*, w^n - u^n \rangle - \|w^n - u^n\|^2] \\ &= (\|w^n - Ax^*\|^2 - \|u^n - Ax^*\|^2) - \|w^n - u^n\|^2 \\ &\leq -\|w^n - u^n\|^2. \end{aligned} \quad (13)$$

**Case 1.**  $A^*(w^n - u^n) = 0$ .

In this case,  $s^n = t^n$ . Also, it follows from (13) that  $\|w^n - u^n\| = 0$ . Thus, the inequalities in Lemma 6 hold.

**Case 2.**  $A^*(w^n - u^n) \neq 0$ .

From (13) and  $\{\rho_n\} \subset [a, b] \subset (0, 1)$ , we get for all  $n \geq n_0$  that

$$\begin{aligned}
\|s^n - x^*\|^2 &= \|(t^n - x^*) + \delta_n A^*(w^n - u^n)\|^2 \\
&= \|t^n - x^*\|^2 + \delta_n^2 \|A^*(w^n - u^n)\|^2 + 2\delta_n \langle t^n - x^*, A^*(w^n - u^n) \rangle \\
&\leq \|t^n - x^*\|^2 + \delta_n^2 \|A^*(w^n - u^n)\|^2 - \delta_n \|w^n - u^n\|^2 \\
&= \|t^n - x^*\|^2 + \frac{\rho_n^2 \|w^n - u^n\|^4}{\|A^*(w^n - u^n)\|^2} - \frac{\rho_n \|w^n - u^n\|^4}{\|A^*(w^n - u^n)\|^2} \\
&= \|t^n - x^*\|^2 - \frac{\rho_n^2 \|w^n - u^n\|^4}{\|A^*(w^n - u^n)\|^2} \cdot \frac{1 - \rho_n}{\rho_n} \\
&\leq \|t^n - x^*\|^2 - \frac{\rho_n^2 \|w^n - u^n\|^4}{\|A^*(w^n - u^n)\|^2} \cdot \frac{1 - b}{b}.
\end{aligned} \tag{14}$$

Then, from (14), we have

$$\begin{aligned}
\|s^n - t^n\|^2 &= \delta_n^2 \|A^*(w^n - u^n)\|^2 = \frac{\rho_n^2 \|w^n - u^n\|^4}{\|A^*(w^n - u^n)\|^2} \\
&\leq \frac{b}{1 - b} (\|t^n - x^*\|^2 - \|s^n - x^*\|^2).
\end{aligned} \tag{15}$$

On the other hand,

$$0 < \|A^*(w^n - u^n)\| \leq \|A^*\| \|w^n - u^n\| = \|A\| \|w^n - u^n\| \leq (\|A\| + 1) \|w^n - u^n\|.$$

Taking into account the last inequality together with (15), we find

$$\begin{aligned}
\|s^n - t^n\|^2 &\geq \frac{\rho_n^2 \|w^n - u^n\|^4}{(\|A\| + 1)^2 \|w^n - u^n\|^2} = \frac{\rho_n^2}{(\|A\| + 1)^2} \|w^n - u^n\|^2 \\
&\geq \frac{a^2}{(\|A\| + 1)^2} \|w^n - u^n\|^2.
\end{aligned}$$

□

**Lemma 7.** Let  $\{x^n\}$ ,  $\{y^n\}$ ,  $\{t^n\}$  and  $\{s^n\}$  be the sequences generated by Algorithm 1. Then the sequences  $\{x^n\}$ ,  $\{y^n\}$ ,  $\{t^n\}$ ,  $\{s^n\}$ , and  $\{F(s^n)\}$  are bounded.

*Proof.* By the  $\eta$ -strong monotonicity and the  $L$ -Lipschitz continuity of  $F$  on  $\mathcal{H}_1$ , we have

$$\begin{aligned}
&\|s^n - x^* - \varepsilon(F(s^n) - F(x^*))\|^2 \\
&= \|s^n - x^*\|^2 - 2\varepsilon \langle s^n - x^*, F(s^n) - F(x^*) \rangle + \varepsilon^2 \|F(s^n) - F(x^*)\|^2 \\
&\leq \|s^n - x^*\|^2 - 2\varepsilon \eta \|s^n - x^*\|^2 + \varepsilon^2 L^2 \|s^n - x^*\|^2
\end{aligned}$$

$$= [1 - \varepsilon(2\eta - \varepsilon L^2)] \|s^n - x^*\|^2. \quad (16)$$

From (16), we obtain, for all  $n \geq n_0$ ,

$$\begin{aligned} & \|s^n - \varepsilon_n F(s^n) - (x^* - \varepsilon_n F(x^*))\| \\ &= \|(s^n - x^*) - \varepsilon_n (F(s^n) - F(x^*))\| \\ &= \left\| \left(1 - \frac{\varepsilon_n}{\varepsilon}\right) (s^n - x^*) + \frac{\varepsilon_n}{\varepsilon} [s^n - x^* - \varepsilon(F(s^n) - F(x^*))] \right\| \\ &\leq \left(1 - \frac{\varepsilon_n}{\varepsilon}\right) \|s^n - x^*\| + \frac{\varepsilon_n}{\varepsilon} \|s^n - x^* - \varepsilon(F(s^n) - F(x^*))\| \\ &\leq \left(1 - \frac{\varepsilon_n}{\varepsilon}\right) \|s^n - x^*\| + \frac{\varepsilon_n}{\varepsilon} \sqrt{1 - \varepsilon(2\eta - \varepsilon L^2)} \|s^n - x^*\| \\ &= \left[1 - \frac{\varepsilon_n}{\varepsilon} \left(1 - \sqrt{1 - \varepsilon(2\eta - \varepsilon L^2)}\right)\right] \|s^n - x^*\| \\ &= \left(1 - \frac{\varepsilon_n^\tau}{\varepsilon}\right) \|s^n - x^*\|, \end{aligned} \quad (17)$$

where

$$\tau = 1 - \sqrt{1 - \varepsilon(2\eta - \varepsilon L^2)} \in (0, 1].$$

Alternatively, we have

$$0 \leq \alpha_n \|x^n - x^{n-1}\| \leq \eta_n, \quad \text{for all } n \geq 0 \quad (18)$$

and

$$\lim_{n \rightarrow \infty} \frac{\alpha_n}{\varepsilon_n} \|x^n - x^{n-1}\| = 0. \quad (19)$$

Indeed, if  $x^n = x^{n-1}$ , then inequality (18) holds. Otherwise, we get

$$\begin{aligned} 0 \leq \alpha_n &= \min \left\{ \frac{\eta_n}{\|x^n - x^{n-1}\|}, \gamma_n \right\} \leq \frac{\eta_n}{\|x^n - x^{n-1}\|} \\ &\Rightarrow 0 \leq \alpha_n \|x^n - x^{n-1}\| \leq \eta_n. \end{aligned}$$

From (18), we have

$$0 \leq \frac{\alpha_n}{\varepsilon_n} \|x^n - x^{n-1}\| \leq \frac{\eta_n}{\varepsilon_n}, \quad \text{for all } n \geq 0.$$

Since  $\lim_{n \rightarrow \infty} \frac{\eta_n}{\varepsilon_n} = 0$ , it can be inferred from the above inequality that  $\lim_{n \rightarrow \infty} \frac{\alpha_n}{\varepsilon_n} \|x^n - x^{n-1}\| = 0$ .

Using a similar argument, we arrive at

$$0 \leq \beta_n \|x^{n-2} - x^{n-1}\| \leq \eta_n, \quad \text{for all } n \geq 0 \quad (20)$$

and

$$\lim_{n \rightarrow \infty} \frac{\beta_n}{\varepsilon_n} \|x^{n-2} - x^{n-1}\| = 0. \quad (21)$$

From  $\lim_{n \rightarrow \infty} \frac{\alpha_n}{\varepsilon_n} \|x^n - x^{n-1}\| = 0$  and  $\lim_{n \rightarrow \infty} \frac{\beta_n}{\varepsilon_n} \|x^{n-2} - x^{n-1}\| = 0$ , we can infer that there exist positive constants  $K_1$  and  $K_2$  such that  $\frac{\alpha_n}{\varepsilon_n} \|x^n - x^{n-1}\| \leq K_1$  and  $\frac{\beta_n}{\varepsilon_n} \|x^{n-2} - x^{n-1}\| \leq K_2$  for all  $n \geq 0$ . So, we have

$$\begin{aligned} \|y^n - x^*\| &= \|(x^n - x^*) + \alpha_n(x^n - x^{n-1}) + \beta_n(x^{n-2} - x^{n-1})\| \\ &\leq \|x^n - x^*\| + \varepsilon_n \cdot \frac{\alpha_n}{\varepsilon_n} \|x^n - x^{n-1}\| + \varepsilon_n \cdot \frac{\beta_n}{\varepsilon_n} \|x^{n-2} - x^{n-1}\| \\ &\leq \|x^n - x^*\| + \varepsilon_n K_1 + \varepsilon_n K_2 \\ &= \|x^n - x^*\| + \varepsilon_n K_3, \quad \text{for all } n \geq 0, \end{aligned} \quad (22)$$

where  $K_3 = K_1 + K_2$ .

From Lemma 6, (12) and (22), we get

$$\|s^n - x^*\| \leq \|t^n - x^*\| \leq \|y^n - x^*\| \leq \|x^n - x^*\| + \varepsilon_n K_3, \quad \text{for all } n \geq n_0. \quad (23)$$

Employing (17) and (23), we derive, for all  $n \geq n_0$ ,

$$\begin{aligned} \|x^{n+1} - x^*\| &= \|s^n - \varepsilon_n F(s^n) - (x^* - \varepsilon_n F(x^*)) - \varepsilon_n F(x^*)\| \\ &\leq \|s^n - \varepsilon_n F(s^n) - (x^* - \varepsilon_n F(x^*))\| + \varepsilon_n \|F(x^*)\| \\ &\leq \left(1 - \frac{\varepsilon_n \tau}{\varepsilon}\right) \|s^n - x^*\| + \varepsilon_n \|F(x^*)\| \end{aligned} \quad (24)$$

$$\begin{aligned} &\leq \left(1 - \frac{\varepsilon_n \tau}{\varepsilon}\right) (\|x^n - x^*\| + \varepsilon_n K_3) + \varepsilon_n \|F(x^*)\| \\ &\leq \left(1 - \frac{\varepsilon_n \tau}{\varepsilon}\right) \|x^n - x^*\| + \varepsilon_n K_3 + \varepsilon_n \|F(x^*)\| \\ &= \left(1 - \frac{\varepsilon_n \tau}{\varepsilon}\right) \|x^n - x^*\| + \frac{\varepsilon_n \tau}{\varepsilon} \cdot \frac{\varepsilon (K_3 + \|F(x^*)\|)}{\tau}. \end{aligned} \quad (25)$$

From (25), we have, for every  $n \geq n_0$ ,

$$\|x^{n+1} - x^*\| \leq \max \left\{ \|x^n - x^*\|, \frac{\varepsilon (K_3 + \|F(x^*)\|)}{\tau} \right\}.$$

So, by induction, we obtain

$$\|x^n - x^*\| \leq \max \left\{ \|x^{n_0} - x^*\|, \frac{\varepsilon(K_3 + \|F(x^*)\|)}{\tau} \right\}, \text{ for all } n \geq n_0.$$

Therefore, the sequence  $\{x^n\}$  is bounded, and so are the sequences  $\{y^n\}$ ,  $\{t^n\}$ ,  $\{s^n\}$ , and  $\{F(s^n)\}$  due to  $\{\varepsilon_n\} \subset (0, 1)$ , (23), and the Lipschitz continuity of  $F$ .  $\square$

**Lemma 8.** Let  $\{x^n\}$  be the sequence generated by Algorithm 1. Then, there exists a constant  $K > 0$  such that for all  $n \geq n_0$ , we have

$$\begin{aligned} \|x^{n+1} - x^*\|^2 &\leq \left(1 - \frac{\varepsilon_n \tau}{\varepsilon}\right) \|x^n - x^*\|^2 + 2\varepsilon_n \langle F(x^*), x^* - x^{n+1} \rangle \\ &\quad + K\alpha_n \|x^n - x^{n-1}\| + K\beta_n \|x^{n-2} - x^{n-1}\|. \end{aligned}$$

*Proof.* From (23), we have

$$\begin{aligned} \|y^n - x^*\|^2 &\leq (\|x^n - x^*\| + \varepsilon_n K_3)^2 \\ &= \|x^n - x^*\|^2 + \varepsilon_n (2K_3 \|x^n - x^*\| + \varepsilon_n K_3^2) \\ &\leq \|x^n - x^*\|^2 + \varepsilon_n K_4, \end{aligned} \tag{26}$$

where  $K_4 = \sup_{n \geq 0} \{2K_3 \|x^n - x^*\| + \varepsilon_n K_3^2\}$ .

From  $\frac{\alpha_n}{\varepsilon_n} \|x^n - x^{n-1}\| \leq K_1$ ,  $\frac{\beta_n}{\varepsilon_n} \|x^{n-2} - x^{n-1}\| \leq K_2$  for all  $n \geq 0$  and  $\{\varepsilon_n\} \subset (0, 1)$ , we deduce that  $\alpha_n \|x^n - x^{n-1}\| \leq K_1$ ,  $\beta_n \|x^{n-2} - x^{n-1}\| \leq K_2$  for all  $n \geq 0$ . This, in conjunction with the boundedness of the sequence  $\{x^n\}$ , implies the existence of a constant  $K > 0$  such that

$$2\|x^n - x^*\| + \alpha_n \|x^n - x^{n-1}\| + \beta_n \|x^{n-2} - x^{n-1}\| \leq K, \text{ for all } n \geq 0. \tag{27}$$

From (27), we get

$$\begin{aligned} \|y^n - x^*\|^2 &= \|(x^n - x^*) + \alpha_n(x^n - x^{n-1}) + \beta_n(x^{n-2} - x^{n-1})\|^2 \\ &= \|x^n - x^*\|^2 + 2\alpha_n \langle x^n - x^*, x^n - x^{n-1} \rangle \\ &\quad + 2\beta_n \langle x^n - x^*, x^{n-2} - x^{n-1} \rangle \\ &\quad + \|\alpha_n(x^n - x^{n-1}) + \beta_n(x^{n-2} - x^{n-1})\|^2 \\ &\leq \|x^n - x^*\|^2 + 2\alpha_n \|x^n - x^*\| \cdot \|x^n - x^{n-1}\| \\ &\quad + 2\beta_n \|x^n - x^*\| \cdot \|x^{n-2} - x^{n-1}\| \\ &\quad + \alpha_n^2 \|x^n - x^{n-1}\|^2 + 2\alpha_n \beta_n \|x^n - x^{n-1}\| \cdot \|x^{n-2} - x^{n-1}\| \end{aligned}$$

$$\begin{aligned}
& + \beta_n^2 \|x^{n-2} - x^{n-1}\|^2 \\
& = \|x^n - x^*\|^2 + \alpha_n \|x^n - x^{n-1}\| (2\|x^n - x^*\| \\
& \quad + \alpha_n \|x^n - x^{n-1}\| + \beta_n \|x^{n-2} - x^{n-1}\|) \\
& \quad + \beta_n \|x^{n-2} - x^{n-1}\| (2\|x^n - x^*\| + \alpha_n \|x^n - x^{n-1}\| \\
& \quad + \beta_n \|x^{n-2} - x^{n-1}\|) \\
& \leq \|x^n - x^*\|^2 + K\alpha_n \|x^n - x^{n-1}\| + K\beta_n \|x^{n-2} - x^{n-1}\|. \quad (28)
\end{aligned}$$

From (17), (23), and (28), we obtain, for all  $n \geq n_0$ ,

$$\begin{aligned}
\|x^{n+1} - x^*\|^2 & \leq \|x^{n+1} - x^*\|^2 + \varepsilon_n^2 \|F(x^*)\|^2 \\
& = \|x^{n+1} - x^* + \varepsilon_n F(x^*)\|^2 - 2\langle \varepsilon_n F(x^*), x^{n+1} - x^* \rangle \\
& = \|s^n - \varepsilon_n F(s^n) - (x^* - \varepsilon_n F(x^*))\|^2 - 2\varepsilon_n \langle F(x^*), x^{n+1} - x^* \rangle \\
& \leq \left[ \left(1 - \frac{\varepsilon_n \tau}{\varepsilon}\right) \|s^n - x^*\| \right]^2 - 2\varepsilon_n \langle F(x^*), x^{n+1} - x^* \rangle \\
& \leq \left(1 - \frac{\varepsilon_n \tau}{\varepsilon}\right) \|s^n - x^*\|^2 - 2\varepsilon_n \langle F(x^*), x^{n+1} - x^* \rangle \quad (29) \\
& \leq \left(1 - \frac{\varepsilon_n \tau}{\varepsilon}\right) \|y^n - x^*\|^2 + 2\varepsilon_n \langle F(x^*), x^* - x^{n+1} \rangle \\
& \leq \left(1 - \frac{\varepsilon_n \tau}{\varepsilon}\right) (\|x^n - x^*\|^2 + K\alpha_n \|x^n - x^{n-1}\| \\
& \quad + K\beta_n \|x^{n-2} - x^{n-1}\|) + 2\varepsilon_n \langle F(x^*), x^* - x^{n+1} \rangle \\
& \leq \left(1 - \frac{\varepsilon_n \tau}{\varepsilon}\right) \|x^n - x^*\|^2 + 2\varepsilon_n \langle F(x^*), x^* - x^{n+1} \rangle \\
& \quad + K\alpha_n \|x^n - x^{n-1}\| + K\beta_n \|x^{n-2} - x^{n-1}\|.
\end{aligned}$$

□

The theorem presented here establishes the validity and convergence of Algorithm 1.

**Theorem 1.** Assume that Assumption 1 is satisfied. Then the sequence  $\{x^n\}$  generated by Algorithm 1 converges strongly to the unique solution of the BSVIP (4), provided the solution set  $\Omega_{\text{SVIP}} = \{x^* \in \text{Sol}(C, F_1) : Ax^* \in \text{Sol}(Q, F_2)\}$  of the SVIP (1)–(2) is nonempty.

*Proof.* We prove that the sequence  $\{x^n\}$  converges strongly to the unique solution  $x^*$  of the problem (4). Let us consider two cases.

**Case 1.** There exists  $n_1 \in \mathbb{N}$  such that  $\{\|x^n - x^*\|\}$  is decreasing for all  $n \geq n_1$ . Consequently, the limit of  $\|x^n - x^*\|$  exists. Therefore, it follows from (23), (26), and (29), for all  $n \geq n_0$ , that

$$\begin{aligned} -\varepsilon_n K_4 &\leq \|y^n - x^*\|^2 - \|s^n - x^*\|^2 - \varepsilon_n K_4 \\ &\leq \|x^n - x^*\|^2 - \|s^n - x^*\|^2 \\ &\leq (\|x^n - x^*\|^2 - \|x^{n+1} - x^*\|^2) - \frac{\varepsilon_n \tau}{\varepsilon} \|s^n - x^*\|^2 \\ &\quad - 2\varepsilon_n \langle F(x^*), x^{n+1} - x^* \rangle. \end{aligned}$$

Given the limit of  $\|x^n - x^*\|$  exists, along with  $\lim_{n \rightarrow \infty} \varepsilon_n = 0$ , and both  $\{x^n\}$  and  $\{s^n\}$  being bounded sequences, the above inequalities imply that

$$\begin{aligned} \lim_{n \rightarrow \infty} (\|y^n - x^*\|^2 - \|s^n - x^*\|^2 - \varepsilon_n K_4) &= 0 \\ \Rightarrow \lim_{n \rightarrow \infty} (\|y^n - x^*\|^2 - \|s^n - x^*\|^2) &= 0, \quad (30) \\ \lim_{n \rightarrow \infty} (\|x^n - x^*\|^2 - \|s^n - x^*\|^2) &= 0. \quad (31) \end{aligned}$$

From (23), we get

$$0 \leq \|y^n - x^*\|^2 - \|t^n - x^*\|^2 \leq \|y^n - x^*\|^2 - \|s^n - x^*\|^2, \quad \text{for all } n \geq 0,$$

from which, by (30), it follows that

$$\lim_{n \rightarrow \infty} (\|y^n - x^*\|^2 - \|t^n - x^*\|^2) = 0. \quad (32)$$

From Lemma 5 and (10), we have

$$\frac{1 - \lambda^2}{2} \|y^n - z^n\|^2 \leq \|y^n - x^*\|^2 - \|t^n - x^*\|^2, \quad \text{for all } n \geq n_0,$$

which together with (32) implies

$$\lim_{n \rightarrow \infty} \|y^n - z^n\| = 0. \quad (33)$$

From (30) and (32), it follows that

$$\lim_{n \rightarrow \infty} (\|t^n - x^*\|^2 - \|s^n - x^*\|^2) = 0.$$

Hence, by combining Lemma 6, we obtain

$$\lim_{n \rightarrow \infty} \|w^n - u^n\| = 0, \quad (34)$$

$$\lim_{n \rightarrow \infty} \|s^n - t^n\| = 0. \quad (35)$$

Using the triangle inequality and the  $L_1$ -Lipschitz continuity of  $F_1$  on  $\mathcal{H}_1$ , we get

$$\begin{aligned} \|y^n - s^n\| &\leq \|y^n - z^n\| + \|z^n - t^n\| + \|t^n - s^n\| \\ &= \|y^n - z^n\| + \|\lambda_n(F_1(z^n) - F_1(y^n))\| + \|t^n - s^n\| \\ &\leq \|y^n - z^n\| + \lambda_n L_1 \|z^n - y^n\| + \|t^n - s^n\| \\ &\leq (1 + \lambda_0 L_1) \|y^n - z^n\| + \|t^n - s^n\|, \end{aligned}$$

which together with (33), (35) implies

$$\lim_{n \rightarrow \infty} \|y^n - s^n\| = 0. \quad (36)$$

Now, observe that

$$\begin{aligned} \|w^n - Ax^*\|^2 &= \|u^n - Ax^* + (w^n - u^n)\|^2 \\ &= \|u^n - Ax^*\|^2 + 2\langle u^n - Ax^*, w^n - u^n \rangle + \|w^n - u^n\|^2 \\ &= \|u^n - Ax^*\|^2 + 2\langle A(t^n - x^*), w^n - u^n \rangle + \|w^n - u^n\|^2 \\ &\geq \|u^n - Ax^*\|^2 - 2\|A(t^n - x^*)\| \|w^n - u^n\| + \|w^n - u^n\|^2 \\ &\geq \|u^n - Ax^*\|^2 - 2\|A\| \|t^n - x^*\| \|w^n - u^n\|. \end{aligned} \quad (37)$$

Combining Lemma 5, (9) and (37) yields

$$\frac{1 - \mu^2}{2} \|u^n - v^n\|^2 \leq 2\|A\| \|t^n - x^*\| \|w^n - u^n\|, \quad \text{for all } n \geq n_0. \quad (38)$$

From (34), (38), and the boundedness of the sequence  $\{t^n\}$ , we obtain

$$\lim_{n \rightarrow \infty} \|u^n - v^n\| = 0. \quad (39)$$

We now prove that

$$\limsup_{n \rightarrow \infty} \langle F(x^*), x^* - s^n \rangle \leq 0. \quad (40)$$

Select a subsequence  $\{s^{n_k}\}$  of  $\{s^n\}$  such that  $\limsup_{n \rightarrow \infty} \langle F(x^*), x^* - s^n \rangle = \lim_{k \rightarrow \infty} \langle F(x^*), x^* - s^{n_k} \rangle$ . Given that  $\{s^{n_k}\}$  is bounded, we may assume that  $\{s^{n_k}\}$  converges weakly to some  $\bar{s} \in \mathcal{H}_1$ .

Therefore

$$\limsup_{n \rightarrow \infty} \langle F(x^*), x^* - s^n \rangle = \lim_{k \rightarrow \infty} \langle F(x^*), x^* - s^{n_k} \rangle = \langle F(x^*), x^* - \bar{s} \rangle. \quad (41)$$

We deduce from  $s^{n_k} \rightharpoonup \bar{s}$  and (35), (36) that  $t^{n_k} \rightharpoonup \bar{s}$  and  $y^{n_k} \rightharpoonup \bar{s}$ . From (33), we have  $\lim_{k \rightarrow \infty} \|y^{n_k} - z^{n_k}\| = 0$ . Since  $z^{n_k} = P_C(y^{n_k} - \lambda_{n_k} F_1(y^{n_k}))$ ,  $y^{n_k} \rightharpoonup \bar{s}$ ,  $\lambda_{n_k} \geq \min\left(\frac{\lambda}{L_1}, \lambda_0\right) > 0$ . By Lemma 2, we get  $\bar{s} \in \text{Sol}(C, F_1)$ .

From  $t^{n_k} \rightharpoonup \bar{s}$ , we get  $u^{n_k} = A(t^{n_k}) \rightharpoonup A(\bar{s})$ . This, together with (39), where  $v^{n_k} = P_Q(u^{n_k} - \mu_{n_k} F_2(u^{n_k}))$  and  $\mu_{n_k} \geq \min\left(\frac{\mu}{L_2}, \mu_0\right) > 0$ , along with Lemma 2, implies that  $A(\bar{s}) \in \text{Sol}(Q, F_2)$ .

With  $\bar{s} \in \text{Sol}(C, F_1)$  and  $A(\bar{s}) \in \text{Sol}(Q, F_2)$ , we conclude that  $\bar{s} \in \Omega_{\text{SVIP}}$ . Consequently, it follows from  $x^* \in \text{Sol}(\Omega_{\text{SVIP}}, F)$  that  $\langle F(x^*), \bar{s} - x^* \rangle \geq 0$ , which together with (41) implies (40).

From the boundedness of  $\{F(s^n)\}$ ,  $\lim_{n \rightarrow \infty} \varepsilon_n = 0$  and (40), we have

$$\begin{aligned} \limsup_{n \rightarrow \infty} \langle F(x^*), x^* - x^{n+1} \rangle &= \limsup_{n \rightarrow \infty} \langle F(x^*), x^* - s^n + \varepsilon_n F(s^n) \rangle \\ &= \limsup_{n \rightarrow \infty} \left[ \langle F(x^*), x^* - s^n \rangle + \varepsilon_n \langle F(x^*), F(s^n) \rangle \right] \\ &= \limsup_{n \rightarrow \infty} \langle F(x^*), x^* - s^n \rangle \leq 0. \end{aligned} \quad (42)$$

From Lemma 8, we get

$$\|x^{n+1} - x^*\|^2 \leq (1 - a_n) \|x^n - x^*\|^2 + a_n b_n, \quad \text{for all } n \geq n_0, \quad (43)$$

where  $a_n = \frac{\varepsilon_n \tau}{\varepsilon}$  and

$$b_n = \frac{2\varepsilon \langle F(x^*), x^* - x^{n+1} \rangle}{\tau} + \frac{K\varepsilon}{\tau} \cdot \frac{\alpha_n}{\varepsilon_n} \|x^n - x^{n-1}\| + \frac{K\varepsilon}{\tau} \cdot \frac{\beta_n}{\varepsilon_n} \|x^{n-2} - x^{n-1}\|.$$

Given (19), (21), and (42), it follows that  $\limsup b_n \leq 0$ . From  $0 < \varepsilon_n < \varepsilon$  for all  $n \geq n_0$  and  $0 < \tau \leq 1$ , we get  $\left\{a_n = \frac{\varepsilon_n \tau}{\varepsilon}\right\}_{n \geq n_0} \subset (0, 1)$ . So, from (43),  $\sum_{n=0}^{\infty} \varepsilon_n = \infty$ ,  $\limsup_{n \rightarrow \infty} b_n \leq 0$  and Lemma 4, we have  $\lim_{n \rightarrow \infty} \|x^n - x^*\|^2 = 0$ , that is,  $x^n \rightarrow x^*$  as  $n \rightarrow \infty$ .

**Case 2.** Suppose that for any integer  $m$ , there exists an integer  $n$  such that  $n \geq m$  and  $\|x^n - x^*\| \leq \|x^{n+1} - x^*\|$ . In this situation, it follows from Lemma 3 that there exists a nondecreasing sequence  $\{\tau(n)\}_{n \geq n_2}$  of  $\mathbb{N}$  such

that  $\lim_{n \rightarrow \infty} \tau(n) = \infty$  and the following inequalities are true:

$$\|x^{\tau(n)} - x^*\| \leq \|x^{\tau(n)+1} - x^*\|, \quad \|x^n - x^*\| \leq \|x^{\tau(n)+1} - x^*\|, \quad \text{for all } n \geq n_2. \quad (44)$$

Choose  $n_3 \geq n_2$  such that  $\tau(n) \geq n_0$  for all  $n \geq n_3$ . From (23), (44) and (24), we get, for all  $n \geq n_3$ ,

$$\begin{aligned} -\varepsilon_{\tau(n)} K_3 &\leq \|y^{\tau(n)} - x^*\| - \|s^{\tau(n)} - x^*\| - \varepsilon_{\tau(n)} K_3 \\ &\leq \|x^{\tau(n)} - x^*\| - \|s^{\tau(n)} - x^*\| \\ &\leq \|x^{\tau(n)+1} - x^*\| - \|s^{\tau(n)} - x^*\| \\ &\leq -\frac{\varepsilon_{\tau(n)} \tau}{\varepsilon} \|s^{\tau(n)} - x^*\| + \varepsilon_{\tau(n)} \|F(x^*)\|. \end{aligned}$$

Thus, from the boundedness of  $\{s^n\}$  and  $\lim_{n \rightarrow \infty} \varepsilon_n = 0$ , we have

$$\begin{aligned} \lim_{n \rightarrow \infty} (\|y^{\tau(n)} - x^*\| - \|s^{\tau(n)} - x^*\| - \varepsilon_{\tau(n)} K_3) &= 0 \\ \Rightarrow \lim_{n \rightarrow \infty} (\|y^{\tau(n)} - x^*\| - \|s^{\tau(n)} - x^*\|) &= 0, \end{aligned} \quad (45)$$

$$\lim_{n \rightarrow \infty} (\|x^{\tau(n)} - x^*\| - \|s^{\tau(n)} - x^*\|) = 0. \quad (46)$$

From (45), (46), and the boundedness of  $\{x^n\}$ ,  $\{y^n\}$ ,  $\{s^n\}$ , we obtain

$$\lim_{n \rightarrow \infty} (\|y^{\tau(n)} - x^*\|^2 - \|s^{\tau(n)} - x^*\|^2) = 0, \quad \lim_{n \rightarrow \infty} (\|x^{\tau(n)} - x^*\|^2 - \|s^{\tau(n)} - x^*\|^2) = 0.$$

Applying a similar line of reasoning as in the first case, we can arrive at the conclusion that

$$\limsup_{n \rightarrow \infty} \langle F(x^*), x^* - s^{\tau(n)} \rangle \leq 0.$$

Therefore, the boundedness of  $\{F(s^n)\}$  and  $\lim_{n \rightarrow \infty} \varepsilon_n = 0$  yield

$$\begin{aligned} \limsup_{n \rightarrow \infty} \langle F(x^*), x^* - x^{\tau(n)+1} \rangle &= \limsup_{n \rightarrow \infty} \langle F(x^*), x^* - s^{\tau(n)} + \varepsilon_{\tau(n)} F(s^{\tau(n)}) \rangle \\ &= \limsup_{n \rightarrow \infty} \left[ \langle F(x^*), x^* - s^{\tau(n)} \rangle \right. \\ &\quad \left. + \varepsilon_{\tau(n)} \langle F(x^*), F(s^{\tau(n)}) \rangle \right] \\ &= \limsup_{n \rightarrow \infty} \langle F(x^*), x^* - s^{\tau(n)} \rangle \leq 0. \end{aligned} \quad (47)$$

From Lemma 8 and (44), we have, for all  $n \geq n_3$ ,

$$\begin{aligned}
\|x^{\tau(n)+1} - x^*\|^2 &\leq \left(1 - \frac{\varepsilon_{\tau(n)}^\tau}{\varepsilon}\right) \|x^{\tau(n)} - x^*\|^2 + 2\varepsilon_{\tau(n)} \langle F(x^*), x^* - x^{\tau(n)+1} \rangle \\
&\quad + K\alpha_{\tau(n)} \|x^{\tau(n)} - x^{\tau(n)-1}\| + K\beta_{\tau(n)} \|x^{\tau(n)-2} - x^{\tau(n)-1}\| \\
&\leq \left(1 - \frac{\varepsilon_{\tau(n)}^\tau}{\varepsilon}\right) \|x^{\tau(n)+1} - x^*\|^2 + 2\varepsilon_{\tau(n)} \langle F(x^*), x^* - x^{\tau(n)+1} \rangle \\
&\quad + K\alpha_{\tau(n)} \|x^{\tau(n)} - x^{\tau(n)-1}\| + K\beta_{\tau(n)} \|x^{\tau(n)-2} - x^{\tau(n)-1}\|.
\end{aligned}$$

In particular, since  $\varepsilon_{\tau(n)} > 0$ , we have, for all  $n \geq n_3$

$$\begin{aligned}
\|x^{\tau(n)+1} - x^*\|^2 &\leq \frac{2\varepsilon}{\tau} \langle F(x^*), x^* - x^{\tau(n)+1} \rangle + \frac{K\varepsilon}{\tau} \cdot \frac{\alpha_{\tau(n)}}{\varepsilon_{\tau(n)}} \|x^{\tau(n)} - x^{\tau(n)-1}\| \\
&\quad + \frac{K\varepsilon}{\tau} \cdot \frac{\beta_{\tau(n)}}{\varepsilon_{\tau(n)}} \|x^{\tau(n)-2} - x^{\tau(n)-1}\|.
\end{aligned}$$

From (44) and the inequality given above, we derive, for all  $n \geq n_3$ ,

$$\begin{aligned}
\|x^n - x^*\|^2 &\leq \frac{2\varepsilon}{\tau} \langle F(x^*), x^* - x^{\tau(n)+1} \rangle + \frac{K\varepsilon}{\tau} \cdot \frac{\alpha_{\tau(n)}}{\varepsilon_{\tau(n)}} \|x^{\tau(n)} - x^{\tau(n)-1}\| \\
&\quad + \frac{K\varepsilon}{\tau} \cdot \frac{\beta_{\tau(n)}}{\varepsilon_{\tau(n)}} \|x^{\tau(n)-2} - x^{\tau(n)-1}\|. \tag{48}
\end{aligned}$$

By taking the limit in (48) as  $n \rightarrow \infty$  and utilizing (47), (19), and (21), we deduce that

$$\limsup_{n \rightarrow \infty} \|x^n - x^*\|^2 \leq 0,$$

which implies  $x^n \rightarrow x^*$ .  $\square$

From Algorithm 1, if we choose  $\gamma_n = 0$  and  $\xi_n = 0$  for all  $n \geq 0$ , it is evident that  $\alpha_n = 0$  and  $\beta_n = 0$  for all  $n \geq 0$ . In this case, Algorithm 1 reduces to the following algorithm. This algorithm, which we will refer to as Algorithm 2, closely resembles [25, Algorithm 3.1], as described in (8). The key difference between the two algorithms lies in the order in which the modified Tseng's extragradient method is applied in the two spaces  $\mathcal{H}_1$  and  $\mathcal{H}_2$ .

**Assumption 2.** Let the following hold

- i)  $F : \mathcal{H} \longrightarrow \mathcal{H}$  is strongly monotone and Lipschitz continuous on  $\mathcal{H}$ .
- ii)  $G : \mathcal{H} \longrightarrow \mathcal{H}$  is pseudomonotone on  $C$ , Lipschitz continuous on  $\mathcal{H}$ .

**Algorithm 2**

**Step 0.** Choose  $\mu_0 > 0$ ,  $\lambda_0 > 0$ ,  $\mu \in (0, 1)$ ,  $\lambda \in (0, 1)$ ,  $\{\rho_n\} \subset [a, b] \subset (0, 1)$ ,  $\{\varepsilon_n\} \subset (0, 1)$  such that  $\lim_{n \rightarrow \infty} \varepsilon_n = 0$ ,  $\sum_{n=0}^{\infty} \varepsilon_n = \infty$ .

**Step 1.** Let  $x^0 \in \mathcal{H}_1$ . Set  $n := 0$ .

**Step 2.** Compute

$$\begin{aligned} y^n &= P_C(x^n - \lambda_n F_1(x^n)), \\ z^n &= y^n - \lambda_n (F_1(y^n) - F_1(x^n)), \end{aligned}$$

where

$$\lambda_{n+1} = \begin{cases} \min \left\{ \frac{\lambda \|x^n - y^n\|}{\|F_1(x^n) - F_1(y^n)\|}, \lambda_n \right\} & \text{if } F_1(x^n) \neq F_1(y^n), \\ \lambda_n & \text{if } F_1(x^n) = F_1(y^n). \end{cases}$$

**Step 3.** Compute  $u^n = A(z^n)$  and

$$\begin{aligned} v^n &= P_Q(u^n - \mu_n F_2(u^n)), \\ w^n &= v^n - \mu_n (F_2(v^n) - F_2(u^n)), \end{aligned}$$

where

$$\mu_{n+1} = \begin{cases} \min \left\{ \frac{\mu \|u^n - v^n\|}{\|F_2(u^n) - F_2(v^n)\|}, \mu_n \right\} & \text{if } F_2(u^n) \neq F_2(v^n), \\ \mu_n & \text{if } F_2(u^n) = F_2(v^n). \end{cases}$$

**Step 4.** Compute

$$t^n = z^n + \delta_n A^*(w^n - u^n),$$

where the stepsize  $\delta_n$  is chosen in such a way that

$$\delta_n = \begin{cases} \frac{\rho_n \|w^n - u^n\|^2}{\|A^*(w^n - u^n)\|^2} & \text{if } A^*(w^n - u^n) \neq 0, \\ 0 & \text{if } A^*(w^n - u^n) = 0. \end{cases}$$

**Step 5.** Compute

$$x^{n+1} = t^n - \varepsilon_n F(t^n).$$

**Step 6.** Set  $n := n + 1$ , and go to **Step 2**.

- iii)  $\limsup_{n \rightarrow \infty} \langle G(x^n), y - y^n \rangle \leq \langle G(\bar{x}), y - \bar{y} \rangle$  holds for any sequences  $\{x^n\}$  and  $\{y^n\}$  in  $\mathcal{H}$  that converge weakly to  $\bar{x}$  and  $\bar{y}$ , respectively.

When  $F_2 = 0$  and  $Q = \mathcal{H}_2$ , the SVIP defined by (1) and (2) reduces to the VIP given by (1). Consequently, according to Algorithm 1 and Theorem 1 (where  $\mathcal{H}_1 = \mathcal{H}$  and  $F_1 = G$ ), we obtain the following result for solving the BVIP specified by (5). It is important to note that the proposed algorithm only requires a single projection onto the feasible set at each iteration

and does not necessitate any knowledge of the Lipschitz constants for the mappings  $F$  and  $G$ , nor the modulus of strong monotonicity of  $F$ .

---

**Algorithm 3**


---

**Step 0.** Choose  $\lambda_0 > 0$ ,  $\lambda \in (0, 1)$ ,  $\{\gamma_n\} \subset [0, \infty)$ ,  $\{\xi_n\} \subset [0, \infty)$ ,  $\{\eta_n\} \subset (0, \infty)$ ,  $\{\varepsilon_n\} \subset (0, 1)$  such that  $\lim_{n \rightarrow \infty} \frac{\eta_n}{\varepsilon_n} = 0$ ,  $\lim_{n \rightarrow \infty} \varepsilon_n = 0$ ,  $\sum_{n=0}^{\infty} \varepsilon_n = \infty$ .

**Step 1.** Let  $x^{-2}, x^{-1}, x^0 \in \mathcal{H}$ . Set  $n := 0$ .

**Step 2.** Compute  $y^n = x^n + \alpha_n(x^n - x^{n-1}) + \beta_n(x^{n-2} - x^{n-1})$ , where

$$\alpha_n = \begin{cases} \min \left\{ \frac{\eta_n}{\|x^n - x^{n-1}\|}, \gamma_n \right\} & \text{if } x^n \neq x^{n-1}, \\ \gamma_n & \text{if } x^n = x^{n-1}, \end{cases}$$

and

$$\beta_n = \begin{cases} \min \left\{ \frac{\eta_n}{\|x^{n-2} - x^{n-1}\|}, \xi_n \right\} & \text{if } x^{n-2} \neq x^{n-1}, \\ \xi_n & \text{if } x^{n-2} = x^{n-1}. \end{cases}$$

**Step 3.** Compute

$$z^n = P_C(y^n - \lambda_n G(y^n)),$$

$$t^n = z^n - \lambda_n(G(z^n) - G(y^n)),$$

where

$$\lambda_{n+1} = \begin{cases} \min \left\{ \frac{\lambda \|y^n - z^n\|}{\|G(y^n) - G(z^n)\|}, \lambda_n \right\} & \text{if } G(y^n) \neq G(z^n), \\ \lambda_n & \text{if } G(y^n) = G(z^n). \end{cases}$$

**Step 6.** Compute

$$x^{n+1} = t^n - \varepsilon_n F(t^n).$$

**Step 7.** Set  $n := n + 1$ , and go to **Step 2**.

---

**Corollary 1.** Suppose that Assumption 2 holds. Then the sequence  $\{x^n\}$  generated by Algorithm 3 converges strongly to the unique solution of the BVIP (5), provided the solution set  $\text{Sol}(C, G)$  of the VIP (6) is nonempty.

**Assumption 3.** Consider the functions  $f$  and  $g$  which satisfy the following conditions:

- i)  $f : \mathbb{R}^n \rightarrow \mathbb{R}$  is continuously differentiable and strongly convex, and its gradient is Lipschitz continuous.
- ii)  $g : \mathbb{R}^n \rightarrow \mathbb{R}$  is convex and continuously differentiable such that its gradient is Lipschitz continuous.

Assuming that all conditions stated in Assumption 3 are satisfied, we find that the gradient mapping  $\nabla f : \mathbb{R}^n \rightarrow \mathbb{R}^n$  is strongly monotone and Lipschitz continuous on  $\mathbb{R}^n$ . Similarly,  $\nabla g : \mathbb{R}^n \rightarrow \mathbb{R}^n$  is monotone and Lipschitz continuous on  $\mathbb{R}^n$ . By taking  $F = \nabla f$ ,  $G = \nabla g$ , and  $C = \mathbb{R}^n$  in Algorithm 3 and Corollary 1, we derive the following algorithm and corollary for the bilevel optimization problem:

$$\text{Find } x^* \in \Omega \text{ such that } f(x) \geq f(x^*), \text{ for all } x \in \Omega, \quad (49)$$

in which  $\Omega$  represents the nonempty set of minimizers associated with the classical convex optimization problem  $\min_{x \in \mathbb{R}^n} g(x)$ .

---

**Algorithm 4**


---

**Step 0.** Choose  $\lambda_0 > 0$ ,  $\lambda \in (0, 1)$ ,  $\{\gamma_n\} \subset [0, \infty)$ ,  $\{\xi_n\} \subset [0, \infty)$ ,  $\{\eta_n\} \subset (0, \infty)$ ,  $\{\varepsilon_n\} \subset (0, 1)$  such that  $\lim_{n \rightarrow \infty} \frac{\eta_n}{\varepsilon_n} = 0$ ,  $\lim_{n \rightarrow \infty} \varepsilon_n = 0$ ,  $\sum_{n=0}^{\infty} \varepsilon_n = \infty$ .

**Step 1.** Let  $x^{-2}, x^{-1}, x^0 \in \mathcal{H}_1$ . Set  $n := 0$ .

**Step 2.** Compute  $y^n = x^n + \alpha_n(x^n - x^{n-1}) + \beta_n(x^{n-2} - x^{n-1})$ , where

$$\alpha_n = \begin{cases} \min \left\{ \frac{\eta_n}{\|x^n - x^{n-1}\|}, \gamma_n \right\} & \text{if } x^n \neq x^{n-1}, \\ \gamma_n & \text{if } x^n = x^{n-1}, \end{cases}$$

and

$$\beta_n = \begin{cases} \min \left\{ \frac{\eta_n}{\|x^{n-2} - x^{n-1}\|}, \xi_n \right\} & \text{if } x^{n-2} \neq x^{n-1}, \\ \xi_n & \text{if } x^{n-2} = x^{n-1}. \end{cases}$$

**Step 3.** Compute

$$\begin{aligned} z^n &= y^n - \lambda_n \nabla g(y^n), \\ t^n &= z^n - \lambda_n (\nabla g(z^n) - \nabla g(y^n)), \end{aligned}$$

where

$$\lambda_{n+1} = \begin{cases} \min \left\{ \frac{\lambda \|y^n - z^n\|}{\|\nabla g(y^n) - \nabla g(z^n)\|}, \lambda_n \right\} & \text{if } \nabla g(y^n) \neq \nabla g(z^n), \\ \lambda_n & \text{if } \nabla g(y^n) = \nabla g(z^n). \end{cases}$$

**Step 6.** Compute

$$x^{n+1} = t^n - \varepsilon_n \nabla f(t^n).$$

**Step 7.** Set  $n := n + 1$ , and go to **Step 2**.

---

**Corollary 2.** Assuming that Assumption 3 is satisfied. Then the sequence  $\{x^n\}$  produced by Algorithm 4 converges strongly to the unique optimal

solution of (49), given that the set  $\Omega$  of all optimal solutions for the problem  $\min_{x \in \mathbb{R}^n} g(x)$  is nonempty.

From Algorithm 1 and Theorem 1, by setting  $F_1 = F_2 = 0$ , we derive the following algorithm and corollary:

---

**Algorithm 5**


---

**Step 0.** Choose  $\{\rho_n\} \subset [a, b] \subset (0, 1)$ ,  $\{\gamma_n\} \subset [0, \infty)$ ,  $\{\xi_n\} \subset [0, \infty)$ ,  $\{\eta_n\} \subset (0, \infty)$ ,  $\{\varepsilon_n\} \subset (0, 1)$  such that  $\lim_{n \rightarrow \infty} \frac{\eta_n}{\varepsilon_n} = 0$ ,  $\lim_{n \rightarrow \infty} \varepsilon_n = 0$ ,  $\sum_{n=0}^{\infty} \varepsilon_n = \infty$ .

**Step 1.** Let  $x^{-2}, x^{-1}, x^0 \in \mathcal{H}_1$ . Set  $n := 0$ .

**Step 2.** Compute  $y^n = x^n + \alpha_n(x^n - x^{n-1}) + \beta_n(x^{n-2} - x^{n-1})$ , where

$$\alpha_n = \begin{cases} \min \left\{ \frac{\eta_n}{\|x^n - x^{n-1}\|}, \gamma_n \right\} & \text{if } x^n \neq x^{n-1}, \\ \gamma_n & \text{if } x^n = x^{n-1}, \end{cases}$$

and

$$\beta_n = \begin{cases} \min \left\{ \frac{\eta_n}{\|x^{n-2} - x^{n-1}\|}, \xi_n \right\} & \text{if } x^{n-2} \neq x^{n-1}, \\ \xi_n & \text{if } x^{n-2} = x^{n-1}. \end{cases}$$

**Step 3.** Compute

$$\begin{cases} z^n = P_C(y^n), & u^n = A(z^n), & v^n = P_Q(u^n), \\ t^n = z^n + \delta_n A^*(v^n - u^n), \end{cases}$$

where the stepsize  $\delta_n$  is chosen in such a way that

$$\delta_n = \begin{cases} \frac{\rho_n \|v^n - u^n\|^2}{\|A^*(v^n - u^n)\|^2} & \text{if } A^*(v^n - u^n) \neq 0, \\ 0 & \text{if } A^*(v^n - u^n) = 0. \end{cases}$$

**Step 4.** Compute

$$x^{n+1} = t^n - \varepsilon_n F(t^n).$$

**Step 5.** Set  $n := n + 1$ , and go to **Step 2**.

---

**Corollary 3.** Let  $C$  and  $Q$  be two nonempty closed convex subset of two real Hilbert spaces  $\mathcal{H}_1$  and  $\mathcal{H}_2$ , respectively. Let  $F : \mathcal{H}_1 \rightarrow \mathcal{H}_1$  be a strongly monotone and Lipschitz continuous mapping. Then the sequence  $\{x^n\}$  generated by Algorithm 5 converges strongly to  $x^* \in \Gamma$ , which is the unique solution of the VIP  $\langle F(x^*), x - x^* \rangle \geq 0$ , for all  $x \in \Gamma$ , provided the solution set  $\Gamma = \{x^* \in C : Ax^* \in Q\}$  of the SFP is nonempty.

We now apply Corollary 3 with  $F(x) = x$  for all  $x \in \mathcal{H}_1$ . It is clear that the identity mapping  $F : \mathcal{H}_1 \rightarrow \mathcal{H}_1$  is 1-Lipschitz continuous and 1-strongly monotone on  $\mathcal{H}_1$ . This leads us to the following result:

**Corollary 4.** Let  $C$  and  $Q$  be two nonempty closed convex subsets of two real Hilbert spaces  $\mathcal{H}_1$  and  $\mathcal{H}_2$ , respectively. The sequence  $\{x^n\}$  generated by Algorithm 5, in which step 4 specifies  $x^{n+1} = (1 - \varepsilon_n)t^n$ , converges strongly to the minimum-norm solution of the SFP, assuming that the solution set  $\Gamma = \{x^* \in C : Ax^* \in Q\}$  is nonempty.

Next, we will analyze how Corollary 4 can be utilized in discrete optimal control problems.

Let  $A_i$  and  $B_i$  be real matrices of size  $q \times q$  and  $q \times p$ , respectively, for  $i = 0, 1, \dots, N-1$ . We are examining a linear discrete optimal control problem

$$\begin{cases} x_{i+1} = A_{i+1}x_i + B_{i+1}u_i, \\ u_i \in C_i, \\ x_0 = 0, \\ J(x, u) := \|u_0\|^2 + \|u_1\|^2 + \dots + \|u_{N-1}\|^2 \longrightarrow \min_{u_i}, \end{cases} \quad \begin{matrix} i = 0, 1, \dots, N-1, \\ \\ x_N \in \mathcal{Q}, \end{matrix} \quad (50)$$

where  $C_i \subset \mathbb{R}^p$  for  $i = 0, 1, \dots, N-1$ , and  $\mathcal{Q} \subset \mathbb{R}^q$  are nonempty closed convex subsets that define the control and state constraints, respectively.

Establish a matrix of dimension  $q \times Np$

$$\mathcal{A} = [\mathcal{D}_0 \quad \mathcal{D}_1 \quad \dots \quad \mathcal{D}_{N-1}],$$

where  $\mathcal{D}_i := A_N A_{N-1} \dots A_{i+2} B_{i+1}$ ,  $i = 0, 1, \dots, N-2$ , and  $\mathcal{D}_{N-1} = B_N$ .

Let  $u := (u_0, u_1, \dots, u_{N-1})$ ,  $\|u\|^2 := \|u_0\|^2 + \|u_1\|^2 + \dots + \|u_{N-1}\|^2$  and  $\mathcal{C} := C_0 \times C_1 \times \dots \times C_{N-1}$ . Then, (50) transforms into finding the minimum-norm solution of the following SFP:

$$\text{Find } u \in \mathcal{C} \text{ such that } \mathcal{A}u \in \mathcal{Q}.$$

Thus, we can utilize Algorithm 5, where step 4 is given by  $x^{n+1} = (1 - \varepsilon_n)t^n$ , to solve the problem.

## 4 Applications in production and consumption systems

A variation of the SVIP, defined by (1) and (2) and referred to as the SMNP, arises when each  $F_i$  is the identity mapping on  $\mathcal{H}_i$  for all  $i = 1, 2$ . In the SMNP framework, the goal is to determine a solution  $x^* \in C$  that minimizes its norm while ensuring that its image  $y^* = Ax^*$  belongs to  $Q$  and also has the smallest possible norm. Mathematically, this is expressed as follows:

$$\text{Find } x^* \in C : \|x^*\| \leq \|x\| \quad \text{for all } x \in C$$

subject to the condition:

$$y^* = Ax^* \in Q : \|y^*\| \leq \|y\| \quad \text{for all } y \in Q.$$

In many practical applications, particularly in supply chain management and production planning, there is a need to achieve efficiency in both production and distribution. In this context, we consider a system where production and consumption are intrinsically linked via a linear transformation. Let  $x \in \mathbb{R}^N$  represent the production vector, quantifying the goods produced, and let  $y \in \mathbb{R}^M$  denote the consumption vector, representing the goods delivered to the market. The connection between production and consumption is modeled by the matrix  $A \in \mathbb{R}^{M \times N}$ , such that  $y = Ax$ . The production process is constrained by various operational factors, including capacity and resource limitations. These are encapsulated in the feasible set  $C \subset \mathbb{R}^N$ . For instance, one may define

$$C = \{x \in \mathbb{R}_+^N : Bx \leq b\},$$

where the matrix  $B$  and the vector  $b$  represent production constraints such as available resources or maximum production capacities. On the other hand, the consumption or distribution process must satisfy market demand or quality requirements, which are modeled by the feasible set  $Q \subset \mathbb{R}^M$ . One common formulation is

$$Q = \{y \in \mathbb{R}_+^M : y \geq d\},$$

with  $d$  being the vector of minimum demand requirements ensuring that the market receives at least the prescribed quantities.

In the production set  $C$ , selecting  $x^*$  with the smallest norm is crucial because it ensures that among all feasible production plans,  $x^*$  consumes the least resources or incurs the lowest production cost. This minimality directly translates into enhanced efficiency in production. Similarly, in the consumption set  $Q$ , requiring that  $y^* = Ax^*$  has the smallest norm means that the corresponding distribution of goods is accomplished with minimal overhead or waste. This condition is essential for achieving an efficient distribution process. Thus, the overall objective is to select a production plan  $x^* \in C$  that minimizes the production norm:

$$\|x^*\| \leq \|x\| \quad \text{for all } x \in C,$$

thereby reducing production costs, resource usage, or energy consumption. Simultaneously, the corresponding consumption vector  $y^* = Ax^*$  must belong to  $Q$  and minimize the consumption norm:

$$\|y^*\| \leq \|y\| \quad \text{for all } y \in Q.$$

The SMNP model provides an integrated framework for addressing the challenges of simultaneously optimizing production and distribution. By merging the operational constraints of production with the market's consumption requirements and enforcing minimal norm conditions, the SMNP formulation successfully reduces costs while enhancing overall supply chain efficiency.

## 5 Numerical illustration

In this section, we present numerical experiments to assess the performance of the proposed algorithms and provide results from various comparisons. All Python code was executed on a 2017 MacBook Pro featuring a 2.3 GHz Intel Core i5 processor, an Intel Iris Plus Graphics 640 GPU with 1536 MB of memory, and 8 GB of 2133 MHz LPDDR3 RAM. The experiments were conducted using Python version 3.11.

**Example 1.** (see [25, Example 4.1]). Let  $\mathbb{R}^K$  be equipped with the standard norm  $\|x\| = \sqrt{x_1^2 + x_2^2 + \cdots + x_K^2}$  for all  $x = (x_1, x_2, \dots, x_K)^T \in \mathbb{R}^K$ . Let

$A(x) = (x_1 + x_3 + x_4, x_2 + x_3 - x_4)^T$  for all  $x = (x_1, x_2, x_3, x_4)^T \in \mathbb{R}^4$ . This shows that  $A$  is a bounded linear operator from  $\mathbb{R}^4$  into  $\mathbb{R}^2$ .

Now, define the set

$$C = \{(x_1, x_2, x_3, x_4)^T \in \mathbb{R}^4 : x_1 - 3x_2 - 2x_3 + x_4 \geq -2\},$$

and let the mapping  $F_1 : \mathbb{R}^4 \rightarrow \mathbb{R}^4$  be defined by  $F_1(x) = (\sin \|x\| + 4)b^0$  for all  $x \in \mathbb{R}^4$ , where  $b^0 = (1, -3, -2, 1)^T \in \mathbb{R}^4$ . It is easy to verify that  $F_1$  is pseudomonotone and Lipschitz continuous on  $\mathbb{R}^4$ .

Now, let  $Q = \{(u_1, u_2)^T \in \mathbb{R}^2 : u_1 - 2u_2 \geq -1\}$ , and define another mapping  $F_2 : \mathbb{R}^2 \rightarrow \mathbb{R}^2$  by  $F_2(u) = (\sin \|u\| + 2)c^0$  for all  $u \in \mathbb{R}^2$ , where  $c^0 = (1, -2)^T \in \mathbb{R}^2$ . Similarly,  $F_2$  is pseudomonotone and Lipschitz continuous on  $\mathbb{R}^2$ .

Consider the mapping  $F : \mathbb{R}^4 \rightarrow \mathbb{R}^4$  defined by  $F(x) = 2x + a^0$  for all  $x \in \mathbb{R}^4$ , where  $a^0 = (-2, 0, 4, -6)^T \in \mathbb{R}^4$ . It is straightforward to verify that  $F$  is strongly monotone and Lipschitz continuous on  $\mathbb{R}^4$ . In [25], the authors demonstrated that the unique solution to the BSVIP (4) is given by  $x^* = \left(\frac{4}{27}, \frac{44}{27}, -\frac{11}{9}, \frac{8}{27}\right)^T$ .

Table 1: A comparison between Algorithm 1 and [25, Algorithm 3.1] with different tolerances  $\varepsilon$  and the stopping criterion  $\|x^n - x^*\| \leq \varepsilon$

$\varepsilon = 10^{-3}$		
	Iter( $n$ )	CPU time(s)
Algorithm 1	9945	1.7804
[25, Algorithm 3.1]	14611	2.4233
$\varepsilon = 10^{-4}$		
	Iter( $n$ )	CPU time(s)
Algorithm 1	99490	19.1084
[25, Algorithm 3.1]	146159	24.8806

We will now assess the performance of Algorithm 1 in comparison to [25, Algorithm 3.1], as outlined in [25]. Both algorithms use the termination

criterion  $\|x^n - x^*\| \leq \varepsilon$  and start with the same initial point,  $x^0$ , where its components are randomly generated within the closed interval  $[-10, 10]$ . Additionally, for Algorithm 1, the components of the initial points  $x^{-2}$  and  $x^{-1}$  are also randomly selected from the same interval. The parameters for each algorithm are specified as follows:

- Algorithm 1:  $\lambda_0 = 3$ ,  $\mu_0 = 2$ ,  $\lambda = 0.3$ ,  $\mu = 0.4$ ,  $\gamma_n = 0.1$ ,  $\xi_n = 0.2$ ,  $\rho_n = 0.99$ ,  $\eta_n = \frac{1}{(n+2)^{1.01}}$  and  $\varepsilon_n = \frac{1}{n+2}$ .
- [25, Algorithm 3.1]:  $\lambda_0 = 3$ ,  $\mu_0 = 2$ ,  $\lambda = 0.3$ ,  $\mu = 0.4$  and  $\varepsilon_n = \frac{1}{n+2}$ .

The results presented in Table 1 indicate that Algorithm 1 outperforms [25, Algorithm 3.1] in terms of both runtime and iteration count.

**Example 2.** Let  $a^0 = (1, -6, -3, 2, -3, 6, -1, -2)^T \in \mathbb{R}^8$ , and consider the set  $C$  defined as  $C = \{x = (x_1, x_2, \dots, x_8)^T \in \mathbb{R}^8 : \langle a^0, x \rangle \geq -2\}$ . Now, let us define a mapping  $G : \mathbb{R}^8 \rightarrow \mathbb{R}^8$  by  $G(x) = (\sin \|x\| + 4)a^0$  for all  $x \in \mathbb{R}^8$ . It can be easily verified that  $G$  is pseudomonotone on  $\mathbb{R}^8$  and Lipschitz continuous on  $\mathbb{R}^8$ . Furthermore, it is evident that the solution set  $\text{Sol}(C, G)$  of the VIP  $VIP(C, G)$  is given by

$$\text{Sol}(C, G) = \{x = (x_1, x_2, \dots, x_8)^T \in \mathbb{R}^8 : \langle a^0, x \rangle = -2\}.$$

Let us consider the mapping  $F : \mathbb{R}^8 \rightarrow \mathbb{R}^8$  defined as  $F(x) = x$  for all  $x \in \mathbb{R}^8$ . This mapping  $F$  is strongly monotone with  $\eta = 1$  and Lipschitz continuous with  $L = 1$  on  $\mathbb{R}^8$ . In this context, problem (5) transforms into finding the minimum-norm solution of the  $VIP(C, G)$ . The resulting minimum-norm solution  $x^*$  for the  $VIP(C, G)$  is  $x^* = P_{\text{Sol}(C, G)}(0) = (-0.02, 0.12, 0.06, -0.04, 0.06, -0.12, 0.02, 0.04)^T$ .

We are set to compare the performance of Algorithm 3 with [25, Algorithm 3.6], as presented in [25], for solving the BVIP problem (5). Both algorithms start with the same initial point,  $x^0$ , whose components are randomly generated within the closed interval  $[-10, 10]$ , and both use the termination criterion  $\|x^n - x^*\| \leq \varepsilon$ . Additionally, for Algorithm 3, the components of the initial points  $x^{-2}$  and  $x^{-1}$  are also randomly chosen from the same closed interval  $[-10, 10]$ . The parameter settings for these methods are as follows:

Table 2: A comparison between Algorithm 3 and [25, Algorithm 3.6] with different tolerances  $\varepsilon$  and the stopping criterion  $\|x^n - x^*\| \leq \varepsilon$

	$\varepsilon = 10^{-3}$	
	Iter( $n$ )	CPU time(s)
Algorithm 3	1343	0.1411
[25, Algorithm 3.6]	13359	1.1035
	$\varepsilon = 10^{-4}$	
	Iter( $n$ )	CPU time(s)
Algorithm 3	12969	1.1198
[25, Algorithm 3.6]	133599	10.9758

• Algorithm 3:  $\lambda_0 = 3$ ,  $\lambda = 0.6$ ,  $\gamma_n = 10^4$ ,  $\xi_n = 10^{-2}$ ,  $\eta_n = \frac{1}{(n+2)^{1.01}}$  and  $\varepsilon_n = \frac{1}{n+2}$ .

• [25, Algorithm 3.6]:  $\lambda_0 = 3$ ,  $\lambda = 0.6$  and  $\varepsilon_n = \frac{1}{n+2}$ .

The results shown in Table 2 suggest that Algorithm 3 demonstrates superior performance when compared to [25, Algorithm 3.6].

**Example 3.** Let  $\mathcal{H}_1 = \mathbb{R}^K$  and let  $\mathcal{H}_2 = \mathbb{R}^L$ , where  $K = 200$  and  $L = 150$ . We consider the SFP with the sets  $C = \{x \in \mathbb{R}^K : \langle c, x \rangle \geq 0\}$ ,  $Q = \{y \in \mathbb{R}^L : \langle q, y \rangle \geq 0\}$  and the bounded linear operator  $A : \mathbb{R}^K \rightarrow \mathbb{R}^L$  defined by  $A(x) = Mx$  for all  $x \in \mathbb{R}^K$ , where  $M$  is an  $L \times K$  real matrix. We generate the elements of  $M$  randomly within the closed interval  $[-10, 10]$ , and the coordinates of  $c$  and  $q$  within the closed interval  $[2, 10]$ . It is straightforward to observe that  $0 \in C$  and  $A(0) = 0 \in Q$ . Therefore,  $0 \in \Gamma = \{x^* \in C : Ax^* \in Q\}$ . Thus, the minimum-norm solution  $x^*$  of the SFP is  $x^* = 0$ .

We aim to compare the performance of Algorithm 5, where  $F$  is the identity mapping, with the algorithm described in [18, Corollary 3.2] for solving the minimum-norm solution of the SFP. Both algorithms begin with the same initial point,  $x^0$ , whose components are randomly generated within the closed interval  $[-10, 10]$ . They also both use the same stopping criterion,

$\|x^n - x^*\| \leq \varepsilon$  and the same  $\varepsilon_n = \frac{1}{n+2}$  (in [18, Corollary 3.2], this is denoted as  $\alpha_n$ ). Additionally, in Algorithm 5, the components of the initial points  $x^{-2}$  and  $x^{-1}$  are also randomly selected from the same closed interval  $[-10, 10]$ . The parameter values in Algorithm 5 are chosen as  $\gamma_n = 10^6$ ,  $\xi_n = 10^{-4}$ ,  $\rho_n = 0.99$ , and  $\eta_n = \frac{1}{(n+2)^{1.01}}$ .

Table 3: A comparison between Algorithm 5, where  $F$  is the identity mapping, and the algorithm described in [18, Corollary 3.2], with different tolerances  $\varepsilon$  and the stopping criterion  $\|x^n - x^*\| \leq \varepsilon$

	$\varepsilon = 10^{-3}$	
	Iter( $n$ )	CPU time(s)
Algorithm 5	189	0.0148
Algorithm in [18, Corollary 3.2]	79652	4.9645
	$\varepsilon = 10^{-4}$	
	Iter( $n$ )	CPU time(s)
Algorithm 5	2498	0.1653
Algorithm in [18, Corollary 3.2]	776266	45.2463

Table 3 illustrates that our Algorithm 5 significantly outperforms the algorithm in [18, Corollary 3.2] in terms of both iteration count and CPU time.

## 6 Conclusions

This paper presented an iterative algorithm for addressing BSVIPs. We established that the iterative sequence strongly converges to the unique solution of the BSVIP without needing to compute or estimate the norm of a bounded linear operator. Moreover, the algorithm can be implemented without requiring any calculations or estimations of the Lipschitz and strongly monotone constants of the mappings involved. We also applied this algorithm to specific cases, including the bilevel VIPs, the bilevel optimization problems, and strongly monotone VIPs with split feasibility constraints. Finally,

we provided an application of the SMNP in production and consumption systems and presented several numerical experiments to demonstrate the implementability of the proposed algorithms.

As a potential direction for future research, it would be interesting to investigate the extension of our results to Banach spaces. This generalization may present new challenges, particularly in handling the lack of Hilbert space structure, but it could also broaden the applicability of our approach to a wider class of problems.

## Acknowledgements

We acknowledge Ho Chi Minh City University of Technology (HCMUT), VNU-HCM for supporting this study.

## Funding

Le Huynh My Van was funded by the Master, PhD Scholarship Programme of Vingroup Innovation Foundation (VINIF), code VINIF.2023.TS.146

## References

- [1] Anh, P.K., Anh, T.V., and Muu, L.D. *On bilevel split pseudomonotone variational inequality problems with applications*, Acta Math. Vietnam, 42 (2017), 413–429.
- [2] Anh, P.N., Kim, J.K., and Muu, L.D. *An extragradient algorithm for solving bilevel pseudomonotone variational inequalities*, J. Glob. Optim., 52 (2012), 627–639.
- [3] Anh, T.V. *A parallel method for variational inequalities with the multiple-sets split feasibility problem constraints*, J. Fixed Point Theory Appl., 19 (2017), 2681–2696.

- [4] Anh, T.V. *A strongly convergent subgradient extragradient-Halpern method for solving a class of bilevel pseudomonotone variational inequalities*, Vietnam J. Math., 45 (2017), 317–332.
- [5] Buong, N. *Iterative algorithms for the multiple-sets split feasibility problem in Hilbert spaces*, Numer. Algorithms, 76 (2017), 783–798.
- [6] Ceng, L.C., Ansari, Q.H., and Yao, J.C. *Some iterative methods for finding fixed points and for solving constrained convex minimization problems*, Nonlinear Anal., 74 (2011), 5286–5302.
- [7] Ceng, L.C., Ansari, Q.H., and Yao, J.C. *An extragradient method for solving split feasibility and fixed point problems*, Comput. Math. Appl., 64 (2012), 633–642.
- [8] Ceng, L.C., Ansari, Q.H., and Yao, J.C. *Relaxed extragradient methods for finding minimum-norm solutions of the split feasibility problem*, Nonlinear Anal., 75 (2012), 2116–2125.
- [9] Ceng, L.C., Coroian, I., Qin, X., and Yao, J.C. *A general viscosity implicit iterative algorithm for split variational inclusions with hierarchical variational inequality constraints*, Fixed Point Theory, 20 (2019), 469–482.
- [10] Ceng, L.C., Liou, Y.C., and Sahu, D.R. *Multi-step hybrid steepest-descent methods for split feasibility problems with hierarchical variational inequality problem constraints*, J. Nonlinear Sci. Appl., 9 (2016), 4148–4166.
- [11] Ceng, L.C., Wong, N.C., and Yao, J.C. *Hybrid extragradient methods for finding minimum-norm solutions of split feasibility problems*, J. Nonlinear Convex Anal., 16 (2015), 1965–1983.
- [12] Censor, Y., Bortfeld, T., Martin, B., et al. *A unified approach for inversion problems in intensity-modulated radiation therapy*, Phys. Med. Biol., 51 (2006), 2353–2365.
- [13] Censor, Y., and Elfving, T. *A multiprojection algorithm using Bregman projections in a product space*, Numer. Algorithms, 8 (1994), 221–239.

- [14] Censor, Y., Elfving, T., Kopf, N. and Bortfeld, T. *The multiple-sets split feasibility problem and its applications for inverse problems*, Inverse Prob., 21 (2005), 2071–2084.
- [15] Censor, Y., Gibali, A., and Reich, S. *Algorithms for the split variational inequality problem*, Numer. Algorithms, 59 (2012), 301–323.
- [16] Censor, Y., and Segal, A. *Iterative projection methods in biomedical inverse problems*, in: Censor, Y., Jiang, M., and Louis, A.K., eds., *Mathematical methods in biomedical imaging and intensity-modulated therapy*, Edizioni della Norale, Pisa, Italy, IMRT, (2008) 65–96.
- [17] Combettes, P.L. and Hirstoaga, S.A. *Equilibrium Programming in Hilbert Spaces*, J. Nonlinear Convex Anal., 6 (2005), 117–136.
- [18] Cuong, T.L., Anh, T.V., and Van, L.H.M. *A self-adaptive step size algorithm for solving variational inequalities with the split feasibility problem with multiple output sets constraints*, Numer. Funct. Anal. Optim., 43 (2022), 1009–1026.
- [19] Daniele, P., Giannessi, F., and Maugeri, A. *Equilibrium problems and variational models*, Dordrecht, Kluwer Academic, 2003.
- [20] Dempe, S. *Foundations of bilevel programming*, Dordrecht, Kluwer Academic Press, 2002.
- [21] Facchinei, F., and Pang, J.S. *Finite-dimensional variational inequalities and complementarity problems*, Berlin, Springer, 2002.
- [22] Giannessi, F., Maugeri, A., and Pardalos, P.M. *Equilibrium problems: nonsmooth optimization and variational inequality models*, Dordrecht, Kluwer Academic, 2004.
- [23] He, S., Zhao, Z., and Luo, B. *A relaxed self-adaptive CQ algorithm for the multiple-sets split feasibility problem*, Optimization, 64 (2015), 1907–1918.
- [24] Huy, P.V., Hien, N.D., and Anh, T.V. *A strongly convergent modified Halpern subgradient extragradient method for solving the split variational inequality problem*, Vietnam J. Math., 48 (2020), 187–204.

- [25] Huy, P.V., Van, L.H.M., Hien, N.D., and Anh, T.V. *Modified Tseng's extragradient methods with self-adaptive step size for solving bilevel split variational inequality problems*, Optimization, 71 (2022), 1721–1748.
- [26] Iyiola, O. S., and Shehu, Y. *Convergence results of two-step inertial proximal point algorithm*, Appl. Numer. Math., 182 (2022), 57–75.
- [27] Izuchukwu, C., Aphane, M., and Aremu, K. O. *Two-step inertial forward-reflected-anchored-backward splitting algorithm for solving monotone inclusion problems*, Comput. Appl. Math., 42 (2023), 351.
- [28] Kinderlehrer, D., and Stampacchia, G. *An introduction to variational inequalities and their applications*, New York, Academic,; 1980.
- [29] Konnov, I.V. *Combined Relaxation Methods for Variational Inequalities*, Springer, Berlin (2000).
- [30] Korpelevich, G.M. *The extragradient method for finding saddle points and other problems*, Ekonomikai Matematicheskie Metody, 12 (1976), 747–756.
- [31] Maingé, P.E. *A hybrid extragradient-viscosity method for monotone operators and fixed point problems*, SIAM J. Control Optim., 47 (2008), 1499–1515.
- [32] Okeke, C. C., Jolaoso, L. O., and Shehu, Y. *Inertial accelerated algorithms for solving split feasibility with multiple output sets in Hilbert spaces*, Int. J. Nonlinear Sci. Numer. Simul., 24 (2021), 769–790.
- [33] Polyak, B. T. *Some methods of speeding up the convergence of iteration methods*, USSR Comput. Math. Math. Phys., 4 (1964), 1–17.
- [34] Raeisi, M., Zamani Eskandani, G., and Eslamian, M. *A general algorithm for multiple-sets split feasibility problem involving resolvents and Bregman mappings*, Optimization, 67 (2018), 309–327.
- [35] Rudin, W. *Functional Analysis*, 2nd ed., McGraw-Hill, New York, 1991.
- [36] Shehu, Y. *Strong convergence theorem for multiple sets split feasibility problems in Banach spaces*, Numerical Funct. Anal. Optim., 37 (2016), 1021–1036.

- [37] Solodov, M.V. *An explicit descent method for bilevel convex optimization*, J. Convex Anal., 14 (2007), 227–237.
- [38] Thong, D.V., and Hieu, D.V. *A strong convergence of modified subgradient extragradient method for solving bilevel pseudomonotone variational inequality problems*, Optimization, 69 (2020), 1313–1334.
- [39] Tseng, P. *A modified forward–backward splitting method for maximal monotone mappings*, SIAM J. Control Optim., 38 (2000), 431–446.
- [40] Uzor, V., Alakoya, T., and Mewomo, O. T. *On split monotone variational inclusion problem with multiple output sets with fixed point constraints*, Comput. Methods Appl. Math., 23 (2023), 729–749.
- [41] Xu, H.K. *Iterative algorithms for nonlinear operators*, J. London Math. Soc., 66 (2002), 240–256.
- [42] Yao, Y., Marino, G., and Muglia, L. *A modified Korpelevich's method convergent to the minimum-norm solution of a variational inequality*, Optimization, 63 (2014), 559–569.
- [43] Yao, Y., Postolache, M., and Zhu, Z. *Gradient methods with selection technique for the multiple-sets split feasibility problem*, Optimization, 69 (2020), 269–281.
- [44] Zegeye, H., Shahzad, N., and Yao, Y. *Minimum-norm solution of variational inequality and fixed point problem in Banach spaces*, Optimization, 64 (2015), 453–471.



## Approximate symmetries of the perturbed KdV-KS equation

A. Mohammadpouri\*, M.S. Hashemi, R. Abbasi and R. Abbasi

### Abstract

The analysis of approximate symmetries in perturbed nonlinear partial differential equations (PDEs) stands as a cornerstone for unraveling complex physical behaviors and solution patterns. This paper delves into the investigation of approximate symmetries inherent in the perturbed

---

\*Corresponding author

Received 18 November 2024; revised 30 January 2025; accepted 1 February 2025

Akram Mohammadpouri

Faculty of Mathematics, Statistics and Computer Sciences, University of Tabriz, Tabriz, Iran. e-mail: [pouri@tabrizu.ac.ir](mailto:pouri@tabrizu.ac.ir)

Mir Sajjad Hashemi

Department of Mathematics, Basic Science Faculty, University of Bonab, Bonab, Iran. e-mail: [hashemi\\_math396@yahoo.com](mailto:hashemi_math396@yahoo.com)

Roya Abbasi

Faculty of Mathematics, Statistics and Computer Sciences, University of Tabriz, Tabriz, Iran. e-mail: [royaabbasi479@gmail.com](mailto:royaabbasi479@gmail.com)

Rana Abbasi

Faculty of Mathematics, Statistics and Computer Sciences, University of Tabriz, Tabriz, Iran. e-mail: [rana.abbasi.1400111@gmail.com](mailto:rana.abbasi.1400111@gmail.com)

### How to cite this article

Mohammadpouri, A., Hashemi, M.S., Abbasi, R. and Abbasi, R., Approximate symmetries of the perturbed KdV-KS equation. *Iran. J. Numer. Anal. Optim.*, 2025; 15(3): 914-929. <https://doi.org/10.22067/ijnao.2025.90854.1551>

Korteweg-de Vries and Kuramoto-Sivashinsky (KdV-KS) equation, fundamental models in the realm of fluid dynamics and wave phenomena. Our study commences by detailing the method to derive approximate vector Lie symmetry generators that underpin the approximate symmetries of the perturbed KdV-KS equation. These generators, while not exact, provide invaluable insights into the equation's dynamics and solution characteristics under perturbations. A comprehensive approximate commutator table is subsequently constructed, elucidating the relationships and interplay between these approximate symmetries and shedding light on their algebraic structure. Leveraging the power of the adjoint representation, we examine the stability of these approximate symmetries when subjected to perturbations. This analysis enables us to discern the most resilient symmetries, instrumental in identifying intrinsic features that persist even in the face of disturbances. Furthermore, we harness the concept of approximate symmetry reductions, a pioneering technique that allows us to distill crucial dynamics from the complexity of the perturbed equation. Through this methodology, we uncover invariant solutions and reduced equations that serve as effective surrogates for the original system, capturing its essential behavior and facilitating analytical and numerical investigations. In summary, our exploration into the approximate symmetries of the perturbed KdV-KS equation not only advances our comprehension of the equation's intricate dynamics but also offers a comprehensive framework for studying the impact of perturbations on approximate symmetries, all while opening new avenues for tackling nonlinear PDEs in diverse scientific disciplines.

**AMS subject classifications (2020):** Primary 53A10; Secondary 22E70, 26M60.

**Keywords:** Approximate Lie symmetries; Commutator table; Adjoint representation; Reductions.

## 1 Introduction

The Korteweg-de Vries and Kuramoto-Sivashinsky (KdV-KS) equation is a notable partial differential equation (PDE) that amalgamates the Korteweg-de Vries equation, renowned for describing long, weakly nonlinear waves, with the Kuramoto-Sivashinsky equation, which captures spatiotemporal chaos in pattern-forming systems. This fusion yields a versatile equation capable of

modeling a diverse array of physical phenomena. The KdV-KS equation finds application in various fields, including fluid dynamics, combustion, and nonlinear optics. In fluid dynamics, it can depict the evolution of complex wave patterns on fluid interfaces, while in combustion processes, it may illuminate the behavior of flame fronts and combustion instabilities. Additionally, the equation's presence in the realm of nonlinear optics can aid in understanding pulse propagation in optical fibers. Its broad applicability underscores the KdV-KS equation's significance as a tool for investigating intricate dynamics in real-world systems and its role in advancing our comprehension of nonlinear phenomena across multiple scientific disciplines [10, 5, 12].

Analytical methods for solving PDEs constitute a vital framework in understanding the behavior of various physical, mathematical, and engineering systems. These methods encompass a range of techniques, such as Lie symmetry method [14, 2, 18, 19, 17, 16, 22], Kudryashov's method [16, 20, 9], Nucci's reduction method [23, 24], invariant subspace method [8, 21, 4], and Tanh method [7, 1, 6].

In cases where PDEs possess specific geometrical or algebraic properties, separation of variables can yield exact solutions by decomposing the equation into simpler ordinary differential equations. Similarity transformations assist in reducing complex PDEs to canonical forms that admit analytical solutions. Integral transforms, like the Fourier and Laplace transforms, provide a powerful means to convert differential equations into algebraic equations that can be more easily solved. Perturbation methods, including the method of matched asymptotic expansions and multiple scales analysis, are particularly useful when dealing with systems that exhibit small parameter deviations from simpler cases, allowing the derivation of approximate solutions.

These analytical techniques not only offer insights into the underlying dynamics of diverse systems but also serve as benchmarks for numerical methods. However, their applicability is often constrained by the complexity of the equations and the presence of nonlinear terms. In such cases, a combination of these methods, along with innovations in mathematical analysis, plays a crucial role in uncovering solutions that enrich our understanding of the intricate interplay between mathematics and the physical world.

We shall research the perturbed KdV-KS equations's vector fields, approximate symmetry, and symmetry reductions. A perturbed form of the KdV-KS equations are

$$u_t + uu_x + u_{xxx} + \epsilon(u_{xx} + u_{xxxx}) = 0, \quad (1)$$

where  $0 < \epsilon \ll 1$  is a small parameter,  $x \in \mathbb{R}$ , and  $t \geq 0$ .

The structure of this work is as follows. We find the approximate symmetry and optimal system of the perturbed KdV-KS equation in section 2. Ordinary differential equation symmetry reductions are covered in section 3. Finally, section 4 will provide the conclusions.

## 2 Analysis of the approximate Lie symmetries

Approximate Lie symmetries play a crucial role in various scientific and mathematical contexts, particularly in the study of dynamical systems and differential equations. Unlike exact symmetries, which lead to conserved quantities and well-defined transformations, approximate Lie symmetries emerge in situations where the underlying system's behavior is influenced by small perturbations or deviations from ideal conditions. These symmetries provide insights into the system's response to fluctuations and disturbances, contributing to our understanding of stability, chaos, and the emergence of complex patterns. By analyzing the behavior of systems under approximate Lie symmetries, researchers gain valuable insights into the underlying dynamics and are better equipped to model real-world phenomena with a more comprehensive perspective.

Let  $\Delta(t, x, u, \epsilon) = \Delta_0(t, x, u) + \epsilon\Delta_1(t, x, u) = u_t + uu_x + u_{xxx} + \epsilon(u_{xx} + u_{xxxx})$ . If an operator  $Y = Y_0 + \epsilon Y_1$  satisfies

$$\left[ Y^{(4)} \Delta(t, x, u, \epsilon) \right]_{\Delta(t, x, u, \epsilon)=0} = 0, \quad (2)$$

then it is referred to as an approximation Lie symmetry generator. Here,  $Y^{(4)}$  is the forth-order prolongation of the forth-order approximate Lie symmetry  $Y$ , and

$$Y_0 = \xi_0^1(t, x, u) \frac{\partial}{\partial t} + \xi_0^2(t, x, u) \frac{\partial}{\partial x} + \eta_0(t, x, u) \frac{\partial}{\partial u},$$

$$Y_1 = \xi_1^1(t, x, u) \frac{\partial}{\partial t} + \xi_1^2(t, x, u) \frac{\partial}{\partial x} + \eta_1(t, x, u) \frac{\partial}{\partial u}.$$

The prolongation formula is a fundamental tool within the realm of Lie symmetry methods, a powerful mathematical approach used to analyze and solve differential equations. In this context, the prolongation formula extends the Lie derivative to higher-order derivatives and introduces a systematic way of calculating symmetries of a given differential equation. By iteratively applying the prolongation formula, one can uncover hidden symmetries that may not be immediately apparent. This process allows researchers to determine transformations that leave the equation invariant and identify conserved quantities or transformations that simplify its solutions.

Equation (2) divides two parts into

$$\left[ Y_0^{(4)} \Delta_0(t, x, u, \epsilon) \right]_{\Delta_0(t, x, u, \epsilon)=0} = 0, \quad (3)$$

$$\left[ Y_1^{(4)} \Delta_0(t, x, u, \epsilon) + Y_0^{(4)} \Delta_1(t, x, u, \epsilon) \right]_{\Delta(t, x, u, \epsilon)=0} = 0. \quad (4)$$

By conditions (3) and (4), we arrive at the set of determining equations below:

$$\begin{aligned} \xi_{0,t}^1 &= \xi_{0,x}^1 = \xi_{0,u}^1 = \xi_{0,x}^2 = \xi_{0,u}^2 = \eta_{0,u} = 0, & \eta_{0,xxx} + \eta_{0,t} + u\eta_{0,x} &= 0, \\ \eta_0 - \xi_{0,t}^2 &= 0, & \xi_{1,x}^1 &= \xi_{1,u}^1 = \xi_{1,u}^2 = 0, & \eta_{1,uu} &= 0, & \eta_{1,xxx} + \eta_{1,t} + u\eta_{1,x} &= 0, \\ 2u\xi_{1,x}^2 - \xi_{1,xxx}^2 - \xi_{1,t}^2 + 3\eta_{1,xxu} + \eta_1 &= 0, & \xi_{1,t}^1 - 3\xi_{1,x}^2 &= 0, & \eta_{1,xu} - \xi_{1,xx}^2 &= 0. \end{aligned}$$

Solving this PDE system gives us

$$\begin{aligned} \xi_0^1 &= a_0, & \xi_0^2 &= b_0t + c_0, & \eta_0 &= b_0, \\ \xi_1^1 &= -\frac{3}{2}a_1t + b_1, & \xi_1^2 &= -\frac{1}{2}a_1x + c_1t + d_1, & \eta_1 &= a_1u + c_1. \end{aligned}$$

Therefore

$$\begin{aligned} X &= (a_0 + \epsilon(-\frac{3}{2}a_1t + b_1))\partial_t + (b_0t + c_0 + \epsilon(-\frac{1}{2}a_1x + c_1t + d_1))\partial_x \\ &\quad + (b_0 + \epsilon(a_1u + c_1))\partial_u \end{aligned}$$

where  $a_0, b_0, c_0, a_1, b_1, c_1$ , and  $d_1$  are constants. Consequently, the following seven independent approximate operators span infinitesimal symmetries of

Table 1: Approximate commutator table for symmetries in (1)

$[y_i, y_j]$	$y_1$	$y_2$	$y_3$	$y_4$	$y_5$	$y_6$	$y_7$
$y_1$	0	0	$y_2$	0	$-\frac{3}{2}y_4$	0	$y_6$
$y_2$	0	0	0	0	$-\frac{1}{2}y_6$	0	0
$y_3$	$-y_2$	0	0	$-y_6$	$y_7$	0	0
$y_4$	0	0	$y_6$	0	0	0	0
$y_5$	$\frac{3}{2}y_4$	$\frac{1}{2}y_6$	$-y_7$	0	0	0	0
$y_6$	0	0	0	0	0	0	0
$y_7$	$-y_6$	0	0	0	0	0	0

equation (1)

$$y_1 = \partial_t, \quad y_2 = \partial_x, \quad y_3 = t\partial_x + \partial_u, \quad y_4 = \epsilon\partial_t, \quad y_5 = \epsilon\left(-\frac{3}{2}t\partial_t - \frac{1}{2}x\partial_x + u\partial_u\right),$$

$$y_6 = \epsilon\partial_x, \quad y_7 = \epsilon(t\partial_x + \partial_u).$$

In the field of Lie symmetry methods, a commutator table serves as a fundamental tool for analyzing the algebraic structure of Lie symmetries associated with a system of differential equations. The commutator of two vector fields, representing different symmetry transformations, is calculated and organized in a table format. This table provides valuable information about the Lie algebra generated by these vector fields, revealing how they interact and combine. By determining the commutators, researchers can discern the algebraic relationships between symmetries, uncover hidden patterns, and ultimately construct a Lie algebra that captures the system's inherent symmetries. The commutator table thus acts as a guiding compass in the exploration of differential equations, aiding in the classification, solution, and deeper comprehension of complex dynamical systems. Table 1 contains the approximate commutator table for symmetries in equation (1). The adjoint representation involves mapping each element of a Lie group to an associated automorphism of its corresponding Lie algebra. This representation provides insights into how transformations in the group relate to transformations in the algebra, offering a way to study the Lie group through its associated

Table 2: Adjoint representation spanned by the the basis approximate symmetries of the KdV-KS equation

$Ad(\exp(ay_i))y_j$	$y_1$	$y_2$	$y_3$	$y_4$	$y_5$	$y_6$	$y_7$
$y_1$	$y_1$	$y_2$	$y_3 - ay_2$	$y_4$	$y_5 + \frac{3}{2}ay_4$	$y_6$	$y_7 - ay_6$
$y_2$	$y_1$	$y_2$	$y_3$	$y_4$	$y_5 + \frac{1}{2}ay_6$	$y_6$	$y_7$
$y_3$	$y_1 + ay_2$	$y_2$	$y_3$	$y_4 + ay_6$	$y_5 - ay_7$	$y_6$	$y_7$
$y_4$	$y_1$	$y_2$	$y_3 - ay_6$	$y_4$	$y_5$	$y_6$	$y_7$
$y_5$	$y_1 - \frac{3}{2}ay_4$	$y_2 - \frac{1}{2}ay_6$	$y_3 - \frac{1}{2}ay_7$	$y_4$	$y_5$	$y_6$	$y_7$
$y_6$	$y_1$	$y_2$	$y_3$	$y_4$	$y_5$	$y_6$	$y_7$
$y_7$	$y_1 + ay_6$	$y_2$	$y_3$	$y_4$	$y_5$	$y_6$	$y_7$

Lie algebra. By analyzing the adjoint representation, researchers can explore the relationships between different Lie group elements and understand the symmetries and transformations that underlie a given system of differential equations. This representation is a key tool for investigating the symmetries, conservation laws, and invariants inherent in complex dynamical systems, ultimately facilitating the application of Lie symmetry methods to a wide range of scientific and mathematical problems. Each  $y_i$ ,  $i = 1, \dots, 7$  of the basis approximate infinitesimal symmetries spans an adjoint representation  $Ad(\exp(ay_i))y_j$ ,  $a$  is a parameter, given by

$$Ad(\exp(ay_i))y_j = y_j - a[y_i, y_j] + \frac{a^2}{2}[y_i, [y_i, y_j]] - \dots$$

Table 2 lists each adjoint representation of the Lie approximate symmetry of the KdV-KS equation. Here, in the following, we find that the group approximate transformation  $h_i$ , which is generated by the  $y_i$  for  $i = 1, 2, \dots, 7$  for the KdV-KS equation (1)

$$\left\{ \begin{array}{l} h_1.(t, x, u) \mapsto (t + a, x, u), \\ h_2.(t, x, u) \mapsto (t, x + a, u), \\ h_3.(t, x, u) \mapsto (t, x + ta, u + a), \\ h_4.(t, x, u) \mapsto (t + a\epsilon, x, u), \\ h_5.(t, x, u) \mapsto ((1 - \frac{3}{2}a\epsilon)t, ((1 - \frac{1}{2}a\epsilon)x, ((1 + a\epsilon)u), \\ h_6.(t, x, u) \mapsto (t, x + a\epsilon, u), \\ h_7.(t, x, u) \mapsto (t, x + a\epsilon t, u + a\epsilon). \end{array} \right.$$

Consequently, the invariant solutions of a solution  $u = g(t, x)$  for the KdV-KS equation is given by

$$\left\{ \begin{array}{l} h_1.g(t, x) = g(t - a, x), \\ h_2.g(t, x) = g(t, x - a), \\ h_3.g(t, x) = g(t, x - ta) + a, \\ h_4.g(t, x) = g(t - a\epsilon, x), \\ h_5.g(t, x) = (1 + a\epsilon)g((1 + \frac{3}{2}a\epsilon)t, (1 + \frac{1}{2}a\epsilon)x), \\ h_6.g(t, x) = g(t, x - \epsilon a), \\ h_7.g(t, x) = \epsilon g(t, x - a\epsilon t) + a. \end{array} \right.$$

It would be useful to determine the minimal collection of subgroups that will produce all potential group invariant solutions since a solution can be utilized to construct additional solutions using various groups. An optimal system, which is a collection of such solutions, is created by analyzing the manner in which group invariant solutions change one another via the adjoint operation. Thus we will construct a one-dimensional optimal system of approximate Lie subalgebra of perturbed KdV-KS equation by considering an arbitrary element  $y = \sum_{i=1}^7 s_i y_i$  of KdV-KS equation lie algebra  $\mathfrak{g}$ . The map  $G_i^{a_i} : \mathfrak{g} \rightarrow \mathfrak{g}$  given by  $y \rightarrow \text{Ad}(\exp(a_i y_i))y$  is a linear,  $i = 1, \dots, 7$ . By using Table 2 the matrix  $M_i^{a_i}$  of  $G_i^{a_i}$  with respect to the approximate basis is given by

$$\begin{aligned}
 M_1^{a_1} &= \begin{bmatrix} 1 & 0 & 0 & 0 & 0 & 0 & 0 \\ 0 & 1 & 0 & 0 & 0 & 0 & 0 \\ 0 & -a_1 & 1 & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 1 & 0 & 0 & 0 \\ 0 & 0 & 0 & \frac{3}{2}a_1 & 1 & 0 & 0 \\ 0 & 0 & 0 & 0 & 0 & 1 & 0 \\ 0 & 0 & 0 & 0 & 0 & -a_1 & 1 \end{bmatrix}, & M_2^{a_2} &= \begin{bmatrix} 1 & 0 & 0 & 0 & 0 & 0 & 0 \\ 0 & 1 & 0 & 0 & 0 & 0 & 0 \\ 0 & 0 & 1 & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 1 & 0 & 0 & 0 \\ 0 & 0 & 0 & 1 & \frac{1}{2}a_2 & 0 & 0 \\ 0 & 0 & 0 & 0 & 0 & 1 & 0 \\ 0 & 0 & 0 & 0 & 0 & 0 & 1 \end{bmatrix}, \\
 M_3^{a_3} &= \begin{bmatrix} 1 & a_3 & 0 & 0 & 0 & 0 & 0 \\ 0 & 1 & 0 & 0 & 0 & 0 & 0 \\ 0 & 0 & 1 & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 1 & 0 & a_3 & 0 \\ 0 & 0 & 0 & 0 & 1 & 0 & -a_3 \\ 0 & 0 & 0 & 0 & 0 & 1 & 0 \\ 0 & 0 & 0 & 0 & 0 & 0 & 1 \end{bmatrix}, & M_4^{a_4} &= \begin{bmatrix} 1 & 0 & 0 & 0 & 0 & 0 & 0 \\ 0 & 1 & 0 & 0 & 0 & 0 & 0 \\ 0 & 0 & 1 & 0 & 0 & -a_4 & 0 \\ 0 & 0 & 0 & 1 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & 1 & 0 & 0 \\ 0 & 0 & 0 & 0 & 0 & 1 & 0 \\ 0 & 0 & 0 & 0 & 0 & 0 & 1 \end{bmatrix}, \\
 M_5^{a_5} &= \begin{bmatrix} 1 & 0 & 0 & -\frac{3}{2}a_5 & 0 & 0 & 0 \\ 0 & 1 & 0 & 0 & 0 & -\frac{1}{2}a_5 & 0 \\ 0 & 0 & 1 & 0 & 0 & 0 & -\frac{1}{2}a_5 \\ 0 & 0 & 0 & a_5 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & 1 & 0 & 0 \\ 0 & 0 & 0 & 0 & 0 & 1 & 0 \\ 0 & 0 & 0 & 0 & 0 & 0 & 1 \end{bmatrix}, & M_6^{a_6} &= I_7, & M_7^{a_7} &= \begin{bmatrix} 1 & 0 & 0 & 0 & 0 & a_7 & 0 \\ 0 & 1 & 0 & 0 & 0 & 0 & 0 \\ 0 & 0 & 1 & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 1 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & 1 & 0 & 0 \\ 0 & 0 & 0 & 0 & 0 & 1 & 0 \\ 0 & 0 & 0 & 0 & 0 & 0 & 1 \end{bmatrix}.
 \end{aligned}$$

Then it is seen that

$$\begin{aligned}
 G_7^{a_7} \circ G_6^{a_6} \circ \dots \circ G_1^{a_1} : y \mapsto & s_1 y_1 + (a_3 s_1 + s_2 - a_1 s_3) y_2 + s_3 y_3 \\
 & + \left( \frac{3}{2} (a_1 s_5 + a_5 s_1) + s_4 \right) y_4 + s_5 y_5 + \left[ a_3 \left( s_4 + \frac{3}{2} a_1 s_5 \right) \right. \\
 & - \frac{1}{2} a_5 (a_3 s_1 + s_2 - a_1 s_3) - a_4 s_3 \\
 & + \frac{1}{2} a_2 s_5 + s_6 - a_1 s_7 + a_7 s_1 \Big] y_6 \\
 & + \left( s_7 - a_3 s_5 - \frac{1}{2} a_5 s_3 \right) y_7.
 \end{aligned}$$

Now, by setting suitable  $a_i$ , we can easily omit the coefficient of  $y_j$  in several cases, so  $y$  can be reduced and one-dimensional optimal system is provided by

$$y_1 + \delta y_3 + \beta y_4, \quad y_1 + \alpha y_3 + \gamma y_5, \quad y_1 + \alpha y_3 + \lambda y_7, \quad y_2 + v y_4 + \lambda y_7, \quad y_2 + \gamma y_5, \\ y_3 + \beta y_4 + \gamma y_5, \quad y_4 + \lambda y_7, \quad y_5, \quad y_6, \quad y_7,$$

where  $\alpha, \beta, \gamma, \lambda$ , and  $\delta, v \neq 0$  are real numbers.

### 3 Approximate symmetry reductions

In this section, we consider some reductions of equation (1) corresponding to some approximate Lie symmetries [13, 11, 15, 25]. The reduction of PDEs by approximate symmetries is a valuable technique used to simplify the complexity of solving these equations while retaining essential features of their behavior.

**Reduction 3.1.** Similarity variable respect to the symmetry  $y_1$ , is  $u(t, x) = s_0(x) + \epsilon s_1(x) + O(\epsilon^2)$ , substituting into equation (1). Comparing of the constant and  $\epsilon$  coefficients, we get the following ordinary differential equation (ODE) system:

$$\begin{cases} (\frac{s_0^2}{2})' + s_0^{(3)} = 0, \\ (s_0 s_1)' + s_1^{(3)} + s_0'' - (\frac{s_0^2}{2})'' = 0, \end{cases}$$

where  $'$  shows the derivative respect to  $x$ .

Note that the symmetry  $y_2$  produces the trivial solutions.

**Reduction 3.2.** Similarity variable of  $y_3$  is  $u(t, x) = \frac{x}{t} + s_0(t) + \epsilon s_1(t) + O(\epsilon^2)$ , where  $s_0(t)$  and  $s_1(t)$  admit the following first order ODE system:

$$\begin{cases} t s_0' + s_0 = 0, \\ t s_1' + s_1 = 0, \end{cases}$$

where  $'$  shows the derivative respect to  $t$ . Therefore, we find that  $s_0(t) = \frac{c_0}{t}$  and  $s_1(t) = \frac{c_1}{t}$ . Thus we have

$$u(t, x) = \frac{x}{t} + \frac{c_0}{t} + \epsilon \frac{c_1}{t} + O(\epsilon^2).$$

**Reduction 3.3.** Similarity variable respect to the symmetry  $y_4$ , is  $u(t, x) = s_0(x) + \epsilon s_1(t, x) + O(\epsilon^2)$ . Substituting into (1), it satisfies the following PDE:

$$\begin{cases} (\frac{s_0^2}{2})' + s_0^{(3)} = 0, \\ s_{1,t} + (s_0 s_1)' + s_1^{(3)} + s_0'' - (\frac{s_0^2}{2})'' = 0, \end{cases}$$

where  $'$  shows the derivative respect to  $x$ .

**Reduction 3.4.** For the approximate operator  $y_5$  the similarity variables are  $\eta = \frac{t}{x^3}$ ,  $u(t, x) = \frac{s_0(\eta)}{x^2} + \epsilon \frac{s_1(\eta)}{x^\alpha} + O(\epsilon^2)$ . Therefore,  $s_0$  satisfies the following reduced equation:

$$27\eta^3 s_0^{(3)} + 90\eta^2 s_0'' + 186\eta s_0' + 3\eta s_0' s_0 - s_0' + 2s_0^2 + 24s_0 = 0,$$

where  $s_0' = \frac{ds_0}{d\eta}$ . Also,  $s_1$  and  $\alpha$  may be determined by following equation:

$$(\frac{s_1}{x^\alpha})_{,t} + (\frac{s_0 s_1}{x^{\alpha+2}})_{,x} + (\frac{s_1}{x^\alpha})_{,xxx} + (\frac{s_0}{x^2})_{,xx} + (\frac{s_0}{x^2})_{,xxx} = 0.$$

**Reduction 3.5.** Similarity variable of  $y_6$  is  $u(t, x) = c_0 + \epsilon s_1(t, x) + O(\epsilon^2)$ , where  $s_1(t, x)$  admits the following PDE equation:

$$s_{1,t} + c_0 s_{1,x} + s_{1,xxx} = 0.$$

**Reduction 3.6.** Similarity variable of  $y_7$  is  $u(t, x) = \frac{x}{t} + \frac{c_0}{t} + \epsilon s_1(t, x) + O(\epsilon^2)$ , where  $s_1(t, x)$  admits the following PDE equation:

$$s_1 + t s_{1,t} + (c_0 + x) s_{1,x} + t s_{1,xxx} = 0.$$

**Reduction 3.7.** Similarity variables respect to the symmetry  $y_1 + y_6$  are  $u(t, x) = s(\eta)$ ,  $\eta = x - \epsilon t$ . We see that the parameter  $\epsilon$  does not appear directly, but it is instead implicitly contained within the relevant variables. Substituting it into equation (1), we obtain the following reduced approximate ODE:

$$s s' + s^{(3)} + \epsilon(-s' + s'' + s^{(4)}) = 0,$$

where  $s' = \frac{ds}{d\eta}$ . Integrating this ODE under the conditions  $s(\mp\infty) = 0$ ,  $s'(\mp\infty) = 0$ ,  $s''(\mp\infty) = 0$ , and  $s^{(3)}(\mp\infty) = 0$ , and setting the integral constant to zero result in

$$\frac{s^2}{2} + s'' + \epsilon(-s + s' + s^{(3)}) = 0.$$

This equation can be given as

$$\begin{cases} \frac{ds}{d\eta} = v, \\ \frac{dv}{d\eta} = w, \\ \epsilon \frac{dw}{d\eta} = -\frac{s^2}{2} - w - \epsilon v + \epsilon s. \end{cases} \quad (5)$$

By putting  $\epsilon = 0$ , in the above slow system, the critical manifold  $M_0$  is any compact subset contained in the set of critical points  $\{(s, v, w) \mid w = -\frac{s^2}{2}\}$  (see a complete information about critical manifold and Fenichel's theorems in [3]). Therefore, the slow flow on  $M_0$  is given by the following system:

$$\begin{cases} \frac{ds}{d\eta} = v, \\ \frac{dv}{d\eta} = -\frac{s^2}{2}. \end{cases} \quad (6)$$

Figure 1 shows the orbit of (6) to the critical point  $(0, 0)$ .

By Fenichel's invariant manifold theorem, for sufficiently small  $\epsilon$ , the slow manifold  $M_\epsilon$  located within  $O(\epsilon)$  of  $M_0$ , that is,

$$w = -\frac{s^2}{2} + \epsilon g(s, v) + O(\epsilon^2).$$

Substituting this relation into the last equation of slow system (5) gives  $g(s, v) = (s - 1)v + s$ .

## 4 Conclusion

In this research, the perturbed KdV-KS equation was studied using Lie approximation symmetry analysis. We were able to reduce this problem using similarity Lie approximation algebra. Based on the optimal system approach, all of the group-invariant solutions to equation (1) are taken into consideration. Wide classes of nonlinear differential equations can be effectively solved using the fundamental concept provided in this study.

## Conflict of interest

The authors have no conflict of interest to declare that are relevant to this article.

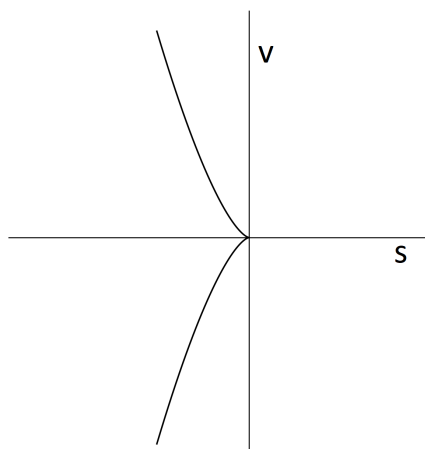


Figure 1: The orbit of (6) to the critical point  $(0, 0)$

## References

- [1] Abbagari, S., Houwe, A., Saliou, Y., Douvagai, D., Chu, Y.M., Inc, M., Rezazadeh, H. and Doka, S.Y. *Analytical survey of the predator–prey model with fractional derivative order*, AIP Advances, 11 (3) (2021) 035127.
- [2] Akbulut, A.R.Z.U., Mirzazadeh, M., Hashemi, M.S., Hosseini, K., Salahshour, S. and Park, C. *Triki–biswas model: Its symmetry reduction, Nucci’s reduction and conservation laws*, Int. J. Mod. Phys. B. 37 (07) (2023) 2350063.
- [3] Arnold, L., Jones, C.K., Mischaikow, K., Raugel, G. and Jones, C.K., *Geometric singular perturbation theory*, Dynamical Systems: Lectures Given at the 2nd Session of the Centro Internazionale Matematico Estivo (CIME) held in Montecatini Terme, Italy, June 13–22, 1994 (1995): 44–118.

- [4] Cheng, X., Hou, J. and Wang, L. *Lie symmetry analysis, invariant subspace method and q-homotopy analysis method for solving fractional system of single-walled carbon nanotube*, Comput. Appl. Math. 40 (2021) 1–17.
- [5] Chentouf, B. and Guesmia, A. *Well-posedness and stability results for the Korteweg-de vries–Burgers and Kuramoto–Sivashinsky equations with infinite memory: A history approach*, Nonlinear Analysis: Real World Applications 65 (2022) 103508.
- [6] Chu, Y., Khater, M.M. and Hamed, Y. *Diverse novel analytical and semi-analytical wave solutions of the generalized  $(2+1)$ -dimensional shallow water waves model*, AIP Advances ,11 (1) (2021) 015223.
- [7] Chu, Y., Shallal, M.A., Mirhosseini-Alizamini, S.M., Rezazadeh, H., Javeed, S. and Baleanu, D. *Application of modified extended tanh technique for solving complex ginzburg-landau equation considering Kerr law nonlinearity*, Comput. Mater. Contin. 66 (2) (2021) 1369–1377.
- [8] Chu, Y.-M., Inc, M., Hashemi, M.S., and Eshaghi, S. *Analytical treatment of regularized prabhakar fractional differential equations by invariant subspaces*, Comput. Appl. Math. 41 (6) (2022) 271.
- [9] Cinar, M., Secer, A., Ozisik, M. and Bayram, M. *Optical soliton solutions of  $(1+1)$ -and  $(2+1)$ -dimensional generalized Sasa–Satsuma equations using new kudryashov method*, Int. J. Geom. Methods Mod. Phys. 20 (02) (2023) 2350034.
- [10] Du, Z. and Li, J. *Geometric singular perturbation analysis to Camassa–Holm Kuramoto–Sivashinsky equation*, J. Differ. Equ. 306 (2022) 418–438.
- [11] Euler, N., Shul’ga, M.W. and Steeb, W.-H. *Approximate symmetries and approximate solutions for a multidimensional Landau–Ginzburg equation*, J. Phys. A Math. Gen. 25 (18) (1992) L1095.
- [12] Fan, X. and Tian, L. *The existence of solitary waves of singularly perturbed MKdV–KS equation*, Chaos Solit. Fract. 26 (4) (2005) 1111–1118.

- [13] Grebenev, V. and Oberlack, M. *Approximate lie symmetries of the Navier-stokes equations*, J. Nonlinear Math. Phys. 14 (2) (2007) 157–163.
- [14] Hashemi, M.S. and Mirzazadeh, M. *Optical solitons of the perturbed nonlinear Schrödinger equation using lie symmetry method*, Optik 281 (2023) 170816.
- [15] Kara, A., Mahomed, F. and Unal, G. *Approximate symmetries and conservation laws with applications*, Int. J. Theor. Phys. 38 (9) (1999) 2389–2399.
- [16] Malik, S., Hashemi, M.S., Kumar, S., Rezazadeh, H., Mahmoud, W. and Osman, M. *Application of new Kudryashov method to various nonlinear partial differential equations*, Opt. Quantum Electron. 55 (1) (2023) 8.
- [17] Mohammadpour, A., Hasannejad, S. and Haji Badali, A. *The study of maximal surfaces by Lie symmetry*, Comput. Method. Differ. Equ. (2024) 1–8.
- [18] Mohammadpour, A., Hashemi, M.S. and Samaei, S. *Noether symmetries and isometries of the area-minimizing Lagrangian on vacuum classes of pp-waves*, Eur. Phys. J. Plus, 138 (2) (2023) 1–7.
- [19] Mohammadpour, A., Hashemi, M.S., Samaei, S. and Salar Anvar, S. *Symmetries of the minimal Lagrangian hypersurfaces on cylindrically symmetric static space-times*, Comput. Method. Differ. Equ. 13(1) (2025) 249–257.
- [20] Ozisik, M., Secer, A., Bayram, M., Sulaiman, T.A. and Yusuf, A. *Acquiring the solitons of inhomogeneous Murnaghan’s rod using extended Kudryashov method with Bernoulli–Riccati approach*, Int. J. Modern Phys. B 36 (30) (2022) 2250221.
- [21] Prakash, P., Priyendhu, K. and Lakshmanan, M. *Invariant subspace method for  $(m+1)$ -dimensional non-linear time-fractional partial differential equations*, Commun. Nonlinear Sci. Numer. Simul. 111 (2022) 106436.

- [22] Sahoo, S., Saha Ray, S., Abdou, M.A.M., Inc, M. and Chu, Y.M. *New soliton solutions of fractional Jaulent-Miodek system with symmetry analysis*, Symmetry 12 (6) (2020) 1001.
- [23] Triki, H., Mirzazadeh, M., Ahmed, H.M., Samir, I. and Hashemi, M.S., *Higher-order Sasa-Satsuma equation: Nucci's reduction and soliton solutions*, Eur. Phys. J. Plus, 138 (5) (2023) 1–10.
- [24] Xia, F.L., Jarad, F., Hashemi, M.S. and Riaz, M.B. *A reduction technique to solve the generalized nonlinear dispersive mk (m, n) equation with new local derivative*, Results Phys. 38 (2022) 105512.
- [25] Yen, T.C., Lang, R.A. and Izmaylov, A.F. *Exact and approximate symmetry projectors for the electronic structure problem on a quantum computer*, J. Chem. Phys. 151 (16) (2019) 164111 .



# Convergence analysis of triangular and symmetric splitting method for fuzzy stochastic linear systems

B. Harika, D. Rajaiah\*, A. Shivaji and L.P. Rajkumar

## Abstract

In this article, the triangular and symmetric splitting iterative method is suggested for solving linear homogeneous systems of equations  $\pi Q = 0$ ,

---

\*Corresponding author

Received 8 December 2024; revised 17 March 2025; accepted 19 March 2025

Bolledla Harika

Department of Mathematics, Kakatiya University, Telangana, India. e-mail: hbolledla@gmail.com

Rajaiah Dasari

Department of Mathematics, Kakatiya Institute of Technology and Science, Warangal, India. e-mail: dr.mh@kitsw.ac.in

Shivaji Arepelly

Department of Mathematics, Kakatiya University, Telangana, India. e-mail: sivarajarepalli@gmail.com

Ladalla Rajkumar

Department of Mathematics, Kakatiya University, Telangana, India. e-mail: ladalla.raj@gmail.com

---

## How to cite this article

Harika, B., Rajaiah, D., Shivaji, A. and Rajkumar, L.P., Convergence analysis of triangular and symmetric splitting method for fuzzy stochastic linear systems. *Iran. J. Numer. Anal. Optim.*, 2025; 15(3): 930-951. <https://doi.org/10.22067/ijnao.2025.91136.1561>

where  $Q$  is the stochastic rate matrix and  $\pi$  is the steady state vector. The homogeneous system is converted to the nonhomogeneous regularized fuzzy linear system  $Ax = b$  with the small perturbation parameter  $0 < r \leq 1$ . The regularized fuzzy linear system is converted into an embedded linear system. The iterative scheme is established; convergence criteria and its sensitivity analysis are analyzed using the numerical examples and convergence theorems. From the numerical results, it is evident to conclude that the proposed method is effective and efficient compared to the theoretical results.

**AMS subject classifications (2020):** Primary 65F10; Secondary 08A72.

**Keywords:** Stochastic rate matrix; Triangular and symmetric splitting Method; Fuzzy liner system; Error analysis.

## 1 Introduction

Fuzzy linear systems (FLS) and Fully Fuzzy linear systems (FFLS) have great applications in various areas of engineering, science, and social sciences, such as physics, statistics, operational research, control problems, neural networks, communication systems, sensors, and economics. The mathematical modeling of a physical problem is formulated into the system of fuzzy linear homogeneous or nonhomogeneous equations. There are many methods in the literature to solve the nonhomogeneous FLSs. The homogeneous FLS has either a trivial solution or an infinite number of solutions. For a unique non-trivial solution, the homogeneous system  $\pi Q = 0$ , where  $\pi$  is the steady state vector and  $Q$  is the stochastic rate matrix, is converted into the regularized nonhomogeneous fuzzy linear system  $Ax = b$  with the small perturbation  $0 < r \leq 1$ . The regularized FLS is converted into an embedded linear system  $\Theta X = \Upsilon$ , where  $\Theta$  and  $\Upsilon$  are fuzzy matrices and  $X$  is an unknown fuzzy vector. Many straight forward methods are existing in the literature to find the unique non-zero solution of pertinent to linear systems when the coefficient matrix is a crisp matrix. However, in actual cases, the parameters may be uncertain or vague. So, to overcome the uncertainty and vagueness,

the coefficient matrix of the system  $Ax = b$  is assumed as a fuzzy stochastic matrix instead of the crisp matrix.

Many researchers proposed direct and iterative methods to solve FLSs. The first iterative model with an embedding technique for computing a class of  $n \times n$  FLS was triggered by Friedman, Ming, and Kandel [11]. For solving a system of fuzzy linear equations, a few numerical methods were developed and discussed for the existence of solution, provided that the diagonal elements are positive and satisfy the diagonal dominance property by Dehghan, Hashemi and Ezzati [6, 9]. The steepest descent method and LU decomposition method were developed in [1, 2]. Allahviranloo [3, 4] used the Jacobi, Gauss–Seidel, SOR, iterative methods for finding the approximate solution of the FLS. A fuzzy system of linear equations with crisp coefficients was proposed by Chakraverty and Behera [5]. The inherited LU factorization method was proposed by Fariborzi Araghi and Fallahzadeh [10] for solving a fuzzy systems of linear equations. Koam et al. [13] used the LU decomposition scheme for solving  $m$ -polar fuzzy system of linear equations. Block SOR method for FLSs was proposed by Miao, Zheng, and Wang [14], and the QR-decomposition method was developed by S.H. Nasseri, Matinfar, and Sohrabi [15]; Wang and Wu [17] introduced the Uzawa-SOR method. Symmetric successive over relaxation method, block iterative method, and splitting iterative methods were established by Wang, Zheng and Yin [18, 19, 21]. Wang and Chen [20] suggested a modified Jacobi iterative method for large-size linear systems, and a new method based on Jacobi iteration was proposed to solve the FLSs by Zhen et al. [12]. If the coefficient matrix is crisp, then it will restrict the modeling of the real-time problems. In the system of linear equations, both the coefficient matrix and right-hand side matrices are fuzzy matrices. then it is defined as an FFLS. FFLS gives wide scope in real-time applications by removing the crispness in the left-hand side coefficient matrix. The iterative solution of general FFLS is proposed in [7]. Edalatpanah [8] proposed a modified iterative method for finding the solution of FFLS. Classical triangular and symmetric (TS) splitting methods are simple to implement and suitable to find the steady state probability vector and performance measures in many real time systems [16]. In this paper, a new improved method based on TS iteration is provided for solving FFLSs. The

rest of the paper is organized as follows: Section 2 gives some basic definitions and results of FLS. In section 3, the new method is established with convergence theorems. Perturbation analysis is discussed in section 4. Numerical examples are presented in section 5, and the conclusions are in section 6.

## 2 Basic definitions of FLSs and convergence analysis of TS method

In this section, we have defined the FLS and some basic definitions such as fuzzy numbers, fuzzy solutions, arithmetic operations on fuzzy numbers, and embedded model of FFLS, which are useful in the numerical solution of FLS.

**Fuzzy number:** An arbitrary form of fuzzy number is an ordered pair of functions  $(\underline{v}(r), \bar{v}(r))$ ,  $0 < r \leq 1$ , satisfying

- $\underline{v}(r)$  is a bounded monotonic increasing left continuous function over  $[0, 1]$ ,
- $\bar{v}(r)$  is a bounded monotonic decreasing left continuous function over  $[0, 1]$ ,
- $\underline{v}(r) \leq \bar{v}(r)$ ,  $0 < r \leq 1$ .

**Arithmetic operations on fuzzy numbers:** If  $u = (\underline{u}(r), \bar{u}(r))$  and  $v = (\underline{v}(r), \bar{v}(r))$  are arbitrary fuzzy numbers, then the arithmetic operations of arbitrary fuzzy numbers for  $0 < r \leq 1$  and real number  $k$ , are defined as follows:

- $u = v$  if and only if  $\underline{u}(r) = \underline{v}(r)$  and  $\bar{u}(r) = \bar{v}(r)$ ,
- $u + v = (\underline{u}(r) + \underline{v}(r), \bar{u}(r) + \bar{v}(r))$ , and
- $ku = \begin{cases} (k\underline{u}(r), k\bar{u}(r)), & k > 0, \\ (k\bar{u}(r), k\underline{u}(r)), & k < 0. \end{cases}$

**Fuzzy linear system:** The  $n \times n$  FLS is defined as

$$\begin{aligned} a_{11}x_1 + a_{12}x_2 + \cdots + a_{1n}x_n &= b_1, \\ a_{21}x_1 + a_{22}x_2 + \cdots + a_{2n}x_n &= b_2, \end{aligned}$$

$$\begin{array}{c} \vdots \\ a_{n1}x_1 + a_{n2}x_2 + \cdots + a_{nn}x_n = b_n. \end{array}$$

The matrix form of the above linear system is

$$Ax = b, \quad (1)$$

where

$$A = \begin{bmatrix} a_{11} & a_{12} & \cdots & a_{1n} \\ a_{21} & a_{22} & \cdots & a_{2n} \\ \vdots & \vdots & \ddots & \vdots \\ a_{n1} & a_{n2} & \cdots & a_{nn} \end{bmatrix}.$$

is a crisp matrix,  $b = [b_1, b_2, \dots, b_n]^T$  is a fuzzy vector, and  $x = [x_1, x_2, \dots, x_n]^T$  is unknown.

**Solution of an FLS:** A fuzzy vector  $x = (x_1, x_2, \dots, x_n)^T$  given by  $x_i = (\underline{x}_i(r), \bar{x}_i(r))$ ,  $1 \leq i \leq n$ ,  $0 < r \leq 1$ , is called a solution of the FLS (1) if

$$\left\{ \begin{array}{l} \overline{\sum_{j=1}^n a_{ij}x_j} = \sum_{j=1}^n \overline{a_{ij}x_j} = \underline{b}_i, \\ \underline{\sum_{j=1}^n a_{ij}x_j} = \sum_{j=1}^n \underline{a_{ij}x_j} = \bar{b}_i. \end{array} \right. \quad (2)$$

**Embedded Model of FLS:** The embedding model of extended FLS (1) into the  $2n \times 2n$  crisp linear system is defined as

$$\begin{array}{l} \theta_{1,1}\underline{x}_1 + \cdots + \theta_{1,n}\underline{x}_n + \theta_{1,n+1}(-\bar{x}_1) + \cdots + \theta_{1,2n}(-\bar{x}_n) = \underline{b}_1, \\ \theta_{2,1}\underline{x}_1 + \cdots + \theta_{2,n}\underline{x}_n + \theta_{2,n+1}(-\bar{x}_1) + \cdots + \theta_{2,2n}(-\bar{x}_n) = \underline{b}_2, \\ \vdots \\ \theta_{n,1}\underline{x}_1 + \cdots + \theta_{n,n}\underline{x}_n + \theta_{n,n+1}(-\bar{x}_1) + \cdots + \theta_{n,2n}(-\bar{x}_n) = \underline{b}_n, \\ \theta_{n+1,1}\underline{x}_1 + \cdots + \theta_{n+1,n}\underline{x}_n + \theta_{n+1,n+1}(-\bar{x}_1) + \cdots + \theta_{n+1,2n}(-\bar{x}_n) = \bar{b}_1, \\ \vdots \\ \theta_{2n,1}\underline{x}_1 + \cdots + \theta_{2n,n}\underline{x}_n + \theta_{2n,n+1}(-\bar{x}_1) + \cdots + \theta_{2n,2n}(-\bar{x}_n) = \bar{b}_n. \end{array}$$

The matrix form of above  $2n \times 2n$  linear system is

$$\Theta X = \Upsilon, \quad (3)$$

where  $\Theta = (\theta_{kl})$ ,  $\theta_{kl}$  are determined as follows:

1. For  $a_{ij} > 0$ ,  $\theta_{ij} = a_{ij}$ ,  $\theta_{n+i, n+j} = a_{ij}$ .
2. For  $a_{ij} < 0$ ,  $\theta_{i, n+j} = a_{ij}$ ,  $\theta_{n+i, j} = a_{ij}$ ,  $1 \leq i, j \leq 2n$ .
3.  $\theta_{kl} = 0$  if it is not presented in above system, and

$$X = \begin{bmatrix} \underline{x}_1 \\ \vdots \\ \underline{x}_n \\ \overline{x}_1 \\ \vdots \\ \overline{x}_n \end{bmatrix} \quad \text{and} \quad \Upsilon = \begin{bmatrix} \underline{b}_1 \\ \vdots \\ \underline{b}_n \\ \overline{b}_1 \\ \vdots \\ \overline{b}_n \end{bmatrix}.$$

Furthermore, the matrix  $\Theta$  has the structure  $\begin{bmatrix} \Theta_1 & \Theta_2 \\ \Theta_2 & \Theta_1 \end{bmatrix}$ ,  $\Theta = \Theta_1 + \Theta_2$ , and (2) can be written as

$$\begin{cases} \Theta_1 \underline{X} + \Theta_2 \overline{X} = \underline{\Upsilon}, \\ \Theta_2 \underline{X} + \Theta_1 \overline{X} = \overline{\Upsilon}, \end{cases} \quad (4)$$

where

$$\underline{X} = \begin{bmatrix} \underline{x}_1 \\ \underline{x}_2 \\ \vdots \\ \underline{x}_n \end{bmatrix}, \quad \overline{X} = \begin{bmatrix} \overline{x}_1 \\ \overline{x}_2 \\ \vdots \\ \overline{x}_n \end{bmatrix},$$

$$\underline{\Upsilon} = \begin{bmatrix} \underline{b}_1 \\ \underline{b}_2 \\ \vdots \\ \underline{b}_n \end{bmatrix}, \quad \text{and} \quad \overline{\Upsilon} = \begin{bmatrix} \overline{b}_1 \\ \overline{b}_2 \\ \vdots \\ \overline{b}_n \end{bmatrix}.$$

In the next section, a new iterative scheme based on TS iteration is presented for regularized linear system with nonsingular coefficient matrix [16].

### 3 Fuzzy TS splitting iterative method for regularized linear system

In this section, we find the steady state probability vector  $\pi$  of a homogeneous equation  $\pi Q = 0$ . The solution of the homogeneous system is either a trivial solution or an infinite number of solutions. For a unique, nonzero solution, the above homogeneous system is converted to the regularized FLS and is equivalent embedded crisp system  $\Theta x = b$ , using the small perturbation  $0 < r \leq 1$ . Now, the TS splitting method for the transition matrix is adopted in a fuzzy environment. Let the coefficient matrix  $\Theta$  of the embedded regularized linear system can be split in the form:

$$\Theta = (L + D - U^T) + (U + U^T) = T + S,$$

where  $T = L + D - U^T$  and  $S = (U + U^T)$  are TS matrices.

The regularized system, (3) can be expressed as  $(T + S)X = b$ . Consider

$$D = \begin{bmatrix} D_1 & 0 \\ 0 & D_1 \end{bmatrix}, L = \begin{bmatrix} L_1 & 0 \\ -S_2 & L_1 \end{bmatrix}, U = \begin{bmatrix} U_1 & -S_2 \\ 0 & U_1 \end{bmatrix},$$

where  $D_1 = \text{diag}(s_{ii})$ ,  $L_1$ , and  $U_1$  are diagonal, lower, and upper triangular matrices, respectively. Now,

$$\begin{aligned} T = L + D - U^T &= \begin{bmatrix} D_1 & 0 \\ 0 & D_1 \end{bmatrix} + \begin{bmatrix} L_1 & 0 \\ -S_2 & L_1 \end{bmatrix} - \begin{bmatrix} U_1 & 0 \\ -S_2 & U_1 \end{bmatrix} \\ &= \begin{bmatrix} L_1 + D_1 - U_1 & 0 \\ 0 & L_1 + D_1 - U_1 \end{bmatrix}, \end{aligned}$$

and

$$S = U + U^T = \begin{bmatrix} 2U_1 & -S_2 \\ -S_2 & 2U_1 \end{bmatrix}.$$

The TS splitting iterative scheme is as follows [16]:

$$\begin{aligned} (\alpha I + T)X^{(k+1/2)} &= (\alpha I - S)X^{(k)} + b, \\ (\alpha I + S)X^{(k+1)} &= (\alpha I - T)X^{(k+1/2)} + b. \end{aligned}$$

The above iterative scheme could be written as

$$X^{(k+1)} = M(\alpha)X^{(k)} + N(\alpha)b, \quad \text{for } k = 0, 1, 2, \dots,$$

where

$$X^{(k+1)} = \left[ \frac{X^{(k+1)}}{X^{(k+1)}} \right]$$

$$M(\alpha) = (\alpha I_n + S)^{-1}(\alpha I_n - T)(\alpha I_n + T)^{-1}(\alpha I_n - S)$$

and

$$N(\alpha) = 2\alpha(\alpha I_n + S)^{-1}(\alpha I_n + T)^{-1}.$$

We have

$$\begin{aligned} \alpha I_n + S &= \begin{bmatrix} \alpha I_n + 2U_1 & -S_2 \\ -S_2 & \alpha I_n + 2U_1 \end{bmatrix}, \\ \alpha I_n - S &= \begin{bmatrix} \alpha I_n - 2U_1 & S_2 \\ S_2 & \alpha I_n - 2U_1 \end{bmatrix}, \\ \alpha I_n + T &= \begin{bmatrix} \alpha I_n + T_1 & 0 \\ 0 & \alpha I_n + T_1 \end{bmatrix}, \\ \alpha I_n - T &= \begin{bmatrix} \alpha I_n - T_1 & 0 \\ 0 & \alpha I_n - T_1 \end{bmatrix}. \end{aligned}$$

Thus,

$$M(\alpha) = \frac{\alpha I_n - T_1}{(\alpha I_n + T_1)[(\alpha I_n + 2U_1)^2 - S_2^2]} \begin{bmatrix} (\alpha I_n)^2 - 4U_1^2 + S_2^2 & 2\alpha I_n S_2 \\ 2\alpha I_n S_2 & (\alpha I_n)^2 - 4U_1^2 + S_2^2 \end{bmatrix}$$

$$\text{and } N(\alpha) = \frac{2\alpha}{(\alpha I_n + 2U_1)^2 - S_2^2} \begin{bmatrix} \frac{\alpha I_n + 2U_1}{\alpha I_n + T_1} & \frac{S_2}{\alpha I_n + T_1} \\ \frac{S_2}{\alpha I_n + T_1} & \frac{\alpha I_n + 2U_1}{\alpha I_n + T_1} \end{bmatrix}.$$

**Theorem 1.** If  $\Theta X = b$  is the regularized FLS, then the convex solution  $X = \{r\underline{X}_j + (1-r)\overline{X}_j/0 < r \leq 1\}$  is the solution.

*Proof.* Let  $\underline{X}_j$  and  $\overline{X}_j$  be the solutions corresponding to right-hand side vector  $\underline{b}_j$  and  $\overline{b}_j$  of

$$\sum_{j=1}^n a_{ij} X_j = b_j,$$

which implies,

$$\sum_{j=1}^n a_{ij} \underline{X}_j = \underline{b}_j,$$

$$\sum_{j=1}^n a_{ij} \overline{X}_j = \overline{b}_j.$$

We prove that  $X_j = r\underline{X}_j + (1-r)\overline{X}_j$  is the solution.

Now,

$$\begin{aligned} \sum_{j=1}^n a_{ij} X_j &= \sum_{j=1}^n a_{ij} [r\underline{X}_j + (1-r)\overline{X}_j] \\ &= r \sum_{j=1}^n a_{ij} \underline{X}_j + (1-r) \sum_{j=1}^n a_{ij} \overline{X}_j \\ &= r\underline{b}_j + (1-r)\overline{b}_j \\ &= b_j. \end{aligned}$$

Therefore,  $X_j = r\underline{X}_j + (1-r)\overline{X}_j$  is the solution of the given system.  $\square$

**Theorem 2.** If  $\Theta X = b$ , then the convex solution  $X = \{r\underline{X} + (1-r)\overline{X} / 0 < r \leq 1\}$  is the solution vector of the system  $\Theta\{\underline{X} + \overline{X}\} = \{\underline{b} + \overline{b}\}$ .

*Proof.* The regularized linear system  $\Theta X = b$  can be written as

$$\sum_{j=1}^n a_{ij} X_j = b_i \quad \text{for } i = 1, 2, \dots, n.$$

Let

$$X_j = r\underline{X}_j + (1-r)\overline{X}_j$$

and

$$b_j = [\underline{b}_j, \overline{b}_j] \quad \text{for } i = 1, 2, \dots, n.$$

Now

$$\begin{aligned}\sum_{j=1}^n a_{ij} X_j &= \sum_{j=1}^n a_{ij} \left[ r \underline{X}_j + (1-r) \overline{X}_j \right] = [\underline{b}_i, \overline{b}_i], \\ \sum_{a_{ij}>0} r a_{ij} \underline{X}_j + \sum_{a_{ij}<0} (1-r) a_{ij} \overline{X}_j &= \underline{b}_i, \\ \text{and} \\ \sum_{a_{ij}>0} (1-r) a_{ij} \underline{X}_j + \sum_{a_{ij}<0} r a_{ij} \overline{X}_j &= \overline{b}_i.\end{aligned}$$

Consider

$$\begin{aligned}\Theta [\underline{X} + \overline{X}] &= \sum_{a_{ij}>0} a_{ij} \left[ r \underline{X}_j + (1-r) \overline{X}_j \right] + \sum_{a_{ij}<0} a_{ij} \left[ (1-r) \overline{X}_j + r \underline{X}_j \right] \\ &= \left[ \sum_{a_{ij}>0} r a_{ij} \underline{X}_j + \sum_{a_{ij}<0} (1-r) a_{ij} \overline{X}_j \right] \\ &\quad + \left[ \sum_{a_{ij}>0} (1-r) a_{ij} \overline{X}_j + \sum_{a_{ij}<0} r a_{ij} \underline{X}_j \right] \\ &= [\underline{b}_i + \overline{b}_i], \\ \Theta [\underline{X} + \overline{X}] &= [\underline{b} + \overline{b}].\end{aligned}$$

This proves that  $X$  is the solution.  $\square$

**Theorem 3.** If  $\Theta X = b$ , then the convex solution  $X = \{r \underline{X} - (1-r) \overline{X}\}$ ,  $0 < r \leq 1$  is the solution vector of the system  $\Theta\{\underline{X} - \overline{X}\} = \{\underline{b} - \overline{b}\}$

*Proof.* The system  $\Theta X = b$  can be written as

$$\sum_{j=1}^n a_{ij} X_j = b_i \quad \text{for } i = 1, 2, \dots, n.$$

Now, we may write the real fuzzy unknown and the right-hand real fuzzy number vectors as

$$X_j = r \underline{X}_j - (1-r) \overline{X}_j,$$

and

$$b_j = [\underline{b}_j, \overline{b}_j] \quad \text{for } j = 1, 2, \dots, n,$$

which implies

$$\begin{aligned} \sum_{j=1}^n [a_{ij} r \underline{X}_j + (1-r) \overline{X}_j] &= [\underline{b}_i, \overline{b}_i], \\ \Rightarrow \sum_{a_{ij}>0} r a_{ij} \underline{X}_j + \sum_{a_{ij}<0} (1-r) a_{ij} \overline{X}_j &= \underline{b}_i \\ &\text{and} \\ \sum_{a_{ij}>0} (1-r) a_{ij} \underline{X}_j + \sum_{a_{ij}<0} r a_{ij} \overline{X}_j &= \overline{b}_i. \end{aligned}$$

Consider

$$\begin{aligned} \Theta [\underline{X} - \overline{X}] &= \sum_{a_{ij}>0} a_{ij} [r \underline{X}_j + (1-r) \overline{X}_j] - \sum_{a_{ij}<0} a_{ij} [(1-r) \overline{X}_j + r \underline{X}_j] \\ &= \left[ \sum_{a_{ij}>0} r a_{ij} \underline{X}_j + \sum_{a_{ij}<0} (1-r) a_{ij} \overline{X}_j \right] \\ &\quad - \left[ \sum_{a_{ij}>0} (1-r) a_{ij} \overline{X}_j + \sum_{a_{ij}<0} r a_{ij} \underline{X}_j \right] \\ &= [\underline{b}_i - \overline{b}_i], \\ \Theta [\underline{X} - \overline{X}] &= [\underline{b} - \overline{b}]. \end{aligned}$$

□

**Theorem 4.** If  $\Theta X = b$ , then the mid point solution  $X = \frac{X_j + \overline{X}_j}{2}$ , is the solution vector of the system  $\Theta X = b$ .

*Proof.* The system  $\Theta X = b$  can be written as

$$\sum_{j=1}^n a_{ij} X_j = b_k \quad \text{for } k = 1, 2, \dots, n.$$

The real fuzzy unknown and the right-hand real fuzzy number vectors can be written as

$$\begin{aligned} X_j &= \frac{X_j + \overline{X}_j}{2} \quad \text{and} \quad b_j = [\underline{b}_j, \overline{b}_j] \quad \text{for } j = 1, 2, \dots, n, \\ \sum_{j=1}^n a_{ij} \frac{X_j + \overline{X}_j}{2} &= [\underline{b}_j, \overline{b}_j], \end{aligned}$$

$$\text{which implies } \frac{1}{2} \left[ \sum_{a_{ij}>0} a_{ij} \underline{X}_j + \sum_{a_{ij}<0} a_{ij} \overline{X}_j \right] = \underline{b}_j,$$

and

$$\frac{1}{2} \left[ \sum_{a_{ij}>0} a_{ij} \overline{X}_j + \sum_{a_{ij}<0} a_{ij} \underline{X}_j \right] = \overline{b}_j.$$

Consider

$$\begin{aligned} \Theta \frac{[\underline{X} + \overline{X}]}{2} &= \sum_{j=1}^n a_{ij} \frac{[\underline{x}_j + \overline{x}_j]}{2} \\ \sum_{j=1}^n a_{ij} \frac{[\underline{X}_j + \overline{X}_j]}{2} &= \sum_{a_{ij}>0} a_{ij} \frac{[\underline{x}_j + \overline{x}_j]}{2} + \sum_{a_{ij}<0} a_{ij} \frac{[\overline{x}_j + \underline{x}_j]}{2} \\ &= \frac{\left[ \sum_{a_{ij}>0} a_{ij} \underline{x}_j + \sum_{a_{ij}<0} a_{ij} \overline{x}_j \right]}{2} + \frac{\left[ \sum_{a_{ij}>0} a_{ij} \overline{x}_j + \sum_{a_{ij}<0} a_{ij} \underline{x}_j \right]}{2} \\ &= \frac{[\underline{b}_j + \overline{b}_j]}{2}, \\ \Theta \frac{[\underline{X} + \overline{X}]}{2} &= \frac{[\underline{b} + \overline{b}]}{2}. \end{aligned}$$

Hence,  $X = \frac{\underline{x}_j + \overline{x}_j}{2}$  is solution of the regularized linear system.  $\square$

#### 4 Perturbation analysis of FLS

As discussed in the previous section, the homogeneous system  $\pi Q = 0$ , where the coefficient matrix  $Q$  is circulant stochastic rate matrix, is converted into the regularized FLS  $Ax = b$  using small perturbation  $0 < r \leq 1$ . In this section, perturbation to the FLS is added, in which both the coefficient matrix and right-hand side matrices are perturbed and the sensitivity analysis of  $Ax = b$  is discussed by using the FFTS method. The well-posed and ill-posed solution of the system  $\Theta X = \Upsilon$  depends on the membership value  $0 < r \leq 1$ . If the coefficient matrix  $\Theta$  or the right-hand side fuzzy vector  $\Upsilon$  or both are slightly disturbed with the membership value  $r$ , then the solu-

tion will be changed as well. The relative error and absolute by the FFTS method are evaluated between the exact solution and numerical solution with the perturbed FLS. The following theorems are proved in preparation for investigating the sensitivity analysis of the regularized fuzzy system.

**Theorem 5.** If  $A \in R^{n \times n}$  is a circulant stochastic matrix of the regularized linear system  $Ax = b$  and  $\Theta$  is the embedded matrix of the embedded system  $\Theta X = \Upsilon$ , then  $\Theta$  is positive definite.

*Proof.* For proving the matrix  $\Theta$  is positive definite, it is sufficient to prove that  $\frac{\Theta + \Theta^T}{2}$  is positive definite.

We have

$$\Theta = \begin{pmatrix} c_1 + r & 0 & \dots & 0 & 0 & c_2 & \dots & c_n \\ 0 & c_1 + r & \dots & 0 & c_n & 0 & \dots & c_{n-1} \\ \vdots & \vdots & \vdots & \vdots & \vdots & \ddots & \vdots & \vdots \\ 0 & 0 & \dots & c_1 + r & c_2 & c_3 & 0 \dots & 0 \\ 0 & c_2 & 0 \dots & c_n & c_1 + r & 0 & \dots & 0 \\ c_n & 0 & 0 \dots & c_{n-1} & 0 & c_1 + r & \dots & 0 \\ \vdots & \vdots & \vdots & \vdots & \vdots & \ddots & \vdots & \vdots \\ c_2 & c_3 & \dots & 0 & 0 & 0 & \dots & c_1 + r \end{pmatrix},$$

$$\frac{\Theta + \Theta^T}{2} = \begin{pmatrix} c_1 + r & 0 & \dots & 0 & 0 & \frac{c_2 + c_n}{2} & \dots & \frac{c_2 + c_n}{2} \\ 0 & c_1 + r & \dots & 0 & \frac{c_2 + c_n}{2} & 0 & \dots & \frac{c_3 + c_{n-1}}{2} \\ \vdots & \vdots & \vdots & \vdots & \vdots & \ddots & \vdots & \vdots \\ 0 & 0 & \dots & c_1 + r & \frac{c_2 + c_n}{2} & \frac{c_3 + c_{n-1}}{2} & 0 \dots & 0 \\ 0 & \frac{c_2 + c_n}{2} & 0 \dots & \frac{c_2 + c_n}{2} & c_1 + r & 0 & \dots & 0 \\ \frac{c_2 + c_n}{2} & 0 & 0 \dots & \frac{c_3 + c_{n-1}}{2} & 0 & c_1 + r & \dots & 0 \\ \vdots & \vdots & \vdots & \vdots & \vdots & \ddots & \vdots & \vdots \\ \frac{c_2 + c_n}{2} & \frac{c_3 + c_{n-1}}{2} & \dots & 0 & 0 & 0 & \dots & c_1 + r \end{pmatrix}$$

$$= (c_1 + r)I_{n^2} - R,$$

where

$$R = \begin{pmatrix} 0 & 0 & \dots & 0 & 0 & \frac{c_2+c_n}{2} & \dots & \frac{c_2+c_n}{2} \\ 0 & 0 & \dots & 0 & \frac{c_2+c_n}{2} & 0 & \dots & \frac{c_3+c_{n-1}}{2} \\ \vdots & \vdots & \vdots & \vdots & \vdots & \ddots & \vdots & \vdots \\ 0 & 0 & \dots & 0 & \frac{c_2+c_n}{2} & \frac{c_3+c_{n-1}}{2} & 0 \dots & 0 \\ 0 & \frac{c_2+c_n}{2} & 0 \dots & \frac{c_2+c_n}{2} & 0 & 0 & \dots & 0 \\ \frac{c_2+c_n}{2} & 0 & 0 \dots & \frac{c_3+c_{n-1}}{2} & 0 & 0 & \dots & 0 \\ \vdots & \vdots & \vdots & \vdots & \vdots & \ddots & \vdots & \vdots \\ \frac{c_2+c_n}{2} & \frac{c_3+c_{n-1}}{2} & \dots & 0 & 0 & 0 & \dots & 0 \end{pmatrix} \geq 0.$$

From the theorem in [16], we have

$$\begin{aligned} \Rightarrow \rho(R) &= c_2 + c_3 + \dots + c_n = c_1, \\ \Rightarrow c_1 + \epsilon &> \rho(R). \end{aligned}$$

Therefore,  $\frac{\Theta + \Theta^T}{2}$  is positive definite.

□

**Theorem 6.** [16] For any nonsymmetric stochastic circulant rate matrix  $Q \in R^{n \times n}$ , there exists a constant  $\epsilon > 0$  such that  $A = Q^T + \epsilon I_n$  is positive definite if and only if all its eigenvalues are nonnegative real numbers.

**Theorem 7.** [16] Let  $A \in R^{n \times n}$  be the regularized matrix, and splitting into TS matrices. Then the spectral radius of the iterative matrix  $M(\alpha)$  is less than one.

**Theorem 8.** Let  $\Theta \in R^{2n \times 2n}$  be a fuzzy matrix of regularized linear system and let  $M(\alpha)$  be the iteration matrix of the FFTS iteration method. Then the spectral radius of  $M(\alpha)$  is less than 1.

*Proof.* The proof of the theorem is on the similar lines of Theorems 6 and 7. □

## 5 Numerical results

In this section, we examine the effectiveness of the FFTS iteration method with the numerical solution of stochastic matrices under a fuzzy environment

and compare the error analysis of fuzzy iterative solution with the TS and TSS iteration methods. For the numerical illustration, we consider the homogeneous system  $\pi Q = 0$ , where  $Q$  is the  $3 \times 3$  doubly stochastic rate matrix given below:

$$Q = \begin{bmatrix} 0.6 & -0.35 & -0.25 \\ -0.25 & 0.6 & -0.35 \\ -0.35 & -0.25 & 0.6 \end{bmatrix}.$$

We convert the above homogeneous system  $\pi Q = 0$ , into a regularized linear system  $Ax = b$  with the small perturbation parameter  $0 < r \leq 1$ . The regularized linear system is converted to a  $6 \times 6$  embedded linear system  $\Theta X = \Upsilon$ , where

$$\Theta = \begin{bmatrix} 0.6+r & 0 & 0 & 0 & -0.35 & -0.25 \\ 0 & 0.6+r & 0 & -0.25 & 0 & -0.35 \\ 0 & 0 & 0.6+r & -0.35 & -0.25 & 0 \\ 0 & -0.35 & -0.25 & 0.6+r & 0 & 0 \\ -0.25 & 0 & -0.35 & 0 & 0.6+r & 0 \\ -0.35 & -0.25 & 0 & 0 & 0 & 0.6+r \end{bmatrix}.$$

Let the initial distribution vector be  $x^{(0)} = [0 \ 0 \ 0 \ 0 \ 0 \ 1]^T$  and let right-hand side vector be  $\Upsilon = [0 \ 0 \ 1-r \ 0 \ 0 \ r-1]$ , where  $0 < r \leq 1$  is the membership function. Only one case  $\Theta = (L + D - U^T) + (U + U^T) = T_1 + S_1$  of FFTS splitting method is considered, and other methods would follow the same. A fuzzy iterative solution to the system (1) is computed, and a classical solution of TS and TSS iterative methods is evaluated. It is illustrated the result for the case of contraction factor  $\alpha = 0.6 + r$ , which is numerically equivalent to the diagonal elements of the matrix  $\Theta$ , for different values of  $r$ . The steady state distribution vector  $x$  of the preconditioned linear system is obtained, and the results are presented in Figures 1–7. The numerical solutions of the FFTS method are presented in Figure 1. From this figure, it is concluded that the numerical solutions FFTS method coincides with the theoretical results. The average and linear convex solutions of lower bound and upper bound solutions are depicted in Figures 2–4. From these two figures, it is concluded that the center and convex solution curves lie within the monotonically increasing and monotonically decreasing curve. The error analysis of the FFTS

method and classical iterative solution is presented in Figure 5. From this figure, it concluded that the FFTS iterative solution converges rapidly when compared with classical TS, TSS methods. The convergence and sensitivity analysis of the FFTS method are presented in Figures 6–7. From Figure 6, it is evidently concluded that the condition number of the FFTS method is very low compared with classical TS and TSS methods. From this figure, one can conclude that the iterative solution obtained using the FFTS method is well conditioned and the regularized matrix is nonsingular for larger membership values. It is depicted in Figure 7 that the spectral radius of the FFTS method and spectral radius are clearly less than one. From this figure, it is concluded that the FFTS method converges to a unique nonzero solution, and it is evident by the theoretical solution.

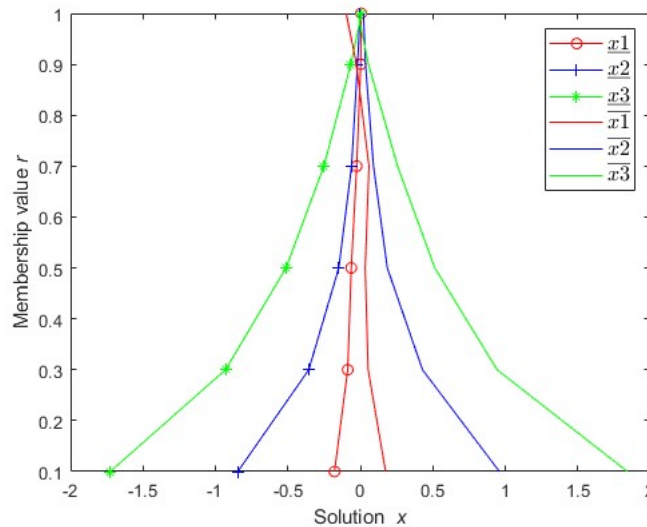


Figure 1: Solution  $x$  for the contraction factor  $\alpha = 0.6$  over the membership value  $r$ .

## 6 Conclusions

In this paper, a new iterative method was suggested based on the TS iteration to the solution of a class of fuzzy linear systems of equations with a coeffi-

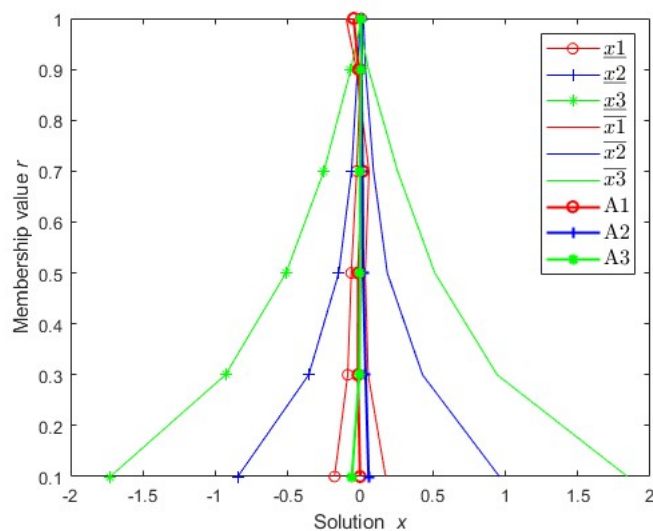


Figure 2: Solution  $x$  and average solution for the contraction factor  $\alpha = 0.6$  over the membership value  $r$ .

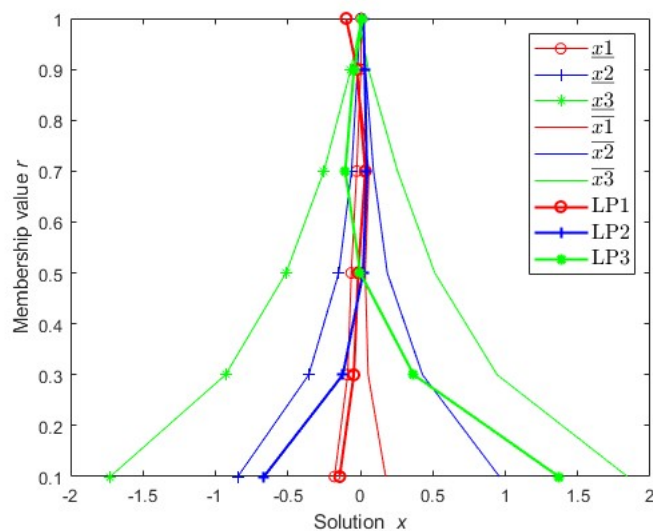


Figure 3: Solution  $x$  and LP solution for the contraction factor  $\alpha = 0.6$  over the membership value  $r$ .

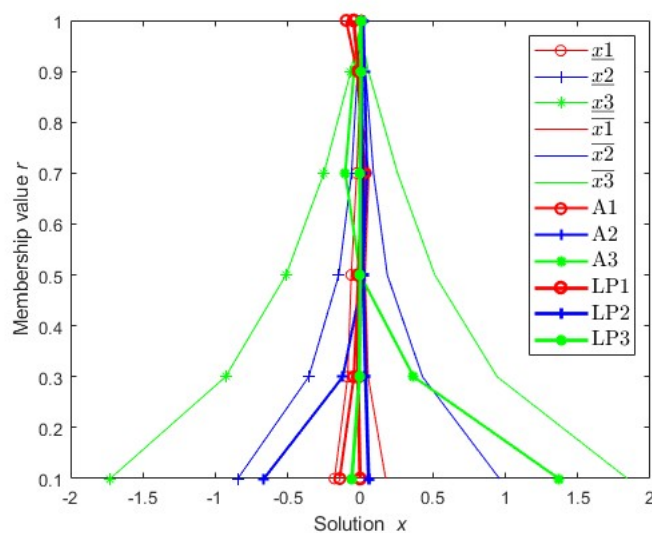


Figure 4: Solution  $x$ , average solution, LP solution for the contraction factor  $\alpha = 0.6$  over the membership value  $r$ .

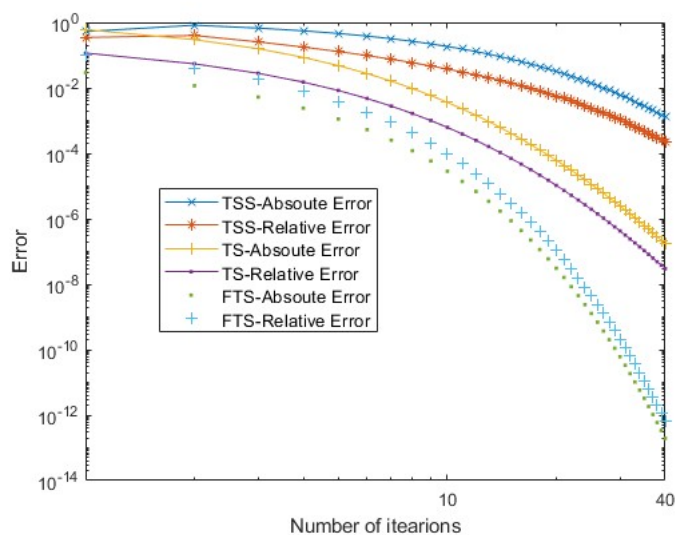


Figure 5: Absolute error and Relative error of the TSS, TS, and FTS iterative solution for the contraction factor  $\alpha = 0.6$

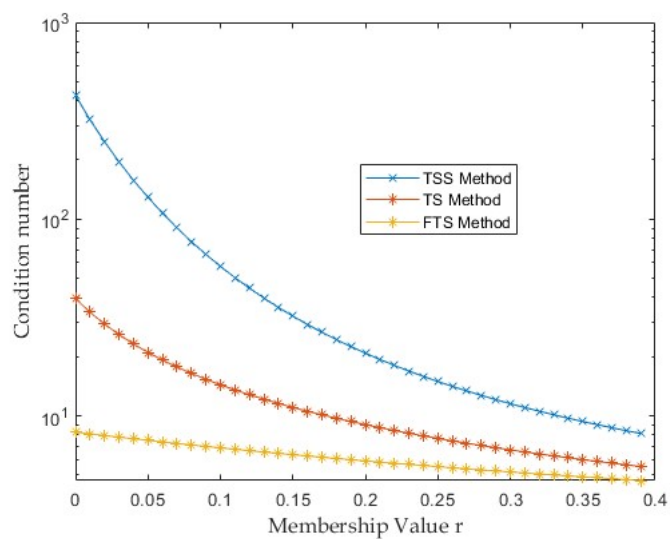


Figure 6: Condition number of TSS, TS, and FTS methods for the contraction factor  $\alpha = 0.6$  over the membership values  $r$ .

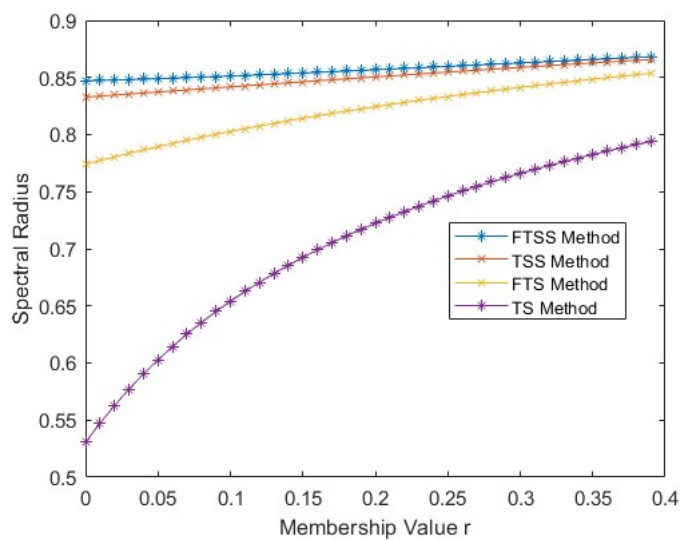


Figure 7: Spectral radius of TSS, FTSS, TS, FTS methods for the contraction factor  $\alpha = 0.6$  over the membership values  $r$ .

cient matrix as a fuzzy stochastic matrix and fuzzy right-hand side matrix. The iterative scheme was established, and the convergence theorems were presented. FTS iterative solutions with classical TS and TSS methods were compared. Numerical examples showed that the method is effective and efficient when compared with the classical iterative methods. The convergence and sensitivity analysis were discussed. The numerical value of spectral radius concluded that the solution of FTS method converges to unique nonzero solution. The numerical value of condition number gave the sensitivity analysis of regularized linear system and concluded that the iterative solution of regularized FLS is well conditioned.

**Disclosure statement:** No potential conflict of interest was reported by the author(s).

**Authors' contributions:** All authors' contribute equally in the preparation of this manuscript.

**Data availability statement:** No data were used to support this study.

## References

- [1] Abbasbandy, S., Ezzati, R. and Jafarian, A. *LU decomposition method for solving fuzzy system of linear equations*, Appl. Math. Comput. 172(1) (2006), 633–43.
- [2] Abbasbandy, S. and Jafarian, A. *Steepest descent method for system of fuzzy linear equations*, Appl. Math. Comput. 175(1) (2006), 823–33.
- [3] Allahviranloo, T. *Numerical methods for fuzzy system of linear equations*, Appl. Math. Comput. 155(2)(2004), 493–502.
- [4] Allahviranloo, T. *The Adomian decomposition method for fuzzy system of linear equations*, Appl. Math. Comput. 163(2)(2005), 553–563.
- [5] Chakraverty, S. and Behera, D. *Fuzzy system of linear equations with crisp coefficients*, J. Intell. Fuzzy Syst. 25(1) (2013), 201–207.
- [6] Dehghan, M. and Hashemi, B. *Iterative solution of fuzzy linear systems*, Appl. Math. Comput. 175(1) (2006), 645–74.

- [7] Dehghan, M., Hashemi, B. and Ghatee, M. *Solution of the Fully Fuzzy Linear Systems Using Iterative Techniques*, Chaos Solitons Fractals. 34(2)(2007), 316–36.
- [8] Edalatpanah, S.A. *Modified iterative methods for solving fully fuzzy linear systems*, Fuzzy Syst. 3 (2017), 55–73.
- [9] Ezzati, R. *Solving fuzzy linear systems*, Soft. Comput. 15(1)(2011), 193–197.
- [10] Fariborzi Araghi, M.A. and Fallahzadeh, A. *Inherited LU factorization for solving fuzzy system of linear equations*, Soft. Comput. 17(1) (2014), 159–163.
- [11] Friedman, M., Ming, M. and Kandel A. *Fuzzy linear systems*, Fuzzy Sets Syst. 96(2) (1998), 201–209.
- [12] Huang, Z., Chen, Z., Zhang, S., Wang, S. and Wang, K. *A new method based on Jacobi iteration for fuzzy linear systems*, Thai J. Math. 21(1) (2023), 29–37.
- [13] Koam, A.N.A., Akram, M., Muhammad, G. and Hussain, N. *LU decomposition scheme for solving m-polar fuzzy system of linear equations*, Math. Probl. Eng. 20(1)(2020), 1–19.
- [14] Miao, S. X., Zheng, B. and Wang, K. *Block SOR methods for fuzzy linear systems*, Jour. Appl. Math. Comput., 26(1)(2008), 201–18.
- [15] Nasseri, S. H., Matinfar, M. and Sohrabi, M. *QR-decomposition method for solving fuzzy system of linear equations*, Int. J. Math. Comput. 4(1)(2009), 129–136.
- [16] Rajaiah, D., Rajkumar, L.P. and Malla Reddy, P. *Steady state probability vector of positive definite regularized linear systems of circulant stochastic matrices*, Linear MultiLinear Algebra, 65(1) (2016), 140–155.
- [17] Wang, K. and Wu, Y. *Uzawa-SOR method for fuzzy linear system*, Int. J. of Inf. and Com. Sci., 1(2)(2012), 30–3.

- [18] Wang, K. and Zheng, B. *Symmetric successive over relaxation methods for fuzzy linear systems*, J. Appl. Math. Comput. 175(2) (2006), 891–901.
- [19] Wang, K. and Zheng, B. *Block iterative methods for fuzzy linear systems*, J. Appl. Math. Comput. 25(1) (2006), 119–36.
- [20] Wang, Y.R. and Chen, Y.L. *A modified Jacobi iterative method for large-size linear systems*, J. Shandong Univ. (Natural Science) 55(6) (2020), 122–6.
- [21] Yin, J.F. and Wang, K. *Splitting iterative methods for fuzzy system of linear equations*, Comput. Math. Model. 20(2)(2009), 326–35.



# Mathematical modeling of COVID-19 spread with media coverage and optimal control analysis

G.P. Sahu<sup></sup> and A.S. Thakur\*,<sup></sup>

## Abstract

The COVID-19 pandemic, initiated by the SARS-CoV-2 virus, first emerged in Wuhan, China and quickly propagated worldwide. In India, lacking immediate access to effective vaccines and antiviral drugs, the response primarily relied on nonpharmaceutical interventions. These strategies, extensively covered by the media, were vital in promoting preventive behaviors to limit viral transmission. This research introduces a new mathematical model,  $SAEI_aIRUM$ , to analyze COVID-19's transmission dynamics. It includes a saturation functional response to depict the media's role in influencing public behavior. The control reproduction number ( $R_c$ ) is determined, and both local and global stability of the disease-free equilibrium

---

\*Corresponding author

Received 12 March 2025; revised 30 April 2025; accepted 29 May 2025

Govind Prasad Sahu

Center for Basic Sciences, Pt. Ravishankar Shukla University, Raipur, Chhattisgarh.

e-mail: govind3012@gmail.com

Amit Singh Thakur

School of Studies in Mathematics, Pt. Ravishankar Shukla University, Raipur, Chhattisgarh.

e-mail: amitsinghprsu@gmail.com

## How to cite this article

Sahu, G.P. and Thakur, A.S., Mathematical modeling of COVID-19 spread with media coverage and optimal control analysis. *Iran. J. Numer. Anal. Optim.*, 2025; 15(3): 952-992. <https://doi.org/10.22067/ijnao.2025.92605.1612>

are analyzed. Using the least-squares method, the model fits daily case data from India from March 30, 2020, to January 24, 2021. We evaluate the impact of various control parameters on disease progression through numerical simulations and employ normalized forward sensitivity analysis to identify critical parameters affecting  $R_c$ . The study advances by formulating an optimal control problem, incorporating the cost of preventive actions as control variables. Findings indicate that an early optimal control strategy could lessen the severity of epidemic peaks by distributing their effects over a longer duration. Simulations demonstrate that combining four control measures outperforms a single or no control.

**AMS subject classifications (2020):** Primary 92D30; Secondary 92C60.

**Keywords:** COVID-19; Environmental transmission; Saturated awareness; Optimal control.

## 1 Introduction

Coronaviruses are a group of single-stranded, positive-sense RNA viruses classified under the Coronaviridae family [14]. They were first classified in 1960, with the name “corona” inspired by their distinctive crown-like structure observed under a microscope [29]. Over time, these viruses have been responsible for three significant outbreaks: The Severe Acute Respiratory Syndrome (SARS) outbreak in China (2003), the Middle East Respiratory Syndrome (MERS) outbreak in Saudi Arabia (2012) [30], and its resurgence in South Korea (2015) [60].

The World Health Organization (WHO) formally named the illness resulting from the novel coronavirus SARS-Cov-2 as Coronavirus Disease 2019 (COVID-19) [36, 17, 33]. Widely acknowledged as one of the most severe public health crises of the 21st century, the COVID-19 pandemic has exerted profound and widespread effects across the globe. Coronaviruses are RNA-based enveloped viruses known to infect both mammals and birds, frequently causing respiratory infections [51, 59]. Recognizable symptoms such as fever, dry cough, and fatigue became widely acknowledged early in the pandemic [26]. COVID-19 has exhibited a rapid transmission rate and high mortal-

ity, surpassing the severity of SARS and MERS. Beyond its physical impact, the pandemic has inflicted profound psychological distress, contributing to heightened anxiety, loneliness, and reduced resilience. The economic losses attributed to infectious diseases during this period are estimated to have exceeded those incurred in all historical wars [52]. Global tracking initiatives, including platforms like Worldometers and the Center for Systems Science and Engineering (CSSE) at Johns Hopkins University, have played a pivotal role in monitoring the virus's spread [27]. As COVID-19 swiftly spread worldwide, India experienced its impact as well. The nation confirmed its first locally transmitted case on January 30, 2020, in Kerala's Thrissur district, where a student returning from Wuhan University tested positive [22]. Concerns regarding insufficient testing rates raised alarms about potential widespread infections [9]. With a transmission rate of 1.7 [54], India's spread was comparatively lower than other global hotspots, though its estimated basic reproduction number ( $R_c$ ) ranged from 2 to 3.5 [4, 63]. The high viral loads in the upper respiratory tract of symptomatic and asymptomatic individuals facilitated silent transmission, akin to influenza [63].

During the early phases of a pandemic, when healthcare resources and biomedical interventions are insufficient, public education on preventive measures becomes the most effective strategy for controlling disease spread. Non-pharmaceutical interventions (NPIs) such as social distancing, mask mandates, and quarantine protocols have been widely disseminated through social media, television, radio, and the internet [49, 48]. Modern research underscores the power of media coverage as a behavioral influence mechanism, capable of shaping public responses without direct economic investment [21]. Several studies have explored the role of media in mitigating infectious disease outbreaks [40, 12, 2, 42, 58, 15, 44]. Misra, Sharma, and Shukla [40] analyzed a nonlinear SIS model demonstrating that media-driven awareness campaigns can significantly reduce transmission by encouraging individuals to self-isolate. Chang et al. [12] examined the impact of media coverage during the COVID-19 outbreak in Hubei, China, finding that reduced media attention delayed the infection peak but ultimately increased overall case numbers. Aldila [2] approached media and rapid testing interventions as an optimal control problem, demonstrating their effectiveness in minimizing in-

fections and economic disruptions in East Java, Indonesia. In India, Rai et al. [42] assessed the influence of social media advertisements on COVID-19 transmission. Wang et al. [58] and Chen, Li, and Zhang [15] conducted sensitivity analyses showing that intensified media campaigns can reduce the adequate reproduction number and curb infection rates. The findings emphasize the necessity of NPIs in reducing the basic reproduction number below one. Regular public health campaigns and digital outreach efforts are essential in encouraging symptomatic individuals to seek hospitalization and asymptomatic carriers to enter quarantine, thereby mitigating viral spread. Media coverage significantly influences human behavior, prompting adherence to precautionary measures such as lockdowns [19, 31], social distancing [13, 18], mask usage [7, 37, 8], quarantine enforcement [7, 34, 1], and hospitalization protocols [47, 25]. Studies have demonstrated the effectiveness of NPIs in controlling COVID-19 transmission. Sardar et al. [45] evaluated lockdown measures across Indian states, revealing a decline in virus transmission. Aldila et al. [3] employed mathematical modeling to assess the impact of social distancing and rapid assessments in Jakarta, Indonesia. Memon, Qureshi, and Memon [39] examined the efficacy of quarantine and isolation in mitigating outbreaks. Srivastav et al. [50] analyzed the effects of face masks, hospitalization, and asymptomatic quarantine on disease transmission in India, concluding that these strategies significantly reduce infection rates. Additionally, Wang and Ruan [57] introduced an epidemic model incorporating constant removal of infectives through treatment, revealing complex transmission dynamics. Yuan et al. [61] proposed an  $SEII_aHR$  model investigating the impact of asymptomatic carriers and isolation measures on global COVID-19 transmission.

While most studies focus on human-to-human transmission, SARS-CoV-2 also spreads through contaminated environments. Infected individuals introduce the virus into their surroundings via respiratory secretions from coughing or sneezing [23], and the virus can persist on surfaces for several days [56]. Several mathematical models have explored the role of environmental contamination in disease spread [53, 46, 5]. For instance, Sarkar, Mondal, and Khajanchi [46] assessed COVID-19 transmission in India, demonstrating that higher environmental contamination correlates with increased infection

rates. Asamoah et al. [5] conducted a similar study in Ghana, emphasizing the need for sanitation measures. These findings highlight the critical role of hygiene and disinfection practices in controlling the virus.

Motivated by the work of Asamoah et al. [5], this study develops a novel mathematical model to analyze COVID-19 transmission dynamics and evaluate intervention strategies in India. Rai et al. [42] illustrated how individuals adjust their behavior based on perceived susceptibility to infection. This research incorporates a saturation-type incidence function [41] to account for adaptive responses such as mask-wearing, hand hygiene, and social distancing. Using dynamical systems theory, numerical simulations, and sensitivity analyses, the study provides critical insights into the effectiveness of various control strategies. Model parameters are estimated using data from India collected between March 2020 and January 2021 [27]. Optimal control theory offers a robust mathematical framework for identifying the most effective strategies to manage infectious disease outbreaks. This approach has been widely utilized to design public health policies aimed at minimizing transmission [7, 5, 32, 62, 16, 20]. The present study applies optimal control theory to refine and evaluate the proposed model, ensuring its practical applicability in pandemic management.

The article is structured as follows: Section 2 presents the mathematical model describing COVID-19 transmission dynamics. In Section 3, the model's well-posedness is demonstrated, followed by an analysis of equilibrium points, stability assessment, and the control reproduction number evaluation. Section 4 is dedicated to numerical investigations, including parameter estimation based on empirical data, normalized sensitivity analysis, and simulations that examine the effects of NPIs and environmental contamination on the spread of the disease. Section 5 addresses the formulation and solution of the optimal control problem. Lastly, Section 6 concludes the study and outlines possible directions for future research.

## 2 Mathematical model

In this section, we introduce a novel mathematical model for COVID-19. The human population is categorized into eight groups: Susceptible indi-

viduals ( $S$ ), exposed individuals ( $E$ ), symptomatic infected individuals ( $I$ ), asymptomatic infected individuals ( $I_a$ ), aware individuals ( $A$ ), and recovered individuals ( $R$ ). Additionally,  $U$  represents the density of the coronavirus in the environment, while  $M$  denotes the media coverage of COVID-19, encompassing social media, print, electronic media, radio, and similar platforms. The model is based on the following assumptions:

1. The population's composition stays unchanged, as new individuals enter the region at a constant rate  $\Lambda$  and are immediately added to the susceptible group upon arrival.
2. Disease transmission occurs when a susceptible individual comes into contact with an infected individual, transitioning into the exposed category at a rate represented by  $\beta$ .
3.  $1/\sigma$  represents the latent period. The exposed individuals who do not exhibit clinical symptoms join the asymptomatic infected class at a rate of  $(1 - \sigma)k_2$ . In contrast, those who exhibit clinical symptoms join the symptomatic infected class at a rate of  $\sigma k_1$ .
4. During an endemic outbreak, health authorities and media outlets leverage social media platforms such as Facebook, Twitter, and WhatsApp to share information with the public. The spread of information is influenced by both the frequency of its dissemination and the severity of the outbreak it pertains to. This suggests that the pace at which information campaigns grow is closely tied to the scale of the affected population [40].
5. Media coverage has a limited impact on how the disease spreads among susceptible individuals. As a result, the rate at which susceptible people become aware is modeled by  $\frac{\lambda MS}{c+M}$ , where  $c$  represents the half-saturation constant indicating the media exposure level at which awareness reaches half its maximum effect. When the level of media coverage reaches  $c$ , it reaches half its maximum value  $\lambda S$  as in [41]. Even those in the aware class can lose their awareness and return to the susceptible class at the rate  $\lambda_0$ . Additionally, the region consistently receives a minimum level of media attention.

6. A proportion  $\alpha$  of people in public places consistently and correctly wear face masks. When face masks are worn properly, they reduce the spread of disease [28].
7. Individuals showing symptoms move into the recovered class at the rate  $\gamma_3$ . Over time, immunity in the recovered group wanes, causing them to re-enter the susceptible group at a rate of  $\xi$ . Moreover, symptomatic infected individuals die due to the disease at a rate given by  $\delta$ .
8. In the model, a fixed proportion  $\phi$  of newly asymptomatic infections is assumed to progress to the symptomatic class, whereas the complementary fraction  $1 - \phi$  recovers without ever developing symptoms. Individuals in the  $\phi$ -branch leave the asymptomatic compartment  $I_a$  at rate  $\gamma_1$  (mean waiting time  $1/\gamma_1$ ) and join the symptomatic class  $I$ ; strictly asymptomatic cases exit  $I_a$  at rate  $\gamma_2$  (mean infectious period  $1/\gamma_2$ ) and enter the recovered class  $R$  [5].  
Asymptomatic individuals who do not display symptoms join the recovered class at the rate  $\gamma_2$ .
9. A person infected with COVID-19 releases the virus into the surroundings by sneezing or coughing. The emission rates of the virus by asymptomatic and symptomatic individuals are represented by  $\theta_1$  and  $\theta_2$ , respectively. However, the virus does not persist indefinitely in the environment; it is gradually removed through natural decay and human efforts such as cleaning and disinfection. The rate at which the virus is eliminated from the surroundings is denoted by  $\epsilon$ .
10. The rate of information growth, symbolized as  $r_1$ , is assumed to be directly linked to the number of infected individuals. The growth rate decreases by a factor of

$$f(A) = \frac{r_1 \theta A}{\omega + A}.$$

Consequently, the net growth rate of TV and social media advertisements aimed at raising awareness within the population is expressed as

$$r_1 \left( 1 - \frac{\theta A}{\omega + A} \right).$$

Here,  $\theta$  represents the rate of decline in advertisement effectiveness as the number of aware individuals increases. The parameter  $\omega$  denotes the half-saturation point, where  $f(A)$  achieves half of its maximum value  $r_1\theta$ , occurring when the aware population reaches  $\omega$ . For the model to remain valid, the value of  $\theta$  must lie between 0 and 1. Meanwhile, the rate of information decay, indicated by  $r_0$ , quantifies how quickly memories of the information naturally diminish over time.

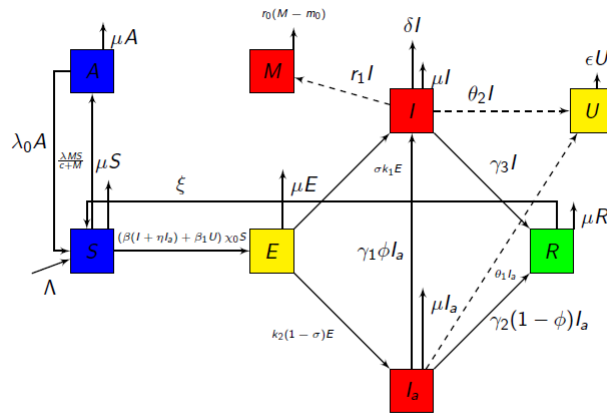


Figure 1: Flow diagram of model (1)

The spread of COVID-19, based on the stated assumptions, can be modeled using the following system of nonlinear ordinary differential equations:

$$\left\{ \begin{array}{l} \frac{dS}{dt} = \Lambda - \beta(1 - \alpha)(I + \eta I_a)S - \beta_1(1 - \alpha)SU + \lambda_0 A + \xi R - \frac{\lambda MS}{c+M} - \mu S, \\ \frac{dA}{dt} = \frac{\lambda MS}{c+M} - \lambda_0 A - \mu A, \\ \frac{dE}{dt} = \beta(1 - \alpha)(I + \eta I_a)S + \beta_1(1 - \alpha)SU - k_2(1 - \sigma)E - k_1 \sigma E - \mu E, \\ \frac{dI_a}{dt} = k_2(1 - \sigma)E - \gamma_1 \phi I_a - \gamma_2(1 - \phi)I_a - \mu I_a, \\ \frac{dI}{dt} = k_1 \sigma E + \gamma_1 \phi I_a - \gamma_3 I - \delta I - \mu I, \\ \frac{dR}{dt} = \gamma_2(1 - \phi)I_a + \gamma_3 I - \xi R - \mu R, \\ \frac{dU}{dt} = \theta_1 I_a + \theta_2 I - \epsilon U, \\ \frac{dM}{dt} = r_1(1 - \frac{\theta A}{\omega + A})I - r_0(M - m_0). \end{array} \right. \quad (1)$$

A flow diagram of the model (1) is depicted in Figure 1. The initial circumstances of model (1) are regarded as follows:

$$S(0) = S_0 > 0, \quad E(0) \geq 0, \quad I_a(0) \geq 0, \quad I(0) \geq 0, \quad A(0) = A_0 \geq 0, \quad R(0) \geq 0, \quad U(0) \geq 0, \quad M(0) \geq m_0.$$

Let us use the notations:  $\chi_0 = (1 - \alpha)$ ,  $\chi_1 = k_2(1 - \sigma) + k_1\sigma + \mu$ ,  $\chi_2 = \gamma_1\phi + \gamma_2(1 - \phi) + \mu$ ,  $\chi_3 = \gamma_3 + \delta + \mu$ ,  $\chi_4 = \lambda_0 + \mu$ ,  $\chi_5 = \xi + \mu$ .

Consequently, the above model (1) can be expressed as

$$\left\{ \begin{array}{l} \frac{dS}{dt} = \Lambda - \beta\chi_0(I + \eta I_a)S - \beta_1\chi_0SU + \lambda_0A + \xi R - \frac{\lambda MS}{c+M} - \mu S, \\ \frac{dA}{dt} = \frac{\lambda MS}{c+M} - \chi_4A, \\ \frac{dE}{dt} = \beta\chi_0(I + \eta I_a)S + \beta_1\chi_0SU - \chi_1E, \\ \frac{dI_a}{dt} = k_2(1 - \sigma)E - \chi_2I_a, \\ \frac{dI}{dt} = k_1\sigma E + \gamma_1\phi I_a - \chi_3I, \\ \frac{dR}{dt} = \gamma_2(1 - \phi)I_a + \gamma_3I - \chi_5R, \\ \frac{dU}{dt} = \theta_1I_a + \theta_2I - \epsilon U, \\ \frac{dM}{dt} = r_1(1 - \frac{\theta A}{\omega+A})I - r_0(M - m_0). \end{array} \right. \quad (2)$$

A description of all parameters is given in Table 1.

### 3 Mathematical analysis

In this section, we provide essential analytical findings for model (1), demonstrating that its solutions remain positive and bounded. We also identify disease-free and endemic equilibrium states and examine their stability. In addition, we derive a theoretical expression for the critical biological parameter known as the control reproduction number.

Table 1: Parameters and their interpretation for models (1)

Parameters	Description
$\eta$	Modification Parameter for asymptomatic class
$\beta$	Human-to-human transmission rate
$\beta_1$	Maximum transmission rate due to environmental contamination
$k_1$	Progression rate from exposed to the symptomatic class
$k_2$	Progression rate from exposed to the asymptomatic class
$\gamma_3$	Recovery rate for symptomatic infected individuals
$\lambda$	The rate at which awareness spreads among susceptible individuals
$c, \omega$	Half saturation constants
$\gamma_1$	Rate of transition from asymptomatic to symptomatic class
$\gamma_2$	Rate of transition from asymptomatic to recovered class
$\delta$	Disease-induced mortality rate for symptomatic individuals
$\phi$	The fraction of asymptomatic patients who subsequently develop symptoms and transition to the symptomatic infected class
$\Lambda$	Recruitment rate
$r_1$	Development rate of information dissemination
$r_0$	Reduction rate of advertisements due to inefficacy and psychological barriers
$\mu$	Natural death rate
$m_0$	Baseline number of media coverage
$\lambda_0$	Rate of transition of aware individuals to the susceptible class
$\alpha$	The proportion of the population wearing face masks
$\theta_1$	The speed at which asymptomatic individuals emit the virus into their surroundings
$\theta_2$	The rate at which symptomatic individuals emit the virus into their surroundings
$\epsilon$	Natural decay rate of virus from the environment
$\theta$	Represent the decline in the effectiveness of advertisements as the number of individuals who are already aware increases
$1/\xi$	Average time it takes for immunity to wear off
$\sigma$	Fraction of exposed individuals joint to $I$ class

### 3.1 Positivity and boundedness

In this section, we demonstrate that the solutions of the model (1) are positive and bounded.

$$\begin{aligned} \left. \frac{dS}{dt} \right|_{S=0} &= \Lambda + \lambda_0 A + \xi R > 0, & \left. \frac{dA}{dt} \right|_{A=0} &= \frac{\lambda MS}{c + M} \geq 0, \\ \left. \frac{dI_a}{dt} \right|_{I_a=0} &= k_2(1 - \sigma)E \geq 0, & \left. \frac{dI}{dt} \right|_{I=0} &= k_1\sigma E + \gamma_1\phi I_a \geq 0, \\ \left. \frac{dE}{dt} \right|_{E=0} &= \beta\chi_0(I + \eta I_a)S + \beta_1\chi_0 SU \geq 0, & \left. \frac{dR}{dt} \right|_{R=0} &= \gamma_3 I + \gamma_2(1 - \phi)I_a > 0, \\ \left. \frac{dU}{dt} \right|_{U=0} &= \theta_2 I + \theta_1 I_a \geq 0, & \left. \frac{dM}{dt} \right|_{M=0} &= r_1 I + r_0 m_0 > 0. \end{aligned}$$

As all rates are nonnegative in this scenario, any solution starting within the interior of the nonnegative bounding cone  $\mathbb{R}_+^8$  will remain confined to this cone, as the vector field consistently points inward along all bounding planes. As a result, it is assured that none of the model (1) solutions are negative. We add equations of the model (1) to demonstrate the boundedness of its solutions, which results in  $\frac{dN}{dt} = \Lambda - \mu N - \delta I$ . Then,  $\frac{dN}{dt} < \Lambda - \mu N$ , Applying Birkhoff's and Rota's theorems on differential inequality [10], as  $t \rightarrow \infty$ , we have  $0 \leq N(t) \leq \frac{\Lambda}{\mu} (= N_0)$ . As  $I(t), I_a(t) \leq N(t)$  at any time, so  $I(t), I_a(t) \leq \frac{\Lambda}{\mu}$ . Now, from the density of the virus in the environment,  $\frac{dU}{dt} = \theta_2 I + \theta_1 I_a - \epsilon U \leq (\theta_2 + \theta_1) \frac{\Lambda}{\mu} - \epsilon U$ . For the initial conditions, when applying the Gronwall inequality, we get  $0 \leq U(t) \leq \frac{(\theta_2 + \theta_1)\Lambda}{\epsilon\mu}$ . Additionally, from the eighth equation of model (1), we obtain  $\frac{dM}{dt} + r_0 M \leq r_0 m_0 + \frac{r_1 \Lambda}{\mu}$ . Using differential inequality theory, we can derive

$$\limsup M(t) \leq m_0 + \frac{r_1 \Lambda}{r_0 \mu},$$

when applying the Gronwall inequality  $0 \leq M(t) \leq m_0 + \frac{r_1 \Lambda}{r_0 \mu}$ . Hence, the feasible region for the model (1) is

$$\Theta = \left\{ (S, A, E, I_a, I, R, U, M) \in \mathbb{R}_+^8 : \right. \\ \left. 0 \leq S, A, E, I_a, I, R, N \leq \frac{\Lambda}{\mu}; \right.$$

$$0 \leq U \leq \frac{(\theta_2 + \theta_1) \Lambda}{\epsilon \mu}; 0 < M \leq m_0 + \frac{r_1 \Lambda}{r_0 \mu} \Bigg\}. \quad (3)$$

Therefore,  $\Theta$  defines the region enclosing the model's solutions.

**Theorem 1.** Solutions of the model under the given initial conditions continue to be nonnegative as time goes on. Furthermore, the closed region remains unchanged and preserved under the dynamics of the model (1).

### 3.2 Disease-free equilibrium (DFE) and basic reproduction number

The disease-free equilibrium (DFE) of the model (1) is given by

$$\zeta_0 = (S_0, A_0, 0, 0, 0, 0, 0, m_0), \text{ where } S_0 = \frac{(c+m_0)\Lambda(\lambda_0+\mu)}{\mu(m_0\lambda+c\lambda_0+m_0\lambda_0+c\mu+m_0\mu)} \text{ and } A_0 = \frac{m_0\Lambda\lambda}{\mu(m_0\lambda+c\lambda_0+m_0\lambda_0+c\mu+m_0\mu)}.$$

The control reproduction number for model (2) is determined using the following generation matrix approach [55]. By defining the state vector as  $x = (E, I_a, I, U)$ , the system in model (2) can be reformulated as  $\frac{dx}{dt} = \mathcal{F}(x) - \mathcal{V}(x)$ , where  $\mathcal{F}$  represents the nonnegative matrix of new infection terms, and the matrix  $\mathcal{V}$  of the remaining terms are given by

$$\mathcal{F} = \begin{bmatrix} \beta\chi_0(I + \eta I_a)S + \beta_1\chi_0SU \\ 0 \\ 0 \\ 0 \end{bmatrix}, \mathcal{V} = \begin{bmatrix} \chi_1 E \\ -k_2(1-\sigma)E + \chi_2 I_a \\ -k_1\sigma E - \gamma_1\phi I_a + \chi_3 I \\ -\theta_1 I_a - \theta_2 I + \epsilon U \end{bmatrix}.$$

The corresponding linearized matrices evaluated at the DFE  $\zeta_0$  are

$$F_1 = \begin{bmatrix} 0 & S_0\chi_0\beta\eta & S_0\chi_0\beta & S_0\chi_0\beta_1 \\ 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 \end{bmatrix}, V_1 = \begin{bmatrix} \chi_1 & 0 & 0 & 0 \\ -k_2(1-\sigma) & \chi_2 & 0 & 0 \\ -k_1\sigma & -\gamma_1\phi & \chi_3 & 0 \\ 0 & -\theta_1 & -\theta_2 & \epsilon \end{bmatrix}.$$

We get control reproduction number  $R_c = \rho(F_1 V_1^{-1})$ , where  $\rho$  is the spectral radius.

$$R_c = \frac{S_0\chi_0(k_1(\beta\epsilon+\beta_1\theta_2)\phi\chi_2+k_2(1-\phi)(\beta\epsilon(\eta\chi_3+\gamma_1\phi)+\beta_1(\theta_1\chi_3+\gamma_1\theta_2\phi)))}{\epsilon\chi_1\chi_2\chi_3}.$$

In the absence of any intervention and behavioral awareness,

$\alpha = 0, \lambda = 0, \lambda_0 = 0, m_0 = 0$ , implies  $A_0 = 0, S_0 = \frac{A}{\mu} = N_0$ . Hence, the basic reproduction number  $R_0$  of the model (2) is given by

$$R_0 = \frac{N_0 \left[ k_1 (\beta \epsilon + \beta_1 \theta_2) \phi \chi_2 + k_2 (1 - \phi) (\beta \epsilon (\eta \chi_3 + \gamma_1 \phi) + \beta_1 (\theta_1 \chi_3 + \gamma_1 \theta_2 \phi)) \right]}{\epsilon \chi_1 \chi_2 \chi_3}.$$

It is clear that  $R_c = \frac{S_0 \chi_0 R_0}{N_0}$ . Since  $\frac{S_0}{N_0} \leq 1$  and  $0 \leq \chi_0 \leq 1$ , this implies that  $R_c \leq R_0$ .

**Theorem 2.** The equilibrium state  $\zeta_0$ , representing the absence of disease in the model (2), remains locally asymptotically stable provided that the control reproduction number  $R_c$  is less than one. Conversely, it loses stability when  $R_c$  exceeds one.

### 3.3 Global stability

This section examines the global stability of the disease-free steady state for a specific case.

**Theorem 3.** The DFE  $\zeta_0$  in model (2) is globally asymptotically stable when  $R_c \leq 1$ ; however, it becomes unstable if  $R_c > 1$ .

*Proof.* We construct a suitably defined Lyapunov function to establish the global stability of the DFE  $\zeta_0$ . Specifically, we consider a continuously differentiable, positive definite function  $\mathcal{G}$  such that

$$\mathcal{G} = d_1 E + d_2 I_a + d_3 I + d_4 U. \quad (4)$$

The constants  $d_j \geq 0$  for  $j = 1, 2, \dots, 4$  will be determined later. Furthermore, by utilizing the equations from model (1) and differentiating equation (4), we obtain

$$\begin{aligned} \mathcal{G}' &= d_1 E' + d_2 I_a' + d_3 I' + d_4 U' \\ &= d_1 (\beta \chi_0 (I + \eta I_a) S + \beta_1 \chi_0 S U - \chi_1 E) + d_2 (k_2 (1 - \sigma) E - \chi_2 I_a) \\ &\quad + d_4 (\theta_1 I_a + \theta_2 I - \epsilon U) + d_3 (k_1 \sigma E + \gamma_1 \phi I_a - \chi_3 I) \\ &\leq d_1 (\beta \chi_0 (I + \eta I_a) S_0 + \beta_1 \chi_0 S_0 U - \chi_1 E) + d_2 (k_2 (1 - \sigma) E - \chi_2 I_a) \\ &\quad + d_4 (\theta_1 I_a + \theta_2 I - \epsilon U) + d_3 (k_1 \sigma E + \gamma_1 \phi I_a - \chi_3 I) \\ &= (-\chi_1 d_1 + d_2 k_2 (1 - \sigma) + d_3 k_1 \sigma) E + (d_1 \beta \eta \chi_0 S_0 - d_2 \chi_2 + d_3 \gamma_1 \phi + d_4 \theta_1) I_a \end{aligned}$$

$$+ (d_1\beta\chi_0S_0 - d_3\chi_3 + d_4\theta_2)I + (d_1\chi_0\beta_1S_0 - d_4\epsilon)U. \quad (5)$$

Let us choose the constant values as follows:  $d_4 = 1, d_3 = \frac{(\beta\epsilon + \beta_1\theta_2)}{\beta_1\chi_3}, d_1 = \frac{\epsilon}{\chi_0S_0\beta_1}, d_2 = \frac{\beta\epsilon\eta\chi_3 + \beta_1\theta_1\chi_3 + \beta\gamma_1\epsilon\phi + \beta_1\gamma_1\theta_2\phi}{\beta_1\chi_2\chi_3}$ , using the aforementioned inequality (5), we obtain the following:

$$\mathcal{G}' \leq \frac{(R_c - 1)\chi_1\epsilon E}{S_0\beta_1\chi_0}. \quad (6)$$

Clearly,  $\mathcal{G}' \leq 0$  whenever  $R_c \leq 1$ , and  $\mathcal{G}' = 0$  if and only if either  $E = 0$  or  $R_c = 1$ , at DFE. Therefore, by LaSalle's invariance principle [35], the equilibrium point  $\zeta_0$  is globally asymptotically stable.  $\square$

### 3.4 Presence and persistence of endemic equilibria

To determine the possible endemic equilibrium points of the proposed model, the system of nonlinear equations derived from the model (2) is solved by setting all derivatives to zero. The endemic equilibrium  $\zeta^* = \{S^*, A^*, E^*, I_a^*, I^*, R^*, U^*, M^*\}$ , of the model (2) is given by

$$E^* = b_1I_a^*, I^* = b_2I_a^*, A^* = \frac{b_5M^*\lambda}{(M^* + c)\chi_4}, R^* = b_3I_a^*, U^* = b_4I_a^*,$$

$$M^* = b_1I^* + m_0, S^* = b_5.$$

Putting the values of  $\{S^*, A^*, E^*, I^*, R^*, U^*\}$  in the first, second, and eighth equation of system of equations (1), we have

$$\Lambda + b_3\xi I_a^* + b_5 \left( -\mu - (b_4\beta_1 + \beta(b_2 + \eta))\chi_0 I_a^* + \frac{\lambda(\lambda_0 - \chi_4)M^*}{(M^* + c)\chi_4} \right) = 0, \quad (7)$$

$$m_0r_0 + b_2r_1I_a^* = \left( r_0 + \frac{b_2b_5r_1\theta\lambda I_a^*}{b_5\lambda M^* + (M^* + c)\chi_4\omega} \right) M^*, \quad (8)$$

where

$$b_1 = \frac{\chi_2}{k_2(1 - \phi)}, \quad b_2 = \frac{b_1k_1\phi + \gamma_1\phi}{\chi_3}, \quad b_3 = \frac{\gamma_2(1 - \phi) + b_2\gamma_3}{\chi_5},$$

$$b_4 = \frac{\theta_1 + b_2\theta_2}{\epsilon}, b_5 = \frac{b_1\chi_1}{\chi_0(b_4\beta_1 + \beta(b_2 + \eta))}.$$

Equations (7) and (8) represent two isoclines in  $I_a^*$  and  $M^*$ . Analyzing the behavior of these isoclines through mathematical methods is challenging. Let  $(I_a^*, M^*)$  denote the unique point where these isoclines intersect.

We have seen that at least one endemic equilibrium always exists. Additionally, we investigate the occurrence of transcritical bifurcation through the application of center manifold theory, as detailed in previous studies [11, 43]. To simplify the process, we modify the variables accordingly and employ a similar approach described in those references [11]:

$S = x_1 + S_0$ ,  $E = x_2$ ,  $I_a = x_3$ ,  $I = x_4$ ,  $A = x_5 + A_0$ ,  $R = x_6$ ,  $U = x_7$ ,  $M = x_8 + m_0$ . As a result, it is possible to rewrite model (1) compactly, as follows:

$$\left\{ \begin{array}{l} \frac{dx_1}{dt} = \Lambda - \chi_0\beta(x_4 + \eta x_3)(x_1 + S_0) - \beta_1\chi_0(x_1 + S_0)x_7 + \xi x_6 \\ \quad + \lambda_0(x_5 + A_0 - \frac{\lambda(x_8+m_0)(x_1+S_0)}{c+(x_8+m_0)} - \mu(x_1 + S_0), \\ \frac{dx_2}{dt} = \beta\chi_0(x_4 + \eta x_3)(x_1 + S_0) + \beta_1\chi_0(x_1 + S_0)x_7 - \chi_1x_2, \\ \frac{dx_3}{dt} = k_2(1 - \sigma)x_2 - \chi_2x_3, \\ \frac{dx_4}{dt} = k_1\sigma x_2 + \gamma_1\phi x_3 - \chi_3x_4, \\ \frac{dx_5}{dt} = \frac{\lambda(x_8+m_0)(x_1+S_0)}{c+(x_8+m_0)} - \chi_5(x_5 + A_0), \\ \frac{dx_6}{dt} = \gamma_2(1 - \phi)x_3 + \gamma_3x_4 - \chi_6x_6, \\ \frac{dx_7}{dt} = \theta_1x_3 + \theta_2x_4 - \epsilon x_7, \\ \frac{dx_{10}}{dt} = r_1 \left( 1 - \frac{\theta(x_5+A_0)}{\omega+(x_5+A_0)} \right) x_4 - r_0x_8. \end{array} \right. \quad (9)$$

The Jacobian matrix of model (9) at the corresponding DFE  $P^0$  is given by

$$J_{P^0} = \begin{bmatrix} -c_2 & 0 & -S_0\beta_c\eta\chi_0 & -S_0\beta_c\chi_0 & \lambda_0 & \xi & -S_0\beta_1\chi_0 & c_4 \\ 0 & -\chi_1 & s_0\beta_c\eta\chi_0 & S_0\beta_c\chi_0 & 0 & 0 & s_0\beta_1\chi_0 & 0 \\ 0 & k_2(1-\phi) & -\chi_2 & 0 & 0 & 0 & 0 & 0 \\ 0 & k_1\phi & \gamma_1\phi & -\chi_3 & 0 & 0 & 0 & 0 \\ c_1 & 0 & 0 & 0 & -\chi_4 & 0 & 0 & -c_4 \\ 0 & 0 & \gamma_2(1-\phi) & \gamma_3 & 0 & -\chi_5 & 0 & 0 \\ 0 & 0 & \theta_1 & \theta_2 & 0 & 0 & -\epsilon & 0 \\ 0 & 0 & 0 & c_3 & 0 & 0 & 0 & -r_0 \end{bmatrix},$$

where

$$c_1 = \frac{\lambda m_0}{c+m_0}, c_2 = \mu + c_1, c_3 = r_1 \left(1 - \frac{\theta A_0}{\omega + A_0}\right), c_4 = -\frac{c S_0 \lambda}{(c+m_0)^2}, c_5 = -c_4.$$

At  $R_c = 1$ , the bifurcation parameter  $\beta$  gives a critical  $\beta_c$  as

$$\beta_c = \frac{-k_1 S_0 \beta_1 \theta_2 \phi \chi_0 \chi_2 + \epsilon \chi_1 \chi_2 \chi_3 - k_2 S_0 \beta_1 (1-\phi) \chi_0 (\theta_1 \chi_3 + \gamma_1 \theta_2 \phi)}{S_0 \epsilon \chi_0 (k_1 \phi \chi_2 + k_2 (1-\phi) (\eta \chi_3 + \gamma_1 \phi))}.$$

Confirming that the Jacobian evaluated at  $\beta = \beta_c$  possesses a right eigenvector associated with the zero eigenvalue is

$$\mathbf{W} = (w_1, w_2, w_3, w_4, w_5, w_6, w_7, w_8)^T, \text{ where } w_2 = 1, \quad w_3 = a_1, w_4 = a_2,$$

$$w_7 = a_4, \quad w_6 = a_5, \quad w_8 = a_3,$$

$$w_1 = -\frac{-a_3 c_4 \lambda_0 + a_3 c_4 \chi_4 + a_5 \xi \chi_4 - a_4 s_0 \beta_1 \chi_0 \chi_4 - a_2 s_0 \beta_c \chi_0 \chi_4 - a_1 s_0 \beta_c \eta \chi_0 \chi_4}{c_1 \lambda_0 - c_2 \chi_4},$$

$$w_5 = -\frac{-a_3 c_1 c_4 + a_3 c_2 c_4 - a_5 c_1 \xi + a_4 c_1 s_0 \beta_1 \chi_0 + a_2 c_1 s_0 \beta_c \chi_0 + a_1 c_1 s_0 \beta_c \eta \chi_0}{-c_1 \lambda_0 + c_2 \chi_4},$$

$$\text{where } a_1 = \frac{k_2(1-\phi)}{\chi_2}, a_2 = \frac{k_1\phi + a_1\gamma_1\phi}{\chi_3}, \quad a_3 = \frac{a_2 c_3}{r_0}, \quad a_4 = \frac{a_1 \theta_1 + a_3 \theta_2}{\epsilon}, \quad a_5 = \frac{a_1 \gamma_2 (1-\phi) + a_3 \gamma_3}{\chi_5}.$$

The elements of the left eigenvector, which correspond to the zero eigenvalues, are also  $\mathbf{V} = (v_1, v_2, v_3, v_4, v_5, v_6, v_7, v_8)$  and must satisfy the equalities  $V.J = 0$  and  $V.W = 1$ , so that we obtain

$$v_1 = 0, \quad v_5 = 0, \quad v_6 = 0, \quad v_8 = 0, \quad v_3 = \frac{k_1 S_0 v_2 \beta_c \epsilon \phi \chi_0 + k_1 S_0 v_2 \beta_1 \theta_2 \phi \chi_0 - v_2 \epsilon \chi_1 \chi_3}{k_2 \epsilon (-1+\phi) \chi_3},$$

$$v_4 = \frac{S_0 v_2 \beta_c \epsilon \chi_0 + S_0 v_2 \beta_1 \theta_2 \chi_0}{\epsilon \chi_3}, \quad v_7 = \frac{s_0 v_2 \beta_1 \chi_0}{\epsilon},$$

$$v_2 = \frac{k_2 \epsilon (-1+\phi) \chi_3}{a_2 k_2 S_0 (\beta_c \epsilon + \beta_1 \theta_2) (-1+\phi) \chi_0 + a_1 k_1 S_0 (\beta_c \epsilon + \beta_1 \theta_2) \phi \chi_0 + k_2 (-1+\phi) (\epsilon + a_4 S_0 \beta_1 \chi_0) \chi_3 - a_1 \epsilon \chi_1 \chi_3}.$$

As outlined in [11, Theorem 4.1], the bifurcation coefficients  $a$  and  $b$  can be determined using the following expressions, where  $f_k$  represents the  $k^{\text{th}}$  component of the vector function  $f$ :

$$\begin{aligned}
a &= \sum_{k,i,j=1}^{10} v_k w_i w_j \frac{\partial^2 f_k}{\partial x_i \partial x_j}(0,0) \\
&= v_2 (2w_1 w_7 \beta_1 \chi_0 + 2w_1 w_4 \beta_c \chi_0 + 2w_1 w_3 \beta_c \eta \chi_0), \\
b &= \sum_{k,i,j=1}^{10} v_k w_i \frac{\partial^2 f_k}{\partial x_i \partial \beta}(0,0) \\
&= S_0 w_4 \chi_0 + S_0 w_3 \eta \chi_0.
\end{aligned} \tag{10}$$

If  $a < 0$  and  $b > 0$  at  $\beta = \beta_c$ , then according to [11, Theorem 4.1 and Remark 1], a transcritical bifurcation occurs at  $R_c = 1$ . Moreover, when  $R_c > 1$ , the unique endemic equilibrium remains locally asymptotically stable.

## 4 Numerical simulation

### 4.1 Parameter estimation

In this section, the proposed model is calibrated against observed data to evaluate its accuracy and predictive capabilities, offering valuable insights into the pandemic's progression and supporting effective response strategies. Initially, baseline values for the model parameters are established using COVID-19 data, relevant information, and published literature. Specifically, data on the total number of COVID-19 cases in India from March 30, 2020, to January 24, 2021, were considered [27]. The least squares method is used to align the observed data points,  $Y_i$ , with the estimated values,  $X_i$ , by minimizing the total squared differences between the observed values and the predicted curve [38]. This process involves minimizing the *sum of squared errors* (*SSE*), expressed as

$$SSE = \sum_{i=1}^n (Y_i - X_i)^2$$

Table 2 and Figure 3 present the fitted model developed using MATLAB, along with the estimated parameter values. In Figure 3, the curve represents the fitted model, while the star points indicate the total number of daily confirmed cases. The estimated reproduction number ( $R_c$ ) is 1.94, suggesting

a moderate transmission rate. The model simulation closely follows the actual data, demonstrating its reliability. Table 2 provides a detailed summary of the estimated and fitted parameters. Following this parameter estimation, we explore hypothetical scenarios where individuals neither wear masks nor are aware of COVID-19. A more in-depth discussion follows below.

Figure 2a illustrates the model's evaluation of India's response to COVID-19. Many disregarded safety measures during the prolonged lockdown and economic crisis, resulting in  $\alpha = 0$  in Figure 2a. However, public awareness of COVID-19 remained high, leading to precautionary behaviors such as wearing face masks and self-quarantining after traveling from high-risk or red-alert zones. Figure 2b provides an alternative perspective, showing that even in cases where individuals were unaware of COVID-19's severity, many still adhered to protective measures like mask-wearing (i.e.,  $A = 0$  in 2b within 1). As illustrated in Figure 3, the model's predictions align closely with the observed data on daily new cases, reinforcing its applicability in understanding the pandemic's progression.

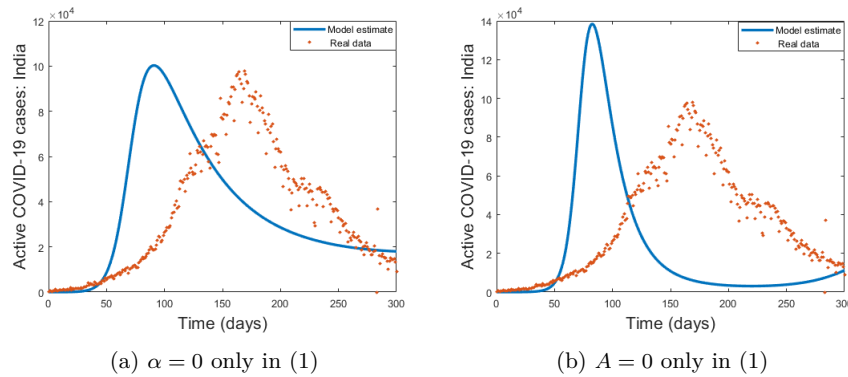


Figure 2: Fitted curve of confirmed cases in India and proposed model

## 4.2 Sensitivity analysis

Examining the sensitivity of parameters in the control reproduction number,  $R_c$ , is essential for understanding the dynamics of infectious diseases. This

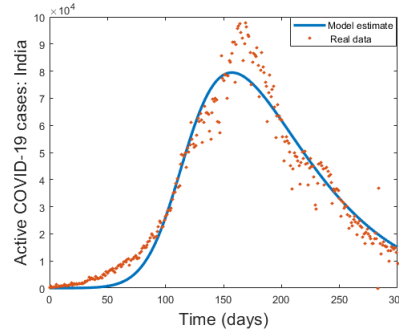


Figure 3: Fitted curve of confirmed case in India and model (1)

process enables the rapid identification of critical factors that drive disease transmission, which is pivotal for designing effective interventions. Modifying these parameters allows for more efficient pandemic management. The normalized forward sensitivity index of  $R_c$  with respect to a parameter  $p$  is defined as  $\Gamma_{R_c}^p = \frac{\partial R_c}{\partial p} \frac{p}{R_c}$  [38, 24].

The sensitivity indices of  $R_c$ , derived using parameter values from Table 2, are presented in Table 3. It shows that  $R_c$  increases with increase in the values of  $\Lambda, \beta, \beta_1, \sigma, \lambda_0, c, \theta_1, \theta_2, \eta, k_1, \gamma_1$ , and  $\phi$ . Conversely, parameters  $m_0, \lambda, \epsilon, \delta, k_2, \gamma_3, \gamma_2, \alpha$ , and  $\mu$ , have negative impact on  $R_c$ . Figure 4 indicates that  $\Lambda$  ( $\lambda$ ) has the maximum positive (negative) impact on  $R_c$ . Lower  $R_c$  values are preferred for disease control. Reducing  $R_c$  to control disease transmission requires increasing control parameters with negative indices and decreasing those with positive indices. Furthermore, it can be seen that  $R_c$  is not affected by the model parameters  $\xi, r_1, r_0, \omega, \theta$ , that is,

$$\Gamma_{R_c}^{\xi} = \Gamma_{R_c}^{r_1} = \Gamma_{R_c}^{r_0} = \Gamma_{R_c}^{\omega} = \Gamma_{R_c}^{\theta} = 0.$$

Figure 5a displays a two-dimensional contour plot, while Figure 5b displays a three-dimensional contour plot of  $R_c(\lambda, \lambda_0)$ . Figure 5 demonstrates that as awareness rates increase over time, there is a substantial reduction in the incidence of COVID-19.

Table 2: Parameter values for the model (1)

Parameters	Range	Baseline	Source
$\eta$	$(0.6281, 0.6366) \text{ day}^{-1}$	0.6364	[5]
$\beta$	$(6.038 \cdot 10^{-8}, 6.988 \cdot 10^{-8}) \text{ day}^{-1}$	$6.933 \cdot 10^{-8}$	Estimated
$\beta_1$	$(3.00199 \cdot 10^{-8}, 4.10199 \cdot 10^{-8}) \text{ day}^{-1}$	$4.00199 \cdot 10^{-8}$	[46]
$k_1$	$(0.0623, 0.0745) \text{ day}^{-1}$	0.0723	Estimated
$k_2$	$(0.066, 0.08) \text{ day}^{-1}$	0.068	Estimated
$\sigma$	$(0.065, 0.077) \text{ day}^{-1}$	0.0749	Estimated
$\gamma_1$	$(0.15, 0.25) \text{ day}^{-1}$	0.2	[5]
$\gamma_2$	$(0.159, 0.46) \text{ day}^{-1}$	0.4599	Estimated
$\gamma_3$	$(0.018, 0.0668) \text{ day}^{-1}$	0.066	Estimated
$\lambda$	$(0.0011, 0.0187) \text{ day}^{-1}$	0.0186	[41]
$\lambda_0$	$(0.00001, 0.008) \text{ day}^{-1}$	0.001	[41]
$c$	$(400, 2000) \text{ day}^{-1}$	430	Estimated
$\delta$	$(0.0066, 0.01) \text{ day}^{-1}$	0.0099	[5]
$\phi$	$(0.006999, 0.0099) \text{ day}^{-1}$	0.00900005	Estimated
$\Lambda$	$(100, 3000) \text{ day}^{-1}$	1319.294	[5]
$r_1$	$(0.001, 0.01) \text{ day}^{-1}$	0.006	[41]
$r_0$	$(0.001, 0.01) \text{ day}^{-1}$	0.005	[41]
$\mu$	$(0.00001, 0.0001) \text{ day}^{-1}$	0.0000425	[50]
$m_0$	$(100, 2000) \text{ day}^{-1}$	500	[41]
$\theta_1$	$(0.0158, 0.0178) \text{ day}^{-1}$	0.0178	[5]
$\theta_2$	$(0.1215, 9315) \text{ day}^{-1}$	0.9215	[5]
$\epsilon$	$(0.1, 0.2) \text{ day}^{-1}$	0.333	[5]
$\theta$	$(0.01, 0.034) \text{ day}^{-1}$	0.0005	[41]
$\xi$	$(0.009, 0.01) \text{ day}^{-1}$	0.008	Assumed
$\alpha$	$(0.1, 0.2) \text{ day}^{-1}$	0.3	Estimated
$\omega$	$(0, 10000) \text{ day}^{-1}$	6000	Estimated

### 4.3 Impact of control parameters

Figure 6a demonstrates that as awareness spreads more effectively, the number of symptomatic infections decreases, mainly because media coverage attracts susceptible people's attention. The rate at which people lose awareness, represented by  $\lambda_0$ , increases symptomatic infections, so efforts should be made to prevent this loss of awareness (see Figure 6b). To keep infection

Table 3: Normalized sensitivity index for each parameter for the COVID-19 model (1), for parameters values given in 2

Parameter	Sensitivity indices	Parameter	Sensitivity indices	Parameter	Sensitivity indices
$\Lambda$	1	$\eta$	0.65	$\gamma_3$	-0.26
$\beta$	0.76	$\sigma$	0.24	$\delta$	-0.03
$\beta_1$	0.23	$\phi$	0.07	$\alpha$	-0.20
$k_1$	0.22	$\gamma_1$	0.06	$k_2$	-0.22
$\theta_1$	0.03	$\epsilon$	-0.23	$\lambda$	-0.88
$\theta_2$	0.19	$m_0$	-0.25	$\mu$	-0.55
$\lambda_0$	0.55	$\gamma_2$	-0.75		
$c$	0.25				

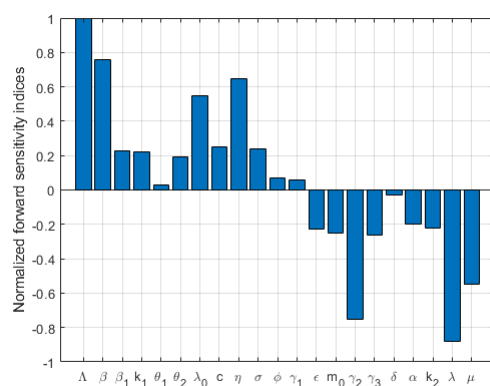
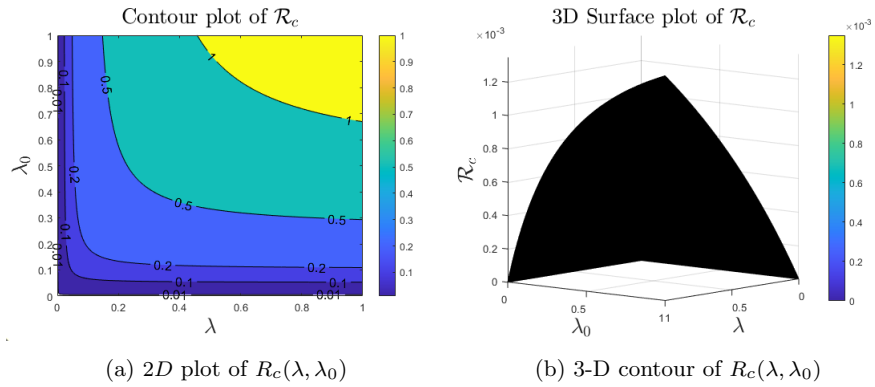
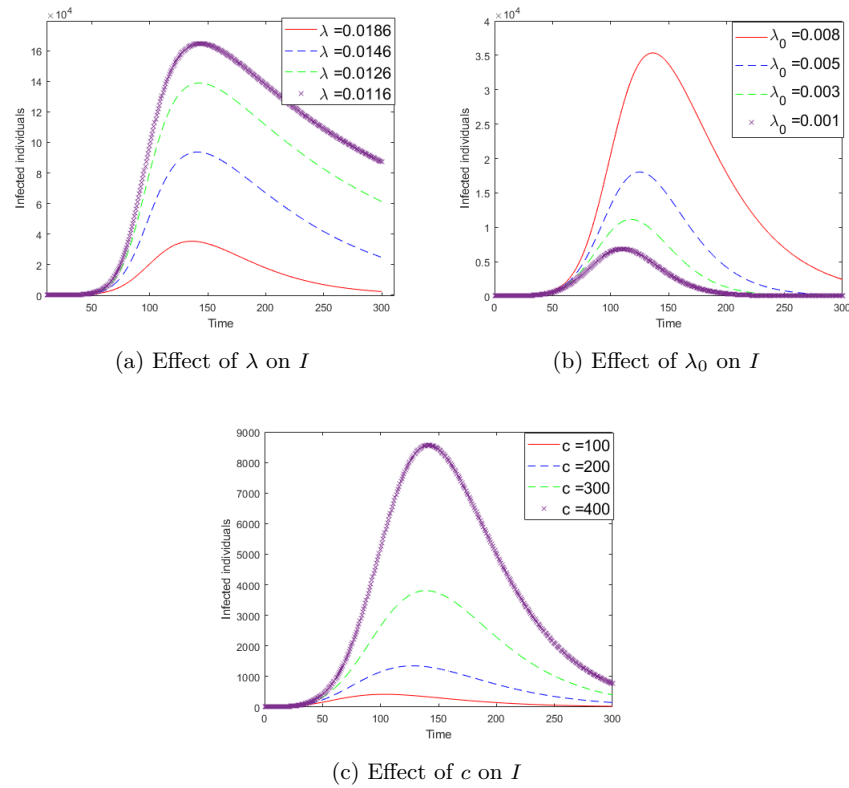
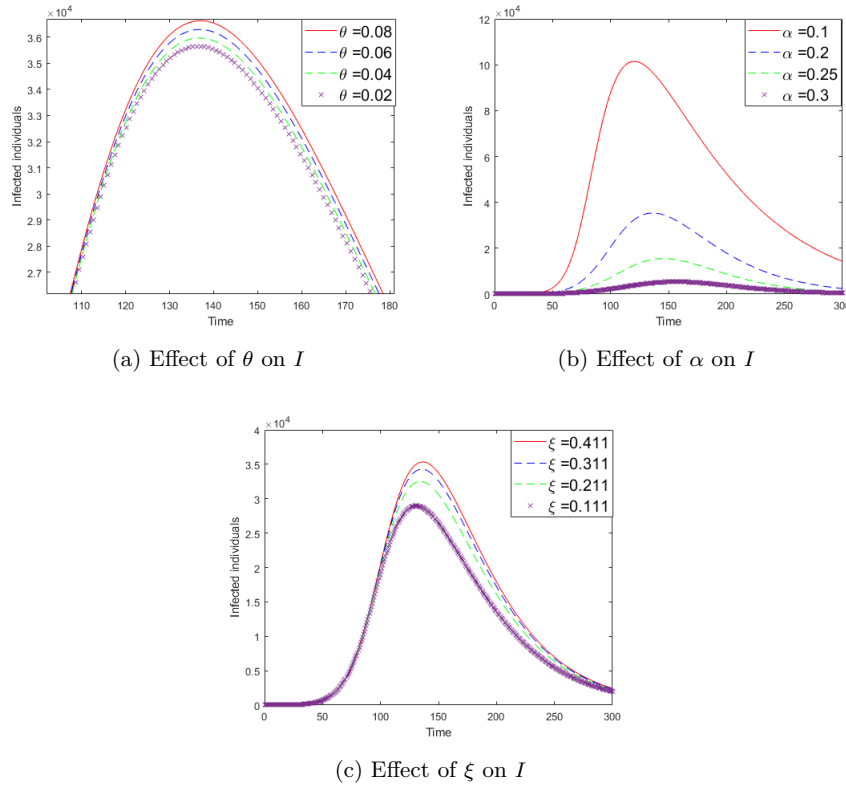


Figure 4: Normalized forward sensitivity indices of  $R_c$

levels low, it is crucial to maintain a steady level of baseline awareness,  $m_0$ . Finally, to reduce the transmission rates  $\beta$  and  $\beta_1$ , measures like wearing masks, and so on (refer to Figure 7b) should be taken.

Figure 5: contour plots of  $R_c$ Figure 6: The effects of varying  $\lambda$ ,  $\lambda_0$  and  $c$  on  $I$ .

Figure 7: The effects of varying  $\theta, \alpha$  and  $\xi$  on  $I$ .

#### 4.4 Impact of environment contamination

Environmental contamination plays a significant role in the transmission of COVID-19. Studies show that the virus can remain viable on copper surfaces for up to four hours, on cardboard for as long as 24 hours, and on stainless steel and plastic surfaces for up to 72 hours [56]. This study examines how environmental contamination affects the dynamics of the proposed model, specifically analyzing the model (1) to assess the impact of COVID-19 on environmental contamination caused by infected individuals. The time series presented in Figures 8a, 8b, and 8c demonstrate the effects of environmental contamination parameters on infection levels.

The number of infected individuals in the class  $I$  increases as the infection rate  $\beta_1$ , associated with the contaminated environment increases. Furthermore, as the rate of environmental contamination ( $\theta_1$  and  $\theta_2$ ) increases, the populations in class  $I$  also grow, as shown in the time series Figures (8a and 8b). Consequently, an increase in the factors  $\theta_1$  and  $\theta_2$  results in a shorter duration of the pandemic. If the contaminated environment is sanitized effectively (i.e., by increasing  $\epsilon$ ), the number of infected individuals in class  $I_a$  and  $I$  remains relatively stable, as demonstrated in Figure 8c. Consequently, eliminating the novel coronavirus from environments can help shorten the pandemic's duration and lower infection rates.

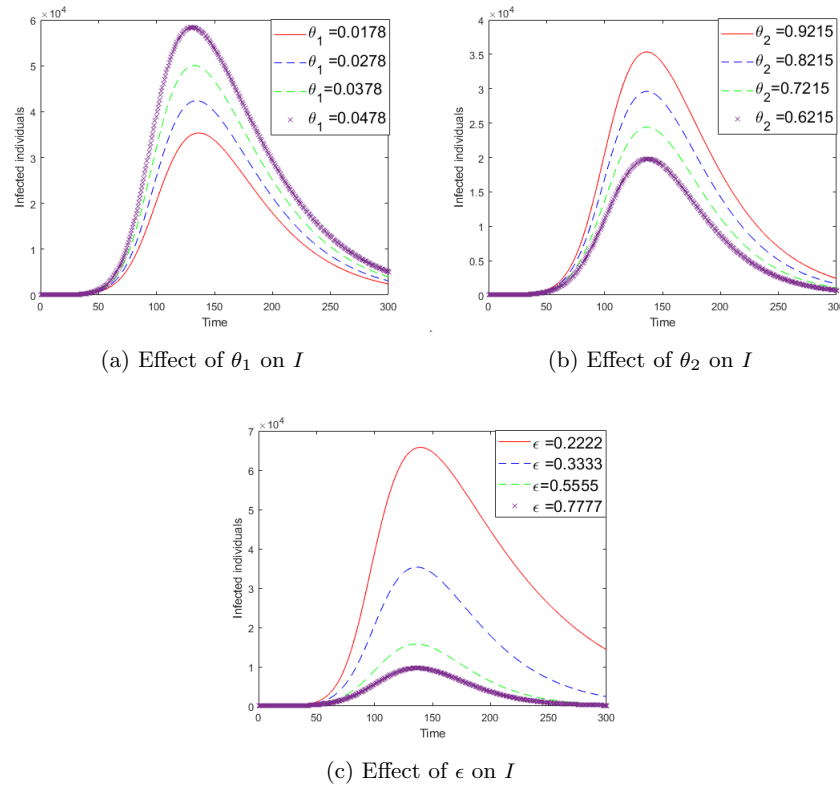


Figure 8: The effects of varying  $\theta_1, \theta_2, \epsilon$  on infected population  $I$ .

## 5 Optimal Control

In this section, we develop and evaluate an optimal control problem that integrates multiple strategies, including the proportion of individuals using face masks, the level of awareness, treatment rates, and the natural decline of the virus in the environment. These measures aim to mitigate the spread of the disease while accounting for economic consequences, such as productivity losses caused by both the disease and the interventions. The subsequent discussion focuses on these control strategies:

1. Control variable  $u_1(t)$ : The susceptible population is continuously exposed to a proportion of individuals wearing face masks, represented by the rate  $\alpha$ . Since implementing mask-wearing incurs costs, optimizing these costs is essential for policymakers. To achieve this, the mask-wearing rate  $\alpha$  is treated as a control variable,  $u_1(t)$ , in the context of model (1), to determine the optimal intervention strategy.
2. Control variable  $u_2(t)$ : As information about the virus disseminates, susceptible individuals transition to an aware class, with the speed of this transition being affected by the concentration of the infected population. Since promoting awareness incurs certain costs, policymakers should aim to optimize these initiatives. Therefore, the awareness rate, denoted by  $\lambda$ , is modeled as a control variable  $u_2(t)$ .
3. Control variable  $u_3(t)$ : The asymptomatic population transitions to the symptomatic state at a rate denoted by  $\gamma_1$ , where  $\phi$  represents the proportion of asymptomatic individuals who undergo this progression. Since treatment and isolation measures involve financial costs, policymakers must strategically manage these expenditures. To facilitate this optimization, the rate  $\gamma_1$  is modeled as a time-dependent control variable,  $u_3(t)$ . In this context,  $u_3$  represents interventions such as medical treatment or isolation efforts, which impact the progression from the asymptomatic to symptomatic states, in the context of model (1).
4. Control variable  $u_4(t)$ : Natural decay rate of the virus,  $\epsilon$  is treated as a control variable  $u_4(t)$ .

The control variables  $u_1(t)$ ,  $u_2(t)$ ,  $u_3(t)$ , and  $u_4(t)$  must be chosen from a set of allowable control functions defined by

$U = \{(u_1(t), u_2(t), u_3(t), u_4(t)) \mid 0 \leq u_1(t) \leq u_{1\max}, 0 \leq u_2(t) \leq u_{2\max}, 0 \leq u_3(t) \leq u_{3\max}, 0 \leq u_4(t) \leq u_{4\max}, t \in [t_0, t_f]\}$  [38]. Here,  $u_1(t)$ ,  $u_2(t)$ ,  $u_3(t)$ , and  $u_4(t)$  are measurable and bounded, and  $t_f$  is the final time for the intervention strategies. It is important to note that this final time  $t_f$  may vary for different diseases and applied interventions, depending on the goals of the control policy. So the following objective function is to minimize both the total number of infections and the related costs, which is expressed as

$$J(u) = \min \int_{t_0}^{t_f} \left( C + D_1 \frac{u_1^2}{2} + D_2 \frac{u_2^2}{2} + D_3 \frac{u_3^2}{2} + D_4 \frac{u_4^2}{2} \right) dt, \quad (11)$$

where  $C = C_1E + C_2I_a + C_3I + C_4U - C_5A$ ,  $u = (u_1, u_2, u_3, u_4)$ , and subject to constraints

$$\begin{cases} \frac{dS}{dt} &= \Lambda - \beta(1 - u_1)(I + \eta I_a)S - \beta_1(1 - u_1)SU + \lambda_0A + \xi R - \frac{u_2MS}{c+M} - \mu S \\ \frac{dA}{dt} &= \frac{u_2MS}{c+M} - \lambda_0A - \mu A \\ \frac{dE}{dt} &= \beta(1 - u_1)(I + \eta I_a)S + \beta_1(1 - u_1)SU - (k_2(1 - \sigma) + k_1\sigma + \mu)E \\ \frac{dI_a}{dt} &= k_2(1 - \sigma)E - u_3\phi I_a - \gamma_2(1 - \phi)I_a - \mu I_a \\ \frac{dI}{dt} &= k_1\sigma E + u_3\phi I_a - \gamma_3I - \delta I - \mu I \\ \frac{dR}{dt} &= \gamma_2(1 - \phi)I_a + \gamma_3I - \xi R - \mu R \\ \frac{dU}{dt} &= \theta_1I_a + \theta_2I - u_4U \\ \frac{dM}{dt} &= r_1(1 - \frac{\theta A}{\omega + A})I - r_0(M - m_0). \end{cases} \quad (12)$$

We presume the initial circumstances:

$S(0) = S_0 > 0$ ,  $E(0) \geq 0$ ,  $I_a(0) \geq 0$ ,  $I(0) \geq 0$ ,  $A(0) = A_0 > 0$ ,  $R(0) \geq 0$ ,  $U(0) \geq 0$ . In the objective function, the constants  $C_1, C_2, C_3, C_4$ , and  $C_5$  denote the weighting factors assigned to the exposed class, infected classes ( $I_a, I$ , and  $U$ ), and the aware class, respectively. The time-dependent control variables  $u_1, u_2, u_3$  and  $u_4$  are associated with the quadratic costs  $D_1u_1^2, D_2u_2^2, D_3u_3^2$  and  $D_4u_4^2$ , respectively, where the square terms indicate the severity of the costs.

The Filippov–Cesari Theorem [6] guarantees that the necessary conditions for achieving an optimal solution to the formulated optimal control problem are fulfilled. The Hessian matrix associated with the given cost functional is expressed as  $\mathbf{D} = \text{diag}(D_1, D_2, D_3, D_4)$ . Since this Hessian matrix is positive definite at all points, the objective functional  $J(u_1, u_2, u_3, u_4)$  is strictly convex. Consequently, there exists a constant  $D = \min D_1, D_2, D_3, D_4 > 0$  such that this lower bound applies to the integrand of the objective functional

$$\begin{aligned} C + D_1 \frac{u_1^2}{2} + D_2 \frac{u_2^2}{2} + D_3 \frac{u_3^2}{2} + D_4 \frac{u_4^2}{2} \\ \geq D (u_1^2 + u_2^2 + u_3^2 + u_4^2), \end{aligned}$$

holds if  $E + I_a + I + A + U \geq 0$ . We apply Pontryagin's maximum principle with the state variables  $S = S^*, A = A^*, E = E^*, I_a = I_a^*, I = I^*, R = R^*, U = U^*, M = M^*$ . We get the Hamiltonian function:

$$\begin{aligned} H = & C_1 E^* + C_2 I_a^* + C_3 I^* + C_4 U^* - C_5 A^* + D_1 \frac{u_1^2}{2} + D_2 \frac{u_2^2}{2} + D_3 \frac{u_3^2}{2} + D_4 \frac{u_4^2}{2} \\ & + \lambda_1 \frac{dS}{dt} + \lambda_2 \frac{dA}{dt} + \lambda_3 \frac{dE}{dt} + \lambda_4 \frac{dI_a}{dt} + \lambda_5 \frac{dI}{dt} + \lambda_6 \frac{dR}{dt} + \lambda_7 \frac{dU}{dt} + \lambda_8 \frac{dM}{dt}. \end{aligned} \quad (13)$$

The corresponding adjoint functions  $\lambda_i, i = 1, 2, \dots, 8$ , satisfy the equations:

$$\left\{ \begin{aligned} \frac{d\lambda_1}{dt} &= -\frac{\partial H}{\partial S} \\ &= -\left((1-u_1)\beta_1 U + (1-u_1)\beta(I+I_a\eta)(\lambda_2-\lambda_1) + \frac{u_2(\lambda_5-\lambda_1)M}{M+c} - \lambda_1\mu\right); \\ \frac{d\lambda_2}{dt} &= -\frac{\partial H}{\partial A} \\ &= -(-C_4 + \lambda_0\lambda_1 - \lambda_5(\lambda_0 + \mu) - \frac{r_1\theta\lambda_8\omega I}{(A+\omega)^2}); \\ \frac{d\lambda_3}{dt} &= -\frac{\partial H}{\partial E} \\ &= -(C_1 + k_2(\lambda_2 - \lambda_3)(-1 + \phi) + k_1\lambda_4\phi - \lambda_2(\mu + k_1\phi)); \\ \frac{d\lambda_4}{dt} &= -\frac{\partial H}{\partial I_a} \\ &= -(C_2 + (-1 + u_1)\beta\eta(\lambda_1 - \lambda_2)S - \lambda_3\mu + \gamma_2(\lambda_3 - \lambda_6)(-1 + \phi)) \\ &\quad -(\theta_1\lambda_7 - u_3(\lambda_3 - \lambda_4)\phi); \\ \frac{d\lambda_5}{dt} &= -\frac{\partial H}{\partial I} \\ &= -(C_3 + (-1 + u_1)\beta(\lambda_1 - \lambda_2)S + \gamma_3(-\lambda_4 + \lambda_6) + \theta_2\lambda_7 - \lambda_4(\delta + \mu)) \\ &\quad -\left(r_1\lambda_8\left(1 - \frac{A\theta}{A+\omega}\right)\right); \\ \frac{d\lambda_6}{dt} &= -\frac{\partial H}{\partial R} \\ &= -(\lambda_1\xi + \lambda_6(\mu + \xi)); \\ \frac{d\lambda_7}{dt} &= -\frac{\partial H}{\partial U} \\ &= (1-u_1)\beta_1(\lambda_1 - \lambda_2)S + u_4\lambda_7; \\ \frac{d\lambda_8}{dt} &= -\frac{\partial H}{\partial M} \\ &= \left(\frac{cu_2S(\lambda_1-\lambda_5)}{(M+c)^2} + r_0\lambda_8\right). \end{aligned} \right.$$

Under the universality condition  $\lambda_i(t_f) = 0$ , and considering that for all control inputs  $u_i$ , where  $i = 1, \dots, 4$ , the following condition holds:

$$\frac{\partial H}{\partial u_i} = 0,$$

The optimal control strategy, according to the appropriate variational principle, is determined as follows:

$$\begin{aligned} u_1^* &= \min \left\{ \max \left( 0, -\frac{(\beta I + \beta_1 U + \beta \eta I_a)(\lambda_1 - \lambda_2)}{D_1} \right), u_{1 \max} \right\}, \\ u_2^* &= \min \left\{ \max \left( 0, \frac{(\lambda_1 - \lambda_5)MS}{D_2(M+c)} \right), u_{2 \max} \right\}, \\ u_3^* &= \min \left\{ \max \left( 0, \frac{(\lambda_3 - \lambda_4)\phi I_a}{D_4} \right), u_{3 \max} \right\}, \\ u_4^* &= \min \left\{ \max \left( 0, \frac{\lambda_7 U^*}{D_5} \right), u_{4 \max} \right\}. \end{aligned}$$

## 5.1 Numerical solution of the model with optimal control

In this subsection, we conduct numerical simulations of the optimal control model to investigate how different time-varying control strategies influence the dynamics of disease spread. The simulations are based on the parameter values listed in Table 2, with some parameters obtained through data fitting using COVID-19 data from India. The analysis is conducted assuming a total population of around 1.40 billion. An initial estimate for the control functions is proposed for the specified period. Additionally, we apply the forward-backward sweep method, as outlined in [6], to numerically simulate the optimal control solution.

The trajectories shown in Figures 9a–9d reveal how each control variable  $u_i$  (for  $i = 1, 2, 3, 4$ ) uniquely influences the state variables  $E$ ,  $I_a$ ,  $I$ , and  $U$ . Solid lines represent the system's behavior without control, whereas dotted lines illustrate the impact of implementing control strategies. These interventions effectively reduce the number of infected individuals and the overall viral load, highlighting their success in significantly lowering the rate of virus introduction into the environment and limiting the progression into the in-

fectured class  $I$ . Figure 10 presents the optimal control levels for  $u_1$ ,  $u_2$ ,  $u_3$ , and  $u_4$  in the context of COVID-19 spread. In particular, Figure 10a shows that  $u_1$  reaches its highest level between days 250 and 270, then gradually declines until day 300. Similarly, Figure 10b illustrates that  $u_2$  peaks near day 290 and then slowly decreases up to day 300. Meanwhile, Figure 10c indicates that  $u_3$  peaks between days 200 and 250 before tapering off by day 300. Lastly, Figure 10d shows  $u_4$  reaching its maximum between days 270 and 290, then gradually declining until day 300.

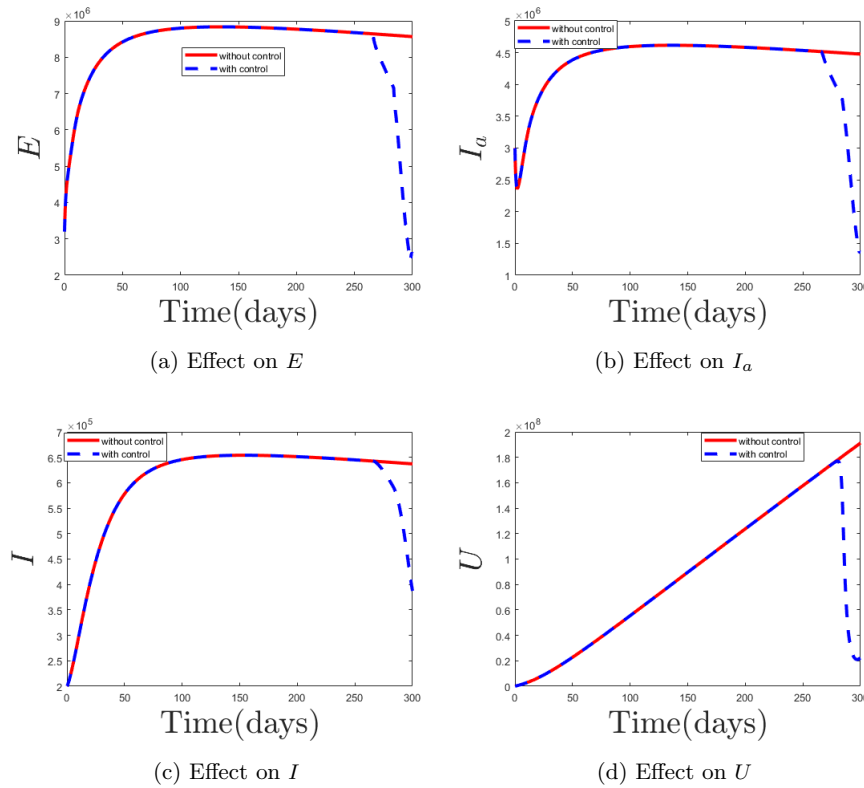


Figure 9: Effect of control measures:  $u_1, u_2, u_3, u_4$

To improve resource utilization and lower the costs associated with managing COVID-19 dynamics, we adopt selective strategies that concentrate on particular combinations of time-dependent control variables rather than employing all five control parameters simultaneously. This method allows us

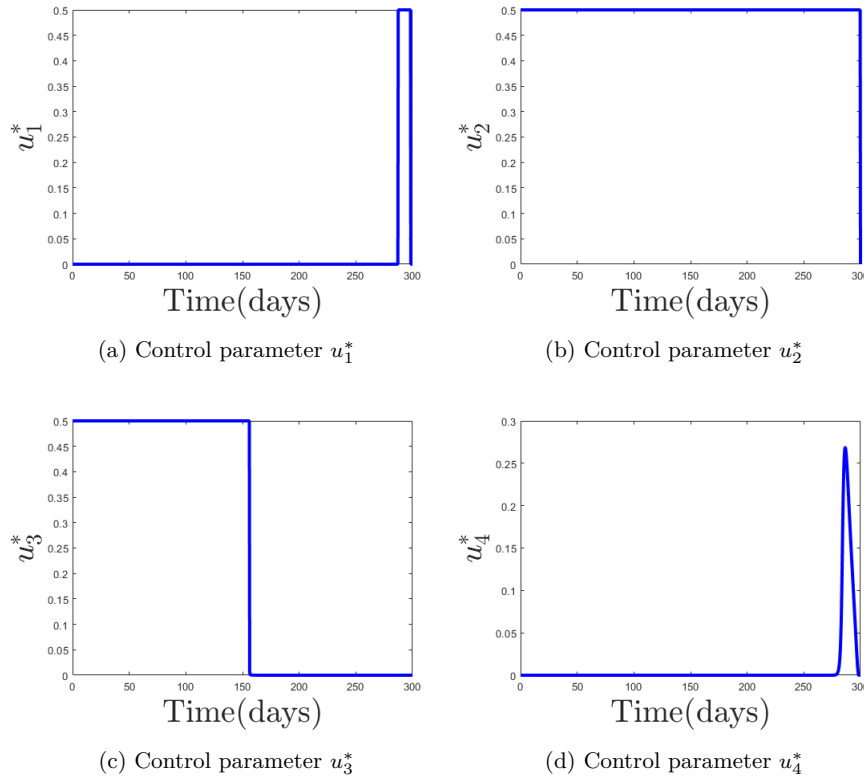
Figure 10: Control profiles:  $u_1, u_2, u_3$  and  $u_4$ 

Table 4: Scenarios with their combination strategy

Scenario	Strategies
A	S-1. ( $u_1 \neq 0, u_2 = 0, u_3 = 0, u_4 = 0$ )
	S-2. ( $u_1 = 0, u_2 \neq 0, u_3 = 0, u_4 = 0$ )
	S-3. ( $u_1 = 0, u_2 = 0, u_3 \neq 0, u_4 = 0$ )
	S-4. ( $u_1 \neq 0, u_2 \neq 0, u_3 \neq 0, u_4 = 0$ )
B	S-5. ( $u_1 = 0, u_2 = 0, u_3 = 0, u_4 \neq 0$ )
C	S-6. ( $u_1 \neq 0, u_2 \neq 0, u_3 \neq 0, u_4 \neq 0$ )

to evaluate the impact of different control combinations (see Table 4), pro-

viding valuable insights into the trade-offs and cost-effectiveness of targeted interventions.

We assess three scenarios, A, B, and C, according to the control strategies outlined in Table 4 and illustrated in Figure 10. Scenario A explores the influence of control variables  $u_1$ ,  $u_2$ , and  $u_3$  while excluding  $u_4$  (refer to Figures 11a–11d). Scenario B focuses on the effect of  $u_4$  in the absence of  $u_1$ ,  $u_2$ , and  $u_3$  (see Figure 11e). Lastly, Scenario C examines the combined impact of all control variables  $u_1$ ,  $u_2$ ,  $u_3$ , and  $u_4$  (Figure 11f).

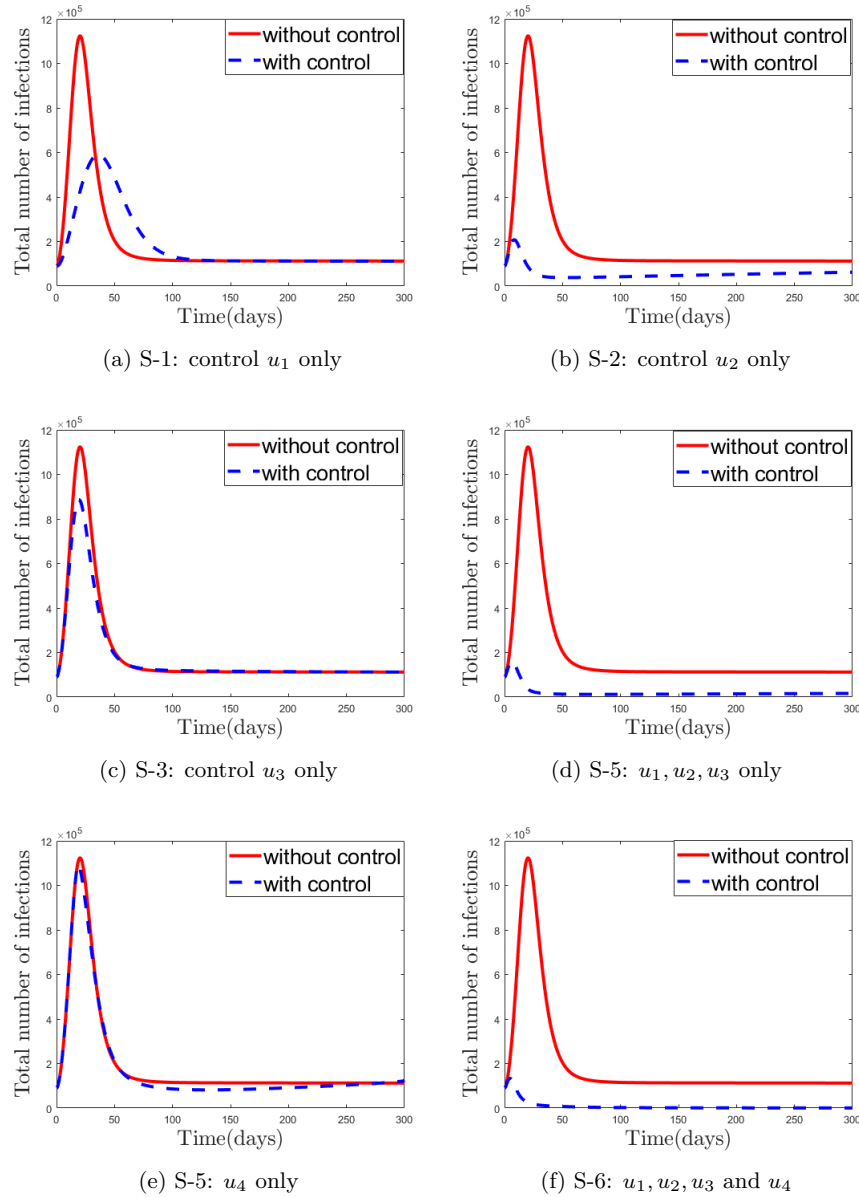
Numerical simulations suggest that targeting infective groups ( $u_1$ ,  $u_2$ , and  $u_3$ ) in Scenario A is more effective in reducing disease transmission than implementing environmental controls ( $u_4$ ) in Scenario B. Additionally, applying control measures to all infective groups simultaneously (Figures 11a–11d) yields a more significant impact compared to implementing them individually (Figures 11a–11c). Among the analyzed scenarios, Scenario C, which combines all control strategies, emerges as the most effective in limiting disease spread, as shown in Figure 11f.

Our analysis of optimal control indicates that successfully applying these strategies can greatly reduce transmission among vulnerable populations, leading to a marked decrease in the pandemic's overall impact.

## 6 Discussion and conclusion

The COVID-19 pandemic has presented significant public health challenges while exerting considerable economic pressure worldwide. With no pharmaceutical treatments initially available, nonpharmaceutical interventions like wearing face masks have been essential in curbing the virus's spread. Moreover, media coverage has played a key role in increasing public awareness and distributing critical information on preventive measures. This study examined the  $SAEI_aIRUM$  model, which integrates nonlinear functional responses to capture the effects of media coverage influence.

We theoretically analyzed the model within the dynamical systems framework, ensuring the solutions remain positive and bounded. Furthermore, we explored the biological significance of the control reproduction number, which was derived using the next-generation matrix. The identified control repro-

Figure 11: Effect of control  $u_1, u_2, u_3, u_4$  and  $u_5$  on total number of infections

duction number ( $R_c$ ) and model equilibrium points include disease-free and

endemic states. Additionally, we examined the local stability of the DFE under the assumption that  $R_c$  is less than one.

Section 4 presents the numerical simulations of the proposed model, utilizing COVID-19 data from India, covering the period from March 30, 2020, to January 24, 2021, as obtained from Johns Hopkins University [27]. As depicted in Figure 3, our model demonstrates superior predictive accuracy compared to the Asamoah et al. [5] model, reinforcing our assertion that it provides the most precise forecasts. A numerical analysis examines the impact of various control parameters on disease prevalence. The practical implementation of nonpharmaceutical interventions (NPIs), such as wearing face masks and self-administering treatment, contributes to reducing the control reproduction number, as illustrated in Figure 7b. Figures 6a–6b demonstrate that increasing the values of  $\lambda$  results in a decline in the number of infected individuals. Furthermore, the gradual waning of immunity acquired through infection increases the risk of reinfection, underscoring the importance of booster vaccinations in maintaining immunity and mitigating disease transmission, as shown in Figure 7c. Conversely, Figures 8a–8b highlight the effects of environmental contamination. Figure 8c illustrates how variations in  $\epsilon$  influence infection peaks, showing a decline as viral removal efforts intensify.

The control reproduction number ( $R_c$ ) determines whether the disease persists or diminishes. A normalized sensitivity analysis (Figure 4) explores the influence of different parameters on  $R_c$ . Normalized forward sensitivity analyses indicate that the recruitment rate ( $\Lambda$ ) has the most significant positive impact. In contrast, the proportion of susceptible individuals who become aware ( $\lambda$ ) exerts the most substantial negative effect on  $R_c$ .

We enhanced the  $SAEI_aIRUM$  model by embedding it within an optimal control framework, incorporating key interventions such as face masks, public awareness campaigns, medical treatment or isolation, and disinfection efforts. The impact of these measures was evaluated through simulations using the forward-backward sweep method. To assess the effectiveness and cost-efficiency of different strategies, we explored three distinct scenarios: Scenario A prioritizes managing infected individuals, Scenario B focuses on minimizing environmental contamination, and Scenario C combines both strategies.

Numerical simulations, illustrated in Figure 11, indicate that while Scenario A significantly reduces disease transmission compared to Scenario B, Scenario C, by combining all control measures, emerges as the most effective and cost-efficient strategy for controlling the spread of the disease.

Time delays significantly impact system dynamics, including delays in reporting confirmed cases caused by incubation periods and other influencing factors. Expanding this method could reveal more intricate dynamics in future research.

## Declarations

**Funding** This study was conducted without financial support from any particular funding organization.

**Data availability** All the data used in this article can be accessed freely through the website  
<https://data.humdata.org/dataset/novel-coronavirus-2019-ncov-cases#data-resources-0>.

**Author contribution** Both authors contributed equally.

**Conflict of interest** No potential conflicts of interest are associated with this research.

## References

- [1] Alanazi, K.M. *The asymptotic spreading speeds of COVID-19 with the effect of delay and quarantine*, AIMS Math. 9(7) (2024), 19397–19413.
- [2] Aldila, D. *Analyzing the impact of the media campaign and rapid testing for COVID-19 as an optimal control problem in East Java, Indonesia*, Chaos, Solitons Fractals, 141 (2020), 110364.

- [3] Aldila, D., Khoshnaw, S.H.A., Safitri, E., Anwar, Y.R., Bakry, A.R.Q., Samiadji, B.M., Anugerah, D.A., Gh, M.F. A., Ayulani, I.D. and Salim, S.N. *A mathematical study on the spread of COVID-19 considering social distancing and rapid assessment: The case of Jakarta, Indonesia*, Chaos Solitons Fractals, 139 (2020), 110042.
- [4] Anderson, R.M., Heesterbeek, H., Klinkenberg, D. and Hollingsworth, D.T. *How will country-based mitigation measures influence the course of the COVID-19 epidemic ?*, The Lancet, 395 (10228) (2020), 931–934.
- [5] Asamoah, J.K.K., Owusu, M.A., Jin, Z., Oduro, F.T., Abidemi, A. and Gyasi, E.O. *Global stability and cost-effectiveness analysis of COVID-19 considering the impact of the environment: using data from Ghana*, Chaos, Solitons Fractals, 140 (2020), 110103.
- [6] Atangana, A. and İğret, A.S. *Mathematical model of COVID-19 spread in Turkey and South Africa: theory, methods, and applications*, Adv. Differ. Equ. 2020(1) (2020), 1–89.
- [7] Bajiyya, V.P., Bugalia, S., Tripathi, J.P. and Martcheva, M. *Deciphering the transmission dynamics of COVID- 19 in India: optimal control and cost effective analysis*, J. Biol. Dyn. 16(1) (2022), 665–712.
- [8] Baroudi, M., Laarabi, H., Zouhri, S., Rachik, M. and Abta, A. *Stochastic optimal control model for COVID-19: mask wearing and active screening/testing*, J. Appl. Math. Comput. 70(6) (2024), 6411–6441.
- [9] BBC, News <https://www.bbc.com/news/world-asia-india-52077395> (Accessed: June, 2022).
- [10] Birkhoff, G. and Rota, G. *Ordinary Differential Equations*, Wiley, United Kingdom, 1978.
- [11] Castillo-Chavez, C. and Song, B. *Dynamical models of tuberculosis and their applications*, Math. Biosci. Eng. 1(2) (2004), 361–404.
- [12] Chang, X., Liu, M., Jin, Z. and Wang, J. *Studying on the impact of media coverage on the spread of COVID-19 in Hubei Province, China*, Math. Biosci. Eng. 17(4) (2020), 3147–3159.

- [13] Chen, K., Pun, C.S. and Wong, H.Y. *Efficient social distancing during the COVID-19 pandemic: integrating economic and public health considerations*, European J. Oper. Res. 304(1) (2023), 84–98.
- [14] Chen, N., Zhou, M., Dong, X., Qu, J., Gong, F., Han, Y., Qiu, Y., Wang, J., Liu, Y., Wei, Y. and Xia, J.A. *Epidemiological and clinical characteristics of 99 cases of 2019 novel coronavirus pneumonia in Wuhan, China: a descriptive study*, The Lancet, 395(10223) (2020), 507–513.
- [15] Chen, T., Li, Z. and Zhang, G. *Analysis of a COVID-19 model with media coverage and limited resources*, Math. Biosci. Eng. 21(4) (2024), 5283–5307.
- [16] Cheneke, K. *Optimal control analysis for modeling HIV transmission*, Iran. J. Numer. Anal. Optim. 13(4) (2023), 747–762.
- [17] Cucinotta, D. and Vanelli, M. *WHO declares COVID-19 a pandemic*, Acta Biomed. 91(1) (2020), 157.
- [18] d’Onofrio, A., Iannelli, M., Manfredi, P. and Marinoschi, G. *Epidemic control by social distancing and vaccination: optimal strategies and remarks on the COVID-19 Italian response policy*, Math. Biosci. Eng. 21(7) (2024), 6493–6520.
- [19] Dwivedi, S., Perumal, S.K., Kumar, S., Bhattacharyya, S. and Kumari, N. *Impact of cross border reverse migration in Delhi- UP region of India during COVID-19 lockdown*, Comput. Math. Biophys. 11 (2023), 1–26.
- [20] Gholami, M., Mirhosseini, A.S. and Heidari, A. *Designing a sliding mode controller for a class of multi-controller COVID-19 disease model*, Iran. J. Numer. Anal. Optim. 15(1) (2025), 27–53.
- [21] Ghosh, I., Tiwari, P.K., Samanta, S., Elmojtaba, I.M., Al-Salti, N. and Chattopadhyay, J. *A simple SI-type model for HIV/AIDS with media and self-imposed psychological fear*, Math. Biosci. 306 (2018), 160–169.
- [22] Government of India <https://www.mygov.in/covid-19> (Accessed: June, 2022).

- [23] Guo, Y. and Li, T. *Modeling the competitive transmission of the Omicron strain and Delta strain of COVID-19*, J. Math. Anal. Appl. 526(2) (2023), 127283.
- [24] Gupta, S., Rajoria, Y.K. and Sahu, G.P. *Mathematical Modelling on Dynamics of Multi-variant SARS-CoV-2 Virus: Estimating Delta and Omicron Variant Impact on COVID-19*, IJAM 55(1) (2025), 180–188.
- [25] Hao, J., Huang, L., Liu, M. and Ma, Y. *Analysis of the COVID-19 model with self-protection and isolation measures affected by the environment*, Math. Biosci. Eng. 21(4) (2024), 4835–4852.
- [26] Huang, C., Wang, Y., Li, X., Ren, L., Zhao, J., Hu, Y., Zhang, L., Fan, G., Xu, J., Gu, X. and Cheng, Z. *Clinical features of patients infected with 2019 novel coronavirus in Wuhan, China*, The lancet, 395(10223) (2020), 497–506.
- [27] Humanitarian Data Exchange. Novel Coronavirus 2019 (COVID-19) Cases <https://data.humdata.org/dataset/novel-coronavirus-2019-ncov-cases#data-resources-0> (Accessed: June, 2023).
- [28] Iboi, E., Sharomi, O.O., Ngonghala, C. and Gumel, A.B. *Mathematical modeling and analysis of COVID-19 pandemic in Nigeria*, MedRxiv, (2020), 1–24.
- [29] Kahn, J.S. and McIntosh, K. *History and recent advances in coronavirus discovery* Pediatr. Infect. Dis. J. 24(11) (2005), S223–S227.
- [30] Killerby, M.E., Biggs, H.M., Midgley, C.M., Gerber, S.I. and Watson, J.T. *Middle East respiratory syndrome coronavirus transmission* Emerg. Infect. Dis. 26(2) (2020), 191.
- [31] Koura, A.F., Raslan, K.R., Ali, K.K. and Shaalan, M.A. *A numerical investigation for the COVID-19 spatiotemporal lockdown-vaccination model*, Comput. Methods Differ. Equ. 12(4) (2024), 669–686.
- [32] Kumar, A., Srivastava, P.K., Dong, Y., and Takeuchi, Y. *Optimal control of infectious disease: Informationinduced vaccination and limited treatment*, Phys. A: Stat. Mech. Appl. 542 (2020), 123196.

- [33] Kurkina, E. and Koltsova, E. *Mathematical modeling of the propagation of Covid-19 pandemic waves in the World*, Comput. Math. Model. 32(2021), 147–170.
- [34] Lakhal, M., Taki, R., El F.M. and El, G.T. *Quarantine alone or in combination with treatment measures to control COVID-19*, J. Anal. 31(4) (2023), 2347–2369.
- [35] LaSalle, J.P. *Stability theory and invariance principles*, Elsevier, New York, 1976.
- [36] Li, Q., Guan, X., Wu, P., Wang, X., Zhou, L., Tong, Y., Ren, R., Leung, K.S., Lau, E.H., Wong, J.Y. and Xing, X. *Early transmission dynamics in Wuhan, China, of novel coronavirus infected pneumonia*, N. Engl. J. Med. 382(13) (2020), 1199–1207.
- [37] Liu, J. and Wang, X.S. *Dynamic optimal allocation of medical resources: a case study of face masks during the first COVID-19 epidemic wave in the United States*, Math. Biosci. Eng. 20(7) (2023), 12472–12485.
- [38] Martcheva, M. *An introduction to mathematical epidemiology*, Springer, United States, 2015.
- [39] Memon, Z., Qureshi, S. and Memon, B.R. *Assessing the role of quarantine and isolation as control strategies for COVID-19 outbreak: a case study*, Chaos Solitons Fractals, 144 (2021), 110655.
- [40] Misra, A., Sharma, A. and Shukla, J. *Modeling and analysis of effects of awareness programs by media on the spread of infectious diseases*, Math. Comput. Model. 53(5-6) (2011), 1221–1228.
- [41] Misra, A.K., Rai, R.K. and Takeuchi, Y. *Modeling the control of infectious diseases: Effects of TV and social media advertisements*, Math. Biosci. Eng. 15(6) (2018), 1315–1343.
- [42] Rai, R.K., Khajanchi, S., Tiwari, P.K., Venturino, E. and Misra, A.K. *Impact of social media advertisements on the transmission dynamics of COVID-19 pandemic in India*, J. Appl. Math. Comput. (2022), 1–26.


- [43] Sahu, G.P. and Dhar, J. *Analysis of an SVEIS epidemic model with partial temporary immunity and saturation incidence rate*, Appl. Math. Model. 36(3) (2012), 908–923.
- [44] Sahu, G.P. and Dhar, J. *Dynamics of an SEQIHRs epidemic model with media coverage, quarantine and isolation in a community with preexisting immunity*, J. Math. Anal. Appl. 421(2) (2015), 1651–1672.
- [45] Sardar, T., Nadim, S.k.S., Rana, S. and Chattopadhyay, J. *Assessment of lockdown effect in some states and overall India: a predictive mathematical study on COVID-19 outbreak*, Chaos Solitons Fractals, 139 (2020), 1-10.
- [46] Sarkar, K., Mondal, J. and Khajanchi, S. *How do the contaminated environment influence the transmission dynamics of COVID-19 pandemic?*, Eur. Phys. J: Spec. Top. 231(18-20) (2022), 3697–3716.
- [47] Senapati, A., Rana, S., Das, T. and Chattopadhyay, J. *Impact of intervention on the spread of COVID-19 in India: A model based study*, J. Theor. Biol. 523 (2021) 110711.
- [48] Sooknanan, J. and Comissiong, D. *Trending on social media: integrating social media into infectious disease dynamics*, Bull. Math. Biol. 82(7) (2020), 86.
- [49] Sooknanan, J. and Mays, N. *Harnessing social media in the modelling of pandemics challenges and opportunities*, Bull. Math. Biol. 83(5) (2021), 57.
- [50] Srivastav, A.K., Tiwari, P.K., Srivastava, P.K., Ghosh, M. and Kang, Y. *A mathematical model for the impacts of face mask, hospitalization and quarantine on the dynamics of COVID-19 in India: deterministic vs. stochastic*, Math. Biosci. Eng. 18(1) (2021), 182–213.
- [51] Su, S., Wong, G., Shi, W., Liu, J., Lai, A.C.K., Zhou, J., Liu, W., Bi, Y. and Gao, G.F. *Epidemiology, genetic recombination, and pathogenesis of coronaviruses*, Trends Microbiol. 24(6) (2016), 490–502.

- [52] Sun, D., Li, Y., Teng, Z., Zhang, T., and Lu, J. *Dynamical properties in an SVEIR epidemic model with age-dependent vaccination, latency, infection, and relapse*, Math. Methods Appl. Sci. 44(17) (2021), 12810–12834.
- [53] Thakur, A.S. and Sahu, G.P. *Modeling the COVID-19 Dynamics with Omicron Variant, Non-pharmaceutical Interventions, and Environmental Contamination* Differ. Equations Dyn. Syst. (2025), 1–25.
- [54] The Indian Express <https://indianexpress.com/article/coronavirus/coronavirus-india-infection-rate-china-6321154/> (Accessed: June, 2023).
- [55] Van den Driessche, P. and Watmough, J. *Reproduction numbers and subthreshold endemic equilibria for compartmental models of disease transmission*, Math. Biosci. 180(1-2) (2002), 29–48.
- [56] Van Doremalen, N., Bushmaker, T., Morris, D.H., Holbrook, M.G., Gamble, A., Williamson, B.N., Tamin, A., Harcourt, J.L., Thornburg, N.J., Gerber, S.I. and Lloyd-Smith, J.O. *Aerosol and surface stability of SARS-CoV-2 as compared with SARS-CoV-1*, N. Engl. J. Med. 382(16) (2020), 1564–1567.
- [57] Wang, W. and Ruan, S. *Bifurcation in an epidemic model with constant removal rate of the infectives*, J. Math. Anal. Appl. 291(2) (2004), 775–793.
- [58] Wang, X., Liang, Y., Li, J. and Liu, M. *Modeling COVID-19 transmission dynamics incorporating media coverage and vaccination*, Math. Biosci. Eng. 20 (2023), 10392–10403.
- [59] Wardeh, M., Baylis, M. and Blagrove, M. S. *Predicting mammalian hosts in which novel coronaviruses can be generated*, Nat. Commun. 12(1) (2021), 780.
- [60] Willman, M., Kobasa, D. and Kindrachuk, J.A. *Comparative analysis of factors influencing two outbreaks of Middle Eastern respiratory syndrome (MERS) in Saudi Arabia and South Korea*, Viruses 11(12) (2019), 1119.

- [61] Yuan, R., Ma, Y., Shen, C., Zhao, J., Luo, X. and Liu, M. *Global dynamics of COVID-19 epidemic model with recessive infection and isolation*, Math. Biosci. Eng. 18(2) (2021), 1833–1844.
- [62] Yuan, Y. and Li, N. *Optimal control and cost-effectiveness analysis for a COVID-19 model with individual protection awareness*, Phys. A: Stat. Mech. Appl. 603 (2022), 127804.
- [63] Zhao, S., Lin, Q., Ran, J., Musa, S.S., Yang, G., Wang, W., Lou, Y., Gao, D., Yang, L., He, D. and Wang, M.H. *Preliminary estimation of the basic reproduction number of novel coronavirus (2019-nCoV) in China, from 2019 to 2020: A data driven analysis in the early phase of the outbreak*, Int. J. Inf. Dis. 92 (2020), 214–217.



## Space-time localized scheme to solve some partial integro-differential equations

M. Hamaidi\*, , M. Briki, A. Nouara and B. Hamdi

### Abstract

It has been demonstrated that the space-time localized radial basis functions collocation method has very good accuracy in several research studies. In this paper, we extend the method to solve the partial integro-differential equations. Since the unknowns of the localized scheme are the values of the

---

\*Corresponding author

Received 19 March 2025; revised 17 April 2025; accepted 26 April 2025

Mohammed Hamaidi

Department of Mathematics, Faculty of exact sciences and computer science, Ziane Achour University, Djelfa, Algeria. e-mail: [hamaidi@yahoo.com](mailto:hamaidi@yahoo.com)

Mabrouk Briki

Department of Mathematics, Faculty of exact sciences and computer science, Ziane Achour University, Djelfa, Algeria. e-mail: [mabroukbriki@yahoo.fr](mailto:mabroukbriki@yahoo.fr)

Ahmed Nouara

Department of Mathematics, Faculty of exact sciences and computer science, Ziane Achour University, Djelfa, Algeria. e-mail: [ahmednouara150@gmail.com](mailto:ahmednouara150@gmail.com)

Brahim Hamdi

Department of Mathematics, Faculty of exact sciences and computer science, Ziane Achour University, Djelfa, Algeria. e-mail: [brahim.hamdi@univ-djelfa.dz](mailto:brahim.hamdi@univ-djelfa.dz)

### How to cite this article

Hamaidi, M., Briki, M., Nouara, A. and Hamdi, B., Space-time localized scheme to solve some partial integro-differential equations. *Iran. J. Numer. Anal. Optim.*, 2025; 15(3): 993-1011. <https://doi.org/10.22067/ijnao.2025.92731.1617>

interpolated function, the method can be easily combined with the trapezoidal rule to find the numerical solution. The main advantages of such formulation are as follows: The time discretization is not applied; the time stability analysis is not discussed; and the recomputation of the resulting matrix at each time level is avoided because the matrix is computed once. Different examples are solved to show the accuracy of such a method.

**AMS subject classifications (2020):** Primary 45K05, 65M99; Secondary 65N22.

**Keywords:** Partial integro-differential equation; Localized radial basis functions; Space-time scheme; Collocation method; Trapezoidal rule.

## 1 Introduction

A partial integro-differential equation (PIDE) is an equation in which the unknown function appears under the sign of integration and contains the unknown function and its derivatives with respect to the space and time variables. Many problems in various fields of physical, engineering, biological, and epidemiology models are described by PIDEs.

The numerical solution of the PIDEs has recently gained much attention from researchers. To our best knowledge, in all published works, the solution methods are based on first discretizing the time variable by applying any time-stepping algorithms as implicit, explicit, Runge–Kutta or others, and seeking the approximate solution at each instant  $t$  in a space domain problem. Siddiqi and Arshed [14] employed cubic b-spline functions for spatial derivatives and the Euler backward formula for time derivatives to solve the PIDE. In [16, 15], it was used the 2-point Euler backward finite difference method was used for the discretization in time with a combination of the finite difference method and the trapezoidal rule to solve the PIDE. El-Sayed, Helal and El-Azab [4] implemented the implicit and explicit finite difference schemes for the time discretization. In most published works, these methods are based on differentiating between time and space variables. All methods start by discretizing the time variable using implicit, explicit, Runge–Kutta, or any other known method, and then solving the problem by

computing the approximate solution at each time  $t$ . The global radial basis function (RBF) method for solving the linear integro-differential equations was investigated by Golbabai and Seifollahi [6, 7]. Parand and Rad [12] presented the RBF collocation method for one-dimensional Volterra-Fredholm-Hammerstein integral equations. All works used the global formulation of RBF [6, 7, 12, 1, 17]. Therefore, in this paper, we develop an RBF-based space-time localized meshless collocation method combined with the trapezoidal rule to solve the space and time PIDE as space-time one, without differentiating between space and time variables. The posed problem can be solved once to approximating the solution at any space-time point  $(x, t)$ . The main advantages of the considered technique are as follows:

- (a). The discussion of the time stability analysis of the discrete system is avoided [8].
- (b). The computational time when dealing with PIDEs with time-dependent coefficients is reduced as there is no need to recompute the matrix for the resulting algebraic system at each time level.
- (c). The method uses the sparse matrices to store only the nonzero elements, so we save a significant amount of memory and speed up the resolution of the linear system.

The paper is organized as follows. In section 2, we introduce the formulation of the PIDE as a space-time problem and the space-time localized RBF method implementation. Section 3 is devoted to the discussion of results obtained by solving different PIDE examples. We conclude in Section 4.

## 2 Numerical details and discretization schemes

In this section, we describe the discretization scheme and the methodology used to solve the PIDE. The considered PIDE has the following form:

$$\left\{ \begin{array}{ll} \mathcal{D}_{(x,t)}u + \mathcal{I}_{(x,t)}u = f(x,t) & \text{for all } x \in (a,b), \text{ for all } t \in (0,T], \\ u(a,t) = g_1(t) & \text{for all } t \in (0,T), \\ u(b,t) = g_2(t) & \text{for all } t \in (0,T), \\ u(x,0) = u_0(x) & \text{for all } x \in [a,b], \end{array} \right. \quad (1)$$

where  $\mathcal{D}_{(x,t)}$  is a differential operator of second order with variable coefficients defined by

$$\mathcal{D}_{(x,t)}u = \frac{\partial u}{\partial t} - a(x,t)\frac{\partial^2 u}{\partial x^2} + b(x,t)\frac{\partial u}{\partial x} + c(x,t)u, \quad (2)$$

$\mathcal{I}_{(x,t)}$  is an integral operator of the form

$$\mathcal{I}_{(x,t)}u = \int_0^t k(x,t,s)u(x,s)ds, \quad (3)$$

and  $f, g_1, g_2, u_0$ , and  $k$  are given smooth functions.

## 2.1 Space-time problem methodology

The formulation of the time-depends problem given by the system (1) as a space-time one starts by combining the space variable  $x$  and the time variable  $t$  in one vector  $\hat{x} = (x, t)$ . The constructed variable vector belongs to the space-time domain  $\Omega_T = [a, b] \times [0, T]$  represented by Figure 1. The boundary of the new formulated domain  $\Omega_T$  is given by  $\partial\Omega_T = \Gamma_1 \cup \Gamma_2 \cup \Gamma_3 \cup \Gamma_4$ , where  $\Gamma_1 = \{a\} \times [0, T]$ ,  $\Gamma_2 = \{b\} \times [0, T]$ ,  $\Gamma_3 = [a, b] \times \{0\}$ , and  $\Gamma_4 = [a, b] \times \{T\}$ .

Then, the problem has the new form:

$$\begin{cases} \mathcal{D}_{(x,t)}u + \mathcal{I}_{(x,t)}u = f(x,t) & \text{for all } (x,t) \in \Omega_T \\ u(x,t) = g_1(t) & \text{for all } (x,t) \in \Gamma_1, \\ u(x,t) = g_2(t) & \text{for all } (x,t) \in \Gamma_2, \\ u(x,t) = u_0(x) & \text{for all } (x,t) \in \Gamma_3, \\ \mathcal{D}_{(x,t)}u + \mathcal{I}_{(x,t)}u = f(x,t) & \text{for all } (x,t) \in \Gamma_4, \end{cases} \quad (4)$$

or in a reduced form, by setting  $\Omega_T = \Omega_T \cup \Gamma_4$  and  $\partial\Omega_T = \Gamma_1 \cup \Gamma_2 \cup \Gamma_3$ , we have

$$\begin{cases} \mathcal{D}_{\hat{x}}u + \mathcal{I}_{\hat{x}}u = f(\hat{x}) & \text{for all } \hat{x} \in \Omega_T, \\ u(\hat{x}) = g(\hat{x}) & \text{for all } \hat{x} \in \partial\Omega_T, \end{cases} \quad (5)$$

where

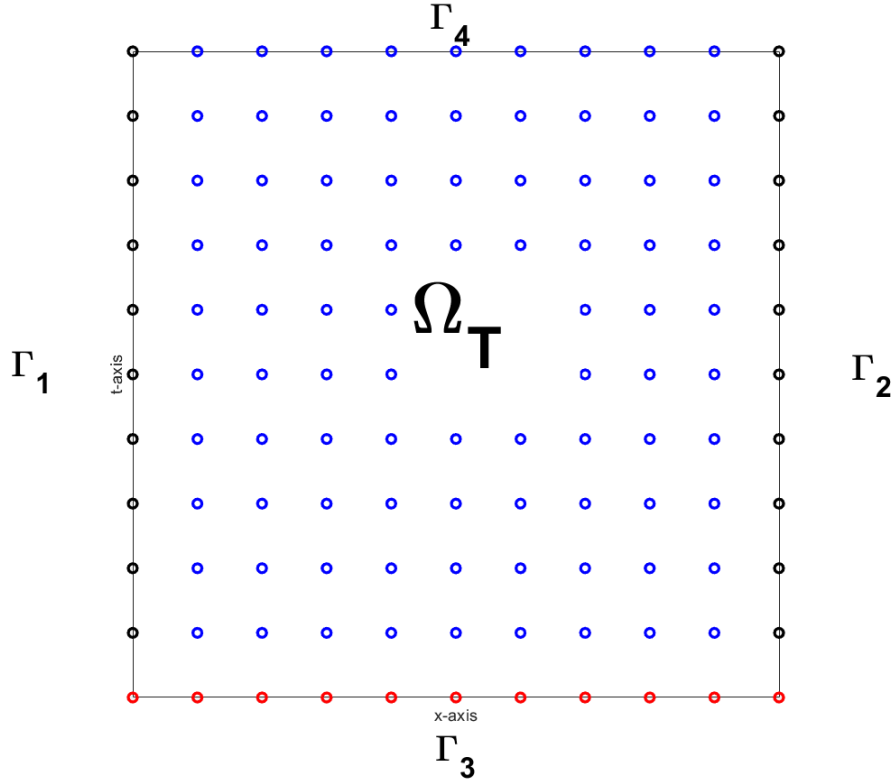


Figure 1: Space-time domain

$$\begin{cases} g(\hat{x}) = g_1(\hat{x}) & \text{for all } \hat{x} \in \Gamma_1, \\ g(\hat{x}) = g_2(\hat{x}) & \text{for all } \hat{x} \in \Gamma_2, \\ g(\hat{x}) = u_0(\hat{x}) & \text{for all } \hat{x} \in \Gamma_3. \end{cases} \quad (6)$$

## 2.2 The space-time localized RBFs scheme

To recall the technique, let  $\{\hat{x}_i\}_{i=1}^{N_i}$  and  $\{\hat{x}_i\}_{i=N_i+1}^N$  be center nodes in  $\Omega_T$  and  $\partial\Omega_T$ , respectively (Interior and boundary nodes, where  $N$  is the total number of nodes in the space-time domain  $\Omega_T$ ). To approximate the differential operator, using the localized RBF method, we first need to derive the local approximation of the unknown function  $u$  [2, 3]. Then the local approx-

imation of  $\mathcal{D}_{(x,t)}u(x, t)$  can be determined easily based on the components of the function  $u$ . So, the local approximation of  $u$  in an influence domain  $\Omega_T^j$  associated with a selecting collocation point  $\hat{x}_j = (x_j, t_j)$  and containing a number  $n_j$  of nearest neighboring points  $\{\hat{x}_k^{[j]} = (x_k^{[j]}, t_k^{[j]})\}_{k=1}^{n_j} \in \Omega_T^j$ , is given by

$$u(\hat{x}_j) \simeq \hat{u}(\hat{x}_j) = \sum_{k=1}^{n_j} \alpha_k \phi(\|\hat{x}_j - \hat{x}_k^{[j]}\|), \quad (7)$$

where  $\{\alpha_k\}_{k=1}^{n_j}$  are the unknown coefficients,  $\|\cdot\|$  is the Euclidean norm, and  $\phi$  is the chosen RBF. There are many different RBFs to choose from. Among them we can mention the multiquadric function  $\phi(r) = \sqrt{1 + (\epsilon r)^2}$ , Which has been proven in many references [11, 5, 13] to be the most effective over the past few decades (The real  $\epsilon$  is the shape parameter of the RBF).

Using the collocation method, (7) is then applied to all collocation points  $\{\hat{x}_k^{[j]}\}_{k=1}^{n_j}$  belonging to the influence domain  $\Omega_T^j$  of  $\hat{x}_j$ . Then we have the following  $n_j \times n_j$  linear system:

$$\hat{\mathbf{u}}^{[j]} = \mathbf{\Phi}^{[j]} \boldsymbol{\alpha}^{[j]}, \quad (8)$$

where  $\mathbf{\Phi}^{[j]} = \left[ \phi(\|\hat{x}_m^{[j]} - \hat{x}_n^{[j]}\|) \right]_{1 \leq m, n \leq n_j}$ ,  $\boldsymbol{\alpha}^{[j]} = [\alpha_1^{[j]}, \alpha_2^{[j]}, \dots, \alpha_{n_j}^{[j]}]$ , and  $\hat{\mathbf{u}}^{[j]} = [u(\hat{x}_1^{[j]}), u(\hat{x}_2^{[j]}), \dots, u(\hat{x}_{n_j}^{[j]})]$ .

Then, the problem of seeking the expansion coefficients  $\{\alpha_k\}_{k=1}^{n_j}$  is transformed into a determination of the values of solution  $\hat{u}^{[j]}$  at each center point  $\{\hat{x}_k^{[j]}\}_{k=1}^{n_j} \subset \Omega_T^j$  by using the equation

$$\boldsymbol{\alpha}^{[j]} = (\mathbf{\Phi}^{[j]})^{-1} \cdot \hat{\mathbf{u}}^{[j]}. \quad (9)$$

The local approximation of  $\mathcal{D}_{\hat{x}}u$  can be determined by applying the differential operators  $\mathcal{D}_{\hat{x}}$  to the equation (7) for any selected center point  $\hat{x}_j$  in any sub-domain  $\Omega_j$  (Figure 2). For  $\hat{x}_j \in \Omega_j$ , we obtain the following equation:

$$\begin{aligned} \mathcal{D}_{\hat{x}}\hat{u}(\hat{x}_j) &= \sum_{k=1}^{n_j} \alpha_k \mathcal{D}_{\hat{x}}\phi(\|\hat{x}_j - \hat{x}_k^{[j]}\|) \\ &= \mathbf{D}^{[j]} \cdot \hat{\mathbf{u}}^{[j]} \\ &= \mathbf{D}^{[j]} \cdot \hat{\mathbf{u}}, \end{aligned} \quad (10)$$

where  $\hat{\mathbf{u}}^{[j]} = [u(\hat{x}_1^{[j]}), u(\hat{x}_2^{[j]}), \dots, u(\hat{x}_{n_j}^{[j]})]$  and  $\mathbf{D}^{[j]} = (\mathcal{D}_{\hat{x}} \Phi^{[j]}) \cdot (\Phi^{[j]})^{-1}$ .

To switch from the local system (10) to the global one, the vector  $\hat{\mathbf{u}} = [u(\hat{x}_1), u(\hat{x}_2), \dots, u(\hat{x}_N)]$  is incorporated in (10) by adding zeros at the proper locations based on the mapping of  $\hat{\mathbf{u}}^{[j]}$  to  $\hat{\mathbf{u}}$ , and considering  $\mathbf{D}_{1 \times N}^{[j]}$  as the global expansions of  $\mathbf{D}_{1 \times n_j}^{[j]}$ . For more details on space-time and localized RBFs collocation method, see [8, 9, 10].

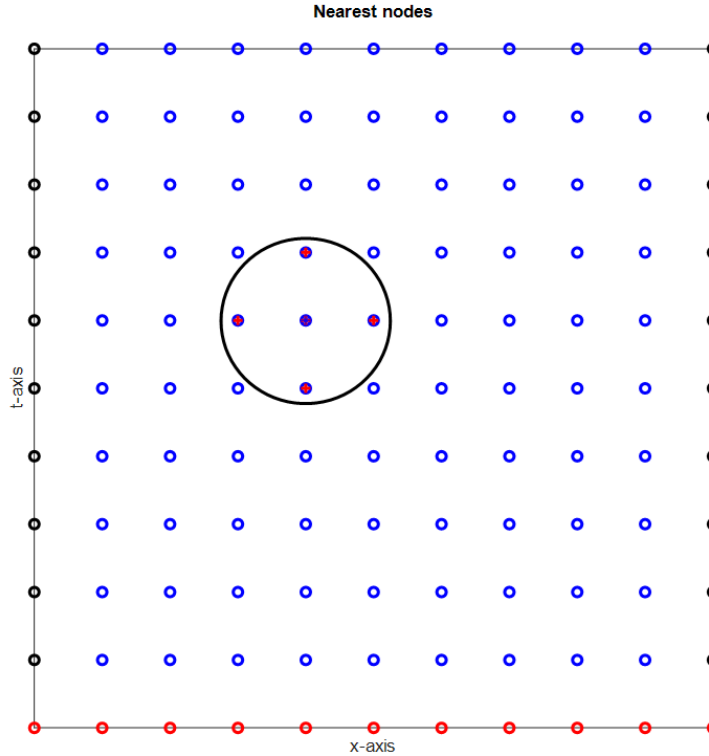


Figure 2: Nearest nodes

### 2.3 The trapezoidal rule on space-time

In the space-time domain, the integral part of (1) can be achieved by the trapezoidal rule. First, for each node  $\hat{x}_i = (x_i, t_i)$ , we determine the in-

tegration nodes in a space-time influence domain  $I_T^i$  having  $p_i$  elements. With uniform distribution nodes, we have  $I_T^i = \{(x_i, t_i), (x_i, t_i - h_t), (x_i, t_i - 2h_t), \dots, (x_i, 0)\}$  and  $h_t = \frac{T}{N_t}$ , where  $N_t$  is the number of nodes on the time axis. Then the integral can be calculated by the trapezoidal rule. The integration nodes are shown in Figure 3.

We have

$$\mathcal{I}_{\hat{x}} u = \int_0^t k(x, t, s) u(x, s) ds. \quad (11)$$

Then, the discretization of  $\mathcal{I}_{\hat{x}} u$  is done as follows:

$$\begin{aligned} \mathcal{I}_{\hat{x}} u(\hat{x}_i) &\approx \hat{\mathcal{I}}_{\hat{x}} u_i \\ &= \frac{h_t}{2} k(x_i, t_i, 0) u(x_i, 0) + h_t \sum_{k=1}^{p_i-1} k(x_i, t_i, t_k) u(x_i, t_k) \\ &\quad + \frac{h_t}{2} k(x_i, t_i, t_i) u(x_i, t_i) \\ &= \frac{h_t}{2} k(x_i, t_i, 0) u_1^{[i]} + h_t \sum_{k=1}^{p_i-1} k(x_i, t_i, t_k) u_k^{[i]} + \frac{h_t}{2} k(x_i, t_i, t_i) u_{p_i}^{[i]}. \end{aligned} \quad (12)$$

In reduced form, we have

$$\hat{\mathcal{I}}_{\hat{x}} u_i = \mathbf{I}^{[i]} \cdot \hat{\mathbf{u}}^{[i]}, \quad (13)$$

where  $\mathbf{I}^{[i]} = [\frac{h_t}{2} k(x_i, t_i, 0), \{h_t k(x_i, t_i, t_k)\}_{1 \leq k \leq p_i-1}, \frac{h_t}{2} k(x_i, t_i, t_i)]$  and  $\hat{\mathbf{u}}^{[i]} = [u(\hat{x}_0^{[i]}), u(\hat{x}_1^{[i]}), \dots, u(\hat{x}_{p_i}^{[i]})]$ .

Similar to the localized scheme, the vector  $\hat{\mathbf{u}} = [u(\hat{x}_1), u(\hat{x}_2), \dots, u(\hat{x}_N)]$  is incorporated in (13) by adding zeros at the proper locations based on the mapping of  $\hat{\mathbf{u}}^{[i]}$  to  $\hat{\mathbf{u}}$ , and considering  $\mathbf{I}_{1 \times N}$  as the global expansions of  $\mathbf{I}_{1 \times p_i}^{[i]}$ .

## 2.4 The combined scheme

For each node  $\hat{x}_i$  in the space-time domain  $\Omega_T$ , we determine the nearest nodes  $\{\hat{x}_k \in \Omega_T^i\}$  and the integration nodes  $\{\hat{x}_k \in \mathbf{I}^{[i]}\}$ . Then we compute the coefficients of  $\hat{\mathbf{u}}$  by

$$\begin{cases} \mathcal{D}_{\hat{x}} \hat{u}_i = \mathbf{D}^{[i]} \cdot \hat{\mathbf{u}}, \\ \hat{\mathcal{I}}_{\hat{x}} u_i = \mathbf{I}^{[i]} \cdot \hat{\mathbf{u}}, \end{cases} \quad (14)$$

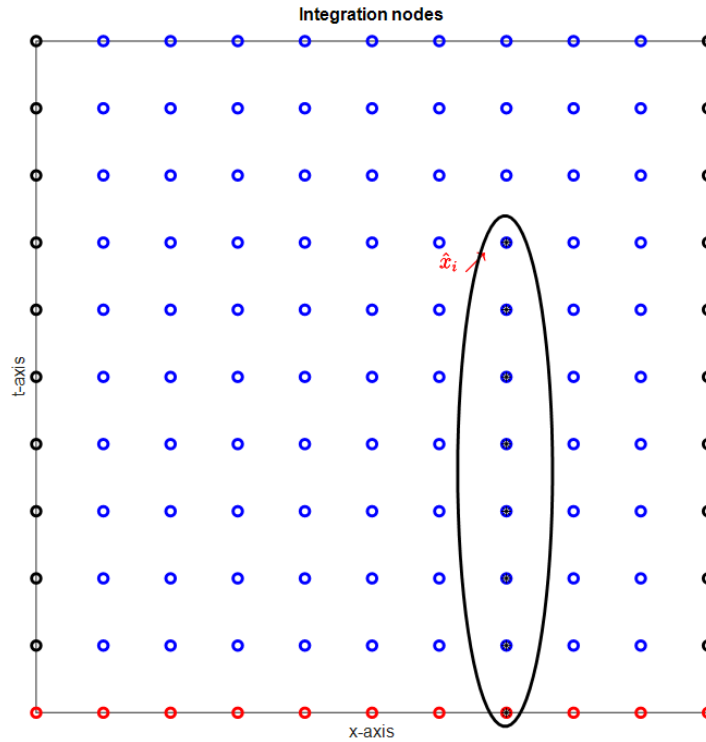


Figure 3: Trapezoidal rule nodes

and we get

$$\begin{cases} \mathcal{D}_{\hat{x}} \hat{u}_i + \hat{\mathcal{I}}_{\hat{x}} u_i = \mathbf{D}^{[i]} \cdot \hat{\mathbf{u}} + \mathbf{I}^{[i]} \cdot \hat{\mathbf{u}} \\ \quad \quad \quad = (\mathbf{D}^{[i]} + \mathbf{I}^{[i]}) \cdot \hat{\mathbf{u}}. \end{cases} \quad (15)$$

By substituting (15) into (5) for  $x_i \in \Omega_T$ , we obtain

$$f(\hat{x}_i) = \mathcal{D}_{\hat{x}} \hat{u}_i + \hat{\mathcal{I}}_{\hat{x}} u_i = (\mathbf{D}^{[i]} + \mathbf{I}^{[i]}) \cdot \hat{\mathbf{u}}. \quad (16)$$

For  $x_i \in \partial\Omega_T$ , we have

$$g(\hat{x}_i) = \hat{\mathbf{u}}_i. \quad (17)$$

By collocating all the interpolation points  $\{x_j\}_{j=1}^N$  and using (16) and (17), we get the following sparse linear system:

$$\mathcal{A}\mathbf{U} = \mathcal{B}, \quad (18)$$

$$\text{where } \mathcal{A} = \begin{bmatrix} (\mathbf{D} + \mathbf{I})(\hat{x}_1) \\ (\mathbf{D} + \mathbf{I})(\hat{x}_2) \\ \vdots \\ (\mathbf{D} + \mathbf{I})(\hat{x}_{N_i}) \\ \mathbb{1}_{N_i+1} \\ \vdots \\ \mathbb{1}_N \end{bmatrix}, \mathbf{U} = \begin{bmatrix} \hat{u}_1 \\ \hat{u}_2 \\ \vdots \\ \hat{u}_{N_i} \\ \hat{u}_{N_i+1} \\ \vdots \\ \hat{u}_N \end{bmatrix}, \text{ and } \mathcal{B} = \begin{bmatrix} f(\hat{x}_1) \\ f(\hat{x}_2) \\ \vdots \\ f(\hat{x}_{N_i}) \\ g(\hat{x}_{N_i+1}) \\ \vdots \\ g(\hat{x}_N) \end{bmatrix}, \text{ where}$$

$$\mathbb{1}_j \cdot \hat{u} = \hat{u}_j, \quad \text{for all } j \in [N_i + 1, N].$$

Note that the linear algebraic system (18) is square since the number of unknowns (the values of the approximate function) and the collocation points are equal. The approximate solution  $\{\hat{u}(\hat{x}_j)\}_{j=1}^N$  at the interpolation points  $\{\hat{x}_j\}_{j=1}^N$  can be obtained by solving the above sparse linear system of equations.

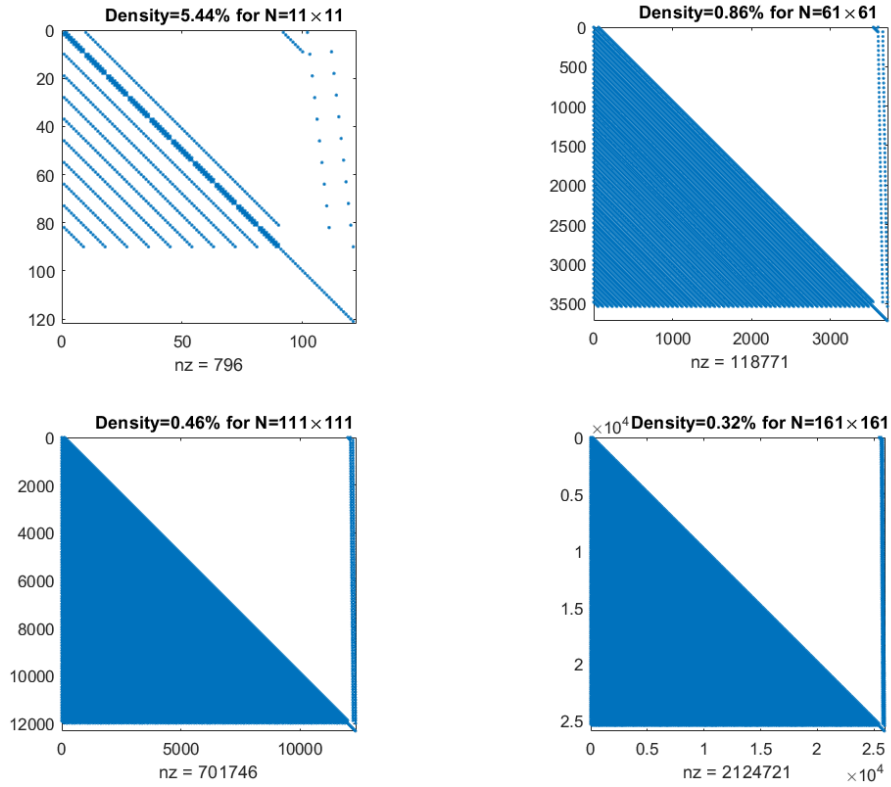
In practice, the mapping from  $\mathbf{D}^{[i]}$  to  $\mathbf{D}$  and  $\mathbf{I}^{[i]}$  to  $\mathbf{I}$  is automatic without the need for inserting zeros, if we make good use of the index vector and store the sparse matrix properly. Figure 4 shows the sparse matrix of the system for some values of  $N$ . The **nz** value designates the nonzero elements in the matrix.

## 2.5 Rate of convergence of the scheme

It is well known that the trapezoidal rule has a quadratic accuracy, and we [8] demonstrated numerically that the rate of the localized space-time also has a quadratic convergence. So we can assert that the proposed scheme in this work also has a numerical quadratic rate of convergence.

The experimental rates of convergence with respect to the mesh size  $h = \sup_{\hat{x} \in \Omega_T} \min_{\hat{x}_j} \|\hat{x} - \hat{x}_j\|$  are calculated using the following formula:

$$ROC = \frac{\log\left(\frac{E_{i+1}}{E_i}\right)}{\log\left(\frac{h_{i+1}}{h_i}\right)},$$

Figure 4: Density of the sparse matrix for some values of  $N$ 

where  $E_i$  is one of the specified errors  $MAE$ ,  $RMSE$ , or  $L_{er}^1$  corresponding to the mesh size  $h_i$ .

### 3 Numerical results and discussions

In this section, we investigate the numerical solution of the PIDE using a spacetime localized RBF collocation method to show its efficiency and accuracy for solving such a problem.

To measure the numerical accuracy, we consider the maximum absolute error (MAE), the root mean squared error (RMSE), and the  $L_{er}^1$  relative error defined as follows:

$$\begin{aligned}
 MAE &= \max_{1 \leq j \leq N} |\hat{u}(\hat{x}_j) - u(\hat{x}_j)|, \\
 RMSE &= \sqrt{\frac{1}{N} \sum_{j=1}^N (\hat{u}(\hat{x}_j) - u(\hat{x}_j))^2}, \\
 L_{er}^1 &= \frac{\sum_{j=1}^N |\hat{u}(\hat{x}_j) - u(\hat{x}_j)|}{\sum_{j=1}^N |u(\hat{x}_j)|},
 \end{aligned} \tag{19}$$

where  $u(\hat{x}_j)$  and  $\hat{u}(\hat{x}_j)$  are the exact and the approximate solutions at the node  $\hat{x}_j$ , respectively. We considered the maximum absolute error to show that there is no big error. The RMSE is more significant; it measures the average magnitude of the error. Moreover,  $L_{er}^1$  shows that even for “big” errors, it is relatively small. Since the norms are equivalent, RMSE suffices.

For all treated examples, the uniform node distribution is adopted. The choice of the shape parameter is not discussed, and it is fixed at  $\epsilon = 1$ . The number of nearest nodes is chosen  $n_j = 5$  as the problems treated are in two-dimensional space-time domains [8].

### 3.1 Example 1

The first example treated is a diffusion integro-differential problem of the form:

$$\frac{\partial u}{\partial t}(x, t) - \frac{\partial^2 u}{\partial x^2}(x, t) - \int_0^t k(t, s)u(x, s)ds = f(x, t) \text{ on } \Omega_T = (0, 1)^2,$$

where  $k(t, s) = st$ . The function  $f$  and the boundary conditions on space-time domain are chosen according to the analytical solution:

$$u(x, t) = \sin(\pi x)e^{-\pi^2 t}.$$

Table 1 shows errors for different values of  $N$ , the total interpolation points in the space-time domain. The CPU time and matrix size are also given.

The experimental rate of convergence in this simulation shows that nearly quadratic convergence is achieved. Figure 5 shows the exact and numerical solution and the absolute error on the entire domain  $\Omega_T = (0, 1)^2$  with  $N = 101^2$ . It can be noted that the results have good accuracy.

Table 1: Errors for Example 3.1 for some values of  $N = N_x \times N_t$

$N$	$h$	MAE	ROC	RMSE	ROC	$L_{er}^1$	ROC	Time	Size
$11^2$	0.100	3.76E-02	0.000	1.05E-02	0.000	6.40E-02	0.000	0.010	$121^2$
$21^2$	0.050	1.12E-02	1.751	3.27E-03	1.686	2.39E-02	1.418	0.016	$441^2$
$31^2$	0.033	5.15E-03	1.904	1.53E-03	1.872	1.20E-02	1.709	0.031	$961^2$
$41^2$	0.025	2.94E-03	1.950	8.80E-04	1.926	7.11E-03	1.812	0.094	$1681^2$
$51^2$	0.020	1.90E-03	1.970	5.69E-04	1.951	4.69E-03	1.862	0.125	$2601^2$
$61^2$	0.017	1.32E-03	1.980	3.98E-04	1.964	3.32E-03	1.891	0.250	$3721^2$
$71^2$	0.014	9.73E-04	1.985	2.94E-04	1.972	2.48E-03	1.910	0.422	$5041^2$
$81^2$	0.013	7.46E-04	1.989	2.26E-04	1.977	1.92E-03	1.924	0.563	$6561^2$
$91^2$	0.011	5.90E-04	1.991	1.79E-04	1.980	1.53E-03	1.934	0.766	$8281^2$
$101^2$	0.010	4.78E-04	1.994	1.45E-04	1.984	1.24E-03	1.942	0.969	$10201^2$
$111^2$	0.009	3.95E-04	1.997	1.20E-04	1.988	1.03E-03	1.950	1.328	$12321^2$
$121^2$	0.008	3.33E-04	1.988	1.01E-04	1.980	8.71E-04	1.946	1.703	$14641^2$
$131^2$	0.008	2.83E-04	2.008	8.60E-05	2.001	7.44E-04	1.969	2.203	$17161^2$
$141^2$	0.007	2.44E-04	1.982	7.43E-05	1.975	6.44E-04	1.946	2.719	$19881^2$
$151^2$	0.007	2.13E-04	2.006	6.47E-05	1.999	5.62E-04	1.972	3.391	$22801^2$
$161^2$	0.006	1.87E-04	2.015	5.68E-05	2.009	4.95E-04	1.984	4.109	$25921^2$

### 3.2 Example 2

As a second example, the following advection-diffusion integro-differential equation is considered [14]:

$$\frac{\partial u}{\partial t}(x, t) - a \frac{\partial^2 u}{\partial x^2}(x, t) + b \frac{\partial u}{\partial x}(x, t) - \int_0^t k(t, s) u(x, s) ds = f(x, t) \text{ on } \Omega_T = (0, 1)^2,$$

where  $a = 0.4$ ,  $b = 0.05$ , and  $k(t, s) = \sqrt{t - s}$ . The other parameters are taken according to the analytical solution:

$$u(x, t) = (t^2 + 1) \sin(\pi x).$$

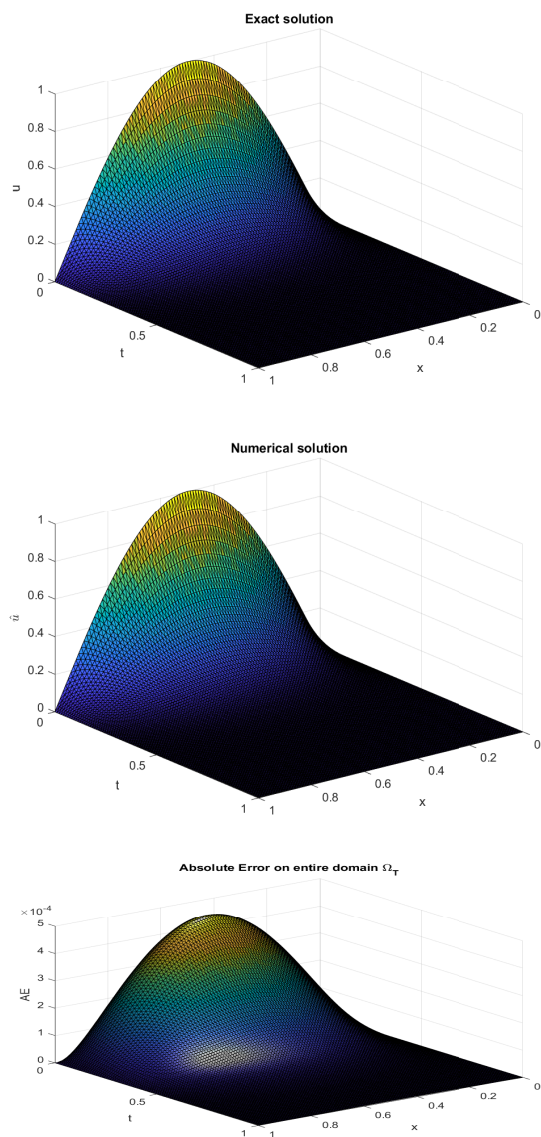


Figure 5: Exact and numerical solutions and the absolute error for the example 3.1 with  $N = 101^2$ .

Table 2 shows the results of some simulations for Example 3.2. Same remarks are noted as in Example 3.1 concerning the accuracy and the rate of convergence.

Table 2: Errors for Example 3.2 for some values of  $N$ 

$N$	$h$	$MAE$	ROC	$RMSE$	ROC	$L_{er}^1$	ROC
$21^2$	0.0500	8.889E-03		2.25E-03		1.69E-03	
$41^2$	0.0250	2.474E-03	1.84508	6.01E-04	1.90682	4.49E-04	1.91278
$61^2$	0.0167	1.157E-03	1.87405	2.79E-04	1.89447	2.09E-04	1.88484
$81^2$	0.0125	6.740E-04	1.87914	1.62E-04	1.88032	1.22E-04	1.86126
$101^2$	0.0100	4.432E-04	1.87820	1.07E-04	1.86698	8.11E-05	1.83983
$121^2$	0.0083	3.148E-04	1.87711	7.63E-05	1.85763	5.82E-05	1.82456
$141^2$	0.0071	2.361E-04	1.86681	5.74E-05	1.83785	4.41E-05	1.79623
$161^2$	0.0063	1.842E-04	1.85918	4.50E-05	1.82052	3.48E-05	1.77233

### 3.3 Example 3

For this third example, more challenging problem is treated. The advection-diffusion-reaction integro-differential equation with variable coefficients is defined by

$$\begin{aligned} & \frac{\partial u}{\partial t}(x, t) - a(x, t) \frac{\partial^2 u}{\partial x^2}(x, t) + b(x, t) \frac{\partial u}{\partial x}(x, t) + c(x, t)u(x, t) \\ & - \int_0^t k(t, s)u(x, s)ds = f(x, t), \end{aligned}$$

where  $a(x, t) = e^{-t}$ ,  $b = \sin(x)$ ,  $c(x, t) = e^t$ , and  $k(t, s) = (t - s)$ . The other parameters are chosen according to the analytical solution:

$$u(x, t) = e^{-t} \cos(x)$$

and the space-time domain is  $\Omega_T = [0, \frac{\pi}{2}] \times [0, 1]$ .

We can observe from Table 3 that even for this kind of complex problem, the results are very accurate. The same rate of convergence is observed. Because we arrived at  $N = 201^2 = 40401$  nodes and the size of the matrix  $\mathcal{A}$  defined in (18) is  $40401 \times 40401$ , without losing accuracy, we can assert that the scheme is also stable.

Table 3: Errors for Example 3.3 for some values of  $N$

$N$	$h$	$MAE$	ROC	$RMSE$	ROC	$L_{er}^1$	ROC
$11^2$	0.1000	3.451E-03		1.65E-03		3.19E-03	
$21^2$	0.0500	9.272E-04	1.89609	4.65E-04	1.82568	9.47E-04	1.75046
$31^2$	0.0333	4.195E-04	1.95598	2.13E-04	1.92592	4.40E-04	1.88872
$41^2$	0.0250	2.374E-04	1.97869	1.21E-04	1.95451	2.53E-04	1.92983
$51^2$	0.0200	1.522E-04	1.99211	7.82E-05	1.96759	1.64E-04	1.94926
$61^2$	0.0167	1.060E-04	1.98399	5.46E-05	1.97497	1.14E-04	1.96039
$71^2$	0.0143	7.794E-05	1.99569	4.02E-05	1.97968	8.45E-05	1.96754
$81^2$	0.0125	5.972E-05	1.99384	3.09E-05	1.98292	6.49E-05	1.97237
$91^2$	0.0111	4.721E-05	1.99551	2.44E-05	1.98531	5.15E-05	1.97605
$101^2$	0.0100	3.824E-05	1.99960	1.98E-05	1.98753	4.18E-05	1.97932
$111^2$	0.0091	3.163E-05	1.99372	1.64E-05	1.98745	3.46E-05	1.98002
$121^2$	0.0083	2.657E-05	2.00062	1.38E-05	1.98999	2.91E-05	1.98332
$131^2$	0.0077	2.265E-05	1.99703	1.18E-05	1.99127	2.48E-05	1.98521
$141^2$	0.0071	1.952E-05	2.00184	1.01E-05	1.99476	2.14E-05	1.98928
$151^2$	0.0067	1.701E-05	1.99955	8.84E-06	1.99140	1.87E-05	1.98606
$161^2$	0.0063	1.495E-05	2.00023	7.77E-06	1.99566	1.64E-05	1.99084
$171^2$	0.0059	1.326E-05	1.98254	6.89E-06	1.97510	1.46E-05	1.96957
$181^2$	0.0056	1.181E-05	2.02565	6.14E-06	2.02247	1.30E-05	2.01969
$191^2$	0.0053	1.060E-05	1.98603	5.52E-06	1.98652	1.17E-05	1.98243
$201^2$	0.0050	9.585E-06	1.97164	4.99E-06	1.95884	1.06E-05	1.95271

## 4 Conclusion

In this paper, we presented a local space-time RBFs collocation method combined with the trapezoidal rule to solve the PIDE as space-time one without differentiating between space and time variables. The problem is solved once to approximate the solution at any point  $(x, t)$ . The main advantages of the considered technique are as follows:

- (1). The discussion of the time stability analysis of the discrete system is avoided.
- (2). The computational time when dealing with PIDEs with time-dependent coefficients is reduced as there is no need to recompute the matrix for the resulting algebraic system at each time level.

- (3). The method uses the sparse matrices to store only the nonzero elements, so we save a significant amount of memory and speed up the resolution of the linear system.
- (4). The formulation is the same for any form of the linear differential operator of second order  $\mathcal{D}_{\hat{x}}$  and any form of the function  $k(x, t, s)$ , only some changes in the programming script can be made according to the problem to be solved.

It has been demonstrated that our technique is simple, straightforward, and applicable to a large type of problems as it is shown in this paper. The application of the developed technique to equations with an integral boundary condition is under investigation. Further work will focus on developing a method without the trapezoidal rule to solve such a problem.

## Conflicts of Interest

We have no conflicts of interest to disclose. All authors declare that they have no conflicts of interest.

## Funding

This research received no specific grant from any funding agency in the public, commercial, or not-for-profit sectors.

## Data availability

No new data were created or analysed in this study. Data sharing is not applicable to this article.

## Acknowledgements

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

## References

- [1] Alipanah, A. and Esmaeili, S. *Numerical solution of the two-dimensional Fredholm integral equations using Gaussian radial basis function*, J. Comput. Appl. Math. 235(18) (2011) 5342–5347.
- [2] Chen, C.S., Fan, C.M. and Wen, P.H. *The method of particular solutions for solving elliptic problems with variable coefficients*, Int. J. Comput. Method. 8 (2011) 545–559.
- [3] Chen, C.S., Fan, C.M. and Wen, P.H. *The method of particular solutions for solving certain partial differential equations*, Numer. Method. Partial Differ. Equ. 28 (2012) 506–522.
- [4] El-Sayed, A.M.A., Helal, S.M. and El-Azab, M.S. *Solution of a parabolic weakly-singular partial integro-differential equation with multi-point non-local boundary conditions*, J. Fract. Calc. Appl. 7(1) (2016) 1–11.
- [5] Fornberg, B. and Flyer, N. *Accuracy of radial basis function interpolation and derivative approximations on 1-D infinite grids*, Adv. Comput. Math. 23 (2005) 5–20.
- [6] Golbabai, A. and S. Seifollahi, *Numerical solution of the second-kind integral equations using radial basis function networks*, Appl. Math. Comput. 174 (2) (2006) 877–883.
- [7] Golbabai, A. and Seifollahi, S. *Radial basis function networks in the numerical solution of linear integro-differential equations*, Appl. Math. Comput. 188 (1) (2007) 427–432.
- [8] Hamaidi, M., Naji, A. and Charafi, A. *Space-time localized radial basis function collocation method for solving parabolic and hyperbolic equations*, Eng. Anal. Bound. Elem. 67 (2016) 152–163.
- [9] Hamaidi, M., Naji, A., Ghafrani, F. and Jourhmane, M. *Noniterative localized and space-time localized RBF meshless method to solve the ill-posed and inverse problem*, Model. Simul. Eng. 2020 (2020) 5046286.

- [10] Hamaidi, M., Naji, A. and Taik, A. *Solving parabolic and hyperbolic equations with variable coefficients using space-time localized radial basis function collocation method*, Model. Simul. Eng. 2021 (2021) 6688806.
- [11] Jiang, Z.W., Wang, R.H., Zhu, C.G. and Xu, M. *High accuracy multiquadric quasi-interpolation*, Appl. Math. Model. 35 (5) (2011) 2185–2195.
- [12] Parand, K. and Rad, J.A. *Numerical solution of nonlinear Volterra-Fredholm-Hammerstein integral equations via collocation method based on radial basis functions*, Appl. Math. Comput. 218 (9) (2012) 5292–5309.
- [13] Scott, S. and Kansa, E.J. *Multiquadric radial basis function approximation methods for the numerical solution of partial differential equations*, Adv. Comput. Mech. 2(2) (2009) 220.
- [14] Siddiqi, S.S. and Arshed, S. *Numerical solution of convection-diffusion integro-differential equations with a weakly singular kernel*, J. Basic Appl. Sci. Res. 3(11) (2013) 106–120.
- [15] Soliman, A.F., El-Asyed, A.M.A. and El-Azab, M.S. *On the numerical solution of partial integro-differential equations*, Math. Sci. Lett. 1(1) (2012) 71–80.
- [16] Soliman, A.F., El-Azab, M.S. and El-Asyed, A. *Fourth and sixth order compact finite difference Schemes for partial integro-differential equations*, J. Math. Comput. Sci. 2 (2) (2013) 206–225.
- [17] Zhang, H.Q., Chen, Y., and Nie, X. *Solving the linear integral equations based on radial basis function interpolation*, J. Appl. Math. 2014 (2014), 793582.



# A study on the convergence and error bound of solutions to 2D mixed Volterra–Fredholm integral and integro-differential equations via high-order collocation method

A.A. Shalangwa\*, M.R. Odekunle and S.O. Adee

## Abstract

The integral equation is transformed into systems of algebraic equations using standard collocation points, and then the algebraic equations are solved using matrix inversion. Their solutions are substituted into the approximate equation to give the numerical results. We establish the analysis of

---

\*Corresponding author

Received 29 October 2024; revised 26 January 2025; accepted 26 January 2025

Ayuba Albert Shalangwa

Department of Mathematical science, Gombe State University, Nigeria. e-mail: draashalangwa2@gmail.com

M.R. Odekunle

Department of Mathematics, Modibbo Adama University Yola, Nigeria.

S.O. Adee

Department of Mathematics, Modibbo Adama University Yola, Nigeria.

## How to cite this article

Shalangwa, A.A., Odekunle, M.R. and Adee, S.O., A study on the convergence and error bound of solutions to 2D mixed Volterra–Fredholm integral and integro-differential equations via high-order collocation method. *Iran. J. Numer. Anal. Optim.*, 2025; 15(3): 1012-1035. <https://doi.org/10.22067/ijnao.2025.90535.1544>

the developed method, which shows that the solution is unique, convergent, and error bound. To illustrate the effectiveness, ease of use, and dependability of the approach, illustrative examples are provided. It demonstrates that the method outperforms other methods.

**AMS subject classifications (2020):** Primary 45A05; Secondary 65R20.

**Keywords:** Volterra integral equation; Fredholm integral equation; Mixed Volterra–Fredholm integral equation; Collocation; Two-dimensional integral.

## 1 Introduction

Due to certain scientists' inability to solve differential equations, integral equations first appeared in writing in the middle of the seventeenth century. The numerous applications of integral equations can be found in the fields of elasticity, plasticity, heat and mass transfer, fluid dynamics, filtration theory, electrostatics, electrodynamics, bio-mechanics, game theory, control, queuing theory, electrical engineering, economics, and medicine, among other scientific disciplines. In many branches of natural science, exact (closed-form) solutions to integral equations are essential to comprehending the qualitative aspects of numerous processes and occurrences [13].

The integral equations provide a significant tool for describing diverse processes and for solving several sorts of boundary value issues relating to ordinary and partial differential equations. The topic of integral equations is one of the most useful mathematical tools in both pure and practical mathematics and it has vast applications in a variety of scientific situations.

Two-dimensional integral equations provide an important tool for modeling several problems in engineering and research [5, 8]. Many processes in physics and engineering domains give rise to two-dimensional integral equations and are frequently difficult to solve analytically. In many circumstances, it is needed to find the approximate solutions. As we know, substantial effort has been done on creating and studying numerical methods for solving one-dimensional integral equations of the second sort, but in two-dimensional cases, a very little amount of work has been done [19].

An equation is considered integral if the unknown function appears inside the integral sign. The various forms of integral equations primarily depend on the equation's kernel and the integration's limits. According to [19], an integral equation is referred to as a Volterra integral equation if at least one of the limits is variable and a Fredholm integral equation if the limits of integration are fixed. The Fredholm integral equation is characterized by fixed integration limits, whereas the Volterra integral equation exhibits at least one variable integration limit.

An essential tool for modeling a wide range of phenomena and resolving various boundary value issues involving ordinary and partial differential equations is the integral equation. One of the most helpful mathematical fields in both pure and applied mathematics is integral equations, which has numerous applications in science, engineering, and so on [11]. An equation that combines the Fredholm integral and the Volterra integral in one equation is known as the Volterra–Fredholm integral equation.

Numerous methods have been developed for solving one-dimensional integral equations and two-dimensional mixed Volterra–Fredholm integral equations (2D MVFIEs). These methods include perturbed collocation method [18], collocation method [2] and [3], boukakar collocation method [1] and [1], multiquadric radial basis functions [4], Two-dimensional Legendre wavelets method [6], applications of two-dimensional triangular functions [12], series solution methods [15], successive approximation method and method of successive substitutions [16], and Adomian decomposition method [17]. In this study, we develop the polynomial collocation method to solve 2D MVFIE of the form:

$$m(x, t) = h(x, t) + \rho \int_0^t \int_a^b k(x, t, y, z) m(y, z) dy dz \quad (1)$$

and

$$m^n(x, t) = h(x, t) + \rho \int_0^t \int_a^b k(x, t, y, z) m(y, z) dy dz, \quad (2)$$

where  $m(x, t)$  is considered an unknown function to be determined, the functions  $h(x, t)$  is analytic on  $C([0, 1]^2, \mathbb{R})$ ,  $k(x, t, y, z)$  is analytic on  $C([0, 1]^4, \mathbb{R})$ ,  $m(y, z)$  is a continuous function with respect to  $m(y, z)$ , and  $\rho$  is a constant coefficient.

**Definition 1.** In order to apply the Bernstein polynomials in the interval  $[0, 1]$ ,  $B_{i,n}(x)$  is defined as [10]

$$B_{i,n}(x) = \binom{n}{i} x^i (1-x)^{n-i}, \quad i = 0, 1, 2, \dots, n. \quad (3)$$

**Definition 2.** Bernstein polynomials of degree  $n$  in the interval  $[0, 1]$  can also be written in the following equivalent form:

$$B_{i,n}(x) = \sum_{p=0}^{n-i} \binom{n}{i} \binom{n-i}{p} (-1)^p x^{i+p}. \quad (4)$$

**Definition 3.** Bernstein polynomials of degree  $n$  can be defined recursively by blending together two Bernstein polynomials of degree  $n-1$ . That is, the  $k$ th  $n$ th-degree Bernstein polynomial can be written as

$$B_{k,n}(x) = (1-x)B_{k,n-1}(x) + xB_{k-1,n-1}(x), \quad k = 0(1)n, \quad n \geq 1. \quad (5)$$

**Definition 4** (Standard Collocation Method (SCM)). This method is used to determine the desired collocation points within an interval, that is,  $[a, b]$  and is given by

$$\begin{aligned} x_i &= a + \frac{(b-a)i}{N}, \quad i = 0(1)N, \\ t_j &= a + \frac{(b-a)j}{N}, \quad j = 0(1)N. \end{aligned} \quad (6)$$

**Definition 5.**

(i) **Lipschitzian** [7]

Let  $(X, \|\cdot\|)$  be a norm space. Mapping  $T : X \rightarrow X$  is L-Lipschitz if there exists  $L > 0$  such that  $\|Tx - Ty\|_\infty \leq L \|x - y\|_\infty, q \in [0, 1]$  for all  $x, y \in X$ .

(ii) **Lipschitz continuity** [14]

A function  $f$  is Lipschitz continuous if there exists  $K < \infty$  such that  $\|f(y) - f(x)\| \leq K \|y - x\|$ .

**Definition 6** (Infinity norm  $\|v\|_\infty$ ). [14]

The infinity norm (also known as the  $L_\infty$ -norm,  $l_\infty$ , max norm, or uniform

norm) of a vector  $v$  is denoted by  $\|v\|_\infty$  and is defined as the maximum of the absolute values of its components, that is,

$$\|v\|_\infty = \max \{|v_i| : i = 1, 2, \dots, n\}$$

## 2 Uniqueness, convergence, error analysis and method of solution

2D MVFIEs can be solved numerically using the polynomial collocation method, which is based on the collocation approach and takes into account the linear combination of the Bernstein polynomial as our approximated solution. In this section, we will develop a method by using standard collocation points to reduce the 2D MVFIE to a system of algebraic equations.

### 2.1 Integral form

Let  $M_N(x, t)$  be the approximate solution of

$$m^n(x, t) = h(x, t) + \rho \int_0^t \int_a^b k(x, t, y, z) m(y, z) dy dz, \quad (7)$$

with initial condition given as  $m^{n-1}(x_0, t) = m_{n-1}$ , where  $m^n(x, t) = \frac{d^n}{dx^n} m(x, t)$  is the  $n$ th order derivative of  $m(x, t)$ ,  $m(x, t)$  is an unknown function to be determined,  $h(x, t)$  and  $k(x, t, y, z)$  are analytic function on  $[a, b]$ .

Here,  $L$  is an operator defined as  $L = \frac{d^n}{dx^n}$  and  $L^{-1} = \int_0^x \int_0^x \dots \int_0^x dx dx \dots dx$  operating  $L^{-1}$  on both sides of (7) is given by

$$L^{-1}(m^n(x, t)) = L^{-1}(h(x, t)) + L^{-1} \left( \rho \int_0^t \int_a^b k(x, t, y, z) m(y, z) dy dz \right). \quad (8)$$

Integrating (7)  $n$  times from 0 to  $x$  gives

$$\begin{aligned} & \int_0^x \int_0^x \dots \int_0^x m^n(x, t) dx dx \dots dx \\ &= \int_0^x \int_0^x \dots \int_0^x h(x, t) dx dx \dots dx \end{aligned}$$

$$+ \int_0^x \int_0^x \cdots \int_0^x \left( \rho \int_0^t \int_a^b k(x, t, y, z) m(y, z) dy dz \right) dx dx \dots dx \quad (9)$$

$$\begin{aligned} & \int_0^x \int_0^x \cdots \int_0^x m^n(x, t) dx dx \dots dx \\ &= \int_0^x \int_0^x \cdots \int_0^x h(x, t) dx dx \dots dx \\ &+ \rho \int_0^x \int_0^x \cdots \int_0^x \left( \int_0^t \int_a^b k(x, t, y, z) m(y, z) dy dz \right) dx dx \dots dx. \end{aligned} \quad (10)$$

Converting multiple integrals to single integral from (10) gives

$$\begin{aligned} m(x, t) &= \frac{x^{n-1}}{(n-1)!} u_0 + \frac{x^{n-2}}{(n-2)!} u_1 + \frac{x^{n-3}}{(n-3)!} u_2 \\ &+ \cdots + u_{n-1} + \frac{1}{(n-1)!} \int_0^x (x-t)^{n-1} h(x, t) dt \\ &+ \rho \frac{1}{(n-1)!} \int_0^x (x-t)^{n-1} \left( \int_0^t \int_a^b k(x, t, y, z) m(y, z) dy dz \right) dt \end{aligned} \quad (11)$$

Simplifying (11) gives

$$\begin{aligned} m(x, t) &= \sum_{i=1}^{n-1} \frac{1}{i!} u_i x^i + \frac{1}{(n-1)!} \int_0^x (x-t)^{n-1} h(x, t) dt \\ &+ \rho \frac{1}{(n-1)!} \int_0^x (x-t)^{n-1} \left( \int_0^t \int_a^b k(x, t, y, z) m(y, z) dy dz \right) dt, \end{aligned} \quad (12)$$

where

$$\begin{aligned} H(x, t) &= \frac{x^{n-1}}{(n-1)!} u_0 + \frac{x^{n-2}}{(n-2)!} u_1 + \frac{x^{n-3}}{(n-3)!} u_2 \\ &+ \cdots + u_{n-1} + \frac{1}{(n-1)!} \int_0^x (x-t)^{n-1} h(x, t) dt \end{aligned} \quad (13)$$

or

$$H(x, t) = \sum_{i=1}^{n-1} \frac{1}{i!} u_i x^i + \frac{1}{(n-1)!} \int_0^x (x-t)^{n-1} h(x, t) dt$$

and

$$\rho(x, t) = \frac{\rho}{(n-1)!} \int_0^x (x-t)^{n-1} dt, \quad (14)$$

$$m(x, t) = H(x, t) + \rho \int_0^t \int_a^b K(x, t, y, z) m(y, z) dy dz. \quad (15)$$

Therefore, (15) is a 2D MVFIE of the second kind, which is the integral form of (2).

## 2.2 Method of solution to 2D MVFIE

We recall that (1) and (2) can be written as

$$m(x, t) = h(x, t) + \lambda \int_0^t \int_a^b k(x, t, y, z) m(y, z) dy dz. \quad (16)$$

Let  $M_N(x, t)$  be the approximate solution to (15), where

$$m_N(x, t) = \sum_{i=0}^N \sum_{j=0}^N c_{i,j} B_{i,N}(x) B_{j,N}(t) = \phi(x, t) C. \quad (17)$$

Substituting (17) into (16) gives

$$\phi(x, t) C = h(x, t) + \rho \int_0^t \int_a^b k(x, t, y, z) (\phi(y, z) C) dy dz, \quad (18)$$

$$\phi(x, t) C - \rho \int_0^t \int_a^b k(x, t, y, z) (\phi(y, z) C) dy dz = h(x, t), \quad (19)$$

$$\left\{ \phi(x, t) - \rho \int_0^t \int_a^b k(x, t, y, z) \phi(y, z) dy dz \right\} C = h(x, t). \quad (20)$$

Collocating (20) and using standard collocation points at  $x = x_i$  and  $t = t_j$  with

$$\begin{aligned} x_i &= a + \frac{(b-a)i}{N}, \quad i = 0(1)N, \\ t_j &= a + \frac{(b-a)j}{N}, \quad j = 0(1)N, \end{aligned} \quad (21)$$

we have

$$\left\{ \phi(x_i, t_j) - \rho \int_0^t \int_a^b k(x_i, t_j, y, z) \phi(y, z) dy dz \right\} C = h(x_i, t_j), \quad (22)$$

where  $\gamma(x_i, t_j) = \left\{ \phi(x_i, t_j) - \rho \int_0^t \int_a^b k(x_i, t_j, y, z) \phi(y, z) dy dz \right\}$  and  $C = [c_{0,0}, c_{0,1}, c_{0,2}, \dots, c_{0,N}, \dots, c_{N,0}, c_{N,1}, c_{N,2}, \dots, c_{N,N}]$ ,

$$\gamma(x_i, t_j) C = h(x_i, t_j). \quad (23)$$

Multiplying both sides of (23) by  $\gamma(x_i, t_j)^{-1}$  gives

$$C = \gamma(x_i, t_j)^{-1} h(x_i, t_j). \quad (24)$$

Substituting  $C$  into the approximate solution to (17) gives

$$M_N(x, t) = \phi(x, t) \gamma(x_i, t_j)^{-1} h(x_i, t_j), \quad i, j = 0(1)N. \quad (25)$$

The system of equations is then solved using Maple 18 software and the unknown constants obtained are then substituted back into the approximate solution to get the required solution.

## 2.3 Uniqueness, convergence and error analysis

Hypothesis

The following assumptions were made:

$Z_1$ : Let  $(C([0, 1] \times [0, 1]), \|\cdot\|)$  be the space of all continuous functions on the interval  $[0, 1] \times [0, 1]$  with the norm  $\|M\|_\infty = \underbrace{\max_{\substack{x \in [0, 1] \\ t \in [0, 1]}} |M(x, t)|}$ .

$Z_2$  :  $M(x, t) \neq 0$  .

$Z_3$  :  $|K(x, t, y, z)| \leq L$  ( $L$  is a positive real number) for all  $(x, t) \in [0, 1] \times [0, 1]$ , and

$Z_4$  : for all  $(x, t) \in [0, 1] \times [0, 1]$  and  $\beta = \{(x, t, y, z) : 0 \leq z \leq t \leq 1; 0 \leq y \leq x \leq 1\}$ .

With this conditions, we present the uniqueness and convergence of the solution.

**Theorem 1** (Uniqueness of solution for 2D MVFIE). Let  $M(x, t)$  be an exact solution to (1), and let  $M_{N,N}(x, t)$  be the approximate solution to (1),

where

$$M_{N,N}(x, t) = \sum_{i=0}^N \sum_{j=0}^N c_{i,j} B_{i,N}(x) B_{j,N}(t).$$

Then (1) has a unique solution whenever  $0 \leq \alpha \leq 1$  and  $\alpha = 1 - L_1 \lambda (b - a) t$ .

*Proof.* Equation (1) can be written in the form

$$M(x, t) = h(x, t) + \rho \int_0^t \int_a^b F(x, t, y, z, M(y, z)) dy dz$$

such that the linear term  $F(M)$  is Lipschitz continuous with  $|F(M) - F(V)| \leq L_1 |M - V|$ .

Let  $M_{N,N}$  and  $M'_{N,N}$  be any two different approximate solutions to (1).

Then

$$\begin{aligned} M_{N,N}(x, t) - M'_{N,N}(x, t) &= h(x, t) + \rho \int_0^t \int_a^b F(x, t, y, z, m_{N,N}(y, z)) dy dz - h(x, t) \\ &\quad - \rho \int_0^t \int_a^b F(x, t, y, z, M'_{N,N}(y, z)) dy dz \end{aligned}$$

$$\begin{aligned} &|M_{N,N}(x, t) - M'_{N,N}(x, t)| \\ &= \left| \rho \int_0^t \int_a^b F(x, t, y, z, M_{N,N}(y, z)) dy dz - \rho \int_0^t \int_a^b F(x, t, y, z, M'_{N,N}(y, z)) dy dz \right|, \end{aligned}$$

$$\begin{aligned} &|M_{N,N}(x, t) - M'_{N,N}(x, t)| \\ &\leq |\rho| \int_0^t \int_a^b |F(x, t, y, z, M_{N,N}(y, z)) - F(x, t, y, z, M'_{N,N}(y, z))| dy dz, \end{aligned}$$

$$|M_{N,N}(x, t) - M'_{N,N}(x, t)| \leq |\rho| \int_0^t \int_a^b |F(M_{N,N}) - F(M'_{N,N})| dy dz,$$

$$|m_{N,N} - M'_{N,N}| \leq |\rho| L_1 \int_0^t \int_a^b |M_{N,N} - M'_{N,N}| dy dz,$$

$$|M_{N,N} - M'_{N,N}| - |\rho| L_1 (b - a) t |M_{N,N} - M'_{N,N}| \leq 0,$$

$$\{1 - |\rho| L_1 (b - a) t\} |M_{N,N} - M'_{N,N}| \leq 0.$$

If  $\alpha = \{1 - |\rho| L_1 (b - a) t\}$ , then

$$\alpha |M_{N,N} - M'_{N,N}| \leq 0.$$

As  $0 \leq \alpha \leq 1$ ,  $|M_{N,N} - M'_{N,N}| = 0$ , which implies  $M_{N,N} = M'_{N,N}$ . Hence, the uniqueness proof is complete.

□

**Theorem 2** (Convergence of the method for 2D MVFIE). Let  $U(x, t)$  be an exact solution to (1), and let  $M_{N,N}(x, t)$  be the approximate solution to (1), where

$$M_{N,N}(x, t) = \sum_{i=0}^N \sum_{j=0}^N c_{i,j} B_{i,N}(x) B_{j,N}(t).$$

Then, the solution of L2D-LMVFIE by using Bernstein polynomial as a basis function is unique and convergent if  $0 \leq \eta_1 \leq 1$ .

*Proof.* Since we have already proved for the uniqueness, we now prove the convergence using the definition of norms and our assumptions  $Z_1 - Z_4$ . We have

$$\|M(x, t) - M_{N,N}(x, t)\|_{\infty} = \underbrace{\overbrace{\max_{x \in [0, 1]} |M(x, t) - M_{N,N}(x, t)|}_{t \in [0, 1]}}$$

$$\begin{aligned} & \|M(x, t) - M_{N,N}(x, t)\|_{\infty} \\ & \underbrace{\overbrace{\max_{x \in [0, 1]} }_{t \in [0, 1]}} \\ & = \left| h(x, t) + \rho \int_0^t \int_a^b k(x, t, y, z) M(y, z) dy dz - h(x, t) \right. \\ & \quad \left. - \rho \int_0^t \int_a^b k(x, t, y, z) M_{N,N}(y, z) dy dz \right|, \end{aligned}$$

$$\begin{aligned} & \|M(x, t) - M_{N,N}(x, t)\|_{\infty} \\ & \leq |\rho| \underbrace{\overbrace{\max_{x \in [0, 1]} }_{t \in [0, 1]}} \int_0^t \int_a^b |k(x, t, y, z)| |M(y, z) - M_{N,N}(y, z)| dy dz, \end{aligned}$$

$$\|M(x, t) - M_{N,N}(x, t)\|_{\infty} \leq |\rho| L\beta \|M(y, z) - m_{N,N}(y, z)\|_{\infty},$$

$$\|M(x, t) - M_{N,N}(x, t)\|_{\infty} (1 - |\rho| L\beta) \leq 0.$$

If  $\eta_1 = (|\rho| L\beta)$ , then

$$(1 - \eta_1) \|M(x, t) - M_{N,N}(x, t)\|_{\infty} \leq 0.$$

Then, if  $0 \leq \eta_1 \leq 1$  and  $N \rightarrow \infty$ , then  $\lim_{N \rightarrow \infty} \|M(x, t) - M_{N,N}(x, t)\|_{\infty} = 0$ .  $\square$

**Theorem 3** (Error bound of 2D MVFIE). Let  $U(x, t)$  be an exact solution to (1), and let  $M_{N,N}(x, t)$  be the approximate solution to (1), where

$$M_{N,N}(x, t) = \sum_{i=0}^N \sum_{j=0}^N c_{i,j} B_{i,N}(x) B_{j,N}(t).$$

Then, the error of L2D-LMVFIE by using Bernstein polynomial as a basis function is

$$\frac{\|e_{N,N}(x, t)\|_{\infty}}{\|e_{N,N}(y, z)\|_{\infty}} \leq |\rho| M_{\alpha} \beta_{\alpha}.$$

*Proof.* In establishing the error bound of this method, we substitute the approximate solution into (1), which gives

$$M_{N,N}(x, t) = h(x, t) + \rho \int_0^t \int_a^b k(x, t, y, z) M_{N,N}(y, z) dy dz,$$

and the exact solution is given by

$$M(x, t) = h(x, t) + \rho \int_0^t \int_a^b k(x, t, y, z) M(y, z) dy dz,$$

$$M_{N,N}(x, t) - M(x, t) = e_N(x, t),$$

$$\begin{aligned} M_{N,N}(x, t) - M(x, t) &= h(x, t) + \rho \int_0^t \int_a^b k(x, t, y, z) M_{N,N}(y, z) dy dz \\ &\quad - h(x, t) - \rho \int_0^t \int_a^b k(x, t, y, z) M(y, z) dy dz, \end{aligned}$$

$$|M_{N,N}(x, t) - M(x, t)| = \left| \rho \int_0^t \int_a^b k(x, t, y, z) M_{N,N}(y, z) dy dz \right.$$

$$\begin{aligned}
& \left| -\rho \int_0^t \int_a^b k(x, t, y, z) M(y, z) dy dz \right|, \\
& |M_{N,N}(x, t) - M(x, t)| \leq |\rho| \int_0^t \int_a^b |k(x, t, y, z)| |M_{N,N}(y, z) - M(y, z)| dy dz, \\
& \frac{|M_{N,N}(x, t) - M(x, t)|}{|M_{N,N}(y, z) - M(y, z)|} \leq \frac{|\rho| \int_0^t \int_a^b |k(x, t, y, z)| |M_{N,N}(y, z) - M(y, z)| dy dz}{|M_{N,N}(y, z) - M(y, z)|}, \\
& \frac{|e_{N,N}(x, t)|}{|e_{N,N}(y, z)|} \leq |\rho| \int_0^t \int_a^b |k(x, t, y, z)| dy dz, \\
& \frac{\|e_{N,N}(x, t)\|_\infty}{\|e_{N,N}(y, z)\|_\infty} \leq |\rho| \int_0^t \int_a^b |k(x, t, y, z)| dy dz, \\
& \frac{\|e_{N,N}(x, t)\|_\infty}{\|e_{N,N}(y, z)\|_\infty} \leq |\rho| M_\alpha \beta_\alpha.
\end{aligned}$$

Therefore the error is bounded and hence the solution of the method is convergent.  $\square$

**Theorem 4.** Let  $M(x, t)$  be the solution to (1). Then the solution is

$$M_N(x, t) = \phi(x, t) \gamma(x_i, t_j)^{-1} h(x_i, t_j); \quad i, j = 0(1)N,$$

where

$$\gamma(x_i, t_j) = \left\{ \phi(x_i, t_j) - \rho \int_0^t \int_a^b k(x_i, t_j, y, z) \phi(y, z) dy dz \right\}.$$

*Proof.* The approximate solution to (1) is

$$m_N(x, t) = \sum_{i=0}^N \sum_{j=0}^N c_{i,j} B_{i,N}(x) B_{j,N}(t) = \phi(x, t) C.$$

From (23)

$$C = \gamma(x_i, t_j)^{-1} h(x_i, t_j),$$

where

$$\gamma(x_i, t_j) = \left\{ \phi(x_i, t_j) - \rho \int_0^t \int_a^b k(x_i, t_j, y, z) \phi(y, z) dy dz \right\}.$$

Substituting for  $C$  in the approximate solution gives

$$M_N(x, t) = \phi(x, t)\gamma(x_i, t_j)^{-1}h(x_i, t_j), \quad i, j = 0(1)N.$$

□

### 3 Numerical examples

In this research, numerical examples are utilized to assess the simplicity and efficiency of the method and are presented in tables except where it delivers the exact solution. All computations are done with the help of the MAPLE 18 program. Let  $M_N(x, t)$  and  $M(x, t)$  be the approximate and exact solution, respectively. Then  $Error_N = |M_N(x, t) - M(x, t)|$ . Table 1 gives a brief description of some abbreviations made.

Table 1: Notations

<i>Tag</i>	<i>Description</i>
$Error_{OurMethod}$	$AbsoluteErrorofOurMethod$
$Error_{NKH}$	$AbsoluteErrorFrom[9]$
$Error_{AM}$	$AbsoluteErrorFrom[4]$

**Example 1.** Consider a linear 2D MVFIE of the second kind [9]

$$m(x, t) = x^2 + e^t + \frac{2}{3}x^3t^2 - \int_0^t \int_0^1 t^2 e^{-z} m(y, z) dy dz, \quad (26)$$

which has an exact solution given as  $m(x, t) = x^2 + e^t$  in the interval  $x, t = [0, 1]$ .

Let the approximate solution to (26) for  $N = 5$  be

$$m_N(x, t) = \sum_{i=0}^5 \sum_{j=0}^5 c_{i,j} B_{i,5}(x) B_{j,5}(t). \quad (27)$$

Substituting (27) in (26) gives

$$\sum_{i=0}^5 \sum_{j=0}^5 c_{i,j} B_{i,5}(x) B_{j,5}(t) = x^2 + e^t + \frac{2}{3}x^3t^2 \quad (28)$$

$$\begin{aligned}
& - \int_0^t \int_0^1 t^2 e^{-z} \left( \sum_{i=0}^5 \sum_{j=0}^5 c_{i,j} B_{i,5}(y) B_{j,5}(z) \right) dy dz, \\
& \sum_{i=0}^5 \sum_{j=0}^5 c_{i,j} B_{i,5}(x) B_{j,5}(t) + \int_0^t \int_0^1 t^2 e^{-z} \left( \sum_{i=0}^5 \sum_{j=0}^5 c_{i,j} B_{i,5}(y) B_{j,5}(z) \right) dy dz \\
& = x^2 + e^t + \frac{2}{3} x^3 t^2.
\end{aligned} \tag{29}$$

Collocating (29) and using standard collocation points at  $x = x_i$  and  $t = t_j$  with

$$\begin{aligned}
x_i &= \frac{i}{5}; \quad i = 0(1)5, \\
t_j &= \frac{j}{5}; \quad j = 0(1)5,
\end{aligned}$$

we have

$$\begin{aligned}
& \sum_{i=0}^5 \sum_{j=0}^5 c_{i,j} B_{i,5}(x_i) B_{j,5}(t_j) + \int_0^t \int_0^1 t_j^2 e^{-z} \left( \sum_{i=0}^5 \sum_{j=0}^5 c_{i,j} B_{i,5}(y) B_{j,5}(z) \right) dy dz \\
& = x_i^2 + e^{t_j} + \frac{2}{3} x_i^3 t_j^2.
\end{aligned} \tag{30}$$

The method was implemented using MAPLE 18 software, and  $M_5(x, t)$  was obtained as

$$\begin{aligned}
M_5(x, t) = & -2.912943530 \times 10^{-8} t^3 x + 2.240094596 \times 10^{-7} t^4 x \\
& - 1.948800244 \times 10^{-7} t^5 x + 4.595065836 \times 10^{-8} t^2 x \\
& + 2.000000000 \times 10^{-8} t x^5 + 1.000000000 \times 10^{-8} t x^3 + \\
& - 3.000000000 \times 10^{-8} t x^4 + 1.000082530 t x^2 + .49906830 t^2 x^2 \\
& + 0.4866 \times 10^{-4} t^2 x^4 - 0.4651 \times 10^{-4} t^2 x^3 + x^2 \\
& + 0.1385710011 \times 10^{-1} t^5 x^2 - 0.1187446705 \times 10^{-3} t^4 x^3 \\
& - 0.2236348902 \times 10^{-3} t^3 x^4 + 0.230 \times 10^{-5} t^2 x^5 \\
& + 0.1086523352 \times 10^{-3} t^3 x^3 - 0.1622 \times 10^{-4} t^5 x^5
\end{aligned}$$

$$\begin{aligned}
& -0.8256489016 \times 10^{-4} t^5 x^4 - 0.11620 \times 10^{-3} t^4 x^5 \\
& + 0.5474233524 \times 10^{-4} t^5 x^3 + 0.3259097803 \times 10^{-3} t^4 x^4 \\
& + 0.5115 \times 10^{-4} t^3 x^5 \\
& + 0.1704088951 t^3 x^2 + 0.3486396978 \times 10^{-1} t^4 x^2.
\end{aligned} \tag{31}$$

Table 2: Results using Bernstein polynomial for Example 1

$(x, t)$	$Exact$	$OurMethod_{N=5}$	$Error_{OurMethod}$
(0, 0)	0.0000000000	0.0000000000	0.0000000000
(0.1, 0.1)	0.01105170918	0.01105172941	$2.023 \times 10^{-8}$
(0.2, 0.2)	0.04885611032	0.04885610125	$9.07 \times 10^{-9}$
(0.3, 0.3)	0.1214872927	0.1214871746	$1.181 \times 10^{-7}$
(0.4, 0.4)	0.2386919517	0.2386917676	$1.841 \times 10^{-7}$
(0.5, 0.5)	0.4121803178	0.4121800093	$3.085 \times 10^{-7}$
(0.6, 0.6)	0.6559627680	0.6559619662	$8.018 \times 10^{-7}$
(0.7, 0.7)	0.9867388264	0.9867372530	$1.5734 \times 10^{-6}$
(0.8, 0.8)	1.424346194	1.424344379	$1.815 \times 10^{-6}$
(0.9, 0.9)	1.992278520	1.992276313	$2.207 \times 10^{-6}$
(1.0, 1.0)	2.718281828	2.718268380	$1.3448 \times 10^{-5}$

Table 3: Comparison Absolute Error for Example 1

$(x, t)$	$Exact$	$Error_{OurMethod}$	$Error_{NKH}$
(0.1, 0)	0.01	0.0000000000	0.0000000000
(0.1, 0.1)	0.01105170918	$2.023 \times 10^{-8}$	$3.34691 \times 10^{-6}$
(0.1, 0.3)	0.01349858808	$9.43 \times 10^{-9}$	$3.03472 \times 10^{-5}$
(0.1, 0.5)	0.01648721271	$1.5 \times 10^{-10}$	$8.22639 \times 10^{-5}$
(0.1, 0.7)	0.02013752707	$1.460 \times 10^{-8}$	$1.48971 \times 10^{-4}$
(0.1, 0.9)	0.02459603111	$1.745 \times 10^{-8}$	$2.05545 \times 10^{-4}$

**Example 2.** Consider a linear 2D MVFIE of the second kind [4]

$$m(x, t) = t^2 e^x + \frac{1}{3} t^3 x^2 + \int_0^t \int_0^1 x^2 e^{-y} m(y, z) dy dz, \quad (32)$$

which has an exact solution given as  $m(x, t) = t^2 e^x$  in the interval  $(x, t) = [0, 1]$ .

Let the approximate solution to (32) for  $N = 5$  be

$$m_N(x, t) = \sum_{i=0}^5 \sum_{j=0}^5 c_{i,j} B_{i,5}(x) B_{j,5}(t). \quad (33)$$

Substituting (33) into (32) gives

$$\begin{aligned} & \sum_{i=0}^5 \sum_{j=0}^5 c_{i,j} B_{i,5}(x) B_{j,5}(t) \\ &= t^2 e^x + \frac{1}{3} t^3 x^2 + \int_0^t \int_0^1 x^2 e^{-y} \left( \sum_{i=0}^5 \sum_{j=0}^5 c_{i,j} B_{i,5}(y) B_{j,5}(z) \right) dy dz, \quad (34) \\ & \sum_{i=0}^5 \sum_{j=0}^5 c_{i,j} B_{i,5}(x) B_{j,5}(t) - \int_0^t \int_0^1 x^2 e^{-y} \left( \sum_{i=0}^5 \sum_{j=0}^5 c_{i,j} B_{i,5}(y) B_{j,5}(z) \right) dy dz \\ &= t^2 e^x + \frac{1}{3} t^3 x^2. \end{aligned} \quad (35)$$

Collocating (35) and using standard collocation points at  $x = x_i$  and  $t = t_j$  with

$$\begin{aligned} x_i &= \frac{i}{5}; \quad i = 0(1)5, \\ t_j &= \frac{j}{5}; \quad j = 0(1)5, \end{aligned}$$

we have

$$\begin{aligned} & \sum_{i=0}^5 \sum_{j=0}^5 c_{i,j} B_{i,5}(x_i) B_{j,5}(t_j) - \int_0^t \int_0^1 x_i^2 e^{-y} \left( \sum_{i=0}^5 \sum_{j=0}^5 c_{i,j} B_{i,5}(y) B_{j,5}(z) \right) dy dz \\ &= t_j^2 e^{x_i} + \frac{1}{3} t_j^3 x_i^2. \end{aligned} \quad (36)$$

The method was implemented using MAPLE 18 software, and  $M_5(x, t)$  was obtained as

$$\begin{aligned}
 M_5(x, t) = & 1. \times 10^{-8}t^3x + 1.000082530t^2x + 0.171638e^{-3}t^5x^5 \\
 & - 0.11186e^{-3}t^4x^5 - 0.2793619605e^{-4}t^3x^5 + 0.1385620746e^{-1}t^2x^5 \\
 & - 0.32928e^{-3}t^5x^4 + 0.26711e^{-3}t^5x^3 + 0.24133e^{-3}t^4x^4 \\
 & - 0.18882e^{-3}t^4x^3 + 0.1951239210e^{-4}t^3x^4 + 0.3486292507e^{-1}t^2x^4 \\
 & - 0.5431e^{-4}t^5x^2 - 0.1628619605e^{-4}t^3x^3 + 0.1704115775t^2x^3 \\
 & + 0.2330e^{-4}t^4x^2 + 0.117e^{-5}t^3x^2 + .499067752t^2x^2 \\
 & + 1.468779972 \times 10^{-7}tx^2 + 2.0 \times 10^{-8}t^5x - 3.0 \times 10^{-8}t^4x + 1.0t^2.
 \end{aligned} \tag{37}$$

Table 4: Result of Absolute Error for Example 2

$(x, t)$	<i>Exact</i>	<i>Error<sub>OurMethod</sub></i>	<i>Error<sub>AM</sub></i>
(0, 0)	0.0000000000	0.0000000000	$2.46 \times 10^{-5}$
(0.1, 0.1)	0.01105170918	$2.064 \times 10^{-8}$	$1.46 \times 10^{-5}$
(0.2, 0.2)	0.04885611032	$1.22 \times 10^{-9}$	$3.37 \times 10^{-4}$
(0.3, 0.3)	0.1214872927	$8.31 \times 10^{-8}$	$2.45 \times 10^{-3}$
(0.4, 0.4)	0.2386919517	$1.192 \times 10^{-7}$	$1.00 \times 10^{-2}$
(0.5, 0.5)	0.4121803178	$3.200 \times 10^{-7}$	$3.05 \times 10^{-2}$
(0.6, 0.6)	0.6559627680	$1.2541 \times 10^{-6}$	$7.58 \times 10^{-2}$
(0.7, 0.7)	0.9867388264	$3.1681 \times 10^{-6}$	$1.63 \times 10^{-1}$
(0.8, 0.8)	1.424346194	$5.121 \times 10^{-6}$	$3.17 \times 10^{-1}$
(0.9, 0.9)	1.992278520	$5.352 \times 10^{-6}$	$5.69 \times 10^{-1}$
(1.0, 1.0)	2.718281828	$5.121 \times 10^{-6}$	$5.70 \times 10^{-1}$

**Example 3.** Consider a linear 2D MVFIE of the second kind [19]

$$m'(x, t) = 2x - \frac{1}{4}t^2 + \frac{1}{6}t^4 + \int_0^t \int_0^1 rtm(r, s)drds, \tag{38}$$

with initial condition  $m(0, t) = -t^2$  which has an exact solution given as  $m(x, t) = x^2 - t^2$  in the interval  $(x, t) = [0, 1]$ .

Let the approximate solution to (38) for  $N = 5$  be

$$m_N(x, t) = \sum_{i=0}^5 \sum_{j=0}^5 c_{i,j} B_{i,5}(x) B_{j,5}(t). \quad (39)$$

Integrating both sides of (38) from 0 from  $x$

$$\int_0^x (m'(x, t)) dx = \int_0^x \left( 2x - \frac{1}{4}t^2 + \frac{1}{6}t^4 \right) + \int_0^x \left( \int_0^t \int_0^1 rtm(r, s) dr ds \right) dx, \quad (40)$$

$$m(x, t) - m(0, t) = x^2 - \frac{1}{4}xt^2 + \frac{1}{6}xt^4 + \int_0^x \left( \int_0^t \int_0^1 rtm(r, s) dr ds \right) dx, \quad (41)$$

$$m(x, t) = x^2 - t^2 - \frac{1}{4}xt^2 + \frac{1}{6}xt^4 + \int_0^x \left( \int_0^t \int_0^1 rtm(r, s) dr ds \right) dx. \quad (42)$$

substituting (39) into (42) gives

$$\begin{aligned} \sum_{i=0}^5 \sum_{j=0}^5 c_{i,j} B_{i,5}(x) B_{j,5}(t) = & x^2 - t^2 - \frac{1}{4}xt^2 + \frac{1}{6}xt^4 \\ & + \int_0^x \left( \int_0^t \int_0^1 rt \left( \sum_{i=0}^5 \sum_{j=0}^5 c_{i,j} B_{i,5}(r) B_{j,5}(s) \right) dr ds \right) dx, \end{aligned} \quad (43)$$

$$\begin{aligned} \sum_{i=0}^5 \sum_{j=0}^5 c_{i,j} B_{i,5}(x) B_{j,5}(t) - \int_0^x \left( \int_0^t \int_0^1 rt \left( \sum_{i=0}^5 \sum_{j=0}^5 c_{i,j} B_{i,5}(r) B_{j,5}(s) \right) dr ds \right) dx \\ = x^2 - t^2 - \frac{1}{4}xt^2 + \frac{1}{6}xt^4. \end{aligned} \quad (44)$$

Collocating (44) and using standard collocation points at  $x = x_i$  and  $t = t_j$  with

$$\begin{aligned} x_i &= \frac{i}{5}; \quad i = 0(1)5, \\ t_j &= \frac{j}{5}; \quad j = 0(1)5, \end{aligned}$$

we have

$$\sum_{i=0}^5 \sum_{j=0}^5 c_{i,j} B_{i,5}(x_i) B_{j,5}(t_j)$$

$$\begin{aligned}
& - \int_0^x \left( \int_0^t \int_0^1 r t_j \left( \sum_{i=0}^5 \sum_{j=0}^5 c_{i,j} B_{i,5}(r) B_{j,5}(s) \right) dr ds \right) dx \\
& = x_i^2 - t_j^2 - \frac{1}{4} x_i t_j^2 + \frac{1}{6} x_i t_j^4.
\end{aligned} \tag{45}$$

The method was implemented using MAPLE 18 software and  $M_5(x, t)$  was obtained as

$$M_5(x, t) = x^2 - t^2, \tag{46}$$

is the exact solution.

**Example 4.** Consider a linear 2D MVFIE of the second kind [19]

$$m'(x, t) = 1 - \frac{1}{6}t^2 - \frac{1}{6}t^3 + \int_0^t \int_0^1 r sm(r, s) dr ds \tag{47}$$

with initial condition  $m(0, t) = t$  that has an exact solution given as  $m(x, t) = x + t$  in the interval  $(x, t) = [0, 1]$ .

Let the approximate solution to (47) for  $N = 5$  be

$$m_N(x, t) = \sum_{i=0}^5 \sum_{j=0}^5 c_{i,j} B_{i,5}(x) B_{j,5}(t). \tag{48}$$

Integrating both sides of (47) from 0 to  $x$ , we have

$$\int_0^x (m'(x, t)) dx = \int_0^x \left( 1 - \frac{1}{6}t^2 - \frac{1}{6}t^3 \right) + \int_0^x \left( \int_0^t \int_0^1 r sm(r, s) dr ds \right) dx, \tag{49}$$

$$m(x, t) = x + t - \frac{1}{6}xt^2 - \frac{1}{6}xt^3 + \int_0^x \left( \int_0^t \int_0^1 r sm(r, s) dr ds \right) dx. \tag{50}$$

Substituting (48) into (50) gives

$$\begin{aligned}
& \sum_{i=0}^5 \sum_{j=0}^5 c_{i,j} B_{i,5}(x) B_{j,5}(t) \\
& - \int_0^x \left( \int_0^t \int_0^1 r s \left( \sum_{i=0}^5 \sum_{j=0}^5 c_{i,j} B_{i,5}(r) B_{j,5}(s) \right) dr ds \right) dx \\
& = x + t - \frac{1}{6}xt^2 - \frac{1}{6}xt^3.
\end{aligned} \tag{51}$$

Collocating (51) and using standard collocation points at  $x = x_i$  and  $t = t_j$  with

$$\begin{aligned} x_i &= \frac{i}{5}; & i &= 0(1)5, \\ t_j &= \frac{j}{5}; & j &= 0(1)5, \end{aligned}$$

we have

$$\begin{aligned} & \sum_{i=0}^5 \sum_{j=0}^5 c_{i,j} B_{i,5}(x_i) B_{j,5}(t_j) \\ & - \int_0^x \left( \int_0^t \int_0^1 r s \left( \sum_{i=0}^5 \sum_{j=0}^5 c_{i,j} B_{i,5}(r) B_{j,5}(s) \right) dr ds \right) dx \\ & = x_i + t_j - \frac{1}{6} x_i t_j^2 - \frac{1}{6} x_i t_j^3 \end{aligned} \quad (52)$$

The method was implemented using MAPLE 18 software and  $M_5(x, t)$  was obtained as

$$\begin{aligned} M_5(x, t) = & 0.166e^{-5}t^4x - 0.141e^{-5}t^3x - 0.61e^{-5}t^2x^2 - 0.240e^{-5}tx^3 \\ & + 0.170e^{-4}t^3x^2 + 0.211e^{-4}t^2x^3 + 0.318e^{-5}tx^4 - 0.595e^{-4}t^3x^3 \\ & - 0.285e^{-4}t^2x^4 - 0.1435e^{-5}tx^5 - 0.3660e^{-4}t^3x^5 + 0.681e^{-4}t^4x^3 \\ & + 0.801e^{-4}t^3x^4 + 0.1294e^{-4}t^2x^5 - 0.196e^{-4}t^4x^2 + 0.826e^{-5}t^5x^2 \\ & - 0.16695e^{-4}t^5x^5 + 0.3678e^{-4}t^5x^4 - 0.2778e^{-4}t^5x^3 \\ & + 0.41300e^{-4}t^4x^5 - 0.914e^{-4}t^4x^4 + 1.000000000t + 1.000000000x \\ & - 7.1e^{-7}t^5x + 7.0e^{-7}tx^2 + 5.0e^{-7}t^2x - 6.0e^{-8}tx \end{aligned} \quad (53)$$

## 4 Conclusion

In this section, a new numerical method was developed for solving 2D MV-FIEs of the second kind utilizing polynomial collocation. The findings obtained from each case were compared with the exact solution and some existing studies in the literature, the new approach established is simple, reliable, and efficient to compute. Maple 18 software is utilized for all computations

Table 5: Results using Bernstein polynomial for Example 4

$(x, t)$	<i>Exact</i>	<i>OurMethod</i> $N = 5$	<i>ErrorOurMethod</i>
(0, 0)	0.0000000000	0.0000000000	0.0000000000
(0.1, 0.1)	0.0200000000	0.0200000000	0.0000000000
(0.2, 0.2)	0.0400000000	0.3999999998	$2.0e-10$
(0.3, 0.3)	0.6000000000	0.5999999993	$7.0e-10$
(0.4, 0.4)	0.8000000000	0.7999999978	$2.2e-9$
(0.5, 0.5)	0.1000000000	0.9999999930	$7.0e-9$
(0.6, 0.6)	1.2000000000	1.199999981	$1.9e-8$
(0.7, 0.7)	1.4000000000	1.3999999564	$4.4e-8$
(0.8, 0.8)	1.6000000000	1.599999896	$1.04e-7$
(0.9, 0.9)	1.8000000000	1.799999750	$2.50e-7$
(1.0, 1.0)	2.0000000000	1.999999430	$5.70e-7$

in this work. The accuracy of the method is proved by considering various examples, which shows that the method is efficient and appropriate for this type of situations. We compare our absolute errors of Example 1 with [9] as shown in Table 2 and also absolute errors of Example 2 with [4] as shown in Table 4. We can therefore conclude that our method is superior and more preferable than the existing methods.

The results obtained from problem 1 at  $N = 5$  and at different values of  $(x, t)$  shows clearly that the developed method is better than the method presented by [9]. From Table 3 for  $(x, t) = (0.1, 0.1)$  and  $N = 5$ , for instance the absolute errors are  $Error_B = 2.023 \times 10^{-8}$  and  $Error_{NKH} = 3.3469 \times 10^{-6}$ . Again from Table 3 for  $(x, t) = (0.1, 0.3)$  and  $N = 5$ , the absolute errors are  $Error_B = 9.43 \times 10^{-9}$  and  $Error_{NKH} = 3.03472 \times 10^{-5}$  which shows clearly that the developed method is consistent, reliable, and performs favorably.

The results obtained from problem 2 at  $N = 5$  and at different values of  $(x, t)$  shows clearly that the developed method is better than the method presented by [4]. From Table 4, for instance the absolute errors for  $(x, t) = (0.0, 0.0)$  and at  $N = 5$  gives  $Error_B = 0.0000000$  and  $Error_{AM} = 2.46 \times 10^{-5}$ , for  $(x, t) = (0.1, 0.1)$  and at  $N = 5$ , gives  $Error_B = 2.064 \times 10^{-8}$  and

$Error_{AM} = 1.46 \times 10^{-5}$ . Again From Table 3 for  $(x, t) = (1.0, 1.0)$  and at  $N = 5$ , the absolute errors are  $Error_B = 5.121 \times 10^{-6}$  and  $Error_{AM} = 5.70 \times 10^{-1}$  shows clearly that the developed method is consistent, efficient and converges faster than the method presented by [4].

It was observed that the results obtained for Example 3 at  $N = 5$  give the exact solution, hence the reason it is not in tabular form. This clearly indicates that the method is efficient and convergent.

The solution obtained from Example 4 at  $N = 5$  and at various values of  $(x, t)$  indicates the method is stable and converges to the exact solution. From Table 5 for instance, the result obtained at  $N = 5$  and  $(x, t) = (0, 0)$ ,  $(x, t) = (0.1, 0.1)$ ,  $(x, t) = (0.2, 0.2)$  and  $(x, t) = (0.3, 0.3)$  gives 0.0000000,  $2.0 \times 10^{-10}$ ,  $7.0 \times 10^{-10}$  respectively.

It has been observed and examined that when the values of  $N$  increase, the error decreases and the approximate solution converges rapidly to the exact solution, the value of  $N = 5$  was chosen arbitrarily and for simplicity.

## Acknowledgments

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

## References





- [1] Adesanya, A.O., Yahaya, Y.A., Ahmed, B. and Onsachi R.O. *Numerical solution of linear integral and integro-differential equations using Boubakar collocation method*, International journal of Mathematical Analysis and Optimization: Theory and Applications, 2019 (2) (2019) 592–598.
- [2] Agbolade O.A. and Anake T.A. *Solutions of first-order Volterra type linear integro differential equations by collocation method*, J. Appl. Math. 2017 (1) (2017) 1510267.

- [3] Ajileye G., James A.A., Ayinde A.M. and Oyedepo T. *Collocation approach for the computational solution of Fredholm-Volterra fractional order of integro-differential equations*, J. Nig. Soc. Phys. Sci. 4 (2022) 834.
- [4] Almasied, H. and Meleh, J.N. *Numerical solution of a class of mixed two-dimensional nonlinear Volterra-Fredholm integral equations using multi-quadratic radial basis functions*, J. Comput. Appl. Math. 260 (2014) 173–179.
- [5] Atkinson K.E. *The numerical solution of integral equations of the second kind*, Cambridge University Press, 1997.
- [6] Banifatemi E., Razzaghi M. and Yousefi S. *Two-dimensional Legendre wavelets method for the mixed Volterra-Fredholm integral equations*, J. Vib.d Contr. 13 (2007), 1667.
- [7] Berinde V. *Iterative approximation of fixed points*, Romania, Editura Efermeride, Baia Mare, 2002.
- [8] Chari M.V.K. and Salon S.J. *Numerical methods in electromagnetism*, Academic Press, 2000.
- [9] Darani, N.M., Maleknejad, K. and Mesgarani, H. *A new approach for two dimensional nonlinear mixed Volterra-Fredholm integral equations and its convergence analysis*, TWMS J. Pure Appl. Math. 10 (2019) 132–139.
- [10] Joy K.I. *Bernstein polynomials*, On-Line Geometric Modeling Notes, 2000.
- [11] Khuri S. and Wazwaz A.M. *The decomposition method for solving a second Fredholm second kind integral equation with a logarithmic kernel*, Inter. J. Comput. Math. 61 (1996), 103–110.
- [12] Maleknejad K. and Behbahani Z. J. *Applications of two-dimensional triangular functions for solving nonlinear class of mixed Volterra-Fredholm integral equations*, Math. Comput. Model. 55(2012) 1833–1844.
- [13] Owaied A.K. *Some approximation methods for solving Volterra-Fredholm integral and integro-differential equations*, PHD thesis, technology university , Iraq (2010).

- [14] Robert J.V. *Uniform continuity is almost Lipschitz continuity*, Department of civil Engineering Princeton university NJ 08544, 1991.
- [15] Rostam K.S. and Karzan A.B. *Solving two-dimensional linear Volterra–Fredholm integral equations of the second kind by using series solution methods*, J. Zankoi Sulaimani Pure Appl. Sci. 17 (4) (2015) 253–270.
- [16] Saeed R.K. and Berdawood K.A. *Solving two-dimensional linear Volterra–Fredholm integral equations of the second kind by using successive approximation method and method of successive substitutions*, Zanco J. Pure Appl. Sci. 28 (2) (2016) 35-46.
- [17] Saleh M.H., Mohammed D.S., and Taher R.A. *Approximate solution of two-dimensional Volterra integral equation by Chebyshev polynomial method and Adomian decomposition method*, Math. Theory Model. 6 (2016) 4.
- [18] Uwaheren O.A., Adebisi A.F. and Taiwo O.A. *Perturbed collocation method for solving singular multi-order fractional differential equations of Lane-Emden type*, J. Nigerian Soci. Phys. Sci. 3 (2020) 141.
- [19] Wazwaz A.M. *Linear and nonlinear integral equations: Methods and applications*, Springer. Saint Xavier University Chicago USA (2011), 306–309.



## Cutting-edge spectral solutions for differential and integral equations utilizing Legendre's derivatives

A.M. Abbas<sup>1, </sup>, Y.H. Youssri<sup>2, </sup>, M. El-Kady<sup>1, </sup> and M. Abdelhakem<sup>1,\*, </sup>

### Abstract

This research introduces a spectral numerical method for solving some types of integral equations, which is the pseudo-Galerkin spectral method. The presented method depends on Legendre's first derivative polynomials as basis functions. Subsequently, an operational integration matrix has been constructed to express integrals as a linear combination of these basis functions. This process transforms the given integral equation into a

\*Corresponding author

Received 24 January 2025; revised 30 April 2025; accepted 9 May 2025

Ahmed M. Abbas, (ahmed.m.abbas.96@gmail.com;

ahmedmabbas@science.helwan.edu.eg)

Youssri Hassan Youssri, (yousri@cu.edu.eg; youssri@aucegypt.edu)

Mamdouh El-Kady, (mamdouh\_elkady@cic-cairo.com)

Mohamed Abdelhakem, (mabdelhakem@yahoo.com;

mabdelhakem@science.helwan.edu.eg)

<sup>1</sup>Mathematics Department, Faculty of Science, Helwan University, Helwan 11795, Egypt, Helwan School of Numerical Analysis in Egypt (HSNAE).

<sup>2</sup>Mathematics Department, Faculty of Science, Cairo University, Giza 12613, Egypt.

### How to cite this article

Abbas, A.M., Youssri, Y.H., El-Kady, M. and Abdelhakem, M., Cutting-edge spectral solutions for differential and integral equations utilizing Legendre's derivatives. *Iran. J. Numer. Anal. Optim.*, 2025; 15(3): 1036-1074.  
<https://doi.org/10.22067/ijnao.2025.91804.1590>

system of algebraic equations. The unknowns of the obtained system are the spectral expansion constants. Then, we solve the obtained algebraic system using the Gauss elimination method for linear systems or Newton's iteration method for nonlinear systems. This approach yields the desired semi-analytic approximate solution. Additionally, our method extends to the solution of ordinary differential equations, as every initial value problem can be equivalently represented as a corresponding Volterra integral equation. On the other hand, every boundary value problem can be transformed into a corresponding Fredholm integral equation. This transformation is achieved by incorporating the given conditions. Moreover, convergence and error analyses are thoroughly examined. Finally, to validate the efficiency and accuracy of the proposed method, we conduct numerical test problems.

**AMS subject classifications (2020):** 65R20, 65N35, 45L05, 33C45.

**Keywords:** Legendre polynomials; Pseudo-Galerkin spectral method; Integral equations; Lane–Emden equation; Stable population model.

Abbreviations	
BVPs:	Boundary Value Problems
$f_n$ :	Function
IVPs:	Initial Value Problems
LGL:	Legendre Gauss Lobatto quadrature points
MAE:	Maximum Absolute Error
PGDL:	Pseudo-Galerkin method via Legendre's Derivative polynomials
RSE:	Root Square Error

## 1 Introduction

Integral and differential equations are found in numerous applications across various fields. In fluid mechanics, the modeling of fluid flow involves differential equations [28, 21, 18, 10, 14]. Electromagnetic wave-related problems in electromagnetism are often described using integral equations [22, 11]. Additionally, the behavior of heat flow and temperature distribution in different materials and shapes can be modeled effectively with differential and

integral equations [26, 34]. Some other important applications, such as the Lane–Emden equation [8, 33, 54, 7], are discussed. Integral and differential equations stand as essential mathematical tools for problem modeling in science and engineering [29, 37, 46, 32, 30].

Often, obtaining the exact solution for these equations proves challenging, leading to the development of numerical, or approximation methods for introducing numerical and semi-analytic approximations as viable alternatives [24]. These methods include finite element [31, 25, 36], finite difference [23, 13], and spectral methods, among others [44, 41, 42, 45]. While finite element and finite difference methods yield numerical approximation solutions, providing a set of dependent variable values at independent variable values. On the other hand, spectral methods offer semi-analytic approximation solutions, representing smooth functions of the independent variable or variables.

The inner product space encompassing all polynomials is not a Cauchy space; nevertheless, any smooth function ( $f_n$ ) serves as a limit point within this space. More precisely, any smooth  $f_n$  can be expressed as the limit of a sequence of polynomials. The fundamental concept underlying all spectral methods involves representing the dependent variable as a linear combination of a set of functions that constitute a basis for the polynomial space. In simpler terms, the unknown  $f_n$  is expanded as a series involving unknown constants multiplied by the basis functions. Crucially, these basis functions are required to be orthogonal concerning an inner product defined by a weight function  $w_1(x)$  [12].

Expanding upon this idea, spectral methods leverage the orthogonality of basis functions to efficiently approximate complicated functions, even if they are two-dimensional [20]. Orthogonal basis functions contribute to simplifying the representation of the unknown function, leading to solutions that are both accurate and computationally efficient. The weight function is pivotal in establishing the inner product, impacting the orthogonality of the basis functions, and enhancing the effectiveness of spectral methods. In summary, the incorporation of orthogonal basis functions and a weight function in spectral methods enhances their ability to efficiently represent and approximate functions, making them effective tools for approximating the solution of dif-

ferential and integral equations in various applications related to science and engineering.

The integral and differential equations can be formulated as an operator acting on the unknown function  $y(x)$ , equated to another given function. When the terms of the equation are rearranged such that they all reside on the left-hand side, and  $y(x)$  is expressed as a finite series involving unknown constants multiplied by basis functions, the left-hand side of the equation is referred to as the residue function. The closer the residue function is to zero, the more accurate the approximate solution is to the exact one, and the value of the residue is termed the residual error. Another set of functions is defined, known as the set of test functions, and a distinct inner product is constructed using a suitable weight function  $w_2(x)$ . By setting the inner product of the residue function and the test functions equal to zero, a system of algebraic equations in the unknown expansion coefficients is obtained. Subsequently, the approximate solution is derived by solving this system of algebraic equations.

Spectral methods encompass three main types: the Galerkin [52, 9], Tau [19, 35, 38], and collocation [17, 27, 48] methods. In the Galerkin method, the test functions and the basis functions are identical, satisfying the initial/boundary conditions of the IVPs/BVPs. Additionally, the weight function  $w_2(x)$  should match  $w_1(x)$ . For the collocation method, the weight function  $w_2(x)$  is set to one, and the set of test functions comprises Dirac delta functions shifted to predetermined collocation points. In the Tau method, the test and basis functions differ, and the weight functions  $w_1$  and  $w_2$  may vary. Additionally, there is no requirement to satisfy any (initial/boundary) conditions.

A modified version of spectral methods, known as pseudo-spectral methods [1, 53, 47], employs the Gauss quadrature method to define an alternative inner product expressed as a summation over a set of quadrature points instead of integration. Moreover, the unknown expansion coefficients can be expressed in terms of the dependent variable evaluated at the quadrature points, aided by the orthogonality relation. Consequently, differentiation and integration matrices can be constructed.

At the heart of both spectral and pseudo-spectral methods lies the crucial decision regarding the choice of basis functions. In the work by the authors in [5], the first derivative of Legendre polynomials was chosen as the basis function. They employed both the tau and collocation methods and crafted the operational matrix of the differentiation of the chosen basis functions. Similarly, in the study by the authors in [4], the same basis functions were adopted, and the pseudo-spectral method was employed to formulate the differentiation and integration matrices. Higher derivatives of Legendre polynomials are utilized in [3, 15].

In this paper, we also employ the same basis functions; however, our focus shifts towards constructing the operational matrix of integration for these basis functions. We then utilize the pseudo-Galerkin spectral method [28, 16, 43] in our numerical approach, in which the system of algebraic equations can be formulated by substituting the Legendre Gauss–Lobatto quadrature points into the residue function.

The paper consists as follows. Section 2 provides a comprehensive overview of Legendre and Legendre’s derivative polynomials, establishing the foundational groundwork for the subsequent discussions. We delve into pertinent notations and historical relationships concerning the basis functions crucial to our methodology. The construction of the operational integration matrix, along with the introduction of key equations, is detailed in section 3. Section 4 encapsulates the algorithmic framework of our proposed approach. A meticulous exploration of convergence and error analysis is undertaken in section 5, shedding light on the method’s reliability and precision. Section 6 is dedicated to the presentation of various test problems, strategically chosen to validate the accuracy, efficiency, and stability of our proposed methodology. Finally, section 7 synthesizes the overarching conclusions drawn from our study and outlines potential avenues for future research.

## 2 Preliminaries

Throughout this work, some needed properties of Legendre and Legendre’s derivative polynomials are displayed, such as boundaries, upper bounds, or-

thogonality relations, and recurrence relations. They are essential to get some new desired properties of Legendre's derivative polynomials.

The Legendre polynomial of degree  $q$ , denoted by  $\mathcal{L}_q(x)$ , can be obtained using the following recurrence relations [5, 39]:

$$(q+1)\mathcal{L}_{q+1}(x) = (2q+1)x\mathcal{L}_q(x) - q\mathcal{L}_{q-1}(x), \quad (1)$$

$$(2q+1)\mathcal{L}_q(x) = \mathcal{L}'_{q+1}(x) - \mathcal{L}'_{q-1}(x), \quad (2)$$

$$\mathcal{L}'_q(x) = \sum_{j=0}^{\lfloor \frac{q-1}{2} \rfloor} [2(q-2j-1)+1]\mathcal{L}_{q-2j-1}(x), \quad (3)$$

$$(1-x^2)\mathcal{L}'_q(x) = \frac{q(q+1)}{2q+1}[\mathcal{L}_q(x) - \mathcal{L}_{q+2}(x)], \quad (4)$$

according to the initials:  $\mathcal{L}_0(x) = 1$  and  $\mathcal{L}_1(x) = x$ , where  $q \geq 1$ .

In addition, they can be expanded as follows [39]:

$$\mathcal{L}_q(x) = \sum_{j=0}^{\lfloor \frac{q}{2} \rfloor} (-1)^j \frac{(2q-2j)!}{2^q j! (q-j)! (q-2j)!} x^{q-2j}, \quad (5)$$

where  $x \in [-1, 1]$ , and  $q \geq 0$ .

The following definition describes the introduced basis functions used in this work for non-negative degree  $q$  [5].

**Definition 1.** Legendre's derivative polynomials of degree  $q$ , denoted by  $DL_q(x)$ , are defined to be the derivative of the Legendre polynomial that is one degree higher:

$$DL_q(x) = \frac{d}{dx} \mathcal{L}_{q+1}(x). \quad (6)$$

Legendre's derivative polynomials can be expanded as follows [5]:

$$DL_q(x) = \sum_{j=0}^{\lfloor \frac{q}{2} \rfloor} (-1)^j \frac{(2q-2j+2)!}{2^{q+1} (j)! (q-j+1)! (q-2j)!} x^{q-2j}. \quad (7)$$

The boundary values for Legendre [39] and Legendre's derivative [5] polynomials are given by

$$\mathcal{L}_q(\pm 1) = (\pm 1)^q, \quad (8)$$

$$DL_q(\pm 1) = \frac{(\pm 1)^q (q+1)(q+2)}{2}. \quad (9)$$

Also, it is known that they are bounded as shown in the following relation [39, 5]:

$$|\mathcal{L}_q(x)| \leq 1, \quad (10)$$

$$|DL_q(x)| \leq (q+1)^2. \quad (11)$$

Legendre polynomials are orthogonal under the weight function  $w(x) = 1$ , and the orthogonality relation is given by

$$\int_{-1}^1 \mathcal{L}_q(x) \mathcal{L}_r(x) dx = \frac{2}{2q+1} \delta_{q,r}, \quad (12)$$

where  $\delta_{q,r} = \begin{cases} 1, & q = r, \\ 0, & q \neq r, \end{cases}$  for nonnegative integers  $q$  and  $r$ .

The Legendre's derivative polynomials are orthogonal with respect to the weight function  $1 - x^2$  such that [5]: [?])

$$\int_{-1}^1 (1-x^2) DL_q(x) DL_r(x) dx = \frac{2(q+2)(q+1)}{2q+3} \delta_{q,r}. \quad (13)$$

To solve integral equations, it is necessary to expand the integration of the basis functions as a linear combination of the basis functions themselves. The essential relations and properties required are shown in the forward sections. The following section will introduce the derivation of some important relations and the operational integration matrix construction.

### 3 Legendre's derivative polynomials operational integration matrix

It is undeniable that integrals on the entire domain  $[a, b]$  or a variable domain  $[a, x]$  occur in integral equations. That is why we need to find a suitable representation for these integrals. Therefore, the integrals of our basis functions will be calculated in this section to help expand these integrals as summations. Hence, we can construct an integration operational matrix for Legendre's derivative polynomials.

At the beginning, some essential relations must be introduced. A recurrence relation of the used polynomials,  $DL_q(x)$ , will be constructed using

recurrence relations (1) and (2):

$$x DL_q(x) = \frac{q+1}{2q+3} DL_{q+1}(x) + \frac{q+2}{2q+3} DL_{q-1}(x), \quad (14)$$

where  $q \geq 1$ .

The next lemma introduces the first moment equation.

**Lemma 1.** Consider Legendre's derivative polynomials column matrix  $DL(x)$  of  $N+1$  rows. Then the square matrix  $X$  of rank  $N+1$  satisfies the following:

$$X \cdot DL(x) = XDL(x), \quad (15)$$

where  $DL(x) = [DL_0(x), DL_1(x), DL_2(x), \dots, DL_N(x)]^T$ ,  $XDL(x) = [0, xDL_0(x), xDL_1(x), \dots, xDL_{N-1}(x)]^T$ , and  $X = \{x_{rc}\}_{(N+1) \times (N+1)}$ , such that

$$x_{rc} = \begin{cases} \frac{r}{2r+1}, & r = c, \\ \frac{r+1}{2r+1}, & c = r-2, r > 1, \\ 0, & \text{otherwise,} \end{cases} \quad (16)$$

where  $r, c = 0, 1, 2, \dots, N$ .

*Proof.* By straight forward matrix multiplication with the help of (14), we have

$$X \cdot DL(x) = \left\{ \sum_{c=1}^{N+1} x_{rc} DL_{c-1}(x) \right\}. \quad (17)$$

□

The next lemmas and theorems will be concerning different forms and techniques for the integration of the introduced basis function.

**Lemma 2.** Consider Legendre's derivative polynomials column matrix  $DL(x)$  of  $N+1$  rows and Legendre polynomials column matrix  $\mathcal{L}(x)$  of  $N+1$  rows. Then the square matrix  $M$  of rank  $N+1$  satisfies the following:

$$M \cdot DL(x) = \mathcal{L}(x), \quad (18)$$

where  $DL(x) = [DL_0(x), DL_1(x), DL_2(x), \dots, DL_N(x)]^T$ ,

$\mathcal{L}(x) = [\mathcal{L}_0(x), \mathcal{L}_1(x), \mathcal{L}_2(x), \dots, \mathcal{L}_N(x)]^T$ , and

$M = \{m_{rc}\}_{(N+1) \times (N+1)}$ , such that

$$m_{rc} = \begin{cases} \frac{1}{2r+1}, & r = c, \\ \frac{-1}{2r+1}, & c = r - 2, r > 1, \\ 0, & \text{otherwise,} \end{cases} \quad (19)$$

where  $r, c = 0, 1, 2, \dots, N$ .

*Proof.* The left-hand side of the relation (18) is expanded using the usual matrix multiplication as follows:

$$L.H.S. = M \cdot DL(x) = \left[ \sum_{c=0}^N m_{0c} DL_c(x), \sum_{c=0}^N m_{1c} DL_c(x), \right. \\ \left. \sum_{c=0}^N m_{2c} DL_c(x), \dots, \sum_{c=0}^N m_{Nc} DL_c(x) \right]^T.$$

With the aid of relation (19), we get

$$L.H.S. = \left[ DL_0(x), \frac{1}{3} DL_1(x), \frac{-1}{5} DL_0(x) + \frac{1}{5} DL_2(x), \dots, \right. \\ \left. \frac{-1}{2N+1} DL_{N-2}(x) + \frac{1}{2N+1} DL_N(x) \right]^T.$$

Using relation (2), the proof is completed.  $\square$

**Theorem 1.** Integrals of the introduced basis functions can be presented as a linear combination of the introduced basis functions themselves according to

$$\int_{-1}^x DL_q(t) dt = \sum_{k=0}^2 \left[ \frac{(-1)^k (1 - \delta_{k,2}) (1 - \delta_{q,0} \delta_{k,1})}{2q+3} + (-1)^q \delta_{k,2} \right] \\ \times DL_{(q+1-2k)(1-\delta_{k,2})(1-\delta_{q,0} \delta_{k,1})}(x). \quad (20)$$

*Proof.* For  $q = 0$ , we get

$$\int_{-1}^x DL_0(t) dt = \int_{-1}^x (1) dt = x + 1 = \frac{1}{3} DL_1(x) + DL_0(x). \quad (21)$$

For  $q = 1, 2, \dots$ , and by using Lemma 2 and (9), we can conclude that, for  $N \geq q$ ,

$$\begin{aligned} \int_{-1}^x DL_q(t) dt &= \left[ \sum_{c=0}^N m_{qc} DL_c(t) \right]_{t=-1}^{t=x} \\ &= \frac{1}{2q+3} DL_{q+1}(x) - \frac{1}{2q+3} DL_{q-1}(x) + (-1)^q DL_0(x), \end{aligned} \quad (22)$$

which completes the proof for  $q > 0$ .  $\square$

As a direct result, we can deduce the following corollary.

**Corollary 1.** The integral on the entire domain  $[-1, 1]$ , can be calculated as follows:

$$\int_{-1}^1 DL_q(t) dt = 1 + (-1)^q, \quad (23)$$

where  $q$  is a nonnegative integer.

Now, all the necessary relations that are needed to construct the operational matrix of integration have been introduced.

**Theorem 2.** Consider the column matrix

$DL(x) = [DL_0(x), DL_1(x), DL_2(x), \dots, DL_N(x)]^T$  and the column matrix  $\int_{-1}^x DL(t) dt = \left[ 0, \int_{-1}^x DL_0(t) dt, \int_{-1}^x DL_1(t) dt, \dots, \int_{-1}^x DL_{N-1}(t) dt \right]^T$  of  $N+1$  rows. Then the square matrix  $\mathcal{G}$  of rank  $N+1$  satisfies (24)

$$\mathcal{G} \cdot DL(x) = \int_{-1}^x DL(t) dt, \quad (24)$$

where  $\mathcal{G} = \{\tilde{\theta}_{rc}\}_{(N+1) \times (N+1)}$ , such that

$$\tilde{\theta}_{rc} = \begin{cases} -\frac{6}{5}, & c=0, r=2, \\ (-1)^{r+1}, & c=0, r \geq 1, r \neq 2, \\ \frac{1}{2r+1}, & c=r, r \geq 1, \\ \frac{-1}{2r+1}, & c=r-2, r \geq 3, \\ 0, & \text{otherwise,} \end{cases} \quad (25)$$

where  $r, c = 0, 1, 2, \dots, N$ .

*Proof.* Multiply the matrix  $\mathcal{G}$  defined in (25) by the column matrix  $DL(x)$  using the matrix multiplication to get the left-hand side of (24) as follows:

$$L.H.S. = \mathcal{G} \cdot DL(x) = \left[ \sum_{c=0}^N \mathfrak{d}_{0c} DL_c(x), \sum_{c=0}^N \mathfrak{d}_{1c} DL_c(x), \right. \\ \left. \sum_{c=0}^N \mathfrak{d}_{2c} DL_c(x), \dots, \sum_{c=0}^N \mathfrak{d}_{Nc} DL_c(x) \right]^T.$$

Use Theorem 1 and the definition of  $\mathfrak{d}_{rc}$  to conclude that

$$L.H.S. = \left[ 0, \int_{-1}^x DL_0(t) dt, \int_{-1}^x DL_1(t) dt, \dots, \int_{-1}^x DL_{N-1}(t) dt \right]^T \\ = \int_{-1}^x DL(t) dt = R.H.S. \quad (26)$$

□

**Remark 1.** Note that the shift that occurs in the column matrix  $\int_{-1}^x DL(t) dt$  is due to the integral operator producing a polynomial that is one degree higher than the input, which means that the polynomials in  $\int_{-1}^x DL(t) dt$  should shift a row downward to align with the degree of the polynomials in the rows of the column matrix  $DL(x)$ .

**Theorem 3.** For all  $q \geq 0$  and  $m \geq 1$ , the moment formula  $x^m DL_q(x)$  can be represented as follows:

$$x^m DL_q(x) = \sum_{k=0}^{\min(m, \lfloor \frac{q+m}{2} \rfloor)} F_{m,k+1,q} DL_{q+m-2k}(x), \quad (27)$$

where

$$F_{1,1,q} = \frac{q+1}{2q+3}, \quad F_{1,2,q} = \frac{q+2}{2q+3}, \\ F_{m,k,q} = \begin{cases} F_{1,1,q} F_{m-1,1,q+1}, & k=1, \\ F_{1,1,0} F_{m-1,k,1}, & 1 < k < m+1, q=0, \\ F_{1,1,q} F_{m-1,k,q+1} + F_{1,2,q} F_{m-1,k-1,q-1}, & 1 < k < m+1, q > 0, \\ F_{1,2,q} F_{m-1,m,q-1}, & k=m+1, \end{cases} \quad (28)$$

for  $m > 1$ .

*Proof.* Three cases will be studied,  $q = 0$ ,  $1 \leq q < m$ , and  $q \geq m$ .

For  $q = 0$ :

We study this at  $m = 1$  as the initial for the proof by mathematical induction:

$$R.H.S. = \sum_{k=0}^0 F_{1,k+1,0} DL_{1-2k}(x) = x = x DL_0 = L.H.S.$$

Consider (27) be valid for some  $m \geq 1$  and  $q = 0$ . Thus, for  $m + 1$ , the right-hand side will be

$$\begin{aligned} R.H.S. &= \sum_{k=0}^{\lfloor \frac{m+1}{2} \rfloor} F_{m+1,k+1,0} DL_{m+1-2k}(x) \\ &= \sum_{k=0}^{\lfloor \frac{m+1}{2} \rfloor} F_{1,1,0} F_{m,k+1,1} DL_{m+1-2k}(x) \\ &= \frac{1}{3} x^m DL_1(x) = x^{m+1} DL_0(x) = L.H.S., \end{aligned}$$

which completes the proof of the first case.

Similarly, for  $1 \leq q < m$ , the initial step will be at  $m = 2$ , which is easy to be verified. Consider the induction step for some  $m \geq 2$ , so at  $m + 1$ , the right-hand side will be

$$\begin{aligned} R.H.S. &= \sum_{k=0}^{\lfloor \frac{m+q+1}{2} \rfloor} F_{m+1,k+1,q} DL_{q+m+1-2k}(x) \\ &= F_{1,1,q} F_{m,1,q+1} DL_{q+m+1}(x) \\ &\quad + \sum_{k=1}^{\lfloor \frac{m+q+1}{2} \rfloor} (F_{1,1,q} F_{m,k+1,q+1} + F_{1,2,q} F_{m,k,q-1}) DL_{q+m+1-2k}(x). \end{aligned}$$

$$\begin{aligned}
 L.H.S. &= x^{m+1} DL_q(x) = x^m (F_{1,1,q} DL_{q+1}(x) + F_{1,2,q} DL_{q-1}(x)) \\
 &= F_{1,1,q} \sum_{k=0}^{\lfloor \frac{m+q+1}{2} \rfloor} F_{m,k+1,q+1} DL_{q+m+1-2k}(x) \\
 &\quad + F_{1,2,q} \sum_{k=0}^{\lfloor \frac{m+q-1}{2} \rfloor} F_{m,k+1,q-1} DL_{q+m-1-2k}(x) \\
 &= F_{1,1,q} F_{m,1,q+1} DL_{q+m+1}(x) \\
 &\quad + \sum_{k=1}^{\lfloor \frac{m+q+1}{2} \rfloor} (F_{1,1,q} F_{m,k+1,q+1} + F_{1,2,q} F_{m,k,q-1}) DL_{q+m+1-2k}(x) \\
 &= R.H.S.
 \end{aligned}$$

Finally, for  $q \geq m$ , the initial step is  $m = 1$ . So, the induction step for  $m + 1$ , the right-hand side will be

$$\begin{aligned}
 R.H.S. &= \sum_{k=0}^{m+1} F_{m+1,k+1,q} DL_{q+m+1-2k}(x) = F_{1,1,q} F_{m,1,q+1} DL_{q+m+1}(x) \\
 &\quad + \sum_{k=1}^m (F_{1,1,q} F_{m,k+1,q+1} + F_{1,2,q} F_{m,k,q-1}) DL_{q+m+1-2k}(x) \\
 &\quad + F_{1,2,q} F_{m,m+1,q-1} DL_{q-m-1}(x).
 \end{aligned}$$

$$\begin{aligned}
 L.H.S. &= x^{m+1} DL_q(x) = x^m (F_{1,1,q} DL_{q+1}(x) + F_{1,2,q} DL_{q-1}(x)) \\
 &= F_{1,1,q} \sum_{k=0}^{m+1} F_{m,k+1,q+1} DL_{q+m+1-2k}(x) \\
 &\quad + F_{1,2,q} \sum_{k=0}^{m+1} F_{m,k+1,q-1} DL_{q+m-1-2k}(x) \\
 &= F_{1,1,q} F_{m,1,q+1} DL_{q+m+1}(x) \\
 &\quad + \sum_{k=1}^{m+1} (F_{1,1,q} F_{m,k+1,q+1} + F_{1,2,q} F_{m,k,q-1}) DL_{q+m+1-2k}(x) \\
 &\quad + F_{1,2,q} F_{m,m+1,q-1} DL_{q-m-1}(x) = R.H.S.,
 \end{aligned}$$

which completes the proof.  $\square$

**Theorem 4.** For every  $m, q \in \mathbb{N}$ , we have

$$\int_{-1}^1 t^m DL_q(t) dt = \begin{cases} 1 + (-1)^{q+m}, & q \geq m, \\ \sum_{k=0}^{\min(m, \lfloor \frac{q+m}{2} \rfloor)} F_{m,k+1,q} \left( 1 + (-1)^{q+m-2k} \right), & 0 \leq q < m. \end{cases} \quad (29)$$

*Proof.* For  $q \geq m$ :

When  $m = 0$ , Corollary 1 is a special case of Theorem 4.

So, we begin with  $m = 1$ . Using (14), integrating on  $[-1, 1]$ , and then using Corollary 1, we get

$$\begin{aligned} L.H.S. &= \int_{-1}^1 t DL_q(t) dt = \frac{q+1}{2q+3} \int_{-1}^1 DL_{q+1}(t) dt + \frac{q+2}{2q+3} \int_{-1}^1 DL_{q-1}(t) dt \\ &= 1 + (-1)^{q+1} = R.H.S., \end{aligned} \quad (30)$$

for  $q \geq 1$ .

Let (29) be valid for some  $m$ , where  $q \geq m$ . Thus, for  $m+1$ , we have

$$\begin{aligned} L.H.S. &= \int_{-1}^1 t^{m+1} DL_q(t) dt = \int_{-1}^1 t^m (t DL_q(t)) dt \\ &= \int_{-1}^1 t^m \left( \frac{q+1}{2q+3} DL_{q+1}(t) + \frac{q+2}{2q+3} DL_{q-1}(t) \right) dt \\ &= \frac{q+1}{2q+3} \left[ 1 + (-1)^{q+1+m} \right] + \frac{q+2}{2q+3} \left[ 1 + (-1)^{q-1+m} \right] \\ &= 1 + (-1)^{q+m+1} = R.H.S., \end{aligned}$$

for  $q \geq m+1$ .

The second case, for  $0 \leq q < m$ , use Theorem 3 to get

$$L.H.S. = \int_{-1}^1 t^m DL_q(t) dt = \int_{-1}^1 \sum_{k=0}^{\min(m, \lfloor \frac{q+m}{2} \rfloor)} F_{m,k+1,q} DL_{q+m-2k}(t) dt. \quad (31)$$

Using Corollary 1, we have

$$L.H.S. = \sum_{k=0}^{\min(m, \lfloor \frac{q+m}{2} \rfloor)} F_{m,k+1,q} \left( 1 + (-1)^{q+m-2k} \right) = R.H.S.. \quad (32)$$

□

## 4 Pseudo-Galerkin approach for solving integral equations

This section presents a method by which some types of integral equations can be solved. This method is based on Legendre's derivative polynomials as basis functions. The relations, lemmas, and theorems developed in section 3 are used while creating our solving technique.

### 4.1 Problem formulation

The introduced problem is a linear Fredholm–Volterra-type integral equation, which can be formulated as follows:

$$Q(x, y(x)) + \int_{-1}^1 p_1(x, t) y(t) dt + \int_{-1}^x p_2(x, t) y(t) dt = 0, \quad (33)$$

where  $p_1(x, t), p_2(x, t)$  are polynomials with respect to  $t$  and  $Q(x, y(x))$  is linear in  $y(x)$ , which means that  $Q(x, y(x)) = f_1(x) + f_2(x) y(x)$  for any two arbitrary functions  $f_1$  and  $f_2$ .

### 4.2 Presented method

The introduced method is the Pseudo-Galerkin method via Legendre's derivative polynomials (PGDL), which expands the unknown function  $y$  as a linear combination of the basis functions as follows:

$$y(x) \approx y_N(x) = \sum_{q=0}^N c_q DL_q(x), \quad (34)$$

for some  $N \in \mathbb{N}$ .

Thus, substitute into (33) to get the residue as follows:

$$R_N(x) = f_1(x) + \sum_{q=0}^N c_q \left[ f_2(x) DL_q(x) + \int_{-1}^1 p_1(x, t) DL_q(t) dt + \int_{-1}^x p_2(x, t) DL_q(t) dt \right] \approx 0. \quad (35)$$

Relations (27) and (28) are used to get the residue function. The integral equation can be written in the form:

$$\begin{aligned} R_N(x) = & f_1(x) + \sum_{q=0}^N c_q [f_2(x) DL_q(x) + a_0(x) (1 + (-1)^q) \\ & + \sum_{r=1}^{m_1} a_r(x) \sum_{k=0}^{\min(r, \lfloor \frac{q+r}{2} \rfloor)} F_{r,k+1,q} (1 + (-1)^{q+r-2k}) \\ & + b_0(x) \sum_{k=0}^2 [\mu(q, k) + (-1)^q \delta_{k,2}] DL_{z(q,k)}(x) \\ & + \sum_{r=1}^{m_2} b_r(x) \sum_{k=0}^{\min(r, \lfloor \frac{q+r}{2} \rfloor)} F_{r,k+1,q} \sum_{v=0}^2 [\mu(q, v) + (-1)^q \delta_{v,2}] \times DL_{z(q,v)}(x) ] \\ = & 0, \end{aligned} \quad (36)$$

where  $z(q, v) = (q + 1 - 2v)(1 - \delta_{v,2})(1 - \delta_{q,0} \delta_{v,1})$ ,  $p_1(x, t) = \sum_{r=0}^{m_1} a_r(x) t^r$ ,  $p_2(x, t) = \sum_{r=0}^{m_2} b_r(x) t^r$ , for some positive integers  $m_1, m_2$ ,  $\{a_r(x), b_r(x)\}$  are real valued functions, and  $\mu(q, v) = \frac{(-1)^v (1 - \delta_{v,2})(1 - \delta_{q,0} \delta_{v,1})}{2q+3}$ .

Consequently, we substitute the independent variable with the set of  $N+1$  Legendre–Gauss–Lobatto (LGL) quadrature points in the interval  $[-1, 1]$  into the residue function to get a system of algebraic equations. By solving the algebraic system to get the unknown coefficients' values  $c_q$ . Hence, the semi-analytic approximate solution is ready.

**Remark 2.** Differential equations can be transformed into corresponding integral equations depending on their types. For example, IVPs can be represented as Volterra integral equations. Meanwhile, BVPs can be transformed into Fredholm integral equations. This transformation involves incorporating the given conditions.

In the next section, we will study the error analysis for the PGDL method in detail to ensure that the presented method is accurate and efficient.

## 5 Convergence and error analysis

In this section, we will study the upper bound for the basis function, the upper bound for the expansion coefficients, and the uniform convergence of the presented method.

**Lemma 3.** The Legendre's derivative polynomials are bounded for  $-1 \leq x \leq 1$  such that

$$|DL_q(x)| \leq \frac{(q+1)(q+2)}{2}. \quad (37)$$

*Proof.* Replace  $q$  by  $q+1$  in (3), with the aid of Definition 1 to get

$$|DL_q(x)| \leq \sum_{j=0}^{\lfloor \frac{q}{2} \rfloor} [2(q-2j)+1]. \quad (38)$$

The left-hand side of the inequality (38) represents an arithmetic series with  $\lfloor \frac{q}{2} \rfloor + 1$  terms, the first term  $2q+1$ , and the common difference  $-4$ . Thus the inequality (38) can be written as

$$|DL_q(x)| \leq \frac{\lfloor \frac{q}{2} \rfloor + 1}{2} \left[ 2(2q+1) - 4\lfloor \frac{q}{2} \rfloor \right]. \quad (39)$$

If  $q$  is even, then  $\lfloor \frac{q}{2} \rfloor = \frac{q}{2}$ , which implies

$$|DL_q(x)| \leq \frac{(q+2)(q+1)}{2}. \quad (40)$$

If  $q$  is odd, then  $\lfloor \frac{q}{2} \rfloor = \frac{q-1}{2}$  and

$$|DL_q(x)| \leq \frac{(q+1)(q+2)}{2}, \quad (41)$$

which completes the proof.  $\square$

In [5], the authors proved that  $|DL_q(x)| \leq (q+1)^2$ . As a comparison between the two upper bounds, we found that the upper bound obtained in this work is found to be better than the upper bound in [5].

**Remark 3.** For simplicity and compactness, some notations are introduced as

$$\Lambda_q(x) = \begin{cases} 1, & q = 0, \\ x, & q = 1, \\ \mathcal{L}_q(x) - \mathcal{L}_{q-2}(x), & q \geq 2, \end{cases} \quad (42)$$

$$\Delta_q(x) = \begin{cases} 1, & q = 0, \\ 3x, & q = 1, \\ DL_q(x) - DL_{q-2}(x), & q \geq 2. \end{cases} \quad (43)$$

In addition, we have the following:

$$\Delta_q(x) = \Lambda'_{q+1}(x), \quad (44)$$

and with the aid of (2), we get

$$\Lambda_{q+2}(x) = \eta_{q+2} \Delta_{q+2}(x) - \eta_q \Delta_q(x), \quad (45)$$

where

$$\eta_q = \frac{1}{q+1}. \quad (46)$$

Also, we need to define  $\alpha_{m,k}$  as

$$\alpha_{m,k} = \sum_{i=0}^{m-1} (-1)^{\lfloor \frac{k}{2^{m-1-i}} \rfloor}, \quad (47)$$

where  $k = 0, 1, 2, \dots, 2^m - 1$ , and  $m \geq 1$ .

Moreover, concerning  $\alpha_{m,k}$ , the following can be concluded:

$$\alpha_{m+1,2k} = \alpha_{m,k} + 1, \quad (48)$$

$$\alpha_{m+1,2k+1} = \alpha_{m,k} - 1, \quad (49)$$

$$|\alpha_{m,k}| \leq m, \quad (50)$$

precisely,  $\alpha_{m,k} \in \{-m, 2-m, 4-m, \dots, m-4, m-2, m\}$ .

**Lemma 4.** For  $x \in [-1, 1]$  and  $m, q = 0, 1, 2, \dots$ , we have

$$\psi'_{m+1,q}(x) = -\psi_{m,q}(x), \quad (51)$$

where

$$\begin{aligned} \psi_{m,q}(x) = & (-1)^{m+\lceil \frac{m}{2} \rceil} \frac{1}{2} \sum_{k=0}^{2^{m+1}-1} \left( (-1)^{\lceil \frac{1+\alpha_{m+1,k}}{2} \rceil} \Delta_{q+1+\alpha_{m+1,k}}(x) \right. \\ & \left. \times \prod_{j=1}^{m+1} \eta_{q+1+\alpha_{j, \lfloor \frac{k}{2^{m+1-j}} \rfloor}} \right). \end{aligned} \quad (52)$$

*Proof.* Equation (52) can be written as two collections of even terms and odd terms as follows:

$$\begin{aligned} \psi_{m,q}(x) = & (-1)^{m+\lceil \frac{m}{2} \rceil} \frac{1}{2} \sum_{k=0}^{2^m-1} \left[ \left( (-1)^{\lceil \frac{1+\alpha_{m+1,2k}}{2} \rceil} \Delta_{q+1+\alpha_{m+1,2k}}(x) \right. \right. \\ & \left. \times \prod_{j=1}^{m+1} \eta_{q+1+\alpha_{j, \lfloor \frac{2k}{2^{m+1-j}} \rfloor}} \right) \\ & \left. + \left( (-1)^{\lceil \frac{1+\alpha_{m+1,2k+1}}{2} \rceil} \Delta_{q+1+\alpha_{m+1,2k+1}}(x) \times \prod_{j=1}^{m+1} \eta_{q+1+\alpha_{j, \lfloor \frac{2k+1}{2^{m+1-j}} \rfloor}} \right) \right]. \end{aligned} \quad (53)$$

Simplify it with the aid of (48, 49) as follows:

$$\begin{aligned} \psi_{m,q}(x) = & (-1)^{m+\lceil \frac{m}{2} \rceil+1} \\ & \times \frac{1}{2} \sum_{k=0}^{2^m-1} \left[ (-1)^{\lceil \frac{\alpha_{m,k}}{2} \rceil} \left( \eta_{q+2+\alpha_{m,k}} \Delta_{q+2+\alpha_{m,k}}(x) \right. \right. \\ & \left. \left. - \eta_{q+\alpha_{m,k}} \Delta_{q+\alpha_{m,k}}(x) \right) \prod_{j=1}^m \eta_{q+1+\alpha_{j, \lfloor \frac{k}{2^{m-j}} \rfloor}} \right]. \end{aligned} \quad (54)$$

From (45) and using the fact that  $\alpha_{n,k}$  and  $n$  are both even or both odd simultaneously, we have

$$\begin{aligned} \psi_{m,q}(x) = & (-1)^{m+\lceil \frac{m-1}{2} \rceil} \frac{1}{2} \\ & \times \sum_{k=0}^{2^m-1} \left[ (-1)^{\lceil \frac{1+\alpha_{m,k}}{2} \rceil} \Delta_{q+2+\alpha_{m,k}}(x) \prod_{j=1}^m \eta_{q+1+\alpha_{j, \lfloor \frac{k}{2^{m-j}} \rfloor}} \right] \end{aligned} \quad (55)$$

By replacing  $m$  with  $m+1$  in (55) and differentiating with respect to  $x$ , we have

$$\psi'_{m+1,q}(x) = (-1)^{m+1+\lceil \frac{m}{2} \rceil} \frac{1}{2} \sum_{k=0}^{2^{m+1}-1} \left[ (-1)^{\lceil \frac{1+\alpha_{m+1,k}}{2} \rceil} \Lambda'_{q+2+\alpha_{m+1,k}}(x) \right. \\ \left. \times \prod_{j=1}^{m+1} \eta_{q+1+\alpha_{j, \lfloor \frac{k}{2^{m+1-j}} \rfloor}} \right], \quad (56)$$

which completes the proof.  $\square$

**Lemma 5.** For  $x \in [-1, 1]$  and  $q = 0, 1, 2, \dots$ , we have

$$(1-x^2) DL_q(x) = \frac{2(q+1)(q+2)}{2q+3} \psi_{0,q}(x). \quad (57)$$

*Proof.* Equation (4) can be written as

$$(1-x^2) DL_q(x) = -\frac{(q+1)(q+2)}{2q+3} \Lambda_{q+2}(x), \quad (58)$$

where  $\Lambda_q$  is defined as in (42).

Use (45) to get

$$(1-x^2) DL_q(x) = -\frac{(q+1)(q+2)}{2q+3} [\eta_{q+2} \Delta_{q+2}(x) - \eta_q \Delta_q(x)]. \quad (59)$$

Then, (52) completes the proof.  $\square$

**Lemma 6.** For nonnegative integers  $q$  and  $r$ , we have

$$\left| \prod_{j=1}^{r+1} \eta_{q+1+\alpha_{j, \lfloor \frac{k}{2^{r+1-j}} \rfloor}} \right| \leq \begin{cases} \frac{1}{2^{r+1}(q-r)^{r+1}}, & q \geq r+1, \\ \frac{1}{(2r+1)!!}, & q = r. \end{cases} \quad (60)$$

*Proof.* The first case,  $q \geq r+1$ , from (46), we have

$$|\eta_q| \leq \frac{1}{2^q}, q \neq 0, \quad (61)$$

while  $\eta_0 = 1$ .

Consequently,

$$\max\{\eta_{q_1}, \eta_{q_2}\} = \eta_{\min\{q_1, q_2\}}. \quad (62)$$

Those relations and inequalities, (50), (61), and (62), can be used to get

$$\left| \prod_{j=1}^{r+1} \eta_{q+1+\alpha_{j, \lfloor \frac{k}{2^{r+1-j}} \rfloor}} \right| \leq \frac{1}{2^{r+1}(q-r)^{r+1}}. \quad (63)$$

While for the second case,  $q = r$ , we have

$$\left| \prod_{j=1}^{r+1} \eta_{r+1+\alpha_{j, \lfloor \frac{k}{2^{r+1}-j} \rfloor}} \right| \leq \left| \prod_{j=1}^{r+1} \eta_{r+1-j} \right| = \frac{1}{(2r+1)!!}, \quad (64)$$

which completes the proof.  $\square$

**Lemma 7.** For nonnegative integers  $q$  and  $r$ , we have

$$|\Delta_{q+1+\alpha_{r+1,k}}(x)| \leq \begin{cases} 31r^2 (q-r)^2, & q \geq r+1, \\ 21r^2, & q = r. \end{cases} \quad (65)$$

where  $\Delta_q$  is defined as (43).

*Proof.* Use (43) and Lemma 3 to get

$$\begin{aligned} |\Delta_q(x)| &= |DL_q(x) - DL_{q-2}(x)| \leq |DL_q(x)| + |DL_{q-2}(x)| \\ &\leq \frac{(q+1)(q+1)}{2} + \frac{(q-1)(q)}{2} = q^2 + q + 1. \end{aligned} \quad (66)$$

Thus,

$$\max\{|\Delta_{q_1}(x)|, |\Delta_{q_2}(x)|\} \leq q^2 + q + 1, \quad (67)$$

such that  $q = \max\{q_1, q_2\}$ .

For  $q \geq r+1$ , we can use (67) together with (50) to calculate

$$\begin{aligned} |\Delta_{q+1+\alpha_{r+1,k}}(x)| &\leq (q+r+2)^2 + (q+r+2) + 1 \\ &\leq (q-r)^2 + (4r+5)(q-r)^2 + (4r^2+10r+7)(q-r)^2 \\ &\leq 31r^2 (q-r)^2. \end{aligned} \quad (68)$$

While for  $q = r$ ,

$$|\Delta_{r+1+\alpha_{r+1,k}}(x)| \leq (2r+2)^2 + (2r+2) + 1 \leq 21r^2, \quad (69)$$

which completes the proof.  $\square$

In Lemmas 6 and 7, there is no need to discuss the  $q < r$  case, since the index will be a negative value, which is not defined and will not be used.

The next theorem shows that the spectral expansion's constants are bounded.

**Theorem 5.** Let  $y(x) \in C^r[-1, 1]$  with bounded  $r$ th-derivative, where  $r \geq 2$ , and consider the expansion  $y(x) = \sum_{q=0}^{\infty} c_q DL_q(x)$ . Then

$$|c_q| \leq \begin{cases} \frac{31Mr^2}{(q-r)^{r-1}}, & q \geq r+1, \\ \frac{21Mr^2 2^{r+1}}{(2r+1)!!}, & q = r, \end{cases} \quad (70)$$

where  $M \in \mathbb{R}$  such that  $|y^{(r)}(x)| \leq M$ .

*Proof.* Let  $y \in C^r[-1, 1]$ , where  $r \geq 2$  and  $|y^{(r)}(x)| \leq M$  for some  $M \in \mathbb{R}$ .

Consider the expansion:

$$y(x) = \sum_{q=0}^{\infty} c_q DL_q(x). \quad (71)$$

From the orthogonality relation (13) and Lemmas 5 and 4, we get

$$\begin{aligned} c_q &= \frac{2q+3}{2(q+2)(q+1)} \int_{-1}^1 (1-x^2) DL_q(x) y(x) dx \\ &= - \int_{-1}^1 \psi'_{1,q}(x) y(x) dx. \end{aligned} \quad (72)$$

Using integration by parts for (72) and applying Lemma 4, we get

$$c_q = M_{1,q} - \int_{-1}^1 \psi'_{2,q}(x) y'(x) dx, \quad (73)$$

where  $M_{1,q} = -\psi_{1,q}(1)y(1) + \psi_{1,q}(-1)y(-1)$ .

Apply integration by parts for the second time to get

$$c_q = M_{1,q} + M_{2,q} + \int_{-1}^1 \psi_{2,q}(x) y''(x) dx, \quad (74)$$

where  $M_{2,q} = -\psi_{2,q}(1)y'(1) + \psi_{2,q}(-1)y'(-1)$ .

Repeat the above steps for  $r-2$  times to get

$$c_q = \sum_{i=1}^r M_{i,q} + \int_{-1}^1 \psi_{r,q}(x) y^{(r)}(x) dx, \quad (75)$$

where  $M_{i,q}$  is can be estimated as

$$M_{i,q} = -\psi_{i,q}(1)y^{(i-1)}(1) + \psi_{i,q}(-1)y^{(i-1)}(-1). \quad (76)$$

Using (8), we have

$$\Lambda_q(\pm 1) = \mathcal{L}_q(\pm 1) - \mathcal{L}_{q-2}(\pm 1) = (\pm 1)^q - (\pm 1)^{q-2} = 0, \quad (77)$$

for  $q \geq 2$ .

Thus, from (55), we get

$$\psi_{m,q}(\pm 1) = \begin{cases} 0, & \text{for } q \geq m, \\ (-1)^{m + \lceil \frac{m-1}{2} \rceil + \lceil \frac{b_{m,q}-q-1}{2} \rceil} \Lambda_{b_{m,q}}(\pm 1) \frac{1}{2} \mathcal{A}, & \text{for } 0 \leq q < m, \end{cases} \quad (78)$$

where

$$b_{m,q} = \begin{cases} 0, & m - q \text{ is even,} \\ 1, & m - q \text{ is odd,} \end{cases} \quad (79)$$

and

$$\mathcal{A} = \sum_{\substack{k=0 \\ \alpha_{m,k} = -q-2+b_{m,q}}}^{2^m-1} \prod_{j=1}^m \eta_{q+1+\alpha_{j, \lfloor \frac{k}{2^{m-j}} \rfloor}}. \quad (80)$$

For  $q \geq r+1$ , use (76), (77), and (78) to conclude  $\sum_{i=1}^r M_{i,q} = 0$ .

To calculate the upper bound of  $\psi_{r,q}(x)$  for  $q \geq r+1$ , we apply Lemmas 6 and 7 to (52):

$$\begin{aligned} |\psi_{r,q}(x)| &\leq \frac{1}{2} \sum_{k=0}^{2^{r+1}-1} |\Delta_{q+1+\alpha_{r+1,k}}(x)| \left| \prod_{j=1}^{r+1} \eta_{q+1+\alpha_{j, \lfloor \frac{k}{2^{r+1-j}} \rfloor}} \right| \\ &\leq \frac{1}{2} \sum_{k=0}^{2^{r+1}-1} 31r^2 (q-r)^2 \frac{1}{2^{r+1} (q-r)^{r+1}} = \frac{31r^2}{2(q-r)^{r-1}}. \end{aligned} \quad (81)$$

Hence,

$$|c_q| \leq \int_{-1}^1 |\psi_{r,q}(x)| |y^{(r)}(x)| dx \leq \frac{31Mr^2}{(q-r)^{r-1}}. \quad (82)$$

For the second case,  $q = r$ , the upper bound for  $\psi_{r,q}(x)$  can be calculated using Lemmas 6 and 7:

$$\begin{aligned}
|\psi_{r,r}(x)| &\leq \frac{1}{2} \sum_{k=0}^{2^{r+1}-1} |\Delta_{r+1+\alpha_{r+1,k}}(x)| \left| \prod_{j=1}^{r+1} \eta_{r+1+\alpha_j, \lfloor \frac{k}{2^{r+1-j}} \rfloor} \right| \\
&\leq \frac{1}{2} \sum_{k=0}^{2^{r+1}-1} \frac{21r^2}{(2r+1)!!} = \frac{21r^2 2^r}{(2r+1)!!}.
\end{aligned} \tag{83}$$

From (75) and (83), we have

$$|c_r| \leq \int_{-1}^1 |\psi_r(x)| |y^{(r)}(x)| dx \leq \frac{21Mr^2 2^{r+1}}{(2r+1)!!}. \tag{84}$$

□

**Theorem 6.** Let  $y(x)$  satisfy the conditions of Theorem 5. Then

$$|y(x) - y_N(x)| \lesssim \mathcal{O}\left(\frac{1}{N-r}\right)^{r-4}, \quad N > r > 4, \tag{85}$$

where  $y_N(x)$  is shown in (34).

*Proof.* From (34), we have

$$|y(x) - y_N(x)| = \left| \sum_{q=N+1}^{\infty} c_q DL_q(x) \right| \leq \sum_{q=N+1}^{\infty} |c_q| |DL_q(x)|. \tag{86}$$

Use Theorem 5 for  $q \geq r+1$  and Lemma 3 to get

$$\begin{aligned}
|y(x) - y_N(x)| &\leq \frac{31Mr^2}{2} \sum_{q=N+1}^{\infty} \frac{(q+1)(q+2)}{(q-r)^{r-1}} \\
&= \frac{31Mr^2}{2} \sum_{q=N+1}^{\infty} \left( \frac{r^2+3r+2}{(q-r)^{r-1}} + \frac{2r+3}{(q-r)^{r-2}} + \frac{1}{(q-r)^{r-3}} \right) \\
&\leq \frac{31Mr^2}{2} \sum_{q=N+1}^{\infty} \frac{12r^2}{(q-r)^{r-3}}.
\end{aligned} \tag{87}$$

For any decreasing positive function  $F(q)$ , we have  $\left| \sum_{q=N+1}^{\infty} F(q) \right| \leq \int_N^{\infty} F(q) dq$ , (see [40]). Hence,

$$|y(x) - y_N(x)| \leq \frac{31Mr^2}{2} \int_N^{\infty} \frac{12r^2}{(q-r)^{r-3}} dq \leq \frac{186Mr^4}{(r-4)(N-r)^{r-4}}, \tag{88}$$

which completes the proof. □

Since  $q > N > r$ , which means  $q \neq r$ , so, only the case of  $q \geq r + 1$  is needed, and there is no need for the  $q = r$  case.

In the next section, some test problems will be solved to clarify the accuracy and efficiency of the presented method.

## 6 Examples

This section will apply the presented method to approximate some types of integral equations. Also, a physical application, Lane–Emden, modeled by a value problem, has been approximated. Three error metrics have been calculated to show the accuracy and efficiency of the presented methods: the maximum absolute error (MAE), point-wise absolute error, and root square error ( $RSE_N$ ).

**Test Problem 1.** Consider the following linear Fredholm integral equation [6]:

$$y(x) = \frac{1}{4} - x + \int_0^1 (3t - 6x^2) y(t) dt, \quad (89)$$

where  $x \in [0, 1]$ , with the exact solution  $y = x^2 - x$ . Apply the PGDL method by expanding the unknown function to be

$$y_2(x) = \sum_{k=0}^2 c_k DL_k(x). \quad (90)$$

Shift the given problem to  $[-1, 1]$  to get the residual as follows:

$$\begin{aligned} \frac{1}{4} + \frac{1}{2}x + \sum_{k=0}^2 c_k \left( DL_k(x) - \frac{3}{4} \int_{-1}^1 t DL_k(t) dt \right. \\ \left. + \frac{3}{4} (x^2 + 2x) \int_{-1}^1 DL_k(t) dt \right) = 0. \end{aligned} \quad (91)$$

From (23) and Theorem 4, we get

$$\sum_{k=0}^2 c_k \left( DL_k(x) - \frac{3}{4} (1 + (-1)^{k+1}) + \frac{3}{4} (x^2 + 2x) (1 + (-1)^k) \right) = -\frac{1}{2}x - \frac{1}{4}. \quad (92)$$

Collocate the above residue by the three LGL points,  $x_0 = -1, x_1 = 0, x_2 = 1$  to get

$$\begin{aligned} -\frac{1}{2}c_0 - \frac{9}{2}c_1 + \frac{9}{2}c_2 &= \frac{1}{4}, \\ c_0 - \frac{3}{2}c_1 - \frac{3}{2}c_2 &= -\frac{1}{4}, \\ \frac{11}{2}c_0 + \frac{3}{2}c_1 + \frac{21}{2}c_2 &= -\frac{3}{4}. \end{aligned} \quad (93)$$

Solving the system (93) to get  $c_0 = -\frac{1}{5}, c_1 = 0, c_2 = \frac{1}{30}$ , which is the exact solution in the domain  $[-1, 1]$ .

**Test Problem 2.** Consider the following linear Volterra–Fredholm integral equation [49]:

$$2y(x) - \int_0^1 (x+t)y(t) dt - \int_0^x (x-t)y(t) dt = \frac{1}{12}x^4 - \frac{1}{6}x^3 - \frac{5}{2}x^2 + \frac{5}{6}x + \frac{17}{12}, \quad (94)$$

where  $x \in [0, 1]$  and the exact solution is  $y = -x^2 + x + 1$ .

Applying the PGDL method with  $N = 2$ , after shifting the domain of (94) will be expanded as follows:

$$\begin{aligned} \sum_{q=0}^2 c_q \left[ 2DL_q(x) - \frac{x+2}{4} \int_{-1}^1 DL_q(t) dt - \frac{1}{4} \int_{-1}^1 t DL_q(t) dt \right. \\ \left. - \frac{x}{4} \int_{-1}^x DL_q(t) dt + \frac{1}{4} \int_{-1}^x t DL_q(t) dt \right] \\ = \frac{1}{12} \left( \frac{x+1}{2} \right)^4 - \frac{1}{6} \left( \frac{x+1}{2} \right)^3 - \frac{5}{2} \left( \frac{x+1}{2} \right)^2 \\ + \frac{5}{6} \left( \frac{x+1}{2} \right) + \frac{17}{12}. \end{aligned} \quad (95)$$

There are four integrals; the first integral can be computed easily from (23) as

$$\int_{-1}^1 DL_q(t) dt = 1 + (-1)^q. \quad (96)$$

The second one is from Theorem 4 to be

$$\int_{-1}^1 t DL_q(t) dt = 1 + (-1)^{q+1}. \quad (97)$$

While the third integral can be computed from (20) to get

$$\int_{-1}^x DL_q(t) dt = \sum_{k=0}^2 \left( \frac{(-1)^k (1 - \delta_{k,2}) (1 - \delta_{q,0} \delta_{k,1})}{2q+3} + (-1)^q \delta_{k,2} \right) \times DL_{(q+1-2k)(1-\delta_{k,2})(1-\delta_{q,0} \delta_{k,1})}(x). \quad (98)$$

Finally, for the last one, using (27) at  $m = 1$ , hence the integration can be determined as the previous one to get the following:

$$\begin{aligned} \int_{-1}^x t DL_q(x) &= \sum_{j=1}^{\min(2, \lfloor \frac{q+3}{2} \rfloor)} F_{1,j,q} \\ &\times \sum_{k=0}^2 \left( \frac{(-1)^k (1 - \delta_{k,2}) (1 - \delta_{q+3-2j,0} \delta_{k,1})}{2q+9-4j} + (-1)^{q+3-2j} \delta_{k,2} \right) \\ &\times DL_{(q+4-2k-2j)(1-\delta_{k,2})(1-\delta_{q+3-2j,0} \delta_{k,1})}(x). \end{aligned}$$

Thus, (95) takes the form

$$\begin{aligned} &\sum_{q=0}^2 c_q \left[ 2 DL_q(x) - \frac{x+2}{4} (1 + (-1)^q) - \frac{1}{4} (1 + (-1)^{q+1}) \right. \\ &- \frac{x}{4} \sum_{k=0}^2 \left( \frac{(-1)^k (1 - \delta_{k,2}) (1 - \delta_{q,0} \delta_{k,1})}{2q+3} + (-1)^q \delta_{k,2} \right) \\ &\times DL_{(q+1-2k)(1-\delta_{k,2})(1-\delta_{q,0} \delta_{k,1})}(x) \\ &+ \frac{1}{4} \sum_{j=1}^{\min(2, \lfloor \frac{q+3}{2} \rfloor)} F_{1,j,q} \sum_{k=0}^2 \left( \frac{(-1)^k (1 - \delta_{k,2}) (1 - \delta_{q+3-2j,0} \delta_{k,1})}{2q+9-4j} \right. \\ &\left. \left. + (-1)^{q+3-2j} \delta_{k,2} \right) DL_{(q+4-2k-2j)(1-\delta_{k,2})(1-\delta_{q+3-2j,0} \delta_{k,1})}(x) \right] \\ &= \frac{1}{12} \left( \frac{x+1}{2} \right)^4 - \frac{1}{6} \left( \frac{x+1}{2} \right)^3 - \frac{5}{2} \left( \frac{x+1}{2} \right)^2 + \frac{5}{6} \left( \frac{x+1}{2} \right) + \frac{17}{12}. \end{aligned} \quad (100)$$

Substitute by three LGL points  $x_0 = -1, x_1 = 0, x_2 = 1$  to get a system of algebraic equations and get the values of the spectral constants as  $c_0 = \frac{6}{5}, c_1 = 0, c_2 = -\frac{1}{30}$ . So, the solution is

$$y_2(x) = \frac{6}{5} (1) - \frac{1}{30} \left( \frac{15}{2} x^2 - \frac{3}{2} \right) = \frac{5-x^2}{4}, \quad (101)$$

which is the exact solution on the domain  $[-1, 1]$ .

**Test Problem 3.** Consider the stable population model, which is a Volterra integral equation that describes the number of female births [4]:

$$y(x) = e^x - \int_0^x (x-t)y(t)dt, \quad (102)$$

where  $x \in [0, 1]$ ,  $(x-t)$  is the net maternity function of females class age  $t$  at time  $x$ , and  $e^x$  is the contribution of birth due to female already present at time  $x$ . The exact solution is  $y(x) = \frac{1}{2}[e^x + \cos x + \sin x]$ .

Table 1 presents the best maximum absolute error for PGDL method compared to other methods. Table 2 shows MAE and  $RSE_N$  for different values of  $N$ , where

$$RSE_N = \sqrt{\sum_{i=0}^N (y_{\text{exact}}(x_i) - y_{\text{approximate}}(x_i))^2}, \quad (103)$$

such that the points  $x_i$ 's have been chosen to be LGL quadrature points.

Table 1: MAE for Example 3 compared to other methods.

Method	Best MAE
[50]	2.14E-14
[51]	1.25E-15
[2]	4.44E-16
PGDL Method	7.13E-18

Table 2: MAE and  $RSE_N$  at different values of  $N$  for Example 3.

N	MAE	RSE
2	6.52E-03	6.23E-04
4	2.55E-05	2.97E-07
6	2.36E-08	7.70E-11
8	3.18E-11	4.95E-14
10	1.11E-14	9.72E-18
12	7.13E-18	3.97E-21

Figure 1 shows the log error, which confirms the stability of the presented method.

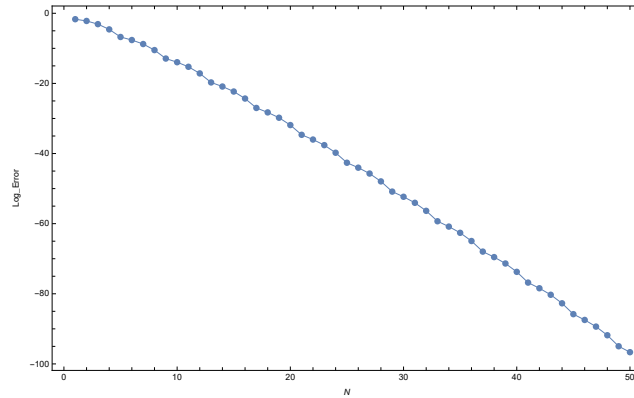


Figure 1: Log error graph for Example 3.

**Test Problem 4.** Consider the following linear Fredholm type integral equation [49]:

$$y(x) + \int_0^1 e^{x-t} y(t) dt = e^x, \quad (104)$$

whose exact solution is given by  $y(x) = \frac{1}{2}e^x$ , where  $x \in [0, 1]$ . In this case, the exponential within the integral part will be expanded using Taylor's expansion.

Using a similar procedure for using PGDL to get the following residual:

$$R_N(x) = \sum_{q=0}^N c_q \left[ DL_q(x) + \frac{1}{2}e^{\frac{x}{2}} \sum_{k=0}^{13} \frac{(-1)^k}{2^k k!} \times \sum_{j=1}^{\min(k+1, \lfloor \frac{q+k+2}{2} \rfloor)} F_{k,j,q} \left( 1 + (-1)^{q+k+2-2j} \right) - e^{\frac{x+1}{2}} \right] \quad (105)$$

Table 3 shows  $RSE_N$  and MAE for various values of  $N$ .

In Figure 2, we can track the log error from  $N = 1$  to  $N = 15$ , which shows the stability of the presented approximate solution.

Table 3:  $\text{RSE}_N$  and MAE for Example 4 at different values of  $N$ .

$N$	[49] RSE	PGDL Method	
		RSE	MAE
5	1.59E-07	6.97E-11	8.58E-07
6	2.29E-09	1.28E-13	3.03E-08
7	2.09E-10	1.03E-16	9.36E-10
8	2.65E-12	7.20E-17	2.58E-11
9	1.98E-13	7.57E-17	6.42E-13
10	2.27E-15	7.93E-17	1.45E-14

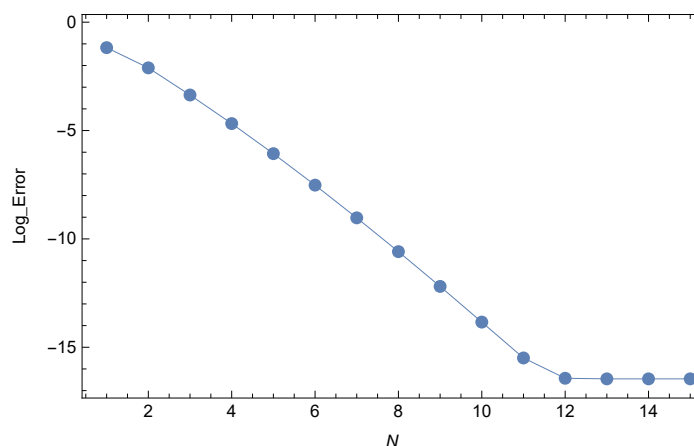


Figure 2: Log error graph for Example 4.

The next example will attempt to approximate the solution of the IVP.

**Test Problem 5.** Consider the following  $2^{nd}$  order Lane–Emden differential equation [2]:

$$y''(x) + \frac{2}{x}y'(x) + y^m(x) = 0, \quad 0 \leq m \leq 5, \quad x \in [0, 3.1], \quad (106)$$

with the following initial conditions,  $y(0) = 1$ , and  $y'(0) = 0$ . While the exact solution at  $m = 1$  is  $y(x) = \frac{\sin x}{x}$ .

Integrate the IVP (106) over the interval  $[0, x]$  transforms into the following Volterra integral equation:

$$x y(x) - x + \int_0^x (x-t) t y^m(t) dt = 0. \quad (107)$$

Table 4 shows the point-wise error compared to the method in [2], which ensures the accuracy and efficiency of our method. On the other hand, Table 5 presents the MAE and  $RSE_N$  for various values of  $N$ .

Table 4: Point-wise absolute error for Example 5.

$x$	[2]	PGDL Method	
	$N = 15$	$N = 14$	$N = 15$
0.0	4.44E-16	1.11E-13	5.97E-16
0.1	2.88E-15	1.33E-14	5.42E-17
0.2	5.77E-15	6.38E-15	3.45E-17
0.3	6.77E-15	2.09E-15	6.17E-18
0.4	6.99E-15	4.33E-15	1.67E-17
0.5	7.54E-15	1.92E-15	4.26E-18
0.6	7.66E-15	2.38E-15	1.12E-17
0.7	7.32E-15	2.67E-15	1.26E-18
0.8	7.43E-15	1.06E-16	6.77E-18
0.9	7.32E-15	2.20E-15	4.69E-18
1.0	7.10E-15	1.72E-15	1.55E-18
1.5	5.88E-15	1.33E-15	8.32E-19
2.0	4.05E-15	1.71E-16	2.16E-19
2.5	2.30E-15	5.73E-16	1.17E-18
3.0	6.45E-16	4.44E-16	1.07E-18
3.1	4.59E-16	3.99E-28	7.85E-31

Table 5: MAE and  $RSE_N$  at different values of  $N$  for Example 5.

$N$	MAE	RSE
5	2.63E-04	2.94E-04
7	2.70E-06	3.04E-06
9	1.74E-08	1.96E-08
11	7.64E-11	8.62E-11
13	2.45E-13	2.76E-13
15	5.97E-16	6.75E-16

Figure 3 represents the log error from  $N = 1$  to  $N = 30$ , which verifies the stability of PGDL method.

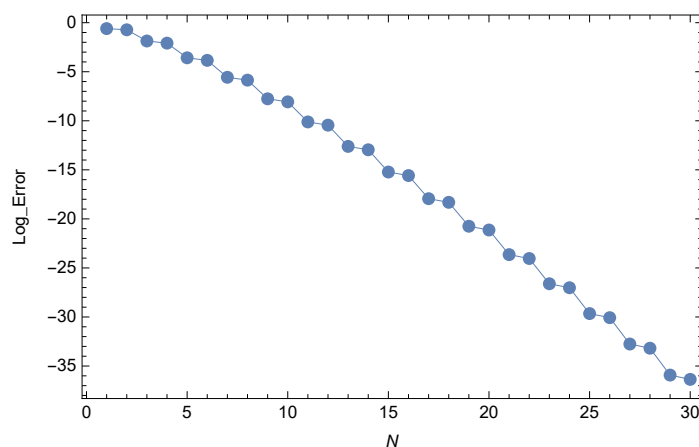


Figure 3: Log error graph for Example 5.

## 7 Concluding remarks

Through this work, we introduced the operational integration matrix for Legendre's derivative polynomials. Essential relations have been investigated, including the moment relation. Also, a technique that solves some types of integral equations has been presented without the need to do any actual integration. Furthermore, we have studied the error analysis and achieved a better upper bound for Legendre's derivative polynomials. In addition, the investigated matrix and technique are applied to approximate some types of differential equations. Some numerical test problems have been solved, and they show the accuracy, efficiency, and stability of the presented method. As a future direction, our results encourage applying our method to more complicated problems, such as nonlinear integral equations, systems of integral equations, and two-dimensional integral equations.

## Conflict of interest

The authors declare that they have no conflict of interest.

## Funding

The authors received no financial support for the research.

## Data availability

No data is associated with this research.

## Author Contributions Statement

YHY, AMA, and MA conducted the mathematical analysis, developed the methodology, verified the results, wrote the initial draft, and reviewed the final version. AMA contributed to the original manuscript, software development, and methodology. ME reviewed and edited the final version of the manuscript.

## Acknowledgements

We would like to express our gratitude to the anonymous reviewers for their valuable feedback and insightful comments, which greatly contributed to the improvement of this manuscript. Their expertise and dedication are deeply appreciated.

## References

- [1] Abdelhakem, M. *Shifted Legendre fractional pseudo-spectral integration matrices for solving fractional Volterra integro-differential equations and Abel's integral equations*, FRACTALS (fractals), 31, (10) (2023) 1–11.
- [2] Abdelhakem, M., Alaa-Eldeen, T., Baleanu, D., Alshehri, M.G. and El-Kady, M. *Approximating real-life BVPs via Chebyshev polynomials' first derivative pseudo-Galerkin method*, Fractal Fract., 5 (4) (2021) 165.

- [3] Abdelhakem, M., Fawzy, M., El-Kady, M. and Moussa, H. *Legendre polynomials' second derivative tau method for solving Lane–Emden and Ricatti equations*, Appl. Math. Inf. Sci., 17 (3) (2023) 437–445.
- [4] Abdelhakem M. and Moussa, H. *Pseudo-spectral matrices as a numerical tool for dealing bvps, based on Legendre polynomials' derivatives*, Alexandria Eng. J., 66, (2023) 301–313.
- [5] Abdelhakem M. and Youssri, Y. *Two spectral Legendre's derivative algorithms for Lane–Emden, bratu equations, and singular perturbed problems*, Appl. Numer. Math., 169, (2021) 243–255.
- [6] Abd-Elhameed W. and Youssri, Y. *Numerical solutions for Volterra–Fredholm–Hammerstein integral equations via second kind Chebyshev quadrature collocation algorithm*, Adv. Math. Sci. Appl., 24, (2014) 129–141.
- [7] Adel, W. and Sabir, Z. *Solving a new design of nonlinear second-order Lane–Emden pantograph delay differential model via Bernoulli collocation method*, Eur. Phys. J. Plus, 135 (5) (2020) 1–12.
- [8] Adel, W., Sabir, Z., Rezazadeh, H. and Aldurayhim, A. *Application of a novel collocation approach for simulating a class of nonlinear third-order Lane–Emden model*, Math. Prob. Eng., 2022 (1) (2022) 5717924.
- [9] Algazaa, S.A.T. and Saeidian, J. *Spectral methods utilizing generalized Bernstein-like basis functions for time-fractional advection–diffusion equations*, Math. Method. Appl. Sci., 2025.
- [10] Alsedaiss, N., Mansour, M.A., Aly, A.M., Abdelsalam, S.I. *Artificial neural network validation of MHD natural bioconvection in a square enclosure: entropic analysis and optimization*, Acta Mechanica Sinica, 41 (2025) 724507.
- [11] Buzov, A., Klyuev, D., Kopylov, D. and Nescheret, A. *Mathematical model of a two-element microstrip radiating structure with a chiral meta-material substrate*, J. Commun. Technol. Electron., 65 (2020) 414–420.

- [12] Doha, E., Youssri, Y. and Zaky, M. *Spectral solutions for differential and integral equations with varying coefficients using classical orthogonal polynomials*, Bull. Iran. Math. Soc., 45, (2019) 527–555.
- [13] Fageehi, Y.A. and Alshoaibi, A.M. *Investigating the influence of holes as crack arrestors in simulating crack growth behavior using finite element method*, Appl. Sci., 14 (2) (2024) 897.
- [14] Farooq, U., Khan, H., Tchier, F., Hincal, E., Baleanu, D., and BinJe-breen, H. *New approximate analytical technique for the solution of time fractional fluid flow models*, Adv. Differ. Equ., 2021, (2021) 1–20.
- [15] Fawzy, M., Moussa, H., Baleanu, D., El-Kady, M. and Abdelhakem, M. *Legendre derivatives direct residual spectral method for solving some types of ordinary differential equations*, Math. Sci. Lett., 11 (3) (2022) 103–108.
- [16] Gamal, M., El-Kady, M. and Abdelhakem, M. *Solving real-life BVPS via the second derivative Chebyshev pseudo-Galerkin method*, Int. J. Mod. Phys. C (IJMPC), 35 (07) (2024) 1–20.
- [17] Ghalini, R.G., Hesameddini, E. and Dastjerdi, H.L. *An efficient spectral collocation method for solving volterra delay integral equations of the third kind*, J. Comput. Appl. Math., 454, (2025) 116138.
- [18] Ghayoor, M., Abbasi, W.S. Majeed, A.H., Alotaibi, H. and Ali, A.R. *Interference effects on wakes of a cluster of pentad square cylinders in a crossflow: A lattice Boltzmann study*, AIP Advances, 14 (12) (2024) 2024.
- [19] Gowtham, K. and Gireesha, B. *Associated Laguerre wavelets: Efficient method to solve linear and nonlinear singular initial and boundary value problems*, Int. J. Appl. Comput. Math., 11 (16) (2025) 2025.
- [20] Hafez, R. and Youssri, Y. *Spectral Legendre-Chebyshev treatment of 2d linear and nonlinear mixed Volterra-Fredholm integral equation*, Math. Sci. Lett., 9 (2) (2020) 37–47.

- [21] Hamza, M.M., Sheriff, A., Isah, B.Y. and Bello, A. *Nonlinear thermal radiation effects on bioconvection nano fluid flow over a convectively heated plate*, Int. J. Non-Linear Mech., 171 (2025) 105010.
- [22] Il'inskii, A. and Galishnikova, T. *Integral equation method in problems of electromagnetic-wave reflection from inhomogeneous interfaces between media*, J. Commun. Technol. Electron., 61, (2016) 981–994.
- [23] Khan, I., Chinyoka, T., Ismail, E.A., Awwad, F.A. and Ahmad, Z. *MHD flow of third-grade fluid through a vertical micro-channel filled with porous media using semi implicit finite difference method*, Alexandria Eng. J., 86 (2024) 513–524.
- [24] Kumar, S., Shaw, P.K., Abdel-Aty, A.-H. and Mahmoud, E.E. *A numerical study on fractional differential equation with population growth model*, Numer. Methods Partial Differ. Equ., 40 (1) (2024) e22684.
- [25] Li, K., Xiao, L., Liu, M. and Kou, Y. *A distributed dynamic load identification approach for thin plates based on inverse finite element method and radial basis function fitting via strain response*, Eng. Struct., 322, (2025) 119072.
- [26] Lighthill, M.J. *Contributions to the theory of heat transfer through a laminar boundary layer*, Proc. R. Soc. A: Math. Phys. Eng. Sci., 202 (1070) (1950) 359–37.
- [27] Mahmoudi, Z., Khalsaraei, M.M., Sahlan, M.N. and Shokri, A. *Laguerre wavelets spectral method for solving a class of fractional order PDEs arising in viscoelastic column modeling*, Chaos. Soliton. Fract., 192, (2025) 116010.
- [28] Mobarak, H.M., Abo-Eldahab, E.M., Adel, R. and Abdelhakem, M. *Mhd 3d nanofluid flow over nonlinearly stretching/shrinking sheet with nonlinear thermal radiation: Novel approximation via Chebyshev polynomials' derivative pseudo-Galerkin method*, Alex. Eng. J., 102 (2024) 119–131.

- [29] Mohammad, M. and Trounev, A. *Implicit Riesz wavelets based-method for solving singular fractional integro-differential equations with applications to hematopoietic stem cell modeling*, Chaos Solitons Fract., 138 (2020) 109991.
- [30] PraveenKumar, P., Balakrishnan, S., Magesh, A., Tamizharasi, P. and Abdelsalam, S.I. *Numerical treatment of entropy generation and Bejan number into an electroosmotically-driven flow of Sutterby nanofluid in an asymmetric microchannel* Numer. Heat Trans. Part B: Fund., 2024 (2024) 1–20.
- [31] Rahmani, S., Baiges, J. and Principe, J. *Anisotropic variational mesh adaptation for embedded finite element methods*, Comput. Method. Appl. Mech. Engin., 433 (2025) 117504.
- [32] Raza, R., Naz, R., Murtaza, S. and Abdelsalam, S.I. *Novel nanostructural features of heat and mass transfer of radiative Carreau nanoliquid above an extendable rotating disk*, Int. J. Mod. Phys. B, 38 (30) (2024) 2450407.
- [33] Sabir, Z., Raja, M.A.Z. and Baleanu, D. *Fractional mayer neuro-swarm heuristic solver for multi-fractional order doubly singular model based on Lane–Emden equation*, Fractals, 29 (05) (2021) 2140017.
- [34] Sadiq, M., Shahzad, H., Alqahtani, H., Tirth, V., Algahtani, A., Irshad, K. and Al-Mughanham, T. *Prediction of cattaneo–christov heat flux with thermal slip effects over a lubricated surface using artificial neural network*, Eur. Phys. J. Plus, 139 (9) (2024) 851.
- [35] Sadri, K., Amilo, D., Hosseini, K., Hincal, E. and Seadawy, A.R. *A Tau-Gegenbauer spectral approach for systems of fractional integro-differential equations with the error analysis*, AIMS Math., 9 (2) (2024) 3850–3880.
- [36] 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*, Iranian Journal of Numerical Analysis and Optimization, 14 (4) (2024) 1309–1335.

- [37] Santamaría, G., Valverde, J., Pérez-Aloe, R., and Vinagre, B. *Microelectronic implementations of fractional-order integro-differential operators*, Comput. Nonlinear Dyn., 3 (2) (2009) 021301.
- [38] Shahmorad, S., Ostadzad, M. and Baleanu, D. *A tau-like numerical method for solving fractional delay integro-differential equations*, Appl. Numer. Math., 151, (2020) 322–336.
- [39] Shen, J., Tang, T. and Wang, L.-L. *Spectral methods: Algorithms, analysis and applications*, 41. Springer Science & Business Media, 2011.
- [40] Stewart, J. *Essential calculus: Early transcendentals*. Brooks/Cole, a part of the Thomson Corporation, 2007.
- [41] Sweis, H. Arqub, O.A. and Shawagfeh, N. *Fractional delay integro-differential equations of nonsingular kernels: Existence, uniqueness, and numerical solutions using Galerkin algorithm based on shifted Legendre polynomials*, Int. J. Modern Phys. C, 34 (04) (2023) 2350052.
- [42] Sweis, H., Arqub, O.A., and Shawagfeh, N. *Hilfer fractional delay differential equations: Existence and uniqueness computational results and pointwise approximation utilizing the shifted-Legendre Galerkin algorithm*, Alexandria Eng. J., 81 (2023) 548–559.
- [43] Sweis, H., Arqub, O.A. and Shawagfeh, N. *Well-posedness analysis and pseudo-Galerkin approximations using tau Legendre algorithm for fractional systems of delay differential models regarding Hilfer  $(\alpha, \beta)$ -framework set*, Plos one, 19 (6) (2024) e0305259.
- [44] Sweis, H., Shawagfeh, N. and Arqub, O.A. *Fractional crossover delay differential equations of mittag-leffler kernel: Existence, uniqueness, and numerical solutions using the Galerkin algorithm based on shifted Legendre polynomials*, Result. Phys., 41 (2022) 105891.
- [45] Sweis, H., Shawagfeh, N. and Arqub, O.A. *Existence, uniqueness, and Galerkin shifted Legendre's approximation of time delays integro-differential models by adapting the Hilfer fractional attitude*, Heliyon, 10, (4) (2024) e25903.

- [46] Tarasov, E. *Fractional integro-differential equations for electromagnetic waves in dielectric media*, Theor. Math. Phys., 158, (2009) 355–359.
- [47] Yang, Y., Yao, P. and Tohidi, E. *Convergence analysis of an efficient multistep pseudo-spectral continuous Galerkin approach for solving Volterra integro-differential equations*, Appl. Math. Comput., 494 (2025) 129284.
- [48] Youssri, Y. and Hafez, R. *Chebyshev collocation treatment of Volterra–Fredholm integral equation with error analysis*, Arab. J. Math., 9 (2020) 471–480.
- [49] Yusufoglu E. and Erbas, B. *Numerical expansion methods for solving Fredholm–Volterra type linear integral equations by interpolation and quadrature rules*, Kybernetes, 37 (6) (2008) 768–785.
- [50] Yüzbaşı, Ş. *Improved Bessel collocation method for linear Volterra integro-differential equations with piecewise intervals and application of a Volterra population model*, Appl. Math. Model., 40 (9-10) (2016) 5349–5363.
- [51] Yüzbaşı, Ş., Sezer, M. and Kemancı, B. *Numerical solutions of integro-differential equations* Appl. Math. Model., 37 (4) (2013) 2086–2101.
- [52] Zavodnik, J. and Brojan, M. *Spherical harmonics-based pseudo-spectral method for quantitative analysis of symmetry breaking in wrinkling of shells with soft cores*, Comput. Method. Appl. Mech. Eng., 433, (2025) 117529.
- [53] Zhang, J., Zhu, X., Chen, T. and Dou, G. *Optimal dynamics control in trajectory tracking of industrial robots based on adaptive Gaussian pseudo-spectral algorithm*, Algorithms, 18 (1) (2025) 18.
- [54] Ziane, D., Cherif, M.H., and Adel, W. *Solving the Lane–Emden and Emden–Fowler equations on cantor sets by the local fractional homotopy analysis method*, Prog. Fract. Differ. Appl., 10, (2024) 241–250.



# Mathematical modeling of Echinococcosis in humans, dogs and livestock with optimal control strategies

I. Sannaky\*, M. Riouali, N. Ouldkhouia, I. El berrai, and K. Adnaoui

\*Corresponding author

Received 29 September 2024; revised 26 January 2025; accepted 15 April 2025

Ibtissam Sannaky

Laboratory of Analysis, Modeling and Simulation, HassanII university, Faculty of sciences Ben M'sik, Casablanca, Morocco. e-mail: [bissamsannaky@gmail.com](mailto:bissamsannaky@gmail.com)

Maryam Riouali

Laboratory of Analysis, Modeling and Simulation, HassanII university, Faculty of sciences Ben M'sik, Casablanca, Morocco. e-mail: [maryam.riouali1@gmail.com](mailto:maryam.riouali1@gmail.com)

Noureddine Ouldkhouia

Laboratory of Analysis, Modeling and Simulation, HassanII university, Faculty of sciences Ben M'sik, Casablanca, Morocco. e-mail: [ouldkhouia220@gmail.com](mailto:ouldkhouia220@gmail.com)

Imane El berrai

Laboratory of Analysis, Modeling and Simulation, HassanII university, Faculty of sciences Ben M'sik, Casablanca, Morocco. e-mail: [im.elberrai@gmail.com](mailto:im.elberrai@gmail.com)

Khalid Adnaoui

Laboratory of Analysis, Modeling and Simulation, HassanII university, Faculty of sciences Ben M'sik, Casablanca, Morocco. e-mail: [khalid.adnaoui@gmail.com](mailto:khalid.adnaoui@gmail.com)

## How to cite this article

Sannaky, I., Riouali, M., Ouldkhouia, N., El berrai, I. and Adnaoui, K., Mathematical modeling of Echinococcosis in humans, dogs and livestock with optimal control strategies. *Iran. J. Numer. Anal. Optim.*, 2025; 15(3): 1075-1115. <https://doi.org/10.22067/ijnao.2025.90038.1530>

### Abstract

In this paper, we present a deterministic model of Cystic Echinococcosis disease of eleven differential equations, describing complex interactions between three types of hosts, specifically humans, livestock and dogs. The model is analyzed; disease-free and endemic equilibrium exist and are globally asymptotically stable if the basic reproduction number is less than or greater than one, respectively. The optimal control represents the efficiency of health education and EG95 sheep vaccination. The existence of optimal controls is proven, the characterization is formulated using Pontryagin's maximum principle, and the optimal system is derived. For the numerical simulation, the data used are from some studies done in Morocco.

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

**Keywords:** Cystic Echinococcosis; deterministic modeling; basic reproduction number; global stability; optimal control.

## 1 Introduction

Mathematical modeling is a branch of science that seeks to illustrate real-life problems using mathematical techniques to better understand them, predict them, and act accordingly. It presents great interest in several fields, especially epidemiology [5, 12, 16, 17].

Cystic Echinococcosis (CE), also known as hydatid disease, is a globally known parasitic disease for its serious medical and economic impacts, it is caused by a parasite called the tapeworm CE. This disease affects humans, domestic and wild animals such as dogs, cats, rodents, livestock, horses, foxes, and wolves. The transmission of the disorder goes through a cycle [27].

Indeed eggs of CE are passed in the environment in the feces of definitive hosts such as dogs. Those eggs are then ingested by intermediate hosts like sheep. Inside those, the eggs hatch and form cysts in the liver, lungs, and sometimes brain, those cysts act like tumors that can disturb the function of the organ where they are found, they can cause poor growth, reduced production of milk and meat, which has severe economic impacts. When the

infected organs of intermediate hosts are eaten by a definitive carnivore host, the cycle restarts again [25].

Humans contract the infection by ingesting parasite eggs by touching contaminated soil, water, or an infected domestic dog for example [26]. There is no direct transmission between humans and those are called to be accidental hosts of the Echinococcosis disease. For them the malady can be dangerous, threatening, and occasionally fatal. Moreover, the treatment is heavy, lengthy, and expensive. In fact, the 2015 WHO Foodborne Disease Burden Epidemiology Reference Group (FERG) predicted Echinococcosis to be the cause for 19 300 deaths and approximately 871,000 disability-adjusted life-years (DALYs) globally each year [25].

Echinococcosis commonly affects rural areas where farming is the main activity, and dogs are kept in great numbers to protect livestock and provide companionship. Their presence is important for keeping safety and managing the pastoral community. So, contact between the three populations (humans, livestock, dogs) speeds the spread of the disease, especially because of the lack of knowledge of pastoralists about how CE is transmitted, and the open access of dogs to uncooked meat or carcasses that can be infectious [20].

In Morocco, a country where many regions depend on sheep farming, Echinococcosis can be dangerous and can threaten food security, animal and human health, and even life. However, studies and states on this disease are poor in Morocco, but the last ones assumed that the disease is critically endemic in regions where livestock are highly produced (Middle Atlas, El Hajeb, Azrou, Timahdit, Ouarzazate, Tiznit-Sidi Ifni, etc.) [9, 19, 2]. Different levels of prevalence of CE are recorded in [9]: 3.6 to 71.4% for cattle, 0.1 to 81.1% in sheep, and 0 to 25.7 % in goats. Concurrently, high degrees of CE infection (up to 70%) were spotted in dogs in various regions of the country. For humans, it has been reported from a study in the Middle Atlas region that CE prevalence is 1.9% out of 5367 people examined in 2017 [6]. Other studies show high infection degrees for humans and animals in Morocco [8, 2, 3, 19].

According to [20, 2], the mortality rate for CE in Morocco ranges between 2% and 3%, and the surgery and treatment of each infected patient costs approximately 17,000 and 32,000 Moroccan dirham MAD (US\$ 1700 and US\$ 3200), for simple and recurrence instances, respectively. Also, according

to a socio-economic study in [18] on the country's CE burden, the DALY ( Disability-adjusted life years ) is 160 years (0.5 years per 100,000 people) per year, with total annual losses of US\$ 73 million (US\$ 54.92 million) due to organ seizures, healthcare bills, and missed income for infected individuals, and all these stats are likely under-notified.

Many deterministic and statistical models are developed to describe and study the transmission of CE disease [4, 5, 14, 23, 24, 11], and so on. Based on these studies, in this article, we propose a developed Echinococcosis model that contains compartments representing humans, dogs, livestock, CE eggs, and infectious meat and that considers relapse after the surgery for humans and dogs. For humans, this rate ranges from 4.65% to 36% of cases in different regions of the world [22]. This recurrence causes a real problem in the management of the disease as stated by Velasco-Tirado et al. [22]. On the other hand, the optimal control theory is applied by integrating two control measures to study their effect on the spread of the disease. Therefore, in this work, based on some results from [13], we investigate the effectiveness of two prevention methods, the first is health education due to the current lack of knowledge of CE among citizens, and the second is the vaccination of livestock with an *E. granulosus* recombinant antigen (EG95) in Morocco [13], an approach that has been previously done in China and Argentina [25], because livestock prevention is probably the best approach to disrupt the CE life cycle.

A comparison between the current article and one of the articles cited above will be made next to show the key differences and innovation in our article.

The rest of this paper is organized as follows: In Section 2, a deterministic model is presented. Then in Section 3, the positivity and boundedness of the solutions are studied, and the basic reproduction number and the two equilibria of the model are computed. Following that we study the stability of the disease-free equilibrium (DFE) and the endemic equilibrium. Thus in Section 4, we apply the optimal control theory to the deterministic model studied. After that, in Section 5, numerical simulations are presented and discussed based on some studies carried out in Morocco. Finally, conclusion is made in Section 6.

## 2 Model formulation

### 2.1 Description of the model

To mathematically describe the spread of CE, we use a deterministic model that takes into account three types of population, humans being accidental hosts, dogs being definitive hosts, and livestock being intermediate hosts according to the diagram in Figure 1.

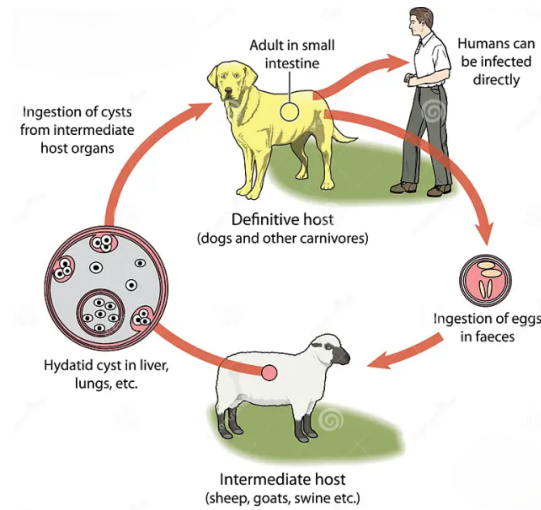


Figure 1: Life cycle of *Echinococcus granulosus*

### Hypothesis

1. The human population is divided into four classes: The susceptible  $H_S$ , the exposed  $H_E$ , the infected  $H_I$ , and the recovered  $H_R$ .

Humans act like accidental hosts, they are infected by touching the eggs of CE, due to lack of personal hygiene. Moreover, there is no direct transmission of CE disease from one human to another, and humans do not get infected from infectious meat.

For the humans submodel, we suppose a constant recruitment rate  $B_H$  due to birth or immigration into the susceptible class  $H_S$ , those decrease owing to exposure to CE parasite's eggs at a rate  $\beta_H$  or to natural death

$\mu_{HS}$ . As relapse of the disease exists even after recovery, we added a rate  $\alpha_S$  that denotes the rate at which recovered individuals become susceptible again.

The exposed class  $H_E$  increases at a rate  $\beta_H$  and decreases when moving to the infected class  $H_I$  at a rate  $\sigma_H$  or due to natural death  $\mu_{HE}$ . Infected humans are regarded to suffer death caused by the CE disease at a rate  $d_{kh}$ , and natural death at a rate  $\mu_{HI}$ , they decrease when moving to the recovered class  $H_R$  at a pace  $r_H$ .

Recovered class  $H_R$  diminishes due to the relapse of the healed individuals into the susceptible class at a pace  $\alpha_S$ , or due to natural death  $\mu_{HR}$ .

2. The dog population is separated into three classes: The susceptible  $D_H$ , the infected  $D_I$ , and the recovered  $D_R$ .

Dogs represent definitive hosts infected by consuming contaminated meat at a rate  $\beta_D$ , and they spread CE eggs present in their feces in the environment at a pace  $B_E$ . Moreover, we suppose that cured dogs can become susceptible again at a rate  $\alpha_D$ . The rest of the parameters of the dog submodel follow the same trend of the human population and are explained in Table 1 and Figure 2.

3. We divide livestock population into two classes: The susceptible  $O_S$  and the infected  $O_I$ .

Livestock population are infected by ingesting CE eggs at a rate  $\beta_O$ , which causes the formation of cysts in some of their organs, those cysts grow with time and stay in the infected organs. Moreover, they are slaughtered at a pace  $\eta_1$  when susceptible, and  $\eta_2$  when infected. The rest of the parameters are described in Table 1 and Figure 2.

4. We also integrate the infected meat compartment denoted by  $V$ , and it grows at a rate  $\eta_2$  when contaminated sheep or cattle are slaughtered and diminishes when it is eaten by humans at a rate  $\gamma$ , and by dogs at a rate  $\beta_D$ .

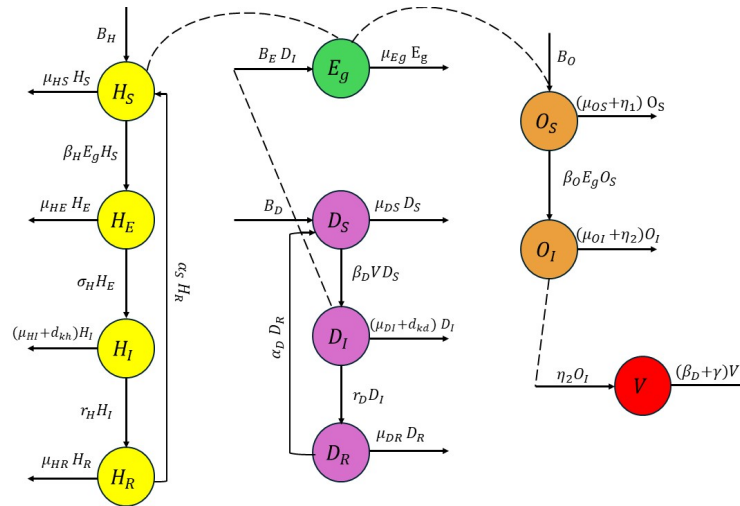


Figure 2: Diagram representation of the CE disease

5. The model also includes the density of Echinococcosis eggs present in the environment denoted by  $E_g$ , they are spread by infected dogs at a rate  $B_E$ , and they face natural death at a rate  $\mu_{Eg}$ .

Table 1: Parameters of the model

Parameters	Details
$B_D$	Dogs birth or immigration rate
$B_O$	Livestock birth or immigration rate
$B_E$	Rate of shedding CE eggs by infected dogs
$\mu_{DS}$	Natural mortality rate of susceptible dogs
$\mu_{DI}$	Natural mortality rate of infected dogs
$\mu_{DR}$	Natural mortality rate of recovered dogs
$\mu_{OS}$	Natural mortality rate of susceptible livestock
$\mu_{OI}$	Natural mortality rate of infected livestock
$\mu_{Eg}$	Natural mortality rate of CE eggs
$\alpha_S$	Relapse of humans after surgery
$\alpha_D$	Relapse of dogs after recovery
$\sigma_H$	Transfer rate to the infected humans class
$d_{kh}$	Humans mortality rate due to the CE disease
$d_{kd}$	Dogs mortality rate due to the CE disease
$\eta_1$	Slaughtering rate of susceptible livestock
$\eta_2$	Slaughtering rate of infected livestock
$r_D$	Recovery rate of dogs

The deterministic model is then given by this system of eleven equations below:

$$\left\{ \begin{array}{l} \frac{dH_S}{dt} = B_H - \beta_H E_g H_S - \mu_{HS} H_S + \alpha_S H_R, \\ \frac{dH_E}{dt} = \beta_H E_g H_S - (\sigma_H + \mu_{HE}) H_E, \\ \frac{dH_I}{dt} = \sigma_H H_E - (\mu_{HI} + d_{kh} + r_H) H_I, \\ \frac{dH_R}{dt} = r_H H_I - (\mu_{HR} + \alpha_S) H_R, \\ \frac{dD_S}{dt} = B_D - \beta_D V D_S - \mu_{DS} D_S + \alpha_D D_R, \\ \frac{dD_I}{dt} = \beta_D V D_S - (r_D + d_{kd} + \mu_{DI}) D_I, \\ \frac{dD_R}{dt} = r_D D_I - (\mu_{DR} + \alpha_D) D_R, \\ \frac{dO_S}{dt} = B_O - \beta_O E_g O_S - (\mu_{OS} + \eta_1) O_S, \\ \frac{dO_I}{dt} = \beta_O E_g O_S - (\mu_{OI} + \eta_2) O_I, \\ \frac{dV}{dt} = \eta_2 O_I - (\beta_D + \gamma) V, \\ \frac{dE_G}{dt} = B_E D_I - \mu_{Eg} E_g. \end{array} \right. \quad (1)$$

In system (1), we assume that all the parameters (described in Table 1) are positive.

## 2.2 Comparison with a previous study

In this section, a comparison is made with the article [5] by Chacha et al.

### - Overview of the article

- In their work, Chacha et al. [5] developed a deterministic then a stochastic CTMC model to describe the dynamic of CE in humans, dogs, and cattle. The results emphasized that disease prevention requires intervention strategies for the populations of dogs and cattle.

### - Comparison of our model with the Chacha et al. model

- Building on the work of Chacha et al. [5], our model also describes CE transmission between three types of population; humans, dogs and livestock and

studies the density of infected meat and CE eggs. Indeed unlike the model in [5], we adopted different types of submodels to address reality more:

- For human submodel, unlike [5], we included the compartment of recovered persons  $H_R$ . This inclusion allows us to give importance to the healing phase and the recovery processes which was neglected in [5]. In fact it has supposed that once the individual is infected, he either stay infected or dies, which is not true in real life, where several people take their treatment or/and surgery to recover. Another reason for adding this compartment is that patients that have experienced the disease and were recovered due to treatment will certainly change and develop some health and hygiene practices to prevent themselves from a new contamination, they will be aware of the danger of the disease and will help raise awareness among their family and neighbors circle conducting to a decrease of being susceptible or exposed to the disease, as improving hygiene is the most effective way to prevent from infection.
- For dog submodel, contrary to [5], we excluded the Exposed dogs compartment, because dogs become infectious almost few time after consuming infected meat of intermediate host, and there is a negligible latent period where dogs doesn't spread CE eggs. Moreover, in real life it is so difficult to detect if a dog is exposed but not yet infectious. Instead, we added a compartment for recovered dogs as there is a category of dogs that get cured by anthelmintics or by removing adult tapeworm in surgery, which was ignored in [5], where infectious dogs either stay infected or die.
- In the same context we added the parameters  $\alpha_S$  and  $\alpha_D$  that stand for the rate of recovered returning susceptible again among humans and dogs, respectively. This parameter aligns the biological behavior of CE, so if a recovered person stops his good hygiene practices, he is at direct risk of being infected again, especially in an endemic area. Similarly, if the immunity of a recovered dog wanes, then it becomes susceptible again to infection at a certain rate.
- For livestock submodel, we disregarded the exposed compartment introduced for cattle in [5], and used only susceptible and infected compartments for livestock. The cause behind this choice is that livestock are intermediate

hosts; they affect the life cycle of the disease only with the cysts present in their infected organs when they are slaughtered or naturally dead, so exposed and infected compartments are considered as one in this case. Adding exposed livestock will not affect the transmission cycle or the dynamic of CE.

- The model in [5] is developed as a stochastic model and our model will be developed as an optimal control model.

### 3 Model analysis

Grounded in the biological foundation of the model (1), we assume that all the initial conditions of solutions' system are positive as

$$\begin{aligned} H_S(0) > 0, \quad H_E(0) > 0, \quad H_I(0) > 0, \quad H_R(0) > 0, \\ D_S(0) > 0, \quad D_I(0) > 0, \quad D_R(0) > 0, \\ O_S(0) > 0, \quad O_I(0) > 0, \\ V(0) > 0, \quad E_g(0) > 0. \end{aligned} \tag{2}$$

Moreover, we suppose  $H_P$  the total human population given by

$$H_P = H_S + H_E + H_I + H_R, \tag{3}$$

and  $D_P$  the total dog population given by

$$D_P = D_S + D_I + D_R. \tag{4}$$

As the same, the total livestock population is modeled by  $O_P$  and given as

$$O_P = O_S + O_I. \tag{5}$$

#### 3.1 Positivity and boundedness of system solutions

**Theorem 1.** 1. The solutions of model system (1) with initial conditions (2) are nonnegative for all  $t \geq 0$ .

2. The solutions of model system (1) with initial conditions (2) are bounded in  $\Gamma_\varepsilon$  when  $t \rightarrow \infty$ .

Here for any  $\varepsilon > 0$ ,  $\Gamma_\varepsilon$  is defined as

$$\Gamma_\varepsilon = \left\{ (H, D, O, V, E_g) \in \mathbb{R}_+^{11} \mid \begin{array}{l} H_N \leq \frac{B_H}{\mu_1} + \varepsilon, \\ D_N \leq \frac{B_D}{\mu_2} + \varepsilon, \\ O_N \leq \frac{B_O}{\mu_3} + \varepsilon, \\ V \leq W, \\ E_g \leq F, \end{array} \right\} \quad (6)$$

with

$$\begin{aligned} H &= (H_S, H_E, H_I, H_R), \quad D = (D_S, D_I, D_R), \quad O = (O_S, O_I), \\ W &= \frac{\eta_2 B_O}{\mu_3(\beta_D + \gamma)} + \left(\frac{\eta_2}{\beta_D + \gamma} + 1\right)\varepsilon, \quad F = \frac{B_E B_D}{\mu_2 \mu_{Eg}} + \varepsilon(B_E + 1), \\ \mu_1 &= \min\{\mu_{HS}, \mu_{HE}, \mu_{HI}, \mu_{HR}\}, \quad \mu_2 = \min\{\mu_{DS}, \mu_{DI}, \mu_{DR}\}, \\ \mu_3 &= \min\{\mu_{OS}, \mu_{OI}\}. \end{aligned}$$

**Proof. Positivity of solutions**

1. First, for convenience, we write a solution of the model (1) as  $(H(t), D(t), O(t), V(t), E_g(t))$ , with  $H(t) = (H_S(t), H_E(t), H_I(t), H_R(t))$ ,  $D(t) = (D_S(t), D_I(t), D_R(t))$ , and  $O(t) = (O_S(t), O_I(t))$ .

We consider  $(H(t), D(t), O(t), V(t), E_g(t))$ , a solution of the model (1) with the initial positive conditions (2).

We proceed by contradiction, and we suppose that there is  $t^* > 0$  such that

$$\min(H(t^*), D(t^*), O(t^*), V(t^*), E_g(t^*)) = 0.$$

So that  $\min(H(t), D(t), O(t), V(t), E_g(t)) > 0$  for all  $t \in [0, t^*[$ .

- **Case 1:**  $\min(H(t^*), D(t^*), O(t^*), V(t^*), E_g(t^*)) = H_S(t^*)$ .

We have in this case from the first equation of the system (1) that

$$\frac{dH_S}{dt} \geq -(\beta_H E_g + \mu_{HS})H_S \quad \text{for all } t \in [0, t^*],$$

and by applying separation method, it yields

$$\frac{dH_S}{H_S} \geq -(\beta_H E_g + \mu_{HS})dt \quad \text{for all } t \in [0, t^*].$$

Then we integrate in initial conditions,

$$0 = H_S(t^*) \geq H_S(0)e^{-\int_0^{t^*} (\beta_H E_g(s) + \mu_{HS})ds} > 0.$$

That results in a contradiction.

- **Case 2:**  $\min(H(t^*), D(t^*), O(t^*), V(t^*), E_g(t^*)) = H_E(t^*)$ .

We obtain from the second equation of the model (1) and by adopting the same method as in case 1, that

$$0 = H_E(t^*) \geq H_E(0)e^{-(\sigma_H + \mu_{HE})t^*} > 0.$$

This leads to a contradiction in this case too.

The same approach is applied for the rest of cases when  $\min(H(t^*), D(t^*), O(t^*), V(t^*), E_g(t^*))$  is equal to  $H_I(t^*)$ ,  $H_R(t^*)$ ,  $D_S(t^*)$ ,  $D_I(t^*)$ ,  $D_R(t^*)$ ,  $O_S(t^*)$ ,  $O_I(t^*)$ ,  $V(t^*)$ , and  $E_g(t^*)$ , respectively, where we have from the corresponding equations of the system (1), respectively:

$$\begin{aligned} 0 &= H_E(t^*) \geq H_E(0)e^{-(\sigma_H + \mu_{HE})t^*} > 0, \\ 0 &= H_I(t^*) \geq H_I(0)e^{-(\mu_{HI} + d_{kh} + r_H)t^*} > 0, \\ 0 &= H_R(t^*) \geq H_R(0)e^{-(\mu_{HR} + \alpha_S)t^*} > 0, \\ 0 &= D_S(t^*) \geq D_S(0)e^{-\int_0^{t^*} (\beta_D V(s) + \mu_{DS})ds} > 0, \\ 0 &= D_I(t^*) \geq D_I(0)e^{-(r_D + d_{kd} + \mu_{DI})t^*} > 0, \\ 0 &= D_R(t^*) \geq D_R(0)e^{-(\mu_{DR} + \alpha_D)t^*} > 0, \\ 0 &= O_S(t^*) \geq O_S(0)e^{-\int_0^{t^*} (\beta_O E_g(s) + \mu_{OS} + \eta_1)ds} > 0, \\ 0 &= O_I(t^*) \geq O_I(0)e^{-(\mu_{OI} + \eta_2)t^*} > 0, \end{aligned}$$

$$0 = V(t^*) \geq V(0)e^{-(\beta_D + \gamma)t^*} > 0,$$

$$0 = E_g(t^*) \geq E_g(0)e^{-\mu_{Eg}t^*} > 0.$$

It yields to a contradiction in each case.

Finally, we conclude that all the solutions of system (1) are nonnegative for all  $t \geq 0$ .

### Boundedness of the solutions

2. In this part of the demonstration, we use the total populations of humans (3), dogs (4), and livestock (5), respectively, to show that the solutions are bounded in  $\Gamma_\varepsilon$  (6).

From (3), and the first four equations of system (1), we got

$$\frac{dH_P}{dt} = B_H - \mu_{HS}H_S - \mu_{HE}H_E - \mu_{HI}H_I - d_{kh}H_I - \mu_{HR}H_R.$$

So

$$\frac{dH_P}{dt} \leq B_H - \mu_{HS}H_S - \mu_{HE}H_E - \mu_{HI}H_I - \mu_{HR}H_R.$$

Let  $\mu_1 = \min\{\mu_{HS}, \mu_{HE}, \mu_{HI}, \mu_{HR}\}$ .

Then

$$\begin{aligned} \frac{dH_P}{dt} &\leq B_H - \mu_1(H_S + H_E + H_I + H_R), \\ \frac{dH_P}{dt} &\leq B_H - \mu_1 H_P. \end{aligned}$$

Thus,

$$\lim_{t \rightarrow \infty} H_P(t) \leq \frac{B_H}{\mu_1}.$$

Hence for any  $\varepsilon > 0$ , there exists  $t_1 > 0$  such that

$$H_P(t) \leq \frac{B_H}{\mu_1} + \varepsilon \quad \text{for all } t \geq t_1.$$

Adopting the same method for the rest of the equation of system (1), we obtain the following results: From (4) and the fifth, sixth, and seventh equations from (1) we obtain that

there exists  $t_2 > 0$  such as

$$D_P(t) \leq \frac{B_D}{\mu_2} + \varepsilon \quad \text{for all } t \geq t_2, \quad (7)$$

with  $\mu_2 = \min\{\mu_{DS}, \mu_{DI}, \mu_{DR}\}$ .

Similarly, we have from (5)

There exists  $t_3 > 0$  such that

$$O_P(t) \leq \frac{B_O}{\mu_3} + \varepsilon \quad \text{for all } t \geq t_3, \quad (8)$$

with  $\mu_3 = \min \{\mu_{OS}, \mu_{OI}\}$ .

Likewise for the tenth equation of system (1) and from (8), we obtain

$$\begin{aligned} \frac{dV(t)}{dt} &= \eta_2 O_I(t) - (\beta_D + \gamma)V(t), \\ \frac{dV(t)}{dt} &\leq \eta_2 \frac{B_O}{\mu_3} + \eta_2 \varepsilon - (\beta_D + \gamma)V(t) \quad \text{for all } t \geq t_3, \end{aligned}$$

so there is  $t_4 > t_3$  such that

$$\begin{aligned} V(t) &\leq \frac{\eta_2 \left( \frac{B_O}{\mu_3} + \varepsilon \right)}{\beta_D + \gamma} + \varepsilon, \\ V(t) &\leq \frac{\eta_2 B_O}{\mu_3(\beta_D + \gamma)} + \left( \frac{\eta_2}{\beta_D + \gamma} + 1 \right) \varepsilon; \quad \text{for all } t \geq t_4. \end{aligned}$$

Finally, from the last equation of system (1) and inequality (7), we obtain the following result:

There exists  $t_5 > t_2$  such that

$$E_g \leq \frac{B_E B_D}{\mu_2 \mu_{Eg}} + \varepsilon(B_E + 1) \quad \text{for all } t \geq t_5.$$

Take  $t^\# = \max \{t_1, t_4, t_5\}$ ; then for all  $t > t^\#$  we have

$$(H(t), D(t), O(t), V(t), E_g(t)) \in \Gamma_\varepsilon.$$

Recall that

$$H = (H_S, H_E, H_I, H_R), \quad D = (D_S, D_I, D_R), \quad O = (O_S, O_I).$$

Therefore the region  $\Gamma_\varepsilon$  is positive and invariant. Moreover, all the solutions of system (1) with initial conditions (2) are bounded.

Finally, the proof of Theorem 1 is concluded.  $\square$

### 3.2 Disease free equilibrium (DFE) and basic reproduction number

Basic computations assume that the system (1) has a DFE given as

$$N_0 = (H_S^0, 0, 0, 0, D_S^0, 0, 0, O_S^0, 0, 0, 0),$$

where

$$\begin{aligned} H_S^0 &= \frac{B_H}{\mu_{HS}}, \\ D_S^0 &= \frac{B_D}{\mu_{DS}}, \\ O_S^0 &= \frac{B_O}{\mu_{OS} + \eta_1}. \end{aligned} \quad (9)$$

The DFE denotes the case when the disease does not occur in the three types of populations.

The basic reproduction number  $R_0$  is the number of secondary infection caused by an infected person during the period of infection [7]. To compute it for the system (1) we use the next generation matrix method [21] and we define then

$$\mathcal{F} = \begin{pmatrix} \beta_H E_g H_S \\ 0 \\ \beta_D V D_S \\ \beta_O E_g O_S \\ 0 \\ 0 \end{pmatrix} \mathcal{V} = \begin{pmatrix} (\sigma_H + \mu_{HE}) H_E \\ -\sigma_H H_E + (\mu_{HI} + d_{kh} + r_H) H_I \\ (r_D + d_{kd} + \mu_{DI}) D_I \\ (\mu_{OI} + \eta_2) O_I \\ -\eta_2 O_I + (\beta_D + \gamma) V \\ -B_E D_I + \mu_{E_g} E_g \end{pmatrix}.$$

Then for the DFE (9), we obtain

$$\mathbb{F} = \begin{pmatrix} 0 & 0 & 0 & 0 & 0 & \frac{\beta_H B_H}{\mu_{HS}} \\ 0 & 0 & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & \frac{\beta_D B_D}{\mu_{DS}} & 0 \\ 0 & 0 & 0 & 0 & 0 & \frac{\beta_O B_O}{\mu_{OS} + \eta_1} \\ 0 & 0 & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & 0 & 0 \end{pmatrix}.$$

Moreover,

$$\mathbb{V} = \begin{pmatrix} a & 0 & 0 & 0 & 0 & 0 \\ -\sigma_H & b & 0 & 0 & 0 & 0 \\ 0 & 0 & c & 0 & 0 & 0 \\ 0 & 0 & 0 & d & 0 & 0 \\ 0 & 0 & 0 & -\eta_2 & e & 0 \\ 0 & 0 & -B_E & 0 & 0 & \mu_{Eg} \end{pmatrix},$$

where

$$a = (\sigma_H + \mu_{HE}), \quad b = (\mu_{HI} + d_{kh} + r_H), \quad c = (r_D + d_{kd} + \mu_{DI}),$$

$$d = (\mu_{OI} + \eta_2), \quad e = (\beta_D + \gamma).$$

Accordingly, as  $R_0 = \rho(\mathbb{F}\mathbb{V}^{-1})$ ,  $R_0$  is given by

$$R_0 = \sqrt{\frac{\beta_D \beta_O B_D B_E \eta_2 B_O}{\mu_{DS} \mu_{Eg} (\eta_2 + \mu_{OI}) (\mu_{OS} + \eta_1) (r_D + d_{kd} + \mu_{DI}) (\beta_D + \gamma)}}. \quad (10)$$

For the seek of simplicity, we can write  $R_0$  as

$$R_0 = \sqrt{\beta_D \beta_O \frac{B_E}{\mu_{Eg}} \frac{B_D}{\mu_{DS}} \frac{\eta_2}{(\eta_2 + \mu_{OI})} \frac{B_O}{(\mu_{OS} + \eta_1)} \frac{1}{(r_D + d_{kd} + \mu_{DI})} \frac{1}{(\beta_D + \gamma)}}. \quad (11)$$

In the expression of  $R_0$  (11), we interpret  $\frac{B_E}{\mu_{Eg}}$  to be the density of CE eggs released in the environment by contaminated dogs,  $\frac{B_D}{\mu_{DS}}$  and  $\frac{B_O}{(\mu_{OS} + \eta_1)}$  to be the size of dog and livestock populations in the beginning respectively. The fraction  $\frac{\eta_2}{(\eta_2 + \mu_{OI})}$  stands for the contaminated population of livestock slaughtered for consuming. The expression  $\frac{1}{(r_D + d_{kd} + \mu_{DI})}$  signifies the outflow rate of dogs from the infected compartment  $D_I$ . Finally, the standard rate describing when infectious meat is disease-spreading, is given by  $\frac{1}{(\beta_D + \gamma)}$ .

It is assumed that when  $R_0 < 1$ , the disease will stop on its own, and when  $R_0 > 1$  it will persist.

### 3.3 Endemic equilibrium

The endemic equilibrium of system (1) denotes the case when the CE disease is endemic. It is represented by

$$N^* = (H_S^*, H_E^*, H_I^*, H_R^*, D_S^*, D_I^*, D_R^*, O_S^*, O_I^*, V^*, E_g^*), \quad (12)$$

where

$$\begin{aligned} H_S^* &= \frac{B_H + \alpha_S H_R^*}{\beta_H E_g^* + \mu_{HS}}, \\ H_E^* &= \frac{\beta_H E_g^* (B_H + \alpha_S H_R^*)}{(\sigma_H + \mu_{HE})(\beta_H E_g^* + \mu_{HS})}, \\ H_I^* &= \frac{\sigma_H \beta_H E_g^* (B_H + \alpha_S H_R^*)}{(\sigma_H + \mu_{HE})(\beta_H E_g^* + \mu_{HS})(\mu_{HI} + d_{kh} + r_H)}, \\ H_R^* &= \frac{B_H r_H \sigma_H \beta_H E_g^*}{(\mu_{HR} + \alpha_S)(\sigma_H + \mu_{HE})(\beta_H E_g^* + \mu_{HS})(\mu_{HI} + d_{kh} + r_H) - \alpha_S r_H \sigma_H \beta_H E_g^*}, \\ D_S^* &= \frac{B_D + \alpha_D D_R^*}{\beta_D V^* + \mu_{DS}}, \\ D_I^* &= \frac{\beta_D V^* (B_D + \alpha_D D_R^*)}{(\beta_D V^* + \mu_{DS})(r_D + d_{kd} + \mu_{DI})}, \\ D_R^* &= \frac{B_D r_D \beta_D V^*}{(\mu_{DR} + \alpha_D)(\beta_D V^* + \mu_{DS})(r_D + d_{kd} + \mu_{DI}) - \alpha_D r_D \beta_D V^*}, \\ O_S^* &= \frac{B_O}{\beta_O E_g^* + \mu_{OS} + \eta_1}, \\ O_I^* &= \frac{\beta_O E_g^* B_O}{(\mu_{OI} + \eta_2)(\beta_O E_g^* + \mu_{OS} + \eta_1)}, \\ V^* &= \frac{\eta_2 \beta_O E_g^* B_O}{(\beta_D + \lambda)(\mu_{OI} + \eta_2)(\beta_O E_g^* + \mu_{OS} + \eta_1)}, \\ E_g^* &= \frac{A}{B} (R_0 - 1)(R_0 + 1), \end{aligned}$$

with  $A = \mu_{DS} \mu_{Eg} (\mu_{DR} + \alpha_D)^2 (r_D + d_{kd} + \mu_{DI}) (\beta_D + \gamma) (\mu_{OI} + \eta_1) (\mu_{OS} + \eta_1)$ ,

and  $B = \mu_{Eg} \beta_D \eta_2 \beta_O B_O [(\mu_{DR} + \alpha_D)(r_D + d_{kd} + \mu_{DI}) - \alpha_D r_D] + \mu_{DS} \mu_{Eg} \beta_O (\mu_{DR} + \alpha_D)(r_D + d_{kd} + \mu_{DI}) (\beta_D + \lambda) (\mu_{OI} + \eta_2)$ .

According to the expression of the endemic equilibrium of system (1), we can say that the CE disease persist in humans, dogs and livestock if  $R_0 > 1$ , which is confirmed by this theorem.

**Theorem 2.** The system (1) modeling the dynamic of transmission of the Echinococcosis disease has an endemic equilibrium when  $R_0 > 1$ .

### 3.4 Stability analysis of the equilibria

#### 3.4.1 Global stability of the DFE

**Theorem 3.** The DFE  $N_0$  is globally asymptotically stable when  $R_0 < 1$ .

*Proof.* To prove the global stability of DFE (9) when  $R_0 < 1$ , we use the Lyapunov function given by

$$L = \frac{1}{(r_D + d_{kd} + \mu_{DI})} D_I + \frac{B_E}{\mu_{DI}\mu_{Eg}} E_g. \quad (13)$$

So from the equation six and eleven of model system (1), we have

$$\begin{aligned} \frac{dL}{dt} &= \frac{1}{r_D + d_{kd} + \mu_{DI}} \frac{dD_I}{dt} + \frac{B_E}{\mu_{DI}\mu_{Eg}} \frac{dE_g}{dt}, \\ \frac{dL}{dt} &= \frac{1}{r_D + d_{kd} + \mu_{DI}} [\beta_D V D_S - (r_D + d_{kd} + \mu_{DI}) D_I] \\ &\quad + \frac{B_E}{\mu_{DI}\mu_{Eg}} (B_E D_I - \mu_{Eg} E_g), \\ \frac{dL}{dt} &= \frac{\beta_D V D_S}{r_D + d_{kd} + \mu_{DI}} - D_I + \frac{B_E^2 D_I}{\mu_{DI}\mu_{Eg}} - \frac{B_E}{\mu_{DI}} E_g, \\ \frac{dL}{dt} &= \frac{\beta_D V D_S}{r_D + d_{kd} + \mu_{DI}} - D_I + \frac{B_E^2 D_I}{\mu_{DI}\mu_{Eg}} - \frac{B_E^2 D_I}{\mu_{DI}\mu_{Eg}}, \\ \frac{dL}{dt} &= \frac{\beta_D V D_S}{r_D + d_{kd} + \mu_{DI}} - D_I, \\ \frac{dL}{dt} &= \frac{\beta_D \eta_2 O_I B_D}{\mu_{DS}(r_D + d_{kd} + \mu_{DI})(\beta_D + \gamma)} - D_I, \\ \frac{dL}{dt} &= \frac{\beta_D \eta_2 B_D \beta_O E_g O_S}{\mu_{DS}(r_D + d_{kd} + \mu_{DI})(\beta_D + \gamma)(\mu_{OI} + \eta_2)} - D_I, \\ \frac{dL}{dt} &= \left( \frac{\beta_D \eta_2 B_D \beta_O B_O B_E}{\mu_{DS}\mu_{Eg}(r_D + d_{kd} + \mu_{DI})(\beta_D + \gamma)(\mu_{OI} + \eta_2)(\mu_{OS} + \eta_1)} - 1 \right) D_I, \\ \frac{dL}{dt} &= (R_0^2 - 1) D_I. \end{aligned}$$

Then

$$\frac{dL}{dt} = (R_0 - 1)(R_0 + 1) D_I. \quad (14)$$

From Theorem 1, we have  $D_I$  is nonnegative, so in (14), if  $R_0 = 1$ , then  $\frac{dL}{dt} = 0$ . If  $R_0 < 1$ , then  $\frac{dL}{dt} < 0$

Based on what precedes we can conclude then that the DFE (9) is globally asymptotically stable in  $\Gamma_\varepsilon(6)$  when  $R_0 < 1$ .

Therefore the proof of this theorem comes to an end.  $\square$

### 3.4.2 Global stability of endemic equilibrium

**Theorem 4.** The system of differential equation (1), is said to have a unique, globally asymptotically stable (GAS), endemic equilibrium (12) if  $R_0 > 1$ .

*Proof.* We know that endemic equilibrium exists if and only if  $R_0 > 1$ , so in this proof we will suppose that we are in the case when  $R_0 > 1$ .

We consider the function  $h(x) = x - 1 - \ln(x)$ ,  $x > 0$ , we have clearly  $h(x) \geq 0$  and that  $h(1) = 0$ .

Moreover, we can easily prove that

$$(x - 1)(y - 1) = h(x) + h(y) - h(xy) \quad \text{for all } x > 0 \text{ and } y > 0. \quad (15)$$

We consider  $\mathcal{L}_\# = \#^* h(\frac{\#}{\#^*})$  where we will replace  $\#$  by  $H_S, H_E, H_I, H_R, D_S, D_I, D_R, O_S, O_I, V$  and  $E_g$ , respectively.

First, we have

$$\frac{d\mathcal{L}_{H_S}}{dt} = \left(1 - \frac{H_S^*}{H_S}\right), \frac{dH_S}{dt} \quad (16)$$

and from the first equation of system (1) and the equilibrium equation, we have

$$B_H - \beta_H E_g^* H_S^* - \mu_{HS} H_S^* + \alpha_S H_R^* = 0, \text{ and we get}$$

$$\begin{aligned} \frac{d\mathcal{L}_{H_S}}{dt} &= \left(1 - \frac{H_S^*}{H_S}\right) (\beta_H E_g^* H_S^* + \mu_{HS} H_S^* - \alpha_S H_R^* - \beta_H E_g H_S - \mu_{HS} H_S + \alpha_S H_R) \\ &= \left(1 - \frac{H_S^*}{H_S}\right) [-\beta_H (E_g H_S - E_g^* H_S^*) - \mu_{HS} (H_S - H_S^*) + \alpha_S (H_R - H_R^*)] \\ &= -\beta_H E_g^* H_S^* \left(1 - \frac{E_g H_S}{E_g^* H_S^*}\right) \left(1 - \frac{H_S^*}{H_S}\right) - \mu_{HS} H_S^* \left(\frac{H_S}{H_S^*} - 1\right) \left(1 - \frac{H_S^*}{H_S}\right) \\ &\quad + \alpha_S H_R^* \left(\frac{H_R}{H_R^*} - 1\right) \left(1 - \frac{H_S^*}{H_S}\right). \end{aligned} \quad (17)$$

Using the expression (15) and the equilibrium equation yields

$$\frac{d\mathcal{L}_{H_S}}{dt} = -\beta_H E_g^* H_S^* h\left(\frac{E_g H_S}{E_g^* H_S^*}\right) - \beta_H E_g^* H_S^* h\left(\frac{H_S}{H_S^*}\right) + \beta_H E_g^* H_S^* h\left(\frac{E_g}{E_g^*}\right)$$

$$\begin{aligned}
& -\mu_{HS}H_S^*h\left(\frac{H_S}{H_S^*}\right) - \mu_{HS}H_S^*h\left(\frac{H_S^*}{H_S}\right) \\
& + \alpha_S H_R^*h\left(\frac{H_R}{H_R^*}\right) + \alpha_S H_R^*h\left(\frac{H_S^*}{H_S}\right) - \alpha_S H_R^*h\left(\frac{H_R H_S^*}{H_S H_R^*}\right) \\
& = -\beta_H E_g^* H_S^* h\left(\frac{E_g H_S}{E_g^* H_S^*}\right) + \beta_H E_g^* H_S^* h\left(\frac{E_g}{E_g^*}\right) - \mu_{HS} H_S^* h\left(\frac{H_S}{H_S^*}\right) \\
& - B_H h\left(\frac{H_S^*}{H_S}\right) + \alpha_S H_R^* h\left(\frac{H_R}{H_R^*}\right) - \alpha_S H_R^* h\left(\frac{H_R H_S^*}{H_S H_R^*}\right). \quad (18)
\end{aligned}$$

Similarly, from the second equation of the system (1), and the endemic equilibrium equation,  $\beta_H E_g^* H_S^* - (\sigma_H + \mu_{HE}) H_E^* = 0$ , we obtain

$$\begin{aligned}
\frac{d\mathcal{L}_{HE}}{dt} &= \left(1 - \frac{H_E^*}{H_E}\right) \frac{dH_E}{dt} \\
&= \left(1 - \frac{H_E^*}{H_E}\right) [\beta_H (E_g H_S - E_g^* H_S^*) - (\sigma_H + \mu_{HE}) (H_E - H_E^*)] \\
&= \beta_H E_g^* H_S^* \left(1 - \frac{H_E^*}{H_E}\right) \left(\frac{E_g H_S}{E_g^* H_S^*} - 1\right) \\
&\quad - (\sigma_H + \mu_{HE}) H_E^* \left(1 - \frac{H_E^*}{H_E}\right) \left(\frac{H_E}{H_E^*} - 1\right). \quad (19)
\end{aligned}$$

Using the expression (15) and the equilibrium equation yields

$$\frac{d\mathcal{L}_{HE}}{dt} = \beta_H E_g^* H_S^* h\left(\frac{E_g H_S}{E_g^* H_S^*}\right) - \beta_H E_g^* H_S^* h\left(\frac{H_E^* E_g H_S}{H_E E_g^* H_S^*}\right) - (\sigma_H + \mu_{HE}) H_E^* h\left(\frac{H_E}{H_E^*}\right).$$

Similarly, we get

$$\begin{aligned}
\frac{d\mathcal{L}_{HI}}{dt} &= \sigma_H H_E^* h\left(\frac{H_E}{H_E^*}\right) - \sigma_H H_E^* h\left(\frac{H_I^* H_E}{H_I H_E^*}\right) - (\mu_{HI} + d_{kh} + r_H) H_I^* h\left(\frac{H_I}{H_I^*}\right), \\
\frac{d\mathcal{L}_{HR}}{dt} &= r_H H_I^* h\left(\frac{H_I}{H_I^*}\right) - r_H H_I^* h\left(\frac{H_R^* H_I}{H_R H_I^*}\right) - (\mu_{HR} + \alpha_S) H_R^* h\left(\frac{H_R}{H_R^*}\right), \\
\frac{d\mathcal{L}_{DS}}{dt} &= -\beta_D V^* D_S^* h\left(\frac{V D_S}{V^* D_S^*}\right) + \beta_D V^* D_S^* h\left(\frac{V}{V^*}\right) - \mu_{DS} D_S^* h\left(\frac{D_S}{D_S^*}\right) - B_D h\left(\frac{D_S^*}{D_S}\right) \\
&\quad + \alpha_D D_R^* h\left(\frac{D_R}{D_R^*}\right) - \alpha_D D_R^* h\left(\frac{D_R D_S^*}{D_S D_R^*}\right), \\
\frac{d\mathcal{L}_{DI}}{dt} &= \beta_D V^* D_S^* h\left(\frac{V D_S}{V^* D_S^*}\right) - \beta_D V^* D_S^* h\left(\frac{V D_S D_I^*}{V^* D_S^* D_I}\right) - (r_D + d_{kd} + \mu_{DI}) D_I^* h\left(\frac{D_I}{D_I^*}\right), \\
\frac{d\mathcal{L}_{DR}}{dt} &= r_D D_I^* h\left(\frac{D_I}{D_I^*}\right) - r_D D_I^* h\left(\frac{D_R^* D_I}{D_R D_I^*}\right) - (\mu_{DR} + \alpha_D) D_R^* h\left(\frac{D_R}{D_R^*}\right), \\
\frac{d\mathcal{L}_{OS}}{dt} &= -\beta_O E_g^* O_S^* h\left(\frac{E_g O_S}{E_g^* O_S^*}\right) + \beta_O E_g^* O_S^* h\left(\frac{E_g}{E_g^*}\right) - (\mu_{OS} + \eta_1) O_S^* h\left(\frac{O_S}{O_S^*}\right)
\end{aligned}$$

$$\begin{aligned}
& -B_O h\left(\frac{O_S^*}{O_S}\right), \\
\frac{d\mathcal{L}_{O_I}}{dt} &= \beta_O E_g^* O_S^* h\left(\frac{E_g O_S}{E_g^* O_S^*}\right) - \beta_O E_g^* O_S^* h\left(\frac{E_g O_S O_I^*}{E_g^* O_S^* O_I^*}\right) - (\mu_{O_I} + \eta_2) O_I^* h\left(\frac{O_I}{O_I^*}\right), \\
\frac{d\mathcal{L}_V}{dt} &= \eta_2 O_I^* h\left(\frac{O_I}{O_I^*}\right) - \eta_2 O_I^* h\left(\frac{V^* O_I}{V O_I^*}\right) - \beta_D V^* h\left(\frac{V}{V^*}\right) - \gamma V^* h\left(\frac{V}{V^*}\right), \\
\frac{d\mathcal{L}_{E_g}}{dt} &= B_E D_I^* h\left(\frac{D_I}{D_I^*}\right) - B_E D_I^* h\left(\frac{D_I E_g^*}{D_I^* E_g}\right) - \mu_{E_g} E_g^* h\left(\frac{E_g}{E_g^*}\right).
\end{aligned}$$

We define the Lyapunov function given by

$$\begin{aligned}
\mathcal{L} &= \frac{\mu_{E_g}}{B_H H_S^*} (\mathcal{L}_{H_S} + \mathcal{L}_{H_E} + \mathcal{L}_{H_I} + \mathcal{L}_{H_R}) + \frac{2B_E}{d_{kd}} (\mathcal{L}_{D_S} + \mathcal{L}_{D_E} + \mathcal{L}_{D_R}) \quad (20) \\
&+ \frac{\mu_{E_g}}{\beta_O O_S^*} (\mathcal{L}_{O_S} + \mathcal{L}_{O_I}) + \frac{2B_E D_S^*}{d_{kd}} \mathcal{L}_V + 2\mathcal{L}_{E_g}.
\end{aligned}$$

The derivative of the Lyapunov function regarding time is then given as

$$\begin{aligned}
\frac{d\mathcal{L}}{dt} &= \frac{\mu_{E_g}}{\beta_H H_S^*} \left[ -\mu_{H_S} H_S^* h\left(\frac{H_S}{H_S^*}\right) - B_H h\left(\frac{H_S^*}{H_S}\right) - \alpha_S H_R^* h\left(\frac{H_R H_S^*}{H_S H_R^*}\right) \right. \\
&\quad - \beta_H E_g^* H_S^* h\left(\frac{H_E E_g H_S}{H_E E_g^* H_S^*}\right) - \mu_{H_E} H_E^* h\left(\frac{H_E}{H_E^*}\right) - \sigma_H H_E^* h\left(\frac{H_I^* H_E}{H_I H_E^*}\right) \\
&\quad \left. - (\mu_{H_I} + d_{kh}) H_I^* h\left(\frac{H_I}{H_I^*}\right) - r_H H_I^* h\left(\frac{H_R^* H_I}{H_R H_I^*}\right) - (\mu_{H_R} + \alpha_S) H_R^* h\left(\frac{H_R}{H_R^*}\right) \right] \\
&+ \frac{2B_E}{d_{kd}} \left[ -\mu_{D_S} D_S^* h\left(\frac{D_S}{D_S^*}\right) - B_D h\left(\frac{D_S^*}{D_S}\right) - \alpha_D D_R^* h\left(\frac{D_R D_S^*}{D_S D_R^*}\right) \right. \\
&\quad - \beta_D V^* D_S^* h\left(\frac{V D_S D_I^*}{V^* D_S^* D_I^*}\right) - (d_{kd} + \mu_{D_I}) D_I^* h\left(\frac{D_I}{D_I^*}\right) \\
&\quad \left. - r_D D_I^* h\left(\frac{D_R^* D_I}{D_R D_I^*}\right) - \mu_{D_R} D_R^* h\left(\frac{D_R}{D_R^*}\right) \right] \\
&+ \frac{\mu_{E_g}}{\beta_O O_S^*} \left[ -(\mu_{O_S} + \eta_1) O_S^* h\left(\frac{O_S}{O_S^*}\right) - B_O h\left(\frac{O_S^*}{O_S}\right) \right. \\
&\quad \left. - \beta_O E_g^* O_S^* h\left(\frac{E_g O_S O_I^*}{E_g^* O_S^* O_I^*}\right) - \mu_{O_I} O_I^* h\left(\frac{O_I}{O_I^*}\right) \right] \\
&+ \frac{2B_E D_S^*}{d_{kd}} \left[ -\eta_2 O_I^* h\left(\frac{V^* O_I}{V O_I^*}\right) - \beta_D V^* h\left(\frac{V}{V^*}\right) - \gamma V^* h\left(\frac{V}{V^*}\right) \right] \\
&- 2 \left[ B_E D_I^* h\left(\frac{D_I E_g^*}{D_I^* E_g}\right) \right].
\end{aligned}$$

We have that all the solutions and parameters of the system (1) are positive. Moreover the function  $h$  is positive, so the derivative of our Lyapunov function

is clearly less than zero.

So  $\frac{d\mathcal{L}}{dt} \leq 0$ , and  $\frac{d\mathcal{L}}{dt} = 0$  if and only if  $H_S = H_S^*, H_E = H_E^*, H_I = H_I^*, H_R = H_R^*, D_S = D_S^*, D_I = D_I^*, D_R = D_R^*, O_S = O_S^*, O_I = O_I^*, V = V^*, E_g = E_g^*$ .

Then the largest invariant set where  $\frac{d\mathcal{L}}{dt} = 0$ , is the singleton  $N^*$ , which is the endemic equilibrium (12).

Hence using Lassale's invariant principle, when  $t \rightarrow +\infty$ , all solutions of (1) approach the endemic equilibrium if  $R_0 > 1$ .

Finally the endemic equilibrium is GAS when  $R_0 > 1$ .  $\square$

## 4 Optimal control

### 4.1 Optimal control problem

In this section, we propose a formulation of an optimal control problem of the CE disease, based on some prevention techniques proposed in some medical studies in Morocco such as [9].

Indeed we introduce to model (1) two controls  $u_1$  and  $u_2$  described as follows:

- The first control  $u_1$  describes health education in the time interval  $[0, T]$ . Indeed, due to the poor knowledge about CE mechanism in rural areas [20], it is a potent approach to help people understand the behavior of CE disease and how it is transmitted among domestic animals and humans to promote good hygiene practices, secure management of livestock, and show the importance of stopping dogs from consuming infected meat. This will help raise awareness among citizens and improve community engagement to fight the disease more effectively.
- The second control  $u_2$  models the vaccination of livestock with an E. granulosus recombinant antigen (EG95) [25] in the interval of time  $[0, T]$ . As livestock are the intermediate hosts in the transmission cycle of CE (see Figure 1), their vaccination will help effectively break the life cycle of the parasite. The vaccine has shown its worth in China where it is highly utilized, and in

Argentina [25]. In Morocco the EG95 vaccine produced in the country was found in a study elaborated in 2019 [25] to be securing from CE in sheep, during an experimental period of 18 months for 4 groups of 20 animals. However, this approach is not yet applied to the majority of livestock in Morocco.

After applying controls to system (1), we obtain the following controlled system of eleven equations:

$$\begin{aligned}
 \frac{dH_S}{dt} &= B_H - (1 - u_1(t)) \beta_H E_g(t) H_S(t) - \mu_H H_S(t) + \alpha_S H_R(t), \\
 \frac{dH_E}{dt} &= (1 - u_1(t)) \beta_H E_g(t) H_S(t) - (\sigma_H + \mu_H) H_E(t), \\
 \frac{dH_I}{dt} &= \sigma_H H_E(t) - (\mu_H + d_{kh} + r_H) H_I(t), \\
 \frac{dH_R}{dt} &= r_H H_I(t) - (\mu_H + \alpha_S) H_R(t), \\
 \frac{dD_S}{dt} &= B_D - \beta_D V(t) D_S(t) + \alpha_D D_R(t) - \mu_D D_S(t), \\
 \frac{dD_I}{dt} &= \beta_D V(t) D_S(t) - (r_D + \mu_D + d_{kd}) D_I(t), \\
 \frac{dD_R}{dt} &= r_D D_I(t) - (\mu_D + \alpha_D) D_R(t), \\
 \frac{dO_S}{dt} &= B_O - (1 - u_2(t)) \beta_O E_g(t) O_S(t) - (\mu_O + \eta_1) O_S(t), \\
 \frac{dO_I}{dt} &= (1 - u_2(t)) \beta_O E_g(t) O_S(t) - (\mu_O + \eta_2) O_I(t), \\
 \frac{dV}{dt} &= \eta_2 O_I(t) - (\beta_D + \gamma) V(t), \\
 \frac{dE_g}{dt} &= \sigma_E D_I(t) - \mu_{E_g} E_g(t).
 \end{aligned} \tag{21}$$

In this model, we suppose that all the compartments of humans, dogs, and livestock have a unique death rate,  $\mu_H, \mu_D$ , and  $\mu_O$ , respectively.

Moreover, the same initial conditions of model (1), given by (2), are also applied to this controlled model.

We associate to system model (21) the objective functional  $J$  given as

$$J(u_1, u_2) = \int_0^T [C_1 H_E(t) + C_2 H_I(t) + C_3 D_I(t) + C_4 O_I(t) + \frac{1}{2} \sum_{i=1}^{i=2} \theta_i u_i^2(t)] dt, \quad (22)$$

where  $C_i \geq 0$ , for  $i = 1, \dots, 4$  are the gains of the populations  $H_E(t)$ ,  $H_I(t)$ ,  $D_I(t)$ , and  $O_I(t)$ , respectively. The constants  $\theta_i$  for  $i = 1, 2$  represent the weights that balance the associated controls.

Furthermore, the integrand of the objective functional is given by

$$\begin{aligned} \mathcal{L}(H_S, H_E, H_I, H_R, D_S, D_I, D_R, O_S, O_I, V, E, u) \\ = C_1 H_E(t) + C_2 H_I(t) + C_3 D_I(t) + C_4 O_I(t) + \frac{1}{2} \sum_{i=1}^{i=2} \theta_i u_i^2(t). \end{aligned} \quad (23)$$

More precisely, the optimal control problem can be defined as follows:

$$J(u_1^*, u_2^*) = \min_{\Omega} J(u_1, u_2), \quad (24)$$

where

$$\Omega = \left\{ (u_1, u_2) \mid \begin{array}{l} u_i(t) \text{ is Lebesgue measurable} \\ 0 \leq u_i(t) \leq 1, \quad t \in [0, T], \text{ for } i = 1, 2, \end{array} \right\} \quad (25)$$

is the set of admissible controls.

Therefore, the optimal control problem is solved when  $(u_1^*, u_2^*) \in \Omega$  that minimize the function (24), are founded.

## 4.2 Existence of optimal control

In order to solve the optimal control problem, it is first necessary to show the existence of the solution of system (21).

Consider the state variables  $H_S, H_E, H_I, H_R, D_S, D_I, D_R, O_S, O_I, V, E_g$  and the control variables  $u_1, u_2$  with nonnegative initial conditions as given in (2), the system (21) can be written as

$$\mathbb{X}_t = \mathbb{A}\mathbb{X} + \mathbb{B}(\mathbb{X}), \quad (26)$$

where  $\mathbb{X}$  and  $\mathbb{B}$  are given bellow and  $\mathbb{A}$  is given by (27):

$$\mathbb{X} = \begin{bmatrix} H_S \\ H_E \\ H_I \\ H_R \\ D_S \\ D_I \\ D_R \\ O_S \\ O_I \\ V \\ E_g \end{bmatrix}, \quad \mathbb{B}(\mathbb{X}) = \begin{bmatrix} B_H - (1 - u_1(t)) \beta_H E_g H_S \\ (1 - u_1(t)) \beta_H E_g H_S \\ 0 \\ 0 \\ B_D - \beta_D V D_S \\ \beta_D V D_S \\ 0 \\ B_O - (1 - u_2(t)) \beta_O E_g O_S \\ (1 - u_2(t)) \beta_O E_g O_S \\ 0 \\ 0 \end{bmatrix},$$

$\mathbb{X}_t$  is the derivative of  $\mathbb{X}$  with respect to time  $t$ . It is clear that the system (26) is a nonlinear system:

$$\mathbb{A} = \begin{bmatrix} \mu_H & 0 & 0 & \alpha_S & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 \\ 0 & -\sigma_H - \mu_H & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 \\ 0 & \sigma_H & -(\mu_H + d_{KH} + r_H) & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 \\ 0 & 0 & r_H & -(\mu_H + \alpha_S) & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & -\mu_D & 0 & \alpha_D & 0 & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & 0 & -(r_D + \mu_D + d_{kd}) & 0 & 0 & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & 0 & 0 & -(\mu_D + \alpha_D) & 0 & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & 0 & 0 & 0 & -(\mu_O + \eta_1) & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & -(\mu_O + \eta_2) & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & -(\beta_D + \gamma) & 0 & 0 \\ 0 & 0 & 0 & 0 & 0 & \sigma_E & 0 & 0 & 0 & 0 & -\mu_{E_g} & 0 \end{bmatrix}. \quad (27)$$

We pose

$$\mathbb{H}(\mathbb{X}) = \mathbb{A}\mathbb{X} + \mathbb{B}(\mathbb{X}). \quad (28)$$

Let  $\mathbb{X}_1$  and  $\mathbb{X}_2$  be two solutions of system (21). Then,

$$\mathbb{B}(\mathbb{X}_1) - \mathbb{B}(\mathbb{X}_2) = \begin{bmatrix} (1 - u_1(t)) \beta_H [-E_{g_1} H_{S_1} + E_{g_2} H_{S_2}] \\ (1 - u_1(t)) \beta_H [E_{g_1} H_{S_1} - E_{g_2} H_{S_2}] \\ 0 \\ 0 \\ \beta_D (-V_1 D_{S_1} + V_2 D_{S_2}) \\ \beta_D (V_1 D_{S_1} - V_2 D_{S_2}) \\ 0 \\ (1 - u_2(t)) \beta_O (-E_{g_1} O_{S_1} + E_{g_2} O_{S_2}) \\ (1 - u_2(t)) \beta_O (E_{g_1} O_{S_1} - E_{g_2} O_{S_2}) \\ 0 \\ 0 \end{bmatrix}.$$

This term satisfies the following:

$$\begin{aligned}
|\mathbb{B}(\mathbb{X}_1) - \mathbb{B}(\mathbb{X}_2)| &= 2\beta_H (1 - u_1(t)) |E_{g_1}(t)H_{S_1}(t) - E_{g_2}(t)H_{S_2}(t)| \\
&\quad + 2\beta_D |V_1 D_{S_1} - V_2 D_{S_2}| \\
&\quad + 2\beta_O (1 - u_2(t)) |E_{g_1} O_{S_1} - E_{g_2} O_{S_2}| \\
&\leq 2\beta_H (1 - u_1(t)) [|E_{g_2}| |H_{S_1}(t) - H_{S_2}(t)| \\
&\quad + |H_{S_1}| |E_{g_1}(t) - E_{g_2}(t)|] \\
&\quad + 2\beta_D [|D_{S_1}| |V_1(t) - V_2(t)| + |V_2| |D_{S_1} - D_{S_2}|] \\
&\quad + 2\beta_O (1 - u_2(t)) [|E_{g_1}| |O_{S_1} - O_{S_2}| + |O_{S_2}| |E_{g_1} - E_{g_2}|] \\
&\leq 2\beta_H (1 - u_1(t)) [F |H_{S_1}(t) - H_{S_2}(t)| \\
&\quad + \frac{B_H}{\mu_H} |E_{g_1}(t) - E_{g_2}(t)|] \\
&\quad + 2\beta_D \left[ \frac{B_D}{\mu_D} |V_1(t) - V_2(t)| + W |D_{S_1} - D_{S_2}| \right] \\
&\quad + 2\beta_O (1 - u_2(t)) \left[ F |O_{S_1} - O_{S_2}| + \frac{B_O}{\mu_O} |E_{g_1} - E_{g_2}| \right] \\
&\leq 2\beta_H F (1 - u_1(t)) |H_{S_1}(t) - H_{S_2}(t)| \\
&\quad + 2\beta_H \frac{B_H}{\mu_H} (1 - u_1(t)) |E_{g_1}(t) - E_{g_2}(t)| \\
&\quad + 2\beta_D \frac{B_D}{\mu_D} |V_1(t) - V_2(t)| + 2\beta_D W |D_{S_1} - D_{S_2}| \\
&\quad + 2\beta_O F (1 - u_2(t)) |O_{S_1} - O_{S_2}| \\
&\quad + 2\beta_O \frac{B_O}{\mu_O} (1 - u_2(t)) |E_{g_1} - E_{g_2}| \\
&\leq M [|H_{S_1}(t) - H_{S_2}(t)| + |E_{g_1}(t) - E_{g_2}(t)| + |V_1(t) - V_2(t)| \\
&\quad + |D_{S_1} - D_{S_2}| + |O_{S_1} - O_{S_2}|],
\end{aligned}$$

where  $M > 0$  is independent of the variables  $H_S$ ,  $E_g$ ,  $V$ ,  $D_S$ , and  $O_S$ . Therefore;

$$|\mathbb{H}(\mathbb{X}_1) - \mathbb{H}(\mathbb{X}_2)| \leq K |\mathbb{X}_1 - \mathbb{X}_2|, \text{ with } K = \|\mathbb{A}\| + M.$$

Thus, it follows that the function  $\mathbb{H}$  satisfies the Lipschitz condition, uniformly with respect to non-negative state variables.

Therefore, there exists a solution of the system (21).

Now, we present a result that will show the existence of an optimal control

that minimizes the objective functional  $J$  in a finite interval  $[0, T]$ , subjected to the system (21).

**Theorem 5.** There exists an optimal control pair  $u^* = \{u_1^*, u_2^*\}$  in  $\Omega$  and a corresponding solution  $\mathbb{X}^* = \{H_S^*, H_E^*, H_I^*, H_R^*, D_S^*, D_I^*, D_R^*, O_S^*, O_I^*, V^*, E_g^*\}$  such that

$$J(u_1^*, u_2^*) = \min_{\Omega} J(u_1, u_2).$$

*Proof.* The existence of an optimal control is proved using Fleming's results (Theorem (III.4.1) and its corresponding corollary in [10]). Indeed we must verify those five conditions:

**P1.** The set solutions to the system (21)–(2) and its corresponding controls are nonempty (proved lastly).

**P2.** The set of controls is convex and closed.

We consider

$$\Omega = \{u \in \mathbb{R}^2 : \|u\| \leq 1\}.$$

Let  $u, v \in \Omega$  such that  $0 \leq \|u\| \leq 1$  and  $0 \leq \|v\| \leq 1$ .

Then, for any  $\varepsilon \in [0, 1]$ , we have

$$0 \leq \|\varepsilon u + (1 - \varepsilon)v\| \leq \varepsilon\|u\| + (1 - \varepsilon)\|v\| \leq 1.$$

**P3.** The state system can be written as linear function of control variables with coefficients depending on time and state variable.

**P4.** The integrand (23) of objective function is convex with respect to controls.

Let  $\mathbb{X} = (H_S, H_E, H_I, H_R, D_S, D_I, D_R, O_S, O_I, V, E)$ ,  $u = (u_1, u_2) \in \Omega$  and  $v = (\{v_1, v_2\}) \in \Omega$ .

Then, for  $\varepsilon \in [0, 1]$ , we have

$$\mathcal{L}(\mathcal{X}, \varepsilon u + (1 - \varepsilon)v) = C_1 H_E(t) + C_2 H_I(t) + C_3 D_I(t) + C_4 O_I(t) + \frac{1}{2} \sum_{i=1}^{i=2} \theta_i (\varepsilon u_i + (1 - \varepsilon)v_i)^2,$$

and

$$\varepsilon \mathcal{L}(\mathcal{X}, u) + (1 - \varepsilon) \mathcal{L}(\mathcal{X}, v) = C_1 H_E(t) + C_2 H_I(t) + C_3 D_I(t) + C_4 O_I(t)$$

$$+ \frac{1}{2}\varepsilon \sum_{i=1}^{i=2} \theta_i u_i^2 + \frac{1}{2}(1-\varepsilon) \sum_{i=1}^{i=2} \theta_i v_i^2.$$

Therefore,

$$\mathcal{L}(\mathcal{X}, \varepsilon u + (1-\varepsilon)v) - (\varepsilon \mathcal{L}(\mathcal{X}, u) + (1-\varepsilon) \mathcal{L}(\mathcal{X}, v)) = \frac{1}{2}(\varepsilon^2 - \varepsilon) \sum_{i=1}^{i=2} \theta_i (u_i - v_i)^2,$$

and

$$\mathcal{L}(\mathcal{X}, \varepsilon u + (1-\varepsilon)v) - (\varepsilon \mathcal{L}(\mathcal{X}, u) + (1-\varepsilon) \mathcal{L}(\mathcal{X}, v)) \leq 0, \quad \text{since } \varepsilon \in [0, 1].$$

**P5.** There exist constants  $\vartheta_1, \vartheta_2 > 0$  and  $\vartheta_3 > 1$  as the integrand (23) is bounded by

$$\vartheta_1(|u_1|^2 + |u_2|^2)^{\vartheta_3/2} - \vartheta_2.$$

In fact it is easy to verify that

$$\begin{aligned} \mathcal{L}(\mathcal{X}, u) &\geq \frac{1}{2} \sum_{i=1}^{i=2} \theta_i u_i^2 \\ &\geq \vartheta_1(|u_1|^2 + |u_2|^2)^{\vartheta_3/2} - \vartheta_2, \end{aligned}$$

where  $\vartheta_1 = \frac{1}{2} \min\{\theta_1, \theta_2\}$ ,  $\vartheta_2 > 0$  and  $\vartheta_3 = 2$ . □

### 4.3 Characterization of optimal control

Theorem 5 assures the existence of an optimal control  $(u_1^*, u_2^*)$  that minimizes the objective functional (22).

The Pontryagin's maximum principle transform system (21) with (22) and (24) to a minimization problem of Hamiltonian, that we define as

$$\begin{aligned} H(\mathbb{X}, u, \Lambda, t) &= C_1 H_E(t) + C_2 H_I(t) + C_3 D_I(t) + C_4 O_I(t) + \frac{1}{2} \sum_{i=1}^{i=2} \theta_i u_i^2(t) \\ &\quad + \lambda_1(t) \frac{dH_S(t)}{dt} + \lambda_2(t) \frac{dH_E(t)}{dt} + \lambda_3(t) \frac{dH_I(t)}{dt} + \lambda_4(t) \frac{dH_R(t)}{dt} \\ &\quad + \lambda_5(t) \frac{dD_S(t)}{dt} + \lambda_6(t) \frac{dD_I(t)}{dt} + \lambda_7(t) \frac{dD_R(t)}{dt} + \lambda_8(t) \frac{dO_S(t)}{dt} \\ &\quad + \lambda_9(t) \frac{dO_I(t)}{dt} + \lambda_{10}(t) \frac{dV(t)}{dt} + \lambda_{11}(t) \frac{dE(t)}{dt}, \end{aligned} \quad (29)$$

with  $\mathbb{X} = \{H_S, H_E, H_I, H_R, D_S, D_I, D_R, O_S, O_I, V, E_g\}$ ,  $u = \{u_1, u_2\} \in \Omega$ , and  $\Lambda = \{\lambda_i, i = 1, \dots, 11\}$  the adjoint variables.

Necessary condition that  $\{\mathbb{X}^*(t), u^*(t)\}$  can be an optimal solution for the optimal control problem is the existence of a nontrivial vector function  $\Lambda(t) = \{\lambda_i(t), i = 1, \dots, 11\}$  such that

$$\begin{aligned} \frac{d\mathbb{X}}{dt} &= \frac{\partial H(\mathbb{X}^*(t), u^*(t), \Lambda(t), t)}{\partial \mathbb{X}}, \\ 0 &= \frac{\partial H(\mathbb{X}^*(t), u^*(t), \Lambda(t), t)}{\partial u}, \\ \Lambda'(t) &= -\frac{\partial H(\mathbb{X}^*(t), u^*(t), \Lambda(t), t)}{\partial \Lambda}. \end{aligned} \quad (30)$$

The next result presents the adjoint system and the characterization of optimal control.

**Theorem 6.** Given an optimal control  $u^* = \{u_1^*, u_2^*\}$  and corresponding solutions

$\mathbb{X}^* = \{H_S^*, H_E^*, H_I^*, H_R^*, D_S^*, D_I^*, D_R^*, O_S^*, O_I^*, V^*, E_g^*\}$  that minimize  $J(u)$

over  $\Omega$ , there exist adjoint variables  $\lambda_i$ ,  $i = 1, 2, \dots, 11$  satisfying

$$\frac{d\lambda_1(t)}{dt} = (1 - u_1(t)) \beta_H E_g(t) [\lambda_1(t) - \lambda_2(t)] + \mu_H \lambda_1(t),$$

$$\frac{d\lambda_2(t)}{dt} = -C_1 + \sigma_H [\lambda_2(t) - \lambda_3(t)] + \mu_H \lambda_2(t),$$

$$\frac{d\lambda_3(t)}{dt} = -C_2 + (\mu_H + d_{KH} + r_H) \lambda_3(t) - r_H \lambda_4(t),$$

$$\frac{d\lambda_4(t)}{dt} = -\alpha_S \lambda_1(t) + (\mu_H + \alpha_S) \lambda_4(t),$$

$$\frac{d\lambda_5(t)}{dt} = \beta_D V [\lambda_5(t) - \lambda_6(t)] + \mu_D \lambda_5(t),$$

$$\frac{d\lambda_6(t)}{dt} = -C_3 + (r_D + \mu_D + d_{KD}) \lambda_6(t) - r_D \lambda_7(t) - \sigma_E \lambda_{11}(t),$$

$$\frac{d\lambda_7(t)}{dt} = -\alpha_D \lambda_5(t) + (\mu_D + \alpha_D) \lambda_7(t),$$

$$\frac{d\lambda_8(t)}{dt} = (1 - u_2(t)) \beta_O E_g(t) [\lambda_8(t) - \lambda_9(t)] + (\eta_1 + \mu_O) \lambda_8(t),$$

$$\frac{d\lambda_9(t)}{dt} = -C_4 + (\eta_2 + \mu_O) \lambda_9(t) - \eta_2 \lambda_{10}(t),$$

$$\frac{d\lambda_{10}(t)}{dt} = (\beta_D + \gamma) \lambda_{10}(t),$$

$$\begin{aligned} \frac{d\lambda_{11}(t)}{dt} = & (1 - u_1(t)) \beta_H H_S(t) [\lambda_1(t) - \lambda_2(t)] \\ & + (1 - u_2(t)) \beta_O O_S(t) [\lambda_8(t) - \lambda_9(t)] + \mu_{E_g} \lambda_{11}(t), \end{aligned}$$

where final time conditions are

$$\lambda_i(T) = 0, \quad i = 1, 2, \dots, 11 \quad (31)$$

Moreover, the following characterization holds:

$$\begin{cases} u_1^*(t) = \max\{\min\{1, \frac{\beta_H[-\lambda_1(t) + \lambda_2(t)]E_g^*(t)H_S^*(t)}{\theta_1}\}, 0\}, \\ u_2^*(t) = \max\{\min\{1, \frac{\beta_O[-\lambda_8(t) + \lambda_9(t)]E_g^*(t)O_S^*(t)}{\theta_2}\}, 0\}. \end{cases} \quad (32)$$

*Proof.* The form of the adjoint system endowed with terminal conditions (31) results from Pontryagin's maximum principle by differentiating the Hamiltonian function (29) at the respective solutions of the state system (21).

Also, to get the characterizations of the optimal control given by (32), we use the optimality conditions.

After solving the equation

$$\frac{\partial H(\mathcal{X}^*(t), u^*(t), \Lambda(t), t)}{du_i} = 0, \text{ for } i = 1, 2,$$

we obtain directly (32) by taking to account the boundedness condition given in (25).  $\square$

## 5 Numerical simulations

In this section, we present numerical simulations done using MATLAB to understand more the comportment of the model of transmission of Cystic Echinococcosis, and the influence of the two controls  $u_1$  and  $u_2$  representing health education and Vaccination of sheep with the EG95 vaccine respectively on the dynamic of the studied populations. Note that in this work, we use the data found in some studies elaborated in some regions of Morocco (Middle Atlas region, Ifrane and El Hajeb ) [2, 1, 18, 19].

The system of stats (1) is resolved in the sens  $t : 0 \rightarrow T$  using the ode45 solver in MATLAB, with the initial conditions  $HS(0) = 5367$ ,  $H_E(0) = 0$ ,  $H_I(0) = 0$ ,  $H_R(0) = 0$ ,  $D_S(0) = 700$ ,  $D_I(0) = 37$ ,  $D_R(0) = 3$ ,  $O_S(0) = 797$ ,  $O_I(0) = 159$ ,  $V(0) = 34$ ,  $E_g(0) = 240$ , where the values of  $H_S(0)$ ,  $D_S(0)$ , and  $O_S(0)$  are derived from [6] and the other initial conditions of (1) are assumed. However the optimality system consisting of control system (21) and the adjoint equations (31), is resolved in the opposite sens  $t : T \rightarrow 0$ , with the transversality conditions  $\lambda_i(T) = 0, 1 \leq i \leq 11$ , where  $T = 60$  months.

We set in (22) also  $C_1 = 50, C_2 = 90, C_3 = 70, C_4 = 60, \theta_1 = 10, \theta_2 = 10$ .

### Parameter selection

- Some parameters of the system (21) have been considered relying on online news, data fitting, or assumption because Echinococcosis is an overlooked disease and only few studies have been done on it in Morocco, and we found it difficult at the time of the study to find all parameters values in the studied areas. The values of the parameters of the system (21) are given in Table 2.
- Some of the table's parameters were based on a study in China because such parameters were unavailable in Morocco, and the choice of those specific values is explained below:
- Parameters  $\mu_{Eg}$  and  $B_E$  are considered universal because they depend on biological and environmental factors, such as humidity and temperature, that are comparable in regions where CE is endemic.
- For the parameter  $r_H$ : the recovery rate is also a universal parameter that depends on the physical context and is common among populations.
- Parameters  $\mu_D$  and  $\mu_O$ : Specific parameters related to the regions studied in Morocco were not available at the time of the study, so we tried to derive those rates from other studies conducted in areas where farming practices and socioeconomic conditions do not vary much.
- So, we assumed that those parameters would be close to their real values in Morocco.

We calibrated the parameters derived from references in Table 2 in months.

-The parameters  $B_H$ ,  $B_D$ , and  $B_O$  are calculated using the same method as in [24].

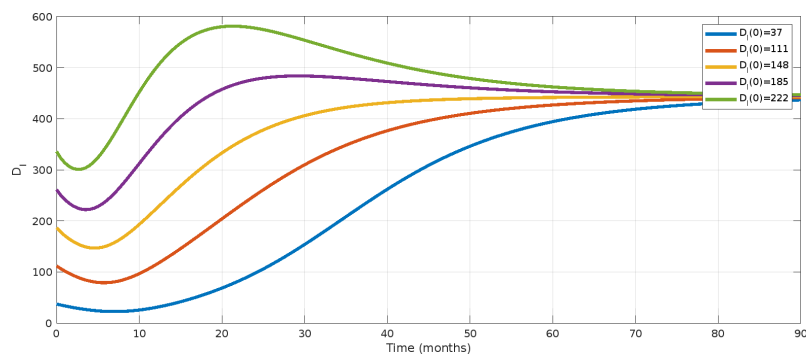


Figure 3: The influence of initial conditions on the number of infected dogs

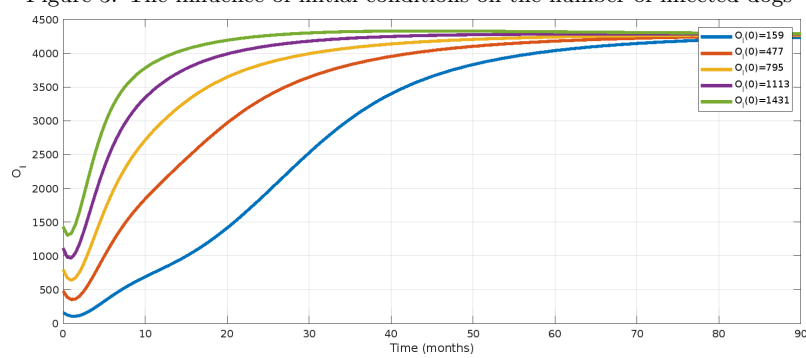


Figure 4: The influence of initial conditions on the number of infected livestock

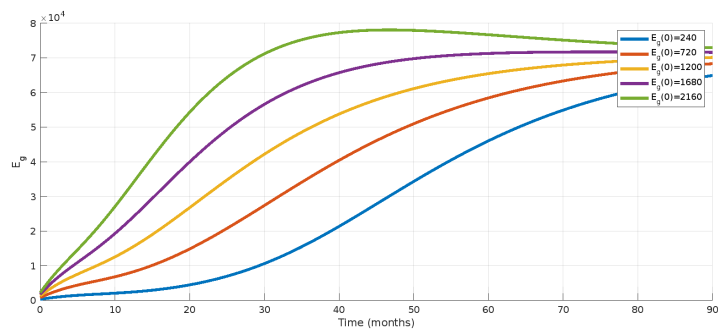


Figure 5: The influence of initial conditions on the number of CE eggs

Table 2: Parameters of system (1)

Parameters	Values	Source
$B_H$	277.3	Calculation
$\beta_H$	1.6% - 2.4%	[6]
$\mu_H$	5.167	[15]
$\sigma_H$	0.0714	Estimated
$d_{kh}$	2% - 3%	[20]
$r_H$	0.075	[11]
$B_D$	56	Calculation
$\beta_D$	2% - 41%	[1]
$\mu_D$	0.08	[24]
$d_{kd}$	0.010	Estimated
$r_D$	22.3%	[1]
$B_O$	121	Calculation
$\beta_O$	0.007	[9]
$\mu_O$	0.152	[12]
$\eta_1$	0.37	Estimated
$\eta_2$	0.37	Estimated
$\gamma$	0.315	Estimated
$B_E$	0.808	[11]
$\mu_{Eg}$	10.42	[11]
$\alpha_S$	8.5%	[22]
$\alpha_D$	4%	[1]

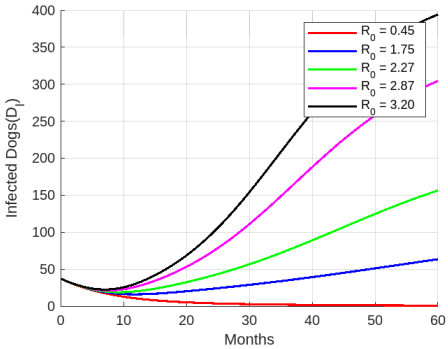


Figure 6: Sensitivity of the infected dogs to  $R_0$

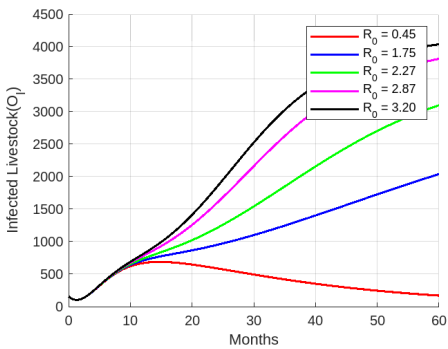


Figure 7: Sensitivity of the infected livestock to  $R_0$

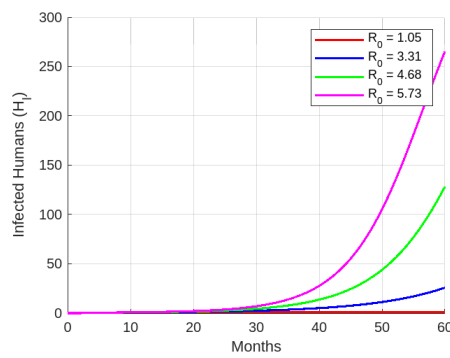
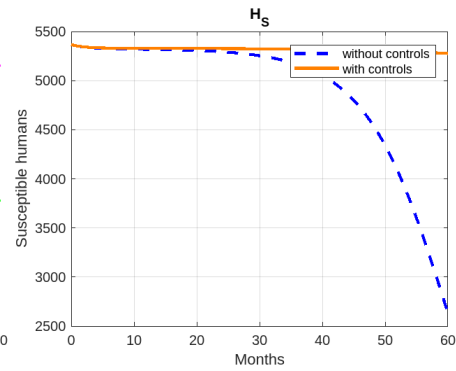
Figure 8: Sensitivity of the infected Humans to  $R_0$ 

Figure 9: Evolution of CE in susceptible humans with and without control

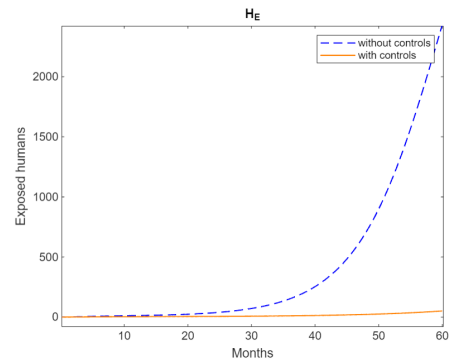


Figure 10: The evolution of CE in exposed humans with and without control

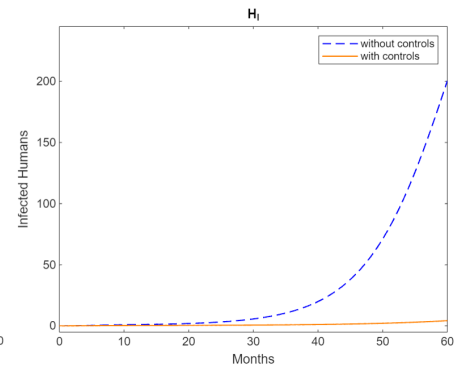


Figure 11: The evolution of CE in infected humans with and without control

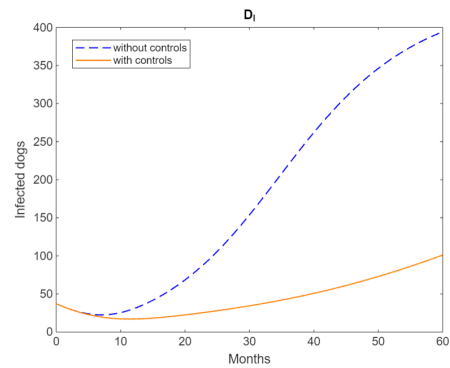


Figure 12: The evolution of CE in infected dogs with and without control

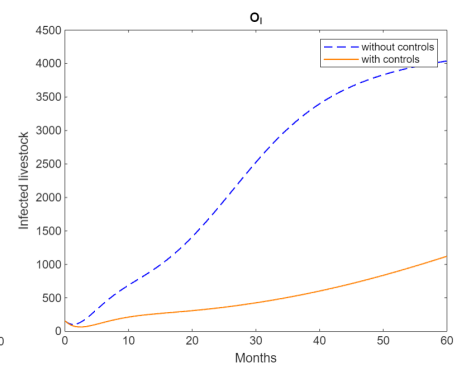


Figure 13: The evolution of CE in infected livestock with and without control

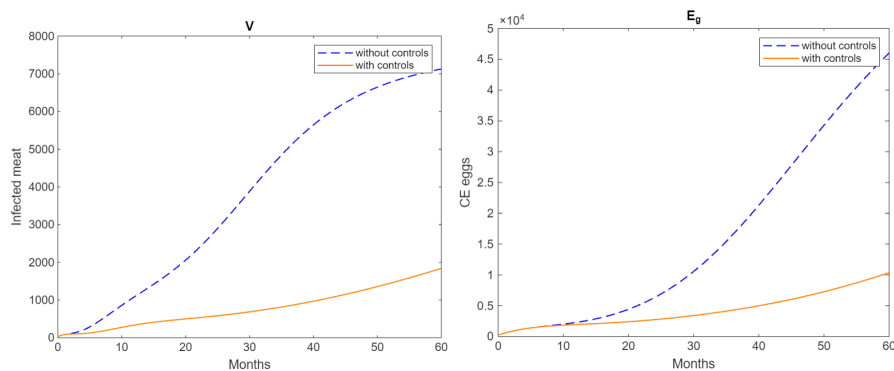


Figure 14: The evolution of infected meat with and without control

Figure 15: The evolution of CE eggs with and without control

## 5.1 Discussion

Figures 3, 4, and 5 show that despite the different values of initial conditions, the system reaches the same steady state, that is, in our case, the endemic equilibrium. The results of the simulations in the figures match the property describing that the endemic equilibrium is globally asymptotically stable. That confirms our theoretical results.

In Figures 8, 6, and 7, we changed the values of the basic reproduction number, and we studied the impact of those variations on the number of infected humans, dogs, and livestock, respectively. The values of  $R_0$  were computed by modifying the rates  $\beta_D$  and  $\beta_O$  at each time.

We remarked that those two rates directly affect the  $R_0$ , and the bigger they get, the greater  $R_0$  becomes. As a result, the infection becomes more important in the three populations with each rise of  $\beta_D$  and  $\beta_O$ .

Based on the results of simulations in Figures 8, 6, and 7, we have chosen two control strategies that directly affect the rates of transmission for humans, dogs and livestock  $\beta_H$ ,  $\beta_D$ , and  $\beta_O$ , by implementing health education and vaccination of livestock, respectively, as described in the previous section.

The purpose of numerical simulations given in Figures 10, 9, 11, 12, 13, 14, 15 is to evaluate the effectiveness and the impact of health education and

livestock vaccination on the spread of CE, those simulations are done while implementing the two controls  $u_1$  and  $u_2$  in the same time.

From the figures above, we remark that in the uncontrolled scenario (dashed blue lines), there is a clear increase in the infected compartments including exposed and infected humans  $H_E, H_I$  (after 20 months) in Figures 10 and 11, infected dogs  $D_I$  (after nearly 7 months) in Figure 12, infected livestock  $O_I$  (after almost 1 month) in Figure 13, infected meat  $V$  (from the start) in Figure 14, and Echinococcosis eggs  $E_g$  (from the start) in Figure 15. That means that without control there is a growing prevalence of the disease in the concerned regions of Morocco starting after month 20 of the study or even from the very beginning for some compartments and reaching the extreme after 60 months. In fact, in this case, the parasite continues its life cycle normally, causing smooth transmission between the intermediate and definitive hosts. In this situation, we spot that the density of the concerned populations begins to increase ( $H_E, H_I, D_I, O_I, V, E_g$ ) until they stabilize at values representing the endemic equilibrium (12) after the 60 months.

However, while implementing the two controls  $u_1$  and  $u_2$  (orange lines), we observe an overall notable decrease in the prevalence of CE. In fact, the curves show that exposed and infected humans remain at a level low to zero (Figures 10 and 11). This means that the control  $u_1$  consisting of health education was very beneficial in stopping so many people from becoming exposed or infected.

From Figure 12, we spot that infected dogs in the controlled case also grow, but at a significantly lower level ( $\sim 100$ ) compared to the uncontrolled case ( $\sim 400$ ). The other compartments  $H_E, H_I, D_I, O_I, V, E_g$  follow the same trend. This means that the second control  $u_2$  of sheep vaccination clearly decreases infected populations over time because it reduces the chance of carrying the parasite inside the intermediate hosts.

Less contamination in sheep leads to lower infection in dogs and also humans due to the break of the life-cycle of the parasite.

Even with the success of the optimal control strategy, the curves representing the Echinococcosis eggs in Figure 15 show that the density of these eggs remains important (approximately 12,000) and is expected to grow more af-

ter the study period, this can be explained by the very high resistance of the eggs to environmental conditions, and it is the source of the remaining infected cases among the populations of dogs and sheep (Figures 12 and 13). This suggests that more long-term control alternatives are needed to stop the disease more effectively by stopping or reducing the continued presence of CE eggs.

The results of the simulations show the effectiveness of health education and sheep vaccination controls in reducing disease in the regions studied in Morocco. However, if no control measures are taken in the country, CE disease will be sustained in human, livestock, and dog populations, which can be very dangerous to the health and food security of Moroccans in the next few years.

## 6 Conclusion

To understand the dynamic of the transmission of CE among humans, dogs, and livestock, we proposed a mathematical model based on the transmission cycle of the disease and taking into consideration the relapse after recovery for humans and dogs. The mathematical analysis of the model allowed us to define a basic reproduction number  $R_0$  that determines the transmission of CE. The results suggested that, when  $R_0 < 1$ , the DFE  $N_0$  is globally asymptotically stable and in this case, the disease dies. When  $R_0 > 1$ , the endemic equilibrium  $N^*$  is globally asymptotically stable, the disease persistently remains. Those theoretical results are confirmed by some numerical simulations. Based on those results, we introduced two control strategies that are health education and sheep vaccination, and we performed the analysis using the optimal control theory. We proved the existence of optimal controls and characterized them using Pontryagin's maximum principle. The plots of the controlled system applied to some regions of Morocco facilitated the comparison between controlled and uncontrolled scenarios. The results revealed that health education combined with the vaccination of sheep with the EG95 vaccine is a must to reduce the prevalence of the disease in Morocco for the next decades. Therefore to sustain the prevention of Cystic

Echinococcosis in Morocco, it is crucial to adopt a continued investment in livestock vaccination, and raise awareness among citizens, especially in rural areas of the kingdom. Due to the relatively high founded level of CE eggs in the environment even with applying the current control strategies, in our future studies, we will integrate into the model other control measures such as inspection of infection in dog populations at least 4 times/year and slaughterhouse monitoring to reduce the environmental contamination. Moreover, we will try to address the randomness in the transmission process of CE disease by adding some stochastic parameters to the deterministic system, and analyze the new stochastic model.

## References

- [1] Amarir, F., Rhalem, A., Sadak, A., Raes, M., Oukessou, M., Saadi, A., Bouslikhane, M., Gauci, C.G., Lightowlers, M.W., Kirschvink, N. and Marcotty, T. *Control of cystic echinococcosis in the Middle Atlas, Morocco: Field evaluation of the EG95 vaccine in sheep and cesticide treatment in dogs*, PLoS Negl. Trop. Dis. 15(3) (2021) e0009253.
- [2] Amarir, F.E., Saadi, A., Marcotty, T., Rhalem, A., Oukessou, M., Sahibi, H. Obtel, M., Bouslikhane, M., Sadak, A. and Kirschvink, N. *Cystic echinococcosis in three locations in the Middle Atlas, Morocco: Estimation of the infection rate in the dog reservoir*, Vector-Borne and Zoonotic Diseases, 20 (6) (2020) 436–443.
- [3] Azlaf, R. and Dakkak, A. *Epidemiological study of the cystic echinococcosis in Morocco*, Vet. Parasitol. 137 (2006) 83–93.
- [4] Birhan, G.B., Munganga, J.M.W. and Hassan, A.S. *Mathematical modeling of echinococcosis in humans, dogs, and sheep*. J. Appl. Math.2020 (2020) Article ID 8482696, 18 pages.
- [5] Chacha, C.S., Stephano, M.A., Irunde, J.I. and Mwasunda, J.A. *Cystic echinococcosis dynamics in dogs, humans and cattle: Deterministic and stochastic modeling*, Results Phys. 51 (2023) 106697.

- [6] Chebli, H., Laamrani El Idrissi, A., Benazzouz, M., Lmimouni, B.E., Nhammi, H., Elabandouni, M., Youbi, M., Afifi, R., Tahiri, S., Essayd El Feydi, A. and Settaf, A. *Human cystic echinococcosis in Morocco: Ultrasound screening in the Mid Atlas through an Italian-Moroccan partnership*, PLoS Negl. Trop. Dis. 11(3) (2017) e0005384.
- [7] Diekmann, O., Heesterbeek, J.A.P. and Metz, J.A.J. *On the definition and the computation of the basic reproduction ratio  $R_0$  in models for infectious diseases in heterogeneous populations*. J. Math. Biol. 28(4) (1990) 365–382.
- [8] El-Berbri, I. *Epidémiologie de l'échinococcose/ hydatidose kystique dans la province de Sidi Kacem et évaluation d'actions de lutte conséquentes dans le cadre d'une approche intégrée associant la leishmaniose viscérale et la rage*, PhD diss., Ph. D Thesis. IAV Hassan II, Rabat, Morocco, 2015.
- [9] El-Berbri, I., Mahir, W., Fihri, O.F., Petavy, A.F., Dakkak, A. and Bouslikhane, M. *Cystic echinococcosis in Morocco: epidemiology, socio-economic impact and control*, J. Vet. Parasitol. 34(2) (2020) 72–78.
- [10] Fleming, W.H. and Rishel, R.W. *Deterministic and stochastic optimal control*, Springer-Verlag, New York, 1975.
- [11] Gong, W. and Wang, Z. *Sliding motion control of Echinococcosis transmission dynamics Model*, Math. Comput. Simul. 205 (2023) 468–482.
- [12] He, Y., Cui, Q. and Hu, Z. *Modeling and analysis of the transmission dynamics of cystic echinococcosis: Effects of increasing the number of sheep*, Math. Biosci. Eng. 20 (2023) 14596–14615.
- [13] Jazouli, M., Lightowlers, M.W., Bamouh, Z., Gauci, C.G., Tadlaoui, K., Ennaji, M.M. and Elharrak, M. *Immunological responses and potency of the EG95NC- recombinant sheep vaccine against cystic echinococcosis*, Parasitol. Inter. 78 (2020) 102149.
- [14] Lahmar, S., Debbek, H., Zhang, L.H., McManus, D.P., Souissi, A., Chelly, S. and Torgerson, P.R. *Transmission dynamics of the Echinococ-*

- cus granulosis sheep-dog strain (G1 genotype) in camels in Tunisia*, Vet. Parasitol. 121(1-2) (2004) 151–156.
- [15] Macrotrends. Morocco death rate 1950-2024. <https://www.macrotrends.net/global-metrics/countries/MAR/morocco/death-rate>
- [16] Ouldkhouia, N., Elberrai, I., Adnaoui, K. and Benhachem, A. *A general stochastic model for tumor growth: Simulating cardiac tumor (Myxoma) development*, Math. Model. Eng. Probl. 11(12) (2024) 3323–3332.
- [17] Ouldkhouia, N., Elberrai, I., Benhachem, A., Sannaky, I., Riouali, M. and Hachoum, S. *On stochastic modelling: The impact of advertisement on the consumption-application on ChatGPT-3*, Math. Model. Eng. Probl. 11 (10) (2024) 2705–2714.
- [18] Saadi, A., Amarir, F., Filali, H., Thys, S., Rhalem, A., Kirschvink, N., Raes, M., Marcotty, T., Oukessou, M., Duchateau, L. and Sahibi, H. *The socio-economic burden of cystic echinococcosis in morocco: A combination of estimation method*, PLoS Negl. Trop. Dis., 14 (2020) 1–20.
- [19] Tahiri, S., Naoui, H., Iken, M., Azelmat, S., Bouchrik, M. and Lmi-mouni, B.E. *RETRAIT: Seroprevalence of cystic echinococcosis in the provinces of Ifrane and El Hajeb in Morocco*, Med. Mal. Infect. (2020) 31928912.
- [20] Thys, S., Sahibi, H., Gabriël, S., Rahali, T., Lefèvre, P., Rhalem, A., Marcotty, T., Boelaert, M. and Dorny, P. *Community perception and knowledge of cystic echinococcosis in the High Atlas Mountains, Morocco*. BMC Public Health 19 (2019) 1-15.
- [21] 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.
- [22] Velasco-Tirado, V., Romero-Alegría, Á., Belhassen-García, M., Alonso-Sardón, M., Esteban-Velasco, C., López-Bernús, A., Carpio-Perez, A.,

- Jimenez López, M.F., Muñoz Bellido, J.L., Muro, A. and Cordero-Sanchez, M. *Recurrence of cystic echinococcosis in an endemic area: a retrospective study*, BMC Infectious Diseases 17 (2017) 1–8.
- [23] Wang, K., Teng, Z. and Zhang, X. *Dynamical behaviors of an echinococcosis epidemic model with distributed delays*, Math. Biosci. Eng. 14 (5&6) (2017) 1425–1445.
- [24] Wang, K., Zhang, X., Jin, Z., Ma, H., Teng, Z. and Wang, L. *Modeling and analysis of the transmission of echinococcosis with application to Xinjiang Uygur Autonomous Region of China.*, J. Theor. Biol. 333 (2013) 78–90.
- [25] Weekly epidemiological record, (48) 2019 29. <https://iris.who.int/bitstream/handle/10665/330007/WER9448-574-579-eng-fre.pdf?sequence=1&isAllowed=y>.
- [26] World Health Organization. (n.d). Cystic and alveolar echinococcosis. [https://www.who.int/docs/default-source/ntds/echinococcosis/cystic-and-alveolar-echinococcosis.pdf?sfvrsn=fa73b7ca\\_8](https://www.who.int/docs/default-source/ntds/echinococcosis/cystic-and-alveolar-echinococcosis.pdf?sfvrsn=fa73b7ca_8).
- [27] World Health Organization. Echinococcosis fact sheet. 2020. <https://www.who.int/news-room/fact-sheets/detail/echinococcosis>.



# A new generalized model of cooperation of advertising companies based on differential games on networks

M. Jashnesade, Z. Nikoeeinejad\* and G.B. Loghmani

## Abstract

In this paper, we reconsider the sustainable cooperation of advertising companies problem on a network of companies and consumers. The aim of this paper is to investigate cooperation and profit distribution within networks involving companies and consumers with asymmetric roles and to compare two scenarios based on advertising efficiency. We extend a model

---

\*Corresponding author

Received 23 February 2025; revised 27 April 2025; accepted 20 May 2025

Marjan Jashnesade

Department of Applied Mathematics, Faculty of Mathematical Sciences, Yazd University, Yazd, Iran. e-mail: [jashnesade@stu.yazd.ac.ir](mailto:jashnesade@stu.yazd.ac.ir)

Zahra Nikoeeinejad

Department of Applied Mathematics, Faculty of Mathematical Sciences, Yazd University, Yazd, Iran. e-mail: [z.nikoeeinejad@yazd.ac.ir](mailto:z.nikoeeinejad@yazd.ac.ir)

Ghasem Barid loghmani

Department of Applied Mathematics, Faculty of Mathematical Sciences, Yazd University, Yazd, Iran. e-mail: [Loghmani@yazd.ac.ir](mailto:Loghmani@yazd.ac.ir)

## How to cite this article

Jashnesade, M. Nikoeeinejad, Z. and Loghmani, G.B. , A new generalized model of cooperation of advertising companies based on differential games on networks. *Iran. J. Numer. Anal. Optim.*, 2025; 15(3): 1116-1144.  
<https://doi.org/10.22067/ijnao.2025.92290.1601>

in the framework of cooperative differential games on network, which derive analytically the construction of the characteristic function. Unlike previous dynamic marketing models that all firms advertise for homogeneous consumers and they view companies as identical, in the proposed generalized model, we assume that companies and consumers are heterogeneous players in the network. The Shapley value, Core and  $\tau$ -value are obtained analytically. A numerical simulation of the proposed model for a graph of advertising firms and consumers with different parameters is presented to obtain sensitivity analysis results. The results show that the optimal cooperative advertisement rate of firms are affected by the sensitivity parameters of consumers to companies. Furthermore, we demonstrate that the allocations of the companies in the cooperative game only for grand coalition will be greater than in the non-cooperative one. We find that the greater the coalition, the more gains from cooperative advertising. We also demonstrated that for the higher product price, the greater Shapley and  $\tau$ -values attributed to the firms.

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

**Keywords:** Differential game; Cooperative advertising; Shapley value; Network.

## 1 Introduction

In the current competitive landscape, neglecting cooperative advertising can result in missed strategic opportunities. Many businesses are finding it difficult to create supply chains that make the most money and keep customers happy. Recently, companies need to work together and agree on things like prices, how much to order, and advertising to succeed in a competitive global market. Instead of competing against each other, businesses are teaming up to make more money and reduce risks. Advertising is increasingly important for businesses to grow market share and revenue. A cooperative advertising strategy is essential for a win-win situation.

Cooperative advertising boosts exposure by allowing more ad outlets and partnering with a major company for increased visibility. Leading firms invest significantly in researching advanced advertising methodologies. Participa-

tion in cooperative advertising programs enables companies to leverage the experiences of more successful firms, contributing to their business growth. Companies use various media like TV, radio, print, Internet and social network for advertising. They also utilize local media and distribute fliers and often share advertising costs with retailers through cooperative advertising program (co-op), providing product images for ads. The cost-sharing depends on the relationship and brand prominence. Cooperative advertising is a crucial strategy for manufacturers to expand market capacity and maximize their self-interest, as per existing literature.

In this study, advertising companies are modeled as players within a game-theoretic framework. According to each company's network structure, we believe that the advantages of cooperative advertising for businesses in networks include growing the number of consumers and boosting profits through the synergy of joint advertising. In this research, ideas from graph theory are applied to construct a cooperative differential game for the network's advertising problem. Also, many cooperative game theory techniques are put out to ensure that the profits from the cooperative advertisements are distributed fairly among the participating companies. Accordingly, this study aims to address the following research questions:

- How can corporations determine the amount of money they make from social network advertising?
- How might more cooperative advertising income be equitably divided among participating agents?
- How will advertising affect the coalition's companies?

The remainder of the paper is arranged as follows. In Section 2, the relevant literature is examined. In addition to defining the mathematical model for cooperative advertising in networks, Section 3 offers a solution methodology. In Section 4, one numerical example is considered, and Section 5 explains the solution imputation and fair distribution of cooperative game models. Section 6 concludes with some recommendations and future research directions.

## 2 Literature review

Three main literature streams serve as the foundation for our work: Cooperative advertising, cooperative game theory applications in networks and cooperative advertising differential game. As a result, we go over these topics in the subsections that follows.

Several techniques based on cooperative game theory have been developed in recent years. In order to communicate the information to their network partners, Zinoviev and Duong [37] expanded the game theoretical model of information diffusion in a star-like network where personality traits are described through a feedback mechanism. In order to calculate the profit margin for their shops and customers, Aust [2] assessed the competition. Increased market competitiveness leads to higher profits and more effective advertising outcomes. In [33], authors investigated three cooperative scenarios for the horizontal firms: one in which a contractor is attempting to create a coalition, as well as cooperative and noncooperative scenarios. For every scenario, the best possible advertising and revenues were determined and contrasted.

The majority of research on cooperative advertising in the literature has concentrated on supply chain advertising campaigns. Jørgensen and Zaccour [19] presented a comprehensive review of cooperative advertising in marketing channels using game-theoretic approaches. Their survey was structured into two parts. The first part focused on simple supply chains consisting of a single supplier and a single retailer, while the second part examined more complex channels involving multiple suppliers and/or retailers. They noted that many findings from static models could be extended to dynamic settings and that static models tended to be homogeneous in terms of assumptions and structure. In contrast, models that incorporated horizontal interactions within or across supply chain levels were relatively scarce. The study also showed that firms' participation in cooperative advertising programs is influenced by both intra-brand and inter-brand competition, and that such participation do not always align with the firms' best interests. Sarkar, Omair, and Kim [25] proposed a model for co-op advertising in supply chains, which considers fluctuating demand based on selling price and advertising expenses. The ideal outcomes improved revenue across the supply chain, overcoming

risks and improving economic analysis and feasibility. In [35], authors studied cooperative advertising in a supply chain involving manufacturers and retailers, examining how power and information structure affect price and advertising decisions. The study found that a dominating member can enhance supply chain performance. In [34], the authors studied how digital platforms and participants balance value generation and appropriation in a Stackelberg game. They emphasized the necessity of cooperative partnerships in strategic alliances. The corona virus epidemic has generated new marketing methods, influencing client purchase behaviors. Ghosh, Seikh, and Chakraborty [13] examined how cooperative advertising and online services affect decision-making in dual channel supply chains. The model revealed the most profitable channel methods, highlighting the benefits of cooperative advertising but downsides for traditional shops. Jørgensen, Signé, and Zaccour [18] examined how a manufacturer and exclusive retailer's advertising affects sales in a two-member channel. The study analyzed four scenarios: no support, support for both forms of advertising, and partial support for one type. Supporting both categories was the most profitable for both sides, followed by moderate support. No assistance was the least profitable option. The purpose was to maximize the manufacturer's profits. The study found that supporting both forms of advertising benefited both members the best. Theoretical analyses provide insight into management.

The Nash bargaining model is used to analyze profit-sharing methods and manufacturer pool rates for cooperative advertising. Chutani and Sethi [6] proposed a Stackelberg differential game model in which the manufacturer allocates advertising shares to  $N$  shops based on their subsidy value, and retailers engage in a Nash differential game to optimize their advertising efforts. In [9], authors employed a game-theoretic model to investigate cooperative advertising in a supply chain, discovering that retailers' concerns regarding Nash bargaining fairness increase local advertising spending, thereby enhancing supply chain efficiency.

Cooperative game theory has been frequently used to assess the collaboration between companies in real-world networks. Studies have shown that cooperative game theory can estimate cost savings and propose fair allocations. Frisk et al. [11] investigated cooperation among forest enterprises in

Sweden's wood sector and offered strategies to disperse savings among participants. Lozano et al. [21] discovered that shippers in a logistics network collaborate by integrating transportation requirements. Razmi, Hassani, and Hafezalkotob [23] discovered that horizontal collaboration in natural gas distribution can lead to cost savings and equitable allocation.

In logistic networks, suppliers may collaborate to make better use of their vehicles. A mathematical model was created to evaluate the advantages of collaboration among distribution businesses, and several cooperative game theory methodologies were compared and examined. Wang et al. [30] investigated a multiple-center truck routing issue in which logistics service providers collaborated in a network, providing cooperative game theory strategies for profit allocation.

Cooperation can also increase flow and reduce costs in maximum flow problems. Reyes [24] showed that logistic companies can increase flow through cooperation, and Hafezalkotob and Makui [15] suggested a mathematical programming approach based on cooperative game theory for player collaboration in logistic networks. It covers coalitional games, cooperative game theory, and its practical applications in communication and wireless networks.

It classifies them into canonical, formation, and graph games, providing a comprehensive treatment for communications and network engineers. Gharehbolagh et al. [12] developed mathematical models to improve reliability and cost in logistics networks. Zhao et al. [36] examined how to handle a large number of community energy consumers (CEPs) in a distribution network. The researchers developed a hybrid game-based optimal operating model that incorporates Stackelberg and cooperative games. It also includes cloud energy storage for economic efficiency.

Differential games are used to handle problems involving several decision-makers making decisions continuously, managing complicated systems, making dynamic decisions. Researchers can refer to [4] for further study. Sorger [27] analyzed a modified version of non-cooperative advertising differential game of Case. Sorger derived Nash equilibria for open-loop controls as well as for feedback strategies for finite and infinite planning horizons. Dockner [7] introduced the differential games for modeling economic and management

issues involving strategic decision making and explored applications in capital accumulation, industrial organization, and more. He, Prasad, and Sethi [17] analyzed co-op advertising as a stochastic Stackelberg differential game, identifying optimal policies for both sides and comparing it to a vertically integrated channel. They also proposed a framework for coordination. Erickson [10] introduced an oligopoly model to determine advertising strategies in an oligopoly. Unit contribution and advertising effectiveness boosted own advertising and sales, while discount and decay rates had negative effects. In asymmetric oligopolies, these factors influenced rivals' advertising and sales differently.

Tur and Petrosyan [29] investigated cooperative differential games on networks, with a focus on optimality principles, characteristic function development, and Shapley value computation. They presented their findings using a differential marketing game. Liu and Wu [20] investigated the shared manufacturing model, a sustainable way to production that involves a producer and a platform with government subsidies. The study analyzed price, collaborative advertising, and decision-making for centralized, decentralized, and bilateral cost-sharing contracts. Centralized decision-making results in cheaper prices, more promotional effort, and higher profits. Despite improved decentralized decision-making, total revenues remained below centralized levels. Wu and Liu [31] examined four models based on differential games to examine pricing and advertising efficiency in shared manufacturing: Classic cooperation, cost-sharing contract, revenue-sharing contract, and bilateral cost-sharing contract. It included suggestions and numerical examples. Du et al. [8] examined the effectiveness of advertising in promoting water-saving products in a two-level supply chain. Contracts for cooperative, noncooperative, and cooperative cooperation were the three scenarios that were distinguished. The outcomes show pareto optimality for market demand and product goodwill. Han et al. [16] studied a dynamic Stackelberg game between a company and retailer to analyze their advertising decisions. The study found that while national advertising can improve advertising behaviors, it may also result in losses. The study also examined retail and brand competitiveness. Co-op advertising helped align manufacturing and retailer decisions in supply chains. The manufacturer determined the participation

rate and wholesale price, while the retailer responded with effective advertising and pricing. Petrosyan, Yeung, and Pankratova [22] introduced a new characteristic function based on the possibility of stopping interaction by players outside the coalition in each time instant or imposing sanction on players from the coalition.

We mention that our research in this article mostly is related to [27, 14, 29].

Tur and Petrosyan [29] investigated cooperative differential games on networks, building characteristic functions, computing Shapley values, and looking at optimality principles. Our article and Tur and Petrosyan [29] article differ in that our model takes into account the roles of producers and consumers. Tur's model, on the other hand, only takes into account the role of distributors in multi-level marketing, wherein distributors collaborate to increase profit. Additionally, while we have looked into the roles of customers and producers, the majority of the models that we described in the introduction take into account the relationship between the retailer and the producer. As a matter of fact, manufacturers work directly together under this strategy to boost revenue through cooperative advertising. In 2018, Hafezalkotob et al. introduced a mathematical model that calculated the profit from cooperative advertising in social networks. They also suggested cooperative game theory techniques for benefit distribution, showing that coalitions yielded higher profits. One way that our study differs from [14] is that we employ a dynamic game (differential game) as opposed to a static game that is solved using nonlinear programming. In contrast to static games, dynamic games take time into account. In contrast to the Hafezalkotob model, we determine each coalition's earnings using the characteristic function and look at two scenarios depending on advertising efficiency variables to see how they impact revenue of coalition.

### 3 Problem description and model assumptions

In this section, a cooperative differential game approach on a network is applied to represent the relationship between players. For this purpose, we assume that a graph  $G = (\mathcal{V}, E)$  is a network with  $\mathcal{V}$  nodes and  $E$  edges with

prescribed duration of  $[0, T]$ . Let  $\{i_1, \dots, i_n\}$  be the set of companies and  $\{k_1, \dots, k_m\}$  be the set of consumers. Each node corresponds to a company  $i$  or a costumer  $j$ , where  $i \in N = \{1, 2, \dots, n\}$  and  $j \in K = \{1, 2, \dots, m\}$ , respectively, and also each edge  $(i, j) \in E$  represents a connection between company  $i$  and costumer  $j$  (as illustrated in Figure 1). We introduce the

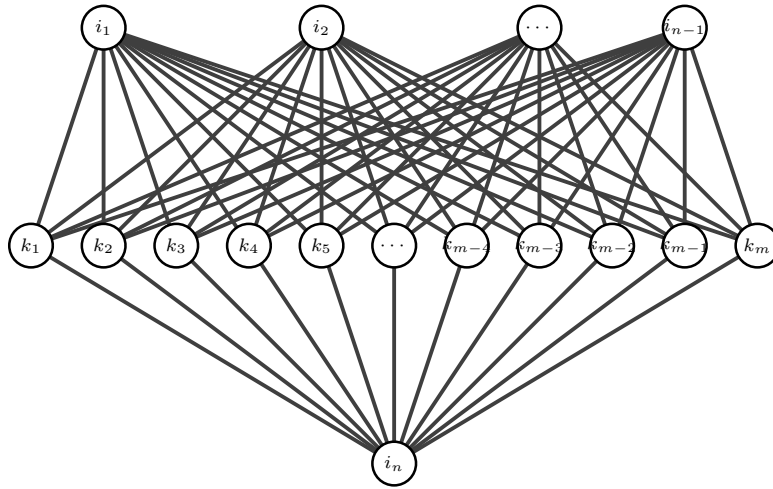


Figure 1: A graph of advertising model with the set  $\{i_1, \dots, i_n\}$  for companies and  $\{k_1, \dots, k_m\}$  for consumers in a network

assumptions of the model as follows:

Assumption 1. We assume that the Players (companies) in our network are reasonable.

Assumption 2. The advertisement of firms is affected by the sensitivity rate of consumers to companies.

Assumption 3. We assume that the utility of this cooperative game is transferable that means the earnings of a coalition can be calculated.

Assumption 4. If the companies cooperate in network, then they cover the consumers of each other.

Assumption 5. It is supposed that all links are undirected.

The dynamics of company  $i$ 's sale rate are governed by

$$\dot{x}_i(t) = v_i(t) + \sum_{k=1}^m \left( \rho_{ik} \sqrt{u_{ik}(t)} - \sum_{j \neq i} \delta_{jk} \sqrt{u_{jk}(t)} \right), \quad i \in \{1, \dots, n\}, \quad (1)$$

where  $x_i(t) \in R^n$  indicates the sales rate of Player  $i$  (state variable) at time  $t$  and  $u_{ik}(t)$  represents the advertisement effort of company  $i$  for consumer  $k$  (control variable). The amount of basis demand from firm  $i$  that is unrelated to the advertising effort is represented by  $v_i(t) \geq 0$ . The values of parameters  $\rho_{ik} \geq 0$  and  $\delta_{ik} \geq 0$  signify the sensitivity of consumer  $k$  to the advertisement of company  $i$  and the sensitivity of consumer  $k$  to the advertisement of company  $j$  (competitor's advertising), respectively.

They discovered that customer responses to advertising which had an impact on the values of  $\rho_{ik}$  and  $\delta_{ik}$ , which stand for specific parameters connected to consumers characteristics [1]. Because customers are more likely to respond favorably to advertisements from a firm they are familiar with than to those from its rivals, it is often assumed that  $\delta_{ik}$  is less than  $\rho_{ik}$ . Furthermore, the sales rate for company  $i$  increases with cooperative advertising efforts and falls in response to competing players' activities.

The objective function of company  $i \in N$  is considered as follows:

$$\max J_i(x_i^0, u_{i1}, \dots, u_{im}) = \int_{t_0}^T \left( p x_i(\tau) - \sum_{k=1}^m u_{ik}(\tau) \right) d\tau, \quad (2)$$

where  $p$  denote the price.

The goal of the problem is to maximize company's profit, which is obtained by subtracting advertising expenses from income. Let  $\Gamma(x_0, T - t_0)$  be a cooperative differential game on network  $(\mathcal{V}, E)$ , let the system dynamics (1) and the sets of feasible controls  $U_i$ ,  $i \in N$  be given, and let the players' payoffs be defined by (2). Assume that companies are able to cooperate together in order to maximum possible joint reward

$$\sum_{i \in N} \int_{t_0}^T \left( p x_i(\tau) - \sum_{k=1}^m u_{ik}(\tau) \right) d\tau, \quad (3)$$

subject to dynamics (1).

The optimal cooperative strategies of companies

$$u^*(t) = (u_{1,1}^*(t), \dots, u_{1,m}^*(t), u_{2,1}^*(t), \dots, u_{2,m}^*(t), \dots, u_{3,1}^*(t), \dots, u_{n,m}^*(t))$$

for  $t \in [t_0, T]$  are defined as follows:

$$u^*(t) = \arg \max_{u_{ik}, i \in N, k \in K} \sum_{i \in N} \int_{t_0}^T (px_i(\tau) - \sum_k u_{ik}(\tau)) d\tau. \quad (4)$$

The optimal cooperative trajectory is  $x^*(t) = (x_1^*(t), x_2^*(t), \dots, x_n^*(t))$ . This trajectory corresponds to the optimal cooperative strategies vector,  $u^*(t)$ .

We express the maximum joint reward as follows:

$$\begin{aligned} V(N, x_0, T - t_0) &= \sum_{i \in N} \int_{t_0}^T (px_i^*(\tau) - \sum_k u_{ik}^*(\tau)) d\tau \\ &= \max_{u_{ik}, i \in N, k \in K} \sum_{i \in N} \int_{t_0}^T (px_i(\tau) - \sum_k u_{ik}(\tau)) d\tau, \end{aligned} \quad (5)$$

subject to dynamics

$$\dot{x}_i^*(t) = v_i(t) + \sum_{k=1}^m \left( \rho_{ik} \sqrt{u_{ik}^*(t)} - \sum_{j \neq i} \delta_{jk} \sqrt{u_{jk}^*(t)} \right), \quad i \in \{1, \dots, n\}. \quad (6)$$

It is important to determine the characteristic function of the problem in order to decide how to distribute the greatest total payment of coalition among the participants under an agreeable scheme. Based on the characteristic function of a cooperative game, two disjoint coalitions of companies are at least as excellent when the companies cooperate as when they work apart, and also the empty set has no value. This concept can be stated mathematically as

1.  $V(\emptyset) = 0$ .
2. If  $S, T \subseteq M$  are disjoint coalitions ( $S \cap T = \emptyset$ ), then  $V(S) + V(T) \leq V(S \cup T)$ .

Let  $S$  be a subset of  $\mathcal{V}$ . A pair  $(\mathcal{V}_S, E_S)$  is called a subnet (subgraph) if it only has subset  $S$  of the set of vertices (companies) of network  $(\mathcal{V}_S, E_S)$ , and  $E_S$  has all connections from  $E$  whose initial and final vertices in network graph are inside subset  $S$ . We assume that the value of a coalition  $S$  is determined along the cooperative trajectory as follows:

$$V(S, x_0, T - t_0) = \sum_{i \in S} \left( \int_{t_0}^T \left( p x_i^*(\tau) - \sum_{k=1}^m u_{ik}^*(\tau) \right) d\tau \right), \quad (7)$$

where  $x_i(t)$  and  $u_{ik}(t)$  are the solutions obtained from (4) and (6), respectively.

Similarly, the cooperative-trajectory characteristic function of the subgame  $\Gamma(x^*(t), T - t)$  starting at time  $t \in [t_0, T]$  can be evaluated as

$$V(S, x^*(t), T - t) = \sum_{i \in S} \left( \int_t^T \left( p x_i^*(\tau) - \sum_{k=1}^m u_{ik}^*(\tau) \right) d\tau \right). \quad (8)$$

Suppose that the game is played in a cooperative scenario and that players have the opportunity to cooperate in order to achieve maximum total payoff:

$$\max_{u_{ik}, i \in N, k \in K} \sum_{i \in N} \int_t^T \left( p x_i(\tau) - \sum_{k=1}^m u_{ik}(\tau) \right) d\tau. \quad (9)$$

We apply the Bellman dynamic programming technique to define the cooperative strategies. The Bellman function in a subgame beginning at moment  $t$  from state  $x(t)$  is denoted by  $V(N, x, T - t)$ :

$$V(N, x, T - t) = \max_{u_{ik}, i \in N, k \in K} \sum_{i \in N} \int_t^T \left( p x_i(\tau) - \sum_{k=1}^m u_{ik}(\tau) \right) d\tau. \quad (10)$$

The Hamilton–Jacobi–Bellman (HJB) equation has the following expression [32]:

$$\begin{aligned} -V_t(N, x, T - t) = & \max_{u_{ik}, i \in N, k \in K} \left\{ \sum_{i \in N} \left( p x_i(\tau) - \sum_{k=1}^m u_{ik}(\tau) \right) + \sum_{i \in N} (v_i(t) \right. \\ & \left. + \sum_{k=1}^n (\rho_{ik} \sqrt{u_{ik}} - \sum_{j \neq i} \delta \sqrt{u_{ik}}) V_{x_i} \right\}, \\ V(T, x(T)) = & 0. \end{aligned} \quad (11)$$

Performing the maximization operator in (11) results:

$$\begin{aligned} u_{ik}^* = & \frac{1}{4} \left( \rho_{ik} V_{x_i} - \sum_{j \neq i} \delta_{ik} V_{x_j} \right)^2, \\ i \in N = & \{1, \dots, n\}, \quad k \in M = \{1, 2, \dots, m\}, \end{aligned} \quad (12)$$

Substituting  $u_{ik}^*$  from (12) into (11) and solving HJB (11), one obtains

$$V(N, x, T - t) = \sum_{i \in N} \left( A_i(t)x_i + B_i(t) \right), \quad (13)$$

where  $A_i(t)$  and  $B_i(t)$  satisfy the following system of ordinary differential equations:

$$\begin{aligned} \dot{A}_i(t) &= -p, \quad A_i(T) = 0, \\ \dot{B}_i(t) &= -\sum_{k=1}^m \frac{A_i(t)^2(4\delta_{i,k}\rho_{ik} - 4\delta_{i,k}^2 - \rho_{ik}^2)}{4} - \sum_{k=1}^m \frac{A_i^2(t)(2\delta_{i,k} - \rho_{ik})\sqrt{(2\delta_{i,k} - \rho_{ik})^2}}{2} \\ &\quad - A_i(t)v_i(t), \\ B_i(T) &= 0. \end{aligned} \quad (14)$$

The solution of ordinary differential equation (14) is obtained as

$$\begin{aligned} A_i(t) &= (T - t)p, \\ B_i(t) &= \frac{p(T - t)^2(\sum_{k=1}^m (2\delta_{ik} - \rho_{ik})^2 p(T - t) + 6v_i(t))}{12}. \end{aligned} \quad (15)$$

The cooperative optimal strategies can be obtained by

$$u_{ik}^* = \frac{1}{4} \left( \rho_{ik} A_i - \sum_{j \neq i} \delta_{ik} A_j \right)^2, \quad i \in \{1, 2, \dots, m\}, \quad k \in \{1, \dots, m\}. \quad (16)$$

Finally, by substituting (16) in (1) and solving the differential equation, we are able to determine the players' sales rate

$$x_i^* = \left( \frac{-pt^2}{4} + \frac{Ttp}{2} \right) \left( \sum_{k=1}^m \rho_{ik}(\rho_{ik} - 2\delta_{ik}) + \sum_{i \neq j} \sum_{k=1}^m 2\delta_{jk}^2 - \delta_{jk}\rho_{jk} \right) + tv_i(t) + x_i^0.$$

Now, the value function for the cooperative joint payout of all  $n$  companies is obtained as

$$V(N, x_0, T - t_0) = \frac{1}{12} \left[ \left\{ \left( \sum_{i=1}^n \sum_{k=1}^m (2\delta_{ik} - \rho_{ik})^2 \right) (T^2 + 4t_0T - 2t_0^2)p \right. \right. \quad (17)$$

$$\left. + 6(T + t_0) \left( \sum_{i=1}^n v_i(t) \right) + 12 \left( \sum_{i=1}^n x_i^0 \right) \right\} p(T - t_0) \right]. \quad (18)$$

Using (7) and substituting  $u_{ik}^*(t)$  and  $x_i^*(t)$  from (16) and (17), respectively, we obtain  $V(S, x_0, T - t_0)$  as follows:

$$V(S, x_0, T - t_0) = \frac{1}{12} \left[ \left\{ \left( \sum_{i \in S} \sum_{k=1}^m (2\delta_{ik} - \rho_{ik})^2 \right) (T^2 + 4t_0T - 2t_0^2) p \right. \right. \\ \left. \left. + 6(T + t_0) \left( \sum_{i \in S} v_i(t) \right) + 12 \left( \sum_{i \in S} x_0^i \right) \right\} p(T - t_0) \right]. \quad (19)$$

## 4 Imputation solution

Exploring coalitions and fair pay distribution in cooperative games is crucial. The characteristic function framework effectively shows coalition possibilities and acceptable payoff distribution schemes among players.

**Definition 1.** ([32]) A vector  $\xi(x_0, T - t_0) = (\xi_1(x_0, T - t_0), \xi_2(x_0, T - t_0), \dots, \xi_n(x_0, T - t_0))$  is solution imputation if it satisfies the following conditions:

- (1)  $\xi_i(x_0, T - t_0) \geq V(\{i\}, x_0, T - t_0), \quad i \in N,$
- (2)  $\sum_{j \in N} \xi_j(x_0, T - t_0) = V(N, x_0, T - t_0).$

Condition (1) states that each element of the imputation vector must be at least as large as the value of the game for each individual player (i.e., individual rational), and condition (2) states that the total of all the elements of the imputation vector must equal the value of the game for the entire coalition (i.e., group rational). In the game  $\Gamma(x_0, T - t_0)$ ,  $L(x_0, T - t_0)$ , the set of all imputations is provided by

$$L(x_0, T - t_0) = \left\{ \xi(x_0, T - t_0) = (\xi_1(x_0, T - t_0), \dots, \xi_n(x_0, T - t_0)) \mid \right. \\ \sum_{i \in N} \xi_i(x_0, T - t_0) = V(N, x_0, T - t_0), \\ \left. \xi_i(x_0, T - t_0) \geq V(i, x_0, T - t_0), \quad i \in N \right\}.$$

The problem in computing optimal profitability for all firm coalitions is identifying how to calculate the contributions of the collaborating companies. It will not be easy because each corporation's coalitional participation in network advertising is uncertain. A theoretical approach such as cooperative game theory provides several ways to efficiently evaluate players' cooperative efforts. A few of them are presented in this section.

The core  $C(x_0, T - t_0)$  of the game  $\Gamma(x_0, T - t_0)$  is the subset of the imputation set  $L(x_0, T - t_0)$ , such that [29]

$$C(x_0, T - t_0) = \left\{ \xi(x_0, T - t_0) = (\xi_1(x_0, T - t_0), \dots, \xi_n(x_0, T - t_0)) \mid \right. \\ \sum_{i \in N} \xi_i(x_0, T - t_0) = V(N, x_0, T - t_0), \\ \left. \sum_{i \in S} \xi_i(x_0, T - t_0) \geq V(S, x_0, T - t_0), \ S \subseteq N \right\}. \quad (20)$$

A new approach to profit allocation in collaboration was developed by [26], it concentrated on the additive, symmetric, efficient, and dummy qualities of participants' contributions. The Shapley value is derived based on the varied shares of players in the coalitions, and it is unique when the core is a set of solution imputation, which is not always unique. If the vector  $\Phi(x_0, T - t_0) = \{\Phi_i(x_0, T - t_0), i = 1, 2, \dots, n\}$  satisfies the following criteria, and the vector is referred to as the Shapley value [32], then

$$\Phi_i(x_0, T - t_0) = \sum_{S \subseteq N, i \in S} \frac{(n-s)!(s-1)!}{n!} [V(S, x_0, T - t_0) - V(S - \{i\}, x_0, T - t_0)], \quad i = 1, 2, \dots, n. \quad (21)$$

In this model, according to the information in the third section and (21), the Shapley value is obtained as follows:

$$\Phi_i(x_0, T - t_0) = \frac{p(T-t)}{12} \left( \left\{ (2t^2 - 4Tt) \left( \sum_{k=1}^m \delta_{ik} (3\rho_{ik} - 2\delta_{ik}) - \rho_{ik}^2 \right) \right. \right. \\ \left. \left. + \sum_{i \neq j} \sum_{k=1}^m \frac{\delta_{jk}(\rho_{jk} - 2\delta_{jk})}{2} \right) \right) + T^2 \left( \sum_{k=1}^m \rho_{ik}^2 - 4\delta_{ik}^2 \right. \\ \left. - \sum_{i \neq j} \sum_{k=1}^m 2\delta_{jk}(\rho_{jk} - 2\delta_{jk}) \right) \Big\} p + 6(T+t)v_i(t) + 12x_0^i \Big\} \quad (22)$$

Another cooperative optimality principle is  $\tau$ -value, which is defined as follows [28]:

A vector  $\tau(x_0, T - t_0) = (\tau_1(x_0, T - t_0), \tau_2(x_0, T - t_0), \dots, \tau_n(x_0, T - t_0))$  is called  $\tau$ -value which applies to the following equation:

$$\tau_i(x_0, T - t_0) = \alpha M_i(x_0, T - t_0) + (1 - \alpha) m_i(x_0, T - t_0),$$

where

$$M_i(x_0, T - t_0) = V(N, x_0, T - t_0) - V(N - \{i\}, x_0, T - t_0), \quad (23)$$

$$m_i(x_0, T - t_0) = \max_{i \in S} \{V(S, x_0, T - t_0) - \sum_{i' \in S - \{i\}} M_{i'}(x_0, T - t_0)\}, \quad (24)$$

and the coefficient  $\alpha \in [0, 1]$  is determined from the equation

$$\sum_{i \in N} (\alpha M_i(x_0, T - t_0) + (1 - \alpha) m_i(x_0, T - t_0)) = V(N, x_0, T - t_0).$$

## 5 Numerical experiments

Although firms are typically seen as independent agents, they also have the ability to form coalitions and dynamically manage their advertising behavior. In general, the aim is to investigate the effect of cooperation in the proposed generalized model on each of the following cases:

1. The characteristic functions (payoffs of coalitions)
2. Advertising and sale rates of companies
3. Sustainability of cooperation
4. Distribution of profit in case of forming a grand coalition

In this section, we consider an example with  $n = 3$  companies and  $k = 12$  consumers, as shown in Figure 2. Let  $\mathcal{S} = \{S_1 = \{1\}, S_2 = \{2\}, S_3 = \{3\}, S_4 = \{1, 2\}, S_5 = \{1, 3\}, S_6 = \{2, 3\}, S_7 = \{1, 2, 3\}\}$  be a set of all possible sub-coalitions between companies. We use several procedures of cooperative games such as core (20), Shapley value (22), and  $\tau$ -value (23) to calculate additional profit of cooperative advertisement in the network (Figure 2).

For numerical simulation, we fix the parameters in the model as shown in Tables 1 and 2 for scenarios 1 and 2, respectively. A sensitivity analysis is provided to check the impact of changing key parameter values  $\rho_{ik} \geq 0$  and  $\delta_{ik} \geq 0$  on sustainability of cooperation between companies.

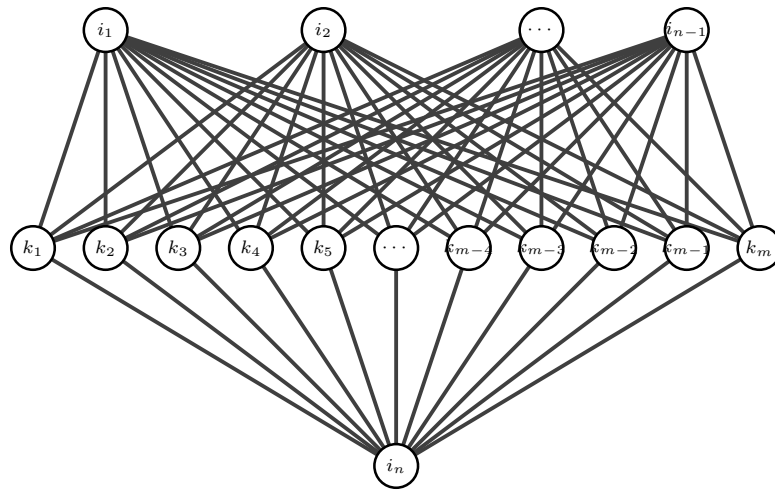


Figure 2: example of cooperative network with 3 advertising company

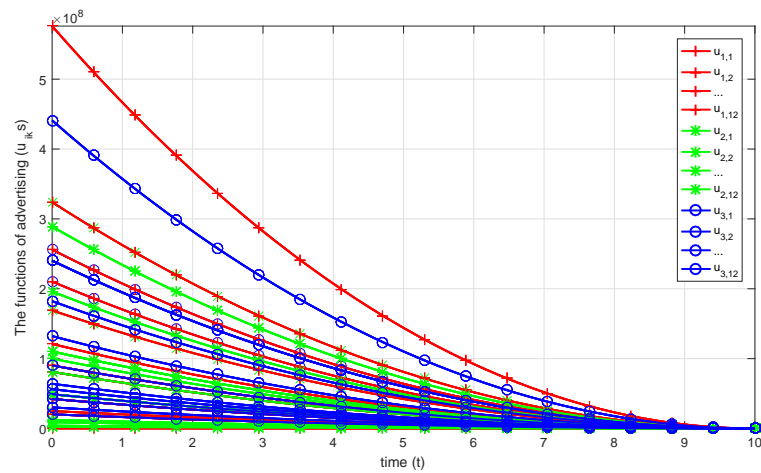


Figure 3: Promotional functions of each player over time

## 5.1 Scenario 1

In this scenario, we solve the problem using the data from Table 1 to generate Figures 3–8 and determine the quantity of coalition revenue and allocation

Table 1: Data of numerical example in scenario 1.  
 Table 2: Data of numerical example in scenario 2.

Players	$k$	$\rho$	$\delta$	$v$	$p$
Company 1	1	5	2	5	100
	2	15	3		
	3	30	4		
	4	50	1		
	5	40	2		
	6	20	5		
	7	20	1		
	8	30	2		
	9	40	4		
	10	25	3		
	11	35	3		
	12	15	1		
Company 2	1	5	1	5	100
	2	10	2		
	3	20	3		
	4	30	1		
	5	25	2		
	6	15	4		
	7	10	3		
	8	20	1		
	9	40	3		
	10	25	2		
	11	30	5		
	12	20	3		
Company 3	1	20	3	5	100
	2	15	2		
	3	25	1		
	4	50	4		
	5	35	2		
	6	15	1		
	7	20	2		
	8	25	3		
	9	35	4		
	10	25	5		
	11	35	2		
	12	15	3		

Players	$k$	$\rho$	$\delta$	$v$	$p$
Company 1	1	1	0.5	5	100
	2	5	3		
	3	7	0.5		
	4	2	1.1		
	5	1	0.5		
	6	6	3		
	7	2	1		
	8	1	0.4		
	9	1	0.9		
	10	1.5	1		
	11	6	5.9		
	12	5	3		
Company 2	1	5	0.25	5	100
	2	4	0.75		
	3	6	0.8		
	4	2	1		
	5	1	0.5		
	6	5	0.25		
	7	1	0.1		
	8	10	1		
	9	3	0.1		
	10	1	0.4		
	11	2.1	1.8		
	12	4	0.2		
Company 3	1	4	0.9	5	100
	2	6	0.75		
	3	2	0.75		
	4	3	0.3		
	5	2	1		
	6	4	2		
	7	6	2		
	8	2	1		
	9	6	0.9		
	10	2	1		
	11	2	1.5		
	12	5	3		

to participants. In Figure 3, we show the advertising function (advertising efforts) over time, for all customers, in red, green, and blue for company 1, 2, and 3, respectively. For every feasible combination, we solve the mathematical model (7)–(10). The observed differences in advertising expenditures can be attributed to varying marketing strategies, differences in budget allocations, or distinct target market characteristics across firms. This figure illustrates the advertising expenditure trends for each company. For example, Company 1 initially invests heavily in advertising but experiences a subsequent decline, whereas Company 2 maintains a more stable pattern.

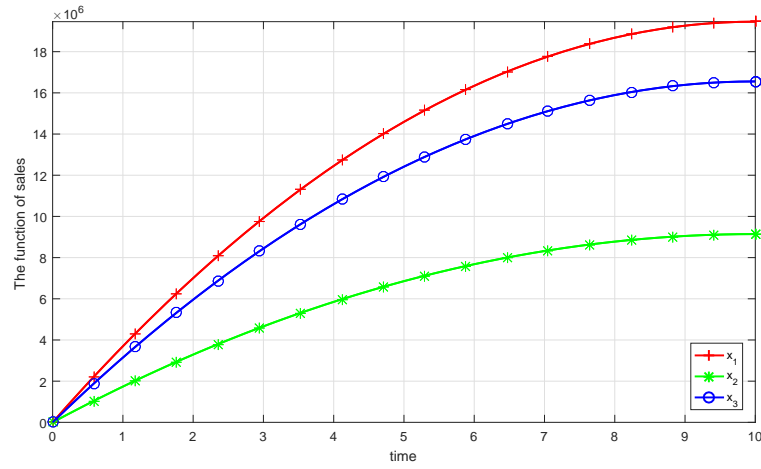


Figure 4: The functions of sales for each player over time

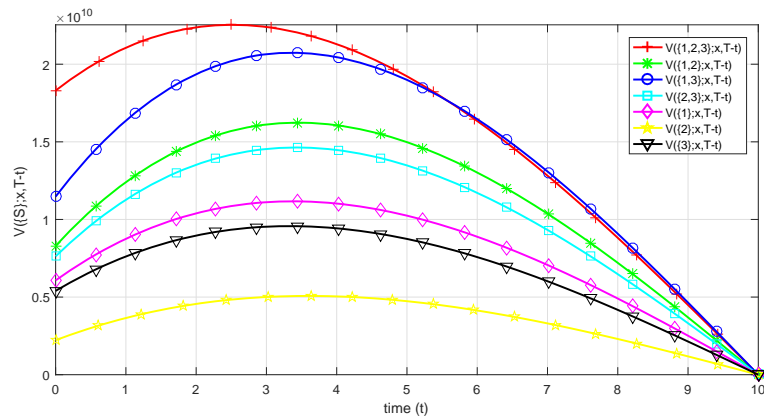


Figure 5: Earnings of all coalitions over time

These differences likely reflect divergent advertising strategies, with firms adopting distinct approaches for customer retention versus new customer acquisition. Table 3 contains a list of all conceivable coalitions' results. Table 3 shows that the larger the coalition, the higher its income will be. For Player 1, joining coalition  $S_5$  is preferable to joining coalition  $S_4$  because  $V(S_5, x_0, T - t_0) \geq V(S_4, x_0, T - t_0)$ , similarly, Player 2 and Player 3 are benefited by joining coalitions  $S_4$  and  $S_5$ , respectively. In addition, this ta-

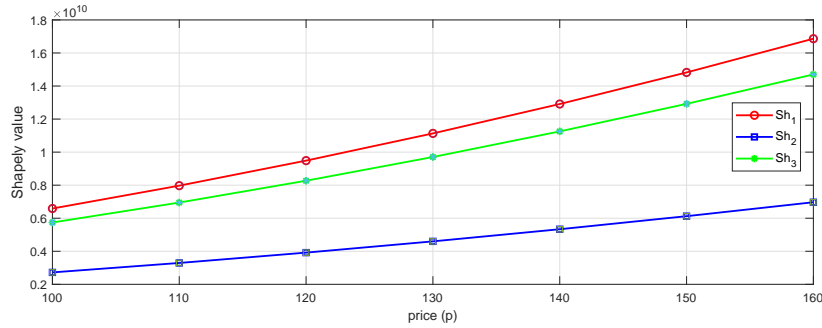
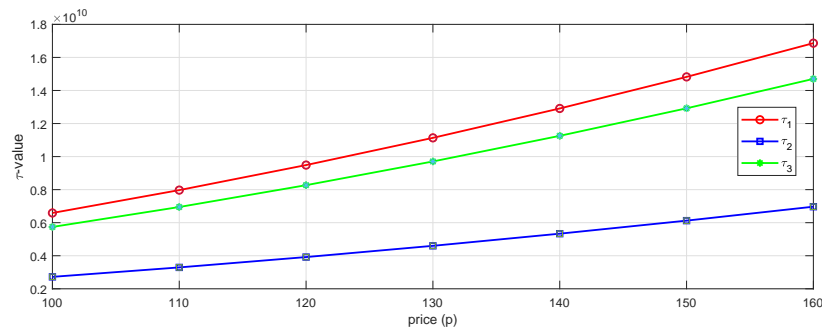


Figure 6: Sensitivity of Shapley value

Figure 7: Sensitivity of  $\tau$ -value

ble shows that the game has an additive property for for all  $S_1, S_2$  we have  $V(\{S_1, S_2\}, x_0, T - t_0) = V(\{S_1\}, x_0, T - t_0) + V(\{S_2\}, x_0, T - t_0)$ .

For fair distribution, several cooperative game theory techniques have been developed. Researchers may refer to [3] and [5] for further details. The imputations derived from different cooperative game methods are shown in Table 4, which includes the Shapley value and the  $\tau$ -value. From Table 4, it can be seen that the Shapley value and the  $\tau$ -value lead to nearly identical allocations. Furthermore, because the game has an additive property, the Shapley value and  $\tau$ -value represent the amount of cash earned by each player individually. Based on the advertising efforts of each company, the sales rate of each player is obtained, as shown in Figure 4. This figure reveals that Company 1 has achieved significantly higher sales volumes, indicating that its advertising campaigns have generated more enduring customer ac-

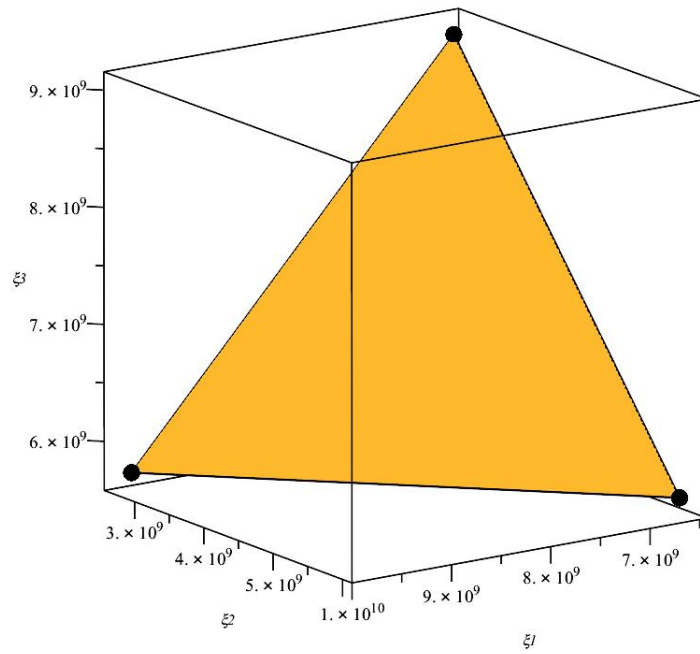


Figure 8: Core for companies in the numerical example.

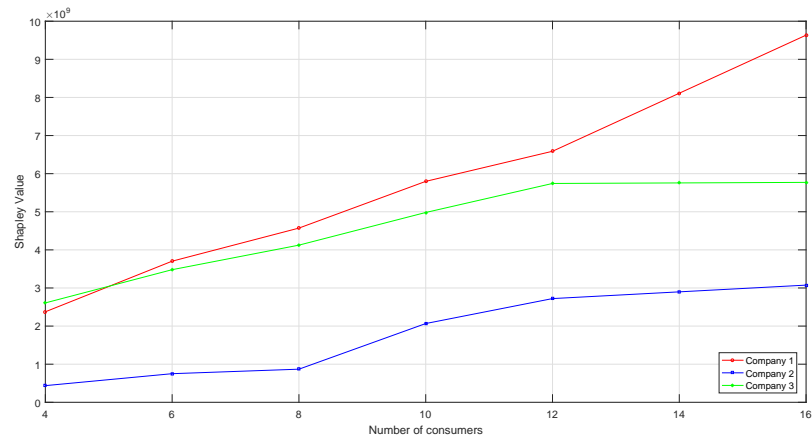


Figure 9: Sensitivity analysis of the Shapley value of each company with respect to the number of consumers (m)

quisition effects. This suggests that Company 1 not only attracts customers temporarily but also fosters long-term customer loyalty. Collectively, Com-

pany 1's extensive advertising efforts, specialized sales strategies, and product quality likely constitute the primary success factors.

Company 3 demonstrates the second highest sales performance after Company 1. While this company initially exhibited favorable growth trends, its expansion rate remained slower than Company 1's. This performance gap may stem from either less effective advertising strategies or reduced effectiveness in new customer acquisition compared to Company 1. Nevertheless, Company 3 has gradually established more stable growth patterns and has frequently outperformed Company 2 across multiple time periods.

Conversely, Company 2 has consistently recorded the lowest sales figures among the three competitors. This under performance could be attributed to deficiencies in advertising strategies, product quality issues, or inadequate new customer acquisition capabilities. It appears that Company 2 has failed to establish a competitive position comparable to Companies 1 and 3, with its initial advertising impacts showing diminishing returns over time. Figure 5 shows the earnings of all possible coalitions over time that illustrates the profit generated by each coalition. It is evident that the grand coalition—comprising all companies—achieves the highest total profit. This outcome reflects the principle of superadditivity, where the collective profit from cooperation exceeds the sum of individual profits gained through independent operations. From a strategic standpoint, the result underscores the potential advantages of forming alliances and coalitions in marketing and advertising contexts. Figures 6 and 7, respectively, illustrate the effect of a company's price sensitivity on Shapley values and  $\tau$ -value obtained from a grand coalition. In these figures, the distribution of profits among coalition members is shown based on the Shapley value and  $\tau$ -value. These methods help to understand the contribution of each company to the overall profit. According to this approach, companies that play a larger role in the success of the coalition will receive a greater share of the profit. For instance, Company 1, due to its marketing efforts and provision of higher-quality products, receives a larger share of the profit. The results show that the higher the price, the higher the amount of revenue allocated to each player. Figure 9 illustrates how the Shapley value allocated to each company in the grand coalition changes as the number of consumers increases. As expected, increasing the number

of consumers from 4 to 16 raises the value assigned to each company, since a larger customer base leads to higher overall revenue, which is then distributed among the members.

For most values of  $m$ , Company 1 receives the highest share. This is likely due to stronger advertising efforts, better market positioning, or greater influence within the coalition.

A notable point in the figure is that the Shapley value of Company 3 remains relatively stable at higher consumer levels (e.g.,  $m = 12, 14$ , and  $16$ ). This indicates that its additional contribution to the coalition diminishes as the market grows—possibly due to limitations such as restricted service capacity, limited advertising strategies, or reduced competitive differentiation.

Interestingly, when  $m = 4$ , Company 3 holds the highest Shapley value, suggesting it may have an advantage in smaller markets. This could result from effective early-stage marketing or closer engagement with a smaller customer base.

As the consumer count rises, however, Company 1's share grows significantly. This trend reflects the competitive advantage of companies with broader reach, higher advertising budgets, or greater scalability in larger markets. The range of core is determined by (20). The equation and inequalities (25) are drawn, and the space between them is taken into consideration as Figure 8, in order to calculate the range of the core using Table 3. Note that if the core is empty, it indicates the instability of the game, and companies may have an incentive to deviate from the coalitions.

$$\begin{aligned}
 \xi_1(x_0, T - t_0) + \xi_2(x_0, T - t_0) + \xi_3(x_0, T - t_0) &= 15052091670, \\
 \xi_1(x_0, T - t_0) &\geq 6587625000, \\
 \xi_2(x_0, T - t_0) &\geq 2721791667, \\
 \xi_3(x_0, T - t_0) &\geq 5742675000, \\
 \xi_1(x_0, T - t_0) + \xi_2(x_0, T - t_0) &\geq 9309416667, \\
 \xi_1(x_0, T - t_0) + \xi_3(x_0, T - t_0) &\geq 12330300000, \\
 \xi_2(x_0, T - t_0) + \xi_3(x_0, T - t_0) &\geq 8464466667.
 \end{aligned} \tag{25}$$

Table 3: Characteristic function for coalitional advertisement.

	$S_1 = \{1\}$	$S_2 = \{2\}$	$S_3 = \{3\}$	$S_4 = \{1, 2\}$	$S_5 = \{1, 3\}$	$S_6 = \{2, 3\}$	$S_7 = \{1, 2, 3\}$
$V(S)$	6587625000	2721791667	5742675000	9309416667	12330300000	8464466667	15052091670

Table 4: Imputation vectors of different cooperative game methods.

Company	Shapley value	$\tau$ -value
$\{1\}$	6587625000	6587625000
$\{2\}$	2721791667	2721791667
$\{3\}$	5742675000	5742675000

## 5.2 Scenario 2

According to the data in Table 2 and scenario 2, the results for the characteristic functions have been reported in Table 5. The results show that for this set of parameters, the profit of coalition  $S_3$  is negative, which means that if Player 3 plays alone, he not only does not make a profit, but also incurs a loss. In other words, Player 3 must enter the coalition game, otherwise he faces the risk of bankruptcy. However, as can be seen, the value of the  $S_6$  coalition is lower than when Player 2 plays alone. Therefore, it is not rational for Player 2 to join this coalition because the condition of individual rationality does not exist, similarly for Player 1 in coalition  $S_5$ . The  $S_4$  coalition is a reasonable two-person coalition.

The payoff of the coalition  $S_4$  is greater than the grand coalition  $S_7$ . Hence, forming a grand coalition is not logical. In such cases, the grand coalition is not stable and the goal of the companies is reduced to finding an optimal sub-coalition. The results show that parameters  $\rho_{ik}$  and  $\delta_{ik}$  are key

Table 5: Characteristic functions for scenario 2.

	$S_1 = \{1\}$	$S_2 = \{2\}$	$S_3 = \{3\}$	$S_4 = \{1, 2\}$	$S_5 = \{1, 3\}$	$S_6 = \{2, 3\}$	$S_7 = \{1, 2, 3\}$
$V(S)$	24141666	80058333	-5516666	104200000	18625000	74541667	98683333

in determining the amount of coalition profits, and changing these parameters have a significant influence on the outcomes of cooperation in the proposed model.

## 6 Conclusion

Most dynamic cooperative advertising research focuses on the interaction between retailers and manufacturers in the framework of cooperative adversarial games. In this paper, we have generalized the relationship between firms and customers in the form of a cooperative differential game in a network. The characteristic functions and benefits of network collaboration for the proposed model were studied. The core, Shapley value, and  $\tau$ -value are obtained for a numerical example. We solved the proposed model for two different sets of parameters to investigate sustainability of cooperation between companies using advertising effectiveness parameters.

In the first scenario, we found that the larger the coalition, the higher its earnings, and we also discovered that the game has the additive property. In the second scenario, it was discovered that a larger coalition does not always yield greater income, and it is possible that players will earn more if they play alone. The influence of price on the model revealed that as the price grows, so does the profit allotted to the firms. According to Table 2, the assigned earnings in the coalition were larger than those in the non-collaborative state, ensuring long-term collaboration. Several topics might be recommended for further investigation. First, the model's collaboration process is explored in the infinite horizon time mode. The second step is to assess the type of communication in the network and its influence on earnings. Third, applying a stochastic dynamic system and solving it for this model would be both interesting and challenging.

## Acknowledgements

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

## References

- [1] Alba, J.W., Broniarczyk, S.M., and Shimp, T.A. *The influence of prior beliefs, frequency cues, and magnitude cues on consumers' perceptions of comparative price data*, J. Consum. Res. 21(2) (1994), 219–235.
- [2] Aust, G., *Vertical cooperative advertising and pricing decisions in a manufacturer-retailer supply chain: a game-theoretic approach*, In Vertical cooperative advertising in supply chain management: a game-theoretic analysis, pp. 65–99. Cham: Springer International Publishing, 2014.
- [3] Barron, E.N. *Game theory: an introduction*, John Wiley & Sons, 2013.
- [4] Basar, T., and Olsder, G.J. *Dynamic noncooperative game theory (2nd ed.)*, Academic Press, London, 1995.
- [5] Branzei, R., Dimitrov, D., and Tijs, S. *Models in cooperative game theory (Vol. 556)*, Springer Science & Business Media, 2008.
- [6] Chutani, A., and Sethi, S.P. *Cooperative advertising in a dynamic retail market oligopoly*, Dyn. Game. Appl. 2 (2012), 347–375.
- [7] Dockner, E. *Differential games in economics and management science*, Cambridge University Press, 2000.
- [8] Du, P., Zhang, S., Wang, H., and Wang, Y. *Research on the optimization of cooperative advertising strategy in the promotion of water-saving products based on differential game*, Water Policy, 24(10) (2022), 1631–1657.
- [9] Du, X., Jiang, S., Tao, S., and Wang, S. *Cooperative advertising and coordination in a supply chain: the role of Nash bargaining fairness concerns*, RAIRO Oper. Res. 58(1) (2024), 1–18.
- [10] Erickson, G.M. *An oligopoly model of dynamic advertising competition*, Eur. J. Oper. Res. 197(1) (2009), 374–388.

- [11] Frisk, M., Göthe-Lundgren, M., Jörnsten, K., and Rönnqvist, M. *Cost allocation in collaborative forest transportation*, Eur. J. Oper. Res. 205(2) (2010), 448–458.
- [12] Gharehbolagh, H.H., Hafezalkotob, A., Makui, A., and Raissi, S. *A cooperative game approach to uncertain decentralized logistic systems subject to network reliability considerations*, Kybernetes, 46(8) (2017), 1452–1468.
- [13] Ghosh, S.K., Seikh, M.R., and Chakraborty, M. *Coordination and strategic decision making in a stochastic dual-channel supply chain based on customers' channel preferences*, Int. J. Syst. Assur. Eng. Manag. 2024, 1–23.
- [14] Hafezalkotob, A., Khodabakhsh, M., Saghaei, A., and Eshghipour, M. *Cooperation of advertising companies in social networks: A graph and game theory approaches*, Comput. Indust. Engin. 125 (2018), 212–220.
- [15] Hafezalkotob, A., and Makui, A. *Cooperative maximum-flow problem under uncertainty in logistic networks*, Appl. Math. Comput., 250 (2015), 593–604.
- [16] Han, J., Sethi, S.P., Siu, C.C., and Yam, S.C.P. *Co-op advertising in randomly fluctuating markets*, Produc. Oper. Manag. 32(6) (2023), 1617–1635.
- [17] He, X., Prasad, A., and Sethi, S.P. *Cooperative advertising and pricing in a dynamic stochastic supply chain: Feedback Stackelberg strategies*, In PICMET'08-2008 Portland International Conference on Management of Engineering and Technology. (2008), 1634–1649.
- [18] Jørgensen, S., Sigué, S. P., and Zaccour, G. *Dynamic cooperative advertising in a channel*, J. Retail. 76(1) (2000), 71–92.
- [19] Jørgensen, S., and Zaccour, G. *A survey of game-theoretic models of cooperative advertising*, Eur. J. Oper. Res. 237(1) (2014), 1–14.
- [20] Liu, P., and Wu, Y. *Differential Game Analysis of Shared Manufacturing Platform Pricing Considering Cooperative Advertising Under Government Subsidies*, IEEE Access, 11 (2023), 132852–132866.

- [21] Lozano, S., Moreno, P., Adenso-Díaz, B., and Algaba, E. *Cooperative game theory approach to allocating benefits of horizontal cooperation*, Eur. J. Oper. Res. 229(2) (2013), 444–452.
- [22] Petrosyan, L., Yeung, D., and Pankratova, Y. *Characteristic functions in cooperative differential games on networks*, J. Dyn. Game. 11(2) (2024), 115–130.
- [23] Razmi, J., Hassani, A., and Hafezalkotob, A. *Cost saving allocation of horizontal cooperation in restructured natural gas distribution network*, Kybernetes, 47(6) (2018), 1217–1241.
- [24] Reyes, P.M. *Logistics networks: A game theory application for solving the transshipment problem*, Appl. Math. Comput. 168(2) (2005), 1419–1431.
- [25] Sarkar, B. , Omair, M., and Kim, N. *A cooperative advertising collaboration policy in supply chain management under uncertain conditions*, Appl. Soft Comput. 88 (2020), 1568–4946.
- [26] Shapley, L.S. *A value for  $n$ -person games*, The Shapley value, 31–40, 1988.
- [27] Sorger, G. *Competitive dynamic advertising: A modification of the case game*, J. Econ. Dyn. Control. 13(1) (1989), 55–80.
- [28] Tur, A., and Petrosyan, L.A. *Cooperative optimality principles in differential games on networks*, Autom. Remote Control. 82(6) (2021), 1095–1106.
- [29] Tur, A., and Petrosyan, L. *Strong time-consistent solution for cooperative differential games with network structure*, Math. 9(7) (2021), 755.
- [30] Wang, Y., Ma, X., Li, Z., Liu, Y., Xu, M., and Wang, Y. *Profit distribution in collaborative multiple centers vehicle routing problem*, J. Clean. Prod. 144 (2017), 203–219.
- [31] Wu, Y., and Liu, P. *Pricing strategies for shared manufacturing platform considering cooperative advertising based on differential game*, PloS one, 19(7) (2024), e0303928.

- [32] Yeung, D.W., and Petrosyan, L.A. *Cooperative stochastic differential games (Vol. 42)*, Springer, New York, 2006.
- [33] Zhang, J., and Xie, J. *A game theoretical study of cooperative advertising with multiple retailers in a distribution channel*, J. Syst. Sci. Syst. Eng. 21(1) (2012), 37–55.
- [34] Zhang, L., Chen, F.W., Xia, S.M., Cao, D.M., Ye, Z., Shen, C.R., Maas, G., and Li, Y.M. *Value co-creation and appropriation of platform-based alliances in cooperative advertising*, Ind. Mark. Manag., 96 (2021), 213–225.
- [35] Zhang, T., Guo, X., Hu, J., and Wang, N. *Cooperative advertising models under different channel power structure*, Ann. Oper. Res. 291(2020), 1103–1125.
- [36] Zhao, B., Duan, P., Fen, M., Xue, Q., Hua, J., and Yang, Z. *Optimal operation of distribution networks and multiple community energy prosumers based on mixed game theory*, Energy, 278 (2023), 128025.
- [37] Zinoviev, D., and Duong, V. *A game theoretical approach to broadcast information diffusion in social networks*, arXiv preprint, arXiv, 1106.5174, 2011.



# Mathematical modeling and optimal control strategies to limit cochineal infestation on cacti plants

K. Sofiane\*,  and B. Omar 

## Abstract

This paper introduces a mathematical model, denoted as *SIRMC*, aimed at understanding the dynamics of cochineal infestation in cacti plants. The model incorporates two control strategies: biological control through *Hypersaspis trifurcata*, a natural predator of cochineal, and chemical control via insecticide spraying. The objective is to reduce the number of infected cacti while also achieving a balance between minimizing infection, maximizing recovery over time, and minimizing the costs associated with the control measures. The proposed framework effectively integrates these strategies to manage cochineal dynamics. Optimal control strategies are

---

\*Corresponding author

Received 27 November 2024; revised 21 April 2025; accepted 16 May 2025

Khassal Sofiane

Department of Mathematics, Faculty of Sciences El Jadida, Chouaib Doukkali University, El Jadida, Morocco. e-mail: [sofiane.k@ucd.ac.ma](mailto:sofiane.k@ucd.ac.ma)

Balatif Omar

Department of Mathematics, Faculty of Sciences El Jadida, Chouaib Doukkali University, El Jadida, Morocco. e-mail: [balatif.maths@gmail.com](mailto:balatif.maths@gmail.com)

## How to cite this article

Sofiane, K. and Omar, B., Mathematical modeling and optimal control strategies to limit cochineal infestation on cacti plants. *Iran. J. Numer. Anal. Optim.*, 2025; 15(3): 1145-1170. <https://doi.org/10.22067/ijnao.2025.90998.1557>

derived using Pontryagin's maximum principle, and numerical simulations conducted in MATLAB validate the theoretical results.

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

**Keywords:** Mathematical modeling; Optimal control theory; Pontryagin Maximum; Cochineal; Cacti plants.

## 1 Introduction

The cactus plant, often referred to as “the fruit of the poor” in some cultures, is quite popular in Morocco. It holds significant importance in both the culinary and natural heritage of many Moroccan regions. The fruit of the cactus is commonly used in cooking and is an integral part of the traditional diet in numerous households [1]. Additionally, the cactus is valued for its medicinal properties in folk and traditional medicine, as it is believed to offer various health benefits. Many also associate the cactus plant with important nutritional and medicinal values [15]. Cacti are among the most readily available and easily cared-for plants, making them a valuable resource in areas with limited resources.

The appearance of the cochineal insect in Morocco for the first time in late 2014, specifically in the village of Saniat Bergig in the Sidi Bennur province, led to significant losses, even causing the disappearance of the plant in certain regions of Morocco [16]. The cochineal insect negatively affects cactus plants primarily by extracting their sap. Through piercing and sucking mouthparts, the insect consumes plant tissues, leading to a substantial loss of sap and hindering the transport of essential nutrients and water within the plant. This depletion of sap results in stunted growth and a reduction in size, particularly affecting younger cactus plants. Observable signs of damage include changes in leaf color and a loss of turgidity, compromising the aesthetic appearance of the cactus [3]. Additionally, the overall health of the plant suffers, diminishing its ability to withstand environmental stressors such as drought or temperature fluctuations. Effectively controlling the cochineal insect is cru-

cial for preserving the health of cactus plants, especially when infestations are widespread and threaten crop vitality.

Cultivating cactus plants presents significant economic potential for farmers due to the plant's versatile applications [17]. Cactus oil, extracted for its moisturizing properties, is a key ingredient in beauty and skincare products, creating a lucrative market [14]. The cactus plant's resilience makes it well-suited for cultivation in arid regions where other crops struggle, providing a sustainable income source for farmers in such environments. Exporting cactus products to international markets, provided quality standards are met, opens opportunities for increased production and income. Additionally, adopting organic farming practices with cactus plants can enhance their value in niche markets, leading to higher prices and better overall income for farmers. In essence, cactus plant cultivation offers farmers a multifaceted opportunity to diversify their income streams and capitalize on the plant's marketable qualities across various industries [12].

Contrary to common belief, the cochineal insect is not native to Morocco. Originally from South America, the cochineal insect has spread to various parts of the world due to factors such as global trade and human movement [5].

The cochineal insect, scientifically known as *Dactylopius coccus*, holds significant historical importance for its role in producing a natural red dye. Its use as a dye dates back to ancient civilizations, where it was highly valued for coloring textiles and fabrics, contributing to various artistic and cultural expressions [4]. The crimson hue derived from cochineal had a notable economic impact, being one of the most valuable products globally during certain historical periods. It played a crucial role in international trade and the economies of specific countries. Beyond its economic value, the cochineal insect found applications in ancient medicine, potentially serving as a source of natural dyes with medicinal properties. The red dye's influence extended to cultural and artistic realms, shaping the colors and patterns in artworks and traditional clothing. The cochineal trade not only influenced economies but also encouraged exploration, contributing to the interconnectedness of the world. Despite its historical significance, the use of cochineal in dye production has declined with technological advancements and the rise of synthetic

alternatives. Nevertheless, its historical impact is still evident in various aspects of human culture and history [5].

The cochineal insect poses a serious threat to cacti, adversely affecting crops and requiring efficient, eco-friendly control measures. Biological control using *Hyperaspis Trifurcata* offers an innovative and safe method to combat this pest without relying on harmful pesticides. As a natural predator, *Hyperaspis Trifurcata* feeds on the cochineal insect without harming agricultural plants, providing a dependable and safe solution for pest control. This method offers significant environmental benefits by reducing pesticide use, maintaining ecological balance, and mitigating the negative impacts on the environment and wildlife. Furthermore, it helps preserve plant health by limiting the spread of cochineal infestations, safeguarding agricultural crops, and minimizing yield losses. The integration of *Hyperaspis Trifurcata* aligns with sustainable agricultural practices, focusing on enhancing crop quality and naturally managing pest populations. However, successful implementation requires effective distribution, continuous monitoring, and proper deployment in agricultural settings, with ongoing impact assessments. In essence, *Hyperaspis Trifurcata* represents an eco-friendly solution for managing cochineal infestations and promoting sustainable agriculture, contributing to environmental and agricultural equilibrium.

A key contribution of this study is the introduction of a novel mathematical model, the *SIRMC* model, designed to understand the dynamics of cochineal infestations in cactus plants. Additionally, our work explores two control strategies for managing the spread of the cochineal insect. The first involves biological control through *Hyperaspis Trifurcata*, a predator that feeds on the cochineal insect without harming the host plant. The second strategy involves insecticide spraying to chemically suppress infestations. The goal is to reduce the number of infected cacti while also achieving a balance between minimizing the infected and maximizing recovery over time.

This paper is structured as follows: Section 2 introduces a deterministic model for the cochineal, outlining its fundamental characteristics. Section 3 constructs a mathematical model that integrates an optimal control strategy for cochineal propagation, presenting results related to the existence of optimal control as defined by Pontryagin's maximum principle. Section 4

discusses a suitable numerical method and presents the corresponding simulation results. Finally, Section 5 concludes with a summary of the insights gained from this study.

## 2 Formulation of the mathematical model

This study presents the nonlinear mathematical model *SIRMC*, which was developed for the purpose of analyzing the control of the cochineal insect that destroys cactus plants. The following section will explain the five sections of the model in turn.

### 2.1 Description of *SIRMC* model dynamic

Before introducing our model, it is essential to understand the life cycle and spread of the cochineal insect. These small, soft-bodied insects reproduce quickly, with females laying eggs that hatch into larvae covered in a white, waxy substance. This wax helps the larvae retain moisture and resist sun exposure. The waxy threads also allow the insects to be carried by the wind to nearby cacti, spreading infestations rapidly. Cochineal insects feed on the fluids of cacti, which can cause significant damage, often leading to the plant's death in severe cases. Farmers typically rely on cutting, burying, or burning infected plants to prevent further spread. However, the introduction of natural predators like *Hyperaspis trifurcata*, a species of lady beetle, offers a biological solution to this pest, reducing the need for more destructive methods.

Our proposed model, denoted as *SIRMC*, is designed to simulate the dynamics of this interaction, focusing on the spread of the cochineal insect and its control within cactus populations. The total cacti population is denoted by  $N(t)$ , which satisfies the equation:

$$N(t) = S(t) + I(t) + R(t).$$

The model is structured into five compartments:

**Compartment  $S$  (Susceptible cacti):** Represents healthy cacti that are vulnerable to cochineal infection.

**Compartment  $I$  (Infectious cacti):** Contains cacti that are currently infested by cochineal insects and can transmit the infestation to other plants.

**Compartment  $R$  (Recovered cacti):** Represents cacti that have recovered from the infestation, either naturally or due to external interventions.

**Compartment  $M$  (Cochineal insect):** This compartment models the population of cochineal insects responsible for spreading the infection among cacti.

**Compartment  $C$  (Hyperaspis trifurcata):** Represents the population of lady beetles that predate on cochineal insects, helping to control their spread and protect the cacti.

The *SIRMC* model we propose is illustrated by the following diagram:

The transition rules between the groups are illustrated as follows:

#### The Susceptible Cacti Population:

Susceptible cacti are introduced into the system through natural recruitment at a rate  $\Lambda$ . However, they may become infested when they come into contact with cochineal insects at a rate  $\lambda SM$ . Additionally, healthy cacti may die due to natural causes at a rate  $\mu S$ . The equation governing this transition is

$$\frac{dS}{dt} = \Lambda - \lambda SM - \mu S.$$

#### The Infected Cacti Population:

Once a healthy cactus becomes infested, it moves into the infected group. The number of infected cacti increases when healthy cacti acquire the infestation

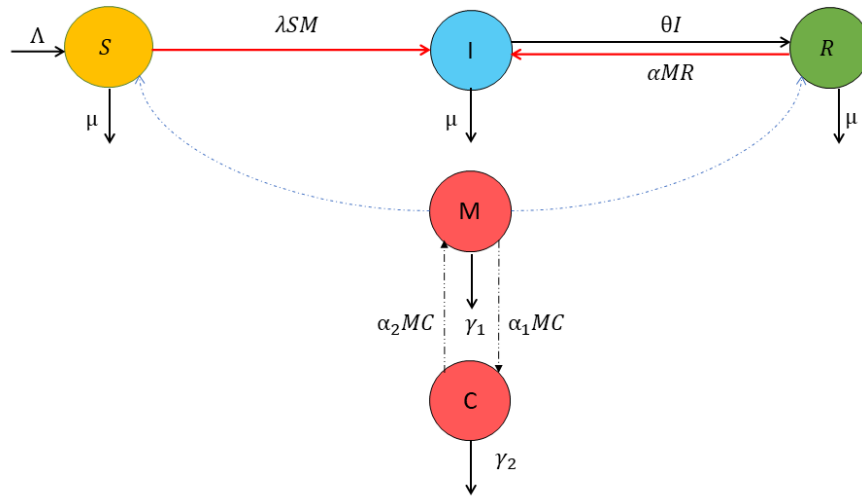


Figure 1: A diagram of the evolution of the transfer in the *SIRM C* model depicts interactions between cactus populations, cochineal insects, and the control measure *Hyperaspis trifurcata*.

( $\lambda SM$ ) or when recovered cacti are reinfected due to residual cochineal insects ( $\alpha MR$ ). Infected cacti may either recover at a rate  $\theta I$  or die naturally at a rate  $\mu I$ . This transition is expressed as

$$\frac{dI}{dt} = \lambda SM + \alpha MR - \theta I - \mu I.$$

#### The Recovered Cacti Population:

Cacti that recover from infestation enter the recovered group at a rate  $\theta I$ . However, some recovered cacti may be reinfected due to the presence of cochineal insects at a rate  $\alpha MR$ , while others die naturally at a rate  $\mu R$ . The equation describing this transition is

$$\frac{dR}{dt} = \theta I - \alpha MR - \mu R.$$

#### The Cochineal Insect Population:

The cochineal insect population follows a logistic growth pattern, increasing at a rate  $\beta_1 M(1 - M/K_1)$ , where  $K_1$  is the carrying capacity. However, their

numbers decrease due to natural mortality at a rate  $\gamma_1 M$  and predation by *Hyperaspis Trifurcata* at a rate  $\alpha_1 MC$ . The governing equation for this dynamic is

$$\frac{dM}{dt} = \beta_1 M \left(1 - \frac{M}{K_1}\right) - \gamma_1 M - \alpha_1 MC.$$

### The Predator Population (*Hyperaspis Trifurcata*):

The predator species, *Hyperaspis Trifurcata*, follows a similar logistic growth pattern, increasing at a rate  $\beta_2 C(1 - C/K_2)$ , where  $K_2$  is its carrying capacity. The predator also benefits from consuming cochineal insects, leading to additional reproduction at a rate  $\alpha_2 MC$ . However, its population declines due to natural mortality at a rate  $\gamma_2 C$ . This dynamic is represented by

$$\frac{dC}{dt} = \beta_2 C \left(1 - \frac{C}{K_2}\right) - \gamma_2 C + \alpha_2 MC.$$

The following system consists of nonlinear ordinary differential equations:

$$\begin{cases} \frac{dS}{dt} = \Lambda - \lambda SM - \mu S, \\ \frac{dI}{dt} = \lambda SM + \alpha MR - \theta I - \mu I, \\ \frac{dR}{dt} = \theta I - \alpha MR - \mu R, \\ \frac{dM}{dt} = \beta_1 M \left(1 - \frac{M}{K_1}\right) - \gamma_1 M - \alpha_1 M(t)C(t), \\ \frac{dC}{dt} = \beta_2 C \left(1 - \frac{C}{K_2}\right) - \gamma_2 C + \alpha_2 M(t)C(t), \end{cases} \quad (1)$$

with the initial conditions:  $S(0) \geq 0$ ,  $I(0) \geq 0$ ,  $R(0) \geq 0$ ,  $M(0) \geq 0$ , and  $C(0) \geq 0$ .

The parameters of model are defined in Table 1.

Table 1: Model parameters

Parameter	Description
$\Lambda$	The recruitment rate for the cacti plants.
$\beta_1$	The growth rate for the cochineal insect.
$\beta_2$	The growth rate for the <i>Hyperaspis Trifurcata</i> .
$\alpha$	The reinfection rate of cacti plants after recovery from the cochineal insect.
$\alpha_1$	The rate of <i>Hyperaspis trifurcata</i> encounters the cochineal insect and preys on it.
$\alpha_2$	The reproduction rate of <i>Hyperaspis trifurcata</i> due to feeding on the cochineal insect.
$\theta$	The recovery rate for cacti plants.
$\gamma_1$	The natural death rate for the cochineal insect.
$\gamma_2$	The natural death rate for <i>Hyperaspis Trifurcata</i> .
$\lambda$	$\lambda$ is the rate of cochineal insect encounters with cacti.
$\mu$	The cochineal insect mortality rate

## 2.2 Positivity of solutions

**Theorem 1.** If  $S(0) \geq 0$ ,  $I(0) \geq 0$ ,  $R(0) \geq 0$ ,  $M(0) \geq 0$  and  $C(0) \geq 0$ , then the solutions  $S(t)$ ,  $I(t)$ ,  $R(t)$ ,  $M(t)$ , and  $C(t)$  of system (1) are positive for all  $t \geq 0$ .

*Proof.* It follows from the first equation of system (1) that

$$\frac{dS}{dt} = \Lambda - \lambda SM - \mu S \geq -(\lambda M + \mu) S. \quad (2)$$

Then, we have

$$\frac{dS}{S} \geq -(\lambda M + \mu) dt. \quad (3)$$

By integrating (2) from 0 to  $t$ , we obtain

$$S(t) \geq S_b(0)e^{-\int_0^t (\lambda M + \mu) ds}.$$

That implies

$$S(t) \geq 0 \text{ for all } t \geq 0.$$

Similarly, we prove that  $I(t) \geq 0$ ,  $R(t) \geq 0$ ,  $M(t) \geq 0$  and  $C(t) \geq 0$  for all  $t \geq 0$ . □

## 2.3 Boundedness of the solutions.

**Theorem 2.** Let  $T_1 = \max\{M(0), K_1\}$ ,  $T_2 = \max\{C(0), K_2\}$ . Then the set  $\Gamma = \left\{ (S, I) \in \mathbb{R}_+^2 : N(t) \leq \frac{\Lambda}{\mu} \right\} \times \{M \in \mathbb{R}_+ : M(t) \leq T_1\} \times \{C \in \mathbb{R}_+ : C(t) \leq T_2\}$  is positively invariant under system (1) with nonnegative initial conditions  $S(0)$ ,  $I(0)$ ,  $R(0)$ ,  $M(0)$ , and  $C(0)$ .

*Proof.* From the initial equations of (1), we derive

$$\frac{dN(t)}{dt} = \Lambda - \mu N(t). \quad (4)$$

Then,

$$N(t) \leq N(0)e^{-\mu t} + \frac{\Lambda}{\mu} [1 - e^{-\mu t}]. \quad (5)$$

If we consider the limit  $t \rightarrow \infty$ , then  $0 \leq N(t) \leq \frac{\Lambda}{\mu}$ .

From the last equation of (1), we have

$$\frac{dM}{dt} \leq \beta M \left( 1 - \frac{M}{K_1} \right). \quad (6)$$

Hence, employing a typical comparison approach yields  $\limsup_{t \rightarrow \infty} M(t) \leq K_1$ .

Similarly, we prove that  $\limsup_{t \rightarrow \infty} C(t) \leq K_2$ .

Finally, the set  $\Gamma$  is positivity invariant for the system (1).  $\square$

## 2.4 Existence of solutions

**Theorem 3.** The system (1) that satisfies a given initial condition  $(S(0), I(0), R(0), M(0), C(0))$  has a unique solution.

*Proof.* The model (1) can be expressed in matrix form as follows:

Let  $X(t) = (S, I, R, M, C)^T$  and  $F(X(t)) = \left( \frac{dS}{dt}, \frac{dI}{dt}, \frac{dR}{dt}, \frac{dM}{dt}, \frac{dC}{dt} \right)^T$ .

The model (1) can be rephrased as

$$F(X(t)) = AX + B(X(t)),$$

where

$$A = \begin{pmatrix} -\mu & 0 & 0 & 0 & 0 \\ 0 & -(\mu + \theta) & 0 & 0 & 0 \\ 0 & \theta & -\mu & 0 & 0 \\ 0 & 0 & 0 & \beta_1 - \gamma_1 & 0 \\ 0 & 0 & 0 & 0 & \beta_2 - \gamma_2 \end{pmatrix}$$

and

$$B(X(t)) = \begin{pmatrix} \Lambda - \lambda SM \\ \lambda SM + \alpha MR \\ -\alpha MR \\ -\frac{\beta_1 M^2}{K_1} - \alpha_1 MC \\ -\frac{\beta_2 C^2}{K_2} + \alpha_2 MC \end{pmatrix}.$$

Let  $X_1$  and  $X_2$  be solutions of model (1). Then

$$\begin{aligned}
|B(X_1) - B(X_2)| &\leq 2 |\lambda(S_2 M_2 - S_1 M_1)| + 2 |\alpha(M_1 R_1 - M_2 R_2)| \\
&\quad + |\alpha_1(M_2 C_2 - M_1 C_1)| + |\alpha_2(M_1 C_1 - M_2 C_2)| \\
&\quad + \left| \frac{\beta_1}{K_1}(M_2^2 - M_1^2) \right| + \left| \frac{\beta_2}{K_2}(C_2^2 - C_1^2) \right| \\
&\leq 2 |\lambda(S_2 M_2 - S_2 M_1)| + 2 |\lambda(S_2 M_1 - S_1 M_1)| \\
&\quad + 2 |\alpha(M_1 R_1 - M_1 R_2)| + 2 |\alpha(M_1 R_2 - M_2 R_2)| \\
&\quad + |\alpha_1(M_2 C_2 - M_2 C_1)| + |\alpha_1(M_2 C_1 - M_1 C_1)| \\
&\quad + |\alpha_2(M_1 C_1 - M_1 C_2)| + |\alpha_2(M_1 C_2 - M_2 C_2)| \\
&\quad + \left| \frac{\beta_1}{K_1}(M_2^2 - M_1^2) \right| + \left| \frac{\beta_2}{K_2}(C_2^2 - C_1^2) \right| \\
&\leq 2\lambda S_2 |M_2 - M_1| + 2\lambda M_1 |S_2 - S_1| \\
&\quad + 2\alpha M_1 |R_1 - R_2| + 2\alpha R_2 |M_1 - M_2| + \alpha_1 M_2 |C_2 - C_1| \\
&\quad + \alpha_1 C_1 |M_2 - M_1| + \alpha_2 M_1 |C_1 - C_2| + \alpha_2 C_2 |M_1 - M_2| \\
&\quad + \frac{\beta_1}{K_1} |M_2 - M_1| |M_2 + M_1| + \frac{\beta_2}{K_2} |C_2 - C_1| |C_2 + C_1| \\
&\leq \frac{2\lambda\Lambda}{\mu} |M_2 - M_1| + 2\lambda T_1 |S_2 - S_1| + 2\alpha T_1 |R_1 - R_2| \\
&\quad + \frac{2\alpha\Lambda}{\mu} |M_1 - M_2| + \alpha_1 T_1 |C_2 - C_1| \\
&\quad + \alpha_1 T_2 |M_2 - M_1| + \alpha_2 T_1 |C_1 - C_2| \\
&\quad + \alpha_2 T_2 |M_1 - M_2| + \frac{2\beta_1 T_1}{K_1} |M_2 - M_1| + \frac{2\beta_2 T_2}{K_2} |C_2 - C_1| \\
&\leq \left( \frac{2\lambda\Lambda}{\mu} + \frac{2\alpha\Lambda}{\mu} + \frac{2\beta_1 T_1}{K_1} + (\alpha_1 + \alpha_2) T_2 \right) |M_1 - M_2| \\
&\quad + 2\lambda T_1 |S_1 - S_2| + 2\alpha T_1 |R_1 - R_2| \\
&\quad + \left( \frac{2\beta_2}{K_2} + (\alpha_1 + \alpha_2) T_1 \right) |C_2 - C_1| \\
&\leq N \|X_1 - X_2\|,
\end{aligned}$$

where

$$N = \max \left( \frac{2\lambda\Lambda}{\mu} + \frac{2\alpha\Lambda}{\mu} + \frac{2\beta_1 T_1}{K_1} + (\alpha_1 + \alpha_2) T_2, 2\lambda T_1, 2\alpha T_1, \frac{2\beta_2}{K_2} + (\alpha_1 + \alpha_2) T_1, \|A\| \right).$$

Therefore,

$$\|F(X_1) - F(X_2)\| \leq N \|X_1 - X_2\|.$$

Thus, it follows that the function  $F$  is uniformly Lipschitz continuous, and the restriction on  $S(t) \geq 0$ ,  $I(t) \geq 0$ ,  $R(t) \geq 0$ ,  $M(t) \geq 0$  and  $C(t) \geq 0$  in  $\mathbb{R}_+^5$ . Therefore, a solution of the model (1) exists [2].

□

### 3 The optimal control problem

Given the ongoing threat of cochineal infestations and their severe economic impact on cactus production, farmers need a cost-effective strategy to control the pest's spread within a specific timeframe. To address this, we develop an optimal control problem that focuses on minimizing the number of infected plants while also achieving a balance between minimizing infection and maximizing recovery over time. A key aspect of our approach is the natural control provided by *Hyperaspis trifurcata*, a predatory beetle that feeds on cochineal insects. By incorporating this biological control agent into the model, we emphasize the beetle's role in naturally reducing cochineal infestations. *Hyperaspis trifurcata* offers a sustainable and environmentally friendly alternative to chemical pesticides, as it directly targets the cochineal population, helping to curb its spread.

The system of equations (1) is adjusted to include two control variables,  $u_1(t)$  and  $u_2(t)$  for  $t \in [t_0, t_f]$ .

$$\begin{cases} \frac{dS}{dt} = \Lambda - \lambda S(t)M(t) - \mu S(t), \\ \frac{dI}{dt} = \lambda S(t)M(t) + \alpha M(t)R(t) - \theta I(t) - \mu I(t) - u_1(t)I(t), \\ \frac{dR}{dt} = \theta I(t) - \alpha M(t)R(t) - \mu R(t) + u_1(t)I(t), \\ \frac{dM}{dt} = \beta_1 M(t) \left(1 - \frac{M(t)}{K_1}\right) - \gamma_1 M(t) - \alpha_1 M(t)C(t) - u_2(t)\sigma_1 M(t)C(t), \\ \frac{dC}{dt} = \beta_2 C(t) \left(1 - \frac{C(t)}{K_2}\right) - \gamma_2 C(t) + \alpha_2 M(t)C(t) + u_2(t)\sigma_2 M(t)C(t), \end{cases} \quad (7)$$

with the initial conditions  $S(0) \geq 0$ ,  $I(0) \geq 0$ ,  $R(0) \geq 0$ ,  $M(0) \geq 0$ , and  $C(0) \geq 0$ .

The control  $u_1(t)$  represents the application of insecticide to combat cochineal, while the control  $u_2(t)$  denotes the use of *Hyperaspis trifurcata*, a

predator that feeds on cochineal.

The problem is to minimize the objective functional:

$$J(u_1, u_2) = I(t_f) + M(t_f) - R(t_f) + \int_{t_0}^{t_f} \left[ I(t) + M(t) - R(t) + \frac{C_1}{2} (u_1(t))^2 + \frac{C_2}{2} (u_2(t))^2 \right] dt, \quad (8)$$

where  $C_1 > 0$  and  $C_2 > 0$ , are chosen to assign the relative importance of  $u_1(t)$  and  $u_2(t)$  at any given time  $t$ , with  $t_f$  representing the final time.

In other words, our goal is to find the optimal controls  $u_1^*$  and  $u_2^*$  such that

$$J(u_1^*, u_2^*) = \min_{(u_1, u_2) \in U} J(u_1, u_2),$$

where  $U$  is the set of admissible controls defined by

$$U = \{(u_1(t), u_2(t)) : 0 \leq u_1(t) \leq 1, 0 \leq u_2(t) \leq 1, / t \in [t_0, t_f]\}.$$

### 3.1 Existence of optimal controls

In this part, we present the theorem which proves the existence of an optimal control  $(u_1^*, u_2^*)$  minimizing the cost function  $J$ .

**Theorem 4.** There exists an optimal control  $(u_1^*, u_2^*) \in U$  such that

$$J(u_1^*, u_2^*) = \min_{(u_1, u_2) \in U} J(u_1, u_2).$$

*Proof.* To use the existence result in [6], we must check the following properties:

(A<sub>1</sub>): The set of controls and the corresponding state variables is nonempty.

(A<sub>2</sub>): The control set  $U$  is convex and closed.

(A<sub>3</sub>): The right-hand side of the state system is bounded by a linear function in the state and control variables.

(A<sub>4</sub>): The integral  $L(I, R, M, u_1, u_2)$  of the objective functional is convex on  $U$ , and there exist constants  $\varkappa_1 > 0$ ,  $\varkappa_2 > 0$ , and  $\varepsilon > 1$  such that

$$L(I, R, M, u_1, u_2) \geq -\varkappa_1 + \varkappa_2 \left( |u_1|^2 + |u_2|^2 \right)^{\frac{\varepsilon}{2}}.$$

The first condition  $(A_1)$  is verified using the result in [11]. The set  $U$  is convex and closed by the definition. Thus the condition  $(A_2)$ . Our state system is linear in  $u_1$  and  $u_2$ . Moreover, the solutions of the system are bounded as proved in model (1), hence the condition  $(A_3)$ . Also, we have the last needed condition  $(A_4)$ ,

$$L(I, R, M, u_1, u_2) \geq -\varkappa_1 + \varkappa_2 \left( |u_1|^2 + |u_2|^2 \right)^{\frac{\varepsilon}{2}},$$

where  $\varkappa_1 = 2 \sup_{t \in [t_0, t_f]} (I(t), R(t), M(t))$ ,  $\varkappa_2 = \inf(\frac{C_1}{2}, \frac{C_2}{2})$ , and  $\varepsilon = 2$ , since  $C_1 > 0$  and  $C_2 > 0$ .

We conclude that there exists an optimal control  $(u_1^*, u_2^*) \in U$  such that

$$J(u_1^*, u_2^*) = \min_{(u_1, u_2) \in U} J(u_1, u_2).$$

□

### 3.2 Characterization of the optimal controls

In this section, we utilize Pontryagin's principle [13]. The central concept is to introduce the adjoint function, which connects the system of differential equations to the objective functional. This connection leads to the formulation of the Hamiltonian. By applying this principle, the task of determining a control that optimizes the objective functional with a specified initial condition is transformed into the problem of finding a control that optimizes the Hamiltonian pointwise.

To derive the optimal control conditions, we apply Pontryagin's maximum principle such that the Hamiltonian  $H$  at time  $t$  is defined by

$$H(t) = I(t) + M(t) - R(t) + \frac{C_1}{2} (u_1(t))^2 + \frac{C_2}{2} (u_2(t))^2 + \sum_{i=1}^5 \lambda_i h_i, \quad (9)$$

where  $h_i$  is the right side of the system of differential equations (7) of  $i$ th state variable.

**Theorem 5.** Given the optimal controls  $(u_1^*, u_2^*)$  and solutions  $S^*, I^*, R^*, M^*$  and  $C^*$  of the corresponding state system (7), there exist adjoint functions  $\lambda_1, \lambda_2, \lambda_3, \lambda_4$ , and  $\lambda_5$  satisfying

$$\begin{cases} \lambda'_1 = -\frac{dH}{dS} = \lambda_1(\lambda M(t) + \mu) - \lambda_2\lambda M(t), \\ \lambda'_2 = -\frac{dH}{dI} = -1 + \lambda_2(\theta + \mu + u_1(t)) - \lambda_3(\theta + u_1(t)), \\ \lambda'_3 = -\frac{dH}{dR} = 1 - \lambda_2\alpha M(t) + \lambda_3(\alpha M(t) + \mu), \\ \lambda'_4 = -\frac{dH}{dM} = -1 + \lambda_1\lambda S(t) - \lambda_2(\lambda S(t) + \alpha R(t)) + \lambda_3\alpha R(t), \\ \quad -\lambda_4\left(\beta_1\left(1 - \frac{2M(t)}{K_1}\right) - \gamma_1 - \alpha_1 C(t) - u_2(t)\sigma_1 C(t)\right) \\ \quad -\lambda_5(\alpha_2 C(t) + u_2(t)\sigma_2 C(t)), \\ \lambda'_5 = -\frac{dH}{dC} = \lambda_4(\alpha_1 M(t) + u_2(t)\sigma_1 M(t)) \\ \quad -\lambda_5\left(\beta_2\left(1 - \frac{2C(t)}{K_2}\right) - \gamma_2 + \alpha_2 M(t) + u_2(t)\sigma_2 M(t)\right), \end{cases} \quad (10)$$

Such that the transversality conditions at time  $t_f$  are

$$\begin{cases} \lambda_1(t_f) = 0, \\ \lambda_2(t_f) = 1, \\ \lambda_3(t_f) = -1, \\ \lambda_4(t_f) = 1, \\ \lambda_5(t_f) = 0. \end{cases} \quad (11)$$

In addition to that we have, for  $t \in [t_0, t_f]$ , optimal controls  $u_1^*(t)$  and  $u_2^*(t)$  are given by

$$\begin{cases} u_1^*(t) = \min\left(1, \max\left(0, \frac{1}{C_1}(\lambda_2 - \lambda_3)I(t)\right)\right), \\ u_2^*(t) = \min\left(1, \max\left(0, \frac{1}{C_2}(\sigma_1\lambda_4 - \sigma_2\lambda_5)C(t)M(t)\right)\right). \end{cases} \quad (12)$$

*Proof.* The Hamiltonian  $H$  is defined as follows:

$$H(t) = I(t) + M(t) - R(t) + \frac{C_1}{2}(u_1(t))^2 + \frac{C_2}{2}(u_2(t))^2 + \sum_{i=1}^5 \lambda_i h_i,$$

where

$$\begin{cases} h_1 = \Lambda - \lambda S(t)M(t) - \mu S(t), \\ h_2 = \lambda S(t)M(t) + \alpha M(t)R(t) - \theta I(t) - \mu I(t) - u_1(t)I(t), \\ h_3 = \theta I(t) - \alpha M(t)R(t) - \mu R(t) + u_1(t)I(t), \\ h_4 = \beta_1 M(t) \left(1 - \frac{M(t)}{K_1}\right) - \gamma_1 M(t) - \alpha_1 M(t)C(t) - u_2(t)\sigma_1 M(t)C(t), \\ h_5 = \beta_2 C(t) \left(1 - \frac{C(t)}{K_2}\right) - \gamma_2 C(t) + \alpha_2 M(t)C(t) + u_2(t)\sigma_2 M(t)C(t). \end{cases}$$

For  $t \in [t_0, t_f]$ , the adjoint equations and transversality conditions can be obtained by using Pontryagin's maximum principle [13] such that

$$\begin{cases} \lambda'_1 = -\frac{dH}{dS} = \lambda_1(\lambda M(t) + \mu) - \lambda_2 \lambda M(t), \\ \lambda'_2 = -\frac{dH}{dI} = -1 + \lambda_2(\theta + \mu + u_1(t)) - \lambda_3(\theta + u_1(t)), \\ \lambda'_3 = -\frac{dH}{dR} = 1 - \lambda_2 \alpha M(t) + \lambda_3(\alpha M(t) + \mu), \\ \lambda'_4 = -\frac{dH}{dM} = -1 + \lambda_1 \lambda S(t) - \lambda_2(\lambda S(t) + \alpha R(t)) + \lambda_3 \alpha R(t), \\ \quad -\lambda_4 \left( \beta_1 \left(1 - \frac{2M(t)}{K_1}\right) - \gamma_1 - \alpha_1 C(t) - u_2(t)\sigma_1 C(t) \right) \\ \quad -\lambda_5 (\alpha_2 C(t) + u_2(t)\sigma_2 C(t)) \\ \lambda'_5 = -\frac{dH}{dC} = \lambda_4(\alpha_1 M(t) + u_2(t)\sigma_1 M(t)) \\ \quad -\lambda_5 \left( \beta_2 \left(1 - \frac{2C(t)}{K_2}\right) - \gamma_2 + \alpha_2 M(t) + u_2(t)\sigma_2 M(t) \right). \end{cases}$$

For  $t \in [t_0, t_f]$ , the optimal controls  $u_1^*$  and  $u_2^*$  can be solved from the optimality condition we have

$$\frac{dH}{du_1} = C_1 u_1(t) - \lambda_2 I(t) + \lambda_3 I(t) = 0.$$

So

$$u_1(t) = \frac{1}{C_1} (\lambda_2 - \lambda_3) I(t),$$

we have

$$\frac{dH}{du_2} = C_2 u_2(t) - \lambda_4 \sigma_1 M(t)C(t) + \lambda_5 \sigma_2 M(t)C(t) = 0.$$

So

$$u_2(t) = \frac{1}{C_2} (\sigma_1 \lambda_4 - \sigma_2 \lambda_5) M(t)C(t).$$

By the bounds in  $U$  of the controls, it is convenient to obtain  $u_1^*$  and  $u_2^*$  in the form of (12).

□

## 4 Numerical simulations

This section begins by introducing an iterative method for numerically solving the optimality system, followed by a presentation of the numerical results obtained using MATLAB.

### 4.1 Discretization and control algorithm

The numerical algorithm presented below uses a semi-implicit finite difference method to discretize the time interval  $[t_0, t_f]$  at the points  $t_i = t_0 + ih$  ( $i = 0, 1, \dots, n$ ), where  $h$  is the time step such that  $t_n = t_f$  [7]. The state variables  $S(t), I(t), R(t), M(t), C(t)$ , and the adjoint variables  $\lambda_1, \lambda_2, \lambda_3, \lambda_4, \lambda_5$ , along with the control variables  $u_1$  and  $u_2$ , are defined at the nodal points as  $S_i, I_i, R_i, M_i, C_i, \lambda_1^i, \lambda_2^i, \lambda_3^i, \lambda_4^i, \lambda_5^i, u_1^i, u_2^i$ .

We proceed with the discretization using a combination of forward and backward difference approximations as follows: The method, developed by [8] and presented in [9] and [10], is then read as

$$\begin{aligned}\frac{S_{i+1} - S_i}{h} &= \Lambda - \lambda S_{i+1} M_{i+1} - \mu S_{i+1}, \\ \frac{I_{i+1} - I_i}{h} &= \lambda S_{i+1} M_{i+1} + \alpha M_{i+1} R_{i+1} - \theta I_{i+1} - \mu I_{i+1} - u_1^i I_{i+1}, \\ \frac{R_{i+1} - R_i}{h} &= \theta I_{i+1} - \alpha M_{i+1} R_{i+1} - \mu R_{i+1} + u_1^i I_{i+1}, \\ \frac{M_{i+1} - M_i}{h} &= \beta_1 M_{i+1} \left(1 - \frac{M_{i+1}}{K_1}\right) - \gamma_1 M_{i+1} - \alpha_1 M_{i+1} C_{i+1} - u_2^i \sigma_1 M_{i+1} C_{i+1}, \\ \frac{C_{i+1} - C_i}{h} &= \beta_2 C_{i+1} \left(1 - \frac{C_{i+1}}{K_2}\right) - \gamma_2 C_{i+1} + \alpha_2 M_{i+1} C_{i+1} + u_2^i \sigma_2 M_{i+1} C_{i+1}.\end{aligned}$$

Using a similar approach, we approximate the time derivative of the adjoint variables by applying a first-order backward difference and then use the corresponding scheme as follows:

$$\begin{aligned}
\frac{\lambda_1^{n-i} - \lambda_1^{n-i-1}}{h} &= (\lambda_1^{n-i-1} - \lambda_3^{n-i}) (\lambda M_{i+1} + \mu) - \lambda_2^{n-i} \lambda M_{i+1}, \\
\frac{\lambda_2^{n-i} - \lambda_2^{n-i-1}}{h} &= -1 + (\lambda_1^{n-i-1} - \lambda_2^{n-i-1}) (\theta + \mu + u_1^i) - \lambda_3^{n-i} (\theta + u_1^i), \\
\frac{\lambda_3^{n-i} - \lambda_3^{n-i-1}}{h} &= 1 - \lambda_2^{n-i-1} \alpha M_{i+1} + \lambda_3^{n-i-1} (\alpha M_{i+1} + \mu), \\
\frac{\lambda_4^{n-i} - \lambda_4^{n-i-1}}{h} &= -1 + \lambda_1^{n-i-1} \lambda S_{i+1} - \lambda_2^{n-i-1} (\lambda S_{i+1} + \alpha R_{i+1}) + \lambda_3^{n-i-1} \alpha R_{i+1} \\
&\quad - \lambda_4^{n-i-1} \left( \beta_1 \left( 1 - \frac{2M_{i+1}}{K_1} \right) - \gamma_1 - \alpha_1 C_{i+1} - u_2^i \sigma_1 C_{i+1} \right) \\
&\quad - \lambda_5^{n-i-1} \left( \alpha_2 C_{i+1} + u_2^i \sigma_2 C_{i+1} \right), \\
\frac{\lambda_5^{n-i} - \lambda_5^{n-i-1}}{h} &= \lambda_4^{n-i-1} \left( \alpha_1 M_{i+1} + u_2^i \sigma_1 M_{i+1} \right) \\
&\quad - \lambda_5^{n-i-1} \left( \beta_2 \left( 1 - \frac{2C_{i+1}}{K_2} \right) - \gamma_2 + \alpha_2 M_{i+1} + u_2^i \sigma_2 M_{i+1} \right).
\end{aligned}$$

The control variables are updated as follows:

## Algorithm 2

*Step 1:*

$$\begin{aligned}
S(0) &= S_0, \quad I(0) = I_0, \quad R(0) = R_0, \quad M(0) = M_0, \quad C(0) = C_0, \\
\lambda_1(t_f) &= 0, \quad \lambda_2(t_f) = 1, \quad \lambda_3(t_f) = -1, \quad \lambda_4(t_f) = 1, \quad \lambda_5(t_f) = 0, \\
u_1(0) &= 0, \quad u_2(0) = 0.
\end{aligned}$$

*Step 2:*

For  $i = 0, \dots, n-1$ , do:

$$\begin{aligned}
S_{i+1} &= \frac{S_i + h\Lambda}{1 + h(\lambda M_{i+1} + \mu)}, \\
I_{i+1} &= \frac{I_i + h(\lambda S_{i+1} M_{i+1} + \alpha M_{i+1} R_{i+1})}{1 + h(\theta + \mu + u_1^i)}, \\
R_{i+1} &= \frac{R_i + h(\theta I_{i+1} - \alpha M_{i+1} R_{i+1} + u_1^i I_{i+1})}{1 + h\mu}, \\
M_{i+1} &= \frac{M_i + h\beta_1 M_{i+1} \left( 1 - \frac{M_{i+1}}{K_1} \right)}{1 + h(\gamma_1 + \alpha_1 C_{i+1} + u_2^i \sigma_1 C_{i+1})},
\end{aligned}$$

$$C_{i+1} = \frac{C_i + h(\beta_2 C_{i+1} \left(1 - \frac{C_{i+1}}{K_2}\right) + \alpha_2 M_{i+1} C_{i+1} + u_2^i \sigma_2 M_{i+1} C_{i+1})}{1 + h\gamma_2}.$$

$$\begin{aligned}\lambda_1^{n-i-1} &= \frac{\lambda_1^{n-i} + h((\lambda_1^{n-i-1} - \lambda_3^{n-i})(\lambda M_{i+1} + \mu) - \lambda_2^{n-i} \lambda M_{i+1})}{1 + h(\lambda M_{i+1} + \mu)}, \\ \lambda_2^{n-i-1} &= \frac{\lambda_2^{n-i} + h(-1 + (\lambda_1^{n-i-1} - \lambda_2^{n-i-1})(\theta + \mu + u_1^i) - \lambda_3^{n-i}(\theta + u_1^i))}{1 + h(\theta + \mu + u_1^i)}, \\ \lambda_3^{n-i-1} &= \frac{\lambda_3^{n-i} + h(1 - \lambda_2^{n-i-1} \alpha M_{i+1} + \lambda_3^{n-i-1}(\alpha M_{i+1} + \mu))}{1 + h(\alpha M_{i+1} + \mu)}, \\ \lambda_4^{n-i-1} &= \frac{\lambda_4^{n-i} + h(-1 + \lambda_1^{n-i-1} \lambda S_{i+1} - \lambda_2^{n-i-1}(\lambda S_{i+1} + \alpha R_{i+1}) + \lambda_3^{n-i-1} \alpha R_{i+1})}{1 + h(\beta_1(1 - \frac{2M_{i+1}}{K_1}) - \gamma_1 - \alpha_1 C_{i+1} - u_2^i \sigma_1 C_{i+1})}, \\ \lambda_5^{n-i-1} &= \frac{\lambda_5^{n-i} + h(\lambda_4^{n-i-1}(\alpha_1 M_{i+1} + u_2^i \sigma_1 M_{i+1}))}{1 + h(\beta_2(1 - \frac{2C_{i+1}}{K_2}) - \gamma_2 + \alpha_2 M_{i+1} + u_2^i \sigma_2 M_{i+1})}.\end{aligned}$$

$$\begin{aligned}T^{i+1} &= \frac{(\lambda_1^{n-i-1} - \lambda_3^{n-i-1})S_{i+1}}{A}, \\ u^{i+1} &= \min(0.9, \max(0, T^{i+1})).\end{aligned}$$

End for.

*Step 3:*

For  $i = 0, \dots, n$ , write:

$$S^*(t_i) = S_i, \quad I^*(t_i) = I_i, \quad R^*(t_i) = R_i, \quad M^*(t_i) = M_i, \quad C^*(t_i) = C_i, \quad u^*(t_i) = u^i.$$

End for.

## 4.2 Numerical results

In this subsection, we present the results obtained by solving the optimality system. For our control problem, we define conditions for the state variables and terminal conditions for the adjoint variables. The optimality system is essentially a two-point boundary value problem, with conditions at the initial time step  $i = t_0$  and the final time step  $i = t_f$ . To solve this system, we ini-

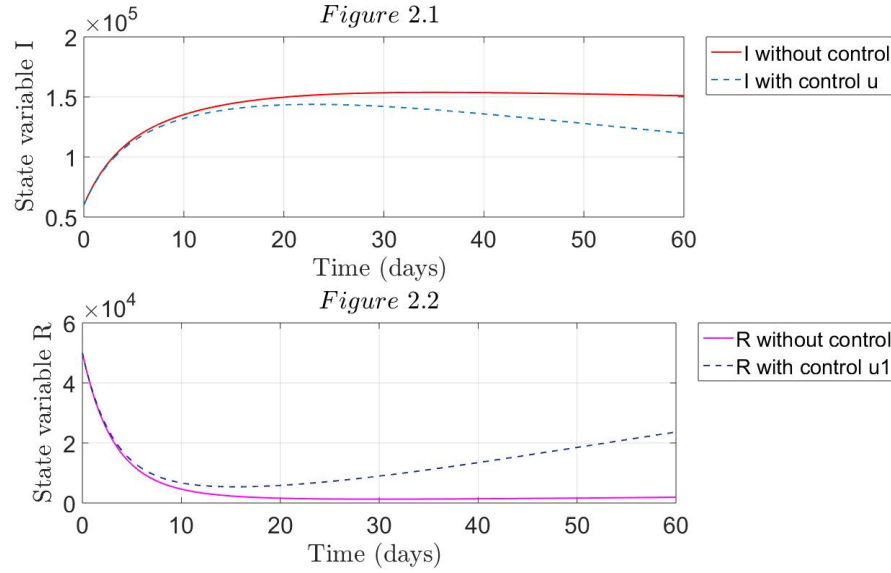
tially solve the state model, followed by solving the adjoint system in reverse order. In the first iteration, we start with an initial guess for the control variables and update them based on a characterization of the optimal controls before moving on to the next iteration. This process is repeated until the iterates converge. To achieve this, we created a MATLAB code utilizing the following parameters. Given the lack of real-world data, the parameter values were chosen hypothetically. The plots illustrating susceptible, infected, and recovered individuals—both with and without control measures—are generated based on these parameter values:  $\Lambda = 1000$ ,  $\lambda = 0.0005$ ,  $\beta_1 = 0.0001$ ,  $\beta_2 = 0.000001$ ,  $\mu = 0.00009$ ,  $\theta = 0.0002$ ,  $\alpha = 0.003$ ,  $\alpha_1 = 0.001$ ,  $\alpha_2 = 0.001$ ,  $\gamma_1 = 0.008$ ,  $\gamma_2 = 0.001$ ,  $\sigma = 1000$ ,  $\sigma_1 = 0.2$ ,  $\sigma_2 = 0.1$ . When analyzing the graphs, please be aware that solid lines represent individuals without control measures, whereas dashed lines indicate those with control measures.

### 4.3 Control Strategy 1: Impacts of Insecticide Application on Cochineal

The goal of this approach is to minimize the function (8), with a primary focus on reducing the cochineal population through insecticide spraying. Figure 2 illustrates the effects of this spraying on the cacti plants.

In Figure 2 (2.1), it is clear that in the absence of control measures, the number of infected cacti steadily increases, reaching approximately  $1.5 \times 10^5$  within the first two months. However, when control measures are applied, the number of infected cacti begins to decrease from day one of implementation, eventually dropping to around  $1.2 \times 10^5$ .

In Figure 2 (2.2), the recovered cacti rises to about  $2 \times 10^4$  with the application of the control strategy, compared to  $0.1 \times 10^4$  birds when control measures are not implemented.

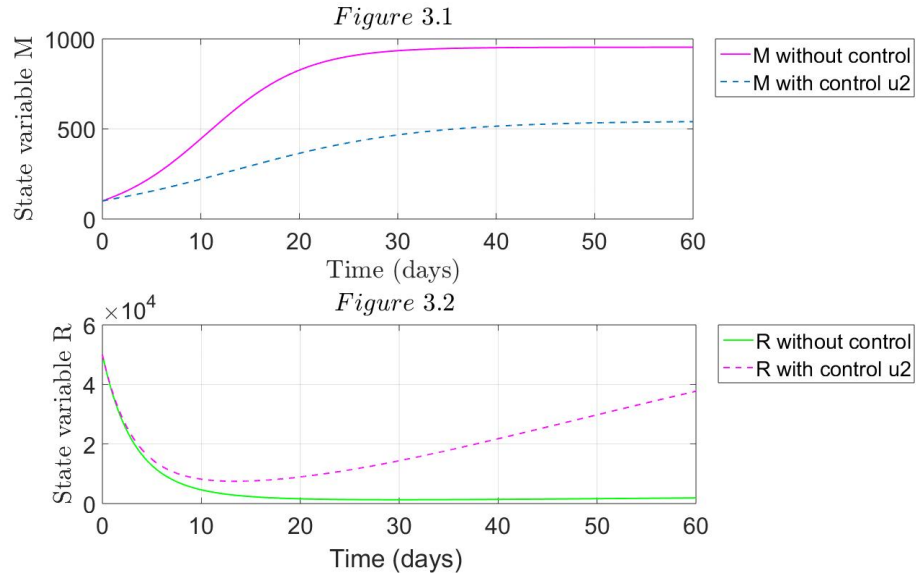
Figure 2: Dynamics with control  $u_1$ .

#### 4.4 Control Strategy 2: Use of Hyperaspis Trifurcata that feeds on the cochineal

The main objective of treating cacti infested with cochineal, within the context of a strategy, is to minimize the function (8) while maintaining other control measures at zero. Figure 3 illustrates the disease dynamics, taking into account the presence or absence of this control measure.

In Figure 3 (3.1), it is clear that without any control measures, the cochineal population steadily increases, peaking at around 1,000 during the first two months. However, with the implementation of controls, the cochineal population begins to decline from the first day and decreases to approximately 500 within two months.

In Figure 3 (3.2), the number of recovered cacti increases to approximately  $4 \times 10^4$  with the use of the control strategy, whereas only  $0.1 \times 10^4$  are recovered when no control measures are in effect.

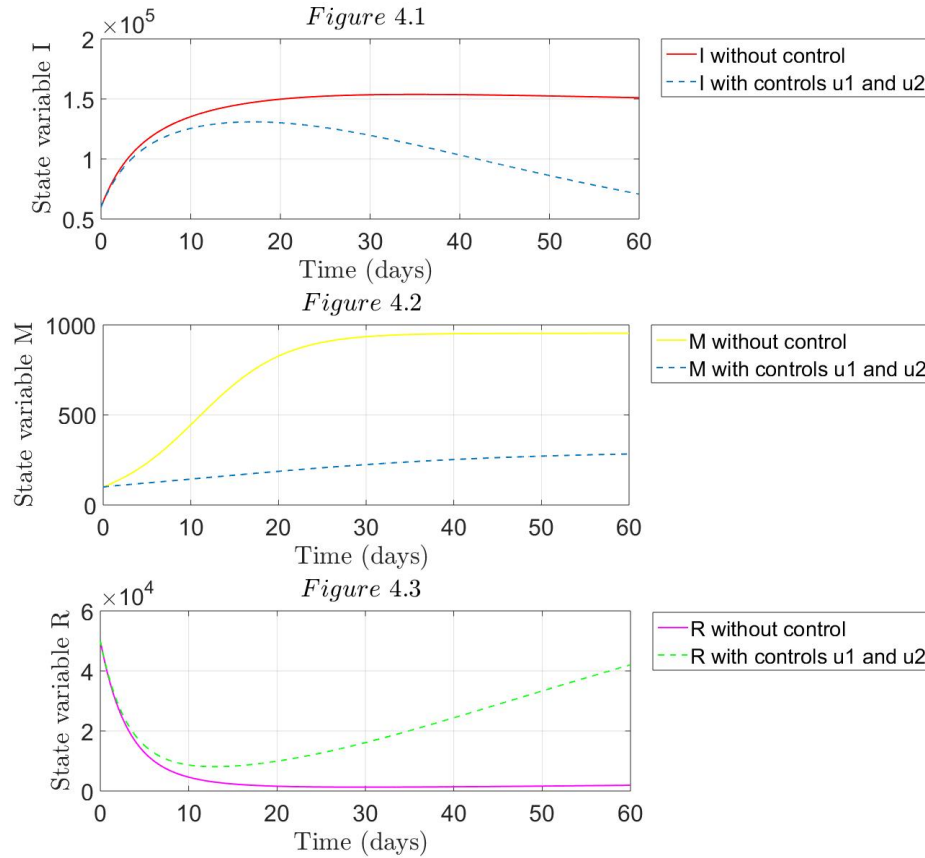
Figure 3: Dynamics with control  $u_2$ .

#### 4.5 Control Strategy 3: insecticide spraying and *Hyperaspis Trifurcata*

This strategy aims to minimize the objective function (8) by implementing both control measures. Figure 4 illustrates the disease progression with both controls in effect, compared to the scenario where no control measures are utilized to manage the disease.

In Figure 4 (4.1), the impact of insecticide as a control measure, along with Use of an insect that feeds on the cochineal, on curbing the propagation of the cochineal are clearly illustrated. It is evident that in the absence of control measures, the number of infected cacti increases, peaking at around  $1.5 \times 10^5$  during the first two months. In contrast, with the implementation of control measures, the infected cacti decreases to approximately  $0.2 \times 10^5$  within two months.

In Figure 4 (4.2), it is observed that in the absence of control measures, the number of the cochineal rises, peaking at around 1000 the first month. In contrast, when control measures are put in place, the number of the cochineal

Figure 4: Dynamics with control  $u_1$  and  $u_2$ .

consistently decreases, ultimately reaching 250 within two months.

In Figure 4 (4.3), the number of recovered cacti rises within the first week of implementing the control strategy, ultimately reaching about  $4.5 \times 10^4$  over the course of two months. In contrast, when no control measures are in effect, the number of recovered cacti decreases to nearly zero.

## 5 Conclusion

In this paper, we introduced a novel model designed to improve the understanding of cochineal dynamics in cactus plants. Our goal is to develop an optimal control strategy based on the *SIRMC* model that minimizes the number of infected cacti while also achieving a balance between minimizing infection and maximizing recovery. We compared scenarios with and without control measures, demonstrating that the implementation of control strategies substantially decreases the number of infected plants. To offer a thorough overview of cochineal dynamics, we presented figures that display the counts of infected, recovered, and cochineal in both scenarios ( $u_1$  and  $u_2$ ), highlighting the differences with and without control measures, as illustrated in Figures 1, 2, 3 and 4. Our results indicate the effectiveness of control measures in reducing the spread of cochineal in cacti.

By incorporating the *SIRMC* model with optimal control strategies, we underscore the potential to decrease disease prevalence and improve recovery rates. These findings highlight the importance of proactive intervention strategies in cactus fields, offering valuable insights for farmers.

## References

- [1] Arba, M., Cherif Benismail, M., and Mokhtari, M. *The cactus pear (Opuntia spp.) in Morocco: Main species and cultivar characterization*, Acta Hortic. 581, (2002), 103–109.
- [2] Birkhoff, G., and Rota, G.C. *Ordinary differential equations*, 4th ed. New York: John Wiley & Sons; 1989.
- [3] Bouharroud, R., Amarraque, A., and Qessaoui, R. *First report of the opuntia cochineal scale Dactylopius opuntiae (Hemiptera: Dactylopiidae) in Morocco*, Bulletin OEPP/EPPO Bulletin, 46(2), (2016), 308–310.
- [4] Chávez-Moreno, C. K., Tecante, A., and Casas, A. *The opuntia (Cactaceae) and dactylopius (Hemiptera: Dactylopiidae) in Mexico: A histor-*

- ical perspective of use, interaction and distribution*, Biodivers. Conserv. 18, (2009) 3337–3355.
- [5] De Jesus Mendez-Gallegos, S., Tiberi, R. and Panzavolta, T. *Carmines cochineal dactylopius coccus Costa (Rhynchota: Dactylopiidae): Significance, production and use*, Adv. Hortic. Sci. 17(3) (2023) 165–171.
- [6] Fleming, W.H., and Rishel, R.W. *Deterministic and stochastic optimal control*, New York, NY, USA, Springer, 1975.
- [7] Gumel, A. B., Patidar, K. C., and Spiteri, R. J. *Asymptotically consistent non-standard finite-difference methods for solving mathematical models arising in population biology*, R.E. Mickens, World Scientific, Singapore, 2005.
- [8] Gumel, A.B., Shivakumar, P.N., and Sahai, B.M. *A mathematical model for the dynamics of HIV-1 during the typical course of infection*, Third World Congress of Nonlinear Analysts, 47, (2001) 2073–2083.
- [9] Hattaf, K., Rachik, M., Saadi, S., Tabit, Y., and Yousfi, N. *Optimal control of tuberculosis with exogenous reinfection*, Appl. Math. Sci. 3(5), (2009) 231–240.
- [10] J. Karakchou, M., Rachik, and Gourari, S. *Optimal control and infectiology: Application to an HIV/AIDS model*, Appl. Math. Comput. 177, (2006) 807–818.
- [11] Lukes, D.L. *Differential Equations: Classical to Controlled*, Academic Press, New York, 1982.
- [12] Mora, M., Cortés, M., Sanhueza, C., and Sáenz, C. *Innovation requirements for the development of cactus pear for export: A new item to be incorporated to the chilean fruit export sector*, Acta Hortic. 995, (2013) 229–236.
- [13] Pontryagin, L.S., Boltyanskii, V.G., Gamkrelidze, R.V., and Mishchenko, E.F. *The mathematical theory of optimal processes*, New York, NY, USA, Wiley, 1962.

- [14] Ramadan, M.F., and Mörsel, J.T. *Oil cactus pear (Opuntia ficus-indica L.)*, Food Chem. (2003) 339–345.
- [15] Shetty, A.A., Rana, M.K., and Preetham, S.P. *Cactus: A medicinal food*, J. Food Sci. Technol. 49(5), (2012) 530–536.
- [16] <https://www.agriculture.gov.ma/ar/actualites/albhth-alzray-tqdm-ham-fy-mkafht-alhshrt-alqrmzyt-almdmrt-llsbar>.
- [17] <https://www.agriculture.gov.ma/ar/filieres-regions/cactus-gon>.



## An efficient Dai-Kou-type method with image de-blurring application

K. Ahmed<sup>1,\*</sup>, M.Y. Waziri<sup>1</sup>, S. Murtala<sup>2</sup>, A.S. Halilu<sup>3,4</sup>, H. Abdullahi<sup>3</sup> and Y.B. Musa<sup>3</sup>

### Abstract

Well-conditioning of matrices has been shown to improve the numerical performance of algorithms by way of ensuring their numerical stability. In this paper, a modified Dai-Kou-type conjugate gradient method is developed for constrained nonlinear monotone systems by employing the well-conditioning approach. The new method ensures that the much required

\*Corresponding author

Received 18 March 2025; revised 20 May 2025; accepted 1 June 2025

Kabiru Ahmed, e-mail: kabiruhungu16@gmail.com

Mohammed Yusuf Waziri, e-mail: mywaziri.mth@buk.edu.ng

Salisu Murtala, e-mail: salisumurtala@gmail.com

Abubakar Sani Halilu, e-mail: abubakars.halilu@slu.edu.ng

Habibu Abdullahi, e-mail: habibmth.slu@gmail.com

Ya'u Balarabe Musa, e-mail: yaumusa.jsu@gmail.com

<sup>1</sup>Department of Mathematical Sciences, Bayero University, Kano, Nigeria.

<sup>2</sup>Department of Mathematics, Federal University, Dutse, Nigeria.

<sup>3</sup>Department of Mathematics, Sule Lamido University, Kafin Hausa, Nigeria.

<sup>4</sup>Faculty of Informatics and Computing, Universiti Sultan Zainal Abidin, Campus Besut, 22200 Terengganu, Malaysia

### How to cite this article

Ahmed, K., Waziri, M.Y., Murtala, S., Halilu, A.S., Abdullahi, H. and Musa, Y.B., An efficient Dai-Kou-type method with image de-blurring application. *Iran. J. Numer. Anal. Optim.*, 2025; 15(3): 1171-1209. <https://doi.org/10.22067/ijnao.2025.92708.1615>

condition for global convergence of iterates generated is satisfied irrespective of the linesearch strategy employed. Another novelty of the scheme is its practical application in image de-blurring problems. The method performs well and converges globally under mild assumptions. Experiments in image de-blurring and convex constrained systems of equations show the scheme to be effective.

**AMS subject classifications (2020):** Primary 90C30; Secondary 90C26, 94A12.

**Keywords:** Nonlinear equations; Eigenvalues; Constrained equations; Convex set; Sparse signals.

## 1 Introduction

Generally, a system of nonlinear monotone equations is given by

$$F(x) = 0, \quad x \in \mathbb{R}^n, \quad (1)$$

with  $F$  from  $\mathbb{R}^n \rightarrow \mathbb{R}^n$ , being a continuous and monotone mapping. Monotonicity of  $F$  means it satisfies the inequality

$$(F(x) - F(y))^T(x - y) \geq 0, \quad \text{for all } x, y \in \mathbb{R}^n. \quad (2)$$

For the constrained version of (1), which is formulated as

$$F(\bar{x}) = 0; \quad \bar{x} \in \mathcal{C}, \quad (3)$$

$\bar{x}$  resides in a closed convex nonempty set  $\mathcal{C} \subseteq \mathbb{R}^n$  for which (3) holds.

The Newton's and quasi-Newton's methods [14, 21, 48, 54] are the famous schemes employed for solving (1) and (3). However, storing the Jacobian or its approximation in every iteration, renders these methods unsuitable for high dimension problems.

The appropriate iterative scheme that conveniently addresses storage requirements is the conjugate gradient (CG) scheme. It is usually designed for the optimization problem

$$\min_{x \in \mathbb{R}^n} f(x), \quad (4)$$

in which  $f$  denotes a smooth real-valued function. The CG method is often applied to solve (4) due to its minimal memory requirement. As with other line search methods, and starting with  $x_0 \in \mathbb{R}^n$ , the CG method's iterates are obtained via

$$x_{k+1} = x_k + s_k, \quad s_k = \vartheta_k d_k, \quad k \geq 0, \quad (5)$$

where  $x_k$  stands for previous iterate,  $\vartheta_k > 0$  is the steplength that is usually obtained using a well-defined formula in the scheme's direction  $d_k$ , namely,

$$d_{k+1} = -g_{k+1} + \beta_k d_k, \quad d_0 = -g_0, \quad (6)$$

where  $g_{k+1} = g(x_{k+1})$ ,  $g_0 = g(x_0)$  represent gradients of  $f$  at  $x_{k+1}$  and  $x_k$ . In addition,  $\beta_k$  in (6) is a parameter that defines the CG scheme and its various formulation exists in the literature (see [31, 42]). The classical ones are proposed in [20, 25, 27, 33, 41, 50, 51] and are given by

$$\beta_k^{FR} = \frac{\|g_{k+1}\|^2}{\|g_k\|^2}, \quad \beta_k^{CD} = \frac{\|g_{k+1}\|^2}{-g_k^T d_k}, \quad \beta_k^{DY} = \frac{\|g_{k+1}\|^2}{d_k^T (g_{k+1} - g_k)}, \quad (7)$$

$$\beta_k^{HS} = \frac{g_{k+1}^T (g_{k+1} - g_k)}{d_k^T (g_{k+1} - g_k)}, \quad \beta_k^{PRP} = \frac{g_{k+1}^T (g_{k+1} - g_k)}{\|g_k\|^2}, \quad \beta_k^{LS} = \frac{g_{k+1}^T (g_{k+1} - g_k)}{-g_k^T d_k}, \quad (8)$$

with  $\|\cdot\|$  being the  $\ell_2$ -norm of vectors.

A typical CG scheme implemented with (5) and (6), generates descent directions if the following inequality holds:

$$d_{k+1}^T g_{k+1} < 0. \quad (9)$$

However, for convergence analysis, the CG methods are required to satisfy the following sufficient descent condition:

$$d_{k+1}^T g_{k+1} \leq -c \|g_{k+1}\|^2, \quad c > 0. \quad (10)$$

By seeking a CG direction such that it will be closest to that of the scaled memoryless BFGS scheme [55], Dai and Kou [18] provided a class of CG schemes (DK) for solving (4) with the update parameter

$$\beta_k^{DK} = \frac{y_k^T g_{k+1}}{y_k^T d_k} - \left( \tau_k + \frac{\|y_k\|^2}{s_k^T y_k} - \frac{y_k^T s_k}{\|s_k\|^2} \right) \frac{g_{k+1}^T s_k}{y_k^T d_k}, \quad (11)$$

where  $y_k = g_{k+1} - g_k$ . The authors in [18] defined  $\tau_k$  in (11) similar to the one given in [55]. Interestingly, other formulations have been provided over the years, which include the ones in [47] provided by Oren and Spedicato, namely,

$$\tau_k^{(1)} = \frac{s_k^T y_k}{y_k^T M_k y_k}, \quad \tau_k^{(2)} = \frac{\|y_k\|^2}{s_k^T y_k},$$

the ones proposed by Oren and Luenberger in [46], that is,

$$\tau_k^{(3)} = \frac{s_k^T M_k^{-1} s_k}{s_k^T y_k}, \quad \tau_k^{(4)} = \frac{s_k^T y_k}{s_k^T Q_k s_k},$$

as well as the choice provided in [6] by Al-Baali, namely,

$$\tau_k^{(5)} = \min \left\{ 1, \frac{\|y_k\|^2}{s_k^T y_k} \right\}, \quad \tau_k^{(6)} = \min \left\{ 1, \frac{s_k^T y_k}{\|s_k\|^2} \right\},$$

where  $M_k$  and  $Q_k$  are matrices. The approximation of  $\tau_k$  given in [18], that is,

$$\tau_k = \frac{s_k^T y_k}{\|s_k\|^2},$$

has so far been taken to be the most effective for implementing the DK scheme. In their work in [18], the authors declared that other efficient approximations of  $\tau_k$  can be obtained by employing different approaches.

Due to the appealing attributes of CG schemes for solving (4) with the knowledge that the optimality condition of (4) and (3) equates both concepts, that is,  $\nabla f = F$ , where  $F$  denotes the gradient of some objective functions, researchers have proposed their versions for solving (1) [34, 58, 59] and (3) [2, 3, 4, 32, 36, 40, 57, 63, 62]. Search directions of these schemes are defined as

$$d_0 = -F_0, \quad d_{k+1} = -F_{k+1} + \beta_k d_k, \quad F_{k+1} = F(x_{k+1}), \quad k = 0, 1, \dots,$$

with  $\beta_k$  representing a modified version of any of the earlier CG parameters in (7) and (8) or their hybrid. To that end, researchers have combined the parameters in (7) and (8) with the projection technique in [54] to solve (1) and (3) (see [3, 34, 40, 58, 59, 62] for details). In response to the issue raised

by the authors in [18] regarding other more effective approximations of the parameter  $\tau_k$  in (11), some research aimed at addressing it have been made in recent years. For example, Ding et al. [22] provided a class of DK schemes for (3) with choices of  $\tau_k$  given as

$$\tau_k^A = \frac{\|y_k\|^2}{s_k^T y_k}, \quad \tau_k^B = \frac{s_k^T y_k}{\|s_k\|^2}, \quad (12)$$

or the convex combination

$$\tau_k = \delta \tau_k^A + (1 - \delta) \tau_k^B, \quad \delta \in [0, 1], \quad (13)$$

in which

$$y_k = \tilde{y}_k - \lambda_k \sigma_k \|F_k\| d_k, \quad \sigma_k d_k = s_k, \quad \sigma_k > 0,$$

with

$$\lambda_k = 1 + \|F_k\|^{-1} \max \left\{ 0, \frac{-\sigma_k (\tilde{y}_k^T d_k)}{\|\sigma_k d_k\|^2} \right\}, \quad \tilde{y}_k = F_{k+1} - F_k, \quad F_k = F(x_k).$$

Following the work in [22] and by exploiting Newton's direction, Waziri et al. [56] presented another DK-type scheme for solving (3) with the choice of  $\tau_k$  given as

$$\tau_k^{MDK} = 1 + \frac{s_k^T w_k}{\|s_k\|^2} - \frac{\|w_k\|^2}{s_k^T w_k}, \quad (14)$$

where

$$w_k = y_k + C_k s_k + D \|F_k\|^r s_k, \quad y_k = F(z_k) - F(x_k), \quad s_k = z_k - x_k = \sigma_k d_k,$$

$$C_k = \max \left\{ -\frac{s_k^T y_k}{\|s_k\|}, 0 \right\}, \quad D > 0, \quad r > 0.$$

In their recent work, Waziri et al. [2] proposed two other types of DK-type methods for (3) with approximations of  $\tau_k$  defined as

$$\bar{\tau}_k = \max \left\{ \tilde{\tau}_k, c_1 \frac{\|\bar{y}_k\|^2}{\bar{s}_k^T \bar{y}_k} \right\}, \quad \hat{\tau}_k = \max \left\{ \tilde{\tau}_k, c_2 + \frac{\|\bar{y}_k\|^2}{\bar{s}_k^T \bar{y}_k} \right\}, \quad (15)$$

in which

$$\tilde{\tau}_k = \frac{3 \bar{s}_k^T \bar{y}_k}{\|\bar{s}_k\|^2} - \frac{\|\bar{y}_k\|^2}{\bar{s}_k^T \bar{y}_k}, \quad (16)$$

where

$$\bar{y}_k = y_k + \psi_k \bar{s}_k + \Lambda \|F_k\|^r \bar{s}_k, \quad \bar{s}_k = w_k - x_k, \quad y_k = F(w_k) - F(x_k), \quad \Lambda > 0, \quad r > 0,$$

with

$$\psi_k = \max \left\{ \frac{-\bar{s}_k^T y_k}{\|\bar{s}_k\|^2}, 0 \right\},$$

and  $c_1 > 1$  and  $c_2 > 0$ .

**Remark 1.** It is worth stating here that only the schemes in [22] with choices of  $\tau_k$  presented in (12) and (13) satisfy the condition (10) necessary for determining global convergence of algorithms for the problem (3) without any adjustments. For instance, the choice of  $\tau_k$  given in (15) was obtained by adjusting the original choice in (16) since adopting the latter may not satisfy (10) automatically. Also, note that the choice in (14) may be negative or zero at some iterative point and may also not always satisfy (10). Lastly, the iteration matrices of the directions in [2, 22, 56] were not shown to be well-conditioned, which could improve the efficiency of the methods.

The article's objectives are listed as follows:

- To derive an efficient DK-type scheme for the constrained problem (3) with an approximation of  $\tau_k$  obtained without any adjustments.
- To present a DK-type scheme for which the inequality (10) necessary in obtaining convergence results of methods for the problem (3) holds.
- To derive a method in which the symmetric form of its direction matrix is well-conditioned.
- To present proof of the scheme's convergence under mild conditions.
- To apply the scheme to image deblurring problems.

The remaining sections of the paper are outlined as follows: Section 2 deals with motivation and derivation of the proposed algorithm. Section 3 discusses the results of the convergence of the scheme. In Section 4, results of

experiments carried out for problem (3) and image deblurring are discussed, while conclusions are made in Section 5.

## 2 Inspiration and Algorithm

We first recall that the most prominent quasi-Newton scheme developed by the researchers Broyden [15], Fletcher [26], Goldfarb [30], and Shanno [53] popularly known as BFGS, where  $B_k$  is usually an  $n \times n$  symmetric positive-definite matrix is formulated as

$$B_{k+1} = B_k - \frac{B_k s_k s_k^T B_k}{s_k^T B_k s_k} + \frac{y_k y_k^T}{y_k^T y_k}, \quad s_k \in \mathbb{R}^n, \quad y_k \in \mathbb{R}^n. \quad (17)$$

From the Woodbury formula presented in [55] for the inverse of the sum of an invertible matrix and a rank-k correction, the inverse of (17) is given as

$$H_{k+1} = H_k - \frac{s_k y_k^T H_k + H_k y_k s_k^T}{s_k^T y_k} + \left(1 + \frac{y_k H_k y_k}{s_k^T y_k}\right) \frac{s_k s_k^T}{s_k^T y_k}, \quad s_k \in \mathbb{R}^n, \quad y_k \in \mathbb{R}^n.$$

To avoid computing and storing the  $n \times n$  matrix  $H_k$  at each iteration, it is replaced by the identity matrix  $I$ , and the so called memoryless update is obtained, that is,

$$H_{k+1} = I - \frac{s_k y_k^T + y_k s_k^T}{s_k^T y_k} + \left(1 + \frac{\|y_k\|^2}{s_k^T y_k}\right) \frac{s_k s_k^T}{s_k^T y_k}, \quad s_k \in \mathbb{R}^n, \quad y_k \in \mathbb{R}^n. \quad (18)$$

As mentioned earlier, the BFGS method implemented with (17) is the most popular and effective quasi-Newton scheme available. The method is guaranteed to satisfy the descent condition (9), since the update (17) satisfies the much required quasi-Newton condition. Other attributes of the BFGS scheme include its correction of eigenvalues mechanism [43]. However, the BFGS's efficiency depends strongly on the structure of eigenvalues of (17) [8]. Powell [52] and Byrd et al. [16] noted that the update (17) better corrects its small eigenvalues than large ones. Also, numerical experiments conducted by Gill and Leonard [29] showed that it is possible for the update (17) to require many iterations or gradient and function evaluations for some problems. The authors in [29] showed that these shortcomings of the BFGS method may result from poor initial Hessian approximations or its ill-conditioning along the

iterations. To overcome these shortfalls of the scheme, a number of scaling techniques have been applied to the BFGS update matrix in (17). This includes the modification by Biggs [13], where the update's third term in (17) was scaled by a positive parameter  $\gamma_k$  to yield

$$B_{k+1} = B_k - \frac{B_k s_k s_k^T B_k}{s_k^T B_k s_k} + \gamma_k \frac{y_k y_k^T}{y_k^T s_k}, \quad s_k \in \mathbb{R}^n, \quad y_k \in \mathbb{R}^n.$$

In Oren and Luenberger [45], the first and second terms of the matrix in (17) were scaled and the resulting modification becomes

$$B_{k+1} = \delta_k \left[ B_k - \frac{B_k s_k s_k^T B_k}{s_k^T B_k s_k} \right] + \frac{y_k y_k^T}{y_k^T s_k}, \quad s_k \in \mathbb{R}^n, \quad y_k \in \mathbb{R}^n,$$

where  $\delta_k > 0$ . Motivated by the strategy of changing structure of eigenvalues [43], Andrei [10] provided a two-parameter scaling BFGS method, where  $B_{k+1}$  is given by

$$B_{k+1} = \delta_k \left[ B_k - \frac{B_k s_k s_k^T B_k}{s_k^T B_k s_k} \right] + \gamma_k \frac{y_k y_k^T}{y_k^T s_k}, \quad s_k \in \mathbb{R}^n, \quad y_k \in \mathbb{R}^n,$$

with  $\gamma_k > 0$  and  $\delta_k > 0$ . In this update,  $\delta_k$  is obtained such that eigenvalues of  $B_{k+1}$  are clustered, while  $\gamma_k$  is computed to have a shift of the eigenvalues to the left. The latter procedure produces a better distribution of the eigenvalues. In other developments, the update matrix defined by (18) has also been modified in order to better distribute the eigenvalues and improve performance of the scheme. To that end, the following self scaled memoryless approximation to the Hessian inverse (18) was presented in [44]

$$H_{k+1} = \theta_k I - \theta_k \frac{s_k y_k^T + y_k s_k^T}{s_k^T y_k} + \left( 1 + \theta_k \frac{\|y_k\|^2}{s_k^T y_k} \right) \frac{s_k s_k^T}{s_k^T y_k}, \quad s_k \in \mathbb{R}^n, \quad y_k \in \mathbb{R}^n, \quad (19)$$

with  $\theta_k$  known as scaling parameter. In line with (19), Babaie-Kafaki [12] proposed the following extension:

$$H_{k+1} = \theta_k I - \theta_k \frac{s_k y_k^T + y_k s_k^T}{s_k^T y_k} + \left( 1 + \gamma_k \frac{\|y_k\|^2}{s_k^T y_k} \right) \frac{s_k s_k^T}{s_k^T y_k}, \quad s_k \in \mathbb{R}^n, \quad y_k \in \mathbb{R}^n, \quad (20)$$

where  $\gamma_k$  and  $\theta_k$  represents positive parameters. Analysis of the scheme obtained with (20) proves that it satisfies (10) and its condition number

remains in an improved condition. A modification of (18) was proposed in [11], namely,

$$H_{k+1} = \frac{1}{\delta_k} \left[ H_k - \frac{H_k y_k s_k^T + s_k y_k^T H_k}{s_k^T y_k} + \left( \frac{\delta_k}{\gamma_k} + \frac{y_k^T H_k y_k}{s_k^T y_k} \right) \frac{s_k s_k^T}{s_k^T y_k} \right],$$

where  $\delta_k$  and  $\gamma_k$  are parameters determined by employing Byrd and Nocedal's measure function in [17]. Now, as stated by Andrei [8], to achieve faster convergence of linear CG methods, the following approaches are employed:

- Clustering eigenvalues of a search direction matrix about a point [9, 60] or about several points [37] in its spectrum.
- Preconditioning of a search direction matrix [35].

Before we proceed to formulate our scheme, we first give the following additional assumptions on the mapping  $F$ :

**Assumption 1.** The solution set  $\bar{\mathcal{C}}$  of (3) is not empty, that is, there exists  $\bar{x} \in \mathcal{C}$  satisfying (3).

**Assumption 2.**  $F$  is Lipschitz continuous, that is,

$$\|F(x) - F(y)\| \leq L\|x - y\|, \quad \text{for all } x, y \in \mathcal{C}, \quad L \text{ a positive constant.} \quad (21)$$

Now, motivated by the shortcomings of the DK-type methods in [2, 22, 56], the scaled double parameter BFGS approximation to the inverse Hessian (20) as well as the need to explore other more effective approximations of the DK parameter, that ensures (10) holds without any adjustment, we propose the following DK-type search direction:

$$d_{k+1} = -\gamma F_{k+1} + \gamma \beta_k^{NHS} d_k - \left( \tau_k + \gamma \frac{\|\bar{y}_k\|^2}{s_k^T \bar{y}_k} - \gamma \frac{s_k^T \bar{y}_k}{\|s_k\|^2} \right) \frac{F_{k+1}^T s_k}{d_k^T \bar{y}_k} d_k, \quad d_0 = -F_0, \quad (22)$$

where

$$\beta_k^{NHS} = \frac{F_{k+1}^T \bar{y}_k}{d_k^T \bar{y}_k}, \quad k = 0, 1, \dots, \quad (23)$$

with

$$\bar{y}_k = y_k + r s_k, \quad y_k = F(w_k) - F(x_k), \quad r > 0, \quad (24)$$

and

$$w_k = x_k + \vartheta_k d_k, \quad s_k = w_k - x_k.$$

From (24) and (2), we have

$$d_k^T \bar{y}_k = \frac{s_k^T y_k}{\vartheta_k} + \frac{r}{\vartheta_k} \|s_k\|^2 \geq \frac{r}{\vartheta_k} \|s_k\|^2 > 0,$$

from which we obtain

$$s_k^T \bar{y}_k = s_k^T y_k + r \|s_k\|^2 \geq r \|s_k\|^2 > 0. \quad (25)$$

Note that the search direction defined by (22) can be written in compact form as

$$d_{k+1} = -M_{k+1} F_{k+1},$$

where

$$M_{k+1} = \gamma I - \gamma \frac{s_k \bar{y}_k^T}{s_k^T \bar{y}_k} + \tau_k \frac{s_k s_k^T}{s_k^T \bar{y}_k} + \gamma \frac{\|\bar{y}_k\|^2 s_k s_k^T}{(s_k^T \bar{y}_k)^2} - \gamma \frac{s_k s_k^T}{\|s_k\|^2}. \quad (26)$$

To proceed, we add rank-one update to (26) to obtain its symmetric form as

$$\bar{M}_{k+1} = \gamma I - \gamma \frac{s_k \bar{y}_k^T}{s_k^T \bar{y}_k} - \gamma \frac{\bar{y}_k s_k^T}{s_k^T \bar{y}_k} + \tau_k \frac{s_k s_k^T}{s_k^T \bar{y}_k} + \gamma \frac{\|\bar{y}_k\|^2 s_k s_k^T}{(s_k^T \bar{y}_k)^2} - \gamma \frac{s_k s_k^T}{\|s_k\|^2}. \quad (27)$$

Better still, we can re-write (27) as

$$\bar{M}_{k+1} = \gamma I Q_{k+1}, \quad (28)$$

in which

$$Q_{k+1} = I - \frac{s_k \bar{y}_k^T}{s_k^T \bar{y}_k} - \frac{\bar{y}_k s_k^T}{s_k^T \bar{y}_k} + \tau_k \frac{s_k s_k^T}{\gamma s_k^T \bar{y}_k} + \frac{\|\bar{y}_k\|^2 s_k s_k^T}{(s_k^T \bar{y}_k)^2} - \frac{s_k s_k^T}{\|s_k\|^2}. \quad (29)$$

We can further express (29) as the rank-two update

$$Q_{k+1} = I - \frac{s_k \bar{y}_k^T}{s_k^T \bar{y}_k} + \frac{(\tau_k \|s_k\|^2 (s_k^T \bar{y}_k) s_k - \gamma \|s_k\|^2 (s_k^T \bar{y}_k) \bar{y}_k + \gamma \|s_k\|^2 \|\bar{y}_k\|^2 s_k - \gamma (s_k^T \bar{y}_k)^2 s_k) s_k^T}{\gamma \|s_k\|^2 (s_k^T \bar{y}_k)^2} \quad (30)$$

Now, since from (25)  $s_k^T \bar{y}_k > 0$ , then  $s_k \neq 0$  and  $\bar{y}_k \neq 0$ . Suppose  $\mathcal{V} = \text{span}\{s_k, \bar{y}_k\}$ . Then  $\dim(\mathcal{V}) \leq 2$  and  $\dim(\mathcal{V}^\perp) \geq n - 2$ , with  $\mathcal{V}^\perp$  being

orthogonal complement of  $\mathcal{V}$ . So, there exists a set of mutually orthogonal vectors  $\{\xi_k^i\}_{i=1}^{n-2} \subset \mathcal{V}^\perp$  such that

$$s_k^T \xi_k^i = \bar{y}_k^T \xi_k^i = 0, \quad i = 1, \dots, n-2,$$

for which we obtain

$$\bar{M}_{k+1} \xi_k^i = \bar{M}_{k+1}^T \xi_k^i = \gamma \xi_k^i, \quad i = 1, \dots, n-2.$$

Therefore,  $\bar{M}_{k+1}$  contains  $n-2$  eigenvalues equal to  $\gamma$  each. We now find the remaining two eigenvalues, which we label as,  $\lambda_k^+$  and  $\lambda_k^-$ .

By applying the fundamental formula of algebra (see [55, inequality (1.2.70)]) for determinant of a rank-two update, namely,

$$\det(I + v_1 v_2^T + v_3 v_4^T) = (1 + v_1^T v_2)(1 + v_3^T v_4) - (v_1^T v_4)(v_2^T v_3), \quad v_1, v_2, v_3, v_4 \in \mathbb{R}^n,$$

and setting  $v_1 = -\frac{s_k}{s_k^T \bar{y}_k}$ ,  $v_2 = \bar{y}_k$ ,

$$v_3 = \frac{(\|s_k\|^2 (s_k^T \bar{y}_k)^2 \tau_k s_k - \gamma \|s_k\|^2 (s_k^T \bar{y}_k) \bar{y}_k + \gamma \|s_k\|^2 \|\bar{y}_k\|^2 s_k - \gamma (s_k^T \bar{y}_k)^2 s_k)}{\gamma \|s_k\|^2 (s_k^T \bar{y}_k)^2}, \text{ and } v_4 = s_k,$$

we get

$$\det(Q_{k+1}) = \tau_k \frac{\|s_k\|^2}{\gamma s_k^T \bar{y}_k} - 1. \quad (31)$$

Note that the matrix  $\bar{M}_{k+1}$  as defined in (28) is the product of two matrices, and

$$\det(\gamma I) = \gamma^n.$$

Combining this result with (31), we obtain

$$\begin{aligned} \det(\bar{M}_{k+1}) &= \gamma^n \left( \tau_k \frac{\|s_k\|^2}{\gamma s_k^T \bar{y}_k} - 1 \right) \\ &= \gamma^{n-2} \cdot \lambda^+ \lambda^-, \end{aligned}$$

which yields

$$\lambda^+ \lambda^- = \gamma^2 \left( \tau_k \frac{\|s_k\|^2}{\gamma s_k^T \bar{y}_k} - 1 \right) = \gamma \tau_k \frac{\|s_k\|^2}{s_k^T \bar{y}_k} - \gamma^2.$$

Since trace of the symmetric matrix  $\bar{M}_{k+1}$  is the summation of all its eigenvalues, we have

$$\begin{aligned}\text{tr}(\bar{M}_{k+1}) &= n\gamma - 2\gamma + \tau_k \frac{\|s_k\|^2}{s_k^T \bar{y}_k} + \gamma \frac{\|\bar{y}_k\|^2 \|s_k\|^2}{(s_k^T \bar{y}_k)^2} - \gamma \\ &= \underbrace{\gamma + \cdots + \gamma}_{(n-2) \text{ times}} + \lambda_k^+ + \lambda_k^-, \end{aligned}$$

which further yields

$$\lambda_k^+ + \lambda_k^- = \tau_k \frac{\|s_k\|^2}{s_k^T \bar{y}_k} + \gamma \frac{\|\bar{y}_k\|^2 \|s_k\|^2}{(s_k^T \bar{y}_k)^2} - \gamma. \quad (32)$$

From (32) and (31), the remaining eigenvalues of  $\bar{M}_{k+1}$  are obtained as solution of the following quadratic polynomial:

$$\lambda^2 - \left( \tau_k \frac{\|s_k\|^2}{s_k^T \bar{y}_k} + \gamma \frac{\|\bar{y}_k\|^2 \|s_k\|^2}{(s_k^T \bar{y}_k)^2} - \gamma \right) \lambda + \gamma \tau_k \frac{\|s_k\|^2}{s_k^T \bar{y}_k} - \gamma^2.$$

Consequently, by setting  $\Phi_k = \frac{\|s_k\|^2}{s_k^T \bar{y}_k}$ ,  $\mu_k = \frac{\|s_k\|^2 \|\bar{y}_k\|^2}{(s_k^T \bar{y}_k)^2}$ ,  $\lambda_k^+$  and  $\lambda_k^-$  are determined by

$$\lambda_k^\pm = \frac{\tau_k \Phi_k + \gamma \mu_k - \gamma \pm \sqrt{(\tau_k \Phi_k + \gamma \mu_k - \gamma)^2 - 4(\gamma \tau_k \Phi_k - \gamma^2)}}{2},$$

or more precisely,

$$\lambda_k^\pm = \frac{\tau_k \Phi_k + \gamma \mu_k - \gamma \pm \sqrt{(\tau_k \Phi_k + \gamma \mu_k - 3\gamma)^2 + 4\gamma^2 \mu_k - 4\gamma^2}}{2}. \quad (33)$$

Clearly, by the Cauchy-Schwarz inequality in (33),  $\lambda_k^+ > 0$ . Also,  $\lambda_k^- > 0$  whenever

$$\tau_k > \frac{\gamma}{\Phi_k} = \frac{\gamma s_k^T \bar{y}_k}{\|s_k\|^2}. \quad (34)$$

Now, we proceed to obtain an approximation of  $\tau_k$  such that (34) is satisfied making  $\bar{M}_{k+1}$  a positive-definite matrix. To achieve this, we employ the clustering of eigenvalues technique. Suppose that  $\lambda_k^+$  and  $\lambda_k^-$  have the same values as the first  $(n-2)$  eigenvalues of  $\bar{M}_{k+1}$ , namely,  $\lambda_k^+ = \lambda_k^- = \gamma$ . Then from determinant of  $\bar{M}_{k+1}$  obtained in (31), we have

$$\tau_k \frac{\|s_k\|^2}{\gamma s_k^T \bar{y}_k} - 1 = 1,$$

which implies that

$$\tau_k = 2 \frac{\gamma s_k^T \bar{y}_k}{\|s_k\|^2}, \quad (35)$$

which clearly satisfies (34) and ensures that all the eigenvalues of  $\bar{M}_{k+1}$  are clustered.

**Lemma 1.** The search direction sequence  $\{d_k\}$  obtained by (22) with (23), (24) and  $\gamma \in (0, 1]$  satisfy the inequality

$$d_{k+1}^T F_{k+1} \leq -c \|F_{k+1}\|^2, \quad (36)$$

where  $c = \frac{3\gamma}{4}$ .

*Proof.* From (22), (35), and by setting  $\Gamma_k = s_k^T \bar{y}_k$  for convenience, we have

$$\begin{aligned} d_{k+1}^T F_{k+1} &= -\gamma \|F_{k+1}\|^2 + \gamma \frac{F_{k+1}^T \bar{y}_k}{\Gamma_k} F_{k+1}^T s_k \\ &\quad - \left( \tau_k + \gamma \frac{\|\bar{y}_k\|^2}{\Gamma_k} - \gamma \frac{\Gamma_k}{\|s_k\|^2} \right) \frac{(F_{k+1}^T s_k)^2}{\Gamma_k} \\ &= -\gamma \|F_{k+1}\|^2 + \gamma \frac{F_{k+1}^T \bar{y}_k}{\Gamma_k} F_{k+1}^T s_k - \left( \gamma \frac{\Gamma_k}{\|s_k\|^2} + \gamma \frac{\|\bar{y}_k\|^2}{\Gamma_k} \right) \frac{(F_{k+1}^T s_k)^2}{\Gamma_k} \\ &\leq -\gamma \|F_{k+1}\|^2 + \gamma \frac{F_{k+1}^T \bar{y}_k}{\Gamma_k} F_{k+1}^T s_k - \gamma \frac{\|\bar{y}_k\|^2}{\Gamma_k^2} (F_{k+1}^T s_k)^2 \\ &= \frac{\gamma F_{k+1}^T \bar{y}_k \Gamma_k F_{k+1}^T s_k - \gamma \Gamma_k^2 \|F_{k+1}\|^2 - \gamma \|\bar{y}_k\|^2 (F_{k+1}^T s_k)^2}{\Gamma_k^2} \\ &\leq \frac{\gamma \frac{\Gamma_k^2 \|F_{k+1}\|^2}{4} + \gamma \|\bar{y}_k\|^2 (F_{k+1}^T s_k)^2 - \gamma \Gamma_k^2 \|F_{k+1}\|^2 - \gamma \|\bar{y}_k\|^2 (F_{k+1}^T s_k)^2}{\Gamma_k^2} \\ &= \gamma \frac{\|F_{k+1}\|^2}{4} - \gamma \|F_{k+1}\|^2 \\ &= -\gamma \left( 1 - \frac{1}{4} \right) \|F_{k+1}\|^2 \\ &= -\frac{3\gamma}{4} \|F_{k+1}\|^2. \end{aligned}$$

We arrived at the last inequality by employing the identity

$$2c_1^T c_2 \leq \|c_1\|^2 + \|c_2\|^2, \quad c_1, c_2 \in \mathbb{R}^n,$$

with  $c_1 = \frac{\Gamma_k F_{k+1}}{\sqrt{2}}$ ,  $c_2 = \sqrt{2}(F_{k+1}^T s_k) \bar{y}_k$ . Hence, setting  $c = \frac{3\gamma}{4}$ , we see that (36) holds.  $\square$

Next, we introduce the projection operator defined by

$$\mathcal{P}_{\mathcal{C}}(x) = \arg \min \|x - y\| : y \in \mathcal{C}, \quad \text{for all } x \in \mathbb{R}^n,$$

with the properties:

$$\|\mathcal{P}_{\mathcal{C}}(x) - \mathcal{P}_{\mathcal{C}}(y)\| \leq \|x - y\|, \quad \text{for all } x, y \in \mathbb{R}^n,$$

and

$$\|\mathcal{P}_{\mathcal{C}}(x) - y\| \leq \|x - y\|, \quad \text{for all } y \in \mathcal{C}, \quad (37)$$

where  $\mathcal{C}$  is as defined earlier.

### Algorithm 1

**Data:** Select  $\epsilon > 0$ ,  $x_0 \in \mathcal{C}$ ,  $\beta \in (0, 1)$ ,  $\delta \in (0, 1)$ ,  $0 < \phi < 2$ ,  $r > 0$ ,  $\gamma \in (0, 1]$ .

**Initialization:** Set  $k = 0$  and  $d_0 = -F_0$ .

- 1: Obtain  $F(x_k)$  and confirm if  $\|F(x_k)\| \leq \epsilon$ . End if yes, otherwise goto 2.
- 2: Determine  $w_k = x_k + \vartheta_k d_k$ , where  $\vartheta_k = \beta^{m_k}$ , with  $m$  being the smallest nonnegative integer for which

$$-F(x_k + \beta^m d_k)^T d_k \geq \delta \beta^m \|d_k\|^2 \quad (38)$$

holds.

- 3: If  $w_k \in \mathcal{C}$  and  $\|F(w_k)\| \leq \epsilon$ , end, otherwise, compute

$$x_{k+1} = \mathcal{P}_{\mathcal{C}}[x_k - \phi \rho_k F(w_k)], \quad \text{where} \quad (39)$$

$$\rho_k = \frac{F(w_k)^T (x_k - w_k)}{\|F(w_k)\|^2}. \quad (40)$$

- 4: Obtain  $d_{k+1}$  by (22) with (23), (24), and (35).

- 5: Set  $k = k + 1$  and proceed to 1.

## 3 Convergence report

First, we show that  $\tau_k$  obtained in (35) is bounded.

From (21), (25), (35) and the Cauchy Schwarz inequality, we have

$$\begin{aligned}
|\tau_k| &\leq \frac{2\|s_k\|\|\bar{y}_k\|}{\|s_k\|^2} \\
&\leq \frac{2L\|s_k\|^2}{\|s_k\|^2} \\
&= 2L \stackrel{\text{def}}{=} \bar{m}.
\end{aligned} \tag{41}$$

**Lemma 2.** The sequence  $\{d_{k+1}\}$  of directions obtained by Algorithm 1 satisfy

$$c\|F_{k+1}\| \leq \|d_{k+1}\| \leq \left( \gamma + \frac{2\gamma L}{r} + \frac{\bar{m}}{r} + \frac{\gamma L^2}{r^2} \right) \|F_{k+1}\|, \tag{42}$$

where  $\gamma \in (0, 1]$ ,  $r > 0$ , and  $L > 0$ .

*Proof.* The first inequality follows from the Cauchy–Schwarz inequality and (22). For  $k = 0$  in (22), we have that  $d_0 = -F_0$ , which indicates that  $\|d_0\| = \|F_0\|$ . Now, we show that the inequality holds for  $k \geq 1$ . From the Cauchy–Schwarz inequality, (21), (22), (25), and (41), we obtain

$$\begin{aligned}
\|d_{k+1}\| &= \left\| -\gamma F_{k+1} + \gamma \frac{F_{k+1}^T \bar{y}_k}{s_k^T \bar{y}_k} s_k - \left( \tau_k + \gamma \frac{\|\bar{y}_k\|^2}{s_k^T \bar{y}_k} - \gamma \frac{s_k^T \bar{y}_k}{\|s_k\|^2} \right) \frac{F_{k+1}^T s_k}{s_k^T \bar{y}_k} s_k \right\| \\
&\leq \gamma \|F_{k+1}\| + \gamma \frac{\|F_{k+1}\| \|\bar{y}_k\| \|s_k\|}{s_k^T \bar{y}_k} + |\tau_k| \frac{\|F_{k+1}\| \|s_k\|^2}{s_k^T \bar{y}_k} \\
&\quad + \gamma \frac{\|F_{k+1}\| \|\bar{y}_k\|^2 \|s_k\|^2}{(s_k^T \bar{y}_k)^2} + \gamma \frac{\|F_{k+1}\| \|s_k\|^3 \|\bar{y}_k\|}{\|s_k\|^2 s_k^T \bar{y}_k} \\
&\leq \gamma \|F_{k+1}\| + \gamma \frac{L \|F_{k+1}\| \|s_k\|^2}{r \|s_k\|^2} + \bar{m} \frac{\|F_{k+1}\| \|s_k\|^2}{r \|s_k\|^2} + \gamma \frac{L^2 \|F_{k+1}\| \|s_k\|^4}{r^2 \|s_k\|^4} \\
&\quad + \gamma \frac{L \|F_{k+1}\| \|s_k\|^4}{r \|s_k\|^4} \\
&= \gamma \|F_{k+1}\| + \gamma \frac{L \|F_{k+1}\|}{r} + \bar{m} \frac{\|F_{k+1}\|}{r} + \gamma \frac{L^2 \|F_{k+1}\|}{r^2} + \gamma \frac{L \|F_{k+1}\|}{r} \\
&= \gamma \|F_{k+1}\| + 2\gamma \frac{L \|F_{k+1}\|}{r} + \bar{m} \frac{\|F_{k+1}\|}{r} + \gamma \frac{L^2 \|F_{k+1}\|}{r^2} \\
&= \left( \gamma + \frac{2\gamma L}{r} + \frac{\bar{m}}{r} + \frac{\gamma L^2}{r^2} \right) \|F_{k+1}\|,
\end{aligned} \tag{43}$$

which proves the second inequality of (42).  $\square$

Next, we prove that the line search (38) is well defined and also terminates after finite iterations:

**Lemma 3.** Let Assumption 2 hold, and suppose that Algorithm 1 is not terminated in step 1. Then there exists a nonnegative integer  $m_k$  such that (38) is satisfied. In addition, the step-size  $\vartheta_k$  obtained in (38) satisfies

$$\vartheta_k \geq \vartheta := \min \left\{ 1, \frac{3\gamma\beta}{4(L+\delta) \left( \gamma + \frac{2\gamma L}{r} + \frac{\bar{m}}{r} + \frac{\gamma L^2}{r^2} \right)^2} \right\}. \quad (44)$$

*Proof.* To show the first part, we assume that there exists  $k_0 \geq 0$  such that (38) is not true in the  $k_0^{th}$  iterate for each value of  $m$ . So, for all  $m \geq 0$ , we have

$$-F(x_{k_0} + \beta^m d_{k_0})^T d_{k_0} < \delta \beta^m \|d_{k_0}\|^2. \quad (45)$$

Since  $F$  is continuous on  $\mathbb{R}^n$ , applying limit to (45) as  $m$  grows to infinity, yields

$$F(x_{k_0})^T d_{k_0} > 0,$$

which is contradicted by (36), namely,

$$F(x_{k_0})^T d_{k_0} \leq -\frac{3\gamma}{4} \|F(x_{k_0})\|^2.$$

Thus, we proved the first part.

Now, suppose that the algorithm is terminated at  $x_k$ , then  $F(x_k) = 0$  or  $F(w_k) = 0$ . This indicates the solution to be  $x_k$ , otherwise  $x_k$  is not a solution. Then, from (36)  $d_k \neq 0$ . Now, from (38) we see that if  $\vartheta_k \neq 1$ , then  $\bar{\vartheta}_k = \beta^{-1}\vartheta_k$  will not satisfy (38), that is,

$$-F(\bar{w}_k)^T d_k < \delta \bar{\vartheta}_k \|d_k\|^2,$$

where,  $\bar{w}_k = x_k + \bar{\vartheta}_k d_k$ . By Assumption 2 and (36), we have

$$\begin{aligned} \frac{3\gamma}{4} \|F_k\|^2 &\leq -F_k^T d_k \\ &= (F(\bar{w}_k) - F_k)^T d_k - F(\bar{w}_k)^T d_k \\ &\leq L\bar{\vartheta}_k \|d_k\|^2 + \delta \bar{\vartheta}_k \|d_k\|^2 \\ &= \beta^{-1}\vartheta_k (L + \delta) \|d_k\|^2. \end{aligned}$$

Hence, we obtain

$$\begin{aligned}
\vartheta_k &\geq \frac{3\gamma\beta}{4(L+\delta)} \frac{\|F_k\|^2}{\|d_k\|^2} \\
&\geq \frac{3\gamma\beta}{4(L+\delta)} \frac{\|F_k\|^2}{\left(\gamma + \frac{2\gamma L}{r} + \frac{\bar{m}}{r} + \frac{\gamma L^2}{r^2}\right)^2 \|F_k\|^2} \\
&= \frac{3\gamma\beta}{4(L+\delta) \left(\gamma + \frac{2\gamma L}{r} + \frac{\bar{m}}{r} + \frac{\gamma L^2}{r^2}\right)^2},
\end{aligned}$$

where (43) was used to obtain the second inequality.  $\square$

**Lemma 4.** Let Assumptions 1, and 2 hold. Then for a solution  $\bar{x}$  of (3) in  $\bar{\mathcal{C}}$ , the sequence  $\{\|x_k - \bar{x}\|\}$  is convergent implying that  $\{x_k\}$  is bounded. Also

$$\lim_{k \rightarrow \infty} \vartheta_k \|d_k\| = 0. \quad (46)$$

*Proof.* From (38) and definition of  $w_k$ , we have

$$(x_k - w_k)^T F(w_k) \geq \delta \vartheta_k^2 \|d_k\|^2. \quad (47)$$

By (2) and for all  $\bar{x} \in \bar{\mathcal{C}}$ , we have

$$\begin{aligned}
(x_k - \bar{x})^T F(w_k) &= (x_k - w_k)^T F(w_k) + (w_k - \bar{x})^T F(w_k) \\
&\geq (x_k - w_k)^T F(w_k) + (w_k - \bar{x})^T F(\bar{x}) \\
&= (x_k - w_k)^T F(w_k).
\end{aligned} \quad (48)$$

From (37), (39), (40), (47) and (48), we have

$$\begin{aligned}
\|x_{k+1} - \bar{x}\|^2 &= \|\mathcal{P}_{\mathcal{C}}[x_k - \phi \rho_k F(w_k)] - \bar{x}\|^2 \\
&\leq \|x_k - \phi \rho_k F(w_k) - \bar{x}\|^2 \\
&= \|(x_k - \bar{x}) - \phi \rho_k F(w_k)\|^2 \\
&= \|x_k - \bar{x}\|^2 - 2\phi \rho_k F(w_k)^T (x_k - \bar{x}) + \phi^2 \rho_k^2 \|F(w_k)\|^2 \\
&\leq \|x_k - \bar{x}\|^2 - 2\phi \rho_k F(w_k)^T (x_k - w_k) + \phi^2 \rho_k^2 \|F(w_k)\|^2 \\
&= \|x_k - \bar{x}\|^2 - \phi(2 - \phi) \frac{(F(w_k)^T (x_k - w_k))^2}{\|F(w_k)\|^2} \\
&\leq \|x_k - \bar{x}\|^2 - \phi(2 - \phi) \frac{\delta^2 \|x_k - w_k\|^4}{\|F(w_k)\|^2},
\end{aligned} \quad (49)$$

which yields

$$0 \leq \|x_{k+1} - \bar{x}\| \leq \|x_k - \bar{x}\| \leq \|x_{k-1} - \bar{x}\| \leq \cdots \leq \|x_0 - \bar{x}\|.$$

So,  $\{\|x_k - \bar{x}\|\}$  is non-increasing and bounded, which indicates that  $\{x_k\}$  is bounded also. This with the fact that  $F$  is Lipschitz continuous implies that a constant  $m_1$  exists for all  $k \geq 0$  such that,

$$\|x_k\| \leq m_1, \quad \|F(x_k)\| \leq m_1. \quad (50)$$

Also, by (43) and (50) a constant  $m_2$  exists for which

$$\|d_k\| \leq \left( \gamma + \frac{2\gamma L}{r} + \frac{\bar{m}}{r} + \frac{\gamma L^2}{r^2} \right) m_1.$$

Setting  $m_2 = \left( \gamma + \frac{2\gamma L}{r} + \frac{\bar{m}}{r} + \frac{\gamma L^2}{r^2} \right) m_1$ , we obtain that  $d_k$  is bounded.

Furthermore, from (50), monotonicity of  $F$ , the Cauchy-Schwarz inequality, and (47), we have

$$m_1 \geq \|F_k\| \geq \frac{F_k^T(x_k - w_k)}{\|x_k - w_k\|} \geq \frac{F(w_k)^T(x_k - w_k)}{\|x_k - w_k\|} \geq \delta \|x_k - w_k\| \geq \delta \|w_k\| - \delta m_1,$$

which consequently implies that

$$\|w_k\| \leq \frac{m_1 + \delta m_1}{\delta}.$$

By setting  $m_3 := \frac{m_1 + \delta m_1}{\delta}$ , we establish boundedness of  $\{w_k\}$ . Hence, from continuity of  $F$ , a constant  $\bar{m}$  exists such that

$$\|F(w_k)\| \leq \bar{m}, \quad \text{for all } k \geq 0.$$

Combining this with (49), we obtain

$$\delta^2 \|x_k - w_k\|^4 \leq \frac{\bar{m}^2}{\phi(2 - \phi)} (\|x_k - \bar{x}\|^2 - \|x_{k+1} - \bar{x}\|^2). \quad (51)$$

Now, following the convergence of  $\{\|x_k - \bar{x}\|\}$  and boundedness of  $\{F(w_k)\}$ , we take limit as  $k$  approaches infinity in (51) to obtain

$$\delta^2 \lim_{k \rightarrow \infty} \vartheta_k^4 \|d_k\|^4 \leq 0,$$

which indicates that

$$\lim_{k \rightarrow \infty} \vartheta_k \|d_k\| = 0.$$

□

**Theorem 1.** Suppose that Assumptions 1 and 2 hold and that  $\{x_k\}$  is obtained by Algorithm 2.1. Then,  $\{x_k\}$  converges to a solution of (3).

*Proof.* Firstly, from (44) and (46), we have that  $0 \leq \vartheta \|d_k\| \leq \vartheta_k \|d_k\| \rightarrow 0$ , which consequently indicates that  $\lim_{k \rightarrow \infty} \|d_k\| = 0$ . This together with (42) yields

$$0 \leq \frac{3}{4\gamma} \|F_k\| \leq \|d_k\| \rightarrow 0,$$

which indicates that  $\lim_{k \rightarrow \infty} \|F_k\| = 0$ . Now, inequality (46) and the boundedness of the sequence  $\{x_k\}$  indicates the existence of a cluster point of  $\{x_k\}$  say  $\tilde{x} \in \bar{\mathcal{C}}$ , where  $\bar{\mathcal{C}}$  denotes solution set of  $F$ . Let  $\mathcal{K} \subseteq \{0, 1, 2, \dots\}$  be an infinite index set for which

$$\lim_{k \rightarrow \infty, k \in \mathcal{K}} x_k = \tilde{x} \in \bar{\mathcal{C}}.$$

Since  $F$  is continuous, we have that

$$0 = \lim_{k \rightarrow \infty} \|F_k\| = \lim_{k \rightarrow \infty, k \in \mathcal{K}} \|F_k\| = \|F(\tilde{x})\|,$$

which indicates that  $\tilde{x}$  is a solution of (3). Also, since  $\{\|x_k - \bar{x}\|\}$  is convergent, setting  $\bar{x} = \tilde{x}$  yields

$$\lim_{k \rightarrow \infty} \|x_k - \bar{x}\| = \lim_{k \rightarrow \infty, k \in \mathcal{K}} \|x_k - \bar{x}\| = 0.$$

which, therefore, indicates that  $\{x_k\}$  converges to  $\bar{x} \in \bar{\mathcal{C}}$ . □

## 4 Results of numerical experiments

To test effectiveness of Algorithm 1, two experiments are conducted and discussed in the next two subsections.

#### 4.1 First experiment: Convex constrained nonlinear monotone systems

For these experiments, the performance of Algorithm 1 is tested against four recent methods for solving the constrained problem (3), namely, ACGD [22], MDKM [56], SCRME [28], and SDYCG [7]. Codes for the algorithms, which are available at <https://github.com/hungugida/hungugida/blob/main/MATLABcodeforconstrainedsystem.zip> was written in MATLAB R2014a and executed using a system configured as (2.30ghz cpu, 4gb RAM). The stoppage criteria for all runs are  $\|F(x_k)\| \leq 10^{-10}$  or  $\|F(w_k)\| \leq 10^{-10}$  or iterations exceed 1000. We set parameters of (38) for Algorithm 1 as  $\beta = 0.6$ ,  $\delta = 0.0001$ ,  $\gamma = 0.27$ ,  $\phi = 1.8$ ,  $r = 0.0001$ . The exact values of the parameters used in the articles for each of the four schemes were also applied here.

The underlisted test examples with dimensions 5000, 10000, and 50000 were used to test Algorithm 1, ACGD, MDKM, SCRME and SDYCG, where  $F$  is given as:  $F = (f_1(x), f_2(x), \dots, f_n(x))^T$ .

**Example 1.** [38] with  $\mathcal{C} = \mathbb{R}_+^n$  added to yield

$$f_i(x) = 2x_i - \sin x_i, \quad i = 1, 2, \dots, n.$$

**Example 2.** [40].

$$\begin{aligned} f_1(x) &= x_1 - \exp\left(\cos\left(\frac{x_1+x_2}{n+1}\right)\right), \\ f_i(x) &= x_i - \exp\left(\cos\left(\frac{x_{i-1}+x_i+x_{i+1}}{n+1}\right)\right), \quad i = 2, 3, \dots, n-1, \\ f_n(x) &= x_n - \exp\left(\cos\left(\frac{x_{n-1}+x_n}{n+1}\right)\right), \\ \text{with } \mathcal{C} &= \mathbb{R}_+^n. \end{aligned}$$

**Example 3.** [38]

$$f_i(x) = 2x_i - \sin |x_i|, \quad i = 1, 2, \dots, n,$$

where  $\mathcal{C} = \mathbb{R}_+^n$ .

**Example 4.** This is a modified version of the example in [39] with  $\mathcal{C} = \mathbb{R}_+^n$  added to yield

$$\begin{aligned} f_1(x) &= e^{\sin x_1} - 1, \\ f_i(x) &= e^{\sin x_i} + x_i - 1, \quad i = 2, \dots, n. \end{aligned}$$

**Example 5.** [64] with  $\mathcal{C} = \mathbb{R}_+^n$  added to yield

$$\begin{aligned} f_1(x) &= 2x_1 + \sin x_1 - 1, \\ f_i(x) &= 2x_{i-1} + 2x_i + 2\sin x_i - 1, \\ f_n(x) &= 2x_n + \sin x_n - 1, \quad i = 2, \dots, n-1. \end{aligned}$$

**Example 6.** This is a modification of test example 4

$$\begin{aligned} f_1(x) &= 3x_1 + e^{\sin x_1} - 1, \\ f_i(x) &= 3x_i + e^{\sin x_i} - 1, \quad i = 2, \dots, n, \\ \text{with } \mathcal{C} &= \mathbb{R}_+^n. \end{aligned}$$

**Example 7.** This is a modification of test example 5

$$\begin{aligned} f_1(x) &= 3x_1 + \cos x_1 - 1, \\ f_i(x) &= 3x_{i-1} + 3x_i + \cos x_i - 1, \\ f_n(x) &= 3x_n + \cos x_n - 1, \quad i = 2, \dots, n-1, \\ \text{with } \mathcal{C} &= \mathbb{R}_+^n. \end{aligned}$$

**Example 8.** Modification of test example 2

$$\begin{aligned} f_1(x) &= x_1 - e^{\left(\cos \frac{x_1+x_2}{2}\right)}, \\ f_i(x) &= x_i - e^{\left(\cos \frac{x_{i-1}+x_i+x_{i+1}}{i}\right)}, \quad i = 2, 3, \dots, n-1, \\ f_n(x) &= x_n - e^{\left(\cos \frac{x_{n-1}+x_n}{n}\right)}. \end{aligned}$$

where  $\mathcal{C} = \mathbb{R}_+^n$ .

The following initial guesses were used:

$$\begin{aligned} x_0^1 &= \left(1, \frac{1}{2}, \dots, \frac{1}{n}\right)^T, \quad x_0^2 = \left(\frac{1}{2}, \frac{3}{2}, \dots, -\frac{[(-1)^n-2]}{2}\right)^T, \quad x_0^3 = \left(1, 3, \dots, -\frac{-2[(-1)^n-2]}{2}\right)^T, \\ x_0^4 &= \left(\frac{n-1}{n}, \frac{n-2}{n}, \dots, 0\right)^T, \quad x_0^5 = \left(\frac{1}{4}, \frac{3}{4}, \dots, \frac{-[(-1)^n-2]}{4}\right)^T, \quad x_0^6 = \left(\frac{1}{n}, \frac{2}{n}, \dots, 1\right)^T. \end{aligned}$$

Table 1: Test results for Examples 1-2.

PN	VAR	SP	Algorithm 1			ACGD			MDKM			SRCME			SDYCG							
			NIT	FE	PT	Norm	NIT	FE	PT	Norm	NIT	FE	PT	Norm	NIT	FE	PT	Norm				
1	5000	$x_9^1$	1	3	0.0054	0	19	41	0.5232	7.42E-11	14	17	0.3490	9.47E-11	38	41	0.1306	9.66E-11	281	1277	1.4496	9.53E-11
	5000	$x_9^1$	1	3	0.0070	0	19	41	0.0632	7.42E-11	15	18	0.0519	1.01E-11	38	41	0.1088	9.65E-11	292	1175	1.2170	1.79E-11
	5000	$x_9^1$	1	4	0.0082	0	20	44	0.0755	7.45E-11	13	18	0.0464	3.37E-11	40	43	0.0970	7.81E-11	347	1593	1.6216	1.92E-11
	5000	$x_9^1$	1	3	0.0092	0	19	41	0.0647	7.42E-11	15	18	0.0444	1.01E-11	38	41	0.0926	9.65E-11	331	1275	1.3538	6.81E-12
	5000	$x_9^1$	1	3	0.0088	0	19	41	0.0668	6.50E-11	12	14	0.0485	1.58E-11	38	41	0.0905	8.95E-11	319	683	0.9334	9.97E-11
	5000	$x_9^1$	1	3	0.0067	0	19	43	0.0717	1.00E-10	13	20	0.0450	1.03E-11	42	44	0.0992	8.05E-11	242	511	0.6694	3.62E-11
	10000	$x_9^1$	1	3	0.0374	0	20	43	0.0929	2.73E-11	15	18	0.0703	1.37E-11	39	42	0.1401	6.84E-11	351	1528	2.3655	1.81E-11
	10000	$x_9^1$	1	3	0.0089	0	20	43	0.0998	2.73E-11	15	18	0.0614	1.42E-11	39	42	0.1427	6.84E-11	285	1187	1.9231	6.45E-11
	10000	$x_9^1$	1	4	0.0101	0	21	46	0.1042	2.75E-11	13	18	0.0598	4.77E-11	41	44	0.1461	5.54E-11	288	1113	1.8137	5.72E-11
	10000	$x_9^1$	1	3	0.0084	0	20	43	0.0973	2.73E-11	15	18	0.0794	1.42E-11	39	42	0.1578	6.84E-11	289	1269	1.9909	3.27E-11
	10000	$x_9^1$	1	3	0.0093	0	19	41	0.1470	9.20E-11	12	14	0.0577	2.23E-11	39	42	0.1445	6.33E-11	283	604	1.3013	1.46E-11
	10000	$x_9^1$	1	3	0.0068	0	20	45	0.1043	3.68E-11	13	20	0.0687	1.46E-11	42	45	0.1496	5.71E-11	240	507	1.0708	6.79E-11
	50000	$x_9^1$	1	3	0.3420	0	20	43	0.3388	6.12E-11	15	18	0.2442	3.12E-11	40	43	0.5318	7.66E-11	296	1369	9.3452	9.80E-11
	50000	$x_9^1$	1	3	0.0260	0	20	43	0.3510	6.12E-11	15	18	0.2356	3.14E-11	40	43	0.5466	7.66E-11	325	1634	10.8721	2.33E-11
	50000	$x_9^1$	1	4	0.0329	0	21	46	0.3673	6.16E-11	14	19	0.2720	1.07E-11	42	45	0.5699	6.22E-11	398	1565	11.1525	9.91E-11
	50000	$x_9^1$	1	3	0.0241	0	20	43	0.3459	5.15E-11	15	18	0.2455	3.14E-11	40	43	0.5385	7.66E-11	411	1712	12.1336	9.77E-11
	50000	$x_9^1$	1	3	0.0261	0	20	43	0.3859	5.35E-11	12	14	0.2001	4.99E-11	40	43	0.5689	7.08E-11	257	552	5.1273	6.16E-11
	50000	$x_9^1$	1	3	0.0233	0	20	45	0.3542	8.23E-11	13	20	0.2499	3.27E-11	44	46	0.6126	6.43E-11	246	519	4.8538	9.24E-11
2	5000	$x_9^2$	11	13	0.1418	1.47E-11	20	43	0.0904	8.19E-11	12	14	0.0497	1.58E-11	40	43	0.1520	8.63E-11	95	193	0.4535	9.42E-11
	5000	$x_9^2$	11	14	0.0487	4.79E-11	20	43	0.1007	8.19E-11	12	14	0.0491	1.58E-11	40	43	0.1289	8.63E-11	95	193	0.4406	9.42E-11
	5000	$x_9^2$	11	14	0.0439	7.30E-11	20	43	0.0921	4.59E-11	11	13	0.0448	8.85E-11	39	42	0.1374	9.66E-11	91	185	0.4374	8.79E-11
	5000	$x_9^2$	12	14	0.0417	2.69E-11	20	43	0.0920	8.19E-11	12	14	0.0515	1.58E-11	40	43	0.1390	8.63E-11	95	193	0.4177	9.42E-11
	5000	$x_9^2$	11	14	0.0474	4.91E-11	20	43	0.0891	8.18E-11	12	14	0.0531	1.58E-11	40	43	0.1374	8.61E-11	99	201	0.4301	8.55E-11
	5000	$x_9^2$	12	14	0.0486	2.69E-11	20	43	0.0959	4.51E-11	11	13	0.0607	8.71E-11	39	42	0.1393	9.50E-11	104	211	0.4846	8.90E-11
	10000	$x_9^2$	10	13	0.0954	9.21E-11	21	45	0.1570	3.01E-11	12	14	0.0859	2.24E-11	41	44	0.2455	6.10E-11	103	209	0.7820	8.88E-11
	10000	$x_9^2$	12	14	0.1105	7.48E-11	21	45	0.1631	3.01E-11	12	14	0.0780	2.24E-11	41	44	0.2176	6.10E-11	103	209	0.7706	8.88E-11
	10000	$x_9^2$	12	14	0.0731	1.03E-11	20	43	0.1622	6.49E-11	12	14	0.0953	1.25E-11	40	43	0.2300	6.83E-11	93	189	0.7331	9.06E-11
	10000	$x_9^2$	11	14	0.0672	6.70E-11	21	45	0.1729	3.01E-11	12	14	0.0799	2.24E-11	41	44	0.2297	6.10E-11	103	209	0.8000	8.88E-11
	10000	$x_9^2$	11	13	0.0703	3.31E-11	21	45	0.1597	3.01E-11	12	14	0.0797	2.23E-11	41	44	0.2452	6.09E-11	109	221	0.8264	9.79E-11
	10000	$x_9^2$	11	14	0.0856	6.70E-11	20	43	0.1617	6.38E-11	12	14	0.0761	1.23E-11	40	43	0.2246	6.72E-11	123	249	0.9785	7.83E-11
	50000	$x_9^2$	9	12	0.3199	8.35E-11	21	45	0.6183	6.74E-11	12	14	0.3359	5.00E-11	42	45	0.9522	6.82E-11	72	147	2.6030	8.18E-11
	50000	$x_9^2$	11	13	0.2748	8.54E-11	21	45	0.6229	6.74E-11	12	14	0.3279	5.00E-11	42	45	0.9546	6.82E-11	72	147	2.6052	8.18E-11
	50000	$x_9^2$	10	13	0.2791	6.32E-11	21	45	0.6132	3.77E-11	12	14	0.3226	2.80E-11	41	44	0.9294	7.64E-11	68	139	2.4719	7.67E-11
	50000	$x_9^2$	10	13	0.2708	6.50E-11	21	45	0.6050	6.74E-11	12	14	0.3376	5.00E-11	42	45	0.9661	6.82E-11	72	147	2.5842	8.18E-11
	50000	$x_9^2$	10	12	0.2800	4.03E-11	21	45	0.6152	6.73E-11	12	14	0.3191	4.99E-11	42	45	0.9721	6.81E-11	74	151	2.6862	9.09E-11
	50000	$x_9^2$	10	13	0.2787	6.50E-11	21	45	0.6079	3.71E-11	12	14	0.3260	2.75E-11	41	44	0.9639	7.51E-11	81	165	2.9192	7.70E-11

Table 2: Test results for Examples 3–4.

PN	VAR	SP	Algorithm 1			ACGD			MDKM			SRCME			SDYCG							
			NIT	FE	PT	Norm	NIT	FE	PT	Norm	NIT	FE	PT	Norm	NIT	FE	PT	Norm				
3	5000	$x_0^1$	1	3	0.0506	0	19	41	0.0575	7.42E-11	12	15	0.0406	1.16E-11	38	41	0.0979	9.66E-11	172	1027	9.9486	5.42E-11
	5000	$x_0^2$	1	3	0.0073	0	19	41	0.0630	7.42E-11	12	15	0.0359	1.16E-11	38	41	0.0998	9.65E-11	153	1041	0.9608	5.83E-11
	5000	$x_0^3$	1	4	0.0082	0	20	44	0.0650	7.45E-11	13	18	0.0387	3.42E-11	40	43	0.1023	7.81E-11	137	1009	0.8833	4.83E-11
	5000	$x_0^4$	1	3	0.0054	0	19	41	0.0718	7.42E-11	12	15	0.0463	1.16E-11	38	41	0.1005	9.87E-11	123	1017	0.8978	2.90E-11
	5000	$x_0^5$	1	3	0.0070	0	19	41	0.0641	6.50E-11	12	14	0.0433	1.60E-11	38	41	0.0914	8.95E-11	106	241	0.3715	8.02E-11
	5000	$x_0^6$	1	3	0.0068	0	19	43	0.0657	1.00E-10	13	20	0.0450	1.04E-11	43	46	0.1046	5.69E-11	33	153	0.1958	7.05E-11
	10000	$x_0^1$	1	3	0.0092	0	20	43	0.0971	2.73E-11	12	15	0.0568	1.65E-11	39	42	0.1443	6.84E-11	179	1020	1.5294	9.73E-11
	10000	$x_0^2$	1	3	0.0085	0	20	43	0.0946	2.73E-11	12	15	0.0553	1.65E-11	39	42	0.1432	6.84E-11	166	928	1.4229	6.49E-11
	10000	$x_0^3$	1	4	0.0108	0	21	46	0.1059	2.75E-11	13	18	0.0625	4.84E-11	41	44	0.1475	5.54E-11	118	918	1.2932	1.66E-11
	10000	$x_0^4$	1	3	0.0093	0	20	43	0.0979	2.73E-11	12	15	0.0640	1.65E-11	39	42	0.1534	6.84E-11	178	989	1.4766	8.44E-11
	10000	$x_0^5$	1	3	0.0091	0	19	41	0.0880	9.20E-11	12	14	0.0640	2.26E-11	39	42	0.1836	6.33E-11	110	249	0.5705	6.12E-11
	10000	$x_0^6$	1	3	0.0078	0	20	45	0.0987	3.68E-11	13	20	0.0682	1.46E-11	43	46	0.1671	8.04E-11	33	153	0.2766	9.37E-11
	50000	$x_0^1$	1	3	0.0249	0	20	43	0.3522	6.12E-11	12	15	0.2029	3.68E-11	40	43	0.5142	7.66E-11	216	1206	8.0604	6.36E-11
	50000	$x_0^2$	1	3	0.0259	0	20	43	0.3571	6.12E-11	12	15	0.2056	3.68E-11	40	43	0.5524	7.66E-11	149	1005	***	***
	50000	$x_0^3$	1	4	0.0326	0	21	46	0.3795	6.16E-11	14	19	0.2440	1.08E-11	42	45	0.5538	6.22E-11	***	***	6.4341	4.96E-11
50000	$x_0^4$	1	3	0.0259	0	20	43	0.3509	6.12E-11	12	15	0.2145	3.68E-11	40	43	0.5434	7.66E-11	210	1545	9.647	7.74E-11	
50000	$x_0^5$	1	3	0.0280	0	20	43	0.3539	5.35E-11	12	14	0.2240	5.05E-11	40	43	0.5291	7.90E-11	143	315	2.9196	8.83E-11	
50000	$x_0^6$	1	3	0.0257	0	20	45	0.3622	8.23E-11	13	20	0.2395	3.27E-11	44	47	0.6443	9.05E-11	37	162	1.2419	2.33E-11	
4	5000	$x_0^1$	1	4	0.0944	0	17	69	0.0637	3.94E-11	12	50	0.0492	2.69E-11	48	51	0.1253	4.73E-11	1	13	0.0193	0
	5000	$x_0^2$	1	4	0.0081	0	17	69	0.0765	3.94E-11	12	50	0.0497	2.69E-11	48	51	0.1203	4.73E-11	1	13	0.0191	0
	5000	$x_0^3$	5	10	0.0322	0	18	73	0.0884	3.23E-11	12	55	0.0602	9.59E-11	50	53	0.1223	9.30E-11	1	13	0.0206	0
	5000	$x_0^4$	1	4	0.0071	0	17	69	0.0724	3.94E-11	12	50	0.0599	2.69E-11	48	51	0.1152	4.73E-11	1	13	0.0219	0
	5000	$x_0^5$	1	4	0.0082	0	16	69	0.0906	3.85E-11	10	46	0.0465	7.37E-11	48	50	0.1100	8.64E-11	1	13	0.0208	0
	5000	$x_0^6$	1	4	0.0079	0	19	78	0.0952	4.11E-11	13	56	0.0639	5.52E-11	53	57	0.1416	8.86E-11	3	26	0.0328	0
	10000	$x_0^1$	1	4	0.0099	0	17	69	0.1228	5.59E-11	12	50	0.0975	3.81E-11	48	51	0.1850	6.69E-11	1	13	0.0296	0
	10000	$x_0^2$	1	4	0.0115	0	17	69	0.1115	5.59E-11	12	50	0.0840	3.81E-11	48	51	0.1988	6.69E-11	1	13	0.0303	0
	10000	$x_0^3$	5	10	0.0277	0	18	73	0.1336	4.59E-11	14	59	0.1005	1.06E-11	52	54	0.2025	9.30E-11	1	13	0.0283	0
	10000	$x_0^4$	1	4	0.0092	0	17	69	0.1294	5.59E-11	12	50	0.0830	3.81E-11	48	51	0.1827	6.69E-11	1	13	0.0302	0
	10000	$x_0^5$	1	4	0.0084	0	16	69	0.1142	5.44E-11	11	50	0.1068	8.17E-12	48	51	0.1862	5.76E-11	1	13	0.0258	0
	10000	$x_0^6$	1	4	0.0106	0	19	78	0.1383	5.85E-11	13	56	0.0997	7.79E-11	55	58	0.2113	8.86E-11	3	26	0.0686	0
	50000	$x_0^1$	1	4	0.0347	0	18	73	0.4537	2.39E-11	12	50	0.3198	8.51E-11	50	53	0.7322	4.99E-11	1	13	0.1211	0
	50000	$x_0^2$	1	4	0.0322	0	18	73	0.4312	2.39E-11	12	50	0.3259	8.51E-11	50	53	0.7229	4.99E-11	1	13	0.1313	0
	50000	$x_0^3$	5	10	0.1218	0	19	77	0.5047	1.96E-11	14	59	0.3879	2.38E-11	52	55	0.7367	9.80E-11	1	13	0.1366	0
50000	$x_0^4$	1	4	0.0425	0	18	73	0.4601	2.39E-11	12	50	0.3517	8.51E-11	50	53	0.7321	4.99E-11	1	13	0.1464	0	
50000	$x_0^5$	1	4	0.0337	0	17	73	0.4407	2.31E-11	11	50	0.3181	1.83E-11	50	52	0.7817	9.11E-11	1	13	0.1235	0	
50000	$x_0^6$	1	4	0.0366	0	20	82	0.5146	2.50E-11	14	60	0.4002	1.36E-11	55	59	0.8389	9.34E-11	3	26	0.2487	0	

Table 3: Test results for Examples 5–6.

PN	VAR	SP	Algorithm 1			ACGD			MDKM			SRCME			SDYCG							
			NIT	FE	PT	Norm	NIT	FE	PT	Norm	NIT	FE	PT	Norm	NIT	FE	PT	Norm				
5	5000	$x_0^1$	39	125	0.1944	7.91E-11	***	***	***	***	47	433	0.3396	9.53E-11	58	62	0.2015	7.56E-11	***	***		
	5000	$x_0^2$	46	167	0.1822	8.03E-11	***	***	***	***	47	433	0.3043	9.93E-11	58	62	0.1707	6.55E-11	***	***		
	5000	$x_0^3$	35	118	0.1484	3.66E-11	167	1532	0.9541	8.43E-11	49	450	0.3283	8.24E-11	62	66	0.1717	6.26E-11	***	***		
	5000	$x_0^4$	43	127	0.1625	9.12E-11	62	558	0.4088	5.30E-11	47	433	0.3098	9.93E-11	58	62	0.1645	6.55E-11	***	***		
	5000	$x_0^5$	38	108	0.1451	4.16E-11	***	***	***	***	49	451	0.3166	6.70E-11	67	72	0.1840	8.07E-11	***	***		
	5000	$x_0^6$	38	129	0.1564	6.50E-11	***	***	***	***	50	458	0.3394	7.72E-11	68	74	0.2063	8.85E-11	***	***		
	10000	$x_0^1$	32	111	0.2280	8.73E-11	***	***	***	***	48	442	0.5151	6.19E-11	58	62	0.2597	9.37E-11	***	***		
	10000	$x_0^2$	46	167	0.2847	9.05E-11	***	***	***	***	48	442	0.5069	6.48E-11	58	62	0.2468	8.44E-11	***	***		
	10000	$x_0^3$	38	117	0.2288	6.22E-11	***	***	***	***	49	450	0.5597	8.85E-11	62	66	0.2643	8.92E-11	***	***		
	10000	$x_0^4$	43	130	0.2462	8.99E-11	134	1221	1.3007	9.29E-11	48	442	0.4975	6.48E-11	58	62	0.2412	8.44E-11	***	***		
	10000	$x_0^5$	34	109	0.2048	8.33E-11	***	***	***	***	49	451	0.5514	8.23E-11	68	73	0.2792	8.30E-11	***	***		
	10000	$x_0^6$	41	134	0.2719	5.37E-11	***	***	***	***	50	458	0.5529	8.52E-11	70	76	0.2798	7.16E-11	***	***		
	50000	$x_0^1$	36	128	0.9632	6.36E-11	***	***	***	***	48	442	2.2058	9.31E-11	60	64	0.9733	7.14E-11	***	***		
	50000	$x_0^2$	47	171	1.2254	8.08E-11	***	***	***	***	48	442	2.1443	9.50E-11	59	63	0.9761	1.00E-10	***	***		
	50000	$x_0^3$	38	124	0.9703	7.24E-11	***	***	***	***	50	459	2.2698	6.54E-11	64	68	1.0506	6.86E-11	***	***		
	50000	$x_0^4$	46	144	1.0945	9.30E-11	***	***	***	***	48	442	2.1490	9.50E-11	59	63	0.9941	1.00E-10	***	***		
	50000	$x_0^5$	43	155	1.1195	5.40E-11	***	***	***	***	50	460	2.2638	8.07E-11	71	76	1.1856	6.75E-11	***	***		
	50000	$x_0^6$	49	149	1.1485	3.40E-11	***	***	***	***	51	467	2.2987	7.15E-11	73	79	1.2177	8.84E-11	***	***		
6	5000	$x_0^1$	1	5	0.0461	0	13	79	0.0566	1.00E-10	11	79	0.0680	1.96E-11	56	60	0.1672	9.23E-11	1	13	0.0167	0
	5000	$x_0^2$	1	5	0.0078	0	13	79	0.0761	1.00E-10	11	79	0.0717	1.96E-11	56	60	0.1466	9.23E-11	1	13	0.0225	0
	5000	$x_0^3$	7	19	0.0257	0	13	85	0.0836	6.70E-11	11	86	0.0811	2.32E-11	61	65	0.1602	6.35E-11	1	13	0.0224	0
	5000	$x_0^4$	1	5	0.0082	0	13	79	0.0868	1.00E-10	11	79	0.0763	1.96E-11	56	60	0.1507	9.23E-11	1	13	0.0210	0
	5000	$x_0^5$	1	5	0.0086	0	12	79	0.0759	9.24E-11	10	79	0.0723	7.83E-12	56	60	0.1641	9.70E-11	1	13	0.0211	0
	5000	$x_0^6$	1	5	0.0093	0	15	96	0.0854	1.14E-11	12	92	0.0850	6.23E-12	62	67	0.1759	6.60E-11	1	13	0.0282	0
	10000	$x_0^1$	1	5	0.0088	0	14	85	0.1316	1.60E-11	11	79	0.1254	2.77E-11	58	62	0.2513	5.22E-11	1	13	0.0357	0
	10000	$x_0^2$	1	5	0.0151	0	14	85	0.1262	1.60E-11	11	79	0.1250	2.77E-11	58	62	0.2652	5.22E-11	1	13	0.0326	0
	10000	$x_0^3$	7	19	0.0485	0	13	85	0.1399	9.53E-11	11	86	0.1047	3.28E-11	61	65	0.2242	8.98E-11	1	13	0.0328	0
	10000	$x_0^4$	1	5	0.0125	0	14	85	0.1159	1.60E-11	11	79	0.1281	2.77E-11	58	62	0.2317	5.22E-11	1	13	0.0348	0
	10000	$x_0^5$	1	5	0.0116	0	13	85	0.1214	1.47E-11	10	79	0.1353	1.11E-11	58	62	0.2324	5.49E-11	1	13	0.0348	0
	10000	$x_0^6$	1	5	0.0090	0	15	96	0.1562	1.64E-11	12	92	0.1333	8.81E-12	62	67	0.2273	9.33E-11	1	13	0.0346	0
	50000	$x_0^1$	1	5	0.0366	0	14	85	0.4488	3.62E-11	11	79	0.3852	6.20E-11	60	64	0.8419	4.67E-11	1	13	0.1447	0
	50000	$x_0^2$	1	5	0.0425	0	14	85	0.4320	3.62E-11	11	79	0.3877	6.20E-11	60	64	0.8647	4.67E-11	1	13	0.1333	0
	50000	$x_0^3$	7	19	0.1931	0	15	91	0.4840	2.42E-11	11	86	0.4735	7.33E-11	63	67	0.9133	8.03E-11	1	13	0.1443	0
	50000	$x_0^4$	1	5	0.0406	0	14	85	0.4272	3.62E-11	11	79	0.4529	6.20E-11	60	64	0.8653	4.67E-11	1	13	0.1435	0
	50000	$x_0^5$	1	5	0.0422	0	13	85	0.4573	3.29E-11	10	79	0.4343	2.48E-11	60	64	0.8796	4.92E-11	1	13	0.1320	0
	50000	$x_0^6$	1	5	0.0384	0	15	96	0.5068	3.72E-11	12	92	0.5133	1.97E-11	64	69	0.9350	8.34E-11	1	13	0.1334	0

Table 4: Test results for Examples 7–8.

PN	VAR	SP	Algorithm 1				ACGD				MDKM				SRCMIE				SDYCG			
			NIT	FE	PT	Norm	NIT	FE	PT	Norm	NIT	FE	PT	Norm	NIT	FE	PT	Norm	NIT	FE	PT	Norm
7	5000	$x_9$	53	200	0.2101	7.56E-11	21	192	0.1532	0	79	712	0.5303	9.73E-11	76	79	0.1909	8.28E-11	2	25	0.0338	0
	5000	$x_9$	46	186	0.2265	6.80E-11	17	157	0.1397	0	104	937	0.7143	9.80E-11	91	94	0.2226	8.75E-11	10	121	0.1459	7.98E-12
	5000	$x_9$	47	190	0.1880	9.99E-11	15	124	0.1260	0	108	973	0.6780	9.80E-11	106	109	0.2492	7.94E-11	10	121	0.1465	2.43E-11
	5000	$x_9$	43	44	0.0587	5.33E-11	17	157	0.1419	0	104	937	0.6370	9.80E-11	91	94	0.2167	8.75E-11	10	121	0.1479	7.98E-12
	5000	$x_9$	44	178	0.1873	7.26E-11	45	408	0.2961	2.83E-13	133	1198	0.8348	8.79E-11	174	177	0.4006	9.98E-11	9	109	0.1424	3.46E-11
	5000	$x_9$	1	6	0.0096	0	73	534	0.3972	0	142	1279	0.8802	9.53E-11	192	195	0.4206	8.61E-11	10	121	0.1454	6.76E-11
	10000	$x_9$	53	200	0.3128	7.56E-11	31	300	0.3714	5.28E-11	79	712	0.8057	9.62E-11	71	74	0.2977	8.85E-11	2	25	0.0625	0
	10000	$x_9$	46	186	0.2824	7.34E-11	26	247	0.2910	5.13E-12	104	937	1.0475	9.46E-11	90	93	0.3259	9.35E-11	10	121	0.2220	7.98E-12
	10000	$x_9$	48	194	0.3304	6.20E-11	12	103	0.1481	0	110	991	1.1046	8.99E-11	105	108	0.4397	9.28E-11	10	121	0.2328	2.43E-11
	10000	$x_9$	7	22	0.0506	8.84E-11	26	247	0.2894	4.18E-12	104	937	1.0184	9.46E-11	90	93	0.3275	9.35E-11	10	121	0.2525	7.98E-12
	10000	$x_9$	44	178	0.3100	7.75E-11	28	231	0.3074	0	134	1207	1.3942	9.34E-11	177	180	0.6494	8.68E-11	9	109	0.2011	3.46E-11
	10000	$x_9$	1	6	0.0157	0	74	544	0.6694	0	144	1297	1.3429	8.44E-11	194	197	0.7516	8.87E-11	10	121	0.2279	6.76E-11
	50000	$x_9$	53	200	1.6891	7.57E-11	17	147	0.7691	2.54E-11	73	658	3.0758	9.06E-11	69	72	1.1170	8.83E-11	2	25	0.2342	0
	50000	$x_9$	46	186	1.3757	7.46E-11	18	165	0.8798	0	107	964	4.5572	9.28E-11	89	92	1.4188	8.98E-11	10	121	0.9147	7.98E-12
	50000	$x_9$	48	194	1.2152	6.57E-11	30	284	1.4076	0	112	1009	4.6808	9.40E-11	104	107	1.5571	9.51E-11	10	121	0.9020	2.43E-11
	50000	$x_9$	3	10	0.0886	1.11E-12	18	165	0.8443	0	107	964	4.4980	9.28E-11	89	92	1.3599	8.98E-11	10	121	0.9147	7.98E-12
	50000	$x_9$	44	178	1.1118	8.02E-11	28	232	1.1933	0	137	1234	5.8379	9.36E-11	181	184	2.7239	9.61E-11	9	109	0.8139	3.46E-11
	50000	$x_9$	1	6	0.0441	0	94	685	3.5413	0	147	1324	6.2662	8.46E-11	199	202	3.0006	8.62E-11	10	121	0.9277	6.76E-11
8	5000	$x_3$	23	25	0.1556	7.52E-11	70	284	0.3487	8.11E-11	24	62	0.1110	2.69E-11	43	46	0.1386	5.92E-11	***	***	***	***
	5000	$x_3$	21	23	0.1221	5.58E-11	28	104	0.1570	3.45E-11	23	61	0.1212	2.62E-11	44	47	0.1399	6.58E-11	***	***	***	***
	5000	$x_9$	21	24	0.0794	9.32E-11	20	59	0.1114	6.93E-11	20	54	0.0959	4.48E-11	42	45	0.1438	5.80E-11	***	***	***	***
	5000	$x_9$	20	23	0.0671	5.55E-11	28	104	0.1447	3.45E-11	23	61	0.1168	2.62E-11	44	47	0.1388	6.58E-11	***	***	***	***
	5000	$x_9$	21	24	0.0736	9.41E-11	29	110	0.1378	7.07E-11	24	62	0.1171	2.67E-11	44	47	0.1518	7.64E-11	***	***	***	***
	5000	$x_9$	21	23	0.1076	9.59E-11	20	59	0.1096	8.60E-11	20	54	0.0993	4.89E-11	42	45	0.1308	5.74E-11	***	***	***	***
	10000	$x_9$	22	25	0.1168	8.12E-11	33	144	0.2716	5.37E-11	24	62	0.1874	2.64E-11	43	46	0.2143	7.26E-11	***	***	***	***
	10000	$x_9$	22	25	0.1308	6.35E-11	27	95	0.2227	3.72E-11	23	61	0.1746	2.58E-11	44	47	0.2515	8.08E-11	***	***	***	***
	10000	$x_3$	21	24	0.1258	5.30E-11	22	66	0.1621	3.90E-11	19	52	0.1560	2.32E-11	42	45	0.2120	7.12E-11	***	***	***	***
	10000	$x_9$	18	21	0.1217	9.01E-11	27	95	0.2210	3.72E-11	23	61	0.1935	2.58E-11	44	47	0.2332	8.08E-11	***	***	***	***
	10000	$x_9$	21	24	0.1117	4.31E-11	73	336	0.6304	9.88E-11	24	62	0.1723	2.63E-11	44	47	0.2345	9.37E-11	***	***	***	***
	10000	$x_9$	21	23	0.1119	5.68E-11	22	66	0.1639	4.96E-11	19	52	0.1725	2.59E-11	42	45	0.2244	7.05E-11	***	***	***	***
	50000	$x_9$	24	26	0.5426	3.53E-11	28	97	0.9419	2.72E-11	24	62	0.7122	2.56E-11	44	47	0.9469	6.43E-11	***	***	***	***
	50000	$x_3$	21	24	0.4761	8.87E-11	24	76	0.7661	2.67E-11	24	62	0.7165	2.53E-11	45	48	0.9792	7.16E-11	***	***	***	***
	50000	$x_9$	22	25	0.5304	4.68E-11	21	63	0.6520	4.01E-11	20	54	0.6841	4.24E-11	43	46	0.9370	6.31E-11	***	***	***	***
	50000	$x_9$	19	22	0.4547	6.54E-11	24	76	0.7589	2.67E-11	24	62	0.7456	2.53E-11	45	48	0.9650	7.16E-11	***	***	***	***
	50000	$x_9$	23	25	0.5032	3.46E-11	25	80	0.8003	3.38E-11	24	62	0.7193	2.55E-11	45	48	0.9796	8.30E-11	***	***	***	***
	50000	$x_9$	23	26	0.5042	5.09E-11	46	183	1.6076	3.29E-11	20	54	0.6270	3.17E-11	43	46	0.9348	6.25E-11	***	***	***	***

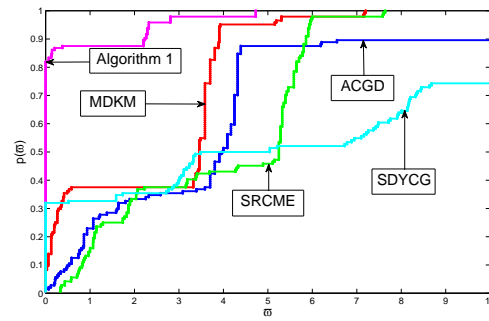


Figure 1: Dolan and More profile for number of iterations

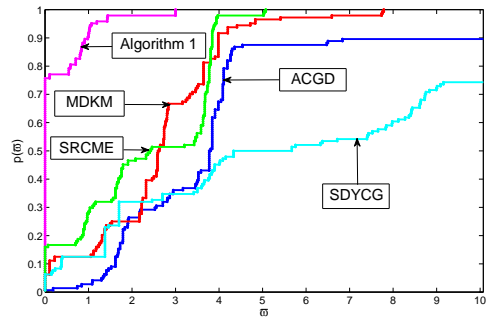


Figure 2: Dolan and More profile for function evaluations

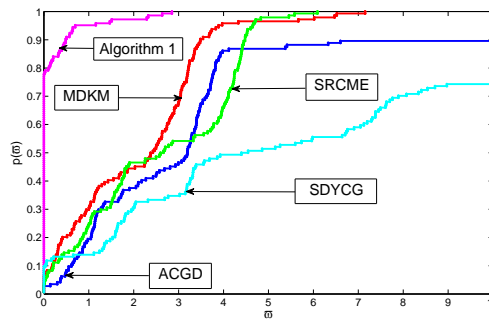


Figure 3: Dolan and More profile for CPU time

We presented results of the first experiment in Tables 1, 2, 3 and 4, where the labels PN, VAR, SP, NIT, FE, PT, and Norm represent number of test

example, Dimension, Initial guess, number of iterations, function evaluations, CPU time, and norm achieved at approximate solution. Also, \* \* \* indicates no solution of (3) was obtained in 1000 iterations. It is clear from the four tables that Algorithm 1 outperformed the other methods in all three metrics considered. These results are further analyzed in Figures 1, 2, and 3, which are plotted by utilizing Dolan and Moré [23] performance profile. In Figure 1, we see that about 83% of the test examples were solved by Algorithm 1 with less iterations, while ACGD, MDKM, SRCME and SDYCG solved 2%, 8%, 0% and 32%. Furthermore, these values include instances where some of the algorithms solved 24% of the test examples with the same minimum number of iterations. Also, from Figure 2, we see that Algorithm 1 solved 77% of the test examples with minimum function evaluations compared to ACGD, MDKM, SRCME and SDYCG that recorded 1%, 5%, 16%, and 6%. Here also, some of the algorithms solved 13% of the test examples with the same minimum function evaluations. Next, we observed from Figure 3 that Algorithm 1 solved 77.78% of the test examples with the least CPU time compared to ACGD, MDKM, SRCME and SDYCG that recorded 2.78%, 6.25%, 3.47%, and 9.72%. In addition, the top curve in all three figures corresponds to that of Algorithm 1, which clearly shows that the scheme is the most effective. Moreover, the average residual for the five algorithms as computed from Tables 1, 2, 3, and 4 are given as follows: Algorithm 1 ( $3.09 \times 10^{-11}$ ), ACGD ( $3.36 \times 10^{-11}$ ), MDKM ( $4.56 \times 10^{-11}$ ), SRCME ( $7.58 \times 10^{-11}$ ), and SDYCG ( $3.90 \times 10^{-10}$ ). This, together with the other metrics analyzed, indicates that Algorithm 1 is more efficient for solving (3) than the other schemes.

## 4.2 Experiment 2: Image De-blurring

We use this subsection to demonstrate the application of Algorithm 1 in deblurring images contaminated by noise. To achieve the desired goal, we compare our scheme with two effective schemes in the literature, namely, HTTCGP [63] and MFRM [1].

As a background for image de-blurring, we briefly discuss sparse signal recovery, which deals with obtaining sparse solutions for the under-determined linear system  $\mathcal{H}x = h$ , where  $\mathcal{H} \in \mathbb{R}^{k \times n}$  ( $k \ll n$ ) is a sampled matrix,  $x$  a sparse signal and  $h \in \mathbb{R}^k$  denotes an observed value. In recovering  $x$  from  $\mathcal{H}x - h$ , the following  $\ell_1$  norm regularization problem is solved:

$$\min_x f(x) := \frac{1}{2} \|h - \mathcal{H}x\|_2^2 + \zeta \|x\|_1, \quad (52)$$

with  $\zeta > 0$ . Careful observation reveals (52) to be a form of the problem represented in (4).

In [24], it was shown that to solve (52), it is first expressed as a convex quadratic model, where  $x \in \mathbb{R}^n$  is written as

$$x = v - \nu, \quad v \geq 0, \quad \nu \geq 0, \quad v, \nu \in \mathbb{R}^n,$$

with  $v_i = (x_i)_+$ ,  $\nu_i = (-x_i)_+$ , for all  $i = 1, 2, \dots, n$  and  $(\cdot)_+ = \max\{0, x\}$ . Using this expression, we have  $\|x\|_1 = E_n^T v + E_n^T \nu$  where  $E_n = (1, 1, \dots, 1)^T \in \mathbb{R}^n$ . Thus, (52) becomes

$$\min \left\{ \frac{1}{2} \|\mathcal{H}(v - \nu) - h\|_2^2 + \zeta (E_n^T v + E_n^T \nu) \mid v \geq 0, \nu \geq 0 \right\}. \quad (53)$$

Now, if we define

$$w = \begin{pmatrix} v \\ \nu \end{pmatrix}, \quad \chi = \zeta E_{2n} + \begin{pmatrix} -\omega \\ \omega \end{pmatrix}, \quad \omega = \mathcal{H}^T h, \quad G = \begin{pmatrix} \mathcal{H}^T \mathcal{H} & -\mathcal{H}^T \mathcal{H} \\ -\mathcal{H}^T \mathcal{H} & \mathcal{H}^T \mathcal{H} \end{pmatrix},$$

then (53) becomes

$$\min \left\{ \frac{1}{2} w^T G w + \chi^T w \mid w \geq 0 \right\}. \quad (54)$$

Moreover, since  $G$  is a positive semi-definite matrix, (54) is a convex quadratic problem [61]. Also, based on the optimality condition mentioned earlier,  $w$  in (54) is a minimizer of (54) if it solves the system of equations

$$F(w) = \min\{w, Gw + \chi\} = 0.$$

Finally, Xiao [61] and Pang [49], showed that  $F$  satisfies (2) and (21). Hence, (52) can be represented as the problem (3), and solved using Algorithm 1.

Next, we apply Algorithm 1 to de-blur three images, which includes Einstein.tif (M1) ( $512 \times 512$ ), Cameraman.png (M2) ( $512 \times 512$ ) and Barbara.png (M3) ( $512 \times 512$ ). In the experiments, the signal-to-noise ratio (SNR)

$$SNR = 20 \times \log_{10} \left( \frac{\|\tilde{x}\|}{\|\bar{x} - \tilde{x}\|} \right),$$

and the peak to signal ratio (PSNR)

$$PSNR = 10 \times \log_{10} \frac{V^2}{MSE},$$

were used to calculate restoration quality, with  $V$  being the maximum absolute value of recovery and (MSE) is defined by

$$MSE = \frac{1}{n} \|\tilde{x} - \bar{x}\|^2, \quad (55)$$

where  $x$  is the signal recovered and  $\tilde{x}$  the actual sparse one. In addition, we use MSE as defined in (55) and structured similarity index (SSIM), which describes the similarity between the original and reconstructed or recovered images to measure numerical efficiency of the algorithms. Performance of Algorithm 1 is compared with that of HTTCGP [63] and MFRM [1], which are also effective for de-blurring images, using the same parameter values in the respective papers. Parameters for Algorithm 1 are set as  $\beta = 0.9$ ,  $\delta = 0.001$ ,  $r = 0.01$  and  $\gamma = 0.25$ , while  $\phi$  retains the value in the first experiment.

**Table 5:** Image de-blurring results for Algorithm 1, HTTCGP, and MFRM under different Gaussian blur kernels

Image	Algorithm 1			HTTCGP			MFRM					
	MSE	SNR	PSNR	SSIM	MSE	SNR	PSNR	SSIM	MSE	SNR	PSNR	SSIM
M1(0.5)	8.6452e+01	20.52	29.11	0.84	9.8875e+01	19.94	28.25	0.82	9.0269e+01	20.33	28.81	0.83
M2(0.5)	1.6736e+02	20.31	26.10	0.83	1.7560e+02	20.10	25.83	0.83	1.7312e+02	20.16	25.97	0.83
M3(0.5)	2.2125e+02	18.79	24.51	0.75	2.2442e+02	18.73	24.43	0.74	2.2872e+02	18.65	24.28	0.73
M1(0.75)	8.8465e+01	20.42	28.93	0.83	9.9917e+01	19.89	28.23	0.82	9.2995e+01	20.21	28.63	0.82
M2(0.75)	1.6960e+02	20.25	26.11	0.82	1.7060e+02	20.23	26.06	0.82	1.7149e+02	20.20	26.06	0.82
M3(0.75)	2.2348e+02	18.75	24.49	0.74	2.2572e+02	18.71	24.31	0.73	2.2573e+02	18.71	24.30	0.73
M1(1.25)	9.5444e+01	20.09	28.62	0.81	1.0258e+02	19.78	28.10	0.81	9.6906e+01	20.03	28.51	0.81
M2(1.25)	1.7558e+02	20.10	25.86	0.78	1.9352e+02	19.68	25.31	0.79	1.7748e+02	20.06	25.81	0.79
M3(1.25)	2.3043e+02	18.62	24.35	0.71	2.3006e+02	18.63	24.34	0.71	2.3312e+02	18.57	24.22	0.71

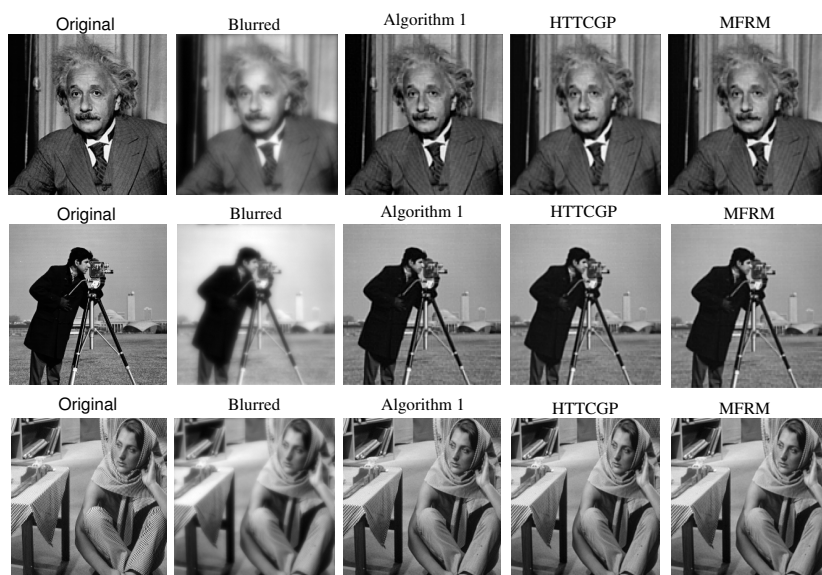


Figure 4: Recovered images under Gaussian blur kernel with standard deviation 0.5

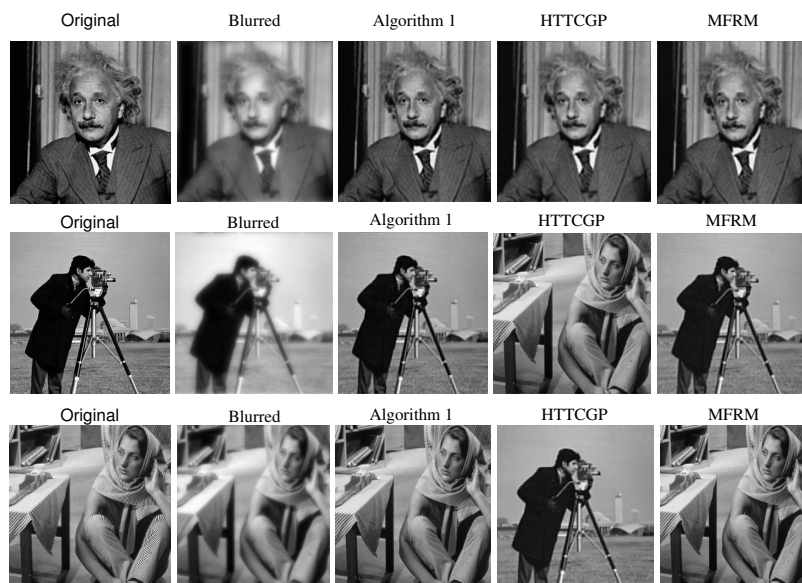


Figure 5: Recovered images under Gaussian blur kernel with standard deviation 0.75

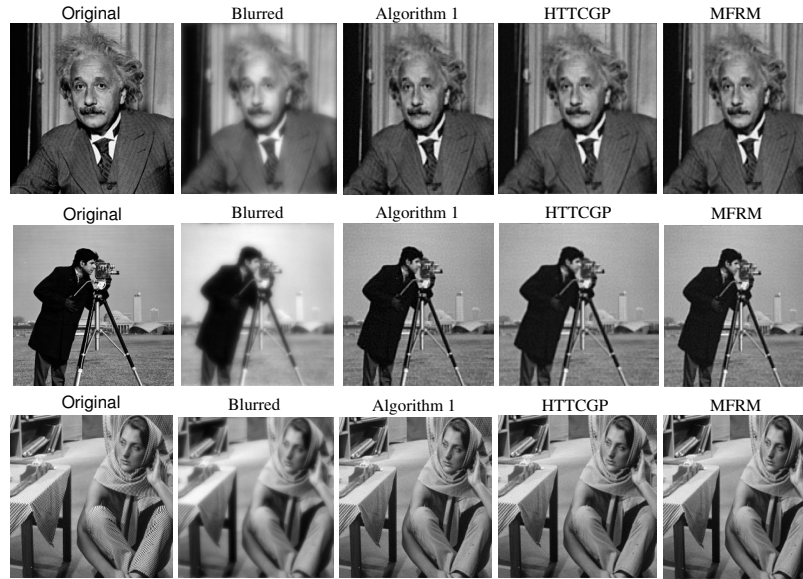


Figure 6: Recovered images under Gaussian blur kernel with standard deviation 1.25

Generally, the restored images from blurry ones by an algorithm with larger values of SNR, PSNR, and SSIM appear much closer to the original ones than algorithms with lower values of the metrics. Also, algorithms with a lower value of MSE yield better quality of restored images than algorithms with larger values of the metrics. In our experiments, Algorithm 1 yields the best values of the aforementioned performance metrics (see underlined values in Table 5). Also, the original, blurry, and recovered images by the three algorithms are presented in Figures 4, 5, and 6. Furthermore, a number of Gaussian blur kernels were used to test robustness of the algorithms (see Table 5). In Table 5, the test problem solved with standard deviation of the Gaussian blur kernel  $\sigma$  is given by  $Mi(\sigma)$ . Therefore, based on this discussion, we conclude that Algorithm 1 is effective for image recovery problems.

## 5 Conclusion

In this work, an adaptive DK method was considered for nonlinear monotone systems and image recovery problems. The novelty of the work is that value

of the parameter of the scheme was obtained such that the eigenvalues of the symmetric form of its iteration matrix are clustered at a point. This strategy helps to ensure that the scheme's directions automatically possess the property for global convergence without any adjustment made to the derived value of the DK parameter. The method can also be used to solve nonsmooth nonlinear problems. Also, analysis of the method's convergence proved that it converges globally, while its effectiveness was shown through experiments with four other effective methods for solving constrained nonlinear problems and image deblurring. As future research, we intend to apply the proposed method to solve signal reconstruction and motion control problems.

## Acknowledgements

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

## References

- [1] Abubakar, A.B., Kumam, P., Mohammed, H. and Sitthithakerngkiet, K. *A modified Fletcher-Reeves conjugate gradient method for monotone nonlinear equations with some applications*, Mathematics, 7 (745) (2019), 1–25.
- [2] Ahmed, K., Waziri, M.Y. and Halilu, A.S. *On two symmetric Dai-Kou type schemes for constrained monotone equations with image recovery application*, Euro J. Comput. Optim. 11 100057 (2023) 1–32.
- [3] Ahmed, K., Waziri, M.Y., Halilu, A.S. and Murtala, S. *Sparse signal reconstruction via Hager-Zhang-type schemes for constrained system of nonlinear equations*, Optimization, 73 (6) (2024), 1949–1980.
- [4] Ahmed, K., Waziri, M.Y., Halilu, A.S., Murtala, S. and Sabi'u, J. *Another Hager-Zhang type method via singular value study for constrained monotone equations with application*, Numer. Algor. 96 (4) (2024), 1583–1623.

- [5] Ahmed, K., Waziri, M.Y., Murtala, S., Halilu, A.S. and Sabi'u, J. *On a scaled symmetric Dai-Liao-type scheme for constrained system of nonlinear equations with applications*, J. Optim. Theory Appl. 200 (2024), 669–702.
- [6] Al-Baali M. *Numerical experience with a class of self-scaling quasi-Newton algorithms*, J. Optim. Theory and Appl. 96(3) (1998), 533–553.
- [7] Althobaiti, A., Sabi'u J., Emadifar, H., Junsawang, P., and Sahoo, S.K. *A scaled Dai-Yuan projection-based conjugate gradient method for solving monotone equations with applications*, Symmetry, 14 (1401) (2023), 1–28.
- [8] Andrei, N. *Eigenvalues versus singular values study in conjugate gradient algorithms for large-scale unconstrained optimization*, Optim. Methods Softw. 32(3)(2017), 534–551.
- [9] Andrei, N. *A Dai-Liao conjugate gradient algorithm with clustering of eigenvalues*, Numer. Algor. 77 (2018), 1273–1282.
- [10] Andrei, N. *A double parameter scaled BFGS method for unconstrained optimization*, J. Comput. Appl. Math. 332 (2018), 26–44.
- [11] Andrei, N. *A double-parameter scaling Broyden-Fletcher-Goldfarb-Shanno method based on minimizing the measure function of Byrd and Nocedal for unconstrained optimization*, J. Optim. Theory Appl. 178 (2018), 191–218.
- [12] Babaie-Kafaki, S. and Aminifard, Z. *Two parameter scaled memoryless BFGS methods with a nonmonotone choice for the initial step length*, Numer. Algor. 82 (2019), 1345–1357.
- [13] Biggs, M.C. *Minimization algorithms making use of non-quadratic properties of the objective function*, J. Inst. Math. Appl. 8 (1971), 315–327.
- [14] Broyden, C.G. *A class of methods for solving nonlinear simultaneous equations*, Math. Comput. 19 (1965), 577–593.
- [15] Broyden, C.G. *The convergence of a class double-rank minimization algorithms*, J. Inst. Math. Appl. 6 (1970) 76–90.

- [16] Byrd, R.H., Liu, D.C. and Nocedal, J. *On the behavior of Broyden's class of quasi-Newton methods*, SIAM J. Optim. 2 (1992), 533–557.
- [17] Byrd, R. and Nocedal, J. *A tool for the analysis of quasi-Newton methods with application to unconstrained minimization*, SIAM J. Numer. Anal. 26 (1989), 727–739.
- [18] Dai, Y.H. and Kou, CX. *A nonlinear conjugate gradient algorithm with an optimal property and an improved wolfe line search*, SIAM J. Optim. 23 (2013), 296–320.
- [19] Dai, Y.H. and Liao, L.Z. *New conjugacy conditions and related nonlinear conjugate gradient methods*, Appl. Math. Optim. 43 (1) (2001), 87–101.
- [20] Dai, Y.H. and Yuan, Y. *A nonlinear conjugate gradient method with a strong global convergence property*, SIAM J. Optim. 10 (1999), 177–182.
- [21] Dennis, J.E. and Schnabel, R.B. *Numerical methods for unconstrained optimization and nonlinear equations*, Englewood Cliffs, NJ: Prentice-Hall, 1983.
- [22] Ding, Y., Xiao, Y. and Li, J. *A class of conjugate gradient methods for convex constrained monotone equations*, Optimization, 66 (12) (2017), 2309–2328.
- [23] Dolan, E.D. and More, J.J. *Benchmarking optimization software with performance profiles*, Math. Program, 91 (2002), 201–2013.
- [24] Figueiredo, M., Nowak, R. and Wright, S.J. *Gradient projection for sparse reconstruction, application to compressed sensing and other inverse problems*, IEEE J-STSP IEEE Press, Piscataway, NJ. (2007), 586–597.
- [25] Fletcher, R. and Reeves, C. *Function minimization by conjugate gradients*, Comput. J. 7 (1964), 149–154.
- [26] Fletcher, R. *A new approach to variable metric algorithms*, Computer J. 13 (1970) 317–322.

- [27] Fletcher, R. *Practical method of optimization*, Volume 1: Unconstrained Optimization, 2nd ed., Wiley, New York, (1997).
- [28] Gao, P., Zheng, W., Wang, T., Li, Y. and Li, F. *Signal recovery with constrained monotone nonlinear equations*, J. Appl. Anal. Comput. 13 (4) (2023), 2006–2025.
- [29] Gill, P.E. and Leonard, M.W. *Reduced-Hessian quasi Newton methods for unconstrained optimization*, SIAM J. Optim. 12 (2001), 209–237.
- [30] Goldfarb, D. *A family of variable metric methods derived by variation mean*, Math. Comput. 23 (1970), 23–26.
- [31] Hager, W.W. and Zhang, H. *A survey of nonlinear conjugate gradient methods*, Pacific J. Optim. 2 (2006), 35–58.
- [32] Halilu, A.S., Majumder, A., Waziri, M.Y., Ahmed, K. and Awwal, A.M. *Motion control of the two joint planar robotic manipulators through accelerated Dai-Liao method for solving system of nonlinear equations*, Eng. Comput. 39 (5) (2021), 1–39.
- [33] Hestenes, M.R. and Stiefel, E.L. *Methods of conjugate gradients for solving linear systems*, J. Research Nat. Bur. Standards, 49 (1952), 409–436.
- [34] Ivanov, B., Milanovic, G.V. and Stanimirovic, P.S. *Accelerated Dai-Liao projection method for solving systems of monotone nonlinear equations with application to image deblurring*, J. Global Optim. 85 (2023) 377–420.
- [35] Kaporin, I.E. *New convergence results and preconditioning strategies for the conjugate gradient methods*, Numer. Linear Alg. Appl. 1(2) (1994), 179–210.
- [36] Kiri, A.I., Waziri, M.Y. and Ahmed, K. *A modified Liu-Storey scheme for nonlinear systems with an application to image recovery*, Iran, J. Numer. Anal. and Optim. 3 (1) (2023), 38–58.
- [37] Kratzer, D., Parter, S.V. and Steuerwalt, M. *Block splittings for the conjugate gradient method*, Comp. Fluid, 11 (1983), 255–279.

- [38] La Cruz, W. *A Spectral algorithm for large-scale systems of nonlinear monotone equations*, Numer. Algor. 76 (2017), 1109–1130.
- [39] La Cruz, W., Martinez, J.M. and Raydan, M. *Spectral residual method without gradient information for solving large-scale nonlinear systems of equations*, Theory and experiments, Technical Report RT-04-08, 2005.
- [40] Liu, J.K., and Li, S.J. *A projection method for convex constrained monotone nonlinear equations with applications*, Comput. Math. Appl. 70 (10) (2015), 2442–2453.
- [41] Liu, Y. and Storey, C. *Efficient generalized conjugate gradient algorithms*, Part 1: Theory, J. Optim. Theory Appl. 69 (1991), 129–137.
- [42] Narushima, Y. and Yabe, H. *A survey of sufficient descent conjugate gradient methods for unconstrained optimization*, SUT J. Math. 50(2) (2014), 167–203.
- [43] Nocedal, J. *Theory of algorithms for unconstrained optimization*, Acta Numer. 1 (1992), 199–242.
- [44] Nocedal, J. and Wright, S.J. *Numerical optimization*, Springer, New York, 2006.
- [45] Oren, S.S. and Luenberger, D.G. *Self-scaling variable metric (SSVM) algorithms*, part I: criteria and sufficient conditions for scaling a class of algorithms, Manag. Sci. 20 (1974), 845–862.
- [46] Oren S.S., and Luenberger D.G. *Self scaling variable metric (SSVM) algorithms*, part I: criteria and sufficient conditions for scaling a class of algorithms, Manag. Sci. 20(5) (1974), 845–862.
- [47] Oren S.S., and Spedicato E. *Optimal conditioning of self scaling variable metric algorithms*, Math. Prog. 10(1) (1976), 70–90.
- [48] Ortega, J.M. and Rheinboldt, W.C. *Iterative solution of nonlinear equations in several variables*, New York: Academic Press, 1970.
- [49] Pang, J.S. *Inexact Newton methods for the nonlinear complementarity problem*, Math. Program. 36 (1986), 54–71.

- [50] Polak, E. and Ribière, G. *Note Sur la convergence de directions conjuguées*, Rev. Francaise Informat. Recherche Operationelle, 3e Annè. 16 (1969), 35–43.
- [51] Polyak, B.T. *The conjugate gradient method in extreme problems*, USSR Comp. Math. Math. Phys. 9 (1969), 94–112.
- [52] Powell, M.J.D. *Some global convergence properties of a variable metric algorithm for minimization without exact line search*, In: Cottle, R.W., Lemke, C.E. (eds.) Nonlinear Programming, SIAM-AMS Proceedings, SIAM, Philadelphia. 9 (1976), 53–72.
- [53] Shanno, D.F. *Conditioning of quasi-Newton methods for function minimization*, Math. Comput. 24 (1970), 647–656.
- [54] Solodov, M.V. and Svaiter, B.F. *A globally convergent inexact Newton method for systems of monotone equations*, in: M. Fukushima, L. Qi (Eds.), Reformulation: Nonsmooth, Piecewise Smooth, Semismooth and Smoothing Methods, Kluwer Academic Publishers. (1998), 355–369.
- [55] Sun, W. and Yuan, Y.X. *Optimization theory and methods*, Nonlinear programming, Springer, New York 2006.
- [56] Waziri, M.Y., Ahmed, K. and Halilu, A.S. *A modified Dai-Kou-type method with applications to signal reconstruction and blurred image restoration*, Comput. Appl. Math. 41(232) (2022), 1–33.
- [57] Waziri, M.Y., Ahmed, K., Halilu, A.S. and Sabi'u, J. *Two new Hager-Zhang iterative schemes with improved parameter choices for monotone nonlinear systems and their applications in compressed sensing*. Rairo Oper. Res. 56 (1) (2021), 239–273.
- [58] Waziri, M.Y., Ahmed, K. and Sabi'u, J. *A Dai-Liao conjugate gradient method via modified secant equation for system of nonlinear equations*, Arab. J. Math. 9 (2020), 443–457.
- [59] Waziri, M.Y., Ahmed, K., Sabi'u, J. and Halilu, A.S. *Enhanced Dai-Liao conjugate gradient methods for systems of monotone nonlinear equations*, SeMA J. 78 (2020), 15–51.

- [60] Winther, R. *Some superlinear convergence results for the conjugate gradient method*, SIAM J. Numer. Anal. 17 (1980), 14–17.
- [61] Xiao, Y., Wang, Q. and Hu, Q. *Non-smooth equations based method for  $\ell_1$  – norm problems with applications to compressed sensing*, Nonlinear Anal. Theory Methods Appl. 74(11) (2011), 3570–3577.
- [62] Xiao, Y. and Zhu, H. *A conjugate gradient method to solve convex constrained monotone equations with applications in compressive sensing*, J. Math. Anal. Appl. 405 (1) (2013), 310–319.
- [63] Yin, J., Jian, J., Jiang, X., Liu, M. and Wang, L. *A hybrid three-term conjugate gradient projection method for constrained nonlinear monotone equations with applications*, Numer. Algor. 88 (2021), 389–418.
- [64] Zhou, W.J. and Li, D.H. *Limited memory BFGS methods for nonlinear monotone equations*, J. Comput. Math. 25 (2007), 89–96.



# Combining the reproducing kernel method with Taylor series expansion to solve systems of nonlinear fractional Volterra integro-differential equations

T. Amoozad, S. Abbasbandy\*, , H. Sahihi, T. Allahviranloo

## Abstract

---

\*Corresponding author

Received 26 October 2024; revised 29 April 2025; accepted 30 April 2025

Taher Amoozad

Department of Mathematics, Science and Research Branch, Islamic Azad University, Tehran, Iran.

Saeid Abbasbandy

Department of Applied Mathematics, Faculty of Science, Imam Khomeini International University, Qazvin, Iran. e-mail: [abbasbandy@yahoo.com](mailto:abbasbandy@yahoo.com), [abbasbandy@sci.ikiu.ac.ir](mailto:abbasbandy@sci.ikiu.ac.ir).

Hussein Sahihi

Department of Mathematics, Science and Research Branch, Islamic Azad University, Tehran, Iran.

Tofiq Allahviranloo

Faculty of Engineering and Natural Sciences, Istinye University, Istanbul, Turkey.

---

## How to cite this article

Amoozad, T., Abbasbandy, S., Sahihi, H. and Allahviranloo, T., Combining the reproducing kernel method with Taylor series expansion to solve systems of nonlinear fractional Volterra integro-differential equations. *Iran. J. Numer. Anal. Optim.*, 2025; 15(3): 1210-1240. <https://doi.org/10.22067/ijnao.2025.90460.1542>

In this article, we present a novel approach for solving systems of nonlinear fractional Volterra integro-differential equations (NFVI-DEs) by reproducing the Hilbert kernel method. Kernel methods are powerful tools for addressing both linear and nonlinear problems. The reproducing kernel method stands out for its wide-ranging applications in solving complex scientific challenges. Our method combines the reproducing kernel method with a truncated Taylor series expansion, resulting in a more precise solution. This transformation converts the original NFVI-DEs into a system of nonlinear fractional differential equations. Our numerical results showcase this approach's effectiveness and align with theorems about error analysis.

**AMS subject classifications (2020):** 65R20, 26A33.

**Keywords:** Reproducing kernel method; Fractional Volterra integro-differential equations; Taylor series expansion; Error analysis.

## 1 Introduction

In the fields of physics, chemistry, biology, and other sciences, many phenomena can be accurately modeled by systems of nonlinear fractional-order Volterra integro-differential equations [1, 11, 13, 18]. Over the years, numerous scientists have attempted to solve these complex equations using various numerical methods, including the discrete Adomian decomposition method, perturbation-based approaches, the Chebyshev wavelet method, the Chebyshev spectral method, block-pulse functions, wavelet methods, the Legendre wavelet method, and multi-step collocation methods [8, 10, 16, 17, 22, 23, 25, 27].

Kernel methods are powerful techniques for solving linear and nonlinear problems. Notably, the reproducing kernel method (RKM) has many applications in tackling challenging scientific problems. Over the past decade, the RKM combined with the Gram-Schmidt orthogonalization process (G-SOP) has been widely used to solve systems of integral equations [15, 26]. However, recent trends have shifted toward using the RKM without the G-SOP due to its advantages, such as easier implementation, lower computational cost, and higher accuracy [19, 21]. Furthermore, a new RKM-based approach

that omits the G-SOP has been developed to solve a wide range of equations, including linear and nonlinear differential equations, integral equations, and systems of fractional-order Volterra integro-differential equations [5, 6, 4].

The RKM relies on several key components: the space, points, inner product, bases, and the chosen solution method. By adjusting these components to suit the specific problem, one can solve complex problems effectively. However, certain problems cannot be resolved simply by modifying these components. In such cases, an innovative approach is required to enhance the numerical results. One such approach involves combining the RKM with a Taylor series expansion. Alvandi and Paripour [2, 3] successfully applied this combined method to solve linear and nonlinear Volterra integro-differential equations. By employing the Taylor series expansion, they transformed the integro-differential equations into a system of differential equations, yielding more accurate numerical solutions.

In this paper, we present a novel method for solving systems of nonlinear fractional Volterra integro-differential equations (NFVI-DEs). Our approach combines the RKM without the G-SOP with Taylor series expansion. Additionally, we address cases where the approximate solution exhibits significant errors without this combined approach. For such problems, we apply Volterra's integral to the nonlinear component and replace it with a Taylor series expansion. This substitution substantially improves numerical accuracy while avoiding the need for complete transformation into a system of nonlinear fractional differential equations (NFDEs). Furthermore, we compare our method with the wavelet method [23], with numerical results demonstrating the superior effectiveness of our approach.

This article is structured as follows: In section 2, we introduce the concept of space and then proceed to prove the basic theorem and lemmas. Next, we present a new algorithm that utilizes linear algebra techniques. In section 3, we evaluate the error of this method. In section 4, we provide four examples that have been solved using this method and demonstrate its efficiency in terms of numerical results compared to other methods. Finally, we conclude in the last section.

Consider the following systems of NFVI-DEs for  $\tau \in [0, 1]$ :

$$\begin{cases} L_{11}\gamma_1(\tau) + L_{12}\gamma_2(\tau) = g_1(\tau) - \lambda_1(\tau, \gamma(\tau), \gamma'(\tau), \int_a^\tau k_1(x, \gamma(x), \gamma'(x)) dx), \\ L_{21}\gamma_1(\tau) + L_{22}\gamma_2(\tau) = g_2(\tau) - \lambda_2(\tau, \gamma(\tau), \gamma'(\tau), \int_a^\tau k_2(x, \gamma(x), \gamma'(x)) dx), \\ \gamma_i(0) = \theta_i, \quad i = 1, 2. \end{cases} \quad (1)$$

The operators  $L_{i,j}$ ,  $i, j = 1, 2$ , and  $\theta_d(\cdot, \cdot)$  are linear and nonlinear operators, respectively. Additionally,  $g_d(\cdot)$  are predetermined functions for  $d = 1, 2$ , and  $\gamma(\cdot) = (\gamma_1(\cdot), \gamma_2(\cdot))^T$  are unknown vector functions to be determined.

In (1), we suppose

$$\begin{cases} L_{11}\gamma_1(\tau) = D^\alpha\gamma_1(\tau) - a_{11}\gamma_1(\tau) - \int_0^\tau k_{11}(\tau, x)\gamma_1(x) dx, \\ L_{12}\gamma_2(\tau) = -a_{12}\gamma_2(\tau) - \int_0^\tau k_{12}(\tau, x)\gamma_2(x) dx, \\ L_{21}\gamma_1(\tau) = -a_{21}\gamma_1(\tau) - \int_0^\tau k_{21}(\tau, x)\gamma_1(x) dx, \\ L_{22}\gamma_2(\tau) = D^\beta\gamma_2(\tau) - a_{22}\gamma_2(\tau) - \int_0^\tau k_{22}(\tau, x)\gamma_2(x) dx. \end{cases}$$

Suppose that  $0 < \alpha, \beta \leq 1$ ,  $D^\alpha\gamma_1(\tau)$  and  $D^\beta\gamma_2(\tau)$  represent Caputo fractional derivatives. Additionally,  $a_{ij}(\cdot)$  are given functions for  $i, j = 1, 2$ . In the nonlinear part of (1), we utilize a truncated Taylor series expansion centered at the point  $x$  within the interval  $[0, 1]$  instead of using  $\gamma(\tau)$  and  $\gamma'(\tau)$ . Therefore, we obtain the following:

$$\begin{cases} L_{11}\gamma_1(\tau) + L_{12}\gamma_2(\tau) = g_1(\tau) \\ \quad - \lambda_1(\tau, \gamma(\tau), \gamma'(\tau), \int_a^\tau k_1(x, \sum_{k=0}^m \frac{\gamma^{(k)}(\tau)(x-\tau)^{(k)}}{k!} \\ \quad , \sum_{k=1}^m \frac{\gamma^{(k)}(\tau)(x-\tau)^{(k-1)}}{(k-1)!} ) dx), \\ L_{21}\gamma_1(\tau) + L_{22}\gamma_2(\tau) = g_2(\tau) \\ \quad - \lambda_2(\tau, \gamma(\tau), \gamma'(\tau), \int_a^\tau k_2(x, \sum_{k=0}^m \frac{\gamma^{(k)}(\tau)(x-\tau)^{(k)}}{k!} \\ \quad , \sum_{k=1}^m \frac{\gamma^{(k)}(\tau)(x-\tau)^{(k-1)}}{(k-1)!} ) dx), \\ \gamma_i(0) = \theta_i, \quad i = 1, 2, \end{cases} \quad (2)$$

where  $\gamma^{(0)}(\tau) = \gamma(\tau)$  and  $\int_a^\tau k_2(x, \sum_{k=0}^m \frac{\gamma^{(k)}(\tau)(x-\tau)^{(k)}}{k!}, \sum_{k=1}^m \frac{\gamma^{(k)}(\tau)(x-\tau)^{(k-1)}}{(k-1)!} ) dx$  in term of  $\gamma(\tau)$  and its derivatives are computable. Therefore, let

$$\begin{cases} H_1(\tau, \gamma(\tau), \gamma'(\tau), \dots, \gamma^{(m)}(\tau)) \\ = \lambda_1(\tau, \gamma(\tau), \gamma'(\tau), \int_a^\tau k_1(x, \sum_{k=0}^m \frac{\gamma^{(k)}(\tau)(x-\tau)^{(k)}}{k!}, \sum_{k=1}^m \frac{\gamma^{(k)}(\tau)(x-\tau)^{k-1}}{(k-1)!}) dx), \\ H_2(\tau, \gamma(\tau), \gamma'(\tau), \dots, \gamma^{(m)}(\tau)) \\ = \lambda_2(\tau, \gamma(\tau), \gamma'(\tau), \int_a^\tau k_2(x, \sum_{k=0}^m \frac{\gamma^{(k)}(\tau)(x-\tau)^k}{k!}, \sum_{k=1}^m \frac{\gamma^{(k)}(\tau)(x-\tau)^{k-1}}{(k-1)!}) dx). \end{cases}$$

Eventually, we can write

$$\begin{cases} L_{11}\gamma_1(\tau) + L_{12}\gamma_2(\tau) = g_1(\tau) - H_1(\tau, \gamma(\tau), \gamma'(\tau), \dots, \gamma^{(m)}(\tau)), \\ L_{21}\gamma_1(\tau) + L_{22}\gamma_2(\tau) = g_2(\tau) - H_2(\tau, \gamma(\tau), \gamma'(\tau), \dots, \gamma^{(m)}(\tau)), \\ \gamma_i(0) = \theta_i, \quad i = 1, 2. \end{cases} \quad (3)$$

Using matrix notation, we define the linear operator  $\mathbf{L}$  as

$$\mathbf{L} = \begin{pmatrix} L_{11} & L_{12} \\ L_{21} & L_{22} \end{pmatrix},$$

and with  $\mathbf{G} = (g_1, g_2)$ ,  $\mathbf{H} = (H_1, H_2)$ , so (2) can be written in the following form:

$$\begin{cases} \mathbf{L}(\gamma(\tau)) = \mathbf{G}(\tau) - \mathbf{H}(\tau, \gamma(\tau), \gamma'(\tau), \dots, \gamma^{(m)}(\tau)), & 0 < \tau \leq 1, \\ \gamma_i(0) = \theta_i, \quad i = 1, 2. \end{cases} \quad (4)$$

In the nonlinear case, where  $\mathbf{H} \neq 0$ , we will examine (4) using the following iterative scheme:

$$\mathbf{L}(\gamma_n(\tau)) = \mathbf{G}(\tau) - \mathbf{H}(\tau, \gamma_{n-1}(\tau), \gamma'_{n-1}(\tau), \dots, \gamma_{n-1}^{(m)}(\tau)), \quad n = 2, 3, \dots, \quad (5)$$

with  $\mathbf{L}(\gamma_1(\tau)) = \mathbf{G}(\tau)$ ; see [9] for more details.

**Definition 1.1.** [20] The Caputo fractional derivative operator of order  $\alpha > 0$ , is

$$D^\alpha u(\tau) = \frac{1}{\Gamma(z-\alpha)} \int_0^\tau (\tau-x)^{z-\alpha-1} u^{(z)}(x) dx, \quad \tau > 0,$$

where  $z-1 < \alpha < z$ ,  $z \in \mathbb{N}$ .

## 2 Main idea

In this section, we will introduce the space and then proceed to prove the basic theorem and lemmas. Additionally, we will present a new algorithm

that utilizes linear algebra techniques. Now, we consider the Hilbert space

$$W_2^k[0, 1] = \{x | x^{(k-1)} \text{ is absolutely continuous, } x^{(k)} \in L^2[0, 1], x(0) = 0\},$$

with the inner product and norm as follows:

$$\langle x(\cdot), y(\cdot) \rangle_{W_2^k} = \sum_{i=0}^{k-1} x^{(i)}(0) y^{(i)}(0) + \int_0^1 x^{(k)}(\tau) y^{(k)}(\tau) d\tau,$$

$$\|x(\cdot)\|_{W_2^k} = \sqrt{\langle x, x \rangle_{W_2^k}}, \quad x(\cdot), y(\cdot) \in W_2^k[0, 1],$$

where  $k$  is a natural number. Also, we consider the Hilbert space

$$\mathbf{W}_2^6[0, 1] = W_2^6[0, 1] \oplus W_2^6[0, 1],$$

with the inner product and norm

$$\langle \mathbf{x}, \mathbf{y} \rangle_{\mathbf{W}_2^6} = \langle x_1, y_1 \rangle_{W_2^6} + \langle x_2, y_2 \rangle_{W_2^6}, \quad \|\mathbf{x}\|_{\mathbf{W}_2^6} = \left( \sum_{i=1}^2 \|x_i\|_{W_2^6}^2 \right)^{1/2},$$

where  $\mathbf{x} = (x_1, x_2)^T$ ,  $\mathbf{y} = (y_1, y_2)^T$ ,  $x_i, y_i \in W_2^6[0, 1]$ .

**Lemma 2.1.** If  $L_{i,j} : W_2^6[0, 1] \rightarrow W_2^1[0, 1]$  in (1) are bounded linear operators, then  $\mathbf{L} : \mathbf{W}_2^6[0, 1] \rightarrow \mathbf{W}_2^1[0, 1]$  is a bounded linear operator, where

$$\mathbf{L} = \begin{pmatrix} L_{11} & L_{12} \\ L_{21} & L_{22} \end{pmatrix},$$

and the boundedness of  $L_{ij}$  implies that  $\mathbf{L}$  is bounded, also the adjoint operator of  $\mathbf{L}$  is

$$\mathbf{L}^* = \begin{pmatrix} L_{11}^* & L_{12}^* \\ L_{21}^* & L_{22}^* \end{pmatrix},$$

where  $L_{ij}^*$  is the adjoint operator of  $L_{ij}$ , [12]. Indeed, according to (5), we have

$$\gamma \in \mathbf{W}_2^6[0, 1], \quad \mathbf{G} - \mathbf{H} \in \mathbf{W}_2^1[0, 1].$$

**Lemma 2.2.** The spaces  $W_2^4[0, 1]$ ,  $W_2^5[0, 1]$ ,  $W_2^6[0, 1]$ , and  $W_2^1[0, 1]$  are all reproducing kernel Hilbert spaces, with their respective reproducing kernels listed in Table 1.

Table 1: The reproducing kernels in the  $W_2^k[0, 1]$  space.

$k$	$W_2^k[0, 1]$
1	$\kappa_y(\tau) = \begin{cases} 1+y, & \tau \geq y, \\ 1+\tau, & \tau < y. \end{cases}$
4	$\kappa_y(\tau) = \begin{cases} -(\tau^7/5040) + \tau y + (\tau^6 y)/720 + (\tau^2 y^2)/4 - (\tau^5 y^2)/240 + (\tau^3 y^3)/36 + (\tau^4 y^3)/144, & \tau \geq y, \\ y^7/5040 + 1/144 \tau^3 (4y^3 + y^4) + 1/240 \tau^2 (60y^2 - y^5) + 1/720 \tau (720y + y^6), & \tau < y. \end{cases}$
5	$\kappa_y(\tau) = \begin{cases} \tau^9/362880 + \tau y - (\tau^8 y)/40320 + (\tau^2 y^2)/4 + (\tau^7 y^2)/10080 + (\tau^3 y^3)/36 - (\tau^6 y^3)/4320 \\ + (\tau^4 y^4)/576 + (\tau^5 y^4)/2880, & \tau \geq y, \\ y^9/362880 + (\tau^4 (5y^4 + y^5))/2880 + (\tau^3 (120y^3 - y^6))/4320 + (\tau^2 (2520y^2 + y^7))/10080 \\ + (\tau (40320y - y^8))/40320, & \tau < y. \end{cases}$
6	$\kappa_y(\tau) = \begin{cases} -(\tau^{11}/39916800) + \tau y + (\tau^{10} y)/3628800 + (\tau^2 y^2)/4 - (\tau^9 y^2)/725760 + (\tau^3 y^3)/36 + (\tau^8 y^3)/241920 \\ + (\tau^4 y^4)/576 - (\tau^7 y^4)/120960 + (\tau^5 y^5)/14400 + (\tau^6 y^5)/86400, & \tau \geq y, \\ -(\tau^{11}/39916800) + (\tau^3 (6y^5 + y^6))/86400 + (\tau^4 (210y^4 - y^7))/120960 + (\tau^3 (6720y^3 + y^8))/241920 \\ + (\tau^2 (181440y^2 - y^9))/725760 + (\tau (3628800y + y^{10}))/3628800, & \tau < y. \end{cases}$

Let  $\{\tau_l\}_{l=1}^\infty$  be a node dense set on  $[0, 1]$ . Then we can deduce that

$$\varphi_{lj}(\tau) = \tilde{\kappa}_\tau(\tau_l) \vec{e}_j^T = \begin{cases} (\tilde{\kappa}_\tau(\tau_l), 0)^T, & j = 1, \\ (0, \tilde{\kappa}_\tau(\tau_l))^T, & j = 2, \end{cases} \quad (6)$$

where  $\phi_{lj}(\tau)$  represents the reproducing kernels of  $\mathbf{W}_2^1[0, 1]$  and  $\mathbf{W}_2^6[0, 1]$ , respectively, and is defined as  $\mathbf{L}^* \varphi_{lj}(\tau)$ . Here,  $\vec{e}_j^T$  is a vector in  $\mathbb{R}^2$  with a value of 1 in the  $j$ th coordinate and 0 in all other coordinates, as stated in [9]. It has been proven in [6] that

$$\langle \phi_{si}(\cdot), \phi_{lj}(\cdot) \rangle_{\mathbf{W}_2^{3,3}} = \begin{cases} 0, & i \neq j, \\ \|\kappa_{\tau_s}\|^2, & s = l, i = j, \\ \kappa_{\tau_s}(\tau_l), & s \neq l, i = j. \end{cases} \quad (7)$$

**Theorem 2.1.** [6] For  $j = 1, 2$  and  $l = 1, 2, \dots$ ,

$$\phi_{lj}(\tau) = \mathbf{L} \kappa_{\tau_l}(\tau) \vec{e}_j^T.$$

**Lemma 2.3.** For each fixed  $N$ ,  $\{\phi_{lj}(\tau)\}_{(1,1)}^{(N,2)}$  is linearly independent in  $\mathbf{W}_2^6[0, 1]$ , [14].

**Theorem 2.2.** If  $\{\tau_s\}_{s=1}^\infty$  is dense on  $[0, 1]$  and the solution of (4) is unique, then this solution is

$$\gamma(\tau) = \sum_{l=1}^\infty \sum_{j=1}^2 c_{j,l} \phi_{lj}(\tau). \quad (8)$$

*Proof.* Substituting (8) into (4), then for  $i = 1$  or 2

$$\begin{aligned}
L\gamma(\tau_s) &= \langle L\gamma(\cdot), \varphi_{si}(\cdot) \rangle_{W_2^1} = \langle \gamma(\cdot), L^* \varphi_{si}(\cdot) \rangle_{W_2^6} \\
&= \left\langle \sum_{l=1}^{\infty} \sum_{j=1}^2 c_{j,l} \phi_{lj}(\cdot), \phi_{si}(\cdot) \right\rangle_{W_2^6} \\
&= \sum_{l=1}^{\infty} \sum_{j=1}^2 c_{j,l} \langle \phi_{lj}(\cdot), \phi_{si}(\cdot) \rangle_{W_2^6} \\
&= \sum_{l=1}^{\infty} \sum_{j=1}^2 c_{j,l} \phi_{lj}(\tau_s) \\
&= G(\tau_s) - H(\tau_s, \gamma(\tau_s), \gamma'(\tau_s), \dots, \gamma^{(m)}(\tau_s)).
\end{aligned}$$

In addition, we have

$$\begin{aligned}
G(\tau_s) - H(\tau_s, \gamma(\tau_s), \dots, \gamma^{(m)}(\tau_s)) &= \left\langle G(\tau) - H(\tau, \gamma(\tau), \dots, \gamma^{(m)}(\tau)), \varphi_{si}(\tau) \right\rangle_{W_2^1} \\
&= \left\langle G(\tau) - H(\tau, \gamma(\tau), \dots, \gamma^{(m)}(\tau)), \tilde{\mathbf{K}}_{\tau}(\tau_s) \vec{e}_i^{\top} \right\rangle_{W_2^1} \\
&= \langle L\gamma(\tau), \tilde{\mathbf{K}}_{\tau}(\tau_s) \vec{e}_i^{\top} \rangle_{W_2^1} \\
&= \langle \gamma(\tau), L^* \tilde{\mathbf{K}}_{\tau}(\tau_s) \vec{e}_i^{\top} \rangle_{W_2^6} \\
&= \langle \gamma(\tau), \phi_{si}(\tau) \rangle_{W_2^6} \\
&= \gamma(\tau_s) \\
&= \sum_{l=1}^{\infty} \sum_{j=1}^2 c_{j,l} \phi_{lj}(\tau_s).
\end{aligned}$$

Therefore,  $\gamma(\cdot)$  is the solution to (4), where  $c_{j,l}$  for  $j = 1, 2$  and  $l = 1, \dots$  are the unknown coefficients to be determined.  $\square$

We denote the numerical solution of  $\gamma$  by

$$\gamma_N(\tau) = \sum_{l=1}^N \sum_{j=1}^2 c_{j,l} \phi_{lj}(\tau), \quad (9)$$

where  $c_{j,l}$  are the unknown numbers to be determined, and  $N$  is the number of collocation points on  $[0, 1]$ . In the following, we aim to obtain a matrix notation for the unknowns in (8) using the iterative scheme in (5) for the nonlinear case. Therefore, the numerical solution is as follows:

$$\gamma_{n,N}(\tau) = \sum_{l=1}^N \sum_{j=1}^2 c_{j,l,n} \phi_{lj}(\tau), \quad n = 2, 3, \dots, \quad (10)$$

where  $n$  represents the iteration number, the coefficients  $c_{j,l,n}$  are obtained as follows: By substituting (10) into (4) and for a sufficiently large value of  $N$ , we obtain the following:

$$\mathbf{L}\gamma_{n,N}(\tau) = \mathbf{G}(\tau) - \mathbf{H}(\tau, \gamma_{n-1,N}(\tau), \gamma'_{n-1,N}(\tau), \dots, \gamma_{n-1,N}^{(m)}(\tau)).$$

According to Theorem 2.2, we can write

$$\sum_{l=1}^N \sum_{j=1}^2 c_{j,l,n} \phi_{lj}(\tau_s) = \mathbf{G}(\tau_s) - \mathbf{H}(\tau_s, \gamma_{n-1,N}(\tau_s), \gamma'_{n-1,N}(\tau_s), \dots, \gamma_{n-1,N}^{(m)}(\tau_s)), \quad (11)$$

where  $s = 1, 2, \dots, N$  is number of collocation points. Now, using Theorem 2.1 we have

$$\begin{aligned} \sum_{l=1}^N \sum_{j=1}^2 c_{j,l,n} \phi_{lj}(\tau_s) &= \sum_{l=1}^N c_{1,l,n} \phi_{l1}(\tau_s) + \sum_{l=1}^N c_{2,l,n} \phi_{l2}(\tau_s) \\ &= \sum_{l=1}^N c_{1,l,n} (\mathbf{L}\boldsymbol{\kappa}_{\tau_l}(\tau_s) \vec{e}_1) + \sum_{l=1}^N c_{2,l,n} (\mathbf{L}\boldsymbol{\kappa}_{\tau_l}(\tau_s) \vec{e}_2) \\ &= \mathbf{L} \sum_{l=1}^N c_{1,l,n} \boldsymbol{\kappa}_{\tau_l}(\tau_s) \vec{e}_1 + \mathbf{L} \sum_{l=1}^N c_{2,l,n} \boldsymbol{\kappa}_{\tau_l}(\tau_s) \vec{e}_2 \\ &= \begin{pmatrix} L_{11} & L_{12} \\ L_{21} & L_{22} \end{pmatrix} \begin{pmatrix} \sum_{l=1}^N c_{1,l,n} \boldsymbol{\kappa}_{\tau_l}(\tau_s) \\ 0 \end{pmatrix} \\ &\quad + \begin{pmatrix} L_{11} & L_{12} \\ L_{21} & L_{22} \end{pmatrix} \begin{pmatrix} 0 \\ \sum_{l=1}^N c_{2,l,n} \boldsymbol{\kappa}_{\tau_l}(\tau_s) \end{pmatrix} \\ &= \begin{pmatrix} L_{11} \sum_{l=1}^N c_{1,l,n} \boldsymbol{\kappa}_{\tau_l}(\tau_s) \\ L_{21} \sum_{l=1}^N c_{1,l,n} \boldsymbol{\kappa}_{\tau_l}(\tau_s) \end{pmatrix} + \begin{pmatrix} L_{12} \sum_{l=1}^N c_{2,l,n} \boldsymbol{\kappa}_{\tau_l}(\tau_s) \\ L_{22} \sum_{l=1}^N c_{2,l,n} \boldsymbol{\kappa}_{\tau_l}(\tau_s) \end{pmatrix} \\ &= \begin{pmatrix} L_{11} \sum_{l=1}^N c_{1,l,n} \boldsymbol{\kappa}_{\tau_l}(\tau_s) + L_{12} \sum_{l=1}^N c_{2,l,n} \boldsymbol{\kappa}_{\tau_l}(\tau_s) \\ L_{21} \sum_{l=1}^N c_{1,l,n} \boldsymbol{\kappa}_{\tau_l}(\tau_s) + L_{22} \sum_{l=1}^N c_{2,l,n} \boldsymbol{\kappa}_{\tau_l}(\tau_s) \end{pmatrix} \\ &= \begin{pmatrix} \sum_{l=1}^N L_{11} \boldsymbol{\kappa}_{\tau_l}(\tau_s) & \sum_{l=1}^N L_{12} \boldsymbol{\kappa}_{\tau_l}(\tau_s) \\ \sum_{l=1}^N L_{21} \boldsymbol{\kappa}_{\tau_l}(\tau_s) & \sum_{l=1}^N L_{22} \boldsymbol{\kappa}_{\tau_l}(\tau_s) \end{pmatrix} \begin{pmatrix} c_{1,l,n} \\ c_{2,l,n} \end{pmatrix}. \end{aligned}$$

Also,

$$\begin{aligned} & \mathbf{G}(\tau_s) - \mathbf{H}(\tau_s, \gamma_{n-1,N}(\tau_s), \gamma'_{n-1,N}(\tau_s), \dots, \gamma_{n-1,N}^{(m)}(\tau_s)) \\ &= \begin{pmatrix} g_1(\tau_s) \\ g_2(\tau_s) \end{pmatrix} - \begin{pmatrix} H_1(\tau_s, \gamma_{n-1,N}(\tau_s), \gamma'_{n-1,N}(\tau_s), \dots, \gamma_{n-1,N}^{(m)}(\tau_s)) \\ H_2(\tau_s, \gamma_{n-1,N}(\tau_s), \gamma'_{n-1,N}(\tau_s), \dots, \gamma_{n-1,N}^{(m)}(\tau_s)) \end{pmatrix}. \end{aligned}$$

So, according to (11) we can deduce that

$$\begin{aligned} & \begin{pmatrix} \sum_{l=1}^N L_{11} \kappa_{\tau_l}(\tau_s) & \sum_{l=1}^N L_{12} \kappa_{\tau_l}(\tau_s) \\ \sum_{l=1}^N L_{21} \kappa_{\tau_l}(\tau_s) & \sum_{l=1}^N L_{22} \kappa_{\tau_l}(\tau_s) \end{pmatrix} \begin{pmatrix} c_{1,l,n} \\ c_{2,l,n} \end{pmatrix} \\ &= \begin{pmatrix} g_1(\tau_s) \\ g_2(\tau_s) \end{pmatrix} - \begin{pmatrix} H_1(\tau_s, \gamma_{n-1,N}(\tau_s), \gamma'_{n-1,N}(\tau_s), \dots, \gamma_{n-1,N}^{(m)}(\tau_s)) \\ H_2(\tau_s, \gamma_{n-1,N}(\tau_s), \gamma'_{n-1,N}(\tau_s), \dots, \gamma_{n-1,N}^{(m)}(\tau_s)) \end{pmatrix}. \end{aligned}$$

Then

$$\begin{aligned} \mathbf{A} &= \begin{pmatrix} \sum_{l=1}^N L_{11} \kappa_{\tau_l}(\tau_s) & \sum_{l=1}^N L_{12} \kappa_{\tau_l}(\tau_s) \\ \sum_{l=1}^N L_{21} \kappa_{\tau_l}(\tau_s) & \sum_{l=1}^N L_{22} \kappa_{\tau_l}(\tau_s) \end{pmatrix}_{2N \times 2N}, \\ \mathcal{C} &= \begin{pmatrix} c_{1,l,n} \\ c_{2,l,n} \end{pmatrix}_{2N \times 1}, \\ \mathbf{M} &= \begin{pmatrix} g_1(\tau_s) \\ g_2(\tau_s) \end{pmatrix}_{2N \times 1} - \begin{pmatrix} H_1(\tau_s, \gamma_{n-1,N}(\tau_s), \gamma'_{n-1,N}(\tau_s), \dots, \gamma_{n-1,N}^{(m)}(\tau_s)) \\ H_2(\tau_s, \gamma_{n-1,N}(\tau_s), \gamma'_{n-1,N}(\tau_s), \dots, \gamma_{n-1,N}^{(m)}(\tau_s)) \end{pmatrix}_{2N \times 1}. \end{aligned}$$

Therefore, we can write

$$\mathbf{A} \mathcal{C} = \mathbf{M}.$$

Finally, according to Lemma 2.3  $\mathbf{A}^{-1}$  exists and

$$\mathcal{C} = \mathbf{A}^{-1} \mathbf{M}.$$

### 3 Error estimation

**Lemma 3.1.** Let  $S = \left\{ \gamma(\cdot) = (\gamma_1(\cdot), \gamma_2(\cdot)) \mid \|\gamma\|_{\mathbf{W}_2^6} \leq \delta \right\}$ . Then  $S$  is a compact set in the space  $C^2[0, 1]$ , where  $\delta$  is a constant [24].

**Lemma 3.2.** Assuming that in system (4), the norm of  $\gamma$  in  $\mathbf{W}_2^6$  is bounded,  $\{\tau_s\}_{s=1}^\infty$  is a dense set on  $[0, 1]$ ,  $\mathbf{L}(\gamma(\cdot))$  is a continuous function of  $\gamma(\cdot)$  that is also invertible, and  $\mathbf{H}(\cdot, \gamma(\cdot), \gamma'(\cdot), \dots, \gamma^{(m)}(\cdot))$  is a continuous function

of  $\gamma(\cdot)$ , then both the analytical solution  $\gamma(\cdot)$  and the numerical solution  $\gamma_{n,N}(\cdot)$  for (4) exist, [24].

**Theorem 3.1.** If  $\gamma(\cdot) = (\gamma_1(\cdot), \gamma_2(\cdot)) \in \mathbf{W}_2^6[0, 1]$  is the solution of (4), then the numerical solution  $\gamma_{n,N}(\cdot) = (\gamma_{1,n,N}(\cdot), \gamma_{2,n,N}(\cdot))$  converges uniformly to  $\gamma(\cdot)$ .

*Proof.* By subtracting the two equations in (3), we can obtain the following form:

$$D^\alpha \gamma_1(\tau) - \bar{a}(\tau) \gamma_1(\tau) - \int_0^\tau k_{11}(\tau, x) \gamma_1(x) dx = \bar{g}(\tau) - \bar{H}(\tau, \gamma(\tau), \gamma'(\tau), \dots, \gamma^{(m)}(\tau)), \quad (12)$$

where  $\bar{a}(\cdot)$ ,  $\bar{g}(\cdot)$  and  $\bar{H}(\cdot)$  are known functions and (12) is a nonlinear equation in the reproducing kernel space  $W_2^6[0, 1]$ . Furthermore, according to Lemma 3.2,  $\gamma_{1,n,N}(\cdot)$  is a numerical solution of  $\gamma_1(\cdot)$ . Hence,

$$\begin{aligned} |\gamma_1(\tau) - \gamma_{1,n,N}(\tau)| &= |\langle \gamma_1 - \gamma_{1,n,N}, \kappa_\tau \rangle_{W_2^6}| \leq \|\gamma_1 - \gamma_{1,n,N}\|_{W_2^6} \|\kappa_\tau\|_{W_2^6} \\ &\leq Q_1 \|\gamma_1 - \gamma_{1,n,N}\|_{W_2^6}, \end{aligned}$$

where  $Q_1$  is constant. Similarly, we have

$$|\gamma_2(\tau) - \gamma_{2,n,N}(\tau)| \leq Q_2 \|\gamma_2 - \gamma_{2,n,N}\|_{W_2^6}.$$

□

**Theorem 3.2.** If  $\gamma_{1,n,N}(\tau) \xrightarrow{\|\cdot\|_{W_2^k}} \gamma_1(\tau)$  and  $\tau_s \rightarrow y (s \rightarrow \infty)$ , then

$$\bar{H}(\tau_s, \gamma_N(\tau_s), \gamma'_N(\tau_s), \dots, \gamma_N^{(m)}(\tau_s)) \rightarrow \bar{H}(y, \gamma(y), \gamma'(y), \dots, \gamma^{(m)}(y))(s \rightarrow \infty).$$

*Proof.* See [3].

□

**Theorem 3.3.** Let  $\gamma(\cdot) = (\gamma_1(\cdot), \gamma_2(\cdot))$  and  $\gamma_{n,N}(\cdot) = (\gamma_{1,n,N}(\cdot), \gamma_{2,n,N}(\cdot))$  be the analytical and numerical solution of (4), respectively. If  $\gamma \in C^6[0, 1]$ ,  $\gamma_{n,N} \in \mathbf{W}_2^6[0, 1]$  and  $\left\| \gamma_{i,n,N}^{(6)} \right\|_\infty \leq M_i$ ,  $i = 1, 2$ , then for  $j = 1, 2$ ,

$$\|\gamma_j - \gamma_{j,n,N}\|_\infty \leq C_j h^6,$$

where  $C_j$  is a constant.

*Proof.* See [7].

□

**Remark 3.1.** The accuracy of the RKM method is affected by the choice of space. Therefore, it is crucial to carefully select the appropriate space based on the specific problem at hand. It is important to note that changing the space can also affect the convergence order. For instance, if we choose  $\mathbf{W}_2^5[0, 1]$  for a particular problem, then the convergence order will be  $Ch^5$ .

**Remark 3.2.** According to Lemma 2.3, if  $\mathbf{A}^{-1}$  exists, then the solution of (4) also exists and is unique. Additionally, we can conclude that the present method is stable in  $\mathbf{W}_2^6[0, 1]$ .

**Remark 3.3.** The formula for convergence order is as follows:

$$C.F_i = \log_2 \frac{\|\gamma_i - \gamma_{i,n,N}\|_\infty}{\|\gamma_i - \gamma_{i,n,2N}\|_\infty},$$

where  $i = 1, 2$ .

## 4 Numerical results

In this section, we will demonstrate the application of the present method in solving four examples of NFVI-DEs. Additionally, we will compare the effectiveness of the present method with the method presented in [23]. In the first example, we will introduce the present method without using the Taylor series expansion and showcase its effectiveness in solving certain problems. However, we will also demonstrate that this method is not effective for solving problems where the Volterra integral is applied to nonlinear components. The following examples have been solved using Mathematica 12 software.

**Example 4.1.** [23] Consider the NFVI-DEs:

$$\begin{cases} D^\alpha \gamma_1(\tau) - \frac{1}{3} \gamma_1(\tau) \gamma_2(\tau) - \frac{1}{2} \gamma_2^2(\tau) - 2\gamma_2(\tau) + \int_0^\tau [\gamma_1(x) - \gamma_2(x)] dx = g_1(\tau), \\ D^\beta \gamma_2(\tau) - \frac{1}{3} \gamma_1(\tau) \gamma_2(\tau) + \gamma_1(\tau) + \int_0^\tau [\gamma_1(x) - 2\gamma_2(x)] dx = g_2(\tau), \\ \gamma_1(0) = 0, \gamma_2(0) = 0, \end{cases}$$

where

$$0 < \alpha, \beta \leq 1,$$

and the analytical solution for  $\alpha = \beta = 1$  is

$$\gamma(\tau) = (\tau^2, \tau).$$

We solved this example in the  $\mathbf{W}_2^5[0, 1]$  space using  $\tau_l = \frac{l}{N+1}$  points. In this example, the Volterra integral is not applied to the nonlinear component, so there is no need to use the Taylor series expansion. As a result, (1) can be rewritten as follows:

$$\begin{cases} L_{11}\gamma_1(\tau) + L_{12}\gamma_2(\tau) = g_1(\tau) - \lambda_1(\tau, \gamma(\tau)), \\ L_{21}\gamma_1(\tau) + L_{22}\gamma_2(\tau) = g_2(\tau) - \lambda_2(\tau, \gamma(\tau)), \\ \gamma_i(0) = \theta_i, \quad i = 1, 2. \end{cases} \quad (13)$$

Therefore, the matrix form of (13) is

$$\begin{cases} \mathbf{L}(\gamma(\tau)) = \mathbf{G}(\tau) - \boldsymbol{\lambda}(\tau, \gamma(\tau)), & 0 < \tau \leq 1, \\ \gamma_i(0) = \theta_i, \quad i = 1, 2, \end{cases} \quad (14)$$

where  $\mathbf{G} = (g_1, g_2)$  and  $\boldsymbol{\lambda} = (\lambda_1, \lambda_2)$ . Finally, the unknown coefficients can be determined using the following equation:

$$\begin{pmatrix} \sum_{l=1}^N L_{11}\kappa_{\tau_l}(\tau_s) & \sum_{l=1}^N L_{12}\kappa_{\tau_l}(\tau_s) \\ \sum_{l=1}^N L_{21}\kappa_{\tau_l}(\tau_s) & \sum_{l=1}^N L_{22}\kappa_{\tau_l}(\tau_s) \end{pmatrix} \begin{pmatrix} c_{1,l,n} \\ c_{2,l,n} \end{pmatrix} = \begin{pmatrix} g_1(\tau_s) \\ g_2(\tau_s) \end{pmatrix} - \begin{pmatrix} \lambda_1(\tau_s, \gamma_{n-1,N}(\tau_s)) \\ \lambda_2(\tau_s, \gamma_{n-1,N}(\tau_s)) \end{pmatrix}.$$

Next, we compared this method with the method proposed in [23], which is based on absolute error and numerical solution. The results are presented in Tables 2, 3, and 4 and Figures 1, 2, and 3. The convergence order is also shown in Table 5. These numerical results demonstrate the efficiency of this method without the use of Taylor series expansion. However, it should be noted that this method may not be as efficient in problems where the Volterra integral is applied to nonlinear components.

**Example 4.2.** [23] Consider the NFVI-DEs:

$$\begin{cases} D^\alpha \gamma_1(\tau) - \gamma_1^2(\tau) - \gamma_2^2(\tau) + \int_0^\tau \gamma_1(x) dx = g_1(\tau), \\ D^\beta \gamma_2(\tau) + \frac{1}{2} \gamma_2^2(\tau) + \gamma_1(\tau) + \int_0^\tau \gamma_1(x) \gamma_2(x) dx = g_2(\tau), \\ \gamma_1(0) = 0, \quad \gamma_2(0) = 1, \end{cases}$$

where

$$0 < \alpha, \beta \leq 1,$$

Table 2: Error comparison in Example 4.1 for  $\alpha = \beta = 1$ .

$\tau$	M in [23] $ \gamma_1(\tau) - \gamma_{1,64}(\tau) $	M in [23] $ \gamma_2(\tau) - \gamma_{2,64}(\tau) $	$\mathbf{W}_2^5[0, 1]$ $ \gamma_1(\tau) - \gamma_{1,10,24}(\tau) $	$\mathbf{W}_2^5[0, 1]$ $ \gamma_2(\tau) - \gamma_{2,10,24}(\tau) $
0	$1.04 \times 10^{-5}$	$2.97 \times 10^{-9}$	0	0
0.1	$1.22 \times 10^{-4}$	$3.04 \times 10^{-9}$	$2.19 \times 10^{-11}$	$2.48 \times 10^{-12}$
0.2	$3.24 \times 10^{-5}$	$4.31 \times 10^{-6}$	$2.10 \times 10^{-11}$	$4.89 \times 10^{-12}$
0.3	$1.07 \times 10^{-4}$	$9.06 \times 10^{-6}$	$1.94 \times 10^{-11}$	$7.41 \times 10^{-12}$
0.4	$1.03 \times 10^{-5}$	$1.41 \times 10^{-5}$	$1.70 \times 10^{-11}$	$1.00 \times 10^{-11}$
0.5	$1.85 \times 10^{-5}$	$1.95 \times 10^{-5}$	$1.39 \times 10^{-11}$	$1.27 \times 10^{-11}$
0.6	$2.48 \times 10^{-6}$	$3.06 \times 10^{-5}$	$9.90 \times 10^{-12}$	$1.55 \times 10^{-11}$
0.7	$7.08 \times 10^{-6}$	$3.62 \times 10^{-5}$	$5.00 \times 10^{-12}$	$1.83 \times 10^{-11}$
0.8	$6.01 \times 10^{-6}$	$4.17 \times 10^{-5}$	$8.83 \times 10^{-13}$	$2.13 \times 10^{-11}$
0.9	$2.94 \times 10^{-6}$	$4.70 \times 10^{-5}$	$8.68 \times 10^{-12}$	$2.46 \times 10^{-11}$
1	$1.97 \times 10^{-4}$	$5.20 \times 10^{-5}$	$2.68 \times 10^{-11}$	$3.22 \times 10^{-11}$

Table 3: Comparison the numerical solutions  $\gamma_1(\cdot)$  for different value of  $\alpha$  in Example 4.1, when  $n = 10$ ,  $N = 16$ .

$\tau$	$\alpha = 0.7$		$\alpha = 0.8$		$\alpha = 0.9$		$\alpha = 1$	
	M in [23]	$\mathbf{W}_2^5[0, 1]$	M in [23]	$\mathbf{W}_2^5[0, 1]$	M in [23]	$\mathbf{W}_2^5[0, 1]$	M in [23]	$\mathbf{W}_2^5[0, 1]$
0	-0.037573	0	-0.0038780	0	-0.0029321	0	-0.0019508	0
0.1	0.0622484	0.0597477	0.0351106	0.0334030	0.0193767	0.0183878	0.0105184	0.0100000
0.2	0.1589985	0.1616368	0.1045106	0.1035549	0.0667343	0.0648684	0.0417159	0.0400000
0.3	0.2689171	0.2834621	0.1938697	0.1986078	0.1349966	0.1350721	0.0916416	0.0900000
0.4	0.3869518	0.4164330	0.2991345	0.3128956	0.2222319	0.2266003	0.1602961	0.1600000
0.5	0.10196899	0.5547315	0.8345105	0.4423304	0.6537138	0.3376242	0.2503604	0.2500000
0.6	0.6369025	0.6939084	0.5492630	0.5834873	0.4526805	0.4665406	0.3600394	0.3600000
0.7	0.7667133	0.8303810	0.6901640	0.7332801	0.5932708	0.6118313	0.4911326	0.4900000
0.8	0.8998356	0.9612530	0.8394357	0.8888207	0.7480303	0.7719908	0.6409619	0.6400000
0.9	1.0371178	1.0842653	0.9968868	1.0473609	0.9165373	0.9454807	0.8095297	0.8100000
1	1.1798913	1.1978056	1.1626464	1.2062802	1.0984737	1.1306994	0.9968391	1.0000000

Table 4: Comparison the numerical solutions  $\gamma_2(\cdot)$  for different values of  $\beta$  in Example 4.1, when  $n = 10$ ,  $N = 16$ .

$\tau$	$\alpha = 0.7$		$\alpha = 0.8$		$\alpha = 0.9$		$\alpha = 1$	
	M in [23]	$\mathbf{W}_2^5[0, 1]$	M in [23]	$\mathbf{W}_2^5[0, 1]$	M in [23]	$\mathbf{W}_2^5[0, 1]$	M in [23]	$\mathbf{W}_2^5[0, 1]$
0	0.355758	0	0.0170683	0	0.0061149	0	$7.21E-07$	0
0.1	0.2153236	0.2046448	0.1689943	0.1624234	0.1306502	0.1277562	0.9999993	0.1000000
0.2	0.3372475	0.3323294	0.2886770	0.2850171	0.2419557	0.2401117	0.1998547	0.2000000
0.3	0.4360154	0.4330255	0.3939580	0.3913621	0.3468649	0.3454015	0.2997735	0.3000000
0.4	0.5189111	0.5172701	0.4886216	0.4869908	0.4468382	0.4457686	0.3996882	0.4000000
0.5	1.1814930	0.5898213	1.1502742	0.5744086	1.0855705	0.5421777	0.4998003	0.5000000
0.6	0.6528540	0.6535326	0.6534302	0.6550868	0.6342915	0.6351784	0.5995105	0.6000000
0.7	0.7104426	0.7104326	0.7271108	0.7300247	0.7230138	0.7251225	0.6994215	0.7000000
0.8	0.7651144	0.7621614	0.7971754	0.7999801	0.8093225	0.8122505	0.7993340	0.8000000
0.9	0.8181686	0.8101849	0.8644942	0.8655891	0.8935603	0.8967336	0.8992493	0.9000000
1	0.8708896	0.8559100	0.9300103	0.9274389	0.9761266	0.9786977	0.9991684	1.0000000

and the analytical solution for  $\alpha = \beta = 1$  is

$$\gamma(\tau) = (\sin(\tau), \cos(\tau)).$$

Table 5: Convergence order in Example 4.1.

$\alpha = \beta = 1, n = 10, \text{ in } \mathbf{W}_2^5[0, 1]$					
$N$	$\ \gamma_1 - \gamma_{1,n,N}\ _\infty$	$C.F_1$	$\ \gamma_2 - \gamma_{2,n,N}\ _\infty$	$C.F_2$	$Cpu\ time(sec)$
4	$2.42 \times 10^{-6}$	—	$3.12 \times 10^{-6}$	—	2
8	$2.74 \times 10^{-8}$	6.46	$3.44 \times 10^{-8}$	6.34	4
16	$3.36 \times 10^{-10}$	6.50	$4.12 \times 10^{-10}$	6.38	8

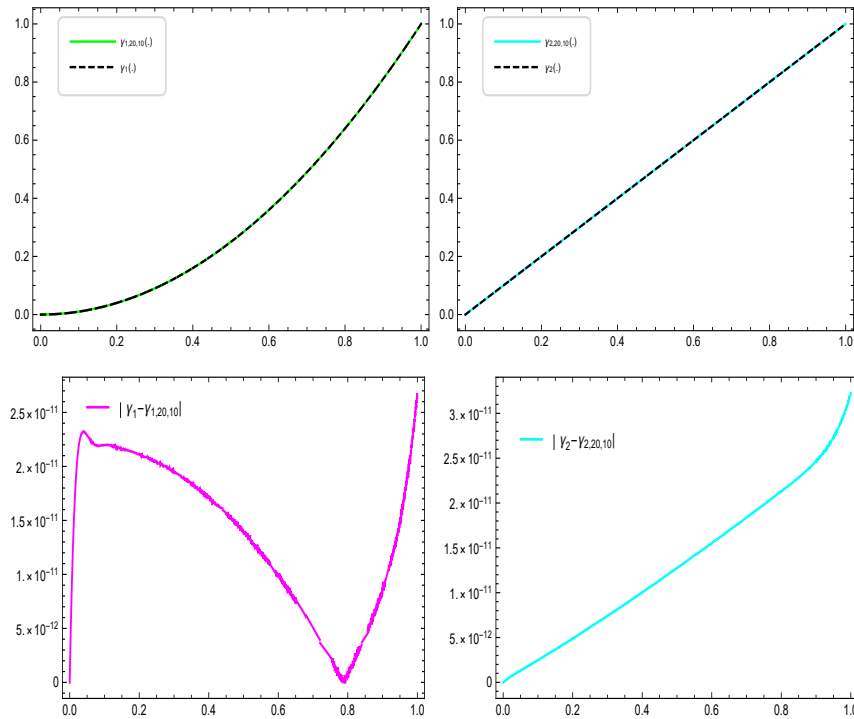


Figure 1: Numerical solution and absolute error without using the Taylor series expansion in Example 4.1.

First, we solved this example in the  $\mathbf{W}_2^6[0, 1]$  space without using the Taylor expansion, which utilizes  $\tau_l = \frac{l}{N+1}$  points. The numerical solution and absolute error are shown in Figure 4. However, this method is not effective. To improve the results, we compared the Present method, which uses the Taylor expansion, with the method proposed in [23]. This comparison was

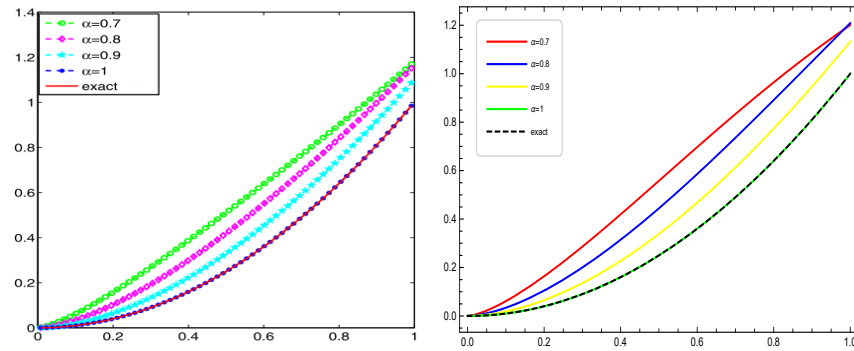


Figure 2: Comparison of the numerical solutions without using the Taylor series expansion in Example 4.1 for  $\alpha = 0.65, 0.75, 0.85, 1$ . (Left:  $\gamma_{1,64}(\cdot)$ , M in [23]; Right:  $\gamma_{1,20,24}(\cdot)$ , Present method).

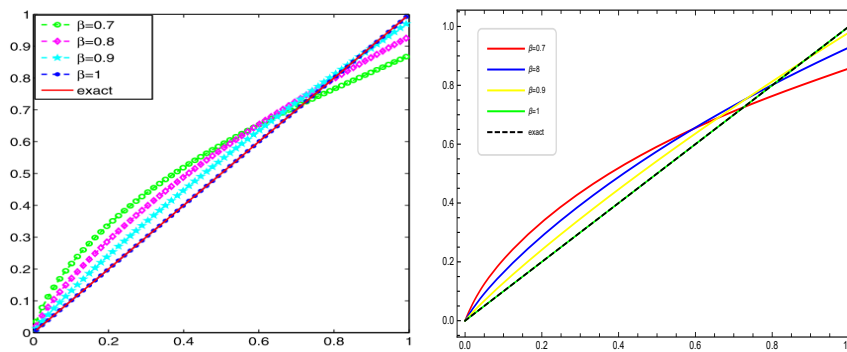


Figure 3: Comparison of the numerical solutions without using Taylor series expansion in Example 4.1 for  $\beta = 0.65, 0.75, 0.85, 1$ . (Left:  $\gamma_{2,64}(\cdot)$ , M in [23]; Right:  $\gamma_{2,20,24}(\cdot)$ , Present method).

based on the absolute error and numerical solution, as shown in Table 6 and Figures 5, 6, and 7. The convergence order is also shown in Table 7. Finally, to solve this example using (2) and (3) instead of  $\gamma_1(\cdot)$  and  $\gamma_2(\cdot)$ , we utilized the truncated Taylor series expansion around the point  $x$  in the interval  $[0, 1]$  for  $m = 3$ . This allowed us to obtain the nonlinear parts of  $H_2(\tau, \gamma(\cdot), \gamma'(\cdot), \gamma''(\cdot), \gamma^{(3)}(\cdot))$  as follows:

$$\frac{1}{2}\gamma_2^2(\tau) + \tau\gamma_1(\tau)\gamma_2(\tau) - \frac{1}{2}\tau^2\gamma_2(\tau)\gamma_1'(\tau) - \frac{1}{2}\tau^2\gamma_1(\tau)\gamma_2'(\tau) + \frac{1}{3}\tau^3\gamma_1(\tau)\gamma_2(\tau)$$

$$\begin{aligned}
& + \frac{1}{6} \tau^3 \gamma_2(\tau) \gamma_1''(\tau) - \frac{1}{8} \tau^4 \gamma_2(\tau) \gamma_1''(\tau) + \frac{1}{6} \tau^3 \gamma_1(\tau) \gamma_2''(\tau) - \frac{1}{8} \tau^4 \gamma_1(\tau) \gamma_2''(\tau) \\
& + \frac{1}{20} \tau^5 \gamma_1''(\tau) \gamma_2''(\tau) - \frac{1}{24} \tau^4 \gamma_2(\tau) \gamma_1^{(3)}(\tau) + \frac{1}{30} \tau^5 \gamma_2'(\tau) \gamma_1^{(3)}(\tau) - \frac{1}{72} \tau^6 \gamma_2''(\tau) \gamma_1^{(3)}(\tau) \\
& - \frac{1}{24} \tau^4 \gamma_1(\tau) \gamma_2^{(3)}(\tau) + \frac{1}{30} \tau^5 \gamma_1'(\tau) \gamma_2^{(3)}(\tau) - \frac{1}{72} \tau^6 \gamma_1''(\tau) \gamma_2^{(3)}(\tau) + \frac{1}{252} \tau^7 \gamma_1^{(3)}(\tau) \gamma_2^{(3)}(\tau).
\end{aligned}$$

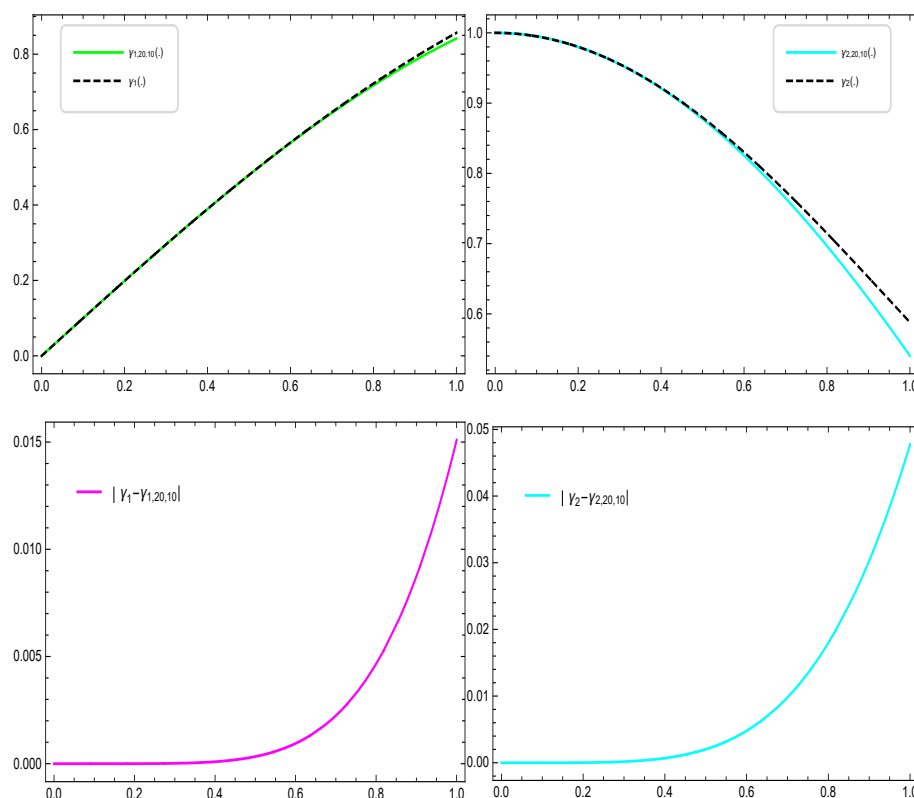


Figure 4: Numerical solution and absolute error without using the Taylor series expansion in Example 4.2.

**Example 4.3.** Consider the NFVI-DEs:

$$\begin{cases} D^\alpha \gamma_1(\tau) - \int_0^\tau \gamma_1^2(x) dx = g_1(\tau), \\ D^\beta \gamma_2(\tau) + \int_0^\tau (\gamma_1^2(x) + \gamma_2^2(x)) dx = g_2(\tau), \\ \gamma_1(0) = 0, \gamma_2(0) = 0, \end{cases}$$

where

Table 6: Error comparison in Example 4.2 for  $\alpha = \beta = 1$ .

$\tau$	$\frac{M \text{ in [23]}}{ \gamma_1(\tau) - \gamma_{1,32}(\tau) }$	$\frac{M \text{ in [23]}}{ \gamma_2(\tau) - \gamma_{1,32}(\tau) }$	$\frac{W_2^0[0, 1]}{ \gamma_1(\tau) - \gamma_{1,10,10}(\tau) }$	$\frac{W_2^0[0, 1]}{ \gamma_2(\tau) - \gamma_{2,10,10}(\tau) }$	$\tau$	$\frac{M \text{ in [23]}}{ \gamma_1(\tau) - \gamma_{1,32}(\tau) }$	$\frac{M \text{ in [23]}}{ \gamma_2(\tau) - \gamma_{1,32}(\tau) }$	$\frac{W_2^0[0, 1]}{ \gamma_1(\tau) - \gamma_{1,10,10}(\tau) }$	$\frac{W_2^0[0, 1]}{ \gamma_2(\tau) - \gamma_{2,10,10}(\tau) }$
0.015625	$5.68 \times 10^{-6}$	$1.21 \times 10^{-4}$	$3.42 \times 10^{-9}$	$9.26 \times 10^{-9}$	0.515625	$1.67 \times 10^{-4}$	$3.05 \times 10^{-5}$	$1.19 \times 10^{-8}$	$6.54 \times 10^{-9}$
0.046875	$1.70 \times 10^{-5}$	$1.18 \times 10^{-4}$	$5.51 \times 10^{-9}$	$1.70 \times 10^{-8}$	0.546875	$1.76 \times 10^{-4}$	$2.18 \times 10^{-5}$	$1.13 \times 10^{-8}$	$5.72 \times 10^{-9}$
0.078125	$2.81 \times 10^{-5}$	$1.15 \times 10^{-4}$	$6.07 \times 10^{-9}$	$1.81 \times 10^{-8}$	0.578125	$1.85 \times 10^{-4}$	$1.27 \times 10^{-5}$	$1.02 \times 10^{-8}$	$4.58 \times 10^{-9}$
0.109375	$3.91 \times 10^{-5}$	$1.11 \times 10^{-4}$	$7.14 \times 10^{-9}$	$1.73 \times 10^{-8}$	0.609375	$1.94 \times 10^{-4}$	$3.26 \times 10^{-6}$	$8.75 \times 10^{-9}$	$2.86 \times 10^{-9}$
0.140625	$4.99 \times 10^{-5}$	$1.07 \times 10^{-4}$	$8.60 \times 10^{-9}$	$1.64 \times 10^{-8}$	0.640625	$2.02 \times 10^{-4}$	$6.51 \times 10^{-6}$	$7.03 \times 10^{-9}$	$1.19 \times 10^{-11}$
0.171875	$6.05 \times 10^{-5}$	$1.03 \times 10^{-4}$	$9.83 \times 10^{-9}$	$1.55 \times 10^{-8}$	0.671875	$2.10 \times 10^{-4}$	$1.66 \times 10^{-5}$	$5.21 \times 10^{-9}$	$5.10 \times 10^{-9}$
0.203125	$7.10 \times 10^{-5}$	$9.79 \times 10^{-5}$	$1.05 \times 10^{-8}$	$1.47 \times 10^{-8}$	0.703125	$2.19 \times 10^{-4}$	$2.71 \times 10^{-5}$	$3.05 \times 10^{-9}$	$1.45 \times 10^{-8}$
0.234375	$8.14 \times 10^{-5}$	$8.14 \times 10^{-5}$	$1.08 \times 10^{-8}$	$1.38 \times 10^{-8}$	0.734375	$2.27 \times 10^{-4}$	$3.78 \times 10^{-5}$	$1.20 \times 10^{-10}$	$3.18 \times 10^{-8}$
0.265625	$9.15 \times 10^{-5}$	$8.73 \times 10^{-5}$	$1.12 \times 10^{-8}$	$1.29 \times 10^{-8}$	0.765625	$2.35 \times 10^{-4}$	$4.89 \times 10^{-5}$	$3.91 \times 10^{-9}$	$6.21 \times 10^{-8}$
0.296875	$1.02 \times 10^{-4}$	$8.15 \times 10^{-5}$	$1.17 \times 10^{-8}$	$1.21 \times 10^{-8}$	0.796875	$2.43 \times 10^{-4}$	$6.04 \times 10^{-5}$	$8.78 \times 10^{-9}$	$1.13 \times 10^{-7}$
0.328125	$1.11 \times 10^{-4}$	$7.53 \times 10^{-5}$	$1.24 \times 10^{-8}$	$1.14 \times 10^{-8}$	0.828125	$2.51 \times 10^{-4}$	$7.21 \times 10^{-5}$	$1.34 \times 10^{-8}$	$1.96 \times 10^{-7}$
0.359375	$1.21 \times 10^{-4}$	$6.87 \times 10^{-5}$	$1.28 \times 10^{-8}$	$1.06 \times 10^{-8}$	0.859375	$2.59 \times 10^{-4}$	$8.42 \times 10^{-5}$	$1.62 \times 10^{-8}$	$3.28 \times 10^{-7}$
0.390625	$1.31 \times 10^{-4}$	$6.18 \times 10^{-5}$	$1.29 \times 10^{-8}$	$9.75 \times 10^{-9}$	0.890625	$2.67 \times 10^{-4}$	$9.66 \times 10^{-5}$	$1.58 \times 10^{-8}$	$5.32 \times 10^{-7}$
0.421875	$1.40 \times 10^{-4}$	$5.45 \times 10^{-5}$	$1.26 \times 10^{-8}$	$8.84 \times 10^{-9}$	0.921875	$2.75 \times 10^{-4}$	$1.09 \times 10^{-4}$	$1.26 \times 10^{-8}$	$8.44 \times 10^{-7}$
0.453125	$1.49 \times 10^{-4}$	$4.69 \times 10^{-5}$	$1.24 \times 10^{-8}$	$7.99 \times 10^{-9}$	0.953125	$2.83 \times 10^{-4}$	$1.22 \times 10^{-4}$	$1.08 \times 10^{-8}$	$1.31 \times 10^{-6}$
0.484375	$1.58 \times 10^{-4}$	$3.88 \times 10^{-5}$	$1.21 \times 10^{-8}$	$7.25 \times 10^{-9}$	0.984375	$2.92 \times 10^{-4}$	$1.36 \times 10^{-4}$	$2.16 \times 10^{-8}$	$2.00 \times 10^{-6}$

Table 7: Convergence order in Example 4.2.

$\alpha = \beta = 1, n = 10, \text{ in } \mathbf{W}_2^6[0, 1]$					
$N$	$\ \gamma_1 - \gamma_{1,n,N}\ _\infty$	$C.F_1$	$\ \gamma_2 - \gamma_{2,n,N}\ _\infty$	$C.F_2$	$Cpu\ time(sec)$
2	$6.79 \times 10^{-3}$	—	$1.71 \times 10^{-2}$	—	3
4	$4.64 \times 10^{-5}$	7.19	$1.52 \times 10^{-4}$	6.81	5
8	$2.75 \times 10^{-7}$	7.39	$2.05 \times 10^{-6}$	6.21	9

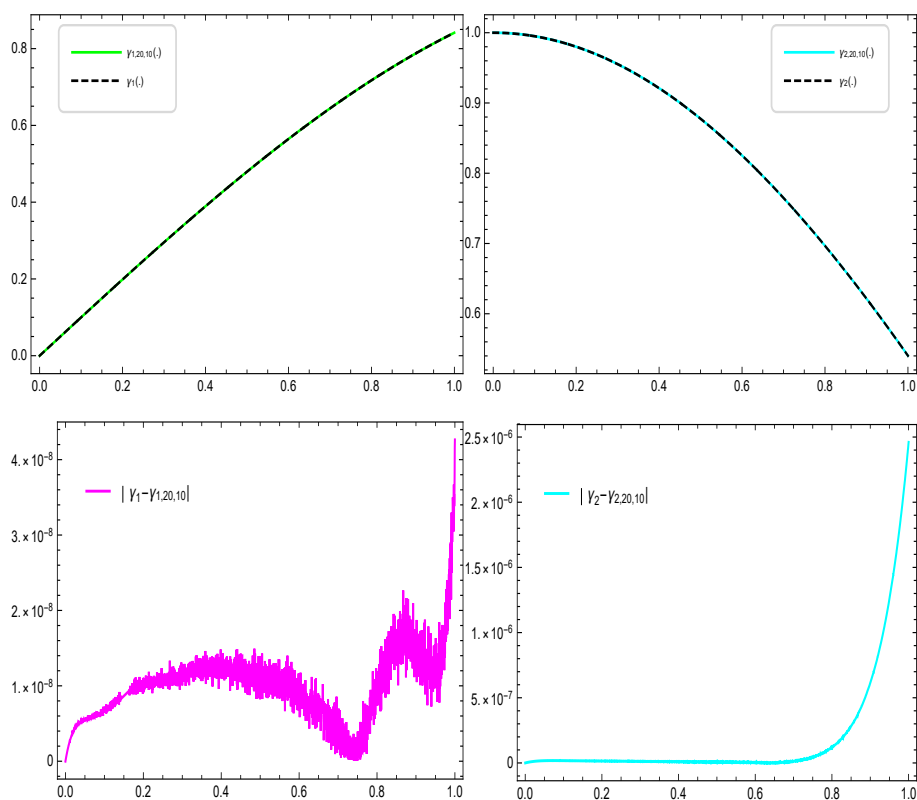


Figure 5: Numerical solution and absolute error using the Present method with the Taylor series expansion in Example 4.2.

$$0 < \alpha, \beta \leq 1,$$

and the analytical solution for  $\alpha = \beta = 1$  is

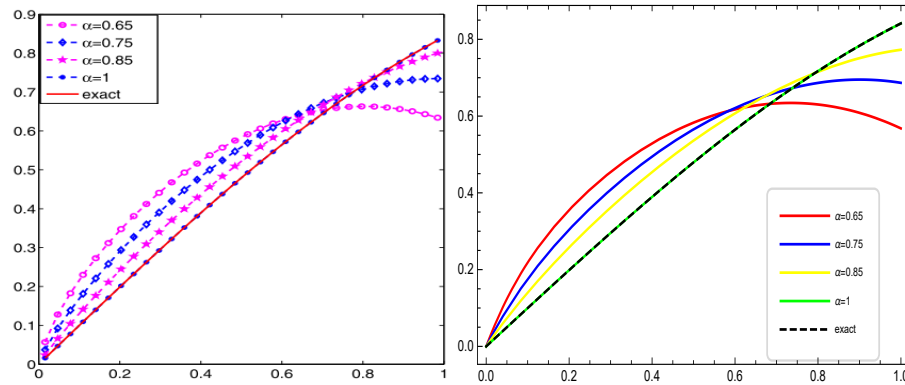


Figure 6: Comparison of the numerical solutions using the Present method with the Taylor series expansion in Example 4.2 for  $\alpha = 0.65, 0.75, 0.85, 1$ . (Left:  $\gamma_{1,32}(\cdot)$ , M in [23]; Right:  $\gamma_{1,20,10}(\cdot)$ , Present method).

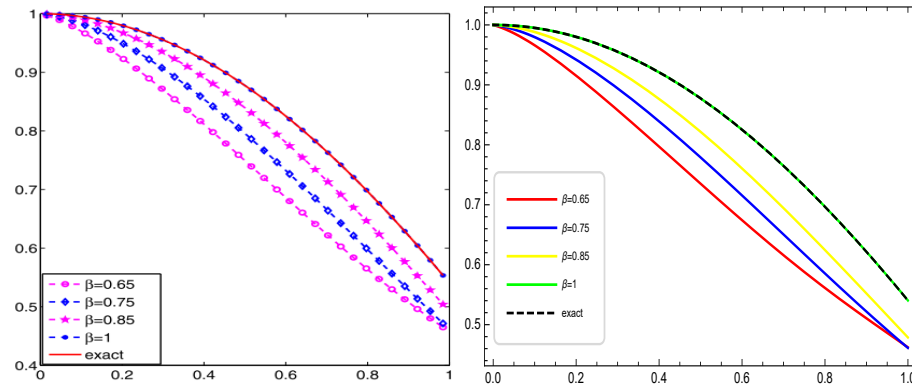


Figure 7: Comparison of the numerical solutions using the Present method with the Taylor series expansion in Example 4.2 for  $\beta = 0.65, 0.75, 0.85, 1$ . (Left:  $\gamma_{2,32}(\cdot)$ , M in [23]; Right:  $\gamma_{2,20,10}(\cdot)$ , Present method).

$$\gamma(\tau) = (\tau \sin(\tau), 1 - \cos(\tau)).$$

To solve this example, we first need to move the nonlinear part to the right side of the equation. Then, we can create a coefficient matrix using the linear part. This will give us the following equation:

$$\begin{cases} D^\alpha \gamma_1(\tau) = g_1(\tau) + \int_0^\tau \gamma_1^2(x) dx, \\ D^\beta \gamma_2(\tau) = g_2(\tau) - \int_0^\tau (\gamma_1^2(x) + \gamma_2^2(x)) dx, \\ \gamma_1(0) = 0, \gamma_2(0) = 0. \end{cases} \quad (15)$$

In the first equation of (15), the linear operator is not applied to  $\gamma_2(\cdot)$ . As a result, the coefficient matrix does not depend on  $\gamma_2(\cdot)$ , and we have substituted it with an  $N \times N$  zero matrix. Similarly, in the second equation of (15), the coefficient matrix related to  $\gamma_1(\cdot)$  is not used, and we have replaced it with an  $N \times N$  zero matrix.

First, we solved this example in the  $\mathbf{W}_2^6[0, 1]$  space without using the Taylor expansion, which utilizes  $\tau_l = \frac{l}{N+1}$  points. The numerical solution and absolute error are shown in Figure 8. However, this method is not effective. To improve the results, we compared the Present method, which uses the Taylor expansion, with the method proposed in [23]. This comparison was based on the absolute error and numerical solution, as shown in Table 8 and Figures 9 and 10. The convergence order is also shown in Table 9. Finally, to solve this example according to (2) and (3) instead of  $\gamma_1(\cdot)$  and  $\gamma_2(\cdot)$ , we use the truncated Taylor series expansion around the point  $x$  in the interval  $[0, 1]$  for  $m = 3$ . As a result, the nonlinear parts of  $H_1$  as,

$$\begin{aligned} & \tau \gamma_1(\tau)^2 - \tau^2 \gamma_1(\tau) \gamma_1'(\tau) + \frac{1}{3} \tau^3 \gamma_1'(\tau)^2 + \frac{1}{3} \tau^3 \gamma_1(\tau) \gamma_1''(\tau)^2 - \frac{1}{4} \tau^4 \gamma_1'(\tau) \gamma_1''(\tau) \\ & + \frac{1}{20} \tau^5 \gamma_1''(\tau)^2 - \frac{1}{12} \tau^4 \gamma_1(\tau) \gamma_1^{(3)}(\tau) + \frac{1}{15} \tau^5 \gamma_1'(\tau) \gamma_1^{(3)}(\tau) \\ & - \frac{1}{36} \tau^6 \gamma_1''(\tau) \gamma_1^{(3)}(\tau) + \frac{1}{252} \tau^7 \gamma_1^{(3)}(\tau)^2, \end{aligned}$$

and  $H_2$  as,

$$\begin{aligned} & \tau \gamma_1(\tau)^2 - \tau^2 \gamma_1(\tau) \gamma_1'(\tau) + \frac{1}{3} \tau^3 \gamma_1'(\tau)^2 + \frac{1}{3} \tau^3 \gamma_1(\tau) \gamma_1''(\tau)^2 - \frac{1}{4} \tau^4 \gamma_1'(\tau) \gamma_1''(\tau) \\ & + \frac{1}{20} \tau^5 \gamma_1''(\tau)^2 - \frac{1}{12} \tau^4 \gamma_1(\tau) \gamma_1^{(3)}(\tau) + \frac{1}{15} \tau^5 \gamma_1'(\tau) \gamma_1^{(3)}(\tau) \\ & - \frac{1}{36} \tau^6 \gamma_1''(\tau) \gamma_1^{(3)}(\tau) + \frac{1}{252} \tau^7 \gamma_1^{(3)}(\tau)^2 \\ & + \tau \gamma_2(\tau)^2 - \tau^2 \gamma_2(\tau) \gamma_2'(\tau) + \frac{1}{3} \tau^3 \gamma_2'(\tau)^2 + \frac{1}{3} \tau^3 \gamma_2(\tau) \gamma_2''(\tau)^2 \\ & - \frac{1}{4} \tau^4 \gamma_2'(\tau) \gamma_2''(\tau) + \frac{1}{20} \tau^5 \gamma_2''(\tau)^2 - \frac{1}{12} \tau^4 \gamma_2(\tau) \gamma_2^{(3)}(\tau) \end{aligned}$$

$$+ \frac{1}{15} \tau^5 \gamma_2'(\tau) \gamma_2^{(3)}(\tau) - \frac{1}{36} \tau^6 \gamma_2''(\tau) \gamma_2^{(3)}(\tau) + \frac{1}{252} \tau^7 \gamma_2^{(3)}(\tau)^2.$$

Table 8: Error in Example 4.3 for  $\alpha = \beta = 1$ .

$\tau$	$\mathbf{W}_2^6[0, 1]$ $ \gamma_1(\tau) - \gamma_{1,20,10}(\tau) $	$\mathbf{W}_2^6[0, 1]$ $ \gamma_2(\tau) - \gamma_{2,20,10}(\tau) $	$\tau$	$\mathbf{W}_2^6[0, 1]$ $ \gamma_1(\tau) - \gamma_{1,20,10}(\tau) $	$\mathbf{W}_2^6[0, 1]$ $ \gamma_2(\tau) - \gamma_{2,20,10}(\tau) $
0.015625	$9.80 \times 10^{-8}$	$1.13 \times 10^{-7}$	0.515625	$2.35 \times 10^{-7}$	$2.54 \times 10^{-7}$
0.046875	$2.10 \times 10^{-7}$	$2.29 \times 10^{-7}$	0.546875	$2.34 \times 10^{-7}$	$2.53 \times 10^{-7}$
0.078125	$2.46 \times 10^{-7}$	$2.63 \times 10^{-7}$	0.578125	$2.37 \times 10^{-7}$	$2.55 \times 10^{-7}$
0.109375	$2.46 \times 10^{-7}$	$2.62 \times 10^{-7}$	0.609375	$2.42 \times 10^{-7}$	$2.57 \times 10^{-7}$
0.140625	$2.35 \times 10^{-7}$	$2.54 \times 10^{-7}$	0.640625	$2.45 \times 10^{-7}$	$2.59 \times 10^{-7}$
0.171875	$2.29 \times 10^{-7}$	$2.50 \times 10^{-7}$	0.671875	$2.44 \times 10^{-7}$	$2.58 \times 10^{-7}$
0.203125	$2.30 \times 10^{-7}$	$2.51 \times 10^{-7}$	0.703125	$2.38 \times 10^{-7}$	$2.54 \times 10^{-7}$
0.234375	$2.34 \times 10^{-7}$	$2.53 \times 10^{-7}$	0.734375	$2.32 \times 10^{-7}$	$2.50 \times 10^{-7}$
0.265625	$2.37 \times 10^{-7}$	$2.55 \times 10^{-7}$	0.765625	$2.33 \times 10^{-7}$	$2.49 \times 10^{-7}$
0.296875	$2.36 \times 10^{-7}$	$2.54 \times 10^{-7}$	0.796875	$2.45 \times 10^{-7}$	$2.53 \times 10^{-7}$
0.328125	$2.33 \times 10^{-7}$	$2.53 \times 10^{-7}$	0.828125	$2.67 \times 10^{-7}$	$1.61 \times 10^{-7}$
0.359375	$2.32 \times 10^{-7}$	$2.52 \times 10^{-7}$	0.859375	$2.95 \times 10^{-7}$	$2.60 \times 10^{-7}$
0.390625	$2.33 \times 10^{-7}$	$2.53 \times 10^{-7}$	0.890625	$3.16 \times 10^{-7}$	$2.21 \times 10^{-7}$
0.421875	$2.36 \times 10^{-7}$	$2.54 \times 10^{-7}$	0.921875	$3.21 \times 10^{-7}$	$9.44 \times 10^{-8}$
0.453125	$2.38 \times 10^{-7}$	$2.55 \times 10^{-7}$	0.953125	$3.05 \times 10^{-7}$	$2.00 \times 10^{-7}$
0.484375	$2.37 \times 10^{-7}$	$2.55 \times 10^{-7}$	0.984375	$2.85 \times 10^{-7}$	$7.78 \times 10^{-7}$

Table 9: Convergence order in Example 4.3.

$\alpha = \beta = 1, n = 20, \text{ in } \mathbf{W}_2^6[0, 1]$					
$N$	$\ \gamma_1 - \gamma_{1,n,N}\ _\infty$	$C.F_1$	$\ \gamma_2 - \gamma_{2,n,N}\ _\infty$	$C.F_2$	$Cpu\ time(sec)$
2	$6.64 \times 10^{-2}$	—	$1.72 \times 10^{-2}$	—	1
4	$7.99 \times 10^{-4}$	6.37	$2.29 \times 10^{-4}$	10.62	3
8	$5.06 \times 10^{-7}$	6.23	$2.78 \times 10^{-6}$	6.36	6

**Example 4.4.** Consider the NFVI-DEs:

$$\begin{cases} D^\alpha \gamma_1(\tau) - \int_0^\tau \gamma_1^2(x) dx = g_1(\tau), \\ D^\beta \gamma_2(\tau) - \int_0^\tau \gamma_2^2 dx = g_2(\tau), \\ \gamma_1(0) = 0, \gamma_2(0) = 0, \end{cases}$$

where

$$0 < \alpha, \beta \leq 1,$$

and the analytical solution for  $\alpha = \beta = 1$  is

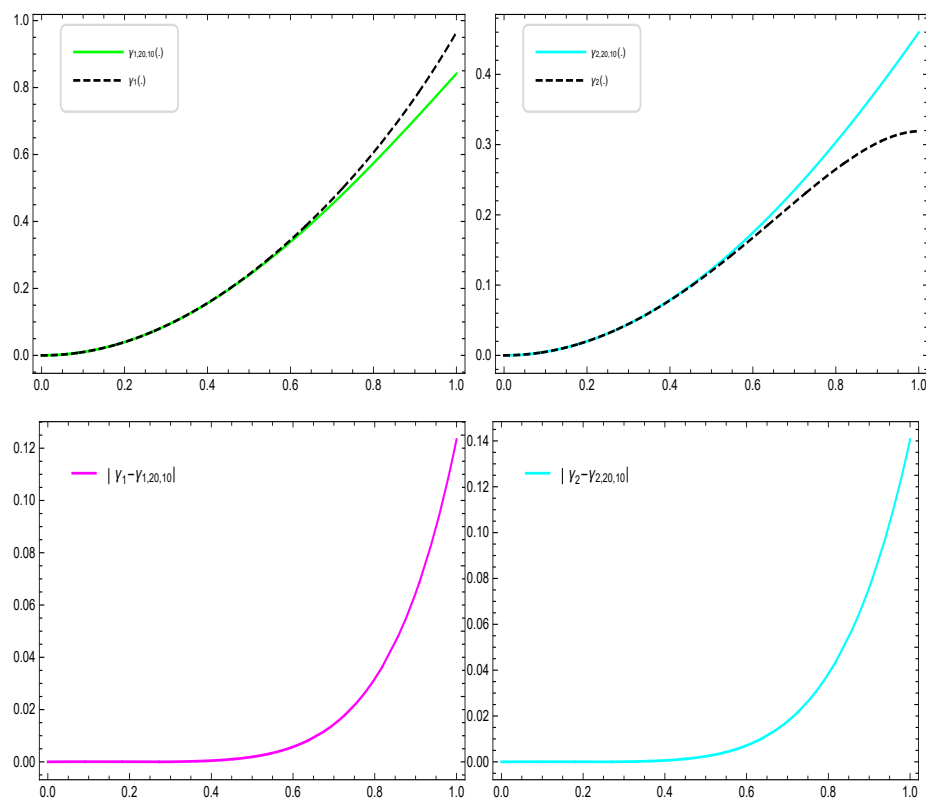


Figure 8: Numerical solution and absolute error without using the Taylor series expansion in Example 4.3.

$$\gamma(\tau) = \left(\tau + \frac{\tau^3}{2}, \tau - \frac{\tau^3}{2}\right).$$

First, we solved this example in the  $\mathbf{W}_2^4[0, 1]$  space without using the Taylor expansion, which utilizes  $\tau_l = \frac{l}{2N+1}$  points. The numerical solution and absolute error are shown in Figure 11. However, this method is not effective. To improve the results, we compared the Present method, which uses the Taylor expansion, with the method proposed in [23]. This comparison was based on the absolute error and numerical solution, as shown in Table 10 and Figures 12 and 13. The convergence order is also shown in Table 11. Finally, to solve this example according to (2) and (3) instead of  $\gamma_1(\cdot)$  and  $\gamma_2(\cdot)$ , we use the truncated Taylor series expansion around the point  $x$  in the interval

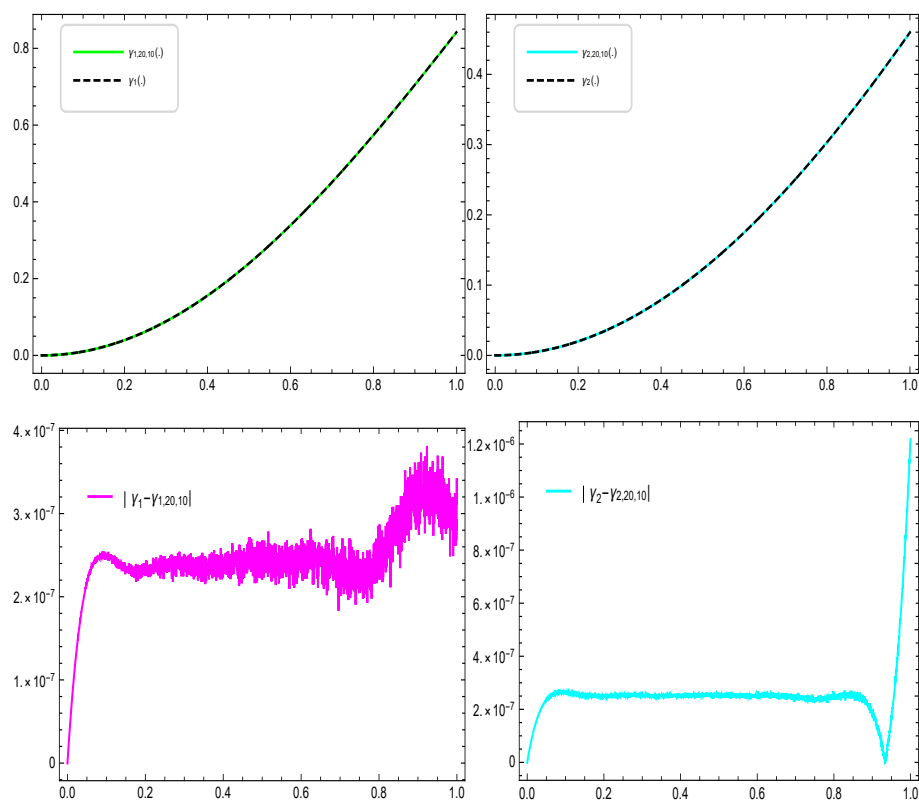


Figure 9: Numerical solution and absolute error with using the Taylor series expansion in Example 4.3.

$[0, 1]$  for  $m = 3$ . As a result, the nonlinear parts of  $H_1$  as

$$\tau\gamma_1(\tau)^2 - \tau^2\gamma_1(\tau)\gamma_1'(\tau) + \frac{1}{3}\tau^3\gamma_1'(\tau)^2 + \frac{1}{3}\tau^3\gamma_1(\tau)\gamma_1''(\tau)^2 - \frac{1}{4}\tau^4\gamma_1'(\tau)\gamma_1''(\tau) + \frac{1}{20}\tau^5\gamma_1''(\tau)^2,$$

and  $H_2$  as

$$\tau\gamma_2(\tau)^2 - \tau^2\gamma_2(\tau)\gamma_2'(\tau) + \frac{1}{3}\tau^3\gamma_2'(\tau)^2 + \frac{1}{3}\tau^3\gamma_2(\tau)\gamma_2''(\tau)^2 - \frac{1}{4}\tau^4\gamma_2'(\tau)\gamma_2''(\tau) + \frac{1}{20}\tau^5\gamma_2''(\tau)^2.$$

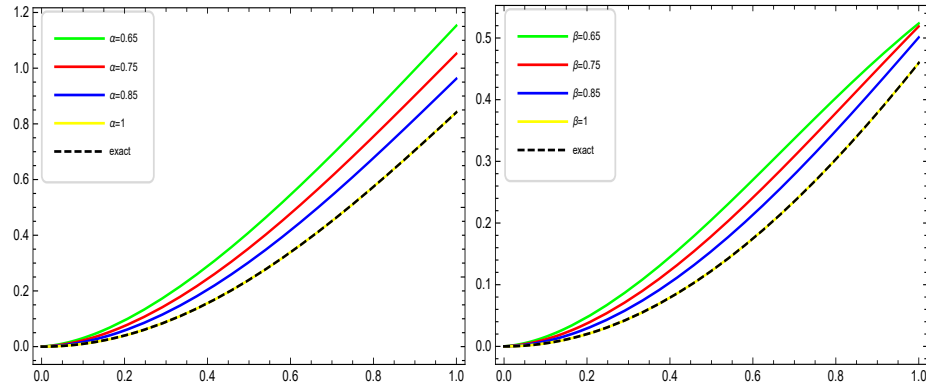


Figure 10: Comparison of the numerical solutions using the Present method with the Taylor series expansion in Example 4.3 for different values  $\alpha$  and  $\beta$ . (Left:  $\gamma_{1,20,10}(\cdot)$ ; Right:  $\gamma_{2,20,10}(\cdot)$ ).

Table 10: Error in Example 4.4 for  $\alpha = \beta = 1$ .

$\tau$	$\mathbf{W}_2^4[0, 1]$ $ \gamma_1(\tau) - \gamma_{1,20,16}(\tau) $	$\mathbf{W}_2^4[0, 1]$ $ \gamma_2(\tau) - \gamma_{2,20,16}(\tau) $	$\tau$	$\mathbf{W}_2^4[0, 1]$ $ \gamma_1(\tau) - \gamma_{1,20,16}(\tau) $	$\mathbf{W}_2^4[0, 1]$ $ \gamma_2(\tau) - \gamma_{2,20,16}(\tau) $
0.015625	$2.82 \times 10^{-7}$	$2.82 \times 10^{-7}$	0.515625	$3.28 \times 10^{-7}$	$2.99 \times 10^{-7}$
0.046875	$3.14 \times 10^{-7}$	$3.14 \times 10^{-7}$	0.546875	$3.31 \times 10^{-7}$	$2.97 \times 10^{-7}$
0.078125	$3.12 \times 10^{-7}$	$3.12 \times 10^{-7}$	0.578125	$3.35 \times 10^{-7}$	$2.94 \times 10^{-7}$
0.109375	$3.12 \times 10^{-7}$	$3.12 \times 10^{-7}$	0.609375	$3.39 \times 10^{-7}$	$2.91 \times 10^{-7}$
0.140625	$3.12 \times 10^{-7}$	$3.12 \times 10^{-7}$	0.640625	$3.44 \times 10^{-7}$	$2.88 \times 10^{-7}$
0.171875	$3.13 \times 10^{-7}$	$3.12 \times 10^{-7}$	0.671875	$3.49 \times 10^{-7}$	$2.85 \times 10^{-7}$
0.203125	$3.13 \times 10^{-7}$	$3.11 \times 10^{-7}$	0.703125	$3.55 \times 10^{-7}$	$2.82 \times 10^{-7}$
0.234375	$3.13 \times 10^{-7}$	$3.11 \times 10^{-7}$	0.734375	$3.61 \times 10^{-7}$	$2.78 \times 10^{-7}$
0.265625	$3.14 \times 10^{-7}$	$3.10 \times 10^{-7}$	0.765625	$3.69 \times 10^{-7}$	$2.74 \times 10^{-7}$
0.296875	$3.15 \times 10^{-7}$	$3.09 \times 10^{-7}$	0.796875	$3.76 \times 10^{-7}$	$2.70 \times 10^{-7}$
0.328125	$3.16 \times 10^{-7}$	$3.09 \times 10^{-7}$	0.828125	$3.85 \times 10^{-7}$	$2.65 \times 10^{-7}$
0.359375	$3.17 \times 10^{-7}$	$3.07 \times 10^{-7}$	0.859375	$3.95 \times 10^{-7}$	$2.60 \times 10^{-7}$
0.390625	$3.19 \times 10^{-7}$	$3.06 \times 10^{-7}$	0.890625	$4.05 \times 10^{-7}$	$2.55 \times 10^{-7}$
0.421875	$3.20 \times 10^{-7}$	$3.05 \times 10^{-7}$	0.921875	$4.16 \times 10^{-7}$	$2.50 \times 10^{-8}$
0.453125	$3.22 \times 10^{-7}$	$3.03 \times 10^{-7}$	0.953125	$4.28 \times 10^{-7}$	$2.44 \times 10^{-7}$
0.484375	$3.25 \times 10^{-7}$	$3.01 \times 10^{-7}$	0.984375	$4.41 \times 10^{-7}$	$2.38 \times 10^{-7}$

## 5 Conclusions

In this article, we proposed a novel approach for solving systems of NFVI-DEs by combining the RKM without the G-SOP with the Taylor series expansion. In Example 4.1, we presented an alternative method that does not employ Taylor series expansion and demonstrated its effectiveness for certain sys-

Table 11: Convergence order in Example 4.4.

$\alpha = \beta = 1, n = 20, \text{ in } \mathbf{W}_2^4[0, 1]$					
$N$	$\ \gamma_1 - \gamma_{1,n,N}\ _\infty$	$C.F_1$	$\ \gamma_2 - \gamma_{2,n,N}\ _\infty$	$C.F_2$	$Cpu\ time(sec)$
4	$2.32 \times 10^{-4}$	—	$1.85 \times 10^{-4}$	—	1
8	$6.28 \times 10^{-6}$	5.20	$4.45 \times 10^{-6}$	3.80	4
16	$4.48 \times 10^{-7}$	5.37	$3.25 \times 10^{-7}$	3.77	8

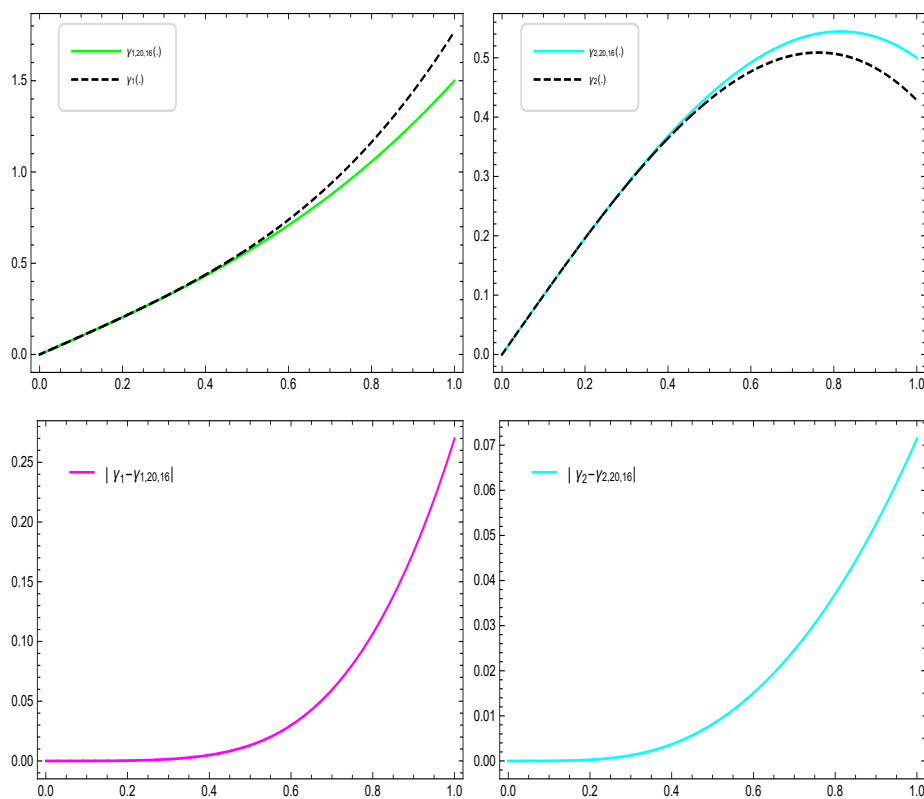


Figure 11: Numerical solution and absolute error without using the Taylor series expansion in Example 4.4.

tems of integro-differential equations. However, we found that this method proves less effective for problems involving Volterra's integral applied to non-linear components. Even modifications to the space and points within this

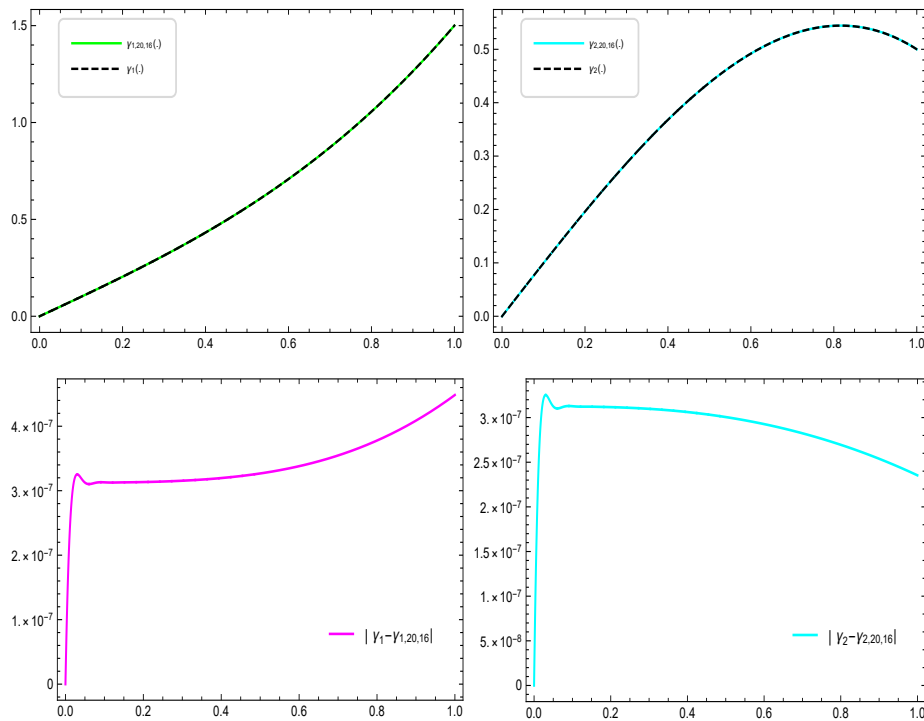


Figure 12: Numerical solution and absolute error with using the Taylor series expansion in Example 4.4.

method fail to yield improved results. Consequently, to solve such problems effectively, we must integrate the RKM with Taylor series expansion. Our numerical results validate the efficacy of this combined approach.

## Declarations

**Author Contributions:** I confirm that all authors listed on the title page have contributed significantly to the work, have read the manuscript, attest to the validity and legitimacy of the data and its interpretation, and agree to its submission.

**Funding:** The authors declare that this research received no grant from any funding agency in the public, commercial, or not-for-profit sectors.

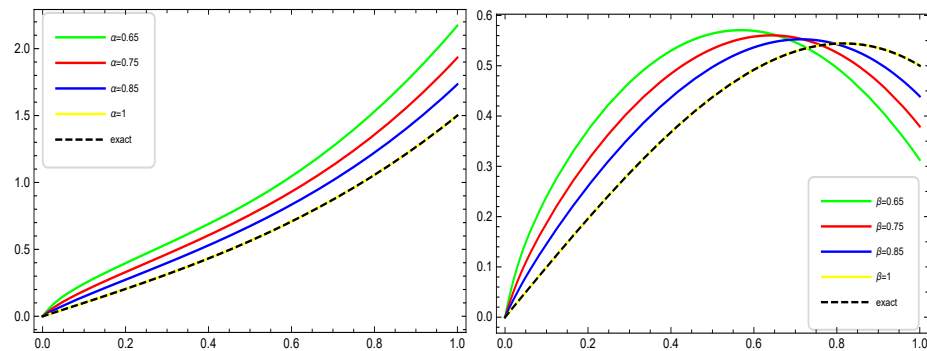


Figure 13: Comparison of the numerical solutions using the Present method with the Taylor series expansion in Example 4.4 for different values  $\alpha$  and  $\beta$ . (Left:  $\gamma_{1,20,16}(\cdot)$ ; Right:  $\gamma_{2,20,16}(\cdot)$ ).

**Competing interests:** The authors declare that they have no conflict of interest.

**Data availability:** Our manuscript has no associated data.

## Acknowledgements

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

## References

- [1] Akbar M., Nawaz R., Ahsan S., Nisar K.S., Abdel-Aty A.H. and Eleuch H. *New approach to approximate the solution for the system of fractional order Volterra integro-differential equations*, Results Phys. 19 (2020) 103453.
- [2] Alvandi A. and Paripour M. *Reproducing kernel method with Taylor expansion for linear Volterra integro-differential equations*, Commun. Numer. Anal. 1 (2017) 1–10.
- [3] Alvandi A. and Paripour M. *The combined reproducing kernel method and Taylor series for handling nonlinear Volterra integro-differential equations with derivative type kernel*, Appl. Math. Comput. 355 (2019) 151–160.
- [4] Amoozad T., Abbasbandy S., Sahihi H. and Allahviranloo T. *A new application of the reproducing kernel method for solving linear systems of fractional order Volterra integro-differential equations*, Phys. Scripta. 99 (2024) 075209.
- [5] Amoozad T., Allahviranloo T., Abbasbandy S. and Rostamy Malkhalifeh M. *Using a new implementation of reproducing kernel Hilbert space method to solve a system of second-order BVPs*, Int. J. Dynam. Control. 12(6) (2024) 1694–1706.
- [6] Amoozad T., Allahviranloo T., Abbasbandy S. and Rostamy Malkhalifeh M. *Application of the reproducing kernel method for solving linear Volterra integral equations with variable coefficients*, Phys. Scripta 99 (2024) 025246.
- [7] Babolian E., Javadi S. and Moradi E. *Error analysis of reproducing kernel Hilbert space method for solving functional integral equations*, Comput. Appl. Math. 300 (2016) 300–311.
- [8] Bakodah H.O., Al-Mazmumy M. and Almuhalbedi S.O. *Solving system of integro differential equations using discrete Adomian decomposition method*, Taibah Univ. Sci. 13 (2019) 805–812.

- [9] Cui M.G. and Lin Y. *Nonlinear numerical analysis in the reproducing kernel space*, Nova Science, Hauppauge, Inc., Hauppauge, 2009.
- [10] Das P., Rana S. and Ramos H. *A perturbation-based approach for solving fractional-order Volterra–Fredholm integro differential equations and its convergence analysis*, Int. J. Comput. Math. 97 (2020) 1994–2014.
- [11] Das P., Rana S. and Ramos H. *On the approximate solutions of a class of fractional order nonlinear Volterra integro-differential initial value problems and boundary value problems of first kind and their convergence analysis*, Comp. Appl. Math. 404 (2022) 113116.
- [12] Geng F.Z. and Cui M.G. *Solving a nonlinear system of second order boundary value problems*, Math. Anal. Appl. 327 (2007) 1167–1181.
- [13] Ghanbari F., Mokhtary P. and Ghanbari K. *Numerical solution of a class of fractional order integro-differential algebraic equations using Müntz–Jacobi Tau method*, Comp. Appl. Math. 362 (2019) 172–184.
- [14] Hansen V.L. *Functional analysis, entering Hilbert space*, World Scientific Publishing Co. Pte. Ltd., 2006.
- [15] Jiang W. and Chen Z. *Solving a system of linear Volterra integral equations using the new reproducing kernel method*, Appl. Math. Comput. 219 (2013) 10225–10230.
- [16] Kumar S., Nieto J.J. and Ahmad B. *Chebyshev spectral method for solving fuzzy fractional Fredholm–Volterra integro-differential equation*, Math. Comput. Simul. 192 (2022) 501–513.
- [17] Li H. and Ma J. *On generalized multistep collocation methods for Volterra integro-differential equations*, Partial Differ. Equ. Appl. Math. 226 (2024) 399–412.
- [18] Mandal M. *Convergence analysis and numerical implementation of projection methods for solving classical and fractional Volterra integro-differential equations*, Math. Comput. Simul. 225 (2024) 889–913.

- [19] Mei L. and Lin Y. *Simplified reproducing kernel method and convergence order for linear Volterra integral equations with variable coefficients*, Comput. Appl. Math. 346 (2019) 390–398.
- [20] Podlubny I. *Fractional differential equations*, Academic Press, Academic Press, 1999.
- [21] Sahihi H., Allahviranloo T. and Abbasbandy S. *Solving system of second-order BVPs using a new algorithm based on reproducing kernel Hilbert space*, Appl. Num. Math. 151 (2020) 27–39.
- [22] Shen L., Zhu S., Liu B., Zhang Z. and Cui Y. *Numerical implementation of nonlinear system of fractional Volterra integral-differential equations by Legendre wavelet method and error estimation*, Numer. Methods Partial Differ. Equ. 37 (2021) 1344–1360.
- [23] Wang J., Xu T.Z., Wei Y.Q. and Xie J.Q. *Numerical simulation for coupled systems of nonlinear fractional order integro-differential equations via wavelets method*, Appl. Math. Comput. 324 (2018) 36–50.
- [24] Wang Y., Chaolu T. and Chen Z. *Using reproducing kernel for solving a class of singular weakly nonlinear boundary value problems*, Comput. Appl. Math. 87 (2010) 367–380.
- [25] Xie J. and Yi M. *Numerical research of nonlinear system of fractional Volterra-Fredholm integral-differential equations via Block-Pulse functions and error analysis*, Comput. Appl. Math. 345 (2019) 159–167.
- [26] Yang, L.H. Shen, J.H. and Wang, Y. *The reproducing kernel method for solving the system of the linear Volterra integral equations with variable coefficients*, Comput. Appl. Math. 236 (2012) 2398–2405.
- [27] Zhou, F. and Xu, X. *Numerical solution of fractional Volterra-Fredholm integro-differential equations with mixed boundary conditions via Chebyshev wavelet method*, Int. J. Comput. Math. 96 (2019) 436–456.



## Convex-hull based two-phase algorithm to solve capacitated vehicle routing problem

M. Afsharirad\*, and A. Hashemi Borzabadi

### Abstract

The goal of this paper is to present a two-phase convex hull-based algorithm for the capacitated vehicle routing problem (CVRP), consisting of clustering and routing phases. First, a K-means-based algorithm is proposed for the clustering phase, where the centroids are updated according to the convex hull of the assigned points. Furthermore, a convex-hull-based algorithm is suggested for the routing phase, which iteratively inserts unrouted points into the convex hull. To improve the routes, an ant colony optimization algorithm is applied. It is shown that the proposed method has a time complexity of order  $o(n^2 \log n)$ , where  $n$  is the number of customers. For performance evaluation, we utilize CVRP benchmark samples and compare

---

\*Corresponding author

Received 7 May 2025; revised 11 June 2025; accepted 30 June 2025

Maria Afsharirad

Department of Applied Mathematics, University of Science and Technology of Mazandaran, Behshahr, P.O. Box: 48518-78195, Mazandaran, Iran. e-mail: [maria.afsharirad@gmail.com](mailto:maria.afsharirad@gmail.com)

Akbar Hashemi Borzabadi

Department of Applied Mathematics, University of Science and Technology of Mazandaran, Behshahr, P.O. Box: 48518-78195, Mazandaran, Iran. e-mail: [akbar.h.borzabadi@gmail.com](mailto:akbar.h.borzabadi@gmail.com)

### How to cite this article

Afsharirad, M. and Hashemi Borzabadi, A. , Convex-hull based two-phase algorithm to solve capacitated vehicle routing problem. *Iran. J. Numer. Anal. Optim.*, 2025; 15(3): 1241-1274. <https://doi.org/10.22067/ijnao.2025.93411.1643>

the results to those of other two-phase CVRP algorithms. The proposed clustering method combined with common routing techniques, as well as the K-means clustering method paired with the proposed routing approach, yields highly favorable results in some instances. Moreover, the proposed two-phase method outperforms other approaches in certain instances.

**AMS subject classifications (2020):** Primary 90C27; Secondary , 90C59.

**Keywords:** Capacitated vehicle routing problem, K-means algorithm, Convex hull, Ant colony optimization

## 1 Introduction

The vehicle routing problem (VRP) is one of the most well-known problems in combinatorial optimization due to its wide applicability in fields such as public transportation, waste collection, and drone routing. In its typical form, the VRP involves finding routes for a fleet of vehicles with limited capacity to serve customers. These routes start from a central node called the “depot”, return to it after visiting customers (within vehicle capacity constraints), and aim to optimize an objective function—commonly minimizing total distance or total service time.

CVRP might be divided into two phases, the first phase is “clustering”, in which customers are assigned to vehicles. The second phase is “routing”, which determines the optimal route of all vehicles in their cluster. It is clear that the second phase is the well-known traveling salesman problem (TSP). Since CVRP contains TSP and Bin packing problem as its special case, therefore it is classified as an NP-hard problem.

A two-phase convex-hull based (CHB) heuristic method is provided in this paper. First, customers are clustered by a convex-hull based K-means (CH-means) algorithm. Since the length of the whole TSP tour should be minimized in CVRP, in CH-means algorithm, the centers are updated as the mean point of the convex hull of the points assigned to each cluster. Accordingly, any cluster contain all points on the line segment joining any two points in it. However adding points in any step is the same with the

K-means algorithm and is based on the minimum distance from the center, but since closest points to the center are located inside the respective convex hull, they will be assigned to that cluster.

Furthermore, the fitness of any step's solution is calculated by a proposed routing algorithm, and it is saved. At the end of the clustering phase, the best found solution is chosen for improvement, which is not necessarily the last solution found. This increases the computational complexity of the algorithm, but it allows the selection of the best clustering.

The routes are built by the proposed convex-hull based routing algorithm (CH-Insertion). Our idea for routing is to construct the convex hull of the points and to insert unassigned points by breaking an edge of the polygon into two edges. This is based on the well known property for Euclidean TSP: The order in which the points appear in an optimal TSP tour must be the same as the order in which these points appear on the convex hull, [30]. Finally, the routes are improved with a meta-heuristic algorithm. Ant colony optimization algorithm has been chosen, due to its satisfactory performance on TSP.

The paper is organized as follows: The literature is reviewed in Section 2 in three categories: Variations, applications and approaches. The problem is defined in Section 3. Section 4 is devoted to the two-phase CHB heuristic. In its first subsection, the CH-means algorithm is presented, and the second subsection explains the CH-insertion algorithm. The last subsection discusses ant colony optimization algorithm. Finally, Section 6 presents numerical results and concludes the paper.

## 2 Related works

The problem was first introduced by Dantzig and Ramser [15]. They applied the problem to petrol deliveries and proposed the first approximation algorithm based on matching. In the following, we review related literature in three categories. First, we state variations of the problem. Then we list the most major applications of VRP. Finally, we discuss different approaches of the problem in literature.

## 2.1 Variations of VRP

Numerous variations have been provided by researchers for VRP due to its variety of applications. Some famous variations are as follows:

Capacitated vehicle routing problem (CVRP) is the most closely related version to the original VRP in which a positive number is assigned to each customer as its capacity or quantity of its demand, see [57].

The VRP with time windows (VRPTW) is a variation of CVRP in which serving any customer must be done within a determined time interval, [52, 53]. There is also another closely related variation of VRP, dealing with online real-time demands. A hybrid meta-heuristic algorithm based on genetic algorithm and tabu search for on-line VRP, is provided in [31].

The VRP with profits (VRPP) is a maximization problem, where all customers do not have to be serviced. The problem is to visit all customers at most once in order to maximize the sum of collected profits according to a vehicle time limitation, [25].

Open VRP (OVRP) is another variation of the problem in which vehicles do not have to return to the depot at the end of their route, [48].

Multi-depot VRP (MDVRP) has multiple nodes as the depot, and the problem is also to assign each vehicle to each depot, [34]. The classic form of the MDVRP in which all vehicles start and end their route at the same depot, is considered in [46].

The VRP with Drones (VRPD) is an extension of CVRP, where not only trucks but also drones are used to service customers. One distinctive feature of the VRPD is that a drone may travel with a truck, take off from its stop to serve customers, and land at a service hub to travel with another truck as long as the flying range and loading capacity limitations are satisfied, [19, 59]

For comprehensive reviews of VRP refer to [5, 4, 35].

## 2.2 Applications

Recent applications of the VRP have expanded into specialized domains. Authors in [16] provided a comprehensive review of the police patrol rout-

ing problem, highlighting how VRP models ensure balanced patrol allocation while minimizing response times and workload disparities across patrol zones. Their work underscores VRP's flexibility in adapting to public-safety contexts, where route equity and real-time adjustments are vital in urban environments. A multi-objective VRP is applied in [36] to optimize real-world postal delivery at scale. They balance delivery efficiency and service fairness—incorporating objectives such as minimizing total distance and regulating driver workload. Through meta-heuristics, they demonstrate substantial cost savings while maintaining operational equity across a large delivery network.

Unmanned aerial vehicle (UAV) applications are reviewed in [56], in VRP contexts—emphasizing disaster relief, surveillance, and agricultural logistics. Their meta-analysis captures how drone-based VRPs address reach limitations of ground fleets, introducing constraints like battery life and communication reliability. This study solidified VRP's extension into UAV-coordinated systems.

Building on these foundations, recent work has advanced VRP in multi-modal and uncertain environments. VRPD-DT is introduced in [27], a vehicle-and-drone routing framework that integrates dynamic traffic prediction using machine learning. Their real-time VND heuristic outperformed static models, showing improved delivery times under fluctuating conditions. On the same front, authors in [14] survey truck-drone cooperative VRPs, categorizing operational modes—from synchronous to independent operations—and summarizing over 200 studies with implications for last-mile delivery, reconnaissance, and patrol. A PRISMA-based review of satellite depots in urban logistics is performed in [54], finding that roughly half of VRP designs incorporate cross-docking via intermediate warehouses. Their review also highlights significant gaps in stochastic and dynamic VRP modeling.

Together, these studies illustrate VRP's evolution beyond classical delivery models toward multi-objective, multi-modal, and dynamic frameworks. Integrating drones (see [27, 14]) and satellite depots (see [54]) supports granular, responsive logistics systems. In public-safety and postal services, VRP's ability to balance equity and efficiency remains critical, as demonstrated in [16, 36]. The confluence of these advancements reflects VRP's growing rel-

evance in addressing complex, real-world routing challenges that demand adaptability, real-time response capabilities, and multi-objective optimization.

## 2.3 Approaches

VRP is the process of selecting feasible routes out of exponentially many selections of any combination of customers with determined demands. There exist three types of integer programming formulations in literature for VRP, which are based on: Commodity flow formulations [10], vehicle flow models [29], and set partition problem [1]. According to its NP-hardness, exact methods are suitable for small instances only. Branch and bound, branch and cut and dynamic programming algorithms are exact methods applied by researchers. A comprehensive overview of exact methods for CVRP and its other variations, is provided in [58].

The first algorithmic approach for VRP, has been provided in [15]. Their algorithm was a simple matching-based heuristic. Afterwards, the Dantzig and Ramser's approach was improved by an effective greedy approach called the savings algorithm, [12]. Generally heuristics for VRP are clustered into two categories:

1. Constructive methods: In this type of methods, tours are constructed gradually by adding nodes to them or by combining subtours in a way not to exceed the capacity. Savings algorithm in [12] is the most famous heuristic of this type.
2. Two-phase methods: These methods solve the problem in two phases, the clustering phase and the routing phase. According to [50], there are two types of two-phase methods, the cluster-first, route-second and the route-first, cluster-second.

This paper suggests a second type heuristic in “cluster-first, route-second” order. So we focus on the same two-phase algorithms in literature. Clustering phase might be done by different approaches. A sweep algorithm is proposed in [22]. The authors consider the depot as the origin of the plane and order

customer points according to their argument in polar coordination system. Then the points are assigned respectively to vehicles up to fulfilling its capacity. In the second phase, they propose an iterative procedure to improve the route of any vehicle. Sweep algorithm has been widely used for the first phase of the algorithm.

A popular two-phase algorithm is provided in [21]. The authors apply the generalized assignment problem for clustering phase and find routes of any cluster, using any TSP method. Sweep algorithm for clustering phase and nearest neighbor approach for routing phase for public transportation problem is applied in [42], in which capacity of the vehicle may vary during their tour, since some passengers may end their trip before the route is completed. The problem of routing drones, in which the combination of sweep algorithm and genetic algorithm is applied to solve VRPD, [19]. The original sweep algorithm in [22] for VRPTW, is applied in [26]. Authors in [17] implemented first-stage customer clustering, then second-stage routing subproblems separately for conventional and electric fleets, enhancing mixed-fleet planning. A two-stage approach is suggested in [32]: first, reduce the network using A\* shortest paths; second, route using an enhanced GA with large-neighborhood search for vehicles with charging considerations. Moreover, a novel 2-phase approach is suggested in [44], which strategically separates customer location/routing decisions and operational routing, integrating a hybrid MILP and MCDM approach for service-option VRPs.

Other commonly used methods for clustering are based on data-mining algorithms. The K-means algorithm is a common procedure for the clustering phase of CVRP. The goal is to divide data into K clusters, to minimize the inner-group dissimilarity. Cluster's dissimilarity is measured by the average distance between cluster center and dataset points.

The multi-depot heterogeneous fleet VRPTW is provided in [18]. Authors provide a 3-phase hierarchical procedure. In clustering as the first phase, a heuristic is integrated into an optimization framework. Clusters of nodes are defined first, then points of the clusters are sequenced on the related tours and finally, the routing and scheduling for each tour are separately found, in terms of the original nodes. A multi-phase algorithm based on K-means clustering is developed in [33] for multi-depot VRP. The savings algorithm for

the cumulative VRP with limited duration is developed in [11] where the load is also considered in the objective function as well as distance. The authors provide a K-means algorithm for clustering in which centroids are updated iteratively, according to a square error function, based on the angle of the line from the depot and any point of the cluster. The K-means method is used for clustering in [38], in order to adjust size of the clusters, decide whether to exchange points between two clusters or not, by calculate the value of the objective function. The authors in [13] used three hierarchical algorithms: K-means, K-medoids and DBScan. Generally, their clustering method is to randomly determine the first  $K$  centers and then assign each customer point to the closest center. New center of any cluster switch to the point possessing the mean value of all objects in that cluster. The procedure repeats till center points remain unchanged. Since clusters may not be feasible at the end, capacity control of the clusters is done by an MILP to make them ready for the routing phase.

Authors in [51] initialized a number of clusters and centroids manually and assign any customer to the closest center and in any iteration update centers as the mean value of the cluster. They also apply saving matrix method in the routing phase. K-medoids clustering is evaluated for CVRP in [6]. An existing meta-heuristic is applied for routing of each generated cluster. Recently, the bi-objective green delivery and pick-up problem has been considered in [20]. The authors used a K-means based algorithm for clustering and a genetic algorithm, for routing phase.

An overview of papers targeting different variations of VRP with two-phase methods is given in Table 1.

### 2.3.1 CHB approaches for routing phase

Since the order of nodes in the optimal tour of TSP with Euclidean distance, is the same order in their convex-hull, there are plenty of algorithms in the routing phase of VRP, based on convex-hull of the points. Accordingly, a fast algorithm is proposed in [23], in which the convex hull of the points are constructed and then for each point not contained in the convex-hull find the edge of it, such that the saving introduced in [12] from the non-contained

Table 1: An overview of cluster-first, route-second two-phase methods for VRP

Authors	The first phase approach	The second phase approach	Problem
Gillet and Miller (1974), [22]	Sweep algorithm	An iterative procedure to improve routes	VRP
Fisher and Jaikumar (1981), [21]	Generalized Assignment Problem(GAP)	Any TSP method	CVRP
Nurkhalo et al. (2002), [42]	Sweep algorithm	Nearest neighbor search	VRP for public transport
Dondo and Cerdá (2007), [18]	Hierarchical procedure	Optimization framework	Multi-depot VRPTW
Nallusamy et al. (2009), [38]	K-Means	Genetic algorithm	Multi-VRP
Luo and Chen (2014), [33]	K-means	3-phase algorithm: Local search + Binary tournament + Cluster adjustment	Multi-depot VRP
Cinar, Gakis, and Pardalos (2016), [11]	K-means	Modified Savings Matrix	Cumulative VRP
Comert et al. (2017), [13]	K-means, K-medoids and DBScan	MIP	VRPTW
Singanamala, Reddy, and Venkataramaiah (2018), [51]	K-means, Savings matrix method	Ant colony	Multi-depot VRP
Bruwer (2018), [6]	K-medoids	Ruin and Recreate method	CVRP
Hertrich, Hungerländer, and Truden (2019), [26]	Sweep algorithm	The same order in sweep algorithm	VRPTW
Euchi and Saduk (2020), [19]	Sweep algorithm	Genetic algorithm	VRPD
Chen, Gu, and Gao (2020), [9]	Exact assignment algorithm	Genetic algorithm	Multi-depot VRPTW
Fatemi-Anaraki et al. (2021), [20]	K-means	Genetic algorithm	Delivery and pick-up problem
Ding, Li, and Hao (2023), [17]	Clustering algorithm (Nearest ID)	Heuristic routing	VRP with mixed fleet
Liu et al. (2023), [32]	Improved sweep algorithm	The same order in sweep	2-depot CVRP
Pournohammadreza and Jokar (2023), [44]	A-star search algorithm	enhanced genetic algorithm	VRP with charging relief
Sehita and Thakar (2023), [49]	sweep algorithm	Genetic algorithm	CVRP

node is minimal. In this way, all non-contained nodes are assigned to an edge of the convex-hull, and among them, a node with the minimum saving number is added to the convex-hull. The procedure is repeated until all nodes are covered. They also showed that the computational complexity of the algorithm is of the order  $O(n^2 \log n)$ , while its worst case is unknown. The algorithm has been modified in [55] to improve the results.

Authors in [28] proposed a strange heuristic for TSP based on convex hull. A blob is located over the set of nodes which are projected into the lattice. The blob is gradually reduced until it passes all nodes in its edge. The initial shape is the convex hull of the points. It is then shrunk by systematically removing some of its constituent particle components. The points act as attractants to the material, effectively “snagging” the material at the locations of uncovered nodes and affecting its subsequent morphological adaptation. As the material continues to shrink all data points, it is becoming a concave area covering all nodes. The classic TSP solutions are enhanced with a convex hull insertion method, in [24], by providing a systematic and fast way to construct near-optimal tours. Authors in [41] enhanced classic TSP solutions with a convex hull insertion method, providing a systematic and fast way to construct near-optimal tours.

A CHB method has also been applied for VRP. The mathematical model for CVRP is considered in [45]. Authors provide a decomposition algorithm for capacity constraints and apply a separation problem to identify nodes violating this constraint. They use convex hull of incidence vectors of all TSP tours, and these tours are tested to find the violated capacity constraints. The convex hull of the points is used in [43] to measure the visual attractiveness of the solution. A saving based algorithm is applied to solve an extended version of VRP. The min-max multi VRP is introduced in [39], in which there are multiple depots, and the objective is to minimize the maximum length of the tour traversed by vehicles. The author uses the convex hull of all nodes containing customer and depot points to find the whole region at hand and then applies Carlsson algorithm in [8] to partition the region for multiple depots.

Authors in [47] consider four criteria to measure visual attractiveness of the routes containing “number of nodes belong to more than one convex hull”.

They propose a heuristic in which the farthest node from the depot is chosen as the seed of a new route, then the surrounding nodes are added to it, until the capacity limit is reached. The whole procedure is repeated until all nodes are routed. After building routes, they find the nodes locating in the convex hull of another route and apply a “Merge-And-Rebuild” process to fix it. In a recent research, the convex hull of customer locations is applied to select initial seed clients, followed by an exchange operator to improve solutions in VRPTW, [49].

### 3 Problem statement

The problem addressed in this paper consists of designing efficient routes for  $K$  identical vehicles to service a set of customers with known demands. More precisely, CVRP is described as an undirected weighted graph  $G = (V, A, c)$ , where  $V = \{0, 1, \dots, n\}$  is the set of vertices, in which point 0 is the depot point and  $\{1, \dots, n\}$  is the set of customers, and  $A = \{(i, j) : i \in V, j \in V, i \neq j\}$  shows the set of arcs. homogeneous vehicle fleets, each with capacity  $Q$ , start their route from the depot and end to it after visiting a subset of customers according to their limited capacity. Moreover,  $d_{ij}$  is the Euclidean distance between nodes  $i$  and  $j$ , while  $q_i > 0$  shows the customer  $i$ 's demand. The problem is solved under the following constraints:

1. Each customer is serviced only once by one vehicle.
2. Each vehicle must start and end its route at the depot.
3. Total demand met by each vehicle cannot exceed  $Q$ .

### 4 The two-phase CHB heuristic

CVRP is solved using a proposed two-phase CHB heuristic.

**Clustering Phase:** This phase is to assign customers to clusters.

- Customers are partitioned into  $K$  clusters using a novel CH-means algorithm, where each cluster is assigned to a vehicle.

- Cluster centroids are updated based on the convex hull of the nodes within each cluster.

**Routing Phase:** This phase is to find efficient routes between nodes of any cluster.

- Efficient routes between nodes in each cluster are constructed using the proposed CH-insertion heuristic.
- The routing procedure is executed at the end of any iteration of the clustering phase to retain the best solution by the end of Phase 1.
- The fitness of a solution is defined as the total length of all routes in that solution.
- Finally, an ant colony optimization procedure refines the routes for further improvement.

Algorithm 1 explains the CHB heuristic.

#### 4.1 Clustering phase: CH-means method

A CH-means algorithm is provided here, while the number of clusters is already known, and it is equal to the number of vehicles. Centroids are updated after a complete iteration of the algorithm. The procedure is repeated until the distance between centroids in two consecutive iterations in all clusters, does not exceed a predetermined threshold. The steps of the CH-means algorithm are described in the following.

##### Step 1: Create initial centroids and distance matrix

Initially, the whole region is divided into  $K$  regions by plotting  $K$  lines originating from depot. To do this, customer points are sorted according to their arguments in polar coordination, where the depot is assumed as the origin. The whole region containing customers is identified by polar coordination  $[\theta_{\min}, \theta_{\max}]$ . The region is divided to  $K$  cones by  $K - 1$  equidistant rays

---

**Algorithm 1** Two-phase CHB algorithm for CVRP

---

**Input:**  $V = \{0\} \cup \{1, \dots, n\}$  depot point and customer points,  $q$  demand vector,  $Q$  vehicle capacity,  $[X, Y]$  Cartesian coordination of vertices and  $K$ : No. of vehicles, parameter  $\epsilon$ .

**Output:**  $K$  routes start from and end to depot point  $0 \in V$ , s.t. all customer points exists in exactly one route and the sum of customer demands in all routes do not exceed  $Q$ .

- 1: **while** The distance between centroids is more than  $\epsilon$  **do**
  - 2:     Cluster customer points into  $K$  clusters  $clus_j, j = 1, \dots, K$ , by Algorithm 2.
  - 3:     **for**  $j = 1, \dots, K$  **do**
  - 4:         Find initial route of  $Clus = clus_j \cup \{0\}$  by Algorithm 3, call it  $tour_j$ .
  - 5:     **end for**
  - 6:     Calculate the fitness of  $Clus$ . Save it if it is the best clustering found by now.
  - 7: **end while**
  - 8: Find the best clustering and its routing and call it  $tour^*$ .
  - 9: Improve  $tour^*$  by the ant colony optimization of Algorithm 4.
- 

originating from depot. The middle points of cones  $\frac{r}{2}$  far from the depot, are forming initial centroids, where  $r$  is the half of the maximum distance from depot. Figure 1 shows two different examples.

Distance matrix  $D$  is calculated with entities  $d_{ij}$  as the Euclidean distance between customer point  $i$  and centroid  $j$ .

### Step 2: Assign customers

The first assignment is related to the minimum entity of  $D$ , say  $d_{pq}$ , if vehicle  $q$  has enough empty capacity for customer  $p$ . This step is repeated till all points are assigned.

**Remark 1.** In some cases, some points might be remained unassigned, while no other cluster has enough empty capacity to meet their demand. There are

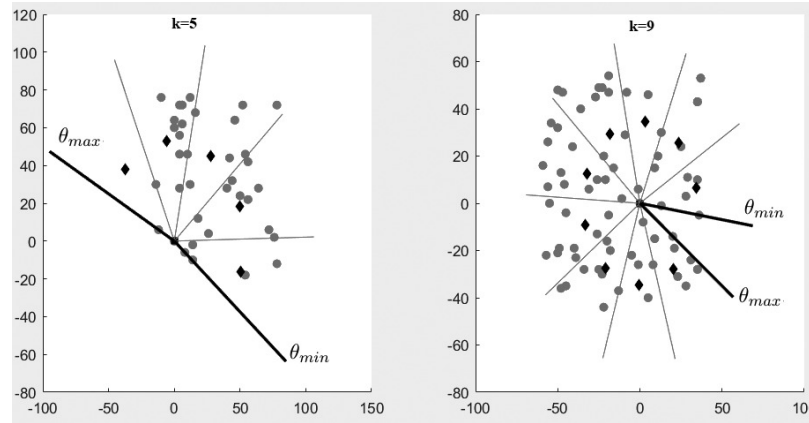


Figure 1: Dividing the whole region by  $K$  rays, diamond black points are initial centroids.

few approaches in literature proposing some methods to convert uncapacitated clusters to capacitated ones, [2, 7]. Most researches open a new cluster for remained points, [40, 37]. Since with the proposed method, this case was so rare, a new cluster is opened for these points instead of undertaking the cost of cluster improving methods. However, this kind of samples will be pointed out in Section 6.

### Step 3: Calculate fitness

At the end of step 2, the fitness of the obtained clustering is calculated. This is done by creating a route for any cluster according to the proposed procedure in subsection 4.2. The fitness is saved and at the end of the whole algorithm, the best solution is selected to be improved by an ant colony optimization method.

### Step 4: Update centroids and distance matrix

Centroids are updated, after a complete assignment. The centroid of any cluster is the mean point of the vertices of the convex hull of all points assigned to that cluster. Subsequently, distance matrix  $D$  is updated according

to new centroids. Steps 2 and 3 are repeated until the distance between all corresponding centroids in two consecutive iterations do not exceed a predetermined threshold. This is mentioned by *approximately unchanged centroids* in Algorithm 2. This algorithm explains the whole procedure in CH-means algorithm.

## 4.2 Routing phase: CH-insertion method

This section presents the CH-insertion algorithm for routing points within a cluster, which is clearly a TSP. Firstly, the convex hull of the current cluster's points is computed to form the initial polygonal route. Afterwards, for each unrouted point on that cluster say  $p$ , all polygon edges say  $(i, j)$  are identified, for which the triangle  $i - p - j$  is acute. Next, any unrouted point is assigned to its closest edge having this property. The insertion cost  $height_p$  is calculated as the perpendicular height from  $p$  to edge  $(i, j)$ , guaranteed to lie within the triangle due to the acute angle condition. Among all unrouted points  $p$ , the one, possessing minimum  $height_p$  is found and its associated edge  $(i, j)$  is broken into two sides  $(i, p)$  and  $(p, j)$ . The procedure is repeated until all nodes are incorporated into the route. The complete algorithm is formalized in Algorithm 3.

Figure 2 shows steps of Algorithm 3 for an instance. Part (a) is the initial step in which the convex hull of the points is found. In the next step, distance of any point in  $\bar{N}$  to its associated side is calculated, and the nearest one (E) is selected and is added to the polygon. Point B is also added in the next step. Part (f) shows the final rout. It should be noted that the clusters are feasible at the beginning of the CH-insertion algorithm, since these points are resulted from the previous clustering step.

**Algorithm 2** CH-means algorithm for clustering

**Input:**  $V = \{0\} \cup \{1, \dots, n\}$  depot point and customer pints,  $q$  demand vector,  $Q$  vehicle capacity,  $[X, Y]$  Cartesian coordination of vertices,  $K$  number of vehicles.

**Output:** Clusters  $clust_j$  for  $j = 1, \dots, K$ .

- 1: Set initial centroids as the middle point of cones  $\frac{r}{2}$  far from the depot.  $\triangleright$   
 $r$ : half of the maximum distance from depot
- 2: Construct distance matrix  $D$ . Let  $best = \infty$   $\triangleright$  (Initialization)
- 3: Let  $cap_j = 0$  for  $j = 1, \dots, K$ .  $\triangleright$  ( $cap_j$  is the occupied amount of capacity of vehicle  $K$ ).
- 4: **while** All centroids approximately remain unchanged **do**
- 5:   let  $AC = \phi$ .  $\triangleright$  ( $AC$  is the set of assigned customers).
- 6:   **while**  $AC \neq \{1, \dots, n\}$  **do**
- 7:      $d_{rp} = \min\{d_{ij}, i = 1, \dots, n, j = 1, \dots, K\}$ .
- 8:     **if**  $d_{rp} < \infty$  **then**
- 9:       **if**  $cap_p + q_r \leq Q$  **then**
- 10:          assign customer  $r$  to vehicle  $p$ :  $clust_p = clust_p \cup \{r\}$ ,
- 11:           $cap_p \leftarrow cap_p + q_r$   $\triangleright$  (Update occupied capacity of vehicle  $p$ )
- 12:          Let  $d_{rj} = \infty$ , for  $j = 1, \dots, K$   $\triangleright$  (Avoid reassigning customer  $r$ )
- 13:          Let  $AC \leftarrow AC \cup \{r\}$ ,
- 14:       **else**
- 15:          Let  $d_{rp} = \infty$   $\triangleright$  (Avoid reconsidering customer  $r$  for vehicle  $p$ )
- 16:       **end if**
- 17:     **else**
- 18:       Open a new cluster  $clust_{K+1}$ .
- 19:     **end if**
- 20:   **end while**
- 21:   Find the convex hull of all clusters.
- 22:   Update centroids of clusters as the mean value of vertices of their convex hull.
- 23:   Update distance matrix  $D$ , according to new centroids.
- 24: **end while**

**Algorithm 3** CH-insertion algorithm for routing

**Input:** Clusters  $clust_j$  for  $j = 1, \dots, K$ .  $[X, Y]$ : Cartesian coordination of vertices.

**Output:** Tours  $tour_j$  for  $j = 1, \dots, K$ , each starts and end to depot.

```

1: for  $j = 1, \dots, K$  do
2:   Let  $H$  = the convex hull of points in  $clust_j$ ,
3:   let  $N$  be the points in  $H$ ,  $\bar{N} = clust_j \setminus N$ .    ▷ (Initialize routed and
   unrounded points)
4:   while  $\bar{N} \neq \phi$  do
5:     for  $h \in \bar{N}$  do
6:       Let  $(i, j)_h = \arg \min_{(i, j) \text{ is a side of } H} \{height_h^{ij} \text{ s.t. triangle i-h-j is acute}\}$ .
7:     end for
8:     Let  $height_p = \min\{height_h^{(i, j)_h}, h \in \bar{N}\}$ .
9:     Update  $H$  by breaking the side  $(i, j)_p$  into two sides  $(i, p)$  and
    $(p, j)$ .
10:    Update  $N \leftarrow N \cup \{p\}$  and  $\bar{N} = \bar{N} \setminus \{p\}$ .
11:   end while
12:    $tour_j$  is the set of ordered points in polygon  $H$  starting from 0.
13: end for

```

### 4.3 Computational complexity of the two-phase CHB heuristic

This section shows that the complexity of the CHB algorithm is polynomial in terms of the number of customers.

**Lemma 1.** The computational complexity of the proposed two-phase CHB heuristic algorithm is  $O(n \times IT \log(n))$ , where  $IT$  is the maximum number of iterations and  $IT \approx \gamma \frac{1}{\epsilon}$ , with  $\gamma \in [5, 10]$  depending on the benchmark samples used.

*Proof.* The computational complexity is analyzed in two phases:

Clustering Phase:

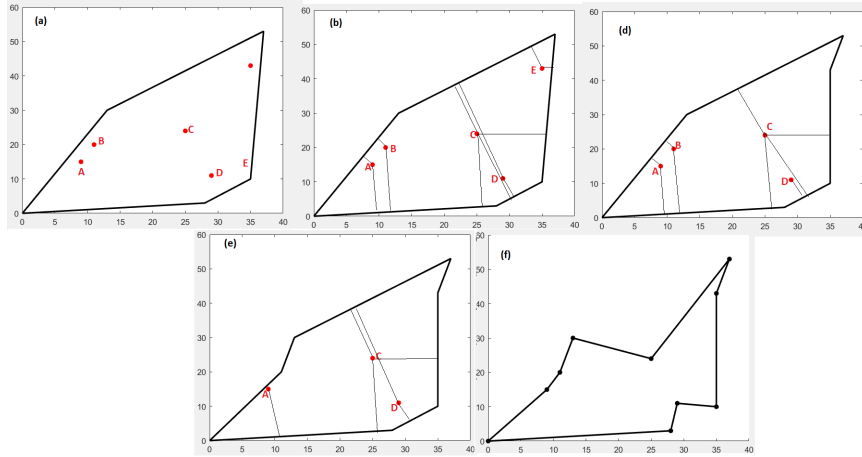


Figure 2: Steps of CH-insertion Algorithm 3

- Constructing the distance matrix for  $n$  nodes and  $K$  clusters costs  $O(nK)$ .
- Finding the minimum element in the distance matrix, which contains  $nK$  elements, requires  $O(nK \log(nK))$ .
- Assigning nodes to vehicles until all points are assigned takes  $O(n)$ .

Hence, the complexity of clustering in one iteration is  $O(nK \log(nK))$ , which simplifies to  $O(n^2 \log(n))$  when  $K = O(n)$ . At the end of any single iteration of clustering, routing procedure is implemented. First, we compute the number of computations in one single cluster.

Routing Phase:

- In one cluster, each vehicle serves approximately  $h = \frac{n}{K}$  customers.
- Constructing the convex hull of  $h$  points using standard algorithms (e.g., CGAL or SciPy) costs  $O(h \log(h))^*$ .
- Calculating the distances of all inner points from all sides of the convex polygon takes  $O(h^2)$ .
- Sorting these distances requires  $O(h^2 \log(h))$ .

\* Available at: <https://www.scipy.org>.

The total complexity for routing a single cluster is  $O(h^2 \log(h) + h^2) = O(h^2 \log(h))$ . For  $K$  clusters, the overall complexity of the routing phase is  $O(K \cdot h^2 \log(h))$ , which reduces to  $O(n \log(n))$  when  $h \approx \frac{n}{K}$ .

#### Overall Complexity:

The clustering and routing phases are repeated until the centroids' displacement across iterations does not exceed a threshold  $\epsilon$ . Since  $IT$  denotes the maximum number of iterations, the overall complexity of the two-phase CHB algorithm is then  $O(n \times IT \log(n))$ .  $\square$

## 5 Improving routes

To further improve the routes found by Algorithm 3, an ant colony optimization method in [3] is followed. The difference here is that the initial routes are the same routes found by CHB heuristic, rather than random ones. To be more precise, the amount of initial pheromones are not random numbers. The algorithm in [3] is provided to solve CVRP, while we apply it on any single route as TSP.

Ant colony optimization algorithm has several parameters that should be determined by user, and there is no deterministic certificate to show which value is the best. In different applications, different values may behave better. Table 2 introduces parameters of ant colony optimization algorithm.

Table 2: Parameters of ant colony optimization algorithm

Parameter	Definition	Parameter	Definition
$MaxIT$	Maximum no. of iterations	$\alpha$	Pheromone importance
$\tau_0$	Initial pheromone on each arc	$\beta$	Distance importance
$\eta_{ij}$	Inverse of $d_{ij}$	$\rho$	Evaporation coefficient
$\tau_{ij}$	Amount of pheromone on arc $(i, j)$	$r_0$	A constant
$d_{ij}$	Distance between $i$ and $j$	$a$	Index of ant, $a \in \{1, \dots, m\}$
$E_P$	Arcs of tour P		

It should be noted that the initial pheromone  $\tau_0$  is usually supposed to be the inverse of the best-known route distance found for that particular problem. Here  $\tau_0$  is evaluated in a way to strengthen the arcs in the tour found by Algorithm 3. So initially, we set  $\tau_0 = 0.1$  for arcs not included in the tour

and  $\tau_0 = 1$  for arcs in it. Following steps explain ant colony method in details.

---

**Algorithm 4** Ant colony optimization algorithm to improve routes of CHB heuristic

---

**Input:** Tour  $P$  starts and ends to depot.

**Output:** Improved tour  $Best_{tour}$  starts and ends to depot.

```

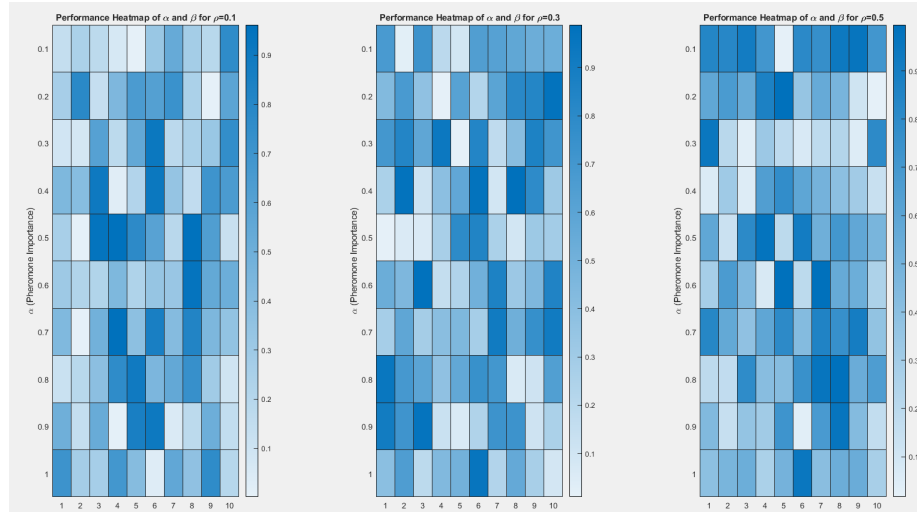
1: Initialize parameters according to Table 2,  $Best = \inf$ . ▷ (Initialization)
2: Let  $\tau_{ij} = 1$ , for all  $(i, j) \in E_P$  and  $\tau_{ij} = 0.1$  for all  $(i, j) \notin E_P$ .
3: for  $Iter = 1, \dots, MaxIT$  do
4:   for  $a = 1, \dots, m$  do
5:     Locate ant  $a$  in depot. Let  $i = 0$ . ▷ (The first location of ant  $a$ )
6:      $Tour = \{0\}$ 
7:     while  $|Tath| < n + 1$  do
8:       Choose random number  $r \in [0, 1]$  by uniform distribution.
9:       if  $r \leq r_0$  then
10:        let  $j = \underset{u \in Tour}{a} \operatorname{rg} \max(\tau_{iu})^\alpha (\eta_{iu})^\beta$  ▷ (Choose the next node by
            instinct)
11:       else
12:        For any unvisited customer  $w$ , let  $P_{iw} = \frac{(\tau_{iw})^\alpha (\eta_{iw})^\beta}{\sum_{u \notin Tour} (\tau_{iu})^\alpha (\eta_{iu})^\beta}$ 
13:        Choose next node  $j$  according to distribution function of  $P$ .
14:       end if
15:       Update  $Tour := [Tour, j]$ 
16:       Update pheromone on arc  $(i, j)$  by  $\tau_{ij} = (1 - \rho)\tau_{ij} + \rho\eta_{ij}$ 
17:       Let  $i = j$  ▷ (Update current node)
18:     end while
19:     Let  $Tour = [Tour, 0]$ ,  $L_a$  =sum of lengths of arcs in  $Tour$ .
20:     if  $L_a < Best$  then
21:        $Best = L_a$  and  $Best_{Tour} = Tour$ 
22:     end if
23:   end for
24:   for  $(i, j) \in Best_{Tour}$  do
25:      $\tau_{ij} = (1 - \rho)\tau_{ij} + \rho \sum_{a=1}^m \Delta_{ij}^a$ , where  $\Delta_{ij}^a$  is obtained by (1) ▷ (Global
        pheromone update)
26:   end for
27: end for

```

---

The formula (1), globally updates the pheromone.

$$\Delta_{ij}^a = \begin{cases} \frac{1}{L_a} & \text{If ant } a \text{ passes arc } (i, j) \\ 0 & \text{Otherwise} \end{cases} \quad (1)$$

Figure 3: Performance heat map of  $\alpha$ ,  $\beta$ , and  $\rho$ .

The ant colony optimization algorithm may not change the solution by CHB heuristic, or even it may make it worse in some cases. However in most cases it improves the solution. There exists 3 parameters in this algorithm which should be initialized,  $\alpha$ , performance importance,  $\beta$ , distance importance and  $\rho$ , evaporation coefficient. All test problems were solved with all values  $\alpha \in \{0, 0.1, \dots, 1\}$ ,  $\beta \in \{1, 2, \dots, 10\}$  and  $\rho \in \{0.1, 0.3, 0.5\}$  for ten times and the average value of the final objective function is the performance of each triple  $(\alpha, \beta, \rho)$ . Figure 3 shows the results of these implementations. Note that the value of the performance is the value of the objective function, normalized between 0 and 1. Values 0 and 1 possess the lightest and the darkest color, respectively.

It can be seen that  $(\alpha, \beta, \rho) = (0.1, 5, 0.1)$  provides the best average performance among all other choices. These values are fixed through all implementations.

## 6 Computational experiments

The proposed method is implemented in MATLAB 2020, 64 bit, and is run on Intel(R) Core(TM) i7-5500U CPU @ 2.4 GHz and 8 GB of RAM. We implement the proposed CHB heuristic in this way: First, nodes are clustered according to the CH-means algorithm 2, and find the respective routing by applying the CH-insertion algorithm 3 in any iteration. Afterwards, the best solution is improved by the ant colony algorithm 4 on CVRP benchmark problems from *CVRPLib*\*\*.

We apply the CH-means method with other insertion methods in literature. Furthermore, we apply the K-means method with the proposed CH-insertion method and compare the results. Table 3 explains all implemented approaches in this section.

Table 3: Implemented approaches for comparison

Clustering Method:	CH-means			K-means		
Routing Method: (insertion)	Clark & Wright insertion	Convex hull Nearest insertion	CH-insertion	Clark & Wright insertion	Convex hull nearest insertion	CH-insertion
Name:	CHmean-ClarkInsert	CHmean-NearestInsert	CHmean-CHInsert	Kmean-ClarkInsert	Kmean-NearestInsert	Kmean-CHInsert

The selection of K-means for comparison, is backed by its wide applicability in clustering problems. Two insertion methods, Clark & Wright method and convex hull nearest insertion method are also selected to be compared to CH-insertion method, for the sake of their speed and accuracy in compare with other insertion methods, [23]. All methods for TSP mentioned in [23] were tested on samples, and these methods resulted in best solutions among others. Therefore results show that Clark & Wright and convex hull nearest insertion methods are appropriate methods to be compared to the proposed CH-Insertion. First of all, we briefly explain the Clarke and Wright savings method and convex hull nearest insertion.

---

\*\* available at: <https://neo.lcc.uma.es/vrp/vrp-instances/capacitated-vrp-instances/>

### Clarke and Wright savings method

In the typical method in [12], savings  $s_{ij} = c_{1i} + c_{1j} - c_{ij}$  for all customer points  $i$  and  $j$  are computed. Savings are ordered from largest to smallest. Initially, subtours  $(0, i, 0)$  are formed for all customer points  $i$ . In any iteration, two subtours containing  $(0, i)$  and  $(j, 0)$ , are merged, possessing the maximum amount of  $s_{ij}$  in the savings matrix. The new merged subtour contains  $(i, j)$ . The procedure is repeated until all points are routed.

### Convex hull nearest insertion

In this method, the convex hull of nodes in the given cluster is formed as an initial subtour, [55]. To each node  $d$  not yet contained, side  $(i, j)$  of the hull is assigned if minimizes  $c_{di} + c_{dj} - c_{ij}$ . Next,  $(i^*, d^*, j^*)$  is determined, which minimizes  $(c_{i^*d^*} + c_{d^*j^*})/c_{i^*j^*}$ . Finally, node  $d^*$  is inserted in the subtour between nodes  $i^*$  and  $j^*$ . The procedure is repeated until all points of the cluster are routed.

We test the proposed CHB heuristic and the methods mentioned in Table 1, on three groups of benchmark problems: set A (Augerat, 1995), set E (Christofides and Eilon, 1969) and set P (Augerat, 1995). Tables 4, 5, and 6 show the results for set A, set B and set C, respectively. The last column of all tables show the optimal value. All optimal solutions and values are accessible from *CVRPlib*. The underlined instances are those with remained nodes without any enough capacity, left in vehicles as mentioned in Remark 1.

Based on the results in most instances, the proposed CHB heuristic is as good as other methods. The CHB heuristic, in some instances such as A-n44-k6 results in the worst cost, while in some other cases such as P-n76-k5 in both phases or in E-n101-k8 in routing phase, it achieves the best solution among other approaches.

Figure 4 shows the results more clearly. Any point in this figure shows the relative error of the corresponding method based on optimal value of the corresponding instance. The black line in left-hand figures is related to the

Table 4: Cost comparison of CVRP's instances: Set A

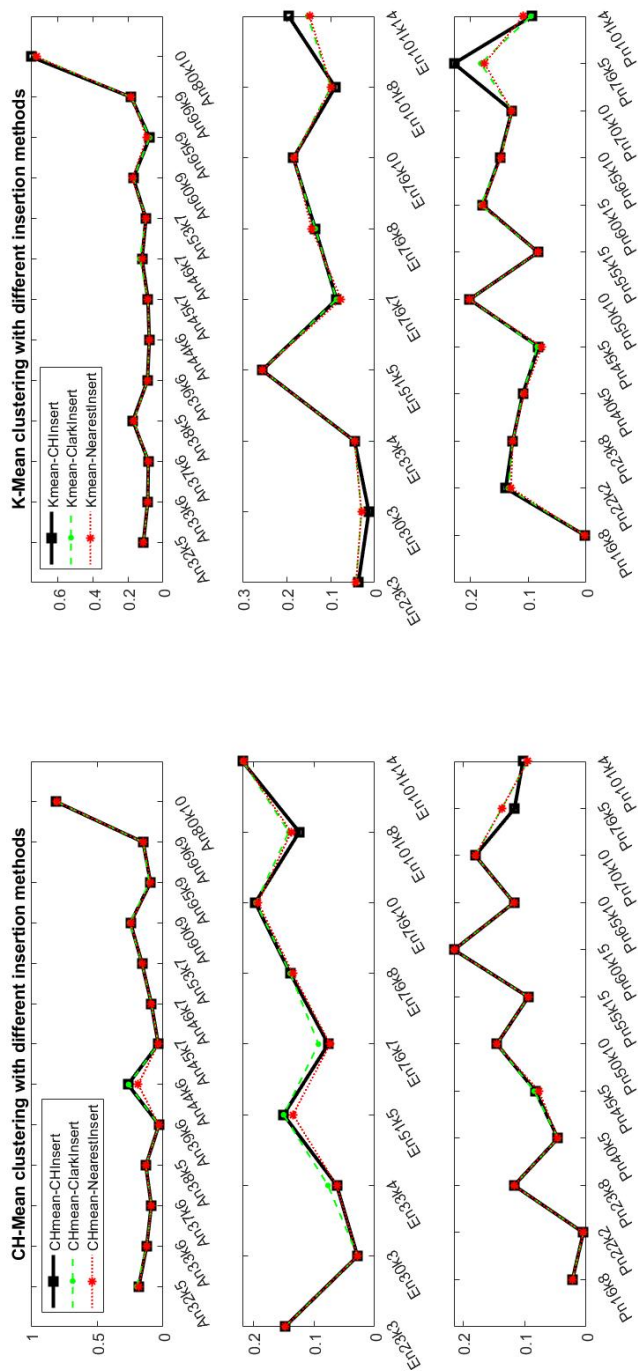
Benchmark instance	Clark Insert	CHmean-Nearest Insert	CH Insert	Clark Insert	Kmean-Nearest Insert	CH Insert	Optimal Value
<b>A-n32-k5</b>	934	928	928	872	872	872	784
<b>A-n33-k6</b>	834	834	834	807	807	807	742
<b>A-n37-k6</b>	1034	1034	1034	1029	1029	1029	949
<b>A-n38-k5</b>	825	825	825	853	853	856	730
<b>A-n39-k6</b>	857	857	857	905	905	905	831
<b>A-n44-k6</b>	1182	1114	1186	1011	1011	1011	937
<b>A-n45-k7</b>	1188	1188	1188	1245	1245	1245	1146
<b>A-n46-k7</b>	995	995	995	1025	1024	1024	914
<b>A-n53-k7</b>	1172	1171	1171	1109	1109	1190	1010
<b>A-n60-k9</b>	1623	1621	1625	1526	1529	1526	1354
<b>A-n65-k9</b>	1439	1437	1439	1415	1421	1415	1174
<b>A-n69-k9</b>	1350	1350	1351	1386	1388	1388	1159
<b>A-n80-k10</b>	2094	2089	2096	2000	1999	2039	1763

Table 5: Cost comparison of CVRP's instances: Set E

Benchmark instance	Clark Insert	CHmean-Nearest Insert	CH Insert	Clark Insert	Kmean-Nearest Insert	CH Insert	Best Found
<b>E-n23-k3</b>	653	653	653	592	592	592	569
<b>E-n30-k3</b>	549	549	549	550	550	547	534
<b>E-n33-k4</b>	889	887	887	873	873	873	835
<b>E-n51-k5</b>	599	591	599	654	654	654	521
<b>E-n76-k7</b>	735	732	734	741	734	741	682
<b>E-n76-k8</b>	837	834	837	834	835	834	735
<b>E-n76-k10</b>	992	989	993	983	982	983	830
<b>E-n101-k8</b>	930	927	926	896	895	891	815
<b>E-n101-k14</b>	1298	1297	1299	1229	1224	1275	1067

two-phase CHB method and the black line in right-hand figures is related to K-means with CH-insertion method. It is clear that black line manages to achieve the least relative error in some instances, both in right-hand and left-hand figures. This means that the proposed two-phase CHB heuristic is efficient in both predicting clusters and routing customers. Therefore, both proposed phases are efficient to be combined with other methods and also to be applied independently, as well.

Figure 5 also shows the percentage of obtaining the best cost solution among other approaches for different three sets A, E and P. In some cases two approaches result in the same solution, such as CH-means clustering with Clark routing and nearest insertion routing. This is shown by phrase "CHmean-Clark & NearestInsert" in the figure. Moreover, the phrase "CHmeans-Anycluster" means that all three methods for routing with CH-



ing problem

Figure 4: Comparison of relative error of cost different approaches according to optimal value, divided by clustering methods.

Table 6: Cost comparison of CVRP's instances: Set P

Benchmark instance	Clark Insert	CHmean-Nearest Insert	CH Insert	Clark Insert	Kmean-Nearest Insert	CH Insert	Best Found
<b>P-n16-k8</b>	460	460	460	451	451	451	450
<b>P-n22-k2</b>	217	217	217	245	244	246	216
<b>P-n23-k8</b>	591	591	591	596	596	596	529
<b>P-n40-k5</b>	479	479	479	508	507	508	458
<b>P-n45-k5</b>	553	549	552	553	549	552	510
<b>P-n50-k10</b>	797	797	797	836	836	836	696
<b>P-n55-k15</b>	1082	1082	1082	1071	1071	1071	989
<b>P-n60-k15</b>	1176	1176	1176	1144	1143	1143	968
<b>P-n65-k10</b>	885	885	885	909	908	910	792
<b>P-n70-k10</b>	976	976	976	933	933	933	827
<b>P-n76-k5</b>	713	713	712	741	737	770	627
<b>P-n101-k4</b>	746	746	751	746	755	749	681

means clustering, managed to achieve the best solution among other approaches. Based on Figure 5, the proposed CH-means method is successful in instances of set A and P, while CH-insertion method behaves efficiently in instances of set E.

Finally, Figure 6 compares the proposed CH-means method to the K-means method. Generally, it can be seen that the CH-means clustering methods is better in instances of set E, while in other two sets its vice versa.

## 7 Conclusion

In this paper, a two-phase CHB heuristic method has been introduced for CVRP, with computational complexity of  $o(n^2 \log n)$ , where  $n$  is the number of customers. Customers are first clustered by a CH-means algorithm. In a typical K-means algorithm, points are clustered in order to minimize the total dissimilarity between the points in a cluster. In CVRP, the length of the whole TSP tour should be minimized, because finally a TSP will be solved in any cluster. Hence, unlike typical K-means algorithm, in CH-means algorithm, initial centers are set in equidistant locations inside the region surrounded by customers. Moreover, the center of each cluster is updated according to the convex hull of the points assigned to that cluster. The

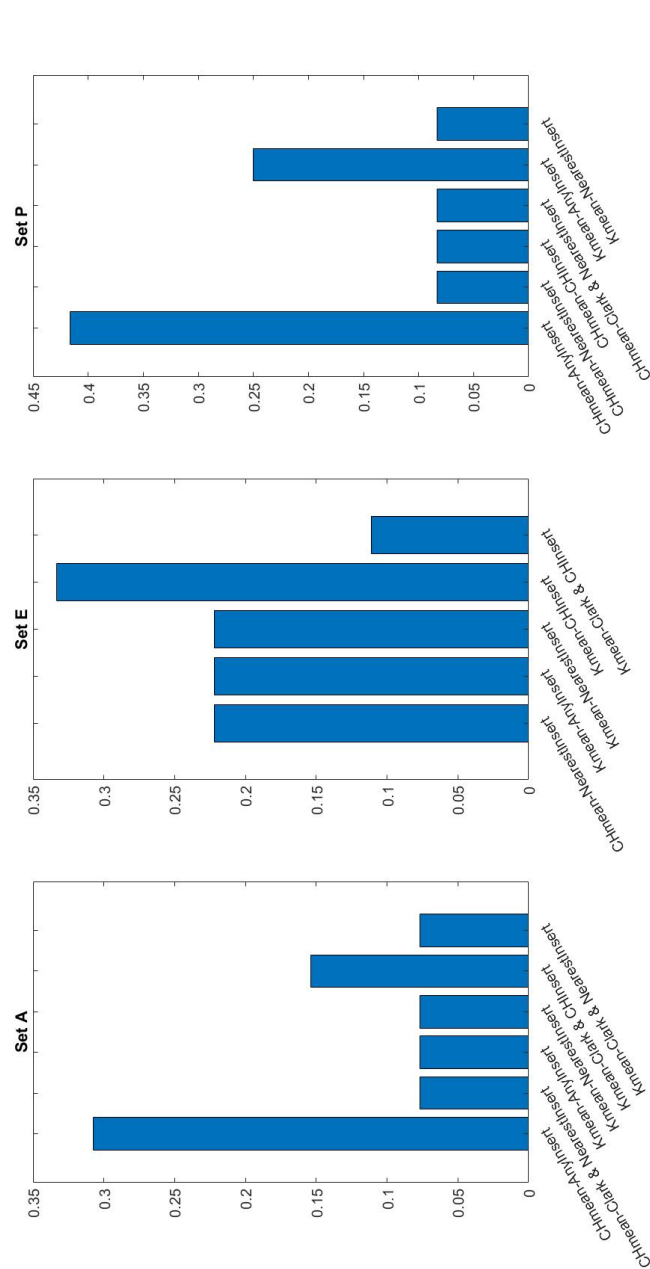


Figure 5: Comparison of relative error of cost different approaches according to optimal value, divided by clustering methods.

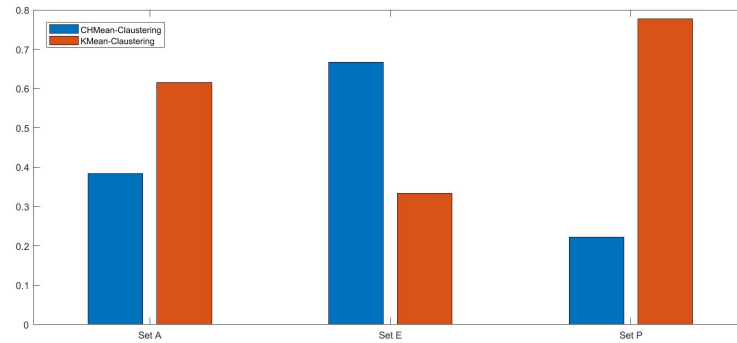


Figure 6: Percentage of achieving the best solution in approaches by CH-means clustering and K-means.

new center is the mean point of that convex hull. The reason of this choice for updating centroids, is to make any cluster contain all points on the line segment joining any two points in it. However, adding points in any step is the same with the K-means algorithm and is based on the minimum distance from center, but since closest points to the center are located inside the respective convex hull, they will be assigned to that cluster.

Compared to the typical two-phase algorithms for CVRP, the novelty of CHB heuristic is that the fitness of any step's solution is calculated by the CHB routing algorithm, which is called CH-insertion, and it is saved. At the end of the clustering phase, the best found solution is chosen for improvement, which is not necessarily the last solution found. Although this increases the computational complexity of the algorithm, but it allows the selection of the best clustering.

The routes are built through the CH-insertion method. The idea for routing is to construct the convex hull of the points and to insert unassigned points by breaking an edge of the polygon into two edges. This is based on the well-known property for Euclidean TSP by [30]. Finally, the routes are improved with a meta-heuristic algorithm. Ant colony optimization algorithm has been chosen, by virtue of its satisfactory performance on TSP.

Implementing the CHB heuristic on benchmark samples and comparing it to other two-phase heuristics show that the proposed two-phase CHB

heuristic is efficient in both clustering -even if combined with other routing methods- and routing -even if combined with other clustering methods. Moreover, the proposed CHB heuristic results in the best solution among other implemented methods in this paper.

## References

- [1] Baldacci, R., Christofides, N. and Mingozzi, A. *An exact algorithm for the vehicle routing problem based on the set partitioning formulation with additional cuts*, Math. Program., 115 (2008), 351–385.
- [2] Barreto, S., Ferreira, C., Paixao, J. and Santos, B.S. *Using clustering analysis in a capacitated location-routing problem*, Eur. J. Oper. Res., 179 (2007), 968–977.
- [3] Bell, J. and McMullen, P. *Ant colony optimization techniques for the vehicle routing problem*, Adv. Eng. Inform., 18 (2004), 41–48.
- [4] Bodin, L.D. and Golden, B.L. *Classification in vehicle routing and scheduling*, Networks, 11 (1981), 97–108.
- [5] Bodin, L.D. *The state of the art in the routing and scheduling of vehicles and crews*, Urban Mass Transp. Adm., Vol. 1, 1983.
- [6] Bruwer, F. *Petal-shaped clustering for the capacitated vehicle routing problem*, Master Thesis, Univ. Witwatersrand, Johannesburg, 2018.
- [7] Buhrmann, J.H., Campbell, I. and Ali, M. *A capacitated clustering heuristic for large datasets*, Proc. Int. Conf. Ind. Eng. Oper. Manag., 2018.
- [8] Carlsson, J., Ge, D., Subramaniam, A., Wu, A. and Ye, Y. *Solving min-max multi-depot vehicle routing problem*, Lect. Glob. Optim., 55 (2009), 31–46.
- [9] Chen, A., Gu, X. and Gao, Z. *Two-phase algorithm to multiple depots vehicle routing problem with soft time windows*, IOP Conf. Ser.: Earth Environ. Sci., 2020.

- [10] Christofides, N., Mignozzi, A. and Toth, P. *The vehicle routing problem*, In: Christofides, N. (ed.), Combinatorial Optimization, Wiley, Hoboken (1979), 315–338.
- [11] Cinar, D., Gakis, K. and Pardalos, P.M. *A 2-phase constructive algorithm for cumulative vehicle routing problems with limited duration*, Expert Syst. Appl., 56 (2016), 48–58.
- [12] Clarke, G. and Wright, J.R. *Scheduling of vehicles from a central depot to a number of delivery points*, Oper. Res., 12(4) (1964), 568–581.
- [13] Comert, S., Yazgan, H., Sertvuran, I. and Sengul, H. *A new approach for solution of vehicle routing problem with hard time window: An application in a supermarket chain*, Indian Acad. Sci., 42(12) (2017), 2067–2080.
- [14] Dang, S., Liu, Y., Luo, Z., Liu, Z. and Shi, J., *Survey of the routing problem for cooperated trucks and drones*, Drones, 8(10) (2024), 550.
- [15] Dantzig, G. and Ramser, J., *The truck dispatching problem*, Manag. Sci., 6(1) (1959), 80–91.
- [16] Dewinter, M., Vandeviver, Ch., Vander Beken, T. and Witlox, F., *Analyzing the police patrol routing problem: A review*, ISPRS Int. J. Geo-Inf., 9(3) (2020), 157.
- [17] Ding, N., Li, M. and Hao, J., *A two-phase approach to routing a mixed fleet with intermediate depots*, Mathematics, 11(8) (2023), 1924.
- [18] Dondo, R. and Cerda, J., *A cluster-based optimization approach for the multi-depot heterogeneous fleet vehicle routing problem with time windows*, Eur. J. Oper. Res., 176 (2007), 1478–1507.
- [19] Euchì, J. and Saduk, A., *Hybrid Genetic-Sweep algorithm to solve the vehicle routing problem with drones*, Phys. Commun., 44 (2020), 101236.
- [20] Fatemi-Anaraki, S., Mokhtarzadeh, M., Rabbani, M. and Abdolhamidi, D., *A hybrid of K-means and genetic algorithm to solve a bi-objective green delivery and pick-up problem*, J. Ind. Prod. Eng., (2021), 1–12.

- [21] Fisher, M.L. and Jaikumar, R., *A generalized assignment heuristic for vehicle routing*, Networks, 11(2) (1981), 109–124.
- [22] Gillett, B. and Miller, L., *A heuristic algorithm for the vehicle-dispatch problem*, Oper. Res., 22(2) (1974), 340–349.
- [23] Golden, B., Bodin, L., Doyle, T. and Stewart, W., JR., *Approximate traveling salesman algorithms*, Oper. Res., 28(3), part 2 (1979), 694–711.
- [24] Goutham, M., Menon, M., Garrow, S. and Stockar, S., *A convex hull cheapest insertion heuristic for the non-euclidean TSP*. arXiv preprint arXiv:2302.06582, 2023.
- [25] Hammami, F., Rekik, M. and Coelho, L.C., *A hybrid adaptive large neighborhood search heuristic for the team orienteering problem*, Comput. Oper. Res., 123 (2020), 105034.
- [26] Hertrich, C., Hungerländer, P. and Truden, C., *Sweep algorithms for the capacitated vehicle routing problem with structured time windows*, Oper. Res. Proc. 2018, (2019), 127–133.
- [27] Imran, N. and Won, M., *VRPD-DT: Vehicle routing problem with drones under dynamically changing traffic conditions*, arXiv preprint arXiv:2404.09065 (2024).
- [28] Jones, J. and Adamatzky, A., *Computation of the traveling salesman problem by a sShrinking blob*, Nat. Comput., 13(1) (2014), 1–16.
- [29] Laporte, G., and Nobert, Y., *A branch and bound algorithm for the capacitated vehicle routing problem*, Oper.-Res.-Spektrum, 5(2) (1983), 77–85.
- [30] Larson, R. and Odoni, A., *Urban operations research*, Prentice Hall, Englewood Cliffs, N.J., 1981.
- [31] Liao, T.-Y., *On-line vehicle routing problems for carbon emissions reduction*, Comput.-Aided Civ. Infrastruct. Eng., 32(12) (2017), 1047–1063.

- [32] Liu, Q., Xu, P., Wu, Y., and Shen, T., *A two-stage algorithm for vehicle routing problem with charging relief in post-disaster*, IET. Intel. Transp. Sys., 45(3) (2023), 123–137.
- [33] Luo, J., and Chen, M.R., *Multi-phase modified shuffled frog leaping algorithm with extremal optimization for the MDVRP and the MDVRPTW*, Comput. Ind. Eng., 72 (2014), 84–97.
- [34] Mahmud, N., and Haque, M.M., *Solving multiple depot vehicle routing problem (MDVRP) using genetic algorithm*, 2019 Int. Conf. Electr. Comput. Commun. Eng., (2019), 1–6.
- [35] Marinakis Y., Marinaki M. and Migdalas A., *Variants and formulations of the vehicle routing problem*, Open Probl. Optim. Data Anal., Springer Optim. Its Appl., (2018), 91–127.
- [36] Meira, L., Martins, P., Menzori, M. and Zeni, G., *Multi-objective vehicle routing problem applied to large scale post office deliveries*, arXiv preprint arXiv:1801.00712, (2017).
- [37] Nadizadeh, A., Sahraeian, R., Sabzevarizadeh, A. and Homayouni, S.M., *Using greedy clustering method to solve capacitated location-routing problem*, Afr. J. Bus. Manag., 5(17) (2011), 7499–7506.
- [38] Nallusamy, R., Duraiswamy, K., Dhalanaksmi, R. and Parthiban, P., *Optimization of multiple vehicle routing problems using approximation algorithms*, Int. J. Eng. Sci. Technol., 1(3) (2009), 129–135.
- [39] Narasimha, K., *Ant colony optimization technique to solve min-max multi depot vehicle routing problem*, Master Thesis, Natl. Inst. Technol. Karnataka, India, 2011.
- [40] Negreiros, M. and Palhano, A., *The capacitated centred clustering problem*, Comput. Oper. Res., 33(6) (2006), 1639–1663.
- [41] Nuriyeva, F. and Kutucu, H., *A convex hull based algorithm for solving the traveling salesman problem*, TWMS J. App. Eng. Math. (2025).

- [42] Nurkahyo, G., Alias, R., Shamsuddin, S.M. and Noor, Md., *Sweep algorithm in vehicle routing problem for public transport*, J. Antarabangsa, 2 (2002), 51–64.
- [43] Poot, A., Kant, G. and Wagelmans, A., *A savings based method for real-life vehicle routing problems*, J. Oper. Res. Soc., 53 (2002), 57–68.
- [44] Pourmohammadreza, N. and Jokar, M.R.A., *A novel two-phase approach for optimization of the last-mile delivery problem with service options*, Sustainability, 15(10), (2023), 8098.
- [45] Ralphs, T.K., Kopman, L., Pulleyblank, W.R. and Trotter, L.E., JR., *On the capacitated vehicle routing problem*, Math. Program., 94 (2003), 343–359.
- [46] Renaud, J., Laporte, G. and Boctor, F., *A Tabu search heuristic for the multi-depot vehicle routing problem*, Comput. Oper. Res., 23(3) (1996), 229–235.
- [47] Rossit, D., Vigo, D., Tohme, F. and Frutos, M., *Improving visual attractiveness in capacitated vehicle routing problems: A heuristic algorithm*, XVIII Lat.-Iberoam. Conf. Oper. Res.-CLAIO, 749, 2016.
- [48] Sariklis, D. and Powell, S., *A heuristic method for the open vehicle routing problem*, J. Oper. Res. Soc., 51(5) (2000), 564–573.
- [49] Sehta, N. and Thakar, U., *Capacitated vehicle routing problem: A solution using convex hull based sweep algorithm and genetic algorithm*, In AIP Conference Proceedings (Vol. 2705, No. 1). AIP Publishing, 2023.
- [50] Simchi-Levi, D. and Bramel, J., *The logic of logistics: Theory, algorithms and applications for logistics management*, Springer-Verlag, New York, 2nd ed., 1997.
- [51] Singanamala, P., Reddy, K. and Venkataramaiah, P., *Solution to a multi depot vehicle routing problem using K-means algorithm, Clarke and Wright algorithm and ant colony optimization*, Int. J. Appl. Eng. Res., 13(21) (2018), 15236–15246.

- [52] Solomon, M., *Algorithms for the vehicle routing and scheduling problems with time window constraints*, Oper. Res., 35(2) (1987), 254–265.
- [53] Solomon, M.M. and Desrosiers, J., *Time window constrained routing and scheduling problems*, Transp. Sci., 22(1) (1988), 1–13.
- [54] Soto-Concha, R., Escobar, J.W., Morillo-Torres, D. and Linfati, R., *The vehicle-routing problem with satellites utilization: A systematic review of the literature.*, Mathematics, 13(7) (2025), 1–29.
- [55] Stewart, W.R., *Computational comparison of five heuristic algorithms for the Euclidean traveling salesman problem*, Lect. Notes Econ. Math. Syst., 199 (1982), Springer, Berlin, Heidelberg.
- [56] Thibbotuwawa, A., Bocewicz, G., Nielsen, P. and Banaszak, Z., *Unmanned aerial vehicle routing problems: A literature review*, Appl. Sci., 10(13) (2020), 4504.
- [57] Toth, P. and Vigo, D., *Branch-and-bound algorithms for the capacitated VRP*, SIAM, (2002), 29–51.
- [58] Toth, P. and Vigo, D., Eds., *Vehicle routing: Problems, methods, and applications*, SIAM, 2014.
- [59] Wang, Z. and Sheu, J.B., *Vehicle routing problem with drones*, Transp. Res. Part B Methodol., 122 (2019), 350–364.



## Accurate ENO-like schemes for the model of fluid flows in a nozzle with variable cross-section

D.H. Cuong and M.D. Thanh\*, 

### Abstract

A class of accurate high-order ENO-like schemes for the model of fluid flows in a nozzle with variable cross-section is presented. The model contains a nonconservative source term, which causes unsatisfactory results to standard numerical schemes, even for low-order ones. The proposed schemes rely on exact Riemann solvers and the reconstructed piecewise polynomials which are nonoscillatory. These schemes inherit the high-order precision of the ENO schemes like many existing ENO-type schemes, and possess a good accuracy. The ENO-like scheme corresponding to  $k = 3$  can get

---

\*Corresponding author

Received 22 April 2024; revised 1 April 2025; accepted 31 May 2025

Dao Huy Cuong

Department of Mathematics, Ho Chi Minh City University of Education, Ho Chi Minh City, Vietnam. Email: cuongdh@hcmue.edu.vn

Mai Duc Thanh

Corresponding author, Department of Mathematics, International University, Ho Chi Minh City, Vietnam. Email: mdthanh@hcmiu.edu.vn  
Vietnam National University, Ho Chi Minh City, Vietnam.

### How to cite this article

Cuong, D.H. and Thanh, M.D., Accurate ENO-like schemes for the model of fluid flows in a nozzle with variable cross-section. *Iran. J. Numer. Anal. Optim.*, 2025; 15(3): 1275-1309. <https://doi.org/10.22067/ijnao.2025.87744.1431>

the precision as good as van Leer-type schemes, and is numerically stable. Moreover, the ENO-like schemes for larger  $k$  may suffer from oscillations.

**AMS subject classifications (2020):** 65M08, 35L65, 76B15

**Keywords:** Nonconservative; Numerical approximation; Essentially nonoscillatory scheme; Accuracy; Resonance.

## 1 Introduction

Essentially nonoscillatory (ENO) schemes for hyperbolic conservation laws were first constructed by Harten et al. [24] in 1987. These high-resolution schemes have been developed by many authors for nonconservative systems, most use approximate Riemann solvers. In this paper, we aim to construct an ENO-like scheme relying on *exact Riemann solvers* for the following non-conservative system of balance laws, which models fluid flows in a nozzle with variable cross-section:

$$\begin{aligned}\partial_t(a\rho) + \partial_x(a\rho u) &= 0, \\ \partial_t(a\rho u) + \partial_x(a(\rho u^2 + p)) &= p\partial_x a, \quad x \in \mathbb{R}, \quad t > 0,\end{aligned}\tag{1}$$

where  $\rho = \rho(x, t)$ ,  $u = u(x, t)$ ,  $p = p(x, t)$  denote the density, particle velocity, and pressure of the fluid, respectively, and  $a = a(x)$  denotes the cross-section area of the nozzle.

The term  $p\partial_x a$  on the right-hand side of the system (1) makes it non-conservative. Recall that the formulation of weak solutions of nonconservative systems of balance laws was introduced in [18]. Often, nonconservative terms cause unsatisfactory results for standard schemes. Therefore, numerical approximations of solutions of nonconservative systems of balance laws are one of the most challenging problems. Recently, we built a Godunov-type scheme for the model (1) in [15]. Our aim in this paper is to construct a high-resolution ENO-like scheme for (1). We will also demonstrate by numerical tests that the ENO-like scheme corresponding to  $k = 3$  can provide us with a second order accurate approximation to a smooth stationary wave and a much better accuracy than the Godunov-type scheme. However, the

ENO-like schemes for larger  $k$  may not be convergent, since oscillations can be observed.

For simplicity, we assume throughout that the fluid is isentropic and ideal, so that an equation of state is given by

$$p = \kappa \rho^\gamma, \quad (2)$$

where  $\kappa > 0$  and  $1 < \gamma < 5/3$  are constants.

Observe that there are many works in the literature about the study of nonconservative hyperbolic systems of balance laws. Basic theory of Riemann problem for models of fluid flows in a nozzle with discontinuous cross-section was considered in [25, 26, 32, 29, 38, 21]. In recent years, a related traffic flow model was formulated in [20], and a related model of shallow water flows over movable bottom with suspended and bedload transport was proposed in [8]. We refer the reader to the book [31] for the Riemann problem for various models in continuum physics. Regarding numerical methods, many results and schemes for models of fluid flows in a nozzle with variable cross-section were studied in [28, 27, 14, 5]. Recently, a second-order scheme based on the first-order Price-T scheme and the MUSCL-Hancock strategy for arterial blood flow models with viscoelasticity was constructed in [10]. Numerical schemes for the model of shallow water equations were studied in [22, 11, 30, 41]. Well-balanced finite difference WENO schemes using approximate solvers for the Ripa model were constructed in [23]. A set of arbitrarily high-order ENO-type schemes is constructed in a recent work [42] using a typical five-point smoothness measurement as the shock-detector, which are able to detect discontinuities before spatial reconstructions. Furthermore, ENO schemes with adaptive order which select a polynomial from several candidates that are reconstructed on stencils of unequal sizes are designed in [36]. A review on ENO schemes was given in [37]. Well-balanced numerical schemes for a single conservation law with source term were presented in

[6, 7, 3]. Numerical schemes for two-phase flow models were built in [4, 13, 12, 19, 1, 9, 33, 40, 39, 17]. The reader is referred to

[2, 34, 35] for Godunov-type schemes for hyperbolic systems of balance laws in nonconservative forms. See also the references therein.

The organization of this paper is as follows. Section 2 provides us with the background of the model (1). Section 3 is devoted to constructing an ENO-like scheme to calculate the approximate solution of the Cauchy problem for (1). Numerical tests and discussions for all types of initial data are presented in Section 4. Finally, in section 5, we draw several conclusions.

## 2 Backgrounds

Let us set

$$c = \sqrt{dp/d\rho}.$$

Then the system (1) with supplementing a trivial equation

$$\partial_t a = 0, \quad (3)$$

can be rewritten in the nonconservative form as

$$\partial_t \mathbf{U} + \mathbf{A}(\mathbf{U})\partial_x \mathbf{U} = \mathbf{0}, \quad (4)$$

where

$$\mathbf{U} = \begin{bmatrix} \rho \\ u \\ a \end{bmatrix}, \quad \mathbf{A}(\mathbf{U}) = \begin{bmatrix} u & \rho & \rho u/a \\ c^2/\rho & u & 0 \\ 0 & 0 & 0 \end{bmatrix}.$$

The matrix  $\mathbf{A}(\mathbf{U})$  has three eigenvalues

$$\lambda_1 = u - c, \quad \lambda_2 = u + c, \quad \lambda_3 = 0. \quad (5)$$

The corresponding eigenvectors can be chosen as

$$\mathbf{r}_1 = \begin{bmatrix} \rho \\ -c \\ 0 \end{bmatrix}, \quad \mathbf{r}_2 = \begin{bmatrix} \rho \\ c \\ 0 \end{bmatrix}, \quad \mathbf{r}_3 = \begin{bmatrix} -\rho u^2 \\ u c^2 \\ a(u^2 - c^2) \end{bmatrix}.$$

The first and the third characteristic speeds coincide on the *upper sonic surface*

$$\mathcal{C}^+ = \{\mathbf{U} | \lambda_1(\mathbf{U}) = \lambda_3(\mathbf{U})\}. \quad (6)$$

The second and the third characteristic speeds coincide on the *lower sonic surface*

$$\mathcal{C}^- = \{\mathbf{U} | \lambda_2(\mathbf{U}) = \lambda_3(\mathbf{U})\}. \quad (7)$$

Therefore, the system (4) is strictly hyperbolic on following regions:

$$\begin{aligned} G_1 &= \{\mathbf{U} | \lambda_1(\mathbf{U}) > \lambda_3(\mathbf{U})\}, \\ G_2 &= \{\mathbf{U} | \lambda_1(\mathbf{U}) < \lambda_3(\mathbf{U}) < \lambda_2(\mathbf{U})\}, \\ G_3 &= \{\mathbf{U} | \lambda_2(\mathbf{U}) < \lambda_3(\mathbf{U})\}. \end{aligned}$$

The set  $G_1 \cup G_3$  is called the *supersonic region*, while  $G_2$  is called the *subsonic region*. The third characteristic field is *linearly degenerate*, since

$$\nabla \lambda_3 \cdot \mathbf{r}_3 = 0.$$

The first and the second characteristic fields are *genuinely nonlinear*, since

$$-\nabla \lambda_1 \cdot \mathbf{r}_1 = \nabla \lambda_2 \cdot \mathbf{r}_2 = \frac{(\gamma + 1)c}{2} > 0.$$

## 2.1 Shock wave curves

Recall that a *discontinuity wave* of (4) connecting a left-hand state  $\mathbf{U}_-$  to a right-hand state  $\mathbf{U}_+$  is a weak solution of the form

$$\mathbf{U}(x, t) = \begin{cases} \mathbf{U}_-, & \text{if } x < \sigma t, \\ \mathbf{U}_+, & \text{if } x > \sigma t, \end{cases} \quad (8)$$

where the speed of discontinuity wave  $\sigma$  must satisfy the Rankine–Hugoniot relations. The Rankine–Hugoniot relation associated with (3) takes the form

$$-\sigma[a] = 0, \quad (9)$$

where  $[a] = a_+ - a_-$  denotes the jump of the quantity  $a$ . As discussed in [29], across a discontinuity wave there are two possibilities:

(i) either  $[a] = 0$ ,

(ii) or  $\sigma = 0$ .

For the first case (i), a discontinuity wave (8) is called a *shock wave*. It called an *i*-Lax shock, if the shock speed  $\sigma = \sigma_i(\mathbf{U}_-, \mathbf{U}_+)$  satisfies the Lax shock inequalities,

$$\lambda_i(\mathbf{U}_+) < \sigma_i(\mathbf{U}_-, \mathbf{U}_+) < \lambda_i(\mathbf{U}_-), \quad i = 1, 2.$$

Given a state  $\mathbf{U}_0 = [\rho_0, u_0, a_0]^T$ , the set of all right-hand states  $\mathbf{U} = [\rho, u, a]^T$  that can be connected to  $\mathbf{U}_0$  by a 1-Lax shock is given by

$$\mathcal{S}_1(\mathbf{U}_0) = \left\{ \mathbf{U} \mid u = u_0 - \sqrt{-(p - p_0) \left( \frac{1}{\rho} - \frac{1}{\rho_0} \right)}, \rho > \rho_0, a = a_0 \right\}, \quad (10)$$

and, in the backward way, the set of all left-hand states  $\mathbf{U} = [\rho, u, a]^T$  that can be connected to  $\mathbf{U}_0$  by a 2-Lax shock is given by

$$\mathcal{S}_{2B}(\mathbf{U}_0) = \left\{ \mathbf{U} \mid u = u_0 + \sqrt{-(p - p_0) \left( \frac{1}{\rho} - \frac{1}{\rho_0} \right)}, \rho > \rho_0, a = a_0 \right\}. \quad (11)$$

We call these set *the forward curve of 1-shock waves* and *the backward curve of 1-shock waves*, respectively.

Furthermore, we have the following result about the sign of the 1-shock speeds along the wave curve  $\mathcal{S}_1(\mathbf{U}_0)$ , which is shown in [29].

**Lemma 1.** If  $\mathbf{U}_0 \in G_2 \cup G_3$ , then  $\sigma_1(\mathbf{U}_0, \mathbf{U})$  remains negative, that is,

$$\sigma_1(\mathbf{U}_0, \mathbf{U}) < 0, \quad \mathbf{U} \in \mathcal{S}_1(\mathbf{U}_0).$$

If  $\mathbf{U}_0 \in G_1$ , then there is exactly one state, denoted by  $\mathbf{U}_0^\# = [\rho_0^\#, u_0^\#, a_0]^\top \in \mathcal{S}_1(\mathbf{U}_0)$ , such that

$$\begin{aligned} \mathbf{U}_0^\# &\in G_2, \quad u_0^\# > 0, \\ \sigma_1(\mathbf{U}_0, \mathbf{U}_0^\#) &= 0, \\ \sigma_1(\mathbf{U}_0, \mathbf{U}) &> 0, \quad \rho_0 < \rho < \rho_0^\#, \\ \sigma_1(\mathbf{U}_0, \mathbf{U}) &< 0, \quad \rho > \rho_0^\#. \end{aligned}$$

## 2.2 Stationary waves

For the case (ii), a discontinuity wave (8) is called a *stationary wave*, and the two states  $\mathbf{U}_\pm$  are called the two *equilibrium states*. As shown in [29], two equilibrium states  $\mathbf{U}_\pm$  must satisfy the jump relations

$$\begin{aligned} [a\rho u] &= 0, \\ \left[ \frac{u^2}{2} + \frac{c^2}{\gamma - 1} \right] &= 0. \end{aligned} \quad (12)$$

Given the state  $\mathbf{U}_0 = [\rho_0, u_0, a_0]^T$  and the cross-section level  $a \neq a_0$ . As discussed in [29], a stationary wave connecting from  $\mathbf{U}_0$  to some state  $\mathbf{U} = [\rho, u, a]^T$  exists if and only if  $a \geq a_{\min}$ , where

$$a_{\min} = \frac{a_0 \rho_0 |u_0|}{\sqrt{\kappa \gamma} (\rho_{\max})^{\frac{\gamma+1}{2}}}, \quad \rho_{\max} = \left( \frac{\gamma - 1}{\kappa \gamma (\gamma + 1)} (u_0^2 + \mu \rho_0^{\gamma-1}) \right)^{\frac{1}{\gamma-1}}. \quad (13)$$

Moreover, if  $a > a_{\min}$ , then there are exactly two states  $\mathbf{U}_0^s, \mathbf{U}_0^b$  that can be connected to  $\mathbf{U}_0$  by a stationary wave,

$$\begin{aligned} \mathbf{U}_0^s &= \left[ \rho_0^s, \frac{a_0 \rho_0 u_0}{a \rho_0^s}, a \right]^T, \\ \mathbf{U}_0^b &= \left[ \rho_0^b, \frac{a_0 \rho_0 u_0}{a \rho_0^b}, a \right]^T, \end{aligned} \quad (14)$$

where  $\rho_0^s < \rho_{\max} < \rho_0^b$  are two roots of the nonlinear equation

$$-\frac{2\kappa\gamma}{\gamma-1} \mu \rho^{\gamma+1} + (u_0^2 + \frac{2\kappa\gamma}{\gamma-1} \rho_0^{\gamma-1}) \rho^2 - (a_0 u_0 \rho_0 / a)^2 = 0. \quad (15)$$

Precisely, we have the following lemma about stationary waves.

**Lemma 2.** [29, Lem. 2.3] The following conclusions hold:

a)

$$\begin{aligned} \rho_{\max} &> \rho_0, & \mathbf{U}_0 &\in G_1 \cup G_3, \\ \rho_{\max} &< \rho_0, & \mathbf{U}_0 &\in G_2, \\ \rho_{\max} &= \rho_0, & \mathbf{U}_0 &\in \mathcal{C}^\pm. \end{aligned}$$

b) The state  $\mathbf{U}_0^s$  belongs to  $G_1$  if  $u_0 > 0$ , and belongs to  $G_3$  if  $u_0 < 0$ , while the state  $\mathbf{U}_0^b$  always belongs to  $G_2$ . In addition, it holds that

(i) If  $a > a_0$ , then

$$\rho_0^s < \rho_0 < \rho_0^b.$$

(ii) If  $a < a_0$ , then

$$\begin{aligned} \rho_0 &< \rho_0^s < \rho_0^b & \text{for } \mathbf{U}_0 &\in G_1 \cup G_3, \\ \rho_0^s &< \rho_0^b < \rho_0 & \text{for } \mathbf{U}_0 &\in G_2. \end{aligned}$$

It follows from Lemma 2 that there are two possible stationary waves from a given state  $\mathbf{U}_0$  to a state with a new level cross-section  $a$ . Thus, it is necessary to impose some condition to select a unique physical stationary state as follows.

(MC) Any stationary jump must not cross the sonic curve in the  $(\rho, u)$ -plane.

### 2.3 Rarefaction wave curves

Recall that the  $i$ -rarefaction wave ( $i = 1, 2$ ) of (4) connecting a left-hand state  $\mathbf{U}_-$  to a right-hand state  $\mathbf{U}_+$  is a weak solution of the form

$$\mathbf{U}(x, t) = \begin{cases} \mathbf{U}_-, & \text{if } x < \lambda_i(\mathbf{U}_-)t, \\ \mathbf{V}_i(x/t), & \text{if } \lambda_i(\mathbf{U}_-)t \leq x \leq \lambda_i(\mathbf{U}_+)t, \\ \mathbf{U}_+, & \text{if } x > \lambda_i(\mathbf{U}_+)t, \end{cases}$$

where  $\mathbf{V}_i(\cdot)$  is the solution of following problem:

$$\begin{aligned} \frac{d\mathbf{V}_i(\xi)}{d\xi} &= \frac{1}{\nabla \lambda_i(\mathbf{V}(\xi)) \cdot \mathbf{r}_i(\mathbf{V}(\xi))} \mathbf{r}_i(\mathbf{V}(\xi)), \quad \lambda_i(\mathbf{U}_-) < \xi < \lambda_i(\mathbf{U}_+), \\ \mathbf{V}_i(\lambda_i(\mathbf{U}_-)) &= \mathbf{U}_-, \quad \mathbf{V}_i(\lambda_i(\mathbf{U}_+)) = \mathbf{U}_+. \end{aligned}$$

Given a state  $\mathbf{U}_0 = [\rho_0, u_0, a_0]^T$ , the set of all right-hand states  $\mathbf{U} = [\rho, u, a]^T$  that can be connected to  $\mathbf{U}_0$  by a 1-rarefaction wave forms the *forward curve of 1-rarefaction waves*, denoted by  $\mathcal{R}_1(\mathbf{U}_0)$ . In a backward way, the set of all left-hand states  $\mathbf{U} = [\rho, u, a]^T$  that can be connected to  $\mathbf{U}_0$  by a 2-rarefaction wave forms the *backward curve of 2-rarefaction wave*, denoted by  $\mathcal{R}_{2B}(\mathbf{U}_0)$ . These curves are given by

$$\begin{aligned} \mathcal{R}_1(\mathbf{U}_0) &= \left\{ \mathbf{U} \mid u = u_0 - \frac{2\sqrt{\kappa\gamma}}{\gamma-1} (\rho^{(\gamma-1)/2} - \rho_0^{(\gamma-1)/2}), \rho \leq \rho_0, a = a_0 \right\}, \\ \mathcal{R}_{2B}(\mathbf{U}_0) &= \left\{ \mathbf{U} \mid u = u_0 + \frac{2\sqrt{\kappa\gamma}}{\gamma-1} (\rho^{(\gamma-1)/2} - \rho_0^{(\gamma-1)/2}), \rho \leq \rho_0, a = a_0 \right\}. \end{aligned} \tag{16}$$

From (10), (11), and (16), we have the forward and backward wave curves in the nonlinear characteristic fields as follows:

$$\begin{aligned}\mathcal{W}_1(\mathbf{U}_0) &= \mathcal{R}_1(\mathbf{U}_0) \cup \mathcal{S}_1(\mathbf{U}_0), \\ \mathcal{W}_{2B}(\mathbf{U}_0) &= \mathcal{R}_{2B}(\mathbf{U}_0) \cup \mathcal{S}_{2B}(\mathbf{U}_0).\end{aligned}\tag{17}$$

## 2.4 Computing the Riemann solution

Observe that by the transformation  $x \mapsto -x$ ,  $u \mapsto -u$ , a left-hand (right-hand) state  $\mathbf{U} = [\rho, u, a]^T$  in  $G_2$  (in  $G_3$ ) will be transformed to the right-hand (left-hand, respectively) state  $\mathbf{V} = [\rho, -u, a]^T$  in  $G_2$  (in  $G_1$ , respectively). Therefore, the construction of the Riemann solutions for Riemann data around  $\mathcal{C}^-$  can be obtained from the one for Riemann data around  $\mathcal{C}^+$ . Thus, without loss of generality, we consider only the case where Riemann data are in  $G_1 \cup G_2$ . We call construction A if  $\mathbf{U}_L$  belongs to  $G_1$ , and construction B if  $\mathbf{U}_L$  belongs to  $G_2$ . In each construction, we divide it into 3 cases depending on the relative position of the backward curve  $\mathcal{W}_{2B}(\mathbf{U}_R)$  compared to the composite wave curves established in each construction.

In this subsection, we use some following notations:

- (i)  $W_k(\mathbf{U}_-, \mathbf{U}_+)$  ( $S_k(\mathbf{U}_-, \mathbf{U}_+)$ ,  $R_k(\mathbf{U}_-, \mathbf{U}_+)$ ) denotes the  $k$ -wave ( $k$ -shock,  $k$ -rarefaction wave, respectively) connecting the left-hand state  $\mathbf{U}_-$  to the right-hand state  $\mathbf{U}_+$ , for  $k = 1, 2, 3$ .
- (ii)  $\mathbf{U}_0^\#$  denotes the state on the forward curve of 1-shock waves  $\mathcal{S}_1(\mathbf{U}_0)$  such that  $\sigma_1(U_0, U_0^\#) = 0$ ; see Lemma 1.
- (iii)  $\mathbf{U}_0^s, \mathbf{U}_0^b$  denote the states resulted by stationary contact wave from  $\mathbf{U}_0$ ; see (14) and Lemma 2.
- (iv)  $\mathbf{U}_0^\pm = \mathcal{W}_1(\mathbf{U}_0) \cap \mathcal{C}^\pm$ , where  $\mathcal{W}_1(\mathbf{U}_0)$  is defined by (17) and  $\mathcal{C}^\pm$  are defined by (6), (7).
- (v)  $\mathbf{U}^{\text{Rie}}(x/t; \mathbf{U}_L, \mathbf{U}_R)$  is the exact solution of (4) with the Riemann initial data

$$\mathbf{U}(x, 0) = \begin{cases} \mathbf{U}_L, & \text{if } x < 0, \\ \mathbf{U}_R, & \text{if } x > 0. \end{cases}\tag{18}$$

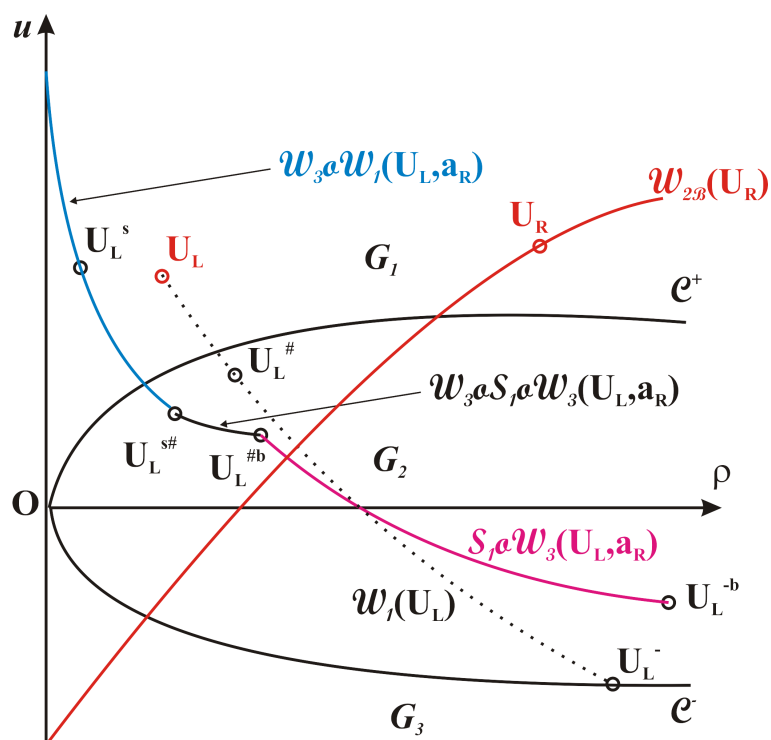


Figure 1: The composite wave curves  $\mathcal{W}_3 \circ \mathcal{W}_1(\mathbf{U}_L, a_R)$ ,  $\mathcal{W}_3 \circ \mathcal{S}_1 \circ \mathcal{W}_3(\mathbf{U}_L, a_R)$ ,  $\mathcal{S}_1 \circ \mathcal{W}_3(\mathbf{U}_L, a_R)$ , and the backward curve of 2-wave  $\mathcal{W}_{2B}(\mathbf{U}_R)$

### 2.4.1 Construction A1

Given a left-hand state  $\mathbf{U}_L \in G_1$ . If  $\mathbf{U}_R$  is a state such that the backward curve  $\mathcal{W}_{2B}(\mathbf{U}_R)$  intersects the composite wave curve  $\mathcal{W}_3 \circ \mathcal{W}_1(\mathbf{U}_L, a_R)$ , see Figure 1, then the Riemann solution of (4) with initial data (18) is

$$\mathbf{U}^{\text{Rie}}(x/t; \mathbf{U}_L, \mathbf{U}_R) = \begin{cases} \mathbf{U}_L, & \text{if } x/t < 0, \\ \mathbf{U}_L^s, & \text{if } 0 < x/t < \min\{\lambda_1(\mathbf{U}_L^s), \lambda_1(\mathbf{U}_*)\}, \\ W_1(\mathbf{U}_L^s, \mathbf{U}_*), & \text{if } \min\{\lambda_1(\mathbf{U}_L^s), \lambda_1(\mathbf{U}_*)\} < x/t < \min\{\lambda_2(\mathbf{U}_*), \lambda_2(\mathbf{U}_R)\}, \\ W_2(\mathbf{U}_*, \mathbf{U}_R), & \text{if } x/t > \min\{\lambda_2(\mathbf{U}_*), \lambda_2(\mathbf{U}_R)\}, \end{cases} \quad (19)$$

where  $\mathbf{U}_*$  is found by

$$\mathcal{W}_{2B}(\mathbf{U}_R) \cap \mathcal{W}_3 \circ \mathcal{W}_1(\mathbf{U}_L, a_R) = \{\mathbf{U}_*\},$$

with  $\mathcal{W}_{2B}(\mathbf{U}_R)$  is defined by (17), and  $\mathcal{W}_3 \circ \mathcal{W}_1(\mathbf{U}_L, a_R)$  is defined by

$$\mathcal{W}_3 \circ \mathcal{W}_1(\mathbf{U}_L, a_R) = \left\{ \mathbf{U} \mid \mathbf{U} \in \mathcal{W}_1(\mathbf{U}_L^s) \text{ and } \mathbf{U} \text{ is located above } \mathbf{U}_L^{s\#} \right\}. \quad (20)$$

The set  $\mathcal{W}_3 \circ \mathcal{W}_1(\mathbf{U}_L, a_R)$  is called *the curve of composite 3-wave and 1-wave*, where  $\mathbf{U}_L^s = \left[ \rho_L^s, \frac{a_L \rho_L u_L}{a_R \rho_L^s}, a_R \right]^T$  is defined by Lemma 2,  $\mathcal{W}_1(\mathbf{U}_L^s)$  is defined by (17), and  $\mathbf{U}_L^{s\#} = (\mathbf{U}_L^s)^\#$  is defined by Lemma 1.

### 2.4.2 Construction A2

Given  $\mathbf{U}_L \in G_1$ . Whenever the backward curve  $\mathcal{W}_{2B}(\mathbf{U}_R)$  intersects the composite wave curve  $\mathcal{W}_3 \circ \mathcal{S}_1 \circ \mathcal{W}_3(\mathbf{U}_L, a_R)$ , see Figure 1, the Riemann solution of (4) with initial data (18) is

$$\mathbf{U}^{\text{Rie}}(x/t; \mathbf{U}_L, \mathbf{U}_R) = \begin{cases} \mathbf{U}_L, & \text{if } x/t < 0, \\ \mathbf{U}_*, & \text{if } 0 < x/t < \min\{\lambda_2(\mathbf{U}_*), \lambda_2(\mathbf{U}_R)\}, \\ W_2(\mathbf{U}_*, \mathbf{U}_R), & \text{if } x/t > \min\{\lambda_2(\mathbf{U}_*), \lambda_2(\mathbf{U}_R)\}, \end{cases} \quad (21)$$

where  $\mathbf{U}_*$  is computed by

$$\mathcal{W}_{2B}(\mathbf{U}_R) \cap \mathcal{W}_3 \circ \mathcal{S}_1 \circ \mathcal{W}_3(\mathbf{U}_L, a_R) = \{\mathbf{U}_*\},$$

with  $\mathcal{W}_{2B}(\mathbf{U}_R)$  is defined by (17), and  $\mathcal{W}_3 \circ \mathcal{S}_1 \circ \mathcal{W}_3(\mathbf{U}_L, a_R)$  is defined by

$$\begin{aligned} \mathcal{W}_3 \circ \mathcal{S}_1 \circ \mathcal{W}_3(\mathbf{U}_L, a_R) = & \left\{ \mathbf{U}_L^{s\#b} \mid a_M \text{ is between } a_L \text{ and } a_R, \right. \\ & \mathbf{U}_L^s = \left[ \rho_L^s, \frac{a_L \rho_L u_L}{a_M \rho_L^s}, a_M \right]^T, \quad \mathbf{U}_L^{s\#} = \left( \mathbf{U}_L^s \right)^\#, \\ & \left. \mathbf{U}_L^{s\#b} = \left[ (\rho_L^{s\#})^b, \frac{a_M \rho_L^{s\#} u_L^{s\#}}{a_R (\rho_L^{s\#})^b}, a_R \right]^T \right\}. \end{aligned} \quad (22)$$

We call the set  $\mathcal{W}_3 \circ \mathcal{S}_1 \circ \mathcal{W}_3(\mathbf{U}_L, a_R)$  the curve of composite 3-wave, 1-shock, and 3-wave.

### 2.4.3 Construction A3

Given  $\mathbf{U}_L \in G_1$ . If  $\mathbf{U}_R$  is a state such that the backward curve  $\mathcal{W}_{2B}(\mathbf{U}_R)$  intersects the composite wave curve  $\mathcal{S}_1 \circ \mathcal{W}_3(\mathbf{U}_L, a_R)$  defined as below, see Figure 1, then the Riemann solution of (4) with initial data (18) is

$$\mathbf{U}^{\text{Rie}}(x/t; \mathbf{U}_L, \mathbf{U}_R) = \begin{cases} \mathbf{U}_L, & \text{if } x/t < \sigma_1(\mathbf{U}_L, \mathbf{U}_*), \\ \mathbf{U}_*, & \text{if } \sigma_1(\mathbf{U}_L, \mathbf{U}_*) < x/t < 0, \\ \mathbf{U}_*^b, & \text{if } 0 < x/t < \min\{\lambda_2(\mathbf{U}_*^b), \lambda_2(\mathbf{U}_R)\}, \\ \mathcal{W}_2(\mathbf{U}_*^b, \mathbf{U}_R), & \text{if } x/t > \min\{\lambda_2(\mathbf{U}_*^b), \lambda_2(\mathbf{U}_R)\}, \end{cases} \quad (23)$$

where  $\mathbf{U}_*$  and  $\mathbf{U}_*^b$  are found by

$$\mathcal{W}_{2B}(\mathbf{U}_R) \cap \mathcal{S}_1 \circ \mathcal{W}_3(\mathbf{U}_L, a_R) = \{\mathbf{U}_*^b\},$$

with  $\mathcal{W}_{2B}(\mathbf{U}_R)$  is defined by (17), and  $\mathcal{S}_1 \circ \mathcal{W}_3(\mathbf{U}_L, a_R)$  is defined by

$$\begin{aligned} \mathcal{S}_1 \circ \mathcal{W}_3(\mathbf{U}_L, a_R) = & \left\{ \mathbf{U}_*^b \mid \mathbf{U}_* \in \mathcal{S}_1(\mathbf{U}_L), \mathbf{U}_* \text{ is located between } \mathbf{U}_L^\# \text{ and } \mathbf{U}_L^-, \right. \\ & \left. \mathbf{U}_*^b = \left[ \rho_*^b, \frac{a_L \rho_* u_*}{a_R \rho_*^b}, a_R \right]^T \right\}. \end{aligned} \quad (24)$$

The set  $\mathcal{S}_1 \circ \mathcal{W}_3(\mathbf{U}_L, a_R)$  is called *the curve of composite 1-shock and 3-wave*, where  $\mathcal{S}_1(\mathbf{U}_L)$  is defined by (10),  $\mathbf{U}_L^\#$  is defined by Lemma 1, and  $\mathbf{U}_L^- = \mathcal{S}_1(\mathbf{U}_L) \cap \mathcal{C}^-$ .

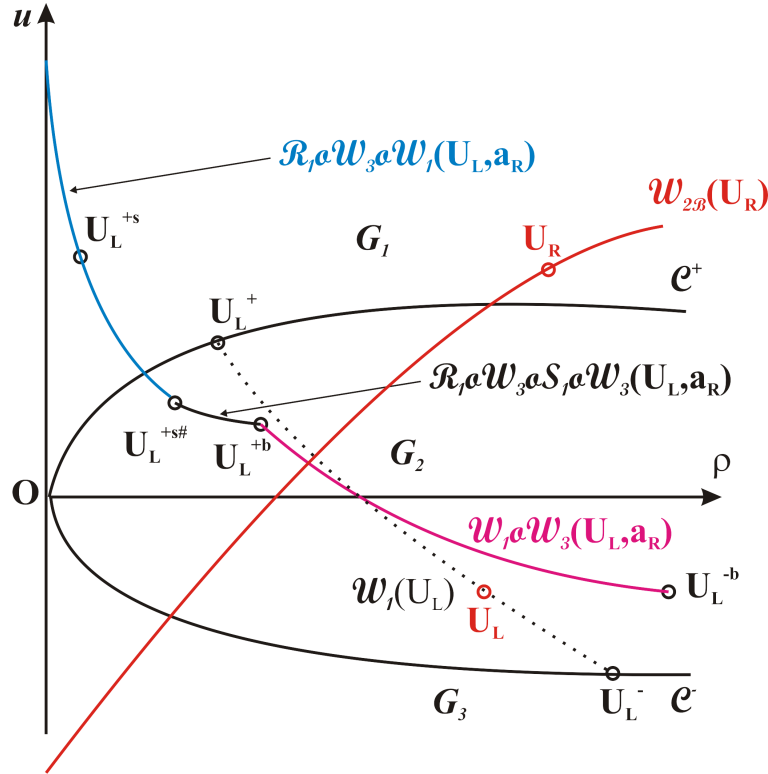


Figure 2: The composite wave curves  $\mathcal{R}_1 \circ \mathcal{W}_3 \circ \mathcal{W}_1(\mathbf{U}_L, a_R)$ ,  $\mathcal{R}_1 \circ \mathcal{W}_3 \circ \mathcal{S}_1 \circ \mathcal{W}_3(\mathbf{U}_L, a_R)$ ,  $\mathcal{W}_1 \circ \mathcal{W}_3(\mathbf{U}_L, a_R)$ , and the backward curve of 2-wave  $\mathcal{W}_{2B}(\mathbf{U}_R)$

#### 2.4.4 Construction B1

Given a left-hand state  $\mathbf{U}_L \in G_2$ . If a right-hand state  $\mathbf{U}_R$  satisfies that the backward curve  $\mathcal{W}_{2B}(\mathbf{U}_R)$  intersects the composite wave curve  $\mathcal{R}_1 \circ \mathcal{W}_3 \circ \mathcal{W}_1(\mathbf{U}_L, a_R)$  defined as below, see Figure 2, then the Riemann solution of (4) with initial data (18) is

$$\mathbf{U}^{\text{Rie}}(x/t; \mathbf{U}_L, \mathbf{U}_R) \quad (25)$$

$$= \begin{cases} R_1(\mathbf{U}_L, \mathbf{U}_L^+), & \text{if } x/t < 0, \\ \mathbf{U}_L^{+s}, & \text{if } 0 < x/t < \min\{\lambda_1(\mathbf{U}_L^{+s}), \lambda_1(\mathbf{U}_*)\}, \\ W_1(\mathbf{U}_L^{+s}, \mathbf{U}_*), & \text{if } \min\{\lambda_1(\mathbf{U}_L^{+s}), \lambda_1(\mathbf{U}_*)\} < x/t < \min\{\lambda_2(\mathbf{U}_*), \lambda_2(\mathbf{U}_R)\}, \\ W_2(\mathbf{U}_*, \mathbf{U}_R), & \text{if } x/t > \min\{\lambda_2(\mathbf{U}_*), \lambda_2(\mathbf{U}_R)\}, \end{cases}$$

where  $\mathbf{U}_*$  is found by

$$\mathcal{W}_{2B}(\mathbf{U}_R) \cap \mathcal{R}_1 \circ \mathcal{W}_3 \circ \mathcal{W}_1(\mathbf{U}_L, a_R) = \{\mathbf{U}_*\},$$

with  $\mathcal{W}_{2B}(\mathbf{U}_R)$  is defined by (17), and  $\mathcal{R}_1 \circ \mathcal{W}_3 \circ \mathcal{W}_1(\mathbf{U}_L, a_R)$  is defined by

$$\mathcal{R}_1 \circ \mathcal{W}_3 \circ \mathcal{W}_1(\mathbf{U}_L, a_R) = \left\{ \mathbf{U} \mid \mathbf{U} \in \mathcal{W}_1(\mathbf{U}_L^{+s}) \text{ and } \mathbf{U} \text{ is located above } \mathbf{U}_L^{+s\#} \right\}. \quad (26)$$

The set  $\mathcal{R}_1 \circ \mathcal{W}_3 \circ \mathcal{W}_1(\mathbf{U}_L, a_R)$  is called *the curve of composite 1-rarefaction wave, 3-wave, and 1-wave*, where  $\mathbf{U}_L^+ = \mathcal{R}_1(\mathbf{U}_L) \cap \mathcal{C}^+$ ,  $\mathbf{U}_L^{+s} = \left[ (\rho_L^+)^s, \frac{a_L \rho_L^+ u_L^+}{a_R (\rho_L^+)^s}, a_R \right]^T$  is defined by Lemma 2,  $\mathcal{R}_1(\mathbf{U}_L)$  and  $\mathcal{W}_1(\mathbf{U}_L^{+s})$  are defined by (16), (17), and  $\mathbf{U}_L^{+s\#} = \left( \mathbf{U}_L^{+s} \right)^\#$  is defined by Lemma 1.

#### 2.4.5 Construction B2

Given a left-hand state  $\mathbf{U}_L \in G_2$ . Whenever the backward curve  $\mathcal{W}_{2B}(\mathbf{U}_R)$  intersects the composite wave curve  $\mathcal{R}_1 \circ \mathcal{W}_3 \circ \mathcal{S}_1 \circ \mathcal{W}_3(\mathbf{U}_L, a_R)$ , see Figure 2, the Riemann solution of (4) with initial data (18) will be

$$\mathbf{U}^{\text{Rie}}(x/t; \mathbf{U}_L, \mathbf{U}_R) = \begin{cases} R_1(\mathbf{U}_L, \mathbf{U}_L^+), & \text{if } x/t < 0, \\ \mathbf{U}_*, & \text{if } 0 < x/t < \min\{\lambda_2(\mathbf{U}_*), \lambda_2(\mathbf{U}_R)\}, \\ W_2(\mathbf{U}_*, \mathbf{U}_R), & \text{if } x/t > \min\{\lambda_2(\mathbf{U}_*), \lambda_2(\mathbf{U}_R)\}, \end{cases} \quad (27)$$

where  $\mathbf{U}_*$  is calculated by

$$\mathcal{W}_{2B}(\mathbf{U}_R) \cap \mathcal{R}_1 \circ \mathcal{W}_3 \circ \mathcal{S}_1 \circ \mathcal{W}_3(\mathbf{U}_L, a_R) = \{\mathbf{U}_*\},$$

with  $\mathcal{W}_{2B}(\mathbf{U}_R)$  is defined by (17), and  $\mathcal{R}_1 \circ \mathcal{W}_3 \circ \mathcal{S}_1 \circ \mathcal{W}_3(\mathbf{U}_L, a_R)$  is defined by

$$\begin{aligned}
\mathcal{R}_1 \circ \mathcal{W}_3 \circ \mathcal{S}_1 \circ \mathcal{W}_3(\mathbf{U}_L, a_R) &= \left\{ \mathbf{U}_L^{+s\#b} \mid a_M \text{ is between } a_L \text{ and } a_R, \right. \\
\mathbf{U}_L^{+s} &= \left[ (\rho_L^+)^s, \frac{a_L \rho_L^+ u_L^+}{a_M (\rho_L^+)^s}, a_M \right]^T, \\
\mathbf{U}_L^{+s\#} &= \left( \mathbf{U}_L^{+s} \right)^\#, \\
\mathbf{U}_L^{+s\#b} &= \left[ (\rho_L^{+s\#})^b, \frac{a_M \rho_L^{+s\#} u_L^{+s\#}}{a_R (\rho_L^{+s\#})^b}, a_R \right]^T \}.
\end{aligned} \tag{28}$$

We refer the set  $\mathcal{R}_1 \circ \mathcal{W}_3 \circ \mathcal{S}_1 \circ \mathcal{W}_3(\mathbf{U}_L, a_R)$  as *the curve of composite 1-rarefaction wave, 3-wave, 1-shock, and 3-wave*, where  $\mathbf{U}_L^+ = \mathcal{R}_1(\mathbf{U}_L) \cap \mathcal{C}^+$ .

### 2.4.6 Construction B3

Given a left-hand state  $\mathbf{U}_L \in G_2$ . If  $\mathbf{U}_R$  is a right-hand state such that the backward curve  $\mathcal{W}_{2B}(\mathbf{U}_R)$  intersects the composite wave curve  $\mathcal{W}_1 \circ \mathcal{W}_3(\mathbf{U}_L, a_R)$ , see Figure 2, then the Riemann solution of (4) with initial data (18) is

$$\mathbf{U}^{\text{Rie}}(x/t; \mathbf{U}_L, \mathbf{U}_R) = \begin{cases} W_1(\mathbf{U}_L, \mathbf{U}_*), & \text{if } x/t < 0, \\ \mathbf{U}_*^b, & \text{if } 0 < x/t < \min\{\lambda_2(\mathbf{U}_*^b), \lambda_2(\mathbf{U}_R)\}, \\ W_2(\mathbf{U}_*^b, \mathbf{U}_R), & \text{if } x/t > \min\{\lambda_2(\mathbf{U}_*^b), \lambda_2(\mathbf{U}_R)\}, \end{cases} \tag{29}$$

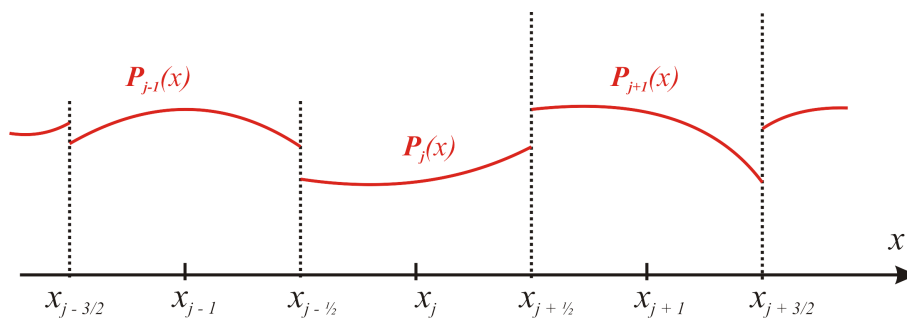
where  $\mathbf{U}_*$  and  $\mathbf{U}_*^b$  are found by

$$\mathcal{W}_{2B}(\mathbf{U}_R) \cap \mathcal{W}_1 \circ \mathcal{W}_3(\mathbf{U}_L, a_R) = \{\mathbf{U}_*^b\},$$

with  $\mathcal{W}_{2B}(\mathbf{U}_R)$  is defined by (17), and  $\mathcal{W}_1 \circ \mathcal{W}_3(\mathbf{U}_L, a_R)$  is defined by

$$\begin{aligned}
\mathcal{W}_1 \circ \mathcal{W}_3(\mathbf{U}_L, a_R) &= \left\{ \mathbf{U}_*^b \mid \mathbf{U}_* \in \mathcal{W}_1(\mathbf{U}_L), \mathbf{U}_* \text{ is located between } \mathbf{U}_L^+ \text{ and } \mathbf{U}_L^-, \right. \\
\mathbf{U}_*^b &= \left[ \rho_*^b, \frac{a_L \rho_*^b u_*^b}{a_R \rho_*^b}, a_R \right]^T \}.
\end{aligned} \tag{30}$$

The set  $\mathcal{W}_1 \circ \mathcal{W}_3(\mathbf{U}_L, a_R)$  is called *the curve of composite 1-wave and 3-wave*, where  $\mathcal{W}_1(\mathbf{U}_L)$  is defined by (17), and  $\mathbf{U}_L^\pm = \mathcal{W}_1(\mathbf{U}_L) \cap \mathcal{C}^\pm$ .

Figure 3: The piecewise polynomial  $\mathbf{U}_{p,pol}(x)$ 

### 3 Building an ENO-type scheme

Relying on the constructions of Riemann solutions in the previous section, we are now in a position to construct an ENO-type scheme for (4). Let us set

$$\mathbf{U} = \begin{bmatrix} a\rho \\ a\rho u \\ a \end{bmatrix}, \quad \mathbf{F}(\mathbf{U}) = \begin{bmatrix} a\rho u \\ a(\rho u^2 + p) \\ 0 \end{bmatrix}, \quad \mathbf{H}(\mathbf{U}) = \begin{bmatrix} 0 \\ p \\ 0 \end{bmatrix}. \quad (31)$$

Then, the system (4) can be written in form

$$\partial_t \mathbf{U} + \partial_x \mathbf{F}(\mathbf{U}) = \mathbf{H}(\mathbf{U}) \partial_x a, \quad x \in \mathbb{R}, \quad t > 0. \quad (32)$$

Given the initial condition

$$\mathbf{U}(x, 0) = \mathbf{U}_0(x), \quad x \in \mathbb{R}, \quad (33)$$

we define the discrete initial values  $\{\mathbf{U}_j^0\}_{j \in \mathbb{Z}}$  are given by

$$\mathbf{U}_j^0 = \frac{1}{\Delta x} \int_{x_{j-1/2}}^{x_{j+1/2}} \mathbf{U}_0(x) dx, \quad j \in \mathbb{Z}. \quad (34)$$

Suppose that the approximation  $\{\mathbf{U}_j^n\}_{j \in \mathbb{Z}}$  of  $\mathbf{U}$  at the time  $t_n$  is known. Recently, the Godunov-type scheme is built in [15] as

$$\mathbf{U}_j^{n+1} = \mathbf{U}_j^n - \frac{\Delta t}{\Delta x} \left( \mathbf{F}(\mathbf{U}^{\text{Rie}}(0-; \mathbf{U}_j^n, \mathbf{U}_{j+1}^n)) - \mathbf{F}(\mathbf{U}^{\text{Rie}}(0+; \mathbf{U}_{j-1}^n, \mathbf{U}_j^n)) \right), \quad (35)$$

where  $\Delta t$  must satisfy the C.F.L condition

$$\frac{\Delta t}{\Delta x} \max\{|\lambda_i(\mathbf{U}_j^n)| : j \in \mathbb{Z}, i = 1, 2\} \leq \text{CFL}. \quad (36)$$

Then, the van Leer-type scheme is built in [16] as

$$\mathbf{U}_j^{n+1} = \mathbf{U}_j^n - \frac{\Delta t}{\Delta x} \left( \mathbf{F}(\mathbf{U}^{\text{Rie}}(0-; \mathbf{U}_{j+1/2,-}^{n+1/2}, \mathbf{U}_{j+1/2,+}^{n+1/2})) \right. \quad (37)$$

$$\left. - \mathbf{F}(\mathbf{U}^{\text{Rie}}(0+; \mathbf{U}_{j-1/2,-}^{n+1/2}, \mathbf{U}_{j-1/2,+}^{n+1/2})) \right), \quad (38)$$

where

$$\mathbf{U}_{j+1/2,-}^{n+1/2} = \mathbf{U}_{j+1/2,-}^n - \frac{\Delta t}{2\Delta x} (\mathbf{F}(\mathbf{U}_{j+1/2,-}^n) - \mathbf{F}(\mathbf{U}_{j-1/2,+}^n)),$$

$$\mathbf{U}_{j-1/2,+}^{n+1/2} = \mathbf{U}_{j-1/2,+}^n - \frac{\Delta t}{2\Delta x} (\mathbf{F}(\mathbf{U}_{j+1/2,-}^n) - \mathbf{F}(\mathbf{U}_{j-1/2,+}^n)),$$

$$\mathbf{U}_{j+1/2,-}^n = \mathbf{U}_j^n + \frac{1}{2} \mathbf{S}_j^n,$$

$$\mathbf{U}_{j-1/2,+}^n = \mathbf{U}_j^n - \frac{1}{2} \mathbf{S}_j^n,$$

$$\mathbf{S}_j^n = (\mathbf{U}_{j+1}^n - \mathbf{U}_j^n) \Phi(\theta_j^n),$$

$$\theta_j^n = \frac{\mathbf{U}_j^n - \mathbf{U}_{j-1}^n}{\mathbf{U}_{j+1}^n - \mathbf{U}_j^n},$$

$$\Phi(\theta) = \frac{|\theta| + \theta}{1 + |\theta|}.$$

Now, in this paper, we construct an ENO-type scheme as follows:

- (1) From the sequence  $\mathbf{U}^n$ , we construct a piecewise polynomial  $\mathbf{U}_{\text{p.pol}}(\cdot)$  as follows:

$$\mathbf{U}_{\text{p.pol}}(x) = \mathbf{P}_j(x) = \begin{bmatrix} \rho_j(x) \\ u_j(x) \\ a_j(x) \end{bmatrix}, \quad x_{j-1/2} < x < x_{j+1/2}, \quad j \in \mathbb{Z}, \quad (39)$$

where for each  $j$ ,  $\mathbf{P}_j(x)$  is a polynomial of degree at most  $k-1$ , and there exist  $r, s \in \mathbb{N}$  (depending on  $j$ ) such that

$$\begin{aligned} \frac{1}{\Delta x} \int_{x_{i-1/2}}^{x_{i+1/2}} \mathbf{P}_j(x) dx &= \mathbf{U}_i^n, \quad i \in \{j-r, \dots, j, \dots, j+s\}, \\ s+r+1 &= k; \end{aligned} \quad (40)$$

see Figure 3. For each  $j$ , to achieve the polynomial  $\mathbf{P}_j(x)$  satisfying (40), we first look for the primitive function  $\mathbf{Q}_j(x)$  of  $\mathbf{P}_j(x)$ , that is,

$$\mathbf{Q}_j(x) = \int_{-\infty}^x \mathbf{P}_j(x) dx,$$

in the following way:

- (i) We start with the two node stencil for  $\mathbf{Q}_j(x)$

$$\{x_{j-1/2}, x_{j+1/2}\},$$

and compute the first order divided difference

$$\mathbf{Q}_j[x_{j-1/2}, x_{j+1/2}] = \frac{\mathbf{Q}_j(x_{j+1/2}) - \mathbf{Q}_j(x_{j-1/2})}{\Delta x},$$

where

$$\begin{aligned} \mathbf{Q}_j(x_{j+1/2}) &= \sum_{l=-\infty}^j \mathbf{U}_l^n \Delta x, \\ \mathbf{Q}_j(x_{j-1/2}) &= \sum_{l=-\infty}^{j-1} \mathbf{U}_l^n \Delta x. \end{aligned}$$

- (ii) Assume that  $l$ -node stencil for  $\mathbf{Q}_j(x)$  ( $l = 2, 3, \dots, k$ )

$$\{x_{i+1/2}, \dots, x_{i+l-1/2}\}$$

is known. To add one of two neighboring nodes,  $x_{i-1/2}$  or  $x_{i+l+1/2}$ , to the stencil, we use the following ENO procedure:

\* If

$$\begin{aligned} &\left| \mathbf{Q}_j[x_{i-1/2}, x_{i+1/2}, \dots, x_{i+l-1/2}] \right| \\ &< \left| \mathbf{Q}_j[x_{i+1/2}, \dots, x_{i+l-1/2}, x_{i+l+1/2}] \right|, \end{aligned}$$

then we add  $x_{i-1/2}$  to the stencil, where the  $l$ th order divided differences are defined recursively by

$$\begin{aligned} &\mathbf{Q}_j[x_{i+1/2}, \dots, x_{i+l+1/2}] \\ &= \frac{\mathbf{Q}_j[x_{i+3/2}, \dots, x_{i+l+1/2}] - \mathbf{Q}_j[x_{i+1/2}, \dots, x_{i+l-1/2}]}{l\Delta x}; \end{aligned}$$

\* Otherwise, we add  $x_{i+l+1/2}$  to the stencil.

- (iii) After the  $(k+1)$ -node stencil

$$\{x_{j-r-1/2}, x_{j-r+1/2}, \dots, x_{j-1/2}, x_{j+1/2}, \dots, x_{j+s-1/2}, x_{j+s+1/2}\},$$

is found, where  $s + r + 1 = k$ , we use Newton interpolation formula to obtain  $\mathbf{Q}_j(x)$ , which is a polynomial of degree at most  $k$ , satisfying

$$\mathbf{Q}_j(x_{i+1/2}) = \sum_{l=-\infty}^i \mathbf{U}_l^n \Delta x, \quad i = j - r - 1, \dots, j + s.$$

Then, we obtain

$$\mathbf{P}_j(x) = \frac{d}{dx} \mathbf{Q}_j(x),$$

which is a polynomial of degree at most  $k - 1$  satisfying (40).

(2) We solve the Cauchy problem for (32) with the initial condition

$$\mathbf{U}(x, 0) = \mathbf{U}_{\text{p.pol}}(x), \quad x \in \mathbb{R}, \quad (41)$$

to find the solution  $\mathbf{U}(\cdot, \Delta t)$ .

(3) We project  $\mathbf{U}(\cdot, \Delta t)$  onto the piecewise constant functions, that is, we set

$$\mathbf{U}_j^{n+1} = \frac{1}{\Delta x} \int_{x_{j-1/2}}^{x_{j+1/2}} \mathbf{U}(x, \Delta t) dx, \quad j \in \mathbb{Z}. \quad (42)$$

In order to obtain an explicit scheme, we integrate the equation (32) over the rectangle  $(x_{j-1/2}, x_{j+1/2}) \times (0, \Delta t)$ , we obtain

$$\begin{aligned} & \int_{x_{j-1/2}}^{x_{j+1/2}} (\mathbf{U}(x, \Delta t) - \mathbf{U}(x, 0)) dx \\ & + \int_0^{\Delta t} \left( \mathbf{F}(\mathbf{U}(x_{j+1/2} - 0, t)) - \mathbf{F}(\mathbf{U}(x_{j-1/2} + 0, t)) \right) dt \\ & = \int_{x_{j-1/2}}^{x_{j+1/2}} \int_0^{\Delta t} \mathbf{H}(\mathbf{U}) \partial_x a dt dx. \end{aligned} \quad (43)$$

Using (40), (39), (41) and (42), we get

$$\begin{aligned} & \Delta x (\mathbf{U}_j^{n+1} - \mathbf{U}_j^n) + \int_0^{\Delta t} \left( \mathbf{F}(\mathbf{U}(x_{j+1/2} - 0, t)) - \mathbf{F}(\mathbf{U}(x_{j-1/2} + 0, t)) \right) dt \\ & = \int_{x_{j-1/2}}^{x_{j+1/2}} \int_0^{\Delta t} \mathbf{H}(\mathbf{U}(x, t)) \partial_x a_j(x) dt dx. \end{aligned} \quad (44)$$

Approximating (44) by using the Midpoint Rule, we obtain

$$\begin{aligned} & \Delta x(\mathbf{U}_j^{n+1} - \mathbf{U}_j^n) + \Delta t \left( \mathbf{F}(\mathbf{U}(x_{j+1/2} - 0, \Delta t/2)) - \mathbf{F}(\mathbf{U}(x_{j-1/2} + 0, \Delta t/2)) \right) dt \\ &= \Delta x \Delta t \mathbf{H}(\mathbf{U}(x_j, \Delta t/2)) \cdot \partial_x a_j(x) \Big|_{x=x_j}. \end{aligned} \quad (45)$$

To approximate  $\mathbf{F}(\mathbf{U}(x_{j+1/2} - 0, \Delta t/2))$ ,  $\mathbf{F}(\mathbf{U}(x_{j-1/2} + 0, \Delta t/2))$ , and  $\mathbf{H}(\mathbf{U}(x_j, \Delta t/2))$ , we use a predictor-corrector method as follows:

(i) We compute the updated values  $\mathbf{U}_{j+1/2,\pm}^{n+1/2}$  by

$$\begin{aligned} & \mathbf{U}_{j+1/2,-}^{n+1/2} \\ &= \mathbf{U}_{j+1/2,-}^n - \frac{\Delta t}{2\Delta x} (\mathbf{F}(\mathbf{U}_{j+1/2,-}^n) - \mathbf{F}(\mathbf{U}_{j-1/2,+}^n)) \\ & \quad + \frac{\Delta t}{2} \mathbf{H}(\mathbf{U}_{j+1/2,-}^n) \cdot \partial_x a_j(x) \Big|_{x=x_{j+1/2}-}, \\ & \mathbf{U}_{j-1/2,+}^{n+1/2} \\ &= \mathbf{U}_{j-1/2,+}^n - \frac{\Delta t}{2\Delta x} (\mathbf{F}(\mathbf{U}_{j+1/2,-}^n) - \mathbf{F}(\mathbf{U}_{j-1/2,+}^n)) \\ & \quad + \frac{\Delta t}{2} \mathbf{H}(\mathbf{U}_{j-1/2,+}^n) \cdot \partial_x a_j(x) \Big|_{x=x_{j-1/2}+}, \end{aligned} \quad (46)$$

where

$$\begin{aligned} \mathbf{U}_{j+1/2,-}^n &= \mathbf{P}_j(x_{j+1/2}), \\ \mathbf{U}_{j-1/2,+}^n &= \mathbf{P}_j(x_{j-1/2}). \end{aligned} \quad (47)$$

(ii) We solve the Riemann problem of (32) with initial data

$$\mathbf{U}(x, 0) = \begin{cases} \mathbf{U}_{j+1/2,-}^{n+1/2}, & \text{if } x < x_{j+1/2}, \\ \mathbf{U}_{j+1/2,+}^{n+1/2}, & \text{if } x > x_{j+1/2}, \end{cases} \quad x_j < x < x_{j+1}, \quad j \in \mathbb{Z}, \quad (48)$$

to obtain the exact solution

$$\mathbf{U}(x, t) = \mathbf{U}^{\text{Rie}}\left(\frac{x - x_{j+1/2}}{t}; \mathbf{U}_{j+1/2,-}^{n+1/2}, \mathbf{U}_{j+1/2,+}^{n+1/2}\right), \quad x_j < x < x_{j+1}, \quad j \in \mathbb{Z}. \quad (49)$$

(iii) We approximate

$$\begin{aligned}
\mathbf{F}(\mathbf{U}(x_{j+1/2} - 0, \Delta t/2)) &\approx \mathbf{F}(\mathbf{U}^{\text{Rie}}(0-; \mathbf{U}_{j+1/2,-}^{n+1/2}, \mathbf{U}_{j+1/2,+}^{n+1/2})), \\
\mathbf{F}(\mathbf{U}(x_{j-1/2} + 0, \Delta t/2)) &\approx \mathbf{F}(\mathbf{U}^{\text{Rie}}(0+; \mathbf{U}_{j-1/2,-}^{n+1/2}, \mathbf{U}_{j-1/2,+}^{n+1/2})), \\
\mathbf{H}(\mathbf{U}(x_j, \Delta t/2)) &\approx \frac{1}{2} \left( \mathbf{H}(\mathbf{U}^{\text{Rie}}(0-; \mathbf{U}_{j+1/2,-}^{n+1/2}, \mathbf{U}_{j+1/2,+}^{n+1/2})) \right. \\
&\quad \left. + \mathbf{H}(\mathbf{U}^{\text{Rie}}(0+; \mathbf{U}_{j-1/2,-}^{n+1/2}, \mathbf{U}_{j-1/2,+}^{n+1/2})) \right).
\end{aligned}$$

Thus, the scheme (45) becomes

$$\begin{aligned}
\mathbf{U}_j^{n+1} = & \mathbf{U}_j^n - \frac{\Delta t}{\Delta x} \left( \mathbf{F}(\mathbf{U}^{\text{Rie}}(0-; \mathbf{U}_{j+1/2,-}^{n+1/2}, \mathbf{U}_{j+1/2,+}^{n+1/2})) \right. \\
& \left. - \mathbf{F}(\mathbf{U}^{\text{Rie}}(0+; \mathbf{U}_{j-1/2,-}^{n+1/2}, \mathbf{U}_{j-1/2,+}^{n+1/2})) \right) \\
& + \frac{\Delta t}{2} \left( \mathbf{H}(\mathbf{U}^{\text{Rie}}(0-; \mathbf{U}_{j+1/2,-}^{n+1/2}, \mathbf{U}_{j+1/2,+}^{n+1/2})) \right. \\
& \left. + \mathbf{H}(\mathbf{U}^{\text{Rie}}(0+; \mathbf{U}_{j-1/2,-}^{n+1/2}, \mathbf{U}_{j-1/2,+}^{n+1/2})) \right) \cdot \partial_x a_j(x) \Big|_{x=x_j}.
\end{aligned} \tag{50}$$

To complete the ENO-type scheme (50), we must visit the Riemann problem for (1) to define the values  $\mathbf{U}^{\text{Rie}}(0\pm; \mathbf{U}_L, \mathbf{U}_R)$  as follows:

- For construction A1 (19):  $\mathbf{U}^{\text{Rie}}(0-; \mathbf{U}_L, \mathbf{U}_R) = \mathbf{U}_L$ , and  $\mathbf{U}^{\text{Rie}}(0+; \mathbf{U}_L, \mathbf{U}_R) = \mathbf{U}_L^s$ , where  $\mathbf{U}_L^s = \left[ \rho_L^s, \frac{a_L \rho_L u_L}{a_R \rho_L^s}, a_R \right]^T$  is defined by Lemma 2.
- For construction A2 (21):  $\mathbf{U}^{\text{Rie}}(0-; \mathbf{U}_L, \mathbf{U}_R) = \mathbf{U}_L$ , and  $\mathbf{U}^{\text{Rie}}(0+; \mathbf{U}_L, \mathbf{U}_R) = \mathbf{U}_*$ , where  $\mathbf{U}_*$  is the intersection point of  $\mathcal{W}_{2B}(\mathbf{U}_R)$  defined by (17), and  $\mathcal{W}_3 \circ \mathcal{S}_1 \circ \mathcal{W}_3(\mathbf{U}_L, a_R)$  defined by (22).
- For construction A3 (23):  $\mathbf{U}^{\text{Rie}}(0-; \mathbf{U}_L, \mathbf{U}_R) = \mathbf{U}_*$ , and  $\mathbf{U}^{\text{Rie}}(0+; \mathbf{U}_L, \mathbf{U}_R) = \mathbf{U}_*^b$ , where  $\mathbf{U}_*$  belongs to  $\mathcal{S}_1(\mathbf{U}_L)$  defined by (10), and  $\mathbf{U}_*^b$  is the intersection point of  $\mathcal{W}_{2B}(\mathbf{U}_R)$  defined by (17), and  $\mathcal{S}_1 \circ \mathcal{W}_3(\mathbf{U}_L, a_R)$  defined by (24).
- For construction B1 (25):  $\mathbf{U}^{\text{Rie}}(0-; \mathbf{U}_L, \mathbf{U}_R) = \mathbf{U}_L^+$ , and  $\mathbf{U}^{\text{Rie}}(0+; \mathbf{U}_L, \mathbf{U}_R) = \mathbf{U}_L^{+s}$ , where  $\mathbf{U}_L^+ = \mathcal{R}_1(\mathbf{U}_L) \cap \mathcal{C}^+$ , and  $\mathbf{U}_L^{+s} = \left[ (\rho_L^+)^s, \frac{a_L \rho_L^+ u_L^+}{a_R (\rho_L^+)^s}, a_R \right]^T$  is defined by Lemma 2.
- For construction B2 (27):  $\mathbf{U}^{\text{Rie}}(0-; \mathbf{U}_L, \mathbf{U}_R) = \mathbf{U}_L^+$ , and  $\mathbf{U}^{\text{Rie}}(0+; \mathbf{U}_L, \mathbf{U}_R) = \mathbf{U}_*$ , where  $\mathbf{U}_*$  is the intersection point of  $\mathcal{W}_{2B}(\mathbf{U}_R)$  defined by (17), and  $\mathcal{R}_1 \circ \mathcal{W}_3 \circ \mathcal{S}_1 \circ \mathcal{W}_3(\mathbf{U}_L, a_R)$  defined by (28).

- For construction B3 (29):  $\mathbf{U}^{\text{Rie}}(0-; \mathbf{U}_L, \mathbf{U}_R) = \mathbf{U}_*$ , and  $\mathbf{U}^{\text{Rie}}(0+; \mathbf{U}_L, \mathbf{U}_R) = \mathbf{U}_*^b$ , where  $\mathbf{U}_*$  belongs to  $\mathcal{W}_1(\mathbf{U}_L)$  defined by (17), and  $\mathbf{U}_*^b$  is the intersection point of  $\mathcal{W}_{2B}(\mathbf{U}_R)$  defined by (17), and  $\mathcal{W}_1 \circ \mathcal{W}_3(\mathbf{U}_L, a_R)$  defined by (30).

## 4 Numerical experiments with discussions

This section is aimed to demonstrate the accuracy of our scheme (50) with some numerical tests with MATLAB. For each test, we find the numerical solutions  $\mathbf{U}_h$  by our scheme (50) with taking

$$\kappa = 1.0, \quad \gamma = 1.6,$$

and we then compare  $\mathbf{U}_h$  with the corresponding exact solution  $\mathbf{U}$ .

### 4.1 Test for well-balanced property

**Test 1.** In this test, we aim to demonstrate that the ENO-like scheme (50) can capture a smooth stationary wave with second order accuracy. Let us consider the Cauchy problem for system (4) with the initial smooth data given by

$$\mathbf{U}(x, 0) = \left[ \rho(x), u(x), a(x) \right]^T, \quad x \geq 0, \quad (51)$$

where  $a(x) = 1 + \frac{1}{2}x^3$ , and  $(\rho(\cdot), u(\cdot))$  is the solution of the following problem:

$$\begin{aligned} \frac{d}{dx}(a\rho u) &= 0, \\ \frac{d}{dx}\left(\frac{u^2}{2} + \frac{\kappa\gamma\rho^{\gamma-1}}{\gamma-1}\right) &= 0, \\ (\rho, u)\Big|_{x=0} &= (0.5, 1.5), \\ \lambda_1(\rho, u) &= u - \sqrt{\kappa\gamma\rho^{\gamma-1}} > 0, \quad x \geq 0. \end{aligned} \quad (52)$$

The exact solution of this problem is just a smooth stationary wave

$$\mathbf{U}(x, t) = \left[ \rho(x), u(x), a(x) \right]^T, \quad x \geq 0, \quad t \geq 0.$$

Figure 4 displays the exact solution and the approximate solution by the ENO-like scheme (50) with  $k = 3$  for the mesh size  $h = 1/80$  at time  $t = 0.1$  and on spatial domain  $x \in [0, 1]$ . The errors, orders of convergence are reported by Table 1.

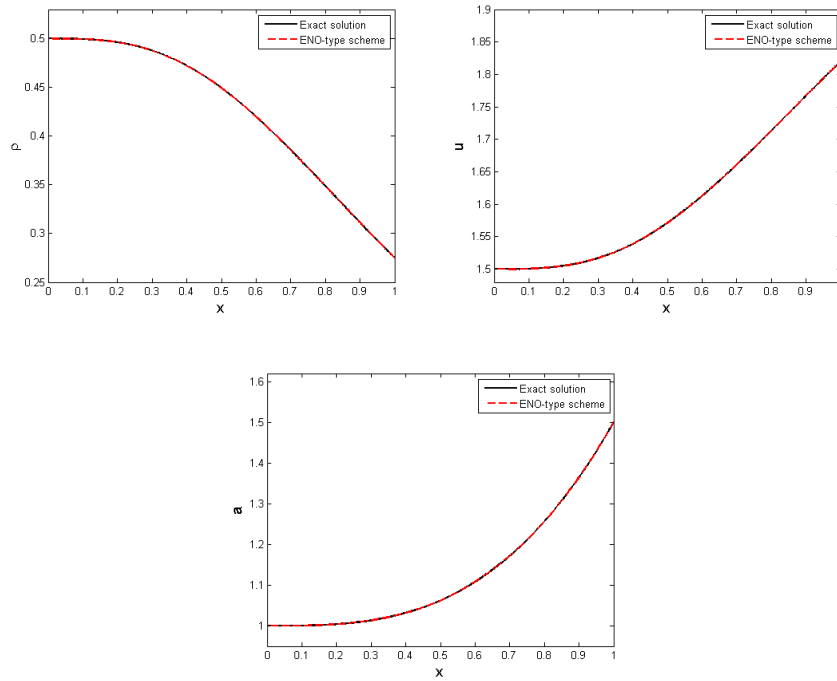


Figure 4: Exact solution and approximate solution by the ENO-like scheme (50) with  $k = 3$  for the mesh size  $h = 1/80$  at time  $t = 0.1$  and on spatial domain  $x \in [0, 1]$  of Test 1

Table 1: Errors and orders of convergence for Test 1

$h$	$L^1$ -error	Order
1/10	$0.20912 \times 10^{-3}$	—
1/20	$0.05769 \times 10^{-3}$	1.86
1/40	$0.013918 \times 10^{-3}$	2.05
1/80	$0.003443 \times 10^{-3}$	2.02
1/160	$0.000917 \times 10^{-3}$	1.91

## 4.2 Test for a complete Riemann solution when initial data belongs to same region

**Test 2.** In this test, we approximate the Riemann solution of the problem (4) with initial data

$$\begin{array}{c|c|c} & \mathbf{U}_L & \mathbf{U}_R \\ \hline \rho & 0.5 & 0.7 \\ u & 1.5 & 2.0 \\ a & 2.0 & 2.5 \end{array} \quad (53)$$

where  $\mathbf{U}_L, \mathbf{U}_R$  are in the same supersonic region  $G_1$ . According to the Construction A1, the Riemann solution is

$$\mathbf{U}(x, t) = \begin{cases} \mathbf{U}_L, & \text{if } x/t < 0, \\ \mathbf{U}_L^s, & \text{if } 0 < x/t < \sigma_1(\mathbf{U}_L^s, \mathbf{U}_*), \\ \mathbf{U}_*, & \text{if } \sigma_1(\mathbf{U}_L^s, \mathbf{U}_*) < x/t < \lambda_2(\mathbf{U}_*), \\ R_2(\mathbf{U}_*, \mathbf{U}_R), & \text{if } x/t > \lambda_2(\mathbf{U}_*), \end{cases}$$

where

$$\begin{array}{c|c|c} & \mathbf{U}_L^s & \mathbf{U}_* \\ \hline \rho & 0.350918 & 0.436769 \\ u & 1.709803 & 1.50012 \\ a & 2.5 & 2.5 \end{array}$$

Figure 5 displays the exact solution and its approximate solutions by the ENO-like scheme (50) with  $k = 3$  and  $k = 7$  for the mesh size  $h = 1/320$  at time  $t = 0.1$  and on spatial domain  $x \in [-1, 1]$ . The errors, orders of convergence are reported in Table 2. This table shows that the errors of the ENO-like scheme (50) are much smaller than the ones of the Godunov-type scheme (35), and the orders of convergence of the ENO-like scheme (50) are higher than the ones of the Godunov-type scheme (35) for all  $k = 2, 3, 4, 5, 6, 7$ . Specially, the errors of the ENO-like scheme (50) with  $k = 3$  are smaller than the ones of the van Leer-type scheme (38), although very small.

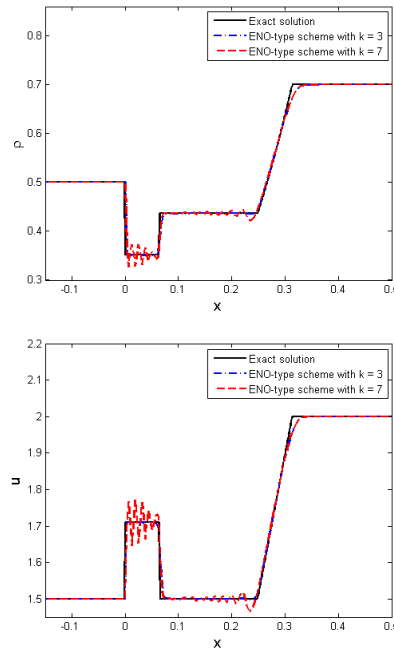


Figure 5: Exact solution and its approximate solutions by the ENO-like scheme (50) with  $k = 3$  and  $k = 7$  for the mesh size  $h = 1/320$  at time  $t = 0.1$  of Test 2

Table 2: Errors and orders of convergence for Test 2

$h$	Godunov-type		van Leer-type		ENO-like $k = 2$		ENO-like $k = 3$	
	$L^1$ -error	Order	$L^1$ -error	Order	$L^1$ -error	Order	$L^1$ -error	Order
1/10	0.147210	—	0.133470	—	0.137800	—	0.136490	—
1/20	0.092721	0.67	0.071526	0.90	0.075273	0.87	0.070917	0.94
1/40	0.056977	0.70	0.037260	0.94	0.039782	0.92	0.035419	1.00
1/80	0.036229	0.65	0.017500	1.09	0.020400	0.96	0.016978	1.06
1/160	0.023050	0.65	0.008817	0.99	0.010675	0.93	0.008495	1.00
1/320	0.014581	0.66	0.004534	0.96	0.005647	0.92	0.004427	0.94
$h$	ENO-like $k = 4$		ENO-like $k = 5$		ENO-like $k = 6$		ENO-like $k = 7$	
	$L^1$ -error	Order	$L^1$ -error	Order	$L^1$ -error	Order	$L^1$ -error	Order
1/10	0.138800	—	0.141660	—	0.144350	—	0.135280	—
1/20	0.072094	0.95	0.074614	0.92	0.076848	0.91	0.072276	0.90
1/40	0.034414	1.07	0.035282	1.08	0.036509	1.07	0.036410	0.99
1/80	0.016349	1.07	0.016685	1.08	0.017567	1.06	0.019454	0.90
1/160	0.008891	0.88	0.009398	0.83	0.011317	0.63	0.012965	0.59
1/320	0.004822	0.88	0.005238	0.84	0.006964	0.70	0.007721	0.75

### 4.3 Test for a complete Riemann solution when initial data belongs to different regions

**Test 3.** Consider the Riemann data

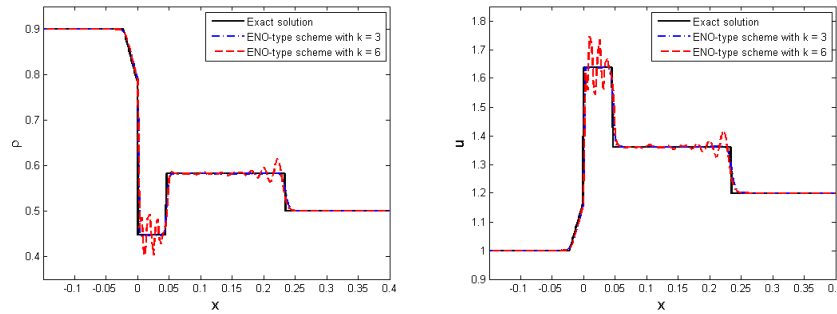


Figure 6: Exact solution and approximate solutions by the ENO-like scheme (50) with  $k = 3$  and  $k = 6$  for the mesh size  $h = 1/320$  at time  $t = 0.1$  of Test 3

$$\begin{array}{c|c|c} & \mathbf{U}_L & \mathbf{U}_R \\ \hline \rho & 0.9 & 0.5 \\ \hline u & 1.0 & 1.2 \\ \hline a & 2.0 & 2.5 \end{array} \quad (54)$$

where  $\mathbf{U}_L \in G_2$ , and  $\mathbf{U}_R \in G_1$ . According to the Construction B1, the exact solution is

$$\mathbf{U}(x, t) = \begin{cases} R_1(\mathbf{U}_L, \mathbf{U}_L^+), & \text{if } x/t < 0, \\ \mathbf{U}_L^{+s}, & \text{if } 0 < x/t < \sigma_1(\mathbf{U}_L^{+s}, \mathbf{U}_*), \\ \mathbf{U}_*, & \text{if } \sigma_1(\mathbf{U}_L^{+s}, \mathbf{U}_*) < x/t < \sigma_2(\mathbf{U}_*, \mathbf{U}_R), \\ \mathbf{U}_R, & \text{if } x/t > \sigma_2(\mathbf{U}_*, \mathbf{U}_R), \end{cases}$$

where

	$\mathbf{U}_L^+$	$\mathbf{U}_L^{+s}$	$\mathbf{U}_*$
$\rho$	0.778780	0.446692	0.582528
$u$	1.173504	1.636746	1.360876
$a$	2.0	2.5	2.5

Figure 6 displays the exact solution and its approximate solutions by the ENO-like scheme (50) with  $k = 3$  and  $k = 6$  for the mesh size  $h = 1/320$  at time  $t = 0.1$  and on spatial domain  $x \in [-1, 1]$ . The errors, orders of convergence are reported in Table 3. This test indicates that the errors of the ENO-like scheme (50) are smaller than those of the Godunov-type scheme (35) only for  $k = 2, 3, 4$ . Specially, the errors, orders of convergence of the ENO-like scheme (50) with  $k = 3$  are approximate to the ones of the van Leer-type scheme (38).

Table 3: Errors and orders of convergence for Test 3

$h$	Godunov-type		van Leer-type		ENO-like $k = 2$		ENO-like $k = 3$	
	$L^1$ -error	Order	$L^1$ -error	Order	$L^1$ -error	Order	$L^1$ -error	Order
1/10	0.145170	—	0.140000	—	0.143600	—	0.141710	—
1/20	0.088237	0.72	0.083727	0.74	0.085501	0.75	0.085796	0.72
1/40	0.045084	0.97	0.038502	1.12	0.038707	1.14	0.039072	1.13
1/80	0.026477	0.77	0.019911	0.95	0.020432	0.92	0.020352	0.94
1/160	0.015181	0.80	0.009689	1.04	0.010287	0.99	0.009670	1.07
1/320	0.009123	0.73	0.005674	0.77	0.005962	0.79	0.005467	0.82
$h$	ENO-like $k = 4$		ENO-like $k = 5$		ENO-like $k = 6$		ENO-like $k = 7$	
	$L^1$ -error	Order	$L^1$ -error	Order	$L^1$ -error	Order	$L^1$ -error	Order
1/10	0.141280	—	0.141260	—	0.141420	—	0.141640	—
1/20	0.086031	0.72	0.086124	0.71	0.086201	0.71	0.086289	0.71
1/40	0.040304	1.09	0.041208	1.06	0.041770	1.05	0.042117	1.03
1/80	0.023236	0.79	0.025029	0.72	0.026564	0.65	0.027639	0.61
1/160	0.011882	0.97	0.015186	0.72	0.017823	0.58	0.019678	0.49
1/320	0.006408	0.89	0.009008	0.75	0.010765	0.73	0.013075	0.59

#### 4.4 Test for a resonant phenomenon case

**Test 4.** This test is conducted to show that the scheme (50) can work well in regions of resonance, where three waves propagate at same speed. Consider the Riemann initial data

$$\begin{array}{c|cc} & \mathbf{U}_L & \mathbf{U}_R \\ \hline \rho & 0.5 & 1.2 \\ u & 1.5 & 0.9 \\ a & 2.0 & 2.5 \end{array} \quad (55)$$

where  $\mathbf{U}_L \in G_1$ , and  $\mathbf{U}_R \in G_2$ . According to the Construction A2, the exact solution is

$$\mathbf{U}(x, t) = \begin{cases} \mathbf{U}_L & \text{if } x/t < 0, \\ \mathbf{U}_L^{s\#b}, & \text{if } 0 < x/t < \lambda_2(\mathbf{U}_L^{s\#b}), \\ R_2(\mathbf{U}_L^{s\#b}, \mathbf{U}_R), & \text{if } x/t > \lambda_2(\mathbf{U}_L^{s\#b}), \end{cases} \quad (56)$$

where

$$\begin{array}{c|ccc} & \mathbf{U}_L^s & \mathbf{U}_L^{s\#} & \mathbf{U}_L^{s\#b} \\ \hline \rho & 0.458944 & 0.886495 & 0.966873 \\ u & 1.557664 & 0.806412 & 0.620557 \\ a & 2.098252 & 2.098252 & 2.5 \end{array}$$

In this Riemann solution (56), we can see that it contains three waves propagating at zero speed, that is,  $W_3(\mathbf{U}_L, \mathbf{U}_L^s)$ ,  $S_1(\mathbf{U}_L^s, \mathbf{U}_L^{s\#})$ , and  $W_3(\mathbf{U}_L^{s\#}, \mathbf{U}_L^{s\#b})$ .

Figure 7 shows the exact solution and its approximate solutions by the ENO-like scheme (50) with  $k = 3$  and  $k = 5$  for the mesh size  $h = 1/320$  at time  $t = 0.1$  and on spatial domain  $x \in [-1, 1]$ . This figure demonstrates

the convergence of the approximate solutions by the ENO-like scheme (50) when the Riemann data belongs to regions of resonance. The errors, orders of convergence are reported in Table 4. We can see from this table that the ENO-like scheme (50) with  $k = 2, 3, 4, 5, 6$  has a better accuracy than the Godunov-type scheme (35). Again, we also see that the errors, orders of convergence of the ENO-like scheme (50) with  $k = 3$  are the same as those of the van Leer-type scheme (38).

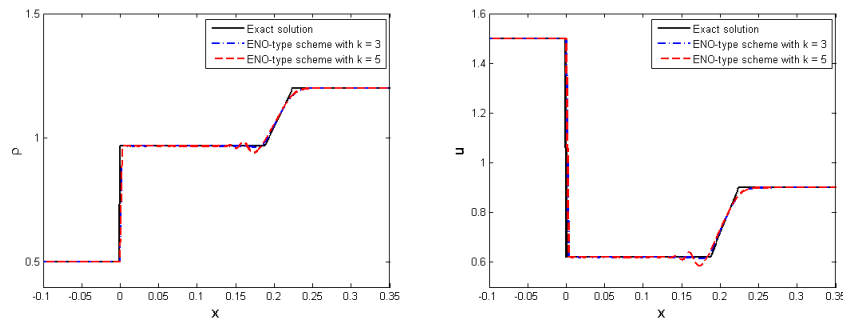


Figure 7: Exact solution and approximate solutions by the ENO-like scheme (50) with  $k = 3$  and  $k = 5$  for the mesh size  $h = 1/320$  at the time  $t = 0.1$  of Test 4

Table 4: Errors and orders of convergence for Test 4

	Godunov-type		van Leer-type		ENO-like $k = 2$		ENO-like $k = 3$	
$h$	$L^1$ -error	Order	$L^1$ -error	Order	$L^1$ -error	Order	$L^1$ -error	Order
1/10	0.156310	—	0.149430	—	0.146500	—	0.147310	—
1/20	0.089646	0.80	0.078463	0.93	0.078737	0.90	0.076385	0.95
1/40	0.053216	0.75	0.042118	0.90	0.043172	0.87	0.042851	0.83
1/80	0.030277	0.81	0.021359	0.98	0.022269	0.96	0.021173	1.02
1/160	0.017451	0.79	0.010617	1.01	0.011415	0.96	0.010297	1.04
1/320	0.010445	0.74	0.005360	0.99	0.005884	0.96	0.005193	0.99
	ENO-like $k = 4$		ENO-like $k = 5$		ENO-like $k = 6$		ENO-like $k = 7$	
$h$	$L^1$ -error	Order	$L^1$ -error	Order	$L^1$ -error	Order	$L^1$ -error	Order
1/10	0.147400	—	0.147240	—	0.147090	—	0.147890	—
1/20	0.076007	0.96	0.076504	0.94	0.076737	0.94	0.080435	0.88
1/40	0.044977	0.76	0.046037	0.73	0.046879	0.71	0.049658	0.70
1/80	0.023064	0.96	0.024237	0.93	0.025491	0.88	0.030240	0.72
1/160	0.011652	0.99	0.012472	0.96	0.013368	0.93	0.025279	0.26
1/320	0.006026	0.95	0.006472	0.95	0.007058	0.92	0.018523	0.45

#### 4.5 Test for interaction of waves

**Test 5.** In this test, we approximate a Cauchy problem with initial condition:

$$\mathbf{U}(x, 0) = \begin{cases} \mathbf{U}_L = [0.8, 2.0, 2.0]^T, & \text{if } x < 0, \\ \mathbf{U}_M = [0.5, 1.5, 2.5]^T, & \text{if } 0 < x < 1, \\ \mathbf{U}_R = [0.372067, 1.679806, 3.0]^T, & \text{if } x > 1, \end{cases} \quad (57)$$

where  $\mathbf{U}_L, \mathbf{U}_M, \mathbf{U}_R \in G_1$ , and  $\mathbf{U}_R = \mathbf{U}_M^s$ . At time  $t = 0.4$ , we can check that the Riemann solution at  $x = 0$  interacts with the one at  $x = 1$ . Therefore, the exact solution at  $t = 0.4$  is

$$\mathbf{U}(x, t) = \begin{cases} \mathbf{U}_L, & \text{if } x/t < 0, \\ \mathbf{U}_1, & \text{if } 0 < x/t < \sigma_1(\mathbf{U}_1, \mathbf{U}_2), \\ \mathbf{U}_2, & \text{if } \sigma_1(\mathbf{U}_1, \mathbf{U}_2) < x/t \text{ and } (x-1)/t < 0, \\ \mathbf{U}_3, & \text{if } 0 < (x-1)/t < \sigma_1(\mathbf{U}_3, \mathbf{U}_4), \\ \mathbf{U}_4, & \text{if } \sigma_1(\mathbf{U}_3, \mathbf{U}_4) < (x-1)/t < \sigma_2(\mathbf{U}_4, \mathbf{U}_R), \\ \mathbf{U}_R, & \text{if } (x-1)/t > \sigma_2(\mathbf{U}_4, \mathbf{U}_R), \end{cases}$$

where

	$\mathbf{U}_1$	$\mathbf{U}_2$	$\mathbf{U}_3$	$\mathbf{U}_4$
$\rho$	0.584096	0.738236	0.567757	0.562968
$u$	2.191420	1.929233	2.090433	2.099463
$a$	2.5	2.5	3.0	3.0

Figure 8 displays the exact solution and its approximate solutions by the ENO-like scheme (50) with  $k = 3$  and  $k = 5$  for the mesh size  $h = 1/320$  at time  $t = 0.4$  and on spatial domain  $x \in [-2, 2]$ . The errors, orders of convergence are reported in Table 5. Like all tests above, this test also indicates that the errors of the ENO-like scheme (50) with  $k = 2, 3, 4, 5$  are much smaller than the ones of the Godunov-type scheme (35), and the orders of convergence of the ENO-like scheme (50) with  $k = 2, 3, 4, 5$  are higher than the ones of the Godunov-type scheme (35). We also see that the accuracy of the ENO-like scheme (50) is less than the van Leer-type scheme (38) for all  $k = 2, 3, 4, 5, 6, 7$ .

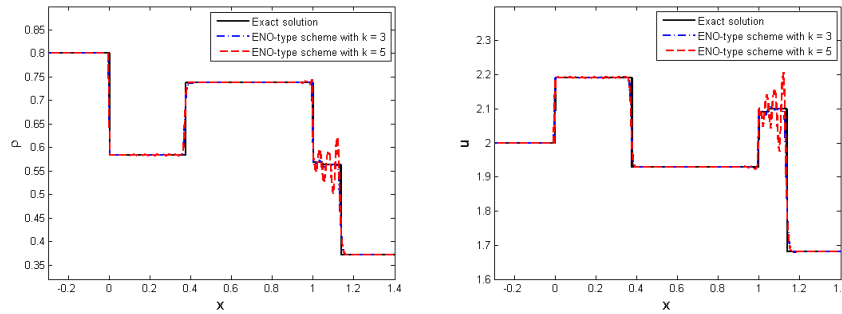


Figure 8: Exact solution and approximate solutions by the ENO-like scheme (50) with  $k = 3$  and  $k = 5$  for the mesh size  $h = 1/160$  at time  $t = 0.4$  of Test 5

Table 5: Errors and orders of convergence for Test 5

$h$	Godunov-type		van Leer-type		ENO-like $k = 2$		ENO-like $k = 3$	
	$L^1$ -error	Order	$L^1$ -error	Order	$L^1$ -error	Order	$L^1$ -error	Order
1/5	0.218000	—	0.164300	—	0.180440	—	0.170460	—
1/10	0.161200	0.44	0.095635	0.78	0.112080	0.69	0.102300	0.74
1/20	0.107390	0.59	0.045529	1.07	0.057382	0.97	0.052002	0.98
1/40	0.067436	0.67	0.024359	0.90	0.029569	0.96	0.026578	0.97
1/80	0.041364	0.71	0.011765	1.05	0.015053	0.97	0.012180	1.13
1/160	0.023396	0.82	0.005772	1.03	0.007526	1.00	0.006322	0.95
$h$	ENO-like $k = 4$		ENO-like $k = 5$		ENO-like $k = 6$		ENO-like $k = 7$	
	$L^1$ -error	Order	$L^1$ -error	Order	$L^1$ -error	Order	$L^1$ -error	Order
1/5	0.187650	—	0.169700	—	0.200110	—	0.191990	—
1/10	0.111190	0.76	0.115720	0.55	0.121310	0.72	0.121630	0.66
1/20	0.056829	0.97	0.073652	0.65	0.086137	0.49	0.091910	0.40
1/40	0.028911	0.98	0.045233	0.70	0.059927	0.52	0.067928	0.44
1/80	0.012443	1.22	0.023754	0.93	0.035052	0.77	0.047736	0.51
1/160	0.006982	0.83	0.015361	0.63	0.028422	0.30	0.036563	0.38

## 5 Conclusions and discussion

The high-resolution ENO-like schemes for the model (1) constructed in this work can approximate exact solutions very well for all kinds of data: supersonic, subsonic, or both. The ENO-like scheme corresponding to  $k = 3$  still works well even in the resonant regime, where the exact solution containing multiple waves associated with different characteristic fields propagates with the same shock speed. This scheme still maintains some valuable properties of the original one for hyperbolic systems of conservation laws: it is oscillatory and has high order accuracy. Numerical tests show that the scheme has a much better accuracy than the Godunov-type scheme and can approx-

imate smooth stationary waves with a second-order accuracy. The ENO-like scheme corresponding to  $k = 3$  works as good as the van Leer-type scheme. However, the ENO-like schemes for larger  $k$  may suffer oscillations.

## Acknowledgment

We are very grateful to the reviewers for their very constructive comments and helpful suggestions.

## References

- [1] Ambroso, A., Chalons, C., Coquel, F. and Galié, T. *Relaxation and numerical approximation of a two-fluid two-pressure diphasic model*, Math. Mod. Numer. Anal., 43 (2009), 1063–1097.
- [2] Ambroso, A., Chalons, C. and Raviart, P.-A. *A Godunov-type method for the seven-equation model of compressible two-phase flow*, Computers & Fluids, 54 (2012), 67–91.
- [3] Audusse, E., Bouchut, F., Bristeau, M.-O., Klein, R. and Perthame, B. *A fast and stable well-balanced scheme with hydrostatic reconstruction for shallow water flows*, SIAM J. Sci. Comput., 25 (2004), 2050–2065.
- [4] Baudin, M., Coquel, F. and Tran, Q.-H. *A semi-implicit relaxation scheme for modeling two-phase flow in a pipeline*, SIAM J. Sci. Comput., 27 (2005), 914–936.
- [5] Ben-Artzi, M. and Falcovitz, J. *An upwind second-order scheme for compressible duct flows*, SIAM J. Sci. and Stat. Comput., 7(2006), 744–768.
- [6] Botchorishvili, R., Perthame, B. and Vasseur, A. *Equilibrium schemes for scalar conservation laws with stiff sources*, Math. Comput., 72 (2003), 131–157.

- [7] Botchorishvili, R. and Pironneau, O. *Finite volume schemes with equilibrium type discretization of source terms for scalar conservation laws*, J. Comput. Phys., 187 (2003), 391–427.
- [8] Castro, M.J., Chalons, C., Del Grosso, A. and Morales de Luna, T. *Lagrange-projection methods for shallow water equations with movable bottom and erosion-deposition processes* Num. Math.: Theory, Meth. Appl., 16(2023), 1087–1126.
- [9] Castro, C.E., and Toro, E.F. *A Riemann solver and upwind methods for a two-phase flow model in non-conservative form*, Internat. J. Numer. Methods Fluids, 50 (2006), 275–307.
- [10] Chalons, C., Del Grosso, A. and Toro, E.F. *Numerical approximation and uncertainty quantification for arterial blood flow models with viscoelasticity*, J. Comput. Phys., 457 (2022), 111071.
- [11] Chinnayya, A., LeRoux, A.-Y. and Seguin, N. *A well-balanced numerical scheme for the approximation of the shallow water equations with topography: the resonance phenomenon*, Int. J. Finite Vol., 1(4), 2004, 1–33.
- [12] Coquel, F., El Amine, K., Godlewski, E., Perthame, B. and P. Rascle, *A numerical method using upwind schemes for the resolution of two-phase flows*, J. Comput. Phys. 136 (1997), 272–288.
- [13] Coquel, F., Hérard, J.-M., Saleh, K. and Seguin, N. *Two properties of two-velocity two-pressure models for two-phase flows*, Commun. Math. Sci. 12 (2014), 593–600.
- [14] Coquel, F., Saleh, K. and Seguin, N. *A robust and entropy-satisfying numerical scheme for fluid flows in discontinuous nozzles*, Math. Mod. Meth. Appl. Sci., 24 (2014), 2043–2083.
- [15] Cuong, D.H. and Thanh, M.D. *A Godunov-type scheme for the isentropic model of a fluid flow in a nozzle with variable cross-section*, Appl. Math. Comput., 256 (2015) 602–629.

- [16] Cuong, D.H. and Thanh, M.D. *A high-resolution van Leer-type scheme for a model of fluid flows in a nozzle with variable cross-section*, J. Korean Math. Soc., Vol. 54 (1) (2017), 141–175.
- [17] Cuong, D.H. and Thanh, M.D. *Computing algorithms in resonant regime for a two-phase flow model*, Taiwan. J. Math. Dec. (2023), 1135–1168
- [18] Dal Maso, G., LeFloch, P.G. and Murat, F. *Definition and weak stability of nonconservative products*, J. Math. Pures Appl., 74 (1995), 483–548.
- [19] Gallouët, T., Hérard, J.-M. and Seguin, N. *Numerical modeling of two-phase flows using the two-fluid two-pressure approach*, Math. Models Methods Appl. Sci., 14 (2004) 663–700.
- [20] Goatin, P. *Macroscopic traffic flow modelling: from kinematic waves to autonomous vehicles*, Commun. Appl. Ind. Math., 14(1) (2023), 1–16.
- [21] Goatin, P. and LeFloch, P.G. *The Riemann problem for a class of resonant nonlinear systems of balance laws*, Ann. Inst. H. Poincaré Anal. NonLinéaire, 21 (2004), 881–902.
- [22] Greenberg, J.M. and Leroux, A.Y. *A well-balanced scheme for the numerical processing of source terms in hyperbolic equations*, SIAM J. Numer. Anal., 33 (1996) 1–16.
- [23] Han, X. and Li, G. *Well-balanced finite difference WENO schemes for the Ripa model*, Comput. Fluid. 134-135 (2016), 1–10.
- [24] Harten, A., Engquist, B., Osher, S. and Chakravarthy, S. *Uniformly high order essentially non-oscillatory schemes, III*, J. Comput. Phys., 71(1987), 231–303.
- [25] Isaacson, E. and Temple, B. *Nonlinear resonance in systems of conservation laws*, SIAM J. Appl. Math., 52 (1992) 1260–1278.
- [26] Isaacson, E. and Temple, B. *Convergence of the  $2 \times 2$  Godunov method for a general resonant nonlinear balance law*, SIAM J. Appl. Math., 55 (1995) 625–640.

- [27] Kröner, D., LeFloch, P.G. and Thanh, M.D. *The minimum entropy principle for fluid flows in a nozzle with discontinuous cross-section*, Math. Mod. Numer. Anal., 42 (2008), 425–442.
- [28] Kröner, D. and Thanh, M.D. *Numerical solutions to compressible flows in a nozzle with variable cross-section*, SIAM J. Numer. Anal., 43 (2005), 796–824.
- [29] LeFloch, P.G. and Thanh, M.D. *The Riemann problem for fluid flows in a nozzle with discontinuous cross-section*, Comm. Math. Sci., 1 (2003), 763–797.
- [30] LeFloch, P.G. and Thanh, M.D. *A Godunov-type method for the shallow water equations with variable topography in the resonant regime*, J. Comput. Phys., 230 (2011), 7631–7660.
- [31] LeFloch, P.G. and Thanh, M.D. *The Riemann problem in continuum physics*, Appl. Math. Sci., Springer, 2024.
- [32] Marchesin, D. and Paes-Leme, P.J. *A Riemann problem in gas dynamics with bifurcation. Hyperbolic partial differential equations III*, Comput. Math. Appl. (Part A), 12 (1986) 433–455.
- [33] Munkejord, S.T. *Comparison of Roe-type methods for solving the two-fluid model with and without pressure relaxation*, Computers & Fluids, 36 (2007), 1061–1080.
- [34] Saurel, R. and Abgrall, R. *A multi-phase Godunov method for compressible multifluid and multiphase flows*, J. Comput. Phys., 150 (1999), 425–467.
- [35] Schwendeman, D.W., Wahle, C.W. and Kapila, A.K. *The Riemann problem and a high-resolution Godunov method for a model of compressible two-phase flow*, J. Comput. Phys., 212 (2006), 490–526.
- [36] Shen, H. *A class of ENO schemes with adaptive order for solving hyperbolic conservation laws*, Computers & Fluids, 266, (2023), 106050.

- [37] Shu, C.W. *Essentially non-oscillatory and weighted essentially non-oscillatory schemes for hyperbolic conservation laws*. In: *Quarteroni A. (eds) Advanced Numerical Approximation of Nonlinear Hyperbolic Equations*. Lecture Notes in Mathematics, vol 1697, Springer, Berlin, Heidelberg (1998), 325–432.
- [38] Thanh, M.D. The Riemann problem for a non-isentropic fluid in a nozzle with discontinuous cross-sectional area, *SIAM J. Appl. Math.*, 69 (2009), 1501–1519.
- [39] Thanh, M.D., Kröner, D. and C. Chalons, *A robust numerical method for approximating solutions of a model of two-phase flows and its properties*, *Appl. Math. Comput.*, 219 (2012), 320–344.
- [40] Thanh, M.D., Kröner, D. and N.T. Nam, *Numerical approximation for a Baer-Nunziato model of two-phase flows*, *Appl. Numer. Math.*, 61 (2011), 702–721.
- [41] Thanh, N.X., Thanh, M.D. and Cuong, D.H. *A well-balanced high-order scheme on van Leer-type for the shallow water equations with temperature gradient and variable bottom topography*, *Adv. Comput. Math.*, 47 (2021) 1–53.
- [42] Xu, C., Zhang, F., Dong, H. and Jiang, H. *Arbitrary high-order extended essentially non-oscillatory schemes for hyperbolic conservation laws*, *Int. J. for Num. Meth. Fluids*, 93(7) (2021), 2136–2154.

## **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 opti-

mization. Articles must be written in English.

All submitted papers will be refereed and the authors may be asked to revise their manuscripts according to the referee's reports. The Editorial Board of the Journal keeps the right to accept or reject the papers for publication.

The papers with more than one authors, should determine the corresponding author. The e-mail address of the corresponding author must appear at the end of the manuscript or as a footnote of the first page.

It is strongly recommended to set up the manuscript by Latex or Tex, using the template provided in the web site of the Journal. Manuscripts should be typed double-spaced with wide margins to provide enough room for editorial remarks.

References should be arranged in alphabetical order by the surname of the first author as examples below:

- [1] Brunner, H. *A survey of recent advances in the numerical treatment of Volterra integral and integro-differential equations*, J. Comput. Appl. Math. 8 (1982), 213-229.
- [2] Stoer, J. and Bulirsch, R. *Introduction to Numerical Analysis*, Springer-verlag, New York, 2002.

# Iranian Journal of Numerical Analysis and Optimization

CONTENTS

Vol. 15, No. 3, pp 852-1309

<b>The analysis of the mathematical stability of a cholera disease model . . . . .</b>	<b>852</b>
I. Sahib, M. Baroudi, H. Gourram, B. Khajji, A. Labzai and M. Belam	
<b>Two-step inertial Tseng's extragradient methods for a class of bilevel split variational inequalities . . . . .</b>	<b>877</b>
L.H.M. Van and T.V. Anh	
<b>Approximate symmetries of the perturbed KdV-KS equation</b>	<b>914</b>
A. Mohammadpouri, M.S. Hashemi, R. Abbasi and R. Abbasi	
<b>Convergence analysis of triangular and symmetric splitting method for fuzzy stochastic linear systems . . . . .</b>	<b>930</b>
B. Harika, D. Rajaiah, A. Shivaji, and L.P. Rajkumar	
<b>Mathematical modeling of COVID-19 spread with media coverage and optimal control analysis . . . . .</b>	<b>952</b>
G.P. Sahu and A.S. Thakur	
<b>Space-time localized scheme to solve some partial integro-differential equations . . . . .</b>	<b>993</b>
M. Hamaidi, M. Briki, A. Nouara and B. Hamdi	
<b>A study on the convergence and error bound of solutions to 2D mixed Volterra–Fredholm integral and integro-differential equations via high-order collocation method . . . . .</b>	<b>1012</b>
A.A. Shalangwa, M.R. Odekunle and S.O. Adeo	
<b>Cutting-edge spectral solutions for differential and integral equations utilizing Legendre's derivatives . . . . .</b>	<b>1036</b>
A.M. Abbas, Y.H. Youssri, M. El-Kady and M. Abdelhakem	
<b>Mathematical modeling of Echinococcosis in humans, dogs and livestock with optimal control strategies . . . . .</b>	<b>1075</b>
I. Sannaky, M. Riouali, N. Ouldkhouia, I. El berrai, and K. Adnaoui	
<b>A new generalized model of cooperation of advertising companies based on differential games on networks . . . . .</b>	<b>1116</b>
M. Jashnesade, Z. Nikoeeinejad and G B. Loghmani	
<b>Mathematical modeling and optimal control strategies to limit cochineal infestation on cacti plants . . . . .</b>	<b>1145</b>
K. Sofiane and B. Omar	
<b>An efficient Dai-Kou-type method with image de-blurring application . . . . .</b>	<b>1171</b>
K. Ahmed, M.Y. Waziri, S. Murtala, A.S. Halilu, H. Abdullahi and Y.B. Musa	

**Combining the reproducing kernel method with Taylor series expansion to solve systems of nonlinear fractional Volterra integro-differential equations . . . . . 1210**  
T. Amoozad, S. Abbasbandy, H. Sahihi, T. Allahviranloo

**Convex-hull based two-phase algorithm to solve capacitated vehicle routing problem . . . . . 1241**  
M. Afsharirad and A. Hashemi Borzabadi

**Accurate ENO-like schemes for the model of fluid flows in a nozzle with variable cross-section . . . . . 1275**  
D.H. Cuong and M.D. Thanh

**web site: <https://ijnao.um.ac.ir>**  
**Email: [ijnao@um.ac.ir](mailto:ijnao@um.ac.ir)**  
**ISSN-Print: 2423-6977**  
**ISSN-Online: 2423-6969**